Essays in Econometrics with focus on smooth minimum distance inference

Inauguraldissertation

zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften

durch die

Rechts- und Staatswissenschaftliche Fakultät der
Rheinischen Friedrich-Wilhelms-Universität
Bonn

vorgelegt von

Daniel Becker

aus Brachtendorf

Bonn 2021

Dekan:	Prof. Dr. Jürgen von Hagen
Erstreferent:	Prof. Dr. Alois Kneip
Zweitreferent:	Prof. Dr. Valentin Patilea

Tag der mündlichen Prüfung: 23. Juli 2021

Acknowledgments

This thesis would not exist without the support I received from many people. First of all, I would like to thank my supervisors Alois Kneip and Valentin Patilea. I'm very happy that I could have them as my supervisors and coauthors. Their advice and guidance were invaluable for my research and helped me to structure my thoughts. I also want to thank Jörg Breitung for his support and feedback. He helped me a lot during the first time of my PhD, especially in starting my first research project.

For advice and endless hours of discussion, I thank all my office mates in Bonn, Alexander Gleim, Heiko Wagner, Dominik Poß, Fabian Walders and Sven Otto. Special thanks go to Sebastian Schaube and Benjamin Hartung for their friendship and for all the time we spent over the last years.

During my frequent visits at ENSAI in the last years I always felt welcomed and enjoyed the fruitful research environment. Especially, I would like to thank Nicolas Lepage-Saucier and Lionel Truquet for helpful discussions and comments on my research.

I want to thank the Bonn Graduate School of Economics and the faculty at the Institute of Finance and Statistics for the excellent research environment. I would also like to acknowledge the financial support which I received from the German Research Foundation (Deutsche Forschungsgesellschaft). I furthermore thank the Regional Computing Center of the University of Cologne (RRZK) for providing computing time on the DFG-funded High Performance Computing (HPC) system CHEOPS as well as support.

I thank my family for all the support throughout the years. Finally, I owe my deepest gratitude to Miriam. Her support and love helped me to get back on track whenever I felt lost or was in doubt.

Contents

List of Figures

List of Tables

Introduction

This thesis consists of three self-contained essays in econometrics and statistics. It discusses methodological topics in semiparametric statistics as well as dynamic panel data models. In the first part, I focus on relaxing assumptions with respect to the model structure in the considered semiparametric model. The main goal is to provide an estimation technique that enables an applied researcher to answer research questions having at hand a reasonable amount of observed data points. Therefore, I provide reliable testings procedures that lead to trustworthy inference results. In the second part, I consider a well known dynamic panel data model. In order to estimate the model parameters I refine an existing likelihood approach and contrast it with a second likelihood approach. The idea of this project is to give a comprehensive overview of the estimation possibilities with likelihood approaches for this panel data model.

In chapters 1 and 2 the smooth minimum distance (SmoothMD) approach proposed by Lavergne and Patilea [56] is considered in the context of a partially linear model. The motivation for the SmoothMD estimator is that models nonlinear in parameters that are based on conditional moment restrictions can render inconsistent parameter estimates when the generalized method of moments (GMM) is used for estimation. The reason is that the conditional moment restrictions, that identify the model, imply an infinite number of unconditional moment restrictions if the conditioning variables have a support with infinite cardinality. GMM relies only on a finite number of instruments and, thus, might lead to inconsistent estimates. See Dominguez and Lobato [29]. Therefore, there have recently been proposed several approaches that account for conditional equations at the outset to obtain more efficient estimators. All these approaches share a common feature. The sensitivity to user-chosen parameters, that remains largely unknown. This is one key motivation for the alternative estimator of Lavergne and Patilea [56].

CHAPTER 1 is joint work with Alois Kneip und Valentin Patilea. We consider a semiparametric partially linear model in the spirit of Robinson [70] with Box-Cox transformed dependent variable. Transformation regression models are widely used in applied econometrics to avoid misspecification and a partially linear semiparametric model is an intermediate strategy that tries to balance advantages and disadvantages of a fully parametric model and nonparametric models. We combine both estimation strategies to allow for a more flexible model structure. Our study seems to be the first consideration of this model. The model parameters are estimated by the SmoothMD approach. The main difference to the estimator of Lavergne and Patilea [56] is that due to the structure of the partially linear model we need to estimate the regressand and the regressors of the model. We show that the SmoothMD estimator can handle the estimation bias stemming from these estimates and consider the asymptotic behavior under general conditions. In addition, new inference methods are proposed. A simulation experiment illustrates the performance of the methods for finite samples. Finally, we show the usefulness of the proposed estimator by applying it to investigate the returns of social and cognitive skills in a labor market context.

CHAPTER 2 considers again the semiparametric partially linear model in the spirit of Robinson [70] but without transformation of the dependent variable as in chapter 1. In addition, we allow for endogenous covariates. Apart from the estimator of Robinson [70] there exist other approaches, consider for instance Li [57] and Li and Stengos [59], that provide consistent and asymptotically normally distributed estimates. Here, it is proposed to employ again the SmoothMD estimator. This seems to be counterintuitive as the considered model is linear in parameters. However, we will show that the SmoothMD estimator captures a part of the estimation bias of the estimated regressand and regressors. This is in contrast to the existing approaches and a simulation study suggests that the SmoothMD estimator improves results especially for inference in finite samples.

CHAPTER 3 is based on work with Jörg Breitung. We consider dynamic panel data models with individual fixed effects and compare the transformed maximum likelihood approach of Hsiao et al. [48] and the factor analytical approach proposed by Bai [17]. This is interesting as the first approach considers the model in differences whereas the latter approach focuses on the model in levels. In addition, we extend the factor analytical approach to models with additional exogenous covariates.

Chapter 1

Smooth minimum distance inference for semiparametric partially linear regressions with Box-Cox transformation

1.1. Introduction

The data consists of independent copies of a response variable Y and a random covariate vector $\left(\boldsymbol{X}^T, \boldsymbol{Z}^T\right)^T \in \mathbb{R}^p \times \mathbb{R}^q$.[1] To model the relationship between the response and the covariate vector, we consider a *transformation partially linear* mean regression model given by

$$T(Y,\lambda) = \boldsymbol{X}^T\boldsymbol{\beta} + m(\boldsymbol{Z}) + \varepsilon, \tag{1.1}$$

where $T(\cdot,\lambda)$ is a known function depending on an unknown finite-dimensional parameter λ, $m(\cdot)$ is an unknown function and

$$E[\varepsilon \mid \boldsymbol{X}, \boldsymbol{Z}] = 0. \tag{1.2}$$

We impose no further assumption on the conditional distribution of ε. In particular, we allow for heteroscedasticity of unknown form. The vector $\boldsymbol{Z}$ contains only continuous variables, but the components of $\boldsymbol{X}$ need not be continuous. Let $\boldsymbol{\beta}_0$ and λ_0 denote the true values of the parameters. Our transformation partially linear model extends the standard partially linear model which corresponds to the case where the true value λ_0 of λ is known. See Robinson [70]. It seems to be the first extension of this kind under the general condition (1.2).

The transformation function $T(Y,\lambda)$ is usually assumed to be strictly increasing in Y. In the literature, many different parametric transformation functions have been proposed. In this study we consider $T(\cdot,\lambda)$ to be the Box-Cox transformation, though our analysis extends to other parametric transformations with similar properties. The transformation proposed by Box and Cox [22] is defined as

$$T(Y,\lambda) = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & ,\lambda \neq 0 \\ \log(Y) & ,\lambda = 0. \end{cases}$$

To use this transformation, one should have positive responses Y. Hence, in the following it will implicitly be assumed that Y is positive. The Box-Cox transformation is widely used in applications. It has become standard in the statistical and econometric literature and is discussed in various textbooks, e.g. Amemiya [6], Greene [33], Horowitz [46], Showalter [75], Wooldridge [81]. Furthermore, there exist several empirical studies that employ the Box-Cox transformation. See, for instance, Berndt et al. [19], Heckman and Polachek [40] or Keane et al. [50]. For an overview of the Box-Cox transformation consider Horowitz [46] and Sakia [71]. The reason for applying the Box-Cox transformation is to increase the flexibility and avoid misspecification of the model. In economic applications the dependent variable is frequently log-transformed, see for example Acemoglu et al. [1] and Autor and Handel [16]. However, in general one does not have guarantees that this transformation leads to the correct model. If the regression specification is true for some different transformation, employing the log-transformation for fitting the regression model might give misleading estimation and inference results. Specifying the transformation up to a

[1] Herein, vectors are column matrices and for any matrix $\boldsymbol{A}$, $\boldsymbol{A}^T$ denotes its transpose.

parameter and estimating the parameter together with the regression parameters leads to more reliable results at the cost of having to estimate only one additional parameter. In addition, the common log-transformation is nested in the Box-Cox transformation and, thus, can be confirmed by a statistical test. Of course, one could aim to extending the framework and considering that the transformation belongs to a nonparametric family of transformations. See, for instance, Horowitz [45] and Zhou et al. [82] for such high-level assumptions on the transformation. However, such more general approaches pay the price of more stringent independence assumptions on the error term ε. Moreover, checking whether a given transformation, such as the log-transformation, is validated by the data becomes much more challenging. For all these reasons the extension of our framework to the case of nonparametric transformations will remain beyond the scope of this study. Despite its popularity, even in a purely parametric framework, estimation and inference in a Box-Cox transformation model is a difficult statistical problem and is usually based on quite restrictive assumptions on the conditional law of the response. See for instance chapter 5 of Horowitz [46] for an illuminating discussion. The problem becomes even more complex in case of the semiparametric regression (1.1) where one only assumes the minimal identification condition (1.2).

The semiparametric partially linear specification of the conditional mean of the response is quite appealing as it allows a linear dependency on a subvector $\boldsymbol{X}$ of covariates, which could include discrete variables, and meanwhile allows a nonparametric additive effect of the covariates $\boldsymbol{Z}$. These features could help practitioners faced with a large cross-sectional data set with independent observations including many candidate explanatory variables, who, on the basis of economic theory or past experience with similar data, feel able to parameterize only some of them.

There are many studies in the literature, where $\boldsymbol{Z}$ in (1.1) is either assumed to be a scalar nonstochastic design variable, or $\boldsymbol{Z}$ is a stochastic vector and of arbitrary fixed dimension. The dimension of $\boldsymbol{Z}$ might influence the estimation accuracy as we will see in the following. It is well-known that, given model (1.1) with the true transformation, an ordinary least squares (OLS) regression of $T(Y,\lambda_0)$ on $\boldsymbol{X}$ alone consistently estimates $\boldsymbol{\beta}_0$, provided that $\boldsymbol{X}$ and $m(\boldsymbol{Z})$ are orthogonal. However, this orthogonality condition cannot be expected to hold true in most situations, and thus in general this OLS estimator is biased, as usually happens with OLS in the presence of nonorthogonal omitted variables. See e.g. White [80]. Consistent estimation of $\boldsymbol{\beta}_0$ in the presence of an unknown function $m(\cdot)$ is possible, however. For instance, one may consider a nonparametric estimator of $e(\boldsymbol{x},\boldsymbol{z}) = E[T(Y,\lambda_0) \mid \boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}=\boldsymbol{z}]$. When $\boldsymbol{X}$ and $\boldsymbol{Z}$ do not overlap, the derivative of this estimator with respect to $\boldsymbol{x}$, denoted by $\boldsymbol{e}_x$, yields a consistent estimate of $\boldsymbol{\beta}_0$ under quite general conditions; see, e.g., Robinson [70]. Unfortunately, the estimators $\hat{e}(\boldsymbol{x},\boldsymbol{z})$ and $\hat{\boldsymbol{e}}_x$ are not $\sqrt{n}$-consistent, $\hat{\boldsymbol{e}}_x$ converging even slower than $\hat{e}(\boldsymbol{x},\boldsymbol{z})$. Moreover, the greater the dimensions of $\boldsymbol{X}$ and $\boldsymbol{Z}$, the further both estimators fall short of $\sqrt{n}$-consistency.

Robinson [70] proposed an alternative and more effective approach. Given the true value λ_0, under the condition (1.2), one gets $E[Y^* \mid \boldsymbol{Z}] = E[\boldsymbol{X} \mid \boldsymbol{Z}]^T\boldsymbol{\beta} + m(\boldsymbol{Z})$, where $Y^* = T(Y,\lambda_0)$. Next, one can rewrite the model as

$$Y^* - E[Y^* \mid \boldsymbol{Z}] = (\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}])^T\boldsymbol{\beta} + \varepsilon, \qquad E[\varepsilon|\boldsymbol{X},\boldsymbol{Z}] = 0.$$

The estimator of $\boldsymbol{\beta}_0$ proposed by Robinson [70] is then a feasible version of the unfeasible OLS estimator of $Y^* - E[Y^* \mid \boldsymbol{Z}]$ on $\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}]$. The regressand and regressors $E[Y^* \mid \boldsymbol{Z}]$ and $E[\boldsymbol{X} \mid \boldsymbol{Z}]$ being unknown, they need to be estimated by some nonparametric procedure. Robinson [70] proposed to estimate them by the Nadaraya-Watson (NW) estimator. He showed that, under suitable regularity assumptions and conditions on the kernel and the bandwidth, the OLS estimator with response $Y^* - E[Y^* \mid \boldsymbol{Z}]$ and covariate vector $\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}]$ yields a $\sqrt{n}$-consistent, asymptotically normally distributed and efficient estimator if the conditional expectations given $\boldsymbol{Z}$ are replaced by their kernel estimates. The quite straightforward way to build efficient estimators made the partially linear model quite a popular. Versions of this model have also been studied by Engle et al. [31], Heckman [42], Shiller [73] and Wahba [79]. For an overview consider Härdle et al. [37] and Li and Racine [58]. In order to avoid the trimming introduced by Robinson [70] to ensure that the estimate of the density of $\boldsymbol{Z}$, $f_z(\boldsymbol{Z})$, stays away from zero, Li [57] considered as starting point the unfeasible OLS regression of $(Y^* - E[Y^* \mid \boldsymbol{Z}])f_z(\boldsymbol{Z})$ on $(\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}])f_z(\boldsymbol{Z})$. Premultiplying by the density of $\boldsymbol{Z}$ does not break the consistency of the unfeasible OLS estimator since $E[f_z(\boldsymbol{Z})\varepsilon \mid \boldsymbol{X},\boldsymbol{Z}] = f_z(\boldsymbol{Z})E[\varepsilon \mid \boldsymbol{X},\boldsymbol{Z}] = 0$. Next, Li [57] proposed to build OLS estimates using standard kernel estimators instead of the unfeasible response and covariates. This new estimator is still $\sqrt{n}$-consistent and asymptotically normally distributed. Moreover, Li [57] relaxed the condition on the bandwidth with the consequence that the smoothing requires higher order kernels only if the dimension of $\boldsymbol{Z}$ is larger than 5, instead of larger than 3 as required in Robinson [70].

The least squares approach fails when the response variable is a transformation depending on an unknown parameter. This is well-known in parametric Box-Cox transformation regression models, and is inherited by our semiparametric extension introduced in equation (1.1). However, the model we consider herein belongs to the large class of models defined by conditional moment restrictions. Therefore, we propose to estimate the finite dimensional parameters λ and $\boldsymbol{\beta}$ of model (1.1) combining the estimation strategy of Li [57] with the smooth minimum distance (SmoothMD) estimator proposed by Lavergne and Patilea [56].

The motivation for the SmoothMD estimator is that models nonlinear in parameters like (1.1) that are based on conditional moment restrictions as condition (1.2) can render inconsistent parameter estimates when the generalized method of moments (GMM) is used for estimation. The reason is that the conditional moment restrictions, that identify the model, imply an infinite number of unconditional moment restrictions if the conditioning variables have a support with infinite cardinality. However, GMM relies only on a finite number of instruments and, thus, in general consistency of GMM relies on additional assumptions. See Dominguez and Lobato [29]. This problem has been pointed out for the Box-Cox transformation by Foster et al. [32] and Shin [74] in the linear case. See also Horowitz [46]. More recent work focuses on accounting for conditional equations at the outset to obtain more efficient estimators. Some methods rely on increasing the number of considered unconditional estimating equations (or instruments) with the sample size, such as the sieve minimum distance (SMD) approach of Ai and Chen [3], or generalizations of GMM and empirical likelihood (EL) by Donald et al. [30] and Hjort et al. [43]. Carrasco and Florens [24] use a regularization approach to generalize the GMM approach to a continuum of estimating equations. Other EL-type estimators use nonparametric smoothing to estimate conditional equations, such as Antoine et al. [13], Kitamura et al. [52], and Smith [76, 77]. All these approaches share one common feature. The estimators' sensitivity to the user-chosen parameter (number of estimating equations, regularization parameter, or smoothing parameter) remains largely unknown. This is one key motivation for the alternative estimator of Lavergne and Patilea [56], the SmoothMD estimator. The asymptotic representation of their estimator is established as a process indexed by a tuning parameter, the user-chosen parameter, which can vary within a wide range including values independent of the sample size.

Let us briefly recall the SmoothMD approach. Consider a general conditional moment restrictions model

$$E[g(\boldsymbol{U};\boldsymbol{\theta}) \mid \boldsymbol{W}] = 0, \tag{1.3}$$

where $g(\cdot)$ is a given function, $\boldsymbol{U}$ is a vector of observed variables, $\boldsymbol{W}$ is a subvector of $\boldsymbol{U}$ and $\boldsymbol{\theta}$ is the finite-dimensional parameter of interest. The components of $\boldsymbol{W}$ need not be continuous random variables. It is assumed that there exists a $\boldsymbol{\theta}_0$ such that $E[g(\boldsymbol{U};\boldsymbol{\theta}_0) \mid \boldsymbol{W}] = 0$ and $\boldsymbol{\theta}_0$ is unique with this property. The SmoothMD approach is based on an equivalent rewriting of equation (1.3) under the form of a suitable unconditional moment. For this purpose, let $\omega(\cdot)$ be a symmetric function of $\boldsymbol{W}$ with positive Fourier transform. The typical example we have in mind is $\omega(\boldsymbol{W}) = \exp\left\{-\boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W}\right\}$ where $\boldsymbol{D}$ is some positive definite matrix. Typically, $\boldsymbol{D}$ is a diagonal matrix with diagonal elements playing the role of standardizing constants. Lavergne and Patilea [56] list many other possible choices for $\omega(\cdot)$. Then condition (1.3) is satisfied if and only if

$$Q(\boldsymbol{\theta}) = E[g(\boldsymbol{U}_1;\boldsymbol{\theta})g(\boldsymbol{U}_2;\boldsymbol{\theta})\omega(\boldsymbol{W}_1 - \boldsymbol{W}_2)] = 0,$$

where $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are independent copies of $\boldsymbol{U}$ (and $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the corresponding subvectors). Whenever $E[g(\boldsymbol{U};\boldsymbol{\theta}) \mid \boldsymbol{W}] \neq 0$, one has $Q(\boldsymbol{\theta}) > 0$. Finally, the SmoothMD estimator is defined as the minimum of a sample based approximation of $Q(\boldsymbol{\theta})$. The SmoothMD estimator is $\sqrt{n}$-consistent and asymptotically normal. Meanwhile, estimators based on instruments may be inconsistent if their number is kept fixed, as pointed out by Dominguez and Lobato [29]. Hence, the SmoothMD estimator bridges a gap between Dominguez and Lobato's method, which does not require a user-chosen parameter, and the competing SMD estimator and EL and GMM-type methods that rely on smoothing. Indeed, Lavergne and Patilea [56] obtained their asymptotic results uniformly with respect to the diagonal of the matrix $\boldsymbol{D}$ in an interval with the right endpoint allowed to grow to infinity. Then, a diagonal element of $\boldsymbol{D}$ could be viewed as the inverse of a kernel smoothing bandwidth tending to zero at a suitable rate.

In the context of our model defined in (1.1), we have to extend the SmoothMD approach to the case where the model contains an infinite-dimensional nuisance parameter. Let $\boldsymbol{W} = \left(\boldsymbol{X}^T, \boldsymbol{Z}^T\right)^T \in \mathbb{R}^p \times \mathbb{R}^q$ and $\boldsymbol{U} = \left(Y, \boldsymbol{W}^T\right)^T$. Moreover, let $\boldsymbol{\theta} = \left(\lambda, \boldsymbol{\beta}^T\right)^T$ and

$$g(\boldsymbol{U};\boldsymbol{\theta},\gamma,\boldsymbol{\eta}) = \left(T(Y,\lambda) - E[T(Y,\lambda) \mid \boldsymbol{Z}] - (\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}])^T\boldsymbol{\beta}\right) f_z(\boldsymbol{Z}) - \gamma. \tag{1.4}$$

Here, $\boldsymbol{\eta}$ is an infinite-dimensional nuisance parameter containing the three unknown functions of $\boldsymbol{Z}$ appearing in the definition of $g(\boldsymbol{U};\boldsymbol{\theta},\gamma,\boldsymbol{\eta})$ and $\gamma \in \mathbb{R}$ is an intercept nuisance parameter. Then, our transformation partially linear mean regression model could be written under the form of a conditional moment equation $E[g(\boldsymbol{U};\boldsymbol{\theta},\gamma,\boldsymbol{\eta}) \mid \boldsymbol{W}] = 0$. The true value of the intercept γ is known to be equal to zero. However, this artificial parameter will be helpful to diminish the amplitude of the variance coming from the nonparametric estimators of the unknown functions in the asymptotic representation of the estimator of $\boldsymbol{\theta}_0 = \left(\lambda_0, \boldsymbol{\beta}_0^T\right)^T$. The generalized SmoothMD approach we propose is based on the equivalence

$$E[g(\boldsymbol{U};\boldsymbol{\theta},\gamma,\boldsymbol{\eta}) \mid \boldsymbol{W}] = 0 \iff Q(\boldsymbol{\theta},\gamma) = E[g(\boldsymbol{U}_1;\boldsymbol{\theta},\gamma,\boldsymbol{\eta}_1)g(\boldsymbol{U}_2;\boldsymbol{\theta},\gamma,\boldsymbol{\eta}_2)\omega(\boldsymbol{W}_1-\boldsymbol{W}_2)] = 0,$$

where $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are independent copies of $\boldsymbol{U}$ (and $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the corresponding subvectors). Whenever, $E[g(\boldsymbol{U};\boldsymbol{\theta},\gamma,\boldsymbol{\eta}) \mid \boldsymbol{W}] \neq 0$, one has $Q(\boldsymbol{\theta},\gamma) > 0$. Next, the idea is to define the SmoothMD estimator as the minimum of a sample based version of $Q(\boldsymbol{\theta},\gamma)$. To take advantage of the structure of our model, we propose to define our SmoothMD estimator using a profiling approach. Given the i.i.d. sample $\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n$ and nonparametric estimates $\widehat{\boldsymbol{\eta}}_1,\ldots,\widehat{\boldsymbol{\eta}}_n$ of the values of the nuisance parameter, for each λ, we define the map

$$(\gamma,\boldsymbol{\beta}^T)^T \mapsto \widehat{Q}_n\left(\left(\lambda,\boldsymbol{\beta}^T\right)^T,\gamma\right) = \frac{1}{n^2}\sum_{1\leq i,j\leq n} g(\boldsymbol{U}_i;\boldsymbol{\theta},\gamma,\widehat{\boldsymbol{\eta}}_i)g(\boldsymbol{U}_j;\boldsymbol{\theta},\gamma,\widehat{\boldsymbol{\eta}}_j)\omega(\boldsymbol{W}_i-\boldsymbol{W}_j),$$

which is quadratic with an explicit unique minimum $(\widehat{\gamma}(\lambda),\widehat{\boldsymbol{\beta}}(\lambda)^T)^T$. Thus, we define a profile SmoothMD estimator of λ_0 as

$$\widehat{\lambda} = \arg\min_{\lambda} \widehat{Q}_n\left(\left(\lambda,\widehat{\boldsymbol{\beta}}(\lambda)^T\right)^T,\widehat{\gamma}(\lambda)\right),$$

and, with at hand the estimate $\widehat{\lambda}$, eventually we calculate $\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$, the semiparametric SmoothMD estimate of $\boldsymbol{\beta}_0$. Let us point out that our SmoothMD approach could also be used in classic particular cases. Indeed, the model of Robinson [70] is nested in the model (1.2) in the sense that it corresponds to the situation where the transformation parameter λ is known. In this case our $\widehat{\boldsymbol{\beta}}(\lambda)$ is an alternative to Robinson [70]'s estimator for which we require weaker technical conditions. The classical linear model of Box and Cox [22], as well as any parametric extension, could also be estimated by SmoothMD, without any further assumptions on the law of ε. In that case the smoothing with respect to $\boldsymbol{Z}$ is unnecessary as the model does not contain any unknown function.

In addition, we can estimate the transformation partially linear model if we have endogeneity in $\boldsymbol{X}$, i.e. $E[\varepsilon \mid \boldsymbol{X}] \neq 0$ but $E[\varepsilon \mid \boldsymbol{Z}] = 0$. In order to be able to estimate the model we need to find a vector $\widetilde{\boldsymbol{W}}$ such that $E[g(\boldsymbol{U};\boldsymbol{\theta},\gamma,\boldsymbol{\eta}) \mid \widetilde{\boldsymbol{W}}] = 0$. Here, one needs a vector of instruments $\boldsymbol{V}$ that is correlated with $\boldsymbol{X}$ but $E[\varepsilon \mid \boldsymbol{V}] = 0$ such that $\widetilde{\boldsymbol{W}} = \left(\boldsymbol{V}^T,\boldsymbol{Z}^T\right)^T$. See Li and Stengos [59] and chapter 2 for a discussion in the standard model of Robinson [70] with known λ_0. If we have endogeneity in $\boldsymbol{Z}$ the estimation problem becomes more delicate and cannot be conducted in the way we propose in this chapter. This problem is left for future research.

The remainder of the chapter is organized as follows. In section 1.2, we present our new estimation method and establish identification of the model parameters. In section 1.3, we develop our uniform-in-bandwidth theory, including consistency and $\sqrt{n}-$consistency of our estimator. In section 1.4, we investigate a distance-metric procedure for testing restrictions on parameters. In section 1.5, we study the finite sample behavior by a simulation study and apply the estimator to a real data sample. Our estimator performs well in our experiments and our tests yield accurate levels and good power in moderate samples. Section 1.6 concludes. Technical assumptions are stated in section 1.7.

1.2. The semiparametric SmoothMD approach

In this section we formally define our semiparametric estimator. First, we investigate two issues. On the one hand, we prove identification of the true value $\boldsymbol{\theta}_0 = (\lambda_0,\boldsymbol{\beta}_0^T)^T$ of the parameter of interest. Next, we discuss the recommendation appearing in the literature for normalizing the response variable. This issue is specific to the Box-Cox transformation, though similar problems occur with other families of transformations. Finally, we define our semiparametric SmoothMD estimator.

We use the following notation throughout the remaining of the chapter. For $d_l, d_c \geq 1$, let $\mathbb{R}^{d_l\times d_c}$ denote the set of $d_l \times d_c$ $\boldsymbol{A}$ matrices with real elements. Let $\mathbf{1}_{d_l}$ (resp. $\mathbf{0}_{d_l}$) denote the vector with all components equal to 1 (resp. 0), $\mathbf{0}_{d_l\times d_c}$ the $d_l \times d_c-$null matrix and $\boldsymbol{I}_{d_l\times d_l}$ the identity matrix with dimension $d_l \times d_l$. For a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|$ is the Frobenius norm and $\|\boldsymbol{A}\|_{\text{Sp}}$ the spectral norm. Below, $\boldsymbol{D} = \text{diag}(\boldsymbol{d})$ is some positive definite diagonal matrix with $\boldsymbol{d} \in \mathcal{D} \subset \mathbb{R}_+^{p+q}$ being a diagonal vector with strictly positive components. Herein, $\mathcal{D}$ is a compact set and our asymptotic results are derived uniformly with respect to $\boldsymbol{d} \in \mathcal{D}$.

1.2.1. Identification

Let $-\infty < \lambda_{\min} < \lambda_0 < \lambda_{\max} < \infty$, with $\lambda_{\min} < 0$ and $\lambda_{\max} > 0$. For any $\lambda \in \Lambda = [\lambda_{\min}, \lambda_{\max}]$, let

$$(\gamma(\lambda), \boldsymbol{\beta}(\lambda)^T)^T = \arg \min_{\gamma \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} E[g(\boldsymbol{U}_1; \boldsymbol{\theta}, \gamma, \boldsymbol{\eta}_1) g(\boldsymbol{U}_2; \boldsymbol{\theta}, \gamma, \boldsymbol{\eta}_2) \omega(\boldsymbol{W}_1 - \boldsymbol{W}_2)], \tag{1.5}$$

with $g(\boldsymbol{U}_1; \boldsymbol{\theta}, \gamma, \boldsymbol{\eta}_1)$ and $g(\boldsymbol{U}_2; \boldsymbol{\theta}, \gamma, \boldsymbol{\eta}_2)$ being independent copies of $g(\boldsymbol{U}; \boldsymbol{\theta}, \gamma, \boldsymbol{\eta})$ defined in equation (1.4) with $\boldsymbol{U} = \left(Y, \boldsymbol{W}^T\right)^T$, $\boldsymbol{W} = \left(\boldsymbol{X}^T, \boldsymbol{Z}^T\right)^T$, $\boldsymbol{\theta} = (\lambda, \boldsymbol{\beta}^T)^T$ and

$$\boldsymbol{\eta} = (f_z(\cdot), E[T(Y, \lambda) \mid \boldsymbol{Z} = \cdot\,], E[\boldsymbol{X} \mid \boldsymbol{Z} = \cdot\,]^T)^T.$$

With all this in hand we can now state the following identification result.

Lemma 1.1. *Suppose that Assumptions 1.1 and 1.2 hold true and* $\max(|\lambda_{min}|, \lambda_{max}) < \infty$*. Let* $\gamma(\lambda)$ *and* $\boldsymbol{\beta}(\lambda)$ *be defined as in equation* (1.5)*. Then,* $\gamma(\lambda_0) = 0$ *and* $\boldsymbol{\beta}(\lambda_0) = \boldsymbol{\beta}_0$ *and*

$$\mathbb{P}\left(E\left[(T(Y, \lambda) - E[T(Y, \lambda) \mid \boldsymbol{Z}]) f_z(\boldsymbol{Z}) - \gamma - (\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}])^T \boldsymbol{\beta} f_z(\boldsymbol{Z}) \mid \boldsymbol{X}, \boldsymbol{Z}\right] = 0\right) < 1,$$

for all $\gamma \in \mathbb{R}$ *and* $\boldsymbol{\theta} = (\lambda, \boldsymbol{\beta}^T)^T \in \Lambda \times \mathbb{R}^p$ *such that* $(\gamma, \boldsymbol{\theta}^T)^T \neq (0, \boldsymbol{\theta}_0^T)^T$*. Moreover, for any* $\varepsilon > 0$*,*

$$\inf_{\lambda \in \Lambda,\, |\lambda - \lambda_0| \geq \varepsilon} \inf_{\boldsymbol{d} \in \mathcal{D}} E\left[g\left(\boldsymbol{U}_1; (\lambda, \boldsymbol{\beta}(\lambda)^T)^T, \gamma(\lambda), \boldsymbol{\eta}_1\right) g\left(\boldsymbol{U}_2; (\lambda, \boldsymbol{\beta}(\lambda)^T)^T, \gamma(\lambda), \boldsymbol{\eta}_2\right) \right.$$
$$\left. \times \exp\left\{-(\boldsymbol{W}_1 - \boldsymbol{W}_2)^T \boldsymbol{D} (\boldsymbol{W}_1 - \boldsymbol{W}_2)\right\}\right] > 0. \tag{1.6}$$

1.2.2. Box-Cox transformation and standardized responses

Let us note that

$$\lim_{\lambda \uparrow \infty} \frac{y^\lambda - 1}{\lambda} = 0 \quad \text{if } 0 < y < 1 \qquad \textit{and} \qquad \lim_{\lambda \downarrow -\infty} \frac{y^\lambda - 1}{\lambda} = 0 \quad \text{if } y > 1.$$

In classical estimation approaches for parametric regression models with Box-Cox transformed response, this is likely to induce instability for the estimation of the parameter λ. See, e.g., Khazzoom [51], Powell [69] and Showalter [75] for a discussion of this well-known issue. In order to avoid such problems, the common recommendation is to standardize the response by some constant, say s, such that

$$\mathbb{P}\left(Y/s < 1\right) > 0 \qquad \text{and} \qquad \mathbb{P}\left(Y/s > 1\right) > 0.$$

The constant s could be for instance the mean of Y or the geometric mean of Y.[2] With finite samples, the practitioner would first estimate such a constant using the sample and next would normalize the responses. The same type of problems might occur in our semiparametric extension of the Box-Cox transformation model. For this reason, we will replace our function $g(\boldsymbol{U}; \boldsymbol{\theta}, \gamma, \boldsymbol{\eta})$ by a family of functions $s^{-\lambda} g(\boldsymbol{U}; \boldsymbol{\theta}, \gamma, \boldsymbol{\eta})$ indexed also by s in some interval on the positive half-line that we will let depend on the sample size. This change of the family of functions is equivalent to changing Y to Y/s in the definition (1.4), and a rescaling of the parameters $\boldsymbol{\beta}$ and γ.

By the profiling-based construction of our SmoothMD estimator, the replacement of the response Y by Y/s matters only for computing $\widehat{\lambda}$. Clearly, the identifiability property established in Lemma 1.1 is preserved. In the following we provide asymptotic results that are uniform with respect to s in order to allow for a data-driven choice of s, such as for instance the sample geometric mean of the response.

1.2.3. The estimator

Given an independent sample $\left(Y_1, \boldsymbol{X}_1^T, \boldsymbol{Z}_1^T\right)^T, \ldots, \left(Y_n, \boldsymbol{X}_n^T, \boldsymbol{Z}_n^T\right)^T$ from $\left(Y, \boldsymbol{X}^T, \boldsymbol{Z}^T\right)^T \in \mathbb{R} \times \mathbb{R}^{p+q}$, let us define

$$\widehat{\mathbb{Y}}_n(\lambda) = \left((T(Y_1, \lambda) - \widehat{E}[T(Y_1, \lambda) \mid \boldsymbol{Z}_1]) \widehat{f}_z(\boldsymbol{Z}_1), \ldots, (T(Y_n, \lambda) - \widehat{E}[T(Y_n, \lambda) \mid \boldsymbol{Z}_n]) \widehat{f}_z(\boldsymbol{Z}_n)\right)^T \in \mathbb{R}^n,$$

[2] The geometric mean is defined as $G_n = \prod_{i=1}^n Y_i^{1/n} = \exp\left\{\frac{1}{n}\sum_{i=1}^n \log(Y_i)\right\}$. Let $y_{\max}$ be the largest and $y_{\min}$ the smallest observed value. We get that $\frac{y_{\max}}{G_n} = \prod_{i=1}^n \left(\frac{y_{\max}}{y_i}\right)^{1/n} > 1$ and $\frac{y_{\min}}{G_n} = \prod_{i=1}^n \left(\frac{y_{\min}}{y_i}\right)^{1/n} < 1$ as long as $y_{\max} > y_{\min}$. The reasoning for $\overline{y}$ is similar.

and

$$\widehat{\mathbb{X}}_n = \left((\boldsymbol{X}_1 - \widehat{E}[\boldsymbol{X}_1 \mid \boldsymbol{Z}_1])\widehat{f}_z(\boldsymbol{Z}_1), \ldots, (\boldsymbol{X}_n - \widehat{E}[\boldsymbol{X}_n \mid \boldsymbol{Z}_n])\widehat{f}_z(\boldsymbol{Z}_n)\right)^T \in \mathbb{R}^{n\times p}.$$

For $1 \leq i \leq n$, $\widehat{\boldsymbol{\eta}}_i = (\widehat{f}_z(\boldsymbol{Z}_i), \widehat{E}[T(Y_i,\lambda) \mid \boldsymbol{Z}_i], \widehat{E}[\boldsymbol{X}_i \mid \boldsymbol{Z}_i]^T)^T$ are nonparametric estimates of $\boldsymbol{\eta}_i = (f_z(\boldsymbol{Z}_i), E[T(Y_i,\lambda) \mid \boldsymbol{Z}_i], E[\boldsymbol{X}_i \mid \boldsymbol{Z}_i]^T)^T$. For the unknown values we use the kernel estimates

$$\widehat{f}_z(\boldsymbol{Z}_i) = \frac{1}{nh^q}\sum_{j=1}^n K\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_j}{h}\right), \quad \widehat{E}[T(Y_i,\lambda) \mid \boldsymbol{Z}_i]\widehat{f}_z(\boldsymbol{Z}_i) = \frac{1}{nh^q}\sum_{j=1}^n T(Y_j,\lambda) K\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_j}{h}\right),$$

and

$$\widehat{E}[\boldsymbol{X}_i \mid \boldsymbol{Z}_i]\widehat{f}_z(\boldsymbol{Z}_i) = \frac{1}{nh^q}\sum_{j=1}^n \boldsymbol{X}_j K\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_j}{h}\right).$$

Here $K(\cdot)$ is a multivariate kernel function and h is the bandwidth. Let $\boldsymbol{\Omega}_n$ be the $n\times n-$ symmetric matrix with elements

$$\boldsymbol{\Omega}_{n,ij} = \exp\{-(\boldsymbol{X}_i^T - \boldsymbol{X}_j^T, \boldsymbol{Z}_i^T - \boldsymbol{Z}_j^T)\boldsymbol{D}(\boldsymbol{X}_i - \boldsymbol{X}_j, \boldsymbol{Z}_i - \boldsymbol{Z}_j)\}, \qquad 1 \leq i,j \leq n.$$

Typically, the components of the vector $\boldsymbol{d}$ defining the diagonal matrix $\boldsymbol{D}$ are proportional to the standard deviation of the components of the vectors $(\boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$. The definition of $\boldsymbol{\Omega}_{n,ij}$ allows also to take into account discrete components of $\boldsymbol{X}$. For finite support discrete covariates, one could set some large value for the corresponding diagonal element of $\boldsymbol{D}$, which in practice would be equivalent to an indicator of the event that the observations i and j have the same value for that covariate.

We can now define, for any λ, the estimates of $(\gamma(\lambda), \boldsymbol{\beta}(\lambda)^T)^T \in \mathbb{R}^{1+p}$ introduced in equation (1.5). For any $s > 0$, let

$$\widehat{Q}_n\left(\left(\lambda, \boldsymbol{\beta}^T\right)^T, \gamma; s\right) = n^{-2}s^{-2\lambda}\left(\widehat{\mathbb{Y}}_n(\lambda) - \gamma\boldsymbol{1}_n - \widehat{\mathbb{X}}_n\boldsymbol{\beta}\right)^T \boldsymbol{\Omega}_n \left(\widehat{\mathbb{Y}}_n(\lambda) - \gamma\boldsymbol{1}_n - \widehat{\mathbb{X}}_n\boldsymbol{\beta}\right).$$

For fixed s and λ, consider the generalized least-squares problem

$$\min_{\gamma,\boldsymbol{\beta}} \widehat{Q}_n\left(\left(\lambda, \boldsymbol{\beta}^T\right)^T, \gamma; s\right). \tag{1.7}$$

The solution of this problem does not depend on $s^{-\lambda}$ and has the form of standard generalized least-squares estimators:

$$\widehat{\gamma}(\lambda, \boldsymbol{\beta}(\lambda)) = \frac{1}{\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n}\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\left(\widehat{\mathbb{Y}}_n(\lambda) - \widehat{\mathbb{X}}_n\boldsymbol{\beta}(\lambda)\right),$$

and

$$\widehat{\boldsymbol{\beta}}(\lambda) = \left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{Y}}_n(\lambda),$$

with

$$\mathbb{D}_n = \boldsymbol{\Omega}_n - \frac{1}{\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n}\boldsymbol{\Omega}_n\boldsymbol{1}_n\boldsymbol{1}_n^T\boldsymbol{\Omega}_n \in \mathbb{R}^{n\times n}. \tag{1.8}$$

Next, plugging $(\widehat{\gamma}(\lambda, \widehat{\boldsymbol{\beta}}(\lambda)), \widehat{\boldsymbol{\beta}}(\lambda)^T)^T$ into the problem (1.7), for given s, we define the SmoothMD estimator of λ_0 as

$$\widehat{\lambda} = \widehat{\lambda}(s) = \arg\min_{\lambda\in\Lambda} s^{-\lambda}\widehat{\mathbb{Y}}_n(\lambda)^T\ \widehat{\mathbb{B}}_n\ s^{-\lambda}\widehat{\mathbb{Y}}_n(\lambda), \tag{1.9}$$

with

$$\widehat{\mathbb{B}}_n = \mathbb{D}_n - \mathbb{D}_n\widehat{\mathbb{X}}_n\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n \in \mathbb{R}^{n\times n}.$$

Note that, by construction,

$$\mathbb{D}_n\boldsymbol{1}_n = \widehat{\mathbb{B}}_n\boldsymbol{1}_n = \boldsymbol{0}_n \quad \text{and} \quad \widehat{\mathbb{B}}_n\widehat{\mathbb{X}}_n = \boldsymbol{0}_{n\times p}.$$

Finally, the SmoothMD estimator of $\boldsymbol{\beta}_0$ is $\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$. We close this section showing that our estimator is well-defined.

Lemma 1.2. *If Assumptions 1.1.3 and 1.2 hold true, then, for each $n \geq 1$,*

1. *the matrices $\boldsymbol{\Omega}_n$ and $\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n$ are positive definite with probability 1. In particular, $\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n > 0$ and $\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n$ is invertible with probability 1.*
2. *the matrix $\widehat{\mathbb{B}}_n$ is positive semi-definite with probability 1.*

Remark 1. The matrix $\mathbb{D}_n$ is defined in equation (1.8) and has dimension $n \times n$. Therefore, it becomes difficult to work with this matrix when the sample size is large. However, it is not necessary to estimate the matrix $\mathbb{D}_n$ itself but it suffices to estimate $\widehat{\mathbb{X}}_n^T \mathbb{D}_n$ and $\widehat{\mathbb{Y}}_n(\lambda)^T \mathbb{D}_n$ to be able to calculate $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\lambda}$. In section 1.5 we show that the estimator can be applied for $n > 100,000$.

1.3. Consistency and asymptotic normality

The estimator $\widehat{\lambda}(s)$ depends on the fixed value s, a value that could, in practice, be data driven and calculated from the sample. For this reason, our asymptotic results are stated uniformly with respect to s. In the simulations we will use the geometric mean, $\exp\left\{\frac{1}{n}\sum_{i=1}^{n}\log(Y_i)\right\}$, for s. Our asymptotic results are also stated uniformly with respect to the diagonal of the matrix $\boldsymbol{D}$. This ensures that we can use a data driven estimate of $\boldsymbol{D}$ proportional to the empirical standard deviation of $\boldsymbol{X}$ or $\boldsymbol{Z}$.

Let's introduce some more notation: for each $\lambda \in \Lambda$, let

$$\mathbb{Y}_n(\lambda) = ((T(Y_1,\lambda) - E[T(Y_1,\lambda) \mid \boldsymbol{Z}_1])f_z(\boldsymbol{Z}_1), \ldots, (T(Y_n,\lambda) - E[T(Y_n,\lambda) \mid \boldsymbol{Z}_n])f_z(\boldsymbol{Z}_n))^T \in \mathbb{R}^n,$$

and

$$\mathbb{X}_n = ((\boldsymbol{X}_1 - E[\boldsymbol{X}_1 \mid \boldsymbol{Z}_1])f_z(\boldsymbol{Z}_1), \ldots, (\boldsymbol{X}_n - E[\boldsymbol{X}_n \mid \boldsymbol{Z}_n])f_z(\boldsymbol{Z}_n))^T \in \mathbb{R}^{n\times p}.$$

Moreover,

$$\mathbb{B}_n = \mathbb{D}_n - \mathbb{D}_n\mathbb{X}_n\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\mathbb{X}_n^T\mathbb{D}_n \in \mathbb{R}^{n\times n},$$

with $\mathbb{D}_n$ defined in equation (1.8). Again, by construction $\mathbb{B}_n\mathbf{1}_n = \mathbf{0}_n$ and $\mathbb{B}_n\mathbb{X}_n = \mathbf{0}_{n\times p}$. With all this in hand we can now state consistency of our estimator.

Theorem 1.1 (Consistency)**.** *Assume that Assumptions 1.1, 1.2 and 1.3 hold true. Let s_0 be some normalizing value such that $\mathbb{P}(Y/s_0 < 1) > 0$ and $\mathbb{P}(Y/s_0 > 1) > 0$ and let S_n be an arbitrary $o_{\mathbb{P}}(1)$ neighborhood of s_0. Then*

$$\sup_{h\in\mathcal{H}_{c,n}}\sup_{s\in S_n}\sup_{\boldsymbol{d}\in\mathcal{D}}\left|\widehat{\lambda}-\lambda_0\right| = o_{\mathbb{P}}(1) \quad \textit{and} \quad \sup_{h\in\mathcal{H}_{c,n}}\sup_{s\in S_n}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\widehat{\boldsymbol{\beta}}(\widehat{\lambda})-\boldsymbol{\beta}_0\right\| = o_{\mathbb{P}}(1).$$

In Theorem 1.1 we require that $h \in \mathcal{H}_{c,n}$, where $\mathcal{H}_{c,n} = [c_{min}n^{-\alpha}, c_{max}n^{-\alpha}]$, with $0 < \alpha < 1/q$ and c_{min}, c_{max} are positive constants. This implies that $nh^q \to \infty$ and $h \to 0$ for $n \to \infty$ which is in line with Robinson [70] and Li [57].

Next, we prove asymptotic normality of our estimator. For this purpose, we first derive the asymptotic linear representation of $\widehat{\lambda}$ and $\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$ from which the $\sqrt{n}-$asymptotic normality follows. In the following result, we show that $\widehat{\lambda}$ and $\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$ are asymptotically not equivalent to the infeasible estimators of λ_0 and $\boldsymbol{\beta}_0$ one would obtain when the infinite-dimensional parameter $\boldsymbol{\eta}$ is given and the intercept γ is equal to 0. This is in contrast to the results of Li [57] and Robinson [70]. The reason is that they can use the fact that $E[\mathbb{X}_{n,i} \mid \boldsymbol{Z}_i] = 0$ when controlling higher order terms. In our case, we weight the observations by $\boldsymbol{\Omega}_{n,ij}$ such that $E[\mathbb{X}_{n,i}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{Z}_i] \neq 0$ for $i \neq j$. This is also the reason why we need to ask for $q < 4$ instead of $q < 6$ as in Li [57]. Therefore, we require that $h \in \mathcal{H}_{sc,n}$, where $\mathcal{H}_{sc,n} = [c_{min}n^{-\alpha}, c_{max}n^{-\alpha}]$, with $\alpha \in (1/4, 1/q)$.

The results are again obtained uniformly with respect to the elements on the diagonal of the matrix $\boldsymbol{D}$ that determines $\boldsymbol{\Omega}_n$ and with respect to the scaling factor s that could be used for numerical stability, as mentioned in Section 1.2.2. In addition, let $K_h(\cdot) = h^{-q}K(\cdot/h)$ and, for any $1 \leq i, j \leq n$, let

$$K_{h,ij} = K_h(\boldsymbol{Z}_i - \boldsymbol{Z}_j).$$

Proposition 1.1 (Asymptotic representation)**.** *Assume that the conditions of Theorem 1.1 hold true. Moreover, Assumption 1.4 holds true. Then, uniformly with respect to $h \in \mathcal{H}_{sc,n}$, $\boldsymbol{d} \in \mathcal{D}$ and $s \in S_n$,*

$$\widehat{\lambda}-\lambda_0 = -\left[\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\,\mathbb{B}_n\,\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\,\mathbb{B}_n\left[(\boldsymbol{\varepsilon}\boldsymbol{f_z})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right] + o_{\mathbb{P}}(n^{-1/2}) = O_{\mathbb{P}}(n^{-1/2}),$$

and

$$\widehat{\boldsymbol{\beta}}(\widehat{\lambda})-\boldsymbol{\beta}_0 = \left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\mathbb{X}_n^T\mathbb{D}_n\left[(\boldsymbol{\varepsilon}\boldsymbol{f_z})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n + \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\left(\widehat{\lambda}-\lambda_0\right)\right] + o_{\mathbb{P}}(n^{-1/2}) = O_{\mathbb{P}}(n^{-1/2}),$$

where $(\boldsymbol{\varepsilon}\boldsymbol{f_z})_n = (\varepsilon_1 f_z(\boldsymbol{Z}_1), \ldots, \varepsilon_n f_z(\boldsymbol{Z}_n))^T$ and $\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n = \left(\frac{1}{n}\sum_{k=1,k\neq 1}^{n}\varepsilon_k K_{h,1k}, \ldots, \frac{1}{n}\sum_{k=1,k\neq n}^{n}\varepsilon_k K_{h,nk}\right)^T.$

Note that the asymptotic representation of $\widehat{\lambda}$ does not depend on s_0, i.e. the choice of s_0 does not influence the asymptotic behavior of $\widehat{\lambda}$. This result is in line with the result of Powell [69].

In the following we state asymptotic normality of our estimator. Therefore, we use the notation $\boldsymbol{\Omega}_{n,ij}(\boldsymbol{d}) = \boldsymbol{\Omega}_{n,ij}$ and

$$\mathbb{D}_n(\boldsymbol{d}) = \boldsymbol{\Omega}_n(\boldsymbol{d}) - \frac{1}{\boldsymbol{1}_n^T\boldsymbol{\Omega}_n(\boldsymbol{d})\boldsymbol{1}_n}\boldsymbol{\Omega}_n(\boldsymbol{d})\boldsymbol{1}_n\boldsymbol{1}_n^T\boldsymbol{\Omega}_n(\boldsymbol{d}),$$

to make the dependence of $\boldsymbol{\Omega}_n$ on $\boldsymbol{d}$ explicit. Note that with

$$\begin{aligned}\boldsymbol{\Omega}^X_{n,ij}(\boldsymbol{d}) = \boldsymbol{\Omega}^X_{n,ij} &= \exp\{-(\boldsymbol{X}_i - \boldsymbol{X}_j)^T \operatorname{diag}(d_1, \ldots, d_p)(\boldsymbol{X}_i - \boldsymbol{X}_j)\} \qquad \text{and}\\ \boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d}) = \boldsymbol{\Omega}^Z_{n,ij} &= \exp\{-(\boldsymbol{Z}_i - \boldsymbol{Z}_j)^T \operatorname{diag}(d_{p+1}, \ldots, d_{p+q})(\boldsymbol{Z}_i - \boldsymbol{Z}_j)\}, \quad 1 \le i,j \le n,\end{aligned}$$

$\boldsymbol{\Omega}_{n,ij}(\boldsymbol{d}) = \boldsymbol{\Omega}^X_{n,ij}(\boldsymbol{d})\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d})$. Furthermore, we define, for $1 \le i \le n$,

$$\boldsymbol{\tau}_i(\boldsymbol{d}) := \left(\left(\frac{\partial}{\partial\lambda}\mathbb{Y}_{n,i} - \frac{1}{E\left[\boldsymbol{1}_n^T\boldsymbol{\Omega}_n(\boldsymbol{d})\boldsymbol{1}_n\right]} E\left[\frac{\partial}{\partial\lambda}\mathbb{Y}_n^T\boldsymbol{\Omega}_n(\boldsymbol{d})\boldsymbol{1}_n\right]\right), - \left(\mathbb{X}^T_{n,i} - \frac{1}{E\left[\boldsymbol{1}_n^T\boldsymbol{\Omega}_n(\boldsymbol{d})\boldsymbol{1}_n\right]} E\left[\boldsymbol{1}_n^T\boldsymbol{\Omega}_n(\boldsymbol{d})\mathbb{X}_n\right]\right)\right)^T,$$

where $\frac{\partial}{\partial\lambda}\mathbb{Y}_{n,i}(\lambda) = \left(\frac{\partial}{\partial\lambda}T(Y_i,\lambda) - E[\frac{\partial}{\partial\lambda}T(Y_i,\lambda) \mid \boldsymbol{Z}_i]\right) f_z(\boldsymbol{Z}_i)$ and $\mathbb{X}_{n,i} = (\boldsymbol{X}_i - E[\boldsymbol{X}_i \mid \boldsymbol{Z}_i])f_z(\boldsymbol{Z}_i)$. In addition, let

$$\boldsymbol{\Phi}^X_{n,ij}(\boldsymbol{d}) = \boldsymbol{\Omega}^X_{n,ij}(\boldsymbol{d}) - E\left[\boldsymbol{\Omega}^X_{n,ik}(\boldsymbol{d}) \mid \boldsymbol{X}_i\right].$$

With all this in hand we can state the following Theorem.

Theorem 1.2 (Asymptotic normality)**.** *Assume that the conditions of Proposition 1.1 hold true. Then, uniformly with respect to $h \in \mathcal{H}_{sc,n}$, $\boldsymbol{d} \in \mathcal{D}$ and $s \in S_n$,*

$$\sqrt{n}\left((\widehat{\lambda}, \widehat{\boldsymbol{\beta}}(\widehat{\lambda})^T)^T - (\lambda_0, \boldsymbol{\beta}_0^T)^T\right) = -\boldsymbol{V}(\boldsymbol{d})^{-1}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^n \varepsilon_j f_z(\boldsymbol{Z}_j) E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\,\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d})\boldsymbol{\Phi}^X_{n,ij}(\boldsymbol{d}) \mid \boldsymbol{X}_j, \boldsymbol{Z}_j\right]\right) + o_{\mathbb{P}}(1),$$

converges in distribution to a tight random process whose marginal distribution is zero-mean normal with covariance function $\boldsymbol{V}(\boldsymbol{d}_1)^{-1}\boldsymbol{\Delta}(\boldsymbol{d}_1,\boldsymbol{d}_2)\boldsymbol{V}(\boldsymbol{d}_2)^{-1}$ where

$$\boldsymbol{V}(\boldsymbol{d}) = \begin{pmatrix} E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{D}_n(\boldsymbol{d})\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right] & -E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{D}_n(\boldsymbol{d})\mathbb{X}_n\right] \\ -E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n(\boldsymbol{d})\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right] & E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n(\boldsymbol{d})\mathbb{X}_n\right]\end{pmatrix}$$

and $\boldsymbol{\Delta}(\boldsymbol{d}_1,\boldsymbol{d}_2) = E\left[Var\left[\varepsilon_j f_z(\boldsymbol{Z}_j) \mid \boldsymbol{X}_j, \boldsymbol{Z}_j\right]\boldsymbol{\tau}_i(\boldsymbol{d}_1)\,\boldsymbol{\tau}_k(\boldsymbol{d}_2)^T\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d}_1)\boldsymbol{\Omega}^Z_{n,kj}(\boldsymbol{d}_2)\boldsymbol{\Phi}^X_{n,ij}(\boldsymbol{d}_1)\boldsymbol{\Phi}^X_{n,kj}(\boldsymbol{d}_2)\right].$

Due to the estimation error coming from the estimation of $\boldsymbol{\eta}$ we need $\boldsymbol{\Phi}^X_{n,ij}(\boldsymbol{d})$ to state the asymptotic variance of our estimators. If $\boldsymbol{\eta}$ was known $\boldsymbol{\Phi}^X_{n,ij}(\boldsymbol{d})$ should be replaced by $\boldsymbol{\Omega}^X_{n,ij}(\boldsymbol{d})$.

We can estimate the covariance matrix by $\widehat{\boldsymbol{V}}(\boldsymbol{d}_1)^{-1}\widehat{\boldsymbol{\Delta}}(\boldsymbol{d}_1,\boldsymbol{d}_2)\widehat{\boldsymbol{V}}(\boldsymbol{d}_2)^{-1}$, where

$$\begin{aligned}\widehat{\boldsymbol{V}}(\boldsymbol{d}) &= \begin{pmatrix} n^{-2}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\widehat{\lambda})^T\mathbb{D}_n(\boldsymbol{d})\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\widehat{\lambda}) & -n^{-2}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\widehat{\lambda})^T\mathbb{D}_n(\boldsymbol{d})\widehat{\mathbb{X}}_n \\ -n^{-2}\widehat{\mathbb{X}}_n^T\mathbb{D}_n(\boldsymbol{d})\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\widehat{\lambda}) & n^{-2}\widehat{\mathbb{X}}_n^T\mathbb{D}_n(\boldsymbol{d})\widehat{\mathbb{X}}_n\end{pmatrix} \\ &\textit{and} \\ \widehat{\boldsymbol{\Delta}}(\boldsymbol{d}_1,\boldsymbol{d}_2) &= n^{-3}\left(\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\widehat{\lambda}), -\widehat{\mathbb{X}}_n\right)^T \mathbb{D}_{n,inf}(\boldsymbol{d}_1)\widehat{\boldsymbol{\Phi}}_n(\boldsymbol{d}_1)\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\Phi}}_n^T(\boldsymbol{d}_2)\mathbb{D}^T_{n,inf}(\boldsymbol{d}_2)\left(\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\widehat{\lambda}), -\widehat{\mathbb{X}}_n\right).\end{aligned} \tag{1.10}$$

Here, $\widehat{\boldsymbol{\Phi}}^X_n$ and $\widehat{\boldsymbol{\Phi}}_n$ are the $n \times n-$ symmetric matrices with elements

$$\begin{aligned}\widehat{\boldsymbol{\Phi}}^X_{n,ij}(\boldsymbol{d}) &= \boldsymbol{\Omega}^X_{n,ij}(\boldsymbol{d}) - \frac{1}{n}\sum_{k=1}^n \boldsymbol{\Omega}^X_{n,ik}(\boldsymbol{d}), \quad && 1 \le i,j \le n \\ \widehat{\boldsymbol{\Phi}}_{n,ij}(\boldsymbol{d}) &= \widehat{\boldsymbol{\Phi}}^X_{n,ij}(\boldsymbol{d})\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d}), && 1 \le i,j \le n \\ \textit{and} \quad \mathbb{D}_{n,inf}(\boldsymbol{d}) &= \boldsymbol{I}_{n\times n} - \frac{1}{\boldsymbol{1}_n^T\boldsymbol{\Omega}_n(\boldsymbol{d})\boldsymbol{1}_n}\boldsymbol{\Omega}_n(\boldsymbol{d})\boldsymbol{1}_n\boldsymbol{1}_n^T.\end{aligned}$$

$\widehat{\boldsymbol{\Sigma}}_n = \operatorname{diag}\Big(\widehat{Var}\left[\varepsilon_1 f_z(\boldsymbol{Z}_1) \mid \boldsymbol{X}_1, \boldsymbol{Z}_1\right], \ldots, \widehat{Var}\left[\varepsilon_n f_z(\boldsymbol{Z}_n) \mid \boldsymbol{X}_n, \boldsymbol{Z}_n\right]\Big)$ is an estimator of the error variance. One can use a nonparametric estimator for the conditional variance or alternatively use an estimate of the error terms to approximate the conditional variance in the spirit of the Eiker-White variance estimator. Consistency of the above estimators is straightforward to establish.

Remark 2. It is also possible to estimate the unknown parameters λ and $\boldsymbol{\beta}$ without the intercept nuisance parameter γ. In that case $\mathbb{D}_n$ is replaced by $\boldsymbol{\Omega}_n$ in the estimation of λ and $\boldsymbol{\beta}$. The estimator is than still $\sqrt{n}-$consistent and is asymptotically normally distributed. In the variance $\boldsymbol{V}(\boldsymbol{d}_1)^{-1}\boldsymbol{\Delta}(\boldsymbol{d}_1,\boldsymbol{d}_2)\boldsymbol{V}(\boldsymbol{d}_2)^{-1}$ $\mathbb{D}_n$ is replaced by $\boldsymbol{\Omega}_n$ and $\boldsymbol{\tau}_i(\boldsymbol{d})$ by $\widetilde{\boldsymbol{\tau}}_i(\boldsymbol{d}) = \left(\frac{\partial}{\partial\lambda}\mathbb{Y}_{n,i}(\lambda), -\mathbb{X}_{n,i}^T\right)^T$. When estimating the variance $\mathbb{D}_{n,inf}$ has to be replaced by $\boldsymbol{I}_{n\times n}$ and $\mathbb{D}_n$ by $\boldsymbol{\Omega}_n$. However, when the model parameters are estimated with intercept nuisance parameter γ the impact of estimating $\boldsymbol{\eta}$ on the asymptotic variance becomes small. In the proof of Theorem 1.2 it was established that

$$\left((\widehat{\lambda},\widehat{\boldsymbol{\beta}}(\widehat{\lambda})^T)^T-(\lambda_0,\boldsymbol{\beta}_0^T)^T\right) = -\boldsymbol{V}(\boldsymbol{d})^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\,\boldsymbol{\Omega}_{n,ij}(\boldsymbol{d})\mid \boldsymbol{X}_j,\boldsymbol{Z}_j\right]\right.$$
$$\left.-\frac{1}{n}\sum_{k=1}^{n}\varepsilon_k f_z(\boldsymbol{Z}_k)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ik}^Z(\boldsymbol{d})\boldsymbol{\Omega}_{n,ij}^X(\boldsymbol{d})\mid \boldsymbol{Z}_k\right]\right)+o_{\mathbb{P}}\left(n^{-1/2}\right).$$

The second sum $\frac{1}{n}\sum\limits_{k=1}^{n}\varepsilon_k f_z(\boldsymbol{Z}_k)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ik}^Z\boldsymbol{\Omega}_{n,ij}^X\mid \boldsymbol{Z}_k\right]$ in the asymptotic representation of $(\widehat{\lambda},\widehat{\boldsymbol{\beta}}(\widehat{\lambda})^T)^T$ is due to the estimation of $\boldsymbol{\eta}$. If we would know $\boldsymbol{\eta}$ this sum would not be present. In the following we consider $E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ik}^Z\boldsymbol{\Omega}_{n,ij}^X\mid \boldsymbol{Z}_k\right]$. If we could replace the index k by j in $\boldsymbol{\Omega}_{n,ik}^Z$ we would get that

$$E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ij}^Z\boldsymbol{\Omega}_{n,ij}^X\mid \boldsymbol{Z}_k\right]=E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ij}^Z\boldsymbol{\Omega}_{n,ij}^X\right]=\boldsymbol{0}_{p+1},$$

such that the second sum would not be present as well. Of course this is not possible. However, here we consider $\boldsymbol{d}$ as a vector with elements playing the role of standardizing constants. If the elements of $\boldsymbol{d}$ are the inverse of a kernel smoothing bandwidth tending to zero at a suitable rate $E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ik}^Z\boldsymbol{\Omega}_{n,ij}^X\mid \boldsymbol{Z}_k\right]$ tends to zero for $n\to\infty$. The exact rate is over the scope of this study, but we will show in the simulation section that even in case of $\boldsymbol{d}$ being a vector of constants it seems that we can forget about the second part in the asymptotic representation. The estimator is labeled SmoothMD* in the simulation section. If we do not consider the second part in the asymptotic representation the variance is estimated by replacing $\widehat{\boldsymbol{\Phi}}_{n,ij}(\boldsymbol{d})$ with $\boldsymbol{\Omega}_{n,ij}(\boldsymbol{d})$.

Note that the constant γ is necessary in order to be able to remove the second part in the representation as $E\left[\widetilde{\boldsymbol{\tau}}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ij}^Z\boldsymbol{\Omega}_{n,ij}^X\mid \boldsymbol{Z}_k\right]=E\left[\widetilde{\boldsymbol{\tau}}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ij}^Z\boldsymbol{\Omega}_{n,ij}^X\right]\neq\boldsymbol{0}_{p+1}$.

1.4. Testing based on SmoothMD for parameter restrictions

In section 1.3 we established consistency and asymptotic normality of our estimator. The asymptotic behavior of our estimator is not influenced by the standardization with s but the asymptotic variance is affected by the estimation of $\boldsymbol{\eta}$. In addition, the behavior of our estimator is, even asymptotically, influenced by the vector $\boldsymbol{d}$. When developing a test theory we should take that influence into account in order to get reliable results. That's what we do in the following.

1.4.1. Testing the transformation parameter

When it comes to testing parameter restrictions in the semiparametric partially linear regression model with Box-Cox transformation we might be mainly interested in testing if λ is zero or not and if the components of $\boldsymbol{\beta}$ are zero. However, we will consider here a more general approach to allow for more complex hypotheses as well. We separate the discussion into two parts. In the first part, we consider only restrictions for λ and in the second part we consider restrictions for $\boldsymbol{\beta}$ with and without restricting λ.

Suppose we want to test the restriction for λ given by

$$H_0:\lambda_0=\lambda_R. \tag{1.11}$$

In order to test this restriction, we can use the distance metric statistic proposed by Lavergne and Patilea [56]. Adapted to our case and for testing (1.11) we consider the distance

$$DM_\lambda=\frac{1}{n}\widehat{\mathbb{Y}}_n(\lambda_R)^T\,\widehat{\mathbb{B}}_n\widehat{\mathbb{Y}}_n(\lambda_R)-\frac{1}{n}\widehat{\mathbb{Y}}_n(\widehat{\lambda})^T\,\widehat{\mathbb{B}}_n\widehat{\mathbb{Y}}_n(\widehat{\lambda}).$$

The distance metric is based on the object that is minimized to get the estimate for λ, see equation (1.9). However, the test statistic is not standardized by $s^{-\lambda}$ as we need this only for the estimation of λ.

Let

$$\boldsymbol{A}_n = \begin{pmatrix} \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T \\ -\mathbb{X}_n^T \end{pmatrix} \mathbb{D}_n \left((\boldsymbol{\varepsilon}\boldsymbol{f_z})_n - \left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}} \right)_n \right).$$

Therefore, we can now state the following Proposition.

Proposition 1.2. *Assume that the conditions of Proposition 1.1 hold true. Then, uniformly with respect to $h \in \mathcal{H}_{sc,n}$, $\boldsymbol{d} \in \mathcal{D}$ and $s \in S_n$,*

$$DM_\lambda - (1, \boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1} n^{-3/2} \boldsymbol{A}_n n^{-3/2} \boldsymbol{A}_n^T \boldsymbol{V}(\boldsymbol{d})^{-1}(1, \boldsymbol{0}_p^T)^T E\left[n^{-2} \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T \mathbb{B}_n \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0) \right] = o_{\mathbb{P}}(1),$$

under H_0 and $\mathbb{P}(n^{-1}DM_\lambda > c) \to 1$ for some $c > 0$ if H_0 does not hold.

The process $(1, \boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1} n^{-3/2} \boldsymbol{A}_n n^{-3/2} \boldsymbol{A}_n^T \boldsymbol{V}(\boldsymbol{d})^{-1}(1, \boldsymbol{0}_p^T)^T E\left[n^{-2} \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T \mathbb{B}_n \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0) \right]$ is asymptotically tight and for each $\boldsymbol{d}$ behaves asymptotically as a chi-square times $(1, \boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}\boldsymbol{\Delta}(\boldsymbol{d}, \boldsymbol{d})\boldsymbol{V}(\boldsymbol{d})^{-1}(1, \boldsymbol{0}_p^T)^T E\left[n^{-2} \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T \mathbb{B}_n \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0) \right]$, see Johnson et al. [49]. The distribution of the distance metric statistic is, thus, in general non-pivotal. Determining critical values requires the estimation of $(1, \boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}\boldsymbol{\Delta}(\boldsymbol{d}, \boldsymbol{d})\boldsymbol{V}(\boldsymbol{d})^{-1}(1, \boldsymbol{0}_p^T)^T E\left[n^{-2} \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T \mathbb{B}_n \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0) \right]$, which can rely on the estimators stated in (1.10).

1.4.2. Testing the slope coefficients

In the next part we consider restrictions for $\boldsymbol{\beta}$. Suppose we want to test r linear restrictions for $\boldsymbol{\beta}$ given by

$$H_0 : \boldsymbol{R}\boldsymbol{\beta}_0 = \boldsymbol{c}, \tag{1.12}$$

where $\boldsymbol{R}$ is a $r \times p-$ matrix of full rank and $\boldsymbol{c} \in \mathbb{R}^r$. In order to test the restrictions, we need to find the restricted estimators for $\boldsymbol{\beta}_0$, $\widehat{\boldsymbol{\beta}}_R(\lambda)$, and λ_0, $\widehat{\lambda}_R$. We minimize

$$n^{-2}s^{-2\lambda}\left(\widehat{\mathbb{Y}}_n(\lambda) - \widehat{\mathbb{X}}_n\boldsymbol{\beta}\right)^T \mathbb{D}_n \left(\widehat{\mathbb{Y}}_n(\lambda) - \widehat{\mathbb{X}}_n\boldsymbol{\beta}\right) \quad s.t. \quad \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{c},$$

with respect to $\boldsymbol{\beta}$ and get that

$$\widehat{\boldsymbol{\beta}}_R(\lambda) = \widehat{\boldsymbol{\beta}}(\lambda) - \left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\boldsymbol{R}^T\right)^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{c}\right).$$

The restricted estimator for λ_0 is then given by

$$\widehat{\lambda}_R = \widehat{\lambda}_R(s) = \arg\min_{\lambda\in\Lambda} s^{-\lambda}\left(\widehat{\mathbb{Y}}_n(\lambda) - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R(\lambda)\right)^T \mathbb{D}_n\, s^{-\lambda}\left(\widehat{\mathbb{Y}}_n(\lambda) - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R(\lambda)\right). \tag{1.13}$$

With all the estimators in hand we can now define our distance metric statistic for testing (1.12).

$$DM_{\boldsymbol{\beta}} = \frac{1}{n}\left(\widehat{\mathbb{Y}}_n(\widehat{\lambda}_R) - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R(\widehat{\lambda}_R)\right)^T \mathbb{D}_n\left(\widehat{\mathbb{Y}}_n(\widehat{\lambda}_R) - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R(\widehat{\lambda}_R)\right) - \frac{1}{n}\widehat{\mathbb{Y}}_n(\widehat{\lambda})^T\,\widehat{\mathbb{B}}_n\widehat{\mathbb{Y}}_n(\widehat{\lambda}).$$

The distance metric is based on the object that is minimized to get the restricted estimate for λ, see equation (1.13). Once again the test statistic is not standardized by $s^{-\lambda}$ as we need this only for the estimation of λ. Let,

$$\mathbb{B}_{n,R} = \mathbb{B}_n + \mathbb{D}_n\mathbb{X}_n\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\mathbb{X}_n^T\mathbb{D}_n,$$

and

$$\boldsymbol{V}_R(\boldsymbol{d}) = \left(E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_{n,R}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]^{-1}, E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_{n,R}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]^{-1} E\left[\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{D}_n\mathbb{X}_n\mathbb{B}_n^+\right]\right),$$

where

$$\mathbb{B}_n^+ = \left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1} - \left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}.$$

Therefore, we can now state the following proposition.

Proposition 1.3. *Assume that the conditions of Proposition 1.1 hold true. Then, uniformly with respect to $h \in \mathcal{H}_{sc,n}$, $\boldsymbol{d} \in \mathcal{D}$ and $s \in S_n$,*

$$DM_{\boldsymbol{\beta}}$$

$$-n^{-3/2}\boldsymbol{A}_n^T\Bigg((\boldsymbol{0}_{p\times 1}, \boldsymbol{I}_{p\times p})^T E\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}E\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}(\boldsymbol{0}_{p\times 1}, \boldsymbol{I}_{p\times p})$$
$$-\boldsymbol{V}_R(\boldsymbol{d})^T\boldsymbol{V}_R(\boldsymbol{d})E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_{n,R}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]$$
$$+\boldsymbol{V}(\boldsymbol{d})^{-1}(1,\boldsymbol{0}_p^T)^T(1,\boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\Bigg)\boldsymbol{A}_n n^{-3/2} = o_{\mathbb{P}}(1),$$

under H_0 and $\mathbb{P}(n^{-1}DM_{\boldsymbol{\beta}} > c) \to 1$ for some $c > 0$ if H_0 does not hold.

The process in Proposition 1.3 is asymptotically tight and for each $\boldsymbol{d}$ behaves asymptotically as a weighted sum of $p+1-r$ independent chi-squares, where the weights are the positive eigenvalues of

$$(\boldsymbol{0}_{p\times 1}, \boldsymbol{I}_{p\times p})^T E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}(\boldsymbol{0}_{p\times 1}, \boldsymbol{I}_{p\times p})\,\boldsymbol{\Delta}(\boldsymbol{d},\boldsymbol{d})$$
$$-\boldsymbol{V}_R(\boldsymbol{d})^T\boldsymbol{V}_R(\boldsymbol{d})\boldsymbol{\Delta}(\boldsymbol{d},\boldsymbol{d})E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_{n,R}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]$$
$$+\boldsymbol{V}(\boldsymbol{d})^{-1}(1,\boldsymbol{0}_p^T)^T(1,\boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}\boldsymbol{\Delta}(\boldsymbol{d},\boldsymbol{d})E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right],$$

see Johnson et al. [49]. Determining critical values requires the estimation of the last display. We can use the estimators stated in (1.10) and for all other components we simply replace the unknown expressions by their sample equivalence, e.g. estimate $\mathbb{B}_{n,R}$ by

$$\widehat{\mathbb{B}}_n + \mathbb{D}_n\widehat{\mathbb{X}}_n\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n.$$

1.4.3. Testing the transformation parameter and the slope coefficients

Finally, we consider the combined restrictions for $\boldsymbol{\beta}$ and λ. Suppose we want to test

$$H_0 : \boldsymbol{R}\boldsymbol{\beta}_0 = \boldsymbol{c} \quad \textit{and} \quad \lambda_0 = \lambda_R.$$

In contrast to the hypothesis stated in (1.12) we do not need to estimate $\widehat{\lambda}_R$. Therefore, the distance metric statistic is for this case given by

$$DM_{\boldsymbol{\beta},\lambda} = \frac{1}{n}\left(\widehat{\mathbb{Y}}_n(\lambda_R) - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R(\lambda_R)\right)^T\mathbb{D}_n\left(\widehat{\mathbb{Y}}_n(\lambda_R) - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R(\lambda_R)\right) - \frac{1}{n}\widehat{\mathbb{Y}}_n(\widehat{\lambda})^T\,\widehat{\mathbb{B}}_n\widehat{\mathbb{Y}}_n(\widehat{\lambda}).$$

In addition, $DM_{\boldsymbol{\beta},\lambda}$ does not converge to the same expression as $DM_{\boldsymbol{\beta}}$ as λ_R is fixed. Therefore, we state the following proposition.

Proposition 1.4. *Assume that the conditions of Proposition 1.1 hold true. Then, uniformly with respect to $h \in \mathcal{H}_{sc,n}$, $\boldsymbol{d} \in \mathcal{D}$ and $s \in S_n$,*

$$DM_{\boldsymbol{\beta},\lambda}$$

$$-n^{-3/2}\boldsymbol{A}_n^T\Bigg((\boldsymbol{0}_{p\times 1}, \boldsymbol{I}_{p\times p})^T E\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}E\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}(\boldsymbol{0}_{p\times 1}, \boldsymbol{I}_{p\times p})$$
$$+\boldsymbol{V}(\boldsymbol{d})^{-1}(1,\boldsymbol{0}_p^T)^T(1,\boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\Bigg)\boldsymbol{A}_n n^{-3/2} = o_{\mathbb{P}}(1),$$

under H_0 and $\mathbb{P}(n^{-1}DM_{\boldsymbol{\beta},\lambda} > c) \to 1$ for some $c > 0$ if H_0 does not hold.

The process in Proposition 1.4 is asymptotically tight and for each $\boldsymbol{d}$ behaves asymptotically as a weighted sum of $p - r$ independent chi-squares, where the weights are the positive eigenvalues of

$$(\mathbf{0}_{p\times 1}, \boldsymbol{I}_{p\times p})^T E\left[\mathbb{X}_n^T \mathbb{D}_n \mathbb{X}_n\right]^{-1} \boldsymbol{R}^T \left(\boldsymbol{R} E\left[\mathbb{X}_n^T \mathbb{D}_n \mathbb{X}_n\right]^{-1} \boldsymbol{R}^T\right)^{-1} \boldsymbol{R} E\left[n^{-2}\mathbb{X}_n^T \mathbb{D}_n \mathbb{X}_n\right]^{-1} (\mathbf{0}_{p\times 1}, \boldsymbol{I}_{p\times p})\, \boldsymbol{\Delta}(\boldsymbol{d}, \boldsymbol{d})$$
$$+ \boldsymbol{V}(\boldsymbol{d})^{-1}(1, \mathbf{0}_p^T)^T(1, \mathbf{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}\boldsymbol{\Delta}(\boldsymbol{d}, \boldsymbol{d}) E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T \mathbb{B}_n \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right],$$

see Johnson et al. [49]. Determining critical values requires the estimation of the last display. We can use the estimators stated in (1.10) and for all other components we simply replace the unknown expressions by their sample equivalence.

Remark 3. The Propositions of section 1.4 are also valid if the unknown parameters λ and $\boldsymbol{\beta}$ are estimated without the intercept nuisance parameter γ. In that case $\mathbb{D}_n$ is replaced by $\boldsymbol{\Omega}_n$ in the statements. Moreover, when estimating the unknown variance $\mathbb{D}_{n,inf}$ has to be replace by $\boldsymbol{I}_{n\times n}$.

The usual chi-square distribution might appear when we use an efficient estimator reaching the semiparametric efficiency bound, i.e. we would need an optimal weighting matrix and $\boldsymbol{d}$ tending to zero. If this is possible and how the weighting matrix would need to look like is left for future research.

1.5. Small sample study and real data application

In this section we consider the small sample behavior of our estimator. We conduct several simulation experiments to consider bias and standard deviation for the estimated parameters. In addition, we conduct hypothesis tests as discussed in section 1.4. We begin with a consideration of the simulation setup. We then state our simulation results and, finally, close the section with a real data application.

1.5.1. Simulation setup

During the simulation, we consider four different models. The models are given by

Model 1: $T(Y, \lambda_0) = X\beta_0 + m(Z) + \varepsilon$, $m(Z) = \frac{\exp\{Z\}}{1+\exp\{Z\}} + \frac{1}{3}$ with $Z \sim N(1,1)$, $\lambda_0 = 0$ and $\beta_0 = 1$, $X = -\frac{2}{3}Z + u$ with $u \sim N(0,1)$ and $\varepsilon = \sqrt{\frac{1+X^2}{2}}\, \widetilde{u}$ with $\widetilde{u} \sim N\left(0, \frac{1}{13}\right)$.

Model 2: $T(Y, \lambda_0) = X\beta_0 + m(Z) + \varepsilon$, $m(Z) = \frac{\exp\{Z\}}{1+\exp\{Z\}} + 3$ with $Z \sim N(1,1)$, $\lambda_0 = 0.5$ and $\beta_0 = 1$, $X = -\frac{2}{3}Z + u$ with $u \sim N(0,1)$ and $\varepsilon \sim N\left(0, \frac{1}{9}\right)$.

Model 3: $T(Y, \lambda_0) = X\beta_0 + m(Z) + \varepsilon$, $m(Z) = \frac{\exp\{Z\}}{1+\exp\{Z\}} - 1$ with $Z \sim U(-3,-1)$, $\lambda_0 = -1$ and $\beta_0 = 1$, $X = \frac{2}{3}Z + u$ with $u \sim U(-1,1)$ and $\varepsilon \sim U\left(-\sqrt{1/9}, \sqrt{1/9}\right)$.

Model 4: $T(Y, \lambda_0) = X_1\beta_{10} + \boldsymbol{X}_2\boldsymbol{\beta}_{20} + m(Z_1, Z_2) + \varepsilon$, $m(Z_1, Z_2) = \frac{1}{3} + Z_1 + Z_2 + Z_1 Z_2$ with $Z_1, Z_2 \sim N(0,1)$, $\lambda_0 = 0$, $\beta_{10} = 1$, $X_1 = -\frac{1}{3}(Z_1 + Z_2) + u$ with $u \sim N(0,1)$, $X_{2,l} \overset{i.i.d.}{\sim} Ber(0.2)$ and $\beta_{2,l} \overset{i.i.d.}{\sim} U(-1,1)$ for $l = 1, \ldots, 30$, $\varepsilon \sim N\left(0, \frac{1}{9}\right)$.

The main difference of the models is the transformation parameter λ. Model 1 and Model 4 have $\lambda_0 = 0$, whereas Model 2 has $\lambda_0 = 0.5$ and Model 3 $\lambda_0 = -1$. To ensure that $Y > 0$ in Model 3 we draw the random variables from uniform distributions. In all other models positivity of Y is ensured as well. Model 1 has heteroskedastic error terms which is captured by the developed theory. Model 4 contains 30 dummy variabels, $\boldsymbol{X}_2$, which take the value 1 with probability 20%.

The estimators are computed by employing a normal kernel for $K(\cdot)$. $\boldsymbol{Z}$ is standardized componentwise by the corresponding standard deviations and $h \propto n^{-1/3.5}$. This bandwidth choice satisfies the assumptions of Theorem 1.2. The components of $\boldsymbol{d}$ defining the diagonal matrix $\boldsymbol{D}$ in $\boldsymbol{\Omega}_n$ are set equal to the componentwise standard deviations of X and $\boldsymbol{Z}$ when X is continuous. In case of the dummy variables $\boldsymbol{X}_2$ an indicator of the event that the observations have the same value is employed. For Model 4 we ensure in the simulations that for every observation there exist at least 4 observations with the same dummy variable combination.

In the estimation we define a grid for values of λ that are considered during the optimization. This optimization grid for λ is given in our simulation by the grid $[\lambda_0 - 0.8, \lambda_0 + 0.8]$ with step size 0.001. We minimize $G_n^{-\lambda}\widehat{\mathbb{Y}}_n(\lambda)^T\, \widehat{\mathbb{B}}_n\, G_n^{-\lambda}\widehat{\mathbb{Y}}_n(\lambda)$ over the defined grid to get $\widehat{\lambda}$ and $\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$, where $G_n = \prod_{i=1}^{n} Y_i^{1/n}$ is the geometric mean.

In the simulation we compare the proposed estimator where γ is employed with the estimator that does not use γ. As explained in section 1.3 both estimators converge asymptotically to a normal distribution. Due to the fact that the estimator with γ reduces the influence on the variance coming from the estimation of $\boldsymbol{\eta}$ it is interesting to consider both estimators.

We consider bias and standard deviation of the estimators as well as power and size of the distance metric statistics proposed in section 1.4. In addition, we test by a simple Z-Test if the estimated parameters are significantly different from the true value. Therefore, we employ the variance estimator stated in equation (1.10) for both estimators with the necessary adjustments for the estimator without γ. To estimate the error variance we employ the Eiker-White variance estimator. In order to see the influence of the estimated $\boldsymbol{\eta}$ on the variance we consider all tests also without taking the estimation error of $\boldsymbol{\eta}$ into account. Therefore, we replace $\widehat{\boldsymbol{\Phi}}_n$ by $\boldsymbol{\Omega}_n$ in the variance estimator.

In addition, the Nonlinear two-stage Least Squares (NL2SLS) estimator for the Box-Cox model introduced by Amemiya and Powell [8] is considered as competitor. In order to be able to employ this estimator it is assumed that the function $m(\cdot)$ is known and, thus, $m(\boldsymbol{Z})$ can be added as additional regressor. The instruments are, therefore, given by $\boldsymbol{W}_i = (1, \boldsymbol{X}_i, \boldsymbol{X}_i^2, m(\boldsymbol{Z}_i), m(\boldsymbol{Z}_i)^2)$. We consider the Z-Test for the NL2SLS estimator as well where we employ the Eiker-White variance estimator again.

1.5.2. Simulation results

Table 1.1 states the results for bias and standard deviation for λ and β in Model 1. All three estimators have comparable results for bias and the bias decreases with sample size for β, whereas it is the lowest for $n = 500$ for the SmoothMD estimators in case of λ. Surprisingly, the standard deviation is also comparable for all three estimators even though $m(\cdot)$ is given for the NL2SLS estimator.

Table 1.2 states the results for bias and standard deviation for λ, β_1 and β_2, one representative parameter out of the 30 parameters in $\boldsymbol{\beta}_2$, in Model 4. All three estimators have comparable results for bias and the bias decreases with sample size for all three parameters. In contrast to the results for Model 1, the standard deviation is smaller in case of the NL2SLS estimator for β_1 and β_2. This result should be expacted as $m(\cdot)$ is given for the NL2SLS estimator. The standard deviations for the SmoothMD estimators with and without γ are as in Model 1 nearly the same.

Table 1.1: *Bias and Standard Deviation of the estimators for λ and β in Model 1.*

	s	Bias			St. dev.		
n		250	500	1000	250	500	1000
λ estimator							
SmoothMD with γ	G_n	0.003	0.0001	0.001	0.042	0.03	0.021
SmoothMD without γ	G_n	0.002	0.0001	0.001	0.041	0.029	0.02
NL2SLS	G_n	-0.003	-0.001	0.0001	0.042	0.029	0.02
β estimator							
SmoothMD with γ	G_n	-0.001	-0.001	0.0004	0.036	0.025	0.017
SmoothMD without γ	G_n	-0.001	-0.001	0.0004	0.035	0.024	0.017
NL2SLS	G_n	-0.002	-0.001	0.0002	0.035	0.024	0.016

Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The grid for λ is $[\lambda_0 - 0.8, \lambda_0 + 0.8]$. For all simulations 2000 Monte Carlo samples were used.

Table 1.2: *Bias and Standard Deviation of the estimators for λ, β_1 and β_2 in Model 4.*

	s	Bias			St. dev.		
n		250	500	1000	250	500	1000
λ estimator							
SmoothMD with γ	G_n	0.0001	-0.0002	-0.0001	0.015	0.01	0.008
SmoothMD without γ	G_n	0.0001	-0.0002	-0.0001	0.015	0.01	0.008
NL2SLS	G_n	-0.0004	-0.0002	-0.0002	0.014	0.01	0.005
β_1 estimator							
SmoothMD with γ	G_n	-0.002	-0.002	-0.001	0.036	0.023	0.017
SmoothMD without γ	G_n	-0.002	-0.002	-0.001	0.036	0.023	0.017
NL2SLS	G_n	0.0004	-0.0001	-0.0001	0.025	0.015	0.011
β_2 estimator							
SmoothMD with γ	G_n	0.004	-0.001	-0.0002	0.133	0.065	0.042
SmoothMD without γ	G_n	0.004	-0.001	-0.0002	0.133	0.065	0.042
NL2SLS	G_n	-0.001	0.002	-0.0006	0.091	0.046	0.028

Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all continuous variables and for the dummy variables an indicator of the event that the observations have the same value is employed. The grid for λ is $[\lambda_0 - 0.8, \lambda_0 + 0.8]$. For all simulations 2000 Monte Carlo samples were used.

Table 1.3: *Empirical Level for distance metric statistics of the estimators for λ and β in Model 2.*

	s	5% level			10% level		
n		250	500	1000	250	500	1000
Test for λ							
SmoothMD with γ	G_n	10.3	8.15	7.8	12.75	10.65	11.45
SmoothMD* with γ	G_n	10.15	7.95	7.7	12.75	10.6	11.45
SmoothMD without γ	G_n	9.55	7.0	6.0	12.45	10.85	10.35
SmoothMD* without γ	G_n	3.75	1.75	1.15	6.35	3.45	2.85
Test for β							
SmoothMD with γ	G_n	10.4	8.25	7.95	12.55	11.1	11.9
SmoothMD* with γ	G_n	10.2	8.3	7.7	12.55	11.3	11.1
SmoothMD without γ	G_n	9.7	6.85	6.7	12.55	10.2	11.2
SmoothMD* without γ	G_n	3.7	1.75	1.45	6.5	3.7	2.9

Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

Table 1.3 states the empirical level for distance metric statistics of the estimators for λ and β in Model 2. Here we state the results for the SmoothMD estimator with correctly estimated variance as well as with variance estimate that does not account for the estimation of $\boldsymbol{\eta}$. In this setup the results of the SmoothMD estimators with and without γ differ. In addition, the estimation of $\boldsymbol{\eta}$ has an influence on the results of the SmoothMD estimator without γ. If we do not consider the additional variance term coming from the estimation of $\boldsymbol{\eta}$ in the estimation of the variance the tests have the wrong size. This does not happen for the SmoothMD estimator with γ. As explained in section 1.3 the constant γ reduces the influence of the estimation error on the variance. For the first three estimators the empirical levels converge to the nominal levels if the sample size increases and β seems to need a larger sample size than λ to get close to the nominal level.

Table 1.4 states the empirical level for the Z-Tests for λ, β_1 and β_2 in Model 4. The fact that we do not consider the estimation error has almost no influence on the results. In addition, both SmoothMD versions lead to similar results. However, in order to get close to the nominal level the sample size needs to be large as only for $n = 1000$ the SmoothMD estimators get close to the nominal level. The NL2SLS estimator gives more convincing results for smaller sample sizes. Note that the dummy variable coefficient β_2 seems to require a larger sample size than the other two parameters to get close to the nominal level when employing the SmoothMD estimators.

Table 1.4: *Empirical Level for Z-Tests of the estimators for λ, β_1 and β_2 in Model 4.*

	s	5% level			10% level		
n		250	500	1000	250	500	1000
Test for λ							
SmoothMD with γ	G_n	8.4	7.1	5.5	15.45	13.5	10.55
SmoothMD* with γ	G_n	8.5	7.1	5.45	15.3	13.45	10.55
SmoothMD without γ	G_n	9.6	8.15	5.2	16.45	12.6	11.65
SmoothMD* without γ	G_n	9.45	8.1	5.2	16.4	12.6	11.75
NL2SLS	G_n	9.6	7.6	6.0	16.8	12.5	11.1
Test for β_1							
SmoothMD with γ	G_n	11.8	8.55	6.4	18.25	14.05	11.75
SmoothMD* with γ	G_n	11.8	8.55	6.45	18.25	14.05	11.8
SmoothMD without γ	G_n	11.75	8.9	6.4	18.35	14.8	12.1
SmoothMD* without γ	G_n	11.75	8.95	6.45	18.35	14.8	12.1
NL2SLS	G_n	8.25	4.95	5.1	13.95	10.4	10.35
Test for β_2							
SmoothMD with γ	G_n	13.6	8.25	7.05	20.7	15.65	12.7
SmoothMD* with γ	G_n	13.7	8.25	7.05	20.75	15.55	12.75
SmoothMD without γ	G_n	13.75	8.35	6.65	20.6	14.55	12.25
SmoothMD* without γ	G_n	13.7	8.35	6.65	20.6	14.55	12.25
NL2SLS	G_n	7.65	6.55	4.7	13.45	13.05	9.55

Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all continuous variables and for the dummy variables an indicator of the event that the observations have the same value is employed. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. β_2 is one representative parameter out of the 30 parameters in $\boldsymbol{\beta}_2$. For all simulations 2000 Monte Carlo samples were used.*

Figure 1.1: *Power function of the distance metric statistic for λ of Model 3 with $n = 250$.*

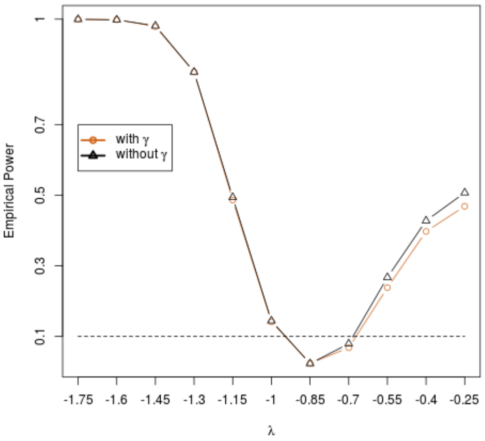
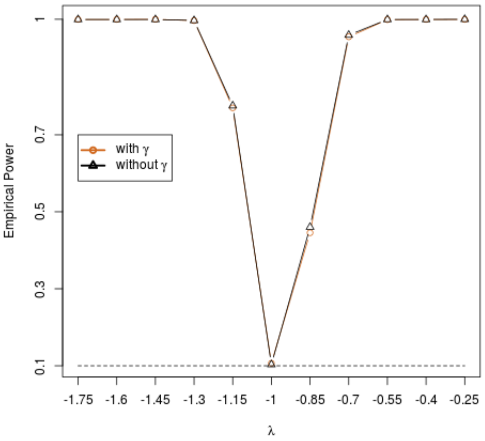


Figure 1.2: *Power function of the distance metric statistic for λ of Model 3 with $n = 1000$.*



Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. Only the SmoothMD estimators that take the additional variance part due to the estimation of $\boldsymbol{\eta}$ into account are considered. For all simulations 2000 Monte Carlo samples were used. The nominal level is 10%.

Figures 1.1 and 1.2 state the power functions of the distance metric statistic for λ in Model 3 with $n = 250$ and $n = 1000$. In case of $n = 250$ the power function is skewed and the power for values larger than -1 is small. In addition, the power function is smaller than the nominal value at -0.85 and -0.7. For the SmoothMD estimator without γ the power function is larger than for the SmoothMD estimator with γ at values larger than -1. These issues disappear for the larger sample size $n = 1000$.

Figures 1.3 and 1.4 state the power functions of the distance metric statistic for β in Model 2 with $n = 250$ and $n = 500$. As in Figure 1.1 the power function for $n = 250$ is skewed but the effect is less distinct. However, the power function is smaller than the nominal value at 0.8. For the SmoothMD estimator without γ the power function is larger than for the SmoothMD estimator with γ at values smaller than 1 for both sample sizes. For $n = 500$ the skewness is less pronounced and the power function has no values lower than the nominal value. The main conclusion from both power functions is that the samples size should not be too small so that the tests have a reasonable power.

Figure 1.3: *Power function of the distance metric statistic for β of Model 2 with $n = 250$.*

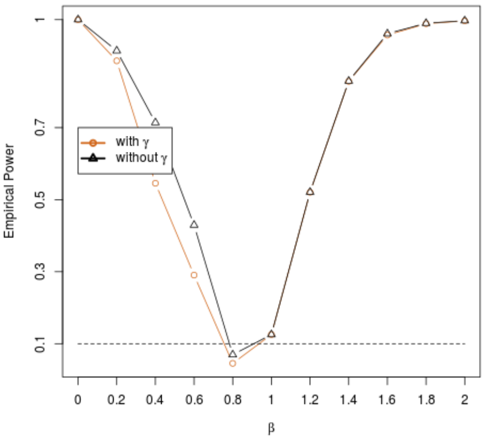
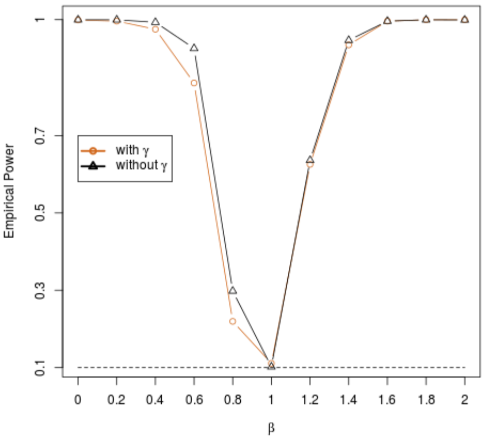


Figure 1.4: *Power function of the distance metric statistic for β of Model 2 with $n = 500$.*



Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. Only the SmoothMD estimators that take the additional variance part due to the estimation of $\boldsymbol{\eta}$ into account are considered. For all simulations 2000 Monte Carlo samples were used. The nominal level is 10%.

Before we consider a real data application we close the discussion with Figures 1.5 and 1.6. The figures state the estimated $m(Z)$ for Model 1 with $n = 250$ and for Model 3 with $n = 500$. For the estimation the NW estimator with same kernel and bandwidth as for the SmoothMD estimator was used. No matter if the SmoothMD estimator with or without γ is employed the results are very accurate. In practice one can of course use cross validation to choose the bandwidth or employ the local linear estimator instead of the NW estimator.

Figure 1.5: *Estimated $m(Z)$ for Model 1 with $n = 250$.*

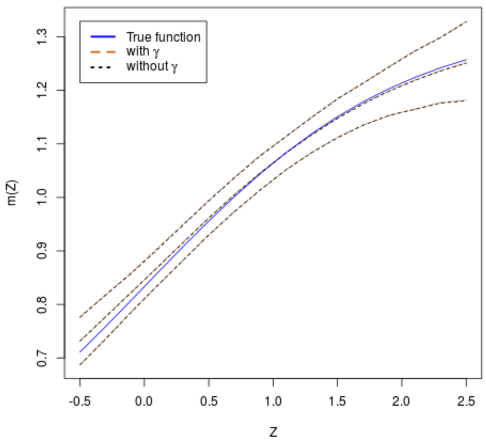
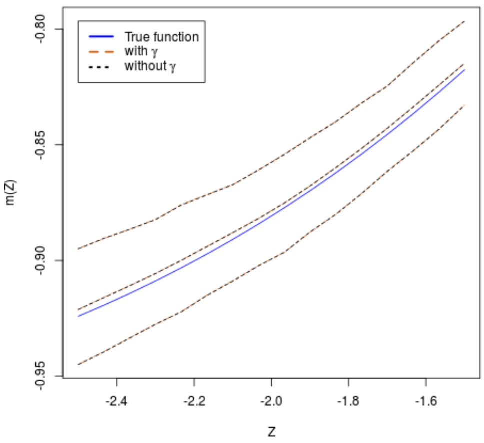


Figure 1.6: *Estimated $m(Z)$ for Model 3 with $n = 500$.*



Notes: For the estimation the NW estimator with normal kernel and $h \propto n^{-1/3.5}$ is employed. The 25% and 75% quantiles as well as the mean are reported. For all simulations 2000 Monte Carlo samples were used.

1.5.3. Real data application

We consider in this section an application of our estimator to investigate the returns of social and cognitive skills in the labor market. For this purpose we apply the proposed transformation partially linear estimator to a dataset studied in Deming [28]. In particular, we consider regression (4) and (5) in TABLE I of Deming [28] that is based on the National Longitudinal Survey of Youth 1979 (NLSY79). The NLSY79 is a nationally representative sample of youth aged 14 to 22 in 1979 conducted in the US. The survey was conducted yearly from 1979 to 1993 and biannually from 1994 through 2012. Deming [28] estimates the model

$$\begin{aligned} \log(wage_{ijt}) = \alpha + \beta_1 \cdot COG_i + \beta_2 \cdot SS_i + \beta_3 \cdot COG_i \times SS_i + \beta_4 \cdot NCOG_i \\ + \boldsymbol{C}_{ijt}^T \boldsymbol{\rho} + \delta_j + \zeta_t + \varepsilon_{ijt}, \end{aligned} \tag{1.14}$$

where COG, SS and $NCOG$ denote measures of cognitive, social and noncognitive skills. The model includes controls $\boldsymbol{C}$ for race-by-gender indicators, indicators for region and urbanicity as well as age (indexed by j) and year (indexed by t) fixed effects.

In his paper Deming [28] develops a theoretical model that is written in levels instead of logs as in equation (1.14). Nevertheless, he estimates the log-linearized model in his paper to follow standard practice in the literature, as he argues. Results for the model in levels are stated in an online appendix. Therefore, it makes sense to use the Box-Cox transformation for $wage$ and estimate the transformation paramter λ together with the remaining model parameters to decide whether the model in logs or in levels is more appropriate.

Furthermore, we consider an unknown functional form for cognitive and social skills to see if the linear form $\beta_1 \cdot COG + \beta_2 \cdot SS + \beta_3 \cdot COG \times SS$ used by Deming [28] is reasonable. The transformation partially linear model is, thus, given by

$$T(wage_{ijt}, \lambda) = m(COG_i, SS_i) + \beta \cdot NCOG_i + \boldsymbol{C}_{ijt}^T \boldsymbol{\rho} + \delta_j + \zeta_t + \varepsilon_{ijt}, \tag{1.15}$$

where $m(\cdot)$ is an unknown function.

As proxy for cognitive skills the Armed Forces Qualifying Test (AFQT) is employed. Deming [28] uses raw scores from Altonji et al. [4] and normalizes them to have mean 0 and standard deviation 1. The social skill measure is constructed from the following four variables of the NLSY79:

1. Self-reported sociability in 1981 (extremely shy, somewhat shy, somewhat outgoing, extremely outgoing)
2. Self-reported sociability in 1981 at age 6 (retrospective)
3. The number of clubs in which the respondent participated in high school
4. Participation in high school sports (yes/no).

Each variable is normalized to have mean 0 and standard deviation 1. The social skill measure is the average of this four normalized variables (also normalized to standard deviation 1). In addition to social and cognitive skill measures Deming [28] includes a noncognitive skill measure in his regression. He uses the Rotter Locus of Control and the Rosenberg Self-Esteem Scale which are also used by Heckman et al. [41]. In the following discussion we use these variables to estimate the models stated in (1.14) and (1.15).

Deming [28] uses a weighted log-linearized OLS estimator to estimate the returns of cognitive and social skills on wage and excludes respondents under the age of 23 or who are enrolled in school. The weighting is necessary as in each survey year of the NLSY79 a set of sampling weights is constructed. These weights provide the researcher with an estimate of how many individuals in the United States each respondent's answers represent. We employ these weights in our analysis as well.

Table 1.5 shows the regression results. The first column, (4), provides the results of Deming [28] estimating equation (1.14). The second and third column state the transformation partially linear estimator of equation (1.15) with and without employing γ. The fourth and fifth column state the transformation partially linear estimator of equation (1.15) with and without employing γ imposing $\lambda = 0$. This is a standard partially linear model as studied by Robinson [70] and Li [57].

For the inner smoothing of the estimations in column 2-5 we use a normal kernel with $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ defining the diagonal matrix $\boldsymbol{D}$ in $\boldsymbol{\Omega}_n$ are set equal to the componentwise standard deviations for all continuous variables. In case of the controls and fixed effects an indicator of the event that the observations have the same value is employed.

Table 1.5: *Labor Market Returns to Cognitive and Social Skills in the NLSY79*

Outcome: (log) hourly wage (in 2012 dollars)	(4)	SmoothMD with γ	SmoothMD without γ	SmoothMD with γ, $\lambda = 0$	SmoothMD without γ, $\lambda = 0$
λ	-	-0.007	-0.007	-	-
		[0.005]	[0.005]		
Cognitive skills	0.189***	-	-	-	-
	[0.007]				
Social skills	0.043***	-	-	-	-
	[0.006]				
Cognitive × Social	0.019***	-	-	-	-
	[0.006]				
Noncognitive skills	0.048***	0.047***	0.047***	0.048***	0.048***
	[0.006]	[0.004]	[0.004]	[0.004]	[0.004]
Demographics and age/ year fixed effects	X	X	X	X	X
Number of Observations	126191	126191	126191	126191	126191

*Notes: The data source is the National Longitudinal Survey of Youth 1979 cohort (NLSY79). (4) denotes the OLS regression proposed by Deming [28]. In all SmoothMD estimations, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all continuous variables and for controls and fixed effects an indicator of the event that the observations have the same value is employed. The grid for λ is $[-0.1, 0.1]$ and $s = G_n$. Cognitive skills are measured by each NLSY79 respondent's score on the Armed Forces Qualifying Test (AFQT) and are normalized to have mean 0 and standard deviation 1. The AFQT score crosswalk of Altonji et al. [4] is used. Social skill is a standardized composite of four variables (i) sociability in childhood, (ii) sociability in adulthood, (iii) participation in high school clubs and (iv) participation in team sports; see the text and Deming [28] for details on construction of the social skill measure. The noncognitive skill measure is the normalized average of the Rotter and Rosenberg scores in the NLSY. The regressions also control for race-by-gender indicator variables, age, year, census region and urbanicity. Standard errors are in brackets and are clustered at the individual level for (4). The remaining standard errors are estimated by the Eiker-White variance estimator. ***$p < .01$, **$p < .05$, *$p < .1$*

The optimization grid for λ is given by $[-0.1, 0.1]$ with step size 0.001.[3] We minimize $G_n^{-\lambda}\widehat{\mathbb{Y}}_n(\lambda)^T\, \widehat{\mathbb{B}}_n\, G_n^{-\lambda}\widehat{\mathbb{Y}}_n(\lambda)$ over the defined grid to get $\widehat{\lambda}$ and the estimates of the remaining coefficients, where $G_n = \prod_{i=1}^{n} Y_i^{1/n}$ is the geometric mean.

The results in the first column of Table 1.5 show that all by Deming [28] estimated coefficients are significantly different from 0. In the remaining four columns we cannot state parameter estimates for cognitive and social skills and the interaction of both as these variables are contained in $m(\cdot)$. However, the parameter estimates for noncognitive skills are comparable to the estimate from the first column. In addition, the estimates for λ with and without γ are equal and close to zero which would imply that a log-transformation of the dependent variable is appropriate. The estimated coefficient for noncognitive skills is significantly different from 0 in all SmoothMD estimations whereas both estimates for λ are not significantly different from 0.

In order to check if the linear specification for cognitive and social skills employed by Deming [28] is reasonable we proceed as follows. We estimate the parameters of the transformation partially linear model as stated in (1.15) to get the residuals

$$\widehat{\varepsilon}_{ijt} = T(wage_{ijt}, \widehat{\lambda}) - \widehat{\beta}\cdot NCOG_i - \boldsymbol{C}_{ijt}^T\widehat{\boldsymbol{\rho}} - \widehat{\delta}_j - \widehat{\zeta}_t.$$

We estimate now the unknown function $m(\cdot)$ by smoothing $\widehat{\varepsilon}$ with the NW estimator. In addition, we also regress $\widehat{\varepsilon}$ on COG, SS and $COG \times SS$. To see if the linear specification is appropriate we compare the MSE of the linear and nonlinear estimates. We employ a normal density kernel for the NW estimator and let $h \propto n^{-1/6}$.

Table 1.6 states the results where OLS indicates that we used the linear model to fit the residuals. The MSE of the linear and nonlinear estimates are identical no matter if we use the SmoothMD estimator with or without γ to estimate the unknown model parameters. The same holds true for the SmoothMD estimator with or without γ where $\lambda = 0$ is imposed. All results show that the linear representation of Deming [28] seems to be reasonable.

Table 1.6: *MSE of estimated nonlinear part in the transformation partially linear model*

	SmoothMD with γ		SmoothMD without γ		SmoothMD with γ, $\lambda = 0$		SmoothMD without γ, $\lambda = 0$	
	OLS	NW	OLS	LL	OLS	NW	OLS	LL
MSE	0.282	0.277	0.282	0.277	0.293	0.288	0.293	0.288
Number of Observations	126191	126191	126191	126191	126191	126191	126191	126191

Notes: For the NW estimator a normal kernel with $h \propto n^{-1/6}$ is employed. OLS indicates that the linear model is used to fit the residuals.

In a second step we include *years of completed education* as additional explanatory variable in the regression models. In one of his estimations Deming [28] controls for years of education as well. Table 1.7 states the regression results for all considered models. The first column, (5), provides the results of Deming [28] estimating equation (1.14) with *years of completed education* as control. The results show that all by Deming [28] estimated coefficients are significantly different from 0. However, the coefficients become smaller compared to the first specification. In addition, the coefficient of the interactive effect is only significant at the 10% level. In the remaining four columns the parameter estimates for noncognitive skills are comparable to the estimate from the first column. In addition, the estimates for λ with and without γ are equal and close to zero which would imply that a log-transformation of the dependent variable is appropriate. The estimated coefficient for noncognitive skills is significantly different from 0 in all SmoothMD estimations whereas both estimates for λ are not significantly different from 0.

Table 1.8 states the MSE of the estimated nonlinear part in the transformation partially linear models. The MSE of the linear and nonlinear estimates are identical no matter if we use the SmoothMD estimator with or without γ to estimate the unknown model parameters. The same holds true for the SmoothMD estimator with or without γ where $\lambda = 0$ is imposed. All results show that the linear representation of Deming [28] seems to be reasonable.

[3] We evaluated subsamples of the dataset before we conducted the final estimation. The estimated λ's in the subsamples are contained in the employed grid.

Table 1.7: *Labor Market Returns to Cognitive and Social Skills in the NLSY79 controlling for education*

Outcome: (log) hourly wage (in 2012 dollars)	(5)	SmoothMD with γ	SmoothMD without γ	SmoothMD with γ, $\lambda = 0$	SmoothMD without γ, $\lambda = 0$
λ	-	0.002	0.002	-	-
		[0.005]	[0.005]		
Cognitive skills	0.126***	-	-	-	-
	[0.008]				
Social skills	0.029***	-	-	-	-
	[0.006]				
Cognitive × Social	0.011*	-	-	-	-
	[0.006]				
Noncognitive skills	0.040***	0.037***	0.037***	0.037***	0.037***
	[0.006]	[0.004]	[0.004]	[0.004]	[0.004]
Demographics and age/ year fixed effects	X	X	X	X	X
Years of completed education	X	X	X	X	X
Number of Observations	126191	126191	126191	126191	126191

*Notes: The data source is the National Longitudinal Survey of Youth 1979 cohort (NLSY79). (5) denotes the OLS regression proposed by Deming [28]. In all SmoothMD estimations, $h \propto n^{-1/3.5}$. The components of $\mathbf{d}$ are set equal to the componentwise standard deviations for all continuous variables and for controls and fixed effects an indicator of the event that the observations have the same value is employed. The grid for λ is $[-0.1, 0.1]$ and $s = G_n$. Cognitive skills are measured by each NLSY79 respondent's score on the Armed Forces Qualifiying Test (AFQT) and are normalized to have mean 0 and standard deviation 1. The AFQT score crosswalk of Altonji et al. [4] is used. Social skill is a standardized composite of four variables (i) sociability in childhood, (ii) sociability in adulthood, (iii) participation in high school clubs and (iv) participation in team sports; see the text and Deming [28] for details on construction of the social skill measure. The noncognitive skill measure is the normalized average of the Rotter and Rosenberg scores in the NLSY. The regressions also control for race-by-gender indicator variables, age, year, census region, urbanicity and years of completed education. Standard errors are in brackets and are clustered at the individual level for (5). The remaining standard errors are estimated by the Eiker-White variance estimator. ***$p < .01$, **$p < .05$, *$p < .1$*

Before we close the section we plot the estimated labor market returns to cognitive and social skills of model (1.15) with and without controlling for years of completed education. The returns are estimated with the NW estimator employing a normal kernel with $h \propto n^{-1/6}$. Figures 1.7 and 1.8 state the results. Note that the mean was subtracted. The return increases no matter if the social or cognitive indicator is increased. However, the cognitive effect seems to be stronger. In addition, if we control for education it seems that being too social might sometimes lower the wage a bit. Nevertheless, both plots confirm that the linear model used by Deming [28] is reasonable.

Table 1.8: *MSE of estimated nonlinear part in the transformation partially linear model controlling for education*

	SmoothMD with γ		SmoothMD without γ		SmoothMD with γ, $\lambda = 0$		SmoothMD without γ, $\lambda = 0$	
	OLS	NW	OLS	LL	OLS	NW	OLS	LL
MSE	0.288	0.283	0.288	0.283	0.284	0.280	0.284	0.280
Number of Observations	126191	126191	126191	126191	126191	126191	126191	126191

Notes: For the NW estimator a normal kernel with $h \propto n^{-1/6}$ is employed. OLS indicates that the linear model is used to fit the residuals.

Figure 1.7: *Estimated Labor Market Returns to Cognitive and Social Skills in the NLSY79.*

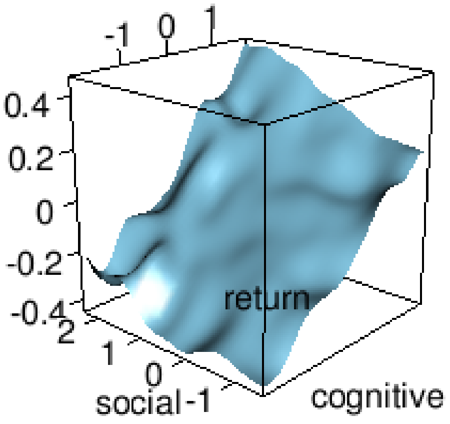
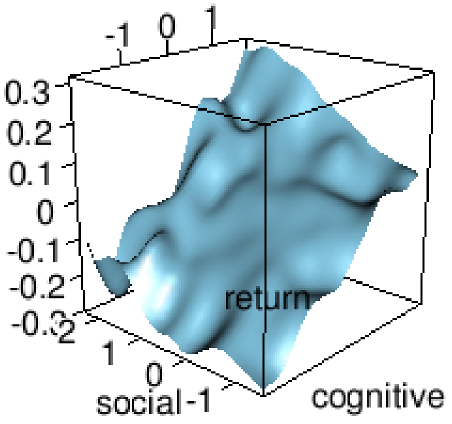


Figure 1.8: *Estimated Labor Market Returns to Cognitive and Social Skills in the NLSY79 controlling for education.*



Notes: The coefficients are estimated by the SmoothMD estimator with γ. For the NW estimator a normal kernel with $h \propto n^{-1/6}$ is employed.

1.6. Conclusion

In this paper we introduced the semiparametric partially linear model with Box-Cox transformed dependent variable. We employed the SmoothMD estimation technique to ensure identification based on conditional moment restrictions. This is new as in the literature typically either a transformation model or a semiparametric partially linear model is studied and estimation is based on GMM methods.

We established consistency as well as $\sqrt{n}$-asymptotic normality. In addition, we proposed a distance metric statistic to test the model parameters. A Monte Carlo experiment showed the usefulness of the proposed estimator in small samples as well as an application to a real data sample.

1.7. Assumptions

Assumption 1.1. *Data Generating Process*

1. The observations $\left(Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T\right)^T$, $1 \leq i \leq n$, are i.i.d. copies of $\left(Y, \boldsymbol{X}^T, \boldsymbol{Z}^T\right)^T \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$. Moreover, the response Y is bounded away from zero with probability 1, *i.e.* there exists a constant $c > 0$ such that $\mathbb{P}(Y > c) = 1$.
2. The covariate vector $\boldsymbol{Z}$ admits a bounded density with respect to the Lebesgue measure in $\mathbb{R}^q$. The covariate vector $\boldsymbol{X}$ is split in two subvectors $\boldsymbol{X}_c \in \mathbb{R}^{p_c}$ and $\boldsymbol{X}_d \in \mathbb{R}^{p_d}$ with $0 \leq p_c, p_d \leq p$ and $p_c + p_d = p$. The subvector $\boldsymbol{X}_c$ admits a bounded density with respect to the Lebesgue measure in $\mathbb{R}^{p_c}$. The subvector $\boldsymbol{X}_d$ takes values in a finite set.
3. The $(p+q)$ diagonal components of the matrix $\boldsymbol{D}$ belong to the set $\mathcal{D} = [d_L, d_U] \times \cdots \times [d_L, d_U] \subset \mathbb{R}_+^{p+q}$, with some fixed $0 < d_L < d_U < \infty$.

The assumption that the discrete components of $\boldsymbol{X}$ take values in a finite set is a technical condition that simplifies the proofs without significant restriction of the generality for the applications.

Assumption 1.2. *Identification*

1. $E\left[\|\boldsymbol{X}\|^2\right] < \infty$ and $Var\left[\boldsymbol{X} - E[\boldsymbol{X}|\boldsymbol{Z}]\right]$ has full rank.
2. The true value $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is not equal to $\mathbf{0}_p$.
3. The continuous random subvector $\boldsymbol{X}_c$ is such that, for any $\boldsymbol{b} \in \mathbb{R}^{p_c}$, $\boldsymbol{b} \neq \mathbf{0}_{p_c}$, the variable $\boldsymbol{X}_c^T \boldsymbol{b}$ is continuous with the support equal to the whole real line.
4. Whenever $\lambda \neq \lambda_0$, for any $\boldsymbol{z}$ in the support of $\boldsymbol{Z}$ and $\boldsymbol{x}_d$ in the support of the discrete subvector $\boldsymbol{X}_d$, the set of values of the map
$$\boldsymbol{x}_c \mapsto E\left[T(Y,\lambda) - T(Y,\lambda_0) \mid \boldsymbol{X}_c = \boldsymbol{x}_c, \boldsymbol{X}_d = \boldsymbol{x}_d, \boldsymbol{Z} = \boldsymbol{z}\right], \quad \boldsymbol{x}_c \in \mathbb{R}^{p_c},$$
is infinite.
5. We have that $E\left[(Y \vee 1)^{4(C_\lambda + |\lambda_0|)}\right] < \infty$, where $\max(|\lambda_{min}|, \lambda_{max}) < C_\lambda$.

Note that $Var\left[(\boldsymbol{X}^T, \boldsymbol{Z}^T)^T\right]$ has necessarily full rank, by Assumption 1.2.1 and the fact that $\boldsymbol{Z}$ admits a density. Indeed, for any $\boldsymbol{u} \in \mathbb{R}^p$ and $\boldsymbol{v} \in \mathbb{R}^q$ such that $(\boldsymbol{u}^T, \boldsymbol{v}^T)^T \neq \mathbf{0}_{p+q}$, we can write

$$Var\left[\boldsymbol{u}^T\boldsymbol{X} + \boldsymbol{v}^T\boldsymbol{Z}\right] = E\left[Var\left[\boldsymbol{u}^T(\boldsymbol{X} - E\left[\boldsymbol{X} \mid \boldsymbol{Z}\right]) \mid \boldsymbol{Z}\right]\right] + Var\left[\boldsymbol{u}^T E\left[\boldsymbol{X} \mid \boldsymbol{Z}\right] + \boldsymbol{v}^T\boldsymbol{Z}\right].$$

If $\boldsymbol{u} \neq \mathbf{0}_p$,

$$\begin{aligned} Var\left[\boldsymbol{u}^T\boldsymbol{X} + \boldsymbol{v}^T\boldsymbol{Z}\right] &\geq E\left[Var\left[\boldsymbol{u}^T(\boldsymbol{X} - E\left[\boldsymbol{X} \mid \boldsymbol{Z}\right]) \mid \boldsymbol{Z}\right]\right] \\ &= \boldsymbol{u}^T E\left[Var\left[\boldsymbol{X} - E\left[\boldsymbol{X} \mid \boldsymbol{Z}\right] \mid \boldsymbol{Z}\right]\right]\boldsymbol{u} = \boldsymbol{u}^T Var\left[\boldsymbol{X} - E\left[\boldsymbol{X} \mid \boldsymbol{Z}\right]\right]\boldsymbol{u} > 0, \end{aligned}$$

where the last inequality is guaranteed by Assumption 1.2.1. When $\boldsymbol{u} = \mathbf{0}_p$, we obtain

$$Var\left[\boldsymbol{u}^T\boldsymbol{X} + \boldsymbol{v}^T\boldsymbol{Z}\right] = Var\left[\boldsymbol{v}^T\boldsymbol{Z}\right] = \boldsymbol{v}^T Var\left[\boldsymbol{Z}\right]\boldsymbol{v} > 0,$$

where the last inequality holds because $\boldsymbol{v} \neq \mathbf{0}_q$ and $Var\left[\boldsymbol{Z}\right]$ has necessarily full rank provided $\boldsymbol{Z}$ admits a density.

Assumption 1.3. *Consistency*

1. The kernel $K(\cdot)$ is the product of q univariate kernel functions $\widetilde{K}$ of bounded variation. Moreover, $\widetilde{K}$ is a symmetric function with integral equal to one and $\int_{\mathbb{R}} t^2\widetilde{K}(t)dt < \infty$.
2. The functions $f_z(\cdot)$, $(mf_z)(\cdot)$, $E[\|\boldsymbol{X}\|^2 \mid \boldsymbol{Z} = \cdot\,]f_z(\cdot)$ and $\sup_{\lambda\in\Lambda}(\partial^2/\partial\lambda^2)E[T(Y,\lambda) \mid \boldsymbol{Z} = \cdot\,]f_z(\cdot)$ have Hölder continuous partial derivatives of order four.
3. The bandwidth h belongs to a range $\mathcal{H}_{c,n} = [c_{min}n^{-\alpha}, c_{max}n^{-\alpha}]$, with $0 < \alpha < 1/q$ and c_{min}, c_{max} positive constants.
4. It holds true that $E\left[\|\boldsymbol{X}\|^4\right] < \infty$, $E\left[\|\boldsymbol{Z}\|\right] < \infty$ and $E\left[\left((Y \vee 1)^{C_\lambda} \log{(Y \vee e)}^4\right)^4\right] < \infty$.

Assumption 1.4. *Asymptotic Normality*

1. $Var\left[\frac{\partial}{\partial\lambda}T(Y,\lambda_0)\right] > 0$.
2. The bandwidth h belongs to a range $\mathcal{H}_{sc,n} = [c_{min}n^{-\alpha}, c_{max}n^{-\alpha}]$, with $\alpha \in (1/4, 1/q)$ and c_{min}, c_{max} positive constants.
3. $E\left[\varepsilon^2 \mid \boldsymbol{X}, \boldsymbol{Z}\right] = \sigma^2(\boldsymbol{X}, \boldsymbol{Z})$ is in $L^1 \cap L^2$.
4. $E\left[\varepsilon^4\right] < \infty$ as well as $E\left[m(\boldsymbol{Z})^4\right] < \infty$.

Appendix

Appendix A: Main proofs

Proof of Lemma 1.1.

It is quite clear that, by construction, $\gamma(\lambda_0) = 0$ and $\boldsymbol{\beta}(\lambda_0) = \boldsymbol{\beta}_0$. In order to ensure global identification, we extend the proof in Shin [74]. For any $(\gamma, \lambda, \boldsymbol{\beta}^T)^T$ we have that

$$\begin{aligned}\mathbb{P}\Big(& E\left[(T(Y,\lambda) - E[T(Y,\lambda) \mid \boldsymbol{Z}])\, f_z(\boldsymbol{Z}) - \gamma - (\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}])^T \boldsymbol{\beta} f_z(\boldsymbol{Z}) \mid \boldsymbol{X}, \boldsymbol{Z}\right] = 0\Big) \\ &= \mathbb{P}\Big(E\left[T(Y,\lambda) - T(Y,\lambda_0) | \boldsymbol{X}, \boldsymbol{Z}\right] - \boldsymbol{X}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = E[T(Y,\lambda) \mid \boldsymbol{Z}] - E[T(Y,\lambda_0)|\boldsymbol{Z}] \\ &\qquad\qquad + \gamma f_z^{-1}(\boldsymbol{Z}) - E[\boldsymbol{X} \mid \boldsymbol{Z}]^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\Big).\end{aligned}$$

Hence, it suffices to prove that the last probability could not be equal to 1 when $(\gamma, \boldsymbol{\theta}^T)^T \neq (0, \boldsymbol{\theta}_0^T)^T$. Note that

$$E[T(Y,\lambda) \mid \boldsymbol{Z}] - E[T(Y,\lambda_0) \mid \boldsymbol{Z}] + \gamma f_z^{-1}(\boldsymbol{Z}) - E[\boldsymbol{X} \mid \boldsymbol{Z}]^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0),$$

does not depend on $\boldsymbol{X}$ anymore but only on $\boldsymbol{Z}$.

If $\lambda = \lambda_0$ the result follows immediately from the full rank condition in Assumption 1.2.1. Indeed, by the variance decomposition formula and Assumption 1.2.1, for any $\boldsymbol{a} \in \mathbb{R}^p$, $\boldsymbol{a} \neq \boldsymbol{0}_p$,

$$\boldsymbol{a}^T Var\left[\boldsymbol{X} - E[\boldsymbol{X}|\boldsymbol{Z}]\right] \boldsymbol{a} = E\left[Var\left[\boldsymbol{a}^T(\boldsymbol{X} - E[\boldsymbol{X}|\boldsymbol{Z}]) \mid \boldsymbol{Z}\right]\right] > 0.$$

This implies

$$\begin{aligned}\boldsymbol{a}^T Var\left[f_z(\boldsymbol{Z})\,(\boldsymbol{X} - E[\boldsymbol{X}|\boldsymbol{Z}])\right] \boldsymbol{a} &= E\left[f_z^2(\boldsymbol{Z})\boldsymbol{a}^T Var\left[(\boldsymbol{X} - E[\boldsymbol{X}|\boldsymbol{Z}]) \mid \boldsymbol{Z}\right] \boldsymbol{a}^T\right] \\ &= E\left[f_z^2(\boldsymbol{Z}) Var\left[\boldsymbol{a}^T\left(\boldsymbol{X} - E[\boldsymbol{X}|\boldsymbol{Z}]\right) \mid \boldsymbol{Z}\right]\right] > 0.\end{aligned}$$

Thus, $f_z(\boldsymbol{Z})(\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}^T)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ cannot be equal to a constant almost surely, as is necessarily the case when $\lambda = \lambda_0$. Next, consider the case $\lambda \neq \lambda_0$. Without loss of generality, we could assume that $\lambda > \lambda_0$.

1. Consider the case $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ and let introduce the event notation

$$\begin{aligned}\mathcal{E} = \big\{ & E\left[T(Y,\lambda) - T(Y,\lambda_0) \mid \boldsymbol{X}, \boldsymbol{Z}\right] - \boldsymbol{X}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\ &= E[T(Y,\lambda) \mid \boldsymbol{Z}] - E[T(Y,\lambda_0)|\boldsymbol{Z}] + \gamma f_z^{-1}(\boldsymbol{Z}) - E[\boldsymbol{X} \mid \boldsymbol{Z}]^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\big\}.\end{aligned}$$

   Taking conditional expectation given $\boldsymbol{Z}$ on both sides, we deduce that necessarily $\gamma = 0$. Thus it suffices to investigate the probability of the event

$$\begin{aligned}\mathcal{E}' = \big\{ & E\left[T(Y,\lambda) - T(Y,\lambda_0) \mid \boldsymbol{X}, \boldsymbol{Z}\right] - \boldsymbol{X}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\ &= E[T(Y,\lambda) \mid \boldsymbol{Z}] - E[T(Y,\lambda_0)|\boldsymbol{Z}] - E[\boldsymbol{X} \mid \boldsymbol{Z}]^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\big\}.\end{aligned}$$

   Note that the right hand side equality does not depend on $\boldsymbol{X}$. We distinguish two sub-cases. First, the case where the components of $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ corresponding to $\boldsymbol{X}_c$ are equal to zero. Thus, the linear combination $\boldsymbol{X}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ does not include any of the continuous components of $\boldsymbol{X}$. In this case, for any value of $\boldsymbol{Z}$, the support of $\boldsymbol{X}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ is finite and independent of the value of $\boldsymbol{X}_c$. Then Assumption 1.2.4 guarantees that the probability of the event $\mathcal{E}'$ could not be equal to 1. Next, consider the case where $\boldsymbol{X}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ includes continuous components of $\boldsymbol{X}$. In this case, by Assumption 1.2.3, the support of the variable $\boldsymbol{X}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ is the whole real line and, by the monotonicity of $\lambda \mapsto T(y;\lambda)$ for each value $y > 0$, $E\left[T(Y,\lambda) - T(Y,\lambda_0)|\boldsymbol{X}, \boldsymbol{Z}\right] \geq 0$ almost surely, the statement follows again.

2. Consider the case $\gamma \neq 0$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. In this case

$$f_z(\boldsymbol{Z})\left(E\left[T(Y,\lambda) - T(Y,\lambda_0) \mid \boldsymbol{X}, \boldsymbol{Z}\right] - E\left[T(Y,\lambda) - T(Y,\lambda_0) \mid \boldsymbol{Z}\right]\right) = \gamma.$$

   Taking expectation on both sides, we deduce that necessarily $\gamma = 0$.

3. Consider the case $(\gamma, \boldsymbol{\beta}^T)^T = (0, \boldsymbol{\beta}_0^T)^T$. Then necessarily

$$E\Bigg[T(Y,\lambda) - T(Y,\lambda_0) | \boldsymbol{X}, \boldsymbol{Z}\Bigg] = E[T(Y,\lambda)|\boldsymbol{Z}] - E[T(Y,\lambda_0)|\boldsymbol{Z}] \quad \text{almost surely.}$$

   Once again the right hand side does not depend on $\boldsymbol{X}$, and thus the probability of the event $\mathcal{E}$ could not be equal to 1 because of Assumption 1.2.3.

Therefore, the first statement follows. Consider now the second statement of the Lemma. First, note that the maps $\lambda \mapsto \gamma(\lambda)$, $\lambda \mapsto \beta(\lambda)$ and

$$\lambda \mapsto E\left[g\left(\boldsymbol{U}_1; (\lambda, \boldsymbol{\beta}(\lambda)^T)^T, \gamma(\lambda), \boldsymbol{\eta}_1\right) g\left(\boldsymbol{U}_2; (\lambda, \boldsymbol{\beta}(\lambda)^T)^T, \gamma(\lambda), \boldsymbol{\eta}_2\right) \exp\left\{-(\boldsymbol{W}_1 - \boldsymbol{W}_2)^T \boldsymbol{D}(\boldsymbol{W}_1 - \boldsymbol{W}_2)\right\}\right],$$

$\lambda \in \Lambda$ are continuous. Indeed, it will become clear from the following that $\inf_{\boldsymbol{d}\in\mathcal{D}} E[\omega(\boldsymbol{W}_1 - \boldsymbol{W}_2)] > 0$. Next, we have

$$\gamma(\lambda) = \frac{1}{E[\omega(\boldsymbol{W}_1 - \boldsymbol{W}_2)]} E\left[\left(T(Y_1,\lambda) - E[T(Y_1,\lambda) \mid \boldsymbol{Z}_1] - (\boldsymbol{X}_1 - E[\boldsymbol{X}_1 \mid \boldsymbol{Z}_1])^T \boldsymbol{\beta}\right) f_z(\boldsymbol{Z}_1) \omega(\boldsymbol{W}_1 - \boldsymbol{W}_2)\right],$$

and

$$\begin{aligned}\boldsymbol{\beta}(\lambda) = E&\left[(\boldsymbol{X}_1 - E[\boldsymbol{X}_1 \mid \boldsymbol{Z}_1])(\boldsymbol{X}_2 - E[\boldsymbol{X}_2 \mid \boldsymbol{Z}_2])^T f_z(\boldsymbol{Z}_1) f_z(\boldsymbol{Z}_2) \omega(\boldsymbol{W}_1 - \boldsymbol{W}_2)\right]^{-1} \\ &E\left[(\boldsymbol{X}_1 - E[\boldsymbol{X}_1 \mid \boldsymbol{Z}_1]) f_z(\boldsymbol{Z}_1) \left((T(Y_2,\lambda) - E[T(Y_2,\lambda) \mid \boldsymbol{Z}_2]) f_z(\boldsymbol{Z}_2) - \gamma(\lambda)\right) \omega(\boldsymbol{W}_1 - \boldsymbol{W}_2)\right],\end{aligned}$$

which are clearly continuous. The continuity of the third map follows by Lebesgue Dominated Convergence Theorem. Finally, by the same inverse Fourier Transform argument used by Lavergne and Patilea [56] we get that

$$\begin{aligned}E\left[g\left(\boldsymbol{U}_1; \boldsymbol{\theta}, \gamma, \boldsymbol{\eta}_1\right) g\left(\boldsymbol{U}_2; \boldsymbol{\theta}, \gamma, \boldsymbol{\eta}_2\right) \omega(\boldsymbol{W}_1 - \boldsymbol{W}_2)\right] &= \frac{\pi^{-(p+q)/2}}{\sqrt{d_1 \cdots d_{p+q}}} \\ &\times \int_{\mathbb{R}^{p+q}} \left|E\left[E[g\left(\boldsymbol{U}; \boldsymbol{\theta}, \gamma, \boldsymbol{\eta}\right) \mid \boldsymbol{X}, \boldsymbol{Z}] \exp\left\{2i\boldsymbol{w}^T \left(\boldsymbol{X}^T, \boldsymbol{Z}^T\right)^T\right\}\right]\right|^2 \exp\left\{-\boldsymbol{w}^T \boldsymbol{D}^{-1} \boldsymbol{w}\right\} d\boldsymbol{w}.\end{aligned}$$

Next, let $\boldsymbol{\theta}(\lambda) = (\lambda, \boldsymbol{\beta}(\lambda)^T)^T$. For any $\boldsymbol{x}, \boldsymbol{z}$, the map $\lambda \mapsto E[g\left(\boldsymbol{U}; \boldsymbol{\theta}(\lambda), \gamma(\lambda), \boldsymbol{\eta}\right) \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}]$ is continuous. By Lebesgue Dominated Convergence Theorem, the map

$$\lambda \mapsto \int_{\mathbb{R}^{p+q}} \left|E\left[E[g\left(\boldsymbol{U}; \boldsymbol{\theta}(\lambda), \gamma(\lambda), \boldsymbol{\eta}\right) \mid \boldsymbol{X}, \boldsymbol{Z}] \exp\left\{2i\boldsymbol{w}^T \left(\boldsymbol{X}^T, \boldsymbol{Z}^T\right)^T\right\}\right]\right|^2 \exp\left\{-\boldsymbol{w}^T \mathrm{diag}(d_L, \ldots, d_L)^{-1} \boldsymbol{w}\right\} d\boldsymbol{w},$$

is continuous, and thus attains its minimum on the compact set $\Lambda \setminus [\lambda_0 - \varepsilon, \lambda_0 + \varepsilon]$. The minimum value is necessarily positive. Since $(d_1 \cdots d_{p+q})^{-1/2} \exp\left\{-\boldsymbol{w}^T \boldsymbol{D} \boldsymbol{w}\right\} \geq d_U^{-(p+q)/2} \exp\left\{-\boldsymbol{w}^T \mathrm{diag}(d_L, \ldots, d_L)^{-1} \boldsymbol{w}\right\}$, the last statement in the Lemma follows.

□

Proof of Lemma 1.2.

1. First, we note that for any $\boldsymbol{u} \in \mathbb{R}^p$ and $\boldsymbol{v} \in \mathbb{R}^q$ such that $(\boldsymbol{u}^T, \boldsymbol{v}^T)^T \neq \boldsymbol{0}_{p+q}$, and any $c \in \mathbb{R}$,

$$\mathbb{P}\left(\boldsymbol{u}^T \boldsymbol{X} + \boldsymbol{v}^T \boldsymbol{Z} = c\right) = 0. \tag{1.16}$$

   This is a consequence of the fact that $Var\left[(\boldsymbol{X}^T, \boldsymbol{Z}^T)^T\right]$ has full rank, by Assumption 1.2. Given a sample $\left(\boldsymbol{X}_1^T, \boldsymbol{Z}_1^T\right)^T, \ldots, \left(\boldsymbol{X}_n^T, \boldsymbol{Z}_n^T\right)^T$, and a vector $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$, using the inverse Fourier Transform, we could write

$$\boldsymbol{a}^T \boldsymbol{\Omega}_n \boldsymbol{a} = \frac{\pi^{-(p+q)/2}}{\sqrt{d_1 \cdots d_{p+q}}} \int_{\mathbb{R}^{p+q}} \left|\sum_{j=1}^n a_j \exp\left\{2i\boldsymbol{w}^T \left(\boldsymbol{X}_j^T, \boldsymbol{Z}_j^T\right)^T\right\}\right|^2 \exp\left\{-\boldsymbol{w}^T \boldsymbol{D}^{-1} \boldsymbol{w}\right\} d\boldsymbol{w},$$

   where $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_{p+q})$ with $d_1, \ldots, d_{p+q} \in [d_L, d_U]$; see Assumption 1.1.3. Then, necessarily

$$\sum_{j=1}^n a_j \exp\left\{2i\boldsymbol{w}^T \left(\boldsymbol{X}_j^T, \boldsymbol{Z}_j^T\right)^T\right\} = 0, \qquad \forall \boldsymbol{w} \in \mathbb{R}^{p+q}. \tag{1.17}$$

Equation (1.16) indicates that, with probability 1, equation (1.17) admits the unique solution $\boldsymbol{a} = \mathbf{0}_n$ $\forall \boldsymbol{w} \in \mathbb{R}^{p+q}$. This means that, with probability 1, the matrix $\mathbf{\Omega}_n$ is positive definite. Next, we use the following Cauchy-Schwarz[4] inequality for matrices: let $\boldsymbol{A} \in \mathbb{R}^{n\times p_1}$ such that $\boldsymbol{A}^T\boldsymbol{A}$ is invertible and let $\boldsymbol{B} \in \mathbb{R}^{n\times p_2}$. Then

$$\boldsymbol{B}^T\boldsymbol{B} - \boldsymbol{B}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{B} \text{ is positive semi-definite.}$$

Moreover, the equality $\boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{B}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{B}$ is equivalent to the relationship $\boldsymbol{B} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{B}$. For any non null vector $\boldsymbol{u} \in \mathbb{R}^p$, taking

$$\boldsymbol{B} = \mathbf{\Omega}_n^{1/2} \quad \text{ and } \boldsymbol{A} = \mathbf{\Omega}_n^{1/2}\mathbf{1}_n,$$

we deduce that $\mathbb{D}_n$ is positive semi-definite and thus $\boldsymbol{u}^T\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\boldsymbol{u} \geq 0$. (Herein, $\mathbf{\Omega}_n^{1/2}$ is the positive definite square root of $\mathbf{\Omega}_n$.) Meanwhile, by elementary matrix algebra, we deduce that, for any $\boldsymbol{a} \in \mathbb{R}^n$,

$$\boldsymbol{a}^T\mathbb{D}_n\boldsymbol{a} = 0 \quad \Leftrightarrow \quad \left(\boldsymbol{a}^T\mathbf{\Omega}_n\boldsymbol{a}^T\right)\left(\mathbf{1}_n^T\mathbf{\Omega}_n\mathbf{1}_n\right) = \left(\boldsymbol{a}^T\mathbf{\Omega}_n\mathbf{1}_n\right)^2.$$

Then, the Cauchy-Schwarz inequality indicates that $\boldsymbol{a}^T\mathbb{D}_n\boldsymbol{a} = 0$ if and only if $\boldsymbol{a} = a\mathbf{1}_n$ for some scalar $a \neq 0$. Thus, $\boldsymbol{u}^T\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\boldsymbol{u} = 0$ if and only if $\widehat{\mathbb{X}}_n\boldsymbol{u} = a\mathbf{1}_n$ for some $a \neq 0$. By Assumption 1.2, the probability of such an event is equal to zero. Thus, $\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n$ is almost surely invertible. Note that we could also write

$$\mathbb{D}_n = \left[\boldsymbol{I}_{n\times n} - \frac{1}{\mathbf{1}_n^T\mathbf{\Omega}_n\mathbf{1}_n}\mathbf{1}_n\mathbf{1}_n^T\mathbf{\Omega}_n\right]^T \mathbf{\Omega}_n \left[\boldsymbol{I}_{n\times n} - \frac{1}{\mathbf{1}_n^T\mathbf{\Omega}_n\mathbf{1}_n}\mathbf{1}_n\mathbf{1}_n^T\mathbf{\Omega}_n\right],$$

and deduce the positive semi-definiteness of $\mathbb{D}_n$ from the positive definiteness of $\mathbf{\Omega}_n$.

2. We could rewrite $\widehat{\mathbb{B}}_n$ under the form

$$\widehat{\mathbb{B}}_n = \left[\boldsymbol{I}_{n\times n} - \widehat{\mathbb{X}}_n\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\right]^T \mathbb{D}_n \left[\boldsymbol{I}_{n\times n} - \widehat{\mathbb{X}}_n\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\right],$$

and deduce its positive semi-definiteness from the positive definiteness of $\mathbb{D}_n$.

□

Proof of Theorem 1.1.

Let

$$\widehat{M_n}(\lambda) = n^{-2}\widehat{\mathbb{Y}}_n(\lambda)^T\ \widehat{\mathbb{B}}_n\ \widehat{\mathbb{Y}}_n(\lambda),$$

such that

$$s^{-2\lambda}\widehat{M_n}(\lambda) = n^{-2}s^{-\lambda}\widehat{\mathbb{Y}}_n(\lambda)^T\ \widehat{\mathbb{B}}_n\ s^{-\lambda}\widehat{\mathbb{Y}}_n(\lambda),$$

and, thus, $\widehat{\lambda} = \arg\min_{\lambda\in\Lambda} s^{-2\lambda}\widehat{M_n}(\lambda)$. Next, let

$$M_n(\lambda) = n^{-2}\mathbb{Y}_n(\lambda)^T\ \mathbb{B}_n\ \mathbb{Y}_n(\lambda).$$

By construction,

$$M_n(\lambda) = Q_n((\lambda, \boldsymbol{\beta}_n(\lambda)^T)^T, \gamma_n(\lambda)),$$

where

$$Q_n((\lambda, \boldsymbol{\beta}^T)^T, \gamma) = n^{-2}\left(\mathbb{Y}_n(\lambda) - \gamma\mathbf{1}_n - \mathbb{X}_n\boldsymbol{\beta}\right)^T \mathbf{\Omega}_n \left(\mathbb{Y}_n(\lambda) - \gamma\mathbf{1}_n - \mathbb{X}_n\boldsymbol{\beta}\right),$$

and

$$\gamma_n(\lambda) = \frac{1}{\mathbf{1}_n^T\mathbf{\Omega}_n\mathbf{1}_n}\mathbf{1}_n^T\mathbf{\Omega}_n\left(\mathbb{Y}_n(\lambda) - \mathbb{X}_n\boldsymbol{\beta}(\lambda)\right),$$

and

$$\boldsymbol{\beta}_n(\lambda) = \left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\mathbb{X}_n^T\mathbb{D}_n\mathbb{Y}_n(\lambda).$$

Let

$$Q((\lambda, \boldsymbol{\beta}^T)^T, \gamma) = E\left[Q_n((\lambda, \boldsymbol{\beta}^T)^T, \gamma)\right], \qquad \lambda \in \Lambda, \boldsymbol{\beta} \in \mathbb{R}^p, \gamma \in \mathbb{R}.$$

[4] A similar so-called Cauchy-Schwarz inequality was proposed by Lavergne [55]. To justify the statement, it suffices to notice that $\boldsymbol{B}^T\boldsymbol{B} - \boldsymbol{B}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{B} = \mathbf{\Gamma}^T\mathbf{\Gamma}$ with $\mathbf{\Gamma} = \boldsymbol{B} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{B}$.

Next, let $c > 0$ be a lower bound of the support of Y. Then, necessarily $c < s_0$ and we could work on the event $c \leq \inf S_n$, that is s stays away from zero. In order to prove the uniform consistency it will suffice to prove

$$\sup_{h \in \mathcal{H}_{c,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} \sup_{\lambda \in \Lambda} \left| \widehat{M}_n(\lambda) - M_n(\lambda) \right| = o_{\mathbb{P}}(1), \tag{1.18}$$

$$\sup_{h \in \mathcal{H}_{c,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} \sup_{\lambda \in \Lambda} \left| Q_n((\lambda, \boldsymbol{\beta}_n(\lambda)^T)^T, \gamma_n(\lambda)) - Q_n((\lambda, \widehat{\boldsymbol{\beta}}(\lambda)^T)^T, \widehat{\gamma}(\lambda)) \right| = o_{\mathbb{P}}(1), \tag{1.19}$$

$$\sup_{\boldsymbol{d} \in \mathcal{D}} \sup_{\lambda \in \Lambda} \sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \sup_{\gamma \in \mathbb{R}} \left| Q_n((\lambda, \boldsymbol{\beta}^T)^T, \gamma) - Q((\lambda, \boldsymbol{\beta}^T)^T, \gamma) \right| = o_{\mathbb{P}}(1), \tag{1.20}$$

and to show that λ_0 is a uniformly well-separated minimum value of $\lambda \mapsto s_0^{-2\lambda} Q((\lambda, \boldsymbol{\beta}(\lambda)^T)^T, \gamma(\lambda))$, that is for any $\varepsilon > 0$,

$$\inf_{\lambda \in \Lambda, |\lambda - \lambda_0| \geq \varepsilon} \inf_{\boldsymbol{d} \in \mathcal{D}} s_0^{-2\lambda} Q((\lambda, \boldsymbol{\beta}(\lambda)^T)^T, \gamma(\lambda)) > 0, \tag{1.21}$$

with $(\gamma(\lambda), \boldsymbol{\beta}(\lambda)^T)^T$ defined in equation (1.5).

For the uniform convergence (1.18), we first decompose

$$\begin{aligned}
\left| \widehat{M}_n(\lambda) - M_n(\lambda) \right| &\leq \left| n^{-1} \mathbb{Y}_n(\lambda)^T \left(\widehat{\mathbb{B}}_n - \mathbb{B}_n \right) n^{-1} \mathbb{Y}_n(\lambda) \right| + 2 \left| n^{-1} \left[\widehat{\mathbb{Y}}_n(\lambda) - \mathbb{Y}_n(\lambda) \right]^T \widehat{\mathbb{B}}_n n^{-1} \mathbb{Y}_n(\lambda) \right| \\
&\quad + n^{-1} \left[\widehat{\mathbb{Y}}_n(\lambda) - \mathbb{Y}_n(\lambda) \right]^T \widehat{\mathbb{B}}_n n^{-1} \left[\widehat{\mathbb{Y}}_n(\lambda) - \mathbb{Y}_n(\lambda) \right] \\
&\leq \left\| \widehat{\mathbb{B}}_n \right\|_{\mathrm{Sp}} \left(2 \left\| n^{-1} \left[\widehat{\mathbb{Y}}_n(\lambda) - \mathbb{Y}_n(\lambda) \right] \right\| \left\| n^{-1} \mathbb{Y}_n(\lambda) \right\| + \left\| n^{-1} \left[\widehat{\mathbb{Y}}_n(\lambda) - \mathbb{Y}_n(\lambda) \right] \right\|^2 \right) \\
&\quad + \left\| n^{-1} \mathbb{Y}_n(\lambda) \right\| \left\| \widehat{\mathbb{B}}_n - \mathbb{B}_n \right\|_{\mathrm{Sp}} \left\| n^{-1} \mathbb{Y}_n(\lambda) \right\|.
\end{aligned}$$

Next, from Lemma 1.3 and 1.7 we obtain that

$$\sup_{h \in \mathcal{H}_{c,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} \left\| \widehat{\mathbb{B}}_n \right\|_{\mathrm{Sp}} = O_{\mathbb{P}}(n) \qquad \text{and} \qquad \sup_{h \in \mathcal{H}_{c,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} \left\| \widehat{\mathbb{B}}_n - \mathbb{B}_n \right\|_{\mathrm{Sp}} = o_{\mathbb{P}}(n).$$

Moreover, by Lemma 1.12

$$\sup_{\lambda \in \Lambda} \left\| n^{-1} \mathbb{Y}_n(\lambda) \right\| = O_{\mathbb{P}}(n^{-1/2}),$$

and by Lemma 1.13

$$\sup_{h \in \mathcal{H}_{c,n}} \sup_{\lambda \in \Lambda} \left\| n^{-1} \left[\widehat{\mathbb{Y}}_n(\lambda) - \mathbb{Y}_n(\lambda) \right] \right\| = o_{\mathbb{P}}(n^{-1/2}).$$

Therefore, the uniform convergence (1.18) follows. Similarly, by a suitable decomposition and elementary matrix algebra calculations

$$\begin{aligned}
Q_n((\lambda, \boldsymbol{\beta}_n(\lambda)^T)^T, \gamma_n(\lambda)) &- Q_n((\lambda, \widehat{\boldsymbol{\beta}}(\lambda)^T)^T, \widehat{\gamma}(\lambda)) \\
&= \left(\gamma_n^2(\lambda) - \widehat{\gamma}^2(\lambda) \right) n^{-2} \mathbf{1}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n \\
&\quad + n^{-2} \left(\boldsymbol{\beta}_n(\lambda)^T \mathbb{X}_n^T \boldsymbol{\Omega}_n \mathbb{X}_n \boldsymbol{\beta}_n(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda)^T \mathbb{X}_n^T \boldsymbol{\Omega}_n \mathbb{X}_n \widehat{\boldsymbol{\beta}}(\lambda) \right) \\
&\quad + 2n^{-2} \left(\gamma_n(\lambda) \mathbf{1}_n^T \boldsymbol{\Omega}_n \mathbb{X}_n \boldsymbol{\beta}_n(\lambda) - \widehat{\gamma}(\lambda) \mathbf{1}_n^T \boldsymbol{\Omega}_n \mathbb{X}_n \widehat{\boldsymbol{\beta}}(\lambda) \right) \\
&\quad - 2n^{-2} \left(\gamma_n(\lambda) \mathbf{1}_n^T \boldsymbol{\Omega}_n \mathbb{Y}_n(\lambda) - \widehat{\gamma}(\lambda) \mathbf{1}_n^T \boldsymbol{\Omega}_n \mathbb{Y}_n(\lambda) \right) \\
&\quad - 2n^{-2} \left(\mathbb{Y}_n(\lambda)^T \boldsymbol{\Omega}_n \mathbb{X}_n \boldsymbol{\beta}_n(\lambda) - \mathbb{Y}_n(\lambda)^T \boldsymbol{\Omega}_n \mathbb{X}_n \widehat{\boldsymbol{\beta}}(\lambda) \right) \\
&= O_{\mathbb{P}}(1) \times \sup_{h \in \mathcal{H}_{c,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} \sup_{\lambda \in \Lambda} \left(\left\| \widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}_n(\lambda) \right\| + |\widehat{\gamma}(\lambda) - \gamma_n(\lambda)| \right).
\end{aligned}$$

By the results of Sherman [72], the rate $O_{\mathbb{P}}(1)$ is uniform with respect to $\boldsymbol{d}$ and λ. See also below for an example of application of the results in Sherman [72]. The uniform convergence of $\left\| \widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}_n(\lambda) \right\|$ and $\widehat{\gamma}(\lambda) - \gamma_n(\lambda)$ follows by the same type of matrix algebra calculations and uniform rates of convergence for $U-$processes. Thus, the uniform convergence (1.19) holds true.

Next, by the properties of Euclidean families, see Nolan and Pollard [67] and Sherman [72], the families of functions

$$\{g\left(\boldsymbol{u}_1;\boldsymbol{\theta},\gamma,\boldsymbol{\eta}_1\right)g\left(\boldsymbol{u}_2;\boldsymbol{\theta},\gamma,\boldsymbol{\eta}_2\right)\exp\left\{-(\boldsymbol{w}_1-\boldsymbol{w}_2)^T\boldsymbol{D}(\boldsymbol{w}_1-\boldsymbol{w}_2)\right\}:\boldsymbol{\theta}=(\lambda,\boldsymbol{\beta}^T)^T\in\Lambda\times\mathbb{R}^p,\gamma\in\mathbb{R},\boldsymbol{d}\in\mathcal{D}\},$$

and $\{g^2\left(\boldsymbol{u};\boldsymbol{\theta},\gamma,\boldsymbol{\eta}\right):\boldsymbol{\theta}\in\Lambda\times\mathbb{R}^p,\gamma\in\mathbb{R}\}$ are Euclidean for a squared envelope. Thus, decomposing $Q_n((\lambda,\boldsymbol{\beta}^T)^T,\gamma)$ is a $U-$process plus the sum of the diagonal terms, and using Corollary 4 of Sherman [72], the uniform convergence (1.20) holds true.

By construction, condition (1.6) in Lemma 1.1 is equivalent with

$$\inf_{\lambda\in\Lambda,|\lambda-\lambda_0|\geq\varepsilon}\inf_{\boldsymbol{d}\in\mathcal{D}}s_0^{-2\lambda}\left(Q((\lambda,\boldsymbol{\beta}(\lambda)^T)^T,\gamma(\lambda))-n^{-1}E\left[g^2\left(\boldsymbol{U};(\lambda,\boldsymbol{\beta}(\lambda)^T)^T,\gamma(\lambda),\boldsymbol{\eta}\right)\right]\right)>0.$$

Since the family $\{g^2\left(\boldsymbol{u};\boldsymbol{\theta},\gamma,\boldsymbol{\eta}\right):\boldsymbol{\theta}\in\Lambda\times\mathbb{R}^p,\gamma\in\mathbb{R}\}$ has an integrable envelope, the expectation in the last display is finite. Thus, we deduce (1.21) and λ_0 is a uniformly well-separated minimum.

Finally, to derive the uniform consistency of $\widehat{\lambda}$, we adapt the steps in the proof of Theorem 5.7 of Van der Vaart [78]. First, for any sequence $s_n\in S_n$, $n\geq1$, and $\widehat{\lambda}=\widehat{\lambda}(s_n)$ defined as in equation (1.9),

$$\begin{aligned}0\leq s_n^{-2\widehat{\lambda}}\widehat{M}_n(\widehat{\lambda})&\leq s_n^{-2\lambda_0}\widehat{M}_n(\lambda_0)\\&=s_0^{-2\lambda_0}M_n(\lambda_0)+s_n^{-2\lambda_0}\left(\widehat{M}_n(\lambda_0)-M_n(\lambda_0)\right)+\left(s_n^{-2\lambda_0}-s_0^{-2\lambda_0}\right)M_n(\lambda_0)\\&=s_0^{-2\lambda_0}Q_n((\lambda_0,\boldsymbol{\beta}_n(\lambda_0)^T)^T,\gamma_n(\lambda_0))+o_{\mathbb{P}}(1)\\&\leq s_0^{-2\lambda_0}Q_n((\lambda_0,\boldsymbol{\beta}(\lambda_0)^T)^T,\gamma(\lambda_0))+o_{\mathbb{P}}(1)\\&=s_0^{-2\lambda_0}Q((\lambda_0,\boldsymbol{\beta}(\lambda_0)^T)^T,\gamma(\lambda_0))+o_{\mathbb{P}}(1)=o_{\mathbb{P}}(1),\end{aligned}\tag{1.22}$$

uniformly with respect to h and $\boldsymbol{d}$. (Note that $\widehat{\lambda}$ depends on s_n, but also on $\boldsymbol{d}$ and h.) For the last inequality in the last display we use the fact that, by definition, $\boldsymbol{\beta}_n(\lambda)$ and $\gamma_n(\lambda)$ minimize $Q_n((\lambda,\boldsymbol{\beta}^T)^T,\gamma)$ with respect to $\boldsymbol{\beta}$ and γ given λ.

Meanwhile, from (1.18), (1.19) and (1.20) and the fact that S_n is a $o_{\mathbb{P}}(1)$ neighborhood of s_0 that is contained in the support of Y, for any s_n,

$$\begin{aligned}&\left|s_n^{-2\widehat{\lambda}}\widehat{M}_n(\widehat{\lambda})-s_n^{-2\widehat{\lambda}}Q((\widehat{\lambda},\widehat{\boldsymbol{\beta}}(\widehat{\lambda})^T)^T,\widehat{\gamma}(\widehat{\lambda}))\right|\\&\quad\leq\left|s_n^{-2\widehat{\lambda}}\widehat{M}_n(\widehat{\lambda})-s_n^{-2\widehat{\lambda}}Q_n((\widehat{\lambda},\boldsymbol{\beta}_n(\widehat{\lambda})^T)^T,\gamma_n(\widehat{\lambda}))\right|\\&\qquad+s_n^{-2\widehat{\lambda}}\left|Q_n((\widehat{\lambda},\boldsymbol{\beta}_n(\widehat{\lambda})^T)^T,\gamma_n(\widehat{\lambda}))-Q_n((\widehat{\lambda},\widehat{\boldsymbol{\beta}}(\widehat{\lambda})^T)^T,\widehat{\gamma}(\widehat{\lambda}))\right|\\&\qquad+s_n^{-2\widehat{\lambda}}\left|Q_n((\widehat{\lambda},\widehat{\boldsymbol{\beta}}(\widehat{\lambda})^T)^T,\widehat{\gamma}(\widehat{\lambda}))-Q((\widehat{\lambda},\widehat{\boldsymbol{\beta}}(\widehat{\lambda})^T)^T,\widehat{\gamma}(\widehat{\lambda}))\right|\\&\quad\leq\sup_{s\in S_n}\sup_{\lambda\in\Lambda}s^{-2\lambda}\times\sup_{h\in\mathcal{H}_{c,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\sup_{\lambda\in\Lambda}\left|\widehat{M}_n(\lambda)-M_n(\lambda)\right|\\&\qquad+\sup_{s\in S_n}\sup_{\lambda\in\Lambda}s^{-2\lambda}\times\sup_{h\in\mathcal{H}_{c,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\sup_{\lambda\in\Lambda}\left|Q_n((\lambda,\boldsymbol{\beta}_n(\lambda)^T)^T,\gamma_n(\lambda))-Q_n((\lambda,\widehat{\boldsymbol{\beta}}(\lambda)^T)^T,\widehat{\gamma}(\lambda))\right|\\&\qquad+\sup_{s\in S_n}\sup_{\lambda\in\Lambda}s^{-2\lambda}\times\sup_{\boldsymbol{d}\in\mathcal{D}}\sup_{\lambda\in\Lambda}\sup_{\boldsymbol{\beta}\in\mathbb{R}^p}\sup_{\gamma\in\mathbb{R}}\left|Q_n((\lambda,\boldsymbol{\beta}^T)^T,\gamma)-Q((\lambda,\boldsymbol{\beta}^T)^T,\gamma)\right|\\&\quad=o_{\mathbb{P}}(1).\end{aligned}\tag{1.23}$$

Next, by property (1.21), for any $\varepsilon>0$ there exists $\zeta>0$ (depending on ε, but also on the endpoints of the sets S_n and Λ) such that the probability of the event

$$E_n=\left\{\inf_{\lambda\in\Lambda,|\lambda-\lambda_0|\geq\varepsilon}\inf_{\boldsymbol{d}\in\mathcal{D}}\inf_{s\in S_n}s^{-2\lambda}Q((\lambda,\boldsymbol{\beta}(\lambda)^T)^T,\gamma(\lambda))>s_0^{-2\lambda_0}Q((\lambda_0,\boldsymbol{\beta}(\lambda_0)^T)^T,\gamma(\lambda_0))+\zeta=\zeta\right\},$$

tends to 1. Moreover, the event

$$\left\{\sup_{h\in\mathcal{H}_{c,n}}\sup_{s\in S_n}\sup_{\boldsymbol{d}\in\mathcal{D}}\left|\widehat{\lambda}(s)-\lambda_0\right|\geq\varepsilon\right\},$$

is contained in the event

$$\left\{\inf_{h\in\mathcal{H}_{c,n}}\inf_{s\in S_n}\inf_{\boldsymbol{d}\in\mathcal{D}}s^{-2\widehat{\lambda}(s)}Q((\widehat{\lambda}(s),\boldsymbol{\beta}(\widehat{\lambda}(s))^T)^T,\gamma(\widehat{\lambda}(s)))>\zeta\right\}\cap E_n.$$

By (1.22) and (1.23), the probability of the intersection event tends to zero. Now the proof for $\widehat{\lambda}$ is complete.

Consider now the convergence of $\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$. Given the uniform convergence of $\left\|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}_n(\lambda)\right\|$ and the continuity of $\lambda \mapsto \boldsymbol{\beta}(\lambda)$, it suffices to obtain the convergence of $\|\boldsymbol{\beta}_n(\lambda) - \boldsymbol{\beta}(\lambda)\|$ uniformly over $o_{\mathbb{P}}(1)$ neighborhoods of λ_0. By construction,

$$\begin{aligned} 0 \leq s_n^{-2\lambda} M_n(\lambda) &= s_n^{-2\lambda} Q_n((\lambda, \boldsymbol{\beta}_n(\lambda)^T)^T, \gamma_n(\lambda)) \\ &\leq s_n^{-2\lambda} Q_n((\lambda, \boldsymbol{\beta}(\lambda)^T)^T, \gamma(\lambda)) = s_n^{-2\lambda} Q((\lambda, \boldsymbol{\beta}(\lambda)^T)^T, \gamma(\lambda)) + o_{\mathbb{P}}(1), \end{aligned} \tag{1.24}$$

uniformly with respect to s_n, λ, h and $\boldsymbol{d}$. Moreover, since $Q((\lambda_0, \boldsymbol{\beta}(\lambda_0)^T)^T, \gamma(\lambda_0)) = 0$, we have

$$Q((\lambda, \boldsymbol{\beta}(\lambda)^T)^T, \gamma(\lambda)) = o_{\mathbb{P}}(1), \tag{1.25}$$

uniformly over $o_{\mathbb{P}}(1)$ neighborhoods of λ_0. Meanwhile, by (1.20), for any s_n and any λ,

$$\begin{aligned} &\left| s_n^{-2\lambda} M_n(\lambda) - s_n^{-2\lambda} Q((\lambda, \boldsymbol{\beta}_n(\lambda)^T)^T, \gamma_n(\lambda)) \right| \\ &\qquad = s_n^{-2\lambda} \left| Q_n((\lambda, \boldsymbol{\beta}_n(\lambda)^T)^T, \gamma_n(\lambda)) - Q((\lambda, \boldsymbol{\beta}_n(\lambda)^T)^T, \gamma_n(\lambda)) \right| \\ &\qquad \leq \sup_{s \in S_n} \sup_{\lambda \in \Lambda} s^{-2\lambda} \times \sup_{\boldsymbol{d} \in \mathcal{D}} \sup_{\lambda \in \Lambda} \sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \sup_{\gamma \in \mathbb{R}} \left| Q_n((\lambda, \boldsymbol{\beta}^T)^T, \gamma) - Q((\lambda, \boldsymbol{\beta}^T)^T, \gamma) \right| \\ &\qquad = o_{\mathbb{P}}(1). \end{aligned} \tag{1.26}$$

Next, by the proof of property (1.21) and continuity arguments, for any $\varepsilon > 0$ there exists $\upsilon > 0$ such that the probability of the event

$$F_n = \left\{ \inf_{\lambda \in \Lambda, |\lambda - \lambda_0| = o_{\mathbb{P}}(1)} \inf_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| > \varepsilon} \inf_{\gamma \in \mathbb{R}} \inf_{\boldsymbol{d} \in \mathcal{D}} \inf_{s \in S_n} s^{-2\lambda} Q((\lambda, \boldsymbol{\beta}^T)^T, \gamma) > \upsilon \right\},$$

tends to 1. Finally, note that the event

$$\left\{ \sup_{\lambda \in \Lambda, |\lambda - \lambda_0| = o_{\mathbb{P}}(1)} \sup_{s \in S_n} \sup_{\boldsymbol{d} \in \mathcal{D}} \|\boldsymbol{\beta}_n(\lambda) - \boldsymbol{\beta}(\lambda)\| \geq \varepsilon \right\},$$

is contained in the intersection

$$\left\{ \inf_{\lambda \in \Lambda, |\lambda - \lambda_0| = o_{\mathbb{P}}(1)} \inf_{s \in S_n} \inf_{\boldsymbol{d} \in \mathcal{D}} s^{-2\lambda} Q((\lambda, \boldsymbol{\beta}_n(\lambda)^T)^T, \gamma_n(\lambda)) > \upsilon \right\} \cap F_n,$$

which, by (1.24), (1.25) and (1.26), has a probability tending to zero. Now the proof is complete.

□

Proof of Proposition 1.1.

As $\widehat{\lambda} - \lambda_0 = o_{\mathbb{P}}(1)$ uniformly with respect to $h \in \mathcal{H}_{sc,n}$, $\boldsymbol{d} \in \mathcal{D}$ and $s \in S_n$, we get that

$$\begin{aligned} 0 = n^{-1} s^{-\widehat{\lambda}} \widehat{\mathbb{Y}}_n(\widehat{\lambda})^T \, \widehat{\mathbb{B}}_n \frac{\partial}{\partial \lambda} \left\{ n^{-1} s^{-\widehat{\lambda}} \widehat{\mathbb{Y}}_n(\widehat{\lambda}) \right\} &= n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0)^T \, \widehat{\mathbb{B}}_n \frac{\partial}{\partial \lambda} \left\{ n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0) \right\} \\ &+ \left[\frac{\partial}{\partial \lambda} \left\{ n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0) \right\}^T \widehat{\mathbb{B}}_n \frac{\partial}{\partial \lambda} \left\{ n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0) \right\} \right. \\ &\left. + n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0)^T \, \widehat{\mathbb{B}}_n \frac{\partial^2}{\partial \lambda^2} \left\{ n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0) \right\} + R_{1,n}(\widetilde{\lambda}, \lambda_0; s) \right] \left(\widehat{\lambda} - \lambda_0 \right), \end{aligned}$$

where $\widetilde{\lambda} = c\widehat{\lambda} + (1 - c)\lambda_0$ for some $c \in (0, 1)$. We have that $\sup_{h \in \mathcal{H}_{sc,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} \sup_{s \in S_n} |R_{1,n}(\widetilde{\lambda}, \lambda_0; s)| = o_{\mathbb{P}}(1)$, see Lemma 1.16.

Note that $\frac{\partial}{\partial \lambda}\{n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0)\} = s^{-\lambda_0} \frac{\partial}{\partial \lambda}\{n^{-1} \widehat{\mathbb{Y}}_n(\lambda_0)\} - \log(s) s^{-\lambda_0} n^{-1} \widehat{\mathbb{Y}}_n(\lambda_0)$. First, we show that

$$\begin{aligned} &n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0)^T \, \widehat{\mathbb{B}}_n \frac{\partial}{\partial \lambda} \left\{ n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0) \right\} \\ &\qquad - n^{-1} s_0^{-\lambda_0} \left(\mathbb{Y}_n(\lambda_0) - \left(\widehat{\varepsilon}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}} \right)_n \right)^T \, \mathbb{B}_n s_0^{-\lambda_0} \frac{\partial}{\partial \lambda} \left\{ n^{-1} \mathbb{Y}_n(\lambda_0) \right\} = o_{\mathbb{P}}\left(n^{-1/2} \right), \end{aligned} \tag{1.27}$$

uniformly with respect to s, $\boldsymbol{d}$ and h. We start by showing that

$$\begin{aligned}&n^{-1}s^{-\lambda_0}\widehat{\mathbb{Y}}_n(\lambda_0)^T\,\widehat{\mathbb{B}}_n n^{-1}s^{-\lambda_0}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)\\ &\qquad - n^{-1}s_0^{-\lambda_0}\left(\mathbb{Y}_n(\lambda_0)-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)^T\,\mathbb{B}_n n^{-1}s_0^{-\lambda_0}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0) = o_{\mathbb{P}}\left(n^{-1/2}\right),\end{aligned}$$

uniformly with respect to s, $\boldsymbol{d}$ and h. By the property $\widehat{\mathbb{X}}_n^T\widehat{\mathbb{B}}_n = \mathbb{X}_n^T\mathbb{B}_n = \mathbf{0}_n$, we could equivalently prove that

$$\begin{aligned}&n^{-1}s^{-\lambda_0}\left(\widehat{\mathbb{Y}}_n(\lambda_0)-\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0\right)^T\,\widehat{\mathbb{B}}_n n^{-1}s^{-\lambda_0}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)\\ &\qquad - n^{-1}s_0^{-\lambda_0}\left(\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)^T\,\mathbb{B}_n n^{-1}s_0^{-\lambda_0}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0) = o_{\mathbb{P}}\left(n^{-1/2}\right),\end{aligned} \quad (1.28)$$

uniformly with respect to $s\in S_n$, $\boldsymbol{d}\in\mathcal{D}$ and $h\in\mathcal{H}_{sc,n}$. To obtain (1.28), we decompose the difference in a sum of the following four terms:

$$\begin{aligned}R_{n1} &= n^{-1}s^{-\lambda_0}\left(\left[\widehat{\mathbb{Y}}_n(\lambda_0)-\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0\right]-\left[\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0\right]+\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)^T\,\widehat{\mathbb{B}}_n n^{-1}s^{-\lambda_0}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0),\\ R_{n2} &= n^{-1}\left(s^{-2\lambda_0}-s_0^{-2\lambda_0}\right)\left(\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)^T\,\widehat{\mathbb{B}}_n n^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)\\ R_{n3} &= n^{-1}s_0^{-\lambda_0}\left(\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)^T\,\left[\widehat{\mathbb{B}}_n-\mathbb{B}_n\right]n^{-1}s_0^{-\lambda_0}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0),\\ \text{and}\qquad R_{n4} &= n^{-1}s_0^{-\lambda_0}\left(\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)^T\,\mathbb{B}_n n^{-1}s_0^{-\lambda_0}\left(\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)-\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right).\end{aligned}$$

Note that

$$\widehat{\mathbb{B}}_n = \mathbb{S}_n^T\left(\boldsymbol{I}_{n\times n}-P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}\right)\mathbb{S}_n \quad\text{and}\quad \mathbb{B}_n = \mathbb{S}_n^T\left(\boldsymbol{I}_{n\times n}-P_{\mathbb{S}_n\mathbb{X}_n}\right)\mathbb{S}_n,$$

where $P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}$ and $P_{\mathbb{S}_n\mathbb{X}_n}$ are the orthogonal projectors on the subspaces generated by $\mathbb{S}_n\widehat{\mathbb{X}}_n$ and $\mathbb{S}_n\mathbb{X}_n$, that is

$$P_{\mathbb{S}_n\widehat{\mathbb{X}}_n} = \mathbb{S}_n\widehat{\mathbb{X}}_n\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\widehat{\mathbb{X}}_n^T\mathbb{S}_n^T \quad\text{and}\quad P_{\mathbb{S}_n\mathbb{X}_n} = \mathbb{S}_n\mathbb{X}_n\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\mathbb{X}_n^T\mathbb{S}_n^T,$$

with

$$\mathbb{S}_n = \left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\boldsymbol{\Omega}_n^{1/2}.$$

$P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}$ is the projector on the subspace generated by the vector $\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n$, that is

$$P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n} = \frac{1}{\mathbf{1}_n^T\boldsymbol{\Omega}_n\mathbf{1}_n}\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{\Omega}_n^{1/2}.$$

Here, $\boldsymbol{\Omega}_n^{1/2}$ is the positive definite square root of $\boldsymbol{\Omega}_n$. Deduce that

$$\begin{aligned}|R_{n2}| \leq{}& \left|s^{-2\lambda_0}-s_0^{-2\lambda_0}\right|\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left[\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right]\right\|\\ &\times\left\|\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\left(\boldsymbol{I}_{n\times n}-P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}\right)\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\right\|_{\mathrm{Sp}}\\ &\times\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)\right\|_{\mathrm{Sp}},\end{aligned}$$

$$\begin{aligned}|R_{n3}| \leq{}& s_0^{-2\lambda_0}\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left[\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right]\right\|\\ &\times\left\|\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\left(P_{\mathbb{S}_n\mathbb{X}_n}-P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}\right)\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\right\|_{\mathrm{Sp}}\\ &\times\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)\right\|_{\mathrm{Sp}},\end{aligned}$$

and

$$|R_{n4}| \leq s_0^{-2\lambda_0} \left\| \boldsymbol{\Omega}_n^{1/2} n^{-1} \left[\mathbb{Y}_n(\lambda_0) - \mathbb{X}_n \boldsymbol{\beta}_0 - \left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}} \right)_n \right] \right\|$$
$$\times \left\| \left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n} \right) \left(\boldsymbol{I}_{n\times n} - P_{\mathbb{S}_n\mathbb{X}_n} \right) \left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n} \right) \right\|_{\mathrm{Sp}}$$
$$\times \left\| \boldsymbol{\Omega}_n^{1/2} n^{-1} \left(\frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0) - \frac{\partial}{\partial\lambda} \mathbb{Y}_n(\lambda_0) \right) \right\|_{\mathrm{Sp}} .$$

The uniform rate $o_{\mathbb{P}}\left(n^{-1/2}\right)$ as in equation (1.28) follows for R_{n2}, R_{n3} and R_{n4} from the fact that the spectral norm of a product of projectors is at most equal to 1, the spectral norm of $P_{\mathbb{S}_n\mathbb{X}_n} - P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}$ tends to zero, $\sup_{s\in S_n} \left| s^{-2\lambda_0} - s_0^{-2\lambda_0} \right| = o_{\mathbb{P}}(1)$ as well as $\sup_{s\in S_n} s^{-2\lambda_0} = O_{\mathbb{P}}(1)$ and from Lemmas 1.14, 1.15 and 1.19. For the term R_{n1}, we could write

$$|R_{n1}| \leq s^{-2\lambda_0} \left\| \boldsymbol{\Omega}_n^{1/2} n^{-1} \left([\mathbb{Y}_n(\lambda_0) - \mathbb{X}_n\boldsymbol{\beta}_0] - \left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}} \right)_n - \left[\widehat{\mathbb{Y}}_n(\lambda_0) - \widehat{\mathbb{X}}_n\boldsymbol{\beta}_0 \right] \right) \right\|$$
$$\times \left\| \left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n} \right) \left(\boldsymbol{I}_{n\times n} - P_{\mathbb{S}_n\widehat{\mathbb{X}}_n} \right) \left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n} \right) \right\|_{\mathrm{Sp}}$$
$$\times \left\| \boldsymbol{\Omega}_n^{1/2} n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0) \right\|_{\mathrm{Sp}} ,$$

and use Lemmas 1.15 and 1.18 and again the facts that the spectral norm of a product of projectors is at most equal to 1 and $\sup_{s\in S_n} s^{-2\lambda_0} = O_{\mathbb{P}}(1)$ to deduce that it is of rate $o_{\mathbb{P}}\left(n^{-1/2}\right)$, uniformly with respect to s, $\boldsymbol{d}$ and h. Now the proof of the property (1.28) is complete. Next, we show that

$$n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0)^T \, \widehat{\mathbb{B}}_n n^{-1} \log(s) s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0) = o_{\mathbb{P}}\left(n^{-1/2}\right),$$

uniformly with respect to s, $\boldsymbol{d}$ and h. By the property $\widehat{\mathbb{X}}_n^T \widehat{\mathbb{B}}_n = \mathbf{0}_n$, we could equivalently prove that

$$n^{-1} s^{-\lambda_0} \left(\widehat{\mathbb{Y}}_n(\lambda_0) - \widehat{\mathbb{X}}_n\boldsymbol{\beta}_0 \right)^T \widehat{\mathbb{B}}_n n^{-1} \log(s) s^{-\lambda_0} \left(\widehat{\mathbb{Y}}_n(\lambda_0) - \widehat{\mathbb{X}}_n\boldsymbol{\beta}_0 \right) = o_{\mathbb{P}}\left(n^{-1/2}\right), \qquad (1.29)$$

uniformly with respect to $s \in S_n$, $\boldsymbol{d} \in \mathcal{D}$ and $h \in \mathcal{H}_{sc,n}$. By a similar decomposition as in the proof of (1.28) we get that

$$n^{-1} s^{-\lambda_0} \left(\widehat{\mathbb{Y}}_n(\lambda_0) - \widehat{\mathbb{X}}_n\boldsymbol{\beta}_0 \right)^T \widehat{\mathbb{B}}_n n^{-1} \log(s) s^{-\lambda_0} \left(\widehat{\mathbb{Y}}_n(\lambda_0) - \widehat{\mathbb{X}}_n\boldsymbol{\beta}_0 \right)$$
$$-n^{-1} s^{-\lambda_0} \left(\mathbb{Y}_n(\lambda_0) - \mathbb{X}_n\boldsymbol{\beta}_0 - \left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}} \right)_n \right)^T \mathbb{B}_n n^{-1} \log(s) s^{-\lambda_0} \left(\mathbb{Y}_n(\lambda_0) - \mathbb{X}_n\boldsymbol{\beta}_0 - \left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}} \right)_n \right) = o_{\mathbb{P}}\left(n^{-1/2}\right).$$

We obtain

$$n^{-1} s^{-\lambda_0} \left(\mathbb{Y}_n(\lambda_0) - \mathbb{X}_n\boldsymbol{\beta}_0 - \left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}} \right)_n \right)^T \mathbb{B}_n n^{-1} \log(s) s^{-\lambda_0} \left(\mathbb{Y}_n(\lambda_0) - \mathbb{X}_n\boldsymbol{\beta}_0 - \left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}} \right)_n \right) = o_{\mathbb{P}}\left(n^{-1/2}\right),$$

from Lemmas 1.14 and 1.19 and the facts that the spectral norm of a product of projectors is at most equal to 1 and $\sup_{s\in S_n} s^{-\lambda_0} = O_{\mathbb{P}}(1)$ as well as $\sup_{s\in S_n} \log(s) s^{-\lambda_0} = O_{\mathbb{P}}(1)$ such that (1.29) follows. (1.27) follows now from (1.28) and (1.29). Next, we show that

$$n^{-1} \frac{\partial}{\partial\lambda} \left\{ s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0) \right\}^T \widehat{\mathbb{B}}_n n^{-1} \frac{\partial}{\partial\lambda} \left\{ s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0) \right\}$$
$$- n^{-1} s_0^{-\lambda_0} \frac{\partial}{\partial\lambda} \mathbb{Y}_n(\lambda_0)^T \mathbb{B}_n n^{-1} s_0^{-\lambda_0} \frac{\partial}{\partial\lambda} \mathbb{Y}_n(\lambda_0) = o_{\mathbb{P}}(1), \qquad (1.30)$$

uniformly with respect to s, $\boldsymbol{d}$ and h. We start by showing that

$$n^{-1} s^{-\lambda_0} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0)^T \widehat{\mathbb{B}}_n n^{-1} s^{-\lambda_0} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0)$$
$$- n^{-1} s_0^{-\lambda_0} \frac{\partial}{\partial\lambda} \mathbb{Y}_n(\lambda_0)^T \mathbb{B}_n n^{-1} s_0^{-\lambda_0} \frac{\partial}{\partial\lambda} \mathbb{Y}_n(\lambda_0) = o_{\mathbb{P}}(1), \qquad (1.31)$$

uniformly with respect to $s \in S_n$, $\boldsymbol{d} \in \mathcal{D}$ and $h \in \mathcal{H}_{sc,n}$. To obtain (1.31), we decompose the difference in a sum of the following four terms:

$$\begin{aligned} \widetilde{R}_{n1} &= n^{-1}s^{-\lambda_0}\left(\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0) - \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right)^T \widehat{\mathbb{B}}_n n^{-1}s^{-\lambda_0}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0), \\ \widetilde{R}_{n2} &= n^{-1}\left(s^{-2\lambda_0} - s_0^{-2\lambda_0}\right)\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\, \widehat{\mathbb{B}}_n n^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0) \\ \widetilde{R}_{n3} &= n^{-1}s_0^{-\lambda_0}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\,\left[\widehat{\mathbb{B}}_n - \mathbb{B}_n\right] n^{-1}s_0^{-\lambda_0}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0), \\ \text{and}\qquad \widetilde{R}_{n4} &= n^{-1}s_0^{-\lambda_0}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\, \mathbb{B}_n n^{-1}s_0^{-\lambda_0}\left(\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0) - \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right). \end{aligned}$$

It follows now that

$$\begin{aligned} \left|\widetilde{R}_{n1}\right| \le s^{-2\lambda_0} &\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left(\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0) - \frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)\right)\right\|_{\mathrm{Sp}} \\ &\times\left\|\left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\left(\boldsymbol{I}_{n\times n} - P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}\right)\left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\right\|_{\mathrm{Sp}} \\ &\times\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)\right\|_{\mathrm{Sp}}, \end{aligned}$$

$$\begin{aligned} \left|\widetilde{R}_{n2}\right| \le \left|s^{-2\lambda_0} - s_0^{-2\lambda_0}\right| &\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right\|_{\mathrm{Sp}} \\ &\times\left\|\left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\left(\boldsymbol{I}_{n\times n} - P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}\right)\left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\right\|_{\mathrm{Sp}} \\ &\times\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)\right\|_{\mathrm{Sp}}, \end{aligned}$$

$$\begin{aligned} \left|\widetilde{R}_{n3}\right| \le s_0^{-2\lambda_0} &\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right\|_{\mathrm{Sp}} \\ &\times\left\|\left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\left(P_{\mathbb{S}_n\mathbb{X}_n} - P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}\right)\left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\right\|_{\mathrm{Sp}} \\ &\times\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)\right\|_{\mathrm{Sp}}, \end{aligned}$$

and

$$\begin{aligned} \left|\widetilde{R}_{n4}\right| \le s_0^{-2\lambda_0} &\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right\|_{\mathrm{Sp}} \\ &\times\left\|\left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\left(\boldsymbol{I}_{n\times n} - P_{\mathbb{S}_n\mathbb{X}_n}\right)\left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\right\|_{\mathrm{Sp}} \\ &\times\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left(\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0) - \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right)\right\|_{\mathrm{Sp}}. \end{aligned}$$

The uniform rate $o_{\mathbb{P}}(1)$ as in equation (1.31) follows for $\widetilde{R}_{n1}$, $\widetilde{R}_{n2}$, $\widetilde{R}_{n3}$ and $\widetilde{R}_{n4}$ from the fact that the spectral norm of a product of projectors is at most equal to 1, the spectral norm of $P_{\mathbb{S}_n\mathbb{X}_n} - P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}$ tends to zero, $\sup_{s\in S_n}\left|s^{-2\lambda_0} - s_0^{-2\lambda_0}\right| = o_{\mathbb{P}}(1)$ as well as $\sup_{s\in S_n} s^{-2\lambda_0} = O_{\mathbb{P}}(1)$ and from Lemma 1.15. Now the proof of (1.31) is complete. In order to proof (1.30) it remains to be shown that

$$n^{-1}\log(s)s^{-\lambda_0}\widehat{\mathbb{Y}}_n(\lambda_0)^T\widehat{\mathbb{B}}_n n^{-1}s^{-\lambda_0}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0) = o_{\mathbb{P}}(1), \tag{1.32}$$

and

$$n^{-1}\log(s)s^{-\lambda_0}\widehat{\mathbb{Y}}_n(\lambda_0)^T\widehat{\mathbb{B}}_n n^{-1}\log(s)s^{-\lambda_0}\widehat{\mathbb{Y}}_n(\lambda_0) = o_{\mathbb{P}}(1), \tag{1.33}$$

uniformly with respect to $s \in S_n$, $\boldsymbol{d} \in \mathcal{D}$ and $h \in \mathcal{H}_{sc,n}$. We obtain (1.32) and (1.33) by a similiar reasoning as in the proof of (1.29). The details are omitted. Now the proof of (1.30) is complete.

Next, we show that

$$n^{-1}s^{-\lambda_0}\widehat{\mathbb{Y}}_n(\lambda_0)^T\widehat{\mathbb{B}}_n\frac{\partial^2}{\partial\lambda^2}\left\{n^{-1}s^{-\lambda_0}\widehat{\mathbb{Y}}_n(\lambda_0)\right\}=o_{\mathbb{P}}\left(1\right), \tag{1.34}$$

uniformly with respect to $s\in S_n$, $\boldsymbol{d}\in\mathcal{D}$ and $h\in\mathcal{H}_{sc,n}$. Note that $\frac{\partial^2}{\partial\lambda^2}\{n^{-1}s^{-\lambda_0}\widehat{\mathbb{Y}}_n(\lambda_0)\}=s^{-\lambda_0}n^{-1}\frac{\partial^2}{\partial\lambda^2}\widehat{\mathbb{Y}}_n(\lambda_0)-2\log(s)s^{-\lambda_0}n^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)+\log(s)^2s^{-\lambda_0}\widehat{\mathbb{Y}}_n(\lambda_0)$. We start by showing that

$$n^{-1}s^{-\lambda_0}\widehat{\mathbb{Y}}_n(\lambda_0)^T\ \widehat{\mathbb{B}}_ns^{-\lambda_0}n^{-1}\frac{\partial^2}{\partial\lambda^2}\widehat{\mathbb{Y}}_n(\lambda_0)=o_{\mathbb{P}}\left(1\right), \tag{1.35}$$

uniformly with respect to $s\in S_n$, $\boldsymbol{d}\in\mathcal{D}$ and $h\in\mathcal{H}_{sc,n}$. Once again we can equivalently consider

$$n^{-1}s^{-\lambda_0}\left(\widehat{\mathbb{Y}}_n(\lambda_0)-\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0\right)^T\widehat{\mathbb{B}}_ns^{-\lambda_0}n^{-1}\frac{\partial^2}{\partial\lambda^2}\widehat{\mathbb{Y}}_n(\lambda_0)=o_{\mathbb{P}}\left(1\right), \tag{1.36}$$

uniformly with respect to $s\in S_n$, $\boldsymbol{d}\in\mathcal{D}$ and $h\in\mathcal{H}_{sc,n}$. To obtain (1.36), we consider

$$\begin{aligned}
&\left|n^{-1}s^{-\lambda_0}\left(\widehat{\mathbb{Y}}_n(\lambda_0)-\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0\right)^T\ \mathbb{B}_ns^{-\lambda_0}n^{-1}\frac{\partial^2}{\partial\lambda^2}\widehat{\mathbb{Y}}_n(\lambda_0)\right|\\
&\qquad\le s^{-2\lambda_0}\Bigg(\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left(\left[\widehat{\mathbb{Y}}_n(\lambda_0)-\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0\right]-\left[\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0\right]+\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\right\|\\
&\qquad\qquad\qquad+\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left[\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right]\right\|\Bigg)\\
&\qquad\qquad\times\left\|\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\left(\boldsymbol{I}_{n\times n}-P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}\right)\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\right\|_{\mathrm{Sp}}\times\left\|n^{-1}\boldsymbol{\Omega}_n^{1/2}\frac{\partial^2}{\partial\lambda^2}\widehat{\mathbb{Y}}_n(\lambda_0)\right\|_{\mathrm{Sp}}.
\end{aligned}$$

The uniform rate $o_{\mathbb{P}}\left(1\right)$ as in equation (1.36) follows from the fact that the spectral norm of a product of projectors is at most equal to 1, $\sup_{s\in S_n}s^{-2\lambda_0}=O_{\mathbb{P}}(1)$ and from Lemmas 1.14, 1.15, 1.18 and 1.19. Now the proof of property (1.35) is complete. (1.34) follows now together with (1.32) and (1.33). Therefore, the proof of the first part of the first statement in the Proposition is complete.

In addition we have that

$$\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)-E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\right\|=O_{\mathbb{P}}(n^{-1/2}),$$

by Lemma 1.5, where the expectation tends to a positive constant. Furthermore, it follows from the fact that the spectral norm of a product of projectors is at most equal to 1 and from Lemmas 1.14, 1.15 and 1.19 that

$$\sup_{\boldsymbol{d}\in\mathcal{D}}\left|n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_n\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\right|=O_{\mathbb{P}}\left(n^{-1/2}\right),$$

such that $(\widehat{\lambda}-\lambda_0)=O_{\mathbb{P}}\left(n^{-1/2}\right)$ uniformly with respect to $s\in S_n$, $\boldsymbol{d}\in\mathcal{D}$ and $h\in\mathcal{H}_{sc,n}$. Therefore, the proof of the second part of the first statement in the Proposition is complete.

We consider now

$$\widehat{\boldsymbol{\beta}}(\widehat{\lambda})=\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{Y}}_n(\widehat{\lambda})=\left(n^{-2}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_nn^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda}).$$

Once again we can write that

$$n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_nn^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda})=n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_nn^{-1}\widehat{\mathbb{Y}}_n(\lambda_0)+\left(n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_nn^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)+R_{2,n}(\widetilde{\lambda},\lambda_0)\right)(\widehat{\lambda}-\lambda_0),$$

where $\widetilde{\lambda}=c\widehat{\lambda}+(1-c)\lambda_0$ for some $c\in(0,1)$. We have that $\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\sup_{s\in S_n}|R_{2,n}(\widetilde{\lambda},\lambda_0)|=o_{\mathbb{P}}(1)$, see Lemma 1.17. In addition, we get that

$$\begin{aligned}
\left(n^{-2}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_nn^{-1}\widehat{\mathbb{Y}}_n(\lambda_0)&=\left(n^{-2}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_nn^{-1}\left(\widehat{\mathbb{Y}}_n(\lambda_0)-\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0+\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0\right)\\
&=\boldsymbol{\beta}_0+\left(n^{-2}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_nn^{-1}\left(\widehat{\mathbb{Y}}_n(\lambda_0)-\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0\right).
\end{aligned}$$

In the first step we show that

$$n^{-1}\widehat{\mathbb{X}}_n^T\,\mathbb{D}_n\,n^{-1}\left(\widehat{\mathbb{Y}}_n(\lambda_0)-\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0\right)-n^{-1}\mathbb{X}_n^T\mathbb{D}_n\,n^{-1}\left(\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)=o_{\mathbb{P}}(n^{-1/2}) \quad (1.37)$$

uniformly with respect to $\boldsymbol{d}\in\mathcal{D}$ and $h\in\mathcal{H}_{sc,n}$. We have that

$$\begin{aligned}
&n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\,n^{-1}\left(\widehat{\mathbb{Y}}_n(\lambda_0)-\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0\right)-n^{-1}\mathbb{X}_n^T\mathbb{D}_n\,n^{-1}\left(\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\\
&\quad=n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\,n^{-1}\left(\left(\widehat{\mathbb{Y}}_n(\lambda_0)-\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0\right)-\left(\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\right)\\
&\qquad+n^{-1}\left(\widehat{\mathbb{X}}_n-\mathbb{X}_n\right)^T\mathbb{D}_n\,n^{-1}\left(\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\left\|n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\left(n^{-1}\left(\widehat{\mathbb{Y}}_n(\lambda_0)-\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0\right)-n^{-1}\left(\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\right)\right\|_{\mathrm{Sp}}\\
&\qquad\le\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\widehat{\mathbb{X}}_n\right\|_{\mathrm{Sp}}\times\left\|\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\right\|_{\mathrm{Sp}}\\
&\qquad\quad\times\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left(\left(\widehat{\mathbb{Y}}_n(\lambda_0)-\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0\right)-\left(\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\right)\right\|,
\end{aligned}$$

and

$$\begin{aligned}
&\left\|n^{-1}\left(\widehat{\mathbb{X}}_n-\mathbb{X}_n\right)^T\mathbb{D}_n\,n^{-1}\left(\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\right\|_{\mathrm{Sp}}\\
&\le\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left(\widehat{\mathbb{X}}_n-\mathbb{X}_n\right)\right\|_{\mathrm{Sp}}\times\left\|\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\right\|_{\mathrm{Sp}}\times\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left(\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\right\|.
\end{aligned}$$

The uniform rate $o_{\mathbb{P}}\left(n^{-1/2}\right)$ as in equation (1.37) follows from the fact that the spectral norm of a product of projectors is at most equal to 1 and from Lemmas 1.8, 1.14, 1.18 and 1.19.

In the next step we show that

$$n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_nn^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)(\widehat{\lambda}-\lambda_0)-n^{-1}\mathbb{X}_n^T\mathbb{D}_nn^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)(\widehat{\lambda}-\lambda_0)=o_{\mathbb{P}}(n^{-1/2}), \quad (1.38)$$

uniformly with respect to $s\in S_n$, $\boldsymbol{d}\in\mathcal{D}$ and $h\in\mathcal{H}_{sc,n}$. We have that

$$\begin{aligned}
&n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_nn^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)(\widehat{\lambda}-\lambda_0)-n^{-1}\mathbb{X}_n^T\mathbb{D}_nn^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)(\widehat{\lambda}-\lambda_0)\\
&\qquad=n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_nn^{-1}\left(\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)-\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right)(\widehat{\lambda}-\lambda_0)\\
&\qquad\quad+n^{-1}\left(\widehat{\mathbb{X}}_n-\mathbb{X}_n\right)^T\mathbb{D}_nn^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)(\widehat{\lambda}-\lambda_0).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\left\|n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\,n^{-1}\left(\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)-\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right)\right\|_{\mathrm{Sp}}\\
&\qquad\le\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\widehat{\mathbb{X}}_n\right\|_{\mathrm{Sp}}\times\left\|\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\right\|_{\mathrm{Sp}}\times\left\|\boldsymbol{\Omega}_n^{1/2}\,n^{-1}\left(\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)-\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right)\right\|_{\mathrm{Sp}},
\end{aligned}$$

and

$$\begin{aligned}
&\left\|n^{-1}\left(\widehat{\mathbb{X}}_n-\mathbb{X}_n\right)^T\mathbb{D}_nn^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right\|_{\mathrm{Sp}}\\
&\qquad\le\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left(\widehat{\mathbb{X}}_n-\mathbb{X}_n\right)\right\|_{\mathrm{Sp}}\times\left\|\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\right\|_{\mathrm{Sp}}\times\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right\|_{\mathrm{Sp}}.
\end{aligned}$$

The uniform rate $o_{\mathbb{P}}\left(n^{-1/2}\right)$ as in equation (1.38) follows from the fact that the spectral norm of a product of projectors is at most equal to 1, from Lemma 1.8 and 1.15 and the fact that $(\widehat{\lambda}-\lambda_0)=O_{\mathbb{P}}\left(n^{-1/2}\right)$ uniformly with respect to $s\in S_n$, $\boldsymbol{d}\in\mathcal{D}$ and $h\in\mathcal{H}_{sc,n}$. In addition, it follows from Lemmas 1.5 and 1.7

that

$$\left\| \left(n^{-2}\widehat{\mathbb{X}}_n^T \mathbb{D}_n \widehat{\mathbb{X}}_n\right)^{-1} - \left(n^{-2}\mathbb{X}_n^T \mathbb{D}_n \mathbb{X}_n\right)^{-1} \right\| = o_{\mathbb{P}}(1).$$

Furthermore, we get that

$$\left\| n^{-1}\mathbb{X}_n^T \mathbb{D}_n n^{-1}\left(\mathbb{Y}_n(\lambda_0) - \mathbb{X}_n \boldsymbol{\beta}_0 - \left(\widehat{\varepsilon}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right) \right\|_{\mathrm{Sp}} = O_{\mathbb{P}}\left(n^{-1/2}\right),$$

and

$$\left\| n^{-1}\mathbb{X}_n^T \mathbb{D}_n n^{-1}\frac{\partial}{\partial \lambda}\mathbb{Y}_n(\lambda_0)(\widehat{\lambda} - \lambda_0) \right\|_{\mathrm{Sp}} = O_{\mathbb{P}}\left(n^{-1/2}\right),$$

from Lemmas 1.8, 1.14, 1.15, 1.19 and $(\widehat{\lambda} - \lambda_0) = O_{\mathbb{P}}\left(n^{-1/2}\right)$ uniformly with respect to $s \in S_n$, $\boldsymbol{d} \in \mathcal{D}$ and $h \in \mathcal{H}_{sc,n}$. Therefore, the second statement follows.

□

Proof of Theorem 1.2.

Let

$$\begin{aligned} \boldsymbol{V}_n(\boldsymbol{d}) &= \begin{pmatrix} \frac{\partial}{\partial \lambda}\mathbb{Y}_n(\lambda_0)^T \mathbb{D}_n(\boldsymbol{d}) \frac{\partial}{\partial \lambda}\mathbb{Y}_n(\lambda_0) & -\frac{\partial}{\partial \lambda}\mathbb{Y}_n(\lambda_0)^T \mathbb{D}_n(\boldsymbol{d})\mathbb{X}_n \\ -\mathbb{X}_n^T \mathbb{D}_n(\boldsymbol{d}) \frac{\partial}{\partial \lambda}\mathbb{Y}_n(\lambda_0) & \mathbb{X}_n^T \mathbb{D}_n(\boldsymbol{d})\mathbb{X}_n \end{pmatrix} \quad \text{and} \\ \boldsymbol{A}_n &= \begin{pmatrix} \frac{\partial}{\partial \lambda}\mathbb{Y}_n(\lambda_0)^T \\ -\mathbb{X}_n^T \end{pmatrix} \mathbb{D}_n\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right). \end{aligned}$$

We get from Proposition 1.1 that

$$\left((\widehat{\lambda}, \widehat{\boldsymbol{\beta}}(\widehat{\lambda})^T)^T - (\lambda_0, \boldsymbol{\beta}_0^T)^T\right) = -\boldsymbol{V}_n(\boldsymbol{d})^{-1}\boldsymbol{A}_n + o_{\mathbb{P}}\left(n^{-1/2}\right).$$

Furthermore, it follows from Lemma 1.5 that

$$\left((\widehat{\lambda}, \widehat{\boldsymbol{\beta}}(\widehat{\lambda})^T)^T - (\lambda_0, \boldsymbol{\beta}_0^T)^T\right) = -\boldsymbol{V}(\boldsymbol{d})^{-1}n^{-2}\boldsymbol{A}_n + o_{\mathbb{P}}\left(n^{-1/2}\right).$$

Both results hold uniformly with respect to $s \in S_n$, $\boldsymbol{d} \in \mathcal{D}$ and $h \in \mathcal{H}_{sc,n}$. Note that $\boldsymbol{V}(\boldsymbol{d})$ is invertible as $E\left[n^{-2}\mathbb{X}_n^T \mathbb{D}_n \mathbb{X}_n\right]$ tends to a positive definite matrix and $E\left[n^{-2}\frac{\partial}{\partial \lambda}\mathbb{Y}_n(\lambda_0)^T \mathbb{B}_n \frac{\partial}{\partial \lambda}\mathbb{Y}_n(\lambda_0)\right]$ to a positive constant, see Lemma 1.5. We consider now $\boldsymbol{A}_n$ and start with $\mathbb{X}_n^T \mathbb{D}_n\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)$. Recall that

$$\begin{aligned} &\mathbb{X}_n^T \mathbb{D}_n\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right) \\ &\qquad = \mathbb{X}_n^T \boldsymbol{\Omega}_n\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right) - \frac{1}{\mathbf{1}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n}\mathbb{X}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n \mathbf{1}_n^T \boldsymbol{\Omega}_n\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right). \end{aligned}$$

It follows from the results of Lemma 1.5 that

$$\sup_{\boldsymbol{d} \in \mathcal{D}} \left\| \frac{1}{\mathbf{1}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n}\mathbb{X}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n - \frac{1}{E\left[\mathbf{1}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n\right]}E\left[\mathbb{X}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n\right] \right\|_{\mathrm{Sp}} = O_{\mathbb{P}}(n^{-1/2}),$$

and together with the results of Lemmas 1.3, 1.14 and 1.19 we get that

$$\begin{aligned} \sup_{\boldsymbol{d} \in \mathcal{D}} \Bigg\| &\frac{1}{\mathbf{1}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n}\mathbb{X}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n \frac{1}{n^2}\mathbf{1}_n^T \boldsymbol{\Omega}_n\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right) \\ &- \frac{1}{E\left[\mathbf{1}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n\right]}E\left[\mathbb{X}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n\right] \frac{1}{n^2}\mathbf{1}_n^T \boldsymbol{\Omega}_n\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right) \Bigg\|_{\mathrm{Sp}} = o_{\mathbb{P}}(n^{-1/2}). \end{aligned}$$

In the next step we consider

$$\frac{1}{n^2}\mathbb{X}_n^T\boldsymbol{\Omega}_n(\boldsymbol{\varepsilon f_z})_n = \frac{1}{2n^2}\sum_{1\le i\neq j\le n}\left(\mathbb{X}_{n,i}\varepsilon_j f_z(\boldsymbol{Z}_j)\boldsymbol{\Omega}_{n,ij} + \mathbb{X}_{n,j}\varepsilon_i f_z(\boldsymbol{Z}_i)\boldsymbol{\Omega}_{n,ji}\right) + \frac{1}{n^2}\sum_{i=1}^n \mathbb{X}_{n,i}\varepsilon_i f_z(\boldsymbol{Z}_i).$$

It's easy to check that

$$\left\|\frac{1}{n^2}\sum_{i=1}^n \mathbb{X}_{n,i}\varepsilon_i f_z(\boldsymbol{Z}_i)\right\| = o_{\mathbb{P}}(n^{-1}).$$

In addition, we have that $E\left[\mathbb{X}_{n,i}\varepsilon_j f_z(\boldsymbol{Z}_j)\boldsymbol{\Omega}_{n,ij}\right] = 0$ and $E\left[\mathbb{X}_{n,i}\varepsilon_j f_z(\boldsymbol{Z}_j)\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{X}_i, \boldsymbol{Z}_i\right] = 0$ as well as

$$E\left[\mathbb{X}_{n,i}\varepsilon_j f_z(\boldsymbol{Z}_j)\boldsymbol{\Omega}_{n,ij} \mid Y_j, \boldsymbol{X}_j, \boldsymbol{Z}_j\right] = \varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\mathbb{X}_{n,i}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{X}_j, \boldsymbol{Z}_j\right].$$

Therefore, we get by applying Hoeffding's decomposition that

$$\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\frac{1}{n^2}\mathbb{X}_n^T\boldsymbol{\Omega}_n(\boldsymbol{\varepsilon f_z})_n - \frac{1}{n}\sum_{j=1}^n \varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\mathbb{X}_{n,i}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{X}_j, \boldsymbol{Z}_j\right]\right\| = O_{\mathbb{P}}(n^{-1}).$$

By Lemma 1.14 it suffices to consider $\boldsymbol{d} = \operatorname{diag}(\mathrm{d_U}, \ldots, \mathrm{d_U})$ such that the uniform result in the last display follows. By the same reasoning we get that

$$\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\frac{1}{n^2}\mathbf{1}_n^T\boldsymbol{\Omega}_n(\boldsymbol{\varepsilon f_z})_n - \frac{1}{n}\sum_{j=1}^n \varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{X}_j, \boldsymbol{Z}_j\right]\right\| = O_{\mathbb{P}}(n^{-1}).$$

In the next step we consider

$$\frac{1}{n^2}\mathbb{X}_n^T\boldsymbol{\Omega}_n\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n = \frac{1}{n^2}\sum_{1\le i\neq j\le n}\mathbb{X}_{n,i}\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_{n,j}\boldsymbol{\Omega}_{n,ij} + \frac{1}{n^2}\sum_{i=1}^n \mathbb{X}_{n,i}\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_{n,i}.$$

It's easy to check that

$$\left\|\frac{1}{n^2}\sum_{i=1}^n \mathbb{X}_{n,i}\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_{n,i}\right\| = o_{\mathbb{P}}(n^{-1/2}).$$

In addition, we have that

$$\begin{aligned}\frac{1}{n^2}\sum_{1\le i\neq j\le n}\mathbb{X}_{n,i}\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_{n,j}\boldsymbol{\Omega}_{n,ij} &= \frac{1}{n^2}\sum_{1\le i\neq j\le n}\mathbb{X}_{n,i}\frac{1}{n}\sum_{k=1,k\neq j}^n \varepsilon_k K_{h,jk}\boldsymbol{\Omega}_{n,ij}\\ &= \frac{1}{n^3}\sum_{1\le i\neq j\neq k\le n}\mathbb{X}_{n,i}\varepsilon_k K_{h,jk}\boldsymbol{\Omega}_{n,ij}\\ &\quad + \frac{1}{n^3}\sum_{1\le i\neq j\le n}\mathbb{X}_{n,i}\varepsilon_i K_{h,ij}\boldsymbol{\Omega}_{n,ij}\\ &= A_n(h) + B_n(h).\end{aligned}$$

In the following we compute the mean and use the Hoeffding decomposition for the $U-$process $A_n(h)$. The kernel of $A_n(h)$ is not symmetric in its arguments. However, we could apply the usual symmetrization idea. Thus, by abuse, we will proceed as if the kernel of the $U-$statistic we handle is symmetric. For instance, for a second order $U-$statistic defined by a kernel $h(\boldsymbol{U}_i, \boldsymbol{U}_j)$, we could replace it by the symmetric kernel $\frac{1}{2}\left[h(\boldsymbol{U}_i, \boldsymbol{U}_j) + h(\boldsymbol{U}_j, \boldsymbol{U}_i)\right]$ from which we get the same $U-$statistic. Here, $\boldsymbol{U}_i = \left(Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T\right)^T$.

In addition, we have that the kernel of $A_n(h)$ is Euclidean for a squared integrable envelope. See Lemma 22 in Nolan and Pollard [67] and Lemma 2.14 in Pakes and Pollard [68]. Therefore, we can in the following repeatedly apply Corollary 7 and the Maximal Inequality of Sherman [72]. All remainder terms are controlled by Assumption 1.3.2.

Recall that by assumption $E\left[\varepsilon_k \mid \boldsymbol{X}_k, \boldsymbol{Z}_k\right] = 0$. Therefore, we get that $E\left[A_n(h)\right] = 0$ as well as

$$E\left[\mathbb{X}_{n,i}\varepsilon_k K_{h,jk}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_p, p\in\{i,j\}\right] = 0.$$

Furthermore we get that

$$\begin{aligned}E\left[\mathbb{X}_{n,i}\varepsilon_k K_{h,jk}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_k\right] &= \varepsilon_k E\left[\mathbb{X}_{n,i} K_{h,jk}\boldsymbol{\Omega}^X_{n,ij}\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{Z}_k\right] \\ &= \varepsilon_k E\left[\mathbb{X}_{n,i} E\left[K_{h,jk}\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{Z}_k, \boldsymbol{Z}_i, \boldsymbol{X}_j\right]\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_k\right] \\ &= \varepsilon_k E\left[\mathbb{X}_{n,i}\left(f_z(\boldsymbol{Z}_k)\boldsymbol{\Omega}^Z_{n,ik} + O_{\mathbb{P}}(h^2)\right)\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_k\right] \\ &= \varepsilon_k f_z(\boldsymbol{Z}_k) E\left[\mathbb{X}_{n,i}\boldsymbol{\Omega}^Z_{n,ik}\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_k\right] \\ &\quad + \varepsilon_k E\left[\mathbb{X}_{n,i}\boldsymbol{\Omega}^X_{n,ij}\right] O_{\mathbb{P}}(h^2).\end{aligned}$$

It follows from the results that the first order $U-$process of the Hoeffding decomposition of $A_n(h)$ is of order $O_{\mathbb{P}}(n^{-1/2})$ uniformly with respect to h and $\boldsymbol{d}$.

We consider now the three second order $U-$processes of the Hoeffding decomposition of $A_n(h)$. We get that

$$E\left[\mathbb{X}_{n,i}\varepsilon_k K_{h,jk}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_i, \boldsymbol{U}_j\right] = 0.$$

In addition,

$$\begin{aligned}E\left[\mathbb{X}_{n,i}\varepsilon_k K_{h,jk}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_i, \boldsymbol{U}_k\right] &= \mathbb{X}_{n,i}\varepsilon_k E\left[K_{h,jk}\boldsymbol{\Omega}^X_{n,ij}\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{U}_i, \boldsymbol{Z}_k\right] \\ &= \mathbb{X}_{n,i}\varepsilon_k E\left[\left(f_z(\boldsymbol{Z}_k)\boldsymbol{\Omega}^Z_{n,ik} + O_{\mathbb{P}}(h^2)\right)\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{U}_i, \boldsymbol{Z}_k\right] \\ &= \mathbb{X}_{n,i}\varepsilon_k f_z(\boldsymbol{Z}_k)\boldsymbol{\Omega}^Z_{n,ik} E\left[\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{X}_i\right] \\ &\quad + \mathbb{X}_{n,i} E\left[\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{X}_i\right] O_{\mathbb{P}}(h^2).\end{aligned}$$

The last conditional expectation that we need to consider is given by

$$\begin{aligned}E\left[\mathbb{X}_{n,i}\varepsilon_k K_{h,jk}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_j, \boldsymbol{U}_k\right] &= \varepsilon_k K_{h,jk} E\left[\mathbb{X}_{n,i}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_j\right] \\ &= h^{-q}h^q K_{h,jk}\tau(\boldsymbol{U}_j, \boldsymbol{U}_k).\end{aligned}$$

Now, we apply the Maximal Inequality of Sherman [72], page 448, for the degenerate $U-$process given by the kernel $h^q K_{h,jk}\tau(\boldsymbol{U}_j, \boldsymbol{U}_k)$, indexed by $h \in \mathcal{H}_{sc,n}$, with envelope $\|K\|_\infty \tau(\cdot, \cdot)$. (Herein, $\|\cdot\|_\infty$ denotes the uniform norm.) We take $p = 1$ and $\beta \in (0, 1)$ arbitrarily close to 1 to stand for Sherman's quantity α. Since $K(\cdot)$ is of bounded variation and symmetric, without loss of generality we could consider that $K(\cdot)$ is nonincreasing on $[0, \infty)$. In this case, $0 \leq K(\cdot/h) \leq K(\cdot/\overline{h})$ with $\overline{h} = \sup \mathcal{H}_{sc,n} =: c_{max}n^{-\alpha}$. Hence, using Jensen's inequality, we could bound the right-hand side of the Maximal Inequality of Sherman [72] by a universal constant times

$$E^{\beta/2}\left[K^2\left(\frac{\boldsymbol{Z}_j - \boldsymbol{Z}_k}{c_{max}n^{-\alpha}}\right)\tau^2(\boldsymbol{U}_j, \boldsymbol{U}_k)\right].$$

By standard changes of variables and suitable integrability conditions, this integral is bounded by a constant times $n^{-\alpha\beta q/2}$. Consequently, the uniform rate of the $U-$process obtained conditioning on $\boldsymbol{U}_j, \boldsymbol{U}_k$ is $n^{-1} \times O_{\mathbb{P}}(n^{\alpha q\{1-\beta/2\}})$. As $1/2 - \alpha q(1 - \beta/2) > 0$ under our assumptions we get that $n^{-1} \times O_{\mathbb{P}}(n^{\alpha q\{1-\beta/2\}}) = o_{\mathbb{P}}(n^{-1/2})$. From all the results it follows that the second order $U-$processes of the Hoeffding decomposition of $A_n(h)$ are of order $o_{\mathbb{P}}(n^{-1/2})$ uniformly with respect to h and $\boldsymbol{d}$.

Finally, we need to consider the third order $U-$process. We get that

$$\mathbb{X}_{n,i}\varepsilon_k K_{h,jk}\boldsymbol{\Omega}_{n,ij} = h^{-q}h^q K_{h,jk}\tau_1(\boldsymbol{U}_i, \boldsymbol{U}_j, \boldsymbol{U}_k).$$

We can again use the Maximal Inequality of Sherman [72] to argue that this process is of order $o_{\mathbb{P}}(n^{-1/2})$ uniformly with respect to h and $\boldsymbol{d}$. The details are omitted.

It remains to consider $B_n(h)$. One can argue in a similar way as for $A_n(h)$ to get that $B_n(h)$ is of order $o_{\mathbb{P}}(n^{-1/2})$ uniformly with respect to h and $\boldsymbol{d}$. The details are omitted.

From all the results it follows now that

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\frac{1}{n^2}\mathbb{X}_n^T\boldsymbol{\Omega}_n\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n - \frac{1}{n}\sum_{k=1}^n \varepsilon_k f_z(\boldsymbol{Z}_k) E\left[\mathbb{X}_{n,i}\boldsymbol{\Omega}^Z_{n,ik}\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_k\right]\right\| = o_{\mathbb{P}}(n^{-1/2}).$$

By the same reasoning we get that

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\frac{1}{n^2}\mathbf{1}_n^T\boldsymbol{\Omega}_n\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n - \frac{1}{n}\sum_{k=1}^n \varepsilon_k f_z(\boldsymbol{Z}_k) E\left[\boldsymbol{\Omega}^Z_{n,ik}\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_k\right]\right\| = o_{\mathbb{P}}(n^{-1/2}).$$

Therefore, we get that

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\Bigg\|\frac{1}{n^2}\mathbb{X}_n^T\mathbb{D}_n\left((\boldsymbol{\varepsilon}\boldsymbol{f_z})_n-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)$$
$$-\frac{1}{n}\sum_{j=1}^n\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\left(\mathbb{X}_{n,i}-\frac{1}{E\left[\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n\right]}E\left[\mathbb{X}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n\right]\right)\boldsymbol{\Omega}_{n,ij}\mid\boldsymbol{X}_j,\boldsymbol{Z}_j\right]$$
$$+\frac{1}{n}\sum_{k=1}^n\varepsilon_k f_z(\boldsymbol{Z}_k)E\left[\left(\mathbb{X}_{n,i}-\frac{1}{E\left[\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n\right]}E\left[\mathbb{X}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n\right]\right)\boldsymbol{\Omega}_{n,ik}^Z\boldsymbol{\Omega}_{n,ij}^X\mid\boldsymbol{Z}_k\right]\Bigg\|=o_{\mathbb{P}}(n^{-1/2}).$$

By the same arguments we get that

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\Bigg\|\frac{1}{n^2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{D}_n\left((\boldsymbol{\varepsilon}\boldsymbol{f_z})_n-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)$$
$$-\frac{1}{n}\sum_{j=1}^n\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\left(\frac{\partial}{\partial\lambda}\mathbb{Y}_{n,i}(\lambda_0)-\frac{1}{E\left[\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n\right]}E\left[\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\boldsymbol{\Omega}_n\boldsymbol{1}_n\right]\right)\boldsymbol{\Omega}_{n,ij}\mid\boldsymbol{X}_j,\boldsymbol{Z}_j\right]$$
$$+\frac{1}{n}\sum_{k=1}^n\varepsilon_k f_z(\boldsymbol{Z}_k)E\left[\left(\frac{\partial}{\partial\lambda}\mathbb{Y}_{n,i}(\lambda_0)-\frac{1}{E\left[\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n\right]}E\left[\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\boldsymbol{\Omega}_n\boldsymbol{1}_n\right]\right)\boldsymbol{\Omega}_{n,ik}^Z\boldsymbol{\Omega}_{n,ij}^X\mid\boldsymbol{Z}_k\right]\Bigg\|=o_{\mathbb{P}}(n^{-1/2}).$$

The details are omitted.

Therefore, we get that

$$\begin{aligned}\left((\widehat{\lambda},\widehat{\boldsymbol{\beta}}(\widehat{\lambda})^T)^T-(\lambda_0,\boldsymbol{\beta}_0^T)^T\right)&=-\boldsymbol{V}(\boldsymbol{d})^{-1}\Bigg(\frac{1}{n}\sum_{j=1}^n\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\,\boldsymbol{\Omega}_{n,ij}(\boldsymbol{d})\mid\boldsymbol{X}_j,\boldsymbol{Z}_j\right]\\&\qquad-\frac{1}{n}\sum_{k=1}^n\varepsilon_k f_z(\boldsymbol{Z}_k)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ik}^Z(\boldsymbol{d})\boldsymbol{\Omega}_{n,ij}^X(\boldsymbol{d})\mid\boldsymbol{Z}_k\right]\Bigg)+o_{\mathbb{P}}\left(n^{-1/2}\right)\\&=-\boldsymbol{V}(\boldsymbol{d})^{-1}\left(\frac{1}{n}\sum_{j=1}^n\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\,\boldsymbol{\Omega}_{n,ij}^Z(\boldsymbol{d})\left(\boldsymbol{\Omega}_{n,ij}^X(\boldsymbol{d})-\boldsymbol{\Omega}_{n,ik}^X(\boldsymbol{d})\right)\mid\boldsymbol{X}_j,\boldsymbol{Z}_j\right]\right)\\&\qquad+o_{\mathbb{P}}\left(n^{-1/2}\right)\\&=-\boldsymbol{V}(\boldsymbol{d})^{-1}\left(\frac{1}{n}\sum_{j=1}^n\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\,\boldsymbol{\Omega}_{n,ij}^Z(\boldsymbol{d})\left(\boldsymbol{\Omega}_{n,ij}^X(\boldsymbol{d})-E\left[\boldsymbol{\Omega}_{n,ik}^X(\boldsymbol{d})\mid\boldsymbol{X}_i\right]\right)\mid\boldsymbol{X}_j,\boldsymbol{Z}_j\right]\right)\\&\qquad+o_{\mathbb{P}}\left(n^{-1/2}\right)\\&=-\boldsymbol{V}(\boldsymbol{d})^{-1}\left(\frac{1}{n}\sum_{j=1}^n\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\,\boldsymbol{\Omega}_{n,ij}^Z(\boldsymbol{d})\boldsymbol{\Phi}_{n,ij}^X(\boldsymbol{d})\mid\boldsymbol{X}_j,\boldsymbol{Z}_j\right]\right)\\&\qquad+o_{\mathbb{P}}\left(n^{-1/2}\right),\end{aligned}$$

uniformly over $h\in\mathcal{H}_{sc,n}$ and $\boldsymbol{d}\in\mathcal{D}$.

We consider now the behavior of $\frac{1}{n}\sum_{j=1}^n\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\,\boldsymbol{\Omega}_{n,ij}^Z(\boldsymbol{d})\boldsymbol{\Phi}_{n,ij}^X(\boldsymbol{d})\mid\boldsymbol{X}_j,\boldsymbol{Z}_j\right]$ in detail by applying Theorem 19.28 of Van der Vaart [78]. The needed Lindeberg condition follows from our assumptions. In the following we will show that

$$\begin{aligned}\sup_{\|\boldsymbol{d}_1-\boldsymbol{d}_2\|<\delta}&E\Big[\Big\|\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d}_1)\,\boldsymbol{\Omega}_{n,ij}^Z(\boldsymbol{d}_1)\boldsymbol{\Phi}_{n,ij}^X(\boldsymbol{d}_1)\mid\boldsymbol{X}_j,\boldsymbol{Z}_j\right]\\&\qquad-\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d}_2)\,\boldsymbol{\Omega}_{n,ij}^Z(\boldsymbol{d}_2)\boldsymbol{\Phi}_{n,ij}^X(\boldsymbol{d}_2)\mid\boldsymbol{X}_j,\boldsymbol{Z}_j\right]\Big\|^2\Big]\to0,\end{aligned}\tag{1.39}$$

whenever $\delta\to0$.

We get that

$$
\begin{aligned}
&E\left[\left\|\varepsilon_j f_z(\boldsymbol{Z}_j) E\left[\boldsymbol{\tau}_i(\boldsymbol{d}_1)\, \boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d}_1)\boldsymbol{\Phi}^X_{n,ij}(\boldsymbol{d}_1) \mid \boldsymbol{X}_j, \boldsymbol{Z}_j\right] - \varepsilon_j f_z(\boldsymbol{Z}_j) E\left[\boldsymbol{\tau}_i(\boldsymbol{d}_2)\, \boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d}_2)\boldsymbol{\Phi}^X_{n,ij}(\boldsymbol{d}_2) \mid \boldsymbol{X}_j, \boldsymbol{Z}_j\right]\right\|^2\right] \\
&\quad = E\Big[E\left[\varepsilon_j^2 f_z(\boldsymbol{Z}_j)^2 \mid \boldsymbol{X}_j, \boldsymbol{Z}_j\right] \big(\boldsymbol{\tau}_i(\boldsymbol{d}_1)^T\, \boldsymbol{\tau}_k(\boldsymbol{d}_1)\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d}_1)\boldsymbol{\Phi}^X_{n,ij}(\boldsymbol{d}_1)\boldsymbol{\Omega}^Z_{n,kj}(\boldsymbol{d}_1)\boldsymbol{\Phi}^X_{n,kj}(\boldsymbol{d}_1) \\
&\qquad\qquad - 2\boldsymbol{\tau}_i(\boldsymbol{d}_1)^T\, \boldsymbol{\tau}_k(\boldsymbol{d}_2)\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d}_1)\boldsymbol{\Phi}^X_{n,ij}(\boldsymbol{d}_1)\boldsymbol{\Omega}^Z_{n,kj}(\boldsymbol{d}_2)\boldsymbol{\Phi}^X_{n,kj}(\boldsymbol{d}_2) \\
&\qquad\qquad + \boldsymbol{\tau}_i(\boldsymbol{d}_2)^T\, \boldsymbol{\tau}_k(\boldsymbol{d}_2)\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d}_2)\boldsymbol{\Phi}^X_{n,ij}(\boldsymbol{d}_2)\boldsymbol{\Omega}^Z_{n,kj}(\boldsymbol{d}_2)\boldsymbol{\Phi}^X_{n,kj}(\boldsymbol{d}_2)\big)\Big].
\end{aligned}
$$

By the same Fourier transformation arguments as in the proof of Lemma 1.1 and the dominated convergence theorem the statement in (1.39) follows. Therefore,

$$
\sqrt{n}\left((\widehat{\lambda}, \widehat{\boldsymbol{\beta}}(\widehat{\lambda})^T)^T - (\lambda_0, \boldsymbol{\beta}_0^T)^T\right) = -\boldsymbol{V}(\boldsymbol{d})^{-1}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^n \varepsilon_j f_z(\boldsymbol{Z}_j) E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\, \boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d})\boldsymbol{\Phi}^X_{n,ij}(\boldsymbol{d}) \mid \boldsymbol{X}_j, \boldsymbol{Z}_j\right]\right) + o_{\mathbb{P}}(1),
$$

converges in distribution to a tight random process whose marginal distribution is zero-mean normal with covariance function $\boldsymbol{V}(\boldsymbol{d}_1)^{-1}\boldsymbol{\Delta}(\boldsymbol{d}_1, \boldsymbol{d}_2)\boldsymbol{V}(\boldsymbol{d}_2)^{-1}$.

□

Proof of Proposition 1.2.

We have that

$$
\begin{aligned}
n^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda})^T\, \widehat{\mathbb{B}}_n\, n^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda}) &= n^{-1}\widehat{\mathbb{Y}}_n(\lambda_0)^T\, \widehat{\mathbb{B}}_n\, n^{-1}\widehat{\mathbb{Y}}_n(\lambda_0) \\
&\quad + 2n^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)^T\, \widehat{\mathbb{B}}_n\, n^{-1}\widehat{\mathbb{Y}}_n(\lambda_0)\left(\widehat{\lambda} - \lambda_0\right) \\
&\quad + \Bigg[n^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)^T\, \widehat{\mathbb{B}}_n\, n^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0) \\
&\quad + n^{-1}\widehat{\mathbb{Y}}_n(\lambda_0)^T\, \widehat{\mathbb{B}}_n\, n^{-1}\frac{\partial^2}{\partial\lambda^2}\widehat{\mathbb{Y}}_n(\lambda_0) + R_{1,n}(\widetilde{\lambda}, \lambda_0)\Bigg]\left(\widehat{\lambda} - \lambda_0\right)^2,
\end{aligned}
$$

where $\widetilde{\lambda} = c\widehat{\lambda} + (1-c)\lambda_0$ for some $c \in (0,1)$. By the same reasoning as in Proposition 1.1 we get that

$$
\begin{aligned}
n^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda})^T\, \widehat{\mathbb{B}}_n\, n^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda}) &= n^{-1}\widehat{\mathbb{Y}}_n(\lambda_0)^T\, \widehat{\mathbb{B}}_n\, n^{-1}\widehat{\mathbb{Y}}_n(\lambda_0) \\
&\quad + 2n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\, \mathbb{B}_n\, n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\left(\widehat{\lambda} - \lambda_0\right) \\
&\quad + n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\, \mathbb{B}_n\, n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\left(\widehat{\lambda} - \lambda_0\right)^2 \\
&\quad + o_{\mathbb{P}}(1/n),
\end{aligned}
$$

uniformly with respect to $s \in S_n$, $\boldsymbol{d} \in \mathcal{D}$ and $h \in \mathcal{H}_{sc,n}$. Therefore, it follows that under H_0

$$
\begin{aligned}
&n^{-1}\widehat{\mathbb{Y}}_n(\lambda_R)^T\, \widehat{\mathbb{B}}_n\, n^{-1}\widehat{\mathbb{Y}}_n(\lambda_R) - n^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda})^T\, \widehat{\mathbb{B}}_n\, n^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda}) \\
&\qquad = \frac{1}{n^2}\left(\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\, \mathbb{B}_n\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\right)^2\left[\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]^{-1} \\
&\qquad\quad + o_{\mathbb{P}}(1/n) \\
&\qquad = (1, \boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}n^{-2}\boldsymbol{A}_n n^{-2}\boldsymbol{A}_n^T\boldsymbol{V}(\boldsymbol{d})^{-1}(1, \boldsymbol{0}_p^T)^T\left[n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_n n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right] \\
&\qquad\quad + o_{\mathbb{P}}(1/n) \\
&\qquad = (1, \boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}n^{-2}\boldsymbol{A}_n n^{-2}\boldsymbol{A}_n^T\boldsymbol{V}(\boldsymbol{d})^{-1}(1, \boldsymbol{0}_p^T)^T E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right] \\
&\qquad\quad + o_{\mathbb{P}}(1/n).
\end{aligned}
$$

When H_0 does not hold it follows by the same arguments as in the proof of Proposition 1.1 that $n^{-1}DM_\lambda$ converges in probability to a positive constant.

□

Proof of Proposition 1.3.

Under H_0 we get that

$$\begin{aligned}
&\left(\widehat{\mathbb{Y}}_n(\lambda) - \widehat{\mathbb{X}}_n \widehat{\boldsymbol{\beta}}_R(\lambda)\right)^T \mathbb{D}_n \left(\widehat{\mathbb{Y}}_n(\lambda) - \widehat{\mathbb{X}}_n \widehat{\boldsymbol{\beta}}_R(\lambda)\right) \\
&\qquad = \widehat{\mathbb{Y}}_n(\lambda)^T \widehat{\mathbb{B}}_n \widehat{\mathbb{Y}}_n(\lambda) + \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{c}\right)^T \left(\boldsymbol{R}\left(\widehat{\mathbb{X}}_n^T \mathbb{D}_n \widehat{\mathbb{X}}_n\right)^{-1} \boldsymbol{R}^T\right)^{-1} \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{c}\right) \\
&\qquad = \widehat{\mathbb{Y}}_n(\lambda)^T \widehat{\mathbb{B}}_n \widehat{\mathbb{Y}}_n(\lambda) + \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{R}\boldsymbol{\beta}_0\right)^T \left(\boldsymbol{R}\left(\widehat{\mathbb{X}}_n^T \mathbb{D}_n \widehat{\mathbb{X}}_n\right)^{-1} \boldsymbol{R}^T\right)^{-1} \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{R}\boldsymbol{\beta}_0\right) \\
&\qquad = \left(\widehat{\mathbb{Y}}_n(\lambda) - \widehat{\mathbb{X}}_n \boldsymbol{\beta}_0\right)^T \\
&\qquad\qquad \left(\widehat{\mathbb{B}}_n + \mathbb{D}_n \widehat{\mathbb{X}}_n \left(\widehat{\mathbb{X}}_n^T \mathbb{D}_n \widehat{\mathbb{X}}_n\right)^{-1} \boldsymbol{R}^T \left(\boldsymbol{R}\left(\widehat{\mathbb{X}}_n^T \mathbb{D}_n \widehat{\mathbb{X}}_n\right)^{-1} \boldsymbol{R}^T\right)^{-1} \boldsymbol{R}\left(\widehat{\mathbb{X}}_n^T \mathbb{D}_n \widehat{\mathbb{X}}_n\right)^{-1} \widehat{\mathbb{X}}_n^T \mathbb{D}_n\right) \\
&\qquad\qquad\qquad \left(\widehat{\mathbb{Y}}_n(\lambda) - \widehat{\mathbb{X}}_n \boldsymbol{\beta}_0\right) \\
&\qquad = \left(\widehat{\mathbb{Y}}_n(\lambda) - \widehat{\mathbb{X}}_n \boldsymbol{\beta}_0\right)^T \widehat{\mathbb{B}}_{n,R} \left(\widehat{\mathbb{Y}}_n(\lambda) - \widehat{\mathbb{X}}_n \boldsymbol{\beta}_0\right),
\end{aligned}$$

where

$$\widehat{\mathbb{B}}_{n,R} = \widehat{\mathbb{B}}_n + \mathbb{D}_n \widehat{\mathbb{X}}_n \left(\widehat{\mathbb{X}}_n^T \mathbb{D}_n \widehat{\mathbb{X}}_n\right)^{-1} \boldsymbol{R}^T \left(\boldsymbol{R}\left(\widehat{\mathbb{X}}_n^T \mathbb{D}_n \widehat{\mathbb{X}}_n\right)^{-1} \boldsymbol{R}^T\right)^{-1} \boldsymbol{R}\left(\widehat{\mathbb{X}}_n^T \mathbb{D}_n \widehat{\mathbb{X}}_n\right)^{-1} \widehat{\mathbb{X}}_n^T \mathbb{D}_n.$$

Therefore, we get by the same reasoning as in the proof of Proposition 1.1 that

$$\begin{aligned}
\widehat{\lambda}_R - \lambda_0 &= -\left[\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T \mathbb{B}_{n,R} \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]^{-1} \frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T \mathbb{B}_{n,R} \left[(\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right] + o_{\mathbb{P}}(n^{-1/2}) \\
&= -\boldsymbol{V}_R(\boldsymbol{d}) n^{-2} \boldsymbol{A}_n + o_{\mathbb{P}}(n^{-1/2}),
\end{aligned}$$

uniformly with respect to $h \in \mathcal{H}_{sc,n}$, $\boldsymbol{d} \in \mathcal{D}$ and $s \in S_n$. Furthermore, we get that

$$\begin{aligned}
&\left(\widehat{\mathbb{Y}}_n(\widehat{\lambda}_R) - \widehat{\mathbb{X}}_n \widehat{\boldsymbol{\beta}}_R(\widehat{\lambda}_R)\right)^T \mathbb{D}_n \left(\widehat{\mathbb{Y}}_n(\widehat{\lambda}_R) - \widehat{\mathbb{X}}_n \widehat{\boldsymbol{\beta}}_R(\widehat{\lambda}_R)\right) = \left(\widehat{\mathbb{Y}}_n(\widehat{\lambda}_R) - \widehat{\mathbb{X}}_n \boldsymbol{\beta}_0\right)^T \widehat{\mathbb{B}}_{n,R} \left(\widehat{\mathbb{Y}}_n(\widehat{\lambda}_R) - \widehat{\mathbb{X}}_n \boldsymbol{\beta}_0\right) \\
&\quad = \left(\widehat{\mathbb{Y}}_n(\lambda_0) - \widehat{\mathbb{X}}_n \boldsymbol{\beta}_0\right)^T \widehat{\mathbb{B}}_{n,R} \left(\widehat{\mathbb{Y}}_n(\lambda_0) - \widehat{\mathbb{X}}_n \boldsymbol{\beta}_0\right) + 2\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)^T\, \widehat{\mathbb{B}}_{n,R} \left(\widehat{\mathbb{Y}}_n(\lambda_0) - \widehat{\mathbb{X}}_n \boldsymbol{\beta}_0\right)\left(\widehat{\lambda}_R - \lambda_0\right) \\
&\qquad + \left[\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)^T\, \widehat{\mathbb{B}}_{n,R}\, \frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0) + \left(\widehat{\mathbb{Y}}_n(\lambda_0) - \widehat{\mathbb{X}}_n \boldsymbol{\beta}_0\right)^T \widehat{\mathbb{B}}_{n,R}\, \frac{\partial^2}{\partial\lambda^2}\widehat{\mathbb{Y}}_n(\lambda_0) + R_{1,n}(\widetilde{\lambda}, \lambda_0)\right]\left(\widehat{\lambda}_R - \lambda_0\right)^2,
\end{aligned}$$

where $\widetilde{\lambda} = c\widehat{\lambda}_R + (1-c)\lambda_0$ for some $c \in (0,1)$. By the same reasoning as in Proposition 1.1 and using the asymptotic representation of $\left(\widehat{\lambda}_R - \lambda_0\right)$ we get that

$$\begin{aligned}
&n^{-1}\left(\widehat{\mathbb{Y}}_n(\widehat{\lambda}_R) - \widehat{\mathbb{X}}_n \widehat{\boldsymbol{\beta}}_R(\widehat{\lambda}_R)\right)^T \mathbb{D}_n\, n^{-1}\left(\widehat{\mathbb{Y}}_n(\widehat{\lambda}_R) - \widehat{\mathbb{X}}_n \widehat{\boldsymbol{\beta}}_R(\widehat{\lambda}_R)\right) \\
&\qquad = n^{-1}\left(\widehat{\mathbb{Y}}_n(\widehat{\lambda}_R) - \widehat{\mathbb{X}}_n \boldsymbol{\beta}_0\right)^T \widehat{\mathbb{B}}_{n,R}\, n^{-1}\left(\widehat{\mathbb{Y}}_n(\widehat{\lambda}_R) - \widehat{\mathbb{X}}_n \boldsymbol{\beta}_0\right) \\
&\qquad = n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)^T \mathbb{B}_{n,R}\, n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right) \\
&\qquad\quad + 2n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\, \mathbb{B}_{n,R}\, n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\left(\widehat{\lambda}_R - \lambda_0\right) \\
&\qquad\quad + \left[n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\, \mathbb{B}_{n,R}\, n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\left(\widehat{\lambda}_R - \lambda_0\right)^2 + o_{\mathbb{P}}(1/n) \\
&\qquad = n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)^T \mathbb{B}_{n,R}\, n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right) \\
&\qquad\quad - n^{-2}\boldsymbol{A}_n^T \boldsymbol{V}_R(\boldsymbol{d})^T \boldsymbol{V}_R(\boldsymbol{d}) \boldsymbol{A}_n n^{-2}\left[n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T \mathbb{B}_{n,R} n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right] + o_{\mathbb{P}}(1/n),
\end{aligned}$$

uniformly with respect to $s \in S_n$, $\boldsymbol{d} \in \mathcal{D}$ and $h \in \mathcal{H}_{sc,n}$. We know from the proof of Proposition 1.2

that

$$\begin{aligned} n^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda})^T\,\widehat{\mathbb{B}}_n\,n^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda}) = {} & n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)^T\,\mathbb{B}_n n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\\ & -n^{-2}\boldsymbol{A}_n^T\boldsymbol{V}(\boldsymbol{d})^{-1}(1,\boldsymbol{0}_p^T)^T(1,\boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}\boldsymbol{A}_n n^{-2}\\ & \qquad\qquad\left[n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\,\mathbb{B}_n\,n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\\ & +o_{\mathbb{P}}\left(1/n\right). \end{aligned}$$

Therefore, we get that

$$\begin{aligned} & n^{-1}\left(\widehat{\mathbb{Y}}_n(\widehat{\lambda}_R)-\widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R(\widehat{\lambda}_R)\right)^T\mathbb{D}_n\,n^{-1}\left(\widehat{\mathbb{Y}}_n(\widehat{\lambda}_R)-\widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R(\widehat{\lambda}_R)\right)-n^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda})^T\,\widehat{\mathbb{B}}_n\,n^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda})\\ &= n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)^T\left(\mathbb{B}_{n,R}-\mathbb{B}_n\right)\,n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\\ &\quad -n^{-2}\boldsymbol{A}_n^T\boldsymbol{V}_R(\boldsymbol{d})^T\boldsymbol{V}_R(\boldsymbol{d})\boldsymbol{A}_n n^{-2}\left[n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\,\mathbb{B}_{n,R}\,n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\\ &\quad +n^{-2}\boldsymbol{A}_n^T\boldsymbol{V}(\boldsymbol{d})^{-1}(1,\boldsymbol{0}_p^T)^T(1,\boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}\boldsymbol{A}_n n^{-2}\left[n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\,\mathbb{B}_n\,n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\\ &\quad +o_{\mathbb{P}}\left(1/n\right)\\ &= n^{-2}\boldsymbol{A}_n^T\left(\boldsymbol{0}_{p\times1},\boldsymbol{I}_{p\times p}\right)^T\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\left(\boldsymbol{0}_{p\times1},\boldsymbol{I}_{p\times p}\right)\boldsymbol{A}_n\\ &\quad -n^{-2}\boldsymbol{A}_n^T\boldsymbol{V}_R(\boldsymbol{d})^T\boldsymbol{V}_R(\boldsymbol{d})\boldsymbol{A}_n n^{-2}\left[n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\,\mathbb{B}_{n,R}\,n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\\ &\quad +n^{-2}\boldsymbol{A}_n^T\boldsymbol{V}(\boldsymbol{d})^{-1}(1,\boldsymbol{0}_p^T)^T(1,\boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}\boldsymbol{A}_n n^{-2}\left[n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\,\mathbb{B}_n\,n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\\ &\quad +o_{\mathbb{P}}\left(1/n\right)\\ &= n^{-2}\boldsymbol{A}_n^T\left(\boldsymbol{0}_{p\times1},\boldsymbol{I}_{p\times p}\right)^T E\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}E\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\left(\boldsymbol{0}_{p\times1},\boldsymbol{I}_{p\times p}\right)\boldsymbol{A}_n n^{-2}\\ &\quad -n^{-2}\boldsymbol{A}_n^T\boldsymbol{V}_R(\boldsymbol{d})^T\boldsymbol{V}_R(\boldsymbol{d})\boldsymbol{A}_n n^{-2}E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\,\mathbb{B}_{n,R}\,\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\\ &\quad +n^{-2}\boldsymbol{A}_n^T\boldsymbol{V}(\boldsymbol{d})^{-1}(1,\boldsymbol{0}_p^T)^T(1,\boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}\boldsymbol{A}_n n^{-2}E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\,\mathbb{B}_n\,\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\\ &\quad +o_{\mathbb{P}}\left(1/n\right), \end{aligned}$$

uniformly with respect to $h\in\mathcal{H}_{sc,n}$, $\boldsymbol{d}\in\mathcal{D}$ and $s\in S_n$. When H_0 does not hold it follows by the same arguments as in the proof of Proposition 1.1 that $n^{-1}DM_{\boldsymbol{\beta}}$ converges in probability to a positive constant.

□

Proof of Proposition 1.4.

We can use the arguments as in the proof of Proposition 1.3. The only difference is that we do not need to taylor

$$n^{-1}\left(\widehat{\mathbb{Y}}_n(\lambda_R)-\widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R(\lambda_R)\right)^T\mathbb{D}_n\,n^{-1}\left(\widehat{\mathbb{Y}}_n(\lambda_R)-\widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R(\lambda_R)\right),$$

as λ_R is fixed. Therefore, we get that under H_0

$$\begin{aligned} & n^{-1}\left(\widehat{\mathbb{Y}}_n(\lambda_R)-\widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R(\lambda_R)\right)^T\mathbb{D}_n\,n^{-1}\left(\widehat{\mathbb{Y}}_n(\lambda_R)-\widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R(\lambda_R)\right)-n^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda})^T\,\widehat{\mathbb{B}}_n\,n^{-1}\widehat{\mathbb{Y}}_n(\widehat{\lambda})\\ &= n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)^T\left(\mathbb{B}_{n,R}-\mathbb{B}_n\right)\,n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)\\ &\quad +n^{-2}\boldsymbol{A}_n^T\boldsymbol{V}(\boldsymbol{d})^{-1}(1,\boldsymbol{0}_p^T)^T(1,\boldsymbol{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}\boldsymbol{A}_n n^{-2}\left[n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\,\mathbb{B}_n\,n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\\ &\quad +o_{\mathbb{P}}\left(1/n\right), \end{aligned}$$

$$
\begin{aligned}
&= n^{-2}\boldsymbol{A}_n^T\left(\mathbf{0}_{p\times 1},\boldsymbol{I}_{p\times p}\right)^T\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\left(\mathbf{0}_{p\times 1},\boldsymbol{I}_{p\times p}\right)\boldsymbol{A}_n\\
&\quad + n^{-2}\boldsymbol{A}_n^T\boldsymbol{V}(\boldsymbol{d})^{-1}(1,\mathbf{0}_p^T)^T(1,\mathbf{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}\boldsymbol{A}_n n^{-2}\left[n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\,\mathbb{B}_n\,n^{-1}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\\
&\quad + o_{\mathbb{P}}\left(1/n\right)\\
&= n^{-2}\boldsymbol{A}_n^T\left(\mathbf{0}_{p\times 1},\boldsymbol{I}_{p\times p}\right)^T E\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}E\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\left(\mathbf{0}_{p\times 1},\boldsymbol{I}_{p\times p}\right)\boldsymbol{A}_n n^{-2}\\
&\quad + n^{-2}\boldsymbol{A}_n^T\boldsymbol{V}(\boldsymbol{d})^{-1}(1,\mathbf{0}_p^T)^T(1,\mathbf{0}_p^T)\boldsymbol{V}(\boldsymbol{d})^{-1}\boldsymbol{A}_n n^{-2}E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\,\mathbb{B}_n\,\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\\
&\quad + o_{\mathbb{P}}\left(1/n\right),
\end{aligned}
$$

uniformly with respect to $h \in \mathcal{H}_{sc,n}$, $\boldsymbol{d} \in \mathcal{D}$ and $s \in S_n$. When H_0 does not hold it follows by the same arguments as in the proof of Proposition 1.1 that $n^{-1}DM_{\boldsymbol{\beta},\lambda}$ converges in probability to a positive constant.

□

Lemma 1.3. *Let Assumptions 1.1.1 and 1.1.3 hold. Then*

$$\sup_{\boldsymbol{d}\in\mathcal{D}} \|\mathbb{D}_n\|_{\mathrm{Sp}} \leq \sup_{\boldsymbol{d}\in\mathcal{D}} \|\boldsymbol{\Omega}_n\|_{\mathrm{Sp}} \leq n.$$

Moreover,

$$\sup_{\boldsymbol{d}\in\mathcal{D}} \|\mathbb{B}_n\|_{\mathrm{Sp}} \leq \sup_{\boldsymbol{d}\in\mathcal{D}} \|\mathbb{D}_n\|_{\mathrm{Sp}} \qquad \text{and} \qquad \sup_{h>0}\sup_{\boldsymbol{d}\in\mathcal{D}} \left\|\widehat{\mathbb{B}}_n\right\|_{\mathrm{Sp}} \leq \sup_{\boldsymbol{d}\in\mathcal{D}} \|\mathbb{D}_n\|_{\mathrm{Sp}}\,.$$

Proof of Lemma 1.3.

For all vectors $\boldsymbol{d}$, the matrix $\boldsymbol{\Omega}_n$ is positive definite, see Lemma 1.2. This implies that its spectral norm is equal to the largest eigenvalue. On the other hand, for all vectors $\boldsymbol{d}$, the trace of $\boldsymbol{\Omega}_n$ is equal to n. Necessarily, the spectral norm of $\boldsymbol{\Omega}_n$ is at most equal to n, uniformly with respect to $\boldsymbol{d}\in\mathcal{D}$. Next, it is easy to see that $\|\boldsymbol{A}_1\|_{\mathrm{Sp}} \leq \|\boldsymbol{A}_2\|_{\mathrm{Sp}}$ whenever $\boldsymbol{A}_1$ and $\boldsymbol{A}_2-\boldsymbol{A}_1$ are positive semi-definite real matrices. Using repeatedly this property and the fact that $\mathbb{D}_n$, $\mathbb{B}_n$ and $\widehat{\mathbb{B}}_n$ are positive semi-definite (cf. proof of Lemma 1.2), we deduce the remaining inequalities, that clearly hold uniformly.

□

Lemma 1.4.

1. *For any positive definite real matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$*

$$\|\boldsymbol{A}_2^{-1/2} - \boldsymbol{A}_1^{-1/2}\|_{\mathrm{Sp}} \leq \frac{1}{2}\left[\max\{\|\boldsymbol{A}_1^{-1}\|_{\mathrm{Sp}}, \|\boldsymbol{A}_2^{-1}\|_{\mathrm{Sp}}\}\right]^{3/2} \|\boldsymbol{A}_2-\boldsymbol{A}_1\|_{\mathrm{Sp}}.$$

2. *Let $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ be $n\times p-$matrices such that $\boldsymbol{A}_1^T\boldsymbol{A}_1 = \boldsymbol{A}_2^T\boldsymbol{A}_2 = \boldsymbol{I}_{p\times p}$. Then*

$$\left\|\boldsymbol{A}_1\boldsymbol{A}_1^T - \boldsymbol{A}_2\boldsymbol{A}_2^T\right\|_{\mathrm{Sp}} \leq 2\left\|\boldsymbol{A}_1-\boldsymbol{A}_2\right\|_{\mathrm{Sp}}.$$

Proof of Lemma 1.4.

1. For any positive definite real matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$

$$\|\boldsymbol{A}_2^{1/2} - \boldsymbol{A}_1^{1/2}\|_{\mathrm{Sp}} \leq \frac{1}{2}\left[\max\{\|\boldsymbol{A}_1^{-1}\|_{\mathrm{Sp}}, \|\boldsymbol{A}_2^{-1}\|_{\mathrm{Sp}}\}\right]^{1/2} \|\boldsymbol{A}_2-\boldsymbol{A}_1\|_{\mathrm{Sp}},$$

(see for instance Horn and Johnson [44], page 557). Moreover, for any invertible matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ we have the identity $\boldsymbol{A}_2^{-1}-\boldsymbol{A}_1^{-1} = \boldsymbol{A}_2^{-1}(\boldsymbol{A}_1-\boldsymbol{A}_2)\boldsymbol{A}_1^{-1}$. Apply this identity with $\boldsymbol{A}_1^{1/2}$ and $\boldsymbol{A}_2^{1/2}$ and, using the fact that the spectral norm of a product of two matrices is smaller or equal to the product of the matrices' spectral norms, we deduce the statement.

2. We could write

$$\begin{aligned}\left\|(\boldsymbol{A}_1\boldsymbol{A}_1^T - \boldsymbol{A}_2\boldsymbol{A}_2^T)\boldsymbol{u}\right\| &= \left\|\boldsymbol{A}_1(\boldsymbol{A}_1-\boldsymbol{A}_2)^T\boldsymbol{u} + (\boldsymbol{A}_1-\boldsymbol{A}_2)\boldsymbol{A}_2^T\boldsymbol{u}\right\| \\ &\leq \left(\|\boldsymbol{A}_1\|_{\mathrm{Sp}} + \|\boldsymbol{A}_2\|_{\mathrm{Sp}}\right)\|\boldsymbol{A}_1-\boldsymbol{A}_2\|_{\mathrm{Sp}}\|\boldsymbol{u}\|\,.\end{aligned}$$

Moreover, $\|\boldsymbol{A}_1\boldsymbol{u}\|^2 = \boldsymbol{u}^T\boldsymbol{A}_1^T\boldsymbol{A}_1\boldsymbol{u} = \|\boldsymbol{u}\|^2$, and thus $\|\boldsymbol{A}_1\|_{\mathrm{Sp}} = \|\boldsymbol{A}_2\|_{\mathrm{Sp}} = 1$. Thus, $2\|\boldsymbol{A}_1-\boldsymbol{A}_2\|_{\mathrm{Sp}}$ is a bound for the norm of the difference between the orthogonal projectors defined respectively by $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$.

□

Lemma 1.5. *If the Assumptions 1.1.1, 1.1.3, 1.2.1 and 1.2.5 hold true, $E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]$ tends to a positive definite $p\times p-$matrix and*

$$\sup_{\boldsymbol{d}\in\mathcal{D}} \left\|n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n - E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]\right\|_{\mathrm{Sp}} = O_{\mathbb{P}}(n^{-1/2}).$$

If in addition Assumption 1.4.1 holds true, $E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]$ tends to a positive constant and

$$\sup_{\boldsymbol{d}\in\mathcal{D}} \left\| n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{D}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0) - E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{D}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\right\| = O_{\mathbb{P}}(n^{-1/2}) \qquad \textit{and}$$
$$\sup_{\boldsymbol{d}\in\mathcal{D}} \left\| n^{-2}\mathbb{X}_n^T\mathbb{D}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0) - E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\right\| = O_{\mathbb{P}}(n^{-1/2}).$$

Proof of Lemma 1.5.

First, we investigate the behavior of $n^{-2}\mathbb{X}_n^T\boldsymbol{\Omega}_n\mathbb{X}_n$ that we decompose.

$$\frac{1}{n^2}\mathbb{X}_n^T\boldsymbol{\Omega}_n\mathbb{X}_n = \frac{n-1}{n}\frac{1}{n(n-1)}\sum_{1\le i\ne j\le n}\boldsymbol{O}_{ij} + \frac{1}{n^2}\sum_{1\le i\le n}(\boldsymbol{X}_i - E[\boldsymbol{X}_i \mid \boldsymbol{Z}_i])(\boldsymbol{X}_i - E[\boldsymbol{X}_i \mid \boldsymbol{Z}_i])^T f_z^2(\boldsymbol{Z}_i),$$

where $\boldsymbol{O}_{ij} = \boldsymbol{O}_{ij}(\boldsymbol{d}) = (\boldsymbol{X}_i - E[\boldsymbol{X}_i \mid \boldsymbol{Z}_i])(\boldsymbol{X}_j - E[\boldsymbol{X}_j \mid \boldsymbol{Z}_j])^T f_z(\boldsymbol{Z}_i) f_z(\boldsymbol{Z}_j)\boldsymbol{\Omega}_{n,ij}$. It is obvious that under our assumptions the second sum, corresponding to the diagonal terms of the quadratic form $\mathbb{X}_n^T\boldsymbol{\Omega}_n\mathbb{X}_n$, has the rate $O_{\mathbb{P}}(n^{-1})$. On the other hand, for any $\boldsymbol{d}\in\mathcal{D}$ and any $\boldsymbol{u}\in\mathbb{R}^p$, using the Fourier Transform and the monotonicity of the exponential function, we have

$$\begin{aligned}
E[\boldsymbol{u}^T\boldsymbol{O}_{ij}(\boldsymbol{d})\boldsymbol{u}] &= E\left[\boldsymbol{u}^T(\boldsymbol{X}_i - E[\boldsymbol{X}_i \mid \boldsymbol{Z}_i])(\boldsymbol{X}_j - E[\boldsymbol{X}_j \mid \boldsymbol{Z}_j])^T\boldsymbol{u} f_z(\boldsymbol{Z}_i) f_z(\boldsymbol{Z}_j)\boldsymbol{\Omega}_{n,ij}\right] \\
&= \frac{\pi^{-(p+q)/2}}{\sqrt{d_1\cdots d_{p+q}}}\int_{\mathbb{R}^{p+q}}\left|E\left[\boldsymbol{u}^T(\boldsymbol{X} - E[\boldsymbol{X}\mid\boldsymbol{Z}])f_z(\boldsymbol{Z})\exp\left\{2i\boldsymbol{w}^T\left(\boldsymbol{X}^T,\boldsymbol{Z}^T\right)^T\right\}\right]\right|^2\exp\left\{-\boldsymbol{w}^T\boldsymbol{D}^{-1}\boldsymbol{w}\right\}d\boldsymbol{w} \\
&\ge \frac{\pi^{-(p+q)/2}}{d_U^{(p+q)/2}} \\
&\quad\times\int_{\mathbb{R}^{p+q}}\left|E\left[\boldsymbol{u}^T(\boldsymbol{X} - E[\boldsymbol{X}\mid\boldsymbol{Z}])f_z(\boldsymbol{Z})\exp\left\{2i\boldsymbol{w}^T\left(\boldsymbol{X}^T,\boldsymbol{Z}^T\right)^T\right\}\right]\right|^2\exp\left\{-\boldsymbol{w}^T\mathrm{diag}(d_L,\ldots,d_L)^{-1}\boldsymbol{w}\right\}d\boldsymbol{w} \\
&= \frac{d_L^{(p+q)/2}}{d_U^{(p+q)/2}}E\left[\boldsymbol{u}^T\boldsymbol{O}_{ij}(\mathrm{diag}(d_L,\ldots,d_L)^{-1})\boldsymbol{u}\right],
\end{aligned}$$

where d_U is the upper bound and d_L the lower bound of the values on the diagonal of $\boldsymbol{D}$. Since by Assumption 1.2.1 the variable $\boldsymbol{u}^T(\boldsymbol{X} - E[\boldsymbol{X}\mid\boldsymbol{Z}])$ could not be equal to zero almost surely, we necessarily have $E[\boldsymbol{u}^T\boldsymbol{O}_{ij}(\boldsymbol{d})\boldsymbol{u}] > 0$ and thus, $E[\boldsymbol{O}_{ij}(\boldsymbol{d})]$ is positive definite. Moreover, it is clear from the last display that there exists a constant $C > 0$ such that $E[\boldsymbol{O}_{ij}(\boldsymbol{d})] - C\boldsymbol{I}_{p\times p}$ is positive definite for each $\boldsymbol{d}\in\mathcal{D}$. By the uniform convergence results of Sherman [72],

$$\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\frac{1}{n^2}\mathbb{X}_n^T\boldsymbol{\Omega}_n\mathbb{X}_n - E[\boldsymbol{O}_{ij}(\boldsymbol{d})]\right\|_{\mathrm{Sp}} = O_{\mathbb{P}}(n^{-1/2}).$$

Next, we derive the convergence of $E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]$. Let us decompose

$$\mathbb{D}_n = \boldsymbol{\Omega}_n^{1/2}\left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\boldsymbol{1}_n}\right)\boldsymbol{\Omega}_n^{1/2},$$

where

$$P_{\boldsymbol{\Omega}_n^{1/2}\boldsymbol{1}_n} = \frac{1}{\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n}\boldsymbol{\Omega}_n^{1/2}\boldsymbol{1}_n\boldsymbol{1}_n^T\boldsymbol{\Omega}_n^{1/2}.$$

(Here, $\boldsymbol{\Omega}_n^{1/2}$ is the positive definite square root of $\boldsymbol{\Omega}_n$.) Let us define

$$P^0_{\boldsymbol{\Omega}_n^{1/2}\boldsymbol{1}_n} = \frac{1}{n^{-2}E\left[\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n\right]}\boldsymbol{\Omega}_n^{1/2}n^{-1}\boldsymbol{1}_n n^{-1}\boldsymbol{1}_n^T\boldsymbol{\Omega}_n^{1/2}.$$

It is clear from above that $n^{-2}E\left[\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n\right]$ converges at the rate $O_{\mathbb{P}}(n^{-1})$ to a strictly positive limit and $n^{-2}\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n - n^{-2}E\left[\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n\right] = O_{\mathbb{P}}(n^{-1/2})$, uniformly with respect to $\boldsymbol{d}\in\mathcal{D}$. Thus

$$\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\frac{1}{\left(n^{-2}E\left[\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n\right]\right)^{1/2}}\boldsymbol{\Omega}_n^{1/2}n^{-1}\boldsymbol{1}_n - \frac{1}{\left(n^{-2}\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n\right)^{1/2}}\boldsymbol{\Omega}_n^{1/2}n^{-1}\boldsymbol{1}_n\right\| = O_{\mathbb{P}}(n^{-1/2}).$$

Then, it follows that

$$\left\|P_{\boldsymbol{\Omega}_n^{1/2}\boldsymbol{1}_n} - P^0_{\boldsymbol{\Omega}_n^{1/2}\boldsymbol{1}_n}\right\|_{\mathrm{Sp}} = O_{\mathbb{P}}(n^{-1/2}).$$

Hence, in order to show that asymptotically the spectrum of $E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]$ stays away from zero, it suffices to show that the spectrum of the $p\times p-$matrix $E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n^0\mathbb{X}_n\right]$ stays away from zero, where

$$\mathbb{D}_n^0 = \boldsymbol{\Omega}_n^{1/2}\left(\boldsymbol{I}_{n\times n} - P^0_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\boldsymbol{\Omega}_n^{1/2}.$$

For any $\boldsymbol{u}\in\mathbb{R}^p$ we have $E\left[n^{-2}\boldsymbol{u}^T\mathbb{X}_n^T\mathbb{D}_n^0\mathbb{X}_n\boldsymbol{u}\right] = \Delta_n/E\left[n^{-2}\mathbf{1}_n^T\boldsymbol{\Omega}_n\mathbf{1}_n\right]$ with

$$\Delta_n = \Delta_n(\boldsymbol{u}) = E\left[\left\|n^{-1}\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n\right\|^2\right]E\left[\left\|n^{-1}\boldsymbol{\Omega}_n^{1/2}\mathbb{X}_n\boldsymbol{u}\right\|^2\right] - E\left[\left|\left\langle n^{-1}\boldsymbol{\Omega}_n^{1/2}\mathbb{X}_n\boldsymbol{u}, n^{-1}\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n\right\rangle\right|^2\right].$$

We aim showing that, for any fixed $\boldsymbol{u}\in\mathbb{R}^p$, $\Delta_n/E\left[n^{-2}\mathbf{1}_n^T\boldsymbol{\Omega}_n\mathbf{1}_n\right]$ stays away from zero, uniformly with respect to $\boldsymbol{d}$. This will imply that the limit of $E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n^0\mathbb{X}_n\right]$ is a positive $p\times p-$matrix. Consider the second order polynomial

$$P_n(t) = P_n(t;\boldsymbol{u}) = E\left[\left\|n^{-1}\boldsymbol{\Omega}_n^{1/2}\mathbb{X}_n\boldsymbol{u} + tn^{-1}\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n\right\|^2\right] = E\left[\left\|n^{-1}\boldsymbol{\Omega}_n^{1/2}\mathbb{X}_n\boldsymbol{u}\right\|^2\right]$$
$$+ 2tE\left[\left\langle n^{-1}\boldsymbol{\Omega}_n^{1/2}\mathbb{X}_n\boldsymbol{u}, n^{-1}\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n\right\rangle\right] + t^2E\left[\left\|n^{-1}\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n\right\|^2\right] \geq 0.$$

By elementary properties of second order polynomials, the minimal value of $P_n(t)$ is $\Delta_n/E\left[n^{-2}\mathbf{1}_n^T\boldsymbol{\Omega}_n\mathbf{1}_n\right]$. If the minimal value of $P_n(t)$ goes to zero, then necessarily

$$\inf_t E\left[n^{-2}\left(\mathbb{X}_n\boldsymbol{u}+t\mathbf{1}_n\right)^T\boldsymbol{\Omega}_n\left(\mathbb{X}_n\boldsymbol{u}+t\mathbf{1}_n\right)\right]\to 0,$$

uniformly with respect to $\boldsymbol{d}\in\mathcal{D}$. From the first part of the proof we could deduce that this contradicts Assumption 1.2.1. Thus, necessarily the spectrum of the $p\times p-$matrix $E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n^0\mathbb{X}_n\right]$ stays away from zero. Finally, to derive the rate of uniform convergence of $n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n$, we could use again the uniform convergence results of Sherman [72] after removing the diagonal terms, and next study the part given by the diagonal terms. The details are omitted.

Next, we derive the convergence of $E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]$. We get that

$$\mathbb{D}_n = \boldsymbol{\Omega}_n^{1/2}\left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\boldsymbol{\Omega}_n^{1/2} = \mathbb{S}_n^T\mathbb{S}_n,$$

where

$$\mathbb{S}_n = \left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\boldsymbol{\Omega}_n^{1/2}.$$

In addition, let $\mathbb{W}_n = \mathbb{S}_n\mathbb{X}_n\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\mathbb{X}_n^T\mathbb{D}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)$. Therefore, it follows that

$$E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right] = E\left[\left\|n^{-1}\mathbb{S}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right\|^2\right] - E\left[n^{-2}\mathbb{W}_n^T\mathbb{W}_n\right].$$

Consider now the second order polynomial

$$P_n(t) = E\left[\left\|n^{-1}\mathbb{S}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0) + tn^{-1}\mathbb{W}_n\right\|^2\right] = E\left[\left\|n^{-1}\mathbb{S}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right\|^2\right]$$
$$+ 2tE\left[\left\langle n^{-1}\mathbb{W}_n, n^{-1}\mathbb{S}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right\rangle\right] + t^2E\left[n^{-2}\mathbb{W}_n^T\mathbb{W}_n\right] \geq 0.$$

By elementary properties of second order polynomials, the minimal value of $P_n(t)$ is $E\left[\left\|n^{-1}\mathbb{S}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right\|^2\right]-E\left[n^{-2}\mathbb{W}_n^T\mathbb{W}_n\right]$. If the minimal value of $P_n(t)$ goes to zero, then necessarily

$$\inf_t E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\left(\boldsymbol{I}_{n\times n} + t\mathbb{D}_n\mathbb{X}_n\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\mathbb{X}_n^T\right)\mathbb{D}_n\left(\boldsymbol{I}_{n\times n} + t\mathbb{X}_n\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\mathbb{X}_n^T\mathbb{D}_n\right)\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]\to 0,$$

uniformly with respect to $\boldsymbol{d}\in\mathcal{D}$. Note that by the same reasoning as for $E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]$ we get that $E\left[n^{-2}\boldsymbol{u}^T\mathbb{D}_n\boldsymbol{u}\right] > 0$ for all $\boldsymbol{u}\in\mathbb{R}^p$ with $\boldsymbol{u}\neq\mathbf{0}$. Therefore, we could deduce that the upper statement contradicts Assumption 1.4.1. Thus, necessarily $E\left[n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{B}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)\right]$ stays away from zero.

Finally, to derive the rates of uniform convergence of $n^{-2}\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)^T\mathbb{D}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)$ and $n^{-2}\mathbb{X}_n^T\mathbb{D}_n\frac{\partial}{\partial\lambda}\mathbb{Y}_n(\lambda_0)$, we could use again the uniform convergence results of Sherman [72] after removing the diagonal terms, and next study the part given by the diagonal terms. The details are omitted. Now the proof is complete. □

Lemma 1.6. *Under the conditions of Theorem 1.1,*

$$\sup_{h\in\mathcal{H}_{c,n}} \frac{1}{\sqrt{n}} \left\| \widehat{\mathbb{X}}_n - \mathbb{X}_n \right\|_{\mathrm{Sp}} = o_{\mathbb{P}}(1).$$

Proof of Lemma 1.6.

In order to prove the statement we consider

$$\frac{1}{n} \left\| \left(\widehat{\mathbb{X}}_n - \mathbb{X}_n \right) \boldsymbol{u} \right\|^2,$$

where $\boldsymbol{u} \in \mathbb{R}^p$ and $\|\boldsymbol{u}\| = 1$. In the remaining of the proof we set without loss of generality $p = 1$ to keep the notation simple, i.e. we consider

$$\frac{1}{n} \left\| \widehat{\mathbb{X}}_n - \mathbb{X}_n \right\|^2.$$

We have that, for $1 \le i \le n$,

$$\begin{aligned}
\widehat{\mathbb{X}}_{n,i} - \mathbb{X}_{n,i} &= (X_i - \widehat{E}[X_i \mid \boldsymbol{Z}_i])\widehat{f}_z(\boldsymbol{Z}_i) - (X_i - E[X_i \mid \boldsymbol{Z}_i])f_z(\boldsymbol{Z}_i) \\
&= \frac{1}{n}\sum_{j=1}^{n} (X_i - X_j)\, K_{h,ij} - (X_i - E[X_i \mid \boldsymbol{Z}_i])f_z(\boldsymbol{Z}_i) \\
&= X_i \frac{1}{n} \sum_{j=1, j\neq i}^{n} (K_{h,ij} - f_z(\boldsymbol{Z}_i)) - \frac{1}{n} X_i f_z(\boldsymbol{Z}_i) \\
&\quad + \frac{1}{n} \sum_{j=1,j\neq i}^{n} (E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_j K_{h,ij}) + \frac{1}{n} E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i).
\end{aligned}$$

We start by considering

$$\begin{aligned}
&\frac{1}{n} \left\| \left(X_1 \frac{1}{n} \sum_{j=1,j\neq 1}^{n} (K_{h,1j} - f_z(\boldsymbol{Z}_1)), \ldots, X_n \frac{1}{n} \sum_{j=1,j\neq n}^{n} (K_{h,nj} - f_z(\boldsymbol{Z}_n)) \right)^T \right\|^2 \\
&= \frac{1}{n^3} \sum_{1\le i\neq j\le n}^{n} X_i^2 \left(K_{h,ij} - f_z(\boldsymbol{Z}_i)\right)^2 + \frac{1}{n^3} \sum_{1\le i\neq j\neq k\le n}^{n} X_i^2 \left(K_{h,ij} - f_z(\boldsymbol{Z}_i)\right)\left(K_{h,ik} - f_z(\boldsymbol{Z}_i)\right) \\
&= A_n + B_n.
\end{aligned}$$

It is easy to check that $\sup_{h\in\mathcal{H}_{c,n}} |A_n| = o_{\mathbb{P}}(1)$. We show in the following that

$$\sup_{h\in\mathcal{H}_{c,n}} |B_n| = o_{\mathbb{P}}(1), \tag{1.40}$$

as well. Note that $\frac{n^3}{(n)_3} B_n$ is a $U-$process of order 3, where $(n)_k = n(n-1)\ldots(n-k+1)$.

For this $U-$process we compute the mean and use the Hoeffding decomposition. The kernel of B_n is not symmetric in its arguments. However, we could apply the usual symmetrization idea. For instance, for a second order $U-$statistic defined by a kernel $h(\boldsymbol{U}_i, \boldsymbol{U}_j)$, we could replace it by the symmetric kernel $\frac{1}{2}\left[h(\boldsymbol{U}_i, \boldsymbol{U}_j) + h(\boldsymbol{U}_j, \boldsymbol{U}_i)\right]$ from which we get the same $U-$statistic. Here, $\boldsymbol{U}_i = \left(X_i, \boldsymbol{Z}_i^T\right)^T$. We can proceed in the same way by considering all 3! permutations of the variables for B_n so that we can apply the Hoeffding decomposition. Thus, by abuse, we will proceed as if the kernel of the $U-$statistic we handle is symmetric. For simpler notation, we use E_i, $E_{i,j}, \ldots$ for the conditional expectations $E\left[\cdot \mid \boldsymbol{U}_i\right]$, $E\left[\cdot \mid \boldsymbol{U}_i, \boldsymbol{U}_j\right], \ldots$.

In addition, we have that $\{(x_i, \boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{z}_k) \mapsto x_i^2 \left(K_{h,ij} - f_z(\boldsymbol{z}_i)\right)\left(K_{h,ik} - f_z(\boldsymbol{z}_i)\right) : h \in \mathcal{H}_{c,n}\}$ is Euclidean for a squared integrable envelope. See Lemma 22 in Nolan and Pollard [67] and Lemma 2.14 in Pakes and Pollard [68]. Therefore, we can in the following repeatedly apply Corollary 7 and the Maximal Inequality of Sherman [72]. All remainder terms are controlled by Assumption 1.3.2.

We start by considering the mean. We get that

$$\begin{aligned}E\left[X_i^2\left(K_{h,ij}-f_z(\boldsymbol{Z}_i)\right)\left(K_{h,ik}-f_z(\boldsymbol{Z}_i)\right)\right] &= E\left[X_i^2\left(K_{h,ij}-f_z(\boldsymbol{Z}_i)\right)E_{i,j}\left[\left(K_{h,ik}-f_z(\boldsymbol{Z}_i)\right)\right]\right]\\ &= E\left[X_i^2\left(K_{h,ij}-f_z(\boldsymbol{Z}_i)\right)h^2\gamma_1(\boldsymbol{Z}_i)\right](1+o(1))\\ &= h^4E\left[X_i^2\gamma_1(\boldsymbol{Z}_i)^2\right](1+o(1)),\end{aligned}$$

where

$$\gamma_1(\boldsymbol{Z})=\mu(K)\cdot\operatorname{tr}\{\boldsymbol{H}_{z,z}f_z(\boldsymbol{Z})\},$$

with $\int_{\mathbb{R}^q}\boldsymbol{u}\boldsymbol{u}^TK(\boldsymbol{u})d\boldsymbol{u}=\mu(K)\boldsymbol{I}_{q\times q}$. $\boldsymbol{H}_{z,z}f_z$ denotes the matrix of second derivative of $f_z(\cdot)$ with respect to the components of $\boldsymbol{Z}\in\mathbb{R}^q$ and $\operatorname{tr}\{\cdot\}$ denotes the trace operator. Therefore, it follows that the mean of B_n is of order $o_{\mathbb{P}}(1)$ uniformly with respect to h.

We consider now the three first order $U-$processes of the Hoeffding decomposition of B_n. We get that, by the same reasoning as for the mean,

$$E_i\left[X_i^2\left(K_{h,ij}-f_z(\boldsymbol{Z}_i)\right)\left(K_{h,ik}-f_z(\boldsymbol{Z}_i)\right)\right]=h^4X_i^2\gamma_1(\boldsymbol{Z}_i)^2(1+o_{\mathbb{P}}(1)).$$

In addition, we get that

$$\begin{aligned}E_j\left[X_i^2\left(K_{h,ij}-f_z(\boldsymbol{Z}_i)\right)\left(K_{h,ik}-f_z(\boldsymbol{Z}_i)\right)\right] &= h^2E_j\left[X_i^2\left(K_{h,ij}-f_z(\boldsymbol{Z}_i)\right)\gamma_1(\boldsymbol{Z}_i)\right](1+o_{\mathbb{P}}(1))\\ &= \left(h^2E_j\left[X_i^2K_{h,ij}\gamma_1(\boldsymbol{Z}_i)\right]-h^2E\left[X_i^2f_z(\boldsymbol{Z}_i)\gamma_1(\boldsymbol{Z}_i)\right]\right)(1+o_{\mathbb{P}}(1))\\ &= h^2E\left[X_j^2\mid\boldsymbol{Z}_j\right]\gamma_1(\boldsymbol{Z}_j)f_z(\boldsymbol{Z}_j)+O_{\mathbb{P}}(h^4)+O_{\mathbb{P}}(h^2)+o_{\mathbb{P}}(h^2).\end{aligned}$$

The reasoning when conditioning on $\boldsymbol{U}_k$ is the same. Therefore, it follows together with Corollary 4 of Sherman [72] that the first order $U-$processes of the Hoeffding decomposition of B_n are of order $o_{\mathbb{P}}(1)$ uniformly with respect to h.

We consider now the three second order $U-$processes of the Hoeffding decomposition of B_n. We start by conditioning on $(\boldsymbol{U}_i,\boldsymbol{U}_j)$ the reasoning for $(\boldsymbol{U}_i,\boldsymbol{U}_k)$ being similar.

$$\begin{aligned}E_{i,j}\left[X_i^2\left(K_{h,ij}-f_z(\boldsymbol{Z}_i)\right)\left(K_{h,ik}-f_z(\boldsymbol{Z}_i)\right)\right] &= X_i^2\left(K_{h,ij}-f_z(\boldsymbol{Z}_i)\right)E_i\left[\left(K_{h,ik}-f_z(\boldsymbol{Z}_i)\right)\right]\\ &= X_i^2\left(K_{h,ij}-f_z(\boldsymbol{Z}_i)\right)h^2\gamma_1(\boldsymbol{Z}_i)(1+o_{\mathbb{P}}(1))\\ &= X_i^2K_{h,ij}h^2\gamma_1(\boldsymbol{Z}_i)(1+o_{\mathbb{P}}(1))-X_i^2f_z(\boldsymbol{Z}_i)h^2\gamma_1(\boldsymbol{Z}_i)(1+o_{\mathbb{P}}(1))\\ &= h^{2-q}h^qK_{h,ij}\tau(\boldsymbol{U}_i)+O_{\mathbb{P}}(h^2).\end{aligned}$$

Now, we apply the Maximal Inequality of Sherman [72], page 448, for the degenerate $U-$process given by the kernel $h^qK_{h,ij}\tau(\boldsymbol{U}_i)$, indexed by $h\in\mathcal{H}_{c,n}$, with envelope $\|K\|_\infty\tau(\cdot)$. (Herein, $\|\cdot\|_\infty$ denotes the uniform norm.) We take $p=1$ and $\beta\in(0,1)$ arbitrarily close to 1 to stand for Sherman's quantity α. Since $K(\cdot)$ is of bounded variation and symmetric, without loss of generality we could consider that $K(\cdot)$ is nonincreasing on $[0,\infty)$. In this case, $0\le K(\cdot/h)\le K(\cdot/\overline{h})$ with $\overline{h}=\sup\mathcal{H}_{c,n}=:c_{max}n^{-\alpha}$. Hence, using Jensen's inequality, we could bound the right-hand side of the Maximal Inequality of Sherman [72] by a universal constant times

$$E^{\beta/2}\left[K^2\left(\frac{\boldsymbol{Z}_i-\boldsymbol{Z}_j}{c_{max}n^{-\alpha}}\right)\tau^2(\boldsymbol{U}_i)\right].$$

By standard changes of variables and suitable integrability conditions, this integral is bounded by a constant times $n^{-\alpha\beta q/2}$. Consequently, the uniform rate of the second $U-$processes obtained conditioning by $\boldsymbol{U}_i,\boldsymbol{U}_j$ and $\boldsymbol{U}_i,\boldsymbol{U}_k$, respectively is $n^{-1}\times O_{\mathbb{P}}(n^{-\alpha\{2-q+\beta q/2\}})$. As $1+\alpha(2-q+\beta q/2)>0$ under our assumptions we get that $n^{-1}\times O_{\mathbb{P}}(n^{-\alpha\{2-q+\beta q/2\}})=o_{\mathbb{P}}(1)$.

In addition, we get that

$$\begin{aligned}&E_{j,k}\left[X_i^2\left(K_{h,ij}-f_z(\boldsymbol{Z}_i)\right)\left(K_{h,ik}-f_z(\boldsymbol{Z}_i)\right)\right]\\ &\qquad= h^{-2q}E_{j,k}\left[X_i^2h^{2q}K_{h,ij}K_{h,ik}\right]-2E_j\left[X_i^2K_{h,ij}f_z(\boldsymbol{Z}_i)\right]+E\left[X_i^2f_z(\boldsymbol{Z}_i)^2\right].\end{aligned}$$

Now, we apply the Maximal Inequality of Sherman [72], page 448, for the degenerate $U-$process given by the kernel $E_{j,k}\left[X_i^2h^{2q}K_{h,ij}K_{h,ik}\right]$, indexed by $h\in\mathcal{H}_{c,n}$, with envelope $E_{j,k}\left[X_i^2\|K\|_\infty^2\right]$. We take $p=1$ and $\beta\in(0,1)$ arbitrarily close to 1 to stand for Sherman's quantity α. Using Jensen's inequality, we could bound the right-hand side of the Maximal Inequality of Sherman [72] by a universal constant times

$$E^{\beta/2}\left[E_{j,k}\left[X_i^2K\left(\frac{\boldsymbol{Z}_i-\boldsymbol{Z}_j}{c_{max}n^{-\alpha}}\right)K\left(\frac{\boldsymbol{Z}_i-\boldsymbol{Z}_k}{c_{max}n^{-\alpha}}\right)\right]^2\right].$$

By standard changes of variables and suitable integrability conditions, this integral is bounded by a constant times $n^{-\alpha\beta q}$. Consequently, the uniform rate of the second $U-$process obtained conditioning by $\boldsymbol{U}_j, \boldsymbol{U}_k$ is $n^{-1} \times O_{\mathbb{P}}(n^{\alpha q\{2-\beta\}})$. As $1-\alpha q(2-\beta) > 0$ under our assumptions we get that $n^{-1} \times O_{\mathbb{P}}(n^{\alpha q\{2-\beta\}}) = o_{\mathbb{P}}(1)$. By similar reasoning we can control the remaining two parts. The details are omitted. Therefore, the second order $U-$processes of B_n are of order $o_{\mathbb{P}}(1)$.

In order to finish the proof of (1.40) it remains to consider the rate for the third order $U-$process of B_n. As the reasoning for this part is the same as for the second order $U-$process we omit the details here. Therefore, the statement in (1.40) follows.

In the next part we consider

$$\begin{aligned}
&\frac{1}{n}\left\| \left(\frac{1}{n} \sum_{j=1, j\neq 1}^{n} \left(E[X_1 \mid \boldsymbol{Z}_1] f_z(\boldsymbol{Z}_1) - X_j K_{h,1j}\right), \ldots, \frac{1}{n} \sum_{j=1, j\neq n}^{n} \left(E[X_n \mid \boldsymbol{Z}_n] f_z(\boldsymbol{Z}_n) - X_j K_{h,nj}\right) \right)^T \right\|^2 \\
&\quad = \frac{1}{n^3} \sum_{1\leq i\neq j\leq n}^{n} \left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_j K_{h,ij}\right)^2 \\
&\qquad + \frac{1}{n^3} \sum_{1\leq i\neq j\neq k\leq n}^{n} \left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_j K_{h,ij}\right) \left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_k K_{h,ik}\right) \\
&\quad = \widetilde{A}_n + \widetilde{B}_n.
\end{aligned}$$

It is easy to check that $\sup_{h\in\mathcal{H}_{c,n}} |\widetilde{A}_n| = o_{\mathbb{P}}(1)$. We show in the following that

$$\sup_{h\in\mathcal{H}_{c,n}} |\widetilde{B}_n| = o_{\mathbb{P}}(1), \tag{1.41}$$

as well. Note that $\frac{n^3}{(n)_3}\widetilde{B}_n$ is a U–process of order 3. For this U–process we compute the mean and use the Hoeffding decomposition. The kernel of $\widetilde{B}_n$ is not symmetric in its arguments. However, we apply again the usual symmetrization idea.

In addition, we have that $\{(x_j, x_k, \boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{z}_k) \mapsto \left(E[X_i \mid \boldsymbol{z}_i] f_z(\boldsymbol{z}_i) - x_j K_{h,ij}\right) \left(E[X_i \mid \boldsymbol{z}_i] f_z(\boldsymbol{z}_i) - x_k K_{h,ik}\right) : h \in \mathcal{H}_{c,n}\}$ is Euclidean for a squared integrable envelope. See Lemma 22 in Nolan and Pollard [67] and Lemma 2.14 in Pakes and Pollard [68]. Therefore, we can in the following repeatedly apply Corollary 7 and the Maximal Inequality of Sherman [72]. All remainder terms are controlled by Assumption 1.3.2.

We start by considering the mean of $\widetilde{B}_n$. We get that

$$\begin{aligned}
&E\left[\left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_j K_{h,ij}\right) \left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_k K_{h,ik}\right)\right] \\
&\qquad = E\left[\left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_j K_{h,ij}\right) E_{i,j}\left[\left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - E[X_k \mid \boldsymbol{Z}_k] K_{h,ik}\right)\right]\right] \\
&\qquad = -E\left[\left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_j K_{h,ij}\right) h^2 \gamma_2(\boldsymbol{Z}_i)\right] (1 + o(1)) \\
&\qquad = h^4 E\left[\gamma_2(\boldsymbol{Z}_i)^2\right] (1 + o(1)),
\end{aligned}$$

where

$$\gamma_2(\boldsymbol{Z}) = \mu(K) \cdot \operatorname{tr}\{\boldsymbol{H}_{z,z}\left(E[X \mid \cdot] f_z\right)(\boldsymbol{Z})\}.$$

$\boldsymbol{H}_{z,z}\left(E[X \mid \cdot] f_z\right)$ denotes the matrix of second derivative of $E[X \mid \cdot] f_z(\cdot)$ with respect to the components of $\boldsymbol{Z} \in \mathbb{R}^q$. Therefore, it follows that the mean of $\widetilde{B}_n$ is of order $o_{\mathbb{P}}(1)$ uniformly with respect to h.

We consider now the three first order U–processes of the Hoeffding decomposition of $\widetilde{B}_n$. We get that by the same reasoning as for the mean

$$E_i\left[\left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_j K_{h,ij}\right) \left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_k K_{h,ij}\right)\right] = h^4 \gamma_2(\boldsymbol{Z}_i)^2 (1 + o_{\mathbb{P}}(1)).$$

In addition, we get that

$$\begin{aligned}
&E_j\left[\left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_j K_{h,ij}\right) \left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_k K_{h,ik}\right)\right] \\
&\qquad = -h^2 E_j\left[\left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_j K_{h,ij}\right) \gamma_2(\boldsymbol{Z}_i)\right] (1 + o_{\mathbb{P}}(1)) \\
&\qquad = \left(h^2 E_j\left[E\left[X_j \mid \boldsymbol{Z}_j\right] K_{h,ij} \gamma_2(\boldsymbol{Z}_i)\right] - h^2 E\left[X_i f_z(\boldsymbol{Z}_i) \gamma_2(\boldsymbol{Z}_i)\right]\right) (1 + o_{\mathbb{P}}(1)) \\
&\qquad = h^2 E\left[X_j \mid \boldsymbol{Z}_j\right] \gamma_2(\boldsymbol{Z}_j) f_z(\boldsymbol{Z}_j) + O_{\mathbb{P}}(h^4) + O_{\mathbb{P}}(h^2) + o_{\mathbb{P}}(h^2).
\end{aligned}$$

The reasoning when conditioning on $\boldsymbol{U}_k$ is the same. Therefore, it follows together with Corollary 4 of Sherman [72] that the first order $U-$processes of the Hoeffding decomposition of $\widetilde{B}_n$ are of order $o_{\mathbb{P}}(1)$ uniformly with respect to h.

We consider now the three second order $U-$processes of the Hoeffding decomposition of $\widetilde{B}_n$. We start by conditioning on $(\boldsymbol{U}_i, \boldsymbol{U}_j)$ the reasoning for $(\boldsymbol{U}_i, \boldsymbol{U}_k)$ being similar.

$$\begin{aligned}
E_{i,j}&\left[\left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_j K_{h,ij}\right)\left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_k K_{h,ik}\right)\right] \\
&= -\left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_j K_{h,ij}\right) h^2 \gamma_2(\boldsymbol{Z}_i)(1 + o_{\mathbb{P}}(1)) \\
&= X_j K_{h,ij} h^2 \gamma_2(\boldsymbol{Z}_i)(1 + o_{\mathbb{P}}(1)) - E\left[X_i \mid \boldsymbol{Z}_i\right] f_z(\boldsymbol{Z}_i) h^2 \gamma_2(\boldsymbol{Z}_i)(1 + o_{\mathbb{P}}(1)) \\
&= h^{2-q} h^q K_{h,ij} \tau(\boldsymbol{U}_i, \boldsymbol{U}_j) + O_{\mathbb{P}}(h^2).
\end{aligned}$$

Now, we apply the Maximal Inequality of Sherman [72], page 448, for the degenerate $U-$process given by the kernel $h^q K_{h,ij} \tau(\boldsymbol{U}_i, \boldsymbol{U}_j)$, indexed by $h \in \mathcal{H}_{c,n}$, with envelope $\|K\|_\infty \tau(\cdot, \cdot)$. We take $p = 1$ and $\beta \in (0, 1)$ arbitrarily close to 1 to stand for Sherman's quantity α. Using Jensen's inequality, we could bound the right-hand side of the Maximal Inequality of Sherman [72] by a universal constant times

$$E^{\beta/2}\left[K^2\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_j}{c_{max} n^{-\alpha}}\right) \tau^2(\boldsymbol{U}_i, \boldsymbol{U}_j)\right].$$

By standard changes of variables and suitable integrability conditions, this integral is bounded by a constant times $n^{-\alpha\beta q/2}$. Consequently, the uniform rate of the second $U-$processes obtained conditioning by $\boldsymbol{U}_i, \boldsymbol{U}_j$ and $\boldsymbol{U}_i, \boldsymbol{U}_k$, respectively is $n^{-1} \times O_{\mathbb{P}}(n^{-\alpha\{2-q+\beta q/2\}})$. As $1 + \alpha(2 - q + \beta q/2) > 0$ under our assumptions we get that $n^{-1} \times O_{\mathbb{P}}(n^{-\alpha\{2-q+\beta q/2\}}) = o_{\mathbb{P}}(1)$.

In addition we get that

$$\begin{aligned}
E_{j,k}&\left[\left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_j K_{h,ij}\right)\left(E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) - X_k K_{h,ik}\right)\right] \\
&= h^{-2q} X_j X_k E_{j,k}\left[h^{2q} K_{h,ij} K_{h,ik}\right] - X_j E_j\left[E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) K_{h,ij}\right] \\
&\quad - X_k E_k\left[E[X_i \mid \boldsymbol{Z}_i] f_z(\boldsymbol{Z}_i) K_{h,ik}\right] + E[X_i \mid \boldsymbol{Z}_i]^2 f_z(\boldsymbol{Z}_i)^2.
\end{aligned}$$

Now, we apply the Maximal Inequality of Sherman [72], page 448, for the degenerate $U-$process given by the kernel $X_j X_k E_{j,k}\left[h^{2q} K_{h,ij} K_{h,ik}\right]$, indexed by $h \in \mathcal{H}_{c,n}$, with envelope $X_j X_k \|K\|_\infty^2$. We take $p = 1$ and $\beta \in (0, 1)$ arbitrarily close to 1 to stand for Sherman's quantity α. Using Jensen's inequality, we could bound the right-hand side of the Maximal Inequality of Sherman [72] by a universal constant times

$$E^{\beta/2}\left[X_j^2 X_k^2 E_{j,k}\left[K\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_j}{c_{max} n^{-\alpha}}\right) K\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_k}{c_{max} n^{-\alpha}}\right)\right]^2\right].$$

By standard changes of variables and suitable integrability conditions, this integral is bounded by a constant times $n^{-\alpha\beta q}$. Consequently, the uniform rate of the second $U-$process obtained conditioning by $\boldsymbol{U}_j, \boldsymbol{U}_k$ is $n^{-1} \times O_{\mathbb{P}}(n^{\alpha q\{2-\beta\}})$. As $1 - \alpha q(2 - \beta) > 0$ under our assumptions we get that $n^{-1} \times O_{\mathbb{P}}(n^{\alpha q\{2-\beta\}}) = o_{\mathbb{P}}(1)$. By similar reasoning we can control the remaining three parts. The details are omitted. Therefore, the second order $U-$processes of $\widetilde{B}_n$ are of order $o_{\mathbb{P}}(1)$.

In order to finish the proof of (1.41) it remains to consider the rate for the third order $U-$process of $\widetilde{B}_n$. As the reasoning for this part is the same as for the second order $U-$process we omit the details here. Therefore, the statement in (1.41) follows.

It is obvious that

$$\begin{aligned}
\frac{1}{n}&\left\|\left(n^{-1} E[X_1 \mid \boldsymbol{Z}_1] f_z(\boldsymbol{Z}_1), \ldots, n^{-1} E[X_n \mid \boldsymbol{Z}_n] f_z(\boldsymbol{Z}_n)\right)^T\right\|^2 = o_{\mathbb{P}}(1) \\
&\text{and} \quad \frac{1}{n}\left\|\left(n^{-1} X_1 f_z(\boldsymbol{Z}_1), \ldots, n^{-1} X_n f_z(\boldsymbol{Z}_n)\right)^T\right\|^2 = o_{\mathbb{P}}(1).
\end{aligned}$$

Therefore, the statement follows.

□

Lemma 1.7. *Under the conditions of Theorem 1.1,*

$$\sup_{h\in\mathcal{H}_{c,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\widehat{\mathbb{B}}_n-\mathbb{B}_n\right\|_{\mathrm{Sp}}=o_{\mathbb{P}}(n).$$

Proof of Lemma 1.7.

We could once again write

$$\mathbb{D}_n=\boldsymbol{\Omega}_n^{1/2}\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\boldsymbol{\Omega}_n^{1/2}=\mathbb{S}_n^T\mathbb{S}_n,$$

where

$$\mathbb{S}_n=\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\boldsymbol{\Omega}_n^{1/2},$$

and $P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}$ is the projector on the 1−dimensional subspace generated by the vector $\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n$, that is

$$P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}=\frac{1}{\mathbf{1}_n^T\boldsymbol{\Omega}_n\mathbf{1}_n}\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{\Omega}_n^{1/2}.$$

Here, $\boldsymbol{\Omega}_n^{1/2}$ is the positive definite square root of $\boldsymbol{\Omega}_n$. Next, we could rewrite $\widehat{\mathbb{B}}_n$ and $\mathbb{B}_n$ under the form

$$\widehat{\mathbb{B}}_n=\mathbb{S}_n^T\left(\boldsymbol{I}_{n\times n}-P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}\right)\mathbb{S}_n\quad\text{ and }\quad\mathbb{B}_n=\mathbb{S}_n^T\left(\boldsymbol{I}_{n\times n}-P_{\mathbb{S}_n\mathbb{X}_n}\right)\mathbb{S}_n,$$

with $P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}$ and $P_{\mathbb{S}_n\mathbb{X}_n}$ the orthogonal projectors on the subspaces generated by $\mathbb{S}_n\widehat{\mathbb{X}}_n$ and $\mathbb{S}_n\mathbb{X}_n$, that is

$$P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}=\mathbb{S}_n\widehat{\mathbb{X}}_n\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\widehat{\mathbb{X}}_n^T\mathbb{S}_n^T\quad\text{ and }\quad P_{\mathbb{S}_n\mathbb{X}_n}=\mathbb{S}_n\mathbb{X}_n\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\mathbb{X}_n^T\mathbb{S}_n^T.$$

Thus,

$$\widehat{\mathbb{B}}_n-\mathbb{B}_n=\mathbb{S}_n^T\left(P_{\mathbb{S}_n\mathbb{X}_n}-P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}\right)\mathbb{S}_n.$$

In view of this decomposition, it suffices to control uniformly the norm of the difference between the projectors $P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}$ and $P_{\mathbb{S}_n\mathbb{X}_n}$. Whenever the inverses exist, we decompose

$$\begin{aligned}\mathbb{S}_n\widehat{\mathbb{X}}_n\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1/2}-\mathbb{S}_n\mathbb{X}_n\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1/2}&=\mathbb{S}_n\widehat{\mathbb{X}}_n\left[\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1/2}-\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1/2}\right]\\&\quad+\left[\mathbb{S}_n\widehat{\mathbb{X}}_n-\mathbb{S}_n\mathbb{X}_n\right]\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1/2}.\end{aligned}$$

Meanwhile, for any $\boldsymbol{u}\in\mathbb{R}^p$,

$$\boldsymbol{u}^T\mathbb{X}_n^T\mathbb{X}_n\boldsymbol{u}=n\boldsymbol{u}^TVar\left[\boldsymbol{X}-E[\boldsymbol{X}\mid\boldsymbol{Z}]\right]\boldsymbol{u}+O_{\mathbb{P}}(n^{1/2}),$$

which indicates that the spectral norm of $n^{-1/2}\mathbb{X}_n$ converges at the rate $O_{\mathbb{P}}(n^{-1/2})$ to the largest eigenvalue of the variance of $\boldsymbol{X}-E[\boldsymbol{X}\mid\boldsymbol{Z}]$. From Lemma 1.6 and the triangle inequality, we deduce that the spectral norm of $n^{-1/2}\widehat{\mathbb{X}}_n$ converges also to the largest eigenvalue of the variance of $\boldsymbol{X}-E[\boldsymbol{X}\mid\boldsymbol{Z}]$. Next, let us write

$$\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n-\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n=\left(\widehat{\mathbb{X}}_n-\mathbb{X}_n\right)^T\mathbb{D}_n\mathbb{X}_n+\mathbb{X}_n^T\mathbb{D}_n\left(\widehat{\mathbb{X}}_n-\mathbb{X}_n\right)+\left(\widehat{\mathbb{X}}_n-\mathbb{X}_n\right)^T\mathbb{D}_n\left(\widehat{\mathbb{X}}_n-\mathbb{X}_n\right).$$

Taking spectral norm on both sides and using the bounds of the spectral norms for $\mathbb{D}_n$, $n^{-1/2}\mathbb{X}_n$ and $n^{-1/2}\widehat{\mathbb{X}}_n$, as well as the uniform bound derived in Lemma 1.6, we deduce that

$$\sup_{h\in\mathcal{H}_{c,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\frac{1}{n^2}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n-\frac{1}{n^2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right\|_{\mathrm{Sp}}=o_{\mathbb{P}}(1).\tag{1.42}$$

Let $\delta>0$ and consider the event $\mathcal{A}_n=\mathcal{A}_{1n}\cap\mathcal{A}_{2n}$ where

$$\mathcal{A}_{1n}=\{n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n-(\delta/2)\boldsymbol{I}_{p\times p}\text{ is positive semi-definite}\},$$

and $\mathcal{A}_{2n}$ is defined in a similar way with $\mathbb{X}_n$ replaced by $\widehat{\mathbb{X}}_n$. From Lemma 1.5, we know that $E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]$ tends to a positive definite matrix. From this and equation (1.42), we could fix $\delta>0$ such that the probability of the event $\mathcal{A}_n$ tends to 1. On the event $\mathcal{A}_n$, using Lemma 1.4 and 1.5 and equation (1.42), we deduce that

$$\begin{aligned}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|n\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1/2}\right\|_{\mathrm{Sp}} &\leq \sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\left(n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1/2}-E[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n]^{-1/2}\right\|_{\mathrm{Sp}}\\ &\quad+\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|E[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n]^{-1/2}\right\|_{\mathrm{Sp}}\\ &\leq\sqrt{2}\delta^{-3/2}O_{\mathbb{P}}(n^{-1/2})+\sqrt{2/\delta},\end{aligned}$$

and

$$\sup_{h\in\mathcal{H}_{c,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|n\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1/2}-n\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1/2}\right\|_{\mathrm{Sp}}=o_{\mathbb{P}}(1).$$

Finally, note that

$$\|\mathbb{S}_n\|_{\mathrm{Sp}}=\left\|\mathbb{S}_n^T\right\|_{\mathrm{Sp}}\leq\left\|\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\right\|_{\mathrm{Sp}}\left\|\boldsymbol{\Omega}_n^{1/2}\right\|_{\mathrm{Sp}}\leq n^{1/2}.$$

Gathering facts and using repeatedly the property $\|\boldsymbol{A}_1\boldsymbol{A}_2\|_{\mathrm{Sp}}\leq\|\boldsymbol{A}_1\|_{\mathrm{Sp}}\|\boldsymbol{A}_2\|_{\mathrm{Sp}}$, Lemma 1.4 and Lemma 1.6, we deduce that

$$\begin{aligned}\left\|P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}-P_{\mathbb{S}_n\mathbb{X}_n}\right\|_{\mathrm{Sp}} &\leq 2\left\|n^{-1/2}\mathbb{S}_n n^{-1/2}\widehat{\mathbb{X}}_n\left[n\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1/2}-n\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1/2}\right]\right\|_{\mathrm{Sp}}\\ &\quad+2\left\|n^{-1/2}\mathbb{S}_n n^{-1/2}\left(\widehat{\mathbb{X}}_n-\mathbb{X}_n\right)n\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1/2}\right\|_{\mathrm{Sp}}\\ &=o_{\mathbb{P}}(1).\end{aligned}$$

Finally,

$$\left\|\widehat{\mathbb{B}}_n-\mathbb{B}_n\right\|_{\mathrm{Sp}}\leq\left\|\mathbb{S}_n^T\right\|_{\mathrm{Sp}}\left\|P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}-P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}\right\|_{\mathrm{Sp}}\|\mathbb{S}_n\|_{\mathrm{Sp}}=o_{\mathbb{P}}(n).$$

Now, the proof is complete.

□

Lemma 1.8. *Assume the conditions of Theorem 1.1 hold true. Then,*

$$\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\boldsymbol{\Omega}_n^{1/2}(\boldsymbol{d})n^{-1}\mathbb{X}_n\right\|_{\mathrm{Sp}}=O_{\mathbb{P}}(1),\qquad\textit{and}\qquad\sup_{h\in\mathcal{H}_{c,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\boldsymbol{\Omega}_n^{1/2}(\boldsymbol{d})n^{-1}\left[\widehat{\mathbb{X}}_n-\mathbb{X}_n\right]\right\|_{\mathrm{Sp}}=o_{\mathbb{P}}(1).$$

As a consequence

$$\sup_{h\in\mathcal{H}_{c,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\boldsymbol{\Omega}_n^{1/2}(\boldsymbol{d})n^{-1}\widehat{\mathbb{X}}_n\right\|_{\mathrm{Sp}}=O_{\mathbb{P}}(1).$$

Proof of Lemma 1.8.

We have that

$$\left\|\boldsymbol{\Omega}_n^{1/2}(\boldsymbol{d})n^{-1}\mathbb{X}_n\right\|_{\mathrm{Sp}}\leq\left\|\boldsymbol{\Omega}_n^{1/2}(\boldsymbol{d})n^{-1/2}\right\|_{\mathrm{Sp}}\left\|n^{-1/2}\mathbb{X}_n\right\|,$$

The first rate follows now from Lemma 1.3 and the fact that by our assumptions the expectation of $\|n^{-1/2}\mathbb{X}_n\|^2$ is finite. The second rate follows from Lemma 1.3 and 1.6. The third rate is a direct consequence of the first two rates.

□

Lemma 1.9. *Under Assumption 1.1.1, there exists a constant C, depending on Λ such that*

$$0\leq\frac{\partial}{\partial\lambda}T(Y,\lambda)\leq C\max\left((Y\vee 1)^{\lambda_{\min}}\log^2(Y\vee e),(Y\vee 1)^{\lambda_{\max}}\log^2(Y\vee e)\right).$$

Proof of Lemma 1.9.

The derivative of $T(Y,\lambda)$ with respect to λ is given by

$$\frac{\partial}{\partial\lambda}T(Y,\lambda)=\begin{cases}\lambda^{-2}\left(Y^\lambda(\lambda\log(Y)-1)+1\right) & ,\ \lambda\neq 0\\ 2^{-1}\log^2(Y) & ,\ \lambda=0.\end{cases}$$

The derivative is positive and continuous at $\lambda = 0$. Furthermore, the sign of the derivative for $\lambda \neq 0$ is determined by $Y^\lambda(\lambda \log(Y) - 1) + 1$. We get that

$$\frac{\partial}{\partial \lambda}\{Y^\lambda(\lambda \log(Y) - 1) + 1\} = \lambda Y^\lambda \log^2(Y),$$

which is positive for $\lambda > 0$ and negative for $\lambda < 0$. This implies that the first derivative is always positive. Next, the upper bound is obvious.

□

Lemma 1.10. *Under Assumption 1.1.1, there exists a constant C, depending on Λ such that*

$$\left|\frac{\partial^2}{\partial \lambda^2} T(Y, \lambda)\right| \leq C \max\left((Y \vee 1)^{\lambda_{\min}} |\log(Y \vee e)|^3, (Y \vee 1)^{\lambda_{\max}} |\log(Y \vee e)|^3\right).$$

Proof of Lemma 1.10.

The second derivative of $T(Y, \lambda)$ with respect to λ is given by

$$\frac{\partial^2 T(Y, \lambda)}{\partial \lambda^2} = \begin{cases} \lambda^{-3}\left(Y^\lambda \lambda^2 \log(Y)^2 - 2\left(Y^\lambda(\lambda \log(Y) - 1) + 1\right)\right) & , \lambda \neq 0 \\ 3^{-1} \log(Y)^3 & , \lambda = 0. \end{cases}$$

Once again we consider the derivative of the nominator for $\lambda \neq 0$. The derivative is given by

$$\frac{\partial}{\partial \lambda}\left\{Y^\lambda \lambda^2 \log(Y)^2 - 2\left(Y^\lambda(\lambda \log(Y) - 1) + 1\right)\right\} = \log(Y)^3 \lambda^2 Y^\lambda.$$

This derivative is positive if $Y > 1$ and negative if $Y < 1$. The second derivative of $T(Y, \lambda)$ with respect to λ is continuous at $\lambda = 0$ and positive if $Y > 1$ and negative if $Y < 1$. Therefore, it follows that

$$\frac{\partial^2 T(Y, \lambda)}{\partial \lambda^2} = \begin{cases} > 0 & , Y > 1 \\ < 0 & , Y < 1, \end{cases}$$

for all λ. Next, the upper bound is obvious.

□

Lemma 1.11. *Under Assumption 1.1.1, there exists a constant C, depending on Λ such that*

$$0 \leq \frac{\partial^3}{\partial \lambda^3} T(Y, \lambda) \leq C \max\left((Y \vee 1)^{\lambda_{\min}} \log(Y \vee e)^4, (Y \vee 1)^{\lambda_{\max}} \log(Y \vee e)^4\right).$$

Proof of Lemma 1.11.

The third derivative of $T(Y, \lambda)$ with respect to λ is given by

$$\frac{\partial^3 T(Y, \lambda)}{\partial \lambda^3} = \begin{cases} \lambda^{-4}\left(Y^\lambda \lambda^2 \log(Y)^2(\lambda \log(Y) - 3) + 6\left(Y^\lambda(\lambda \log(Y) - 1) + 1\right)\right) & , \lambda \neq 0 \\ 4^{-1} \log(Y)^4 & , \lambda = 0. \end{cases}$$

Once again we consider the derivative of the nominator for $\lambda \neq 0$. The derivative is given by

$$\frac{\partial}{\partial \lambda}\left\{Y^\lambda \lambda^2 \log(Y)^2(\lambda \log(Y) - 3) + 6\left(Y^\lambda(\lambda \log(Y) - 1) + 1\right)\right\} = \lambda^3 \log(Y)^4 Y^\lambda,$$

which is positive for $\lambda > 0$ and negative for $\lambda < 0$. This implies that the third derivative is always positive. Next, the upper bound is obvious.

□

Lemma 1.12. *Under the conditions of Theorem 1.1,*

$$\sup_{\lambda\in\Lambda}\left\|n^{-1}\mathbb{Y}_n(\lambda)\right\| = O_{\mathbb{P}}(n^{-1/2}).$$

Proof of Lemma 1.12.

First, note that the functions $\{y \mapsto \lambda^{-1}(y^\lambda - 1) : y \geq c > 0, \lambda \in \Lambda\}$, with c a fixed lower bound of the support of Y, are Lipschitz in the index parameter λ. See Lemma 1.9. Deduce that this family of functions of Y is Euclidean for a to the power of four integrable envelope. See Lemma 2.13 in Pakes and Pollard [68]. Since the Euclidean property is preserved by multiplication with a fixed function, the family $\{(y, \boldsymbol{z}) \mapsto \lambda^{-1}(y^\lambda - 1)f_z(\boldsymbol{z}) : y \geq c > 0, \boldsymbol{z} \in \mathbb{R}^q, \lambda \in \Lambda\}$ is also Euclidean for a to the power of four integrable envelope. See Lemma 2.14 in Pakes and Pollard [68]. The Euclidean property is also preserved if the functions of Y and $\boldsymbol{Z}$ are centered by their conditional expectation given $\boldsymbol{Z}$. See Lemma 5 in Sherman [72]. Next, it is also preserved by taking the square of the functions in the family. The envelope is now squared integrable. See Lemma 2.14 in Pakes and Pollard [68]. Deduce from Corollary 7 in Sherman [72] that

$$\sup_{\lambda\in\Lambda}\left|\frac{1}{n}\sum_{i=1}^{n}(T(Y_i,\lambda) - E[T(Y_i,\lambda)\mid \boldsymbol{Z}_i])^2 f_z^2(\boldsymbol{Z}_i)\right| = O_{\mathbb{P}}(1).$$

Then the required rate follows.

□

Lemma 1.13. *Under the conditions of Theorem 1.1,*

$$\sup_{h\in\mathcal{H}_{c,n}}\sup_{\lambda\in\Lambda}\left\|n^{-1}\left[\widehat{\mathbb{Y}}_n(\lambda) - \mathbb{Y}_n(\lambda)\right]\right\| = o_{\mathbb{P}}(n^{-1/2}).$$

Proof of Lemma 1.13.

It suffices to decompose

$$\widehat{\mathbb{Y}}_n(\lambda) - \mathbb{Y}_n(\lambda) = \boldsymbol{R}_{1n} + \boldsymbol{R}_{2n},$$

with

$$\boldsymbol{R}_{1n} = \left(T(Y_1,\lambda)\left(\widehat{f}_z(\boldsymbol{Z}_1) - f_z(\boldsymbol{Z}_1)\right), \ldots, T(Y_n,\lambda)\left(\widehat{f}_z(\boldsymbol{Z}_n) - f_z(\boldsymbol{Z}_n)\right)\right)^T,$$

and

$$\boldsymbol{R}_{2n} = \Big(\Big(E[T(Y_1,\lambda)\mid \boldsymbol{Z}_1]f_z(\boldsymbol{Z}_1) - \widehat{E}[T(Y_1,\lambda)\mid \boldsymbol{Z}_1]\widehat{f}_z(\boldsymbol{Z}_1)\Big), \\ \ldots, \Big(E[T(Y_n,\lambda)\mid \boldsymbol{Z}_n]f_z(\boldsymbol{Z}_n) - \widehat{E}[T(Y_n,\lambda)\mid \boldsymbol{Z}_n]\widehat{f}_z(\boldsymbol{Z}_n)\Big)\Big)^T.$$

We can now use the same arguments as in Lemma 1.6 to show that

$$\sup_{h\in\mathcal{H}_{c,n}}\sup_{\lambda\in\Lambda}\left\|n^{-1}\boldsymbol{R}_{1n}\right\| = o_{\mathbb{P}}(n^{-1/2}) \quad \textit{and} \quad \sup_{h\in\mathcal{H}_{c,n}}\sup_{\lambda\in\Lambda}\left\|n^{-1}\boldsymbol{R}_{2n}\right\| = o_{\mathbb{P}}(n^{-1/2}).$$

The Euclidean properties needed follow from a similar discussion as in Lemma 1.12.

□

Lemma 1.14. *Assume the conditions of Proposition 1.1 hold true. Then*

$$\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\boldsymbol{\Omega}_n^{1/2}(\boldsymbol{d})\,n^{-1}\left[\mathbb{Y}_n(\lambda_0) - \mathbb{X}_n\boldsymbol{\beta}_0\right]\right\| = O_{\mathbb{P}}(n^{-1/2}).$$

Proof of Lemma 1.14.

By definition $\mathbb{Y}_n(\lambda_0) - \mathbb{X}_n\boldsymbol{\beta}_0 = (\boldsymbol{\varepsilon f_z})_n = (\varepsilon_1 f_z(\boldsymbol{Z}_1), \ldots, \varepsilon_n f_z(\boldsymbol{Z}_n))^T$. Next, for any $\boldsymbol{d}$, using the Fourier Transform (see the last part of the proof of Lemma 1.1), we can write

$$0 \leq n^{-2}(\boldsymbol{\varepsilon f_z})_n^T\boldsymbol{\Omega}_n(\boldsymbol{d})(\boldsymbol{\varepsilon f_z})_n \leq \frac{d_U^{(p+q)/2}}{d_L^{(p+q)/2}}n^{-2}(\boldsymbol{\varepsilon f_z})_n^T\boldsymbol{\Omega}_n(\mathrm{diag}(d_U,\ldots,d_U))(\boldsymbol{\varepsilon f_z})_n.$$

Simply calculating the expectation, the last quadratic form in the last display has the rate $O_{\mathbb{P}}\left(n^{-1}\right)$.

The uniform rate follows.

□

Lemma 1.15. *Assume the conditions of Theorem 1.1 hold true and let* Λ_{0n} *be an arbitrary* $o_{\mathbb{P}}(1)$ *neighborhood of* λ_0*. Then, for* $s \in \{1,2,3\}$*,*

$$\sup_{\lambda\in\Lambda_{0n}} \sup_{\boldsymbol{d}\in\mathcal{D}} \left\| \boldsymbol{\Omega}_n^{1/2}(\boldsymbol{d}) n^{-1} \frac{\partial^s}{\partial\lambda^s} \mathbb{Y}_n(\lambda) \right\|_{\mathrm{Sp}} = O_{\mathbb{P}}(1),$$

and

$$\sup_{\lambda\in\Lambda_{0n}} \sup_{h\in\mathcal{H}_{c,n}} \sup_{\boldsymbol{d}\in\mathcal{D}} \left\| \boldsymbol{\Omega}_n^{1/2}(\boldsymbol{d}) n^{-1} \left[\frac{\partial^s}{\partial\lambda^s} \widehat{\mathbb{Y}}_n(\lambda) - \frac{\partial^s}{\partial\lambda^s} \mathbb{Y}_n(\lambda) \right] \right\|_{\mathrm{Sp}} = o_{\mathbb{P}}(1).$$

As a consequence

$$\sup_{\lambda\in\Lambda_{0n}} \sup_{h\in\mathcal{H}_{c,n}} \sup_{\boldsymbol{d}\in\mathcal{D}} \left\| \boldsymbol{\Omega}_n^{1/2}(\boldsymbol{d}) n^{-1} \frac{\partial^s}{\partial\lambda^s} \widehat{\mathbb{Y}}_n(\lambda) \right\|_{\mathrm{Sp}} = O_{\mathbb{P}}(1).$$

Proof of Lemma 1.15.

We have that

$$\left\| \boldsymbol{\Omega}_n^{1/2}(\boldsymbol{d}) n^{-1} \frac{\partial^s}{\partial\lambda^s} \mathbb{Y}_n(\lambda) \right\|_{\mathrm{Sp}} \leq \left\| \boldsymbol{\Omega}_n^{1/2}(\boldsymbol{d}) n^{-1/2} \right\|_{\mathrm{Sp}} \left\| n^{-1/2} \frac{\partial^s}{\partial\lambda^s} \mathbb{Y}_n(\lambda) \right\|.$$

The first rate follows now from Lemma 1.3 and the fact that, by our assumptions, the expectation of $\sup_{\lambda\in\Lambda_{0n}} \| n^{-1/2}(\partial^s/\partial\lambda^s)\mathbb{Y}_n(\lambda)\|^2$ is finite.

The second rate follows again from Lemma 1.3 and the same arguments as in Lemma 1.6 and 1.13. The Euclidean properties needed follow from a similar discussion as in Lemma 1.12. The third rate is a direct consequence of the first two rates.

□

Lemma 1.16. *Assume the conditions of Theorem 1.1 hold true and let* Λ_{0n} *be an arbitrary* $o_{\mathbb{P}}(1)$ *neighborhood of* λ_0*. Then,*

$$\sup_{\lambda\in\Lambda_{0n}} \sup_{h\in\mathcal{H}_{c,n}} \sup_{s\in S_n} \sup_{\boldsymbol{d}\in\mathcal{D}} \Big| \frac{\partial}{\partial\lambda} \left\{ n^{-1} s^{-\lambda} \widehat{\mathbb{Y}}_n(\lambda) \right\}^T \widehat{\mathbb{B}}_n \frac{\partial}{\partial\lambda} \left\{ n^{-1} s^{-\lambda} \widehat{\mathbb{Y}}_n(\lambda) \right\} \\ - \frac{\partial}{\partial\lambda} \left\{ n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0) \right\}^T \widehat{\mathbb{B}}_n \frac{\partial}{\partial\lambda} \left\{ n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0) \right\} \Big| = o_{\mathbb{P}}(1),$$

and

$$\sup_{\lambda\in\Lambda_{0n}} \sup_{h\in\mathcal{H}_{c,n}} \sup_{s\in S_n} \sup_{\boldsymbol{d}\in\mathcal{D}} \Big| n^{-1} s^{-\lambda} \widehat{\mathbb{Y}}_n(\lambda)^T \widehat{\mathbb{B}}_n \frac{\partial^2}{\partial\lambda^2} \left\{ n^{-1} s^{-\lambda} \widehat{\mathbb{Y}}_n(\lambda) \right\} \\ - \left\{ n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0) \right\}^T \widehat{\mathbb{B}}_n \frac{\partial^2}{\partial\lambda^2} \left\{ n^{-1} s^{-\lambda_0} \widehat{\mathbb{Y}}_n(\lambda_0) \right\} \Big| = o_{\mathbb{P}}(1).$$

Proof of Lemma 1.16.

Note that $\frac{\partial}{\partial\lambda}\{n^{-1}s^{-\lambda}\widehat{\mathbb{Y}}_n(\lambda)\} = s^{-\lambda}\frac{\partial}{\partial\lambda}\{n^{-1}\widehat{\mathbb{Y}}_n(\lambda)\} - \log(s)s^{-\lambda}n^{-1}\widehat{\mathbb{Y}}_n(\lambda)$.

We have that

$$\begin{aligned} & s^{-2\lambda} n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda)^T \widehat{\mathbb{B}}_n \, n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda) - s^{-2\lambda_0} n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0)^T \widehat{\mathbb{B}}_n \, n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0) \\ &= s^{-2\lambda} n^{-1} \left[\frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda) - \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0) \right]^T \widehat{\mathbb{B}}_n \, n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda) \\ &\quad - s^{-2\lambda} n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0)^T \widehat{\mathbb{B}}_n \, n^{-1} \left[\frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0) - \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda) \right] \\ &\quad + \left(s^{-2\lambda} - s^{-2\lambda_0} \right) n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0)^T \widehat{\mathbb{B}}_n \, n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0) \end{aligned}$$

$$\begin{aligned}
&= s^{-2\lambda} n^{-1} \frac{\partial^2}{\partial\lambda^2} \widehat{\mathbb{Y}}_n(\widetilde{\lambda})^T \, \widehat{\mathbb{B}}_n \, n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda)(\lambda-\lambda_0) \\
&\quad + s^{-2\lambda} n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0)^T \, \widehat{\mathbb{B}}_n \, n^{-1} \frac{\partial^2}{\partial\lambda^2} \widehat{\mathbb{Y}}_n(\widetilde{\lambda})(\lambda-\lambda_0) \\
&\quad + \left(s^{-2\lambda} - s^{-2\lambda_0}\right) n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0)^T \, \widehat{\mathbb{B}}_n \, n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0),
\end{aligned}$$

where $\widetilde{\lambda} = c\lambda + (1-c)\lambda_0$ for some $c \in (0,1)$. Recall that

$$\widehat{\mathbb{B}}_n = \mathbb{S}_n^T \left(\boldsymbol{I}_{n\times n} - P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}\right) \mathbb{S}_n,$$

where $P_{\mathbb{S}_n\widehat{\mathbb{X}}_n}$ is the orthogonal projector on the subspace generated by $\mathbb{S}_n\widehat{\mathbb{X}}_n$ with

$$\mathbb{S}_n = \left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right) \boldsymbol{\Omega}_n^{1/2},$$

and $P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}$ is the projector on the subspace generated by the vector $\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n$. Deduce that

$$\begin{aligned}
&\left| s^{-2\lambda} n^{-1} \frac{\partial^2}{\partial\lambda^2} \widehat{\mathbb{Y}}_n(\widetilde{\lambda})^T \, \widehat{\mathbb{B}}_n \, n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda)(\lambda-\lambda_0) \right| \\
&\qquad \le s^{-2\lambda} |(\lambda-\lambda_0)| \left\| \boldsymbol{\Omega}_n^{1/2} n^{-1} \frac{\partial^2}{\partial\lambda^2} \widehat{\mathbb{Y}}_n(\widetilde{\lambda}) \right\|_{\mathrm{Sp}} \\
&\qquad\quad \times \left\| \left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right) \left(\boldsymbol{I}_{n\times n} - P_{\mathbb{S}_n\mathbb{X}_n}\right) \left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right) \right\|_{\mathrm{Sp}} \\
&\qquad\quad \times \left\| \boldsymbol{\Omega}_n^{1/2} n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda) \right\|_{\mathrm{Sp}}.
\end{aligned}$$

By the same reasoning we get that

$$\begin{aligned}
&\left| s^{-2\lambda} n^{-1} \frac{\partial^2}{\partial\lambda^2} \widehat{\mathbb{Y}}_n(\widetilde{\lambda})^T \, \widehat{\mathbb{B}}_n \, n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0)(\lambda-\lambda_0) \right| \\
&\qquad \le s^{-2\lambda} |(\lambda-\lambda_0)| \left\| \boldsymbol{\Omega}_n^{1/2} n^{-1} \frac{\partial^2}{\partial\lambda^2} \widehat{\mathbb{Y}}_n(\widetilde{\lambda}) \right\|_{\mathrm{Sp}} \\
&\qquad\quad \times \left\| \left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right) \left(\boldsymbol{I}_{n\times n} - P_{\mathbb{S}_n\mathbb{X}_n}\right) \left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right) \right\|_{\mathrm{Sp}} \\
&\qquad\quad \times \left\| \boldsymbol{\Omega}_n^{1/2} n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0) \right\|_{\mathrm{Sp}}.
\end{aligned}$$

and

$$\begin{aligned}
&\left| \left(s^{-2\lambda} - s^{-2\lambda_0}\right) n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0)^T \, \widehat{\mathbb{B}}_n \, n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0) \right| \\
&\qquad \le |s^{-2\lambda} - s^{-2\lambda_0}| \left\| \boldsymbol{\Omega}_n^{1/2} n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0) \right\|_{\mathrm{Sp}} \\
&\qquad\quad \times \left\| \left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right) \left(\boldsymbol{I}_{n\times n} - P_{\mathbb{S}_n\mathbb{X}_n}\right) \left(\boldsymbol{I}_{n\times n} - P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right) \right\|_{\mathrm{Sp}} \\
&\qquad\quad \times \left\| \boldsymbol{\Omega}_n^{1/2} n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0) \right\|_{\mathrm{Sp}}.
\end{aligned}$$

It follows now from the fact that the spectral norm of a product of projectors is at most equal to 1, $\lambda \in \Lambda_{0n}$, $\sup_{\lambda\in\Lambda_{0n}} \sup_{s\in S_n} \left|s^{-2\lambda} - s^{-2\lambda_0}\right| = o_{\mathbb{P}}(1)$ as well as $\sup_{\lambda\in\Lambda_{0n}} \sup_{s\in S_n} s^{-2\lambda} = O_{\mathbb{P}}(1)$ and from Lemma 1.15 that

$$\begin{aligned}
\sup_{\lambda\in\Lambda_{0n}} \sup_{h\in\mathcal{H}_{c,n}} \sup_{s\in S_n} \sup_{\boldsymbol{d}\in\mathcal{D}} \Bigg| & s^{-2\lambda} n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda)^T \, \widehat{\mathbb{B}}_n \, n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda) \\
& - s^{-2\lambda_0} n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0)^T \, \widehat{\mathbb{B}}_n \, n^{-1} \frac{\partial}{\partial\lambda} \widehat{\mathbb{Y}}_n(\lambda_0) \Bigg| = o_{\mathbb{P}}(1).
\end{aligned}$$

By similar reasoning, we get that

$$\sup_{\lambda\in\Lambda_{0n}}\sup_{h\in\mathcal{H}_{c,n}}\sup_{s\in S_n}\sup_{\boldsymbol{d}\in\mathcal{D}}\Big|\log(s)s^{-\lambda}n^{-1}\widehat{\mathbb{Y}}_n(\lambda)^T\,\widehat{\mathbb{B}}_n\,n^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda) \\ -\log(s)s^{-\lambda_0}n^{-1}\widehat{\mathbb{Y}}_n(\lambda_0)^T\,\widehat{\mathbb{B}}_n\,n^{-1}\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)\Big| = o_{\mathbb{P}}(1).$$

and

$$\sup_{\lambda\in\Lambda_{0n}}\sup_{h\in\mathcal{H}_{c,n}}\sup_{s\in S_n}\sup_{\boldsymbol{d}\in\mathcal{D}}\Big|\log(s)s^{-\lambda}n^{-1}\widehat{\mathbb{Y}}_n(\lambda)^T\,\widehat{\mathbb{B}}_n\,\log(s)s^{-\lambda}n^{-1}\widehat{\mathbb{Y}}_n(\lambda) \\ -\log(s)s^{-\lambda_0}n^{-1}\widehat{\mathbb{Y}}_n(\lambda_0)^T\,\widehat{\mathbb{B}}_n\,\log(s)s^{-\lambda_0}n^{-1}\widehat{\mathbb{Y}}_n(\lambda_0)\Big| = o_{\mathbb{P}}(1).$$

Therefore, the first statement follows. The second statement follows by the same reasoning. The details are omitted.

□

Lemma 1.17. *Assume the conditions of Theorem 1.1 hold true and let Λ_{0n} be an arbitrary $o_{\mathbb{P}}(1)$ neighborhood of λ_0. Then,*

$$\sup_{\lambda\in\Lambda_{0n}}\sup_{h\in\mathcal{H}_{c,n}}\sup_{s\in S_n}\sup_{\boldsymbol{d}\in\mathcal{D}}\left|n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n n^{-1}\left(\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda)-\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)\right)\right| = o_{\mathbb{P}}(1).$$

Proof of Lemma 1.17.

We have that

$$n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n n^{-1}\left(\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda)-\frac{\partial}{\partial\lambda}\widehat{\mathbb{Y}}_n(\lambda_0)\right) = n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n n^{-1}\frac{\partial^2}{\partial\lambda^2}\widehat{\mathbb{Y}}_n(\widetilde{\lambda})(\lambda-\lambda_0),$$

where $\widetilde{\lambda} = c\lambda + (1-c)\lambda_0$ for some $c\in(0,1)$. We get that

$$\left|n^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n n^{-1}\frac{\partial^2}{\partial\lambda^2}\widehat{\mathbb{Y}}_n(\widetilde{\lambda})\right| \\ \leq \left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\widehat{\mathbb{X}}_n\right\|_{\mathrm{Sp}}\times\left\|\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\left(\boldsymbol{I}_{n\times n}-P_{\boldsymbol{\Omega}_n^{1/2}\mathbf{1}_n}\right)\right\|_{\mathrm{Sp}}\times\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\frac{\partial^2}{\partial\lambda^2}\widehat{\mathbb{Y}}_n(\widetilde{\lambda})\right\|_{\mathrm{Sp}}.$$

The statement follows now from the fact that the spectral norm of a product of projectors is at most equal to 1, $\lambda\in\Lambda_{0n}$ and Lemma 1.8 and 1.15.

□

Lemma 1.18. *Assume the conditions of Proposition 1.1 hold true. Then,*

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\boldsymbol{\Omega}_n^{1/2}(\boldsymbol{d})n^{-1}\left([\mathbb{Y}_n(\lambda_0)-\mathbb{X}_n\boldsymbol{\beta}_0]-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n-\left[\widehat{\mathbb{Y}}_n(\lambda_0)-\widehat{\mathbb{X}}_n\boldsymbol{\beta}_0\right]\right)\right\| = o_{\mathbb{P}}(n^{-1/2}).$$

Proof of Lemma 1.18.

By the arguments used for Lemma 1.14, it suffices to consider $\boldsymbol{d} = \mathrm{diag}(d_U,\ldots,d_U)$. Moreover, for simpler notation, we omit the argument $\boldsymbol{d}$ in $\boldsymbol{\Omega}_n(\boldsymbol{d})$. We get that, for $1\leq i\leq n$,

$$\left[\widehat{\mathbb{Y}}_{n,i}(\lambda_0)-\widehat{\mathbb{X}}_{n,i}\boldsymbol{\beta}_0\right]-[\mathbb{Y}_{n,i}(\lambda_0)-\mathbb{X}_{n,i}\boldsymbol{\beta}_0]+\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_{n,i} \\ = \left[\frac{1}{n}\sum_{k=1}^{n}(m(\boldsymbol{Z}_i)-m(\boldsymbol{Z}_k))K_{h,ik}+\frac{1}{n}\sum_{k=1}^{n}(\varepsilon_i-\varepsilon_k)K_{h,ik}\right]-\varepsilon_i f_z(\boldsymbol{Z}_i)+\frac{1}{n}\sum_{k=1,k\neq i}^{n}\varepsilon_k K_{h,ik} \\ = \frac{1}{n}\sum_{k=1}^{n}(m(\boldsymbol{Z}_i)-m(\boldsymbol{Z}_k))K_{h,ik}+\frac{1}{n}\sum_{k=1,k\neq i}^{n}\varepsilon_i(K_{h,ik}-f_z(\boldsymbol{Z}_i))-\frac{1}{n}\varepsilon_i f_z(\boldsymbol{Z}_i).$$

Let

$$\left(\boldsymbol{\varepsilon}\left(\boldsymbol{f}_{\boldsymbol{z}}-\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)\right)_n = \left(\tfrac{1}{n}\sum_{k=1,k\neq 1}^{n}\varepsilon_1(f_z(\boldsymbol{Z}_1)-K_{h,1k}),\ldots,\tfrac{1}{n}\sum_{k=1,k\neq n}^{n}\varepsilon_n(f_z(\boldsymbol{Z}_n)-K_{h,nk})\right)^T \text{ and}$$

$$\left(\boldsymbol{m}\widehat{\boldsymbol{f}}_{\boldsymbol{z}} - \widehat{\boldsymbol{m}\boldsymbol{f}}_{\boldsymbol{z}}\right)_n = \left(\tfrac{1}{n}\sum_{k=1}^{n}\left(m(\boldsymbol{Z}_1) - m(\boldsymbol{Z}_k)\right)K_{h,1k}, \ldots, \tfrac{1}{n}\sum_{k=1}^{n}\left(m(\boldsymbol{Z}_n) - m(\boldsymbol{Z}_k)\right)K_{h,nk}\right)^T.$$

We start by showing that

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left(\boldsymbol{\varepsilon}\left(\boldsymbol{f}_{\boldsymbol{z}} - \widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)\right)_n\right\| = o_{\mathbb{P}}(n^{-1/2}). \tag{1.43}$$

We get that

$$\begin{aligned}
&\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left(\boldsymbol{\varepsilon}\left(\boldsymbol{f}_{\boldsymbol{z}} - \widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)\right)_n\right\|^2 \\
&\quad = \frac{1}{n^2}\sum_{1\le i\neq j\le n}\left(\boldsymbol{\varepsilon}\left(\boldsymbol{f}_{\boldsymbol{z}} - \widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)\right)_{n,i}\left(\boldsymbol{\varepsilon}\left(\boldsymbol{f}_{\boldsymbol{z}} - \widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)\right)_{n,j}\boldsymbol{\Omega}_{n,ij} + \frac{1}{n^2}\sum_{i=1}^{n}\left(\boldsymbol{\varepsilon}\left(\boldsymbol{f}_{\boldsymbol{z}} - \widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)\right)_{n,i}^2 \\
&\quad = A_n + B_n.
\end{aligned}$$

It is easy to check that $\sup_{h\in\mathcal{H}_{sc,n}} B_n = o_{\mathbb{P}}(n^{-1})$. Furthermore, we get that

$$A_n = \frac{1}{n^2}\sum_{1\le i\neq j\le n}\left[\frac{1}{n}\sum_{1\le k\le n, k\neq i}\varepsilon_i\left(f_z(\boldsymbol{Z}_i) - K_{h,ik}\right)\right]\left[\frac{1}{n}\sum_{1\le l\le n, l\neq j}\varepsilon_j\left(f_z(\boldsymbol{Z}_j) - K_{h,jl}\right)\right]\boldsymbol{\Omega}_{n,ij}.$$

We show in the following that

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}|A_n| = \sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}|A_n(h)| = o_{\mathbb{P}}(n^{-1}). \tag{1.44}$$

For this purpose, we define $(n)_k = n(n-1)\ldots(n-k+1)$ and decompose $A_n(h)$ into a sum of four $U-$processes, i.e.

$$A_n(h) = \frac{(n-1)_3}{n^3}A_{1,n}(h) + \frac{(n-1)_2}{n^2}A_{2,n}(h) + 2\frac{(n-1)_2}{n^2}A_{3,n}(h) + \frac{n-1}{n}A_{4,n}(h),$$

where

$$\begin{aligned}
A_{1,n} = A_{1,n}(h) &= \frac{1}{(n)_4}\sum_{1\le i\neq j\neq k\neq l\le n}\varepsilon_i\left(f_z(\boldsymbol{Z}_i) - K_{h,ik}\right)\varepsilon_j\left(f_z(\boldsymbol{Z}_j) - K_{h,jl}\right)\boldsymbol{\Omega}_{n,ij} \\
A_{2,n} = A_{2,n}(h) &= \frac{1}{n(n)_3}\sum_{1\le i\neq j\neq k\le n}\varepsilon_i\left(f_z(\boldsymbol{Z}_i) - K_{h,ik}\right)\varepsilon_j\left(f_z(\boldsymbol{Z}_j) - K_{h,jk}\right)\boldsymbol{\Omega}_{n,ij} \\
A_{3,n} = A_{3,n}(h) &= \frac{1}{n(n)_3}\sum_{1\le i\neq j\neq l\le n}\varepsilon_i\left(f_z(\boldsymbol{Z}_i) - K_{h,ij}\right)\varepsilon_j\left(f_z(\boldsymbol{Z}_j) - K_{h,jl}\right)\boldsymbol{\Omega}_{n,ij} \\
\text{and}\quad A_{4,n} = A_{4,n}(h) &= \frac{1}{n^2(n)_2}\sum_{1\le i\neq j\le n}\varepsilon_i\left(f_z(\boldsymbol{Z}_i) - K_{h,ij}\right)\varepsilon_j\left(f_z(\boldsymbol{Z}_j) - K_{h,ij}\right)\boldsymbol{\Omega}_{n,ij}.
\end{aligned}$$

For each of these U–processes we compute the mean and use the Hoeffding decomposition. The kernels of $A_{1,n}$, $A_{2,n}$ and $A_{3,n}$ are not symmetric in their arguments. However, we could apply the usual symmetrization idea. For instance, for a second order $U-$statistic defined by a kernel $h(\boldsymbol{U}_i, \boldsymbol{U}_j)$, we could replace it by the symmetric kernel $\frac{1}{2}\left[h(\boldsymbol{U}_i, \boldsymbol{U}_j) + h(\boldsymbol{U}_j, \boldsymbol{U}_i)\right]$ from which we get the same $U-$statistic. Here, $\boldsymbol{U}_i = \left(Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T\right)^T$. We can proceed in the same way by considering all 4! permutations of the variables for $A_{1,n}$ and 3! permutations for $A_{2,n}$ and $A_{3,n}$ so that we can apply the Hoeffding decomposition. Thus, by abuse, we will proceed as if the kernels of the $U-$statistics we handle are symmetric.

In addition, we have that the kernels of $A_{1,n}$, $A_{2,n}$, $A_{3,n}$ and $A_{4,n}$ are Euclidean for a squared integrable envelope. See Lemma 22 in Nolan and Pollard [67] and Lemma 2.14 in Pakes and Pollard [68]. Therefore, we can in the following repeatedly apply Corollary 7 and the Maximal Inequality of Sherman [72]. All remainder terms are controlled by Assumption 1.3.2.

We start by considering $A_{1,n}$. Recall that by assumption $E\left[\varepsilon_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i\right] = E\left[\varepsilon_j \mid \boldsymbol{X}_j, \boldsymbol{Z}_j\right] = 0$. Therefore, we get that $E\left[A_{1,n}\right] = 0$ as well as

$$E\left[\varepsilon_i\left(f_z(\boldsymbol{Z}_i) - K_{h,ik}\right)\varepsilon_j\left(f_z(\boldsymbol{Z}_j) - K_{h,jl}\right)\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_p, p\in\{i,j,k,l\}\right] = 0.$$

Note that we need to consider the conditional expectations with respect to all four variables for the first order $U-$process of the Hoeffding decomposition of $A_{1,n}$ as we symmetrized the kernel. It follows from the results that the first order $U-$process of the Hoeffding decomposition of $A_{1,n}$ is 0.

We consider now the six second order $U-$processes of the Hoeffding decomposition of $A_{1,n}$. There are two types of such processes. First, the ones that are 0. This is the case when conditioning by the pairs $(\boldsymbol{U}_i, \boldsymbol{U}_l)$, $(\boldsymbol{U}_i, \boldsymbol{U}_k)$, $(\boldsymbol{U}_j, \boldsymbol{U}_l)$, $(\boldsymbol{U}_j, \boldsymbol{U}_k)$ and $(\boldsymbol{U}_l, \boldsymbol{U}_k)$. The second case occurs when conditioning on $(\boldsymbol{U}_i, \boldsymbol{U}_j)$. We get that

$$\begin{aligned} E[\varepsilon_i\,(f_z(\boldsymbol{Z}_i) - K_{h,ik})\,\varepsilon_j\,(f_z(\boldsymbol{Z}_j) - K_{h,jl})\,\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_i, \boldsymbol{U}_j] \\ = \varepsilon_i\varepsilon_j E\left[(f_z(\boldsymbol{Z}_i) - K_{h,ik})\,(f_z(\boldsymbol{Z}_j) - K_{h,jl}) \mid \boldsymbol{Z}_i, \boldsymbol{Z}_j\right]\boldsymbol{\Omega}_{n,ij} \\ = \varepsilon_i\varepsilon_j E\left[(f_z(\boldsymbol{Z}_i) - K_{h,ik}) \mid \boldsymbol{Z}_i\right] E\left[(f_z(\boldsymbol{Z}_j) - K_{h,jl}) \mid \boldsymbol{Z}_j\right]\boldsymbol{\Omega}_{n,ij} \\ = \varepsilon_i\varepsilon_j h^4\gamma_1(\boldsymbol{Z}_i)\gamma_1(\boldsymbol{Z}_j)(1 + o_{\mathbb{P}}(1))\boldsymbol{\Omega}_{n,ij}, \end{aligned}$$

where

$$\gamma_1(\boldsymbol{Z}) = \mu(K)\cdot \operatorname{tr}\{\boldsymbol{H}_{z,z}f_z(\boldsymbol{Z})\},$$

with $\int_{\mathbb{R}^q} \boldsymbol{u}\boldsymbol{u}^T K(\boldsymbol{u})d\boldsymbol{u} = \mu(K)\boldsymbol{I}_{q\times q}$. $\boldsymbol{H}_{z,z}f_z$ denotes the matrix of second derivative of $f_z(\cdot)$ with respect to the components of $\boldsymbol{Z} \in \mathbb{R}^q$ and $\operatorname{tr}\{\cdot\}$ denotes the trace operator. Therefore, it follows together with Corollary 4 of Sherman [72] that the second order $U-$processes of the Hoeffding decomposition of $A_{1,n}$ are of order $O_{\mathbb{P}}(n^{-1}n^{-4\alpha}) = o_{\mathbb{P}}(n^{-1})$ uniformly with respect to h and $\boldsymbol{d}$.

We consider now the four $U-$processes of order three obtained by conditioning on any subset of three of the four vectors $\boldsymbol{U}_i$, $\boldsymbol{U}_k$, $\boldsymbol{U}_j$ and $\boldsymbol{U}_l$. There are two types of such processes. First, the ones that are 0. This is the case when conditioning by $(\boldsymbol{U}_i, \boldsymbol{U}_l, \boldsymbol{U}_k)$ or $(\boldsymbol{U}_j, \boldsymbol{U}_l, \boldsymbol{U}_k)$. The second case occurs when conditioning on $(\boldsymbol{U}_i, \boldsymbol{U}_j, \boldsymbol{U}_l)$ or $(\boldsymbol{U}_i, \boldsymbol{U}_j, \boldsymbol{U}_k)$, the other one being similar. We get that

$$\begin{aligned} E[\varepsilon_i\,(f_z(\boldsymbol{Z}_i) - K_{h,ik})\,\varepsilon_j(f_z(\boldsymbol{Z}_j) - K_{h,jl})\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_i, \boldsymbol{U}_j, \boldsymbol{U}_k] \\ = \varepsilon_i\,(f_z(\boldsymbol{Z}_i) - K_{h,ik})\,\varepsilon_j E\left[(f_z(\boldsymbol{Z}_j) - K_{h,jl}) \mid \boldsymbol{Z}_j\right]\boldsymbol{\Omega}_{n,ij} \\ = \varepsilon_i\,(f_z(\boldsymbol{Z}_i) - K_{h,ik})\,\varepsilon_j h^2\gamma_1(\boldsymbol{Z}_j)(1 + o_{\mathbb{P}}(1))\boldsymbol{\Omega}_{n,ij} \\ = \varepsilon_i f_z(\boldsymbol{Z}_i)\varepsilon_j h^2\gamma_1(\boldsymbol{Z}_j)(1 + o_{\mathbb{P}}(1))\boldsymbol{\Omega}_{n,ij} \\ - \varepsilon_i K_{h,ik}\varepsilon_j h^2\gamma_1(\boldsymbol{Z}_j)(1 + o_{\mathbb{P}}(1))\boldsymbol{\Omega}_{n,ij} \\ = \varepsilon_i f_z(\boldsymbol{Z}_i)\varepsilon_j h^2\gamma_1(\boldsymbol{Z}_j)(1 + o_{\mathbb{P}}(1))\boldsymbol{\Omega}_{n,ij} \\ - h^{2-q}h^q K_{h,ik}\tau(\boldsymbol{U}_i, \boldsymbol{U}_j)(1 + o_{\mathbb{P}}(1)). \end{aligned}$$

Now, we apply the Maximal Inequality of Sherman [72], page 448, for the degenerate $U-$process given by the kernel $h^q K_{h,ik}\tau(\boldsymbol{U}_i, \boldsymbol{U}_j)$, indexed by $h \in \mathcal{H}_{sc,n}$, with envelope $\|K\|_\infty \tau(\cdot,\cdot)$. (Herein, $\|\cdot\|_\infty$ denotes the uniform norm.) We take $p = 1$ and $\beta \in (0,1)$ arbitrarily close to 1 to stand for Sherman's quantity α. Since $K(\cdot)$ is of bounded variation and symmetric, without loss of generality we could consider that $K(\cdot)$ is nonincreasing on $[0,\infty)$. In this case, $0 \leq K(\cdot/h) \leq K(\cdot/\overline{h})$ with $\overline{h} = \sup\mathcal{H}_{sc,n} =: c_{max}n^{-\alpha}$. Hence, using Jensen's inequality, we could bound the right-hand side of the Maximal Inequality of Sherman [72] by a universal constant times

$$E^{\beta/2}\left[K^2\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_k}{c_{max}n^{-\alpha}}\right)\tau^2(\boldsymbol{U}_i, \boldsymbol{U}_j)\right].$$

By standard changes of variables and suitable integrability conditions, this integral is bounded by a constant times $n^{-\alpha\beta q/2}$. Consequently, the uniform rate of the second $U-$processes obtained conditioning by $\boldsymbol{U}_i, \boldsymbol{U}_j, \boldsymbol{U}_k$ and $\boldsymbol{U}_i, \boldsymbol{U}_j, \boldsymbol{U}_l$, respectively is $n^{-3/2} \times O_{\mathbb{P}}(n^{-\alpha\{2-q+\beta q/2\}})$. As $1/2 + \alpha(2 - q + \beta q/2) > 0$ under our assumptions we get that $n^{-3/2} \times O_{\mathbb{P}}(n^{-\alpha\{2-q+\beta q/2\}}) = o_{\mathbb{P}}(n^{-1})$ such that the third order $U-$processes of the Hoeffding decomposition of $A_{1,n}$ are of order $o_{\mathbb{P}}(n^{-1})$.

Finally, we consider the remaining $U-$process of order four. This process is given by

$$\begin{aligned} \varepsilon_i\,(f_z(\boldsymbol{Z}_i) - K_{h,ik})\,\varepsilon_j(f_z(\boldsymbol{Z}_j) - K_{h,jl})\boldsymbol{\Omega}_{n,ij} \\ = \varepsilon_i\varepsilon_j\,(f_z(\boldsymbol{Z}_i)f_z(\boldsymbol{Z}_j) - f_z(\boldsymbol{Z}_j)K_{h,ik} - f_z(\boldsymbol{Z}_i)K_{h,jl} + K_{h,ik}K_{h,jl})\,\boldsymbol{\Omega}_{n,ij} \\ = \varepsilon_i\varepsilon_j f_z(\boldsymbol{Z}_i)f_z(\boldsymbol{Z}_j)\boldsymbol{\Omega}_{n,ij} - h^{-q}\tau_1(\boldsymbol{U}_i, \boldsymbol{U}_j)h^q K_{h,jl} - h^{-q}\tau_2(\boldsymbol{U}_i, \boldsymbol{U}_j)h^q K_{h,ik} \\ + h^{-2q}\tau_3(\boldsymbol{U}_i, \boldsymbol{U}_j)h^{2q}K_{h,ik}K_{h,jl}. \end{aligned}$$

Now, we apply again the Maximal Inequality of Sherman [72], page 448, for the degenerate $U-$process given by the kernel $\tau_3(\boldsymbol{U}_i, \boldsymbol{U}_j)h^q K_{h,ik}h^q K_{h,jl}$, indexed by $h \in \mathcal{H}_{sc,n}$, with envelope $\|K\|_\infty^2\tau(\cdot,\cdot)$. We take again $p = 1$ and $\beta \in (0,1)$ arbitrarily close to 1 to stand for Sherman's quantity α. Hence, using Jensen's inequality, we could bound the right-hand side of the Maximal Inequality of Sherman [72] by a universal constant times

$$E^{\beta/2}\left[K^2\left(\frac{\boldsymbol{Z}_i-\boldsymbol{Z}_k}{c_{max}n^{-\alpha}}\right)K^2\left(\frac{\boldsymbol{Z}_j-\boldsymbol{Z}_l}{c_{max}n^{-\alpha}}\right)\tau_3^2(\boldsymbol{U}_i,\boldsymbol{U}_j)\right].$$

By standard changes of variables and suitable integrability conditions, this integral is bounded by a constant times $n^{-\alpha\beta q}$. Consequently, the uniform rate of the fourth order $U-$process is $n^{-2}\times O_{\mathbb{P}}(n^{\alpha q\{2-\beta\}})$. Since $1>\alpha q$ under our assumptions we get that $n^{-2}\times O_{\mathbb{P}}(n^{\alpha q\{2-\beta\}})=o_{\mathbb{P}}(n^{-1})$. By the same reasoning we can control $h^{-q}\tau_1(\boldsymbol{U}_i,\boldsymbol{U}_j)h^qK_{h,jl}$ and $h^{-q}\tau_2(\boldsymbol{U}_i,\boldsymbol{U}_j)h^qK_{h,ik}$. The details are omitted.

From all the results it follows that

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}|A_{1,n}|=\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}|A_{1,n}(h)|=o_{\mathbb{P}}(n^{-1}).$$

In the next step we consider $A_{2,n}$. We get that $E\left[A_{2,n}\right]=0$ as well as

$$E\left[\varepsilon_i\left(f_z(\boldsymbol{Z}_i)-K_{h,ik}\right)\varepsilon_j\left(f_z(\boldsymbol{Z}_j)-K_{h,jk}\right)\boldsymbol{\Omega}_{n,ij}\mid\boldsymbol{U}_p,p\in\{i,j,k\}\right]=0.$$

In addition, it is easy to see that the second and third order U–processes of the Hoeffding decomposition of $A_{2,n}$ are of order $o_{\mathbb{P}}(n^{-1})$ if we apply the Maximal Inequality of Sherman [72]. From all the results it follows that

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}|A_{2,n}|=\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}|A_{2,n}(h)|=o_{\mathbb{P}}(n^{-1}).$$

As it follows by the same reasoning as for $A_{2,n}$ that

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}|A_{3,n}|=\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}|A_{3,n}(h)|=o_{\mathbb{P}}(n^{-1}).$$

we omit the details here.

Finally, we get that $E\left[A_{4,n}\right]=0$ as well as

$$E\left[\varepsilon_i\left(f_z(\boldsymbol{Z}_i)-K_{h,ij}\right)\varepsilon_j\left(f_z(\boldsymbol{Z}_j)-K_{h,ij}\right)\boldsymbol{\Omega}_{n,ij}\mid\boldsymbol{U}_p,p\in\{i,j\}\right]=0.$$

In addition, it is easy to see that the second order U–process of the Hoeffding decomposition of $A_{4,n}$ is of order $o_{\mathbb{P}}(n^{-1})$ if we apply the Maximal Inequality of Sherman [72]. Deduce that

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}|A_{4,n}|=o_{\mathbb{P}}(n^{-1}).$$

With all these results (1.44) and, in particular, (1.43) follow.

In the next part we show that

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left(\boldsymbol{m}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}-\widehat{\boldsymbol{mf}}_{\boldsymbol{z}}\right)_n\right\|=o_{\mathbb{P}}(n^{-1/2}).\tag{1.45}$$

We get that

$$\begin{aligned}&\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left(\boldsymbol{m}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}-\widehat{\boldsymbol{mf}}_{\boldsymbol{z}}\right)_n\right\|^2\\&\quad=\frac{1}{n^2}\sum_{1\le i\ne j\le n}\left(\boldsymbol{m}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}-\widehat{\boldsymbol{mf}}_{\boldsymbol{z}}\right)_{n,i}\left(\boldsymbol{m}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}-\widehat{\boldsymbol{mf}}_{\boldsymbol{z}}\right)_{n,j}\boldsymbol{\Omega}_{n,ij}+\frac{1}{n^2}\sum_{i=1}^n\left(\boldsymbol{m}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}-\widehat{\boldsymbol{mf}}_{\boldsymbol{z}}\right)_{n,i}^2\\&\quad=\widetilde{A}_n+\widetilde{B}_n.\end{aligned}$$

It is easy to check that $\sup_{h\in\mathcal{H}_{sc,n}}\widetilde{B}_n=o_{\mathbb{P}}(n^{-1})$. Furthermore, we get that

$$\widetilde{A}_n=\frac{1}{n^2}\sum_{1\le i\ne j\le n}\left[\frac{1}{n}\sum_{1\le k\le n,k\ne i}\left(m(\boldsymbol{Z}_i)-m(\boldsymbol{Z}_k)\right)K_{h,ik}\right]\left[\frac{1}{n}\sum_{1\le l\le n,l\ne j}\left(m(\boldsymbol{Z}_j)-m(\boldsymbol{Z}_l)\right)K_{h,jl}\right]\boldsymbol{\Omega}_{n,ij}.$$

We show in the following that

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}|\widetilde{A}_n|=\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}|\widetilde{A}_n(h)|=o_{\mathbb{P}}(n^{-1}).\tag{1.46}$$

We decompose $\widetilde{A}_n(h)$ into a sum of four $U-$ processes, i.e.

$$\widetilde{A}_n(h) = \frac{(n-1)_3}{n^3}\widetilde{A}_{1,n}(h) + \frac{(n-1)_2}{n^2}\widetilde{A}_{2,n}(h) + 2\frac{(n-1)_2}{n^2}\widetilde{A}_{3,n}(h) - \frac{n-1}{n}\widetilde{A}_{4,n}(h),$$

where

$$\begin{aligned}
\widetilde{A}_{1,n} = \widetilde{A}_{1,n}(h) &= \frac{1}{(n)_4}\sum_{1\le i\ne j\ne k\ne l\le n}\left(m(\boldsymbol{Z}_i)-m(\boldsymbol{Z}_k)\right)K_{h,ik}\left(m(\boldsymbol{Z}_j)-m(\boldsymbol{Z}_l)\right)K_{h,jl}\boldsymbol{\Omega}_{n,ij}\\
\widetilde{A}_{2,n} = \widetilde{A}_{2,n}(h) &= \frac{1}{n(n)_3}\sum_{1\le i\ne j\ne k\le n}\left(m(\boldsymbol{Z}_i)-m(\boldsymbol{Z}_k)\right)K_{h,ik}\left(m(\boldsymbol{Z}_j)-m(\boldsymbol{Z}_k)\right)K_{h,jk}\boldsymbol{\Omega}_{n,ij}\\
\widetilde{A}_{3,n} = \widetilde{A}_{3,n}(h) &= \frac{1}{n(n)_3}\sum_{1\le i\ne j\ne l\le n}\left(m(\boldsymbol{Z}_i)-m(\boldsymbol{Z}_j)\right)K_{h,ij}\left(m(\boldsymbol{Z}_j)-m(\boldsymbol{Z}_l)\right)K_{h,jl}\boldsymbol{\Omega}_{n,ij}\\
\text{and}\quad \widetilde{A}_{4,n} = \widetilde{A}_{4,n}(h) &= \frac{1}{n^2(n)_2}\sum_{1\le i\ne j\le n}\left(m(\boldsymbol{Z}_i)-m(\boldsymbol{Z}_j)\right)^2K_{h,ij}^2\boldsymbol{\Omega}_{n,ij}.
\end{aligned}$$

For each of these $U-$processes we compute the mean and use the Hoeffding decomposition. The kernels of $\widetilde{A}_{1,n}$, $\widetilde{A}_{2,n}$ and $\widetilde{A}_{3,n}$ are not symmetric in their arguments. However, we could apply the usual symmetrization idea. Here, $\widetilde{\boldsymbol{U}}_i = \left(\boldsymbol{X}_i^T, \boldsymbol{Z}_i^T\right)^T$. Thus, by abuse, we will proceed as if the kernels of the $U-$statistics we handle are symmetric. For simpler formulae, we use the short notation $m_i, m_k, \ldots$ instead of $m(\boldsymbol{Z}_i), m(\boldsymbol{Z}_k), \ldots$.

In addition, we have that the kernels of $\widetilde{A}_{1,n}$, $\widetilde{A}_{2,n}$, $\widetilde{A}_{3,n}$ and $\widetilde{A}_{4,n}$ are Euclidean for a squared integrable envelope. See Lemma 22 in Nolan and Pollard [67] and Lemma 2.14 in Pakes and Pollard [68]. Therefore, we can in the following repeatedly apply Corollary 7 and the Maximal Inequality of Sherman [72]. All remainder terms are controlled by Assumption 1.3.2.

We start by considering the expectation of $\widetilde{A}_{1,n}$. We get that

$$\begin{aligned}
E[\widetilde{A}_{1,n}] &= E\left[(m_i - m_k)\,K_{h,ik}\,(m_j - m_l)\,K_{h,jl}\boldsymbol{\Omega}_{n,ij}\right]\\
&= E\left[E\left[(m_i - m_k)\,K_{h,ik}\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \widetilde{\boldsymbol{U}}_j, \boldsymbol{Z}_l\right](m_j - m_l)\,K_{h,jl}\boldsymbol{\Omega}^X_{n,ij}\right].
\end{aligned}$$

Next, by Taylor expansion and Dominated convergence

$$\begin{aligned}
E\left[m_iK_{h,ik}\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \widetilde{\boldsymbol{U}}_j, \boldsymbol{Z}_l\right] &= E\left[m_iE\left[K_{h,ik} \mid \boldsymbol{Z}_i\right]\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \boldsymbol{Z}_j\right]\\
&= E\left[m_i(f_z(\boldsymbol{Z}_i) + h^2\gamma_1(\boldsymbol{Z}_i)(1+o_{\mathbb{P}}(1)))\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \boldsymbol{Z}_j\right]\\
&= E\left[m_if_z(\boldsymbol{Z}_i)\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \boldsymbol{Z}_j\right]\\
&\quad + h^2E\left[m_i\gamma_1(\boldsymbol{Z}_i)\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \boldsymbol{Z}_j\right](1+o_{\mathbb{P}}(1)).
\end{aligned}$$

Similarly,

$$\begin{aligned}
E\left[m_kK_{h,ik}\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \widetilde{\boldsymbol{U}}_j, \boldsymbol{Z}_l\right] &= E\left[E\left[m_kK_{h,ik} \mid \boldsymbol{Z}_i\right]\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \boldsymbol{Z}_j\right]\\
&= E\left[(m_if_z(\boldsymbol{Z}_i) + h^2\gamma_2(\boldsymbol{Z}_i)(1+o_{\mathbb{P}}(1))\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \boldsymbol{Z}_j\right]\\
&= E\left[m_if_z(\boldsymbol{Z}_i)\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \boldsymbol{Z}_j\right]\\
&\quad + h^2E\left[\gamma_2(\boldsymbol{Z}_i)\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \boldsymbol{Z}_j\right](1+o_{\mathbb{P}}(1)),
\end{aligned}$$

where $\gamma_2(\boldsymbol{Z}) = \mu(K)\cdot\operatorname{tr}\{\boldsymbol{H}_{z,z}\,(mf_z)\,(\boldsymbol{Z})\}$. $\boldsymbol{H}_{z,z}\,(mf_z)$ denotes the matrix of second derivative of $mf_z(\cdot)$ with respect to the components of $\boldsymbol{Z}\in\mathbb{R}^q$. Thus,

$$\begin{aligned}
E[\widetilde{A}_{1,n}] &= E\left[E\left[(m_i - m_k)K_{h,ik}\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \widetilde{\boldsymbol{U}}_j, \boldsymbol{Z}_l\right](m_j - m_l)K_{h,jl}\boldsymbol{\Omega}^X_{n,ij}\right]\\
&= h^2E\left[\gamma_3(\boldsymbol{X}_i, \boldsymbol{Z}_j)(m_j - m_l)K_{h,jl}\boldsymbol{\Omega}^X_{n,ij}\right](1+o(1))\\
&= h^2E\left[\gamma_3(\boldsymbol{X}_i, \boldsymbol{Z}_j)E\left[(m_j - m_l)K_{h,jl} \mid \boldsymbol{X}_i, \boldsymbol{X}_j, \boldsymbol{Z}_j\right]\boldsymbol{\Omega}^X_{n,ij}\right](1+o(1))\\
&= h^2E\left[\gamma_3(\boldsymbol{X}_i, \boldsymbol{Z}_j)E\left[(m_j - m_l)K_{h,jl} \mid \boldsymbol{Z}_j\right]\boldsymbol{\Omega}^X_{n,ij}\right](1+o(1)),
\end{aligned}$$

where $\gamma_3(\boldsymbol{X}_i, \boldsymbol{Z}_j) = E\left[m_i\gamma_1(\boldsymbol{Z}_i)\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \boldsymbol{Z}_j\right] - E\left[\gamma_2(\boldsymbol{Z}_i)\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{X}_i, \boldsymbol{Z}_j\right]$.

In the next step we consider $E\left[(m_j - m_l)K_{h,jl} \mid \boldsymbol{Z}_j\right]$. We get that

$$\begin{aligned} E\left[m_j K_{h,jl} \mid \boldsymbol{Z}_j\right] &= m_j E\left[K_{h,jl} \mid \boldsymbol{Z}_j\right] \\ &= m_j \left(f_z(\boldsymbol{Z}_j) + h^2\gamma_1(\boldsymbol{Z}_j)(1+o_{\mathbb{P}}(1))\right) \\ &= m_j f_z(\boldsymbol{Z}_j) + h^2 m_j \gamma_1(\boldsymbol{Z}_j)(1+o_{\mathbb{P}}(1)). \end{aligned}$$

In addition, we get that

$$\begin{aligned} E\left[m_l K_{h,jl} \mid \boldsymbol{Z}_j\right] &= E\left[m_j f_z(\boldsymbol{Z}_j) + h^2\gamma_2(\boldsymbol{Z}_j)(1+o_{\mathbb{P}}(1)) \mid \boldsymbol{Z}_j\right] \\ &= m_j f_z(\boldsymbol{Z}_j) + h^2\gamma_2(\boldsymbol{Z}_j)(1+o_{\mathbb{P}}(1)). \end{aligned}$$

Therefore, we get that

$$\begin{aligned} E[\widetilde{A}_{1,n}] &= h^2 E\left[\gamma_3(\boldsymbol{X}_i,\boldsymbol{Z}_j) E\left[(m_j - m_l)K_{h,jl} \mid \boldsymbol{Z}_j\right] \boldsymbol{\Omega}^X_{n,ij}\right](1+o(1)) \\ &= h^4 E\left[\gamma_3(\boldsymbol{X}_i,\boldsymbol{Z}_j)\left(m_j\gamma_1(\boldsymbol{Z}_j) - \gamma_2(\boldsymbol{Z}_j)\right) \boldsymbol{\Omega}^X_{n,ij}\right](1+o(1)). \end{aligned}$$

This implies that $E[\widetilde{A}_{1,n}] = o_{\mathbb{P}}(n^{-1})$ uniformly with respect to $h \in \mathcal{H}_{sc,n}$.

We consider now the first order $U-$process of the Hoeffding decomposition for $\widetilde{A}_{1,n}$. As we symmetrized the kernel we need to consider the conditional expectations with respect to all four variables. By the same reasoning as for $E[\widetilde{A}_{1,n}]$ we get that

$$\begin{aligned} E\big[(m_i - m_k)K_{h,ik}(m_j - m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij} \mid \widetilde{\boldsymbol{U}}_i\big] &= h^2 E\left[\gamma_3(\boldsymbol{X}_j,\boldsymbol{Z}_i)(m_i - m_k)K_{h,ik}\boldsymbol{\Omega}^X_{n,ij} \mid \widetilde{\boldsymbol{U}}_i\right](1+o_{\mathbb{P}}(1)) \\ &= h^2 E\left[\gamma_3(\boldsymbol{X}_j,\boldsymbol{Z}_i) E\left[(m_i - m_k)K_{h,ik} \mid \boldsymbol{Z}_i\right]\boldsymbol{\Omega}^X_{n,ij} \mid \widetilde{\boldsymbol{U}}_i\right](1+o_{\mathbb{P}}(1)) \\ &= h^2 E\left[(m_i - m_k)K_{h,ik} \mid \boldsymbol{Z}_i\right] E\left[\gamma_3(\boldsymbol{X}_j,\boldsymbol{Z}_i)\boldsymbol{\Omega}^X_{n,ij} \mid \widetilde{\boldsymbol{U}}_i\right](1+o_{\mathbb{P}}(1)) \\ &= h^4\left(m_i\gamma_1(\boldsymbol{Z}_i) - \gamma_2(\boldsymbol{Z}_i)\right) E\left[\gamma_3(\boldsymbol{X}_j,\boldsymbol{Z}_i)\boldsymbol{\Omega}^X_{n,ij} \mid \widetilde{\boldsymbol{U}}_i\right](1+o_{\mathbb{P}}(1)). \end{aligned}$$

Note that the reasoning for $E\left[(m_i - m_k)K_{h,ik}(m_j - m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij} \mid \widetilde{\boldsymbol{U}}_j\right]$ is exactly the same. In addition, we have that

$$E\left[(m_i - m_k)K_{h,ik}(m_j - m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij} \mid \widetilde{\boldsymbol{U}}_k\right] = h^2 E\left[\gamma_3(\boldsymbol{X}_j,\boldsymbol{Z}_i)(m_i - m_k)K_{h,ik}\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_k\right].$$

We get that

$$\begin{aligned} E\big[\gamma_3(\boldsymbol{X}_j,\boldsymbol{Z}_i) m_i K_{h,ik}\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_k\big] &= E\big[K_{h,ik} E\big[\gamma_3(\boldsymbol{X}_j,\boldsymbol{Z}_i) m_i \boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_i\big] \mid \boldsymbol{Z}_k\big] \\ &= m_k f_z(\boldsymbol{Z}_k) E\big[\gamma_3(\boldsymbol{X}_j,\boldsymbol{Z}_k)\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_k\big] + O_{\mathbb{P}}(h^2), \end{aligned}$$

and

$$\begin{aligned} E\big[\gamma_3(\boldsymbol{X}_j,\boldsymbol{Z}_i) m_k K_{h,ik}\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_k\big] &= m_k E\big[K_{h,ik} E\big[\gamma_3(\boldsymbol{X}_j,\boldsymbol{Z}_i)\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_i\big] \mid \boldsymbol{Z}_k\big] \\ &= m_k f_z(\boldsymbol{Z}_k) E\big[\gamma_3(\boldsymbol{X}_j,\boldsymbol{Z}_k)\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_k\big] + O_{\mathbb{P}}(h^2). \end{aligned}$$

Note that the reasoning for $E\left[(m_i - m_k)K_{h,ik}(m_j - m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij} \mid \widetilde{\boldsymbol{U}}_l\right]$ is exactly the same. Therefore, it follows together with Corollary 4 of Sherman [72] that the first order $U-$processes of the Hoeffding decomposition for $\widetilde{A}_{1,n}$ are of order $o_{\mathbb{P}}(n^{-1})$ uniformly with respect to h.

We consider now the six second order $U-$processes of the Hoeffding decomposition for $\widetilde{A}_{1,n}$. There are two types of such processes. First, the ones where the two kernels $K_{h,ik}$ and $K_{h,jl}$ are both integrated with respect to one of the variables they contain. This is the case when conditioning by the pairs $(\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_j)$, $(\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_l)$, $(\widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{U}}_j)$ and $(\widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{U}}_l)$.

We get that

$$\begin{aligned}
&E\left[(m_i-m_k)K_{h,ik}(m_j-m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij}\mid\widetilde{\boldsymbol{U}}_i,\widetilde{\boldsymbol{U}}_j\right]\\
&\quad=E\left[E\left[(m_i-m_k)K_{h,ik}\mid\boldsymbol{Z}_i\right](m_j-m_l)K_{h,jl}\mid\boldsymbol{Z}_i,\boldsymbol{Z}_j\right]\boldsymbol{\Omega}_{n,ij}\\
&\quad=\left[m_i\left(f_z(\boldsymbol{Z}_i)+h^2\gamma_1(\boldsymbol{Z}_i)(1+o_{\mathbb{P}}(1))\right)-\left(m_if_z(\boldsymbol{Z}_i)+h^2\gamma_2(\boldsymbol{Z}_i)(1+o_{\mathbb{P}}(1))\right)\right]\\
&\qquad\left[m_j\left(f_z(\boldsymbol{Z}_j)+h^2\gamma_1(\boldsymbol{Z}_j)(1+o_{\mathbb{P}}(1))\right)-\left(m_jf_z(\boldsymbol{Z}_j)+h^2\gamma_2(\boldsymbol{Z}_j)(1+o_{\mathbb{P}}(1))\right)\right]\boldsymbol{\Omega}_{n,ij}\\
&\quad=h^4\left[m_i\gamma_1(\boldsymbol{Z}_i)-\gamma_2(\boldsymbol{Z}_i)\right]\left[m_j\gamma_1(\boldsymbol{Z}_j)-\gamma_2(\boldsymbol{Z}_j)\right](1+o_{\mathbb{P}}(1))\boldsymbol{\Omega}_{n,ij}.
\end{aligned}$$

In addition, we get that

$$\begin{aligned}
&E\left[(m_i-m_k)K_{h,ik}(m_j-m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij}\mid\widetilde{\boldsymbol{U}}_i,\widetilde{\boldsymbol{U}}_l\right]\\
&\quad=E\left[E\left[(m_i-m_k)K_{h,ik}\mid\boldsymbol{Z}_i\right](m_j-m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij}\mid\widetilde{\boldsymbol{U}}_i,\widetilde{\boldsymbol{U}}_l\right]\\
&\quad=\left[m_i\left(f_z(\boldsymbol{Z}_i)+h^2\gamma_1(\boldsymbol{Z}_i)(1+o_{\mathbb{P}}(1))\right)-\left(m_if_z(\boldsymbol{Z}_i)+h^2\gamma_2(\boldsymbol{Z}_i)(1+o_{\mathbb{P}}(1))\right)\right]\\
&\qquad E\left[(m_j-m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij}\mid\widetilde{\boldsymbol{U}}_i,\widetilde{\boldsymbol{U}}_l\right]\\
&\quad=h^2\left[m_i\gamma_1(\boldsymbol{Z}_i)(1+o_{\mathbb{P}}(1))-\gamma_2(\boldsymbol{Z}_i)(1+o_{\mathbb{P}}(1))\right]\\
&\qquad\left[m_lf_z(\boldsymbol{Z}_l)\boldsymbol{\Omega}^Z_{n,il}E\left[\boldsymbol{\Omega}^X_{n,ij}\mid\boldsymbol{X}_i\right]-m_lf_z(\boldsymbol{Z}_l)\boldsymbol{\Omega}^Z_{n,il}E\left[\boldsymbol{\Omega}^X_{n,ij}\mid\boldsymbol{X}_i\right]+O_{\mathbb{P}}(h^2)\right]=O_{\mathbb{P}}(h^4).
\end{aligned}$$

The reasoning when conditioning on $(\widetilde{\boldsymbol{U}}_k,\widetilde{\boldsymbol{U}}_j)$ is the same. For the fourth part we get that

$$\begin{aligned}
&E\left[(m_i-m_k)K_{h,ik}(m_j-m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij}\mid\widetilde{\boldsymbol{U}}_k,\widetilde{\boldsymbol{U}}_l\right]\\
&=E\left[(m_i-m_k)K_{h,ik}E\left[(m_j-m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij}\mid\boldsymbol{Z}_l,\widetilde{\boldsymbol{U}}_i\right]\mid\widetilde{\boldsymbol{U}}_k,\widetilde{\boldsymbol{U}}_l\right]\\
&=E\left[(m_i-m_k)K_{h,ik}\left(m_lf_z(\boldsymbol{Z}_l)\boldsymbol{\Omega}^Z_{n,il}E\left[\boldsymbol{\Omega}^X_{n,ij}\mid\boldsymbol{X}_i\right]-m_lf_z(\boldsymbol{Z}_l)\boldsymbol{\Omega}^Z_{n,il}E\left[\boldsymbol{\Omega}^X_{n,ij}\mid\boldsymbol{X}_i\right]+O_{\mathbb{P}}(h^2)\right)\mid\widetilde{\boldsymbol{U}}_k,\widetilde{\boldsymbol{U}}_l\right]\\
&=\left(m_kf_z(\boldsymbol{Z}_k)-m_kf_z(\boldsymbol{Z}_k)+O_{\mathbb{P}}(h^2)\right)O_{\mathbb{P}}(h^2)=O_{\mathbb{P}}(h^4).
\end{aligned}$$

Applying the results of Sherman [72], the four $U-$processes for which the two kernels $K_{h,ik}$ and $K_{h,jl}$ are both integrated with respect to one of their variables have the uniform rate $o_{\mathbb{P}}(n^{-1})$.

Next, we investigate one of the two $U-$processes of the Hoeffding decomposition obtained by conditioning on the pairs $(\widetilde{\boldsymbol{U}}_i,\widetilde{\boldsymbol{U}}_k)$ and $(\widetilde{\boldsymbol{U}}_j,\widetilde{\boldsymbol{U}}_l)$, the other one being similar. We have

$$\begin{aligned}
E\left[(m_i-m_k)K_{h,ik}(m_j-m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij}\mid\widetilde{\boldsymbol{U}}_j,\widetilde{\boldsymbol{U}}_l\right]&=(m_j-m_l)K_{h,jl}E\left[(m_i-m_k)K_{h,ik}\boldsymbol{\Omega}_{n,ij}\mid\widetilde{\boldsymbol{U}}_j\right]\\
&=(m_j-m_l)K_{h,jl}E\left[E\left[(m_i-m_k)K_{h,ik}\boldsymbol{\Omega}^Z_{n,ij}\mid\boldsymbol{X}_i,\boldsymbol{Z}_j\right]\boldsymbol{\Omega}^X_{n,ij}\mid\widetilde{\boldsymbol{U}}_j\right]\\
&=h^2(m_j-m_l)K_{h,jl}E\left[\gamma_3(\boldsymbol{X}_i,\boldsymbol{Z}_j)\boldsymbol{\Omega}^X_{n,i,j}\mid\widetilde{\boldsymbol{U}}_j\right](1+o_{\mathbb{P}}(1))\\
&=:h^{2-q}(1+o_{\mathbb{P}}(1))\times h^qK_{h,jl}\tau(\widetilde{\boldsymbol{U}}_j,\widetilde{\boldsymbol{U}}_l).
\end{aligned}$$

Now, we apply the Maximal Inequality of Sherman [72], page 448, for the degenerate $U-$process given by the kernel $h^qK_{h,jl}\tau(\widetilde{\boldsymbol{U}}_j,\widetilde{\boldsymbol{U}}_l)$, indexed by $h\in\mathcal{H}_{sc,n}$, with envelope $\|K\|_\infty\tau(\cdot,\cdot)$. We take $p=1$ and $\beta\in(0,1)$ arbitrarily close to 1 to stand for Sherman's quantity α. Using Jensen's inequality, we could bound the right-hand side of the Maximal Inequality of Sherman [72] by a universal constant times

$$E^{\beta/2}\left[K^2\left(\frac{\boldsymbol{Z}_j-\boldsymbol{Z}_l}{c_{max}n^{-\alpha}}\right)\tau^2(\widetilde{\boldsymbol{U}}_j,\widetilde{\boldsymbol{U}}_l)\right].$$

By standard changes of variables and suitable integrability conditions, this integral is bounded by a constant times $n^{-\alpha\beta q/2}$. Consequently, the uniform rate of the second $U-$processes obtained conditioning by $\widetilde{\boldsymbol{U}}_i,\widetilde{\boldsymbol{U}}_k$ and $\widetilde{\boldsymbol{U}}_j,\widetilde{\boldsymbol{U}}_l$, respectively is $n^{-1}\times O_{\mathbb{P}}(n^{-\alpha\{2-q+\beta q/2\}})$. Since $\beta<1$ could be arbitrarily close to 1, we have $2-q+\beta q/2>0$, and, thus, $n^{-1}\times O_{\mathbb{P}}(n^{-\alpha\{2-q+\beta q/2\}})=o_{\mathbb{P}}(n^{-1})$.

We consider now the four $U-$processes of order three obtained by conditioning on any subset of three of the four vectors $\widetilde{\boldsymbol{U}}_i$, $\widetilde{\boldsymbol{U}}_k$, $\widetilde{\boldsymbol{U}}_j$ and $\widetilde{\boldsymbol{U}}_l$. We start by conditioning on $(\widetilde{\boldsymbol{U}}_i,\widetilde{\boldsymbol{U}}_j,\widetilde{\boldsymbol{U}}_l)$ and $(\widetilde{\boldsymbol{U}}_i,\widetilde{\boldsymbol{U}}_j,\widetilde{\boldsymbol{U}}_k)$, the other one being similar.

$$\begin{aligned}
&E[(m_i - m_k)K_{h,ik}(m_j - m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij} \mid \widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_j, \widetilde{\boldsymbol{U}}_l] \\
&\quad = E\left[(m_i - m_k)K_{h,ik} \mid \widetilde{\boldsymbol{U}}_i\right](m_j - m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij} \\
&\quad = \left[m_i\left(f_z(\boldsymbol{Z}_i) + h^2\gamma_1(\boldsymbol{Z}_i)(1 + o_{\mathbb{P}}(1))\right) - \left(m_i f_z(\boldsymbol{Z}_i) + h^2\gamma_2(\boldsymbol{Z}_i)(1 + o_{\mathbb{P}}(1))\right)\right](m_j - m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij} \\
&\quad = h^{2-q}\tau(\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_j, \widetilde{\boldsymbol{U}}_l)h^q K_{h,jl}(1 + o_{\mathbb{P}}(1)).
\end{aligned}$$

Now, we apply again the Maximal Inequality of Sherman [72], page 448, for the degenerate $U-$process given by the kernel $h^q K_{h,jl}\tau(\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_j, \widetilde{\boldsymbol{U}}_l)$, indexed by $h \in \mathcal{H}_{sc,n}$, with envelope $\|K\|_\infty \tau(\cdot,\cdot,\cdot)$. We take again $p = 1$ and $\beta \in (0,1)$ arbitrarily close to 1 to stand for Sherman's quantity α. Hence, using Jensen's inequality, we could bound the right-hand side of the Maximal Inequality of Sherman [72] by a universal constant times

$$E^{\beta/2}\left[K^2\left(\frac{\boldsymbol{Z}_j - \boldsymbol{Z}_l}{c_{max}n^{-\alpha}}\right)\tau^2(\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_j, \widetilde{\boldsymbol{U}}_l)\right].$$

By standard changes of variables and suitable integrability conditions, this integral is bounded by a constant times $n^{-\alpha\beta q/2}$. Consequently, the uniform rate of the $U-$processes obtained conditioning by $\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_j, \widetilde{\boldsymbol{U}}_l$ and $\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_j, \widetilde{\boldsymbol{U}}_k$, respectively is $n^{-3/2} \times O_{\mathbb{P}}(n^{-\alpha\{2-q+\beta q/2\}})$. Since $1/2 + \alpha\{2 - q + \beta q/2\} > 0$ under our assumptions $q < 4$ and $\alpha \in (1/4, 1/q)$ we get that $n^{-3/2} \times O_{\mathbb{P}}(n^{-\alpha\{2-q+\beta q/2\}}) = o_{\mathbb{P}}(n^{-1})$.

In addition, we get by conditioning on $\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{U}}_l$ and $\widetilde{\boldsymbol{U}}_j, \widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{U}}_l$, the other one being similar, that

$$\begin{aligned}
&E[(m_i - m_k)K_{h,ik}(m_j - m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij} \mid \widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{U}}_l] \\
&\qquad = (m_i - m_k)K_{h,ik}E[(m_j - m_l)K_{h,jl}\boldsymbol{\Omega}_{n,ij} \mid \widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_l] \\
&\qquad = (m_i - m_k)K_{h,ik}\left(m_l f_z(\boldsymbol{Z}_l)\boldsymbol{\Omega}^Z_{n,il}E\left[\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{X}_i\right] - m_l f_z(\boldsymbol{Z}_l)\boldsymbol{\Omega}^Z_{n,il}E\left[\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{X}_i\right] + O_{\mathbb{P}}(h^2)\right) \\
&\qquad = h^{2-q}\tau(\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{U}}_l)h^q K_{h,ik}(1 + O_{\mathbb{P}}(h^2)).
\end{aligned}$$

Now, we apply again the Maximal Inequality of Sherman [72], page 448, for the degenerate $U-$process given by the kernel $h^q K_{h,ik}\tau(\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{U}}_l)$, indexed by $h \in \mathcal{H}_{sc,n}$, with envelope $\|K\|_\infty \tau(\cdot,\cdot,\cdot)$. We take again $p = 1$ and $\beta \in (0,1)$ arbitrarily close to 1 to stand for Sherman's quantity α. Hence, using Jensen's inequality, we could bound the right-hand side of the Maximal Inequality of Sherman [72] by an universal constant times

$$E^{\beta/2}\left[K^2\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_k}{c_{max}n^{-\alpha}}\right)\tau^2(\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{U}}_l)\right].$$

By standard changes of variables and suitable integrability conditions, this integral is bounded by a constant times $n^{-\alpha\beta q/2}$. Consequently, the uniform rate of the $U-$processes obtained conditioning by $\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{U}}_l$ and $\widetilde{\boldsymbol{U}}_j, \widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{U}}_l$, respectively is $n^{-3/2} \times O_{\mathbb{P}}(n^{-\alpha\{2-q+\beta q/2\}})$. Since $1/2 + \alpha\{2 - q + \beta q/2\} > 0$ under our assumptions $q < 4$ and $\alpha \in (1/4, 1/q)$ we get that $n^{-3/2} \times O_{\mathbb{P}}(n^{\alpha q\{1-\beta/2\}}) = o_{\mathbb{P}}(n^{-1})$.

Finally, we consider the remaining $U-$process of order four. This process is given by

$$(m(\boldsymbol{Z}_i) - m(\boldsymbol{Z}_k))K_{h,ik}(m(\boldsymbol{Z}_j) - m(\boldsymbol{Z}_l))K_{h,jl}\boldsymbol{\Omega}_{n,ij} = h^{-2q}\tau(\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_j, \widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{U}}_l)h^q K_{h,ik}h^q K_{h,jl}.$$

Now, we apply again the Maximal Inequality of Sherman [72], page 448, for the degenerate $U-$process given by the kernel $\tau(\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_j, \widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{U}}_l)h^q K_{h,ik}h^q K_{h,jl}$, indexed by $h \in \mathcal{H}_{sc,n}$, with envelope $\|K\|^2_\infty \tau(\cdot,\cdot,\cdot,\cdot)$. We take again $p = 1$ and $\beta \in (0,1)$ arbitrarily close to 1 to stand for Sherman's quantity α. Hence, using Jensen's inequality, we could bound the right-hand side of the Maximal Inequality of Sherman [72] by an universal constant times

$$E^{\beta/2}\left[K^2\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_k}{c_{max}n^{-\alpha}}\right)K^2\left(\frac{\boldsymbol{Z}_j - \boldsymbol{Z}_l}{c_{max}n^{-\alpha}}\right)\tau^2(\widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_j, \widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{U}}_l)\right].$$

By standard changes of variables and suitable integrability conditions, this integral is bounded by a constant times $n^{-\alpha\beta q}$. Consequently, the uniform rate of the fourth order $U-$process is $n^{-2} \times O_{\mathbb{P}}(n^{\alpha q\{2-\beta\}})$. Since $1 > \alpha q$ under our assumptions $q < 4$ and $\alpha \in (1/4, 1/q)$ we get that $n^{-2} \times O_{\mathbb{P}}(n^{\alpha q\{2-\beta\}}) = o_{\mathbb{P}}(n^{-1})$.

From all the results it follows that

$$\sup_{h \in \mathcal{H}_{sc,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} |\widetilde{A}_{1,n}| = \sup_{h \in \mathcal{H}_{sc,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} |\widetilde{A}_{1,n}(h)| = o_{\mathbb{P}}(n^{-1}).$$

In the next step we consider $\widetilde{A}_{2,n}$. We get that

$$\begin{aligned}
&E[(m_i - m_k)K_{h,ik}(m_j - m_k)K_{h,jk}\boldsymbol{\Omega}_{n,ij}] = E[(m_i - m_k)K_{h,ik}E[(m_j - m_k)K_{h,jk}\boldsymbol{\Omega}_{n,ij} \mid \widetilde{\boldsymbol{U}}_i, \widetilde{\boldsymbol{U}}_k]] \\
&= E\left[(m_i - m_k)K_{h,ik}\left(f_z(\boldsymbol{Z}_k)m_k\boldsymbol{\Omega}^Z_{n,ik}E\left[\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{X}_i\right] - f_z(\boldsymbol{Z}_k)m_k\boldsymbol{\Omega}^Z_{n,ik}E\left[\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{X}_i\right] + O_{\mathbb{P}}(h^2)\right)\right] \\
&= E[(m_i - m_k)K_{h,ik}]O(h^2) = O(h^4).
\end{aligned}$$

This implies that $E[\widetilde{A}_{2,n}] = o_{\mathbb{P}}(n^{-1})$ uniformly with respect to $h \in \mathcal{H}_{sc,n}$.

We consider now the first order U–process of the Hoeffding decomposition for $\widetilde{A}_{2,n}$. As we symmetrized the kernel we need to consider the conditional expectations with respect to all four variables. By the same reasoning as for $E[\widetilde{A}_{2,n}]$ we get that

$$\begin{aligned}
&E[(m_i - m_k)K_{h,ik}(m_j - m_k)K_{h,jk}\boldsymbol{\Omega}_{n,ij} \mid \widetilde{\boldsymbol{U}}_i] \\
&= E[(m_i - m_k)K_{h,ik}E[(m_j - m_k)K_{h,jk}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{Z}_k, \widetilde{\boldsymbol{U}}_i] \mid \widetilde{\boldsymbol{U}}_i] \\
&= E\left[(m_i - m_k)K_{h,ik}\left(m_k f_z(\boldsymbol{Z}_k)\boldsymbol{\Omega}^Z_{n,ik}E\left[\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{X}_i\right] - m_k f_z(\boldsymbol{Z}_k)\boldsymbol{\Omega}^Z_{n,ik}E\left[\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{X}_i\right] + O_{\mathbb{P}}(h^2)\right) \mid \widetilde{\boldsymbol{U}}_i\right] \\
&= \left(m_i f_z(\boldsymbol{Z}_i) - m_i f_z(\boldsymbol{Z}_i) + O_{\mathbb{P}}(h^2)\right)O_{\mathbb{P}}(h^2) = O_{\mathbb{P}}(h^4).
\end{aligned}$$

Note that the reasoning when conditioning on $\widetilde{\boldsymbol{U}}_j$ and $\widetilde{\boldsymbol{U}}_k$ is the same. Therefore, we get that the first order U–processes of the Hoeffding decompositions for $\widetilde{A}_{2,n}$ are of order $o_{\mathbb{P}}(n^{-1})$.

It is easy to see that the second and third order U–processes of the Hoeffding decomposition for $\widetilde{A}_{2,n}$ are of order $o_{\mathbb{P}}(n^{-1})$ if we apply the Maximal Inequality of Sherman [72]. From all the results it follows that

$$\sup_{h \in \mathcal{H}_{sc,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} |\widetilde{A}_{2,n}| = \sup_{h \in \mathcal{H}_{sc,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} |\widetilde{A}_{2,n}(h)| = o_{\mathbb{P}}(n^{-1}).$$

As it follows by the same reasoning as for $\widetilde{A}_{2,n}$ that

$$\sup_{h \in \mathcal{H}_{sc,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} |\widetilde{A}_{3,n}| = \sup_{h \in \mathcal{H}_{sc,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} |\widetilde{A}_{3,n}(h)| = o_{\mathbb{P}}(n^{-1}),$$

we omit the details here. Finally, we get by standard change of variables that

$$E[\widetilde{A}_{4,n}] = n^{-2}E[(m_i - m_j)^2K^2_{h,ij}\boldsymbol{\Omega}_{n,ij}] = O(n^{-2}n^{\alpha q}) = o(n^{-1}),$$

as well as

$$n^{-2}E[(m_i - m_j)^2K^2_{h,ij}\boldsymbol{\Omega}_{n,ij} \mid \widetilde{\boldsymbol{U}}_i] = O_{\mathbb{P}}(n^{-2}n^{\alpha q}),$$

and

$$n^{-2}E[(m_i - m_j)^2K^2_{h,ij}\boldsymbol{\Omega}_{n,ij} \mid \widetilde{\boldsymbol{U}}_j] = O_{\mathbb{P}}(n^{-2}n^{\alpha q}).$$

Using the Hoeffding decomposition and applying Corollary 4 of Sherman [72], we deduce that

$$n^2h^{2q}(\widetilde{A}_{4,n} - E[\widetilde{A}_{4,n}]) = O_{\mathbb{P}}(n^{-1}) + O_{\mathbb{P}}(n^{-1/2}n^{-\alpha q}),$$

uniformly with respect to h. Deduce that

$$\sup_{h \in \mathcal{H}_{sc,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} |\widetilde{A}_{4,n}| = O_{\mathbb{P}}(n^{-3}n^{2\alpha q}) + O_{\mathbb{P}}(n^{-5/2}n^{\alpha q}) + O_{\mathbb{P}}(n^{-2}n^{\alpha q}) = o_{\mathbb{P}}(n^{-1}).$$

With all these results (1.46) and, in particular, (1.45) follow.

We know from Lemma 1.14 that

$$\left\| \boldsymbol{\Omega}_n^{1/2} n^{-2} \left(\boldsymbol{\varepsilon f_z}\right)_n \right\| = O_{\mathbb{P}}(n^{-3/2}),$$

such that from all these results the statement follows.

□

Lemma 1.19. *Assume the conditions of Proposition 1.1 hold true. Then,*

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right\| = O_{\mathbb{P}}(n^{-1/2}).$$

Proof of Lemma 1.19.

By the arguments used for Lemma 1.14, it suffices to consider $\boldsymbol{d} = \operatorname{diag}(d_U,\dots,d_U)$. Moreover, for simpler notation, we omit the argument $\boldsymbol{d}$ in $\boldsymbol{\Omega}_n(\boldsymbol{d})$. We get that

$$\begin{aligned}&\left\|\boldsymbol{\Omega}_n^{1/2}n^{-1}\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right\|^2\\ &= \frac{1}{n^2}\sum_{1\le i\ne j\le n}\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_{n,i}\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_{n,j}\boldsymbol{\Omega}_{n,ij} + \frac{1}{n^2}\sum_{i=1}^{n}\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_{n,i}^2\\ &= A_n + B_n.\end{aligned}$$

It is easy to check that $\sup_{h\in\mathcal{H}_{sc,n}} B_n = o_{\mathbb{P}}(n^{-1})$. Furthermore, we get that

$$A_n = \frac{1}{n^2}\sum_{1\le i\ne j\le n}\left[\frac{1}{n}\sum_{1\le k\le n,k\ne i}\varepsilon_k K_{h,ik}\right]\left[\frac{1}{n}\sum_{1\le l\le n,l\ne j}\varepsilon_l K_{h,jl}\right]\boldsymbol{\Omega}_{n,ij}.$$

We show in the following that

$$\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}|A_n| = \sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}|A_n(h)| = O_{\mathbb{P}}(n^{-1}). \tag{1.47}$$

For this purpose, we define $(n)_k = n(n-1)\dots(n-k+1)$ and decompose $A_n(h)$ into a sum of four $U-$processes, i.e.

$$A_n(h) = \frac{(n-1)_3}{n^3}A_{1,n}(h) + \frac{(n-1)_2}{n^2}A_{2,n}(h) + 2\frac{(n-1)_2}{n^2}A_{3,n}(h) + \frac{n-1}{n}A_{4,n}(h),$$

where

$$\begin{aligned}A_{1,n} &= A_{1,n}(h) = \frac{1}{(n)_4}\sum_{1\le i\ne j\ne k\ne l\le n}\varepsilon_k K_{h,ik}\varepsilon_l K_{h,jl}\boldsymbol{\Omega}_{n,ij}\\ A_{2,n} &= A_{2,n}(h) = \frac{1}{n(n)_3}\sum_{1\le i\ne j\ne k\le n}\varepsilon_k^2 K_{h,ik}K_{h,jk}\boldsymbol{\Omega}_{n,ij}\\ A_{3,n} &= A_{3,n}(h) = \frac{1}{n(n)_3}\sum_{1\le i\ne j\ne l\le n}\varepsilon_j K_{h,ij}\varepsilon_l K_{h,jl}\boldsymbol{\Omega}_{n,ij}\\ \text{and}\quad A_{4,n} &= A_{4,n}(h) = \frac{1}{n^2(n)_2}\sum_{1\le i\ne j\le n}\varepsilon_i\varepsilon_j K_{h,ij}^2\boldsymbol{\Omega}_{n,ij}.\end{aligned}$$

For each of these $U-$processes we compute the mean and use the Hoeffding decomposition. The kernels of $A_{1,n}$, $A_{2,n}$ and $A_{3,n}$ are not symmetric in their arguments. However, we could apply the usual symmetrization idea. Thus, by abuse, we will proceed as if the kernels of the $U-$statistics we handle are symmetric. Here, $\boldsymbol{U}_i = \left(Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T\right)^T$.

In addition, we have that the kernels of $A_{1,n}$, $A_{2,n}$, $A_{3,n}$ and $A_{4,n}$ are Euclidean for a squared integrable envelope. See Lemma 22 in Nolan and Pollard [67] and Lemma 2.14 in Pakes and Pollard [68]. Therefore, we can in the following repeatedly apply Corollary 7 and the Maximal Inequality of Sherman [72]. All remainder terms are controlled by Assumption 1.3.2.

We start by considering $A_{1,n}$. Recall that by assumption $E\left[\varepsilon_k \mid \boldsymbol{X}_k, \boldsymbol{Z}_k\right] = E\left[\varepsilon_l \mid \boldsymbol{X}_l, \boldsymbol{Z}_l\right] = 0$. Therefore, we get that $E\left[A_{1,n}\right] = 0$ as well as

$$E\left[\varepsilon_k K_{h,ik}\varepsilon_l K_{h,jl}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_p, p\in\{i,j,k,l\}\right] = 0.$$

Note that we need to consider the conditional expectations with respect to all four variables for the first order $U-$process of the Hoeffding decomposition of $A_{1,n}$ as we symmetrized the kernel. It follows from the results that the first order $U-$process of the Hoeffding decomposition of $A_{1,n}$ is 0.

We consider now the six second order $U-$processes of the Hoeffding decomposition of $A_{1,n}$. There are two types of such processes. First, the ones that are 0. This is the case when conditioning by the pairs $(\boldsymbol{U}_i, \boldsymbol{U}_l)$, $(\boldsymbol{U}_i, \boldsymbol{U}_k)$, $(\boldsymbol{U}_j, \boldsymbol{U}_l)$, $(\boldsymbol{U}_j, \boldsymbol{U}_k)$ and $(\boldsymbol{U}_i, \boldsymbol{U}_j)$. The second case occurs when conditioning on $(\boldsymbol{U}_k, \boldsymbol{U}_l)$. We get that

$$\begin{aligned}
&E[\varepsilon_k K_{h,ik}\varepsilon_l K_{h,jl}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_k, \boldsymbol{U}_l] \\
&\qquad = \varepsilon_k\varepsilon_l E[K_{h,ik}K_{h,jl}\boldsymbol{\Omega}^X_{n,ij}\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{Z}_k, \boldsymbol{Z}_l] \\
&\qquad = \varepsilon_k\varepsilon_l E[E[K_{h,ik}\boldsymbol{\Omega}^X_{n,ij}\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{Z}_k, \boldsymbol{Z}_j]K_{h,jl} \mid \boldsymbol{Z}_k, \boldsymbol{Z}_l] \\
&\qquad = \varepsilon_k\varepsilon_l E[\left(f_z(\boldsymbol{Z}_k)E[\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_j]\boldsymbol{\Omega}^Z_{n,kj} + O_{\mathbb{P}}(h^2)\right) K_{h,jl} \mid \boldsymbol{Z}_k, \boldsymbol{Z}_l] \\
&\qquad = \varepsilon_k\varepsilon_l f_z(\boldsymbol{Z}_k)E[f_z(\boldsymbol{Z}_l)E[\boldsymbol{\Omega}^X_{n,ij}]\boldsymbol{\Omega}^Z_{n,kl} + O_{\mathbb{P}}(h^2) \mid \boldsymbol{Z}_k, \boldsymbol{Z}_l] \\
&\qquad\quad + \varepsilon_k\varepsilon_l O_{\mathbb{P}}(h^2)\left(f_z(\boldsymbol{Z}_l) + O_{\mathbb{P}}(h^2)\right) \\
&\qquad = \varepsilon_k\varepsilon_l f_z(\boldsymbol{Z}_k)f_z(\boldsymbol{Z}_l)E[\boldsymbol{\Omega}^X_{n,ij}]\boldsymbol{\Omega}^Z_{n,kl} + \varepsilon_k\varepsilon_l f_z(\boldsymbol{Z}_k)O_{\mathbb{P}}(h^2) \\
&\qquad\quad + \varepsilon_k\varepsilon_l f_z(\boldsymbol{Z}_l)O_{\mathbb{P}}(h^2) + \varepsilon_k\varepsilon_l O_{\mathbb{P}}(h^4).
\end{aligned}$$

Therefore, it follows together with Corollary 4 of Sherman [72] that the second order $U-$processes of the Hoeffding decomposition of $A_{1,n}$ are of order $O_{\mathbb{P}}(n^{-1})$ uniformly with respect to h and $\boldsymbol{d}$.

We consider now the four $U-$processes of order three obtained by conditioning on any subset of three of the four vectors $\boldsymbol{U}_i$, $\boldsymbol{U}_k$, $\boldsymbol{U}_j$ and $\boldsymbol{U}_l$. There are two types of such processes. First, the ones that are 0. This is the case when conditioning by $(\boldsymbol{U}_i, \boldsymbol{U}_j, \boldsymbol{U}_k)$ or $(\boldsymbol{U}_i, \boldsymbol{U}_j, \boldsymbol{U}_l)$. The second case occurs when conditioning on $(\boldsymbol{U}_i, \boldsymbol{U}_k, \boldsymbol{U}_l)$ or $(\boldsymbol{U}_j, \boldsymbol{U}_k, \boldsymbol{U}_l)$, the other one being similar. We get that

$$\begin{aligned}
&E[\varepsilon_k K_{h,ik}\varepsilon_l K_{h,jl}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_i, \boldsymbol{U}_k, \boldsymbol{U}_l] \\
&\qquad = \varepsilon_k\varepsilon_l K_{h,ik}E[K_{h,jl}\boldsymbol{\Omega}^X_{n,ij}\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{U}_i, \boldsymbol{Z}_l] \\
&\qquad = \varepsilon_k\varepsilon_l K_{h,ik}f_z(\boldsymbol{Z}_l)\boldsymbol{\Omega}^Z_{n,il}E[\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{X}_i] \\
&\qquad\quad + \varepsilon_k\varepsilon_l K_{h,ik}O_{\mathbb{P}}(h^2) \\
&\qquad = h^{-q}h^q K_{h,ik}\tau_1(\boldsymbol{U}_i, \boldsymbol{U}_k, \boldsymbol{U}_l) \\
&\qquad\quad + \tau_2(\boldsymbol{U}_k, \boldsymbol{U}_l)h^{-q}h^q K_{h,ik}O_{\mathbb{P}}(h^2).
\end{aligned}$$

Now, we apply the Maximal Inequality of Sherman [72], page 448, for the degenerate $U-$process given by the kernel $h^q K_{h,ik}\tau_1(\boldsymbol{U}_i, \boldsymbol{U}_k, \boldsymbol{U}_l)$, indexed by $h \in \mathcal{H}_{sc,n}$, with envelope $\|K\|_\infty \tau_1(\cdot,\cdot,\cdot)$. The reasoning for $h^q K_{h,ik}\tau_2(\boldsymbol{U}_k, \boldsymbol{U}_l)$ is the same. (Herein, $\|\cdot\|_\infty$ denotes the uniform norm.) We take $p = 1$ and $\beta \in (0,1)$ arbitrarily close to 1 to stand for Sherman's quantity α. Since $K(\cdot)$ is of bounded variation and symmetric, without loss of generality we could consider that $K(\cdot)$ is nonincreasing on $[0,\infty)$. In this case, $0 \leq K(\cdot/h) \leq K(\cdot/\overline{h})$ with $\overline{h} = \sup \mathcal{H}_{sc,n} =: c_{max}n^{-\alpha}$. Hence, using Jensen's inequality, we could bound the right-hand side of the Maximal Inequality of Sherman [72] by a universal constant times

$$E^{\beta/2}\left[K^2\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_k}{c_{max}n^{-\alpha}}\right)\tau_1^2(\boldsymbol{U}_i, \boldsymbol{U}_k, \boldsymbol{U}_l)\right].$$

By standard changes of variables and suitable integrability conditions, this integral is bounded by a constant times $n^{-\alpha\beta q/2}$. Consequently, the uniform rate of the second $U-$processes obtained conditioning by $\boldsymbol{U}_i, \boldsymbol{U}_k, \boldsymbol{U}_l$ and $\boldsymbol{U}_j, \boldsymbol{U}_k, \boldsymbol{U}_l$, respectively is $n^{-3/2} \times O_{\mathbb{P}}(n^{\alpha q\{1-\beta/2\}})$. As $1/2 - \alpha q(1-\beta/2) > 0$ under our assumptions we get that $n^{-3/2} \times O_{\mathbb{P}}(n^{\alpha q\{1-\beta/2\}}) = o_{\mathbb{P}}(n^{-1})$ such that the third order $U-$processes of the Hoeffding decomposition of $A_{1,n}$ are of order $o_{\mathbb{P}}(n^{-1})$.

Finally, we consider the remaining $U-$process of order four. This process is given by

$$\varepsilon_k K_{h,ik}\varepsilon_l K_{h,jl}\boldsymbol{\Omega}_{n,ij} = h^{-2q}\tau_3(\boldsymbol{U}_i, \boldsymbol{U}_j, \boldsymbol{U}_k, \boldsymbol{U}_l)h^{2q}K_{h,ik}K_{h,jl}.$$

Now, we apply again the Maximal Inequality of Sherman [72], page 448, for the degenerate $U-$process given by the kernel $\tau_3(\boldsymbol{U}_i, \boldsymbol{U}_j, \boldsymbol{U}_k, \boldsymbol{U}_l)h^q K_{h,ik}h^q K_{h,jl}$, indexed by $h \in \mathcal{H}_{sc,n}$, with envelope $\|K\|_\infty^2 \tau(\cdot,\cdot)$. We take again $p = 1$ and $\beta \in (0,1)$ arbitrarily close to 1 to stand for Sherman's quantity α. Hence, using Jensen's inequality, we could bound the right-hand side of the Maximal Inequality of Sherman [72] by an universal constant times

$$E^{\beta/2}\left[K^2\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_k}{c_{max}n^{-\alpha}}\right)K^2\left(\frac{\boldsymbol{Z}_j - \boldsymbol{Z}_l}{c_{max}n^{-\alpha}}\right)\tau_3^2(\boldsymbol{U}_i, \boldsymbol{U}_j, \boldsymbol{U}_k, \boldsymbol{U}_l)\right].$$

By standard changes of variables and suitable integrability conditions, this integral is bounded by a constant times $n^{-\alpha\beta q}$. Consequently, the uniform rate of the fourth order $U-$process is $n^{-2} \times$

$O_{\mathbb{P}}(n^{\alpha q\{2-\beta\}})$. Since $1 > \alpha q$ under our assumptions we get that $n^{-2} \times O_{\mathbb{P}}(n^{\alpha q\{2-\beta\}}) = o_{\mathbb{P}}(n^{-1})$.

From all the results it follows that

$$\sup_{h\in\mathcal{H}_{sc,n}} \sup_{\boldsymbol{d}\in\mathcal{D}} |A_{1,n}| = \sup_{h\in\mathcal{H}_{sc,n}} \sup_{\boldsymbol{d}\in\mathcal{D}} |A_{1,n}(h)| = O_{\mathbb{P}}(n^{-1}).$$

In the next step we consider $A_{2,n}$. We get that

$$\begin{aligned}
nE\left[A_{2,n}\right] &= E\left[\varepsilon_k^2 K_{h,ik}K_{h,jk}\boldsymbol{\Omega}_{n,ij}\right] \\
&= E\left[E\left[\varepsilon_k^2 \mid \boldsymbol{X}_k, \boldsymbol{Z}_k\right] K_{h,ik}K_{h,jk}\boldsymbol{\Omega}_{n,ij}\right] \\
&= E\left[\sigma^2\left(\boldsymbol{X}_k, \boldsymbol{Z}_k\right) E\left[K_{h,ik}\boldsymbol{\Omega}^X_{n,ij}\boldsymbol{\Omega}^Z_{n,ij} \mid \boldsymbol{Z}_j, \boldsymbol{Z}_k\right] K_{h,jk}\right] \\
&= E\left[\sigma^2\left(\boldsymbol{X}_k, \boldsymbol{Z}_k\right) f_z(\boldsymbol{Z}_k)\boldsymbol{\Omega}^Z_{n,kj} E\left[\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_j\right] K_{h,jk}\right] \\
&\quad + E\left[\sigma^2\left(\boldsymbol{X}_k, \boldsymbol{Z}_k\right) K_{h,jk}\right] O(h^2) \\
&= E\left[\sigma^2\left(\boldsymbol{X}_k, \boldsymbol{Z}_k\right) f_z(\boldsymbol{Z}_k) E\left[\boldsymbol{\Omega}^Z_{n,kj} E\left[\boldsymbol{\Omega}^X_{n,ij} \mid \boldsymbol{Z}_j\right] K_{h,jk} \mid \boldsymbol{Z}_k\right]\right] \\
&\quad + E\left[\sigma^2\left(\boldsymbol{X}_k, \boldsymbol{Z}_k\right) E\left[K_{h,jk} \mid \boldsymbol{Z}_k\right]\right] O(h^2) \\
&= E\left[\sigma^2\left(\boldsymbol{X}_k, \boldsymbol{Z}_k\right) f_z(\boldsymbol{Z}_k)^2 E\left[\boldsymbol{\Omega}^X_{n,ij}\right]\right] \\
&\quad + 2E\left[\sigma^2\left(\boldsymbol{X}_k, \boldsymbol{Z}_k\right) f_z(\boldsymbol{Z}_k)\right] O(h^2) \\
&\quad + E\left[\sigma^2\left(\boldsymbol{X}_k, \boldsymbol{Z}_k\right)\right] O(h^4).
\end{aligned}$$

Therefore, $E\left[A_{2,n}\right] = O(n^{-1})$.

In addition, we get by a similar reasoning that the first, second and third order U–processes of the Hoeffding decomposition of $A_{2,n}$ are of order $o_{\mathbb{P}}(n^{-1})$ if we apply the Maximal Inequality of Sherman [72]. The details are omitted. From all the results it follows that

$$\sup_{h\in\mathcal{H}_{sc,n}} \sup_{\boldsymbol{d}\in\mathcal{D}} |A_{2,n}| = \sup_{h\in\mathcal{H}_{sc,n}} \sup_{\boldsymbol{d}\in\mathcal{D}} |A_{2,n}(h)| = O_{\mathbb{P}}(n^{-1}).$$

In the next step we consider $A_{3,n}$. We get that $E\left[A_{3,n}\right] = 0$ as well as

$$E\left[\varepsilon_j K_{h,ij}\varepsilon_l K_{h,jl}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_p, p \in \{i,j,l\}\right] = 0.$$

In addition, it is easy to see that the second and third order U–processes of the Hoeffding decomposition of $A_{3,n}$ are of order $o_{\mathbb{P}}(n^{-1})$ if we apply the Maximal Inequality of Sherman [72]. From all the results it follows that

$$\sup_{h\in\mathcal{H}_{sc,n}} \sup_{\boldsymbol{d}\in\mathcal{D}} |A_{3,n}| = \sup_{h\in\mathcal{H}_{sc,n}} \sup_{\boldsymbol{d}\in\mathcal{D}} |A_{3,n}(h)| = o_{\mathbb{P}}(n^{-1}).$$

Finally, we get that $E\left[A_{4,n}\right] = 0$ as well as

$$E\left[\varepsilon_i\varepsilon_j K^2_{h,ij}\boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{U}_p, p \in \{i,j\}\right] = 0.$$

In addition, it is easy to see that the second order U–process of the Hoeffding decomposition of $A_{4,n}$ is of order $o_{\mathbb{P}}(n^{-1})$ if we apply the Maximal Inequality of Sherman [72]. Deduce that

$$\sup_{h\in\mathcal{H}_{sc,n}} \sup_{\boldsymbol{d}\in\mathcal{D}} |A_{4,n}| = o_{\mathbb{P}}(n^{-1}).$$

From all the results (1.47) follows and, therefore, the statement.

□

Table 1.C.9: *Bias and Standard Deviation of the estimators for λ and β in Model 2.*

	s	Bias			St. dev.		
n		250	500	1000	250	500	1000
λ estimator							
SmoothMD with γ	G_n	0.006	0.004	0.002	0.094	0.066	0.047
SmoothMD without γ	G_n	0.006	0.004	0.002	0.09	0.063	0.045
NL2SLS	G_n	−0.0003	0.041	−0.001	0.059	0.041	0.028
β estimator							
SmoothMD with γ	G_n	0.025	0.014	0.007	0.182	0.125	0.089
SmoothMD without γ	G_n	0.024	0.013	0.007	0.173	0.119	0.084
NL2SLS	G_n	0.005	0.003	0.0003	0.109	0.074	0.051

Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The grid for λ is $[\lambda_0 - 0.8, \lambda_0 + 0.8]$. For all simulations 2000 Monte Carlo samples were used.

Table 1.C.10: *Bias and Standard Deviation of the estimators for λ and β in Model 3.*

	s	Bias			St. dev.		
n		250	500	1000	250	500	1000
λ estimator							
SmoothMD with γ	G_n	−0.002	0.002	−0.003	0.146	0.102	0.073
SmoothMD without γ	G_n	−0.003	0.002	−0.003	0.145	0.101	0.072
NL2SLS	G_n	−0.003	0.002	−0.002	0.123	0.086	0.06
β estimator							
SmoothMD with γ	G_n	0.016	0.004	0.007	0.171	0.118	0.084
SmoothMD without γ	G_n	0.017	0.005	0.007	0.17	0.117	0.084
NL2SLS	G_n	0.013	0.002	0.004	0.145	0.099	0.07

Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The grid for λ is $[\lambda_0 - 0.8, \lambda_0 + 0.8]$. For all simulations 2000 Monte Carlo samples were used.

Table 1.C.11: *Empirical Level for Z-Tests of the estimators for λ and β in Model 1.*

	s	5% level			10% level		
n		250	500	1000	250	500	1000
Test for λ							
SmoothMD with γ	G_n	5.75	6.0	4.55	10.2	11.15	10.75
SmoothMD* with γ	G_n	5.95	6.35	4.65	10.75	11.35	11.1
SmoothMD without γ	G_n	6.45	6.6	5.4	11.1	12.1	10.9
SmoothMD* without γ	G_n	5.1	5.05	4.45	10.05	10.0	9.15
NL2SLS	G_n	9.25	7.75	5.95	15.1	13.9	11.55
Test for β							
SmoothMD with γ	G_n	5.85	4.5	3.85	10.25	9.15	7.9
SmoothMD* with γ	G_n	6.3	5.15	4.3	11.55	10.65	8.55
SmoothMD without γ	G_n	6.2	4.7	3.95	10.65	9.75	7.9
SmoothMD* without γ	G_n	9.2	8.05	7.1	14.85	14.4	12.3
NL2SLS	G_n	7.65	6.1	4.45	12.85	12.15	8.6

Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

Table 1.C.12: *Empirical Level for Z-Tests of the estimators for λ and β in Model 2.*

	s	5% level			10% level		
n		250	500	1000	250	500	1000
Test for λ							
SmoothMD with γ	G_n	7.3	5.85	5.5	12.7	10.5	10.25
SmoothMD* with γ	G_n	6.7	5.85	5.3	12.5	10.15	10.1
SmoothMD without γ	G_n	6.7	5.65	5.1	12.1	10.25	9.95
SmoothMD* without γ	G_n	1.1	0.7	0.55	3.3	2.4	1.85
NL2SLS	G_n	6.2	5.45	4.8	12.3	10.25	10
Test for β							
SmoothMD with γ	G_n	7.15	5.25	5.55	11.75	9.95	10.35
SmoothMD* with γ	G_n	6.95	5.15	5.5	11.45	9.75	10.2
SmoothMD without γ	G_n	6.3	5.25	4.9	11.25	9.9	10.1
SmoothMD* without γ	G_n	1.0	2.45	0.75	3.4	2.45	1.95
NL2SLS	G_n	6.55	4.95	5.1	12.25	10.95	9.75

Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

Table 1.C.13: *Empirical Level for Z-Tests of the estimators for λ and β in Model 3.*

	s	5% level			10% level		
n		250	500	1000	250	500	1000
Test for λ							
SmoothMD with γ	G_n	6.45	5.55	4.85	11.2	9.65	9.5
SmoothMD* with γ	G_n	6.95	6.2	5.55	12.05	10.5	10.55
SmoothMD without γ	G_n	6.4	5.55	5.0	11.5	10.15	9.7
SmoothMD* without γ	G_n	5.3	4.8	3.95	9.65	8.65	8.9
NL2SLS	G_n	5.75	5.6	5.15	11.8	10.75	10.05
Test for β							
SmoothMD with γ	G_n	5.9	5.15	4.95	11.1	9.8	9.25
SmoothMD* with γ	G_n	6.4	5.5	5.5	12.15	10.95	10.3
SmoothMD without γ	G_n	6.05	5.35	4.9	10.9	9.8	9.1
SmoothMD* without γ	G_n	5.1	4.6	4.15	9.7	8.7	8.35
NL2SLS	G_n	6.5	5.75	5.45	11.4	10.95	9.15

Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

Table 1.C.14: *Empirical Level for distance metric statistics of the estimators for λ and β in Model 1.*

	s	5% level			10% level		
n		250	500	1000	250	500	1000
Test for λ							
SmoothMD with γ	G_n	7.0	6.8	5.35	11.9	12.7	11.2
SmoothMD* with γ	G_n	6.95	7.45	5.35	12.1	13.25	11.65
SmoothMD without γ	G_n	6.7	7.1	5.4	12.05	12.8	11.25
SmoothMD* without γ	G_n	5.45	5.35	4.6	10.65	10.1	9.3
Test for β							
SmoothMD with γ	G_n	5.9	4.5	3.9	10.55	9.45	7.9
SmoothMD* with γ	G_n	6.5	5.0	4.35	11.75	10.65	8.5
SmoothMD without γ	G_n	6.35	4.75	4.0	10.8	9.75	7.85
SmoothMD* without γ	G_n	9.3	8.15	7.15	14.95	14.5	12.15

Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

Table 1.C.15: *Empirical Level for distance metric statistics of the estimators for λ and β in Model 3.*

	s	5% level			10% level		
n		250	500	1000	250	500	1000
Test for λ							
SmoothMD with γ	G_n	12.15	9.55	7.05	14.1	12.6	10.65
SmoothMD* with γ	G_n	12.85	10.1	7.85	14.65	12.65	11.0
SmoothMD without γ	G_n	12.3	9.55	6.95	14.4	12.7	10.3
SmoothMD* without γ	G_n	11.75	8.55	6.35	14.25	11.6	9.55
Test for β							
SmoothMD with γ	G_n	12.15	9.25	6.55	14.4	12.05	10.3
SmoothMD* with γ	G_n	12.6	10.05	7.2	14.6	12.4	10.95
SmoothMD without γ	G_n	11.85	8.95	6.7	14.15	11.75	10.7
SmoothMD* without γ	G_n	11.45	8.25	6.05	13.7	11.2	9.65

Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

Figure 1.C.9: *Power function of the distance metric statistic for λ of Model 1 with $n = 500$.*

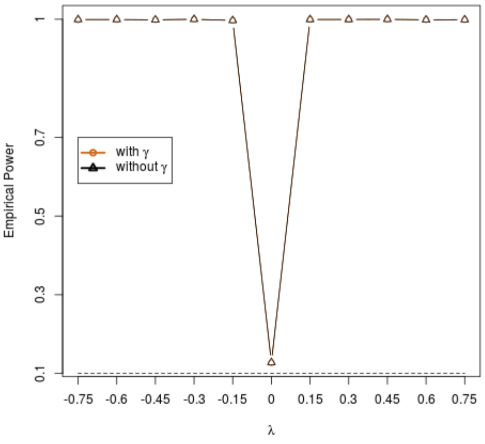
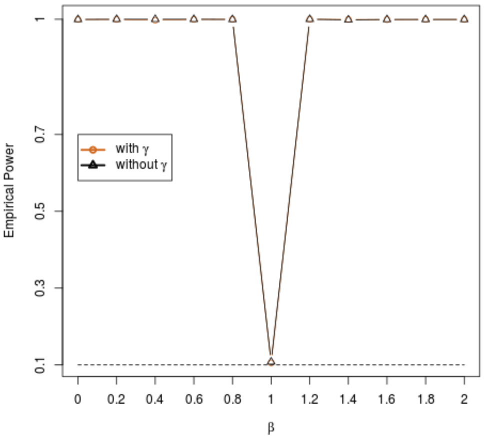


Figure 1.C.10: *Power function of the distance metric statistic for β of Model 1 with $n = 250$.*



Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. Only the SmoothMD estimators that take the additional variance part due to the estimation of $\boldsymbol{\eta}$ into account are considered. For all simulations 2000 Monte Carlo samples were used. The nominal level is 10%.

Figure 1.C.11: *Estimated $m(Z)$ for Model 1 with $n = 500$.*

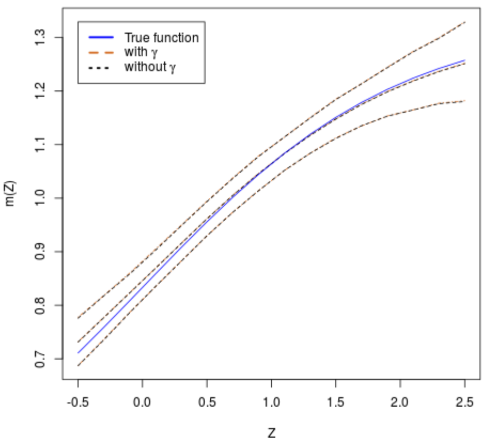
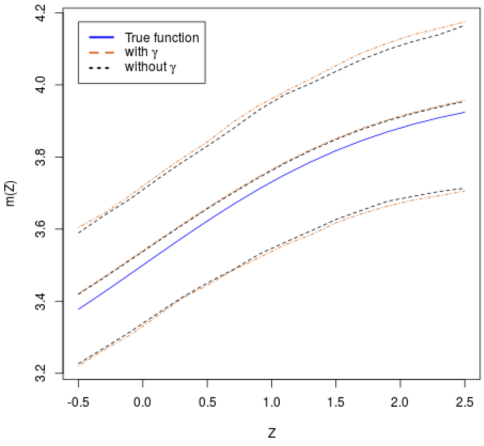


Figure 1.C.12: *Estimated $m(Z)$ for Model 2 with $n = 500$.*



Notes: For the estimation the NW estimator with normal kernel and $h \propto n^{-1/3.5}$ is employed. The 25% and 75% quantiles as well as the mean are reported. For all simulations 2000 Monte Carlo samples were used.

Chapter 2

Hypothesis testing and inference in case of Root-N-consistent semiparametric estimation of partially linear models

2.1. Introduction

We consider a *partially linear* mean regression model given by

$$Y = \boldsymbol{X}^T\boldsymbol{\beta} + m(\boldsymbol{Z}) + \varepsilon, \tag{2.1}$$

where Y is a scalar response variable, $\left(\boldsymbol{X}^T, \boldsymbol{Z}^T\right)^T \in \mathbb{R}^p \times \mathbb{R}^q$ a random covariate vector and $m(\cdot)$ an unknown function. We assume that the error term ε has mean zero conditional on $\boldsymbol{Z}$ and a random vector $\boldsymbol{V} \in \mathbb{R}^s$, i.e.

$$E[\varepsilon|\boldsymbol{V}, \boldsymbol{Z}] = 0. \tag{2.2}$$

The vector $\boldsymbol{V}$ can be considered as a vector of instruments that allows to consider models with endogenous covariate vector $\boldsymbol{X}$. Typically, it is for models like (2.1) assumed that $E[\varepsilon|\boldsymbol{X}, \boldsymbol{Z}] = 0$, see Robinson [70] and Li [57]. However, when $\boldsymbol{X}$ is endogenous this assumption is not met and, thus, an estimator build on it will most likely be inconsistent. If we have an instrument vector $\boldsymbol{V}$ that is correlated with $\boldsymbol{X}$ and fulfills (2.2) consistent estimation is still possible. In case $\boldsymbol{X}$ is not endogenous we can just set $\boldsymbol{V} = \boldsymbol{X}$. For $\boldsymbol{V}$ it is necessary to require that $s \geq p$, see section 2.2.2, as is standard in the instrumental variable literature.

We impose no further assumption on the conditional distribution of ε. In particular, we allow for heteroscedasticity of unknown form. The vector $\boldsymbol{Z}$ contains continuous variables, but the components of $\boldsymbol{X}$ and $\boldsymbol{V}$ need not be continuous. Let $\boldsymbol{\beta}_0$ denote the true value of the parameter.

To estimate the unknown structural parameter $\boldsymbol{\beta}_0$ we employ the smooth minimum distance (SmoothMD) estimator for transformation partially linear models developed in chapter 1. The main difference between the model in chapter 1 and model (2.1) is that in chapter 1 the transformation of Y is unknown, i.e. the model is given by

$$T(Y, \lambda) = \boldsymbol{X}^T\boldsymbol{\beta} + m(\boldsymbol{Z}) + \varepsilon, \tag{2.3}$$

where $T(\cdot, \lambda)$ is the Box-Cox transformation given by

$$T(Y, \lambda) = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & , \lambda \neq 0 \\ \log(Y) & , \lambda = 0. \end{cases}$$

In model (2.3), in addition to $\boldsymbol{\beta}$, the transformation parameter λ is unknown and needs, thus, to be estimated. To identify the model it is in chapter 1 assumed that $E[\varepsilon|\boldsymbol{X}, \boldsymbol{Z}] = 0$ which rules out endogeneity. Here, we allow $\boldsymbol{X}$ to be endogenous.[1]

The SmoothMD estimator in chapter 1 combines estimation techniques from Lavergne and Patilea [56], Li [57] and Robinson [70]. Given condition (2.2) with $\boldsymbol{V} = \boldsymbol{X}$ Robinson [70] proposed to use the

[1] Note that the results stated in chapter 1 apply to model (2.1) as well when $\boldsymbol{X}$ is exogenous.

fact that $E[Y \mid \boldsymbol{Z}] = E[\boldsymbol{X} \mid \boldsymbol{Z}]^T\boldsymbol{\beta} + m(\boldsymbol{Z})$. Therefore, we can rewrite model (2.1) as

$$Y - E[Y \mid \boldsymbol{Z}] = (\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}])^T\boldsymbol{\beta} + \varepsilon. \tag{2.4}$$

The estimator of $\boldsymbol{\beta}_0$ proposed by Robinson [70] is then a feasible version of the unfeasible OLS estimator of $Y - E[Y \mid \boldsymbol{Z}]$ on $\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}]$. The regressand and regressors $E[Y \mid \boldsymbol{Z}]$ and $E[\boldsymbol{X} \mid \boldsymbol{Z}]$ being unknown, they need to be estimated by some nonparametric procedure. Robinson [70] proposed to estimate them by the Nadaraya-Watson estimator. He showed that, under suitable regularity assumptions and conditions on the kernel and the bandwidth, the OLS estimator with response $Y - E[Y \mid \boldsymbol{Z}]$ and covariate vector $\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}]$ yields $\sqrt{n}$-consistent, asymptotically normally and efficient estimators if the conditional expectations given $\boldsymbol{Z}$ are replaced by their kernel estimates. Robinson [70] used a trimming procedure to ensure that the estimated density of $\boldsymbol{Z}$, $f_z(\boldsymbol{Z})$, stays away from zero. To avoid this trimming Li [57] considered the unfeasible OLS regression of $(Y - E[Y \mid \boldsymbol{Z}])f_z(\boldsymbol{Z})$ on $(\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}])f_z(\boldsymbol{Z})$. Premultiplying by the density of $\boldsymbol{Z}$ does not break the consistency of the unfeasible OLS estimator since $E[f_z(\boldsymbol{Z})\varepsilon \mid \boldsymbol{X}, \boldsymbol{Z}] = f_z(\boldsymbol{Z})E[\varepsilon \mid \boldsymbol{X}, \boldsymbol{Z}] = 0$.

However, when $\boldsymbol{X}$ is endogenous the estimators of Li [57] and Robinson [70] will not be consistent as $E[\varepsilon | \boldsymbol{X}, \boldsymbol{Z}] \neq 0$. Therefore, Li and Stengos [59] proposed to employ the instrument vector $\boldsymbol{V}$ to get a consistent estimator. Note that we can rewrite model (2.1) as in (2.4) even if $E[\varepsilon \mid \boldsymbol{V}, \boldsymbol{Z}] = 0$ holds instead of $E[\varepsilon \mid \boldsymbol{X}, \boldsymbol{Z}] = 0$ with $\boldsymbol{V} \neq \boldsymbol{X}$. The estimator is based on the moment equations

$$E\left[f_z(\boldsymbol{Z})(\boldsymbol{V} - E[\boldsymbol{V} \mid \boldsymbol{Z}])\left(f_z(\boldsymbol{Z})\left(Y - E[Y \mid \boldsymbol{Z}]\right) - f_z(\boldsymbol{Z})(\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}])^T\boldsymbol{\beta}\right)\right],$$

where the number of covariates $\boldsymbol{X}$ are the same as the number of instruments $\boldsymbol{V}$, i.e. $s = p$. As before, the regressors $E[\boldsymbol{V} \mid \boldsymbol{Z}]$ are unknown and need to be estimated. Li and Stengos [59] use the Nadaraya-Watson estimator here as well and show that the estimator is $\sqrt{n}$-consistent and asymptotically normally distributed.

We will show in this chapter that it is possible to apply the SmoothMD estimator to estimate $\boldsymbol{\beta}_0$ in model (2.1) even for endogenous $\boldsymbol{X}$. In addition, we argue that the SmoothMD approach has properties that might lead to better testing and inference results compared to the estimators of Li [57], Li and Stengos [59] and Robinson [70]. The reason is that the estimation of $E[\boldsymbol{Y} \mid \boldsymbol{Z}]$, $E[\boldsymbol{X} \mid \boldsymbol{Z}]$ and $E[\boldsymbol{V} \mid \boldsymbol{Z}]$ introduces a small sample bias. The SmoothMD approach is able to capture a part of this bias so that we can hope to get a better small sample behavior with respect to hypothesis testing.

The remainder of the chapter is organized as follows. In section 2.2, we present our new estimation method, establish identification of the model parameter and develop our uniform-in-bandwidth theory, including consistency and $\sqrt{n}-$consistency of our estimator as well as a testing procedure. Section 2.3 discusses the differences between SmoothMD and the established approaches. In section 2.4, we study the small sample behavior of our estimator by a simulation study. Section 2.5 concludes. Technical assumptions are stated in section 2.6.

2.2. The SmoothMD approach

In this section we consider how the SmoothMD approach can be applied in case of endogeneity and formally define our estimator. First, we develop the new SmoothMD approach and prove identification of the parameter of interest. Then, we define our SmoothMD estimator and prove consistency and $\sqrt{n}-$consistency of the estimator. Finally, we state a testing procedure.

We use the following notation throughout the remaining of the chapter. For $d_l, d_c \geq 1$, let $\mathbb{R}^{d_l \times d_c}$ denote the set of $d_l \times d_c$ $\boldsymbol{A}$ matrices with real elements. Let $\mathbf{1}_{d_l}$ (resp. $\mathbf{0}_{d_l}$) denote the vector with all components equal to 1 (resp. 0), $\mathbf{0}_{d_l \times d_c}$ the $d_l \times d_c-$null matrix and $\boldsymbol{I}_{d_l \times d_l}$ the identity matrix with dimension $d_l \times d_l$. For a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|$ is the Frobenius norm.

2.2.1. SmoothMD in case of endogeneity

In this section we consider the SmoothMD approach as proposed by Lavergne and Patilea [56] and extend it for our needs. Consider a general conditional moment restrictions model

$$E[g(\boldsymbol{U}; \boldsymbol{\theta}) \mid \boldsymbol{W}] = 0, \tag{2.5}$$

where $g(\cdot)$ is a given function, $\boldsymbol{U}$ and $\boldsymbol{W}$ are vectors of observed variables, and $\boldsymbol{\theta}$ is the unknown finite-dimensional parameter of interest. The components of $\boldsymbol{W}$ need not be continuous random variables. It is assumed that there exists a unique $\boldsymbol{\theta}_0$ such that $E[g(\boldsymbol{U}; \boldsymbol{\theta}_0) \mid \boldsymbol{W}] = 0$. The SmoothMD approach is based on an equivalent rewriting of equation (2.5) as an unconditional moment. For this purpose, let $\omega(\cdot)$ be a symmetric function of $\boldsymbol{W}$ with positive Fourier transform. We will employ $\omega(\boldsymbol{W}) = \exp\left\{-\boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W}\right\}$, where $\boldsymbol{D} = \operatorname{diag}(\boldsymbol{d})$ is some positive definite diagonal matrix with $\boldsymbol{d} \in \mathcal{D} \subset \mathbb{R}_+^{s+q}$ being a diagonal vector with strictly positive components. $\mathcal{D}$ is a compact set. The condition stated in (2.5) is satisfied if and only if

$$Q(\boldsymbol{\theta}) = E[g(\boldsymbol{U}_1;\boldsymbol{\theta})g(\boldsymbol{U}_2;\boldsymbol{\theta})\omega(\boldsymbol{W}_1 - \boldsymbol{W}_2)] = 0,$$

where $\left(\boldsymbol{U}_1^T, \boldsymbol{W}_1^T\right)^T$ and $\left(\boldsymbol{U}_2^T, \boldsymbol{W}_2^T\right)^T$ are independent copies of $\left(\boldsymbol{U}^T, \boldsymbol{W}^T\right)^T$. Whenever $E[g(\boldsymbol{U};\boldsymbol{\theta}) \mid \boldsymbol{W}] \neq 0$ it follows that $Q(\boldsymbol{\theta}) > 0$. Finally, the SmoothMD estimator is defined as the minimum of a sample based approximation of $Q(\boldsymbol{\theta})$ and yields $\sqrt{n}$-consistent and asymptotically normally distributed estimates for $\boldsymbol{\theta}_0$.

In case of model (2.1), we have to extend the SmoothMD approach to a model that contains an infinite-dimensional nuisance parameter as in chapter 1. Let $\boldsymbol{W} = \left(\boldsymbol{V}^T, \boldsymbol{Z}^T\right)^T$ and $\boldsymbol{U} = \left(Y, \boldsymbol{X}^T, \boldsymbol{Z}^T\right)^T$. Moreover, let

$$g(\boldsymbol{U};\boldsymbol{\beta},\gamma,\boldsymbol{\eta}) = \left(Y - E[Y \mid \boldsymbol{Z}] - (\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}])^T\boldsymbol{\beta}\right) f_z(\boldsymbol{Z}) - \gamma. \tag{2.6}$$

Here, $\boldsymbol{\eta}$ is an infinite-dimensional nuisance parameter containing the three unknown functions of $\boldsymbol{Z}$ appearing in the definition of $g(\boldsymbol{U};\boldsymbol{\beta},\gamma,\boldsymbol{\eta})$ and $\gamma \in \mathbb{R}$ is an intercept nuisance parameter. Now, the partially linear mean regression model can be stated as conditional moment equation by $E[g(\boldsymbol{U};\boldsymbol{\beta},\gamma,\boldsymbol{\eta}) \mid \boldsymbol{W}] = 0$. The true value of the intercept γ is known to be equal to zero. However, as already discussed in chapter 1 this artificial parameter will be helpful to diminish the amplitude of the variance coming from the nonparametric estimators of the unknown functions in the asymptotic representation of the estimator. The SmoothMD approach we employ to get an estimate for $\boldsymbol{\beta}$ is given by

$$E[g(\boldsymbol{U};\boldsymbol{\beta},\gamma,\boldsymbol{\eta}) \mid \boldsymbol{W}] = 0 \iff Q(\boldsymbol{\beta},\gamma) = E[g(\boldsymbol{U}_1;\boldsymbol{\theta},\gamma,\boldsymbol{\eta}_1)g(\boldsymbol{U}_2;\boldsymbol{\theta},\gamma,\boldsymbol{\eta}_2)\omega(\boldsymbol{W}_1 - \boldsymbol{W}_2)] = 0,$$

where $\left(\boldsymbol{U}_1^T, \boldsymbol{W}_1^T\right)^T$ and $\left(\boldsymbol{U}_2^T, \boldsymbol{W}_2^T\right)^T$ are again independent copies of $\left(\boldsymbol{U}^T, \boldsymbol{W}^T\right)^T$. It follows by construction that when $E[g(\boldsymbol{U};\boldsymbol{\beta},\gamma,\boldsymbol{\eta}) \mid \boldsymbol{W}] \neq 0$ we have that $Q(\boldsymbol{\beta},\gamma) > 0$. Therefore, the SmoothMD estimator is defined as the minimum of a sample based version of $Q(\boldsymbol{\beta},\gamma)$.

Given an i.i.d. sample $\left(\boldsymbol{U}_1^T, \boldsymbol{W}_1^T\right)^T, \ldots, \left(\boldsymbol{U}_n^T, \boldsymbol{W}_n^T\right)^T$ and nonparametric estimates $\widehat{\boldsymbol{\eta}}_1, \ldots, \widehat{\boldsymbol{\eta}}_n$ of the values of the nuisance parameter we define

$$\widehat{Q}_n\left(\boldsymbol{\beta},\gamma\right) = \frac{1}{n^2}\sum_{1\leq i,j\leq n} g(\boldsymbol{U}_i;\boldsymbol{\beta},\gamma,\widehat{\boldsymbol{\eta}}_i)g(\boldsymbol{U}_j;\boldsymbol{\beta},\gamma,\widehat{\boldsymbol{\eta}}_j)\omega(\boldsymbol{W}_i - \boldsymbol{W}_j).$$

The sample based version $\widehat{Q}_n\left(\boldsymbol{\beta},\gamma\right)$ of $Q(\boldsymbol{\beta},\gamma)$ is quadratic with an explicit unique minimum $(\widehat{\gamma},\widehat{\boldsymbol{\beta}}^T)^T$.

Note that the SmoothMD estimators with endogenous and exogenous covariate vector $\boldsymbol{X}$ differ only in the weighting $\omega(\boldsymbol{W}_1 - \boldsymbol{W}_2)$, i.e. $\boldsymbol{W} = \left(\boldsymbol{X}^T, \boldsymbol{Z}^T\right)^T$ when $\boldsymbol{X}$ is exogenous and $\boldsymbol{W} = \left(\boldsymbol{V}^T, \boldsymbol{Z}^T\right)^T$ when $\boldsymbol{X}$ is endogenous. In contrast to the estimator of Li and Stengos [59] an estimate of $E[\boldsymbol{V} \mid \boldsymbol{Z}]$ is not needed.

2.2.2. Identification

In this section we show that the parameters $(\gamma,\boldsymbol{\beta}^T)^T$ in our model are uniquely identified. It follows by construction that

$$\left(0,\boldsymbol{\beta}_0^T\right)^T = \arg\min_{\gamma\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p} E[g(\boldsymbol{U}_1;\boldsymbol{\beta},\gamma,\boldsymbol{\eta}_1)g(\boldsymbol{U}_2;\boldsymbol{\beta},\gamma,\boldsymbol{\eta}_2)\omega(\boldsymbol{W}_1 - \boldsymbol{W}_2)], \tag{2.7}$$

where $g(\boldsymbol{U}_1;\boldsymbol{\beta},\gamma,\boldsymbol{\eta}_1)$ and $g(\boldsymbol{U}_2;\boldsymbol{\beta},\gamma,\boldsymbol{\eta}_2)$ are independent copies of $g(\boldsymbol{U};\boldsymbol{\beta},\gamma,\boldsymbol{\eta})$ defined in equation (2.6) and

$$\boldsymbol{\eta} = \left(f_z(\cdot), E[Y \mid \boldsymbol{Z} = \cdot\,], E[\boldsymbol{X} \mid \boldsymbol{Z} = \cdot\,]^T\right)^T.$$

The following statement shows that the minimum in (2.7) is unique such that the model parameters are uniquely identified.

Lemma 2.1. *Suppose that Assumptions 2.1 and 2.2 hold true. Then,*

$$\mathbb{P}\left(E\left[(Y - E[Y \mid \boldsymbol{Z}])f_z(\boldsymbol{Z}) - \gamma - (\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}])^T\boldsymbol{\beta}f_z(\boldsymbol{Z}) \mid \boldsymbol{V},\boldsymbol{Z}\right] = 0\right) < 1,$$

for all $\gamma \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $(\gamma,\boldsymbol{\beta}^T)^T \neq (0,\boldsymbol{\beta}_0^T)^T$.

For identification, it is necessary to assume that $Var\left[E[\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{V}] - E[\boldsymbol{X}|\boldsymbol{Z}]\right]$ has full rank, see Assumption 2.2.1. This ensures that $\boldsymbol{X}$ and $\boldsymbol{V}$ are not independent and also not independent conditional on $\boldsymbol{Z}$, otherwise we would get that $E[\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{V}] - E[\boldsymbol{X}|\boldsymbol{Z}] = E[\boldsymbol{X}|\boldsymbol{Z}] - E[\boldsymbol{X}|\boldsymbol{Z}] = \boldsymbol{0}_p$ and the full rank assumption is not met. In addition, the assumption implies that $s \geq p$ to ensure the full rank condition.

2.2.3. The estimator

In this section we state our estimation strategy that follows the ideas in chapter 1. Given an independent sample $\left(Y_1, \boldsymbol{X}_1^T, \boldsymbol{Z}_1^T, \boldsymbol{V}_1^T\right)^T, \ldots, \left(Y_n, \boldsymbol{X}_n^T, \boldsymbol{Z}_n^T, \boldsymbol{V}_n^T\right)^T$ from $\left(Y, \boldsymbol{X}^T, \boldsymbol{Z}^T, \boldsymbol{V}^T\right)^T \in \mathbb{R} \times \mathbb{R}^{p+q+s}$, let us define

$$\widehat{\mathbb{Y}}_n = \Big((Y_1 - \widehat{E}[Y_1 \mid \boldsymbol{Z}_1])\widehat{f}_z(\boldsymbol{Z}_1), \ldots, (Y_n - \widehat{E}[Y_n \mid \boldsymbol{Z}_n])\widehat{f}_z(\boldsymbol{Z}_n)\Big)^T \in \mathbb{R}^n$$
$$\text{and} \quad \widehat{\mathbb{X}}_n = \Big((\boldsymbol{X}_1 - \widehat{E}[\boldsymbol{X}_1 \mid \boldsymbol{Z}_1])\widehat{f}_z(\boldsymbol{Z}_1), \ldots, (\boldsymbol{X}_n - \widehat{E}[\boldsymbol{X}_n \mid \boldsymbol{Z}_n])\widehat{f}_z(\boldsymbol{Z}_n)\Big)^T \in \mathbb{R}^{n\times p}.$$

For $1 \leq i \leq n$, $\widehat{\boldsymbol{\eta}}_i = (\widehat{f}_z(\boldsymbol{Z}_i), \widehat{E}[Y_i \mid \boldsymbol{Z}_i], \widehat{E}[\boldsymbol{X}_i \mid \boldsymbol{Z}_i]^T)^T$ is a nonparametric estimate of $\boldsymbol{\eta}_i = (f_z(\boldsymbol{Z}_i), E[Y_i \mid \boldsymbol{Z}_i], E[\boldsymbol{X}_i \mid \boldsymbol{Z}_i]^T)^T$. For the unknown values we use the kernel estimates

$$\widehat{f}_z(\boldsymbol{Z}_i) = \frac{1}{nh^q}\sum_{j=1}^n K\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_j}{h}\right), \quad \widehat{E}[Y_i \mid \boldsymbol{Z}_i]\widehat{f}_z(\boldsymbol{Z}_i) = \frac{1}{nh^q}\sum_{j=1}^n Y_j K\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_j}{h}\right)$$
$$\text{and} \quad \widehat{E}[\boldsymbol{X}_i \mid \boldsymbol{Z}_i]\widehat{f}_z(\boldsymbol{Z}_i) = \frac{1}{nh^q}\sum_{j=1}^n \boldsymbol{X}_j K\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_j}{h}\right).$$

Here $K(\cdot)$ is a multivariate kernel function and h is the bandwidth. Let $\boldsymbol{\Omega}_n$ be the $n \times n-$ symmetric matrix with elements

$$\boldsymbol{\Omega}_{n,ij} = \exp\{-(\boldsymbol{V}_i^T - \boldsymbol{V}_j^T, \boldsymbol{Z}_i^T - \boldsymbol{Z}_j^T)\boldsymbol{D}(\boldsymbol{V}_i - \boldsymbol{V}_j, \boldsymbol{Z}_i - \boldsymbol{Z}_j)\}, \qquad 1 \leq i, j \leq n.$$

Typically, the components of the vector $\boldsymbol{d}$ defining the diagonal matrix $\boldsymbol{D}$ are proportional to the standard deviation of the components of the vector $(\boldsymbol{V}_i^T, \boldsymbol{Z}_i^T)^T$. The definition of $\boldsymbol{\Omega}_{n,ij}$ allows also to take into account discrete components of $\boldsymbol{V}$.

We can now define the estimates of $(\gamma, \boldsymbol{\beta}^T)^T \in \mathbb{R}^{1+p}$ introduced in equation (2.7). Let

$$\widehat{Q}_n(\boldsymbol{\beta}, \gamma) = n^{-2}\left(\widehat{\mathbb{Y}}_n - \gamma\mathbf{1}_n - \widehat{\mathbb{X}}_n\boldsymbol{\beta}\right)^T \boldsymbol{\Omega}_n \left(\widehat{\mathbb{Y}}_n - \gamma\mathbf{1}_n - \widehat{\mathbb{X}}_n\boldsymbol{\beta}\right).$$

Finally, consider the generalized least-squares problem

$$\min_{\gamma, \boldsymbol{\beta}} \widehat{Q}_n(\boldsymbol{\beta}, \gamma).$$

The solution of this problem has the form of standard generalized least-squares estimators:

$$\widehat{\gamma}(\boldsymbol{\beta}) = \frac{1}{\mathbf{1}_n^T\boldsymbol{\Omega}_n\mathbf{1}_n}\mathbf{1}_n^T\boldsymbol{\Omega}_n\left(\widehat{\mathbb{Y}}_n - \widehat{\mathbb{X}}_n\boldsymbol{\beta}\right) \quad \text{and} \quad \widehat{\boldsymbol{\beta}} = \left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{Y}}_n,$$

with

$$\mathbb{D}_n = \boldsymbol{\Omega}_n - \frac{1}{\mathbf{1}_n^T\boldsymbol{\Omega}_n\mathbf{1}_n}\boldsymbol{\Omega}_n\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{\Omega}_n \in \mathbb{R}^{n\times n}.$$

The structure of $\mathbb{D}_n$ is as in chapter 1, however, $\boldsymbol{\Omega}_n$ does now depend on $\boldsymbol{V}$ instead of $\boldsymbol{X}$. In addition, we do not need the matrix $\widehat{\mathbb{B}}_n$ for estimation as we do not need to estimate the transformation parameter. Note that again, by construction, $\mathbb{D}_n\mathbf{1}_n = \mathbf{0}_n$.

We close this section showing that our estimator is well-defined.

Lemma 2.2. *If Assumptions 2.1.3 and 2.2 hold true, then, for each $n \geq 1$, the matrices $\boldsymbol{\Omega}_n$ and $\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n$ are positive definite with probability 1. In particular, $\mathbf{1}_n^T\boldsymbol{\Omega}_n\mathbf{1}_n > 0$ and $\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n$ is invertible with probability 1.*

2.2.4. Consistency and asymptotic normality

In this section we consider the asymptotic behavior of the estimator. Our asymptotic results are stated uniformly with respect to the diagonal of the matrix $\boldsymbol{D}$. This ensures that we can use a data driven estimate of $\boldsymbol{D}$ proportional to the empirical standard deviations of $\boldsymbol{V}$ and $\boldsymbol{Z}$.

Let's introduce some more notation:

$$\mathbb{Y}_n = ((Y_1 - E[Y_1 \mid \boldsymbol{Z}_1])f_z(\boldsymbol{Z}_1), \ldots, (Y_n - E[Y_n \mid \boldsymbol{Z}_n])f_z(\boldsymbol{Z}_n))^T \in \mathbb{R}^n,$$

and

$$\mathbb{X}_n = ((\boldsymbol{X}_1 - E[\boldsymbol{X}_1 \mid \boldsymbol{Z}_1])f_z(\boldsymbol{Z}_1), \ldots, (\boldsymbol{X}_n - E[\boldsymbol{X}_n \mid \boldsymbol{Z}_n])f_z(\boldsymbol{Z}_n))^T \in \mathbb{R}^{n\times p}.$$

With all this in hand we can now state consistency of our estimator.

Theorem 2.1 (Consistency)**.** *Assume that Assumptions 2.1, 2.2 and 2.3 hold true. Then*

$$\sup_{h\in\mathcal{H}_{c,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right\| = o_{\mathbb{P}}(1) \quad \textit{and} \quad \sup_{h\in\mathcal{H}_{c,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\widehat{\gamma}(\widehat{\boldsymbol{\beta}}) = o_{\mathbb{P}}(1).$$

In Theorem 2.1 we require that $h\in\mathcal{H}_{c,n}$, where $\mathcal{H}_{c,n}=[c_{min}n^{-\alpha}, c_{max}n^{-\alpha}]$, with $0<\alpha<1/q$ and c_{min}, c_{max} are positive constants. This implies that $nh^q\to\infty$ and $h\to 0$ for $n\to\infty$ which is in line with Li [57], Li and Stengos [59] and Robinson [70] as well as chapter 1.

Next, we prove asymptotic normality of our estimator. For this purpose, we first derive the asymptotic linear representation of $\widehat{\boldsymbol{\beta}}$ from which the $\sqrt{n}-$asymptotic normality follows. In the following result, we show that $\widehat{\boldsymbol{\beta}}$ is asymptotically not equivalent to the infeasible estimator of $\boldsymbol{\beta}_0$ one would obtain when the infinite-dimensional parameter $\boldsymbol{\eta}$ is given and the intercept γ is equal to 0. This is in contrast to the results of Li [57], Li and Stengos [59] and Robinson [70]. The reason is that they can use the fact that $E[\mathbb{X}_{n,i}\mid \boldsymbol{Z}_i]=0$ when controlling higher order terms. In our case, we weight the observations by $\boldsymbol{\Omega}_{n,ij}$ such that $E[\mathbb{X}_{n,i}\boldsymbol{\Omega}_{n,ij}\mid \boldsymbol{Z}_i]\neq 0$ for $i\neq j$. For more details consider section 2.3. This is also the reason why we need to ask for $q<4$ instead of $q<6$ as in Li [57]. Therefore, we require that $h\in\mathcal{H}_{sc,n}$, where $\mathcal{H}_{sc,n}=[c_{min}n^{-\alpha}, c_{max}n^{-\alpha}]$, with $\alpha\in(1/4,1/q)$.

The results are again obtained uniformly with respect to the elements on the diagonal of the matrix $\boldsymbol{D}$ that determines $\boldsymbol{\Omega}_n$. In addition, let $K_h(\cdot)=h^{-q}K(\cdot/h)$ and, for any $1\leq i,j\leq n$, let

$$K_{h,ij}=K_h(\boldsymbol{Z}_i-\boldsymbol{Z}_j).$$

Proposition 2.1 (Asymptotic representation)**.** *Assume that the conditions of Theorem 2.1 hold true. Moreover, Assumption 2.4 holds true. Then, uniformly with respect to $h\in\mathcal{H}_{sc,n}$ and $\boldsymbol{d}\in\mathcal{D}$,*

$$\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0=\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\mathbb{X}_n^T\mathbb{D}_n\left[(\boldsymbol{\varepsilon f_z})_n-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right]+o_{\mathbb{P}}(n^{-1/2})=O_{\mathbb{P}}(n^{-1/2}),$$

where $(\boldsymbol{\varepsilon f_z})_n=(\varepsilon_1 f_z(\boldsymbol{Z}_1),\ldots,\varepsilon_n f_z(\boldsymbol{Z}_n))^T$ *and* $\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n=\left(\frac{1}{n}\sum_{k=1,k\neq 1}^{n}\varepsilon_k K_{h,1k},\ldots,\frac{1}{n}\sum_{k=1,k\neq n}^{n}\varepsilon_k K_{h,nk}\right)^T.$

In the following we state asymptotic normality of our estimator. Therefore, we use the notation $\boldsymbol{\Omega}_{n,i,j}(\boldsymbol{d})=\boldsymbol{\Omega}_{n,i,j}$ and

$$\mathbb{D}_n(\boldsymbol{d})=\boldsymbol{\Omega}_n(\boldsymbol{d})-\frac{1}{\boldsymbol{1}_n^T\boldsymbol{\Omega}_n(\boldsymbol{d})\boldsymbol{1}_n}\boldsymbol{\Omega}_n(\boldsymbol{d})\boldsymbol{1}_n\boldsymbol{1}_n^T\boldsymbol{\Omega}_n(\boldsymbol{d}),$$

to make the dependence of $\boldsymbol{\Omega}_n$ on $\boldsymbol{d}$ explicit. Note that with

$$\begin{aligned}\boldsymbol{\Omega}^V_{n,ij}(\boldsymbol{d})&=\boldsymbol{\Omega}^V_{n,ij}=\exp\{-(\boldsymbol{V}_i-\boldsymbol{V}_j)^T\operatorname{diag}(d_1,\ldots,d_s)(\boldsymbol{V}_i-\boldsymbol{V}_j)\} \qquad \text{and}\\ \boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d})&=\boldsymbol{\Omega}^Z_{n,ij}=\exp\{-(\boldsymbol{Z}_i-\boldsymbol{Z}_j)^T\operatorname{diag}(d_{s+1},\ldots,d_{s+q})(\boldsymbol{Z}_i-\boldsymbol{Z}_j)\},\quad 1\leq i,j\leq n,\end{aligned}$$

$\boldsymbol{\Omega}_{n,ij}(\boldsymbol{d})=\boldsymbol{\Omega}^V_{n,ij}(\boldsymbol{d})\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d})$. As discussed the structure of $\mathbb{D}_n$ is the same as in chapter 1. The only difference is that we replace $\boldsymbol{\Omega}^X_{n,ij}(\boldsymbol{d})$ by $\boldsymbol{\Omega}^V_{n,ij}(\boldsymbol{d})$ when $\boldsymbol{X}$ is endogenous.

Furthermore, we define, for $1\leq i\leq n$,

$$\boldsymbol{\tau}_i(\boldsymbol{d}):=\mathbb{X}_{n,i}-\frac{1}{E\left[\boldsymbol{1}_n^T\boldsymbol{\Omega}_n(\boldsymbol{d})\boldsymbol{1}_n\right]}E\left[\mathbb{X}_n^T\boldsymbol{\Omega}_n(\boldsymbol{d})\boldsymbol{1}_n\right],$$

where $\mathbb{X}_{n,i}=(\boldsymbol{X}_i-E[\boldsymbol{X}_i\mid\boldsymbol{Z}_i])f_z(\boldsymbol{Z}_i)$. In addition, let

$$\boldsymbol{\Phi}^V_{n,ij}(\boldsymbol{d})=\boldsymbol{\Omega}^V_{n,ij}(\boldsymbol{d})-E\left[\boldsymbol{\Omega}^V_{n,ik}(\boldsymbol{d})\mid\boldsymbol{V}_i\right].$$

With all this in hand we can state the following Theorem.

Theorem 2.2 (Asymptotic normality)**.** *Assume that the conditions of Proposition 2.1 hold true. Then, uniformly with respect to $h\in\mathcal{H}_{sc,n}$ and $\boldsymbol{d}\in\mathcal{D}$,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)=E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n(\boldsymbol{d})\mathbb{X}_n\right]^{-1}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\,\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d})\boldsymbol{\Phi}^V_{n,ij}(\boldsymbol{d})\mid\boldsymbol{V}_j,\boldsymbol{Z}_j\right]\right)+o_{\mathbb{P}}(1),$$

converges in distribution to a tight random process whose marginal distribution is zero-mean normal with *covariance function* $\boldsymbol{E}\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n(\boldsymbol{d}_1)\mathbb{X}_n\right]^{-1}\boldsymbol{\Delta}(\boldsymbol{d}_1,\boldsymbol{d}_2)E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n(\boldsymbol{d}_2)\mathbb{X}_n\right]^{-1}$ *with*

$$\boldsymbol{\Delta}(\boldsymbol{d}_1,\boldsymbol{d}_2)=E\left[Var\left[\varepsilon_j f_z(\boldsymbol{Z}_j)\mid \boldsymbol{V}_j,\boldsymbol{Z}_j\right]\boldsymbol{\tau}_i(\boldsymbol{d}_1)\,\boldsymbol{\tau}_k(\boldsymbol{d}_2)^T\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d}_1)\boldsymbol{\Omega}^Z_{n,kj}(\boldsymbol{d}_2)\boldsymbol{\Phi}^V_{n,ij}(\boldsymbol{d}_1)\boldsymbol{\Phi}^V_{n,kj}(\boldsymbol{d}_2)\right].$$

Due to the estimation error coming from the estimation of $\boldsymbol{\eta}$ we need $\boldsymbol{\Phi}^V_{n,ij}(\boldsymbol{d})$ to state the asymptotic variance of our estimators. If $\boldsymbol{\eta}$ was known $\boldsymbol{\Phi}^V_{n,ij}(\boldsymbol{d})$ should be replaced by $\boldsymbol{\Omega}^V_{n,ij}(\boldsymbol{d})$.

We can estimate the covariance matrix by $\left(n^{-2}\widehat{\mathbb{X}}_n^T\mathbb{D}_n(\boldsymbol{d}_1)\widehat{\mathbb{X}}_n\right)^{-1}\widehat{\boldsymbol{\Delta}}(\boldsymbol{d}_1,\boldsymbol{d}_2)\left(n^{-2}\widehat{\mathbb{X}}_n^T\mathbb{D}_n(\boldsymbol{d}_2)\widehat{\mathbb{X}}_n\right)^{-1}$, where

$$\widehat{\boldsymbol{\Delta}}(\boldsymbol{d}_1,\boldsymbol{d}_2)=n^{-3}\widehat{\mathbb{X}}_n^T\mathbb{D}_{n,inf}(\boldsymbol{d}_1)\widehat{\boldsymbol{\Phi}}_n(\boldsymbol{d}_1)\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\Phi}}_n^T(\boldsymbol{d}_2)\mathbb{D}_{n,inf}^T(\boldsymbol{d}_2)\widehat{\mathbb{X}}_n. \tag{2.8}$$

Here, $\widehat{\boldsymbol{\Phi}}_n^V$ and $\widehat{\boldsymbol{\Phi}}_n$ are the $n\times n-$ symmetric matrices with elements

$$\begin{aligned}\widehat{\boldsymbol{\Phi}}^V_{n,ij}(\boldsymbol{d})&=\boldsymbol{\Omega}^V_{n,ij}(\boldsymbol{d})-\frac{1}{n}\sum_{k=1}^n\boldsymbol{\Omega}^V_{n,ik}(\boldsymbol{d}), &&1\le i,j\le n\\ \widehat{\boldsymbol{\Phi}}_{n,ij}(\boldsymbol{d})&=\widehat{\boldsymbol{\Phi}}^V_{n,ij}(\boldsymbol{d})\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d}), &&1\le i,j\le n\\ \text{and}\quad \mathbb{D}_{n,inf}(\boldsymbol{d})&=\boldsymbol{I}_{n\times n}-\frac{1}{\boldsymbol{1}_n^T\boldsymbol{\Omega}_n(\boldsymbol{d})\boldsymbol{1}_n}\boldsymbol{\Omega}_n(\boldsymbol{d})\boldsymbol{1}_n\boldsymbol{1}_n^T.\end{aligned}$$

$\widehat{\boldsymbol{\Sigma}}_n=\operatorname{diag}\left(\widehat{Var}\left[\varepsilon_1 f_z(\boldsymbol{Z}_1)\mid\boldsymbol{V}_1,\boldsymbol{Z}_1\right],\ldots,\widehat{Var}\left[\varepsilon_n f_z(\boldsymbol{Z}_n)\mid\boldsymbol{V}_n,\boldsymbol{Z}_n\right]\right)$ is an estimator of the error variance. One can use a nonparametric estimator for the conditional variance or alternatively use an estimate of the error terms to approximate the conditional variance in the spirit of the Eiker-White variance estimator. Consistency of the above estimators is straightforward to establish.

2.2.5. Testing the slope coefficients

In this section we provide a distance metric statistic to test restrictions for $\boldsymbol{\beta}$. Let

$$\boldsymbol{A}_n=\mathbb{X}_n^T\mathbb{D}_n\left((\boldsymbol{\varepsilon}\boldsymbol{f_z})_n-\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right).$$

Suppose we want to test r linear restrictions for $\boldsymbol{\beta}$ given by

$$H_0:\boldsymbol{R}\boldsymbol{\beta}_0=\boldsymbol{c}, \tag{2.9}$$

where $\boldsymbol{R}$ is a $r\times p-$ matrix of full rank and $\boldsymbol{c}\in\mathbb{R}^r$. In order to test the restrictions, we need to find the restricted estimator for $\boldsymbol{\beta}_0$, $\widehat{\boldsymbol{\beta}}_R$. Therefore, we minimize

$$n^{-2}\left(\widehat{\mathbb{Y}}_n-\widehat{\mathbb{X}}_n\boldsymbol{\beta}\right)^T\mathbb{D}_n\left(\widehat{\mathbb{Y}}_n-\widehat{\mathbb{X}}_n\boldsymbol{\beta}\right)\quad s.t.\quad \boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{c},$$

with respect to $\boldsymbol{\beta}$ and get that

$$\widehat{\boldsymbol{\beta}}_R=\widehat{\boldsymbol{\beta}}-\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\boldsymbol{R}^T\right)^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{c}\right).$$

Given the restricted estimator we can now define our distance metric statistic for testing (2.9).

$$DM=\frac{1}{n}\left(\widehat{\mathbb{Y}}_n-\widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R\right)^T\mathbb{D}_n\left(\widehat{\mathbb{Y}}_n-\widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R\right)-\frac{1}{n}\left(\widehat{\mathbb{Y}}_n-\widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}\right)^T\mathbb{D}_n\left(\widehat{\mathbb{Y}}_n-\widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}\right).$$

The following result shows the validity of the distance metric statistic.

Proposition 2.2. *Assume that the conditions of Proposition 2.1 hold true. Then, uniformly with respect to* $h\in\mathcal{H}_{sc,n}$ *and* $\boldsymbol{d}\in\mathcal{D}$,

$$DM-n^{-3/2}\boldsymbol{A}_n^TE\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}E\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{A}_nn^{-3/2}=o_{\mathbb{P}}(1),$$

under H_0 *and* $\mathbb{P}(n^{-1}DM>c)\to 1$ *for some* $c>0$ *if* H_0 *does not hold.*

The process in Proposition 2.2 is asymptotically tight and for each $\boldsymbol{d}$ behaves asymptotically as a weighted sum of $p-r$ independent chi-squares, where the weights are the positive eigenvalues of

$$E\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}E\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{\Delta}(\boldsymbol{d},\boldsymbol{d}),$$

see Johnson et al. [49]. Determining critical values requires the estimation of the last display. We can use the estimator stated in (2.8) and for all other components we simply replace the unknown expressions by their sample equivalence.

2.3. Why SmoothMD might be preferable

The motivation for the SmoothMD estimator is that models nonlinear in parameters that are based on conditional moment restrictions as condition (2.2) can render inconsistent parameter estimates when the generalized method of moments (GMM) is used for estimation. See Dominguez and Lobato [29]. However, the partially linear model (2.1) is linear in $\boldsymbol{\beta}$. Therefore, it seems to make no sense to apply SmoothMD in this situation as $\boldsymbol{\beta}$ can also be estimated with the estimators proposed by Li [57] or Li and Stengos [59].

The reason for applying SmoothMD is that the estimation of $\boldsymbol{\eta}$ introduces a small sample bias. The SmoothMD approach is able to capture a part of this bias which should lead to an improved small sample behavior.

In the following we assume that $\boldsymbol{X}$ is exogenous such that $\boldsymbol{V}=\boldsymbol{X}$. Recall that $\widehat{\boldsymbol{\beta}}=\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{Y}}_n$. In addition, the estimator of Li [57] is given by $\widehat{\boldsymbol{\beta}}_{Li}=\left(\widehat{\mathbb{X}}_n^T\widehat{\mathbb{X}}_n\right)^{-1}\widehat{\mathbb{X}}_n^T\widehat{\mathbb{Y}}_n$. Both estimators differ only in the weighting matrix $\mathbb{D}_n$. This matrix is defined as

$$\mathbb{D}_n=\boldsymbol{\Omega}_n-\frac{1}{\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\boldsymbol{1}_n}\boldsymbol{\Omega}_n\boldsymbol{1}_n\boldsymbol{1}_n^T\boldsymbol{\Omega}_n.$$

By construction, we have that $\widehat{\mathbb{X}}_n^T\boldsymbol{1}_n=\boldsymbol{0}_p$. Therefore, the estimator $\widehat{\boldsymbol{\beta}}_{Li}$ is equivalent to $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\Omega}_n=\boldsymbol{I}_{n\times n}$.

Recall from Proposition 2.1 that

$$\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0=\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\mathbb{X}_n^T\mathbb{D}_n\left[(\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n-\left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right]+o_{\mathbb{P}}(n^{-1/2}).$$

Furthermore, it was established in the proof of Theorem 2.2 that

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)=E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\Bigg(&\frac{1}{\sqrt{n}}\sum_{j=1}^n\varepsilon_jf_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\,\boldsymbol{\Omega}_{n,ij}(\boldsymbol{d})\mid\boldsymbol{X}_j,\boldsymbol{Z}_j\right]\\&-\frac{1}{\sqrt{n}}\sum_{k=1}^n\varepsilon_kf_z(\boldsymbol{Z}_k)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ik}^Z(\boldsymbol{d})\boldsymbol{\Omega}_{n,ij}^X(\boldsymbol{d})\mid\boldsymbol{Z}_k\right]\Bigg)+o_{\mathbb{P}}(1).\end{aligned}$$

The process in the last display is asymptotically tight. The second sum, $\frac{1}{\sqrt{n}}\sum\limits_{k=1}^n\varepsilon_kf_z(\boldsymbol{Z}_k)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ik}^Z\boldsymbol{\Omega}_{n,ij}^X\mid\boldsymbol{Z}_k\right]$, in the asymptotic representation of $\widehat{\boldsymbol{\beta}}$ occurs due to the estimation of $\boldsymbol{\eta}$. The reason is that

$$\begin{aligned}&\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\frac{1}{n^2}\mathbb{X}_n^T\boldsymbol{\Omega}_n\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n-\frac{1}{n}\sum_{k=1}^n\varepsilon_kf_z(\boldsymbol{Z}_k)E\left[\mathbb{X}_{n,i}\boldsymbol{\Omega}_{n,ik}^Z\boldsymbol{\Omega}_{n,ij}^X\mid\boldsymbol{Z}_k\right]\right\|=o_{\mathbb{P}}(n^{-1/2})\\&\text{as well as}\sup_{h\in\mathcal{H}_{sc,n}}\sup_{\boldsymbol{d}\in\mathcal{D}}\left\|\frac{1}{n^2}\boldsymbol{1}_n^T\boldsymbol{\Omega}_n\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n-\frac{1}{n}\sum_{k=1}^n\varepsilon_kf_z(\boldsymbol{Z}_k)E\left[\boldsymbol{\Omega}_{n,ik}^Z\boldsymbol{\Omega}_{n,ij}^X\mid\boldsymbol{Z}_k\right]\right\|=o_{\mathbb{P}}(n^{-1/2}).\end{aligned}\tag{2.10}$$

In contrast, we get that

$$\sup_{h\in\mathcal{H}_{sc,n}}\left\|\frac{1}{n^2}\mathbb{X}_n^T\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n-\frac{1}{n}\sum_{k=1}^n\varepsilon_kf_z(\boldsymbol{Z}_k)E\left[\mathbb{X}_{n,k}\mid\boldsymbol{Z}_k\right]\right\|=\sup_{h\in\mathcal{H}_{sc,n}}\left\|\frac{1}{n^2}\mathbb{X}_n^T\left(\widehat{\varepsilon}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right\|=o_{\mathbb{P}}(n^{-1/2})\tag{2.11}$$

as $E\left[\mathbb{X}_{n,k}\mid\boldsymbol{Z}_k\right]=\boldsymbol{0}_p$. Therefore, the asymptotic representation of the estimator proposed by Li [57] is given by

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{Li}-\boldsymbol{\beta}_0\right)=E\left[n^{-2}\mathbb{X}_n^T\mathbb{X}_n\right]^{-1}\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j f_z(\boldsymbol{Z}_j)\mathbb{X}_{n,j}+o_{\mathbb{P}}(1).$$

The process in the last display is asymptotically tight.

The main difference between the asymptotic representations of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{Li}$ is that the error due to the estimation of $\boldsymbol{\eta}$ has no influence on the representation of $\widehat{\boldsymbol{\beta}}_{Li}$ in contrast to the one of $\widehat{\boldsymbol{\beta}}$. The reasons are statements (2.10) and (2.11). However, in finite samples the estimator of Li [57] will be influenced by the estimation error of $\boldsymbol{\eta}$ as well. Therefore, employing $\widehat{\boldsymbol{\beta}}$ might improve the accuracy in case of hypothesis testing. The simulation results in section 2.4 support the argumentation.

Note that it is possible to estimate $\boldsymbol{\beta}_0$ without the intercept nuisance parameter γ as well, see the discussion in chapter 1. In that case $\mathbb{D}_n$ is replaced by $\boldsymbol{\Omega}_n$ in the estimation of $\boldsymbol{\beta}$. The estimator is than still $\sqrt{n}-$consistent and asymptotically normally distributed. In the variance

$$\boldsymbol{E}\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n(\boldsymbol{d}_1)\mathbb{X}_n\right]^{-1}\boldsymbol{\Delta}(\boldsymbol{d}_1,\boldsymbol{d}_2)E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n(\boldsymbol{d}_2)\mathbb{X}_n\right]^{-1}$$

we replace $\mathbb{D}_n$ by $\boldsymbol{\Omega}_n$ and $\boldsymbol{\tau}_i(\boldsymbol{d})$ by $\mathbb{X}_{n,i}$. When estimating the variance, $\mathbb{D}_{n,inf}$ has to be replaced by $\boldsymbol{I}_{n\times n}$ and $\mathbb{D}_n$ by $\boldsymbol{\Omega}_n$.

However, when $\boldsymbol{\beta}_0$ is estimated with intercept nuisance parameter γ the impact of estimating $\boldsymbol{\eta}$ on the asymptotic variance might become small. Consider again $\frac{1}{n}\sum\limits_{k=1}^{n}\varepsilon_k f_z(\boldsymbol{Z}_k)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ik}^Z\boldsymbol{\Omega}_{n,ij}^X\mid\boldsymbol{Z}_k\right]$. If we could replace the index k by j in $\boldsymbol{\Omega}_{n,ik}^Z$ we would get that

$$E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ij}^Z\boldsymbol{\Omega}_{n,ij}^X\mid\boldsymbol{Z}_k\right]=E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ij}^Z\boldsymbol{\Omega}_{n,ij}^X\right]=\boldsymbol{0}_p,$$

such that the second sum in the asymptotic representation of $\widehat{\boldsymbol{\beta}}$ would not be present. Of course this is not possible. However, here we consider $\boldsymbol{d}$ as a vector with elements playing the role of standardizing constants. If the elements of $\boldsymbol{d}$ are the inverse of a kernel smoothing bandwidth tending to zero at a suitable rate $E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\boldsymbol{\Omega}_{n,ik}^Z\boldsymbol{\Omega}_{n,ij}^X\mid\boldsymbol{Z}_k\right]$ tends to zero for $n\to\infty$. The exact rate is over the scope of this chapter, but we have seen in the simulations of chapter 1 that even in case of $\boldsymbol{d}$ being a vector of constants it seems that we can forget about the second part in the asymptotic representation. The estimator is labeled SmoothMD* in the simulation section. If we do not consider the second part in the asymptotic representation, the variance is estimated by replacing $\widehat{\boldsymbol{\Phi}}_{n,ij}(\boldsymbol{d})$ with $\boldsymbol{\Omega}_{n,ij}(\boldsymbol{d})$.

Note that the previous discussion does not apply to the case without constant γ as

$$E\left[\mathbb{X}_{n,i}\boldsymbol{\Omega}_{n,ij}^Z\boldsymbol{\Omega}_{n,ij}^X\mid\boldsymbol{Z}_k\right]=E\left[\mathbb{X}_{n,i}\boldsymbol{\Omega}_{n,ij}^Z\boldsymbol{\Omega}_{n,ij}^X\right]\neq\boldsymbol{0}_p.$$

2.4. Small sample study

In this section we consider the small sample behavior of our estimator. We conduct several simulation experiments to consider bias and standard deviation for the estimated parameters. In addition, we conduct hypothesis tests for $\boldsymbol{\beta}$. We begin with a consideration of the simulation setup. Finally, we state our simulation results.

2.4.1. Simulation setup

During the simulation, we consider six different models. The models are given by

Model 1: $\log(Y)=X\beta_0+m(Z)+\varepsilon$, $m(Z)=\frac{\exp\{Z\}}{1+\exp\{Z\}}+\frac{1}{3}$ with $Z\sim N(1,1)$, $\beta_0=1$, $X=-\frac{2}{3}Z+u$ with $u\sim N(0,1)$ and $\varepsilon=\sqrt{\frac{1+X^2}{2}}\,\widetilde{u}$ with $\widetilde{u}\sim N\left(0,\frac{1}{13}\right)$.

Model 2: $\log(Y)=X\beta_0+m(Z_1+Z_2+Z_3)+\varepsilon$, $m(Z)=\frac{\exp\{Z\}}{1+\exp\{Z\}}+\frac{1}{3}$ with $Z_1,Z_2,Z_3\sim N(1,1)$, $\beta_0=1$, $X=-\frac{2}{9}(Z_1+Z_2+Z_3)+u$ with $u\sim N(0,1)$ and $\varepsilon=\sqrt{\frac{1+X^2}{2}}\,\widetilde{u}$ with $\widetilde{u}\sim N\left(0,\frac{3}{35}\right)$.

Model 3: $Y=X\beta_0+m(Z)+\varepsilon$, $m(Z)=\frac{\exp\{Z\}}{1+\exp\{Z\}}-1$ with $Z\sim U(-3,-1)$ and $\beta_0=1$, $X=\frac{2}{3}Z+u$ with $u\sim U(-1,1)$ and $\varepsilon\sim U\left(-\sqrt{1/9},\sqrt{1/9}\right)$.

Model 4: $Y=X\beta_0+m(Z_1+Z_2+Z_3)+\varepsilon$, $m(Z)=\frac{\exp\{Z\}}{1+\exp\{Z\}}-1$ with $Z_1,Z_2,Z_3\sim U(-3,-1)$ and $\beta_0=1$, $X=\frac{2}{9}(Z_1+Z_2+Z_3)+u$ with $u\sim U(-1,1)$ and $\varepsilon\sim U\left(-\sqrt{1/9},\sqrt{1/9}\right)$.

Model 5: $Y = X\beta_0 + m(Z_1 + Z_2 + Z_3) + \varepsilon$, $m(Z) = \frac{\exp\{Z\}}{1+\exp\{Z\}} + \frac{1}{3}$ with $Z_1, Z_2, Z_3 \sim N(1,1)$, $\beta_0 = 1$, $X = -\frac{2}{9}(Z_1 + Z_2 + Z_3) + V + u$ with $V \sim N(2,9)$, $u \sim N(0,1)$, $\varepsilon \sim N\left(0, \frac{1}{9}\right)$ and $Cov(u, \varepsilon) = 0.5$.

Model 6: $Y = X\beta_0 + m(Z_1 + Z_2 + Z_3) + \varepsilon$, $m(Z) = \frac{\exp\{Z\}}{1+\exp\{Z\}} + \frac{1}{3}$ with $Z_1, Z_2, Z_3 \sim N(1,1)$, $\beta_0 = 1$, $X = -\frac{2}{9}(Z_1 + Z_2 + Z_3) + V + u$ with $V \sim N(2,9)$, $u \sim N(0,1)$, $\varepsilon = \sqrt{4.5}\left(\widetilde{\varepsilon}^2 - 1/9\right)$, where $\widetilde{\varepsilon} \sim N\left(0, \frac{1}{9}\right)$ and $Cov\left(u, \widetilde{\varepsilon}\right) = 0.5$.

The models differ in the number of covariates $\boldsymbol{Z}$. In Model 1 and Model 3, $m(\cdot)$ contains only one covariate, whereas it contains three in all other models. In addition, the error terms ε are heteroscedastic in Model 1 and Model 2 in contrast to the remaining models. Furthermore, the error terms in Model 6 have a skewed density function as they are χ^2-distributed.

The estimators are computed by employing a normal kernel for $K(\cdot)$. $\boldsymbol{Z}$ is standardized componentwise by the corresponding standard deviations and $h \propto n^{-1/3.5}$. This bandwidth choice satisfies the assumptions of Theorem 2.2. The components of $\boldsymbol{d}$ defining the diagonal matrix $\boldsymbol{D}$ in $\boldsymbol{\Omega}_n$ are set equal to the componentwise standard deviations of X and $\boldsymbol{Z}$ in Model 1 – Model 4 and to the componentwise standard deviations of V and $\boldsymbol{Z}$ in Model 5 and Model 6 as X is endogenous in the latter models.

In the simulation we compare the proposed estimator where γ is employed with the estimator that does not use γ. Either of the two estimators converges asymptotically to a normal distribution. Therefore, it is interesting to consider which one has the better small sample behavior.

We consider bias and standard deviation of the estimators as well as the size of the distance metric statistics proposed in Section 2.2.5. In addition, we test by a simple Z-Test if the estimated parameters are significantly different from the true value. Therefore, we employ the variance estimator stated in (2.8) for both estimators with the necessary adjustments for the estimator without γ. To estimate the error variance we employ the Eiker-White variance estimator. In order to see the influence of the estimated $\boldsymbol{\eta}$ on the variance we consider all tests also without taking the estimation error of $\boldsymbol{\eta}$ into account. Therefore, we replace $\widehat{\boldsymbol{\Phi}}_n$ by $\boldsymbol{\Omega}_n$ in the variance estimator (2.8).

In addition, we consider the estimator of Li [57] for Model 1 – Model 4 and the estimator of Li and Stengos [59] for Model 5 and Model 6 as competitors. Let

$$\widehat{\mathbb{V}}_n = \Big((\boldsymbol{V}_1 - \widehat{E}[\boldsymbol{V}_1 \mid \boldsymbol{Z}_1])\widehat{f}_z(\boldsymbol{Z}_1), \ldots, (\boldsymbol{V}_n - \widehat{E}[\boldsymbol{V}_n \mid \boldsymbol{Z}_n])\widehat{f}_z(\boldsymbol{Z}_n)\Big)^T \in \mathbb{R}^{n \times p}$$

with $\widehat{E}[\boldsymbol{V}_i \mid \boldsymbol{Z}_i]\widehat{f}_z(\boldsymbol{Z}_i) = \frac{1}{nh^q}\sum_{j=1}^n \boldsymbol{V}_j K\left(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_j}{h}\right)$ for all i. Note that $s = p$ in the considered models and recall that

$$\widehat{\boldsymbol{\beta}}_{Li} = \left(\widehat{\mathbb{V}}_n^T \widehat{\mathbb{X}}_n\right)^{-1} \widehat{\mathbb{V}}_n^T \widehat{\mathbb{Y}}_n,$$

where $\widehat{\mathbb{V}}_n = \widehat{\mathbb{X}}_n$ in case of exogenous $\boldsymbol{X}$. The variance of $\widehat{\boldsymbol{\beta}}_{Li}$ can be estimated by

$$\left(\widehat{\mathbb{V}}_n^T \widehat{\mathbb{X}}_n\right)^{-1} \widehat{\mathbb{V}}_n^T \widehat{\boldsymbol{\Sigma}}_n \widehat{\mathbb{V}}_n \left(\widehat{\mathbb{X}}_n^T \widehat{\mathbb{V}}_n\right)^{-1},$$

where $\widehat{\boldsymbol{\Sigma}}_n = \operatorname{diag}\Big(\widehat{Var}\left[\varepsilon_1 f_z(\boldsymbol{Z}_1) \mid \boldsymbol{V}_1, \boldsymbol{Z}_1\right], \ldots, \widehat{Var}\left[\varepsilon_n f_z(\boldsymbol{Z}_n) \mid \boldsymbol{V}_n, \boldsymbol{Z}_n\right]\Big)$ as in (2.8).

In the estimation of $E[X \mid \boldsymbol{Z}]$ and $E[V \mid \boldsymbol{Z}]$ we employ the same kernel and bandwidth as for the SmoothMD estimator and we use again the Eiker-White variance estimator to estimate the error variance.

2.4.2. Simulation results

Table 2.1 states the results for bias and standard deviation for β in Model 1. All three estimators have comparable results for bias and the bias decreases with sample size. The standard deviation is largest for the estimator of Li [57] but decreases with samples size such that it is almost equal to the SmoothMD estimators for $n = 500$.

Table 2.2 states the results for bias and standard deviation for β in Model 4. Again, all three estimators have comparable results for bias and the bias decreases with sample size. In contrast, the standard deviation is larger as in Model 1 and almost equal for all three estimators. The larger standard deviation is reasonable as $m(\cdot)$ contains three covariates in Model 4 and only one in Model 1.

Table 2.3 states the results for bias and standard deviation for β in Model 5. The bias for $n = 50$ is here a bit larger for SmoothMD than for the estimator of Li and Stengos [59]. However, the bias decreases with sample size. The standard deviation is almost equal for all three estimators.

Table 2.1: *Bias and Standard Deviation of the estimator for β in Model 1.*

	Bias			St. dev.		
n	50	250	500	50	250	500
β estimator						
SmoothMD with γ	−0.002	−0.001	−0.001	0.062	0.027	0.02
SmoothMD without γ	−0.002	−0.001	−0.001	0.063	0.026	0.02
Li	0.001	−0.0001	0.0001	0.068	0.03	0.022

Notes: For the SmoothMD estimators and the estimator of Li, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. For all simulations 2000 Monte Carlo samples were used.

Table 2.2: *Bias and Standard Deviation of the estimator for β in Model 4.*

	Bias			St. dev.		
n	50	250	500	50	250	500
β estimator						
SmoothMD with γ	0.0002	−0.002	−0.001	0.09	0.04	0.02
SmoothMD without γ	0.0002	−0.002	−0.001	0.09	0.04	0.02
Li	−0.0001	−0.002	−0.001	0.1	0.04	0.03

Notes: For the SmoothMD estimators and the estimator of Li, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. For all simulations 2000 Monte Carlo samples were used.

Table 2.3: *Bias and Standard Deviation of the estimator for β in Model 5.*

	Bias			St. dev.		
n	50	250	500	50	250	500
β estimator						
SmoothMD with γ	0.02	0.005	0.003	0.06	0.025	0.018
SmoothMD without γ	0.02	0.005	0.003	0.06	0.025	0.018
Li and Stengos	−0.004	−0.002	−0.001	0.07	0.027	0.019

Notes: For the SmoothMD estimators and the estimator of Li and Stengos, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. For all simulations 2000 Monte Carlo samples were used.

Table 2.4: *Bias and Standard Deviation of the estimator for β in Model 6.*

	Bias			St. dev.		
n	50	250	500	50	250	500
β estimator						
SmoothMD with γ	0.0003	0.0002	0.0001	0.097	0.04	0.03
SmoothMD without γ	0.0003	0.0002	0.0001	0.097	0.04	0.03
Li and Stengos	0.001	−0.0002	−0.0001	0.11	0.05	0.03

Notes: For the SmoothMD estimators and the estimator of Li and Stengos, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. For all simulations 2000 Monte Carlo samples were used.

Table 2.4 states the results for bias and standard deviation for β in Model 6. The bias is for all three estimators smaller than in Model 5, the second model with endogenous covariates. However, the standard deviation is larger as in Model 5. This might be due to the skewed error distribution of Model 6.

In addition, note that the results for SmoothMD with and without γ are close for all considered models.

Table 2.5: *Empirical Level for the Z-Test of the estimator for β in Model 2.*

	5% level			10% level		
n	50	250	500	50	250	500
Test for β						
SmoothMD with γ	19.7	7.6	5.85	27.4	13.4	11.5
SmoothMD* with γ	19.6	8.6	6.2	26.55	14.6	12.25
SmoothMD without γ	19.85	7.7	5.9	27.15	13.55	11.6
SmoothMD* without γ	19.7	8.65	6.25	26.4	14.3	12.25
Li	26.75	15.45	11.45	32.5	23.3	18.85

Notes: For the SmoothMD estimators and the estimator of Li, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

Table 2.6: *Empirical Level for the Z-Test of the estimator for β in Model 3.*

	5% level			10% level		
n	50	250	500	50	250	500
Test for β						
SmoothMD with γ	7.8	4.7	4.0	12.55	9.7	8.65
SmoothMD* with γ	8.45	5.4	4.9	13.25	10.9	10.1
SmoothMD without γ	7.9	4.75	3.95	12.6	9.75	8.6
SmoothMD* without γ	8.15	5.45	4.9	13.1	10.8	10.15
Li	9.65	5.95	5.35	14.3	11.5	10.25

Notes: For the SmoothMD estimators and the estimator of Li, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

Table 2.7: *Empirical Level for the distance metric statistic of the estimator for β in Model 1.*

	5% level			10% level		
n	50	250	500	50	250	500
Test for β						
SmoothMD with γ	6.05	3.45	4.0	10.75	7.4	8.9
SmoothMD* with γ	7.0	4.1	4.9	11.8	8.8	10.5
SmoothMD without γ	6.2	3.4	3.95	11.15	7.55	9.0
SmoothMD* without γ	7.15	3.9	4.9	12.3	9.0	10.65

Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

Table 2.8: *Empirical Level for the Z-Test of the estimator for β in Model 6.*

	5% level			10% level		
n	50	250	500	50	250	500
Test for β						
SmoothMD with γ	16.1	6.5	5.7	24.2	12.05	11.05
SmoothMD* with γ	15.3	6.5	5.8	22.6	11.9	11.05
SmoothMD without γ	16.3	6.45	5.7	24.05	12.05	11.05
SmoothMD* without γ	14.75	6.65	5.45	21.95	11.7	10.85
Li and Stengos	20.5	11.85	9.7	29.2	19.65	16.6

Notes: For the SmoothMD estimators and the estimator of Li and Stengos, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

Table 2.5 states the empirical level for the Z-Tests for β in Model 2. All four considered SmoothMD estimators get close to the nominal levels for increasing n. However, the estimator of Li [57] overrejects for all considered sample sizes. It seems that the results improve for increasing sample size but do not get close to the results of the SmoothMD estimators. SmoothMD with and without γ lead to similar results.

Table 2.6 states the empirical level for the Z-Tests for β in Model 3. All five considered estimators get close to the nominal levels for increasing n. However, for $n = 50$ all estimators overrejct with the estimator of Li [57] performing worst. SmoothMD with and without γ lead to similar results. Surprisingly, the SmoothMD with wrongly estimated variance gets closer to the nominal levels than the SmoothMD with correct variance. Note that the main difference between Model 2 and Model 3 is in the dimension of $\boldsymbol{Z}$. From the results it seems that the estimator of Li [57] needs a larger sample size to control the estimation error of $\boldsymbol{\eta}$ with the results getting worse the larger the dimension of $\boldsymbol{Z}$.

Table 2.8 states the empirical level for the Z-Tests for β in Model 6. The stated results are in line with the ones discussed before. Therefore, it seems that the SmoothMD estimator outperforms the estimator of Li and Stengos [59].

Finally, Table 2.7 states the empirical level for the distance metric statistic of the estimator for β in Model 1. The results are convincing, supporting the theoretical statement. SmoothMD with and without γ lead to similar results.

Figure 2.1: *Power function of the Z-Test for β of Model 1 with $n = 250$.*

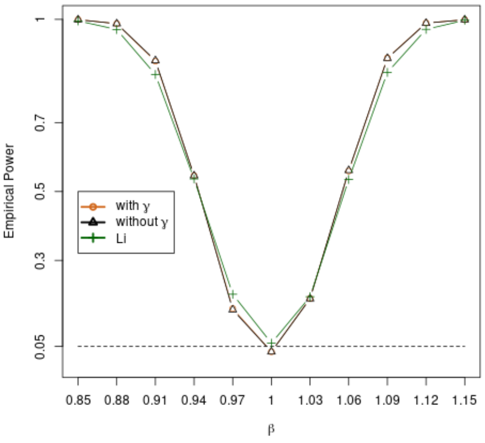
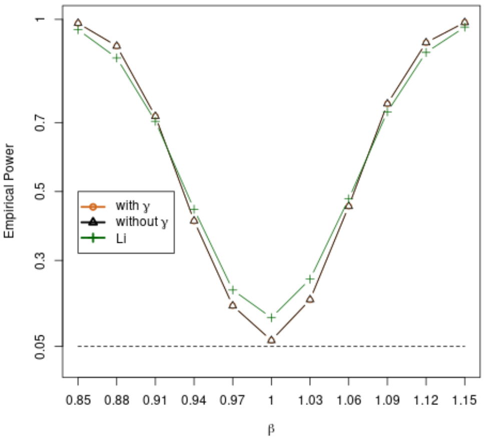


Figure 2.2: *Power function of the Z-Test for β of Model 4 with $n = 250$.*



Notes: For the SmoothMD estimators and the estimator of Li, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. Only the SmoothMD estimators that take the additional variance part due to the estimation of $\boldsymbol{\eta}$ into account are considered. For all simulations 2000 Monte Carlo samples were used. The nominal level is 5%.

Figure 2.3: *Power function of the Z-Test for β of Model 5 with $n = 250$.*

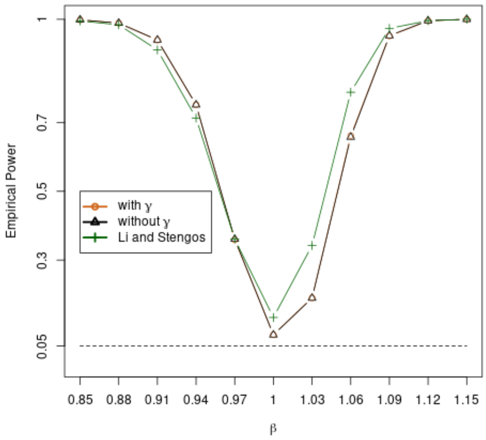
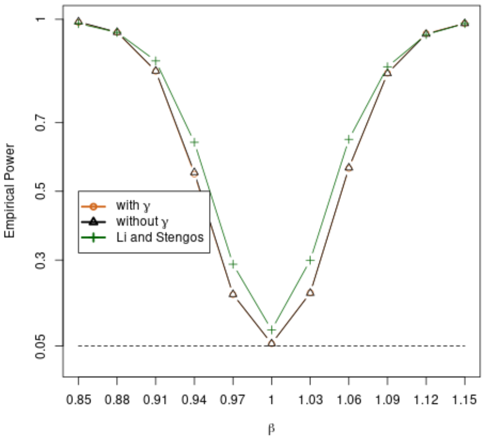


Figure 2.4: *Power function of the Z-Test for β of Model 6 with $n = 500$.*



Notes: For the SmoothMD estimators and the estimator of Li and Stengos, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. Only the SmoothMD estimators that take the additional variance part due to the estimation of $\boldsymbol{\eta}$ into account are considered. For all simulations 2000 Monte Carlo samples were used. The nominal level is 5%.

Figures 2.1 and 2.2 state the power functions of the Z-Test for β in Model 1 and Model 4 with $n = 250$. We consider here only the estimator of SmoothMD with the correct error variance and the estimator of Li [57]. In the first case all three estimators perform reasonable well and it seems that the SmoothMD estimators have a bit larger power. In the second case the estimator of Li [57] overrejects at the true value whereas the SmoothMD estimators only overreject slightly. It seems that in terms of power the SmoothMD outperforms the estimator of Li [57] here as well. Note that both SmoothMD versions lead to similar results.

Figures 2.3 and 2.4 state the power functions of the Z-Test for β in Model 5 with $n = 250$ and Model 6 with $n = 500$. In the first case the power functions for all three estimators are not symmetric anymore. Furthermore, the estimator of Li and Stengos [59] overrejects at the true value whereas the SmoothMD estimators only overreject slightly. In the second case the power functions are symmetric and the SmoothMD estimators reach the nominal value at the true value and the estimator of Li and Stengos [59] overrejects again.

2.5. Conclusion

In this paper we considered the semiparametric partially linear model studied in Li [57] and Robinson [70]. We employed the SmoothMD estimation technique to the partially linear model and argued why SmoothMD might be preferable to the estimators proposed by Li [57] and Li and Stengos [59].

We established consistency as well as $\sqrt{n}$-asymptotic normality. In addition, we proposed a distance metric statistic to test the model parameters. A Monte Carlo experiment showed the usefulness of the proposed estimator in small samples.

2.6. Assumptions

Assumption 2.1. *Data Generating Process*

1. The observations $\left(Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T, \boldsymbol{V}_i^T\right)^T$, $1 \leq i \leq n$, are i.i.d. copies of $\left(Y, \boldsymbol{X}^T, \boldsymbol{Z}^T, \boldsymbol{V}^T\right)^T \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^s$, with $s \geq p$.
2. The covariate vector $\boldsymbol{Z}$ admits a bounded density with respect to the Lebesgue measure in $\mathbb{R}^q$. The covariate vectors $\boldsymbol{X}$ and $\boldsymbol{V}$ are split in two subvectors $\boldsymbol{X}_c \in \mathbb{R}^{p_c}$ and $\boldsymbol{X}_d \in \mathbb{R}^{p_d}$, $\boldsymbol{V}_c \in \mathbb{R}^{s_c}$ and $\boldsymbol{V}_d \in \mathbb{R}^{s_d}$ with $0 \leq p_c, p_d \leq p$, $p_c + p_d = p$, $0 \leq s_c, s_d \leq s$ and $s_c + s_d = s$. The subvectors $\boldsymbol{X}_c$ and $\boldsymbol{V}_c$ admit bounded densities with respect to the Lebesgue measures in $\mathbb{R}^{p_c}$ and $\mathbb{R}^{s_c}$. The subvectors $\boldsymbol{X}_d$ and $\boldsymbol{V}_d$ take values in finite sets.
3. The $(s+q)$ diagonal components of the matrix $\boldsymbol{D}$ belong to the set $\mathcal{D} = [d_L, d_U] \times \cdots \times [d_L, d_U] \subset \mathbb{R}_+^{s+q}$, with some fixed $0 < d_L < d_U < \infty$.

The assumption that the discrete components of $\boldsymbol{X}$ and $\boldsymbol{V}$ take values in finite sets is a technical condition that simplifies the proofs without significant restriction of the generality for the applications.

Assumption 2.2. *Identification*

1. $E\left[\|\boldsymbol{X}\|^2\right] < \infty$ and $Var\left[E[\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{V}] - E[\boldsymbol{X}|\boldsymbol{Z}]\right]$ as well as $Var\left[(\boldsymbol{V}^T, \boldsymbol{Z}^T)^T\right]$ have full rank.
2. The true value $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is not equal to $\mathbf{0}_p$.
3. We have that $E\left[Y\right] < \infty$.

Note that Assumption 2.2.1 implies that we need $s \geq p$. In addition, we have that

$$Var\left[\boldsymbol{u}^T(\boldsymbol{X} - E\left[\boldsymbol{X} \mid \boldsymbol{Z}\right])\right] = Var\left[\boldsymbol{u}^T(\boldsymbol{X} - E\left[\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{V}\right])\right] + Var\left[\boldsymbol{u}^T(E\left[\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{V}\right] - E\left[\boldsymbol{X} \mid \boldsymbol{Z}\right])\right] > 0$$

such that $Var\left[\boldsymbol{X} - E[\boldsymbol{X}|\boldsymbol{Z}]\right]$ has full rank as well. Therefore, it follows from the discussion in chapter 1 that $Var\left[(\boldsymbol{X}^T, \boldsymbol{Z}^T)^T\right]$ has full rank.

Assumption 2.3. *Consistency*

1. The kernel $K(\cdot)$ is the product of q univariate kernel functions $\widetilde{K}$ of bounded variation. Moreover, $\widetilde{K}$ is a symmetric function with integral equal to one and $\int_{\mathbb{R}} t^2 \widetilde{K}(t) dt < \infty$.
2. The functions $f_z(\cdot)$, $(mf_z)(\cdot)$, $E[\|\boldsymbol{X}\|^2 \mid \boldsymbol{Z} = \cdot\,]f_z(\cdot)$ and $E[Y \mid \boldsymbol{Z} = \cdot\,]f_z(\cdot)$ have Hölder continuous partial derivatives of order four.
3. The bandwidth h belongs to a range $\mathcal{H}_{c,n} = [c_{min} n^{-\alpha}, c_{max} n^{-\alpha}]$, with $0 < \alpha < 1/q$ and c_{min}, c_{max} positive constants.
4. It holds true that $E\left[\|\boldsymbol{X}\|^4\right] < \infty$, $E\left[\|\boldsymbol{Z}\|\right] < \infty$, $E\left[Y^4\right] < \infty$ and $E\left[\varepsilon^4\right] < \infty$.

Assumption 2.4. *Asymptotic Normality*

1. The bandwidth h belongs to a range $\mathcal{H}_{sc,n} = [c_{min} n^{-\alpha}, c_{max} n^{-\alpha}]$, with $\alpha \in (1/4, 1/q)$ and c_{min}, c_{max} positive constants.
2. $E\left[\varepsilon^2 \mid \boldsymbol{V}, \boldsymbol{Z}\right] = \sigma^2(\boldsymbol{V}, \boldsymbol{Z})$ is in $L^1 \cap L^2$.
3. $E\left[m(\boldsymbol{Z})^4\right] < \infty$.

Appendix

Appendix A: Proofs

Proof of Lemma 2.1.

In order to ensure global identification, we employ the approach of chapter 1 that is an extension the proof in Shin [74]. For any $(\gamma, \boldsymbol{\beta}^T)^T$ we have that

$$\mathbb{P}\bigg(E\left[(Y - E[Y \mid \boldsymbol{Z}])\, f_z(\boldsymbol{Z}) - \gamma - (\boldsymbol{X} - E[\boldsymbol{X} \mid \boldsymbol{Z}])^T \boldsymbol{\beta} f_z(\boldsymbol{Z}) \mid \boldsymbol{V}, \boldsymbol{Z}\right] = 0 \bigg) \\ = \mathbb{P}\bigg(f_z(\boldsymbol{Z})\, (E[\boldsymbol{X} \mid \boldsymbol{Z}] - E[\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{V}])^T\, (\boldsymbol{\beta} - \boldsymbol{\beta}_0) = \gamma \bigg).$$

Hence, it suffices to prove that the last probability could not be equal to 1 when $(\gamma, \boldsymbol{\beta}^T)^T \neq (0, \boldsymbol{\beta}_0^T)^T$.

The result follows immediately from the full rank condition in Assumption 2.2.1. Indeed, by the variance decomposition formula and Assumption 2.2.1, for any $\boldsymbol{a} \in \mathbb{R}^p$, $\boldsymbol{a} \neq \boldsymbol{0}_p$,

$$\boldsymbol{a}^T Var\left[E[\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{V}] - E[\boldsymbol{X}|\boldsymbol{Z}]\right] \boldsymbol{a} = E\left[Var\left[\boldsymbol{a}^T (E[\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{V}] - E[\boldsymbol{X}|\boldsymbol{Z}]) \mid \boldsymbol{Z}\right]\right] > 0.$$

This implies

$$\boldsymbol{a}^T Var\left[f_z(\boldsymbol{Z})\, (E[\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{V}] - E[\boldsymbol{X}|\boldsymbol{Z}])\right] \boldsymbol{a} = E\left[f_z^2(\boldsymbol{Z}) \boldsymbol{a}^T Var\left[(E[\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{V}] - E[\boldsymbol{X}|\boldsymbol{Z}]) \mid \boldsymbol{Z}\right] \boldsymbol{a}^T\right] \\ = E\left[f_z^2(\boldsymbol{Z}) Var\left[\boldsymbol{a}^T\, (E[\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{V}] - E[\boldsymbol{X}|\boldsymbol{Z}]) \mid \boldsymbol{Z}\right]\right] > 0.$$

Thus, $f_z(\boldsymbol{Z})(E[\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{V}] - E[\boldsymbol{X} \mid \boldsymbol{Z}]^T)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ cannot be equal to a constant almost surely.

□

Proof of Lemma 2.2.

We follow the proof of Lemma 1.2 in chapter 1. First, we note that for any $\boldsymbol{u} \in \mathbb{R}^s$ and $\boldsymbol{v} \in \mathbb{R}^q$ such that $(\boldsymbol{u}^T, \boldsymbol{v}^T)^T \neq \boldsymbol{0}_{s+q}$, and any $c \in \mathbb{R}$,

$$\mathbb{P}\left(\boldsymbol{u}^T \boldsymbol{V} + \boldsymbol{v}^T \boldsymbol{Z} = c\right) = 0. \tag{2.12}$$

This is a consequence of the fact that $Var\left[(\boldsymbol{V}^T, \boldsymbol{Z}^T)^T\right]$ has full rank, by Assumption 2.2. Given a sample $\left(\boldsymbol{V}_1^T, \boldsymbol{Z}_1^T\right)^T, \dots, \left(\boldsymbol{V}_n^T, \boldsymbol{Z}_n^T\right)^T$, and a vector $\boldsymbol{a} = (a_1, \dots, a_n) \in \mathbb{R}^n$, using the inverse Fourier Transform, we could write

$$\boldsymbol{a}^T \boldsymbol{\Omega}_n \boldsymbol{a} = \frac{\pi^{-(s+q)/2}}{\sqrt{d_1 \cdots d_{s+q}}} \int_{\mathbb{R}^{s+q}} \left| \sum_{j=1}^n a_j \exp\left\{2i\boldsymbol{w}^T \left(\boldsymbol{V}_j^T, \boldsymbol{Z}_j^T\right)^T\right\} \right|^2 \exp\left\{-\boldsymbol{w}^T \boldsymbol{D}^{-1} \boldsymbol{w}\right\} d\boldsymbol{w},$$

where $\boldsymbol{D} = \mathrm{diag}(d_1, \dots, d_{s+q})$ with $d_1, \dots, d_{s+q} \in [d_L, d_U]$, see Assumption 2.1.3. Then, necessarily

$$\sum_{j=1}^n a_j \exp\left\{2i\boldsymbol{w}^T \left(\boldsymbol{V}_j^T, \boldsymbol{Z}_j^T\right)^T\right\} = 0, \qquad \forall \boldsymbol{w} \in \mathbb{R}^{s+q}. \tag{2.13}$$

Equation (2.12) indicates that, with probability 1, equation (2.13) admits the unique solution $\boldsymbol{a} = \boldsymbol{0}_n$ $\forall \boldsymbol{w} \in \mathbb{R}^{s+q}$. This means that, with probability 1, the matrix $\boldsymbol{\Omega}_n$ is positive definite.

The remaining arguments are identical to the arguments in the proof of Lemma 1.2 in chapter 1 and are, thus, omitted.

□

Proof of Theorem 2.1.

We start by considering $\widehat{\boldsymbol{\beta}}$. Recall that $\widehat{\boldsymbol{\beta}} = \left(\widehat{\mathbb{X}}_n^T \mathbb{D}_n \widehat{\mathbb{X}}_n\right)^{-1} \widehat{\mathbb{X}}_n^T \mathbb{D}_n \widehat{\mathbb{Y}}_n$. It follows now from Lemmas 1.3, 1.7, 1.8 and 1.13 in chapter 1 that

$$\left\| \widehat{\boldsymbol{\beta}} - \left(\mathbb{X}_n^T \mathbb{D}_n \mathbb{X}_n\right)^{-1} \mathbb{X}_n^T \mathbb{D}_n \mathbb{Y}_n \right\| = o_{\mathbb{P}}(1)$$

uniform with respect to $\boldsymbol{d} \in \mathcal{D}$ and $h \in \mathcal{H}_{c,n}$. Note that

$$\left(\mathbb{X}_n^T \mathbb{D}_n \mathbb{X}_n\right)^{-1} \mathbb{X}_n^T \mathbb{D}_n \mathbb{Y}_n = \boldsymbol{\beta}_0 + \left(\mathbb{X}_n^T \mathbb{D}_n \mathbb{X}_n\right)^{-1} \mathbb{X}_n^T \mathbb{D}_n (\boldsymbol{\varepsilon} \boldsymbol{f}_{\boldsymbol{z}})_n,$$

where $(\boldsymbol{\varepsilon} \boldsymbol{f}_{\boldsymbol{z}})_n = (\varepsilon_1 f_z(\boldsymbol{Z}_1), \dots, \varepsilon_n f_z(\boldsymbol{Z}_n))^T$. It follows now from Lemmas 1.3, 1.5 , 1.8 and 1.14 in chapter 1 that

$$\left\| \left(\mathbb{X}_n^T \mathbb{D}_n \mathbb{X}_n\right)^{-1} \mathbb{X}_n^T \mathbb{D}_n (\boldsymbol{\varepsilon} \boldsymbol{f}_{\boldsymbol{z}})_n \right\| = o_{\mathbb{P}}(1)$$

uniform with respect to $\boldsymbol{d} \in \mathcal{D}$ such that

$$\sup_{h \in \mathcal{H}_{c,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right\| = o_{\mathbb{P}}(1).$$

The second result follows by similar arguments.

□

Proof of Proposition 2.1 .

The result follows by the same arguments as in the proof of Proposition 1.1 in chapter 1. The proof uses Lemmas 1.18 and 1.19 of chapter 1. In the proof of the two Lemmas we need to replace $\boldsymbol{U}_i = (Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$ by $\widetilde{\boldsymbol{U}}_i = (Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T, \boldsymbol{V}_i^T)^T$ for all $i = 1, \dots, n$ when applying the Hoeffding decomposition. Apart from this replacement the arguments remain the same.

□

Proof of Proposition 2.2 .

The result follows by the same arguments as in the proof of Theorem 1.2 in chapter 1. Recall from Proposition 2.1 that

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \left(\mathbb{X}_n^T \mathbb{D}_n \mathbb{X}_n\right)^{-1} \mathbb{X}_n^T \mathbb{D}_n \left[(\boldsymbol{\varepsilon} \boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}} \right)_n \right] + o_{\mathbb{P}}(n^{-1/2})$$

uniformly with respect to $h \in \mathcal{H}_{sc,n}$ and $\boldsymbol{d} \in \mathcal{D}$. It follows now from Lemma 1.5 in chapter 1 that

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = E\left[n^{-2} \mathbb{X}_n^T \mathbb{D}_n \mathbb{X}_n\right]^{-1} n^{-2} \mathbb{X}_n^T \mathbb{D}_n \left[(\boldsymbol{\varepsilon} \boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}} \right)_n \right] + o_{\mathbb{P}}(n^{-1/2})$$

uniformly with respect to $h \in \mathcal{H}_{sc,n}$ and $\boldsymbol{d} \in \mathcal{D}$. Following the arguments in the proof of Theorem 1.2 in chapter 1, where we again replace $\boldsymbol{U}_i = (Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$ by $\widetilde{\boldsymbol{U}}_i = (Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T, \boldsymbol{V}_i^T)^T$ for all $i = 1, \dots, n$ when applying the Hoeffding decomposition, we get that

$$\begin{aligned} \sup_{h \in \mathcal{H}_{sc,n}} \sup_{\boldsymbol{d} \in \mathcal{D}} \Bigg\| & \frac{1}{n^2} \mathbb{X}_n^T \mathbb{D}_n \left((\boldsymbol{\varepsilon} \boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\varepsilon}_{|\boldsymbol{z}} \widehat{\boldsymbol{f}}_{\boldsymbol{z}} \right)_n \right) \\ & - \frac{1}{n} \sum_{j=1}^n \varepsilon_j f_z(\boldsymbol{Z}_j) E\left[\left(\mathbb{X}_{n,i} - \frac{1}{E\left[\mathbf{1}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n\right]} E\left[\mathbb{X}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n\right] \right) \boldsymbol{\Omega}_{n,ij} \mid \boldsymbol{V}_j, \boldsymbol{Z}_j \right] \\ & + \frac{1}{n} \sum_{k=1}^n \varepsilon_k f_z(\boldsymbol{Z}_k) E\left[\left(\mathbb{X}_{n,i} - \frac{1}{E\left[\mathbf{1}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n\right]} E\left[\mathbb{X}_n^T \boldsymbol{\Omega}_n \mathbf{1}_n\right] \right) \boldsymbol{\Omega}_{n,ik}^Z \boldsymbol{\Omega}_{n,ij}^V \mid \boldsymbol{Z}_k \right] \Bigg\| = o_{\mathbb{P}}(n^{-1/2}). \end{aligned}$$

Finally, we get that

$$\begin{aligned} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right) &= E\left[n^{-2} \mathbb{X}_n^T \mathbb{D}_n \mathbb{X}_n\right]^{-1} \Bigg(\frac{1}{n} \sum_{j=1}^n \varepsilon_j f_z(\boldsymbol{Z}_j) E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\, \boldsymbol{\Omega}_{n,ij}(\boldsymbol{d}) \mid \boldsymbol{V}_j, \boldsymbol{Z}_j \right] \\ & \qquad - \frac{1}{n} \sum_{k=1}^n \varepsilon_k f_z(\boldsymbol{Z}_k) E\left[\boldsymbol{\tau}_i(\boldsymbol{d}) \boldsymbol{\Omega}_{n,ik}^Z(\boldsymbol{d}) \boldsymbol{\Omega}_{n,ij}^V(\boldsymbol{d}) \mid \boldsymbol{Z}_k \right] \Bigg) + o_{\mathbb{P}}\left(n^{-1/2}\right) \\ &= E\left[n^{-2} \mathbb{X}_n^T \mathbb{D}_n \mathbb{X}_n\right]^{-1} \left(\frac{1}{n} \sum_{j=1}^n \varepsilon_j f_z(\boldsymbol{Z}_j) E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\, \boldsymbol{\Omega}_{n,ij}^Z(\boldsymbol{d}) \left(\boldsymbol{\Omega}_{n,ij}^V(\boldsymbol{d}) - \boldsymbol{\Omega}_{n,ik}^V(\boldsymbol{d}) \right) \mid \boldsymbol{V}_j, \boldsymbol{Z}_j \right] \right) + o_{\mathbb{P}}\left(n^{-1/2}\right) \end{aligned}$$

$$= E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\,\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d})\left(\boldsymbol{\Omega}^V_{n,ij}(\boldsymbol{d}) - E\left[\boldsymbol{\Omega}^V_{n,ik}(\boldsymbol{d}) \mid \boldsymbol{V}_i\right]\right) \mid \boldsymbol{V}_j, \boldsymbol{Z}_j\right]\right) + o_{\mathbb{P}}\left(n^{-1/2}\right)$$
$$= E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\,\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d})\boldsymbol{\Phi}^V_{n,ij}(\boldsymbol{d}) \mid \boldsymbol{V}_j, \boldsymbol{Z}_j\right]\right) + o_{\mathbb{P}}\left(n^{-1/2}\right),$$

uniformly over $h \in \mathcal{H}_{sc,n}$ and $\boldsymbol{d} \in \mathcal{D}$.

We can now consider the behavior of $\frac{1}{n}\sum\limits_{j=1}^{n}\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\,\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d})\boldsymbol{\Phi}^V_{n,ij}(\boldsymbol{d}) \mid \boldsymbol{V}_j, \boldsymbol{Z}_j\right]$ in detail by applying Theorem 19.28 of Van der Vaart [78] as in the proof of Theorem 1.2 in chapter 1. The Lindeberg condition follows again from our assumptions as well as

$$\sup_{\|\boldsymbol{d}_1-\boldsymbol{d}_2\|<\delta} E\Big[\Big\|\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d}_1)\,\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d}_1)\boldsymbol{\Phi}^V_{n,ij}(\boldsymbol{d}_1) \mid \boldsymbol{V}_j, \boldsymbol{Z}_j\right]$$
$$- \varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d}_2)\,\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d}_2)\boldsymbol{\Phi}^V_{n,ij}(\boldsymbol{d}_2) \mid \boldsymbol{V}_j, \boldsymbol{Z}_j\right]\Big\|^2\Big] \to 0,$$

whenever $\delta \to 0$. Therefore,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n(\boldsymbol{d})\mathbb{X}_n\right]^{-1}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j f_z(\boldsymbol{Z}_j)E\left[\boldsymbol{\tau}_i(\boldsymbol{d})\,\boldsymbol{\Omega}^Z_{n,ij}(\boldsymbol{d})\boldsymbol{\Phi}^V_{n,ij}(\boldsymbol{d}) \mid \boldsymbol{V}_j, \boldsymbol{Z}_j\right]\right) + o_{\mathbb{P}}(1),$$

converges in distribution to a tight random process whose marginal distribution is zero-mean normal with covariance function $E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n(\boldsymbol{d}_1)\mathbb{X}_n\right]^{-1}\boldsymbol{\Delta}(\boldsymbol{d}_1,\boldsymbol{d}_2)E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n(\boldsymbol{d}_2)\mathbb{X}_n\right]^{-1}$.

□

Proof of Proposition 2.2.

Under H_0 we get that

$$\left(\widehat{\mathbb{Y}}_n - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R\right)^T\mathbb{D}_n\left(\widehat{\mathbb{Y}}_n - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R\right)$$
$$= \left(\widehat{\mathbb{Y}}_n - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}\right)^T\mathbb{D}_n\left(\widehat{\mathbb{Y}}_n - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}\right) + \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right)^T\left(\boldsymbol{R}\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\boldsymbol{R}^T\right)^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right)$$
$$= \left(\widehat{\mathbb{Y}}_n - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}\right)^T\mathbb{D}_n\left(\widehat{\mathbb{Y}}_n - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}\right) + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^T\boldsymbol{R}^T\left(\boldsymbol{R}\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right).$$

By the same reasoning as in Proposition 2.1 we get that

$$n^{-1}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^T\boldsymbol{R}^T\left(\boldsymbol{R}\left(\widehat{\mathbb{X}}_n^T\mathbb{D}_n\widehat{\mathbb{X}}_n\right)^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}\,n^{-1}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$$
$$= n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right)^T\mathbb{D}_n\mathbb{X}_n$$
$$\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}$$
$$\mathbb{X}_n^T\mathbb{D}_n\,n^{-1}\left((\boldsymbol{\varepsilon}\boldsymbol{f}_{\boldsymbol{z}})_n - \left(\widehat{\boldsymbol{\varepsilon}}_{|\boldsymbol{z}}\widehat{\boldsymbol{f}}_{\boldsymbol{z}}\right)_n\right) + o_{\mathbb{P}}(1/n)$$

uniformly with respect to $\boldsymbol{d} \in \mathcal{D}$ and $h \in \mathcal{H}_{sc,n}$. Therefore, we get together with the results in Lemma 1.5 of chapter 1 that

$$\frac{1}{n}\left(\widehat{\mathbb{Y}}_n - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R\right)^T\mathbb{D}_n\frac{1}{n}\left(\widehat{\mathbb{Y}}_n - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}_R\right) - \frac{1}{n}\left(\widehat{\mathbb{Y}}_n - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}\right)^T\mathbb{D}_n\frac{1}{n}\left(\widehat{\mathbb{Y}}_n - \widehat{\mathbb{X}}_n\widehat{\boldsymbol{\beta}}\right)$$
$$= n^{-2}\boldsymbol{A}_n^T\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}\left(\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}\left(n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right)^{-1}n^{-2}\boldsymbol{A}_n + o_{\mathbb{P}}(1/n)$$
$$= n^{-2}\boldsymbol{A}_n^T E\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\left(\boldsymbol{R}E\left[\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{R}^T\right)^{-1}\boldsymbol{R}E\left[n^{-2}\mathbb{X}_n^T\mathbb{D}_n\mathbb{X}_n\right]^{-1}\boldsymbol{A}_n n^{-2} + o_{\mathbb{P}}(1/n)$$

uniformly with respect to $h \in \mathcal{H}_{sc,n}$ and $\boldsymbol{d} \in \mathcal{D}$. When H_0 does not hold it follows by the same arguments as in the proof of Proposition 2.1 that $n^{-1}DM$ converges in probability to a positive constant.

□

Table 2.B.9: *Bias and Standard Deviation of the estimator for β in Model 2.*

	Bias			St. dev.		
n	50	250	500	50	250	500
β estimator						
SmoothMD with γ	−0.002	−0.002	0.0001	0.11	0.05	0.03
SmoothMD without γ	−0.002	−0.002	0.0001	0.11	0.05	0.03
Li	−0.002	−0.002	−0.0002	0.12	0.06	0.04

Notes: For the SmoothMD estimators and the estimator of Li, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. For all simulations 2000 Monte Carlo samples were used.

Table 2.B.10: *Bias and Standard Deviation of the estimator for β in Model 3.*

	Bias			St. dev.		
n	50	250	500	50	250	500
β estimator						
SmoothMD with γ	0.002	0.001	0.001	0.06	0.02	0.02
SmoothMD without γ	0.002	0.001	0.001	0.06	0.02	0.02
Li	0.001	0.001	0.001	0.06	0.02	0.02

Notes: For the SmoothMD estimators and the estimator of Li, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. For all simulations 2000 Monte Carlo samples were used.

Table 2.B.11: *Empirical Level for the Z-Test of the estimator for β in Model 1.*

	5% level			10% level		
n	50	250	500	50	250	500
Test for β						
SmoothMD with γ	5.85	3.45	4.0	10.85	7.4	8.95
SmoothMD* with γ	6.55	4.15	4.85	11.65	8.75	10.5
SmoothMD without γ	6.0	3.45	3.95	11.2	7.5	9.0
SmoothMD* without γ	6.65	3.9	4.95	12.2	9.0	10.7
Li	10.95	5.95	5.55	16.55	11.1	10.65

Notes: For the SmoothMD estimators and the estimator of Li, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

Table 2.B.12: *Empirical Level for the Z-Test of the estimator for β in Model 4.*

	5% level			10% level		
n	50	250	500	50	250	500
Test for β						
SmoothMD with γ	18.5	6.75	5.4	26.15	13.0	10.75
SmoothMD* with γ	17.8	7.2	5.9	24.55	13.55	11.15
SmoothMD without γ	18.4	6.75	5.4	26.15	12.95	10.8
SmoothMD* without γ	17.65	7.15	5.9	24.05	13.45	11.15
Li	22.95	13.35	10.75	31.7	20.7	17.7

Notes: For the SmoothMD estimators and the estimator of Li, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

Table 2.B.13: *Empirical Level for the Z-Test of the estimator for β in Model 5.*

	5% level			10% level		
n	50	250	500	50	250	500
Test for β						
SmoothMD with γ	22.95	8.3	7.05	29.5	13.7	12.4
SmoothMD* with γ	21.6	8.4	7.1	28.15	13.45	12.5
SmoothMD without γ	22.85	8.25	7.0	29.55	13.45	12.35
SmoothMD* without γ	20.9	8.35	7.0	27.5	13.5	12.3
Li and Stengos	24.95	13.3	10.65	32.5	20.65	17.65

Notes: For the SmoothMD estimators and the estimator of Li and Stengos, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

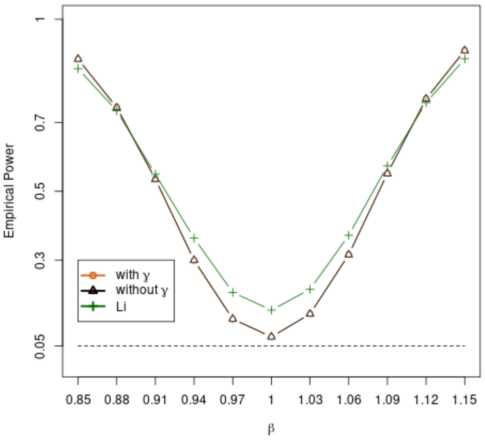
Table 2.B.14: *Empirical Level for the distance metric statistic of the estimator for β in Model 3.*

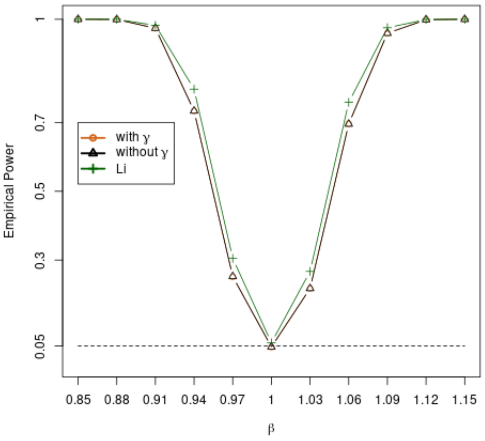
	5% level			10% level		
n	50	250	500	50	250	500
Test for β						
SmoothMD with γ	8.05	4.75	4.0	12.9	9.75	8.65
SmoothMD* with γ	8.95	5.4	4.9	14.05	11.15	10.2
SmoothMD without γ	8.15	4.75	3.95	12.95	9.8	8.6
SmoothMD* without γ	8.35	5.5	4.85	13.6	11.0	10.2

Notes: For the SmoothMD estimators, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. For SmoothMD the additional variance part due to the estimation of $\boldsymbol{\eta}$ is not taken into account. For SmoothMD the additional variance part is taken into account. For all simulations 2000 Monte Carlo samples were used.*

Figure 2.B.5: *Power function of the Z-Test for β of Model 2 with $n = 250$.*

Figure 2.B.6: *Power function of the Z-Test for β of Model 3 with $n = 250$.*





Notes: For the SmoothMD estimators and the estimator of Li, $h \propto n^{-1/3.5}$. The components of $\boldsymbol{d}$ are set equal to the componentwise standard deviations for all variables. The variances are estimated by the Eiker-White variance estimator. Only the SmoothMD estimators that take the additional variance part due to the estimation of $\boldsymbol{\eta}$ into account are considered. For all simulations 2000 Monte Carlo samples were used. The nominal level is 5%.

Chapter 3

Maximum likelihood estimation of dynamic panel data models with additive fixed effects

3.1. Introduction

We consider the linear dynamic panel data model with fixed effects that is given by

$$Y_{it} = \alpha_i + \lambda_t + \rho Y_{i,t-1} + \boldsymbol{X}_{it}^T \boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{3.1}$$

where $\boldsymbol{X}_{it} \in \mathbb{R}^k$ is a vector of explanatory variables and $(\rho, \boldsymbol{\beta}^T)^T \in \mathbb{R}^{k+1}$ are the unknown model parameters. In addition, α_i and λ_t are the unobserved individual- and time-specific effects which are assumed to be constant for given i over t and vice versa. The model is dynamic as the response variable Y_{it} is an explanatory variable of Y_{it+1} and it is a fixed effects model as α_i and λ_t are allowed to be correlated with $\boldsymbol{X}_{it}$. Furthermore, it is assumed that the error term ε_{it} has the following structure: $E[\varepsilon_{it}] = 0$, $E[\varepsilon_{it}\varepsilon_{js}] = \sigma^2$ if $i = j$ and $t = s$ whereas $E[\varepsilon_{it}\varepsilon_{js}] = 0$ otherwise.

Panel data models are frequently employed in empirical analyses to answer research questions in labor and health economics as well as finance and macroeconomics. Frequently, economists consider models with individual specific effects to allow for unobserved heterogeneity, consider for instance Dell et al. [27] and McArthur and McCord [62]. Therefore, model (3.1) plays an important role in applied research.

However, introducing individual- and/or time-specific effects in dynamic panel data models may lead to inconsistent parameter estimates when they are correlated with the exogenous variables. This results in the well known incidental parameter problem, see Neyman and Scott [65] or Nickell [66], that has been discussed extensively in the literature. The problem could be avoided if we would assume that α_i and λ_t are not correlated with $\boldsymbol{X}_{it}$. This is the so-called random effects model. Nevertheless, if the effects are correlated with the exogenous regressors the parameter estimates in the random effects model will be inconsistent.

Let $\lambda_t = 0$. In a linear panel data model without autoregressive part, i.e. $\rho = 0$, the incidental parameter problem can be solved by subtracting the individual mean from Y_{it}. This give us

$$Y_{it} - \overline{Y}_i = \left(\boldsymbol{X}_{it} - \overline{\boldsymbol{X}}_i\right)^T \boldsymbol{\beta} + \varepsilon_{it} - \overline{\varepsilon}_i, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{3.2}$$

where $\overline{Y}_i = 1/T \sum_{t=1}^T Y_{it}$, $\overline{\boldsymbol{X}}_i = 1/T \sum_{t=1}^T \boldsymbol{X}_{it}$ and $\overline{\varepsilon}_i = 1/T \sum_{t=1}^T \varepsilon_{it}$. It is now possible to estimate (3.2) by a standard OLS procedure. This estimator is also called the within-group estimator.

When $\rho \neq 0$ subtracting the individual mean leads to

$$Y_{it} - \overline{Y}_i = \rho \left(Y_{it-1} - \overline{Y}_{i\,-1}\right) + \left(\boldsymbol{X}_{it} - \overline{\boldsymbol{X}}_i\right)^T \boldsymbol{\beta} + \varepsilon_{it} - \overline{\varepsilon}_i, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{3.3}$$

where $\overline{Y}_{i\,-1} = 1/T \sum_{t=1}^T Y_{it-1}$. The OLS or within-group estimates of (3.3) are not consistent anymore for fixed T as $Y_{it-1} - \overline{Y}_{i\,-1}$ is correlated with the error term $\varepsilon_{it} - \overline{\varepsilon}_i$. When $T \to \infty$ and $|\rho| < 1$ the estimates are still consistent. When T increases faster than N it is also possible to get $\sqrt{NT}$ consistency, consider Hahn and Kuersteiner [34] and Alvarez and Arellano [5]. However, it is likely that there still occurs a bias in small samples.

To circumvent the incidental parameter problem the generalized method of moments (GMM) is frequently applied. There exists a huge amount of studies that propose different GMM versions, see among others Anderson and Hsiao [11, 12], Amemiya and MaCurdy [7], Arellano and Bond [14], Arellano and Bover [15], Ahn and Schmidt [2], Blundell and Bond [21], Hayakawa [38] and Han and Phillips [35].

On the one hand, GMM estimation is easy to compute and provides asymptotically valid inference under a minimal set of assumptions. In addition, GMM solves the problem of $Y_{it-1} - \overline{Y}_{i-1}$ being correlated with the error terms $\varepsilon_{it} - \overline{\varepsilon}_i$. On the other hand, GMM estimators suffer from a number of drawbacks including the poor behavior when the autoregressive parameter is close to unity, see e.g. Kiviet [53] and Blundell and Bond [21]. In addition, Bekker [18] showed that the asymptotic theory for GMM estimators possibly breaks down if the number of instruments tends to infinity which is a relevant scenario if T is large relative to the number of cross section units N.

In this study, we address the incidental parameter problem by employing maximum likelihood estimation techniques. Recently, various variants of maximum likelihood estimators for dynamic panel data models have been studied, see Hsiao et al. [48], Kruiniger [54], Bai [17], Han and Phillips [36], Moral-Benito [63] and Hayakawa and Pesaran [39]. Here, we compare the estimators of Hsiao et al. [48] and Bai [17][1]. The main difference between the two approaches is that Hsiao et al. [48] eliminates the fixed effects by taking first differences whereas Bai [17] models the behavior of the effects. Bai [17] focused in his paper mainly on the case without exogenous regressors, therefore we will extend his model to the empirically more relevant case of a model with exogenous explanatory variables.

To keep the discussion simple we will assume that $\lambda_t = 0$ in the remaining of the chapter.[2] The remainder of the chapter is organized as follows. Section 3.2 considers the dynamic model without exogenous regressors. In particular, the importance of the initial value in short panels is discussed. Section 3.3 allows for exogenous regressors in the dynamic model. In section 3.4 the small sample behavior of the considered estimators is studied by a Monte Carlo experiment. Finally, section 3.5 concludes.

3.2. Dynamic models without exogenous regressors

In this section we consider the linear dynamic panel data model with individual fixed effects but without additional exogenous regressors, i.e.

$$Y_{it} = \alpha_i + \rho Y_{it-1} + \varepsilon_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T. \tag{3.4}$$

We compare the estimation strategy of Bai [17] that considers the model in levels with the estimation strategy of Hsiao et al. [48] that considers the model in first differences. In addition, we discuss the importance to model the unobserved initial value Y_{i0} and how this can be done for the two estimation strategies.

We use the following notation throughout the remaining of the chapter. For $d_l \geq 1$, let $\mathbf{1}_{d_l}$ (resp. $\mathbf{0}_{d_l}$) denote the vector with all elements equal to 1 (resp. 0) and $\boldsymbol{I}_{d_l \times d_l}$ the identity matrix with dimension $d_l \times d_l$.

3.2.1. Maximum likelihood estimation of the model in levels

In this section we will first follow the discussion in the paper of Bai [17], i.e we first assume that the initial condition $Y_{i0} = 0$ for all i. However, we will show that consistency of the estimator fails if this condition is not met. Therefore, we will provide an estimation strategy when $Y_{i0} \neq 0$ for all or some i in the second part of this section.

Initial condition is zero

The parameter of interest is the autoregressive coefficient ρ which is assumed to be a fixed and finite constant. Before we state the proposed estimator of Bai [17] note that the model can be written in matrix notation as

$$\begin{aligned} \boldsymbol{B}(\rho)\boldsymbol{Y}_i &= \alpha_i \mathbf{1}_T + \boldsymbol{\varepsilon}_i \\ \boldsymbol{Y}_i &= \alpha_i \boldsymbol{\Gamma}(\rho)\mathbf{1}_T + \boldsymbol{\Gamma}(\rho)\boldsymbol{\varepsilon}_i, \end{aligned} \tag{3.5}$$

where $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{iT})^T$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{iT})^T$ are $T \times 1$ vectors. Furthermore,

$$\boldsymbol{B}(\rho) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ -\rho & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & -\rho & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Gamma}(\rho) = \boldsymbol{B}(\rho)^{-1} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \rho & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \rho^{T-1} & \dots & \rho & 1 \end{pmatrix}$$

are $T \times T$ matrices.

[1] We consider it as maximum likelihood approach even though Bai [17] labels it *factor analytical approach.*

[2] See Bai [17] and Hsiao [47] page 122 for models with $\lambda_t \neq 0$.

Bai [17] considers model (3.5) as a factor model with a single factor where $\boldsymbol{\Gamma}(\rho)\mathbf{1}_T$ is the factor loading and α_i the factor score. This analogy leads Bai [17] to label this estimation method the *Factor Analytical Approach*. This factor structure is identified for $T \geq 3$.

In order to be able to state the objective function that needs to be minimized to estimate the unknown model parameters we define $\boldsymbol{S}_N = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i\mathbf{Y}_i^T$ and state the following assumptions:

Assumption 3.1.

1. $Y_{i0} = 0$ for all i.
2. ε_{it} is i.i.d. over i and t with $E[\varepsilon_{it}] = 0$ and $Var[\varepsilon_{it}] = \sigma^2 > 0$, where σ^2 is finite.
3. $\alpha_1 \ldots, \alpha_N$ are fixed effects and $a_N = \frac{1}{N}\sum_{i=1}^{N}\alpha_i^2$. There exists a positive and finite constant a such that $\lim_{N\to\infty} a_N = a$.

In addition, let $\boldsymbol{\theta}_1 = (\rho, \sigma^2, a)^T$ and note that it follows from Assumption 3.1 that

$$\boldsymbol{S}_N \xrightarrow{p} \boldsymbol{\Gamma}(\rho)\left(a\mathbf{1}_T\mathbf{1}_T^T + \sigma^2\boldsymbol{I}_{T\times T}\right)\boldsymbol{\Gamma}(\rho)^T.^3$$

To estimate the unknown parameters Bai [17] considers a discrepancy function between $\boldsymbol{S}_N$ and the limit of $\boldsymbol{S}_N$, i.e.

$$\log\left(\left|\boldsymbol{\Gamma}(\rho)\left(a\mathbf{1}_T\mathbf{1}_T^T + \sigma^2\boldsymbol{I}_{T\times T}\right)\boldsymbol{\Gamma}(\rho)^T\right|\right) + \operatorname{tr}\left[\boldsymbol{S}_N\left(\boldsymbol{\Gamma}(\rho)\left(a\mathbf{1}_T\mathbf{1}_T^T + \sigma^2\boldsymbol{I}_{T\times T}\right)\boldsymbol{\Gamma}(\rho)^T\right)^{-1}\right].$$

This discrepancy function has the same form as the likelihood function for a central Wishart distribution if we multiply the function by $-N/2$. Furthermore, it has, up to a constant, the same form as the likelihood function in case of a random effects model with α_i and ε_i i.i.d. normal.[4] Therefore, we will in the remaining work with the likelihood function, which is given by

$$L_1(\boldsymbol{\theta}_1) = (2\pi)^{-\frac{NT}{2}}\left|\boldsymbol{\Gamma}(\rho)\left(a\mathbf{1}_T\mathbf{1}_T^T + \sigma^2\boldsymbol{I}_{T\times T}\right)\boldsymbol{\Gamma}(\rho)^T\right|^{-\frac{N}{2}} \exp\left\{-\frac{N}{2}\operatorname{tr}\left[\boldsymbol{S}_N\left(\boldsymbol{\Gamma}(\rho)\left(a\mathbf{1}_T\mathbf{1}_T^T + \sigma^2\boldsymbol{I}_{T\times T}\right)\boldsymbol{\Gamma}(\rho)^T\right)^{-1}\right]\right\}. \quad (3.6)$$

Taking the log of the likelihood function we get[5]

$$\begin{aligned}
\ell_1(\boldsymbol{\theta}_1) = \log\left(L_1(\boldsymbol{\theta}_1)\right) &= -\frac{NT}{2}\log(2\pi) - \frac{NT}{2}\log(\sigma^2) - \frac{N}{2}\log\left(1 + \frac{Ta}{\sigma^2}\right) \\
&\quad - \frac{N}{2}\operatorname{tr}\left[\boldsymbol{S}_N\left(\boldsymbol{\Gamma}(\rho)\left(a\mathbf{1}_T\mathbf{1}_T^T + \sigma^2\boldsymbol{I}_{T\times T}\right)\boldsymbol{\Gamma}(\rho)^T\right)^{-1}\right] \\
&= -\frac{NT}{2}\log(2\pi) - \frac{NT}{2}\log(\sigma^2) - \frac{N}{2}\log\left(1 + \frac{Ta}{\sigma^2}\right) \\
&\quad - \frac{1}{2}\sum_{i=1}^{N}\mathbf{Y}_i^T\boldsymbol{B}(\rho)^T\left(a\mathbf{1}_T\mathbf{1}_T^T + \sigma^2\boldsymbol{I}_{T\times T}\right)^{-1}\boldsymbol{B}(\rho)\mathbf{Y}_i \\
&= -\frac{NT}{2}\log(2\pi) - \frac{NT}{2}\log(\sigma^2) - \frac{N}{2}\log\left(1 + \frac{Ta}{\sigma^2}\right) \\
&\quad - \frac{1}{2}\sum_{i=1}^{N}\mathbf{Y}_i^T\left(\boldsymbol{I}_{T\times T} - \rho\boldsymbol{J}_{T\times T}\right)^T\left(a\mathbf{1}_T\mathbf{1}_T^T + \sigma^2\boldsymbol{I}_{T\times T}\right)^{-1}\left(\boldsymbol{I}_{T\times T} - \rho\boldsymbol{J}_{T\times T}\right)\mathbf{Y}_i,
\end{aligned}$$

[3] Bai [17] employs $\frac{1}{N-1}\sum_{i=1}^{N}\mathbf{Y}_i\mathbf{Y}_i^T$ instead of $\boldsymbol{S}_N$. However, both representations lead to the same estimation results.

[4] For further discussion of dynamic random-effects estimators consider Hsiao [47] Chapter 4.

[5] We used that $|\boldsymbol{\Gamma}(\rho)| = 1$ for all ρ as well as the fact that $|\boldsymbol{I}_{m\times m} + \boldsymbol{A}\boldsymbol{A}^T| = |\boldsymbol{I}_{n\times n} + \boldsymbol{A}^T\boldsymbol{A}|$ for some $m \times n$ matrix $\boldsymbol{A}$.

where

$$\boldsymbol{J}_{T\times T} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 1 & 0 \end{pmatrix}$$

is a $T \times T$ matrix. The estimator $\widehat{\boldsymbol{\theta}}_1 = (\widehat{\rho}_1, \widehat{\sigma}_1^2, \widehat{a}_1)^T$ of $\boldsymbol{\theta}_1$ is obtained by solving the first-order conditions of $\ell_1(\boldsymbol{\theta}_1)$, which are stated in Appendix A, simultaneously. However, there exists no closed form solution of the first-order conditions. It is possible to use the Newton-Raphson procedure to find the solution. Here, we will employ a sequential iterative procedure, see also Hsiao [47] page 45. We define the $T \times T$ matrix $\boldsymbol{Q}_{T\times T} = \boldsymbol{I}_{T\times T} - \frac{1}{T}\boldsymbol{1}_T\boldsymbol{1}_T^T$ and get from the first order-conditions the following estimates:

$$\widehat{\rho}_1 = \frac{\sum_{i=1}^N \boldsymbol{Y}_i^T \boldsymbol{J}_{T\times T}^T \left(a\boldsymbol{1}_T\boldsymbol{1}_T^T + \sigma^2 \boldsymbol{I}_{T\times T}\right)^{-1} \boldsymbol{Y}_i}{\sum_{i=1}^N \boldsymbol{Y}_i^T \boldsymbol{J}_{T\times T}^T \left(a\boldsymbol{1}_T\boldsymbol{1}_T^T + \sigma^2 \boldsymbol{I}_{T\times T}\right)^{-1} \boldsymbol{J}_{T\times T}\boldsymbol{Y}_i}, \tag{3.7}$$

$$\widehat{\sigma}_1^2 = \frac{1}{N(T-1)} \sum_{i=1}^N \boldsymbol{Y}_i^T \boldsymbol{B}(\rho)^T \boldsymbol{Q}_{T\times T} \boldsymbol{B}(\rho) \boldsymbol{Y}_i \tag{3.8}$$

$$\text{and} \quad \widehat{a}_1 = \frac{1}{NT^2} \sum_{i=1}^N \left(\boldsymbol{Y}_i^T \boldsymbol{B}(\rho)^T \boldsymbol{1}_T\right)^2 - \frac{1}{T}\sigma^2. \tag{3.9}$$

Therefore, we obtain the final estimate $\widehat{\boldsymbol{\theta}}_1$ by first substituting an initial estimate of ρ into (3.8) so that we get $\widehat{\sigma}_1^2$. We can now estimate a by substituting the initial estimate of ρ and $\widehat{\sigma}_1^2$ into (3.9). Finally, we get $\widehat{\rho}_1$ by substituting $\widehat{\sigma}_1^2$ and $\widehat{a}_1$ into (3.7). This process is repeated until it converges.

The true value of $\boldsymbol{\theta}_1$ is denoted by $\boldsymbol{\theta}_{01}$. Therefore, we can now state the following corollary.

Corollary 3.1. *Assume that Assumption 3.1 holds true. Then, invoking the results of Browne [23], Amemiya et al. [9] and Anderson and Amemiya [10] we have that*

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{01}\right) \rightsquigarrow N\left(\boldsymbol{0}_3, \boldsymbol{M}_1^{-1}\right)$$

when $N \to \infty$ and T is fixed, with $\boldsymbol{M}_1 = \frac{1}{N}E\left[\frac{\partial^2 \ell_1(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1 \partial\boldsymbol{\theta}_1^T}\right]$. In addition,

$$\sqrt{NT}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{01}\right) \rightsquigarrow N\left(\boldsymbol{0}_3, \boldsymbol{M}_2^{-1}\right)$$

when $N \to \infty$ and $T \to \infty$, with $\boldsymbol{M}_2 = \frac{1}{NT}E\left[\frac{\partial^2 \ell_1(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1 \partial\boldsymbol{\theta}_1^T}\right]$.

The approach of Bai [17] leads to a consistent and asymptotically normally distributed estimator of $\boldsymbol{\theta}_1$. However, the approach depends crucially on the assumption that $Y_{i0} = 0$ for all i. Without this assumption the estimator is not consistent as can be seen from the following discussion. The first order condition of $\ell_1(\boldsymbol{\theta}_1)$ with respect to ρ evaluated at $\boldsymbol{\theta}_{01}$ is given by[6]

$$\begin{aligned} \frac{\partial \ell_1(\boldsymbol{\theta}_1)}{\partial \rho}_{|\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{01}} &= \frac{1}{\sigma_0^2}\sum_{i=1}^N \sum_{t=1}^T Y_{it-1}(Y_{it} - \rho_0 Y_{it-1}) - \frac{T^2 a_0}{\sigma_0^2(\sigma_0^2 + a_0 T)} \sum_{i=1}^N \overline{Y}_{i\,-1}(\overline{Y}_i - \rho_0 \overline{Y}_{i\,-1}) \\ &= \frac{1}{\sigma_0^2}\sum_{i=1}^N \sum_{t=1}^T Y_{it-1}(\varepsilon_{it} + \alpha_i) - \frac{T^2 a_0}{\sigma_0^2(\sigma_0^2 + a_0 T)} \sum_{i=1}^N \overline{Y}_{i\,-1}(\overline{\varepsilon}_i + \alpha_i) \\ &= \frac{1}{\sigma_0^2}\sum_{i=1}^N \sum_{t=1}^T Y_{it-1}(\varepsilon_{it} - \overline{\varepsilon}_i) + \frac{T\phi^2}{\sigma_0^2} \sum_{i=1}^N \overline{Y}_{i\,-1}(\overline{\varepsilon}_i + \alpha_i), \end{aligned} \tag{3.10}$$

where $\phi^2 = \sigma_0^2/(Ta_0 + \sigma_0^2)$. Assume now that T is fixed. Inserting the estimator

[6] We employ the fact that $(\boldsymbol{I}_{m\times m} + \boldsymbol{A}\boldsymbol{A}^{\boldsymbol{T}})^{-1} = \boldsymbol{I}_{m\times m} - \boldsymbol{A}(\boldsymbol{I}_{n\times n} + \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^T$ for some $m \times n$ matrix $\boldsymbol{A}$, see Lütkepohl [60].

$$\widehat{\phi}^2 = \frac{\sum\limits_{i=1}^{N}\sum\limits_{t=1}^{T}\left[Y_{it} - \overline{Y}_i - \rho_0(Y_{it-1} - \overline{Y}_{i\,-1})\right]^2}{(T-1)T\sum\limits_{i=1}^{N}(\overline{Y}_i - \rho_0\overline{Y}_{i\,-1})^2}$$

for ϕ^2 in (3.10) yields the nonlinear first order condition

$$m_N(\rho_0) = \frac{1}{\sigma_0^2}C_N(\rho_0) + \frac{N\widehat{\sigma}^2\sum\limits_{i=1}^{N}\overline{Y}_{i\,-1}\left(\overline{Y}_i - \rho_0\overline{Y}_{i\,-1}\right)}{\sigma_0^2\sum\limits_{i=1}^{N}\left(\overline{Y}_i - \rho_0\overline{Y}_{i\,-1}\right)^2},$$

where

$$C_N(\rho_0) = \sum_{i=1}^{N}\sum_{t=1}^{T}Y_{it-1}\left[Y_{it} - \overline{Y}_i - \rho_0(Y_{it-1} - \overline{Y}_{i\,-1})\right]$$
$$\text{and}\quad \hat{\sigma}^2 = \frac{1}{(T-1)N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[Y_{it} - \overline{Y}_i - \rho_0(Y_{it-1} - \overline{Y}_{i\,-1})\right]^2.$$

Indeed, it is not difficult to see that

$$\begin{aligned} &\frac{1}{N}C_N(\rho_0) \xrightarrow{p} -b_T(\rho_0)\sigma_0^2, && \hat{\sigma}^2 \xrightarrow{p} \sigma_0^2, \\ &\frac{T}{N}\sum_{i=1}^{N}\left(\overline{Y}_i - \rho_0\overline{Y}_{i\,-1}\right)^2 \xrightarrow{p} \sigma_0^2 + Ta_0 \quad \text{and} && \frac{T}{N}\sum_{i=1}^{N}\overline{Y}_{i\,-1}\left(\overline{Y}_i - \rho_0\overline{Y}_{i\,-1}\right) \xrightarrow{p} b_T(\rho_0)(\sigma_0^2 + Ta_0), \end{aligned} \tag{3.11}$$

where $b_T(\rho_0) = T^{-1}\sum_{t=0}^{T-1}\sum_{s=0}^{t-1}\rho_0^s$. Therefore,

$$\frac{1}{N}m_N(\rho_0) \xrightarrow{p} 0. \tag{3.12}$$

It is important to note, however, that the result stated in (3.12) crucially depends on the initial condition $Y_{i0} = 0$. If, for example, the process starts at $t = -1$ such that $Y_{i0} = \alpha_i + \varepsilon_{i0}$, the last statement in (3.11) becomes

$$\frac{T}{N}\sum_{i=1}^{N}\overline{Y}_{i\,-1}\left(\overline{Y}_i - \rho_0\overline{Y}_{i\,-1}\right) \xrightarrow{p} b(\rho_0)(\sigma_0^2 + Ta_0) + \left(\sum_{t=0}^{T-1}\rho_0^t\right)a$$

and, thus, the estimator is inconsistent for such initial values. The next section provides one possible solution for this issue.

Initial condition is not zero

In this section the assumption on the starting value Y_{i0} is relaxed to accommodate models without $Y_{i0} = 0$ for all i. For example, consider the case where the process starts at an arbitrary time period $s \leq 0$ at some arbitrary initial value. Accordingly, the starting value Y_{i0} may depend on the individual effect α_i. As discussed in the last section, minimizing $\ell_1(\boldsymbol{\theta}_1)$ with respect to $\boldsymbol{\theta}_1$ does not lead to consistent estimates in such cases. In order to be able to get consistent estimates for an arbitrary initial value, we impose the following assumptions:

Assumption 3.2. The starting value Y_{i0} is unobserved with

1. $\frac{1}{N}\sum\limits_{i=1}^{N}\rho^2 Y_{i0}^2 \xrightarrow{p} a^*$.
2. $\frac{1}{N}\sum\limits_{i=1}^{N}\rho Y_{i0}\alpha_i \xrightarrow{p} \tau$.
3. ε_{it} is independent of Y_{i0} for all i and $t \geq 1$ such that $\frac{1}{N}\sum\limits_{i=1}^{N}\varepsilon_{it}Y_{i0} \xrightarrow{p} 0$ for all $t \geq 1$.

As we do not assume that $Y_{i0} = 0$ anymore we have that $Y_{i1} = \alpha_i + \rho Y_{i0} + \varepsilon_{i1}$ for all i. The model can now be written in matrix notation as

$$\boldsymbol{B}(\rho)\boldsymbol{Y}_i = \rho \boldsymbol{e}_1 Y_{i0} + \alpha_i \mathbf{1}_T + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{e}_1 = (1, 0, ..., 0)^T$. It follows now from Assumptions 3.1.2, 3.1.3 and Assumption 3.2 that

$$\boldsymbol{S}_N \xrightarrow{p} \boldsymbol{\Gamma}(\rho) \left[\mathbf{1}_T^+ \begin{pmatrix} a & \tau \\ \tau & a^* \end{pmatrix} \mathbf{1}_T^{+T} + \sigma^2 \boldsymbol{I}_T \right] \boldsymbol{\Gamma}(\rho)^T = \boldsymbol{\Sigma}(\boldsymbol{\theta}_2),$$

where $\mathbf{1}_T^+ = (\mathbf{1}_T, \boldsymbol{e}_1)$ and $\boldsymbol{\theta}_2 = (\rho, \sigma^2, a, a^*, \tau)^T$. The model implies two factors, one is attached to the individual effect α_i and the other corresponds to the initial value Y_{i0}.

The idea is to adjust the (pseudo) maximum likelihood estimator of Bai [17] and to consider the more general initial condition by substituting $\boldsymbol{\Gamma}(\rho) \left(a \mathbf{1}_T \mathbf{1}_T^T + \sigma^2 \boldsymbol{I}_{T \times T} \right) \boldsymbol{\Gamma}(\rho)^T$ in (3.6) by $\boldsymbol{\Sigma}(\boldsymbol{\theta}_2)$ such that the resulting objective function is given by

$$L_2(\boldsymbol{\theta}_2) = (2\pi)^{-\frac{NT}{2}} \left| \boldsymbol{\Sigma}(\boldsymbol{\theta}_2) \right|^{-\frac{N}{2}} \exp \left\{ -\frac{N}{2} \text{tr} \left[\boldsymbol{S}_N \boldsymbol{\Sigma}(\boldsymbol{\theta}_2)^{-1} \right] \right\}.$$

Once again we define $\ell_2(\boldsymbol{\theta}_2) = \log(L_2(\boldsymbol{\theta}_2))$. The inverse of $\boldsymbol{\Sigma}(\boldsymbol{\theta}_2)$ is a complicated object that makes it difficult to solve the first order conditions of $\ell_2(\boldsymbol{\theta}_2)$ so that we can set up a sequential iterative procedure as in section 3.2.1. In order to circumvent the problem note that

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}_2) = \boldsymbol{\Gamma}(\rho) \begin{pmatrix} \widetilde{a}^* & \widetilde{\tau} \mathbf{1}_{T-1}^T \\ \widetilde{\tau} \mathbf{1}_{T-1} & a \mathbf{1}_{T-1} \mathbf{1}_{T-1}^T + \sigma^2 \boldsymbol{I}_{T-1 \times T-1} \end{pmatrix} \boldsymbol{\Gamma}(\rho)^T = \widetilde{\boldsymbol{\Sigma}}(\widetilde{\boldsymbol{\theta}}_2),$$

where $\widetilde{a}^* = a^* + 2\tau + a + \sigma^2$, $\widetilde{\tau} = \tau + a$ and $\widetilde{\boldsymbol{\theta}}_2 = (\rho, \sigma^2, a, \widetilde{a}^*, \widetilde{\tau})^T$. In addition, we define

$$\widetilde{L}_2(\widetilde{\boldsymbol{\theta}}_2) = (2\pi)^{-\frac{NT}{2}} \left| \widetilde{\boldsymbol{\Sigma}}(\widetilde{\boldsymbol{\theta}}_2) \right|^{-\frac{N}{2}} \exp \left\{ -\frac{N}{2} \text{tr} \left[\boldsymbol{S}_N \widetilde{\boldsymbol{\Sigma}}(\widetilde{\boldsymbol{\theta}}_2)^{-1} \right] \right\}$$

and $\widetilde{\ell}_2(\widetilde{\boldsymbol{\theta}}_2) = \log \left(\widetilde{L}_2(\widetilde{\boldsymbol{\theta}}_2) \right)$. It is now easy to see that minimizing $\ell_2(\boldsymbol{\theta}_2)$ is equivalent to minimizing $\widetilde{\ell}_2(\widetilde{\boldsymbol{\theta}}_2)$. We use "equivalent" in the sense that the minimum value of $\ell_2(\boldsymbol{\theta}_2)$ is the same as the minimum value of $\widetilde{\ell}_2(\widetilde{\boldsymbol{\theta}}_2)$. As the structure of $\widetilde{\boldsymbol{\Sigma}}(\widetilde{\boldsymbol{\theta}}_2)$ is more handy than the structure of $\boldsymbol{\Sigma}(\boldsymbol{\theta}_2)$ we are now able to state the estimates $\widehat{\boldsymbol{\theta}}_2 = (\widehat{\rho}_2, \widehat{\sigma}_2^2, \widehat{a}_2, \widehat{a}_2^*, \widehat{\tau}_2)^T$ of $\boldsymbol{\theta}_2$ and $\widehat{\widetilde{\boldsymbol{\theta}}}_2 = (\widehat{\rho}_2, \widehat{\sigma}_2^2, \widehat{a}_2, \widehat{\widetilde{a}}_2^*, \widehat{\widetilde{\tau}}_2)^T$ of $\widetilde{\boldsymbol{\theta}}_2$ respectively. First, we define the following three objects:

$$\widehat{m}_\eta(\rho) = \frac{\sum_{i=1}^N \left(\boldsymbol{Y}_i^T \overline{\boldsymbol{B}}(\rho)^T \mathbf{1}_{T-1} \right)^2}{\sum_{i=1}^N Y_{i1}^2}, \qquad \widehat{m}_{\widetilde{\tau}}(\rho) = \frac{\sum_{i=1}^N \boldsymbol{Y}_i^T \overline{\boldsymbol{B}}(\rho)^T \mathbf{1}_{T-1} Y_{i1}}{\sum_{i=1}^N \left(\boldsymbol{Y}_i^T \overline{\boldsymbol{B}}(\rho)^T \mathbf{1}_{T-1} \right)^2}$$

$$\text{and} \quad \widehat{m}_\lambda(\rho) = \frac{1}{N} \sum_{i=1}^N \left(Y_{i1} - \widehat{m}_{\widetilde{\tau}}(\rho) \boldsymbol{Y}_i^T \overline{\boldsymbol{B}}(\rho)^T \mathbf{1}_{T-1} \right)^2,$$

where

$$\overline{\boldsymbol{B}}(\rho) = \begin{pmatrix} -\rho & 1 & ... & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & -\rho & 1 \end{pmatrix}$$

is a $(T-1) \times T$ matrix. It follows now from the discussion in Appendix B that $\widetilde{\ell}_2(\widetilde{\boldsymbol{\theta}}_2)$ is minimized at the following values:

$$\widehat{\rho}_2 = \frac{\sum_{i=1}^N \boldsymbol{Y}_i^T \boldsymbol{J}_{T \times T}^T \widetilde{\boldsymbol{V}}(\widetilde{\boldsymbol{\theta}}_2)^{-1} \boldsymbol{Y}_i}{\sum_{i=1}^N \boldsymbol{Y}_i^T \boldsymbol{J}_{T \times T}^T \widetilde{\boldsymbol{V}}(\widetilde{\boldsymbol{\theta}}_2)^{-1} \boldsymbol{J}_{T \times T} \boldsymbol{Y}_i}, \qquad \widehat{\sigma}_2^2 = \frac{1}{N(T-2)} \sum_{i=1}^N \boldsymbol{Y}_i^T \overline{\boldsymbol{B}}(\rho)^T \boldsymbol{Q}_{T-1 \times T-1} \overline{\boldsymbol{B}}(\rho) \boldsymbol{Y}_i,$$

$$\widehat{a}_2 = \frac{\widehat{\eta}_2 - \sigma^2}{T-1}, \qquad \widehat{\widetilde{a}}_2^* = \frac{\widehat{\lambda}_2 + \widetilde{\tau}^2 (T-1)}{\sigma^2 + (T-1)a}$$

$$\text{and} \quad \widehat{\widetilde{\tau}}_2 = \widehat{m}_{\widetilde{\tau}}(\rho) \widehat{\eta}_2,$$

where

$$\widehat{\eta}_2 = \frac{\widehat{m}_\lambda(\rho)\widehat{m}_\eta(\rho)}{(T-1)(1-\widehat{m}_{\widetilde{\tau}}(\rho)^2\widehat{m}_\eta(\rho))}, \qquad \widehat{\lambda}_2 = \frac{\widehat{m}_\lambda(\rho)^2\widehat{m}_\eta(\rho)}{(T-1)(1-\widehat{m}_{\widetilde{\tau}}(\rho)^2\widehat{m}_\eta(\rho))}$$

$$\text{and} \quad \widetilde{\boldsymbol{V}}(\widetilde{\boldsymbol{\theta}}_2) = \begin{pmatrix} \widetilde{a}^* & \widetilde{\tau}\mathbf{1}_{T-1}^T \\ \widetilde{\tau}\mathbf{1}_{T-1} & a\mathbf{1}_{T-1}\mathbf{1}_{T-1}^T + \sigma^2\boldsymbol{I}_{T-1\times T-1} \end{pmatrix}.$$

From the definitions of $\widetilde{\tau}$ and $\widetilde{a}^*$ it follows now that

$$\widehat{\tau}_2 = \widehat{\widetilde{\tau}}_2 - \widehat{a}_2 \qquad \text{and} \qquad \widehat{a_2^*} = \widehat{\widetilde{a}_2^*} - 2\widehat{\tau}_2 - \widehat{a}_2 - \widehat{\sigma}_2^2. \tag{3.13}$$

We obtain the estimate $\widehat{\boldsymbol{\theta}}_2$ now by first substituting an initial estimate of ρ into the expressions of $\widehat{\sigma}_2^2$ and $\widehat{\widetilde{\tau}}_2$ so that we get first estimates of σ^2 and $\widetilde{\tau}$. We can now obtain $\widehat{a}_2$ by substituting the estimates $\widehat{\sigma}_2^2$ and $\widehat{\eta}_2$ into the expression of $\widehat{a}_2$. Employing the estimates from the previous steps we get $\widehat{\widetilde{a}_2^*}$. In the last step, we get $\widehat{\rho}_2$ by substituting $\widehat{\sigma}_2^2$, $\widehat{a}_2$, $\widehat{\widetilde{a}_2^*}$ and $\widehat{\widetilde{\tau}}_2$ into the expression of $\widehat{\rho}_2$. This process is repeated until it converges. Finally, we get $\widehat{\tau}_2$ and $\widehat{a_2^*}$ by employing the expressions stated in (3.13).

The true value of $\boldsymbol{\theta}_2$ is denoted by $\boldsymbol{\theta}_{02}$. Therefore, we can now state the following corollary.

Corollary 3.2. *Assume that Assumption 3.1.2, 3.1.3 and Assumption 3.2 hold true. Then, invoking the results of Browne [23], Amemiya et al. [9] and Anderson and Amemiya [10] we have that*

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{02}\right) \rightsquigarrow N\left(\mathbf{0}_5, \boldsymbol{M}_3^{-1}\right)$$

when $N \to \infty$ and T is fixed, with $\boldsymbol{M}_3 = \frac{1}{N}E\left[\frac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_2^T}\right]$. In addition,

$$\sqrt{NT}\left(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{02}\right) \rightsquigarrow N\left(\mathbf{0}_5, \boldsymbol{M}_4^{-1}\right)$$

when $N \to \infty$ and $T \to \infty$, with $\boldsymbol{M}_4 = \frac{1}{NT}E\left[\frac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_2^T}\right]$.

So far we followed the approach of Bai [17] and worked with the model (3.4) in levels. However, there exists a second approach studied by Hsiao et al. [48] that considers the model in differences. We will introduce this approach in the next section.

Remark 4. Bai [17] allows for heteroscedastic error terms, i.e. ε_{it} is i.i.d. over i and independent over t but $Var[\varepsilon_{it}] = \sigma_t^2$. It is shown that for $|\rho| < 1$ the estimate of ρ is asymptotically normally distributed even if $N, T \to \infty$, with $N/T^3 \to 0$. The second result in Corollary 3.2 does not apply if $Var[\varepsilon_{it}] = \sigma_t^2$ as the number of unknown parameters increases with T. However, as the influence of the initial condition declines with $T \to \infty$ it can be expected that the results of Bai [17] still hold even if $Y_{i0} \neq 0$.

3.2.2. Maximum likelihood estimation of the model in first differences

In this section we discuss the approach of Hsiao et al. [48] that considers the model in differences instead of levels. The approach is also labeled *Transformed Likelihood Approach*. The reason for taking the differences is that the individual effects α_i are eliminated by this operation. Recall that our model of interest is given by

$$Y_{it} = \alpha_i + \rho Y_{it-1} + \varepsilon_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T.$$

Taking the first differences we get that

$$\Delta Y_{it} = \rho\Delta Y_{it-1} + \Delta\varepsilon_{it}, \quad i = 1, \ldots, N, \quad t = 2, \ldots, T, \tag{3.14}$$

where $\Delta Y_{it} = Y_{it} - Y_{it-1}$ and $\Delta\varepsilon_{it} = \varepsilon_{it} - \varepsilon_{it-1}$. It is now again crucial whether we assume that $Y_{i0} = 0$ or not.

Initial condition is zero

Consider the transformed likelihood approach of Hsiao et al. [48] under the initial condition $Y_{i0} = 0$ for all i. Therefore, the transformed system is for all i given by

$$\begin{aligned} \Delta Y_{i1} &= Y_{i1} = \alpha_i + \varepsilon_{i1} \\ \Delta Y_{it} &= \rho\Delta Y_{it-1} + \Delta\varepsilon_{it}, \quad t = 2, \ldots, T. \end{aligned} \tag{3.15}$$

Here, $\Delta Y_{i1} = Y_{i1}$ is observed as Y_{i0} is assumed to be zero. The system of equations stated in (3.15) is given in matrix notation by

$$\boldsymbol{B}(\rho)\boldsymbol{B}(1)\boldsymbol{Y}_i = \alpha_i \boldsymbol{B}(1)\boldsymbol{1}_T + \boldsymbol{B}(1)\boldsymbol{\varepsilon}_i.$$

Note that $\boldsymbol{B}(1)\boldsymbol{Y}_i = (\Delta Y_{i1}, \dots, \Delta Y_{iT})^T$ and $\boldsymbol{B}(1)\boldsymbol{1}_T = (1, 0, \dots, 0)^T$. It now holds true that

$$\begin{aligned} &\boldsymbol{B}(\rho)\boldsymbol{B}(1)\boldsymbol{Y}_i = \alpha_i \boldsymbol{B}(1)\boldsymbol{1}_T + \boldsymbol{B}(1)\boldsymbol{\varepsilon}_i \\ \Leftrightarrow\ &\boldsymbol{B}(1)\boldsymbol{B}(\rho)\boldsymbol{Y}_i = \alpha_i \boldsymbol{B}(1)\boldsymbol{1}_T + \boldsymbol{B}(1)\boldsymbol{\varepsilon}_i \\ \Leftrightarrow\ &\boldsymbol{B}(\rho)\boldsymbol{Y}_i = \alpha_i \boldsymbol{1}_T + \boldsymbol{\varepsilon}_i. \end{aligned}$$

Therefore, the transformed system is equivalent to the original system (3.5) and the approaches in levels and differences will lead to the same results.

Initial condition is not zero

In this section we discuss the transformed likelihood approach when the initial condition is not assumed to be zero. Recall that the transformed system is given by (3.14) which is a well-defined process for $t \geq 3$ but not for $t = 2$ as ΔY_{i1} is not observed in contrast to the discussion in the last section. Therefore, we need to model ΔY_{i2} in a suitable way. Let $\Delta\widetilde{\varepsilon}_{i2} = \sum_{j=0}^{m-1} \rho^j \Delta\varepsilon_{i\,2-j}$ such that by continuous substitution of (3.14)

$$\Delta Y_{i2} = \rho^m \Delta Y_{i\,-m+2} + \Delta\widetilde{\varepsilon}_{i2}, \quad i = 1, \dots, N. \tag{3.16}$$

Hsiao et al. [48] distinguishes the cases where the process in (3.16) has reached stationarity or not and, thus, states the following assumptions:

Assumption 3.3.

1. $|\rho| < 1$ and the process has been going on for a long time, i.e. $m \to \infty$, with $E[\Delta Y_{i2}] = 0$, $Var[\Delta Y_{i2}] = \frac{2\sigma^2}{1+\rho}$ and $Cov[\Delta\widetilde{\varepsilon}_{i2}, \Delta\varepsilon_{i3}] = -\sigma^2$ for all i. Finally, $Cov[\Delta\widetilde{\varepsilon}_{i2}, \Delta\varepsilon_{it}] = 0$ for all i and $t \geq 4$.
2. The process in (3.16) has started from a finite period in the past not too far back from the first observed period $t = 1$ such that $E[\Delta Y_{i2}] = b$, $Var[\Delta Y_{i2}] = c\sigma^2$, with $c > 0$, and $Cov[\Delta\widetilde{\varepsilon}_{i2}, \Delta\varepsilon_{i3}] = -\sigma^2$ for all i. Finally, $Cov[\Delta\widetilde{\varepsilon}_{i2}, \Delta\varepsilon_{it}] = 0$ for all i and $t \geq 4$.

For a further discussion of Assumption 3.3 consider Hsiao et al. [48] chapter 3. Even so Assumptions 3.3.1 and 3.3.2 are different it is possible to set up a likelihood approach that is consistent with both assumptions. In order to do so, it is assumed that

Assumption 3.4.

1. ε_{it} is i.i.d. normal over i and t with $E[\varepsilon_{it}] = 0$ and $Var[\varepsilon_{it}] = \sigma^2 > 0$.[7]

Furthermore, we define $\Delta\boldsymbol{Y}_i = (\Delta Y_{i2}, \dots, \Delta Y_{iT})^T$ and $\Delta\boldsymbol{Y}_i^- = (0, \Delta Y_{i2}, \dots, \Delta Y_{iT-1})^T$ and $\Delta\boldsymbol{\varepsilon}_i = \Delta\boldsymbol{Y}_i - \boldsymbol{e}_1 b^* - \rho\Delta\boldsymbol{Y}_i^-$, where $b^* = 0$ under Assumption 3.3.1 and $b^* = b$ under Assumption 3.3.2. The covariance matrix of $\Delta\boldsymbol{\varepsilon}_i$ is now given by

$$\boldsymbol{\Omega}(\sigma^2, \omega) = \sigma^2 \begin{pmatrix} \omega & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} = \sigma^2 \widetilde{\boldsymbol{\Omega}}(\omega).$$

Note that $\omega = 2/(1+\rho)^2$ under Assumption 3.3.1 whereas $\omega = c$ under Assumption 3.3.2. Given Assumptions 3.3 and 3.4 we state the likelihood function that we employ to estimate the parameters $\boldsymbol{\theta}_3 = (\rho, \sigma^2, b^*, \omega)^T$:

$$L_3(\boldsymbol{\theta}_3) = (2\pi)^{-\frac{NT}{2}} \mid \boldsymbol{\Omega}(\sigma^2, \omega) \mid^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} \left(\Delta\boldsymbol{Y}_i - \boldsymbol{e}_1 b^* - \rho\Delta\boldsymbol{Y}_i^-\right)^T \boldsymbol{\Omega}(\sigma^2, \omega)^{-1} \left(\Delta\boldsymbol{Y}_i - \boldsymbol{e}_1 b^* - \rho\Delta\boldsymbol{Y}_i^-\right) \right\}$$

[7] Hayakawa and Pesaran [39] consider this model assuming heteroscedasticity in the variance over individuals.

and, thus, the log-likelihood function is given by

$$\begin{aligned}\ell_3(\boldsymbol{\theta}_3) = \log(L_3(\boldsymbol{\theta}_3)) &= -\frac{NT}{2}\log(2\pi) - \frac{N}{2}\log(\mid \boldsymbol{\Omega}(\sigma^2,\omega) \mid) \\ &\quad - \frac{1}{2}\sum_{i=1}^{N}\left(\Delta\boldsymbol{Y}_i - \boldsymbol{e}_1 b^* - \rho\Delta\boldsymbol{Y}_i^-\right)^T \boldsymbol{\Omega}(\sigma^2,\omega)^{-1}\left(\Delta\boldsymbol{Y}_i - \boldsymbol{e}_1 b^* - \rho\Delta\boldsymbol{Y}_i^-\right) \\ &= -\frac{NT}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^{2(T-1)}(1+(T-1)(\omega-1))) \\ &\quad - \frac{1}{2}\sum_{i=1}^{N}\left(\Delta\boldsymbol{Y}_i - \boldsymbol{e}_1 b^* - \rho\Delta\boldsymbol{Y}_i^-\right)^T \boldsymbol{\Omega}(\sigma^2,\omega)^{-1}\left(\Delta\boldsymbol{Y}_i - \boldsymbol{e}_1 b^* - \rho\Delta\boldsymbol{Y}_i^-\right).\end{aligned}$$[8]

We are now able to state the estimates $\widehat{\boldsymbol{\theta}}_3 = (\widehat{\rho}_3, \widehat{\sigma}_3^2, \widehat{b}_3^*, \widehat{\omega}_3)^T$ of $\boldsymbol{\theta}_3$. First, we define $\Delta\boldsymbol{W}_i = (\boldsymbol{e}_1, \Delta\boldsymbol{Y}_i^-)$ and $\boldsymbol{\kappa} = (T-1, T-2, \ldots, 1)^T$. Therefore, we get that

$$\begin{aligned}(\widehat{b}_3^*, \widehat{\rho}_3)^T &= \left(\sum_{i=1}^{N}\Delta\boldsymbol{W}_i^T\widetilde{\boldsymbol{\Omega}}(\omega)^{-1}\Delta\boldsymbol{W}_i\right)^{-1}\sum_{i=1}^{N}\Delta\boldsymbol{W}_i^T\widetilde{\boldsymbol{\Omega}}(\omega)^{-1}\Delta\boldsymbol{Y}_i \\ \widehat{\sigma}_3^2 &= \frac{1}{N(T-1)}\sum_{i=1}^{N}\left(\Delta\boldsymbol{Y}_i - \boldsymbol{e}_1 b^* - \rho\Delta\boldsymbol{Y}_i^-\right)^T\widetilde{\boldsymbol{\Omega}}(\omega)^{-1}\left(\Delta\boldsymbol{Y}_i - \boldsymbol{e}_1 b^* - \rho\Delta\boldsymbol{Y}_i^-\right) \\ \widehat{\omega}_3 &= \frac{T-2}{T-1} + \frac{1}{\sigma^2 N(T-1)^2}\sum_{i=1}^{N}\left(\Delta\boldsymbol{Y}_i - \boldsymbol{e}_1 b^* - \rho\Delta\boldsymbol{Y}_i^-\right)^T\boldsymbol{\kappa}\boldsymbol{\kappa}^T\left(\Delta\boldsymbol{Y}_i - \boldsymbol{e}_1 b^* - \rho\Delta\boldsymbol{Y}_i^-\right),\end{aligned}$$

see also Hsiao et al. [48] page 144.

We obtain the estimate $\widehat{\boldsymbol{\theta}}_3$ again by a sequential iterative procedure. Therefore, we first substitute initial estimates of ρ, b^* and σ^2 into the expression of $\widehat{\omega}_3$ so that we get a first estimate of ω. We now get the estimates $\widehat{\rho}_3$ and $\widehat{b}_3^*$ by substituting $\widehat{\omega}_3$ into the expression of $(\widehat{b}_3^*, \widehat{\rho}_3)^T$. In the last step, we get $\widehat{\sigma}_3^2$ by substituting $\widehat{\omega}_3$, $\widehat{\rho}_3$ and $\widehat{b}_3^*$ into the expression of $\widehat{\sigma}_3^2$. This process is repeated until it converges.

The true value of $\boldsymbol{\theta}_3$ is denoted by $\boldsymbol{\theta}_{03}$. Therefore, we can now state the following corollary.

Corollary 3.3. *Assume that Assumption 3.4 and either Assumption 3.3.1 or 3.3.2 hold true. Then, as the likelihood function $\ell_3(\boldsymbol{\theta}_3)$ is well defined, depends on a fixed number of parameters and satisfies the usual regularity conditions*

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_3 - \boldsymbol{\theta}_{03}\right) \rightsquigarrow N\left(\mathbf{0}_4, \boldsymbol{M}_5^{-1}\right)$$

when $N \to \infty$ and T is fixed, with $\boldsymbol{M}_5 = \frac{1}{N}E\left[\frac{\partial^2\ell_3(\boldsymbol{\theta}_3)}{\partial\boldsymbol{\theta}_3\partial\boldsymbol{\theta}_3^T}\right]$. In addition,

$$\sqrt{NT}\left(\widehat{\boldsymbol{\theta}}_3 - \boldsymbol{\theta}_{03}\right) \rightsquigarrow N\left(\mathbf{0}_4, \boldsymbol{M}_6^{-1}\right)$$

when $N \to \infty$ and $T \to \infty$, with $\boldsymbol{M}_6 = \frac{1}{NT}E\left[\frac{\partial^2\ell_3(\boldsymbol{\theta}_3)}{\partial\boldsymbol{\theta}_3\partial\boldsymbol{\theta}_3^T}\right]$.[9]

3.3. Dynamic models with exogenous regressors

In this section we study the empirically more relevant case of a model with k additional exogenous variables. The exogenous regressors are comprised in the vector $\boldsymbol{X}_{it} = (X_{1,it}, \ldots, X_{k,it})^T$ such that the extended model of (3.5) is given by

$$Y_{it} = \alpha_i + \rho Y_{it-1} + \boldsymbol{\beta}^T\boldsymbol{X}_{it} + \varepsilon_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T. \tag{3.17}$$

In addition, let $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots \boldsymbol{X}_{iT})^T$. The individual effects α_i might be arbitrarily correlated with $\boldsymbol{X}_i$. This is the main difference between the fixed effects model and the random effects model. In the random effects model α_i and $\boldsymbol{X}_i$ are assumed to be independent which is in contrast to the fixed effects model.

[8] Consider Hsiao et al. [48] equation (3.7) for the derivation of $\mid \boldsymbol{\Omega}(\sigma^2,\omega) \mid$.

[9] The second derivatives of $\ell_3(\boldsymbol{\theta}_3)$ are stated in Appendix B of Hsiao et al. [48].

As in section 3.2, we compare the estimation strategy of Bai [17] that considers the model in levels with the estimation strategy of Hsiao et al. [48] that considers the model in first differences. As discussed in the last section it is crucial to model the initial value Y_{i0} to get consistent estimates. Therefore, we consider here only the case were the initial value is not assumed to be zero.

3.3.1. Maximum likelihood estimation of the model in levels

In this section we follow the ideas of Bai [17] that are related to the approaches stated in section 3.2.1. However, in contrast to the model without additional exogenous regressors we need to model the dependence between α_i and $\boldsymbol{X}_i$. In addition, we will extend the approach of Bai [17] such that it is possible to allow for initial values that are not equal to 0. If the process stated in (3.17) has the same structure for $t = 0$ we get that

$$Y_{i0} = \alpha_i + \rho Y_{i-1} + \boldsymbol{\beta}^T \boldsymbol{X}_{i0} + \varepsilon_{i0}, \quad i = 1, \dots, N.$$

Therefore, the initial condition Y_{i0} depends on α_i and is, thus, correlated with $\boldsymbol{X}_i$. This dependence needs to be modeled to get consistent estimates of the unknown parameters.

Let $\boldsymbol{Z}_i = \left(1, \boldsymbol{X}_{i1}^T, \dots, \boldsymbol{X}_{iT}^T\right)^T$ and assume that the dependence between α_i and $\boldsymbol{X}_i$ can be modeled by a Mundlak-Chamberlain projection, i.e.

$$\begin{aligned}\alpha_i &= c_0 + \boldsymbol{c}_1^T \boldsymbol{X}_{i1} + \boldsymbol{c}_2^T \boldsymbol{X}_{i2} + \dots + \boldsymbol{c}_T^T \boldsymbol{X}_{iT} + \xi_i \\ &= \boldsymbol{c}^T \boldsymbol{Z}_i + \xi_i,\end{aligned}$$

where $\boldsymbol{c} = \left(c_0, \boldsymbol{c}_1^T, \dots, \boldsymbol{c}_T^T\right)^T$, see Mundlak [64], Chamberlain [25], Chamberlain and Moreira [26] and Bai [17]. Furthermore, we assume that

$$\begin{aligned}Y_{i0} &= h_0 + \boldsymbol{h}_1^T \boldsymbol{X}_{i1} + \boldsymbol{h}_2^T \boldsymbol{X}_{i2} + \dots + \boldsymbol{h}_T^T \boldsymbol{X}_{iT} + \zeta_i \\ &= \boldsymbol{h}^T \boldsymbol{Z}_i + \zeta_i,\end{aligned}$$

where $\boldsymbol{h} = \left(h_0, \boldsymbol{h}_1^T, \dots, \boldsymbol{h}_T^T\right)^T$. In order to state the model in matrix notation we define

$$\widetilde{\boldsymbol{X}}_i = \begin{pmatrix} \boldsymbol{Z}_i^T & \boldsymbol{0}_{kT+1}^T & 0 & \boldsymbol{X}_{i1}^T \\ \boldsymbol{0}_{kT+1}^T & \boldsymbol{Z}_i^T & Y_{i1} & \boldsymbol{X}_{i2}^T \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0}_{kT+1}^T & \boldsymbol{Z}_i^T & Y_{iT-1} & \boldsymbol{X}_{iT}^T \end{pmatrix}$$

and $\boldsymbol{\gamma} = \left(\boldsymbol{f}^T, \boldsymbol{c}^T, \rho, \boldsymbol{\beta}^T\right)^T$, with $\boldsymbol{f} = \rho \boldsymbol{h} + \boldsymbol{c}$, and get that

$$\boldsymbol{Y}_i = \widetilde{\boldsymbol{X}}_i \boldsymbol{\gamma} + \rho \boldsymbol{e}_1 \zeta_i + \xi_i \boldsymbol{1}_T + \boldsymbol{\varepsilon}_i.$$

In order to be able to state the maximum likelihood function that we need to estimate the unknown model parameters we impose the following assumptions:

Assumption 3.5.

1. ε_{it} is i.i.d. normal over i and t with $E[\varepsilon_{it}] = 0$ and $Var[\varepsilon_{it}] = \sigma^2 > 0$.
2. ζ_i is i.i.d. normal over i with $E[\zeta_i] = 0$ and $Var[\rho\zeta_i] = \sigma_\zeta^2 > 0$.
3. ξ_i is i.i.d. normal over i with $E[\xi_i] = 0$ and $Var[\xi_i] = \sigma_\xi^2 > 0$.
4. $Cov[\varepsilon_{it}, \zeta_i] = Cov[\varepsilon_{it}, \xi_i] = 0$ for all i and t and $Cov[\xi_i, \rho\zeta_i] = \sigma_{\xi\zeta}$.

From the stated assumptions it follows that the covariance matrix of $\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i \boldsymbol{\gamma}$ is given by

$$\left[\boldsymbol{1}_T^+ \begin{pmatrix} \sigma_\xi^2 & \sigma_{\xi\zeta} \\ \sigma_{\xi\zeta} & \sigma_\zeta^2 \end{pmatrix} \boldsymbol{1}_T^{+T} + \sigma^2 \boldsymbol{I}_{T \times T}\right] = \boldsymbol{\Psi}(\boldsymbol{\theta}_4),$$

where $\boldsymbol{\theta}_4 = (\boldsymbol{\gamma}^T, \sigma^2, \sigma_\xi^2, \sigma_\zeta^2, \sigma_{\xi\zeta})^T$. Finally, the likelihood function is given by

$$L_4(\boldsymbol{\theta}_4) = (2\pi)^{-\frac{NT}{2}} \left|\boldsymbol{\Psi}(\boldsymbol{\theta}_4)\right|^{-\frac{N}{2}} \exp\left\{-\frac{1}{2} \sum_{i=1}^{N} \left(\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i \boldsymbol{\gamma}\right)^T \boldsymbol{\Psi}(\boldsymbol{\theta}_4)^{-1} \left(\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i \boldsymbol{\gamma}\right)\right\}$$

and $\ell_4(\boldsymbol{\theta}_4) = \log(L_4(\boldsymbol{\theta}_4))$. Recall from section 3.2.1 that the inverse of $\boldsymbol{\Psi}(\boldsymbol{\theta}_4)$ is a complicated object that makes it difficult to solve the first order conditions of $\ell_4(\boldsymbol{\theta}_4)$. In order to circumvent the problem note that

$$\boldsymbol{\Phi}(\boldsymbol{\theta}_4) = \begin{pmatrix} \widetilde{\sigma}_\zeta^2 & \widetilde{\sigma}_{\xi\zeta}\mathbf{1}_{T-1}^T \\ \widetilde{\sigma}_{\xi\zeta}\mathbf{1}_{T-1} & \sigma_\xi^2\mathbf{1}_{T-1}\mathbf{1}_{T-1}^T + \sigma^2\boldsymbol{I}_{T-1\times T-1} \end{pmatrix} = \widetilde{\boldsymbol{\Phi}}(\widetilde{\boldsymbol{\theta}}_4),$$

where $\widetilde{\sigma}_\zeta^2 = \sigma_\zeta^2 + 2\sigma_{\xi\zeta} + \sigma_\xi^2 + \sigma^2$, $\widetilde{\sigma}_{\xi\zeta} = \sigma_{\xi,\zeta} + \sigma_\xi^2$ and $\widetilde{\boldsymbol{\theta}}_4 = (\boldsymbol{\gamma}^T, \sigma^2, \sigma_\xi^2, \widetilde{\sigma}_\zeta^2, \widetilde{\sigma}_{\xi\zeta})^T$. In addition, we define

$$\widetilde{L}_4(\widetilde{\boldsymbol{\theta}}_4) = (2\pi)^{-\frac{NT}{2}} \left|\widetilde{\boldsymbol{\Phi}}(\widetilde{\boldsymbol{\theta}}_4)\right|^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^N \left(\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i\gamma\right)^T \widetilde{\boldsymbol{\Phi}}(\widetilde{\boldsymbol{\theta}}_4)^{-1}\left(\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i\gamma\right)\right\}$$

and $\widetilde{\ell}_4(\widetilde{\boldsymbol{\theta}}_4) = \log\left(\widetilde{L}_4(\widetilde{\boldsymbol{\theta}}_4)\right)$. It is now easy to see that minimizing $\ell_4(\boldsymbol{\theta}_4)$ is equivalent to minimizing $\widetilde{\ell}_4(\widetilde{\boldsymbol{\theta}}_4)$. We are now able to state the estimates $\widehat{\boldsymbol{\theta}}_4 = (\widehat{\boldsymbol{\gamma}}_4^T, \widehat{\sigma}_4^2, \widehat{\sigma}_{\xi 4}^2, \widehat{\sigma}_{\zeta 4}^2, \widehat{\sigma}_{\xi\zeta 4})^T$ of $\boldsymbol{\theta}_4$ and $\widehat{\widetilde{\boldsymbol{\theta}}}_4 = (\widehat{\boldsymbol{\gamma}}_4^T, \widehat{\sigma}_4^2, \widehat{\sigma}_{\xi 4}^2, \widehat{\widetilde{\sigma}}_{\zeta 4}^2, \widehat{\widetilde{\sigma}}_{\xi\zeta 4})^T$ of $\widetilde{\boldsymbol{\theta}}_4$ respectively. First, we define the following three objects:

$$\widehat{m}_{\eta^*}(\boldsymbol{\gamma}) = \frac{\sum_{i=1}^N \left((\boldsymbol{Y}_i^- - \widetilde{\boldsymbol{X}}_i^-\boldsymbol{\gamma})^T\mathbf{1}_{T-1}\right)^2}{\sum_{i=1}^N \left((\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\gamma})^T\boldsymbol{e}_1\right)^2}, \qquad \widehat{m}_{\widetilde{\sigma}_{\xi\zeta}}(\boldsymbol{\gamma}) = \frac{\sum_{i=1}^N (\boldsymbol{Y}_i^- - \widetilde{\boldsymbol{X}}_i^-\boldsymbol{\gamma})^T\mathbf{1}_{T-1}(\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\gamma})\boldsymbol{e}_1}{\sum_{i=1}^N \left((\boldsymbol{Y}_i^- - \widetilde{\boldsymbol{X}}_i^-\boldsymbol{\gamma})^T\mathbf{1}_{T-1}\right)^2}$$

$$\text{and} \quad \widehat{m}_{\lambda^*}(\boldsymbol{\gamma}) = \frac{1}{N}\sum_{i=1}^N \left((\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\gamma})^T\boldsymbol{e}_1 - \widehat{m}_{\widetilde{\sigma}_{\xi\zeta}}(\boldsymbol{\gamma})(\boldsymbol{Y}_i^- - \widetilde{\boldsymbol{X}}_i^-\boldsymbol{\gamma})^T\mathbf{1}_{T-1}\right)^2,$$

where $\boldsymbol{Y}_i^- = (Y_{i2}, \ldots, Y_{iT})^T$ and

$$\widetilde{\boldsymbol{X}}_i^- = \begin{pmatrix} \mathbf{0}_{kT+1} & \boldsymbol{Z}_i^T & Y_{i1} & \boldsymbol{X}_{i2}^T \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}_{kT+1} & \boldsymbol{Z}_i^T & Y_{iT-1} & \boldsymbol{X}_{iT}^T \end{pmatrix}.$$

It follows now from similar arguments as in Appendix B that $\widetilde{\ell}_4(\widetilde{\boldsymbol{\theta}}_4)$ is minimized at the following values:

$$\widehat{\gamma}_4 = \left(\sum_{i=1}^N \widetilde{\boldsymbol{X}}_i^T\widetilde{\boldsymbol{\Phi}}(\widetilde{\boldsymbol{\theta}}_4)^{-1}\widetilde{\boldsymbol{X}}_i\right)^{-1}\sum_{i=1}^N \widetilde{\boldsymbol{X}}_i^T\widetilde{\boldsymbol{\Phi}}(\widetilde{\boldsymbol{\theta}}_4)^{-1}\boldsymbol{Y}_i, \qquad \widehat{\sigma}_{\xi 4}^2 = \frac{\widehat{\eta}_4^* - \sigma^2}{T-1}$$

$$\widehat{\sigma}_4^2 = \frac{1}{N(T-2)}\sum_{i=1}^N (\boldsymbol{Y}_i^- - \widetilde{\boldsymbol{X}}_i^-\gamma)^T\boldsymbol{Q}_{T-1\times T-1}(\boldsymbol{Y}_i^- - \widetilde{\boldsymbol{X}}_i^-\gamma), \qquad \widehat{\widetilde{\sigma}}_{\zeta 4}^2 = \frac{\widehat{\lambda}_4^* + \widetilde{\sigma}_{\xi\zeta}^2(T-1)}{\sigma^2 + (T-1)\sigma_\xi^2}$$

$$\text{and} \quad \widehat{\widetilde{\sigma}}_{\xi\zeta 4} = \widehat{m}_{\widetilde{\sigma}_{\xi\zeta}}(\boldsymbol{\gamma})\widehat{\eta}_4^*,$$

where

$$\widehat{\eta}_4^* = \frac{\widehat{m}_{\lambda^*}(\boldsymbol{\gamma})\widehat{m}_{\eta^*}(\boldsymbol{\gamma})}{(T-1)(1 - \widehat{m}_{\widetilde{\sigma}_{\xi\zeta}}(\boldsymbol{\gamma})^2\widehat{m}_{\eta^*}(\boldsymbol{\gamma}))}, \qquad \widehat{\lambda}_4^* = \frac{\widehat{m}_{\lambda^*}(\boldsymbol{\gamma})^2\widehat{m}_{\eta^*}(\boldsymbol{\gamma})}{(T-1)(1 - \widehat{m}_{\widetilde{\sigma}_{\xi\zeta}}(\boldsymbol{\gamma})^2\widehat{m}_{\eta^*}(\boldsymbol{\gamma}))}.$$

From the definitions of $\widetilde{\sigma}_{\xi\zeta}$ and $\widetilde{\sigma}_\zeta^2$ it follows now that

$$\widehat{\sigma}_{\xi\zeta 4} = \widehat{\widetilde{\sigma}}_{\xi\zeta 4} - \widehat{\sigma}_{\xi 4}^2 \qquad \text{and} \qquad \widehat{\sigma}_{\zeta 4}^2 = \widehat{\widetilde{\sigma}}_{\zeta 4}^2 - 2\widehat{\sigma}_{\xi\zeta 4} - \widehat{\sigma}_{\xi 4}^2 - \widehat{\sigma}_4^2. \tag{3.18}$$

We obtain the estimate $\widehat{\boldsymbol{\theta}}_4$ now by first substituting an initial estimate of $\boldsymbol{\gamma}$ into the expressions $\widehat{\sigma}_4^2$ and $\widehat{\widetilde{\sigma}}_{\xi\zeta 4}$. With the estimates of $\boldsymbol{\gamma}$ and σ^2 we get $\widehat{\sigma}_{\xi 4}^2$ and can, thus, calculate $\widehat{\widetilde{\sigma}}_{\zeta 4}^2$. We can now obtain a new estimate of $\boldsymbol{\gamma}$. This process is repeated until it converges. Finally, we get $\widehat{\sigma}_{\xi\zeta 4}$ and $\widehat{\sigma}_{\zeta 4}^2$ by employing the expressions stated in (3.18).

The true value of $\boldsymbol{\theta}_4$ is denoted by $\boldsymbol{\theta}_{04}$. Therefore, we can now state the following corollary.

Corollary 3.4. *Assume that Assumption 3.5 holds true. Then, as the likelihood function $\ell_4(\boldsymbol{\theta}_4)$ is well-defined, depends on a fixed number of parameters and satisfies the usual regularity conditions*

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_4 - \boldsymbol{\theta}_{04}\right) \rightsquigarrow N\left(\mathbf{0}_{k(2T+1)+7}, \boldsymbol{M}_7^{-1}\right)$$

when $N \to \infty$ and T is fixed, with $\boldsymbol{M}_7 = \frac{1}{N} E\left[\frac{\partial^2 \ell_4(\boldsymbol{\theta}_4)}{\partial \boldsymbol{\theta}_4 \partial \boldsymbol{\theta}_4^T}\right]$.[10]

In contrast to the models without exogenous regressors asymptotic normality of the estimator follows without further arguments only when T is fixed. The reason is that the number of estimated model parameters increases with T.

3.3.2. Maximum likelihood estimation of the model in first differences

In this section we discuss the transformed likelihood approach when the initial condition is not assumed to be zero. The transformed system of (3.17) is given by

$$\Delta Y_{it} = \rho \Delta Y_{it-1} + \boldsymbol{\beta}^T \Delta \boldsymbol{X}_{it} + \Delta \varepsilon_{it}, \quad i = 1, \ldots, N, \quad t = 2, \ldots, T, \tag{3.19}$$

where $\Delta \boldsymbol{X}_{it} = \boldsymbol{X}_{it} - \boldsymbol{X}_{it-1}$. The process stated in (3.19) is a well-defined process for $t \geq 3$ but not for $t = 2$ as ΔY_{i1} and $\Delta \boldsymbol{X}_{i1}$ are not observed. Therefore, we need to model ΔY_{i2} in a suitable way.

Following the discussion in Hsiao et al. [48] we assume that

$$\begin{aligned}\Delta Y_{i2} &= d_0 + \boldsymbol{d}_1^T \Delta \boldsymbol{X}_{i2} + \boldsymbol{d}_2^T \Delta \boldsymbol{X}_{i3} + \cdots + \boldsymbol{d}_{T-1}^T \Delta \boldsymbol{X}_{iT} + \upsilon_{i2} \\ &= \boldsymbol{d}^T \Delta \boldsymbol{Z}_i + \upsilon_{i2},\end{aligned}$$

where $\Delta \boldsymbol{Z}_i = \left(1, \Delta \boldsymbol{X}_{i2}^T, \ldots, \Delta \boldsymbol{X}_{iT}^T\right)^T$ and $\boldsymbol{d} = \left(d_0, \boldsymbol{d}_1^T, \ldots, \boldsymbol{d}_{T-1}^T\right)^T$. In order to set up a likelihood approach that leads to consistent estimates of the unknown model parameters we state the following assumptions:

Assumption 3.6.

1. υ_{i2} is independent of $\Delta \boldsymbol{Z}_i$ for all i. In addition, $E[\upsilon_{i2}] = 0$, $Var[\upsilon_{i2}] = \sigma_\upsilon^2$ and $Cov[\upsilon_{i2}, \Delta\varepsilon_{i3}] = -\sigma^2$ for all i. Finally, $Cov[\upsilon_{i2}, \Delta\varepsilon_{it}] = 0$ for all i and $t \geq 4$.
2. ε_{it} is i.i.d. normal over i and t with $E[\varepsilon_{it}] = 0$ and $Var[\varepsilon_{it}] = \sigma^2 > 0$.

For a detailed discussion of Assumption 3.6 consider Hsiao et al. [48] chapter 4. We define $\Delta \boldsymbol{X}_i^- = (\mathbf{0}_k, \Delta \boldsymbol{X}_{i2}, \ldots, \Delta \boldsymbol{X}_{iT-1})^T$ and $\Delta \boldsymbol{\varepsilon}_i = \Delta \boldsymbol{Y}_i - \boldsymbol{e}_1 \boldsymbol{d}^T \Delta \boldsymbol{Z}_i - \rho \Delta \boldsymbol{Y}_i^- - \Delta \boldsymbol{X}_i^- \boldsymbol{\beta}$. The covariance matrix of $\Delta \boldsymbol{\varepsilon}_i$ is now given by $\boldsymbol{\Omega}(\sigma^2, \omega) = \sigma^2 \widetilde{\boldsymbol{\Omega}}(\omega)$, where $\omega = \sigma_\upsilon^2 / \sigma^2$.

Given Assumption 3.6 we state the likelihood function that we employ to estimate the parameters $\boldsymbol{\theta}_5 = (\rho, \boldsymbol{\beta}^T, \sigma^2, \boldsymbol{d}^T, \omega)^T$:

$$\begin{aligned}L_5(\boldsymbol{\theta}_5) = (2\pi)^{-\frac{NT}{2}} \mid \boldsymbol{\Omega}(\sigma^2, \omega) \mid^{-\frac{N}{2}} \exp \Bigg\{ -\frac{1}{2} \sum_{i=1}^N \left(\Delta \boldsymbol{Y}_i - \Delta \widetilde{\boldsymbol{W}}_i \left(\boldsymbol{d}^T, \rho, \boldsymbol{\beta}^T\right)^T\right)^T \\ \boldsymbol{\Omega}(\sigma^2, \omega)^{-1} \left(\Delta \boldsymbol{Y}_i - \Delta \widetilde{\boldsymbol{W}}_i \left(\boldsymbol{d}^T, \rho, \boldsymbol{\beta}^T\right)^T\right) \Bigg\},\end{aligned}$$

where

$$\Delta \widetilde{\boldsymbol{W}}_i = \begin{pmatrix} \Delta \boldsymbol{Z}_i^T & 0 & \mathbf{0}_k^T \\ \mathbf{0}_{k(T-1)+1}^T & \Delta Y_{i2} & \Delta \boldsymbol{X}_{i3}^T \\ \vdots & \vdots & \vdots \\ \mathbf{0}_{k(T-1)+1}^T & \Delta Y_{iT-1} & \Delta \boldsymbol{X}_{iT}^T \end{pmatrix}.$$

[10] The second derivatives of $\ell_4(\boldsymbol{\theta}_4)$ follow from the derivatives stated in Appendix C by replacing $\boldsymbol{J}_{T\times T}\boldsymbol{Y}_i$ by $\widetilde{\boldsymbol{X}}_i$ and $\boldsymbol{B}(\rho)\boldsymbol{Y}_i$ by $\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i \boldsymbol{\gamma}$.

Therefore, the log-likelihood function is given by

$$\begin{aligned}\ell_5(\boldsymbol{\theta}_5) = \log(L_5(\boldsymbol{\theta}_5)) &= -\frac{NT}{2}\log(2\pi) - \frac{N}{2}\log(\mid \boldsymbol{\Omega}(\sigma^2,\omega) \mid) \\ &\quad - \frac{1}{2}\sum_{i=1}^{N}\left(\Delta \boldsymbol{Y}_i - \Delta\widetilde{\boldsymbol{W}}_i\left(\boldsymbol{d}^T,\rho,\boldsymbol{\beta}^T\right)^T\right)^T \boldsymbol{\Omega}(\sigma^2,\omega)^{-1}\left(\Delta \boldsymbol{Y}_i - \Delta\widetilde{\boldsymbol{W}}_i\left(\boldsymbol{d}^T,\rho,\boldsymbol{\beta}^T\right)^T\right) \\ &= -\frac{NT}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^{2(T-1)}(1+(T-1)(\omega-1))) \\ &\quad - \frac{1}{2}\sum_{i=1}^{N}\left(\Delta \boldsymbol{Y}_i - \Delta\widetilde{\boldsymbol{W}}_i\left(\boldsymbol{d}^T,\rho,\boldsymbol{\beta}^T\right)^T\right)^T \boldsymbol{\Omega}(\sigma^2,\omega)^{-1}\left(\Delta \boldsymbol{Y}_i - \Delta\widetilde{\boldsymbol{W}}_i\left(\boldsymbol{d}^T,\rho,\boldsymbol{\beta}^T\right)^T\right).\end{aligned}$$

We are now able to state the estimates $\widehat{\boldsymbol{\theta}}_5 = (\widehat{\rho}_5, \widehat{\boldsymbol{\beta}}_5^T, \widehat{\sigma}_5^2, \widehat{\boldsymbol{d}}_5^T, \widehat{\omega}_5)^T$ of $\boldsymbol{\theta}_5$. We get that

$$\begin{aligned}\left(\widehat{\boldsymbol{d}}_5^T, \widehat{\rho}_5, \widehat{\boldsymbol{\beta}}_5^T\right)^T &= \left(\sum_{i=1}^{N}\Delta\widetilde{\boldsymbol{W}}_i^T\widetilde{\boldsymbol{\Omega}}(\omega)^{-1}\Delta\widetilde{\boldsymbol{W}}_i\right)^{-1}\sum_{i=1}^{N}\Delta\widetilde{\boldsymbol{W}}_i^T\widetilde{\boldsymbol{\Omega}}(\omega)^{-1}\Delta\boldsymbol{Y}_i \\ \widehat{\sigma}_5^2 &= \frac{1}{N(T-1)}\sum_{i=1}^{N}\left(\Delta \boldsymbol{Y}_i - \Delta\widetilde{\boldsymbol{W}}_i\left(\boldsymbol{d}^T,\rho,\boldsymbol{\beta}^T\right)^T\right)^T \widetilde{\boldsymbol{\Omega}}(\omega)^{-1}\left(\Delta \boldsymbol{Y}_i - \Delta\widetilde{\boldsymbol{W}}_i\left(\boldsymbol{d}^T,\rho,\boldsymbol{\beta}^T\right)^T\right) \\ \widehat{\omega}_5 &= \frac{T-2}{T-1} + \frac{1}{\sigma^2 N(T-1)^2}\sum_{i=1}^{N}\left(\Delta \boldsymbol{Y}_i - \Delta\widetilde{\boldsymbol{W}}_i\left(\boldsymbol{d}^T,\rho,\boldsymbol{\beta}^T\right)^T\right)^T \boldsymbol{\kappa}\boldsymbol{\kappa}^T\left(\Delta \boldsymbol{Y}_i - \Delta\widetilde{\boldsymbol{W}}_i\left(\boldsymbol{d}^T,\rho,\boldsymbol{\beta}^T\right)^T\right),\end{aligned}$$

see also Hsiao et al. [48] page 144.

We obtain the estimate $\widehat{\boldsymbol{\theta}}_5$ again by a sequential iterative procedure. Therefore, we first substitute initial estimates of ρ, $\boldsymbol{\beta}$, $\boldsymbol{d}$ and σ^2 into the expression of $\widehat{\omega}_5$ so that we get a first estimate of ω. We now get the estimates $\widehat{\rho}_5$, $\widehat{\boldsymbol{\beta}}_5$ and $\widehat{\boldsymbol{d}}_5$ by substituting $\widehat{\omega}_5$ into the expression of $\left(\widehat{\boldsymbol{d}}_5^T, \widehat{\rho}_5, \widehat{\boldsymbol{\beta}}_5^T\right)^T$. In the last step, we get $\widehat{\sigma}_5^2$ by substituting $\widehat{\rho}_5$, $\widehat{\boldsymbol{\beta}}_5$, $\widehat{\boldsymbol{d}}_5$ and $\widehat{\omega}_5$ into the expression of $\widehat{\sigma}_5^2$. This process is repeated until it converges.

The true value of $\boldsymbol{\theta}_5$ is denoted by $\boldsymbol{\theta}_{05}$. Therefore, we can now state the following corollary.

Corollary 3.5. *Assume that Assumption 3.6 holds true. Then, as the likelihood function $\ell_5(\boldsymbol{\theta}_5)$ is well-defined, depends on a fixed number of parameters and satisfies the usual regularity conditions*

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_5 - \boldsymbol{\theta}_{05}\right) \rightsquigarrow N\left(\boldsymbol{0}_{kT+4}, \boldsymbol{M}_8^{-1}\right)$$

when $N \to \infty$ and T is fixed, with $\boldsymbol{M}_8 = \frac{1}{N}E\left[\frac{\partial^2\ell_5(\boldsymbol{\theta}_5)}{\partial\boldsymbol{\theta}_5\partial\boldsymbol{\theta}_5^T}\right]$.[11]

As in the last section asymptotic normality of the estimator follows without imposing further arguments only when T is fixed. The reason is that the number of estimated model parameters increases with T. We end the discussion here and conduct in the next section some Monte Carlo simulations.

3.4. Monte Carlo study

In this section we consider the small sample behavior of the discussed estimators with exogenous regressors. We conduct several simulation experiments to consider bias and standard deviation for the estimated parameters. In addition, we state size and power of hypothesis tests. We begin with a consideration of the simulation setup that follows the setup in Hsiao et al. [48] and finally state our simulation results.

3.4.1. Simulation design

Recall that the model with one exogenous variable X_{it} is given by

$$Y_{it} = \eta_i + \rho Y_{i,t-1} + \beta X_{it} + \varepsilon_{it}.$$

[11] The second derivatives of $\ell_5(\boldsymbol{\theta}_5)$ are stated in Appendix B of Hsiao et al. [48].

We consider two different distributions for the error terms ε_{it}. On the one hand we consider $\varepsilon_{it} \sim N(0,1)$ independent and identical across i and t. On the other hand we generate ε_{it} as

$$\varepsilon_{it} = \frac{1}{\sqrt{4}} \left(\zeta_{1it}^2 + \zeta_{1it}^2 - 2\right), \tag{3.20}$$

where $\zeta_{jit} \sim N(0,1)$ independent over i, t and j. When ε_{it} follows the process in (3.20) the normal assumption of the error terms for the maximum likelihood estimators in section 3.3 is not met. Therefore, we can get an idea how crucial the normal assumption is.

The regressor X_{it} is generated according to

$$X_{it} = \mu_i + 0.01t + \xi_{it},$$

where $\xi_{it} = 0.5\xi_{it-1} + u_{it} + 0.5u_{it-1}$ with $u_{it} \sim N\left(0, \sigma_u^2\right)$. We set $\xi_{i-49} = 0$ and $u_{i-49} = 0$. This ensures that the process is not to much influenced by the initial value. Due to the same reason we set $Y_{i-49} = 0$, i.e. we discard the first 50 observations such that Y_{i1} is the first observed value.

Finally, the fixed effects η_i and μ_i are generated by

$$\begin{aligned}
\mu_i &= e_{1i} + \frac{1}{T+50} \sum_{t=-49}^{T} u_{it}, \quad e_{1i} \sim N\left(0, 1 - \sigma_u^2/(T+50)\right) \\
\eta_i &= e_{2i} + \frac{1}{T+50} \sum_{t=-49}^{T} \Delta X_{it}, \quad e_{2i} \sim N\left(0, 1 - 3\sigma_u^2/\left(3(T+50)\right)\right).
\end{aligned}$$

The way we generate μ_i and η_i ensures that the individual effects are correlated with the exogenous regressors such that a random effects specification would not lead to consistent parameter estimates.

During the simulation, we consider four different models. The models are given by

Model 1: $\rho = 0.4$, $\beta = 0.6$ and $\sigma_u^2 = 0.8$. In addition, $\varepsilon_{it} \sim N(0,1)$.
Model 2: $\rho = 0.8$, $\beta = 0.2$ and $\sigma_u^2 = 1.875$. In addition, $\varepsilon_{it} \sim N(0,1)$.
Model 3: $\rho = 0.4$, $\beta = 0.6$ and $\sigma_u^2 = 0.8$. In addition, ε_{it} follows the process in (3.20).
Model 4: $\rho = 0.8$, $\beta = 0.2$ and $\sigma_u^2 = 1.875$. In addition, ε_{it} follows the process in (3.20).

We compare the maximum likelihood estimators with two existing estimators. First, we consider the IV estimator that is obtained by employing the model in differences, see equation (3.19), and then using Y_{it-2} and ΔX_{it} as instruments. Denote the estimates by $\widehat{\rho}_{IV}$ and $\widehat{\beta}_{IV}$ and let

$$\boldsymbol{W}_i = \begin{pmatrix} \Delta y_{i2} & \Delta y_{i3} & \dots & \Delta y_{iT-1} \\ \Delta x_{i3} & \Delta x_{i4} & \dots & \Delta x_{iT} \end{pmatrix}^T \quad \text{and} \quad \boldsymbol{S}_i = \begin{pmatrix} y_{i1} & y_{i2} & \dots & y_{iT-2} \\ \Delta x_{i3} & \Delta x_{i4} & \dots & \Delta x_{iT} \end{pmatrix}^T.$$

With $\boldsymbol{W} = \left(\boldsymbol{W}_1^T, \dots, \boldsymbol{W}_N^T\right)^T$ and $\boldsymbol{S} = \left(\boldsymbol{S}_1^T, \dots, \boldsymbol{S}_N^T\right)^T$ the estimated covariance of the IV estimator is given by[12]

$$\widehat{\boldsymbol{Var}}_{IV} = \widehat{\sigma}_{IV}^2 \left(\boldsymbol{S}^T \boldsymbol{W}\right)^{-1} \left(\boldsymbol{S}^T \left(\boldsymbol{I}_{N\times N} \otimes \boldsymbol{V}\right) \boldsymbol{S}\right) \left(\boldsymbol{W}^T \boldsymbol{S}\right)^{-1}, \tag{3.21}$$

where $\widehat{\sigma}_{IV}^2 = \frac{1}{2N(T-2)} \sum_{i=1}^{N} \sum_{t=3}^{T} \left(\Delta Y_{it} - \widehat{\rho}_{IV} \Delta Y_{it-1} - \widehat{\beta}_{IV} \Delta X_{it}\right)^2$ and

$$\boldsymbol{V} = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}.$$

Note that $\boldsymbol{V}$ has dimension $(T-2) \times (T-2)$.

As additional benchmark we compute the GMM estimator suggested by Arellano and Bond [14] where the full set of instruments $(Y_{i1}, \dots, Y_{it-2}, X_{i1}, X_{i2}, \dots, X_{iT})$ is used for each of the time periods. In order

[12] Here, ⊗ denotes the Kronecker product.

to state this GMM estimator we define

$$\boldsymbol{D}_i = \begin{pmatrix} \boldsymbol{\delta}_{i1} & \boldsymbol{0}_{T+2}^T & \dots & \boldsymbol{0}_{2T-2}^T \\ \boldsymbol{0}_{T+1}^T & \boldsymbol{\delta}_{i2} & & \vdots \\ \vdots & & \ddots & \boldsymbol{0}_{2T-2}^T \\ \boldsymbol{0}_{T+1}^T & \dots & \boldsymbol{0}_{2T-3}^T & \boldsymbol{\delta}_{iT-2} \end{pmatrix},$$

with $\boldsymbol{\delta}_{il} = (Y_{i1}, \dots, Y_{il}, X_{i1}, \dots, X_{iT})$. The GMM estimator of Arellano and Bond [14] is now given by

$$\left(\widehat{\rho}_{GMM}, \widehat{\beta}_{GMM}\right)^T = \left(\left[\sum_{i=1}^N \boldsymbol{W}_i^T \boldsymbol{D}_i\right] \left[\sum_{i=1}^N \boldsymbol{D}_i^T \boldsymbol{V} \boldsymbol{D}_i\right]^{-1} \left[\sum_{i=1}^N \boldsymbol{D}_i^T \boldsymbol{W}_i\right] \right)^{-1}$$
$$\left(\left[\sum_{i=1}^N \boldsymbol{W}_i^T \boldsymbol{D}_i\right] \left[\sum_{i=1}^N \boldsymbol{D}_i^T \boldsymbol{V} \boldsymbol{D}_i\right]^{-1} \left[\sum_{i=1}^N \boldsymbol{D}_i^T \Delta \boldsymbol{Y}_i^+\right] \right),$$

where $\Delta \boldsymbol{Y}_i^+ = (\Delta Y_{i3}, \dots, \Delta Y_{iT})^T$. The corresponding covariance estimator is given by

$$\widehat{\boldsymbol{Var}}_{GMM} = \widehat{\sigma}_{GMM}^2 \left(\left[\sum_{i=1}^N \boldsymbol{W}_i^T \boldsymbol{D}_i\right] \left[\sum_{i=1}^N \boldsymbol{D}_i^T \boldsymbol{V} \boldsymbol{D}_i\right]^{-1} \left[\sum_{i=1}^N \boldsymbol{D}_i^T \boldsymbol{W}_i\right] \right)^{-1}, \tag{3.22}$$

where $\widehat{\sigma}_{GMM}^2 = \frac{1}{2N(T-2)} \sum_{i=1}^N \sum_{t=3}^T \left(\Delta Y_{it} - \widehat{\rho}_{GMM} \Delta Y_{it-1} - \widehat{\beta}_{GMM} \Delta X_{it}\right)^2$.

We consider bias and standard deviation of the estimators. In addition, we test by a simple Z-Test whether the estimated parameters are significantly different from the true value. In order to get an estimate for the variances of the estimators we employ (3.21) for the IV estimator, (3.22) for the GMM estimator, the second derivatives stated in Appendix C for the maximum likelihood estimator in levels and the second derivatives stated in Appendix B of Hsiao et al. [48] for the maximum likelihood estimator in differences.

For the sequential iterative procedure of the maximum likelihood estimators we need some starting values. For the estimator in levels we set $\boldsymbol{f}_{in} = \boldsymbol{0}_{kT+1}$, $\boldsymbol{c}_{in} = \boldsymbol{0}_{kT+1}$, $\boldsymbol{\rho}_{in} = \widehat{\rho}_{GMM}$ and $\boldsymbol{\beta}_{in} = \widehat{\beta}_{GMM}$, where the suffix *in* denotes the starting value. In addition, we consider for the estimator in differences $\boldsymbol{d}_{in} = \boldsymbol{0}_{k(T-1)+1}$, $\boldsymbol{\rho}_{in} = \widehat{\rho}_{GMM}$, $\boldsymbol{\beta}_{in} = \widehat{\beta}_{GMM}$ and $\sigma_{in}^2 = \widehat{\sigma}_{GMM}^2$.

Results are stated for $T \in \{5, 15, 25\}$ and $N \in \{100, 200, 500\}$. In all experiments 2000 replications are used.

3.4.2. Simulation results

Table 3.1 states the results for bias and standard deviation for ρ and β in Model 2. The IV estimator is biased when T and N are small and the bias is more pronounced when estimating ρ. However, the results improve for increasing sample size. The GMM estimator performs better than the IV estimator with some bias for small N when estimating ρ. Nevertheless, both maximum likelihood estimators (MLE) outperform both mentioned estimators and deliver comparable results. The main difference occurs for small T where the estimator in levels outperforms the estimator in differences. Table 3.2 states the results for bias and standard deviation for ρ and β in Model 3. The IV estimator is here not so severely biased when T and N are small in relation to Model 2. In contrast, the GMM estimator has still some bias for small N when estimating ρ even though the bias seems to be smaller. Once again, both maximum likelihood estimators outperform the IV and the GMM estimator and give almost identical results. The maximum likelihood estimator in differences performs here also comparable to the estimator in levels when T is small in contrast to the results for Model 2.

Table 3.3 states the empirical level for the Z-Tests for ρ and β for Model 1. Both maximum likelihood estimators get close to the nominal level no matter if we test ρ or β. The same holds true for the IV estimator. The GMM estimator performs quite bad when testing for ρ. The null hypothesis is rejected too often. The results improve when N increases but get worse when T increases for fixed N. Table 3.4 states the empirical level for the Z-Tests for ρ and β for Model 4. The IV estimator gets close to the nominal level no matter if we test for ρ or β. In contrast, both maximum likelihood estimators reject too often for small sample sizes when testing for ρ. The results improve for increasing sample size. Recall that the error terms are not normally distributed in Model 4. Therefore, the results might show that we need a higher sample size to get close to the nominal level when the assumption on the error distribution is not met. Note that the IV estimator does not rely on a normal assumption. Again, the GMM estimator performs quite bad when testing for ρ. The null hypothesis is rejected too often. The results improve when N increases but get worse when T increases for fixed N.

Figures 3.1 and 3.2 state the power functions of the Z-Tests for β in Model 2 with $N = 200$, $T = 5$ and $N = 500$, $T = 5$. For both sample sizes the power functions for all estimators are symmetric. The maximum likelihood estimators and the GMM estimator deliver comparable results and outperform the IV estimator. Figures 3.3 and 3.4 state the power functions of the Z-Tests for ρ in Model 3 with $N = 100$, $T = 5$ and $N = 100$, $T = 25$. In Figure 3.3 the power functions of the maximum likelihood estimators are symmetric, attain their minimum at 0.4 and are almost identical. In contrast, the power function of the GMM estimator is shifted to the left and reaches the nominal level at 0.35, i.e. the estimator is biased. The power function of the IV estimator is quite flat and is outperformed by the maximum likelihood estimators. In Figure 3.4 the values of the power functions have increased at all points. In addition, the power function of the GMM estimator is not shifted any more but the power function does not attain the nominal level at 0.4. As before, the maximum likelihood estimators outperform the IV estimator. The main conclusion from Figures 3.3 and 3.4 is that the maximum likelihood estimators give convincing results for both sample sizes and that the power improves if the sample size increases. In addition, all figures show that testing for ρ is the statistically more difficult task than testing for β.

Table 3.1: *Bias and Standard Deviation of the estimators for ρ and β in Model 2.*

	T	Bias			St. dev.		
N		100	200	500	100	200	500
ρ estimator							
IV	5	0.9341	0.0075	0.018	24.98	1.441	0.193
	15	−0.0019	0.0017	−0.0009	0.095	0.065	0.041
	25	−0.0007	−0.001	0.0002	0.056	0.041	0.026
GMM	5	−0.0811	−0.0425	−0.0181	0.097	0.071	0.045
	15	−0.0399	−0.0213	−0.009	0.023	0.017	0.011
	25	−0.0297	−0.0168	−0.007	0.014	0.01	0.006
MLE in levels	5	0.0066	0.0005	−0.0021	0.108	0.068	0.04
	15	−0.0014	−0.0003	−0.0003	0.02	0.014	0.008
	25	−0.0006	−0.0003	0.0002	0.012	0.008	0.005
MLE in differences	5	0.0154	0.0074	0.0003	0.119	0.078	0.044
	15	−0.0011	0.0001	−0.0001	0.02	0.015	0.008
	25	−0.0006	−0.0001	−0.0001	0.012	0.008	0.005
β estimator							
IV	5	0.0696	−0.0014	0.0016	1.91	0.119	0.019
	15	0.0001	−0.0004	0.0001	0.017	0.012	0.007
	25	−0.0003	−0.0001	0.0001	0.012	0.008	0.005
GMM	5	−0.0011	−0.0018	0.0001	0.028	0.02	0.012
	15	0.0031	0.0013	0.0007	0.011	0.008	0.005
	25	0.0042	0.0021	0.0012	0.007	0.006	0.003
MLE in levels	5	0.0014	−0.0006	0.0005	0.028	0.02	0.012
	15	−0.0001	−0.0003	0.0001	0.011	0.008	0.004
	25	0.0003	−0.0001	0.0002	0.007	0.005	0.003
MLE in differences	5	0.0018	−0.0004	0.0005	0.029	0.02	0.012
	15	−0.0001	−0.0004	0.0001	0.011	0.008	0.004
	25	0.0003	−0.0001	0.0002	0.007	0.005	0.003

Notes: For all simulations 2000 Monte Carlo samples were used.

Table 3.2: *Bias and Standard Deviation of the estimators for ρ and β in Model 3.*

	T	Bias			St. dev.		
N		100	200	500	100	200	500
ρ estimator							
IV	5	0.0058	0.0027	0.0004	0.164	0.108	0.067
	15	−0.0003	−0.0001	0.0007	0.05	0.036	0.023
	25	−0.0012	0.0001	−0.0001	0.036	0.026	0.016
GMM	5	−0.0415	−0.0219	−0.0089	0.079	0.057	0.036
	15	−0.0246	−0.0131	−0.0048	0.025	0.018	0.011
	25	−0.0213	−0.0117	−0.0047	0.017	0.012	0.008
MLE in levels	5	−0.0002	0.0002	0.0002	0.069	0.05	0.031
	15	−0.0011	−0.0005	0.0002	0.023	0.016	0.009
	25	−0.0006	−0.0006	−0.0001	0.016	0.011	0.007
MLE in differences	5	0.0001	0.0003	0.0002	0.07	0.051	0.032
	15	−0.0012	−0.0005	0.0002	0.023	0.016	0.009
	25	−0.0008	−0.0006	−0.0001	0.016	0.011	0.007
β estimator							
IV	5	−0.0037	0.002	−0.001	0.089	0.059	0.037
	15	0.0013	0.0008	−0.0005	0.04	0.028	0.018
	25	−0.0006	0.0008	−0.0001	0.029	0.021	0.013
GMM	5	0.0042	0.0043	0.0007	0.07	0.048	0.031
	15	0.0096	0.0056	0.0018	0.027	0.019	0.012
	25	0.0094	0.0053	0.0021	0.019	0.014	0.009
MLE in levels	5	−0.0019	0.0012	−0.0007	0.07	0.048	0.031
	15	0.0001	0.0004	−0.0002	0.027	0.019	0.012
	25	−0.0001	0.0002	−0.0001	0.019	0.014	0.009
MLE in differences	5	−0.0019	0.0012	−0.0007	0.07	0.048	0.031
	15	0.0002	0.0004	−0.0002	0.027	0.019	0.012
	25	0.0001	0.0002	−0.0001	0.019	0.014	0.009

Notes: For all simulations 2000 Monte Carlo samples were used.

Table 3.3: *Empirical Level for Z-Tests of the estimators for ρ and β in Model 1.*

	T	5% level			10% level		
N		100	200	500	100	200	500
ρ estimator							
IV	5	5.45	5.05	5.4	9.65	10.65	9.95
	15	4.35	5.9	5.1	9.6	10.7	9.9
	25	4.65	5.55	4.65	10.3	10.75	9.75
GMM	5	9.6	7.85	6.35	16.35	13.7	11.6
	15	17.7	12.55	6.85	26.4	20.85	12.35
	25	24.3	16.15	9.1	34.35	25.3	15.45
MLE in levels	5	5.25	4.55	5.4	10.35	9.15	10.95
	15	4.8	4.85	4.2	10.05	10.5	8.8
	25	4.15	4.7	4.55	9.0	10.6	9.85
MLE in differences	5	5.4	4.55	5.35	9.8	9.05	10.6
	15	4.85	4.8	4.25	10.3	10.7	9.15
	25	6.2	5.55	4.55	11.0	10.5	9.9
β estimator							
IV	5	4.75	5.1	4.95	10.0	10.75	9.95
	15	4.7	4.15	4.55	9.45	9.15	9.9
	25	5.4	5.15	5.0	10.05	9.95	9.75
GMM	5	6.0	4.95	5.35	12.2	10.75	9.9
	15	6.45	6.45	6.1	12.4	11.4	10.4
	25	8.55	6.2	6.0	15.4	11.95	11.1
MLE in levels	5	5.55	5.0	4.85	11.1	10.1	10.05
	15	4.8	4.85	5.4	9.45	9.6	10.45
	25	6.15	4.05	5.25	10.85	9.35	9.75
MLE in differences	5	5.7	5.05	4.85	11.25	10.05	10.1
	15	4.65	4.8	5.4	9.4	9.65	10.5
	25	6.2	4.05	5.3	10.8	9.35	9.75

Notes: For all simulations 2000 Monte Carlo samples were used.

Table 3.4: *Empirical Level for Z-Tests of the estimators for ρ and β in Model 4.*

	T	5% level			10% level		
N		100	200	500	100	200	500
ρ estimator							
IV	5	4.9	6.1	4.35	8.4	10.1	8.35
	15	4.85	4.75	4.55	8.75	10.0	9.95
	25	4.0	5.3	4.4	7.5	10.9	9.3
GMM	5	17.75	10.6	7.25	25.45	16.75	13.1
	15	43.7	26.9	14.4	54.95	38.25	22.4
	25	62.2	38.4	18.5	72.7	49.9	28.8
MLE in levels	5	12.3	8.0	6.55	16.8	13.15	12.2
	15	5.9	4.85	5.55	10.3	10.45	10.2
	25	5.8	5.6	4.85	11.1	10.4	9.5
MLE in differences	5	14.1	9.7	7.85	18.2	15.2	13.05
	15	5.5	5.0	5.55	11.2	10.5	10.5
	25	6.3	5.5	4.7	11.5	10.4	9.7
β estimator							
IV	5	4.65	5.35	4.85	9.6	10.4	10.95
	15	4.45	5.15	4.8	8.94	9.75	10.3
	25	4.2	5.1	5.5	9.3	11.1	10.5
GMM	5	5.7	5.05	5.1	10.75	10.7	10.5
	15	6.3	7.2	4.55	12.15	13.5	9.9
	25	8.4	6.3	5.4	15.4	11.7	9.85
MLE in levels	5	5.6	4.6	4.8	10.3	10.2	10.0
	15	4.8	5.85	4.4	9.75	12.3	9.5
	25	4.3	4.4	4.1	9.5	9.9	9.1
MLE in differences	5	5.3	4.75	4.65	10.2	10.0	10.2
	15	4.85	6.0	4.45	9.75	12.3	9.4
	25	4.5	4.4	4.1	9.7	9.8	9.0

Notes: For all simulations 2000 Monte Carlo samples were used.

Figure 3.1: *Power function of β in Model 2 with $N = 200$ and $T = 5$.*

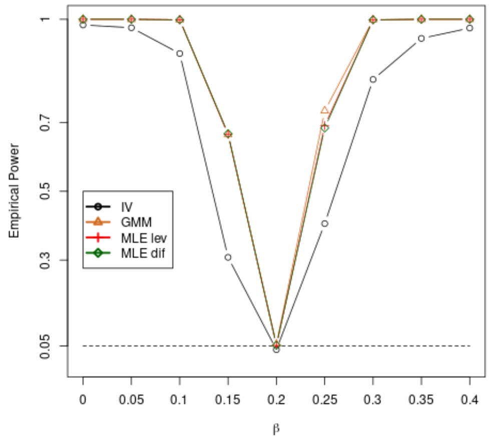
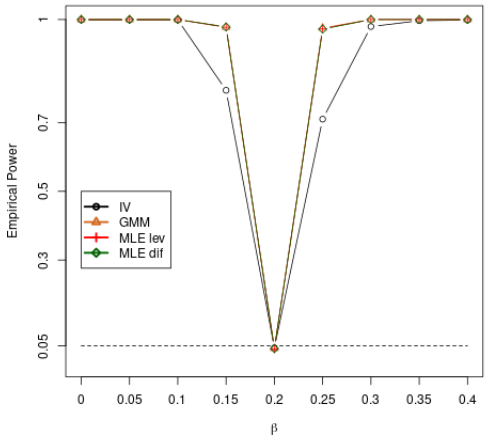


Figure 3.2: *Power function of β in Model 2 with $N = 500$ and $T = 5$.*



Notes: For all simulations 2000 Monte Carlo samples were used. The nominal level is 5%.

Figure 3.3: *Power function of ρ in Model 3 with $N = 100$ and $T = 5$.*

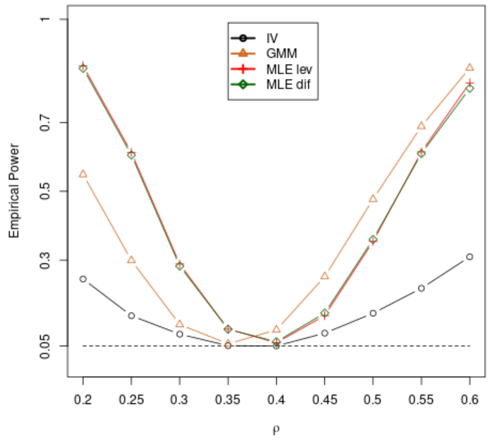
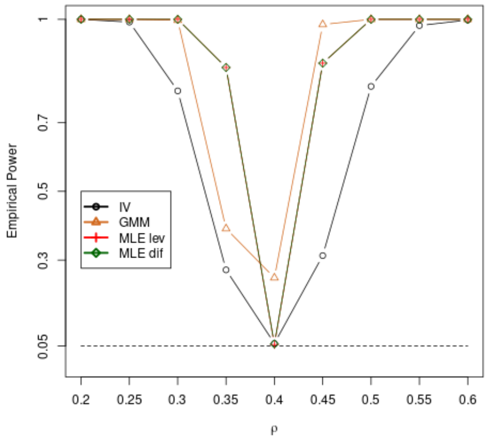


Figure 3.4: *Power function of ρ in Model 3 with $N = 100$ and $T = 25$.*



Notes: For all simulations 2000 Monte Carlo samples were used. The nominal level is 5%.

3.5. Conclusion

In this paper we discussed dynamic panel data models in the presence of incidental parameters for individuals. The transformed maximum likelihood approach was compared with a factor analytical approach and for the model without additional covariates consistent estimators under mild conditions on the initial value were proposed. In addition, we extended the factor analytical approach to models with additional covariates. This estimator controls for the initial value under mild conditions. Monte Carlo results have also been conducted were the transformed maximum likelihood approach, the extended factor analytical approach, a GMM estimator and an IV estimator are compared.

Appendix

Appendix A: Derivatives of the log-likelihood function l_1

Following Maddala [61] we get that

$$\left(a\mathbf{1}_T\mathbf{1}_T^T + \sigma^2\boldsymbol{I}_{T\times T}\right)^{-1} = \frac{1}{\sigma^2}\left(\boldsymbol{Q}_{T\times T} + \frac{\sigma^2}{T(\sigma^2+Ta)}\mathbf{1}_T\mathbf{1}_T^T\right).$$

Therefore, the likelihood function can be stated as

$$\begin{aligned}\ell_1(\boldsymbol{\theta}_1) =& -\frac{NT}{2}\log(2\pi) - \frac{N(T-1)}{2}\log(\sigma^2) - \frac{N}{2}\log\left(\sigma^2+Ta\right)\\ &-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{Q}_{T\times T}\boldsymbol{B}(\rho)\boldsymbol{Y}_i - \frac{1}{2T(\sigma^2+Ta)}\sum_{i=1}^{N}\left(\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\mathbf{1}_T\right)^2.\end{aligned}$$

In addition, recall that $\boldsymbol{B}(\rho) = \boldsymbol{I}_{T\times T} - \rho\boldsymbol{J}_{T\times T}$. With this representation of the likelihood function we can now consider the first and second derivatives. The first derivatives are given by

$$\begin{aligned}\frac{\partial\ell_1(\boldsymbol{\theta}_1)}{\partial\rho} &= \sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{J}_{T\times T}^T\left(a\mathbf{1}_T\mathbf{1}_T^T+\sigma^2\boldsymbol{I}_{T\times T}\right)^{-1}\left(\boldsymbol{I}_{T\times T}-\rho\boldsymbol{J}_{T\times T}\right)\boldsymbol{Y}_i,\\ \frac{\partial\ell_1(\boldsymbol{\theta}_1)}{\partial a} &= -\frac{NT}{2(\sigma^2+Ta)} + \frac{1}{2(\sigma^2+Ta)^2}\sum_{i=1}^{N}\left(\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\mathbf{1}_T\right)^2,\\ \frac{\partial\ell_1(\boldsymbol{\theta}_1)}{\partial\sigma^2} &= -\frac{N(T-1)}{2\sigma^2} - \frac{N}{2(\sigma^2+Ta)} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{Q}_{T\times T}\boldsymbol{B}(\rho)\boldsymbol{Y}_i + \frac{1}{2T(\sigma^2+Ta)^2}\sum_{i=1}^{N}\left(\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\mathbf{1}_T\right)^2.\end{aligned}$$

Finally, the second derivatives are given by

$$\begin{aligned}\frac{\partial^2\ell_1(\boldsymbol{\theta}_1)}{\partial\rho^2} &= -\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{J}_{T\times T}^T\left(a\mathbf{1}_T\mathbf{1}_T^T+\sigma^2\boldsymbol{I}_{T\times T}\right)^{-1}\boldsymbol{J}_{T\times T}\boldsymbol{Y}_i,\\ \frac{\partial^2\ell_1(\boldsymbol{\theta}_1)}{\partial a^2} &= \frac{NT^2}{2(\sigma^2+Ta)^2} - \frac{T}{(\sigma^2+Ta)^3}\sum_{i=1}^{N}\left(\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\mathbf{1}_T\right)^2,\\ \frac{\partial^2\ell_1(\boldsymbol{\theta}_1)}{\partial a\partial\rho} &= -\frac{1}{(\sigma^2+Ta)^2}\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\mathbf{1}_T\boldsymbol{Y}_i^T\boldsymbol{J}_{T\times T}^T\mathbf{1}_T,\\ \frac{\partial^2\ell_1(\boldsymbol{\theta}_1)}{(\partial\sigma^2)^2} &= \frac{N(T-1)}{2\sigma^4} + \frac{N}{2(\sigma^2+Ta)^2} - \frac{1}{\sigma^6}\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{Q}_{T\times T}\boldsymbol{B}(\rho)\boldsymbol{Y}_i - \frac{1}{T(\sigma^2+Ta)^3}\sum_{i=1}^{N}\left(\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\mathbf{1}_T\right)^2,\\ \frac{\partial^2\ell_1(\boldsymbol{\theta}_1)}{\partial\sigma^2\partial\rho} &= -\frac{1}{\sigma^4}\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{J}_{T\times T}^T\boldsymbol{Q}_{T\times T}\boldsymbol{B}(\rho)\boldsymbol{Y}_i - \frac{1}{T(\sigma^2+Ta)^2}\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\mathbf{1}_T\boldsymbol{Y}_i^T\boldsymbol{J}_{T\times T}^T\mathbf{1}_T,\\ \frac{\partial^2\ell_1(\boldsymbol{\theta}_1)}{\partial\sigma^2\partial a} &= \frac{NT}{2(\sigma^2+Ta)^2} - \frac{1}{(\sigma^2+Ta)^3}\sum_{i=1}^{N}\left(\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\mathbf{1}_T\right)^2.\end{aligned}$$

Appendix B: Derivation of the estimates of the log-likelihood function l_2

Let $\lambda = \widetilde{a}^*(\sigma^2 + (T-1)a) - \widetilde{\tau}^2(T-1)$ and $\eta = \sigma^2 + (T-1)a$. We get that

$$\begin{aligned}
\left|\widetilde{\boldsymbol{\Sigma}}(\widetilde{\boldsymbol{\theta}}_2)\right| &= \left|\begin{pmatrix} \widetilde{a}^* & \widetilde{\tau}\mathbf{1}_{T-1}^T \\ \widetilde{\tau}\mathbf{1}_{T-1} & a\mathbf{1}_{T-1}\mathbf{1}_{T-1}^T + \sigma^2\boldsymbol{I}_{T-1\times T-1} \end{pmatrix}\right| \\
&= \left|a\mathbf{1}_{T-1}\mathbf{1}_{T-1}^T + \sigma^2\boldsymbol{I}_{T-1\times T-1}\right| \left(\widetilde{a}^* - \widetilde{\tau}^2\mathbf{1}_{T-1}^T\left(a\mathbf{1}_{T-1}\mathbf{1}_{T-1}^T + \sigma^2\boldsymbol{I}_{T-1\times T-1}\right)^{-1}\mathbf{1}_{T-1}\right) \\
&= \sigma^{2(T-2)}(\sigma^2 + (T-1)a)\left(\widetilde{a}^* - \widetilde{\tau}^2\frac{T-1}{\sigma^2 + (T-1)a}\right) \\
&= \sigma^{2(T-2)}\left(\widetilde{a}^*(\sigma^2 + (T-1)a) - \widetilde{\tau}^2(T-1)\right) \\
&= \sigma^{2(T-2)}\lambda,
\end{aligned}$$

see Bhargava and Sargan [20] page 1642. In addition, we have that

$$\begin{pmatrix} \widetilde{a}^* & \widetilde{\tau}\mathbf{1}_{T-1}^T \\ \widetilde{\tau}\mathbf{1}_{T-1} & a\mathbf{1}_{T-1}\mathbf{1}_{T-1}^T + \sigma^2\boldsymbol{I}_{T-1\times T-1} \end{pmatrix}^{-1} = \begin{pmatrix} \frac{\eta}{\lambda} & -\frac{\widetilde{\tau}}{\lambda}\mathbf{1}_{T-1}^T \\ -\frac{\widetilde{\tau}}{\lambda}\mathbf{1}_{T-1} & \frac{1}{\sigma^2}\boldsymbol{Q}_{T-1\times T-1} + \frac{\lambda+\widetilde{\tau}^2(T-1)}{\lambda\eta(T-1)}\mathbf{1}_{T-1}\mathbf{1}_{T-1}^T \end{pmatrix} = \breve{\boldsymbol{V}}(\breve{\boldsymbol{\theta}}_2)^{-1},$$

with $\breve{\boldsymbol{\theta}}_2 = (\rho, \sigma^2, \lambda, \eta, \widetilde{\tau})^T$. Following the discussion in Appendix A of Bhargava and Sargan [20] we define

$$\begin{aligned}
\breve{\ell}_2(\breve{\boldsymbol{\theta}}_2) = &-\frac{NT}{2}\log(2\pi) - \frac{N(T-2)}{2}\log(\sigma^2) - \frac{N}{2}\log(\lambda) \\
&- \frac{1}{2}\sum_{i=1}^{N}\boldsymbol{Y}_i^T\left(\boldsymbol{I}_{T\times T} - \rho\boldsymbol{J}_{T\times T}\right)^T\breve{\boldsymbol{V}}(\breve{\boldsymbol{\theta}}_2)^{-1}\left(\boldsymbol{I}_{T\times T} - \rho\boldsymbol{J}_{T\times T}\right)\boldsymbol{Y}_i.
\end{aligned}$$

It is now easy to see that minimizing $\widetilde{\ell}_2(\widetilde{\boldsymbol{\theta}}_2)$ is equivalent to minimizing $\breve{\ell}_2(\breve{\boldsymbol{\theta}}_2)$. From the first order conditions of $\breve{\ell}_2(\breve{\boldsymbol{\theta}}_2)$ we get the estimate $\widehat{\breve{\boldsymbol{\theta}}}_2 = (\widehat{\rho}_2, \widehat{\sigma}_2^2, \widehat{\eta}_2, \widehat{\lambda}_2, \widehat{\widetilde{\tau}}_2)^T$ of $\breve{\boldsymbol{\theta}}_2$:

$$\begin{aligned}
\widehat{\rho}_2 &= \frac{\sum_{i=1}^N \boldsymbol{Y}_i^T\boldsymbol{J}_{T\times T}^T\breve{\boldsymbol{V}}(\breve{\boldsymbol{\theta}}_2)^{-1}\boldsymbol{Y}_i}{\sum_{i=1}^N \boldsymbol{Y}_i^T\boldsymbol{J}_{T\times T}^T\breve{\boldsymbol{V}}(\breve{\boldsymbol{\theta}}_2)^{-1}\boldsymbol{J}_{T\times T}\boldsymbol{Y}_i}, \\
\widehat{\sigma}_2^2 &= \frac{1}{N(T-2)}\sum_{i=1}^{N}\boldsymbol{Y}_i^T\overline{\boldsymbol{B}}(\rho)^T\boldsymbol{Q}_{T-1\times T-1}\overline{\boldsymbol{B}}(\rho)\boldsymbol{Y}_i, \\
\widehat{\eta}_2 &= \frac{\widehat{m}_\lambda(\rho)\widehat{m}_\eta(\rho)}{(T-1)(1 - \widehat{m}_{\widetilde{\tau}}(\rho)^2\widehat{m}_\eta(\rho))}, \\
\widehat{\lambda}_2 &= \widehat{m}_\lambda(\rho)\eta, \\
\text{and}\quad \widehat{\widetilde{\tau}}_2 &= \widehat{m}_{\widetilde{\tau}}(\rho)\eta.
\end{aligned}$$

From the definitions of λ and η it follows now that

$$\widehat{a}_2 = \frac{\widehat{\eta}_2 - \widehat{\sigma}_2^2}{T-1} \quad \text{and} \quad \widehat{\widetilde{a}}_2^* = \frac{\widehat{\lambda}_2 + \widehat{\widetilde{\tau}}_2^2(T-1)}{\widehat{\sigma}_2^2 + (T-1)\widehat{a}_2}.$$

Appendix C: Derivatives of the log-likelihood function l_2

It follows from the discussion in Appendix B that

$$\begin{aligned}\ell_2(\boldsymbol{\theta}_2) = &-\frac{NT}{2}\log(2\pi) - \frac{N(T-2)}{2}\log(\sigma^2) - \frac{N}{2}\log\big((a^*+2\tau+a+\sigma^2)(\sigma^2+(T-1)a)-(\tau+a)^2(T-1)\big)\\ &-\frac{1}{2}\sum_{i=1}^{N}\boldsymbol{Y}_i^T\left(\boldsymbol{I}_{T\times T}-\rho\boldsymbol{J}_{T\times T}\right)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\left(\boldsymbol{I}_{T\times T}-\rho\boldsymbol{J}_{T\times T}\right)\boldsymbol{Y}_i,\end{aligned}$$

where

$$\boldsymbol{V}(\boldsymbol{\theta}_2) = \left[\boldsymbol{1}_T^+\begin{pmatrix} a & \tau\\ \tau & a^*\end{pmatrix}\boldsymbol{1}_T^{+T}+\sigma^2\boldsymbol{I}_{T\times T}\right].$$

In addition, recall that $\boldsymbol{B}(\rho)=\boldsymbol{I}_{T\times T}-\rho\boldsymbol{J}_{T\times T}$, $\lambda=\widetilde{a}^*(\sigma^2+(T-1)a)-\widetilde{\tau}^2(T-1)$ and $\eta=\sigma^2+(T-1)a$. With this representation of the likelihood function we can now consider the first and second derivatives of $\ell_2(\boldsymbol{\theta}_2)$. The first derivatives are given by

$$\begin{aligned}\frac{\partial\ell_2(\boldsymbol{\theta}_2)}{\partial\rho} &= \sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{J}_{T\times T}^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{B}(\rho)\boldsymbol{Y}_i,\\ \frac{\partial\ell_2(\boldsymbol{\theta}_2)}{\partial a^*} &= -\frac{N}{2\lambda}\eta+\frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1\right)^2,\\ \frac{\partial\ell_2(\boldsymbol{\theta}_2)}{\partial\sigma^2} &= -\frac{N(T-2)}{2\sigma^2}-\frac{N}{2\lambda}(\widetilde{a}^*+\eta)+\frac{1}{2}\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{B}(\rho)\boldsymbol{Y}_i,\\ \frac{\partial\ell_2(\boldsymbol{\theta}_2)}{\partial a} &= -\frac{N}{2\lambda}(\widetilde{a}^*(T-1)+\eta-2\widetilde{\tau}(T-1))+\frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{1}_T\right)^2,\\ \frac{\partial\ell_2(\boldsymbol{\theta}_2)}{\partial\tau} &= -\frac{N}{\lambda}(\eta-\widetilde{\tau}(T-1))+\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{1}_T\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1.\text{[13]}\end{aligned}$$

Finally, the second derivatives are given by

$$\begin{aligned}\frac{\partial^2\ell_2(\boldsymbol{\theta}_2)}{\partial\rho^2} &= -\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{J}_{T\times T}^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{J}_{T\times T}\boldsymbol{Y}_i,\\ \frac{\partial^2\ell_2(\boldsymbol{\theta}_2)}{\partial\rho\partial a^*} &= -\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{J}_{T\times T}^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1\boldsymbol{e}_1^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{B}(\rho)\boldsymbol{Y}_i,\\ \frac{\partial^2\ell_2(\boldsymbol{\theta}_2)}{\partial\rho\partial\sigma^2} &= -\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{J}_{T\times T}^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{B}(\rho)\boldsymbol{Y}_i,\\ \frac{\partial^2\ell_2(\boldsymbol{\theta}_2)}{\partial\rho\partial a} &= -\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{J}_{T\times T}^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{1}_T\boldsymbol{1}_T^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{B}(\rho)\boldsymbol{Y}_i,\\ \frac{\partial^2\ell_2(\boldsymbol{\theta}_2)}{\partial\rho\partial\tau} &= -\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{J}_{T\times T}^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\left(\boldsymbol{1}_T\boldsymbol{e}_1^T+\boldsymbol{e}_1\boldsymbol{1}_T^T\right)\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{B}(\rho)\boldsymbol{Y}_i,\\ \frac{\partial^2\ell_2(\boldsymbol{\theta}_2)}{(\partial a^*)^2} &= \frac{N}{2\lambda^2}\eta^2-\sum_{i=1}^{N}\left(\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1\right)^2\boldsymbol{e}_1^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1,\\ \frac{\partial^2\ell_2(\boldsymbol{\theta}_2)}{\partial a^*\partial\sigma^2} &= \frac{N}{2\lambda^2}\eta\left(\widetilde{a}^*+\eta\right)-\frac{N}{2\lambda}-\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1,\end{aligned}$$

[13] We used that $\frac{\partial\boldsymbol{A}(t)^{-1}}{\partial t}=-\boldsymbol{A}(t)^{-1}\frac{\partial\boldsymbol{A}(t)}{\partial t}\boldsymbol{A}(t)^{-1}$ for some symmetric matrix $\boldsymbol{A}$ depending on t.

$$\begin{aligned}
\frac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{\partial a^* \partial a} &= \frac{N}{2\lambda^2}\eta\left(\widetilde{a}^*(T-1)+\eta-2\widetilde{\tau}(T-1)\right)-\frac{N(T-1)}{2\lambda} \\
&\quad -\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\mathbf{1}_T\mathbf{1}_T^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1, \\
\frac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{\partial a^* \partial \tau} &= \frac{N}{\lambda^2}\eta\left(\eta-\widetilde{\tau}(T-1)\right)-\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\left(\boldsymbol{e}_1\mathbf{1}_T^T+\mathbf{1}_T\boldsymbol{e}_1^T\right)\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1, \\
\frac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{(\partial \sigma^2)^2} &= \frac{N(T-2)}{2\sigma^4}+\frac{N}{2\lambda^2}(\widetilde{a}^*+\eta)^2-\frac{N}{\lambda}-\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{B}(\rho)\boldsymbol{Y}_i, \\
\frac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{\partial \sigma^2 \partial a} &= \frac{N}{2\lambda^2}(\widetilde{a}^*+\eta)(\widetilde{a}^*(T-1)+\eta-2\widetilde{\tau}(T-1))-\frac{NT}{2\lambda} \\
&\quad -\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\mathbf{1}_T\mathbf{1}_T^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{B}(\rho)\boldsymbol{Y}_i, \\
\frac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{\partial \sigma^2 \partial \tau} &= \frac{N}{\lambda^2}(\widetilde{a}^*+\eta)(\eta-\widetilde{\tau}(T-1))-\frac{N}{\lambda}-\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\left(\boldsymbol{e}_1\mathbf{1}_T^T+\mathbf{1}_T\boldsymbol{e}_1^T\right)\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{B}(\rho)\boldsymbol{Y}_i, \\
\frac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{\partial a^2} &= \frac{N}{2\lambda^2}(\widetilde{a}^*(T-1)+\eta-2\widetilde{\tau}(T-1))^2-\sum_{i=1}^{N}\left(\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\mathbf{1}_T\right)^2\mathbf{1}_T^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\mathbf{1}_T, \\
\frac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{\partial a \partial \tau} &= \frac{N}{\lambda^2}(\widetilde{a}^*(T-1)+\eta-2\widetilde{\tau}(T-1))(\eta-\widetilde{\tau}(T-1))-\sum_{i=1}^{N}\left(\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\mathbf{1}_T\right)^2\boldsymbol{e}_1^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\mathbf{1}_T \\
&\quad -\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\mathbf{1}_T\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1\mathbf{1}_T^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\mathbf{1}_T, \\
\frac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{\partial \tau^2} &= \frac{2N}{\lambda^2}(\eta-\widetilde{\tau}(T-1))^2+\frac{N(T-1)}{\lambda}-2\sum_{i=1}^{N}\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\mathbf{1}_T\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1\boldsymbol{e}_1^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\mathbf{1}_T \\
&\quad -\sum_{i=1}^{N}\left(\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\mathbf{1}_T\right)^2\boldsymbol{e}_1^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1-\sum_{i=1}^{N}\left(\boldsymbol{Y}_i^T\boldsymbol{B}(\rho)^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{e}_1\right)^2\mathbf{1}_T^T\boldsymbol{V}(\boldsymbol{\theta}_2)^{-1}\mathbf{1}_T.
\end{aligned}$$

Table 3.D.5: *Bias and Standard Deviation of the estimators for ρ and β in Model 1.*

	T	Bias			St. dev.		
N		100	200	500	100	200	500
ρ estimator							
IV	5	0.0028	−0.0016	0.0008	0.154	0.108	0.068
	15	−0.0005	−0.0003	−0.0001	0.05	0.036	0.022
	25	0.0005	0.0001	0.0002	0.036	0.026	0.016
GMM	5	−0.0421	−0.0219	−0.0086	0.078	0.057	0.037
	15	−0.0244	−0.0134	−0.0056	0.025	0.019	0.011
	25	−0.0216	−0.0113	−0.0048	0.017	0.013	0.008
MLE in levels	5	0.0019	−0.0001	0.0001	0.069	0.048	0.03
	15	−0.0007	−0.0005	−0.0005	0.023	0.016	0.01
	25	−0.0007	−0.0003	−0.0002	0.016	0.012	0.007
MLE in differences	5	0.0024	0.0001	−0.0001	0.069	0.048	0.03
	15	−0.0007	−0.0005	−0.0005	0.023	0.016	0.01
	25	−0.0008	−0.0002	−0.0001	0.016	0.012	0.007
β estimator							
IV	5	−0.0003	0.0008	−0.0001	0.086	0.061	0.038
	15	0.0005	−0.0009	−0.0001	0.04	0.028	0.018
	25	−0.0006	0.0001	−0.0001	0.029	0.021	0.013
GMM	5	0.008	0.004	0.0011	0.071	0.048	0.03
	15	0.0101	0.005	0.0023	0.027	0.019	0.012
	25	0.0098	0.0052	0.0022	0.02	0.014	0.009
MLE in levels	5	0.0018	0.0009	−0.0002	0.071	0.048	0.03
	15	0.0005	−0.0002	0.0002	0.026	0.019	0.012
	25	0.0001	0.0001	0.0001	0.02	0.014	0.009
MLE in differences	5	0.0018	0.0009	−0.0001	0.071	0.048	0.03
	15	0.0005	−0.0002	0.0002	0.026	0.019	0.012
	25	0.0002	0.0001	0.0001	0.02	0.014	0.009

Notes: For all simulations 2000 Monte Carlo samples were used.

Table 3.D.6: *Bias and Standard Deviation of the estimators for ρ and β in Model 4.*

	T	Bias			St. dev.		
N		100	200	500	100	200	500
ρ estimator							
IV	5	0.1476	0.246	0.0047	11.63	7.13	0.179
	15	−0.0023	−0.0004	−0.0011	0.094	0.065	0.041
	25	−0.0003	0.0005	−0.0001	0.054	0.041	0.025
GMM	5	−0.0836	−0.042	−0.016	0.098	0.068	0.045
	15	−0.0394	−0.0221	−0.0098	0.023	0.016	0.011
	25	−0.03	−0.0166	−0.0072	0.014	0.01	0.007
MLE in levels	5	0.0045	−0.0016	−0.001	0.114	0.071	0.043
	15	−0.0007	−0.0009	−0.0006	0.021	0.014	0.009
	25	−0.0009	0.0001	−0.0001	0.013	0.009	0.005
MLE in differences	5	0.0128	0.0048	0.0008	0.126	0.083	0.049
	15	−0.0003	−0.0005	−0.0003	0.021	0.014	0.009
	25	−0.001	0.0002	0.0001	0.013	0.009	0.005
β estimator							
IV	5	0.0132	0.0182	−0.0005	0.335	0.506	0.018
	15	−0.0004	0.0001	−0.0001	0.016	0.012	0.007
	25	−0.0001	−0.0002	−0.0001	0.012	0.009	0.005
GMM	5	−0.0031	−0.0007	−0.0007	0.028	0.02	0.013
	15	0.003	0.0017	0.0005	0.011	0.008	0.005
	25	0.004	0.0019	0.0009	0.008	0.006	0.003
MLE in levels	5	−0.0005	0.0003	−0.0003	0.029	0.02	0.013
	15	−0.0001	0.0001	−0.0001	0.011	0.008	0.005
	25	0.0001	−0.0003	−0.0001	0.008	0.006	0.003
MLE in differences	5	−0.0002	0.0005	−0.0003	0.029	0.02	0.013
	15	−0.0001	0.0001	−0.0002	0.011	0.008	0.005
	25	0.0002	−0.0003	−0.0001	0.008	0.006	0.003

Notes: For all simulations 2000 Monte Carlo samples were used.

Table 3.D.7: *Empirical Level for Z-Tests of the estimators for ρ and β in Model 2.*

	T	5% level			10% level		
N		100	200	500	100	200	500
ρ estimator							
IV	5	5.35	5.15	4.6	9.0	8.4	8.85
	15	4.2	4.4	5.3	9.3	8.75	10.45
	25	4.9	5.6	6.4	9.5	11.2	10.5
GMM	5	18.05	11.35	6.9	27.05	18.1	12.25
	15	43.85	25.8	12.2	56.3	36.45	21.55
	25	61.1	38.7	19.1	71.8	51.8	28.75
MLE in levels	5	10.95	6.7	4.85	16.35	11.45	8.55
	15	5.75	5.7	4.95	10.8	11.75	9.0
	25	5.3	4.7	5.0	10.3	9.5	9.85
MLE in differences	5	11.7	7.05	4.85	16.65	10.8	9.25
	15	5.1	5.75	4.95	10.4	11.65	8.9
	25	5.7	4.8	4.6	10.9	9.7	9.65
β estimator							
IV	5	3.85	3.95	4.45	7.05	8.35	8.8
	15	5.0	5.1	5.2	10.35	10.0	9.1
	25	5.2	4.3	4.4	10.1	9.5	8.4
GMM	5	6.05	5.6	4.3	10.4	10.9	9.65
	15	6.3	5.3	6.1	12.9	10.15	11.5
	25	8.7	6.8	6.6	14.7	11.7	11.25
MLE in levels	5	5.25	5.2	4.15	9.6	10.25	9.6
	15	5.15	4.7	5.85	10.1	9.55	11.2
	25	4.5	4.5	4.9	9.3	9.5	9.6
MLE in differences	5	5.15	5.2	4.2	9.85	10.4	9.7
	15	5.3	4.65	5.75	10.2	9.8	11.35
	25	4.6	4.4	4.9	9.3	9.5	9.5

Notes: For all simulations 2000 Monte Carlo samples were used.

Table 3.D.8: *Empirical Level for Z-Tests of the estimators for ρ and β in Model 3.*

	T	5% level			10% level		
N		100	200	500	100	200	500
ρ estimator							
IV	5	5.05	5.3	5.35	11.4	10.5	10.2
	15	4.25	4.6	4.9	10.6	9.5	10.55
	25	5.75	5.55	4.8	10.9	11.2	10.05
GMM	5	9.7	7.45	6.95	16.4	13.45	12.1
	15	16.9	10.7	6.2	27.4	18.4	12.5
	25	24.85	16.5	8.95	35.2	25.4	15.3
MLE in levels	5	5.95	6.5	6.15	10.9	12.25	11.55
	15	4.95	5.7	4.45	9.95	10.4	9.5
	25	5.45	6.1	5.25	10.5	10.7	10.7
MLE in differences	5	6.25	6.45	6.3	10.6	12.3	12.05
	15	4.95	5.65	4.6	9.7	10.25	9.55
	25	5.6	6.2	5.25	10.6	10.8	10.55
β estimator							
IV	5	5.95	5.25	5.0	11.25	9.95	10.2
	15	5.2	4.65	5.75	10.05	10.05	10.2
	25	5.1	5.7	4.95	9.95	11.2	9.75
GMM	5	6.6	5.0	5.2	11.2	9.85	10.85
	15	6.65	7.1	5.0	12.95	12.55	9.95
	25	7.65	7.7	5.9	13.35	13.3	10.6
MLE in levels	5	5.9	4.7	5.2	11.35	9.6	10.7
	15	5.2	5.1	4.4	9.2	11.0	9.55
	25	4.2	5.9	5.3	9.05	10.9	9.85
MLE in differences	5	5.95	4.6	5.25	11.35	9.55	10.7
	15	5.25	5.25	4.5	9.35	11.15	9.55
	25	4.35	5.9	5.35	9.0	10.8	9.8

Notes: For all simulations 2000 Monte Carlo samples were used.

Figure 3.D.5: *Power function of β in Model 1 with $N = 100$ and $T = 15$.*

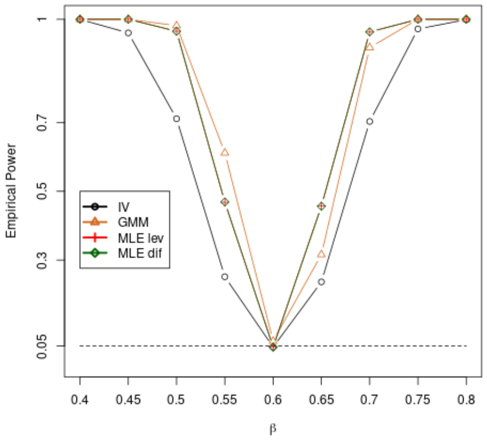
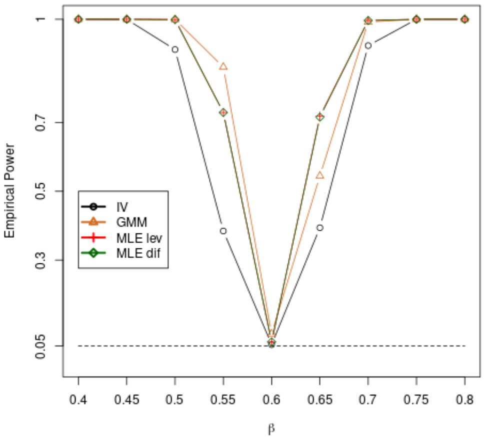


Figure 3.D.6: *Power function of β in Model 1 with $N = 100$ and $T = 25$.*



Notes: For all simulations 2000 Monte Carlo samples were used. The nominal level is 5%.

Figure 3.D.7: *Power function of ρ in Model 4 with $N = 200$ and $T = 5$.*

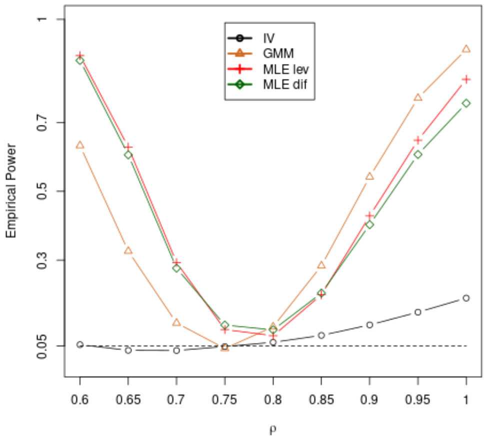
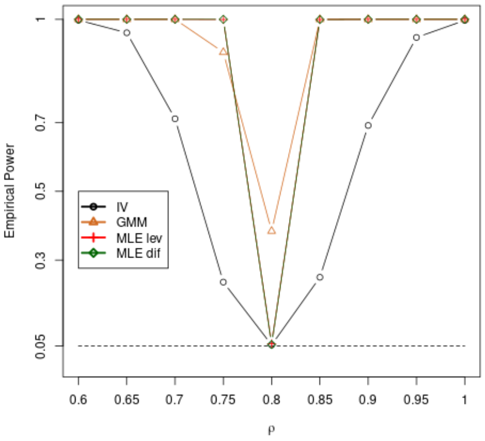


Figure 3.D.8: *Power function of ρ in Model 4 with $N = 200$ and $T = 25$.*



Notes: For all simulations 2000 Monte Carlo samples were used. The nominal level is 5%.

Bibliography

[1] Acemoglu, D., Johnson, S., Robinson, J. A., 2001. The colonial origins of comparative development: An empirical investigation. American Economic Review 91 (5), 1369–1401.

[2] Ahn, S. C., Schmidt, P., 1995. Efficient estimation of models for dynamic panel data. Journal of Econometrics 68 (1), 5–27.

[3] Ai, C., Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 71 (6), 1795–1843.

[4] Altonji, J. G., Bharadwaj, P., Lange, F., 2012. Changes in the characteristics of American youth: Implications for adult outcomes. Journal of Labor Economics 30 (4), 783–828.

[5] Alvarez, J., Arellano, M., 2003. The time series and cross-section asymptotics of dynamic panel data estimators. Econometrica 71 (4), 1121–1159.

[6] Amemiya, T., 1985. Advanced econometrics. Harvard University Press.

[7] Amemiya, T., MaCurdy, T. E., 1986. Instrumental-variable estimation of an error-components model. Econometrica 54, 869–880.

[8] Amemiya, T., Powell, J. L., 1981. A comparison of the Box-Cox maximum likelihood estimator and the non-linear two-stage least squares estimator. Journal of Econometrics 17 (3), 351–381.

[9] Amemiya, Y., Fuller, W. A., Pantula, S. G., 1987. The asymptotic distributions of some estimators for a factor analysis model. Journal of Multivariate Analysis 22 (1), 51–64.

[10] Anderson, T. W., Amemiya, Y., 1988. The asymptotic normal distribution of estimators in factor analysis under general conditions. The Annals of Statistics 16 (2), 759–771.

[11] Anderson, T. W., Hsiao, C., 1981. Estimation of dynamic models with error components. Journal of the American Statistical Association 76 (375), 598–606.

[12] Anderson, T. W., Hsiao, C., 1982. Formulation and estimation of dynamic models using panel data. Journal of Econometrics 18 (1), 47–82.

[13] Antoine, B., Bonnal, H., Renault, E., 2007. On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood. Journal of Econometrics 138 (2), 461–487.

[14] Arellano, M., Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. The Review of Economic Studies 58 (2), 277–297.

[15] Arellano, M., Bover, O., 1995. Another look at the instrumental variable estimation of error-components models. Journal of Econometrics 68 (1), 29–51.

[16] Autor, D. H., Handel, M. J., 2013. Putting tasks to the test: Human capital, job tasks, and wages. Journal of Labor Economics 31 (S1), 59–96.

[17] Bai, J., 2013. Fixed-effects dynamic panel models, a factor analytical method. Econometrica 81 (1), 285–314.

[18] Bekker, P. A., 1994. Alternative approximations to the distributions of instrumental variable estimators. Econometrica 62, 657–681.

[19] Berndt, E. R., Showalter, M. H., Wooldridge, J. M., 1993. An empirical investigation of the Box-Cox model and a nonlinear least squares alternative. Econometric Reviews 12 (1), 65–102.

[20] Bhargava, A., Sargan, J. D., 1983. Estimating dynamic random effects models from panel data covering short time periods. Econometrica 51, 1635–1659.

[21] Blundell, R., Bond, S., 1998. Initial conditions and moment restrictions in dynamic panel data models. Journal of Econometrics 87 (1), 115–143.

[22] Box, G. E., Cox, D. R., 1964. An analysis of transformations. Journal of the Royal Statistical Society: Series B 26 (2), 211–243.

[23] Browne, M. W., 1974. Generalized least squares estimators in the analysis of covariance structures. South African Statistical Journal 8 (1), 1–24.

[24] Carrasco, M., Florens, J.-P., 2000. Generalization of GMM to a continuum of moment conditions. Econometric Theory 16 (6), 797–834.

[25] Chamberlain, G., 1984. Panel data. Handbook of Econometrics 2, 1247–1318.

[26] Chamberlain, G., Moreira, M. J., 2009. Decision theory applied to a linear panel data model. Econometrica 77 (1), 107–133.

[27] Dell, M., Jones, B. F., Olken, B. A., 2009. Temperature and income: Reconciling new cross-sectional and panel estimates. American Economic Review 99 (2), 198–204.

[28] Deming, D. J., 2017. The growing importance of social skills in the labor market. The Quarterly Journal of Economics 132 (4), 1593–1640.

[29] Dominguez, M. A., Lobato, I. N., 2004. Consistent estimation of models defined by conditional moment restrictions. Econometrica 72 (5), 1601–1615.

[30] Donald, S. G., Imbens, G. W., Newey, W. K., 2003. Empirical likelihood estimation and consistent tests with conditional moment restrictions. Journal of Econometrics 117 (1), 55–93.

[31] Engle, R. F., Granger, C. W., Rice, J., Weiss, A., 1986. Semiparametric estimates of the relation between weather and electricity sales. Journal of the American Statistical Association 81 (394), 310–320.

[32] Foster, A., Tian, L., Wei, L., 2001. Estimation for the Box-Cox transformation model without assuming parametric error distribution. Journal of the American Statistical Association 96 (455), 1097–1101.

[33] Greene, W. H., 2003. Econometric analysis. Pearson Education.

[34] Hahn, J., Kuersteiner, G., 2002. Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. Econometrica 70 (4), 1639–1657.

[35] Han, C., Phillips, P. C., 2010. GMM estimation for dynamic panels with fixed effects and strong instruments at unity. Econometric Theory 26 (01), 119–151.

[36] Han, C., Phillips, P. C., 2013. First difference maximum likelihood and dynamic panel estimation. Journal of Econometrics 175 (1), 35–45.

[37] Härdle, W., Liang, H., Gao, J., 2000. Partially linear models. Springer Science & Business Media.

[38] Hayakawa, K., 2007. Small sample bias properties of the system GMM estimator in dynamic panel data models. Economics Letters 95 (1), 32–38.

[39] Hayakawa, K., Pesaran, M. H., 2015. Robust standard errors in transformed likelihood estimation of dynamic panel data models with cross-sectional heteroskedasticity. Journal of Econometrics 188 (1), 111–134.

[40] Heckman, J., Polachek, S., 1974. Empirical evidence on the functional form of the earnings-schooling relationship. Journal of the American Statistical Association 69 (346), 350–354.

[41] Heckman, J., Stixrud, J., Urzua, S., 2006. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. Journal of Labor Economics 24 (3), 411–482.

[42] Heckman, N. E., 1986. Spline smoothing in a partly linear model. Journal of the Royal Statistical Society: Series B 48 (2), 244–248.

[43] Hjort, N. L., McKeague, I. W., Van Keilegom, I., 2009. Extending the scope of empirical likelihood. The Annals of Statistics 37 (3), 1079–1111.

[44] Horn, R., Johnson, C., 1994. Topics in matrix analysis. Cambridge University Press.

[45] Horowitz, J. L., 1996. Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. Econometrica 64 (1), 103–137.

[46] Horowitz, J. L., 1998. Semiparametric methods in econometrics. Vol. 131. Springer Science & Business Media.

[47] Hsiao, C., 2014. Analysis of panel data: Third edition. Cambridge University Press.

[48] Hsiao, C., Pesaran, M. H., Tahmiscioglu, A. K., 2002. Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. Journal of Econometrics 109 (1), 107–150.

[49] Johnson, N. L., Kotz, S., Balakrishnan, N., 1995. Continuous univariate distributions. Vol. 2. Wiley.

[50] Keane, M., Moffitt, R., Runkle, D., 1988. Real wages over the business cycle: Estimating the impact of heterogeneity with micro data. Journal of Political Economy 96 (6), 1232–1266.

[51] Khazzoom, J. D., 1989. A note on the application of the nonlinear two-stage least-squares estimator to a Box-Cox-transformed model. Journal of Econometrics 42 (3), 377–379.

[52] Kitamura, Y., Tripathi, G., Ahn, H., 2004. Empirical likelihood-based inference in conditional moment restriction models. Econometrica 72 (6), 1667–1714.

[53] Kiviet, J. F., 2007. Judging contending estimators by simulation: Tournaments in dynamic panel data models. The Refinement of Econometric Estimation and Test Procedures, 282–318.

[54] Kruiniger, H., 2008. Maximum likelihood estimation and inference methods for the covariance stationary panel AR(1)/unit root model. Journal of Econometrics 144 (2), 447–464.

[55] Lavergne, P., 2008. A Cauchy-Schwarz inequality for expectation of matrices. Department of Economics, Simon Fraser University, Discussion Papers.

[56] Lavergne, P., Patilea, V., 2013. Smooth minimum distance estimation and testing with conditional estimating equations: Uniform in bandwidth theory. Journal of Econometrics 177 (1), 47–59.

[57] Li, Q., 1996. On the root-n-consistent semiparametric estimation of partially linear models. Economics Letters 51 (3), 277–285.

[58] Li, Q., Racine, J. S., 2007. Nonparametric econometrics: Theory and practice. Princeton University Press.

[59] Li, Q., Stengos, T., 1996. Semiparametric estimation of partially linear panel data models. Journal of Econometrics 71 (1-2), 389–397.

[60] Lütkepohl, H., 1996. Handbook of matrices. John Wiley&Sons.

[61] Maddala, G. S., 1971. The use of variance components models in pooling cross section and time series data. Econometrica 39, 341–358.

[62] McArthur, J. W., McCord, G. C., 2017. Fertilizing growth: Agricultural inputs and their effects in economic development. Journal of Development Economics 127, 133–152.

[63] Moral-Benito, E., 2013. Likelihood-based estimation of dynamic panels with predetermined regressors. Journal of Business & Economic Statistics 31 (4), 451–472.

[64] Mundlak, Y., 1978. On the pooling of time series and cross section data. Econometrica 46, 69–85.

[65] Neyman, J., Scott, E. L., 1948. Consistent estimates based on partially consistent observations. Econometrica 16, 1–32.

[66] Nickell, S., 1981. Biases in dynamic models with fixed effects. Econometrica 49, 1417–1426.

[67] Nolan, D., Pollard, D., 1987. U-processes: Rates of convergence. The Annals of Statistics 15 (2), 780–799.

[68] Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica 57 (5), 1027–1057.

[69] Powell, J. L., 1996. Rescaled methods-of-moments estimation for the Box-Cox regression model. Economics Letters 51 (3), 259–265.

[70] Robinson, P. M., 1988. Root-n-consistent semiparametric regression. Econometrica 56 (4), 931–954.

[71] Sakia, R., 1992. The Box-Cox transformation technique: A review. Journal of the Royal Statistical Society: Series D 41 (2), 169–178.

[72] Sherman, R. P., 1994. Maximal inequalities for degenerate U-processes with applications to optimization estimators. The Annals of Statistics 22 (1), 439–459.

[73] Shiller, R. J., 1984. Smoothness priors and nonlinear regression. Journal of the American Statistical Association 79 (387), 609–615.

[74] Shin, Y., 2008. Semiparametric estimation of the Box-Cox transformation model. The Econometrics Journal 11 (3), 517–537.

[75] Showalter, M. H., 1994. A Monte Carlo investigation of the Box-Cox model and a nonlinear least squares alternative. The Review of Economics and Statistics 76 (3), 560–570.

[76] Smith, R. J., 2007. Efficient information theoretic inference for conditional moment restrictions. Journal of Econometrics 138 (2), 430–460.

[77] Smith, R. J., 2007. Local GEL estimation with conditional moment restrictions. In: The refinement of econometric estimation and test procedures: Finite sample and asymptotic analysis. Cambridge University Press, pp. 100–122.

[78] Van der Vaart, A. W., 2000. Asymptotic statistics. Cambridge University Press.

[79] Wahba, G., 1984. Partial spline models for the semiparametric estimation of functions of several variables. In: Statistical Analysis of Time Series, Proceedings of the Japan U.S. Joint Seminar. pp. 319–329.

[80] White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50 (1), 1–25.

[81] Wooldridge, J. M., 1992. Some alternatives to the Box-Cox regression model. International Economic Review 33 (4), 935–955.

[82] Zhou, X.-H., Lin, H., Johnson, E., 2008. Non-parametric heteroscedastic transformation regression models for skewed data with an application to health care costs. Journal of the Royal Statistical Society: Series B 70 (5), 1029–1047.