Essays in Statistics

Inaugural-Dissertation
zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften

durch die

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

vorgelegt von

Alexander Gleim

aus Kamyschin (Russische Föderation)

Bonn 2016

Dekan:	Prof. Dr. Rainer Hüttemann
Erstreferent:	Prof. Dr. Christian Pigorsch
Zweitreferent:	Prof. Dr. Alois Kneip
Tag der mündlichen Prüfung:	21.03.2016

Diese Dissertation ist auf dem Hochschulschriftenserver der ULB Bonn
http://hss.ulb.uni-bonn.de/diss_online elektronisch publiziert.

Contents

List of Authors

The following authors contributed to

Chapter 1: Christian Pigorsch.

Chapter 2: Christian Pigorsch.

Chapter 3: Christian Pigorsch, Nikolaus Schweizer.

Chapter 4: Nazarii Salish.

List of Figures

List of Tables

To my family.

Acknowledgements

Many people have contributed to this thesis and I owe them an enormous amount of gratitude. First and foremost, I would like to thank my main advisor Christian Pigorsch for his guidance, encouragement and steady support. I am very grateful to have jointly worked with him on my first research projects which opened the door to the vast field of Bayesian statistics for me. My greatest appreciation also goes to my second advisor Alois Kneip and his invaluable contribution to the fourth and fifth Chapter of this dissertation. I greatly cherish the discussions with him and his never failing intuition.

There were, however, many other people that have supported me on my academic path throughout the years. Special thanks go to Jörg Breitung, Norbert Christopeit, Matei Demetrescu, Joachim Grammig, Peter Guttorp, Stefan Hoderlein, Richard Nickl and James Powell, and I pride myself on having worked with them in one way or the other.

The past few years would not have been the same without the companionship of fellow students and coauthors. I want to particularly thank Daniel Becker, Thomas Nebeling and Nazarii Salish for many a night spent discussing the countability of certain sets - it was splendid and an experience that I never want to miss!

Abstract

This thesis is comprised of several contributions to the field of mathematical statistics, particularly with regards to computational issues of Bayesian statistics and functional data analysis.

The first two chapters are concerned with computational Bayesian approaches that allow one to generate samples from an approximation to the posterior distribution in settings where the likelihood function of some statistical model of interest is unknown. This has led to a class of Approximate Bayesian Computation (ABC) methods whose performance depends on the ability to effectively summarize the information content of the data sample by a lower-dimensional vector of summary statistics. Ideally, these statistics are sufficient for the parameter of interest. However, it is difficult to establish sufficiency in a straightforward way if the likelihood of the model is unavailable.

In Chapter 1 we propose an indirect approach to select sufficient summary statistics for ABC methods that borrows its intuition from the indirect estimation literature in econometrics. More precisely, we introduce an auxiliary statistical model that is large enough as to contain the structural model of interest. Summary statistics are then identified in this auxiliary model and mapped to the structural model of interest. We show sufficiency of these statistics for Indirect ABC methods based on parameter estimates (ABC-IP), likelihood functions (ABC-IL) and scores (ABC-IS) of the auxiliary model. A detailed simulation study investigates the performance of each proposal and compares it to a traditional, moment-based ABC approach. Particularly, the ABC-IL and ABC-IS algorithms are shown to perform better than both standard ABC and the ABC-IP methods.

In Chapter 2 we extend the notion of Indirect ABC methods by proposing an efficient way of weighting the individual entries of the vector of summary statistics obtained from the score-based Indirect ABC approach (ABC-IS). In particular, the weighting ma-

trix is given by the inverse of the asymptotic covariance matrix of the score vector of the auxiliary model and allows us to appropriately assess the distance between the true posterior distribution and the approximation based on the ABC-IS method. We illustrate the performance gain in a simulation study. An empirical application then implements the weighted ABC-IS method to the problem of estimating a continuous-time stochastic volatility model based on non-Gaussian Ornstein-Uhlenbeck processes. We show how a suitable auxiliary model can be constructed and confirm estimation results from concurring Bayesian estimation approaches suggested in the literature.

In Chapter 3 we consider the problem of sampling from high-dimensional probability distributions that exhibit multiple, well-separated modes. Such distributions arise frequently, for instance, in the Bayesian estimation of macroeconomic DSGE models. Standard Markov Chain Monte Carlo (MCMC) methods, such as the Metropolis-Hastings algorithm, are prone to get trapped in local neighborhoods of the target distribution thus severely limiting the use of these methods in more complex models. We suggest the use of a Sequential Markov Chain Monte Carlo approach to overcome these difficulties and investigate its finite sample properties. The results show that Sequential MCMC methods clearly outperform standard MCMC approaches in a multimodal setting and can recover both the location as well as the mixture weights in a 12-dimensional mixture model. Moreover, we provide a detailed comparison of the effects different choices of tuning parameters have on the approximation to the true sampling distribution. These results can serve as valuable guidelines when applying this method to more complex economic models, such as the (Bayesian) estimation of Dynamic Stochastic General Equilibrium models.

Chapters 4 and 5 study the statistical problem of prediction from a functional perspective. In many statistical applications, data is becoming available at ever increasing frequencies and it has thus become natural to think of discrete observations as realizations of a continuous function, say over the course of one day. However, as functions are generally speaking infinite-dimensional objects, the statistical analysis of such functional data is intrinsically different from standard multivariate techniques.

In Chapter 4 we consider prediction in functional additive models of first-order autoregressive type for a time series of functional observations. This is a generalization of functional linear models that are commonly considered in the literature and has two advantages to be applied in a functional time series setting. First, it allows us to introduce a very general notion of time dependencies for functional data in this modeling framework. Particularly, it is rooted at the correlation structure of functional principal

component scores and even allows for long memory behavior in the score series across the time dimension. Second, prediction in this modeling framework is straightforwardly implemented as it only concerns conditional means of scalar random variables and we suggest a k-nearest neighbors classification scheme. The theoretical contributions of this paper are twofold. In a first step, we verify the applicability of the functional principal components analysis under our notion of time dependence and obtain precise rates of convergence for the mean function and the covariance operator associated with the observed sample of functions. In a second step, we derive precise rates of convergence of the mean squared error for the proposed predictor, taking into account both the effect of truncating the infinite series expansion at some finite integer L as well as the effect of estimating the covariance operator and associated eigenelements based on a sample of N curves.

In Chapter 5 we investigate the performance of functional models in a forecasting study of ground-level ozone-concentration surfaces over the geographical domain of Germany. Our perspective thus differs from the literature on spatially distributed functional processes (which are considered to be (univariate) functions of time that show spatial dependence) in that we consider smooth surfaces defined over some spatial domain that are sampled consecutively over time. In particular, we treat discrete observations that are sampled both over a spatial domain and over time as noisy realizations of some time series of smooth bivariate functions. In a first step we therefore discuss how smooth functions can be reconstructed from such noisy measurements through a finite element spline smoother that is defined over some triangulation of the spatial domain. In a second step we consider two forecasting approaches to functional time series. The first one is a functional linear model of first-order auto-regressive type, whereas the second considers the non-parametric extension to functional additive models discussed in Chapter 4. Both approaches are applied to predicting ground-level ozone concentration measured over the spatial domain of Germany and are shown to yield similar predictions.

1

Approximate Bayesian computation with indirect summary statistics

1.1 Introduction

Bayesian inference has become increasingly popular in empirical research in the last decades. Based on efficient Markov chain Monte Carlo methods and the availability of powerful computational resources, Bayesian problems that, years ago, appeared to be intractable, can now easily be computed on average desktop computers with standard software packages. Simultaneously, however, highly complex stochastic models have been developed for which standard Bayesian methods are insufficient. These models are characterized by the fact that their likelihood function does not admit a closed form expression or is just (computationally) too costly to be evaluated at a specific point. Classical estimators such as *Maximum Likelihood estimators*, but also Bayesian estimators based on *Markov Chain Monte Carlo* methods, heavily rely on the evaluation of the likelihood at any given point in the parameter space. To overcome the complexity of the model, likelihood-free inference methods, called *Approximate Bayesian Computation (ABC)*, have been proposed in the population genetics literature (see Beaumont et al.

[12], Fu and Li [49], Pritchard et al. [89], Tanaka et al. [104], Tavaré et al. [106] and the references therein) and used in different areas. For instance, Bortot et al. [17] and Erhardt and Smith [44] apply this approach to the estimation of models for (spatial) extremes and Peters et al. [88] consider this method for Bayesian inference in α-stable models. For a more detailed overview we refer to Marin et al. [81] and Sisson and Fan [99] and the references therein.

The general idea of ABC methods is (in accordance to Monte Carlo algorithms in general) to generate sampled draws of parameters $\theta \in \Theta$ from the posterior distribution without relying on the computation of values of the likelihood function. Instead, they require that it is possible to *simulate* from the model, i.e. to generate a simulated sample of variables for any given parameter value $\theta \in \Theta$. This is an assumption that is usually met in most applications.

The problem in Bayesian inference is that if the likelihood function $l : \mathcal{Y} \times \Theta \to \mathbb{R}_+$ is unavailable, it is not possible to obtain (draws from) the (exact) posterior distribution with density[1]

$$p(\theta|\tilde{y}) \propto l(\tilde{y}, \theta)\pi(\theta), \tag{1.1}$$

using standard algorithms such as Markov chain Monte Carlo methods, where $\theta \in \Theta$ with Θ denoting the *parameter space*, $\tilde{y} \in \mathcal{Y}$ the observed data, $\pi : \Theta \to \mathbb{R}_+$ the density of the prior distribution and $p : \Theta \times \mathcal{Y} \to \mathbb{R}_+$ the density of the posterior distribution. To overcome this problem, the ABC approach successfully absorbs the well established method of data augmentation from likelihood based Bayesian inference to the likelihood free case. The main idea is to introduce an ancillary variable that can take on values on the same set of possible observations as the observed data, $\hat{y} \in \mathcal{Y}$, such that the joint posterior density of the ancillary variable and the parameter of interest can be written as

$$p(\theta, \hat{y}|\tilde{y}) \propto g(\tilde{y}|\hat{y}, \theta)l(\hat{y}, \theta)\pi(\theta). \tag{1.2}$$

Usually, the interest is in the marginal posterior distribution of the parameter which is obtained from (1.2) by integrating out the ancillary variable, i.e.

$$p(\theta|\tilde{y}) \propto \int_{\mathcal{Y}} g(\tilde{y}|\hat{y}, \theta)l(\hat{y}, \theta)\pi(\theta)\,\mathrm{d}\,\hat{y} = \pi(\theta)\int_{\mathcal{Y}} g(\tilde{y}|\hat{y}, \theta)l(\hat{y}, \theta)\,\mathrm{d}\,\hat{y}. \tag{1.3}$$

[1]Throughout the paper we denote any quantities based on the *observed* data sample by tilde (e.g. $\tilde{y}$) and all quantities based on *simulated* data samples by hat (e.g. $\hat{y}$). Furthermore, we use the letters π, l, and p to denote the density of the prior distribution, the likelihood function, and the density of the posterior distribution, respectively.

Comparing (1.1) with (1.3) it becomes obvious that the posterior density (1.3) is only exact if g puts mass only on $\tilde{y}$.

In its simplest form, any ABC algorithm can be thought of as being a special case of a general accept-reject algorithm. An obvious way to implement likelihood-free rejection sampling based on (1.3) would be to generate a candidate value $\theta^* \in \Theta$ from the prior distribution with density function π and then to simulate a data sample $\hat{y} \in \mathcal{Y}$ from the stochastic model according to the likelihood function for the sampled parameter value θ^*. The candidate value θ^* would then be accepted if the simulated sample $\hat{y}$ matched the observed sample $\tilde{y}$. The resulting (exact) algorithm is outlined in Algorithm 1.

Algorithm 1 ABC Reject Algorithm - Exact Case

1. Generate a proposal θ^* from the prior distribution with density π.
2. Simulate $\hat{y}$ from the model according to the likelihood function for $\theta = \theta^*$.
3. Accept θ^* if $\hat{y} = \tilde{y}$.
4. Return to 1.

Although such an algorithm would result in exact draws from the true posterior distribution with density p without involving the computation of the (unknown) likelihood function, it has some severe drawbacks. An exact match between simulated data $\hat{y}$ and observed data $\tilde{y}$ has a non-prohibitive acceptance probability only in discrete low-dimensional models. In the case of high-dimensional data, the acceptance probability will be too small to allow for a feasible application and will be even strictly zero in a continuous setting. For these reasons *approximate* approaches have been introduced, generalizing the above concept as follows.

Instead of putting all the mass on $\tilde{y}$, the mass is spread out in the neighborhood of $\tilde{y}$, i.e. g becomes g_ϵ with

$$g_\epsilon(\tilde{y}|\hat{y},\theta) = \frac{1}{\epsilon} K_\epsilon \left(\frac{|\tilde{y} - \hat{y}|}{\epsilon} \right), \tag{1.4}$$

where $K_\epsilon : \mathcal{Y} \to \mathbb{R}_+$ is a standard smoothing kernel density and $\epsilon \geq 0$. Moreover, the dimensionality of the data is reduced by summarizing the information in the data by

a vector of summary statistics $S : \mathcal{Y} \to \mathcal{S}$ such that (1.4) becomes

$$g_\epsilon^{(S)}(\tilde{y}|\hat{y},\theta) = \frac{1}{\epsilon} K_\epsilon^{(S)} \left(\frac{|S(\tilde{y}) - S(\hat{y})|}{\epsilon} \right),$$

with $K_\epsilon^{(S)} : \mathcal{S} \to \mathbb{R}_+$ again a standard smoothing kernel, but now defined on the range space of the summary statistics S.

Naturally, different kernels are available and have been considered in the literature, e.g. Ratmann et al. [94] use a nonparameteric density estimator, whereas Beaumont et al. [12] adopt the Epanechnikov kernel. In this paper we present our results based on the uniform kernel density which has been widely used in the literature, e.g. Marjoram et al. [82] and Sisson et al. [100]. This choice allows for a simple representation of $g_\epsilon^{(S)}$ which is given by

$$g_\epsilon^{(S)}(\tilde{y}|\hat{y},\theta) \propto \begin{cases} 1 & \text{if } d\left(S\left(\tilde{y}\right), S\left(\hat{y}\right)\right) < \epsilon, \\ 0 & \text{otherwise,} \end{cases}$$

where $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$ is a metric defined over the summary statistics. A straightforward implementation of the sampling algorithm is outlined in Algorithm 2. For this al-

Algorithm 2 ABC Reject Algorithm - General Case

1. Generate a proposal θ^* from the prior distribution with density π.
2. Simulate $\hat{y}$ from the model according to the likelihood function for $\theta = \theta^*$.
3. Accept θ^* if $d(S(\hat{y}), S(\tilde{y})) \leq \epsilon$.
4. Return to 1.

gorithm, ϵ can be interpreted as a tolerance level which is chosen such that any candidate value θ^* is accepted if the distance between simulated and observed data, as measured by the metric d over the summary statistics, is below ϵ.

For ϵ close to 0 and summary statistics that are (approximately) sufficient, the ABC posterior with density

$$p^{(ABC)}(\theta|S(\tilde{y})) \propto \pi(\theta) \int_{\mathcal{Y}} g_\epsilon^{(S)}(\tilde{y}|\hat{y},\theta) l(\hat{y},\theta) \, \mathrm{d}\, \hat{y}$$

is a reasonable approximation to the true posterior distribution p that is based directly

on the observed data. Naturally, the performance of any ABC algorithm depends crucially on the particular choice of summary statistics S, the choice of the metric d and the tolerance level ϵ. In particular, the (automatic) choice of summary statistics has been an open question in the literature and gave rise to a number of different proposals that we are going to briefly review in Section 1.2. Existing approaches on choosing summary statistics have been mainly based on empirical considerations. Moreover, the existence of sufficient summary statistics has been assumed to not be verifiable, although in the light of the preceding discussion this property is highly desirable, as, intuitively speaking, sufficient statistics summarize all the information contained in the data that is relevant for the estimation of the model.

In this paper we address the question of how to choose sufficient summary statistics S. While searching for sufficient summary statistics, we borrow methods from the indirect estimation literature. In particular, we suggest the use of a suitably-chosen auxiliary model that captures the data generating process well and admits a tractable likelihood function. We develop (indirect) summary statistics that are sufficient for the parameters of the auxiliary model and derive conditions under which sufficiency carries over to the model of interest. This is a novel result in that sufficient summary statistics have long thought to be unidentifiable for ABC, with the exception of the special case of models of the exponential family. We thus refer to the ABC approach using indirect summary statistics as *Approximate Bayesian Computation with Indirect Summary Statistics* or, in short, *Indirect Approximate Bayesian Computation*.

Another important question that is addressed in the literature is how to improve the acceptance rate of proposed candidate values that are sampled from the prior distribution. Even in the approximate case (Algorithm 2), the acceptance probabilities may still be too small. A number of extensions have been recently suggested that address this issue of computational efficiency. Marjoram et al. [82] implemented an MCMC step into the algorithm which can result in more accepted draws from $p^{(ABC)}$ but at the usual cost of retaining serially dependent draws. Based on the work of Del Moral et al. [36] on *Sequential Monte Carlo*, Sisson et al. [100] tried to overcome the issue of serially dependent posterior draws by an adaptive scheme. But, as Beaumont et al. [13] pointed out, their approach results in biased draws from the posterior. Currently, it seems that the Population Monte Carlo approach as introduced by Cappé et al. [22] is one of the computationally most efficient versions of the ABC scheme. In this paper we focus on the selection of summary statistics for the original ABC sampling (Algorithm 2) but, of course, all arguments remain valid if a more efficient sampling scheme is used.

This paper is organized as follows. In Section 1.2 we review the existing literature on the problem of choosing summary statistics and establish linkages to the indirect estimation literature which is also discussed in some detail. Section 1.3 then develops our notion of *Indirect ABC*, shows how to systematically choose summary statistics and establishes sufficiency results for these statistics. In Section 1.4 we assess the performance of Indirect ABC in a simulation study and compare the results to traditional moment based ABC approaches. Section 1.5 concludes.

1.2 Literature review and preliminaries

In this section we review the existing literature on the construction and selection of summary statistics for ABC sampling schemes. We further present in some detail the concept of indirect estimation methods in general and discuss their linkages to ABC methods. This serves the purpose of introducing and establishing the necessary concepts for our notion of Indirect ABC.

1.2.1 Existing approaches to choosing summary statistics

Although the choice and construction of summary statistics is crucial in making ABC methods applicable, the current state of the literature is highly based on empirical considerations (see, e.g., Beaumont et al. [12] and Csilléry et al. [34]). With the notable exception of Gibbs Random Fields, sufficient statistics have not been proven to exist in general. The reason why it is possible to construct sufficient summary statistics in the special case of Gibbs Random Fields is that the model structure is given by an exponential family for which a simple form of sufficient statistics is known to exist - a fact that can be fruitfully exploited (see, e.g., Grelaud et al. [63]). For more general (non-exponential) settings, however, sufficient statistics have not been thought to exist in general and instead it is usually argued on an ad-hoc consideration of the problem at hand which statistics might be suitable to summarize the amount of information contained in the sample of observed variables. Consequently, a number of proposals have been made in the literature that address the question of dimension reduction within ABC sampling schemes. They share the common belief that by including a large number of summary statistics one can depict the main characteristic features of the data. However, the experimenter is facing a trade-off between a good description of the data and the curse of dimensionality as implied by the Kernel smoothing function $g_\epsilon^{(S)}$ from before. A recent review of these proposals has been given in Blum et al. [16] and we are referring to their

work for more details on a comparative analysis. In what follows, we are giving a brief discussion of the main concepts.

Blum et al. [16] categorize the existing methods of choosing summary statistics into three (non-mutually exclusive) classes. The first class concerns *best subset selection techniques* that involve specifying a (possibly very large) initial set of summary statistics subsets of which are subsequently scored as to retain those statistics that contain the most amount of information as measured by some suitable criterion. Joyce and Marjoram [76] consider to this end a notion of sufficiency that they call ϵ-*sufficiency*. A given set of k summary statistics $\{S_1, \ldots, S_k\}$ is said to be ϵ-sufficient relative to some new statistic S_{k+1} if

$$\sup_{\theta \in \Theta} \log l(S_{k+1}|S_1, \ldots, S_k, \theta) - \inf_{\theta \in \Theta} \log l(S_{k+1}|S_1, \ldots, S_k, \theta) \leq \epsilon.$$

Based on this definition, they propose a sequential scoring scheme with which one decides on the inclusion of a new summary statistic S_{k+1} by considering whether this statistic will contribute significantly to the quality of inference as measured by a likelihood ratio statistic. Once the contribution of a statistic is below a certain pre-specified threshold, the inclusion of more statistics is stopped. As Marin et al. [81] point out, however, this method is not only paramount to the *ordering* in which the statistics are considered but, more importantly, it results in highly correlated summary statistics. A related method was introduced by Nunes and Balding [86] in that they consider an entropy measure instead of a sufficiency criterion to measure informativeness of summary statistics.

The second class of proposed methods is comprised of *projection techniques*. Rather than just considering some initial set of candidate statistics one allows for (non-)linear combinations of those statistics. As such one adds a regression layer to the ABC scheme, the rationale of which being that one can considerably decrease the dimensionality of the summary statistics while keeping the information they contain unaltered. Consider for brevity the uni-variate case and a given set of summary statistics $\{S_i\}_{i=1}^k$. Then, in the simplest case, one considers a homoskedastic regression, i.e.

$$\theta_i = m(S_i) + \epsilon_i,$$

where (θ_i, S_i) are draws from the prior predictive distribution $l(S|\theta)\pi(\theta)$, $i = 1, \ldots, k$ and $m(S_i) = \mathbb{E}[\theta|S = S_i]$ is the (conditional) mean function. Beau-

mont et al. [12] assume a linear model, i.e. $m(S_i) = \alpha + \beta^\mathsf{T} S_i$ whereas Blum and François [15] consider a heteroskedastic extension.

The third class of methods consists of *regularization techniques* which are essentially based on projection techniques where one additionally penalizes the outcome for model complexity. Blum et al. [16] propose an approach based on ridge regression where the regression coefficients are shrunk to zero such that uninformative summary statistics are associated with the smallest coefficients.

In a related contribution, Fearnhead and Prangle [46] take a different perspective on ABC methods in that they consider ABC to be an inferential scheme in its own right rather than a simple method to obtain non-parametric estimates of the posterior density. Instead of requiring $p^{(ABC)}$ to be a good approximation to the true posterior distribution *globally*, they only require that $p^{(ABC)}$ is a good approximation *locally*, that is only for certain parameter estimates. As such, the desired scheme should be able to represent the uncertainty in the parameters accurately. In particular, consider the probability that the ABC posterior $p^{(ABC)}$ assigns to the event $\mathcal{A} \subset \Theta$ which can be written as

$$\Pr^{(ABC)}(\theta \in \mathcal{A}|S(\tilde{y})) = \int_{\mathcal{A}} p^{(ABC)}(\theta|S(\tilde{y}))\,\mathrm{d}\,\theta.$$

Fearnhead and Prangle [46] then call an inferential scheme *calibrated* if

$$\Pr(\theta \in \mathcal{A}|\Pr^{(ABC)}(\theta \in \mathcal{A}|S(\tilde{y})) = q) = q,$$

i.e. events that have probability q assigned by the ABC posterior will indeed have probability q to occur. Although standard ABC is not calibrated they show that this can be achieved by considering a slight (randomizing) modification which they call *Noisy ABC*. They show that the optimal choice of such summary statistics is given by the posterior means of the parameters which can be estimated by standard least-squares regression techniques.

In this paper, however, we take a fully automatic approach on finding sufficient summary statistics. We approach this by using an indirect procedure that originated in the frequentist statistics literature. The linkage between ABC and indirect estimation methods has already been recognized in the literature. In independent and concurrent work Creel and Kristensen [32], Fearnhead and Prangle [46] and Drovandi et al. [41] discuss the usage of the parameter estimates of an auxiliary model as summary statistics within ABC. However, other indirect summary statistics can be constructed that have

favorable statistical properties and that are based on the likelihood function or its score of some auxiliary model. In the following subsection we provide a brief review of these indirect estimation methods and discuss in more detail how they can be linked to ABC particularly with regards to choosing summary statistics.

1.2.2 The indirect estimation approach

Like any other (parametric) estimation method, indirect estimation aims to estimate a statistical model

$$\mathcal{M}_S = (\mathcal{Y}_S, \mathcal{P}_S),$$

where $\mathcal{Y}_S$ is the set of *possible observations* and

$$\mathcal{P}_S = \{P_S(\theta), \theta \in \Theta\}$$

is a *family of probability distributions* on this space that is parametrized by θ which can take on values on the *structural parameter space* $\Theta \subset \mathbb{R}^q$. In the following, we call $\mathcal{M}_S$ the *structural model* for reasons that will become clear shortly and denote this with a subscript S. Indirect estimation methods were developed to overcome the problem where the likelihood function $l_S : \mathcal{Y}_S \times \Theta \to \mathbb{R}_+$ of the probability distributions for the structural model, $\mathcal{P}_S$, (with respect to a dominating measure) is not available or too costly to evaluate.

A first attempt to estimate a model without the knowledge of the likelihood function, based on simulations of the model, would be to minimize the distance between some carefully selected empirical moments (e.g. mean, variance or autocovariance) and their population counterparts computed by Monte Carlo methods. This approach was implemented in the *Simulated Method of Moments (SMM)* which was proposed by Duffie and Singleton [42], Lee and Ingram [79], McFadden [83] and Pakes and Pollard [87]. SMM is a simulation based extension of the Generalized Method of Moments (GMM), proposed by Hansen [69], in so far as the population moments can be computed by Monte Carlo methods and must not be available in closed form.[2] Obviously, the ABC approach is the Bayesian counterpart to the SMM. The appropriate selection of informa-

[2]Note that GMM as an extension to the Method of Moments of Karl Pearson does not require the availability of the likelihood function either. However, the moments need to be expressed in terms of the structural parameters, which is often at least as challenging as the formulation of the likelihood function, and is rarely possible in practical applications. Indirect estimation methods, however, are applicable if neither a likelihood function nor closed form moment conditions are available.

tive moment conditions is a challenging task and might for both, ABC and SMM, heavily depend on the model to be estimated. A successful approach for moment selection in the SMM setup is given by indirect estimation methods and we propose to extend this approach to the setting of ABC in the next section.

Indirect estimation methods assume the existence of an *auxiliary model*

$$\mathcal{M}_A = (\mathcal{Y}_A, \mathcal{P}_A),$$

with

$$\mathcal{P}_A = \{P_A(\omega), \omega \in \Omega\},$$

where the *parameter space of the auxiliary model* is given by $\Omega \subset \mathbb{R}^p$ with $q \leq p$. The key assumption for the auxiliary model is the availability of the likelihood function $l_A : \mathcal{Y}_A \times \Omega \to \mathbb{R}_+$. In practical applications it is moreover desirable to use an auxiliary model for which an efficient computation of the likelihood function is feasible, or - even better - for which a maximum likelihood estimator is available in closed form. The parameters of the auxiliary model, $\omega \in \Omega$, are often referred to as *auxiliary parameters*. Furthermore, it is assumed that the auxiliary model provides an adequate representation of the data generating process. Very often it is assumed that the structural model is nested within the auxiliary model, $\mathcal{P}_S \subset \mathcal{P}_A$, and that there exists a mapping $\eta : \Theta \to \Omega$ which maps the parameters of the structural model into the parameter space of the auxiliary model, such that[3]

$$\mathcal{P}_S = \{P_A(\eta(\theta)), \theta \in \Theta\}.$$

In frequentist statistics, the existence of the map can be relaxed to only hold in the neighborhood of the true parameter. However, for Bayesian inference this seems impossible, such that this presentation follows Gallant and McCulloch [51, Assumption 1] and requires the existence of this map for all $\theta \in \Theta$.[4]

Due to the availability of the likelihood function for the auxiliary model, its estimation via maximum likelihood is feasible yielding an estimate $\tilde{\omega}^{(ML)}$. The general idea of indirect estimation is now to choose a structural model (identified by a parameter $\tilde{\theta}^{(IE)} \in \Theta$) that is close to the estimated auxiliary model. As such, these estimation

[3] This map is also called *binding function* in the literature.

[4] A slightly more general assumption would require the existence of the map in the regions of positive prior probability mass.

methods concentrate the available information of the data in the (parametric) auxiliary model. An estimator for the structural parameter is thus found by minimizing some objective function $\mathcal{Q}^{(IE)} : \Theta \to \mathbb{R}$ which measures the distance between the structural model and the auxiliary model, i.e. one has

$$\bar{\theta}^{(IE)} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, \mathcal{Q}^{(IE)}(\theta)$$

for the estimated structural parameter.

The existing indirect estimation methods can be distinguished with respect to the employed objective function, i.e. how the distance between the structural model and the auxiliary model is measured. At least three propositions have been made in the literature:

1. The *Indirect Inference (II)* method was proposed by Gouriéroux et al. [62] and Smith [101] and is based on the distance between the parameters of the auxiliary model, i.e.

$$\mathcal{Q}^{(II)}(\theta) = \left(\tilde{\omega}^{(ML)} - \eta(\theta)\right) W \left(\tilde{\omega}^{(ML)} - \eta(\theta)\right)^{\mathsf{T}}$$

   where $W \in \mathbb{S}_{++}$ is a positive-definite weighting matrix.

2. The *Efficient Method of Moments (EMM)* was introduced by Gallant and Tauchen [54] and is based on the score of the auxiliary model, i.e.

$$\begin{aligned}
&\mathcal{Q}^{(EMM)}(\theta) \\
&= \left(\left.\frac{\partial \log l_A(\tilde{y}, \omega)}{\partial \omega}\right|_{\omega=\tilde{\omega}^{(ML)}} - \mathbb{E}_\theta \left[\left.\frac{\partial \log l_A(Y, \omega)}{\partial \omega}\right|_{\omega=\tilde{\omega}^{(ML)}} \right] \right) W \\
&\times \left(\left.\frac{\partial \log l_A(\tilde{y}, \omega)}{\partial \omega}\right|_{\omega=\tilde{\omega}^{(ML)}} - \mathbb{E}_\theta \left[\left.\frac{\partial \log l_A(Y, \omega)}{\partial \omega}\right|_{\omega=\tilde{\omega}^{(ML)}} \right] \right)^{\mathsf{T}} \\
&= \mathbb{E}_\theta \left[\left.\frac{\partial \log l_A(Y, \omega)}{\partial \omega}\right|_{\omega=\tilde{\omega}^{(ML)}} \right] W\, \mathbb{E}_\theta \left[\left.\frac{\partial \log l_A(Y, \omega)}{\partial \omega}\right|_{\omega=\tilde{\omega}^{(ML)}} \right]^{\mathsf{T}}
\end{aligned}$$

   with $\mathbb{E}_\theta$ denoting the expectation operator with respect to the structural model with parameter θ, i.e. Y is distributed according to the structural model with parameter θ such that

$$\mathbb{E}_\theta \left[\left.\frac{\partial \log l_A(Y, \omega)}{\partial \omega}\right|_{\omega=\tilde{\omega}^{(ML)}} \right] = \int_{\mathcal{Y}_S} \left.\frac{\partial \log l_A(y, \omega)}{\partial \omega}\right|_{\omega=\tilde{\omega}^{(ML)}} l_S(y, \theta)\, \mathrm{d}\, y.$$

Note that the score of the auxiliary model is zero when evaluated at the Maximum Likelihood Estimator $\tilde{\omega}^{(ML)}$. Here $W \in \mathbb{S}_{++}$ is again a positive-definite weighting matrix.

3. The *Simulated Quasi-Maximum Likelihood* (SQML) estimator was proposed by Smith [101] and is based on the distance between the log-likelihood values, i.e.

$$\mathcal{Q}^{(SQML)}(\theta) = \log l_A(\tilde{y}, \tilde{\omega}^{(ML)}) - \log l_A(\tilde{y}, \eta(\theta)) \propto -\log l_A(\tilde{y}, \eta(\theta)).$$

These three indirect estimation approaches differ only in their choice of summary statistics of the auxiliary model. As such, they readily lend themselves to be employed within ABC. Parameter estimates of the auxiliary model have been considered by Drovandi et al. [41] whereas the auxiliary likelihood function was employed in Gallant and McCulloch [51] in an approach that is somewhat related to ABC and that was applied to asset pricing in Aldrich and Gallant [1]. In this paper we particularly propose to follow the Efficient Method of Moments approach of Gallant and Tauchen [54] in using the scores of the auxiliary log-likelihood function as indirect summary statistics.

To facilitate the remaining discussion in the paper we introduce the following notation: we refer to general ABC methods with indirect summary statistics as *ABC-I*. We further differentiate these methods with respect to the employed indirect summary statistics. In particular, *ABC-IP* refers to ABC-I where the summary statistics are given by the parameter estimates of the auxiliary model. *ABC-IL* then refers to ABC-I with the likelihood function of the auxiliary model as summary statistic, while our approach, that makes use of the score of the auxiliary model, is denoted by *ABC-IS*.

1.3 Indirect Approximate Bayesian Computation

After establishing the necessary preliminaries, we now introduce our notion of Indirect ABC and establish sufficiency of these summary statistics for the structural model $\mathcal{M}_S$. To prove sufficiency of our indirect summary statistics we first establish sufficiency for the auxiliary model $\mathcal{M}_A$ and then derive conditions under which sufficiency carries over to the structural model. We end this section with a discussion of implementation details.

1.3.1 Indirect summary statistics

Let us consider a structural model $\mathcal{M}_S$ with likelihood function given by $l_S(y,\theta) : \mathcal{Y}_S \times \Theta \to \mathbb{R}_+$. The likelihood function of the structural model (l_S) may not be available in closed form but we assume that it is possible to simulate from the model, i.e. we can obtain draws $\hat{y} \equiv (\hat{y}_t)_{t=1}^n$ that are a sample from $y \mapsto l_S(y,\theta)$ for any given parameter value $\theta \in \Theta$. For brevity of exposition we consider the special case of a stationary stochastic process $(y_t)_{t=-\infty}^{\infty}$ with transition density given by $l_S^\dagger$ such that the likelihood function l_S can be factorized as

$$l_S(y,\theta) = l_S^\ddagger(x_0,\theta) \prod_{t=1}^{n} l_S^\dagger(y_t|x_{t-1};\theta), \tag{1.5}$$

where x_{t-1} denotes the vector of state variables (e.g. the first L lagged variables of the series itself, $(y_{t-1}, y_{t-2}, \ldots, y_{t-L})$).

We base the estimation of the parameters of the structural model on summary statistics of some auxiliary model $\mathcal{M}_A$. To this end, we consider the (tractable) likelihood function $l_A(y,\omega) : \mathcal{Y} \times \Omega \to \mathbb{R}_+$ of a given auxiliary model $\mathcal{M}_A$ and denote by

$$\tilde{\omega}^{(ML)} = \operatorname*{argmax}_{\omega\in\Omega} l_A(\tilde{y},\omega)$$

the Maximum Likelihood estimator of the parameters of the auxiliary model for the observed data sample $\tilde{y} \equiv (\tilde{y}_t)_{t=1}^n$. As in the indirect estimation literature, we distinguish three choices for indirect summary statistics as discussed in Section 1.2.2.

In analogy to the *Indirect Inference* approach, indirect summary statistics S_{IP} can be identified by considering parameter estimates of the auxiliary model $\mathcal{M}_A$, i.e.

$$S_{IP}(y) = \operatorname*{argmax}_{\omega\in\Omega} l_A(y,\omega) \tag{1.6}$$

which results in the ABC-IP algorithm of Drovandi et al. [41]. Following the *Simulated Quasi-Maximum Likelihood* approach, one can identify indirect summary statistics S_{IL} by considering the likelihood function l_A of the auxiliary model, i.e.

$$S_{IL}(y,\omega) = l_A(y,\omega) \tag{1.7}$$

which yields the ABC-IL algorithm. Finally, indirect summary statistics can be devised

by following the *Efficient Method of Moments* approach in which summary statistics S_{IS} are based on the score of the auxiliary model, i.e.

$$S_{IS}(y,\omega) = \frac{\partial}{\partial \omega'} \log l_A\left(y,\omega'\right)\Big|_{\omega'=\omega} \tag{1.8}$$

which results in the ABC-IS algorithm.

It is important to note that these summary statistics depend on the structural parameter θ through the simulated sample $\hat{y}$. Furthermore, in the case of ABC-IS we observe that (1.8) evaluated at the observed data is zero for all values of θ, i.e. $S_{IS}(\tilde{y},\tilde{\omega}^{(ML)}) = 0$, by construction. A generic implementation of indirect summary statistics within ABC is outlined in Algorithm 3. All ABC-I algorithms are initialized by specifying and estimating an auxiliary model $\mathcal{M}_A$ on the observed data $\tilde{y}$. Steps 4.*a* - 4.*c* then calculate the indirect summary statistics for ABC-IP, ABC-IL and ABC-IS respectively.

1.3.2 Sufficiency results

In this section we show sufficiency of our indirect summary statistics defined in (1.6), (1.7) and (1.8). By the sufficiency principle, any inference about the parameter of interest should depend on the data y only through some statistic $T : \mathcal{Y} \to \mathcal{T}$. In other words, any sufficient statistic T contains as much information as the whole sample y itself, thus providing a convenient method to reduce the complexity of our data. Our approach is to first review some theoretical results on sufficient statistics. We then show sufficiency of the indirect summary statistics for the auxiliary model and discuss conditions under which sufficiency carries over to the structural model.

To make these ideas more precise, it proves insightful to think of sufficiency of a statistic T in terms of σ-fields over the space of observations $\mathcal{Y}$ that are induced by the statistic T (see Bahadur [6] and Halmos and Savage [67] for more details on this view on sufficiency). Consider the measure space $(\mathcal{Y}, \mathfrak{S}, \mathcal{P})$ where $\mathcal{Y}$ and $\mathcal{P}$ are our sample space and family of probability measures, respectively, from before and the σ-field $\mathfrak{S}$ is comprised of all subsets $\mathcal{A} \subset \mathcal{Y}$ such that any member $P \in \mathcal{P}$ assigns a well-defined probability $P(\mathcal{A})$ to any event of the form $\{y \in \mathcal{A}\}$. Corresponding to the σ-field $\mathfrak{S}$ we denote with $\mathfrak{T}$ the σ-field comprised of sets $\mathcal{B}$ such that $T^{-1}(\mathcal{B}) = \{y : T(y) \in \mathcal{B}\}$ is an $\mathfrak{S}$-measurable subset of $\mathcal{Y}$. We write $\mathfrak{S}_0$ for the sub-σ-field of $\mathfrak{S}$ that is *induced* by the statistic T, i.e. it is comprised of all sets $\{T^{-1}(\mathcal{B}), \mathcal{B} \in \mathfrak{T}\}$. We furthermore de-

Algorithm 3 ABC-IP, ABC-IL and ABC-IS Reject Algorithm

1. Compute the MLE of the auxiliary model parameter ($\tilde{\omega}^{(ML)}$) based on observations $\tilde{y} = (\tilde{y}_t)_{t=1}^n$.

2. Generate a proposal θ^* from the prior with density π.

3. Simulate $\hat{y} = (\hat{y}_t)_{t=1}^n$ from the structural model with likelihood function $y \mapsto l_S(y, \theta^*)$.

4.a For ABC-IP, compute $S_{IP}(\hat{y})$ by

$$S_{IP}(\hat{y}) = \underset{\omega \in \Omega}{\operatorname{argmax}}\, l_A(\hat{y}, \omega)$$

and calculate the overall distance d by

$$d = \|S_{IP}(\tilde{y}) - S_{IP}(\hat{y})\|_2.$$

4.b For ABC-IL, compute $S_{IL}(\hat{y}, \tilde{\omega}^{(ML)})$ by

$$S_{IL}(\hat{y}, \tilde{\omega}^{(ML)}) = l_A(\hat{y}, \tilde{\omega}^{(ML)})$$

and calculate the overall distance d by

$$d = |S_{IL}(\tilde{y}, \tilde{\omega}^{(ML)}) - S_{IL}(\hat{y}, \tilde{\omega}^{(ML)})|$$

4.c For ABC-IS, compute $S_{IS}(\hat{y}, \tilde{\omega}^{(ML)})$ by

$$S_{IS}(\hat{y}, \tilde{\omega}^{(ML)}) = \left. \frac{\partial}{\partial \omega} \log l_A(\hat{y}, \omega) \right|_{\omega = \tilde{\omega}^{(ML)}}$$

and calculate the overall distance d by

$$d = S_{IS}(\hat{y}, \tilde{\omega}^{(ML)})^{\mathsf{T}} S_{IS}(\hat{y}, \tilde{\omega}^{(ML)}).$$

5. If $d < \epsilon$

   accept θ^* as a draw from the ABC posterior.

6. Return to 2.

note with $\mathcal{Q}$ the set of probability measures of the form $\{Q = P(T^{-1}(\cdot))\}$ for any $P \in \mathcal{P}$.

Following Bahadur [6, Definition 3.1], we say that a statistic T is *sufficient* for a family of probability measures $\mathcal{P}$ if for any $\mathfrak{S}$-measurable set $\mathcal{A}$ there exists a $\mathfrak{T}$-Q-integrable function $\varphi_{\mathcal{A}}(y)$ such that for all $\mathcal{B} \in \mathfrak{T}$ and all $P \in \mathcal{P}$ one has

$$\int_{\mathcal{A}\cap T^{-1}(\mathcal{B})} \mathrm{d}\, P = \int_{\mathcal{B}} \varphi_{\mathcal{A}}(y)\, \mathrm{d}\, P(T^{-1}(y)).$$

Equivalently, one can define sufficiency in terms of the σ-field $\mathfrak{S}_0$ that is induced by the statistic T alone. Following Bahadur [6, Definition 5.1], we say that a σ-field $\mathfrak{S}_0 \subseteq \mathfrak{S}$ is *sufficient* for a family of probability measures $\mathcal{P}$ if for any $\mathfrak{S}$-measurable set $\mathcal{A}$ there exists an $\mathfrak{S}_0$-measurable function $\varphi_{\mathcal{A}}(y)$ such that one as

$$\varphi_{\mathcal{A}}(y) = \mathbb{E}_P(\mathbb{I}_{\mathcal{A}}(y)|\mathfrak{S}_0)$$

modulo $\mathfrak{S}$-P-nullsets. In other words, a statistic T is sufficient for $\mathcal{P}$ if and only if the sub-σ-field $\mathfrak{S}_0$ induced by T is for which it has to hold that the conditional probability of a set $\mathcal{A}$ given $\mathfrak{S}_0$ is the same for each $P \in \mathcal{P}$. If we furthermore assume that the family of probability measures $\mathcal{P}$ is *dominated* by some measure μ (e.g. Lebesgue measure) we can give a more practical condition for the sufficiency of $\mathfrak{S}_0$ in terms of Radon-Nikodym derivatives. Halmos and Savage [67, Theorem 1] (see also Bahadur [6, Theorem 6.1 (iii)]) showed that a necessary and sufficient condition for a sub-σ-field $\mathfrak{S}_0$ to be *sufficient* for $\mathcal{P}$ is that for any $P \in \mathcal{P}$ there exists a non-negative $\mathfrak{S}_0$-measurable function g_P such that on has

$$g_P = \frac{\mathrm{d}\, P}{\mathrm{d}\, \mu}$$

on $\mathfrak{S}$. In other words, the Radon-Nikodym derivative of any $P \in \mathcal{P}$ with respect to the dominating measure μ has to be $\mathfrak{S}_0$-measurable.

We consider now sufficiency of our indirect summary statistics for the auxiliary model $\mathcal{M}_A = (\mathcal{Y}_A, \mathfrak{S}_A, \mathcal{P}_A)$ to which we added the σ-field $\mathfrak{S}_A$ generated by $\mathcal{Y}_A$ to make $\mathcal{M}_A$ a measure space. Assume that the family of probability measures $\mathcal{P}_A = \{P(\omega) : \omega \in \Omega\}$ is *dominated* by some measure μ (e.g. Lebesgue measure) and write

$$l_A(y, \omega) = \frac{\mathrm{d}\, P(\omega)}{\mathrm{d}\, \mu}(y) \tag{1.9}$$

for some version of the Radon-Nikodym derivative on the right hand side of (1.9). Denote by $\mathfrak{L}_A$ the σ-field generated by the likelihood functions $l_A(\cdot, \omega)$. Obviously, the functions $l_A(\cdot, \omega)$ are $\mathfrak{L}_A$-measurable from which we conclude that the σ-field $\mathfrak{L}_A$ is sufficient for $\mathcal{P}_A$. This provides us with a rigorous formulation of what one would intuitively regard as the *sufficiency of the likelihood function*. It moreover holds that any statistic T that generates the same partition of the sample space as the likelihood function is also sufficient (see Barndorff-Nielsen et al. [11]) and following Barndorff-Nielsen and Cox [8] we can thus argue that the first derivative of the log-likelihood function, and thus our indirect summary statistics, are indeed sufficient for $\mathcal{P}_A$.

What remains to be discussed are conditions under which sufficiency for the auxiliary parameters carries over to the parameters of the structural model. Assume that there exists a map from the parameter space Θ of the structural model to the parameter space Ω of the auxiliary model such that

$$l_S(y, \theta) = l_A(y, \eta(\theta))$$

for all $\theta \in \Theta$ (and $y \in \mathcal{Y}_S$) for which our prior beliefs have positive probability mass. The idea is that given such a map η we can think of the auxiliary model $\mathcal{M}_A$ to be large enough to contain the structural model $\mathcal{M}_S$ as a special case or, in other words, model $\mathcal{M}_S$ is nested in model $\mathcal{M}_A$ in the region where our prior distribution has positive probability mass. Then any statistic $T : \mathcal{Y}_A \to \mathcal{T}$ that is sufficient for $\mathcal{P}_A$ is also sufficient for $\mathcal{P}_S$ such that our sufficient statistic for the auxiliary model is also a sufficient statistic for the structural model (see Gouriéroux and Monfort [61, Property 3.5]). This argument completes the derivation of our sufficiency result for the here proposed ABC-IS method and also applies to the ABC-IL method based on the likelihood functions of the auxiliary model alone. Since, as Barndorff-Nielsen [7] points out, also the Maximum Likelihood estimator is itself *asymptotically sufficient*[5] of order $\mathcal{O}(n^{-1/2})$, the above argument establishes an asymptotic sufficiency result for the ABC-IP method as proposed by Drovandi et al. [41] as well.

[5]Asymptotics are taken here with respect to letting the sample size n tend to infinity. The asymptotic sufficiency of the ML estimator is essentially due to the fact that the (normed) likelihood function can be expressed in a Taylor series to any desired degree.

1.3.3 Discussion

Using an indirect approach to parameter estimation requires the specification of two different models for the same data set. This often causes confusion about how we should think of the two models, namely the structural and the auxiliary model. If we can approximate the (underlying) data generating process (to any desired degree) by some fully parametrized auxiliary model, then why shall we consider to estimate a highly complex structural model in the first place? The answer to this question depends on how the experimenter thinks of the data at hand. If the auxiliary model gave a sensible *explanation* of how the observed data was generated, then we would not have to rely on computationally involved indirect methods to solve the estimation problem since, per assumption, a Maximum Likelihood Estimator is readily available for the auxiliary model.

Indirect estimation methods indeed only make sense when we have good reason to believe that the structural model *explains* the data well, but is too complex to be estimated by standard methods. The auxiliary model is thus merely seen as the best statistical fit on the data and may also be called the *statistical model* in contrast to the *structural*, e.g. economic, model. Especially in research areas where the interest is in structural and causal relationships, these methods allow for the simple and straightforward estimation of otherwise hard to estimate structural models based on easy to estimate auxiliary models. It is therefore not surprising that these methods have found widespread use especially in mathematical biology and economics. Prominent economic examples are the estimation of stochastic differential equations, e.g. see Andersen et al. [3] and Gallant and Long [50]; the estimation of stochastic volatility models, e.g. see Chernov et al. [26]; the estimation of dynamic stochastic general equilibrium models, see Le et al. [78], and the estimation of labor market models, see Magnac et al. [80] and Topa [107].

In comparison to the parameter and likelihood based indirect estimators (i.e. II and SQML in the frequentist setting and ABC-IP and ABC-IL for ABC) the score based approaches (i.e. EMM and ABC-IS) have the advantage that the computation of the objective function is more efficient. Thus, the estimation time is significantly smaller. This is due to the fact that they do not require the computation of the map (i.e. the binding function), which involves the estimation of the auxiliary model. So far, we have assumed that the map η is available and known. However, with the exception of very simple examples, this is rarely the case and the parameter and likelihood based indirect estimators therefore rely on a Monte Carlo approach to estimate the map. In particular, in the ABC-IL, II and SQML approach, M samples $\left\{ (\hat{y}_{m,t})_{t=1}^{n} \right\}_{m=1}^{M}$ of length n are first generated

from the structural model $\mathcal{M}_S$ for a given structural parameter vector θ. Based on these simulated samples, the map is computed as

$$\hat{\eta}(\theta) = \frac{1}{M} \sum_{m=1}^{M} \hat{\omega}_m$$

with

$$\hat{\omega}_m = \underset{\omega \in \Omega}{\operatorname{argmax}}\, l_A(\hat{y}_m, \omega).$$

The ABC-IP method, in contrast, only requires the estimation of the map for one simulated data set (for a given structural parameter). In fact, even if the map is known it is only of limited use in the ABC-IP case. To illustrate this, consider the case where the map is known and the tolerance level is rather small such that the draws from the ABC posterior are concentrated around the structural parameter that yields the smallest value of the objective function which is not necessarily consistent with the posterior distribution. This will not be the case if we allow for sample variation in the computation of the map (for ABC-IP) and the computation of the score (for ABC-IS). It is important to note that for realistic auxiliary models the maximization problem of the likelihood function is rarely solvable in closed form such that the use of numerical optimization methods is often inevitable. This can result in significant computational costs. In contrast, the score based methods only require the computation of the score. For the EMM case, the expectation of the score of the auxiliary model under the structural model can be estimated by

$$\hat{\mathbb{E}}_\theta \left[\frac{\partial \log l_A(Y, \omega)}{\partial \omega} \bigg|_{\omega = \tilde{\omega}^{(ML)}} \right] \approx \frac{1}{M} \sum_{m=1}^{M} \frac{\partial \log l_A(\hat{y}_m, \omega)}{\partial \omega} \bigg|_{\omega = \tilde{\omega}^{(ML)}},$$

whereas in the ABC-IS case only one sample is used since the same arguments apply as in the ABC-IP case.

The missing optimization step in the computation of the objective function leads to a significant reduction of computing time. In the setting of the simulation study of Section 1.4 the difference is of factor 50 for ABC-IP vs ABC-IS, that is for $M = 1$. For more realistic models, e.g. high-dimensional and nonlinear models, this effect will be even more pronounced as for each candidate value from the prior distribution a non-trivial estimation step is involved.

1.4 Illustration and simulation results

In this section we illustrate the usage and performance of indirect approaches to ABC. We consider a simple model that admits a closed form expression for the associated posterior distribution such that exact Bayesian inference can be carried out. We then investigate the accuracy of the indirect ABC algorithms to the true posterior in a simulation study and compare the results to standard ABC methods.

1.4.1 Model setting

We consider the data $y \equiv (y_t)_{t=1}^n$ to be an i.i.d. sample from an exponential distribution $\mathcal{E}(\lambda)$ with parameter $\lambda > 0$ such that the structural model is given by

$$\mathcal{M}_S = \left(\mathbb{R}_+^n, \left\{\mathcal{E}(\lambda)^{\otimes n}, \lambda > 0\right\}\right),$$

where the exponential distribution has a density with respect to Lebesgue measure given by

$$z \mapsto \lambda \exp(-\lambda z)\, \mathbb{I}_{z \geq 0}.$$

We take the conjugate prior on the parameter λ of the structural model $\mathcal{M}_S$ to be $\lambda \sim \mathcal{G}\left(\alpha^{(\pi)}, \beta^{(\pi)}\right)$ with $\mathcal{G}$ denoting the gamma distribution which allows us to perform exact Bayesian inference. The likelihood of the model (with respect to Lebesgue measure) then reads as

$$l_S(y, \lambda) = \lambda^n \exp\left(-\lambda \sum_{t=1}^n y_t\right)$$

such that the closed form solution of the posterior distribution is given by

$$\lambda | y \sim \mathcal{G}\left(\alpha^{(\pi)} + n,\ \beta^{(\pi)} + \sum_{t=1}^n y_t\right). \tag{1.10}$$

To conduct Indirect ABC we have to specify an auxiliary model. In this example we consider the same model structure as for the structural model but with Gamma distributed observations, i.e.

$$\mathcal{M}_A = \left(\mathbb{R}_+^n, \left\{\mathcal{G}(\alpha, \beta)^{\otimes n}, (\alpha, \beta) \in (\mathbb{R}_+ \times \mathbb{R}_+)\right\}\right)$$

with density

$$z \mapsto \frac{\beta^{\alpha}}{\Gamma(\alpha)} z^{\alpha-1} \exp(-\beta z)\ \mathbb{I}_{z>0}.$$

It is obvious that the mapping assumption is satisfied for $\eta\colon \lambda \mapsto (1, \lambda)$ such that the structural model is nested within the auxiliary model, i.e.

$$\left\{\mathcal{E}(\lambda)^{\otimes n}, \lambda > 0\right\} \subset \left\{\mathcal{G}(\alpha, \beta)^{\otimes n}, (\alpha, \beta) \in (\mathbb{R}_+ \times \mathbb{R}_+)\right\}.$$

This setup enables us to illustrate the differences between the summary statistics used in ABC-IP, ABC-IS and ABC-IL and compare them to the true posterior distribution. To this end note that

$$\frac{\partial \log l_A^{\dagger}(y_t, \alpha, \beta)}{\partial \alpha} = \log(\beta) - \psi(\alpha) + \log(y_t),$$

and

$$\frac{\partial \log l_A^{\dagger}(y_t, \alpha, \beta)}{\partial \beta} = \frac{\alpha}{\beta} - y_t,$$

where $\psi(z) = \Gamma(z)'/\Gamma(z)$ is the digamma function. Moreover, the Fisher information for the auxiliary model is given by

$$\begin{bmatrix} \bar{\psi}(\alpha) & -1/\beta \\ -1/\beta & \alpha/\beta^2 \end{bmatrix}$$

with $\bar{\psi}$ the trigamma function, i.e. the first derivative of the digamma function.

Taking expectations with respect to an exponentially distributed random variable (i.e. $Y \sim \mathcal{E}(\lambda)$) we obtain after some algebra

$$\mathbb{E}_{\lambda}\left[\frac{\partial \log l_A^{\dagger}(Y, \alpha, \beta)}{\partial \alpha}\right] = \log(\beta) - \psi(\alpha) - \gamma - \log(\lambda),$$

and

$$\mathbb{E}_{\lambda}\left[\frac{\partial \log l_A^{\dagger}(Y, \alpha, \beta)}{\partial \beta}\right] = \frac{\alpha}{\beta} - \frac{1}{\lambda},$$

with $\gamma = -\psi(1) \approx 0.5772156649015\ldots$ the Euler–Mascheroni constant.

Ignoring in a first analysis the estimation uncertainty for the auxiliary model, i.e. $\tilde{\alpha}^{(ML)} = 1$ and $\tilde{\beta}^{(ML)} = \lambda_0$, it becomes obvious that the objective function $\mathcal{Q}^{(IE)}$ of any indirect estimation procedure is minimized (yields a value of zero) at $\lambda = \lambda_0$. Thus, assuming a known map, the II, EMM and SQML methods yield the same estimate, i.e. $\tilde{\lambda}^{(EMM)} = \tilde{\lambda}^{(II)} = \tilde{\lambda}^{(SQML)} = \lambda_0$. However, in realistic applications the estimation uncertainty can not be ignored such that in general $\tilde{\alpha}^{(ML)} \neq 1$ and $\tilde{\beta}^{(ML)} \neq \lambda_0$. This results in different estimators for the different moment conditions underlying the II, EMM and SQML principle or, correspondingly, the Indirect ABC approaches. Figure 1.1 highlights the difference for a simulated data set with 500 observations from the structural model with $\lambda_0 = 1$. The solid black vertical line depicts the manifold in the parameter space of the auxiliary model induced by the structural model. Without loss of generality, we consider the case of equally weighted moments. The II estimator is obtained by minimizing the distance between the Maximum Likelihood estimate $(\tilde{\alpha}^{(ML)}, \tilde{\beta}^{(ML)})$ and the manifold of the structural model. The SQML estimator is given at the point on the manifold for which the auxiliary likelihood function is maximal. In Figure 1.1, this is the point where the contour lines of the likelihood function are tangential to the manifold. The EMM objective function is indicated by the vectors originating from the manifold. They represent the (expectation) of the score of the auxiliary likelihood function at the manifold. At a first glance one would expect to see vectors pointing towards the steepest ascent of the log likelihood function. However, this intuition is misleading, as the objective function is the score of the auxiliary model *evaluated at the maximum likelihood estimates* of the parameters of the auxiliary model, such that the observed pattern in Figure 1.1 is in fact reasonable. The resulting EMM estimator $\tilde{\theta}^{(EMM)}$ is thus given by the structural parameter for which the length of the vector is minimal.

However, the minimum of the respective objective functions is only of limited interest in Bayesian inference. Even for this simple example an analytical analysis of the ABC approach with different indirect summary statistics is infeasible and we thus proceed with a simulation study.

1.4.2 Simulation study

In this study we analyze the approximation error of different ABC methods. In particular we consider a simple ABC approach using the mean and the variance of the data sample as summary statistics (ABC-M), the indirect ABC approach using the parameter

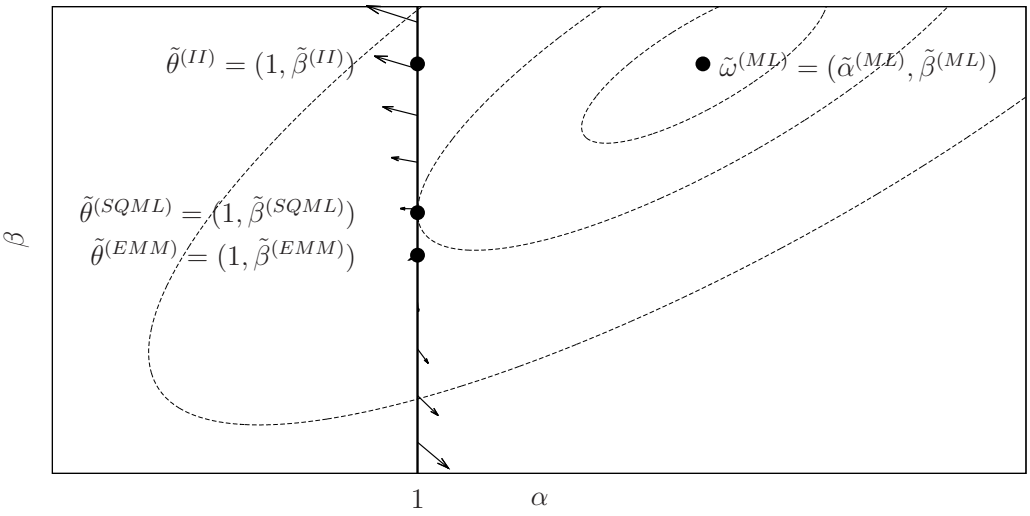


Figure 1.1: **Estimates based on different indirect estimation methods.** This figure illustrates for a simulated data set the parameter estimates of a structural model (exponential distribution) based on different indirect estimation methods using an auxiliary model (gamma distribution). The dashed lines represent the contours of the auxiliary log likelihood function, the horizontal line is the parameter space of the structural model, and the arrows represent the objective function of the EMM estimator. The points give the estimates resulting from the II, SQML and the EMM principles.

estimates of the auxiliary model (ABC-IP), the indirect ABC approach using the likelihood function of the auxiliary model (ABC-IL) and the indirect ABC approach using the score of the auxiliary model as summary statistics (ABC-IS). The corresponding distances are summarized in Table 1.1. In our analysis we consider the following simulation setup. For a given prior distribution (specified by α^{π} and β^{π}) we simulate $1,000$ samples of 100 exponentially distributed random variables with $\lambda_0 = 1$ (implying a mean and variance of one) representing our observed data. For every sample we compute the exact posterior distribution according to (1.10) and approximate the posterior distribution using the different ABC methods. To obtain the approximate posterior distribution we simulate for every observed data set $100,000$ proposal draws from the prior distribution and compute the distance implied by the different ABC approaches to the observed data. Discarding 99% of these proposals leaves us for each ABC method $1,000$ draws from the approximate posterior distribution. Although this (implicit) selection procedure for ϵ is rather crude, it allows for a fair comparison between these methods in terms of computing time and is standard practice in the literature (see Marin et al. [81]).

We then consider different statistics to measure the distance between the true poste-

Model	Distance over summary statistics: $d(S(\tilde{y}), S(\hat{y}))$
ABC-M	$\left\| \begin{pmatrix} \frac{1}{n}\sum_{t=1}^{n} \tilde{y}_t \\ \frac{1}{n}\sum_{t=1}^{n} (\tilde{y}_t - \bar{\tilde{y}})^2 \end{pmatrix} - \begin{pmatrix} \frac{1}{n}\sum_{t=1}^{n} \hat{y}_t \\ \frac{1}{n}\sum_{t=1}^{n} (\hat{y}_t - \bar{\hat{y}})^2 \end{pmatrix} \right\|_2$
ABC-IP	$\left\| \tilde{\omega}^{(ML)} - \hat{\omega}^{(ML)} \right\|_2$
ABC-IL	$\lvert l_A(\tilde{y}, \tilde{\omega}^{(ML)}) - l_A(\hat{y}, \tilde{\omega}^{(ML)}) \rvert$
ABC-IS	$\frac{\partial}{\partial \omega^\intercal} \log l_A(y, \omega)\Big\|_{\omega=\tilde{\omega}^{(ML)}} \frac{\partial}{\partial \omega} \log l_A(y, \omega)\Big\|_{\omega=\tilde{\omega}^{(ML)}}$

Table 1.1: **Different distances over summary statistics.** This table shows the different summary statistics employed and how the respective distances are calculated.

rior distribution and the approximations and compute for each statistic the average over all $1,000$ samples. This is repeated for several combinations of $\alpha^{(\pi)}$ and $\beta^{(\pi)}$ to analyze the effect of different priors, i.e. we compare informative prior distributions against rather flat prior distributions, and prior distributions that are in accordance with the data against prior distributions that are not. The results are presented in Table 1.2 and 1.3. To facilitate the interpretation of the output we order the results by the mean and the variance of the prior distribution which is in a one to one relation to the parameters of the prior distribution.

Table 1.2 reports the average of the (absolute) distance between the exact posterior mean and the respective ABC approximations. Interestingly, the simple ABC approach (ABC-M) that uses the sample mean and variance as summary statistics almost always performs worse than any of the indirect methods and exhibits the largest difference between the true and the approximated posterior mean. This is insofar surprising as the mean is a sufficient statistic for the exponential distribution.

The table also shows that the Indirect ABC approach with parameter estimates as summary statistics (ABC-IP) performs slightly better than plain ABC, but worse than score based approach and mostly worse than the likelihood based approach. Overall, it seems that the best results are obtained by the ABC-IS and ABC-IL with a slightly better performance of the ABC-IS method. Compared to the simple ABC-M approach, the reduction of the distance between the exact posterior mean and the ABC-IS approximation is at least of factor 0.37 (mean of 2 and variance of 0.25) and very often substantially better.

estimator	moments of the prior distribution			
	variance	mean		
		0.5	1	2
ABC-M	0.25	0.0183	0.0090	0.0434
ABC-IP		0.0141	0.0070	0.0560
ABC-IL		0.0110	0.0127	0.0258
ABC-IS		0.0103	0.0057	0.0288
ABC-M	0.5	0.0228	0.0132	0.0290
ABC-IP		0.0175	0.0090	0.0186
ABC-IL		0.0119	0.0117	0.0131
ABC-IS		0.0127	0.0069	0.0122
ABC-M	1	0.0289	0.0149	0.0351
ABC-IP		0.0241	0.0114	0.0135
ABC-IL		0.0126	0.0111	0.0123
ABC-IS		0.0168	0.0086	0.0095
ABC-M	2	0.0419	0.0216	0.0159
ABC-IP		0.0317	0.0164	0.0132
ABC-IL		0.0213	0.0121	0.0118
ABC-IS		0.0209	0.0121	0.0091
ABC-M	4	0.0592	0.0315	0.0175
ABC-IP		0.0469	0.0198	0.0135
ABC-IL		0.0750	0.0166	0.0156
ABC-IS		0.0287	0.0145	0.0101

Table 1.2: **Distance between the true posterior mean and the posterior mean based on different ABC methods.** This table shows the distance between the true posterior mean and the posterior mean based on different ABC methods and different prior distributions that are specified by their mean and variance. ABC-M corresponds to the ABC approach with sample mean and variance as summary statistics, ABC-IP corresponds to the Indirect ABC approach with parameter estimates as summary statistics, ABC-IL corresponds to the Indirect ABC approach with likelihood based summary statistics and ABC-IS corresponds to the Indirect ABC approach with score based summary statistics.

Although the posterior mean is an important functional of the posterior distribution, it is just one statistic and may not be representative for the whole posterior distribution. We therefore also aim to compare the exact posterior *distribution* with the approximation obtained by the different ABC approaches. To this end we repeat the procedure as described above but instead of considering the mean we focus on the chi-squared distance

$$\sum_{i=1}^{20} \frac{(o_i - e_i)^2}{e_i}$$

where e_i is the expected number of observations in cell i (obtained from the exact posterior distribution for a given simulated data set) and o_i is the number of realized observations in cell i. We use 20 cells that are computed in such a way that every cell contains 5% of the data, i.e. for the i-th cell we choose the interval given by

$$[Q(0.05(i-1)), Q(0.05i)]$$

with $Q : [0,1] \to \mathbb{R}_+ \cup \infty$ denoting the quantile function of the posterior distribution (1.10), $Q(0) = 0$ and $Q(1) = \infty$. Table 1.3 shows the averages of the chi-squared statistics over all $1,000$ replications using the same prior distributions as in Table 1.2.

Table 1.3 shows that, again that the ABC-IL and ABC-IS method clearly outperform the standard ABC approach as well as the parameter based ABC-IP method. However, a clear ranking between ABC-IL and ABC-IS seems to depend on the specification of the prior distribution.

The chi-squared statistic also allows us to analyze the approximation quality for different prior distributions. As may be expected, the best results will be obtained if the prior mean is equal to the value of the unknown parameter (note that $\lambda_0 = 1$ implies a mean of one) and if the prior variance is small. This ensures that the proposed draws (from the prior distributions) are very often close to the posterior distribution.

Increasing the variance of the prior distribution has two opposing effects. In fact, an increase of the variance leads to a larger variance of the proposal draws but also increases the support of the prior distribution that receives considerable probability mass. Which of those two effects will dominate depends on the relative position of the prior distribution with respect to the posterior distribution. If the mean of the prior distribution is relatively close to the (unknown) posterior mean, the first effect will clearly dominate: as the prior variance increases, the variance of the proposed draws increases as well

whereas the increase in support is insignificant such that the approximation gets worse. In contrast, if the prior mean is far away from the posterior mean, the second effect will dominate at first: an increase in the prior variance leads to an increase of the support that receives sufficient probability mass, resulting in more proposal draws that are close to the posterior distribution. As such, the approximation gets better by increasing the variance. However, this effect will be eventually dominated by the increased variance of the proposal draws and the approximation gets worse again by further increasing the prior variance once the benefits of an increased prior support have been exploited.

These effects are also seen in Table 1.3. For draws from the prior with a mean of 0.5 (which is relatively close to 1) increasing the variance leads to a worse approximation as does for a mean of 1. Whereas ABC-IL performs better for draws from the prior with mean 0.5, ABC-IS performs better for draws from the prior with mean 1. However, for draws from the prior with a mean of 2, the approximation gets better by increasing the variance up until a value of 2 and gets worse again by increasing the prior beyond that value. In these cases, ABC-IS performs better than ABC-IL.

1.5 Conclusion

In this paper we formalized a selection mechanism for sufficient summary statistics within the ABC framework that is based on an auxiliary model and borrows its intuition from the indirect estimation literature in statistics, particularly in econometrics. Three such mechanisms have been considered, based on the auxiliary parameter estimates (ABC-IP), the auxiliary log-likelihood function (ABC-IL) and on the auxiliary score vector (ABC-IS).

The ABC-IS and ABC-IL proposals have been shown to give rise to (exactly) sufficient summary statistics (whereas the ABC-IP method yields asymptotically sufficient summary statistics) for the structural model under the assumption that the latter is nested within the auxiliary model. A detailed simulation study investigated the performance of each proposal and compared it to a traditional, moment-based ABC approach. Particularly, the ABC-IL and ABC-IS algorithms performed better than both standard ABC and the ABC-IP methods.

estimator	variance	mean		
		0.5	1	2
ABC-M	0.25	141.8813	54.2198	573.2338
ABC-IP		151.6284	47.7336	922.7216
ABC-IL		67.6088	76.9999	211.2614
ABC-IS		46.9136	27.0765	377.0657
ABC-M	0.5	318.7890	100.2351	355.1234
ABC-IP		210.1628	73.0205	198.0316
ABC-IL		68.5202	60.9606	67.7523
ABC-IS		60.3490	31.1686	57.9846
ABC-M	1	301.3400	98.0395	143.1314
ABC-IP		367.2982	107.2002	143.3554
ABC-IL		62.9832	54.8310	70.2524
ABC-IS		94.7804	40.0977	51.4272
ABC-M	2	558.0046	183.6534	111.9762
ABC-IP		567.3480	193.4671	137.2659
ABC-IL		88.7074	57.7679	62.2512
ABC-IS		134.7983	58.0195	43.1016
ABC-M	4	1223.4122	612.3418	201.342
ABC-IP		1109.0594	264.1298	136.0284
ABC-IL		174.9038	61.1863	67.0217
ABC-IS		276.6809	78.0879	45.2486

Table 1.3: **Chi-squared distance between the true posterior distribution and the posterior distribution based on different ABC methods.** This table shows the chi-squared statistics for measuring the distance between the true posterior distribution and the posterior distribution based on different ABC methods and different prior distributions that are specified by their mean and variance. The results are reported for the ABC methods and the prior distributions discussed in Table 1.2.

2

Efficiently weighted Indirect ABC: an application of estimating a stochastic volatility model of OU type

2.1 Introduction

In this paper we apply the Indirect Approximate Bayesian Computation methodology developed in the previous chapter to the problem of estimating a continuous-time stochastic volatility model of Ornstein-Uhlenbeck type which is driven by a non-Gaussian Lévy process. This class of models was introduced by Barndorff-Nielsen and Shephard [9] and Barndorff-Nielsen and Shephard [10] and has gained wide popularity in the literature as it effectively captures stylized facts of financial time series such as volatility clustering, heavy tails in the return distribution and most importantly, jumps in the volatility process. The estimation of these models is challenging as the model structure induces two characteristics that cannot be handled by standard estimation methods. First, the volatility process is unobservable and second, the model is formulated in continuous time whereas the data is observed only at discrete time points. These challenges

have lead to the development of different estimation approaches. Taufer et al. [105], for example, use the affine structure of the characteristic function to estimate the model in a frequentist setup. Bayesian estimation was considered in Frühwirth-Schnatter and Sögner [48], Griffin and Steel [64] and Roberts et al. [97]. In the Bayesian setup, the jump times and jump sizes of the background driving Lévy process are modeled as unobserved variables and are integrated out.

In contrast to these approaches, the proposed IABC method only requires the possibility to generate simulations from the structural model. For the considered stochastic volatility model this simulation is straightforward when the driving Lévy processes is of finite activity and can be implemented to any desired degree if the process is of infinite activity (see Asmussen and Glynn [5, Chapter 12] and the references therein). As the observed data is condensed in an auxiliary model, it is further more required to specify an adequate model which describes the (observed) data sufficiently well. Especially for financial data the semi-nonparametric model of Gallant and Nychka [52] seems to be very successful as indicated by several studies, e.g. see Andersen et al. [3] or Chernov et al. [26].

The paper is organized as follows. Section 2.2 briefly discusses the employed estimation methodology based on Indirect Approximate Bayesian Computation where the summary statistics are given by the score vector of some suitable auxiliary model. Moreover, we consider a modification to the exposition in the previous chapter in that we introduce an efficient weighting scheme for the individual summary statistics. The effect of weighting is illustrated in a simulation study in Section 2.3 where we follow the same setup as in the previous chapter. Section 2.4 reviews the structural model, i.e. the Ornstein-Uhlenbeck type stochastic volatility model, while Section 2.5 presents the employed auxiliary model based on a semi-nonparametric density approach. Our estimation results are presented in Section 2.6. Section 2.7 concludes.

2.2 Approximate Bayesian Computation with weighted indirect score-based summary statistics

As discussed in the previous chapter, Indirect ABC methods consider the interplay between a *structural model* $\mathcal{M}_S = (\mathcal{Y}_S, \mathcal{P}_S)$ and a *statistical model* $\mathcal{M}_A = (\mathcal{Y}_A, \mathcal{P}_A)$. Here, $\mathcal{Y}_S$ and $\mathcal{Y}_A$ denote the sample space of the structural and auxiliary model, respectively whereas $\mathcal{P}_S = \{P_S(\theta), \theta \in \Theta\}$ and $\mathcal{P}_A = \{P_A(\omega), \omega \in \Omega\}$ denote families of probability distributions that are parametrized by some structural and auxiliary pa-

rameter vector θ and ω, respectively. While $\mathcal{M}_S$ is our model of interest, it is thought that its likelihood function $l_S(y,\theta) : \mathcal{Y}_S \times \Theta \to \mathbb{R}_+$ is not available in closed form (or too computationally involved to compute). However, it is assumed that one can generate, for each $\theta^* \in \Theta$, a simulated data sample $(\hat{y}_t)_{t=1}^n$ according to $y \mapsto l_S(y,\theta^*)$. As before, we consider for brevity of exposition the special case of a stationary stochastic process $(y_t)_{t=-\infty}^{\infty}$ with transition density given by $l_S^\dagger$ such that the likelihood function l_S can be factorized as

$$l_S(y,\theta) = l_S^\ddagger(x_0,\theta)\prod_{t=1}^{n} l_S^\dagger(y_t|x_{t-1};\theta), \tag{2.1}$$

where x_{t-1} denotes the vector of state variables (e.g. the first L lagged variables of the series itself, $(y_{t-1}, y_{t-2}, \ldots, y_{t-L})$).

In contrast, $\mathcal{M}_A$ is seen as a purely statistical model that represents the data generating process suitably well and admits a tractable likelihood function which we denote by $l_A(y,\omega) : \mathcal{Y} \times \Omega \to \mathbb{R}_+$. Similarly to above, we can factorize l_A in the stationary case as

$$l_A(y,\omega) = l_A^\ddagger(x_0,\omega)\prod_{t=1}^{n} l_A^\dagger(y_t|x_{t-1};\omega), \tag{2.2}$$

where $l_A^\dagger$ denotes the transition density and x_{t-1} denotes again the vector of state variables (e.g. the first L lagged variables of the series itself, $(y_{t-1}, y_{t-2}, \ldots, y_{t-L})$). Since it is assumed that l_A is available in closed form we can devise the Maximum Likelihood estimator $\tilde{\omega}^{(ML)}$ of the auxiliary parameter ω based on a sample of size n of observed data $(\tilde{y}_t)_{t=1}^n$ as

$$\tilde{\omega}^{(ML)} = \operatorname*{argmax}_{\omega\in\Omega} l_A(\tilde{y},\omega).$$

As mentioned in the previous chapter, indirect summary statistics can be devised by following the *Efficient Method of Moments* (EMM) approach of Gallant and Tauchen [54]. ABC-IS thus bases summary statistics $S_{IS} \equiv S_{IS}(y,\omega)$ on the score of the auxiliary model, i.e.

$$S_{IS}(y,\omega) = \frac{\partial}{\partial\omega'}\log l_A\left(y,\omega'\right)\Big|_{\omega'=\omega}. \tag{2.3}$$

Similarly to EMM, the question arises how the different summary statistics (or more precisely, deviations between summary statistics based on observed and simulated data) should be weighted in the ABC-IS approach. In the EMM approach, the weighting is irrelevant in the exactly identified case, i.e. if the dimensions of the parameter spaces of

the auxiliary model and of the structural model are identical. In contrast, if the structural model is overidentified, i.e. the parameter space of the auxiliary model exceeds that of the structural model, then the specification of the weighting matrix has an impact on the (asymptotic) variance of the estimator and it is therefore crucial to choose a weighting matrix that minimizes the asymptotic variance. In the Indirect ABC setup, however, the specification of the weighting matrix is relevant in both cases: in the case of overidentification as well as in the case of exact identification. The reason is that in contrast to the frequentist interpretation, we are not only interested in the minimum of the objective function but, more generally, in the whole posterior distribution. To see this, note that for a given tolerance level ϵ, a simple weighting of the summary statistics implies an acceptance area that is a p-dimensional ball centered around the summary statistic of the data. However, not all statistics carry the same information content and, thus, different weights may be helpful. Moreover, if two statistics are highly dependent, it is preferable to take this property into account by choosing an appropriate weighting scheme. Hence, a careful weighting of the summary statistics is a relevant issue.

One of the main advantages in using a score based indirect summary statistic (ABC-IS) is the possibility to obtain an efficient weighting scheme for the individual entries of the summary statistic vector S. Note that the score based summary statistic S_{IS} in (2.3) evaluated at the observed data is zero for all values of θ, i.e. $S_{IS}(\tilde{y}, \tilde{\omega}^{(ML)}) = 0$, by construction. Hence, when comparing the distance between score based summary statistics evaluated at observed and simulated data, we determine how far $S_{IS}(\hat{y}, \tilde{\omega}^{(ML)})$ is away from zero. Following the analogy of ABC-IS to EMM, we therefore consider the weighted quadratic form which, in the frequentist setting, is a well established weighting scheme with desirable properties. For $\mathcal{S}$ denoting the space of summary statistics it is given by

$$d_{\mathcal{I}^{-1}} : \mathcal{S} \to \mathbb{R}, \;\; s \mapsto s^{\mathsf{T}} \mathcal{I}_A^{-1} s$$

with

$$\mathcal{I}_A = \mathbb{V}\left(\frac{\partial \ln l_A(Y, \omega)}{\partial \omega}\right) = -\mathbb{E}\left(\frac{\partial^2 \ln l_A(Y, \omega)}{\partial \omega \partial \omega^{\mathsf{T}}}\right) \tag{2.4}$$

denoting the corresponding information matrix (of the auxiliary model). Depending on the auxiliary model, an estimator for the information matrix can very often derived in a straightforward way. Under the assumption of stationarity, an estimator for the in-

formation matrix in (2.4) based on observed data $\tilde{y} = (\tilde{y}_t)_{t=1}^n$ is given by

$$\tilde{\mathcal{I}}_A = \sum_{t=1}^{n} \left[\frac{\partial}{\partial \omega} \log l_A(\tilde{y}_t | \tilde{x}_{t-1}; \omega) \Big|_{\omega=\tilde{\omega}^{(ML)}} \right] \left[\frac{\partial}{\partial \omega} \log l_A(\tilde{y}_t | \tilde{x}_{t-1}; \omega) \Big|_{\omega=\tilde{\omega}^{(ML)}} \right]^{\mathsf{T}}.$$

The resulting distance between summary statistics S for the ABC-IS method is then given by

$$S_{IS}(\hat{y}, \tilde{\omega}^{(ML)})^{\mathsf{T}} \quad \tilde{\mathcal{I}}_A^{-1} \quad S_{IS}(\hat{y}, \tilde{\omega}^{(ML)}). \tag{2.5}$$

This weighting scheme accounts for different variances of the different elements in the score and, moreover, appropriately captures the dependence structure of the summary statistics. In contrast to the case with equal weights (i.e., where $\tilde{\mathcal{I}}_A^{-1}$ is the identity matrix) this weighting implies an (possibly rotated) ellipsoid for the acceptance area.

2.3 Simulation study

To illustrate the effect that the introduced weighting scheme has on the approximation properties of the ABC-IS algorithm we recapture the setup of the simulation study of the previous chapter. As before, we take the structural model to be given by

$$\mathcal{M}_S = \left(\mathbb{R}_+^n, \left\{ \mathcal{E}(\lambda)^{\otimes n}, \lambda > 0 \right\} \right),$$

where the observed data $(\tilde{y}_t)_{t=1}^n$ is sampled independently and identically from the exponential distribution with parameter λ such that the associated likelihood function of the structural model (with respect to Lebesgue measure) reads as

$$l_S(y, \lambda) = \lambda^n \exp\left(-\lambda \sum_{t=1}^{n} y_t \right).$$

We again consider the conjugate prior on the parameter λ of the structural model $\mathcal{M}_S$ to be $\lambda \sim \mathcal{G}\left(\alpha^{(\pi)}, \beta^{(\pi)}\right)$ with $\mathcal{G}$ denoting the gamma distribution such that the closed form solution of the posterior distribution is given by

$$\lambda | y \sim \mathcal{G}\left(\alpha^{(\pi)} + n,\ \beta^{(\pi)} + \sum_{t=1}^{n} y_t \right). \tag{2.6}$$

The auxiliary model is specified in terms of gamma distributed observations, i.e.

$$\mathcal{M}_A = \left(\mathbb{R}_+^n, \left\{\mathcal{G}\left(\alpha,\beta\right)^{\otimes n}, (\alpha,\beta) \in (\mathbb{R}_+ \times \mathbb{R}_+)\right\}\right).$$

The simulation setup is as follows. For a given prior distribution (specified by α^π and β^π) we simulate $1,000$ samples of 100 exponentially distributed random variables with $\lambda_0 = 1$ (implying a mean and variance of one) representing our observed data. For every sample we compute the exact posterior distribution according to (2.6) and approximate the posterior distribution using both the unweighted (ABC-IS*) and the weighted (ABC-IS) version of the score-based indirect ABC methods. To obtain the approximate posterior distribution we simulate for every observed data set $100,000$ proposal draws from the prior distribution and compute the distance implied by the different ABC approaches to the observed data. Discarding 99% of these proposals leaves us for each ABC method $1,000$ draws from the approximate posterior distribution.

As before, different statistics are considered to measure the distance between the true posterior distribution and the approximations and for each statistic the average over all $1,000$ samples is computed. This is repeated for several combinations of $\alpha^{(\pi)}$ and $\beta^{(\pi)}$ to analyze the effect of different priors. The results are presented in Table 2.1 and 2.2. Whereas in Table 2.1 we consider the distance between the true posterior mean and the mean of the posterior approximation based on ABC-IS* and ABC-IS, Table 2.2 considers the chi-squared distance

$$\sum_{i=1}^{20} \frac{(o_i - e_i)^2}{e_i}$$

where e_i is the expected number of observations in cell i (obtained from the exact posterior distribution for a given simulated data set) and o_i is the number of realized observations in cell i. We use 20 cells that are computed in such a way that every cell contains 5% of the data, i.e. for the i-th cell we choose the interval given by

$$[Q(0.05(i-1)), Q(0.05i)]$$

with $Q : [0,1] \to \mathbb{R}_+ \cup \infty$ denoting the quantile function of the posterior distribution (2.6), $Q(0) = 0$ and $Q(1) = \infty$.

As the results in Tables 2.1 and 2.2 indicate, employing a suitable weighting scheme greatly improves the approximation properties of the ABC-IS algorithm. This is indeed

estimator	moments of the prior distribution			
	variance	mean		
		0.5	1	2
ABC-IS*	0.25	0.0103	0.0057	0.0288
ABC-IS		0.0028	0.0028	0.0163
ABC-IS*	0.5	0.0127	0.0069	0.0122
ABC-IS		0.0036	0.0030	0.0089
ABC-IS*	1	0.0168	0.0086	0.0095
ABC-IS		0.0051	0.0033	0.0047
ABC-IS*	2	0.0209	0.0121	0.0091
ABC-IS		0.0074	0.0038	0.0040
ABC-IS*	4	0.0287	0.0145	0.0101
ABC-IS		0.0137	0.0048	0.0039

Table 2.1: **Distance between the true posterior mean and the posterior mean based on different ABC methods.** This table shows the distance between the true posterior mean and the posterior mean based on different ABC methods and different prior distributions that are specified by their mean and variance. ABC-IS* corresponds to the unweighted Indirect ABC approach with score based summary statistics, while ABC-IS considers the inverse information matrix as weighting matrix.

the case uniformly over different specifications of the mean and variance of the prior distributions. If we furthermore compare these results to the ABC-IL method discussed in the previous chapter, weighted ABC-IS shows the best performance of all indirect ABC methods and has the additional advantage of being computationally efficient. These considerations make weighted ABC-IS an ideal contender to be employed in a challenging estimation problem such as the one considered in the next sections.

estimator	variance	mean		
		0.5	1	2
ABC-IS*	0.25	46.9136	27.0765	377.0657
ABC-IS		28.2088	20.9334	285.9329
ABC-IS*	0.5	60.3490	31.1686	57.9846
ABC-IS		34.4208	22.2936	47.4405
ABC-IS*	1	94.7804	40.0977	51.4272
ABC-IS		49.8150	25.0106	29.3161
ABC-IS*	2	134.7983	58.0195	43.1016
ABC-IS		84.0525	32.0681	26.9558
ABC-IS*	4	276.6809	78.0879	45.2486
ABC-IS		204.8259	42.8542	28.1143

Table 2.2: **Chi-squared distance between the true posterior distribution and the posterior distribution based on different ABC methods.** This table shows the chi-squared statistics for measuring the distance between the true posterior distribution and the posterior distribution based on different ABC methods and different prior distributions that are specified by their mean and variance. The results are reported for the ABC methods and the prior distributions discussed in Table 2.1.

2.4 The structural model: An Ornstein-Uhlenbeck type stochastic volatility model

Our structural model $\mathcal{M}_S$ is defined in terms of the following two stochastic differential equations:

$$\mathrm{d}\, x^*(t) = (\mu + \beta\sigma^2(t))\,\mathrm{d}\, t + \sigma(t)\,\mathrm{d}\, W(t) \tag{2.7}$$

$$\mathrm{d}\,\sigma^2(t) = -\lambda\sigma^2(t)\,\mathrm{d}\, t + \mathrm{d}\, Z(\lambda t). \tag{2.8}$$

Here we denote with $(x^*(t))_{t\geq 0}$ the log price process of an asset, $(W(t))_{t\geq 0}$ is a standard Brownian motion and $\left(\sigma^2(t)\right)_{t\geq 0}$ is the underlying latent *instantaneous volatility process* of OU type, independent of $(W(t))_{t\geq 0}$, with $(Z(\lambda t))_{t\geq 0}$ being the *background driving Lévy process* (BDLP). The parameters μ and β are denoting the drift and risk premium, respectively, in the SDE for the log-price $x^*(t)$ (2.7), whereas the parameter λ governs both the exponential decay of $\sigma^2(t)$ and the rate at which jumps in (instanta-

neous) volatility occur in (2.8). As in Barndorff-Nielsen and Shephard [9], we use the unusual timing concept of λt instead of t for the BDLP as in that case the marginal distribution of $\sigma^2(t)$ turns out to be independent of the parameter λ. The distribution (and qualitative properties) of the volatility process depends moreover on the specification of the BDLP of which several proposals have been made in the literature. For example it is possible to specify a process with finite jump activity or with infinite jump activity. A widely used process is the Gamma-OU process which implies a gamma $\mathcal{G}(\alpha, \delta)$ law for the marginal distribution of the process $(\sigma^2(t))_{t\geq 0}$. Several studies show the adequacy of this process for modeling financial time series, e.g. see Griffin and Steel [64] and Frühwirth-Schnatter and Sögner [48], who also consider the Bayesian estimation of these processes.

Consider now *aggregated returns* over an interval of length Δ, given by

$$y_n = \int_{(n-1)\Delta}^{n\Delta} \mathrm{d}\, x^*(t) = x^*(n\Delta) - x^*((n-1)\Delta).$$

Using the discretization of the so-called *actual volatility process* in Barndorff-Nielsen and Shephard [9, Equation 3] we can write

$$\begin{aligned}\sigma_n^2 &= \sigma^{2*}(n\Delta) - \sigma^{2*}((n-1)\Delta) = \int_{(n-1)\Delta}^{n\Delta} \sigma^2(u)\,\mathrm{d}\,u \\ &= \frac{1}{\lambda}\left[Z(\lambda n\Delta) - Z(\lambda(n-1)\Delta) - (\sigma^2(n\Delta) - \sigma^2((n-1)\Delta))\right],\end{aligned}$$

where we denote by $\left(\sigma^{2*}(t)\right)_{t\geq 0}$ the *integrated volatility process*. It can be shown that the conditional distribution of y_n given σ_n^2 is given by

$$y_n|\sigma_n^2 \sim \mathcal{N}(\mu\Delta + \beta\sigma_n^2, \sigma_n^2).$$

Nonetheless, exploiting this result for estimation purposes is difficult since the conditional distribution of y_n, although normal, depends on the latent processes σ^2. Consequently, a standard Maximum Likelihood approach is not feasible in this context since the likelihood function of the structural parameter vector $\theta = (\mu, \beta, \lambda, \alpha, \delta)$ takes no explicit form.

Another problem that arises from the formulation of the stochastic volatility model as in (2.7) - (2.8) is that it fails to capture the dependence structure between squared returns properly. Barndorff-Nielsen and Shephard [9, Equation 44] show that if the OU

process for the instantaneous volatility process $\sigma^2(t)$ admits a finite variance, the correlation between squared returns will be given by

$$cor(y_n^2, y_{n+s}^2) = C \exp\{-\lambda\Delta(s-1)\}, \quad s > 0,$$

for C some constant that depends on the (finite) variance of $\sigma^2(t)$, thus implying an exponential decay in the autocorrelation function of squared returns. Empirical evidence, however, suggests that the autocorrelation function of squared returns initially falls very steeply and decays rather slowly at greater lags (see e.g. Ding and Granger [39]). This dependence structure cannot be modeled by a single OU process. However, a linear combination or *superposition* of several OU processes can give rise to such a dependence structure. We therefore consider the instantaneous volatility process σ^2 to be given by a sum of m Gamma-OU processes σ_i^2, each with a respective BDLP Z_i and parametrized with parameter λ_i and marginal distribution parameters α_i, δ_i. Formally, we have

$$\sigma^2(t) = \sum_{i=1}^{m} \sigma_i^2(t),$$

with σ_i^2 being the solution of

$$\mathrm{d}\,\sigma_i^2(t) = -\lambda_i \sigma_i^2(t)\,\mathrm{d}\,t + \mathrm{d}\,Z_i(\lambda_i t), \quad i = 1, \ldots, m.$$

Particularly, such a formulation leads to an autocorrelation function of squared returns given by

$$cor(y_n^2, y_{n+s}^2) = \sum_{i=1}^{m} C_i \exp\{-\lambda_i\Delta(s-1)\}, \quad s > 0,$$

where the constants $C_i, i = 1, \ldots, m$ depend on the (finite) variances of the processes $\sigma_i^2(t)$. As we will see in Section 2.6, a superposition of only $m = 2$ components is already sufficient to capture the dependence structure appropriately. Naturally, the results for the special case $m = 1$ from above extend easily, and one has

$$y_n | \sigma_n^2 \sim \mathcal{N}(\mu\Delta + \beta\sigma_n^2, \sigma_n^2).$$

where now the actual volatility is given by

$$\sigma_n^2 = \sum_{i=1}^{m} \frac{1}{\lambda_i} \left[Z_i(\lambda_i n\Delta) - Z_i(\lambda_i(n-1)\Delta) - (\sigma_i^2(n\Delta) - \sigma_i^2((n-1)\Delta)) \right]. \tag{2.9}$$

2.5 The auxiliary model: A semi-nonparametric density approach

Using the Indirect ABC procedure requires us to specify an auxiliary model that is both analytically tractable and provides a good approximation of the true data generating process. The SNP model as introduced by Gallant and Nychka [52] seems to meet these requirements: it is certainly analytically tractable and its empirical adequacy for our data set is demonstrated in Section 2.6. In the following we briefly review the SNP model.[1]

To this end consider the location-scale transformation for y_t to be of the form

$$y_t = \mu_{x_{t-1}} + R_{x_{t-1}} z_t \tag{2.10}$$

where the innovation is denoted as z_t and where the subscripts indicate the dependence of R and μ on the lagged state vector x_{t-1} for reasons to become clear below.

The SNP density approach is based on the fact that a Hermite expansion can be used as a general purpose approximation to a density function. More precisely, we expand the square root of an innovation density h in a Hermite expansion and truncate the infinite polynomial at some integer K_z which, together with other tuning parameters of the SNP density, has to be determined through a model selection criterion (such as BIC). Now we take the leading term of the Hermite expansion to follow a Gaussian GARCH model. In other words, we expand $\sqrt{h(z)}$ into a polynomial in z of degree K_z whose coefficients are polynomials in x of degree K_x. The truncated density of an innovation z_t given past values of y_t up to lag L (which we denoted x_{t-1}) can now be written as

[1] Considering our example of a stochastic volatility model for financial data it may be somewhat natural to choose a GARCH model as our auxiliary model. GARCH models allow us to reproduce several stylized facts of financial data such as dependence in conditional variances, skewness and excess kurtosis. However, standard GARCH models for financial data aim at a parsimonious model structure for financial applications, such that these models may not be the appropriate choice for an auxiliary model which should provide a detailed description of the data.

$$h_K(z_t \mid x_{t-1}) = \frac{\mathfrak{P}^2(z_t, x_{t-1})\varphi(z_t)}{\int \mathfrak{P}^2(u, x_{t-1})\varphi(u)\mathrm{d}u}$$
$$= \frac{\left(\sum_{|\alpha|=0}^{K_z}\left(\sum_{|\beta|=0}^{K_x} a_{\alpha\beta} x_{t-1}^{\beta}\right) z_t^{\alpha}\right)^2 \varphi(z_t)}{\int \left(\sum_{|\alpha|=0}^{K_z}\left(\sum_{|\beta|=0}^{K_x} a_{\alpha\beta} x_{t-1}^{\beta}\right) u^{\alpha}\right)^2 \varphi(u)\mathrm{d}u}.$$

The normalization in the denominator is necessary for h_K to integrate to one. In the expression above, φ denotes the standard normal density function, β is an index vector of length L and $x^{\beta} = \prod_{i=1}^{L} x_i^{\beta_i}$. Note furthermore that since

$$\mathfrak{P}^2(z, x) / \int \mathfrak{P}^2(u, x)\varphi(u)\mathrm{d}u$$

is a homogeneous function of the coefficients of the polynomial $\mathfrak{P}$, $\mathfrak{P}$ can only be determined up to a scalar multiple. To achieve a unique representation, the constant term a_{00} of the polynomial $\mathfrak{P}$ is normalized to unity. Therefore, h_K can be interpreted as a series expansion whose leading term is the normal density φ and whose higher-order terms induce departures from normality.

To complete our auxiliary model we have to specify the scale-location transformation in (2.10). As mentioned in Gallant and Tauchen [53] it proves advantageous in applications to allow the scale $R_{x_{t-1}}$ to explicitly depend on the lagged state vector x_{t-1} as that reduces the degree K_x required to obtain a good approximation for our structural model $l_S^{\dagger}$. We follow Chernov et al. [26] and specify $R_{x_{t-1}}$ by a univariate GARCH-like specification:

Therefore, let

$$R_{x_{t-1}}^2 = \beta_0 + \sum_{i=1}^{L_r} \beta_i \left(y_{t-1-L_r+i} - \mu_{x_{t-2-L_r+i}}\right)^2 + \sum_{i=1}^{L_g} \gamma_i R_{x_{t-2-L_g+i}}^2.$$

For the location $\mu_{x_{t-1}}$ we propose an AR process of the form

$$\mu_{x_{t-1}} = \alpha_0 + \sum_{i=1}^{L_\mu} \alpha_i \, x_{t-2-L_\mu+i},$$

where the lag-length of $\mu_{x_{t-1}}$ is denoted by L_μ. In summary, the different lag-lengths L_r, L_g, L_μ govern the location-scale transformation for y_t and hence, determine the nature of the leading term of the Hermite expansion. On the other hand, K_z, K_x govern the degree of the polynomial of the Hermite expansion and, hence, determine the nature of the innovation process z_t. Finally, a change of variable is the last step in our derivation of the SNP density of the auxiliary model and leads to the likelihood of the auxiliary model:

$$l_A(y, \omega) = \prod_{t=1}^{T} \frac{h_K\left(R_{x_{t-1}}^{-1}(y_t - \mu_{x_{t-1}}) \mid x_{t-1}\right)}{R_{x_{t-1}}},$$

where all auxiliary parameters are collected in ω.

2.6 Estimation results

Our study of the OU type stochastic volatility model is based on daily stock returns of the IBM stock ranging from 1990/01/03 to 2011/12/30, yielding 4787 observations. Figure 2.1 illustrates the usual stylized facts, such as volatility clustering and fat tails in the return data.

Specification of the auxiliary model

The key for a successful application of indirect methods in general is the specification of an auxiliary model. It has to both capture the characteristic information of the data, and be computationally tractable[2]. Naturally, there is little hope to find a universal model suited for every application. However, in the context of financial data analysis, the SNP densities introduced above turn out to provide an adequate description (see, e.g. Chernov et al. [26]).

When fitting an SNP auxiliary model to our data set, we rely on the Bayesian information criterion (BIC). In a first step, we increase the order of the auto-regressive polyno-

[2]This has to hold particularly for the ABC-IP and ABC-IL (or SQML and II) approach as they require the estimation of the auxiliary model for *every* proposal from the prior distribution.

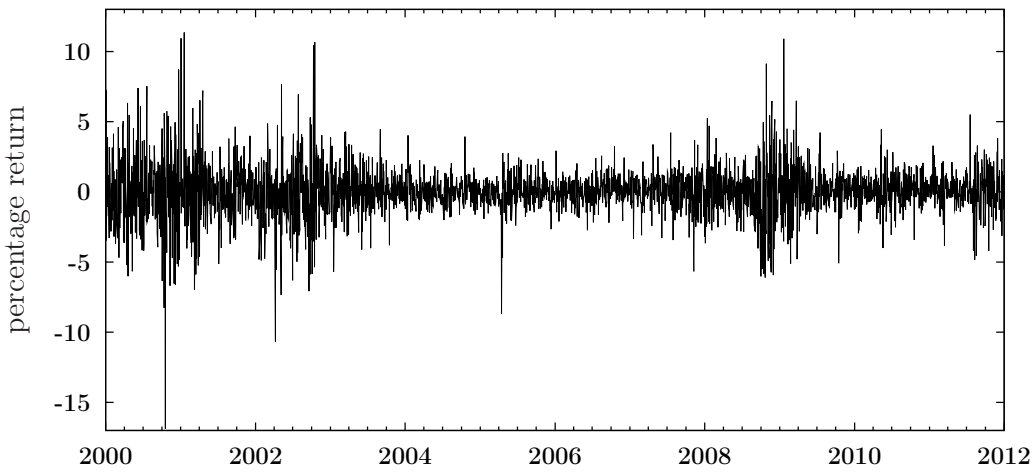


Figure 2.1: **Daily returns.** Time series plot of the daily percentage logarithmic return. The panel shows the evolvement of the return of the IBM stock (January 3rd, 1990 until December 30th, 2011).

mial and then the order of the GARCH model. This results in an AR(1)-GARCH(1,1) model, and the polynomial order K_x is chosen to be 8 whereas K_z is taken to be zero. A more detailed description of the model selection procedure can be found in Andersen et al. [3] and Chernov et al. [26]. Table 2.3 reports the parameter estimates and the corresponding asymptotic t-values.

Prior distributions

In our application we estimate the Ornstein-Uhlenbeck type stochastic volatility model from Section 2.4 where we consider modeling the instantaneous volatility σ^2 both as a single OU process and as a superposition of several OU processes. These OU processes are taken to have marginal gamma distribution so that the structural parameter vectors are given by

$$\theta_1 = (\mu, \beta, \lambda, \alpha, \delta)$$
$$\theta_2 = (\mu, \beta, \lambda_1, \lambda_2, \ldots, \lambda_m, \alpha_1, \alpha_2, \ldots, \alpha_m, \delta_1, \delta_2, \ldots, \delta_m)$$

respectively.

In accordance with the literature (see, e.g. Frühwirth-Schnatter and Sögner [48]; Griffin and Steel [64] and Roberts et al. [97]) we found in preliminary analyses the risk premium β to be negligible and thus exclude it from the structural parameter vectors

parameter	estimate	t-values
μ_0	0.1893	3.4158
β_0	0.0339	6.1020
β_1	0.0626	8.9223
γ_1	0.9364	152.5255
$\alpha_{0,1}$	-0.0881	-2.6380
$\alpha_{0,2}$	-0.2241	-10.5103
$\alpha_{0,3}$	0.0289	2.1246
$\alpha_{0,4}$	0.0514	7.1445
$\alpha_{0,5}$	-0.0033	-1.5877
$\alpha_{0,6}$	-0.0043	-4.9306
$\alpha_{0,7}$	0.0001	1.1581
$\alpha_{0,8}$	0.0002	4.7374

Table 2.3: **Auxiliary model.** Reported are the parameter estimates and t-values of the auxiliary model.

θ_1 and θ_2 in the remainder of this section. Following Griffin and Steel [64], we choose a weakly informative prior on the drift parameter μ, taken to be the $\mathcal{N}(0, 90)$ Normal distribution with mean zero and variance 90. In the case where we consider only one OU process, we follow the above mentioned authors and choose as a prior for the parameter λ the $\mathcal{G}(1, 1)$ Gamma distribution with shape and rate one which implies a $\mathcal{B}(1, 1)$ prior for the autocorrelation $e^{-\lambda\Delta}$ of the instantaneous volatility process σ^2. The priors on the shape (α) and rate (δ) parameters of the marginal Gamma distribution of σ^2 are then taken to be given by the $\mathcal{G}(1, 1)$ and $\mathcal{G}(1, 100)$ Gamma distributions, respectively.

The case where we consider a superposition of m OU processes to model σ^2 is more involved since, as Frühwirth-Schnatter and Sögner [48] remark, the superposition model in Section 2.4 is identified only up to relabeling the indices of the m components since the expression for the actual volatility in (2.9) is invariant under such relabeling. As Frühwirth-Schnatter and Sögner [48] point out, in the case of $m = 2$ the likelihood function of the parameters $\theta = (\mu, \beta, \lambda_1, \lambda_2, \alpha_1, \alpha_2, \delta_1, \delta_2)$ is the same as the one for the relabeled parameters $\theta^\star = (\mu, \beta, \lambda_2, \lambda_1, \alpha_2, \alpha_1, \delta_2, \delta_1)$, thus giving rise to two equivalent modal regions. Using an MCMC sampler based on symmetric independence

priors for the parameters in the superposition model of the form

$$p(\lambda_1, \lambda_2, \ldots, \lambda_m, \alpha_1, \alpha_2, \ldots, \alpha_m, \delta_1, \delta_2, \ldots, \delta_m) = \prod_{i=1}^{m} p(\lambda_i) p(\alpha_i) p(\delta_i)$$

can lead to label-switching if the local modes are not well-separated. Frühwirth-Schnatter and Sögner [48] thus suggest to use an asymmetric prior of the form

$$\begin{aligned} &p(\lambda_1, \lambda_2, \ldots, \lambda_m, \alpha_1, \alpha_2, \ldots, \alpha_m, \delta_1, \delta_2, \ldots, \delta_m) \\ &\qquad = p(\lambda_1, \lambda_2, \ldots, \lambda_m) \prod_{i=1}^{m} p(\alpha_i) p(\delta_i), \end{aligned}$$

where

$$p(\lambda_1, \lambda_2, \ldots, \lambda_m) = p(\lambda_1) \prod_{i=2}^{m} p(\lambda_i | \lambda_{i-1}).$$

The prior on λ_1 is taken to be the $\mathcal{E}(1)$ Exponential distribution, whereas for $i = 2, \ldots, m$ we consider, given the value for λ_{i-1}, the parameter λ_i to follow an $\mathcal{E}(1)$ Exponential distribution left truncated at λ_{i-1}. Concerning the parameters of the marginal Gamma distribution of the OU processes σ_i^2 we adopt the following priors. The shape parameters $(\alpha_1, \alpha_2, \ldots, \alpha_m)$ are taken to independently follow a $\mathcal{G}(1,1)$ Gamma distribution whereas we restrict the rate parameters $\delta_1 = \delta_2 = \ldots = \delta_m \equiv \delta$ which we take to follow a $\mathcal{G}(1, 100)$ Gamma distribution. The prior on the drift parameter μ will, as before, be given by a $\mathcal{N}(0, 90)$ Normal distribution.

Estimation results

We first estimate the stochastic volatility model in (2.7) based on the instantaneous volatility process being modeled by only one non-Gaussian Ornstein-Uhlenbeck process. The structural parameter vector of interest is thus given by $\theta_1 = (\mu, \lambda, \alpha, \delta)$. We generate $3,000,000$ proposal draws of θ_1 and retain the 0.1% percentile of sampled distances, yielding a total of $3,000$ draws from the approximate posterior distribution. Figure 2.2 shows the corresponding (marginal) histograms for the structural parameters as well as the estimated autocorrelation function of the volatility process σ_n^2. As is to be expected, such a simple formulation cannot depict the complex dependence structure appropriately, as indicated by the exponential decay of the autocorrelation function (see

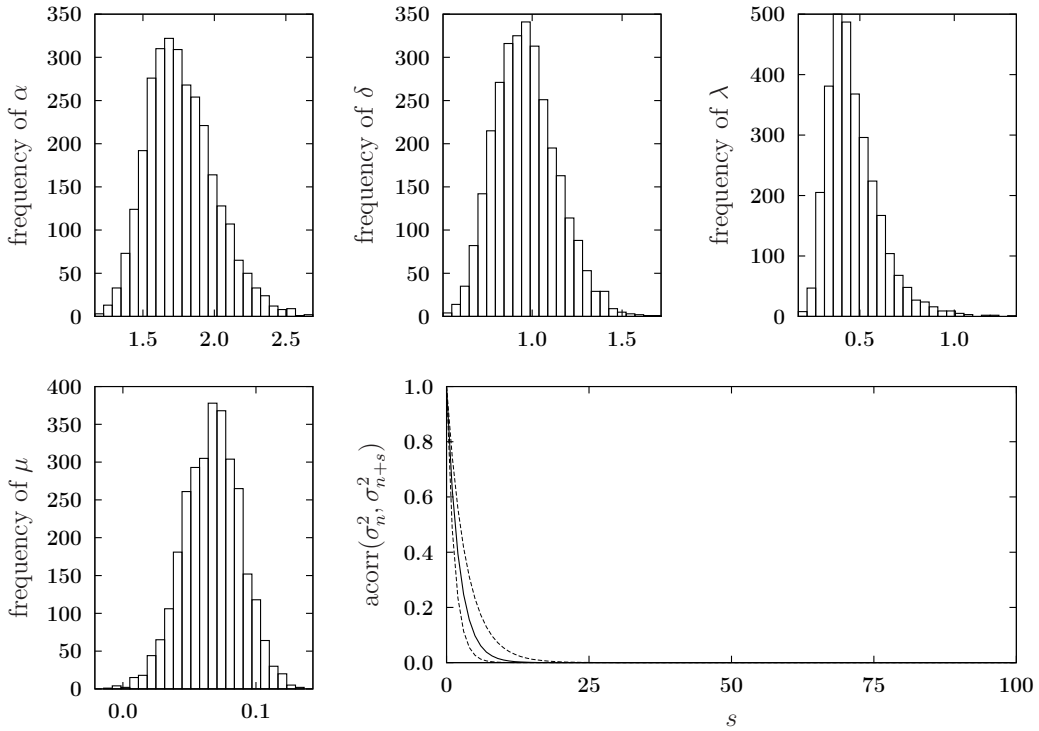


Figure 2.2: **Histograms and autocorrelation function.** This figure shows marginal histograms of the structural parameters $\alpha, \delta, \lambda, \mu$ based on $3,000$ accepted draws from the ABC-IS approximation to the posterior distribution based on a single OU process. The tolerance level ϵ is chosen implicitly through retaining the 0.1% percentile of sampled distances. The estimated autocorrelation function of the volatility process σ_n^2 is plotted.

also Frühwirth-Schnatter and Sögner [48] and Chernov et al. [26]).

We thus proceed with estimating the model in (2.7) with the instantaneous volatility process given by a superposition of two non-Gaussian Ornstein-Uhlenbeck processes. In addition, we follow Frühwirth-Schnatter and Sögner [48] and restrict the rate parameters of the marginal Gamma distribution of the respective OU processes by taking $\delta_1 = \delta_2 \equiv \delta$ such that the structural parameter vector of interest is now given by $\theta_2 = (\mu, \lambda_1, \lambda_2, \alpha_1, \alpha_2, \delta)$. The tolerance level is again chosen implicitly by discarding 99% of the proposed draws of θ_2. Figure 2.3 shows the (marginal) histograms for the sampled structural parameters $\mu, \lambda_1, \lambda_2, \alpha_1, \alpha_2, \delta$ based on $3,000$ accepted draws of the ABC-IS algorithm. Summary statistics of the respective (marginal) posterior distributions are reported in 2.4.

Furthermore, the last row in Figure 2.3 shows the (estimated) autocorrelation function of the individual volatility processes $\sigma_1^2(n)$ and $\sigma_2^2(n)$ as well as their superposition

parameter	feature of the posterior distribution		
	mean	median	standard deviation
δ	0.5562	0.5581	0.1385
α_1	0.6059	0.6071	0.1683
λ_1	0.0140	0.0127	0.0098
α_2	0.6902	0.6808	0.2139
λ_2	0.9666	0.8360	0.4585
μ	0.0685	0.0667	0.0129
α_1/δ	1.1659	1.0837	0.4811
α_1/δ^2	2.4529	1.9283	1.9698
$\lambda_1\alpha_1$	0.0086	0.0073	0.0067
$e^{-\lambda_1}$	0.9861	0.9873	0.0097
α_2/δ	1.2318	1.2239	0.1840
α_2/δ^2	2.3595	2.2254	0.7162
$\lambda_2\alpha_2$	0.6410	0.5918	0.3008
$e^{-\lambda_2}$	0.4158	0.4334	0.1537

Table 2.4: **Summary statistics of posterior distribution.** The mean, median and standard deviation of several parameters are presented, based on draws from the posterior distribution.

$\sigma^2(n) = \sigma_1^2(n) + \sigma_2^2(n)$. As is to be expected, the first OU process clearly models long-term persistence in the volatility process, indicated by a rather small value of the decaying rate λ_1. Opposed to that, the second OU process exhibits an autocorrelation function that decays sharply and thus reflects short-term variation in the volatility process which is indicated by a much larger value of λ_2. As such, a superposition of just two OU processes can give rise to an adequate representation of the dependence structure in that the joint autocorrelation function decays sharply at initial lags whereas the decay is much slower at longer lags.

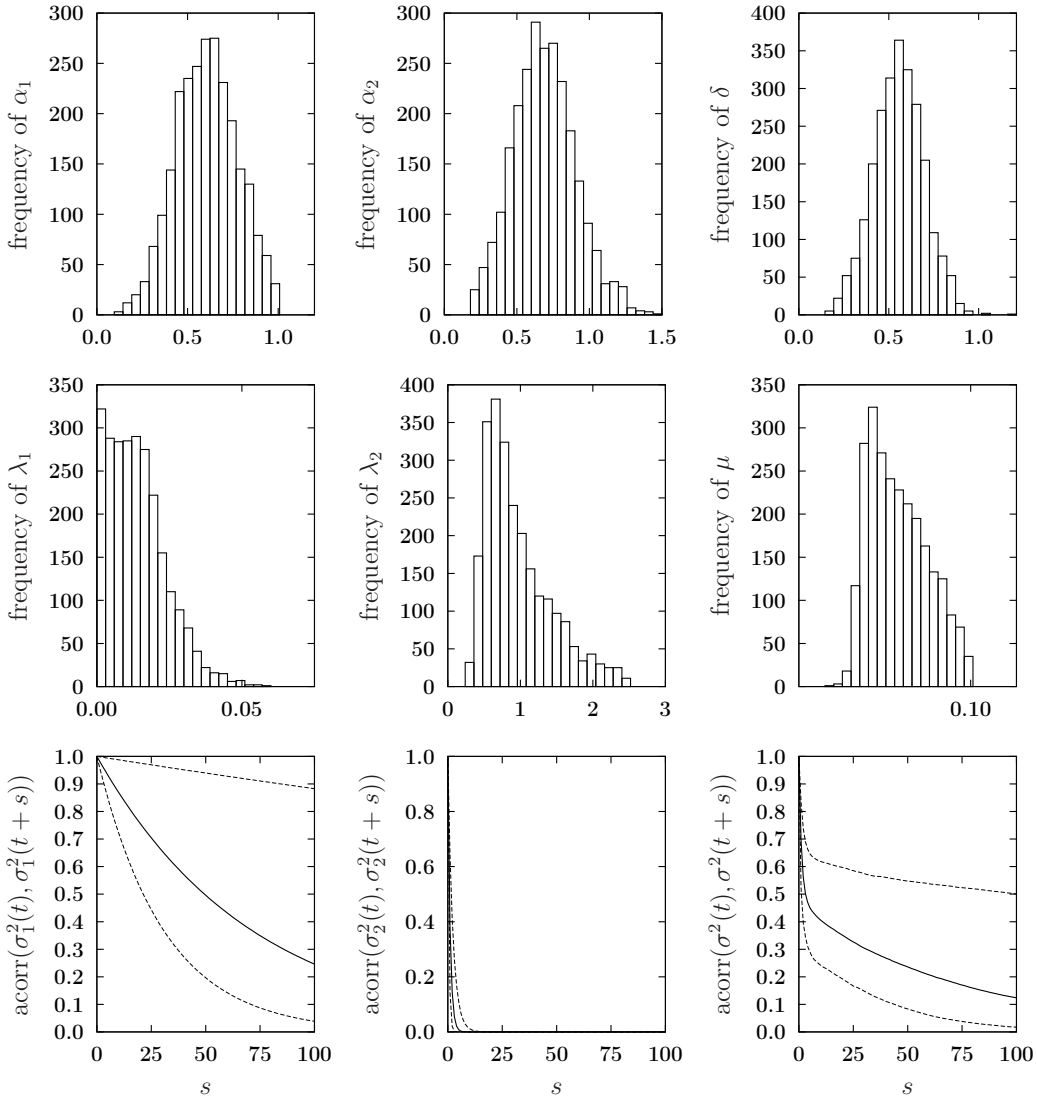


Figure 2.3: **Histograms and autocorrelation functions.** This figure shows marginal histograms of the structural parameters $\alpha_1, \alpha_2, \delta, \lambda_1, \lambda_2, \mu$ based on $3,000$ accepted draws from the ABC-IS approximation to the posterior distribution based on the superposition of two OU processes. The tolerance level ϵ is chosen implicitly through retaining the 0.1% percentile of sampled distances. The third row shows the estimated autocorrelation functions of the individual volatility processes $\sigma_1^2(n)$ and $\sigma_2^2(n)$ as well as their superposition $\sigma^2(n) = \sigma_1^2(n) + \sigma_2^2(n)$.

2.7 Conclusion

In this paper we extended the notion of Indirect ABC methods by proposing an efficient way of weighting the individual entries of the vector of summary statistics obtained from the score-based Indirect ABC approach (ABC-IS). In particular, the weighting matrix was given by the inverse of the asymptotic covariance matrix of the score vector of the auxiliary model and allowed us to appropriately assess the distance between the true posterior distribution and the approximation based on the ABC-IS method. We illustrated the performance gain in a simulation study. An empirical application then implemented the weighted ABC-IS method to the problem of estimating a continuous-time stochastic volatility model based on non-Gaussian Ornstein-Uhlenbeck processes. We showed how a suitable auxiliary model can be constructed and confirmed estimation results from concurring Bayesian estimation approaches suggested in the literature.

3

Exploring multimodal sampling distributions

3.1 Introduction

Economic models have become increasingly more complex over the past decades and their empirical validation therefore heavily relies on efficient and accurate estimation procedures. Traditional econometric approaches adopted a frequentist perspective in that one was seeking to optimize a certain objective function $L : \Theta \times \mathcal{X} \to \mathbb{R}$ over the parameter set Θ and an estimator of the parameter of interest was thus given by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \, L\left(\boldsymbol{\theta}, \boldsymbol{x}\right),$$

with $\boldsymbol{x} \in \mathcal{X}$ denoting the sample. Examples of these estimators include the maximum likelihood (ML) estimator and the generalized method of moments (GMM) estimator. Although these estimators have attractive properties under rather mild assumptions, the computational challenge is to find the global optimum $\hat{\boldsymbol{\theta}}$ in a reasonable amount of time if the estimator is not given as an explicit function. Numerical (standard) optimization algorithms work well if the objective function is *smooth* and if the *global maximum* is the only *local maximum*.

However, the objective function of many important econometric methods and empirical applications is rather inappropriate for these algorithms, and alternative estimation procedures are required. In the light of vastly increasing computational possibilities, Bayesian methods have become more and more popular. Instead of maximizing an objective function they analyze the posterior distribution with density given by

$$\mu\left(\boldsymbol{\theta}|\boldsymbol{x}\right) \propto f\left(\boldsymbol{x}|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right),$$

where $f : \mathcal{X} \times \Theta \to \mathbb{R}$ denotes the likelihood function of a data sample $\boldsymbol{x}$ and π denotes the probability density function (pdf) of the prior distribution on Θ. Parameter estimates are then obtained as functionals of the posterior distribution depending on the loss function, e.g. the posterior mean in the case of a quadratic loss function. Since these quantities are rarely available in closed form, Markov Chain Monte Carlo (MCMC) algorithms, such as the Metropolis-Hastings algorithm or Gibbs-sampling, can be used to sample from the posterior distribution. In this case the posterior distribution is approximated by a sample of simulated parameter values and the estimates are the corresponding statistics of the sample. Informally speaking, MCMC algorithms work by constructing a Markov chain on the parameter space Θ such that its stationary distribution equals the posterior distribution.

In contrast to other Monte Carlo based methods, such as importance sampling, MCMC has the advantage that while the precise construction of an importance function is not required, it is still able to depict the characteristics of $\mu(\boldsymbol{\theta}|\boldsymbol{x})$ (see Robert [95, section 6.3]). These appealing properties can also be used in estimation settings that are not grounded on the Bayesian paradigm. Chernozhukov and Hong [27] for instance proposed the so-called Laplace type estimators (LTEs). Based on a prior distribution with pdf π these estimators are functionals such as the mean, median or other quantiles of the so-called quasi-posterior distribution with density given by

$$\mu\left(\boldsymbol{\theta}|\boldsymbol{x}\right) = \frac{\mathrm{e}^{L(\boldsymbol{\theta},\boldsymbol{x})}\, \pi\left(\boldsymbol{\theta}\right)}{\int_{\Theta} \mathrm{e}^{L(\boldsymbol{\theta},\boldsymbol{x})}\, \pi\left(\boldsymbol{\theta}\right) \mathrm{d}\,\boldsymbol{\theta}} \propto \mathrm{e}^{L(\boldsymbol{\theta},\boldsymbol{x})}\, \pi\left(\boldsymbol{\theta}\right), \tag{3.1}$$

which depends on the choice of the objective function L. If, for example, L is the log likelihood function of the model then μ is the pdf of the posterior distribution in the Bayesian setting. However, the setup is much more general as other objective functions such as GMM-like moment conditions can also be considered. The specific statistic computed from the *quasi-posterior* distribution depends, as in the genuine Bayesian set-

ting, on the loss function that is employed.

The performance of MCMC methods relies on constructing a Markov chain on the parameter space Θ whose empirical distribution gets close to the stationary distribution in a practically reasonable computing time. Of particular importance is the requirement that the Markov chain explores the *relevant sample space* efficiently. If the target distribution of interest is reasonable well behaved (not necessarily unimodal) MCMC methods are not as sensitive as classical estimation/optimization methods. However, this requirement can limit standard MCMC methods severely in cases where the target distribution exhibits areas of high probability mass that are separated from each other by areas of very low probability mass, i.e. such that the relevant areas of the sample space are not connected. For such *multimodal* target distributions MCMC methods are prone to be trapped locally in one of the areas of high probability mass. This behavior is well known in the statistical literature and becomes particularly apparent in the case of mixture models (see Celeux et al. [25]; Jasra et al. [75]; Robert and Casella [96] and the references therein). In an economic context, An and Schorfheide [2] show that the use of unmodified MCMC methods to generate samples from the posterior distribution of dynamic stochastic general equilibrium (DSGE) models may be problematic.

In this paper we propose the use of sequential MCMC methods to generate draws from a possibly ill shaped distribution, e.g. a multimodal (quasi-)posterior distribution. So far, sequential MCMC methods have been primarily used to filter, forecast and smooth in nonlinear state-space models, see e.g. Creal [31], Flury and Shephard [47] and the references therein. In these applications, particles are used to approximate the distribution of the unobserved states. In our setup, however, we generate an artificial sequence of distributions and use particles to approximate these artificial distributions. The initial distribution is chosen such that no distinct modes are available and plain MCMC methods are appropriate to explore the global characteristics of this distribution. Sometimes, it may be even possible to sample directly from the initial distribution. The final distribution is then the original target distribution, e.g. the (quasi-)posterior distribution.

Informally speaking, the proposed method *interpolates* between an easily sampled initial probability distribution μ_0 and the target distribution of interest μ_n in some suitable sense. One starts by generating a large sample of *particles* from the initial probability distribution μ_0 on Θ. These particles are then propagated such that they can be regarded as a sample of particles from the first interpolant μ_1. By iterating this procedure we finally obtain a sample of particles from the target distribution μ_n.

Compared to standard MCMC methods this sequential approach guarantees that

the parameter space is indeed well explored. As we will demonstrate in the simulation study of Section 3.3, the sequential approach successfully both detects high-dimensional modes that are (very) far apart and is able to adjust the probability mass assigned to the respective modes at each iteration of the algorithm. This gain in accuracy is, however, not for free. In fact, the computational burden of the sequential MCMC algorithm is significantly larger than for the plain Metropolis-Hastings method. Nonetheless, the most time-consuming part can be easily and effectively parallelized. Hence, in contrast to classical MCMC methods, this algorithm is computational efficient since it supports modern multicore central processing unit architectures.

This paper is organized as follows. Section 3.2 formally describes the sampling method whereas Section 3.3 investigates the finite sample properties in an extensive simulation study. Section 3.4 concludes.

3.2 The Sequential MCMC sampling method

In this section we formally describe the proposed sequential MCMC sampling procedure. The algorithm we present is based on the Sequential Monte Carlo literature and can be regarded as adaptations or simple special cases of the algorithms proposed in Gordon et al. [60], Cappé et al. [23] and Del Moral et al. [36]. The links of the here proposed method to this literature is discussed below.

3.2.1 Pitfalls of standard sampling approaches

As mentioned in the introduction, using standard sampling approaches to draw from ill-shaped probability distributions can have severe drawbacks. We thus first give a heuristic illustration of possible pitfalls with standard MCMC based sampling schemes. As an illustrative example we consider a bivariate normal mixture model, i.e. our interest is to obtain sampled draws from the distribution with density

$$p(\boldsymbol{\theta}) = \omega f_{\mathcal{N}}\left(\boldsymbol{\theta}; \boldsymbol{\mu}^{(1)}, \boldsymbol{V}^{(1)}\right) + (1-\omega) f_{\mathcal{N}}\left(\boldsymbol{\theta}; \boldsymbol{\mu}^{(2)}, \boldsymbol{V}^{(2)}\right), \qquad (3.2)$$

where $\boldsymbol{\theta} \in \mathbb{R}^2$. This example is constructed in analogy to Chib and Ramamurthy [28] and can be regarded as a simplified representation of the marginal distribution of a DSGE model presented by An and Schorfheide [2]. Of particular interest is to consider the case where the modes are well separated. Following Chib and Ramamurthy [28] we take the modes to be located at $\boldsymbol{\mu}^{(1)} = [1, -1]^\mathsf{T}$ and $\boldsymbol{\mu}^{(2)} = 8 \times \boldsymbol{\mu}^{(1)}$ and we set the

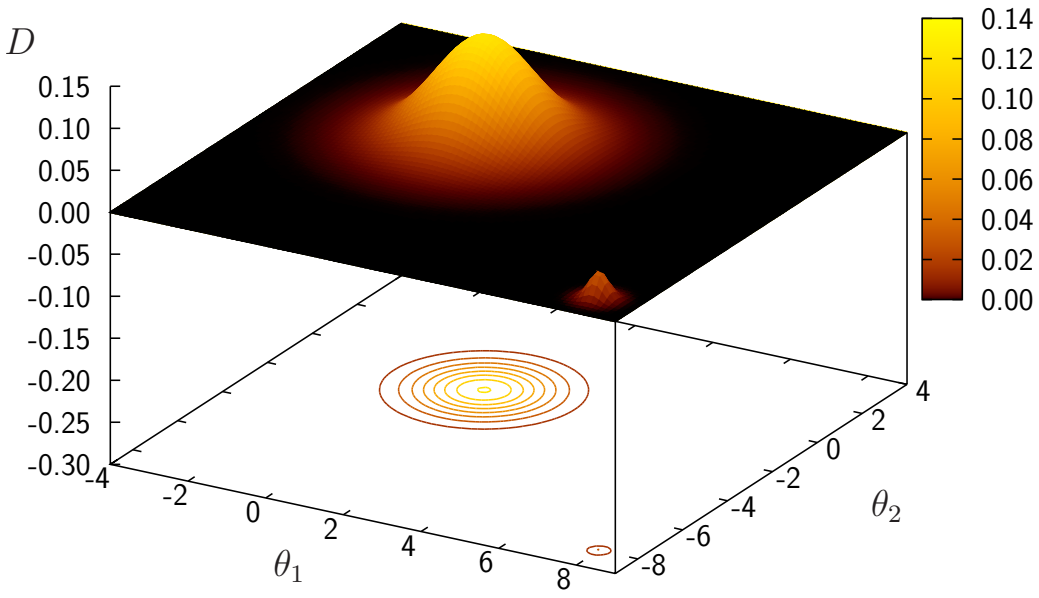


Figure 3.1: **Bivariate surface plot.** This figure shows the surface plot of a bivariate mixture normal distribution in analogy to Chib and Ramamurthy [28].

mixture weight to $\omega = 0.99$. The covariance matrices are set to $\boldsymbol{V}^{(1)} = \text{diag}[1.3, 1.3]$ and $\boldsymbol{V}^{(2)} = \text{diag}[0.05, 0.05]$, respectively. Figure 3.1 depicts the surface of the posterior density $p(\boldsymbol{\theta})$ and the corresponding contour lines.

Using a standard MCMC sampling scheme such as the Metropolis-Hastings algorithm leads to the unfortunate behavior that the Markov chain can get stuck in local regions of high probability mass and fails to explore the entire sampling space. This behavior is illustrated in Figure 3.2 which shows the sampling paths of two Markov chains initialized at different starting values. The chain in the left panel is started at $[1, -1]^\mathsf{T}$ and the chain in the right panel is started at $[8, -8]^\mathsf{T}$. The acceptance rate of both chains is approximately 0.27. Obviously, the chains only explore the mode that is closest to the starting value. While this behavior is not critical for the first chain (left panel) as 99% of the sample is correctly sampled, the consequences for the second chain (right panel) are severe. Only 1% of the random variables are correctly assigned. Bayesian estimates based on this deficient sample are necessarily inconsistent and lead to invalid inference. Of course, there exist several ad-hoc proposals to obtain a more realistic sample. One approach is to simply increase the variance of the proposal distribution. However, in practical applications the variance of the proposal distribution is commonly chosen to meet a rejection rate of approximately 0.2-0.3, such that this kind of behavior is often

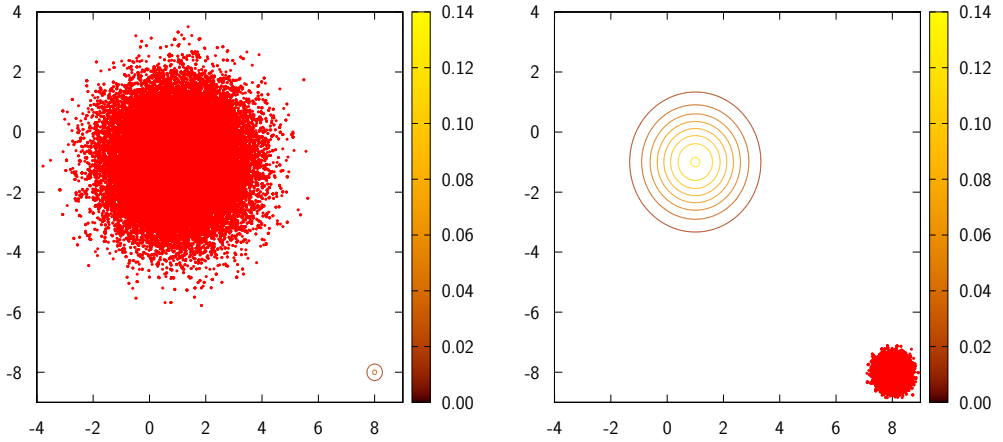


Figure 3.2: **Metropolis-Hastings Algorithm.** This figure shows two samples from a Metropolis-Hastings algorithm with target density (3.2) and a (zero mean) normal random walk proposal. The scaling of the proposal distribution is chosen in such a way that the rejection rate is approximately 0.27. The initial value of the chain in the left (right) panel is $[1, -1]^\mathsf{T}$ ($[8, -8]^\mathsf{T}$), respectively.

undiscovered and realistic. Another idea, proposed by Chib and Ramamurthy [28], is to search for all modes before using a MCMC algorithm. This is a reasonable approach in a low dimensional setup or if the modes are known, but in high dimensional problems the localization of all modes is quite challenging. This strong dependency of the Metropolis-Hastings algorithm on the starting points at which the Markov chain is initialized becomes even more cumbersome in high dimensional settings and standard techniques to monitor the convergence behavior of the Markov chain fail to recognize that there are subspaces that the chain has not visited yet.

Contrary to methods that construct a Markov chain with prescribed stationary distribution, sequential MCMC methods first construct a particle system from some easily sampled distribution μ_0. Each particle is then propagated in such a way as to be regarded as a particle approximation to some intermediate distribution μ_1 which is *closer* to the target distribution μ_n in some suitable sense. This process is iterated until one obtains a system of particles that can be regarded as draws from the the target μ_n. Figure 3.3 illustrates the univariate marginal densities of the sequence of distributions from μ_0 to μ_n. Note that the initial distribution μ_0 has a density that is rather flat thus enabling us to draw samples efficiently by standard techniques. Moreover, the algorithm not only succeeds in detecting both modes but is also able to shift probability mass from one mode

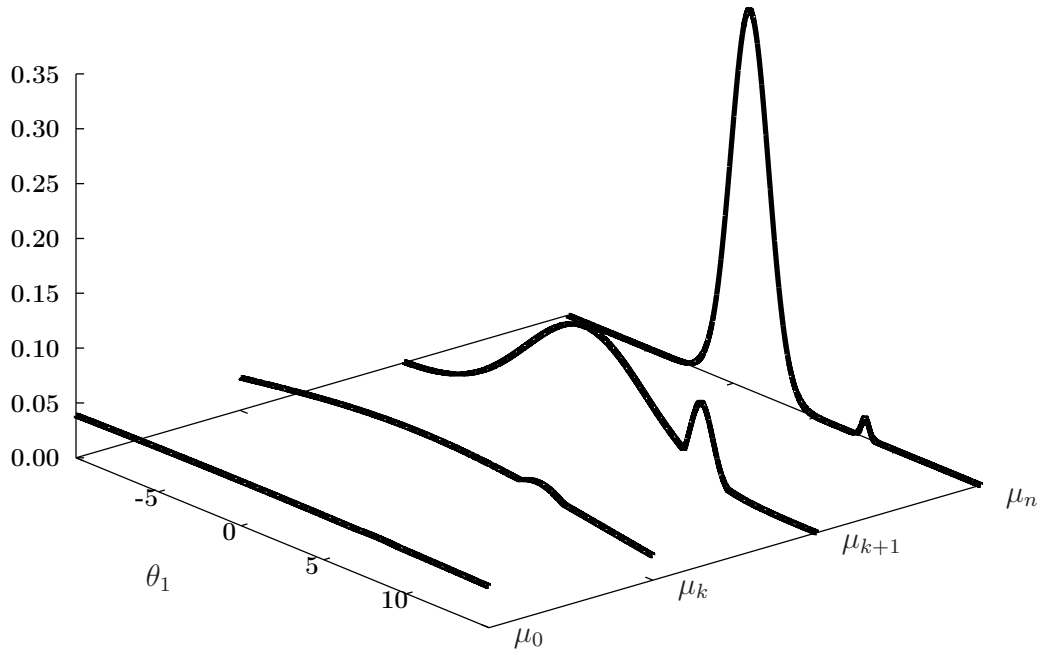


Figure 3.3: **Sequence of marginal densities.** This figure depicts an exemplary sequence of interpolating distributions for the marginal distribution (of θ_1) from the bivariate example (3.2).

to another as to establish the correct mixture weights of the two components. In the following Section we describe the algorithm in more detail.

3.2.2 A sequential sampling approach

In this section we give a formal presentation of the sequential MCMC algorithm and discuss its links to the particle filtering literature.

Formal description

Let the target μ_n be a probability distribution on a state space E that depends on the sample size n and where E can be taken, for example, to be $\mathbb{R}^d$. In our application μ_n is the (quasi-)posterior distribution but in general μ_n can be any probability distribution. For $f : E \to \mathbb{R}$ a measurable function we want to numerically approximate

$$\mu_n(f) = \int_E f(x)\mu_n(\mathrm{d}\, x).$$

Note that for f the identity function and μ_n the (quasi-)posterior distribution we obtain the (quasi-)posterior mean as a standard Bayesian estimator (under quadratic loss).

The approximation of $\mu_n(f)$ can now be based on sampled draws from the distribution μ_n. Assume we are given a collection of N such draws $\left(\xi_n^i\right)_{i=1}^N$, $\xi_n^i \in E$ which we call a *particle system*. The integral $\mu_n(f)$ is then approximated by the Monte Carlo sum

$$\eta_n^N(f) = \frac{1}{N}\sum_{i=1}^{N} f(\xi_n^i),$$

i.e. the empirical mean of the $f(\xi_n^i)$. Monte Carlo methods have found widespread use in econometrics, e.g. for simulation based estimation methods such as the efficient method of moments estimation of Gallant and Tauchen [54], the indirect inference approach Gouriéroux et al. [62] or the simulated maximum likelihood method of Durham and Gallant [43].

We are now interested in settings where we can not readily sample from the target distribution μ_n. A prominent case where this is not possible is, for example, when the distribution μ_n exhibits *multiple well-separated modes*. Contrary to the standard notion of MCMC methods we assume an *initial probability distribution* μ_0 on E that is easily sampled. We generate an initial particle system $\left(\xi_0^i\right)_{i=1}^N$ of draws sampled from μ_0 and propagate each particle ξ_0^i sequentially as to obtain a particle system $\left(\xi_n^i\right)_{i=1}^N$ that can be regarded as a set of draws sampled from μ_n.

To this end assume that there is a sequence of probability distributions $(\mu_k)_{k=1}^{n-1}$ which *interpolate* between μ_0 and μ_n in the following sense: For $k = 0, \ldots, n-1$, the distributions μ_k and μ_{k+1} are mutually absolutely-continuous with $\bar{g}_{k,k+1}$ denoting the relative (Radon-Nikodym) density of μ_{k+1} with respect to μ_k, i.e.

$$\mu_{k+1}(f) = \mu_k(\bar{g}_{k,k+1} f)$$

for any measurable function $f : E \mapsto \mathbb{R}$. To capture the idea of interpolation, we assume that there exists a constant $\gamma > 1$ such that $\bar{g}_{k,k+1}(x) < \gamma$ for all $x \in E$. Thus, the weight assigned to a point in E by μ_{k+1} can be bounded by γ times the weight assigned by μ_k. Note that the availability of this interpolation sequence is essential for the proposed algorithm. This may seem to be a rather strong assumption. However, we present a simple approach below to derive the interpolating sequence for general distributions.

To formulate the algorithm we need to introduce a sequence of *transition kernels* $K_k(x, dy)$ on E where K_k has stationary distribution μ_k. K_k can thus be thought of as many steps of a local Metropolis dynamics with respect to μ_k. We discuss this issue in more detail below.

The algorithm alternates between two steps: (i) an *Importance Sampling Resampling step* which moves each particle from μ_{k-1} to μ_k and (ii) *MCMC steps* with respect to μ_k.

The algorithm is thus given by:

1. For $k = 0$, generate N particles $(\xi_0^i)_{i=1}^N$ from the initial distribution μ_0.

2. For $k = 0, \ldots, n$ generate N preliminary particles $(\hat{\xi}_k^i)_{i=1}^N$ from the distribution μ_k through an Importance Sampling step with (multinomial) Resampling:
   The particles $\hat{\xi}_k^i$ are drawn conditionally independently from the empirical distribution of the particles $(\hat{\xi}_{k-1}^i)_{i=1}^N$ weighted with the relative density $\bar{g}_{k-1,k}$, i.e.

$$\mathbb{P}\left(\hat{\xi}_k^i = \xi_{k-1}^j | \xi_{k-1}^1, \ldots, \xi_{k-1}^N\right) = \frac{\bar{g}_{k-1,k}(\xi_{k-1}^j)}{\sum_{l=1}^N \bar{g}_{k-1,k}(\xi_{k-1}^l)}.$$

3. The (resampled) preliminary particles $\hat{\xi}_k^i$ are each moved conditionally independently in the MCMC step by the kernel K_k to generate the new particle positions ξ_k^i, i.e.

$$\mathbb{P}\left(\xi_k^i \in dx | \hat{\xi}_k^1, \ldots, \hat{\xi}_k^N\right) = K_k(\hat{\xi}_k^i, dx).$$

Let us discuss the individual steps of the algorithm in more detail. Consider the importance sampling step first. Assume we are given a particle approximation $\left(\xi_{k-1}^i\right)_{i=1}^N$ for the measure μ_{k-1}. We now wish to obtain a particle approximation for the measure μ_k. Importance sampling now transforms each particle ξ_{k-1}^i (i.e. an approximate draw from μ_{k-1}) to a new particle ξ_k^i (i.e. an approximate draw from μ_k) by assigning an *importance weight* ω_k^i. These importance weights are equal to the relative density of the measure μ_k with respect to μ_{k-1} evaluated at a certain particle. We thus obtain for the normalized importance weights $\{\omega_k^1, \ldots, \omega_k^N\}$ associated with the particles $\{\xi_k^1, \ldots, \xi_k^N\}$ the expression

$$\omega_k^i = \frac{\bar{g}_{k-1,k}(\xi_{k-1}^i)}{\sum_{l=1}^N \bar{g}_{k-1,k}(\xi_{k-1}^l)}, \quad i = 1, \ldots, N.$$

Next we consider the resampling step which is a crucial component in this algorithm. To see why resampling is necessary we observe that if a particle has a high importance weight at one iteration of the algorithm it is likely to have a high importance weight in the next iteration as well. Consequently, without resampling most probability mass will be eventually concentrated in only a few particles, a phenomenon that is commonly referred to as *weight degeneration*. The resampling step can thus be seen as a method of allocating the particles proportionally to their probability mass at every step of the algorithm. In the above algorithm we have proposed multinomial resampling. There are, however, also other resampling schemes available in the literature which we discuss and compare in detail below.

Although resampling can prevent weight degeneration, it has to be used cautiously. Consider a sample of particles $\xi_k^1, \ldots, \xi_k^N$ distributed as μ_k. Now resample, i.e. generate a new sample $\tilde{\xi}_k^1, \ldots, \tilde{\xi}_k^N$ by drawing N times independently and uniformly from $\{\xi_k^1, \ldots, \xi_k^N\}$. The empirical distribution of the new sample still approximates μ_k, but the quality of the sample is worse since some values from the original sample are lost while others are duplicated. This effect gets even stronger when the procedure is iterated until at some point all ξ_k^i have the same value. Hence, the implications of the resampling step are double-edged. On the one hand it prevents the degeneration of the particle approximation but on the other hand it worsens the approximation.

It is important to note that the Importance Sampling Resampling step alone leaves the initial positions of the particles generated from the initial distribution μ_0 unchanged. Old particles are propagated sequentially to serve as an approximation to the target distribution μ_n only through changes in the assigned importance weights. To overcome the problems with importance sampling and resampling an MCMC step is included at each iteration. After resampling a Metropolis Markov chain is initialized at each particle with stationary distribution μ_k that moves the particles apart but leaves their empirical distribution unchanged. The MCMC steps thus serve at least two purposes in the algorithm: (i) they help to better explore the target distribution, and (ii) they decrease the dependence between the particles by moving apart particles that duplicate the same predecessor.

In what follows we discuss the specification of the key ingredients of the Sequential MCMC algorithm, namely how to choose the sequence of interpolating probability distributions, the Markov Kernel K for the MCMC steps and we discuss and compare different resampling schemes.

Specification of the interpolating sequence of probability distributions

In the formal description of the Sequential MCMC algorithm we have so far assumed that the sequence of interpolating probability distributions $(\mu_k)_{k=1}^{n-1}$ is known. Based on the literature on Simulated Annealing and Tempering algorithms we define μ_k to be given by a slight variation of the quasi-posterior distribution for Laplace type estimators, i.e.

$$\mu_k(dx) \propto e^{\beta_k L(x)} \mu_0(dx).$$

where $(\beta_k)_{k=1}^{n}$ is an increasing sequence of numbers on $(0,1]$. Choosing the sequence of probability distributions, hence, amounts to choosing a sequence of numbers on the interval $(0,1]$. This choice can be conducted in an *adaptive* way. More precisely, we choose for $k = 1, \ldots, n-1$ the corresponding sequence of positive numbers $(\beta_k)_{k=1}^{n}$ such that the maximal ratio between the two interpolating distributions μ_{k+1} and μ_k is bounded by some constant γ, i.e.

$$\max_{x \in E} \bar{g}_{k,k+1}(x) = \max_{x \in E} \frac{d\mu_{k+1}}{d\mu_k}(x) \leq \gamma$$

for $\gamma \in (1, \infty)$. It is obvious that when choosing the constant γ one is facing a trade-off. A larger value of γ results in fewer interpolating distributions and the algorithm reaches the target distribution sooner. This reduction in run time may, however, results in too crude an approximation to the final target distribution. How specific values of γ effect the performance of the Sequential MCMC algorithm will be investigated in Section 3.3.3.

We moreover note that in order to run this algorithm, it is sufficient to know the relative densities only up to a normalizing factor. In what follows we denote by $g_{k-1,k}$ the unnormalized version of $\bar{g}_{k-1,k}$.

Specification of the Markov kernel

For the sake of completeness we give a general description of the MCMC step of our algorithm. For a general treatment we refer the interested reader to Robert and Casella [96]. In particular we discuss the choice of the Markov Kernel K_k. Starting with a resampled particle $\hat{\zeta}_k^i$ at the k-th iteration of the Sequential MCMC algorithm we generate an ergodic Markov chain with stationary distribution μ_k. This method is indeed the defining characteristic feature at the heart of all MCMC methods of which the Metropolis-Hastings algorithm is regarded to be the first and foremost example. Note that, since

any MCMC method generates *serially dependent* draws that approximate the distribution μ_k, ergodicity is a key requirement in that it allows us to use the obtained sample as if it were in fact independently and identically distributed according to μ_k in order to estimate

$$\eta_n^N(f) = \frac{1}{N}\sum_{i=1}^{N} f(\xi_n^i).$$

A generic definition of the Metropolis-Hastings algorithm is given below.

1. For $t = 0$ we start the Markov chain with initial value $\xi_k^i(0) = \hat{\xi}_k^i$.

2. For $t = 1, \ldots$ and given the current state of the Markov chain $\xi_k^i(t)$ we first generate a preliminary random variable X_t from some proposal distribution with density $q(x|\xi_k^i(t))$. The next state of the Markov chain is taken to be

$$\xi_k^i(t+1) = \begin{cases} X_t & \text{with probability } \rho(\xi_k^i(t), X_t) \\ \xi_k^i(t) & \text{with probability } 1 - \rho(\xi_k^i(t), X_t) \end{cases}$$

   where the acceptance probability is given by

$$\rho(\xi_k^i, x) = \min\left\{ \frac{d\mu_k(\xi_k^i)}{d\mu_k(x)} \frac{q(\xi_k^i|x)}{q(x|\xi_k^i)}, 1 \right\}.$$

The transition kernel associated with this algorithm is given by

$$K_k(\xi_k^i, x) = \rho(\xi_k^i, x) q(x|\xi_k^i) + (1 - r(\xi_k^i))\delta_{\xi_k^i}(x)$$

where $r(\xi_k^i) = \int \rho(\xi_k^i, x) q(x|\xi_k^i) dx$ and $\delta_{\xi_k^i}$ denotes the Dirac mass in ξ_k^i. It is easy to see that we have

$$\begin{aligned} \rho(\xi_k^i, x) q(x|\xi_k^i) d\mu_k(\xi_k^i) &= \rho(x, \xi_k^i) q(\xi_k^i|x) d\mu_k(x) \\ (1 - r(\xi_k^i))\delta_{\xi_k^i}(x) d\mu_k(\xi_k^i) &= (1 - r(x))\delta_x(\xi_k^i) d\mu_k(x) \end{aligned}$$

which together establish the *detailed balance equation* with $d\mu_k$ and it follows that μ_k is indeed the stationary distribution of the thus generated Markov chain. It moreover follows that the specification of the Markov kernel K_k essentially entails choosing a suitable proposal distribution with density q, the possible choices of which are nearly endless. Throughout the paper we restrict ourselves to a symmetric Random Walk proposal of

the form

$$X_t = \xi_k^i(t) + \varepsilon, \;\; \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2).$$

Specification of the resampling algorithm

In this subsection we are going to discuss different resampling schemes that have been proposed in the literature. For a detailed discussion we refer to Douc and Cappé [40] and Hol et al. [71]. Generally speaking, the resampling step removes particles with low importance weight and duplicates particles with high importance weight as to obtain a sample of particles that all have equal importance weight $1/N$. Four different resampling schemes have been widely used in applications, namely *Multinomial resampling, Residual Resampling, Stratified resampling* and *Systematic resampling*. The reason why one would consider different resampling schemes is the possibility to decrease the Monte Carlo variance of the particle sample and consequently to improve the quality of any estimator based on these particles. We will denote by $\{\xi_k^1, \ldots, \xi_k^N\}$ the set of particles approximating probability distribution μ_k at the k-th step of the Sequential MCMC algorithm with corresponding set of importance weights $\{\omega_k^1, \ldots, \omega_k^N\}$. We denote by $\{\tilde{\xi}_k^1, \ldots, \tilde{\xi}_k^N\}$ the set of *resampled* particles approximating μ_k and write $\{N_k^1, \ldots, N_k^N\}$ for the set of *duplicate counts*, i.e. $N_k^i = \sharp\{j, 1 \leq j \leq N : \tilde{\xi}_k^j = \xi_k^i\}$ denotes how many times a particle ξ_k^i is going to be duplicated through resampling.

1. *Multinomial resampling:* The general idea is to generate resampled draws independently from the common point mass distribution $\sum_{i=1}^N \omega_k^i \delta_{\xi_k^i}$. This can be be done by using the inversion method:

   First we generate N uniform random variables $\left(U^i\right)_{i=1}^N$ on the interval $(0, 1]$. We define the index $I^i = F_\omega^{-1}(U^i)$ where F_ω^{-1} is the inverse of the cumulative distribution function of the (normalized) importance weights $\left(\omega_k^i\right)_{i=1}^N$, i.e. we have $F_\omega^{-1}(u) = i$ if $u \in \left(\sum_{j=1}^{i-1} \omega_k^j, \sum_{j=1}^{i} \omega_k^j\right]$. We then take as the multinomially resampled particle

   $$\tilde{\xi}_k^i = \xi_k^{I^i}, \; i = 1, \ldots, N.$$

   As such the duplicate counts will be multinomially distributed, i.e.

   $$(N_k^1, \ldots, N_k^N) \sim Mult(N; \omega_k^1, \ldots, \omega_k^N).$$

2. *Residual resampling:* In this resampling scheme the number of duplicates of a particle ξ_k^i is determined by its importance weight ω_k^i plus a remainder term that is

again multinomially distributed. In particular we have for $i = 1, \ldots, N$

$$N_k^i = \lfloor N\omega_k^i \rfloor + \bar{N}_k^i,$$

where $\lfloor \cdot \rfloor$ denotes the integer part of its argument and the $\bar{N}_k^1 \ldots, \bar{N}_k^N$ are multinomially distributed according to $Mult(N - R; \bar{\omega}_k^1, \ldots, \bar{\omega}_k^N)$ where the residual part R is given by

$$R = \sum_{i=1}^{N} \lfloor N\omega_k^i \rfloor$$

with associated residual importance weights

$$\omega_k^i = \frac{N\omega_k^i - \lfloor N\omega_k^i \rfloor}{N - R}, \ i = 1, \ldots, N.$$

In practice, the multinomial counts $\bar{N}_k^1 \ldots, \bar{N}_k^N$ are generated by the inversion method as for Multinomial resampling.

3. *Stratified resampling:* We start with partitioning the interval $(0, 1]$ into N disjoint subsets, i.e.

$$(0, 1] = \left(0, \frac{1}{N}\right] \cup \ldots \cup \left(\frac{N-1}{N}, 1\right].$$

Then a uniformly distributed random variable U^i is drawn from each of these subintervals, i.e.

$$U^i \sim U\left(\left(\frac{i-1}{N}, \frac{i}{N}\right]\right), \ i = 1, \ldots, N.$$

Finally we use again the inversion method as in Multinomial resampling.

4. *Systematic resampling:* This resampling scheme works similar to Stratified resampling with the difference being that all the uniform variables drawn in each of the sub-intervals are deterministically linked. Particularly only one uniform random variable $U \sim U\left(\left(0, \frac{1}{N}\right]\right)$ is drawn and we compute

$$U^i = \frac{i-1}{N} + U, \ i = 1, \ldots, N.$$

Again we use the inversion method as in Multinomial resampling.

Douc and Cappé [40] and Hol et al. [71] compared these four resampling schemes in detail. All algorithms are shown to be unbiased in that, conditional on all particles and

associated importance weights up to and including iteration k, one has $\mathbb{E}(N_k^i) = N\omega_k^i$ for all $i = 1, \ldots, N$. Furthermore, the importance weight of each particle after resampling is equal to $\tilde{\omega}_k^i = 1/N$. Considering the Monte Carlo variance of the estimator

$$\eta_k^N(f) = \frac{1}{N}\sum_{i=1}^{N} f(\xi_k^i),$$

Douc and Cappé [40] show that the conditional variance by using Residual and Stratified Resampling is always lower than the one based on Multinomial resampling. The case of Systematic Resampling is more difficult to analyze theoretically, mainly, since all the systematically resampled particles are (conditionally) dependent. A conjecture that is often raised in the literature is that Systematic resampling does outperform Stratified resampling and as such Multinomial resampling. Douc and Cappé [40] however show with a counter example that this is generally not true and in fact the conditional variance can even become larger than the one obtained through Multinomial resampling. Hol et al. [71] provide a small simulation study on the computational complexity of the four resampling schemes discussed here. They conclude that both Stratified and Systematic resampling perform better than Multinomial resampling. The general message is thus that one can always do better by using any other resampling scheme than Multinomial resampling. In terms of robustness, Residual and Stratified resampling seem to be the best choice for the moment whereas Systematic resampling has to be applied cautiously.

Discussion

As mentioned above, Sequential MCMC is closely related to the literature on Particle Filters (see e.g. Del Moral et al. [36]). In fact, Sequential MCMC and Particle Filters are virtually identical methods that differ only in the question of which parameters are choice parameters and which parameters are part of the problem. Generally speaking, filtering is the problem of extracting information about the current state of some unobservable variable (the so-called signal) from noisy or partially revealing observations. In this interpretation, the state space E is the set of possible values of the signal and the distribution μ_k is the distribution of the signal at time k conditional on all information up to and including time k. It is clear that in filtering problems the entire sequence $(\mu_k)_{k=0}^n$ is given, whereas in Sequential MCMC only the target distribution μ_n is fixed and the sequence $(\mu_k)_{k=0}^{n-1}$ is a choice parameter that can be chosen such that the algorithm works

best. Notably, while the distributions μ_k can be thought of as smoothed versions of μ_n in Sequential MCMC this is typically not the case in filtering problems. Furthermore, in filtering the sequence $(\mu_k)_{k=0}^n$ only becomes available over time. In typical applications, integrals with respect to μ_k need to be approximated at time k before μ_{k+1} is known.

3.3 Finite sample properties - a simulation study

In this section we analyze the finite sample behavior of the proposed algorithm in different setups. In particular, we consider a 12-dimensional mixture model with four well separated modes. This example is constructed in analogy to the second example of Chib and Ramamurthy [28].

3.3.1 The model

Consider the distribution of a 12-dimensional random variable which is modeled as a four component Gaussian mixture with density

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = w^{(1)} f_{\mathcal{N}}\left(\boldsymbol{x}; \boldsymbol{\mu}^{(1)}, \boldsymbol{V}^{(1)}\right) + w^{(2)} f_{\mathcal{N}}\left(\boldsymbol{x}; \boldsymbol{\mu}^{(2)}, \boldsymbol{V}^{(2)}\right) + w^{(3)} f_{\mathcal{N}}\left(\boldsymbol{x}; \boldsymbol{\mu}^{(3)}, \boldsymbol{V}^{(3)}\right) + w^{(4)} f_{\mathcal{N}}\left(\boldsymbol{x}; \boldsymbol{\mu}^{(4)}, \boldsymbol{V}^{(4)}\right) \quad (3.3)$$

where $\boldsymbol{x} \in \mathbb{R}^{12}, w^{(1)} = 0.05, w^{(2)} = 0.75, w^{(3)} = 0.05, w^{(4)} = 0.15, f_{\mathcal{N}}(.; \boldsymbol{m}, \boldsymbol{S})$ is the density of the Gaussian distribution with mean vector $\boldsymbol{m} \in \mathbb{R}^{12}$ and covariance matrix $\boldsymbol{S} \in \mathbb{S}_{12}^{+}$,[1]

$$\boldsymbol{\mu}^{(1)} = \left[\boldsymbol{\mu}^{\mathsf{T}}, 1.5 \times \boldsymbol{\mu}^{\mathsf{T}}\right]^{\mathsf{T}}; \;\; \boldsymbol{\mu}^{(2)} = \left[-10 \times \boldsymbol{\mu}^{\mathsf{T}}, 10 \times \boldsymbol{\mu}^{\mathsf{T}}\right]^{\mathsf{T}};$$
$$\boldsymbol{\mu}^{(3)} = \left[\; 10 \times \boldsymbol{\mu}^{\mathsf{T}}, 5 \times \boldsymbol{\mu}^{\mathsf{T}}\right]^{\mathsf{T}}; \;\; \boldsymbol{\mu}^{(4)} = \left[\; 4 \times \boldsymbol{\mu}^{\mathsf{T}}, -15 \times \boldsymbol{\mu}^{\mathsf{T}}\right]^{\mathsf{T}};$$

with

$$\boldsymbol{\mu} = [1.41, 0.81, 0.49, 0.80, 1.07, 0.30]^{\mathsf{T}}.$$

[1] In the following $\mathbb{R}$ ($\mathbb{S}_d^+$) denotes the set of real numbers (cone of $d \times d$ positive-semidefinite matrices), respectively.

	$\boldsymbol{\mu}^{(2)}$	$\boldsymbol{\mu}^{(3)}$	$\boldsymbol{\mu}^{(4)}$
$\boldsymbol{\mu}^{(1)}$	30.3268	21.0665	36.5859
$\boldsymbol{\mu}^{(2)}$	0	44.9740	62.5084
$\boldsymbol{\mu}^{(3)}$	.	0	45.5523

Table 3.1: **Distance of the modes.** This table shows the Euclidean distance between the different modes of the distribution.

Furthermore,

$$\boldsymbol{V}^{(1)} = \begin{bmatrix} \boldsymbol{\Sigma}^{(1)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}^{(1)} \end{bmatrix}; \quad \boldsymbol{V}^{(2)} = \begin{bmatrix} \boldsymbol{\Sigma}^{(1)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}^{(2)} \end{bmatrix};$$
$$\boldsymbol{V}^{(3)} = \begin{bmatrix} \boldsymbol{\Sigma}^{(2)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}^{(1)} \end{bmatrix}; \quad \boldsymbol{V}^{(4)} = \begin{bmatrix} \boldsymbol{\Sigma}^{(2)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}^{(2)} \end{bmatrix};$$

with

$$\boldsymbol{\Sigma}^{(1)} = \begin{bmatrix} 0.0885 & -0.0023 & 0.0077 & 0.0041 & -0.0229 & -0.0025 \\ -0.0023 & 0.0055 & 0.0015 & 0.0028 & 0.0013 & 0.0001 \\ 0.0077 & 0.0015 & 0.0031 & 0.0018 & -0.0011 & -0.0002 \\ 0.0041 & 0.0028 & 0.0018 & 0.0029 & 0.0004 & -0.0012 \\ -0.0229 & 0.0013 & -0.0011 & 0.0004 & 0.0169 & 0.0004 \\ -0.0025 & 0.0001 & -0.0002 & -0.0012 & 0.0004 & 0.0024 \end{bmatrix}$$

and

$$\boldsymbol{\Sigma}^{(2)} = \begin{bmatrix} 0.1365 & -0.0009 & 0.0063 & 0.0075 & -0.0119 & -0.0101 \\ -0.0009 & 0.0082 & 0.0021 & 0.0048 & 0.0045 & -0.0049 \\ 0.0063 & 0.0021 & 0.0020 & 0.0018 & 0.0011 & -0.0018 \\ 0.0075 & 0.0048 & 0.0018 & 0.0137 & 0.0042 & -0.0078 \\ -0.0119 & 0.0045 & 0.0011 & 0.0042 & 0.0153 & -0.0051 \\ -0.0101 & -0.0049 & -0.0018 & -0.0078 & -0.0051 & 0.0096 \end{bmatrix}.$$

It is important to note that the four components of the distribution are well separated. Table 3.1 shows the Euclidean distances between the modes (means of the components) of the distribution. However, we note that the distance between the modes is not very

	$\mathcal{E}^{(2)}$	$\mathcal{E}^{(3)}$	$\mathcal{E}^{(4)}$
$\mathcal{E}^{(1)}$	26.6997	17.2649	32.4987
$\mathcal{E}^{(2)}$	0	40.8195	57.7571
$\mathcal{E}^{(3)}$	.	0	41.3448

Table 3.2: **Distance of the** 99.99999 99999 9999% **ellipsoids.** This table shows the minimal Euclidean distance between the 99.99999 99999 9999% ellipsoids of the components of the distribution.

informative since the separation also depends on the variances of the components. To illustrate this we also compute the (minimal) distance between the α% ellipsoids of the components. To this end note that for a d-dimensional Gaussian random vector with mean vector $\boldsymbol{m} \in \mathbb{R}^{12}$ and covariance matrix $\boldsymbol{S} \in \mathbb{S}_{12}^{+}$ the quadratic form

$$C = (\boldsymbol{x} - \boldsymbol{m})\, \boldsymbol{S}^{-1}\, (\boldsymbol{x} - \boldsymbol{m})^{\mathsf{T}} \tag{3.4}$$

is chi-squared distributed with d degrees of freedom, $C \sim \chi^2(d)$. The α% ellipsoids $\mathcal{E}^{(i)}(\alpha)$ are defined by

$$\mathcal{E}^{(i)} = \left\{ \boldsymbol{x} \in \mathbb{R}^{12} : \left(\boldsymbol{x} - \boldsymbol{\mu}^{(i)}\right) \left(\boldsymbol{V}^{(i)}\right)^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}^{(i)}\right)^{\mathsf{T}} \leq q_\alpha \right\} \quad \forall i \in \{1, \ldots, 4\}$$

with q_α denoting the α% quantile of the chi-squared distribution with 12 degrees of freedom.

Table 3.2 shows the minimal distance between the 99.99999 99999 9999% ellipsoids. Obviously, the distance between the ellipsoids is decreased relative to the (Euclidean) distance of the modes, but the distance is still huge. In practical terms this means that the masses from the components do not overlap.[2] We use this fact in the following evaluation of the distance between a particle approximation and the continuous target distribution.

[2]The term *practical* here means that there is no point in the support of the mixture distribution such that the density of two components is different from zero for double precision floating point numbers.

3.3.2 Measuring the distance

To evaluate different combinations of particle numbers, target ratios, MCMC draws and resampling algorithms with respect to the quality of the particle approximation it is important to have a reliable measure of distance between the continuous target distribution and the discrete particle approximation. In the univariate case it would be reasonable to consider some moments and compare the approximation to the true moments. However, this approach is not feasible in our 12-dimensional example, as even considering only the mean and the elements of the covariance matrix would result in 90 statistics. Moreover, the sole use of these moments is problematic in most relevant situations, as the first and second moments of multimodal distributions are not highly informative about the distributional shape. Rather than testing whether some statistics are identical to the values under the null hypothesis it seems therefore more appropriate to test whether the particles are sampled from the continuous target distribution.

Two prominent tests have been suggested in the literature. The *Kolmogorov-Smirnov test* uses the distance between the empirical distribution function and the distribution function under the null to test whether the distributions are reasonably close. The test, however, is mainly applied in the univariate case as multivariate extensions (see e.g. Justel et al. [77]) and relies on the computation of conditional distributions which are rarely available. For this reason, the *Pearson's chi-squared test* is adopted here. The Pearson's chi-squared test uses a partition of the support of the target distribution and compares the empirical frequencies with the theoretical frequencies in these disjoint subsets. Let $\mathcal{S}$ be the support of the distribution (under the null) and $\{\mathcal{B}_i\}_{i=1}^{b}$ a partition of $\mathcal{S}$, i.e.

$$\mathcal{S} = \cup_{i=1}^{b} \mathcal{B}_i \quad \text{with} \quad \mathcal{B}_i \cap \mathcal{B}_j = \emptyset \quad \forall i \neq j.$$

Then, the test statistic is given by

$$\mathcal{Z} = \sum_{i=1}^{b} \frac{\left(\sum_{j=1}^{n} \mathbb{I}_{\mathcal{B}_i}\left(\xi_j\right) - n\mathbb{P}\left(\zeta \in \mathcal{B}_i\right)\right)^2}{n\mathbb{P}\left(\zeta \in \mathcal{B}_i\right)}, \tag{3.5}$$

with $\mathbb{I}$ denoting the indicator function[3], $\left\{\xi_j\right\}_{j=1}^{n}$ the sample and ζ a d-dimensional random variable that is distributed according to the distribution under the null. The test statistic $\mathcal{Z}$ is asymptotically chi-squared distributed with $b-1$ degrees of freedom.

[3]Note that $\mathbb{I}_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$ and $\mathbb{I}_{\mathcal{A}}(x) = 0$ if $x \notin \mathcal{A}$.

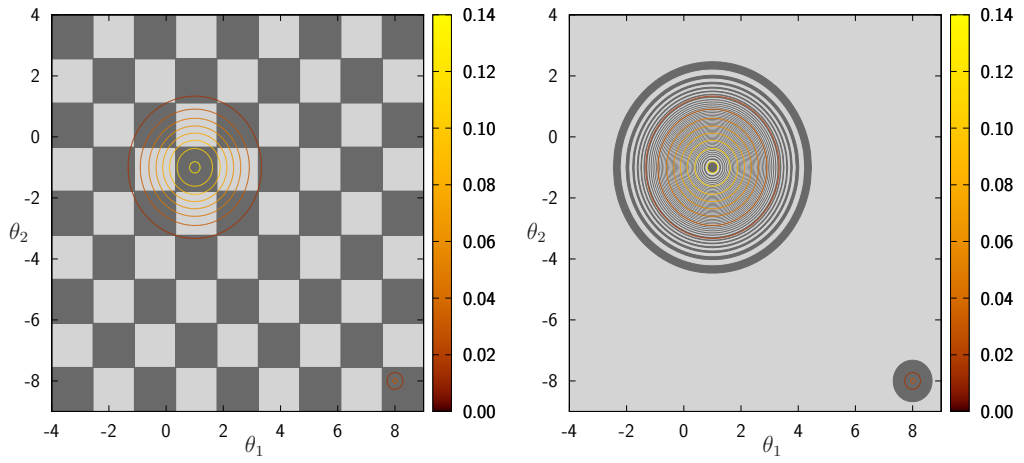


Figure 3.4: **Different partitions of the support.** The left panel shows a partition of the support into rectangular subsets. The right panel shows a partition of the support into ellipsoids.

To obtain a partition of the support one usually divides the support by splitting every dimension in d_i segments resulting in $\prod_{i=1}^{d} d_i$ bins, i.e. rectangular subsets of the support (see left panel of Figure 3.4 for a graphical illustration in terms of the bivariate example from before). This procedure is, however, not applicable in our example. Even if we used only two segments in each dimension ($d_i = 2 \; \forall i$) we would end up with $2^{12} = 4096$ bins. We thus propose a different approach, taking into account the fact that the components of the distribution are well separated as has been shown in the preceding section. We choose $b = 101$ cells and assign 100 cells to the modes and keep one cell for the remaining area. The cells for the modes are allocated among the modes according to their weights, i.e. in our example $b^{(1)} = 5, b^{(2)} = 75, b^{(3)} = 5$, $b^{(4)} = 15$. To obtain cells with approximately[4] equal mass (of $1/101$) we compute $b^{(i)}$ ellipsoids $\{\tilde{\mathcal{B}}_j^{(i)}\}_{j=1}^{b^{(i)}}$ around the corresponding mean $\boldsymbol{\mu}^{(i)}$ such that

$$\mathbb{P}\left(\zeta^{(i)} \in \tilde{\mathcal{B}}_j^{(i)}\right) = \frac{1}{w^{(i)}} \frac{j}{101} \quad \forall j = 1, \ldots, b^{(i)} \text{ and } i \in \{1, \ldots, 4\}$$

[4]"Approximately" means here that we ignore the (practically non existent) overlapping probability masses.

Significance level		Number of particles			
		10,000	100,000	1,000,000	10,000,000
α	0.500	0.4998	0.4998	0.5032	0.5084
	0.250	0.2502	0.2498	0.2524	0.2539
	0.100	0.1007	0.0995	0.1024	0.1053
	0.050	0.0503	0.0494	0.0508	0.0527
	0.025	0.0254	0.0246	0.0246	0.0270
	0.010	0.0102	0.0095	0.0101	0.0105

Table 3.3: **Size of the Pearson's chi-squared test.** Reported are the rejection probabilities of the Pearson's chi-squared test based on our selected cells for different number of particles.

with $\zeta^{(i)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(i)}, \boldsymbol{V}^{(i)}\right)$. Using property (3.4) the ellipsoids are given by

$$\tilde{\mathcal{B}}_j^{(i)} = \left\{ \boldsymbol{x} \in \mathbb{R}^{12} : \left(\boldsymbol{x} - \boldsymbol{\mu}^{(i)}\right) \left(\boldsymbol{V}^{(i)}\right)^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}^{(i)}\right)^{\mathsf{T}} < q_{j/(101 w^{(i)})} \right\}$$
$$\forall j = 1, \ldots, b^{(i)} \text{ and } i \in \{1, \ldots, 4\}$$

where again q_α denotes the α % quantile of the chi-squared distribution with 12 degrees of freedom. The final cells are the ellipsoidal disks given by

$$\mathcal{B}_j^{(i)} = \tilde{\mathcal{B}}_j^{(i)} \setminus \tilde{\mathcal{B}}_{j-1}^{(i)} \quad \forall j = 1, \ldots, b^{(i)} \text{ and } i \in \{1, \ldots, 4\}$$

with $\tilde{\mathcal{B}}_0^{(i)} = \emptyset \; \forall i \in \{1, \ldots, 4\}$ and one additional cell for the remaining support

$$\mathcal{B}^* = \mathbb{R}^{12} \setminus \left(\tilde{\mathcal{B}}_5^{(1)} \cup \tilde{\mathcal{B}}_{75}^{(2)} \cup \tilde{\mathcal{B}}_5^{(3)} \cup \tilde{\mathcal{B}}_{15}^{(4)} \right),$$

(see right panel of Figure 3.4 for a graphical illustration in terms of the bivariate example from before).

Table 3.3 reports the rejection probabilities of the Pearson's chi-squared test (3.5) for different numbers of particles. The results are based on $100,000$ replications. The table nicely illustrates that the asymptotic chi-squared distribution is quite accurate even for finite samples, i.e. for a finite number of particles.

To assess the power of the test we simulate from different alternative distributions. Scenario I considers independent and identically distributed (i.i.d.) random variables

Scenario	Quantiles of the test statistic $\mathcal{Z}$		
	1%	50%	99%
I	9.9916×10^7	9.9925×10^7	9.9934×10^7
II	3.2998×10^5	3.3014×10^5	3.3030×10^5
III	1.1005×10^5	1.1011×10^5	1.1018×10^5
IV	5.2173×10^4	5.2210×10^4	5.2256×10^4
V	5.2173×10^4	5.2210×10^4	5.2257×10^4

Table 3.4: **Power of the Pearson's chi-squared test.** The table reports selected quantiles of the Pearson's chi-squared test statistic associated with different sampling scenarios.

from a Gaussian distribution with mean and covariance matrix equal to the mean and variance of the mixture distribution, i.e. they are given by

$$\tilde{\boldsymbol{\mu}} = \sum_{i=1}^{4} w^{(i)} \boldsymbol{\mu}^{(i)} = \left[-6.350 \times \boldsymbol{\mu}^{\mathsf{T}}, 5.575 \times \boldsymbol{\mu}^{\mathsf{T}} \right]^{\mathsf{T}}$$

and

$$\sum_{i=1}^{4} w^{(i)} \left(\boldsymbol{\mu}^{(i)} \boldsymbol{\mu}^{(i)\mathsf{T}} + \boldsymbol{V}^{(i)} \right) - \tilde{\boldsymbol{\mu}} \tilde{\boldsymbol{\mu}}^{\mathsf{T}} =$$
$$\begin{bmatrix} 42.1275 & -46.0238 \\ -46.0238 & 79.0319 \end{bmatrix} \otimes \boldsymbol{\mu} \boldsymbol{\mu}^{\mathsf{T}} + \begin{bmatrix} 0.8\boldsymbol{\Sigma}^{(1)} + 0.2\boldsymbol{\Sigma}^{(2)} & \mathbf{0} \\ 0 & 0.1\boldsymbol{\Sigma}^{(1)} + 0.9\boldsymbol{\Sigma}^{(2)} \end{bmatrix},$$

respectively, where $\otimes$ denotes the Kronecker product. In Scenario II we sample from the second component of the mixture distribution, which is the component with the largest weight. Adding the component with the second largest weight (component 4) and setting the weights proportional to the original weights gives Scenario III, while adding the first (third) component yields Scenario IV (V).

In Table 3.4 we highlight the power of the Pearson's chi-squared test by reporting selected quantiles of the test statistic associated with the above scenarios. The number of particles used to this end has been set to $1,000,000$. As the table indicates, the test highly rejects at any reasonable error level. Furthermore, it can be observed that for distributions closer to the null distribution the statistic decreases. This makes the Pearson's chi-squared test statistic a reasonable distance measure for our application.

3.3.3 Simulation results

In this section we employ the distance measure proposed in Section 3.3.2 to analyze the effect of changes in the hyperparameters of the algorithm on the approximation quality. To this end, we first consider the different resampling algorithms discussed in Section 3.2.2. Thereafter we assess the effect of different MCMC iterations and, finally, we analyze different bounds for the relative density (which restricts the number of interpolating distributions considered to move from the initial μ_0 to the target μ_n).

Table 3.5 reports selected quantiles of our chi-squared test statistic $\mathcal{Z}$ for different resampling algorithms. As one would expect, the simple multinomial resampling algorithm performs worst. Any other resampling algorithm performs considerably better. It is interesting that it is systematic resampling that delivers an excellent approximation. This seems in line with our previous discussion of the different resampling algorithms and highlights the need to cross-validate the approximation obtained through different resampling schemes. Furthermore, we note that the computational resources for the systematic resampling approach are less demanding than for any other algorithm such that systematic resampling seems recommendable in this setting.

Algorithm	Quantiles of the test statistic $\mathcal{Z}$		
	1%	50%	99%
multinomial	1.5099×10^2	1.1387×10^3	1.2747×10^5
residual	1.1408×10^2	9.2459×10^2	2.4044×10^4
stratified	1.1698×10^2	9.1358×10^2	2.3637×10^4
systematic	1.1599×10^2	8.7246×10^2	2.1869×10^4

Table 3.5: **Effect of resampling.** The table reports selected quantiles of the Pearson's chi-squared test statistic associated with different resampling algorithms.

Table 3.6 shows the results for different MCMC iterations. We observe that a larger number of MCMC iterations leads to a better final approximation. This is not surprising as the MCMC step serves two aims. First, it explores the new (intermediate) target distribution, and second, it removes the dependency between particles with the same anchor particle. Both effects are improved if a larger number of steps is considered. However, even for values that are realistic in most applications, e.g. 40, the approximation is very good and only improves slightly if we consider more MCMC steps.

MCMC iterations	Quantiles of the test statistic $\mathcal{Z}$ 1%	50%	99%
5	8.0336×10^2	2.2599×10^4	1.0440×10^5
10	3.8555×10^2	5.4592×10^3	7.1441×10^4
20	2.5786×10^2	2.5940×10^3	6.3691×10^4
40	2.2558×10^2	1.4254×10^3	3.7246×10^4
80	1.2725×10^2	1.2259×10^3	1.6089×10^5
160	1.6105×10^2	1.1373×10^3	1.1262×10^5

Table 3.6: **Effect of MCMC iterations.** The table reports selected quantiles of the Pearson's chi-squared test statistic associated with different iterations of the MCMC step.

The results with respect to changes in the adaptive scheme are presented in Table 3.7. First we note that the smaller the maximal ratio of the interpolating distribution, the more interpolating distributions are considered. This has two, potentially conflicting, effects on the final approximation. On the one hand, a smaller value leads to an interpolating distribution that is closer to the previous distribution and therefore the approximation error decreases. On the other hand, as the approximating distributions are closer, there are more intermediate distributions required to reach the target distribution and a greater number of intermediate distributions may lead to a higher Monte Carlo variance leading to a worse approximation. However, as shown in Table 3.7, the first effect clearly dominates.

Max. ratio	Quantiles of the test statistic $\mathcal{Z}$ 1%	50%	99%
1.5	1.5982×10^2	1.1565×10^3	1.5723×10^5
2	1.7856×10^2	1.4827×10^3	7.1441×10^5
2.5	2.7304×10^2	1.5522×10^4	3.7076×10^6
5	1.1484×10^3	2.7330×10^4	8.6816×10^6

Table 3.7: **Effect of interpolation sequence.** The table reports selected quantiles of the Pearson's chi-squared test statistic associated with different bounds on the maximal ratio of interpolating distributions.

3.4 Conclusion

In this paper we have discussed a promising approach to overcome the difficulties standard MCMC methods have when facing multimodal sampling distributions. We have studied a Sequential MCMC algorithm in an extensive simulation study and investigated its finite sample properties. As we demonstrated, this algorithm is capable of detecting all modes in a 12-dimensional mixture model where the modes are chosen such that they are considerably far apart from each other. Such models proved very difficult to estimate by standard MCMC methods and so served as our benchmark case for the investigation of the Sequential MCMC approach. We introduced a measure of distance that takes into account the covariance structure of the modes, and, based on this metric, proposed a test statistic that allows us to assess through a simulation study the finite sample properties of the Sequential MCMC sampler.

The results show that Sequential MCMC methods clearly outperform standard MCMC approaches in a multimodal setting. Not only did the Sequential MCMC algorithm detect all four well-separated modes in a $12-$dimensional setting. More importantly, the algorithm is capable of yielding correct estimates for the mixture weights. This is mainly due to the fact that the Sequential MCMC sampler can shift probability mass between different regions of the parameter space efficiently when moving from the initial probability distribution to the final target distribution of interest. This holds true even when for some intermediary distribution the respective modes are already well pronounced. An additional advantage of the Sequential approach is that the resulting particle system approximating the target distribution is not subject to the same serial dependence problems as a standard MCMC chain.

Moreover, we confirmed results in the literature on the use of different resampling schemes and provided a detailed comparison of the effects different choices of tuning parameters have on the approximation to the true sampling distribution. These results can serve as valuable guidelines when applying this method to more complex economic models, such as the (Bayesian) estimation of Dynamic Stochastic General Equilibrium models.

4

Prediction in dynamic functional additive models: a k-nearest neighbors approach

4.1 Introduction

In many statistical applications, data is becoming available at ever increasing frequencies. It has thus become natural to think of discrete observations as realizations of a continuous function, say over the course of one day. The statistical analysis of such *functional data* is intrinsically different from standard multivariate techniques and has seen much attention in the literature recently (see, e.g. the monographs of Ramsay and Silverman [92], Ramsay and Silverman [91] and Horváth and Kokoszka [73] for a general discussion and applications).

In the most general context of *functional regression*, the interest is to model the relationship between a functional response Y and a functional predictor X, i.e. to model

$$M(x) := \mathbb{E}\left[Y|X = x\right]. \tag{4.1}$$

Typically, the framework for $M(x)$ considered in the literature is through functional *linear* models (see Cardot et al. [24], Yao et al. [110], Cai and Hall [21], Hall and Horowitz [66] and Crambes et al. [30]). The scope of functional regression models can, however, be widened by considering generalizations to functional *additive* models that were proposed by Müller and Yao [84]. As the authors point out, this gives rise to far more flexible and essentially nonparametric models, all the while avoiding the curse of dimensionality that is inevitable if no structure on the regression model in (4.1) were imposed.

In this paper, we are interested in modeling functional regression functions of *first-order auto-regressive type*, i.e. where the functional predictor X is a lagged version of the functional response Y such that the setting we are concerned with is the one of *functional time series*. Functional linear models of auto-regressive type have been studied extensively by Bosq [18], and the aim of this paper is to provide an extension to the class of functional additive models (FAM) as proposed by Müller and Yao [84]. Particularly, we are interested in the problem of prediction of some future function X_{N+1} given a sample of N functional observations X_i, $i = 1, \ldots, N$ and the suggested FAM modeling approach has several advantages in this regard.

First, it allows us to introduce a notion of time-dependence for functional data that is rooted directly at the functional principal component scores. As Müller and Yao [84] show, functional additive models emerge naturally in view of uncorrelatedness of the functional principal component scores across the spectral dimension. While we furthermore strengthen this condition to independence across the spectral dimension, we do allow for very general dependencies across the time dimension, and particularly do not rule out long memory behavior.

Second, it allows us to consider the problem of prediction in a very natural way, as only functions of the principal component scores have to be predicted. To this end, we propose a k-nearest neighbors classification approach that is very intuitive and easy to implement. In the finite-dimensional setting, this method is well understood and has been successfully applied to classical time series analysis such that theoretical results are readily available (see Cover and Hart [29], Stone [102], Stute [103], Yakowitz [108] and Rakotomarolahy [90]).

This gives rise to our notion of FAM-knn models that we present in more detail in Section 4.2, where we furthermore state and discuss the relevant assumptions. Prediction, as much as estimation, in functional regression models needs regularization which we achieve by projecting on finitely many eigenfunctions of the covariance operator of the functional observations. In Section 4.3 we thus verify the applicability of functional

principal components analysis, particularly under the assumption of some suitably defined dependence over time between the functional observations. Implementation as well as theoretical results on the consistency of the proposed FAM-knn predictor are then stated in Section 4.4. The proofs of all results are collected in Section 4.5 while Section 4.6 concludes.

4.2 Modeling framework

In this section we present in more detail our modeling framework of functional additive models of first-order auto-regressive type and discuss the relevant assumptions we impose. Consider to this end a sample of N functional observations $(X_i(t))_{i=1}^N$ that are assumed to belong to the Hilbert space $L^2 \equiv L^2(\mathcal{T}, \|\cdot\|)$ of square integrable functions defined over some domain $\mathcal{T} = [0, T]$ and equipped with a norm $\|\cdot\|$ induced by an innerproduct which we denote as $\langle\cdot,\cdot\rangle$. We refer to i as the time index, such that the sample $(X_i)_{i=1}^N$ is given by consecutively observed functions that are defined over the domain $\mathcal{T}$.

4.2.1 A functional additive model of first-order auto-regressive type

It is well known that every function X in L^2 (and as such all functional regression models) admits a spectral decomposition in terms of eigenfunctions of the covariance operator associated with X. Let us define by

$$\mu := \mathbb{E}[X] \tag{4.2}$$

$$\mathcal{C}_X[x] := \mathbb{E}\left[\langle X, x\rangle X\right], \quad x \in L^2 \tag{4.3}$$

the mean function and covariance operator of the random function X, respectively. For future reference we denote the space of Hilbert-Schmidt operators from $L^2 \mapsto L^2$ by $\mathcal{S}$ and equip it with the operator norm $\|\cdot\|_{\mathcal{S}}$, i.e. for some $\Psi \in \mathcal{S}$, $\|\Psi\|_{\mathcal{S}} := \left(\sum_{h=1}^{\infty} \|\Psi[e_h]\|^2\right)^{1/2}$ for any orthonormal basis $(e_h)_{h\geq 1}$. Let us furthermore denote by $\lambda_1 > \lambda_2 > \ldots$ the decreasing sequence of eigenvalues associated with $\mathcal{C}_X$ and denote by $\psi_1, \psi_2, \ldots$ the corresponding eigenfunctions. The eigenfunctions $(\psi_l)_{l\geq 1}$ are referred to as functional principal components and constitute an orthonormal basis system that span the space L^2 and as such any $X \in L^2$ admits the Karhunen-Loève

decomposition of the form

$$X(t) = \mu(t) + \sum_{l=1}^{\infty} \langle X, \psi_l \rangle \psi_l(t) = \mu(t) + \sum_{l=1}^{\infty} \theta_l \psi_l(t) \tag{4.4}$$

where $\theta_l := \langle X, \psi_l \rangle$ denotes the l-th functional principal component score of X and convergence of the right-hand-side in (4.4) is understood to be in L^2. By construction, the sequence of functional principal component scores $(\theta_l)_{l \geq 1}$ is such that the θ_l are uncorrelated across the spectral dimension l, have mean zero and variance λ_l. In what follows, we strengthen uncorrelatedness of the θ_l to independence across l, an assumption that is for example satisfied if X is a Gaussian process (see Müller and Yao [84]).

Now assume we are given a sample of N functional observations $(X_i)_{i=1}^N$ with Karhunen-Loève decomposition $X_i = \mu + \sum_{l \geq 1} \theta_{i,l} \psi_l$ and consider a functional regression model of first-order auto-regressive type, given by

$$M(x) = \mathbb{E}\left[X_i | X_{i-1} = x\right], \quad i = 1, \ldots, N, \tag{4.5}$$

where the realization x of X_{i-1} also admits a Karhunen-Loève decomposition of the form $x = \mu + \sum_{l \geq 1} \theta_l \psi_l$ and is thus characterized by a countable sequence of functional principal component scores $(\theta_l)_{l \geq 1}$.

We propose a functional additive model in analogy to Müller and Yao [84] in which the functional regression model defined in (4.5) takes the form

$$\mathbb{E}\left[X_i | X_{i-1} = x\right] = \mu + \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} m_{k,l}(\theta_k) \psi_l.$$

Note that the relationship between the functional response X_i and the predictor scores θ_k is now modeled additively through the function $m_{k,l}(\theta_k)$. This distinguishes this approach from linear functional models where this relationship is linear in the predictor scores θ_k. Moreover, it follows from Müller and Yao [84] that $m_{k,l}(\theta_k) = \mathbb{E}[\theta_{i,l} | \theta_{i-1,k} = \theta_k]$ and in view of the assumption on independence of the $\theta_{i,l}$ across l we have that $m_{k,l}(\theta_k) = 0$ if $k \neq l$ and $m_{l,l}(\theta_l) \equiv m_l(\theta_l) = \mathbb{E}[\theta_{i,l} | \theta_{i-1,l} = \theta_l]$. Hence,

the functional additive model of first-order auto-regressive type we consider is given by

$$\begin{aligned} M(x) &:= \mu + \sum_{l=1}^{\infty} m_l(\theta_l)\psi_l \\ &= \mu + \sum_{l=1}^{\infty} \mathbb{E}\left[\theta_{i,l} | \theta_{i-1,l} = \theta_l\right] \psi_l, \quad i = 1, \ldots, N, \end{aligned} \tag{4.6}$$

and identifiability is ensured since

$$\mathbb{E}[m_l(\theta_l)] = \mathbb{E}[\mathbb{E}\left[\theta_{i,l} | \theta_{i-1,l} = \theta_l\right]] = \mathbb{E}[\theta_{i,l}] = 0 \quad \forall\, l \geq 1.$$

As already mentioned in the introduction, this modeling framework has several advantages to be considered in the context of predicting functional time series. First it allows for a straightforward implementation of very general notions of time dependencies for functional data as it effectively boils down to considering the correlation structure of the functional principal component scores $\theta_{i,l}$ across the time dimension i. As will become apparent in the proofs of the theoretical results, these time dependencies can be as general as to allow for long memory behavior. Second, as we are interested in prediction, this modeling framework provides us with an intuitive approach, as we can consider prediction of the functional principal component scores component-wise across the spectral dimension. More precisely, we are interested in estimating the conditional mean of a function at some future time point $N + 1$ given that at time point N it takes on some realized value x. Then, in the modeling framework considered here, this functional regression model takes on the form

$$\begin{aligned} M(x) &= \mathbb{E}\left[X_{N+1} | X_N = x\right] \\ &= \mu + \sum_{l=1}^{\infty} \mathbb{E}\left[\theta_{N+1,l} | \theta_{N,l} = \theta_l\right] \psi_l. \end{aligned} \tag{4.7}$$

As a consequence, in order to form a prediction X_{N+1}^f of a future function X_{N+1}, it suffices to estimate the conditional means $\mathbb{E}\left[\theta_{N+1,l} | \theta_{N,l} = \theta_l\right]$, which only involves scalar random variables. To this end, we suggest a k-nearest neighbors classification approach that is very easy to implement. In addition, both the mean function and the eigenelements of the covariance operator are unknown and have to be estimated based on a sample of N functional observations. Regularization in our modeling framework

is achieved through projection on a finite number of functional principal components. We discuss this point in more details in Section 4.3 and give implementation details and theoretical consistency results of the here proposed FAM-knn predictor in Section 4.4.

4.2.2 Assumptions

In this section we discuss the relevant assumptions we make. Particularly, we propose a general notion of time dependencies for functional additive models of auto-regressive type. Time dependent stochastic processes have been considered in the statistical literature in many ways. In the context of classical (i.e. finite dimensional) time series analysis, *ergodicity* and various *mixing* conditions have been very popular (see, e.g. Hamilton [68] and Davidson [35]). In the functional context, however, only few results are available when dealing with time-dependent curves. Among them, Bosq [18] studies the theory of linear functional time series, focusing on functional auto-regressive models while Hörmann and Kokoszka [72] introduce a moment based notion of weak dependence and show that in that case eigenfunctions and eigenvalues can still be consistently estimated.

In this paper we attempt to introduce a more general notion of time dependencies for functional time series which in view of the previous discussion is rooted at the correlation structure of the $\theta_{i,l}$ across the time dimension i. As in classical time series analysis, our starting point is to restrict our attention to the class of stationary processes which we define below.

Definition 4.1. *A functional time series* $(X_i)_{i\geq 1}$ *with* $X_i \in L^2$ *is called* stationary *if there exists a function* $\mu \in L^2$ *and a sequence* $(c_m(t,s))_{m\geq 0}$ *such that for each* $i, j \in \mathbb{N}$ *one has*

$$\begin{aligned}
&\mathbb{E}[X_i](t) = \mu(t), \\
&\mathbb{E}\left[(X_i(t) - \mu(t))(X_j(s) - \mu(s))\right] = c_{|i-j|}(t,s).
\end{aligned}$$

For a given stationary functional time series $(X_i)_{i\geq 1}$ that admits a Karhunen-Loève decomposition as in (4.4), let the random principal component scores $\theta_{i,l}$ have unconditional probability density function $f_l(\theta_{i,l})$, and write $f_l(\theta_{i+1,l}|\theta_{i,l} = \theta_l)$ for the conditional probability density of $\theta_{i+1,l}$ given that $\theta_{i,l} = \theta_l$. Moreover, for $p = 3, \ldots, 6$, let $\kappa_{l(0,\tau_1,\ldots,\tau_{p-1})}$ denote the p-th order cumulant of $(\theta_{i,l}, \theta_{i+\tau_1,l}, \ldots, \theta_{i+\tau_{p-1},l})$, where $\tau_1, \ldots, \tau_{p-1}$ are integers (see, e.g. Brillinger [19, p.19]). As before, we write $m_l(\theta_l) =$

$\mathbb{E}[\theta_{N+1,l}|\theta_{N,l} = \theta_l]$. Finally, we note that in what follows, we adopt the convention that constants will generally be denoted by C, without distinguishing them unless it is required. Then we shall assume the following.

Assumption 4.1.

(**i**) $\mathbb{E}\|X_i\|^6 < \Delta < \infty$ *for all* $i \geq 1$.

(**ii**) *We have for some* $\alpha > 1$ *and all* $l \geq 1$,

$$\lambda_l - \lambda_{l+1} \sim l^{-\alpha-1}.$$

(**iii**) $m_l(\cdot)$ *and* $f_l(\cdot)$ *are twice continuously differentiable and* $f_l(\cdot)$ *is bounded. Furthermore, the functional principal component scores* $\theta_{i,l}$ *are independent across* l.

(**iv**) *Define* $B_{m,l} := \sup_i |\mathbb{E}[\theta_{i,l}\theta_{i-m,l}]|$. *Then there exists a constant* $C > 0$ *and some* $\beta > 0$ *such that for all* $l \geq 1$,

$$B_{m,l} \leq C\, m^{-\beta}\lambda_l.$$

(**v**) *For all* $l \geq 1$ *and* $p = 3, \ldots, 6$, *the score sequence* $(\theta_{i,l})_{i\geq 1}$ *has absolutely summable* p*-th order cumulants with*

$$\sum_{\tau_1,\ldots,\tau_{p-1}=-\infty}^{\infty} \cdots \sum |\kappa_{l(0,\tau_1,\ldots,\tau_{p-1})}| \leq C\lambda_l^{p/2}.$$

Part (**i**) of Assumption 4.1 enhances the standard assumption of finite fourth moments that is usually employed in functional data analysis under i.i.d. sampling (see, e.g. Horváth and Kokoszka [73]) to the existence of sixth moments. This assumption is in line with the standard time series literature where higher moment restrictions are imposed as a trade off for allowing time dependencies.

Condition (**ii**) is standard and prevents the spacing between adjacent eigenvalues λ_l from being too small. It also implies that $\lambda_l \sim l^{-\alpha}$. Furthermore, note that we can restrict ourselves to the case $\alpha > 1$ since it follows from the moment condition in (**i**) that the eigenvalues have to be summable, i.e. $\sum_{l\geq 1}\lambda_l < \infty$. The importance of this spacing property (**ii**) will particularly become apparent in the proofs of Corollary 4.1 and Theorem 4.3. Intuitively, as l increases, it becomes more difficult to estimate

the eigenfunctions ψ_l associated with λ_l as the expected L^2 error is proportional to δ_l^2 where $\delta_l := \max_{1\leq k\leq l} (\lambda_k - \lambda_{k+1})^{-1}$. As a consequence, the sequence of eigenvalues $(\lambda_l)_{l\geq 1}$ cannot decrease too fast for the estimation error of the eigenfunctions ψ_l not to explode.

The principal component scores $\theta_{i,l}$ of the Karhunen-Loève decomposition are uncorrelated across the spectral dimension l by construction. In Part $(\mathbf{iii})$ of Assumption 4.1 we strengthen this condition to independence across l. Moreover, both the principal component regression function m_l and their density f_l need to be sufficiently smooth as to allow for a Taylor expansion of up to second order. This is essentially a requirement for the consistency results of the k-nearest neighbors estimator of m_l to hold.

Part $(\mathbf{iv})$ and $(\mathbf{v})$ restrict the form of time dependencies that we allow for the score series $(\theta_{i,l})_{i\geq 1}$, for all $l \geq 1$. The assumed behavior of the $B_{m,l}$, which represent a measure of absolute autocovariances of the score series $(\theta_{i,l})_{i\geq 1}$, is only a mild restriction. In particular, part $(\mathbf{iv})$ implies a natural restriction on the absolute summability of the m-th autocovariances of the score series across the spectral dimension l, since $\sum_{l\geq 1} B_{m,l} \leq Cm^{-\beta}$. However, absolute summability of the autocovariances of the score series is not required across the time dimension i. More precisely, for $0 < \beta < 1$ one can conclude that $\sum_{m=1}^{N} B_{m,l}$ is of order $N^{1-\beta}\lambda_l$ which has explosive behavior for fixed l and large N. In fact, if one is interested in only establishing consistency of the functional principal components method, part $(\mathbf{iv})$ can be relaxed further to

$(\mathbf{iv})'$ For $B_{m,l}$ defined as before there exists a constant $C > 0$ such that for all $l \geq 1$

$$B_{m,l} \leq Cb_m\lambda_l \quad \text{with} \quad \sum_{m=1}^{\infty} m^{-1}b_m < \infty.$$

Condition $(\mathbf{iv})'$ allows for a very slow decay of the time dependencies (as measured through absolute autocovariances) represented by the component b_m that can even be of logarithmic order (see, e.g. Davidson [35, Theorem 2.31]), i.e.

$$b_m = \mathcal{O}\left(\ln(m)^{-1-\beta}\right) \text{ for } \beta > 0.$$

Finally, condition $(\mathbf{v})$ of Assumption 4.1 in general rules out long range dependence in the p-th moments of the $(\theta_{i,l})_{i\geq 1}$, for $p = 3,\ldots,6$ and all $l \geq 1$. The given cumulant condition is standard in the time series literature (see, e.g. Brillinger [19], Andrews [4] and Demetrescu et al. [38]) and will provide us with a useful measure of

the joint statistical dependence of higher order moments and a convenient tool for deriving rates of convergence. The following combinatorial representation of p-th order moments in terms of joint cumulants will be particularly useful. For a set of random variables $\theta_1, \ldots, \theta_p$ one has

$$\mathbb{E}\left[\theta_1 \cdot \ldots \cdot \theta_p\right] = \sum_{\pi} \prod_{B \in \pi} \kappa\left(\theta_i : i \in B\right), \tag{4.8}$$

were π cycles through all possible partitions of the set $\{1, 2, \ldots, p\}$ and B cycles through all blocks of partition π.

Furthermore, note that the concept of α-mixing is closely related to the form of time dependencies assumed in $(\mathbf{iv})$-$(\mathbf{v})$. In fact, α-mixing together with finite sixth moments implies absolute summability of the joint cumulants up to sixth order (see, e.g. Andrews [4] or Gonçalves and Kilian [59]). Hence, the main difference between the two approaches lies in the way how autocovariances are handled. In general we find that conditions $(\mathbf{iv})$ and $(\mathbf{v})$ have several advantages in a functional setting: first, they allow for a broader scope of time dependencies (in that absolutely summable autocovariances are not necessary which can be controlled through the β parameter); second, incorporating the decay across the spectral dimension l is straighforward, which is crucial for the analysis; and third, the stated conditions have an intuitive interpretation of the employed time dependency concept for functional data when compared to various mixing properties.

4.3 Applicability of functional principal components analysis

In this section we show that functional principal components analysis is applicable under the stated assumptions on time dependent functional data. As mentioned before, regularization in our modeling framework is achieved through projection on finitely many functional principal components. In practice this means that the infinite sum in (4.4) is truncated at some finite integer L, yielding a sample of truncated Karhunen-Loève decompositions $(X_{i,L})_{i=1}^N$ where

$$X_{i,L}(t) := \mu(t) + \sum_{l=1}^{L} \langle X_i, \psi_l \rangle \psi_l(t) = \mu(t) + \sum_{l=1}^{L} \theta_{i,l} \psi_l(t). \tag{4.9}$$

As a consequence, an initially infinite-dimensional object such as $X_{i,L}$ can be reduced to a finite, countable set of functional principal component scores $(\theta_{i,1}, \ldots, \theta_{i,L})$. Naturally, all quantities above that involve an expectations operator (such as the eigenelements of the covariance operator in (4.3)) have to be estimated on the basis of an observed sample of functions $(X_i)_{i=1}^N$ which gives rise to empirical expansions that approximate (4.4). The standard empirical approximations of $\mu(t)$ and $\mathcal{C}_X[x](t)$ are given by the following sample averages,

$$\hat{\mu}(t) = \frac{1}{N}\sum_{i=1}^{N} X_i(t), \tag{4.10}$$

$$\widehat{\mathcal{C}}_X[x](t) = \frac{1}{N}\sum_{i=1}^{N} \langle X_i - \hat{\mu}, x\rangle \left(X_i(t) - \hat{\mu}(t)\right), \quad x \in L^2. \tag{4.11}$$

We denote the eigenelements of $\widehat{C}_X$ by $(\hat{\lambda}_l)_{l\geq 1}$ and $(\widehat{\psi}_l)_{l\geq 1}$, respectively, such that, upon truncation after the first L functional principal components, the Karhunen-Loève approximation of each function X_i is defined as

$$\widehat{X}_{i,L}(t) := \hat{\mu}(t) + \sum_{l=1}^{L} \langle X_i, \widehat{\psi}_l\rangle \widehat{\psi}_l(t) = \hat{\mu}(t) + \sum_{l=1}^{L} \hat{\theta}_{i,l}\widehat{\psi}_l(t) \tag{4.12}$$

where now $\hat{\theta}_{i,l} := \langle X_i, \widehat{\psi}_l\rangle$ denotes the l-th estimated functional principal component score.

The following results show that in the setting presented above and given a sample of curves $(X_i)_{i=1}^N$, the mean function μ and the covariance operator $\mathcal{C}_X$ as defined in (4.2)-(4.3) can be consistently estimated by $\hat{\mu}$ and $\widehat{\mathcal{C}}_X$, respectively. All proofs are given in Section 4.5.

Theorem 4.1. *If a stationary functional time series $(X_i)_{i=1}^N$ fulfills Assumption 4.1 then*

$$(i) \quad \mathbb{E}\,\|\hat{\mu} - \mu\|^2 = \mathcal{O}\left(N^{-\beta^*}\right),$$
$$(ii) \quad \mathbb{E}\left\|\widehat{\mathcal{C}}_X - \mathcal{C}_X\right\|_{\mathcal{S}}^2 = \mathcal{O}\left(N^{-2\beta^{**}}\right),$$

where $\beta^* := \beta\,\mathbb{I}(0 < \beta < 1) + \mathbb{I}(\beta \geq 1)$, $\beta^{**} := \beta\,\mathbb{I}(0 < \beta < 1/2) + \frac{1}{2}\,\mathbb{I}(\beta \geq 1/2)$ and $\mathbb{I}(\cdot)$ denotes the indicator function.

If, instead, we assume Assumption 4.1 with $(\mathbf{iv})'$, then we have the following corollary to Theorem 4.1.

Corollary 4.1. *If a stationary functional time series $(X_i)_{i=1}^N$ fulfills Assumption 4.1 with $(\mathbf{iv})'$ then*

$$(i) \quad \mathbb{E}\left\|\hat{\mu} - \mu\right\|^2 = o(1),$$
$$(ii) \quad \mathbb{E}\left\|\widehat{\mathcal{C}}_X - \mathcal{C}_X\right\|_{\mathcal{S}}^2 = o(1).$$

The results of Theorem 4.1 should be interpreted as follows. The fastest convergence speed that can be achieved for the empirical estimator of the mean function and the covariance operator to their population counterparts is N^{-1} when $\beta > 1$. In other words, as soon as we allow for absolute summability of the autocovariances in the functional principal component score series $(\theta_{i,l})_{i\geq 1}$ across the time dimension i, the speed of convergence will be the same as under i.i.d. sampling (see, e.g. Horváth and Kokoszka [73, Theorems 2.3 and 2.5]).

Our next result gives explicit bounds for the mean squared error of the eigenelement estimators, again under the time dependency assumption stated in Assumption 4.1.

Corollary 4.2. *If a stationary functional time series $(X_i)_{i=1}^N$ fulfills Assumption 4.1 then for every $l \geq 1$,*

$$(i) \quad \mathbb{E}\left\|c_l\widehat{\psi}_l - \psi_l\right\|^2 = \mathcal{O}\left(\delta_l^2 N^{-2\beta^{**}}\right),$$
$$(ii) \quad \mathbb{E}\left|\hat{\lambda}_l - \lambda_l\right|^2 = \mathcal{O}\left(N^{-2\beta^{**}}\right),$$

where $c_l := sign(\langle\widehat{\psi}_l, \psi_l\rangle)$, $\delta_l := \max_{1\leq k\leq l}(\lambda_k - \lambda_{k+1})^{-1}$ and where $(\lambda_l)_{l\geq 1}$ and $(\psi_l)_{l\geq 1}$ are the eigenelements of the covariance operator $\mathcal{C}_X$ defined in (4.3) *and $(\hat{\lambda}_l)_{l\geq 1}$ and $(\widehat{\psi}_l)_{l\geq 1}$ are the eigenelements of the estimated covariance operator $\widehat{\mathcal{C}}_X$ defined in* (4.11).

The results in Corollary 4.2 indicate that the estimator $\widehat{\psi}_l$ of ψ_l is only identified up to a change in sign. As is standard in the literature, we shall tacitly assume that the sign of $\widehat{\psi}_l$ is chosen such that $\int \widehat{\psi}_l\psi_l \geq 0$.

4.4 Theoretical results

In this section we investigate the statistical properties of the functional one-step ahead predictor X_{N+1}^f of some future function X_{N+1}. Assume, without loss of generality, that the functions X_i have mean zero. In the modeling framework considered here, and as discussed in Section 4.2, a future function X_{N+1} satisfies

$$\begin{aligned} M(x) &= \mathbb{E}\left[X_{N+1} | X_N = x\right] \\ &= \sum_{l=1}^{\infty} m_l(\theta_l)\psi_l, \end{aligned}$$

where $m_l(\theta_l) = \mathbb{E}[\theta_{N+1,l}|\theta_{N,l} = \theta_l]$. A predictor X_{N+1}^f can thus be devised by providing estimators for the functional principal components ψ_l as well as for the conditional means $m_l(\theta_l)$ and truncating the infinite sum in the above display at some finite integer L. The predictor X_{N+1}^f then takes on the form

$$X_{N+1}^f = \widehat{M}(\hat{x}) = \sum_{l=1}^{L} \widehat{m}_l(\hat{\theta}_l)\widehat{\psi}_l.$$

While estimation of the functional principal components ψ_l has already been discussed in Section 4.3, we propose to estimate the conditional means $m_l(\theta_l)$ by a k-nearest neighbors classification approach. Informally speaking, this considers estimating $m_l(\theta_l)$ by a sample average of those $\theta_{i+1,l}$'s for which the preceding $\theta_{i,l}$'s are closest to the feature score component θ_l.

In what follows, we make this approach precise and define several quantities that allow us to separately analyze the effects of truncation and estimation.

4.4.1 Definitions and notation

Consider first the effect of truncating the (infinite-dimensional) Karhunen-Loève expansion of the functions X_i at some finite integer L but assuming all quantities in the expansion to be known. We denote these truncated expansions by

$$X_{i,L}(t) := \sum_{l=1}^{L} \theta_{i,l}\psi_l(t)$$

and the truncated realization by

$$x_L(t) := \sum_{l=1}^{L} \theta_l \psi_l(t).$$

As before, the truncated feature element x_L is thus characterized by a finite, countable set of functional principal component scores $(\theta_1, \ldots, \theta_L)$.

Now take, for each $l = 1, \ldots, L$, the number of neighbors to θ_l to depend on the sample size N in that $k_N \to \infty$ as $N \to \infty$. Furthermore, denote the index set of the k_N closest neighbors of the series $(\theta_{i,l})_{i=1}^N$ to the feature score component θ_l by $\mathcal{I}(k_N; \theta_l)$ and define the *infeasible* estimator $m_{l,N}(\theta_l)$ of $m_l(\theta_l)$ to be given by

$$m_{l,N}(\theta_l) := \frac{1}{k_N} \sum_{i \in \mathcal{I}(k_N; \theta_l)} \theta_{i+1,l}. \tag{4.13}$$

This is a straightforward implementation of the k_N-NN classifier to estimate $m_l(\theta_l) = \mathbb{E}[\theta_{N+1,l} | \theta_{N,l} = \theta_l]$. Consequently, the *infeasible* functional predictor $X_{N+1}^f = M_{N,L}(x_L)$ based on a truncated sample $(X_{i,L})_{i=1}^N$ is defined by

$$M_{N,L}(x_L) := \sum_{l=1}^{L} m_{l,N}(\theta_l) \psi_l. \tag{4.14}$$

For future reference we furthermore define

$$\begin{aligned} M_L(x_L) &:= \sum_{l=1}^{L} \mathbb{E}\left[\theta_{N+1,l} | \theta_{N,l} = \theta_l\right] \psi_l \\ &= \sum_{l=1}^{L} m_l(\theta_l) \psi_l, \end{aligned} \tag{4.15}$$

where in comparison to (4.14) the k_N-NN estimators of the scores have been replaced by the corresponding conditional population means.

While the truncated estimator $M_{N,L}(x_L)$ is infeasible since the elements of the Karhunen-Loève decomposition are unknown, we consider a *feasible* version that replaces the elements in the series expansion of the $X_{i,L}$ with estimated quantities. As above, this yields a series of truncated Karhunen-Loève approximations which are given

by

$$\widehat{X}_{i,L}(t) := \sum_{l=1}^{L} \hat{\theta}_{i,l} \widehat{\psi}_l(t).$$

Note that in comparison to $X_{i,L}$, all quantities involved in the series expansion are now estimated on a sample of N functional observations. Similarly to above, we denote the corresponding truncated and estimated realization by

$$\hat{x}_L(t) := \sum_{l=1}^{L} \hat{\theta}_l \widehat{\psi}_l(t)$$

which is now characterized by a finite, countable set of estimated functional principal component scores $(\hat{\theta}_1, \ldots, \hat{\theta}_L)$. Denote the index set of the k_N closest neighbors of the series $(\hat{\theta}_{i,l})_{i=1}^N$ to the estimated feature score component $\hat{\theta}_l$ by $\widehat{\mathcal{I}}(k_N; \hat{\theta}_l)$ and define the *feasible* estimator $\widehat{m}_{l,N}(\hat{\theta}_l)$ of $m_l(\theta_l)$ to be given by

$$\widehat{m}_{l,N}(\theta_l) := \frac{1}{k_N} \sum_{i \in \widehat{\mathcal{I}}(k_N; \hat{\theta}_l)} \hat{\theta}_{i+1,l}. \tag{4.16}$$

Consequently, the *feasible* functional predictor $X_{N+1}^f = \widehat{M}_{N,L}(\hat{x}_L)$ based on a truncated sample of Karhunen-Loève approximations $(\widehat{X}_{i,L})_{i=1}^N$ is defined by

$$\widehat{M}_{N,L}(\hat{x}_L) := \sum_{l=1}^{L} \widehat{m}_{l,N}(\hat{\theta}_l) \widehat{\psi}_l. \tag{4.17}$$

The remainder of this section proceeds by first considering the effect of truncation, i.e. showing that $M_{N,L}(x_L)$ converges to $M(x)$ with a suitable rate. In a second step we show convergence of $\widehat{M}_{N,L}(\hat{x}_L)$ to $M_{N,L}(x_L)$.

4.4.2 The effect of truncation

In this section we consider the infeasible estimator $M_{N,L}(x_L)$ based on a truncated sample $(X_{i,L})_{i=1}^N$. More precisely, we study convergence rates of

$$\mathbb{E}\|M_{N,L}(x_L) - M(x)\|^2. \tag{4.18}$$

Upon adding and subtracting $M_L(x_L)$ to the argument in (4.18), and imposing the standard triangle inequality for L^2-norms, the Cauchy-Schwarz and Jensen's inequality it suffices to consider the terms

$$\mathbb{E}\|M_L(x_L) - M(x)\|^2 \quad \text{and} \quad \mathbb{E}\|M_{N,L}(x_L) - M_L(x_L)\|^2,$$

for which we obtain the following rates of convergence, the proof of which is given in Section 4.5.

Theorem 4.2. *We have for some $\alpha > 1$ and $k_N \sim \lfloor N^{4/5} \rfloor$,*

$$\begin{aligned} &(i) \quad \mathbb{E}\|M_L(x_L) - M(x)\|^2 = \mathcal{O}\left((L+1)^{1-\alpha}\right), \\ &(ii) \quad \mathbb{E}\|M_{N,L}(x_L) - M_L(x_L)\|^2 = \mathcal{O}\left(k_N^{-1}\right). \end{aligned}$$

Here $\lfloor \cdot \rfloor$ denotes the integer part of the argument. As a consequence, we obtain the following convergence rate for the mean squared error of truncation,

$$\mathbb{E}\|M_{N,L}(x_L) - M(x)\|^2 = \mathcal{O}\left(\max\left((L+1)^{1-\alpha},\, k_N^{-1},\, (L+1)^{(1-\alpha)/2} k_N^{-1/2}\right)\right).$$

Note that $\alpha > 1$ in view of summability of the sequence of eigenvalues $(\lambda_l)_{l\geq 1}$ which implies convergence to zero of all terms in the above display as both N and L tend to infinity.

4.4.3 The effect of estimation

We now consider the effect of estimation by showing that the feasible estimator $\widehat{M}_{N,L}(\hat{x}_L)$ based on a sample of approximated functions $(\widehat{X}_{i,L})_{i=1}^N$ converges to the infeasible version $M_{N,L}(x_L)$, i.e we study

$$\mathbb{E}\left\|\widehat{M}_{N,L}(\hat{x}_L) - M_{N,L}(x_L)\right\|^2. \tag{4.19}$$

Using (4.17) and (4.14) and upon adding and subtracting $\sum_{l=1}^L m_{l,N}(\theta_l)\widehat{\psi}_l$ to the argument of the quantity of interest it suffices (again in view of the triangle, Cauchy-Schwarz and Jensen inequalities) to analyze the quantities

$$\mathbb{E}\left\|\sum_{l=1}^L m_{l,N}(\theta_l)\left(\widehat{\psi}_l - \psi_l\right)\right\|^2 \quad \text{and} \quad \mathbb{E}\left\|\sum_{l=1}^L \left(\widehat{m}_{l,N}(\hat{\theta}_l) - m_{l,N}(\theta_l)\right)\widehat{\psi}_l\right\|^2.$$

The convergence rates for these quantities are given in the following theorem, the proof of which can again be found in Section 4.5. Note that if we were not to assume that the X_i have mean zero, then an additional term $\mathbb{E}\|\hat{\mu} - \mu\|^2$ would have to be considered. Convergence rates for this expression have, however, been already derived in Theorem 4.1 and, as the following theorem shows, are of faster order.

Theorem 4.3. *We have for some* $\alpha > 1$, $\beta^* := \beta\, \mathbb{I}(0 < \beta < 1) + \mathbb{I}(\beta > 1)$, $\beta^{**} := \beta\, \mathbb{I}(0 < \beta < 1/2) + \frac{1}{2}\, \mathbb{I}(\beta > 1/2)$ *and* $k_N \sim \lfloor N^{4/5} \rfloor$,

$$
\begin{aligned}
(i) \quad & \mathbb{E} \left\| \sum_{l=1}^{L} m_{l,N}(\theta_l) \left(\widehat{\psi}_l - \psi_l \right) \right\|^2 = \mathcal{O}\left(\frac{L^{3+\alpha}}{k_N^{\beta^*} N^{2\beta^{**}}} \right), \\
(ii) \quad & \mathbb{E} \left\| \sum_{l=1}^{L} \left(\widehat{m}_{l,N}(\hat{\theta}_l) - m_{l,N}(\theta_l) \right) \widehat{\psi}_l \right\|^2 \\
& \qquad = \mathcal{O}\left(\max\left(\frac{L^{3+2\alpha} \log(N)}{N^{2\beta^{**}}}, \frac{L^{\frac{3}{2}-\frac{\alpha}{2}} (\log(N))^{\frac{1}{4}}}{N^{\frac{1}{2}\beta^{**}}} \right) \right).
\end{aligned}
$$

The results of Theorem 4.3 now allow us to precisely state how fast L can grow for the proposed FAM-knn method to still be consistent. Assume that the number of principal components considered depends on the sample size such that $L \sim \lfloor N^{\gamma} \rfloor$ for some $\gamma > 0$. Then from part (i) in Theorem 4.3 we need

$$\gamma(3 + \alpha) < \frac{4}{5}\beta^* + 2\beta^{**}.$$

In the most favorable case, $\beta^* = 1$ and $\beta^{**} = 1/2$ such that for α just larger than 1 one obtains $\gamma < 9/20$. By the same arguments, the first term on the right hand side of part (ii) in Theorem 4.3 dominates. Since moreover the $\log(N)$ term is asymptotically negligible, we need

$$\gamma(3 + 2\alpha) < 2\beta^{**}$$

which in the most favorable case of $\beta^{**} = 1/2$ and α just larger than 1 gives $\gamma < 4/20$. Intuitively, both the sample size and the number of principal components of the Karhunen-Loève decomposition has to go to infinity. However, the number of principal components L cannot grow too fast as it becomes increasingly difficult to estimate the corresponding eigenelements of the covariance operator.

4.5 Proofs

4.5.1 Preliminary results

We first state some preliminary results as a series of lemmata.

Lemma 4.1. *If $(x_n)_{n\geq 1}$ is a real positive sequence with $x_n \sim n^{\alpha}$ then*

(i) if $\alpha > -1$ then $\sum_{m=1}^{n} x_m \sim n^{1+\alpha}$;

(ii) if $\alpha = -1$ then $\sum_{m=1}^{n} x_m \sim \log(n)$;

(iii) if $\alpha < -1$ then $\sum_{m=1}^{\infty} x_m < \infty$ and $\sum_{m=n}^{\infty} x_m = \mathcal{O}(n^{1+\alpha})$.

Proof. The proof can be found in Davidson [35, Theorem 2.27]. □

A direct consequence of Lemma 4.1 is the following Lemma on the behavior of the sequence of eigenvalues $(\lambda_l)_{l\geq 1}$.

Lemma 4.2. *We have for the series of eigenvalues $(\lambda_l)_{l\geq 1}$ with $\lambda_l \sim l^{-\alpha}$ for some $\alpha > 1$*

(i) $\sum_{l=1}^{\infty} \lambda_l < \infty$;

(ii) $\sum_{l=L+1}^{\infty} \lambda_l = \mathcal{O}((L+1)^{1-\alpha})$;

(iii) $\sum_{l=1}^{\infty} \lambda_l^2 < \infty$.

4.5.2 Proof of Theorem 4.1

Proof of statement (i). We have

$$\begin{aligned}\mathbb{E}\left\|\hat{\mu}-\mu\right\|^2 &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left\langle X_i-\mu, X_j-\mu\right\rangle \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left\|X_i-\mu\right\|^2 + \frac{1}{N^2}\sum_{i\neq j}^{N}\sum\mathbb{E}\left\langle X_i-\mu, X_j-\mu\right\rangle.\end{aligned}$$

As a consequence of part $(\mathbf{i})$ of Assumption 4.1 the second moments of X_i are finite for all $i \geq 1$ such that the first term in the last equation above behaves as $\mathcal{O}\left(N^{-1}\right)$.

Rearranging the second term and invoking Assumption 4.1 $(\mathbf{iv})$ gives

$$\begin{aligned}\frac{1}{N^2}\sum_{i\neq j}^{N}\sum \mathbb{E}\left\langle X_i-\mu, X_j-\mu\right\rangle &= \frac{2}{N^2}\sum_{m=1}^{N-1}\sum_{i=m+1}^{N}\sum_{l=1}^{\infty}\mathbb{E}\left(\theta_{i,l},\theta_{i-m,l}\right)\\ &\leq \frac{2}{N^2}\sum_{m=1}^{N-1}\sum_{i=m+1}^{N}\sum_{l=1}^{\infty}B_{m,l}\\ &\leq \frac{C}{N^2}\sum_{m=1}^{N-1}(N-m)m^{-\beta}\sum_{l=1}^{\infty}\lambda_l.\end{aligned}$$

Assumption 4.1 $(\mathbf{i})$ again implies that $\sum_{l=1}^{\infty}\lambda_l$ is bounded. Then using that $(N-m)/N<1$ the result follows in view of Lemma 4.1. □

Proof of statement (ii). Assume for simplicity and without loss of generality that $\mu=0$. Then by definition of the Hilbert-Schmidt norm and orthonormality of the sequence of eigenfunctions $(\psi_h)_{h\geq 1}$,

$$\begin{aligned}&\mathbb{E}\left\|\widehat{\mathcal{C}}_X-\mathcal{C}_X\right\|_{\mathcal{S}}^2\\ &\quad=\sum_{h_1=1}^{\infty}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\left\langle X_i,\psi_{h_1}\right\rangle X_i-\mathbb{E}\left[\left\langle X_i,\psi_{h_1}\right\rangle X_i\right]\right)\right\|^2\\ &\quad\leq\frac{1}{N^2}\sum_{i=1}^{N}\sum_{h_1=1}^{\infty}\mathbb{E}\left\|Y_{i,h_1}\right\|^2+\frac{1}{N^2}\sum_{i\neq j}^{N}\sum\sum_{h_1=1}^{\infty}\mathbb{E}\left\langle Y_{i,h_1}Y_{j,h_1}\right\rangle,\end{aligned}\tag{4.20}$$

where $Y_{i,h}:=\left\langle X_i,\psi_h\right\rangle X_i-\mathbb{E}\left[\left\langle X_i,\psi_h\right\rangle X_i\right]$.

The first term in (4.20) is of order $\mathcal{O}\left(N^{-1}\right)$ since

$$\begin{aligned}
&\frac{1}{N^2}\sum_{i=1}^{N}\sum_{h_1=1}^{\infty}\mathbb{E}\left\|\left\langle X_i,\psi_{h_1}\right\rangle X_i-\mathbb{E}\left[\left\langle X_i,\psi_{h_1}\right\rangle X_i\right]\right\|^2\| \\
&\quad\leq\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\sum_{h_1=1}^{\infty}\left\|\left\langle X_i,\psi_{h_1}\right\rangle X_i\right\|^2\right] \\
&\quad=\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\|X_i\|^2\sum_{h_1=1}^{\infty}\left\langle X_i,\psi_{h_1}\right\rangle^2\right] \\
&\quad=\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\|X_i\|^4\leq\frac{C}{N}.
\end{aligned}$$

The second term in (4.20) is handled as follows. First note that in view of the definition of the $Y_{n,h}$ and the Karhunen-Loève decomposition of each X_i, X_j we have

$$\sum_{h_1=1}^{\infty}\mathbb{E}\left\langle Y_{i,h_1}Y_{j,h_1}\right\rangle=\sum_{h_1,h_2=1}^{\infty}\sum\mathbb{E}\left[\theta_{i,h_1}\theta_{j,h_1}\theta_{i,h_2}\theta_{j,h_2}\right]-\sum_{h_1=1}^{\infty}\lambda_{h_1}^2,$$

which yields

$$\begin{aligned}
&\frac{1}{N^2}\sum_{i\neq j}^{N}\sum_{h_1=1}^{\infty}\mathbb{E}\left\langle Y_{i,h_1}Y_{j,h_1}\right\rangle \\
&\quad=\frac{1}{N^2}\sum_{i\neq j}^{N}\sum\sum_{h_1,h_2=1}^{\infty}\mathbb{E}\left[\theta_{i,h_1}\theta_{j,h_1}\theta_{i,h_2}\theta_{j,h_2}\right]-\frac{1}{N^2}\sum_{i\neq j}^{N}\sum_{h_1=1}^{\infty}\lambda_{h_1}^2.
\end{aligned}$$

Distinguishing the cases $h_1=h_2$ and $h_1\neq h_2$ in the above display then gives

$$\begin{aligned}
&\frac{1}{N^2}\sum_{i\neq j}^{N}\sum_{h_1=1}^{\infty}\mathbb{E}\left\langle Y_{i,h_1}Y_{j,h_1}\right\rangle \\
&\quad=\frac{1}{N^2}\sum_{i\neq j}^{N}\sum_{h_1=1}^{\infty}\mathbb{E}\left[\theta_{i,h_1}^2\theta_{j,h_1}^2\right]+\frac{1}{N^2}\sum_{i\neq j}^{N}\sum\sum_{h_1\neq h_2}^{\infty}\mathbb{E}\left[\theta_{i,h_1}\theta_{j,h_1}\theta_{i,h_2}\theta_{j,h_2}\right] \\
&\qquad-\frac{1}{N^2}\sum_{i\neq j}^{N}\sum_{h_1=1}^{\infty}\lambda_{h_1}^2. \qquad (4.21)
\end{aligned}$$

For the first term in (4.21) we have by the relationship of higher-order moments to joint cumulants (4.8),

$$\begin{aligned}\frac{1}{N^2}\sum_{i\neq j}^{N}\sum_{h_1=1}^{\infty}\mathbb{E}\left[\theta_{i,h_1}^2\theta_{j,h_1}^2\right] &= \frac{1}{N^2}\sum_{i\neq j}^{N}\sum_{h_1=1}^{\infty}\kappa_{h_1}(0,0,|i-j|,|i-j|)\\ &\quad+\frac{1}{N^2}\sum_{i\neq j}^{N}\sum_{h_1=1}^{\infty}\mathbb{E}\left[\theta_{i,h_1}^2\right]\mathbb{E}\left[\theta_{j,h_1}^2\right]\\ &\quad+\frac{2}{N^2}\sum_{i\neq j}^{N}\sum_{h_1=1}^{\infty}\mathbb{E}\left[\theta_{i,h_1}\theta_{j,h_1}\right]^2. \qquad (4.22)\end{aligned}$$

First note that the second term in (4.22) cancels out the third term in (4.21) since $\mathbb{E}[\theta_{i,h_1}^2] = \mathbb{E}[\theta_{j,h_1}^2] = \lambda_{h_1}$. The first term in (4.22) is of order $\mathcal{O}(N^{-1})$ since

$$\begin{aligned}&\frac{1}{N^2}\sum_{i\neq j}^{N}\sum_{h_1=1}^{\infty}\kappa_{h_1}(0,0,|i-j|,|i-j|)\\ &\leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{h_1=1}^{\infty}\sum_{\tau_1,\tau_2,\tau_3=-\infty}^{\infty}\left|\kappa_{h_1}(0,\tau_1,\tau_2,\tau_3)\right|\\ &\leq \frac{C}{N}\sum_{h_1=1}^{\infty}\lambda_{h_1}^2 \leq \frac{C}{N}.\end{aligned}$$

For the second term in (4.22) we have, for some constant $C > 0$,

$$\begin{aligned}\frac{2}{N^2}\sum_{i\neq j}^{N}\sum_{h_1=1}^{\infty}\mathbb{E}\left[\theta_{i,h_1}\theta_{j,h_1}\right]^2 &\leq \frac{4}{N^2}\sum_{m=1}^{N-1}\sum_{i=m+1}^{N}\sum_{h_1=1}^{\infty}B_{m,h_1}^2\\ &\leq \frac{C}{N}\sum_{m=1}^{N-1}m^{-2\beta}\sum_{h_1=1}^{\infty}\lambda_{h_1}^2\\ &= \mathcal{O}\left(N^{-2\beta^{**}}\right),\end{aligned}$$

where $\beta^{**} := \beta\,\mathbb{I}(0 < \beta < 1/2) + \frac{1}{2}\,\mathbb{I}(\beta \geq 1/2)$.

What remains to be discussed is the second term in (4.21) for which we have by similar arguments

$$\begin{aligned}
&\frac{1}{N^2}\sum_{i\neq j}^{N}\sum\sum_{h_1\neq h_2}^{\infty}\mathbb{E}\left[\theta_{i,h_1}\theta_{j,h_1}\right]\mathbb{E}\left[\theta_{i,h_2}\theta_{j,h_2}\right]\\
&\quad\leq \frac{1}{N^2}\sum_{i\neq j}^{N}\sum\sum_{h_1\neq h_2}^{\infty}\left|\mathbb{E}\left[\theta_{i,h_1}\theta_{j,h_1}\right]\right|\left|\mathbb{E}\left[\theta_{i,h_2}\theta_{j,h_2}\right]\right|\\
&\quad\leq \frac{1}{N^2}\sum_{m=1}^{N-1}\sum_{i=m+1}^{N}\sum_{h_1,h_2=1}^{\infty}B_{m,h_1}B_{m,h_2}\\
&\quad\leq \frac{C}{N}\sum_{m=1}^{N-1}m^{-2\beta}\sum_{h_1=1}^{\infty}\lambda_{h_1}\sum_{h_2=1}^{\infty}\lambda_{h_2}\\
&\quad= \mathcal{O}\left(N^{-2\beta^{**}}\right),
\end{aligned}$$

where $\beta^{**} := \beta\,\mathbb{I}(0 < \beta < 1/2) + \frac{1}{2}\,\mathbb{I}(\beta \geq 1/2)$. The desired result now follows upon observing that the dominant convergence rate is precisely of the required order. □

4.5.3 Proof of Corollary 4.1

Proof. It suffices to show convergence of $\frac{1}{N}\sum_{m=1}^{N-1} b_m$ to zero as $N \to \infty$. For this we make use of Kronecker's lemma which states that for positive real sequences $(x_i)_{i\geq 1}$ and $(a_i)_{i\geq 1}$ with $a_i \uparrow \infty$ one has

$$\sum_{i=1}^{N}\frac{x_i}{a_i} \;\to\; C \quad \text{implies} \quad \frac{1}{a_N}\sum_{i=1}^{N}x_i \;\to\; 0,$$

as $N \to \infty$. The corollary now follows from the proof of the Theorem 4.1 together with the fact that $\frac{1}{N}\sum_{m=1}^{N-1} b_m \to 0$. □

4.5.4 Proof of Corollary 4.2

Proof. This proof requires two auxiliary results formulated in the following lemma.

Lemma 4.3. *Under the conditions of Theorem 4.1, one has for each* $l \geq 1$

$$\begin{aligned} |\hat{\lambda}_l - \lambda_l| &\leq \left\| \widehat{\mathcal{C}}_X - \mathcal{C}_X \right\|_{\mathcal{L}}, \quad (4.23) \\ \left\| c_l \widehat{\psi}_l - \psi_l \right\| &\leq C\delta_l \left\| \widehat{\mathcal{C}}_X - \mathcal{C}_X \right\|_{\mathcal{L}}, \quad (4.24) \end{aligned}$$

where $c_l = sign\left(\langle \widehat{\psi}_l, \psi_l \rangle\right)$, $\delta_l = \max_{1 \leq k \leq l} (\lambda_k - \lambda_{k+1})^{-1}$, $C > 0$ *and* $\|\cdot\|_{\mathcal{L}}$ *denotes the operator norm for the space of bounded linear operators* $\mathcal{L}$ *on* $L^2(\mathcal{T}, \|\cdot\|)$.

Both results (4.23) and (4.24) follow from Bosq [18, Lemma 4.2 and 4.3], respectively. Given the result in Lemma 4.3, the proof of Corollary 4.2 is straightforward since the inequalities (4.23) and (4.23) together with the fact that $\|\cdot\|_{\mathcal{L}} \leq \|\cdot\|_{\mathcal{S}}$ yield

$$\begin{aligned} |\hat{\lambda}_l - \lambda_l|^2 &\leq \left\| \widehat{\mathcal{C}}_X - \mathcal{C}_X \right\|_{\mathcal{S}}^2 \\ \left\| c_l \widehat{\psi}_l - \psi_l \right\|^2 &\leq \delta_l^2 \left\| \widehat{\mathcal{C}}_X - \mathcal{C}_X \right\|_{\mathcal{S}}^2. \end{aligned}$$

Then Theorem 4.1 implies the desired results. □

4.5.5 Proof of Theorem 4.2

Since our interest is in analyzing $\mathbb{E}\|M_{N,L}(x_L) - M(x)\|^2$, it suffices, upon adding and subtracting $M_L(x_L)$ in the argument of our object of interest, to consider the two terms

$$\mathbb{E}\|M_{N,L}(x_L) - M_L(x_L)\|^2, \quad (4.25)$$
$$\mathbb{E}\|M_L(x_L) - M(x)\|^2. \quad (4.26)$$

Proof of statement (i). We start with analyzing (4.26) and first recall that we have

$$\begin{aligned} M(x) &= \sum_{l=1}^{\infty} \mathbb{E}\left[\theta_{N+1,l} | \theta_{N,l} = \theta_l\right] \psi_l \\ &= \sum_{l=1}^{\infty} m_l(\theta_l) \psi_l. \quad (4.27) \end{aligned}$$

We then have, using the definitions in (4.15) and (4.27) and orthonormality of the ψ_l

$$\begin{aligned}\mathbb{E}\|M_L(x_L) - M(x)\|^2 &= \mathbb{E}\left\|\sum_{l=1}^{L} m_l(\theta_l)\psi_l - \sum_{l=1}^{\infty} m_l(\theta_l)\psi_l\right\|^2 \\ &= \mathbb{E}\left\|\sum_{l=L+1}^{\infty} m_l(\theta_l)\psi_l\right\|^2 \\ &= \sum_{l=L+1}^{\infty} \mathbb{E}\left[m_l(\theta_l)^2\right]. \end{aligned} \tag{4.28}$$

Now observe that the summand in (4.28) can be expressed as

$$\begin{aligned}\mathbb{E}\left[m_l(\theta_l)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\theta_{N+1,l}|\theta_{N,l} = \theta_l\right]^2\right] \\ &= \mathbb{E}\left[\theta_{N+1,l}^2\right] - \mathbb{E}\left[\mathbb{V}\left[\theta_{N+1,l}|\theta_{N,l} = \theta_l\right]\right] \\ &\leq \lambda_l,\end{aligned}$$

where the last inequality follows in view of $\mathbb{V}\left[\theta_{N+1,l}\right] = \mathbb{E}\left[\theta_{N+1,l}^2\right] = \lambda_l$ and the fact that $0 \leq \mathbb{E}\left[\mathbb{V}\left[\theta_{N+1,l}|\theta_{N,l} = \theta_l\right]\right] \leq \mathbb{V}\left[\theta_{N+1,l}\right]$. Now note that by assumption, $\lambda_l \sim l^{-\alpha}$ for some $\alpha > 1$. We thus have that (4.28) is of order $\mathcal{O}((L+1)^{1-\alpha})$ in view of the second statement of Lemma 4.2. □

Proof of statement (ii). Now we consider (4.25) which, upon using the definitions in (4.14) and (4.15), can be written as

$$\begin{aligned}\mathbb{E}\|M_{N,L}(x_L) - M_L(x_L)\|^2 &= \mathbb{E}\left\|\sum_{l=1}^{L}\left(m_{l,N}(\theta_l) - m_l(\theta_l)\right)\psi_l\right\|^2 \\ &= \sum_{l=1}^{L}\mathbb{E}\left[\left(m_{l,N}(\theta_l) - m_l(\theta_l)\right)^2\right], \end{aligned} \tag{4.29}$$

where the second equality follows again in view of the orthonormality of the sequence of eigenfunctions $(\psi_l)_{l=1}^{L}$. For fixed $l = 1, \ldots, L$, rates of convergence of the mean squared error in (4.29) can be derived by following results in Yakowitz [108]. A careful inspection of the proofs in Yakowitz [108] reveals that analyzing the second moment of the distance between (the given) θ_l and its farthest (of the k_N) neighbor is of key importance. Denote this farthest neighbor to θ_l by $\theta_{N(k_N),l}$ and write $R_{i,l}(\theta_l) :=$

$|\theta_{i,l} - \theta_l|$ such that $R_{(k_N),l}(\theta_l) := |\theta_{N(k_N),l} - \theta_l|$ denotes the k_N-th order statistic of the $R_{i,l}(\theta_l)$. Results in Yakowitz [108] indicate that $\mathbb{E}[R_{(k_N),l}(\theta_l)^2] \leq C_1(l)k_N^{-1/2}$, where $C_1(l)$ is some constant that depends only on l. While this holds true for fixed l, we have to consider asymptotics where L goes to infinity. Now observe that

$$\mathbb{E}\left[R_{(k_N),l}(\theta_l)^2\right] = \mathbb{E}\left[|\theta_{N(k_N),l} - \theta_l|^2\right] \leq C_2(N)\lambda_l$$

for fixed N, where $C_2(N)$ is some constant only depending on N. Combining these results gives us $\mathbb{E}[R_{(k_N),l}(\theta_l)^2] \leq C_3 k_N^{-1/2}\lambda_l$, where now C_3 is a constant that is independent of both l and N. Moreover, Yakowitz [108] shows that the number of neighbors k_N has to grow with the sample size where $k_N \sim \lfloor N^{4/5} \rfloor$.

The desired result now follows in view of Lemma 4.1, the main result of Yakowitz [108, Theorem 2.1] and the arguments presented above.

□

4.5.6 Proof of Theorem 4.3

Proof of statement (i). We assume for simplicity, and without loss of generality, that the functions $(X_i)_{i=1}^N$ have mean zero. As before, we denote, for $i = 1, \ldots, k_N$, by $N(i) \in \mathcal{I}(k_N; \theta_l)$ the index of the i-th nearest neighbor to θ_l. We then have in view of the Cauchy-Schwarz inequality, definition (4.14) and Lemma 4.3

$$\begin{aligned}
&\mathbb{E}\left\|m_{l,N}(\theta_l)\left(\widehat{\psi}_l - \psi_l\right)\right\|^2 \\
&\quad = \mathbb{E}\left[\sum_{l,k=1}^{L} m_{l,N}(\theta_l)m_{k,N}(\theta_k)\left\langle \widehat{\psi}_l - \psi_l, \widehat{\psi}_k - \psi_k \right\rangle\right] \\
&\quad \leq \mathbb{E}\left[\sum_{l,k=1}^{L} m_{l,N}(\theta_l)m_{k,N}(\theta_k)\left\|\widehat{\psi}_l - \psi_l\right\|\left\|\widehat{\psi}_k - \psi_k\right\|\right] \\
&\quad \leq \frac{1}{k_N^2}\sum_{l,k=1}^{L}\sum_{i,j=1}^{k_N} \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,k}\delta_l\delta_k\left\|\widehat{\mathcal{C}}_X - \mathcal{C}_X\right\|_{\mathcal{S}}^2\right]. \qquad (4.30)
\end{aligned}$$

As already detailed in Section 4.3, we define, for any sequence $(e_h)_{h\geq 1}$ of orthonormal basis functions, $Y_{n,h} := \langle X_n, e_h \rangle X_n - \mathbb{E}\left[\langle X_n, e_h \rangle X_n\right]$ such that we have

$$\begin{aligned}\left\|\widehat{\mathcal{C}}_X - \mathcal{C}_X\right\|_{\mathcal{S}}^2 &= \sum_{h_1=1}^{\infty} \left\| \frac{1}{N} \sum_{n=1}^{N} Y_{n,h_1} \right\|^2 \\ &= \frac{1}{N^2} \sum_{h_1=1}^{\infty} \sum_{n,m=1}^{N} \left\langle Y_{n,h_1}, Y_{m,h_1} \right\rangle .\end{aligned}$$

The expression in (4.30) can thus be written as

$$\begin{aligned}&\frac{1}{k_N^2} \sum_{l,k=1}^{L} \sum_{i,j=1}^{k_N} \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,k}\delta_l\delta_k \left\|\widehat{\mathcal{C}}_X - \mathcal{C}_X\right\|_{\mathcal{S}}^2\right] \\ &\quad= \frac{1}{k_N^2} \sum_{l,k=1}^{L} \sum_{i,j=1}^{k_N} \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,k}\delta_l\delta_k \frac{1}{N^2} \sum_{n,m=1}^{N} \sum_{h_1=1}^{\infty} \left\langle Y_{n,h_1}, Y_{m,h_1} \right\rangle\right] \\ &\quad= \frac{1}{k_N^2 N^2} \sum_{l,k=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \sum_{h_1=1}^{\infty} \delta_l\delta_k \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,k} \left\langle Y_{n,h_1}, Y_{m,h_1} \right\rangle\right] .\end{aligned} \tag{4.31}$$

Observe that by definition of the $Y_{n,h}$ and properties of the $\theta_{n,h} = \langle X_n, \psi_h \rangle$ we have upon taking $(e_h)_{h\geq 1} = (\psi_h)_{h\geq 1}$,

$$\begin{aligned}
\sum_{h_1=1}^{\infty} \langle Y_{n,h_1}, Y_{m,h_1} \rangle &= \sum_{h_1=1}^{\infty} \left\langle \langle X_n, \psi_{h_1} \rangle X_n, \langle X_m, \psi_{h_1} \rangle X_m \right\rangle \\
&\quad + \sum_{h_1=1}^{\infty} \left\langle \mathbb{E}\left[\langle X_n, \psi_{h_1} \rangle X_n \right], \mathbb{E}\left[\langle X_m, \psi_{h_1} \rangle X_m \right] \right\rangle \\
&\quad - \sum_{h_1=1}^{\infty} \left\langle \langle X_n, \psi_{h_1} \rangle X_n, \mathbb{E}\left[\langle X_m, \psi_{h_1} \rangle X_m \right] \right\rangle \\
&\quad - \sum_{h_1=1}^{\infty} \left\langle \langle X_m, \psi_{h_1} \rangle X_m, \mathbb{E}\left[\langle X_n, \psi_{h_1} \rangle X_n \right] \right\rangle, \\
&= \sum_{h_1,h_2=1}^{\infty} \theta_{n,h_1}\theta_{n,h_2}\theta_{m,h_1}\theta_{m,h_2} \\
&\quad + \sum_{h_1=1}^{\infty} \lambda_{h_1}^2 - \sum_{h_1=1}^{\infty} \lambda_{h_1}\theta_{n,h_1}^2 - \sum_{h_1=1}^{\infty} \lambda_{h_1}\theta_{m,h_1}^2. \qquad (4.32)
\end{aligned}$$

In view of (4.32), the expression in (4.31) can be decomposed into

$$A_1 + A_2 - 2A_3,$$

where

$$\begin{aligned}
A_1 &:= \frac{1}{k_N^2 N^2} \sum_{l,k=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \sum_{h_1,h_2=1}^{\infty} \delta_l \delta_k \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,k}\theta_{n,h_1}\theta_{n,h_2}\theta_{m,h_1}\theta_{m,h_2} \right], \\
A_2 &:= \frac{1}{k_N^2 N^2} \sum_{l,k=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \sum_{h_1=1}^{\infty} \delta_l \delta_k \lambda_{h_1}^2 \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,k} \right], \\
A_3 &:= \frac{1}{k_N^2 N^2} \sum_{l,k=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \sum_{h_1=1}^{\infty} \delta_l \delta_k \lambda_{h_1} \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,k}\theta_{n,h_1}^2 \right].
\end{aligned}$$

The analysis of the terms above now proceeds by considering the relationship between higher order moments and joint cumulants as defined in (4.8) and noting that the random variables $\theta_{\cdot,h} = \langle X_\cdot, \psi_h \rangle$ have zero mean by construction and are independent across h by assumption.

We start with term A_2 and first note that the relevant case for us to consider is $l = k$ as otherwise $A_2 = 0$ by the above arguments. Distinguishing the cases where $l \neq h_1$ and $l = h_1$ then yields

$$\begin{aligned} A_2 &= \frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \sum_{h_1=L+1}^{\infty} \delta_l^2 \lambda_{h_1}^2 \kappa_l(0,|N(i)-N(j)|) \\ &\quad + \frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \delta_l^2 \lambda_l^2 \kappa_l(0,|N(i)-N(j)|) \\ &=: A_{2,1} + A_{2,2}. \end{aligned} \tag{4.33}$$

Now consider the term A_3 and again note that it suffices to consider only the case $l = k$. Again distinguishing the cases where $l \neq h_1$ and $l = h_1$ we have by (4.8) that

$$\begin{aligned} A_3 &= \frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \sum_{h_1=L+1}^{\infty} \delta_l^2 \lambda_{h_1}^2 \kappa_l(0,|N(i)-N(j)|) \\ &\quad + \frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \delta_l^2 \lambda_l \kappa_l(0,|N(i)-N(j)|,|N(i)+1-n|,|N(i)+1-n|) \\ &\quad + \frac{2}{k_N^2 N^2} \sum_{l=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \delta_l^2 \lambda_l \kappa_l(0,|N(i)+1-n|) \kappa_l(0,|N(j)+1-n|) \\ &\quad + \frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \delta_l^2 \lambda_l \kappa_l(0,|N(i)-N(j)|) \kappa_l(0,0) \\ &=: A_{3,1} + A_{3,2} + A_{3,3} + A_{3,4}. \end{aligned} \tag{4.34}$$

Note that the term A_3 enters the object of interest twice with a negative sign, such that all terms of which A_2 is comprised are canceled in view of $A_{2,1} = A_{3,1}$ and $A_{2,2} = A_{3,4}$ and since $\kappa_l(0,0) = \lambda_l$.

We now tun to term A_1 and first decompose into the cases where $h_1 \neq h_2$ and $h_1 = h_2$. The second case is furthermore decomposed into cases where $l = k$ and $l \neq k$. This yields

$$\begin{aligned}
A_1 &= \frac{1}{k_N^2 N^2} \sum_{l,k=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \sum_{h_1 \neq h_2}^{\infty} \delta_l \delta_k \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,k}\theta_{n,h_1}\theta_{n,h_2}\theta_{m,h_1}\theta_{m,h_2}\right] \\
&\quad + \frac{1}{k_N^2 N^2} \sum_{l \neq k}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \sum_{h_1=1}^{\infty} \delta_l \delta_k \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,k}\theta_{n,h_1}^2\theta_{m,h_1}^2\right] \\
&\quad + \frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \sum_{h_1=1}^{\infty} \delta_l^2 \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,k}\theta_{n,h_1}^2\theta_{m,h_1}^2\right] \\
&=: A_{1,1} + A_{1,2} + A_{1,3}. \qquad (4.35)
\end{aligned}$$

Now note that $A_{1,2} = 0$ by the same arguments as above. For term $A_{1,3}$, we decompose into the cases where $l \neq h_1$ and $l = h_1$ which yields

$$\begin{aligned}
A_{1,3} &= \frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \delta_l^2 \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\theta_{n,l}^2\theta_{m,l}^2\right] \\
&\quad + \frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \sum_{h_1=L+1}^{\infty} \delta_l^2 \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\right] \mathbb{E}\left[\theta_{n,h_1}^2\theta_{m,h_1}^2\right]. \qquad (4.36)
\end{aligned}$$

We consider first the first term of (4.36). By (4.8) and writing, with some abuse of notation, $\kappa^{(p)}$ for the p-th order cumulant, we have

$$\begin{aligned}
&\mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\theta_{n,l}^2\theta_{m,l}^2\right] \\
&= \kappa_l^{(6)} + 15\kappa_l^{(4)}\kappa_l^{(2)} + 10\kappa_l^{(3)}\kappa_l^{(3)} + 15\kappa_l^{(2)}\kappa_l^{(2)}\kappa_l^{(2)}.
\end{aligned}$$

There are 15 instances of $\kappa_l^{(2)}$ which are of the form

$$
\begin{aligned}
&1 \times \kappa_l(0,|N(i)-N(j)|)\\
&2 \times \kappa_l(0,|N(i)+1-n|)\\
&2 \times \kappa_l(0,|N(i)+1-m|)\\
&2 \times \kappa_l(|N(i)-N(j)|,|N(i)+1-n|)\\
&2 \times \kappa_l(|N(i)-N(j)|,|N(i)+1-m|)\\
&4 \times \kappa_l(|N(i)+1-n|,|N(i)+1-m|)\\
&1 \times \kappa_l(|N(i)+1-n|,|N(i)+1-n|)\\
&1 \times \kappa_l(|N(i)+1-m|,|N(i)+1-m|)
\end{aligned}
$$

Now note that there are precisely four instances where $\kappa_l^{(2)}$ is such that the first term in (4.36) takes the form

$$\frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \delta_l^2 \lambda_l \kappa_l(0,|N(i)+1-n|)\kappa_l(0,|N(j)+1-n|)$$

and precisely one instance where $\kappa_l^{(2)}$ is such that the first term in (4.36) takes the form

$$\frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \delta_l^2 \lambda_l^2 \kappa_l(0,|N(i)-N(j)|)$$

which are canceled by $A_{3,3}$ and $A_{3,4}$, respectively, since these terms enters twice with a negative sign. By similar arguments, we have two instances in which $\kappa_l^{(4)}$ is such that the first term in (4.36) takes the form

$$\frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum_{i,j=1}^{k_N} \sum_{n,m=1}^{N} \delta_l^2 \lambda_l \kappa_l(0,|N(i)-N(j)|,|N(i)+1-n|,|N(i)+1-n|)$$

which are canceled by $A_{3,2}$, again since that term enters twice with a negative sign. The remaining terms of the first term in (4.36) do not provide the dominant rate of convergence such that we skip the further analysis and consider next the second term in (4.36).

By (4.8) we have

$$\begin{aligned}&\mathbb{E}\left[\theta^2_{n,h_1}\theta^2_{m,h_1}\right]\\&= \kappa_{h_1}(0,0,|n-m|,|n-m|)+\kappa_{h_1}(0,0)\,\kappa_{h_1}(|n-m|,|n-m|)+2\kappa_{h_1}(0,|n-m|)\,\kappa_{h_1}(0,|n-m|)\end{aligned}$$

such that we obtain for the second term of (4.36)

$$\begin{aligned}&\frac{1}{k_N^2N^2}\sum_{l=1}^{L}\sum_{i,j=1}^{k_N}\sum_{n,m=1}^{N}\sum_{h_1=L+1}^{\infty}\delta_l^2\mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\right]\mathbb{E}\left[\theta^2_{n,h_1}\theta^2_{m,h_1}\right]\\&=\frac{1}{k_N^2N^2}\sum_{l=1}^{L}\sum_{i,j=1}^{k_N}\sum_{n,m=1}^{N}\sum_{h_1=L+1}^{\infty}\delta_l^2\lambda_{h_1}^2\mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\right]\\&\quad+\frac{1}{k_N^2N^2}\sum_{l=1}^{L}\sum_{i,j=1}^{k_N}\sum_{n,m=1}^{N}\sum_{h_1=L+1}^{\infty}\delta_l^2\mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\right]\kappa_{h_1}(0,0,|n-m|,|n-m|)\\&\quad+2\frac{1}{k_N^2N^2}\sum_{l=1}^{L}\sum_{i,j=1}^{k_N}\sum_{n,m=1}^{N}\sum_{h_1=L+1}^{\infty}\delta_l^2\mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\right]\kappa_{h_1}(0,|n-m|)^2.\end{aligned}$$

Observe now that the first term in the above display is canceled by $A_{3,1}$ as it enters twice with a negative sign. As a consequence, the terms A_2, A_3 and parts of A_1 cancel each other out. The dominant rate of convergence is now obtained by considering the third term in the above display for which we have

$$\begin{aligned}&\frac{2}{k_N^2N^2}\sum_{l=1}^{L}\sum_{i,j=1}^{k_N}\sum_{n,m=1}^{N}\sum_{h_1=L+1}^{\infty}\delta_l^2\mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\right]\kappa_{h_1}(0,|n-m|)^2\\&=2\left(\frac{1}{k_NN}\sum_{l=1}^{L}\delta_l^2\sum_{i,j=1}^{k_N}\mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\right]\right)\times\\&\qquad\left(\frac{1}{k_NN}\sum_{h_1=L+1}^{\infty}\sum_{n,m=1}^{N}\kappa_{h_1}(0,|n-m|)^2\right).\end{aligned}\tag{4.37}$$

For the first term in brackets in (4.37) we have, for some constant $C > 0$,

$$\begin{aligned}
(\ldots) &\leq \frac{1}{k_N N} \sum_{l=1}^{L} \delta_l^2 \sum_{i=1}^{k_N} \mathbb{E}\left[\theta_{N(i)+1,l}^2\right] + \frac{1}{k_N N} \sum_{l=1}^{L} \delta_l^2 \sum\sum_{i \neq j}^{k_N} \left|\mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\right]\right| \\
&\leq \frac{1}{k_N N} \sum_{l=1}^{L} \delta_l^2 \sum_{i=1}^{k_N} \lambda_l + \frac{2}{k_N N} \sum_{m=1}^{k_N-1} \sum_{i=m+1}^{k_N} \sum_{l=1}^{L} \delta_l^2 B_{m,l} \\
&\leq \frac{1}{N} \sum_{l=1}^{L} \delta_l^2 \lambda_l + \frac{C}{k_N N} \sum_{m=1}^{k_N-1} (k_N - m) m^{-\beta} \sum_{l=1}^{L} \delta_l^2 \lambda_l \\
&= \mathcal{O}\left(\frac{k_N^{1-\beta^*} L^{3+\alpha}}{N}\right),
\end{aligned}$$

where $\beta^* := \beta\, \mathbb{I}(0 < \beta < 1) + \mathbb{I}(\beta > 1)$ and the last equality follows in view of Assumption 4.1 $(\mathbf{ii})$ and $(\mathbf{iv})$ and Lemma 4.1. For the second term in brackets in (4.37) we have by similar arguments for some constants $C, C^* > 0$,

$$\begin{aligned}
(\ldots) &\leq \frac{1}{k_N N} \sum_{h_1=1}^{\infty} \sum\sum_{n,m=1}^{N} \mathbb{E}\left[\theta_{n,h_1}\theta_{m,h_1}\right]^2 \\
&\leq \frac{1}{k_N N} \sum_{h_1=1}^{\infty} \sum_{n=1}^{N} \mathbb{E}\left[\theta_{n,h_1}^2\right]^2 + \frac{1}{k_N N} \sum_{h_1=1}^{\infty} \sum\sum_{n \neq m}^{N} \left|\mathbb{E}\left[\theta_{n,h_1}\theta_{m,h_1}\right]\right|^2 \\
&\leq \frac{1}{k_N} \sum_{h_1=1}^{\infty} \lambda_{h_1}^2 + \frac{2}{k_N N} \sum_{m=1}^{N-1} \sum_{i=1}^{N} \sum_{h_1=1}^{\infty} B_{m,h_1}^2 \\
&\leq \frac{C}{k_N} + \frac{C^*}{k_N N} \sum_{m=1}^{N-1} \sum_{i=1}^{N} m^{-2\beta} \sum_{h_1=1}^{\infty} \lambda_{h_1}^2 \\
&= \mathcal{O}\left(\frac{N^{1-2\beta^{**}}}{k_N}\right)
\end{aligned}$$

where $\beta^{**} := \beta\, \mathbb{I}(0 < \beta < 1/2) + \frac{1}{2}\, \mathbb{I}(\beta > 1/2)$ and since $\sum_{h_1 \geq 1} \lambda_{h_1}^2$ is bounded in view of Lemma 4.2. Combining these results we obtain the desired rate of convergence

$$\mathcal{O}\left(\frac{L^{3+\alpha}}{k_N^{\beta^*} N^{2\beta^{**}}}\right).$$

Note that we omit the analysis of term $A_{1,1}$ for brevity as it follows by the same arguments presented above and yields the same dominant rate of convergence. □

Proof of statement (ii). We have in view of orthonormality of the $\widehat{\psi}_l$ and using the definitions of $\widehat{m}_{l,N}(\hat{\theta}_l)$ and $m_{l,N}(\theta_l)$

$$\begin{aligned}
&\mathbb{E}\left\|\sum_{l=1}^{L}\left(\widehat{m}_{l,N}(\hat{\theta}_l) - m_{l,N}(\theta_l)\right)\widehat{\psi}_l\right\|^2 \\
&= \sum_{l=1}^{L}\mathbb{E}\left[\left(\widehat{m}_{l,N}(\hat{\theta}_l) - m_{l,N}(\theta_l)\right)^2\right] \\
&= \sum_{l=1}^{L}\mathbb{E}\left[\left(\frac{1}{k_N}\sum_{i\in\widehat{\mathcal{I}}(k_N;\hat{\theta}_l)}\hat{\theta}_{i+1,l} - \frac{1}{k_N}\sum_{i\in\mathcal{I}(k_N;\theta_l)}\theta_{i+1,l}\right)^2\right]
\end{aligned} \tag{4.38}$$

As before, we write $R_{(k_N),l}(\theta_l) := |\theta_{N(k_N),l} - \theta_l|$ for the distance of the farthest of the k_N neighbors $\theta_{N(k_N),l}$ to some θ_l and similarly define $\widehat{R}_{(k_N),l}(\hat{\theta}_l) := |\hat{\theta}_{N(k_N),l} - \hat{\theta}_l|$ in terms of estimated quantities. Now define by $B_{\theta_l}(R_{(k_N),l}(\theta_l))$ and $B_{\hat{\theta}_l}(\widehat{R}_{(k_N),l}(\hat{\theta}_l))$ the balls with centers θ_l and $\hat{\theta}_l$ and radii $R_{(k_N),l}(\theta_l)$ and $\widehat{R}_{(k_N),l}(\hat{\theta}_l)$, respectively. In fact, although the balls just defined correspond to intervals on the real line, we adopt the notion of balls for this proof. It helps for the remainder of the analysis to express the summations in (4.38) in terms of indicator functions of events $\omega_{i,l} := \{\theta_{i,l} \in B_{\theta_l}(R_{(k_N),l}(\theta_l))\}$ and $\widehat{\omega}_{i,l} := \{\hat{\theta}_{i,l} \in B_{\hat{\theta}_l}(\widehat{R}_{(k_N),l}(\hat{\theta}_l))\}$, respectively. Moreover, we note that in view of the Cauchy-Schwarz inequality we have

$$\sup_{i\leq N}\left[|\hat{\theta}_{i,l} - \theta_{i,l}|^2\right] = \sup_{i\leq N}\left\langle X_i, \widehat{\psi}_l - \psi_l\right\rangle^2 \leq \sup_{i\leq N}\|X_i\|^2\left\|\widehat{\psi}_l - \psi_l\right\|^2 =: \Delta_l \tag{4.39}$$

and we shall furthermore assume that for some constant $C > 0$, $\sup_{i\leq N}\|X_i\|^2 \leq C\log(N)$ almost-surely. As a consequence we have $\sup_{i\leq N}|\hat{\theta}_{i,l} - \theta_{i,l}| \leq \Delta_l^{1/2}$ and

$$\left|\widehat{R}_{(k_N),l}(\hat{\theta}_l) - R_{(k_N),l}(\hat{\theta}_l)\right| \leq \left|\hat{\theta}_{N(k_N),l} - \theta_{N(k_N),l}\right| \leq \Delta_l^{1/2} \tag{4.40}$$

in view of the reverse triangle inequality.

We can then write for (4.38)

$$
\begin{aligned}
&\sum_{l=1}^{L} \mathbb{E}\left[\left(\frac{1}{k_N}\sum_{i\in\widehat{\mathcal{I}}(k_N;\hat{\theta}_l)}\hat{\theta}_{i+1,l} - \frac{1}{k_N}\sum_{i\in\mathcal{I}(k_N;\theta_l)}\theta_{i+1,l}\right)^2\right] \\
&= \sum_{l=1}^{L} \mathbb{E}\left[\left(\frac{1}{k_N}\sum_{i=1}^{N-1}\hat{\theta}_{i+1,l}\mathbb{I}(\widehat{\omega}_{i,l}) - \frac{1}{k_N}\sum_{i=1}^{N-1}\theta_{i+1,l}\mathbb{I}(\omega_{i,l})\right)^2\right] \\
&= \frac{1}{k_N^2}\sum_{l=1}^{L} \mathbb{E}\left[\left(\sum_{i=1}^{N-1}\hat{\theta}_{i+1,l}\mathbb{I}(\widehat{\omega}_{i,l}) - \theta_{i+1,l}\mathbb{I}(\omega_{i,l})\right)^2\right] \\
&= \frac{1}{k_N^2}\sum_{l=1}^{L}\sum_{i=1}^{N-1} \mathbb{E}\left[\left(\hat{\theta}_{i+1,l}\mathbb{I}(\widehat{\omega}_{i,l}) - \theta_{i+1,l}\mathbb{I}(\omega_{i,l})\right)^2\right] \\
&\quad + \frac{1}{k_N^2}\sum_{l=1}^{L}\sum_{i\neq j}^{N-1} \mathbb{E}\left[\left(\hat{\theta}_{i+1,l}\mathbb{I}(\widehat{\omega}_{i,l}) - \theta_{i+1,l}\mathbb{I}(\omega_{i,l})\right)\left(\hat{\theta}_{j+1,l}\mathbb{I}(\widehat{\omega}_{j,l}) - \theta_{j+1,l}\mathbb{I}(\omega_{j,l})\right)\right] \\
&=: B + C. \qquad (4.41)
\end{aligned}
$$

We are going to analyze the terms B and C separately and start with the former. Adding and subtracting the quantity $\theta_{i+1,l}\mathbb{I}(\widehat{\omega}_{i,l})$ to the argument in B and expanding the second degree polynomial gives

$$
\begin{aligned}
&\frac{1}{k_N^2}\sum_{l=1}^{L}\sum_{i=1}^{N-1} \mathbb{E}\left[\left((\hat{\theta}_{i+1,l} - \theta_{i+1,l})\mathbb{I}(\widehat{\omega}_{i,l}) + \theta_{i+1,l}(\mathbb{I}(\widehat{\omega}_{i,l}) - \mathbb{I}(\omega_{i,l}))\right)^2\right] \\
&= \frac{1}{k_N^2}\sum_{l=1}^{L}\sum_{i=1}^{N-1} \mathbb{E}\left[(\hat{\theta}_{i+1,l} - \theta_{i+1,l})^2\mathbb{I}(\widehat{\omega}_{i,l})^2\right] \\
&\quad + \frac{1}{k_N^2}\sum_{l=1}^{L}\sum_{i=1}^{N-1} \mathbb{E}\left[\theta_{i+1,l}^2(\mathbb{I}(\widehat{\omega}_{i,l}) - \mathbb{I}(\omega_{i,l}))^2\right] \\
&\quad + \frac{2}{k_N^2}\sum_{l=1}^{L}\sum_{i=1}^{N-1} \mathbb{E}\left[(\hat{\theta}_{i+1,l} - \theta_{i+1,l})\mathbb{I}(\widehat{\omega}_{i,l})\theta_{i+1,l}(\mathbb{I}(\widehat{\omega}_{i,l}) - \mathbb{I}(\omega_{i,l}))\right] \\
&=: B_1 + B_2 + 2B_3. \qquad (4.42)
\end{aligned}
$$

We start with B_1 and first observe that we always have

$$\mathbb{E}\left[(\hat{\theta}_{i+1,l}-\theta_{i+1,l})^2\mathbb{I}(\widehat{\omega}_{i,l})^2\right] \leq \mathbb{E}\left[(\hat{\theta}_{i+1,l}-\theta_{i+1,l})^2\right] \leq \mathbb{E}\left[\Delta_l\right],$$

and that second, there are precisely k_N instances in which $\mathbb{I}(\widehat{\omega}_{i,l}) = 1$. As a consequence, we have in view of Corollary 4.2 and for some constant $C > 0$,

$$\begin{aligned}
B_1 &\leq \frac{1}{k_N^2}\sum_{l=1}^{L} k_N \mathbb{E}\left[\Delta_l\right] \\
&= \frac{1}{k_N}\sum_{l=1}^{L} \mathbb{E}\left[\sup_{i\leq N}\|X_i\|^2\|\widehat{\psi}_l-\psi_l\|^2\right] \\
&\leq \frac{C}{k_N}\sum_{l=1}^{L}\log(N)\delta_l^2 N^{-2\beta^{**}} \\
&= \mathcal{O}\left(\frac{L^{3+2\alpha}\log(N)}{k_N N^{2\beta^{**}}}\right).
\end{aligned}$$

We proceed with B_2 and observe that by the Cauchy-Schwarz inequality we have

$$\mathbb{E}\left[\theta_{i+1,l}^2(\mathbb{I}(\widehat{\omega}_{i,l})-\mathbb{I}(\omega_{i,l}))^2\right] \leq \left(\mathbb{E}\left[\theta_{i+1,l}^4\right]\right)^{\frac{1}{2}}\left(\mathbb{E}\left[(\mathbb{I}(\widehat{\omega}_{i,l})-\mathbb{I}(\omega_{i,l}))^4\right]\right)^{\frac{1}{2}}.$$

Now consider the first term in brackets for which we have, again using the relationship between higher-order moments of $\theta_{i,l}$ and associated cumulants (4.8), and for some constant $C > 0$,

$$\begin{aligned}
\left(\mathbb{E}\left[\theta_{i+1,l}^4\right]\right)^{\frac{1}{2}} &\leq \left(\sum_{\tau_1,\tau_2,\tau_3=-\infty}^{\infty}\sum\kappa_{l(0,\tau_1,\tau_2,\tau_3)}+3\mathbb{E}\left[\theta_{i+1,l}^2\right]\mathbb{E}\left[\theta_{i+1,l}^2\right]\right)^{\frac{1}{2}} \\
&\leq C\sqrt{\lambda_l^2} = C\lambda_l.
\end{aligned}$$

For the second term in brackets we first observe that there are at most $2k_N$ instances in which $(\mathbb{I}(\widehat{\omega}_{i,l})-\mathbb{I}(\omega_{i,l}))^2 = 1$. Now denote by $\triangle$ the symmetric set difference operator, and observe that we have

$$\mathbb{E}\left[(\mathbb{I}(\widehat{\omega}_{i,l})-\mathbb{I}(\omega_{i,l}))^4\right] = \mathbb{E}\left[\mathbb{I}(\widehat{\omega}_{i,l}\triangle\omega_{i,l})\right] \leq \mathbb{E}\left[\mathbb{I}(\widehat{\omega}_{i,l}\cup\omega_{i,l})\right].$$

Recall that the balls defined above correspond to intervals of the real line such that the event $\widehat{\omega}_{i,l} \cup \omega_{i,l}$ is associated with an interval with length bounded by $\mathcal{R} := R_{(k_N),l}(\hat{\theta}_l) + 2\Delta_l^{1/2} + R_{(k_N),l}(\theta_l)$. As $\mathcal{R}$ is random, we make use of the law of total expectations and by arguments similar to those presented in Rakotomarolahy [90] we have, for some constant $C > 0$,

$$\begin{aligned}(\mathbb{E}\left[\mathbb{I}(\widehat{\omega}_{i,l} \cup \omega_{i,l})\right])^{\frac{1}{2}} &= (\mathbb{E}\left[\mathbb{E}\left[\mathbb{I}(\widehat{\omega}_{i,l} \cup \omega_{i,l})|\mathcal{R}\right]\right])^{\frac{1}{2}} \\ &\leq C\left(\mathbb{E}\left[R_{(k_N),l}(\hat{\theta}_l) + 2\Delta_l^{1/2} + R_{(k_N),l}(\theta_l)\right]\right)^{\frac{1}{2}}.\end{aligned}$$

Now observe that in view of the proof of Theorem 4.3 (i), we have upon using Jensen's inequality that for some constant $C > 0$,

$$\begin{aligned}\mathbb{E}\left[R_{(k_N),l}(\hat{\theta}_l)\right] &\leq Ck_N^{-1/4}\lambda_l^{1/2}, \\ \mathbb{E}\left[R_{(k_N),l}(\theta_l)\right] &\leq Ck_N^{-1/4}\lambda_l^{1/2},\end{aligned}$$

such that it is through $\Delta_l^{1/2}$ that the dominant rate of convergence is achieved. Combining the above results then yields, for some constant $C > 0$,

$$\begin{aligned}B_2 &\leq \frac{C}{k_N^2}\sum_{l=1}^{L} k_N\lambda_l(\log(N))^{\frac{1}{4}}\delta_l^{1/2}N^{-\frac{1}{2}\beta^{**}} \\ &= \mathcal{O}\left(\frac{L^{\frac{3}{2}-\frac{\alpha}{2}}(\log(N))^{\frac{1}{4}}}{k_N N^{\frac{1}{2}\beta^{**}}}\right).\end{aligned}$$

We omit the analysis of B_3 for brevity as, in view of the Cauchy-Schwarz inequality, the dominant rate of convergence is determined through the consideration of B_1 and B_2.

Let us now turn to term C in (4.41) and observe that, again in view of the Cauchy-Schwarz inequality, we have to consider essentially the same terms as in B_1 and B_2, the difference being that an additional sum over the time index j enters the expressions. As a consequence, to obtain rates of convergence for the corresponding terms C_1 and C_2 it suffices to multiply the respective rates for B_1 and B_2 by an additional factor of k_N. This gives the following rates of convergence,

$$C_1 = \mathcal{O}\left(\frac{L^{3+2\alpha}\log(N)}{N^{2\beta^{**}}}\right),$$
$$C_2 = \mathcal{O}\left(\frac{L^{\frac{3}{2}-\frac{\alpha}{2}}(\log(N))^{\frac{1}{4}}}{N^{\frac{1}{2}\beta^{**}}}\right).$$

which establishes the desired result. By the same arguments as for B_3 we omit the analysis of C_3 for brevity as, in view of the Cauchy-Schwarz inequality, the dominant rate of convergence is determined through the consideration of C_1 and C_2.

□

4.6 Conclusion

In this paper we were concerned with the statistical analysis of functional time series, particularly with regards to the problem of prediction. Within the framework of first-order auto-regression, we proposed a functional additive model that extends the current literature to the functional time series scenario. The proposed modeling framework provided several advantages. First, it allowed us to introduce a general notion of time dependencies for functional data that is rooted at the correlation structure of the functional principal components scores and borrows its intuition from classical time series analysis. Second, it allowed us to consider a very intuitive and easy to implement predictor of some future function that is based on a k-nearest neighbors classification scheme.

The theoretical contributions in this paper were two-fold. In a first step, we verified the applicability of the functional principal components analysis and obtained precise rates of convergence for the mean function and the covariance operator associated with the observed sample of functions. In a second step, we derived precise rates of convergence of the mean squared error for the proposed predictor, taking into account both the effect of truncating the infinte series expansion at some finite integer L as well as the effect of estimating the covariance operator and associated eigenelements based on a sample of N curves.

5

Forecasting ground-level ozone concentration surfaces: a functional perspective

5.1 Introduction

In this paper we are concerned with the prediction of ground-level ozone concentration *surfaces* over the geographical area of Germany from a functional perspective. Ground-level ozone is a harmful pollutant and the importance of obtaining reliable forecasts has given rise to a large body of literature. The challenge that this problem poses lie in the fact that spatial and temporal information interact and have to be taken into account when developing a forecasting method.

In classical spatial statistics, space-time models have been very popular where the focus of analysis is in modeling the space-time covariance structure of the underlying (finite-dimensional) *spatiotemporal process*

$$\left\{\mathcal{Y}_{s,t} : \boldsymbol{s} \in \mathbb{R}^d, t \in \mathbb{R}\right\} \tag{5.1}$$

(see Cressie and Huang [33] and Gneiting [58] and the references therein). Applications to the spatiotemporal modeling of ozone concentration can be found in Guttorp et al. [65], Huang and Hsu [74] and Bruno et al. [20]. Typically, the process $\mathcal{Y}$ is sampled at spatial locations $\mathbf{s} = \mathbf{s}_1, \ldots, \mathbf{s}_N$ and discrete times $t = 1, \ldots, T$. The question of interest is then in predicting the value of $\mathcal{Y}$ at some unmonitored site $\mathbf{s}^*$ and at a specific point in time t^*.

This problem has also been considered recently from a functional perspective. Analogously to above, a *spatial functional process* is defined as

$$\{\mathcal{Y}_s(t) : s \in \mathcal{D},\ t \in \mathcal{T}\} \tag{5.2}$$

where $\mathcal{D} \subset \mathbb{R}^d$ is a generic set of *spatial locations* and $\mathcal{T} \subset \mathbb{R}$ is the time-horizon of observation (see Delicado et al. [37]). For each fixed spatial location s, $\mathcal{Y}_s(t)$ is thought to be a random function taking values in some suitable function space, such as L^2, i.e. the space of square integrable functions defined on $\mathcal{T}$. Assume one observes a family of N functions $(\mathcal{Y}_{s_i})_{i=1}^N$ at locations $s_1, \ldots, s_N$. As opposed to standard Functional Data Analysis (see Ramsay and Silverman [92] for more information), the curves $\mathcal{Y}_{s_1}, \ldots, \mathcal{Y}_{s_N}$ are thought to exhibit *spatial dependencies*. The interest is thus to take this dependence structure into account when *predicting* a function $\mathcal{Y}_{s^*}(\cdot)$ at some unmonitored site s^*. Several approaches have been suggested in the recent literature that are mainly based on the framework of functional linear regression models. For example, Giraldo et al. [55], Giraldo et al. [56] and Nerini et al. [85] extend the methodology of *kriging* from classical geostatistics to the functional setting while Yamanishi and Tanaka [109] consider additional functional covariates.

In this paper we take an alternative *functional* perspective in that we consider the ozone concentration over some spatial domain $\mathcal{D}$ for some fixed time point t to be given by a smooth surface that belongs to a suitably defined function space, such that we consider a *time series of spatial surfaces* defined by

$$\{\mathcal{X}_t(s) : s \in \mathcal{D},\ t \in \mathcal{T}\}. \tag{5.3}$$

In particular we consider the space of square-integrable functions defined over the spatial domain $\mathcal{D}$ and equipped with the norm $\|\cdot\|$, i.e. $\mathcal{X}_t(s) \in L^2(\mathcal{D}, \|\cdot\|)$.

While the processes defined in (5.2) and (5.3) appear to be related, their interpretation, and the initial prediction problem for which they have been formulated, differs. Whereas the prediction problem in (5.2) is *spatial* in nature, we are interested in a *dy-*

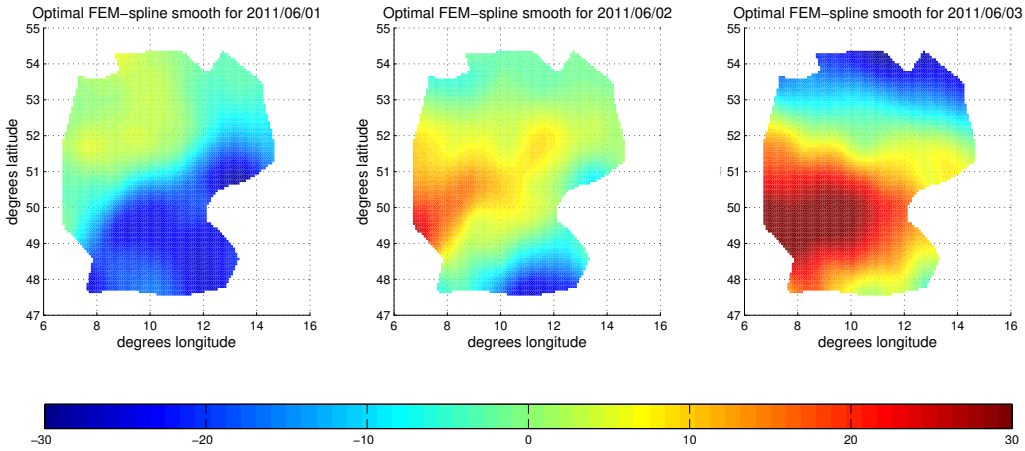


Figure 5.1: **Smoothed ozone concentration surfaces.** This figure shows ozone concentration surfaces over Germany for the first three days in June 2011 that are obtained through a finite element smoothing approach of discrete measurements.

namic perspective. As such, the prediction problem that we have in mind is more related to the classical time series scenario, in that we wish to obtain forecasts of some future surface $\mathcal{X}_{T+1}(\cdot)$, say, based on a *functional time series* $(\mathcal{X}_t(\boldsymbol{s}))_{t=1}^T$, $\boldsymbol{s} \in \mathcal{D}$. In Figure 5.1 we plot the (smoothed) surface of ozone concentration across Germany for three consecutive days from 2011/06/01 until 2011/06/03.

The data that motivated this research consists of daily measurements of ozone concentration made available through AirBase - the European air quality database provided by the European Environment Agency[1]. For the case of Germany, daily measurements of ozone concentration are available at $N = 171$ stations for the year 2011 (see the left panel of Figure 5.2).

Given the nature of the data, we are in a first step concerned with obtaining a spatial smooth over the spatial domain $\mathcal{D}$ of discrete measurements for each fixed time point $t = 1, \ldots, T$. As many other environmental data, ozone concentration displays strong seasonality over the yearly horizon. From a statistical perspective, it is desirable to work with a stationary sequence of surfaces and we thus filter the raw data by means of a Hodrick-Prescott filter (see Hodrick and Prescott [70]). This yields a decomposition of the original time series for each station into a seasonal and a residual component. From here on, the term "ozone concentration" is understood to mean the residual component

[1]http://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-7tab-data-by-country

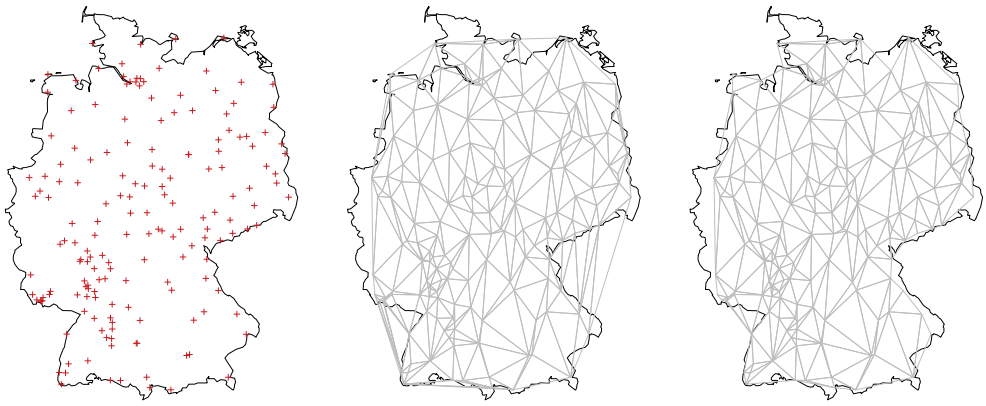


Figure 5.2: **Spatial locations of measurement stations and triangular mesh.** The left panel shows the spatial locations of ozone measurement stations together with the border of Germany. The middle panel shows the Delaunay triangulation where the nodes are placed at ozone sample stations. The right panel shows the corrected Delaunay triangular mesh where triangles that cover the exterior of Germany have been removed.

of some suitable filter applied to the original data. As we are working with spatial data, we are furthermore interested in taking the geographical boundaries of Germany into account and thus opt for a finite element smoothing approach as suggested in Ramsay [93] and Sangalli et al. [98] where they consider finite element basis functions that are locally defined over a triangulation of the spatial domain (see the right panel of Figure 5.2).

As in the case of spatial functional processes, prediction in a dynamic functional setting has also been mainly considered in the context of functional linear regression models of first-order auto-regressive type. Ettinger et al. [45] consider the ozone concentration $\mathcal{Y}_{\boldsymbol{s}^*,t}$ at some particular site $\boldsymbol{s}^* \in \mathcal{D}$ and time point t as the scalar response to a functional linear model where the functional covariate $\mathcal{X}_{t-1}$ is given by the surface over the entire domain $\mathcal{D}$ at the *previous* time point. Tacitly assuming that all stochastic quantities have mean zero, the model then takes the form

$$\mathbb{E}\left[\mathcal{Y}_{\boldsymbol{s}^*,t} | \mathcal{X}_{t-1}\right] = \int_{\mathcal{D}} \gamma(\boldsymbol{s}) \mathcal{X}_{t-1}(\boldsymbol{s}) d\boldsymbol{s}, \quad t = 1, \ldots, T, \tag{5.4}$$

which can be seen as a special case of functional first-order auto-regressive models of the

form

$$\mathbb{E}\left[\mathcal{X}_t(\boldsymbol{s})|\mathcal{X}_{t-1} = x\right]. \tag{5.5}$$

Without imposing additional structure on the modeling framework, estimation of such functional regression models is subject to the curse of dimensionality (as, generally speaking, functions are infinite-dimensional objects). Thus a prominent suggestion in the literature (see Bosq [18]) is to restrict the attention to functional *linear* models of first-order auto-regressive type, FAR(1) for short, of the form

$$\mathbb{E}\left[\mathcal{X}_t|\mathcal{X}_{t-1} = x\right] = \Gamma\left[x\right], \quad t = 1, \ldots, T, \tag{5.6}$$

where Γ is now a linear operator mapping an element of L^2 to L^2 that admits an integral representation of the form

$$\Gamma\left[\mathcal{X}_t\right](\boldsymbol{s}) = \int_{\mathcal{D}} \gamma(\boldsymbol{s}, \boldsymbol{u})\mathcal{X}_t(\boldsymbol{u})d\boldsymbol{u}. \tag{5.7}$$

Once the operator Γ is estimated by some empirical estimator Γ_T based on a sample of functions $(\mathcal{X}_t)_{t=1}^T$, this model lends itself readily for predicting a future value $\mathcal{X}_{T+1}$ given that $\mathcal{X}_T = x$ through

$$\mathcal{X}_{T+1}^f = \Gamma_T\left[x\right].$$

Application of the FAR(1) model can be found in Besse et al. [14] where it performs favorably in a forecasting study of functional climatic variations.

Whereas the FAR(1) model assumes a *linear* first-order auto-regressive structure, an essentially nonparametric extension was recently suggested by Gleim and Salish [57]. It takes the form

$$\mathbb{E}\left[\mathcal{X}_t|\mathcal{X}_{t-1} = x\right] = M(x), \tag{5.8}$$

and the authors suggest modeling $M(x)$ by a functional *additive* model (FAM) as proposed by Müller and Yao [84] and where predictions are formed by means of a k-nearest neighbors classification scheme (FAM-knn). Such a formulation considerably broadens the scope of functional regression models and we compare its performance relative to functional linear models in a forecasting study of ground-level ozone concentration

surfaces over Germany. Similarly to above, the prediction of some future value of the surface $\mathcal{X}_{T+1}$ given that $\mathcal{X}_T = x$ is then given by

$$\mathcal{X}_{T+1}^f = M_T(x),$$

where M_T denotes an empirical estimator of M.

The remainder of the paper is organized as follows. Section 5.2 considers how smooth surfaces can be obtained from noisy discrete spatial measurements by using a finite element spline smoother. The key tool to analyze and estimate both the FAR(1) and FAM-knn models is the functional principal components analysis. This considers the spectral decomposition of the surfaces $\mathcal{X}_t$ in terms of eigenfunctions of the associated covariance operator and regularization in these models is achieved through projection on a finite number of functional principal components. Section 5.3 briefly discusses the so-called Karhunen-Loève decomposition of L^2-functions and details the estimation steps for both FAR and FAM-knn models. Both models are then compared in a forecasting study of ground-level ozone concentration surfaces over the geographical domain of Germany and the results are reported in section 5.4. Section 5.5 concludes.

5.2 Spatial smoothing with finite element splines

In this section we are concerned with the fact that although we think of the $(\mathcal{X}_t)_{t=1}^T$ as a time series of surfaces over some spatial domain $\mathcal{D}$ we only observe $X_t^*(\boldsymbol{s}_i)$ at discrete locations $\boldsymbol{s}_i, i = 1, \ldots, N$ where $X_t^* = \mathcal{X}_t + \eta_t$ is a noisy observation of $\mathcal{X}_t$ with η_t denoting some mean zero error term with finite variance. As a consequence, some statistical smoothing procedure is required to approximate $\mathcal{X}_t$ from the noisy observations $X_t^*(\boldsymbol{s}_i)$. Any smoothing method is employing some penalization term that governs the roughness (or smoothness) of the obtained approximation. A very common approach is to measure roughness through a (suitably defined) notion of squared second derivatives. We thus restrict our attention to L^2-functions that have square-integrable derivatives up to second order and denote the corresponding function space by H^2. In this paper we are going to employ a finite element spline smoother and we denote the thus approximated surface by $\widetilde{X}_t$. We give a detailed account of this approach following the exposition in Ramsay [93] and Sangalli et al. [98].

5.2.1 The minimization problem

In order to develop a finite element basis representation $\widetilde{X}$ of some function $\mathcal{X}$ we consider minimizing the penalized sum of squared residual functional J given by

$$J_\lambda(\mathcal{X}) = \sum_{i=1}^{N} \left(X^*(\boldsymbol{s}_i) - \mathcal{X}(\boldsymbol{s}_i)\right)^2 + \lambda \int_{\mathcal{D}} (\Delta\mathcal{X})^2 \;\rightarrow\; \min! \tag{5.9}$$

where $\lambda > 0$ is a smoothing parameter and Δ denotes a differential operator of second order. Note that we consider a generic $\mathcal{X} \in H^2$ and omit the subscript time index t to lighten the notational load. Moreover, we require that the smoothing problem be independent of the underlying spatial coordinate system so that the roughness penalty should be invariant under translation and rotation of the spatial coordinates. This is ensured if Δ is comprised of polynomials of the Laplacian operator. Since we are working in the Hilbert space H^2 we define Δ, for any $\mathcal{X} \in H^2$, to be given by

$$\Delta\mathcal{X}(\boldsymbol{s}) = \Delta\mathcal{X}(x,y) := \frac{\partial^2\mathcal{X}}{\partial x^2}(x,y) + \frac{\partial^2\mathcal{X}}{\partial y^2}(x,y),$$

where we note that the spatial locations $\boldsymbol{s}$ can be represented by pairs (x,y) of longitude and latitude coordinates.

Sangalli et al. [98] show that a unique solution X to the minimization problem given in (5.9) exists if one imposes a boundary condition on X, such as $X \in H^2_{n0}(\mathcal{D})$ where $H^2_{n0}(\mathcal{D})$ denotes the space of L^2-functions that have square-integrable partial derivatives up to second order and assume zero normal derivatives at the boundary. Let us denote for any function $X \in H^2(\mathcal{D})$ by $\boldsymbol{X}_N := (X(\boldsymbol{s}_1), X(\boldsymbol{s}_2), \ldots, X(\boldsymbol{s}_N))^\mathsf{T}$ the N-vector of evaluations of the function X at the N spatial sampling locations. The solution X is then characterized by

$$\boldsymbol{U}_N^\mathsf{T}\boldsymbol{X}_N + \lambda \int_{\mathcal{D}} (\Delta U)\,(\Delta X) = \boldsymbol{U}_N^\mathsf{T}\boldsymbol{X}_N^* \tag{5.10}$$

for every $U \in H^2_{n0}(\mathcal{D})$.

While the variational problem in (5.9) searches for a solution in the infinite dimensional space H^2, the finite element approach approximates this solution in a finite-dimensional subspace of the larger space H^1. Consequently, the characterization of the solution given in (5.10) has to be reformulated such that it is well defined in the space H^1 while still being in H^2. Sangalli et al. [98] show that (5.10) is equivalent to finding a pair of

functions $(X, Z) \in H^2(\mathcal{D}) \times H^2(\mathcal{D})$ such that

$$\boldsymbol{u}_N^\top \boldsymbol{X}_N - \lambda \int_{\mathcal{D}} (\nabla U)(\nabla Z) = \boldsymbol{u}_N^\top \boldsymbol{X}_N^* \tag{5.11}$$

$$\int_{\mathcal{D}} ZV + \int_{\mathcal{D}} (\nabla X)(\nabla V) = 0 \tag{5.12}$$

for every $(U, V) \in H^2(\mathcal{D}) \times H^2(\mathcal{D})$. Here we denote by ∇ the spatial gradient operator. The importance of this representation stems from the fact that all quantities involved are well-defined in the space $H^1(\mathcal{D})$, all the while the solution X being still in $H^2(\mathcal{D})$.

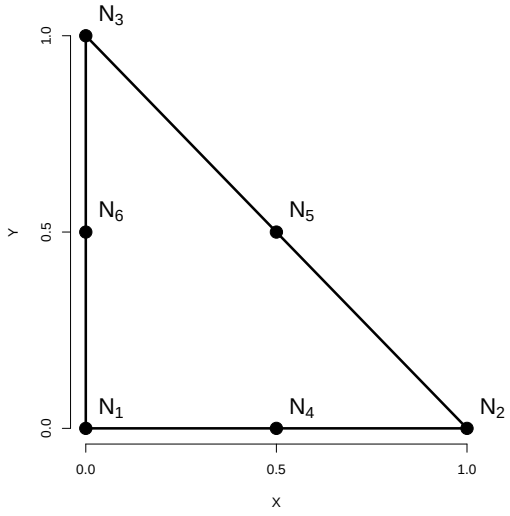
Given this reformulation, the finite element approach proceeds by approximating the solution X in the finite-dimensional space $H^1(\triangle_{\mathcal{D}})$ which is comprised of polynomials defined piecewise over triangles that make up a triangulation $\triangle_{\mathcal{D}}$ of the domain $\mathcal{D}$. In the remainder of this section we detail the construction of the finite element space $H^1(\triangle_{\mathcal{D}})$ and show how to obtain the finite element approximation $\widetilde{X}$ to X.

5.2.2 The finite element space $H^1(\triangle_{\mathcal{D}})$

The finite-dimensional subspace $H^1(\triangle_{\mathcal{D}})$ is constructed by partitioning the spatial domain $\mathcal{D}$ into disjoint sets (i.e. finite elements). We take each of the finite elements to be given by a triangle where the spatial locations $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_N$ correspond to vertices of the resulting triangular mesh. How to choose the set of triangles in practice is difficult and many possibilities have been offered in the literature. For our purposes, the Delaunay triangular mesh seems the most plausible as it chooses the triangulation such that it maximizes the minimum angle over all possible triangulations. As a result, the Delaunay triangulation avoids thin triangles and favors triangles that are as equiangular as possible. As long as no four or more nodes lie on a common circle, the Delaunay triangulation is uniquely defined and implementation routines are readily available for most statistical software. The middle panel of Figure 5.2 shows the Delaunay triangular mesh as the convex hull of the spatial location of ozone measurement stations. However, some triangles also cover the exterior of Germany and have been removed. The resulting triangular mesh we consider in the remainder of the paper is shown in the right panel of Figure 5.2 and we denote the thus approximated spatial domain by $\triangle_{\mathcal{D}}$.

The surface we wish to construct over $\triangle_{\mathcal{D}}$ is assumed to be polynomial of second

order in the spatial coordinates $s = (x, y)$ over any triangle while being continuous over edges and vertices. Note that in order to construct a quadratic polynomial over a triangle, the function value has to be specified at six nodal points which we take to be the vertices and the midpoints of each edge of a triangular finite element as indicated in Figure 5.3 for the right unit triangle. With each of the local nodal points we associate a *shape function* which is a second order polynomial in the spatial coordinates $s = (x, y)$ that takes the value one at one local nodal point and the value zero at all other local nodal points. The six shape functions that are constructed in such a way are plotted in Figure



$$N_1(x, y) = (1 - x - y)(1 - 2x - 2y)$$
$$N_2(x, y) = x(2x - 1)$$
$$N_3(x, y) = y(2y - 1)$$
$$N_4(x, y) = 4x(1 - x - y)$$
$$N_5(x, y) = 4xy$$
$$N_6(x, y) = 4y(1 - x - y)$$

Figure 5.3: **Reference triangular finite element.** This figure shows a reference triangle with corresponding nodal shape functions of a triangular finite reference element.

5.4.

Let us denote the nodal points of the triangulation (i.e. vertices and midpoints of edges) by $\xi_k, k = 1, \ldots, K$, and for ease of notation we number the nodal points in such as way as to have the spatial locations $s_i, i = 1, \ldots, N$, correspond to the first N nodal points. We associate with each node ξ_k, $k = 1, \ldots, K$, a *nodal basis function* ϕ_k that corresponds to the shape function associated with this node when restricted to a triangle which has the k-th node as a vertex. Nodal basis functions are thus implicitly defined by combining the shape functions of those triangles that share a certain node. In Figure 5.5 we plot the resulting finite element nodal basis function associated with the nodal point $(0, 0)$ which is shared by the four unit triangles. This set of K basis functions spans a

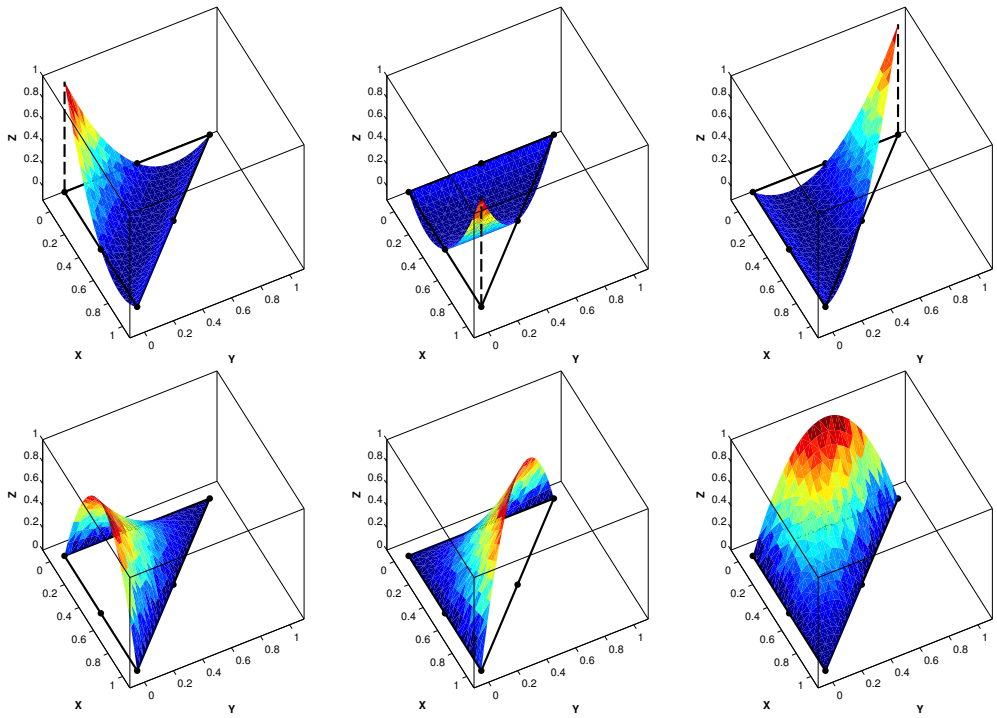


Figure 5.4: **Quadratic finite element shape functions.** This figure shows the six shape functions of a triangular finite reference element.

function space that we denote by $H^1(\triangle_{\mathcal{D}})$, i.e. the space of continuous functions on $\mathcal{D}$ that are piecewise quadratic polynomials when restricted to some triangular finite element.

5.2.3 Spatial finite element approximation

In this section we give a computable representation of the finite element spline approximation $\widetilde{X}$ to the solution X in the finite-dimensional space $H^1(\triangle_{\mathcal{D}})$. Particularly, we show that this estimation problem is equivalent to solving a linear system of equations.

To this end we denote by $\boldsymbol{\phi}_K := (\phi_1, \phi_2, \ldots, \phi_K)^\mathsf{T}$ the K-vector of spatial basis functions ϕ_k. Moreover, for any function $X \in H^1(\mathcal{D})$ we denote by $\boldsymbol{X}_K := (X(\boldsymbol{\xi}_1), X(\boldsymbol{\xi}_2), \ldots, X(\boldsymbol{\xi}_K))^\mathsf{T}$ the K-vector of evaluations of X at the K nodal points $\boldsymbol{\xi}_k$. As the nodal points $\boldsymbol{\xi}_k$ are numbered such that the first N nodes correspond to the spatial locations of measurement stations, we denote by

$$\boldsymbol{X}_K^* := (X^*(\boldsymbol{s}_1), X^*(\boldsymbol{s}_2), \ldots, X^*(\boldsymbol{s}_N), 0, \ldots, 0)^\mathsf{T}$$

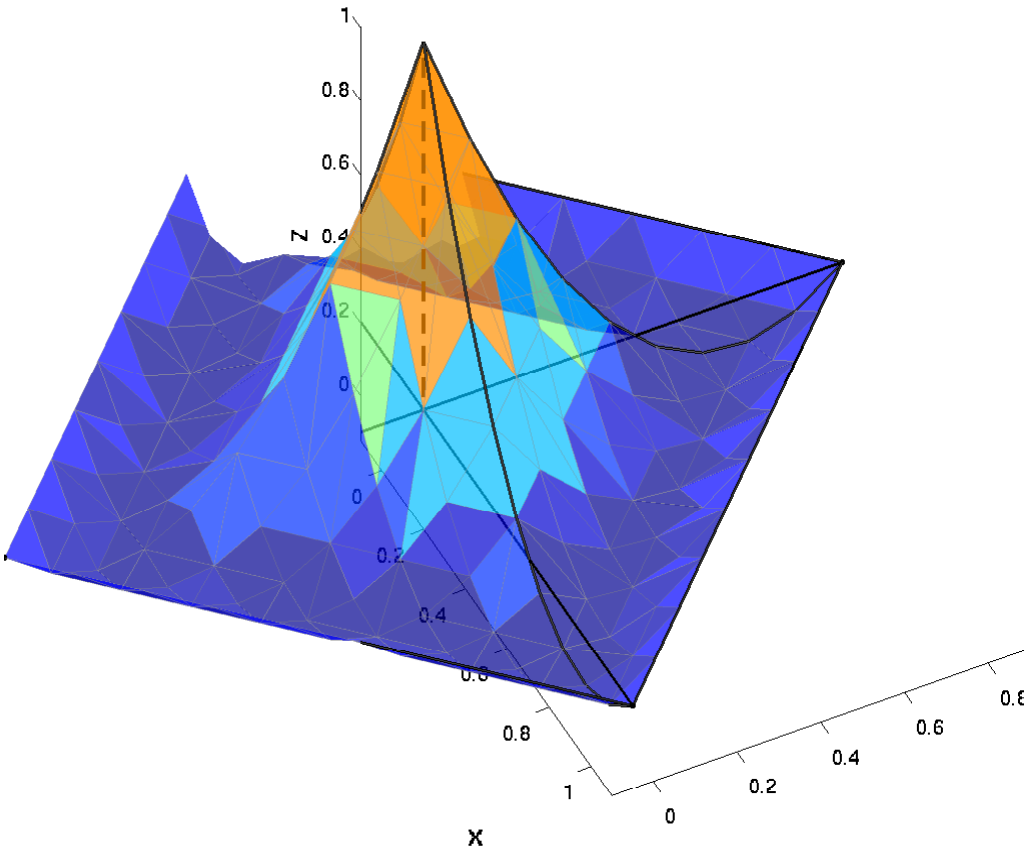


Figure 5.5: **Quadratic finite element basis function.** This figure shows the pyramid plot of a quadratic finite element basis function associated with the nodal point $(0,0)$.

the K-vector which has on the first N entries the observations $\boldsymbol{X}_N^*$ at the N spatial locations and zeros otherwise. By construction of the finite element space we can write any approximation $\widetilde{X} \in H^1(\triangle_{\mathcal{D}})$ to the solution X as

$$\widetilde{X}(\boldsymbol{s}) = \sum_{k=1}^{K} c_k \phi_k(\boldsymbol{s}) = \sum_{k=1}^{K} \widetilde{X}(\boldsymbol{\xi}_k)\phi_k(\boldsymbol{s}) = \left(\widetilde{\boldsymbol{X}}_K\right)^{\mathsf{T}} \boldsymbol{\phi}_K(\boldsymbol{s}). \tag{5.13}$$

Let us furthermore denote the K-vectors of partial derivatives of the nodal basis functions with respect to spatial x and y coordinates as $\boldsymbol{\phi}_K^{(x)} := (\partial\phi_1/\partial x, \ldots, \partial\phi_K/\partial x)^{\mathsf{T}}$

and $\boldsymbol{\phi}_K^{(y)} := (\partial\phi_1/\partial y, \ldots, \partial\phi_K/\partial y)^\mathsf{T}$ and define the order K matrices

$$\begin{aligned}\boldsymbol{A}_{K,K} &:= \int_{\triangle_\mathcal{D}} (\boldsymbol{\phi}_K)\,(\boldsymbol{\phi}_K)^\mathsf{T} \\ \boldsymbol{B}_{K,K} &:= \int_{\triangle_\mathcal{D}} \left(\boldsymbol{\phi}_K^{(x)}\right)\left(\boldsymbol{\phi}_K^{(x)}\right)^\mathsf{T} + \left(\boldsymbol{\phi}_K^{(y)}\right)\left(\boldsymbol{\phi}_K^{(y)}\right)^\mathsf{T}.\end{aligned}$$

Moreover, we denote by $\boldsymbol{D}_{K,K}$ that order K diagonal matrix that has i-th diagonal element 1 if the i-th node is a data point and 0 otherwise. The estimation problem in (5.11)-(5.12) can now be formulated as finding K-vectors $\left(\widetilde{\boldsymbol{X}}_K, \boldsymbol{Z}_K\right) \in \mathbb{R}^K \times \mathbb{R}^K$ that satisfy the equations

$$\boldsymbol{U}_K^\mathsf{T}\boldsymbol{D}_{K,K}\widetilde{\boldsymbol{X}}_K - \lambda\boldsymbol{U}_K^\mathsf{T}\boldsymbol{B}_{K,K}\boldsymbol{Z}_K = \boldsymbol{U}_K^\mathsf{T}\boldsymbol{X}_K^*$$

$$\boldsymbol{V}_K^\mathsf{T}\boldsymbol{A}_{K,K}\boldsymbol{Z}_K + \boldsymbol{V}_K^\mathsf{T}\boldsymbol{B}_{K,K}\widetilde{\boldsymbol{X}}_K = \boldsymbol{0}_K$$

for all $(\boldsymbol{U}_K, \boldsymbol{V}_K) \in \mathbb{R}^K \times \mathbb{R}^K$. Solving this system for $\widetilde{\boldsymbol{X}}_K$ and $\boldsymbol{Z}_K$ is then equivalent to solving the system given by

$$\begin{pmatrix} -\boldsymbol{D}_{K,K} & \lambda\boldsymbol{B}_{K,K} \\ \lambda\boldsymbol{B}_{K,K} & \lambda\boldsymbol{A}_{K,K} \end{pmatrix} \begin{pmatrix} \widetilde{\boldsymbol{X}}_K \\ \boldsymbol{Z}_K \end{pmatrix} = \begin{pmatrix} -\boldsymbol{X}_K^* \\ \boldsymbol{0}_K \end{pmatrix}. \tag{5.14}$$

In view of (5.13) this defines the representation of $\widetilde{X}$ in terms of finite element basis function ϕ_k.

From a computational perspective, this system of linear equations is very fast to solve. Even though the number of basis functions K can be very large ($K = 646$ in our application), the system is highly sparse.

5.2.4 Choosing the smoothing parameter

What remains to be discussed is the optimal choice of the smoothing parameter λ. Let us denote by $\boldsymbol{C}_{2K,2K}$ the order $2K$ matrix on the left-hand side of (5.14) and set $\boldsymbol{S}_{2K,2K} = -\boldsymbol{C}_{2K,2K}^{-1}$. Furthermore we denote by $\boldsymbol{S}_{N,N}$ the order N matrix corresponding to the first N rows and N columns of $\boldsymbol{S}_{2K,2K}$. The order N vector of the spatial smooth $\widetilde{X}$ evaluated at the N sampling stations is then given by $\widetilde{\boldsymbol{X}}_N = \boldsymbol{S}_{N,N}\boldsymbol{X}_N^*$ such that the smoothed values $\widetilde{\boldsymbol{X}}_N$ at the N sampling stations are thus given by a linear transformation of the measurements $\boldsymbol{X}_N^*$.

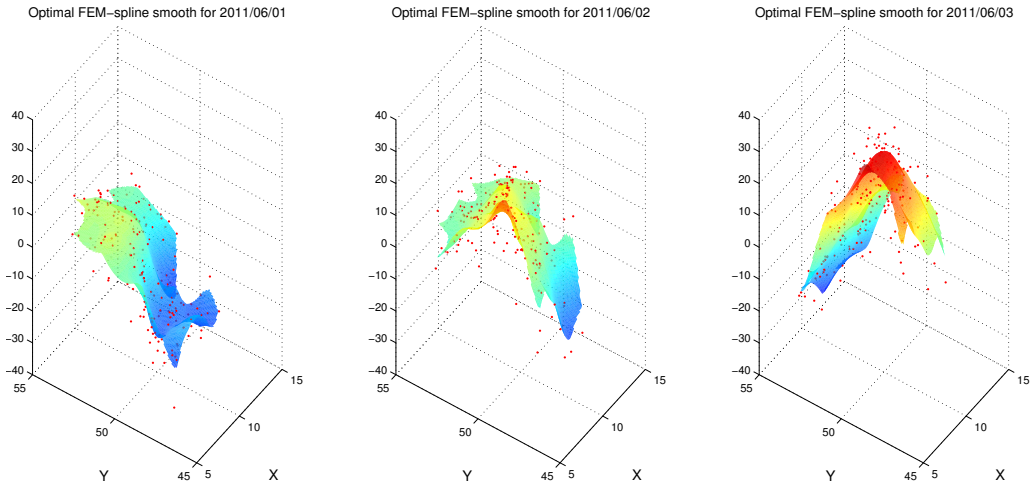


Figure 5.6: **Smoothed ozone concentration surfaces.** This figure shows ozone concentration surfaces over Germany for the first three days in June 2011 that are obtained through a finite element smoothing approach of discrete measurements.

One method commonly used in the literature is to consider minimization of the generalized cross-validation criterion given by

$$GCV(\lambda) = \frac{1}{N\left(1 - \mathrm{tr}(\boldsymbol{S}_{N,N})/N\right)^2} \left(\boldsymbol{X}_N^* - \boldsymbol{S}_{N,N}\boldsymbol{X}_N^*\right)^\top \left(\boldsymbol{X}_N^* - \boldsymbol{S}_{N,N}\boldsymbol{X}_N^*\right) \tag{5.15}$$

(see Ramsay and Silverman [92]). In our application the GCV criterion was minimized for a value of the smoothing parameter given by $\lambda = 0.01$. Associated with this value we call the spatially smoothed surface $\widetilde{X}$ an *optimal FEM-spline smooth*. In Figure 5.6 we plot the optimal FEM-spline smooths of ozone concentration across Germany for three consecutive days from 2011/06/01 until 2011/06/03 with the values of measurements included.

5.3 Forecasting bivariate surfaces with FAR and FAM models

In this section we consider computational implementation and prediction within the frameworks of FAR(1) and FAM-knn models. Our approach is based on a spectral decomposition of the surfaces $(\mathcal{X}_t)_{t=1}^T$ in terms of eigenfunctions of the associated covari-

ance operator. Let us define by

$$\mu := \mathbb{E}[\mathcal{X}] \tag{5.16}$$

$$\mathcal{C}[x] := \mathbb{E}\left[\langle \mathcal{X}, x\rangle \mathcal{X}\right], \quad x \in L^2 \tag{5.17}$$

the mean function and covariance operator of some random (bivariate) function $\mathcal{X} \in L^2(\mathcal{D})$, respectively. Note that since we are working in the space $L^2(\mathcal{D})$, the covariance operator $\mathcal{C}$ defined in (5.17) admits a representation as an integral operator in view of

$$\mathcal{C}[x](\boldsymbol{s}) = \int_{\mathcal{D}} \sigma(\boldsymbol{s}, \boldsymbol{u}) x(\boldsymbol{u}) d\boldsymbol{u},$$

where the kernel of the covariance operator is given by

$$\sigma(\boldsymbol{s}, \boldsymbol{u}) := \mathbb{E}\left[\left(\mathcal{X}(\boldsymbol{s}) - \mu(\boldsymbol{s})\right)\left(\mathcal{X}(\boldsymbol{u}) - \mu(\boldsymbol{u})\right)\right],$$

i.e. the covariance function associated with $\mathcal{X}$. Let us furthermore denote by $\lambda_1 > \lambda_2 > \ldots$ the decreasing sequence of eigenvalues associated with $\mathcal{C}$ and denote by $\psi_1, \psi_2, \ldots$ the corresponding eigenfunctions. Since $\mathcal{C}$ is a symmetric positive-semidefinite Hilbert-Schmidt operator, it admits the singular value decomposition

$$\mathcal{C}[x] = \sum_{l=1}^{\infty} \lambda_l \langle x, \psi_l \rangle \psi_l, \quad x \in L^2(\mathcal{D}). \tag{5.18}$$

Moreover, the eigenfunctions $(\psi_l)_{l \geq 1}$ constitute an orthonormal basis system that span the space L^2 and as such any $\mathcal{X} \in L^2(\mathcal{D})$ admits the Karhunen-Loève decomposition of the form

$$\mathcal{X}(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \sum_{l=1}^{\infty} \theta_l \psi_l(\boldsymbol{s}) \tag{5.19}$$

where now $\theta_l := \langle \mathcal{X}, \psi_l \rangle$ denotes the l-th functional principal component score of $\mathcal{X}$ and convergence of the right-hand-side in (5.19) is understood to be in L^2. By construction, the sequence of functional principal component scores $(\theta_l)_{l \geq 1}$ is such that the θ_l are uncorrelated across the spectral dimension l, have mean zero and variance λ_l. In what follows, we strengthen uncorrelatedness of the θ_l to independence across l, an assumption that is for example satisfied if X is a Gaussian process (see Müller and Yao [84]).

In practice, all quantities above involving an expectation operator need to be estimated on a sample of functions $(\mathcal{X}_t)_{t=1}^T$. Moreover, as discussed in the previous section, the functions $(\mathcal{X}_t)_{t=1}^T$ are reconstructed from discrete observations through finite element splines, yielding a sample of reconstructed functions $(\widetilde{X}_t)_{t=1}^T$ where $\widetilde{X}_t = \sum_{k=1}^K c_{k,t}\phi_k$. Results from Gleim and Salish [57] indicate that, allowing for very general notions of time dependence of the functional time series, the mean function μ and the covariance operator $\mathcal{C}$ as defined in (5.16)-(5.17) can be consistently estimated by

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} \widetilde{X}_t \tag{5.20}$$

$$\widehat{\mathcal{C}}[x] = \frac{1}{T}\sum_{t=1}^{T} \left\langle \widetilde{X}_t - \hat{\mu}, x \right\rangle \left(\widetilde{X}_t - \hat{\mu} \right), \quad x \in L^2. \tag{5.21}$$

As a consequence, for each $t = 1, \ldots, T$, the Karhunen-Loève approximation $\widehat{X}_{t,L}$ of $\widetilde{X}_t$ is then given by a truncated empirical version of expression (5.19), i.e.

$$\widehat{X}_{t,L}(\boldsymbol{s}) := \hat{\mu}(\boldsymbol{s}) + \sum_{l=1}^{L} \hat{\theta}_{t,l}\hat{\psi}_l(\boldsymbol{s}), \tag{5.22}$$

where now the $(\hat{\psi}_l)_{l=1}^L$ are the eigenfunctions associated with the L largest eigenvalues of the estimated covariance operator $\widehat{\mathcal{C}}$ defined in (5.21) and the estimated principal component scores are defined as $\hat{\theta}_{t,l} := \langle \widehat{X}_{t,L}, \hat{\psi}_l \rangle$.

5.3.1 Forecasting with an FAR(1) model

In this section we discuss the estimation and prediction of first-order auto-regressive functional linear models in the space $L^2(\mathcal{D})$. Theoretical results can be found in Bosq [18] whereas computational issues are detailed in Horváth and Kokoszka [73]. We closely follow the exposition in Horváth and Kokoszka [73, Chapters 13.2-13.3] and assume for brevity a mean zero process $(\mathcal{X}_t)_{t=1}^T$ such that the model is given by

$$\mathbb{E}\left[\mathcal{X}_t | \mathcal{X}_{t-1}\right] = \Gamma\left[\mathcal{X}_{t-1}\right], \quad t = 1, \ldots, T \tag{5.23}$$

where Γ denotes a linear operator, which, as we are working in the space L^2, admits an integral representation with respect to some kernel γ in view of

$$\Gamma[x](\boldsymbol{s}) = \int_{\mathcal{D}} \gamma(\boldsymbol{s},\boldsymbol{u})x(\boldsymbol{u})d\boldsymbol{u}, \quad x \in L^2(\mathcal{D}).$$

We apply the Yule-Walker estimation procedure to estimate the auto-regressive operator Γ (or equivalently, its kernel γ) which follows analogously to the classical time series scenario. Note that if Γ is such that there exists an integer j_0 with $\|\Gamma^{j_0}\|_{\mathcal{L}} := \sup(\|\Gamma[x]\| : \|x\| \leq 1) < 1$, we have, for any $x \in L^2$,

$$\mathbb{E}\left[\langle \mathcal{X}_t, x\rangle \mathcal{X}_{t-1}\right] = \mathbb{E}\left[\langle \Gamma\left[\mathcal{X}_{t-1}\right], x\rangle \mathcal{X}_{t-1}\right].$$

Let us define by

$$\mathcal{C}_1 := \mathbb{E}\left[\langle \mathcal{X}_t, x\rangle \mathcal{X}_{t+1}\right]$$

the lag-1 covariance operator. Horváth and Kokoszka [73] show that the following identity holds, i.e.

$$\mathcal{C}_1 = \Gamma\mathcal{C}, \tag{5.24}$$

which suggests a natural approach in estimating Γ by considering a finite sample version of the relation $\Gamma = \mathcal{C}_1\mathcal{C}^{-1}$. However, as the authors point out, the covariance operator $\mathcal{C}$ does not have a bounded inverse on the whole of $L^2(\mathcal{D})$. They argue that as $\mathcal{C}$ admits a singular value decomposition as in (5.18) this implies that $\mathcal{C}^{-1}\left[\mathcal{C}[x]\right] = x$, where

$$\mathcal{C}^{-1}[y] = \sum_{l=1}^{\infty} \lambda_l^{-1}\langle y, \psi_l\rangle\psi_l, \quad y \in L^2(\mathcal{D}),$$

which is then defined if all eigenvalues λ_l of $\mathcal{C}$ are positive. Unboundedness of $\mathcal{C}^{-1}$ now follows in view of $\|\mathcal{C}^{-1}[\psi_l]\| = \lambda_l^{-1} \to \infty$ as $l \to \infty$. As a consequence, estimating the bounded operator Γ through the relationship $\Gamma = \mathcal{C}_1\mathcal{C}^{-1}$ is difficult and some form of regularization has to be employed. A typical approach is, in analogy to the Karhunen-Loève approximation in (5.22), to consider projection on a finite number L of most

important empirical principal components $\hat{\psi}_l$, and to define

$$\widehat{IC}_L[x] = \sum_{l=1}^{L} \hat{\lambda}_l^{-1} \langle x, \hat{\psi}_l \rangle \hat{\psi}_l.$$

This inverse operator is defined on the whole of $L^2(\mathcal{D})$ and it is bounded if $\hat{\lambda}_l > 0$ for $l \leq L$. As in (5.22), the truncation integer L has to be chosen carefully to trade off relevant information in the sample and the explosive behavior of reciprocals of small eigenvalues $\hat{\lambda}_l$.

A computable estimator of Γ is then derived by using an empirical version of the relationship (5.24) and replacing the sample of functions $(\mathcal{X}_t)_{t=1}^T$ with a sample of reconstructed functions $(\widetilde{X}_t)_{t=1}^T$. Since $\mathcal{C}_1$ is estimated by

$$\widehat{\mathcal{C}}_1[x] = \frac{1}{T-1} \sum_{t=1}^{T-1} \langle \widetilde{X}_t, x \rangle \widetilde{X}_{t+1},$$

we obtain, for any $x \in L^2(\mathcal{D})$,

$$\begin{aligned}
\widehat{\Gamma}[x] = \widehat{\mathcal{C}}_1 \widehat{IC}_L[x] &= \widehat{\mathcal{C}}_1 \left(\sum_{l=1}^{L} \hat{\lambda}_l^{-1} \langle x, \hat{\psi}_l \rangle \hat{\psi}_l \right) \\
&= \frac{1}{T-1} \sum_{t=1}^{T-1} \left\langle \widetilde{X}_t, \sum_{l=1}^{L} \hat{\lambda}_l^{-1} \langle x, \hat{\psi}_l \rangle \hat{\psi}_l \right\rangle \widetilde{X}_{t+1} \\
&= \frac{1}{T-1} \sum_{t=1}^{T-1} \sum_{l=1}^{L} \hat{\lambda}_l^{-1} \langle x, \hat{\psi}_l \rangle \langle \widetilde{X}_t, \hat{\psi}_l \rangle \widetilde{X}_{t+1}.
\end{aligned}$$

Whereas this estimator can be used in principle, the authors suggest an additional smoothing step by replacing $\widetilde{X}_t$ with its Karhunen-Loève approximation $\widehat{X}_{t,L} = \hat{\mu} + \sum_{l=1}^{L} \hat{\theta}_{t,l} \hat{\psi}_l$. This leads to the estimator

$$\widehat{\Gamma}_L[x] = \frac{1}{T-1} \sum_{t=1}^{T-1} \sum_{k=1}^{L} \sum_{l=1}^{L} \hat{\lambda}_k^{-1} \langle x, \hat{\psi}_k \rangle \langle \widetilde{X}_t, \hat{\psi}_k \rangle \langle \widetilde{X}_{t+1}, \hat{\psi}_l \rangle \hat{\psi}_l. \tag{5.25}$$

Moreover, the authors show that kernel γ of Γ can be estimated by

$$\hat{\gamma}_L(\boldsymbol{s},\boldsymbol{u}) = \frac{1}{T-1}\sum_{t=1}^{T-1}\sum_{k=1}^{L}\sum_{l=1}^{L}\hat{\lambda}_k^{-1}\langle\widetilde{X}_t,\hat{\psi}_k\rangle\langle\widetilde{X}_{t+1},\hat{\psi}_l\rangle\hat{\psi}_k(\boldsymbol{s})\hat{\psi}_l(\boldsymbol{u}). \tag{5.26}$$

Based on these estimators, the prediction of a new function $\widetilde{X}_{T+1}$ can now be calculated as

$$\widetilde{X}^f_{T+1}(\boldsymbol{s}) = \int_{\mathcal{D}}\hat{\gamma}_L(\boldsymbol{s},\boldsymbol{u})\widetilde{X}_T(\boldsymbol{u})d\boldsymbol{u} = \sum_{l=1}^{L}\left(\sum_{k=1}^{L}\hat{\gamma}_{lk}\langle\widetilde{X}_T,\hat{\psi}_k\rangle\right)\hat{\psi}_l(\boldsymbol{s}) \tag{5.27}$$

where

$$\hat{\gamma}_{lk} = \hat{\lambda}^{-1}(T-1)^{-1}\sum_{t=1}^{T-1}\langle\widetilde{X}_t,\hat{\psi}_k\rangle\langle\widetilde{X}_{t+1},\hat{\psi}_l\rangle.$$

5.3.2 Forecasting with an FAM-knn model

As mentioned before and as can be inferred from (5.25), the FAR(1) model is linear in the predictor scores $\langle x,\hat{\psi}_k\rangle$. Gleim and Salish [57] extend this linear structure by considering a functional additive model (as suggested by Müller and Yao [84]) for the regression function

$$\mathbb{E}\left[\mathcal{X}_{T+1}|\mathcal{X}_T = x\right] = M(x),$$

in which the linear relationship between the functional response and the predictor scores is now allowed to follow a non-parametric model. Now assume that the realization x has a Karhunen-Loève decomposition of the form

$$x = \sum_{l=1}^{\infty}\langle x,\psi_l\rangle\psi_l,$$

where the ψ_l are the functional principal components from before and the function x is thus characterized by a countable sequence of *feature score components* $\theta_l := \langle x,\psi_l\rangle$.

The functional regression model we consider takes then the form

$$M(x) = \sum_{l=1}^{\infty} \mathbb{E}\left[\theta_{T+1,l} | \theta_{T,l} = \theta_l\right] \psi_l, \tag{5.28}$$

where $\theta_{t,l} := \langle \mathcal{X}_t, \psi_l \rangle$ are the true functional principal component scores and where we again tacitly assume that the $\mathcal{X}_t$ have mean zero.

As in the linear FAR(1) model, we approach estimation of (5.28) by regularization where we truncate the infinite sum at some finite integer L. Moreover, the conditional means $\mathbb{E}\left[\theta_{T+1,l} | \theta_{T,l} = \theta_l\right]$ are estimated by considering a k-nearest neighbors classification scheme. This approach was proposed in the functional time series context by Gleim and Salish [57] and consistency results have been derived under very general notions of time dependence between the functional principal component scores $\theta_{t,l}$ across the time dimension t and under the assumption of independence across the spectral dimension l. It proceeds by selecting the k_T neighbors that are closest to the *feature score component* θ_l and we denote the index set of those k_T nearest neighbors to θ_l by $\mathcal{I}(k_T; \theta_l)$[2]. If all quantities in (5.28) were known, this would yield an *infeasible* estimator of $M(x)$ of the form

$$M_{T,L}(x) = \sum_{l=1}^{L} \left(\frac{1}{k_T} \sum_{t \in \mathcal{I}(k_T;\theta_l)} \theta_{t+1,l} \right) \psi_l. \tag{5.29}$$

However, both the functional principal components ψ_l and the functional principal component scores $\theta_{t,l}$ have to be estimated based on a times series of reconstructed surfaces $(\widetilde{X}_t)_{t=1}^T$. As discussed above, this yields Karhunen-Loève approximations $\widehat{X}_{t,L}$, and particularly the feature function x can be written as

$$\widehat{x}_L = \sum_{l=1}^{L} \langle x, \widehat{\psi}_l \rangle \widehat{\psi}_l, \tag{5.30}$$

such that it is characterized by *estimated feature score components* $\hat{\theta}_l := \langle x, \widehat{\psi}_l \rangle$. This

[2]Note that the number of neighbors has to depend on the sample size in that $k_T \to \infty$ as $T \to \infty$ for the estimator to be consistent.

leads to the *feasible* estimator $\widehat{M}_{T,L}(\widehat{x}_L)$ which can be written as

$$\widehat{M}_{T,L}\left(\widehat{x}_L\right) = \sum_{l=1}^{L} \left(\frac{1}{k_T} \sum_{t \in \widehat{\mathcal{I}}\left(k_T;\hat{\theta}_l\right)} \hat{\theta}_{t+1,l} \right) \widehat{\psi}_l,$$

where now $\widehat{\mathcal{I}}(k_T;\hat{\theta}_l)$ denotes the index set of the k_T closest neighbors to the estimated feature score component $\hat{\theta}_l$ (out of a sample $(\hat{\theta}_{t,l})_{t=1}^{T}$ of estimated scores).

The forecasted surface for time point $T+1$ is thus given in terms of its Karhunen-Loève approximation where the corresponding principal component scores are themselves forecasted with k-NN, i.e.

$$\widetilde{X}_{T+1}^{f} = \widehat{M}_{T,L}(\widehat{x}_L).$$

5.4 Empirical results

In this section we present the results of forecasting ground-level ozone concentration surfaces over the geographical area of Germany from a dynamic functional perspective. Both the FAR(1) and FAM-knn method as described above have been implemented.

5.4.1 Data

The data that motivated this research consists of daily measurements of ozone concentration made available through AirBase - the European air quality database provided by the European Environment Agency. This is a public database containing air quality monitoring information for more than 35 countries throughout Europe. We analyzed raw data of daily ozone concentration for Germany which consists of measurements at 1656 stations dating back as far as 1984/01/01. However, not all stations have been operated continuously and continuous measurements are available at $N = 171$ stations for the year 2011. As the raw data set was very large (ca. 5GB), consisting of roughly $20,000$ text files that also contain recordings of other air pollutants, the data was first parsed with a Python script to generate a dataset that could be analyzed further.

5.4.2 Forecasting

In a first step, we reconstructed the sample of $T = 365$ surfaces $(\widetilde{X}_t)_{t=1}^{T}$ from discrete observations at $N = 171$ sample measurement stations using the FEM spline

approach presented in section 5.2. The optimal value of the smoothing parameter λ has been determined through minimizing the generalized cross validation criterion presented in (5.15) which resulted in a parameter value of $\lambda = 0.01$.

In a second step, we obtained the Karhunen-Loève approximations $(\widehat{X}_{t,L})_{t=1}^{T}$ from a functional principal components analysis where the expansion was truncated at $L = 5$ components. The truncation integer L is commonly chosen by means of a scree plot which suggests to find L where the decrease of the (estimated) eigenvalues appears to level off. This truncation adds an additional smoothing step as can be seen when comparing the top left and right panel of Figure 5.7.

In a third step, the forecasts with both the FAR(1) and FAM-knn method have been built using an increasing sample design where the number of neighbors for the FAM-knn method has been taken as $k = 10$. To this end an initial $T_0 = 31$ surfaces $(\widehat{X}_{t,L})_{t=1}^{T_0}$ have been selected as training sample to predict the one-step ahead value at time $T_1 = T_0 + 1$. This corresponds to considering data for the month of January 2011 to predict the value for 2011/02/01. The training sample is then increased by one to contain the surfaces $(\widehat{X}_{t,L})_{t=1}^{T_1}$ and again a one-step ahead prediction is formed to obtain the value at time $T_2 = T_1 + 1$. This procedure continues until we reach the end of the sample at time $T = 365$. In terms of computing time, both methods are equally fast and require ca. 0.56 seconds per forecast step.

5.4.3 Discussion

The bottom two panels of Figure 5.7 show the forecasts obtained with the FAR(1) and FAM-knn method, respectively for 2011/05/15. Both methods provide a good prediction of the general shape of the ozone concentration surface at that date, with the FAM-knn method capturing the range of the true function $\widetilde{X}_t$ considerably better. However, such behavior cannot be inferred uniformly over the forecast horizon. In Figure 5.8 we plot the root mean squared error (RMSE) for the FAM-knn (red line) and FAR(1) (blue line) one-step ahead forecasts when comparing $\widetilde{X}_t^f$ to $\widetilde{X}_t$. Both methods seem to be performing similarly, yielding predictions that are reasonably close to each other.

To get a more differentiated view of forecasting performance, we compare the predictions $\widetilde{X}_t^f$ evaluated at sampling locations at which ozone measurements have been taken. A typical trajectory of forecasts and (smoothed) real data is pictured in Figure 5.9 for station DEBY063 in Regensburg, Germany. Again, the red line corresponds to FAM-knn

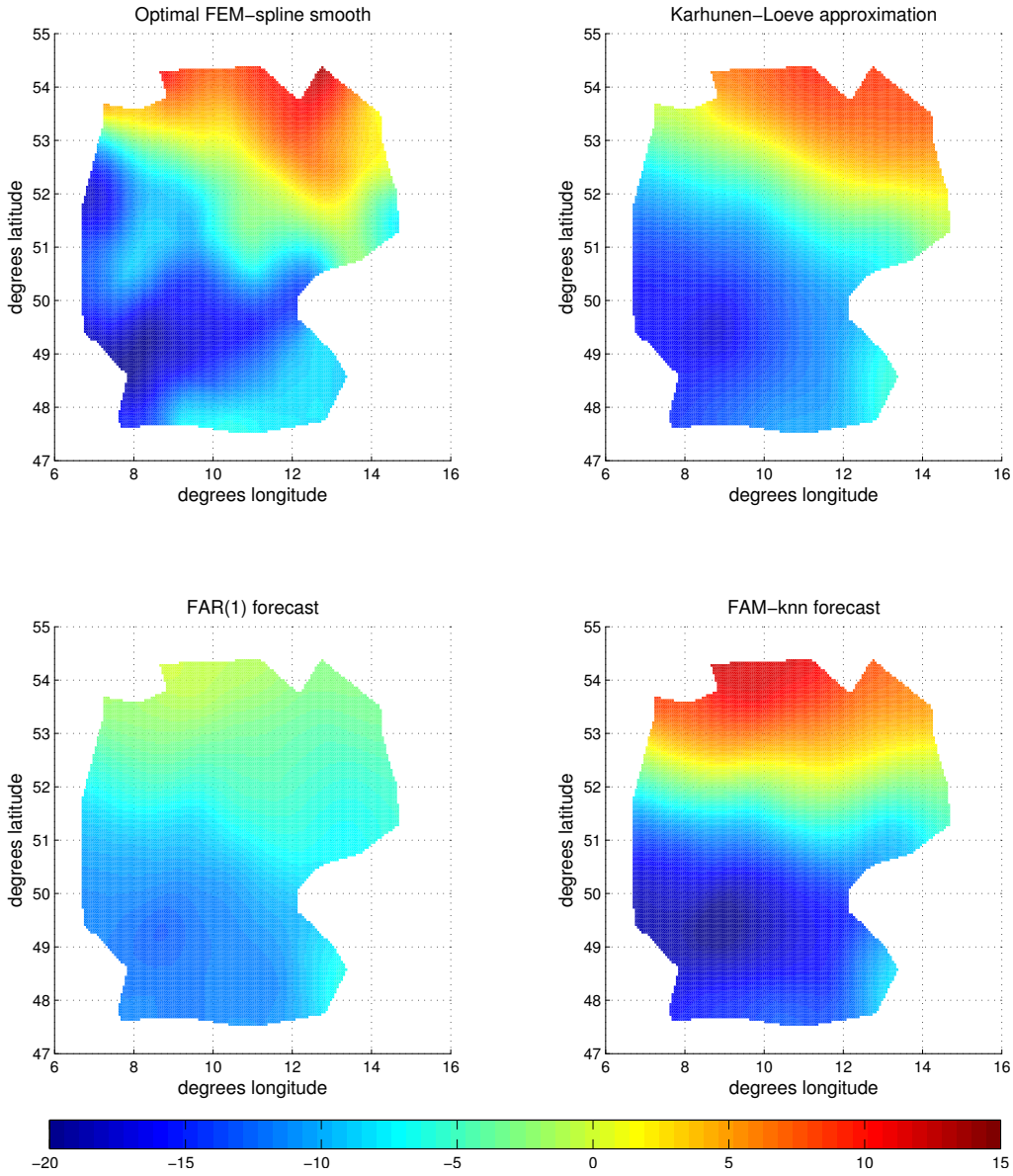


Figure 5.7: **Smoothed ozone concentration surfaces.** The top two panels show the optimal FEM spline smooth and corresponding Karhunen-Loève approximation of the ozone concentration surface over Germany on 2011/05/15. The bottom two panels show the FAR(1) and FAM-knn prediction of ozone concentration at the same date.

forecasts evaluated at at that specific sampling station for the time period 2011/02/01 until 2011/12/31. Similarly, the blue line corresponds to FAR(1) forecasts and the gray line corresponds to smoothed measurements taken at that station (i.e. evaluations of $\widetilde{X}_t$ at sampling locations). As is to be expected, the FAR(1) forecasts exhibit far less

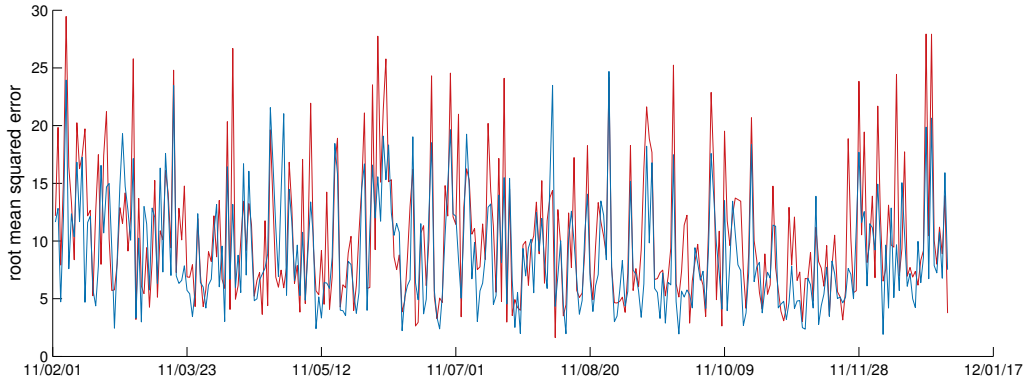


Figure 5.8: **Root mean squared error of functional predictions.** This figure plots the root mean squared error for the FAM-knn (red line) and FAR(1) (blue line) one-step ahead forecasts from 2011/02/01 until 2011/12/31.

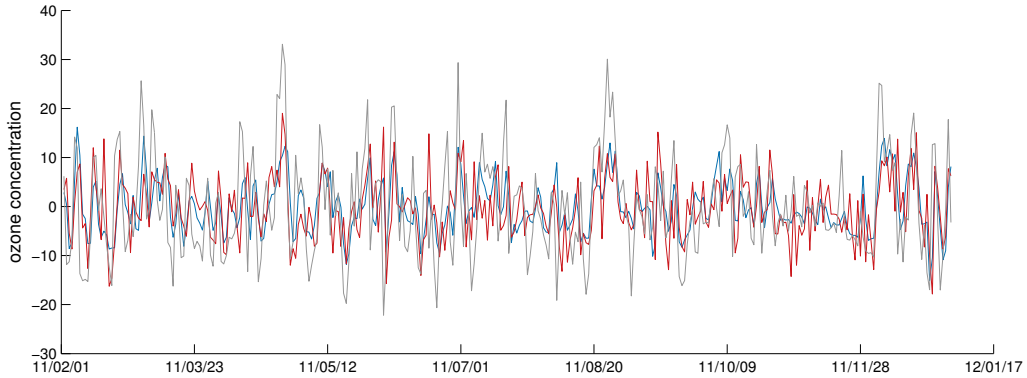


Figure 5.9: **Functional prediction evaluations.** This figure plots the evaluations at sampling station DEBY063 (Regensburg, Germany) for the FAM-knn (red line) and FAR(1) (blue line) one-step ahead forecasts together with smoothed real data (gray line) from 2011/02/01 until 2011/12/31.

variability when compared to real measurements. This is in fact an intrinsic characteristic of forecast performance of the FAR(1) model (see Horváth and Kokoszka [73, Chapter 13.3] for a more detailed discussion). As opposed to the FAR(1) model, the FAM-knn forecasts are somewhat less prone to such behavior. In Figure 5.10 we plot the root mean squared error of forecast evaluations at all sampling locations (where averaging is done over the time dimension). As can be inferred from the plots, the FAR(1) forecasts tend to yield smaller RMSE than the FAM-knn method, even though the differences are small. On average over all sampling locations, the RMSE of the FAR(1) model amounts to 13.8 compared to 14.8 for the FAM-knn method.

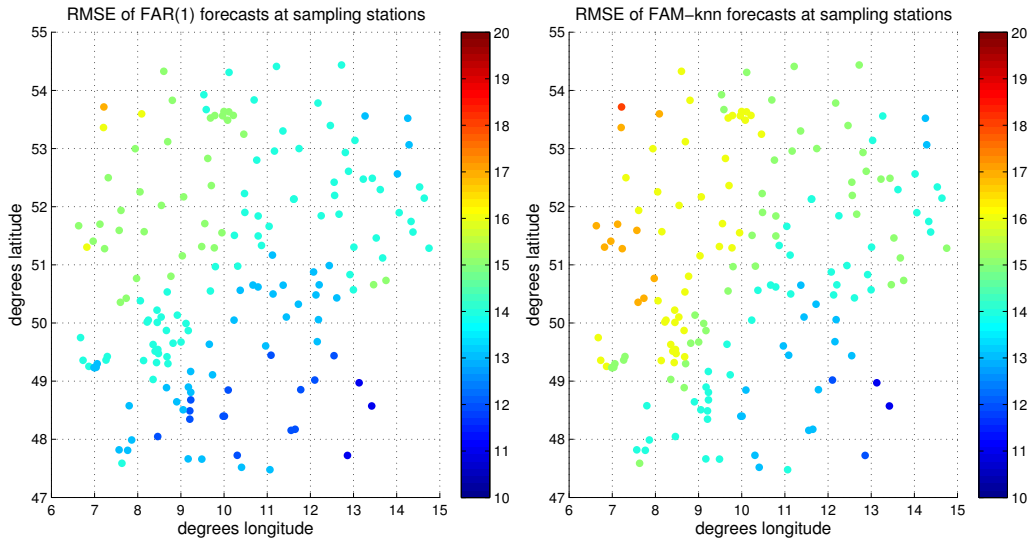


Figure 5.10: **Root mean squared error of functional predictions.** This figure shows the root mean squared error of evaluations of the FAR(1) (left panel) and FAM-knn (right panel) forecast evaluations at all $N = 171$ sampling stations.

5.5 Conclusion

In this paper we were concerned with the problem of forecasting ground-level ozone concentration from a dynamic functional perspective. As opposed to the literature on spatial functional processes, we considered smooth surfaces (i.e. bivariate functions) defined over some spatial domain that are sampled consecutively over time.

As the data was only made available at discrete spatial locations, smooth functions had to be reconstructed by means of a suitable statistical smoothing procedure. In order to take the complex shape of the geographical boundary of the spatial domain into account, we opted for a finite element spline smoother where basis functions are locally defined over a triangulation of the spatial domain.

Two functional first-order auto-regressive models have been applied in a forecasting study, a functional linear and a functional additive model. The non-linear relationship between the predictor scores and the functional response in the FAM model was estimated by means of a k-nearest neighbors classifier. As the results indicate, both the linear FAR(1) and the non-linear FAM-knn model yield predictions that are very close to each other, with the linear model performing slightly better on average.

Bibliography

[1] E. M. Aldrich and A. R. Gallant. Habit, Long-Run Risks, Prospect? A Statistical Inquiry. *Journal of Financial Econometrics*, 9:589–618, 2011.

[2] S. An and F. Schorfheide. Bayesian analysis of DSGE models. *Econometric Reviews*, 26:113–172, 2007.

[3] T. G. Andersen, L. Benzoni, and J. Lund. An empirical investigation of continuous-time equity return models. *Journal of Finance*, 57:1239–1284, 2002.

[4] D. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59:817–858, 1991.

[5] S. Asmussen and P. W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, New York, 2007.

[6] R. R. Bahadur. Sufficiency and statistical decision functions. *The Annals of Mathematical Statistics*, 25:423–462, 1954.

[7] O. E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, Chichester, 1978.

[8] O. E. Barndorff-Nielsen and D. R. Cox. *Inference and Asymptotics*. Chapman & Hall, London, 1994.

[9] O. E. Barndorff-Nielsen and N. Shephard. Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society. Series B*, 63:167–241, 2001.

[10] O. E. Barndorff-Nielsen and N. Shephard. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society. Series B*, 64:253–280, 2002.

[11] O. E. Barndorff-Nielsen, J. Hoffmann-Jørgensen, and K. Pedersen. On the minimal sufficiency of the likelihood function. *Scandinavian Journal of Statistics*, 3: 37–38, 1976.

[12] M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162:2025–2035, 2002.

[13] M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96:983–990, 2009.

[14] P. C. Besse, H. Cardot, and D. B. Stephenson. Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, 27:673–687, 2000.

[15] M. G. B. Blum and D. J. François. Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20:63–73, 2010.

[16] M. G. B. Blum, M. A. Nunes, D. Prangle, and S. A. Sisson. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28:189–208, 2013.

[17] P. Bortot, S. G. Coles, and S. A. Sisson. Inference for stereological extremes. *Journal of the American Statistical Association*, 102:84–92, 2007.

[18] D. Bosq. *Linear Processes in Function Spaces*. Springer, New York, 2000.

[19] D. R. Brillinger. *Time Series: Data Analysis and Theory*. Society for Industrial and Applied Mathematics, Philadelphia, 2001.

[20] F. Bruno, P. Guttorp, P. D. Sampson, and D. Cocchi. A simple non-separable, non-stationary spatiotemporal model for ozone. *Environmental and Ecological Statistics*, 16:515–529, 2009.

[21] T. T. Cai and P. Hall. Prediction in functional linear regression. *The Annals of Statistics*, 34:2159–2179, 2006.

[22] O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert. Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13:907–929, 2004.

[23] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, New York, 2005.

[24] H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statistics & Probability Letters*, 45:11–22, 1999.

[25] G. Celeux, M. Hurn, and C. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95:957–970, 2000.

[26] M. Chernov, A. R. Gallant, E. Ghysels, and G. Tauchen. Alternative models for stock price dynamics. *Journal of Econometrics*, 116:225–257, 2003.

[27] V. Chernozhukov and H. Hong. An MCMC approach to classical estimation. *Journal of Econometrics*, 115:293–346, 2003.

[28] S. Chib and S. Ramamurthy. Tailored randomized block MCMC methods with application to DSGE models. *Journal of Econometrics*, 155:19–38, 2010.

[29] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[30] C. Crambes, A. Kneip, and P. Sarda. Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, 37:35–72, 2009.

[31] D. D. Creal. A survey of sequential Monte Carlo methods for economics and finance. *Econometric Reviews*, 31:245–296, 2012.

[32] M. Creel and D. Kristensen. Indirect likelihood inference. Working Paper, 2013. URL http://pareto.uab.es/wp/2013/93113.pdf.

[33] N. Cressie and H.-C. Huang. Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94:1330–1339, 1999.

[34] K. Csilléry, M. G. B. Blum, O. E. Gaggiotti, and O. François. Approximate Bayesian computation (ABC) in practice. *Trends in Ecology and Evolution*, 25: 410–418, 2010.

[35] J. Davidson. *Stochastic Limit Theory*. Oxford University Press, Oxford, 1994.

[36] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society. Series B*, 68:411–436, 2006.

[37] P. Delicado, R. Giraldo, C. Comas, and J. Mateu. Statistics for spatial functional data: some recent contributions. *Environmetrics*, 21:224–239, 2010.

[38] M. Demetrescu, V. Kuzin, and U. Hassler. Long memory testing in the time domain. *Econometric Theory*, 24(1):176–215, 2008.

[39] Z. Ding and C. W. J. Granger. Modeling volatility persistence of speculative returns: A new approach. *Journal of Econometrics*, 73:185–215, 1996.

[40] R. Douc and O. Cappé. Comparison of resampling schemes for particle filtering. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, pages 64–69, 2005.

[41] C. C. Drovandi, A. N. Pettitt, and M. J. Faddy. Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society. Series C*, 60: 317–337, 2011.

[42] D. Duffie and K. J. Singleton. Simulated moments estimation of Markov models of asset prices. *Econometrica*, 61:929–952, 1993.

[43] G. B. Durham and A. R. Gallant. Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics*, 20(3):297–338, 2002.

[44] R. J. Erhardt and R. L. Smith. Approximate Bayesian computing for spatial extremes. *Computational Statistics & Data Analysis*, 56:1468–1481, 2012.

[45] B. Ettinger, S. Guillas, and M.-J. Lai. Bivariate splines for ozone concentration forecasting. *Environmetrics*, 23:317–328, 2012.

[46] P. Fearnhead and D. Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society. Series B*, 74:419–474, 2012.

[47] T. Flury and N. Shephard. Bayesian inference based only on simulated likelihood: particle filter analysis of dynamic economic models. *Econometric Theory*, 27:933–956, 2011.

[48] S. Frühwirth-Schnatter and L. Sögner. Bayesian estimation of stochastic volatility models based on OU processes with marginal Gamma law. *Annals of the Institute of Statistical Mathematics*, 61:159–179, 2009.

[49] Y.-X. Fu and W.-H. Li. Estimating the age of the common ancestor of a sample of DNA sequences. *Molecular Biology and Evolution*, 14:195–199, 1997.

[50] A. R. Gallant and J. R. Long. Estimating stochastic differential equations efficiently by minimum chi-square. *Biometrika*, 84:125–141, 1997.

[51] A. R. Gallant and R. E. McCulloch. On the determination of general scientific models with application to asset pricing. *Journal of the American Statistical Association*, 104:117–131, 2009.

[52] A. R. Gallant and D. W. Nychka. Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55:363–390, 1987.

[53] A. R. Gallant and G. Tauchen. Seminonparametric estimation of conditionally constrained heterogeneous processes: Asset pricing applications. *Econometrica*, 57:1091–1120, 1989.

[54] A. R. Gallant and G. Tauchen. Which moments to match? *Econometric Theory*, 12:657–681, 1996.

[55] R. Giraldo, P. Delicado, and J. Mateu. Continuous time-varying kriging for spatial prediction of functional data: An environmental application. *Journal of Agricultural, Biological and Environmental Statistics*, 15:66–82, 2010.

[56] R. Giraldo, P. Delicado, and J. Mateu. Ordinary kriging for functional-valued spatial data. *Environmental and Ecological Statistics*, 18:411–426, 2011.

[57] A. Gleim and N. Salish. Prediction in dynamic functional additive models: a k-nearest neighbors approach. Working Paper, 2014.

[58] T. Gneiting. Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97:590–600, 2002.

[59] S. Gonçalves and L. Kilian. Asymptotic and bootstrap inference for AR(∞) processes with conditional heteroskedasticity. *Econometric Reviews*, 26(6):609–641, 2007.

[60] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F, In Radar and Signal Processing*, 140:107–113, 1993.

[61] C. Gouriéroux and A. Monfort. *Statistics and Econometric Models*. Cambridge University Press, Cambridge, 1995.

[62] C. Gouriéroux, A. Monfort, and E. Renault. Indirect inference. *Journal of Applied Econometrics*, 8:S85–S118, 1993.

[63] A. Grelaud, C. P. Robert, J. M. Marin, F. Rodolphe, and J. F. Taly. ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, 4:317–335, 2009.

[64] J. E. Griffin and M. F. J. Steel. Inference with non-Gaussian Ornstein-Uhlenbeck processes for stochastic volatility. *Journal of Econometrics*, 134:605–644, 2006.

[65] P. Guttorp, W. Meiring, and P. D. Sampson. A space-time analysis of ground-level ozone data. *Environmetrics*, 5:241–254, 1994.

[66] P. Hall and J. L. Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91, 2007.

[67] P. R. Halmos and L. J. Savage. Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *The Annals of Mathematical Statistics*, 20:225–241, 1949.

[68] J. D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, 1994.

[69] L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982.

[70] R. J. Hodrick and E. C. Prescott. Postwar U.S. business cycles: An empirical investigation. *Journal of Money, Credit, and Banking*, 29(1):1–16, 1997.

[71] J. Hol, T. Schön, and F. Gustafsson. On resampling algorithms for particle filters. In *Nonlinear Statistical Signal Processing Workshop*, pages 79–82, 2006.

[72] S. Hörmann and P. Kokoszka. Weakly dependend functional data. *The Annals of Statistics*, 38:1845–1884, 2010.

[73] L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer, New York, 2012.

[74] H.-C. Huang and H.-N. Hsu. Modeling transport effects on ground-level ozone using a non-stationary space-time model. *Environmetrics*, 15:251–268, 2004.

[75] A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20:50–67, 2005.

[76] P. Joyce and P. Marjoram. Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7:Article 26, 2008.

[77] A. Justel, D. Peñab, and R. Zamarc. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters*, 35:251–259, 1997.

[78] V. P. M. Le, D. Meenagh, P. Minford, and M. Wickens. How much nominal rigidity is there in the US economy? Testing a new Keynesian DSGE model using indirect inference. *Journal of Economic Dynamics and Control*, 35:2078–2104, 2011.

[79] B.-S. Lee and B. F. Ingram. Simulation estimation of time-series models. *Journal of Econometrics*, 47:197–205, 1991.

[80] T. Magnac, J.-M. Robin, and M. Visser. Analysing incomplete individual employment histories using indirect inference. *Journal of Applied Econometrics*, 10: S153–S169, 1995.

[81] J. M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22:1167–1180, 2012.

[82] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100:15324–15328, 2003.

[83] D. McFadden. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57:995–1026, 1989.

[84] H.-G. Müller and F. Yao. Functional additive models. *Journal of the American Statistical Association*, 103:1534–1544, 2008.

[85] D. Nerini, P. Monestiez, and C. Manté. Cokriging for spatial functional data. *Journal of Multivariate Analysis*, 101:409–418, 2010.

[86] M. A. Nunes and D. J. Balding. On optimal selection of summary statistics for approximate Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 9:Article 34, 2010.

[87] A. Pakes and D. Pollard. Simulation and the asymptotics of optimization estimators. *Econometrica*, 57:1027–1057, 1989.

[88] G. W. Peters, S. A. Sisson, and Y. Fan. Likelihood-free Bayesian inference for α-stable models. *Computational Statistics & Data Analysis*, 56:3743–3756, 2012.

[89] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16:1791–1798, 1999.

[90] P. Rakotomarolahy. Statistical properties of nearest neighbor regression function estimate for strong mixing processes. Working Paper, 2012. URL http://hal.archives-ouvertes.fr/docs/00/72/58/41/PDF/Second_Order_properties_kNN.pdf.

[91] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis: methods and case studies*. Springer, New York, 2002.

[92] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2nd edition, 2005.

[93] T. Ramsay. Spline smoothing over difficult regions. *Journal of the Royal Statistical Society*, 64:307–319, 2002.

[94] O. Ratmann, C. Andrieu, C. Wiuf, and S. Richardson. Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences*, 106:10576–10581, 2009.

[95] C. P. Robert. *The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation*. Springer, New York, 2nd edition, 2001.

[96] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition, 2004.

[97] G. O. Roberts, O. Papaspiliopoulos, and P. Dellaportas. Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility processes. *Journal of the Royal Statistical Society. Series B*, 66:369–393, 2004.

[98] L. M. Sangalli, J. O. Ramsay, and T. O. Ramsay. Spatial spline regression models. *Journal of the Royal Statistical Society, Series B*, 75:681–703, 2013.

[99] S. A. Sisson and Y. Fan. Likelihood-free Markov chain Monte Carlo. In S. P. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 313–336. Chapman & Hall/CRC, 2011.

[100] S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104:1760–1765, 2007.

[101] A. A. Smith. Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics*, 8:S63–S84, 1993.

[102] C. J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5:595–620, 1977.

[103] W. Stute. Asymptotic normality of nearest neighbor regression function estimates. *The Annals of Statistics*, 12(3):917–926, 1984.

[104] M. M. Tanaka, A. R. Francis, F. Luciani, and S. A. Sisson. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, 173:1511–1520, 2006.

[105] E. Taufer, N. Leonenko, and M. Bee. Characteristic function estimation of Ornstein-Uhlenbeck-based stochastic volatility models. *Computational Statistics & Data Analysis*, 55:2525–2539, 2011.

[106] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145:505–518, 1997.

[107] G. Topa. Social interactions, local spillovers and unemployment. *The Review of Economic Studies*, 68:261–295, 2001.

[108] S. Yakowitz. Nearest-neighbour methods for time series analysis. *Journal of Time Series Analysis*, 8:235–247, 1987.

[109] Y. Yamanishi and Y. Tanaka. Geographically weighted functional multiple regression analysis: a numerical investigation. *Journal of the Japanese Society of Computational Statistics*, 15:307–317, 2003.

[110] F. Yao, H.-G. Müller, and J.-L. Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33:2873–2903, 2005.

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 12 point Arno Pro, designed by Robert Slimbach in the style of book types from the Aldine Press in Venice, and issued by Adobe in 2007. A template, which can be used to format a PhD thesis with this look and feel, has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/ or from the author at suchow@post.harvard.edu.