Essays in Applied Microeconomics and Numerical Optimization

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

Mariam Petrosyan

aus Ashtarak, Armenien

Bonn

2024

Dekan:	Prof. Dr. Jürgen von Hagen
Erstreferent:	Prof. Dr. Hans-Martin von Gaudecker
Zweitreferent:	Prof. Dr. Jürgen Maurer
Tag der mündlichen Prüfung:	07.03.2024

Acknowledgements

First and foremost, I want to thank my supervisor Hans-Martin von Gaudecker for his unwavering support in many aspects of my PhD journey. He gave me the opportunity to work on a number of interesting projects with incredibly talented people. Without his academic guidance and personal encouragement this thesis would have not happened.

I am also deeply thankful to my second supervisor, Jürgen Maurer, for his support and guidance. His expertise in health economics has been a beacon of light, guiding me through the complexities of my research and helping me to refine my ideas with a sharper focus.

I also want to thank Julia Mink for agreeing to chair my committee and for all the advice she provided on how to navigate both inside the academia, and outside when I was applying for jobs.

Apart from being my supervisors, Hans-Martin von Gaudecker and Jürgen Maurer are coauthors to two chapters of my thesis, and for that, I am indebted to them. I am also thankful to my other coauthors Janoś Gabler, Tim Mensinger, and Sebastian Gsell.

I am grateful as well for the support from the Bonn Graduate School of Economics, Institute of Macroeconomics and Econometrics, and the Research Training Group 2281.

Finally, I want to thank the people who are closest to me: my family for always making me feel their love despite being so far, and Janoś for offering me his unconditional support and love, and giving me the opportunity to become part of his amazing family.

Contents

List of Figures

List of Tables

Introduction

This thesis combines three self-contained research studies. In the first chapter, we use empirical data to study the dynamics of physical and cognitive capacities during later stages of life within. Concerning a related topic, but in a different setting, the second chapter also uses empirical data to study functional outcomes in the later stages of life, with a particular aim of understanding how environmental factors interact with intrinsic capacities to determine those outcomes. In the third chapter, we propose an optimization algorithm tailored to the needs of economists who deal with "hard" optimization problems when fitting structural models to empirical data.

Chapter 1: Mens Sana in Corpore Sano? Almost by definition, human capital development, encompassing the acquisition of knowledge, cognitive skills, and physical abilities, plays a crucial role in shaping individuals' lives. While the importance of early childhood investments in the formation of human capital has been studied extensively, the literature on the maintenance of human capital reserves in later stages of life, to slow down the decline of cognitive and physical capacities, is scarce. This chapter aims to bridge this gap by employing a dynamic modeling approach to investigate the interdependencies between physical and cognitive capacity throughout individuals' later life stages.

To this end, we adapt the *Technology of Skill Formation* model by Cunha, Heckman, and Schennach (2010) to the context of aging. Our model introduces two latent factors reflecting human capital- physical and cognitive capacities, and two investment factors- physical exercise and cognitive stimulation. As in Cunha, Heckman, and Schennach (2010), we treat the factors as unobservable and estimate their joint distribution by modeling them in a dynamic system of state-space equations together with observable *measurement* variables.

Following Cunha, Heckman, and Schennach (2010), we estimate the model parameters via a maximum likelihood estimator. We apply the estimator on data from the Health and Retirement Study (*Health and Retirement Study (HRS)* no date). To remove the selection bias introduced by the link between poor health and mortality, we incorporate a simple model of mortality into the dynamic latent factor model. To account for gender-specific differences in health trajectories, we perform the estimation separately for females and males.

Our study yields three main results: 1) We estimate substantial noise in all observed variables. No single observable variable can be detected as a perfect measurement for the latent factor it represents, rendering it impossible to use just one measurement variable and ignore the measurement errors in the analysis. 2) Despite a general decline in physical and cognitive capacity with

age, the relative ranking of latent factors remains remarkably stable. 3) Investments in physical and cognitive capacity can influence these latent factors until very late stages of life. The impact of cognitive stimulation is limited to cognitive capacity, while physical exercise primarily enhances physical capacity, albeit with a modest impact on cognitive capacity.

Chapter 2: Intrinsic and External Determinants of Age-Related Decline in Functioning Aging is an inevitable biological process that affects all living organisms. At the biological level, the gradual accumulation of molecular and cellular damage associated with aging (World Health Organization, 2015) can lead to a broad spectrum of impairments that limit an individual's ability to perform daily activities and maintain independence. Yet, age-related functional decline and the associated risk of disability is neither a deterministic nor a linear function of the biological age (World Health Organization, 2015), but is influenced by a complex interplay between *intrinsic* and *environmental* factors.

Deterioration of intrinsic capacities manifested mainly through accumulating multiple chronic conditions in older adults can strain their physical and mental resources, leading to reduced mobility, fatigue, and cognitive impairment. At the same time, environmental factors, including social support, access to healthcare, and living conditions, can either exacerbate or mitigate the effects of intrinsic factors.

Understanding the intricate relationship between intrinsic and environmental factors is crucial for developing effective interventions to promote healthy aging and prevent functional decline. By identifying modifiable environmental factors, we can design strategies to optimize individuals' environments and enhance their ability to maintain independence and quality of life as they age.

While the research on both intrinsic capacities and external factors as possible determinants of old-age disability is rich, and the importance of viewing disability in the context of one's environment has been established, to the best of our knowledge, in the existing literature, the prevalent econometric approach is the modeling of the relevant variables in a somewhat simplified, linear manner.

In this chapter, we estimate *nonlinear* interaction terms between intrinsic and extrinsic factors and their impact on disability rate. To this end, we use the semi-parametric double index binary choice estimator developed in Klein and Vella (2009). In this econometric model, we identify two indices, *intrinsic* and *extrinsic*, which are summary quantification of intrinsic and environmental factors, respectively. Within this framework, we are able to abstain from parametric assumptions regarding the functional form of the link function between the two indices to obtain the predicted probability of being disabled.

We find that the environmental index has a nontrivial impact on the predicted probability of being disabled. We also find considerable nonlinear interaction effects between intrinsic and extrinsic indices. In particular, we find the intrinsic gradient of the predicted probability of disability to be steeper at lower quantiles of the environmental index.

Chapter3: Tranquilo In this study, we propose the tranquilo algorithm, a model-based (derivative-free) trust-region optimizer that aims to facilitate optimization problems that arise during the method of simulated moments estimation (MSM).

Despite the prevalence of MSM estimation in structural papers (see Eisenhauer, Heckman, and Mosso (2015) for a review) and widely available anecdotal evidence that structural researchers

would love to spend less time on solving optimization problems, there are no specialized optimization algorithms that are tailored to the characteristics of MSM estimation problems.

tranquilo has been developed to precisely close this gap by enabling scientists to solve hard optimization problems more frequently arising during MSM estimation with less need for manual intervention. The difficulty here arises both from the computational complexity, often requiring hours or even days of runtime, and from the necessity of many manual interventions to achieve the right configuration of start values and algorithm parameters.

We restrict our attention to economic problems, allowing us to make the critical assumption that most of the computational costs are due to the objective function. The algorithm is particularly suited for this type of problem as it (1) can utilize the least-squares structure of the MSM problem, (2) can be parallelized on the level of the algorithm, and (3) can adaptively deal with noise in the objective function. By comparing benchmark results, we show that *tranquilo* can compete with state-of-the-art algorithms and even outperform them in specific scenarios. At the same time, the usefulness of *tranquilo* is not restricted to the field of economics, as problems the same characteristics are also encountered in other fields. Prime examples are design optimization in engineering or calibrating epidemiological models to empirical data.

In *tranquilo*, we make the following contributions:

First, We adopt a conventional trust-region approach for nonlinear least-squares solvers (see, for instance, Conn, Gould, and Toint (2000)) and restructure it in a modular style that facilitates the substitution of individual algorithm components to tailor it to the specifics of MSM estimation issues.

Second, we add parallelization capabilities to the trust-region framework. While some parts of derivative-free trust-region algorithms have been parallelized in other algorithms, we add two new ideas for a more efficient parallelization: The first is a parallel line search that tries out multiple step lengths in the search direction obtained by solving the trust-region subproblem. The second is speculative sampling: While doing the function evaluation(s) needed to decide whether a candidate point is accepted, we already sample points that would be helpful in the next iteration if the candidate point is accepted, and evaluate the objective function on those points.

Third, we propose novel ways of adaptively determining how many function evaluations are needed to average out the noise just enough so that the optimizer can make progress.

Fourth, we make *tranquilo* (Gabler, Gsell, Mensinger, and Petrosyan, 2024) available as an open-source Python package that can be used in isolation or via the *estimagic* package (Gabler, 2022).

References

Conn, Andrew R., Nicholas I. M. Gould, and Philippe L. Toint. 2000. *Trust Region Methods*. Society for Industrial, and Applied Mathematics. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9780898719857. [3]

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica* 78 (3): 883–931. [1]

Eisenhauer, Philipp, James J. Heckman, and Stefano Mosso. 2015. "ESTIMATION OF DYNAMIC DISCRETE CHOICE MODELS BY MAXIMUM LIKELIHOOD AND THE SIMULATED METHOD OF MOMENTS." *International Economic Review* 56 (2): 331–57. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/iere.12107. [2]

Gabler, Janoś. 2022. "A Python Tool for the Estimation of large scale scientific models." [3]

Gabler, Janoś, Sebastian Gsell, Tim Mensinger, and Mariam Petrosyan. 2024. "Tranquilo." [3]

"Health and Retirement Study (HRS)." No date. "Health and Retirement Study (HRS)." [1]

Klein, Roger, and Francis Vella. 2009. "A semiparametric model for binary response and continuous outcomes under index heteroscedasticity." *Journal of Applied Econometrics* 24 (5): 735–62. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jae.1064. [2]

World Health Organization. 2015. *World report on ageing and health*. World Health Organization, 246 p. [2]

Chapter 1

Mens Sana in Corpore Sano?

Joint with Hans-Martin von Gaudecker, Jürgen Maurer, and Janoś Gabler

1.1 Introduction

Development and maintenance of human capital throughout the life-course enables individuals to lead longer, more productive and more satisfactory lives. The notion of human capital generally comprises a broad range of useful abilities that shape individuals' capabilities, behaviors and wellbeing such as their knowledge, skills, and health among others (World Bank, 2018). While there is a large economic literature on early-life human capital development and its effects on adult outcomes (Heckman and Mosso, 2014), fewer studies in economics have analyzed the roles individual investments and corresponding technologies for the maintenance and depreciation of human capital during later life within an integrated framework to model later-life human capital dynamics (McFadden, 2008).

Physical and cognitive capacity represent two key forms of human capital during adulthood and are perhaps the most important forms of human capital at older ages, especially after retirement. Physical and cognitive capacity are key determinants of many important outcomes in health economics and beyond such as mortality, healthcare use and healthcare cost and spending, falls and disability, long-term care needs and nursing home use, economic and social participation and subjective wellbeing to name but a few. As a result, investments in the maintenance of physical and cognitive capacity are key to ensuring a healthier, longer, and happier old-age. Moreover, since many of these outcomes are highly uncertain, demand for various healthcare and long-term care related insurance products depends on the later-life dynamics of physical and cognitive capacity (Hosseini, Kopecky, and Zhao, 2022). Understanding the later-life dynamics of physical and cognitive capacity is, therefore, a key pre-requisite and input into models aimed at studying

⋆

We would like to thank Johannes Ewald for preparatory work in his M.Sc. thesis at the University of Bonn (March 2020). The authors are grateful for support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1– 390838866 – and through CRC-TR 224 (Project C01).

the role of later-life human capital on these important later-life outcomes and related investment and insurance decisions.

While physical and cognitive capacity tend to decline during later life (Niccoli and Partridge, 2012), there is considerable heterogeneity in the onset and speed of such aging-related declines across individuals, which is often related to individual differences in exposures and investments (Crimmins, 2020). What is more, several studies have in fact shown significant improvements in later life physical and cognitive capacity following targeted investments such as physical exercise programs or cognitive trainings, suggesting that both physical and cognitive function remain malleable even at very high ages (Fiatarone, O'Neill, Ryan, Clements, Solares, et al., 1994; Ball, Berch, Helmers, Jobe, Leveck, et al., 2002). This evidence suggests that aging-related changes in function are not fully pre-determined biologically but can be postponed, slowed down, compensated and in certain instances perhaps even (temporarily) reversed or overcompensated through appropriate later-life investments. These findings highlight the important role of health investments for physical and cognitive capacity throughout the entire life course, even if early-life health investments into health to build up "reserves" for later life may be more efficient due to a higher degree of malleability early in life, the longer time horizon available to capitalize on early investments and potentially important complementarities of health investments over time (Cunha, Heckman, and Schennach, 2010).

Besides documenting the continued malleability of physical and cognitive capacity during later life, the more recent literature in gerontological science has also found for evidence potentially important cross-effects of physical function on cognitive function and vice versa. These cross-effects may go beyond the responses of physical and cognitive function due to common risk factors such as physical inactivity or diseases affecting both physical and cognitive capacities such as Parkinson's disease, and represent more general connections between physical and cognitive capacity (Clouston, Brewster, Kuh, Richards, Cooper, et al., 2013). Evidence for such connections comes from both observational studies and RCTs, often but not always focused on the connection between cognitive and gait (dys-)function (Montero-Odasso, Verghese, Beauchet, and Hausdorff, 2012). In view of these findings, economic models of human capital maintenance and depreciation during later life should thus allow for flexible later-life dynamics of physical and cognitive capacities that can incorporate different forms of investment, and possible cross-effects between physical and cognitive capacities.

Varied existing conceptualizations of physical and cognitive capacity used in the literature and potentially widespread measurement error in physical and cognitive assessments in survey data and self-reported health investments further complicate the already complex task of capturing the joint dynamics of later-life physical and cognitive capacity and related investments (Bound, Brown, and Mathiowetz, 2001; Baker, Stabile, and Deri, 2004; Kapteyn, Banks, Hamer, Smith, Steptoe, et al., 2018; Hosseini, Kopecky, and Zhao, 2022). Physical capacity, for example, is a multifaceted concept that is generally assessed through multiple self-reported and/or performance-based survey items presenting noisy measurements for underlying true physical capacity (Kasper, Chan, and Freedman, 2017). Similarly, cognition comprises a range of different cognitive functions such as such as perception, attention, intelligence, knowledge, memory and working memory, judgement, reasoning, computation, problem solving or comprehension, whose

corresponding measurements have signal value for overall cognitive capacity (Salthouse, 2010; Salthouse, 2012). Perhaps more surprisingly, even commonly used survey items for health investments such as self-reported physical activity contain substantial measurement error relative to actual health investments and, therefore, need to be treated with caution (Kapteyn et al., 2018). Given the large potential for significant measurement error in survey-based assessments of physical and cognitive capacity and corresponding health investments documented in the literature, it seems prudent to employ an analytical framework that can readily accommodate such measurement errors when analyzing the joint dynamics of these outcomes.

The main objective of this paper is to estimate the technology for human capital maintenance and depreciation in later-life focusing on the dynamic interplay between later-life physical and cognitive capacity and corresponding investments among older adults in the US. To this end, we propose the use of a non-linear dynamic latent factor model as first proposed by Cunha, Heckman and Schennach (Cunha, Heckman, and Schennach, 2010) as a framework to model early-life human capital accummulation, to study later-life human capital depreciation processes using longitudinal data from the US Health and Retirement Study (HRS). Applying this framework to investigate the joint dynamics of later-life physical and cognitive capacity and related investments is very attractive as such a non-linear dynamic latent factor model can incorporate the main aforementioned stylized facts about human capital depreciation, i.e., (1) allowing for a joint modelling of physical and cognitive capacity and investments that can incorporate potentially important cross-domain effects; (2) integrating the continued malleability of both physical and cognitive capacity into the model to study dynamically optimal investment paths and (3) accounting for error in the measurement of physical and cognitive function and corresponding investments in a context where there are several measurements of each of these domains in many commonly used data sets, but each measurement is likely to provide only a noisy signal for the underlying construct at hand. In addition to accommodating key stylized facts about human capital maintenance and depreciation into a unified framework, our model also allows us to identify the distribution of latent factors from noisy measurements, simulate the effects of different investment patterns on physical and cognitive capacity, calculate optimal investment patterns, notably the role of investments for human capital maintenance in younger old vs older old individuals, and anchor the results in interpretable metrics such as survival probabilities.

Our paper relates to two strands of research in economics, a methodological one on the use of non-linear dynamic latent factor models for estimating dynamic human capital production, which has-to the best of our knowledge-so far only been applied to the case of human capital accumulation in early life but not to human capital maintenance and depreciation in later life, and a more substantive one on the measurement and modelling of health dynamics during adulthood and later life. From a methodological point of view, our paper transfers widely used methods for the study of early-life human capital accumulation to the study of later-life dynamics of physical and cognitive function and eventual mortality. As a technical contribution, we show how to incorporate mortality into the framework and improve the numerical stability of a well known maximum likelihood estimator. By applying non-linear dynamic latent factor models to questions of aging and later life health dynamics, we show the usefulness of these methods to study human development not just in early life but across the entirely life-course, especially since many

of the modelling and measurement issues mentioned above seem common to both ends of the life-course. As a result, we hope that our paper will aspire a larger group of life-course and aging researchers to consider such models in their research both in health economics and related fields.

Substantively, we contribute to the literature on how to measure and model later-life health dynamics in situations where we observe multiple potentially very noisy measurements for fewer latent concepts such as physical and cognitive capacity, which has long challenged empirical analyses in health economics and beyond. More specifically, one important issue in this literature is how to measure health in a comprehensive yet parsimonious way in view of the multifaceted nature of health on the one hand and the common need for dimensionality reduction in econometric models on the other. To address this trade-off, one set of commonly adopted approach to measuring health is to directly use (usually ordered measurements of) self-rated health as summary measure of health as outcome of interest (Contoyannis, Jones, and Rice, 2004; Heiss, 2011; Latham and Peek, 2012). This approach is generally motivated by a high predictive value of self-rated health for mortality (Idler and Benyamini, 1997). Alternatively to directly using self-reports to measure health, a commonly used approach is to "instrument" health via a larger and "more objective" set of individual health measurements, such as information on specific health conditions, functional limitations, performance test results or anthropometric measures. This approach endogenously derives weights for aggregating the more detailed set of individual health measurements into a single health index that can then be used in further analysis (Cutler and Richardson, 1997; Jürges, 2007). Relative to using self-rated health directly as outcome, the approach aims to improve measurement by using "more objective" measures of health to construct an underlying health index, whereby the weights attributed to each detailed and "more objective" health measure in the final health index is determined by the partial association of the respective detailed health measure with self-rated health. While this approach can address some known issues with self-rated health, such as potential age-, sex- or SES-dependent reporting heterogeneity (Lindeboom and Van Doorslaer, 2004; Dowd and Zajacova, 2007; Dowd and Zajacova, 2010), there is often still considerable measurement error in the "more objective" health measures that cannot be purged using this approach and may require further consideration (Baker, Stabile, and Deri, 2004; Maurer, Klein, and Vella, 2011). A second related approach side-steps the use of self-rated health entirely and instead uses principal component analysis of the more detailed health measurements to derive lower dimensional health indices (Jenicek, Cleroux, and Lamoureux, 1979; Poterba, Venti, and Wise, 2017; Nakazato, Sugiyama, Ohno, Shimoyama, Leung, et al., 2020). A third and increasingly popular approach simplifies the aggregation process for the more detailed health measurements even further by constructing a so-called “frailty index” or “deficit index”, which simply consists of the total number of prevalent "health deficits" divided by the total number of potential "health deficits" (Rockwood and Mitnitski, 2007; Hosseini, Kopecky, and Zhao, 2022). A such constructed "frailty index"/"deficit index" is thus bounded to lie between zero and one and represents the percentage of potential "health deficits" already suffered by a given individual. A final set of studies refrains from performing some form of dimensionality reduction and uses the more detailed health measures directly in their analyses, either in isolation or simultaneously. As this is, for example,, the standard approach of disease-based analyses, most published papers on health adopt this latter approach.

While all of the aforementioned approaches have their respective advantages and disadvantages in measuring and modelling health in economic applications and have been employed with some success in the literature, they have mainly been used to describe the dynamic evolution of health during adulthood as inputs for structural models in health economics concerning retirement, housing or insurance decisions rather than studying the production technology of later life health maintenance or depreciation directly. Regarding the latter, the aforementioned approaches have some potential downsides that we aim to address in this paper. First, to the best of our knowledge, our paper is the first to explicitly study the dynamic interplay between physical capacity, cognitive capacity and related investments in the context of a structural non-linear dynamic latent factor model as first proposed by Cunha, Heckman and Schennach (Cunha, Heckman, and Schennach, 2010), which can generate new insights on the dynamic relationships between physical and cognitive capacity as well as investment into these important facets of human capital. Second, explicitly distinguishing between physical and cognitive capacity is thereby not only important due to increasing evidence for potentially important cross-effects between the two health domains cited above but also in view of likely differences in the consequences of depleted levels of physical vs cognitive capacity for functioning, participation and other important later life outcomes (Crimmins, 2020; Amengual, Bueren, and Crego, 2021). In the economics literature, there is to date only limited evidence on the potential cross-effects between physical and cognitive capacity maintenance with Schiele and Schmitz (Schiele and Schmitz, 2021) being a notable exception studying the effects of adverse physical health shocks on cognitive capacity in later life using non-structural event study methods. Third, our approach can accommodate a situation where information about a few latent factors needs to be extracted from many measurements of the underlying construct which can potentially suffer from severe measurement error.

Our analysis complements the aforementioned approaches to modelling and analyzing later-life health by delivering new insights on the dynamics of later-life human capital and related investments among older adults in the US. Our approach, thereby, highlights the structural production function of older adults concerning the maintenance and depreciation of physical and cognitive capacity. Our key findings are as follows: 1) There is substantial noise in all observed variables. While most measurements have a high correlation with the latent factor they measure, no single measurement dominates to an extent where it would be justified to just use a single variable and ignore the measurement error in the econometric analysis. 2) Despite a strong decline in means of physical and cognitive capacity, the rank order of these latent factors is remarkably stable. 3) Physical and cognitive capacity can be influenced by investments until very high ages. Cognitive stimulation is a specific investment into cognitive capacity. Physical exercise has a larger effect on physical capacity and a small effect on cognitive capacity.

The remainder of the paper is organized as follows: Section 1.2 provides information on our main data source and gives detailed description of the factor measurements. Section 1.3 describes our empirical approach and the challenges associated with it. Section 2.4 presents and discusses our results, and section 1.5 concludes.

1.2 Data and Measurements

We base our empirical analysis on the 1992-2016 waves of the Health and Retirement Study (HRS) conducted by The University of Michigan. The HRS offers longitudinal panel data with representative sample of approximately 40,000 individuals living in the U.S. and aged 50 and above. The HRS core questionnaire offers rich set of measures of physical health, mental status, and behaviors. Measures of physical and cognitive capacities include self-reported diagnoses, subjective assessments, and objective biomedical markers. Additional off-wave surveys offer additional measures that are particularly relevant for our analysis. Specifically, we employ the Consumption and Activities Mail Survey (CAMS) (Health and Retirement Study, 2022b) to extract measurements for Exercise and Cognitive Stimulation.

Wherever possible, we include data prepared by the RAND corporation (Health and Retirement Study, 2022c), which provides a harmonized and easy-to-use version of the core HRS data. Out of the many variables we need, several are not included in the RAND HRS data, however, and we recur to the original core files (Health and Retirement Study, 2022b).

We start our analysis at age 68, when most people are retired and we start to see meaningful variation in the measures at our disposal for physical and cognitive capacity. The last age we consider is 93, after which the sample size becomes small. Since the HRS questionnaire is administered biannually, we work with two-year transitions and age groups. For conciseness, we refer to these age groups by the lower bound included – "age 68" thus includes ages 68 and 69, and at the other end of the spectrum "age 92" comprises ages 92 and 93. Because men and women show very different aging patterns, we present all statistics by gender. We will also estimate the model separately for each gender.

We standardize almost all measures to have mean zero and unit variance in the first age group included in our data. Any age trends are thus preserved. For example, until age 90, the mean of (residualized) grip strength declines by around 1.4 original standard deviations. At the same time, the dispersion of grip strength shrinks to around 80% of its original standard deviation. For categorical variables, all of which have numerical values with spacing 1, we add noise using uniform distributions on $(-0.5, 0.5)$. This preserves the original ordering and add to the numerical stability of the estimator below. Changing the seed of the random number generator did not affect any results; future work will pursue additional robustness exercises.

1.2.1 Physical Capacity

We employ six variables as measurements for physical capacity. Quite naturally, **vital status** is a dummy for being alive, which becomes zero in the first HRS wave after an individual has died. It is set to missing thereafter, so that the average of this variable can be interpreted as the probability of surviving until the next survey wave. The first row of Figure 1.2.1 shows the age trends in our measures of physical capacity.[1] Unsurprisingly, survival probabilities decreases in age both for

1. Figure 1.A.1 in Appendix 1.A shows the same trends for the standard deviations of our measurements.

women (Figure 1.2.1a) and men (Figure 1.2.1b). Note that the level of survival probabilities is depressed because the HRS is very good at tracking respondents' dates of death even when they have not responded to previous waves. In this version of the data preparation, individuals who did not respond to a survey round would not enter the denominator of vital status.

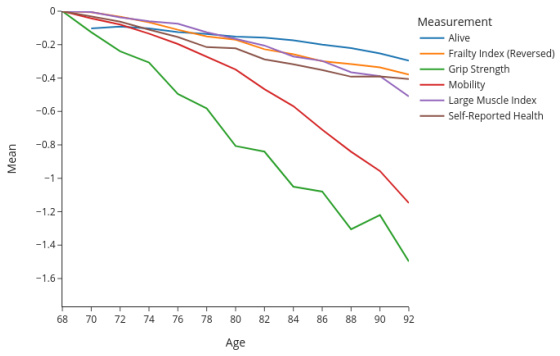
The second measurement shown in Figures 1.2.1a and 1.2.1b is a version of the **frailty index** used, for example, in Hosseini, Kopecky, and Zhao (2022). The frailty index is the unweighted sum of all recorded medical conditions a doctor has diagnosed in an individual. These conditions comprise high blood pressure, diabetes, cancer, lung disease, heart disease, stroke, psychiatric problems, and arthritis. We reverse it so that higher values indicate better health. The reversed frailty index declines by 0.4 (women) and 0.3 (men) original standard deviations until the end of the age range we consider. Note that this trend and all those we will subsequently discuss are conditional on survival. Due to the high predictive power of the frailty index for mortality—as noted by Hosseini, Kopecky, and Zhao (2022) and others—the effect of mortality selection is particularly large here. For individuals still alive at age 80, average frailty at age 68 is 0.39 among women and 0.34 among men. By including vital status among the health measures, our model below will take care of this to some extent, but it is important to keep in mind for the descriptive statistics.

Grip Strength measurements were introduced to the HRS survey in 2006 and consist of in-home physical tests of the hand grip strength, conducted twice for each hand. To obtain our variable of use, we average the four measurements. Our measure of grip strength is then the residual of a regression of average grip strength on individuals' height. We partial height out because of the high correlation between height and grip strength (Steiber, 2016) and we do not expect differences in grip strength associated with differences in height to be indicative of physical capacity. Among all measures pertaining to physical health, grip strength shows the steepest decline.

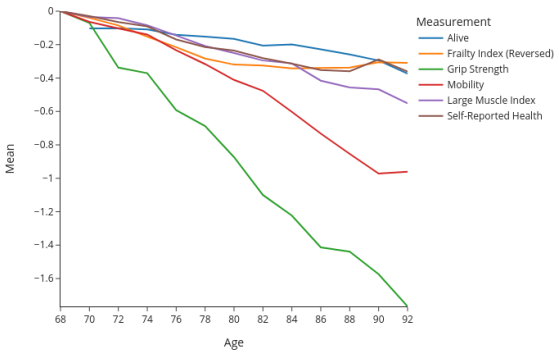
Mobility summarizes difficulties in performing the various activities of daily living: walking several blocks, walking one block, walking across the room, climbing several flights of stairs, and climbing one flight of stairs. As with the frailty index, we add up indicators for each measurement and reverse the scale so that higher values are associated with greater mobility. Mobility declines strongly in age. At the same time, its standard deviation rises as mobility impairments become more frequent over time.

Closely related, the **Large Muscle Index** summarizes difficulties in performing a number of activities associated with large muscles' strength. These activities are sitting for two hours, getting up from a chair, stooping ot kneeling or crouching, and pushing or pulling a large object. Again, we revert the order of the values to have a positive association between the variable and physical capacity.

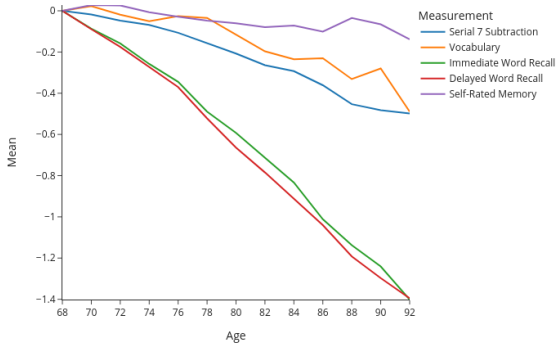
Finally, **Self-Reported Health** is a measure of health that is based on the respondent's self-assessed rating of their general health status. The values range from 1 (poor) to 5 (excellent). It probably is the most common health measure employed by economists as it provides an individuals' summary of her/his health in a single measure.



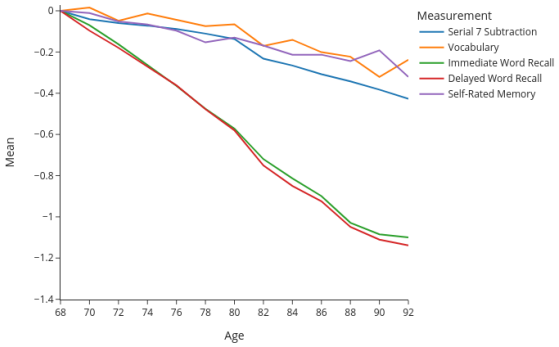
(a) Physical capacity, females



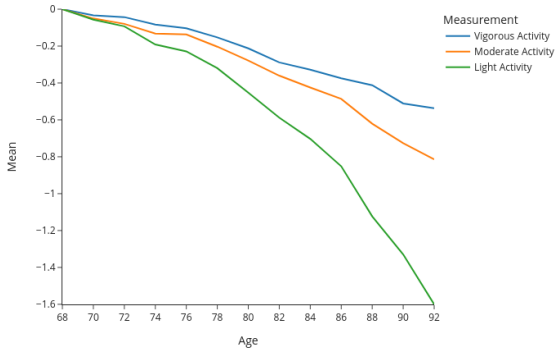
(b) Physical capacity, males



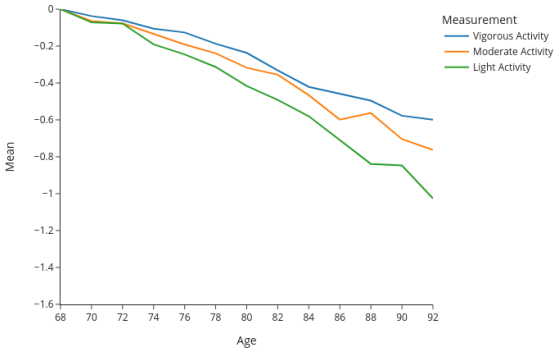
(c) Cognitive capacity, females



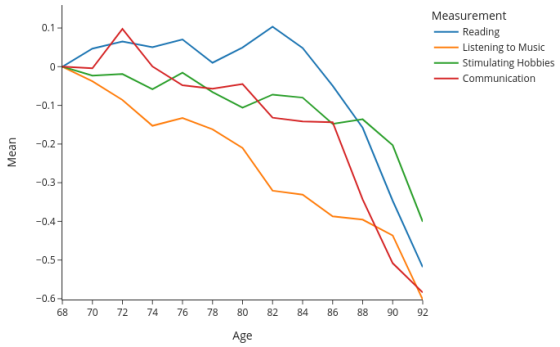
(d) Cognitive capacity, males



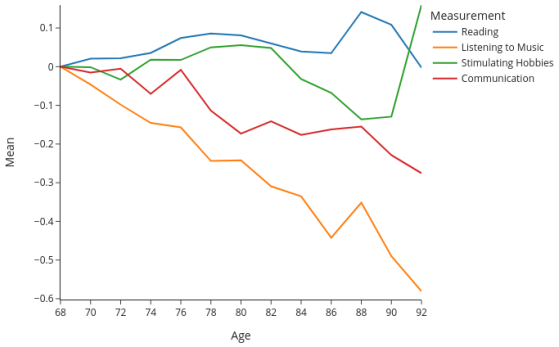
(e) Exercise, females



(f) Exercise, males



(g) Cognitive Stimulation, females



(h) Cognitive Stimulation, males

Figure 1.2.1. Average measurements by age

1.2.2 Cognitive Capacity

During interviews (in-person and via phone) HRS conducts a rich set of tests measuring respondents' cognitive capacity. For respondents that do not answer some of the cognitive test questions, HRS assumes non-random missing values and provides cross-wave imputation data in special data files (Health and Retirement Study, 2022a). Our measures of cognitive capacity are based on these cognitive tests and respondents' subjective ranking of their general memory status. In total, we employ five measures of cognitive capacity.

Serial 7 Subtraction is our first measure and is based on the test of serial sevens (SST) during which respondents are asked to subtract 7 from 100 and from continue subtracting 7 from each resulting number for a maximum of five times. The respondents are then assigned scores based on the total number of correct answers. In psycho-medical literature SST has widely been used to assess mental status of patients with dementia and been generally regarded as a measure of concentration (Karzmark, 2000). Figures 1.2.1c and 1.2.1d demonstrate a steady decline in concentration, as measured by the serial sevens, for both men and women, from our youngest age group to the oldest one being somewhat larger for women (0.5 units of original standard deviaton) than for men (0.4 units of original standard deviation).

Our second measure of cognitive capability is **Vocabulary** which is a test summarizing respondents' ability to provide correct definitions of words from a list of five words. One of two sets of words is assigned randomly at the first interview, and alternating sets are given during subsequent interviews. The two alternating sets of words are 1) repair, fabric, domestic, remorse, plagiarize; and 2) conceal, enormous, perimeter, compassion, audacious. We can see in Figures 1.2.1c and 1.2.1d that Vocabulary test has an age trend similar to that of the Serial 7 Subtraction, both in terms of absolute slopes and relative differences between men and women.

Immediate Word Recall is the third variable in Figures 1.2.1c and 1.2.1d and results from a test that asks the respondents to recall words (in any order) form a list of ten (later waves) or twenty (earlier waves) words, directly after being read the list. Examples of words included in a list are lake, car, army, etc. In in the initial wave, respondents were randomly assigned a list from the set of four lists and during the consequent four waves there were assigned a different list (McCammon, Fisher, Hassan, Faul, Rodgers, et al., 2022). **Delayed Word Recall** has the same structure as immediate word recall. In this task, respondents are asked to recall the same list of words once more, after spending several minutes on answering other survey questions. Word recall tests are widely used as measures of episodic memory frequently administered to patients with alzheimer's disease (see, e.g., Dixon and Frias, 2014; Runge, 2015).

Both of the word recall variables being measures of the same conceptual variable (episodic memory) perhaps explains the similar trends that they display. Of all the measurements of cognitive capacity, word recall variables have the sharpest decline over the age span in our model, and as with other measurements, the decline is larger for women than for men, with the caveat that our data are conditional on survival.

Finally, **Self-Rated Memory**, is our last measure of cognitive capacity and is based on respondents' self-assessed rating of their general memory status. The values range from 1 (poor)

to 5 (excellent). Self-Rated Memory displays a moderate decline in both genders, which has a somewhat more pronounced trend among men.

1.2.3 Exercise and Cognitive Stimulation

We use **Vigorous, Moderate and Light Activities** as measures for investment in physical health. Each of these survey questions asks respondents how often they do vigorous (running, jogging, cycling, etc.), moderate (gardening, cleaning the car, walking at moderate pace, dancing, stretching) and light/mildly energetic (vacuuming, laundry, home repair), respectively. Up until the sixth wave (year 2002) respondents were only asked if they do vigorous activities at least three times a week. Starting from wave seven, this questionnaire item was replaced by the three activity questions that we use in our study. Figures 1.2.1e and 1.2.1f show that with age people do less of all types of physical activities, with largely similar trends for men and women.

To obtain measures for cognitive stimulation, we utilized the CAMS survey which allowed us to construct measures of time respondents spend on different cognitively stimulating activities. Among these, our first measurement of cognitive stimulation is **Reading** that counts weekly hours spent on reading books, newspapers, or magazines. The association between reading and cognitive decline has been studied in psycho-medical literature, and reading has been found to be positively associated with hampered cognitive decline (Chang, Wu, and Hsiung, 2021). In Figures 1.2.1g and 1.2.1h we see that Reading has declining trend among women and is rather stable among men.

The second variable in Figures 1.2.1g and 1.2.1h is **Listening to Music**, and it meausres how many hours weekly respondents listen to music. The effects of music listening on cognitive functioning of at-risk patients have been studied in psycho-medical literature, and listening to music has been found to be beneficial for cognitive functioning (see, e.g., Särkämö, Tervaniemi, Laitinen, Forsblom, Soinila, et al., 2008; Särkämö and Soto, 2012). As with most measurements of cognitive stimulation, we observe a declining age trend for Listening to Music both among men and women.

Our last variables for cogntivie stimulation are **Stimulating Hobbies** and **Communication** which summarize how many hours respondents spend weekly on various hobbies that may be expected to stimulate cognition, and the weekly hours spent on interacting with others, respectively. Stimulating Hobbies aggregates the survey variables that ask how many hours respondents spend on: 1) playing cards or solving jigsaw puzzles, 2) singing or playing instruments, 3) doing arts and crafts, and 4) going to movies or lectures. We construct the Communication variable as the sum of hours spent on visiting with others in person and communication via letters/phone/email. Looking at Figures 1.2.1g and 1.2.1h, Communication has similarly declining tred among men and women, whereas Stimulating Hobbies has a steeper slope for women and than for men.

1.2.4 Raw Correlations in the Data

Figures 1.2.2 and 1.2.3 show correlation matrices for women and men, respectively. Each figure contains two panels. The upper panels show within-period correlations until age 79, the lower

panels do the same ages 80 and above. We show the lower triangular part of the correlation matrix. We leave out the indicator for being alive because we only measure the other variables whenever it is one. In addition to showing the numbers, we color the matrix' elements such that a correlation of 1 is dark red, 0 is white, and −1 is dark blue. Scaling is linear on both sides of the origin. Variables are ordered by factor, which we include in the label of the first measure pertaining to it. The measures in the first five rows and columns—from the reversed frailty index until self-reported health—load on physical capacity. The subsequent block of five rows and columns load on cognitive capacity. In the lower part of the matrix, exercise and cognitive stimulation load on three and four measures, respectively.

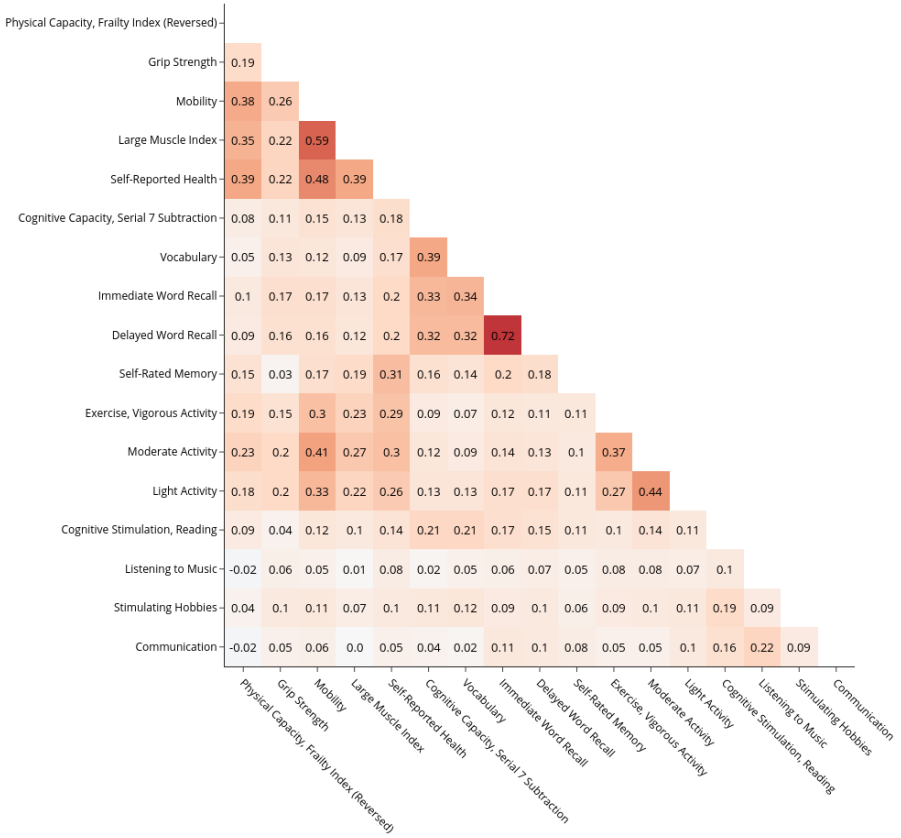
Several patterns are visually apparent in all four correlation matrices. First of all, the blocks of measures pertaining to each factor are clearly visible as having substantial cross-correlation throughout. For example, the first four entries in the first columns are the correlations of the reversed frailty index with the other measures loading on physical capacity. Across all four panels, correlations are at least 0.3 with the exception of the correlation of reversed frailty and grip strength, which is at least 0.1 throughout.

Similarly, the triangle with correlations for measurements pertaining to cognitive capacity—with the three corners (Serial 7 Subtraction, Vocabulary), (Serial 7 Subtraction, Self-Rated Memory), and (Delayed Word Recall, Self-Rated Memory)—has distinctly dark colors throughout. Unsurprisingly, correlations are particularly large between the two word recall tasks. The three correlations between the various types of physical activity are high throughout. The six elements to the bottom right to the matrix contain the correlations among the measures loading on cognitive stimulation. Among all factors, these have the weakest within-factor correlations with values ranging from 0.09 to 0.25. This is not very surprising as the variables do cover a much wider range of activities than, say, the various activity levels that load on exercising.

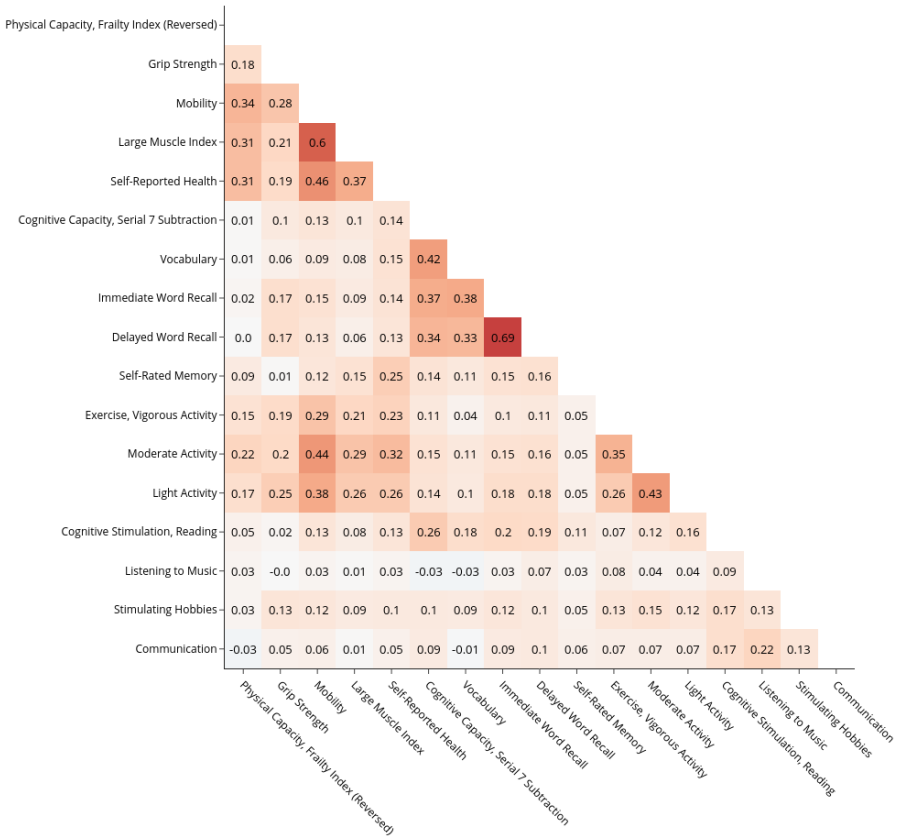
A second salient feature is that almost all elements are positive. This implies that it is important to model physical and cognitive capacity jointly with each other and with the two types of investments. This being written, there are clear level differences. Maybe unsurprisingly, the largest correlations are between measures of exercise and those of physical capacity. Most measures of cognitive capacity are substantially and positively related to variables measuring physical capacity and exercise, respectively. The correlation patterns are somewhat more mixed when it comes to cognitive stimulation and the other three factors.

This is related to our third broad observation: While the general patterns noted so far hold up across age groups and genders, there are some important differences. For example, the correlations of grip strength with other health measures are higher among women than among men, particularly at higher ages. Correlation patterns of individual measures pertaining to cognitive stimulation and cognitive capacity are quite distinct among men and women, particularly at older ages. For example, among individuals aged 80 and above, reading and serial 7 subtraction have a correlation of 0.27 among women whereas it is 0.18 among men. Among women in this age group, listening to music is slightly negatively correlated with serial 7 subtraction and vocabulary scores. For men, the same correlations are small and positive.

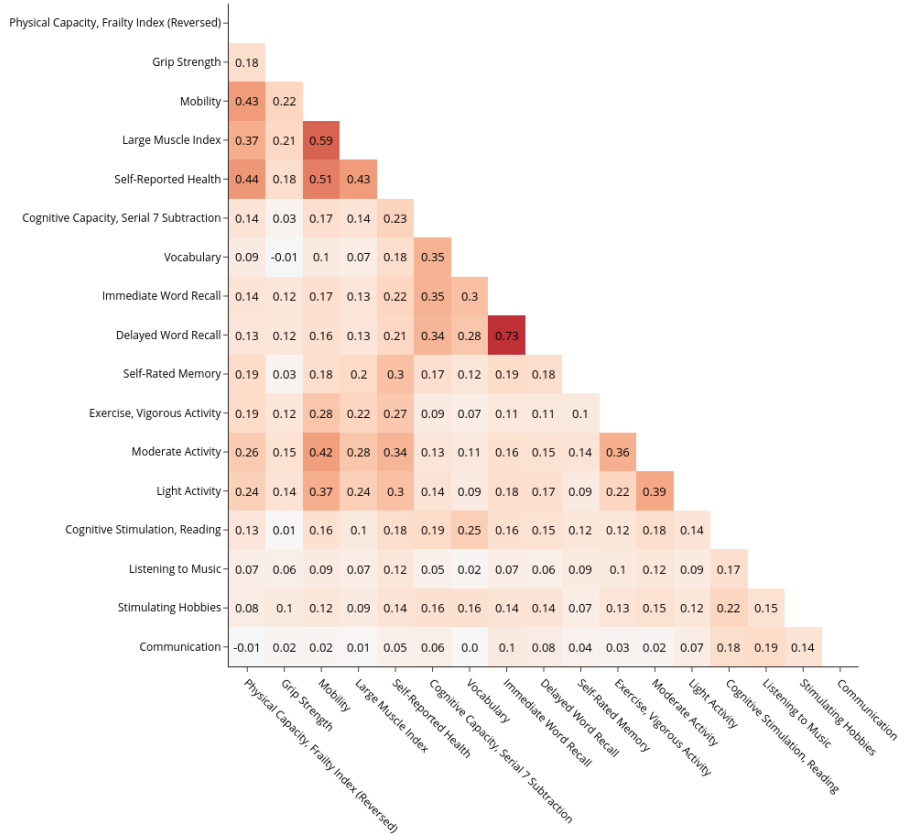
While these patterns are informative, the $2 \times 2 \times 153$ numbers in Figures 1.2.2 and 1.2.3 are clearly too many to make sense of directly – and the matrices already reduce the 13 periods



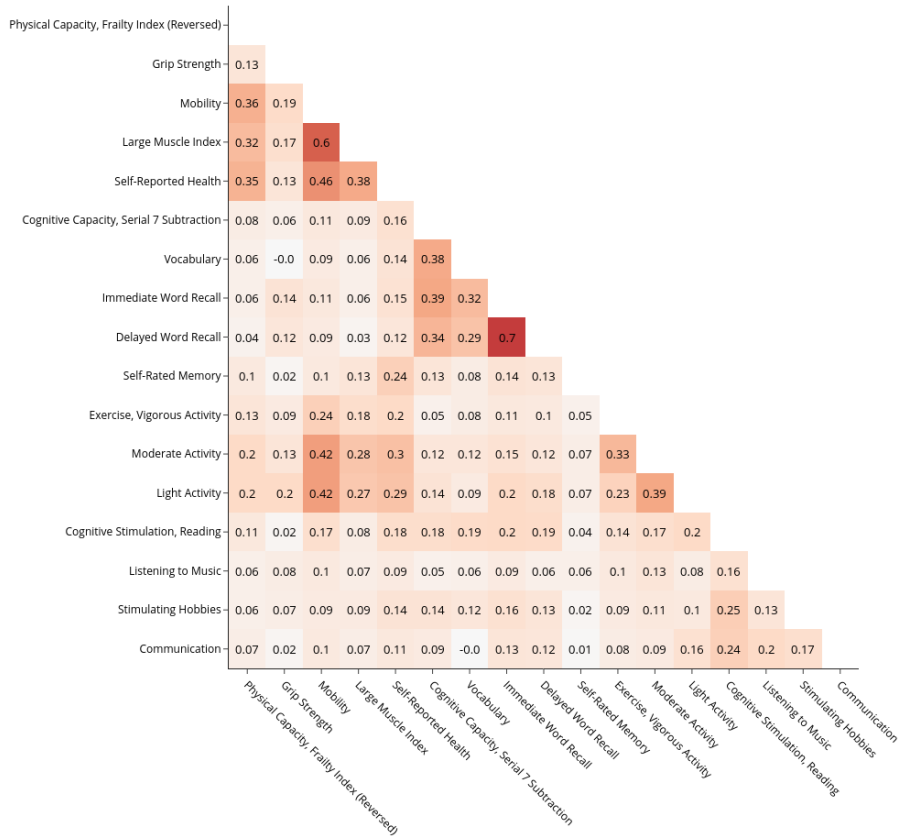
(a) Aged below 80



(b) Aged 80 and above



(a) Aged below 80



(b) Aged 80 and above

we observe in our data to 2. In the next section, we outline a framework that constructs latent variables for our four factors and which allows us to interpret their joint evolution.

1.2.5 Example Transitions

Before going to the formal model, we show a few exemplary trajectories of physical and cognitive capacity. Figure 1.2.4 shows the trajectories of 500 randomly sampled individuals from our dataset. Dots at the end of a trajectory mean that that person died in the next period. Trajectories that do not end in a dot are from individuals whose death was not observed, either because they dropped out of the sample or are still alive in the last wave.

The highlighted lines are hand-picked examples of individuals that had a physical capacity close to the 90th percentile, but very different trajectories afterwards. The blue line shows a person that had a strong decline in physical capacity over two periods and then passed away. The yellow line shows a person with a very volatile trajectory in both physical and cognitive capacity. The red line shows an individual who had a bad health shock at some point but recovered and enjoyed a high level of physical capacity for many years. The right panel of the figure shows the cognitive capacity of the same individuals. All three lines show fluctuations around a robust declining trend.

The plot illustrates that vastly different trajectories of physical and cognitive capacities are possible even for people with similar starting conditions in terms of physical capacity. Answering the question whether such differences can be explained or are the product of random shocks requires ar more rigorous approach.

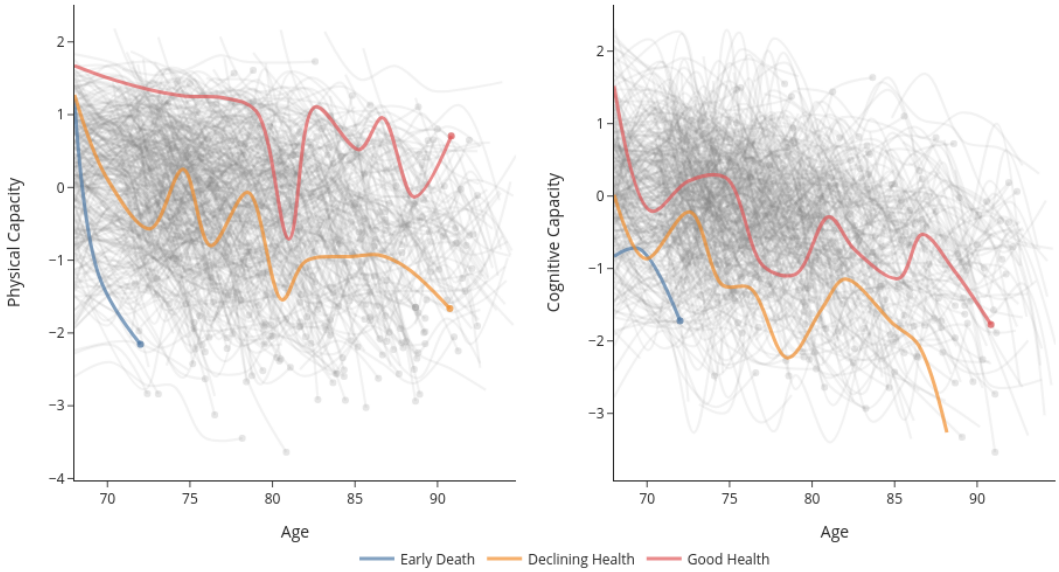


Figure 1.2.4. Trajectories for decline of health and cognitive capacity

1.3 Model

1.3.1 The Technology of Aging

Analyzing the joint evolution of physical and cognitive capacity and the effect physical exercise and cognitive stimulation have on both poses many econometric challenges.

1. As discussed in the previous section, there are many potential observed variables to measure each concept we analyze. In order to make the results interpretable, their dimensionality has to be reduced.
2. All observed variables are subject to measurement error, which is potentially large in many cases.
3. Physical and cognitive capacity, exercise, and cognitive stimulation are dynamically intertwined in the sense that each of them has a potential effect on all others. For example, exercise should improve physical capacity. Conversely, it may well be that the cost of exercise might be higher at low levels of physical capacity because physiotherapy is less enjoyable than a walk in nature.
4. The relationships between variables might change over time.

The Technology of Skill Formation (Cunha and Heckman, 2007; Cunha, Heckman, and Schennach, 2010) is an econometric framework that emerged to deal with very similar challenges in the

context of skill formation during childhood. It distinguishes observed variables—for example an IQ test—from latent factors such as cognitive and non-cognitive skills. The technology is the law of motion of latent factors over multiple discrete time periods. Observed variables are stochastic functions of one or more latent factors. In addition to the latent factors of interest, the framework allows for observed or latent investments such as parental investments in skills or schooling.

To account for the multitude of potential effects, each latent factor may depend on lagged values of itself and all other latent factors. The law of motion of the latent factors is usually nonlinear. This is necessary to allow for different productivity of investments at different levels of skills. Moreover, it allows for dynamic complementarity, i.e., the fact that earlier investments may increase the productivity of later investments (Cunha and Heckman, 2007).

The Technology of Skill Formation maps perfectly on our setting. Instead of cognitive and non-cognitive skills, our Technology of Aging models physical and cognitive capacity. Instead of parental investments, we have exercise and cognitive stimulation. While we separate investments into a physical and cognitive component, we allow each investment factor to influence both latent capacities.

Transition Functions

We assume the following law of motion of our latent factors:

$$
\begin{aligned}
x_{1,t+1} &= \beta_{1,t} + \sum_{i=1}^{4} \gamma_{1,t,i} x_{t,i} + \sum_{i=1}^{4} \sum_{j=1}^{i} \delta_{1,t,i,j} x_{t,i} x_{j,t} + \eta_{1,t} \\
x_{2,t+1} &= \beta_{2,t} + \sum_{i=1}^{4} \gamma_{2,t,i} x_{t,i} + \sum_{i=1}^{4} \sum_{j=1}^{i} \delta_{2,t,i,j} x_{t,i} x_{j,t} + \eta_{2,t} \\
x_{3,t+1} &= \beta_{3,t} + \sum_{i \in \{1,2,3\}} \gamma_{3,t,i} x_{t,i} + \eta_{3,t} \\
x_{4,t+1} &= \beta_{4,t} + \sum_{i \in \{1,2,4\}} \gamma_{4,t,i} x_{t,i} + \eta_{4,t}
\end{aligned}
\tag{1.3.1}
$$

Where x_1, x_2, x_3, and x_4 are physical capacity, cognitive capacity, exercise, and cognitive stimulation, respectively. β, γ and δ denote the technology parameters to be estimated. η denotes a stochastic shock.

The first two equations in (1.3.1) mean that physical and cognitive capacity follow a flexible functional form containing all lagged factors, their squares, and their interaction terms. This is known as the translog function in the skill formation literature (because skills are typically assumed to be measured in logs, not levels) and has been used by, for example, Agostinelli and Wiswall (2016a). The translog function allows for dynamic complementarity but does not assume it. While this functional form is not a standard economic production function, we interpret it as a flexible approximation to an arbitrary underlying production function in the spirit of a nonparametric series estimator.

The bottom two equations in (1.3.1) relate to exercise and cognitive stimulations, respectively. Both investment factors are assumed to depend on their own lagged values along with the lagged values of physical and cognitive capacity.

Measurement System

We assume the measurement equations to be linear with an additively separable and normally distributed error term. All of them thus have the following form:

$$y_{\ell,t} = \alpha_{\ell,t} + \sum_{i=1}^{4} h_{\ell,t,i} x_{t,i} + \epsilon_{\ell,t} \tag{1.3.2}$$

where $y_{\ell,t}$ denotes the ℓ^{th} measurement in period t, α is the intercept of the measurement equation and h are factor loadings. In the empirical application we only have measurements that load on just one factor, so that for all measurements, three out of the potentially four loadings $h_{\ell,t}$ are zero by construction. Subject to identification requirements outlined in Cunha, Heckman, and Schennach (2010), this could easily be relaxed.

In typical applications of the Technology of Skill Formation, the number and type of available measurement variables varies strongly across periods. This is because any test score that is applicable to very young children would not work for older children. In our case, the measurements stay the same across periods and most of them can be assumed to be time-invariant, i.e. to have the same loading, intercept, and standard deviation of measurement error in each period.

1.3.2 Identification and Interpretation of Parameters

The econometric model implied by the Technology of Skill Formation is a Structural Equation Model or dynamic latent factor model. Linear Structural equation models are widely used since the 1970ies to study relationships between latent and observable variables. However, standard identification results and software for Structural Equation Models are not applicable to our setting because they usually require linearity assumptions or put restrictions on the connectedness of the underlying causal graph, which go beyond those encoded in our system (1.3.1).

Cunha, Heckman, and Schennach (2010) provide general nonparametric identification results for nonlinear dynamic latent factor models. The exact conditions for identification depend on the assumptions one is willing to put on the nature of measurement error. Typically, having at least two dedicated continuous measurements for each latent factor in each period is sufficient to identify an arbitrary production function under mild conditions. Doing so requires normalizations of location and scale in each period because latent factors do not have a natural unit of measurement.

A subsequent literature (Agostinelli and Wiswall, 2016b; Freyberger, 2021) has shown that much fewer normalizations are required when empirical applications assume the popular constant-elasticity-of-substitution (CES) form, which implies restrictions on the location and scale of its outputs (see Appendix 1.C.1 for details). Our specification of the production function (1.3.1) does not impose any such restrictions. However, as discussed previously, we have at least one age

invariant measurement for each latent factor. We always use such measurements for normalizations, which pin down the location and scale of each corresponding factor in all periods.

The lack of natural units for the latent factors and the requirement for normalizations also poses challenges for the interpretation of the results. In short: any outcome that depends on transformations of measurements outside of the model, the choice of the measurement being normalized, or the values of the normalized parameters cannot be interpreted without further information. For details and a more formal definition see Freyberger (2021).

In practical applications, different ways of dealing with this have emerged. Cunha and Heckman (2008) and Cunha, Heckman, and Schennach (2010) propose to anchor the latent factors in terms of observable cardinal variables. For example, they anchor cognitive and non-cognitive skills in terms of years of schooling, wages or the probability to commit a criminal offense. For each anchoring outcome, they re-estimate the model to obtain estimated production function parameters in terms of anchored factors. Attanasio, Meghir, and Nix (2020) do not have access to adult outcomes. Instead they communicate the variables that were normalized and state that results have to be interpreted with respect to the normalizations. Del Bono, Kinsler, and Pavan (2020) propose to simply standardize the variance of the latent factors in logs. This allows for statements such as increasing investment by 1 % increases skills by x %. While this is invariant to any normalization of location and scale in the measurement system, the approach is only valid if one defines that skills are measured in logs not levels. Due to the ordinality of skills, this is a valid but arbitrary definition and thus the approach falls short of its goal to be completely objective. Freyberger (2021) proposes to translate inputs and outputs of the production functions into ranks. This is invariant to any normalization of location and scale, assumptions on whether latent factors are measured in levels or logs and transformations of the measurements outside of the model.

We acknowledge that there is no single natural scale for latent factors and thus see value in all of the above approaches. For example, translating everything to ranks is a natural way of solving a problem that is caused by ordinality. Moreover, it makes the results completely invariant to many decisions made by the econometrician. However, it might not be as interpretable as anchoring approaches. For example, it destroys any time trend that was present in the measurements. To address the shortcomings of any single method, we thus use a combination of all of them.

We standardize age invariant measures with respect to their mean and standard deviation at age 68. We estimate the parameters of the production function, normalizing one age-invariant measure for each factor in period zero. The normalized measures are the reversed Frailty Index, Serial 7 Subtraction, Moderate Activity, and Reading. This preserves the time trend in the measurement variables and means that our estimated parameters and the time trend can roughly be interpreted in terms of standard deviations at age 68. For reference, we also show the marginal distributions of each latent factor and the joint distributions of each factor pair at multiple ages (see 1.D.3).

1.3.3 Estimation

Multiple estimators for nonlinear dynamic latent factor models are available. Agostinelli and Wiswall (2016a) estimate the fist period factor loadings from ratios of covariances between measurements. To estimate production function parameters, they subsequently employ an iterative IV approach. Their method is very tractable; it comes at the cost of statistical efficiency. Our own experiments on simulated data suggest that it works well for models with few periods but becomes imprecise if there are ten or more periods, especially when the correlation between latent factors is high.

Attanasio, Cunha, and Jervis (2019) use linear regression on Bartlett factor scores with a correction approach. This estimator is computationally very attractive. However, it does not deal well with missing observations. Several of our variables are not contained in the core HRS questionnair; they are available for subsets of individuals at different points in time. Because of this, the estimator of Attanasio, Cunha, and Jervis (2019) is unsuitable for our application.

Attanasio, Meghir, and Nix (2020) first estimate the distribution of the latent factors as a mixture of normal distributions and then estimate the parameters of the production functions on a simulated sample from that distribution. This approach is computationally harder than the two previous ones but simpler than the maximum likelihood estimator by Cunha, Heckman, and Schennach (2010). The required assumptions are the same as for the likelihood estimator.

Cunha, Heckman, and Schennach (2010) use a maximum likelihood estimator. For computational tractability, they use nonlinear Kalman Filters to factorize the likelihood function into a product of conditional likelihoods. This estimator is computationally more difficult than the others. In its original formulation, numerical stability is often compromised. However, the estimator is statistically efficient and it can deal well with observations that are missing at random.

We derive a mathematically equivalent but numerically stable version of the likelihood estimator used by Cunha, Heckman, and Schennach (2010). Our version replaces standard filters by square-root Kalman filters (Prvan and Osborne, 1988; van der Merwe and Wan, 2001), which are numerically more robust. The computational cost is similar to the original approach. The details of the original and the reformulated estimator as well as the exact assumptions required for estimation are described in Appendix 1.B.

To account for mortality, we add a dummy variable for being alive as an additional measurement of physical capacity. This is analogous to a linear probability model of survival. Thus, the estimated health state of survivors is adjusted upwards, while the health state of everyone who has passed away is adjusted downwards compared to a state estimation that ignores mortality. In future work we plan to replace the linear probability model of mortality by a Probit model.

A flexible implementation of the new estimator can be found in the Python package skillmodels (Gabler, 2022). It uses JAX (Bradbury, Frostig, Hawkins, Johnson, Leary, et al., 2018) for just in time compilation and automatic differentiation. This reduces the computational cost drastically. We use estimagic (Gabler, Raabe, Röhrl, and Gaudecker, 2022) for numerical optimization and the calculation of standard errors. To generate good start values for the optimization, we first decompose the model into four single factor model with much fewer free parameters. In a second step we estimate a linear model. In the third step we estimate the full nonlinear model. We

use pytask (Raabe, 2020) and the Templates for Reproducible Research Projects in Economics (Gaudecker, 2019) to automate our research project and to parallelize many tasks. The full estimation takes approximately four hours on a laptop.

1.4 Results

We present our results in three stages. First, we describe the measurement system. Next, we describe broad patterns for the transition equations. Finally, we dig deeper into the dynamic effects of changing factors along their distribution.

1.4.1 Measurement System

Table 1.4.1 shows exemplary parameter estimates of the measurement system. The first panel shows the parameters that we constrain to be time-invariant. The three panels below display time-varying parameters of the system at ages 70, 80, and 90. We show loadings and standard deviations for women and men, respectively. Tables 1.D.1–1.D.8 in Appendix 1.D.1 show the complete set of parameter estimates, including the intercepts. Remember from Section 1.2 that we scale all measures—except for dummy measuring vital status, which retains its natural form—to have mean zero and unit variance in the initial period.

For the measurements loading on physical capacity, we normalize the reversed frailty index to have intercept zero and unit loading. We also restrict the parameters relating to mobility, the large muscle index, and self-reported health to be time-invariant – all of these have fairly similar time trends as seen in Figure 1.2.1 (note that mobility has a steeper trend than the others, but making the measurement system time-varying did not change results). All four measurement have similar factor loadings in the 0.93–1.33 range and the standard deviation in their measurement errors is very similar, too (0.75–0.8). The correlations between these four measurements are high throughout in the 0.6-0.85 range (see the correlation matrices in Section 1.D.2 of the Appendix).

We leave the measurement systems for vital status and grip strength unrestricted across age groups. The standard deviation of measurement error in grip strength decreases over time; the loadings decrease for females and stay roughly constant for males. In sum, this means that the correlation between grip strength and the latent factor representing physical capacity stays constant with age for women at 0.3 and increases for men from 0.35 to 0.45. The loading on vital status increases for both genders. Due to the fact that the dummy for being alive has its natural scale, the coefficient has a meaningful interpretation in terms of survival probabilities. At age 70, the interquartile range of physical capacity is 0.95 for women and 0.78 for men (see Appendix Section 1.D.3). Changing physical capacity from its first to its third quartile thus increases the probability of survival by $0.95 \times 4.2\% = 4\%$ for women and $0.78 \times 5.8\% = 4.5\%$ for men. At age 80, the interquartile ranges are just below 1 and the loadings of 0.09 for both genders directly measure changes in survival chances as one moves across the outer quartiles. The same is true at age 90 for men ($\Delta_{\text{survival}} = 0.2$), for women the distribution is less dispersed at that age and an interquartile range of 0.8 implies a increase in survival probabilities of 11%. This is in line with

Table 1.4.1. Loadings and Measurement Standard Deviations

			Female		Male	
			Loading	Meas. Std.	Loading	Meas. Std.
Age	Factor	Measurement				
All	Physical Capacity	Frailty Index (Reversed)	1.000	0.707***	1.000	0.796***
				(0.001)		(0.002)
		Mobility	1.228***	0.766***	1.331***	0.750***
			(0.005)	(0.003)	(0.007)	(0.003)
		Large Muscle Index	0.929***	0.750***	1.032***	0.761***
			(0.005)	(0.002)	(0.006)	(0.003)
		Self-Reported Health	0.950***	0.765***	0.963***	0.793***
			(0.004)	(0.002)	(0.006)	(0.003)
	Cognitive Capacity	Serial 7 Subtraction	1.000	0.890***	1.000	0.907***
				(0.003)		(0.004)
		Vocabulary	0.839***	0.923***	0.960***	0.900***
			(0.013)	(0.004)	(0.016)	(0.004)
		Immediate Word Recall	1.801***	0.583***	1.684***	0.599***
			(0.015)	(0.003)	(0.016)	(0.003)
		Delayed Word Recall	1.805***	0.595***	1.648***	0.605***
			(0.014)	(0.002)	(0.015)	(0.003)
	Exercise	Vigorous Activity	0.682***	0.809***	0.741***	0.814***
			(0.010)	(0.004)	(0.012)	(0.005)
		Moderate Activity	1.000	0.794***	1.000	0.816***
				(0.004)		(0.004)
		Light Activity	1.076***	0.933***	0.927***	0.861***
			(0.012)	(0.004)	(0.013)	(0.004)
	Cognitive Stimulation	Reading	1.000	0.780***	1.000	0.683***
				(0.006)		(0.007)
		Listening to Music	0.512***	0.980***	0.229***	1.004***
			(0.010)	(0.006)	(0.010)	(0.007)
		Stimulating Hobbies	0.578***	0.925***	0.375***	0.969***
			(0.011)	(0.005)	(0.012)	(0.005)
		Communication	0.523***	0.999***	0.325***	0.989***
			(0.010)	(0.005)	(0.011)	(0.006)
70	Physical Capacity	Alive	0.042***	0.303***	0.058***	0.303***
			(0.011)	(0.039)	(0.013)	(0.035)
		Grip Strength	0.489***	0.933***	0.578***	0.978***
			(0.042)	(0.015)	(0.053)	(0.020)
	Cognitive Capacity	Self-Rated Memory	0.576***	0.961***	0.626***	0.937***
			(0.031)	(0.009)	(0.035)	(0.011)
80	Physical Capacity	Alive	0.091***	0.353***	0.089***	0.367***
			(0.023)	(0.047)	(0.031)	(0.068)
		Grip Strength	0.367***	0.882***	0.571***	0.891***
			(0.052)	(0.021)	(0.061)	(0.023)
	Cognitive Capacity	Self-Rated Memory	0.470***	1.013***	0.589***	0.988***
			(0.038)	(0.012)	(0.048)	(0.015)
90	Physical Capacity	Alive	0.137*	0.425***	0.204*	0.430***
			(0.081)	(0.133)	(0.121)	(0.128)
		Grip Strength	0.357***	0.736***	0.508***	0.766***
			(0.099)	(0.032)	(0.120)	(0.055)
	Cognitive Capacity	Self-Rated Memory	0.459***	1.081***	0.386***	1.080***
			(0.097)	(0.026)	(0.120)	(0.038)

Note: ***p<0.01;**p<0.05;*p<0.1

the intuition that physical capacity is more predictive of death at older ages, as deterioration of overall health becomes a more important cause of death than fairly sudden shocks such as cancer or heart attacks (Gill, Gahbauer, Han, and Allore, 2010).

For measures pertaining to cognitive capacity, we normalize the results from the serial 7 subtraction task to have intercept zero and unit loading. This measure along with the vocabulary score and the two word recall tasks are restricted to have the same factor loading and measurement error variance across all ages. Serial 7 subtraction and the vocabulary score look very similar in terms of loading and measurement error. For the word recall tasks, loadings are substantially higher and measurement errors are lower than this. Consequently, all correlations between these measures and the cognitive capacity factor are high throughout – around 0.5 for serial 7 subtraction and vocabulary; exceeding 0.8 for the word recall tasks. The measurement system of self-rated memory is allowed to vary with age. For both genders, its loading is estimated to be about 0.6 initially and decreases over time. The standard deviation of measurement error is around unity, with a slightly increasing trend. Consequently, the correlation of self-rated memory with cognitive capacity us declining with age which is consistent with Huang and Maurer (2019)

Given the similarity of our measurements for exercise, it is unsurprising that all three of them load substantially on the underlying factor. Moderate activity—the normalized measurement—has the largest correlation with the exercise factor at all ages. The correlation of vigorous activity and exercise declines over time whereas light activity goes the other direction. Both of these trends are more pronounced among women than among men.

Among the measurements loading on cognitive stimulation, we normalize the parameters on the time spent reading. This is also the dominant one among the four measurements with a standard deviation of its error around 0.78 (women) and 0.68 (men) and correlations with the factor exceeding 0.7 throughout. The errors on the other three measurements are between 0.9 and 1; their loadings are estimated to be around 0.5 for women and 0.3 for men. For women, these coefficients translates into correlations with the cognitive stimulation factor of around 0.4, which are roughly stable over time. Among men, they start from a level around 0.2-0.3. While communication activities maintain a constant correlation with the factor, listening to music or pursuing stimulating hobbies have hardly any relation left with it by age 90.

In sum, the measurements show a high correlation with the factors they are supposed to identify. For many measurements, it is sensible to restrict the model parameters to be time invariant and we do so. Measurements that are allowed to be changing with age vary in a way that makes sense in the light of prior literature. Differences between genders are not dramatic, but large enough to command separate estimation. Having established these direct relations to the data, we now turn to the core contribution of our paper: The joint evolution of physical and cognitive capacity and the impact of exercise and cognitive stimulation.

1.4.2 Transition Equations

The translog production functions for physical and cognitive capacity have many parameters. In total, we have 15 coefficients per factor, which needs to be multiplied with four age groups (or“stages” in the terminonlogy of Cunha, Heckman, and Schennach, 2010) and two genders.

Furthermore, the parameters do not have intuitive interpretations without referring to precise values of the four factors in our model. We thus refrain from listing the parameters in the main text and relegate them to Tables 1.D.9–1.D.16 in Appendix 1.D.4. We note that the vast majority of parameters is very precisely estimated. The set of model parameters is completed with the initial distribution of states and the standard deviation of period-by-period innovations, which we relegate to Appendix 1.D.5.

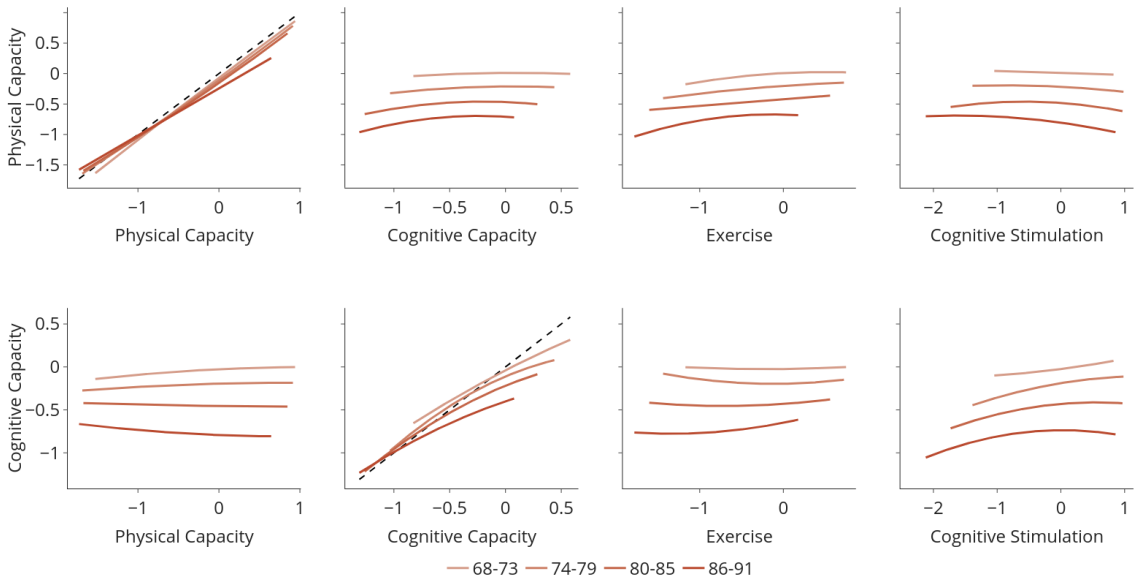
As a first pass, Figure 1.4.1 shows transition equations for physical capacity (first row of each subfigure referring to women and men, respectively) and cognitive capacity as a function of the input factors. Each of the sixteen panels contains four lines, one for each age group or stage. Input factors are kept at their median except for the one on the x-axis, which is varied from the 1^{st} to the 99^{th} percentile of its distribution in the respective age group.

The top left panel in Figure 1.4.1a thus shows the result of the following thought experiment: Conditional on current age, what is a woman's expected value of physical capacity in two years as a function of her current physical capacity while fixing cognitive capacity, exercise, and cognitive stimulation at their median values. The results show that there is a high degree of persistence in all age groups. For the upper part of the distribution of physical capacity, the lines are below the 45°-line (the distributions at ages 70, 80, and 90 are shown in in Appendix 1.D.3, Figures 1.D.7–1.D.9; as a rough guide to interpret the first panel of Figure 1.4.1a, the first quartile at age 90 has a value of −1.2). The transition function is below 45°-line everywhere in the youngest age group, which has the steepest slope throughout. This means that at median levels of cognitive capacity, exercise, and cognitive stimulation, physical capacity will unambiguously decline in expectation regardless of the initial level. In contrast, for very low values of physical capacity at older ages, there would be some mean reversion – if all other factors were at their median. Of course, cognitive capacity is not a (direct) choice and there might be substantial costs to reaching median levels of exercise or cognitive stimulation if physical capacity is very low, for example.

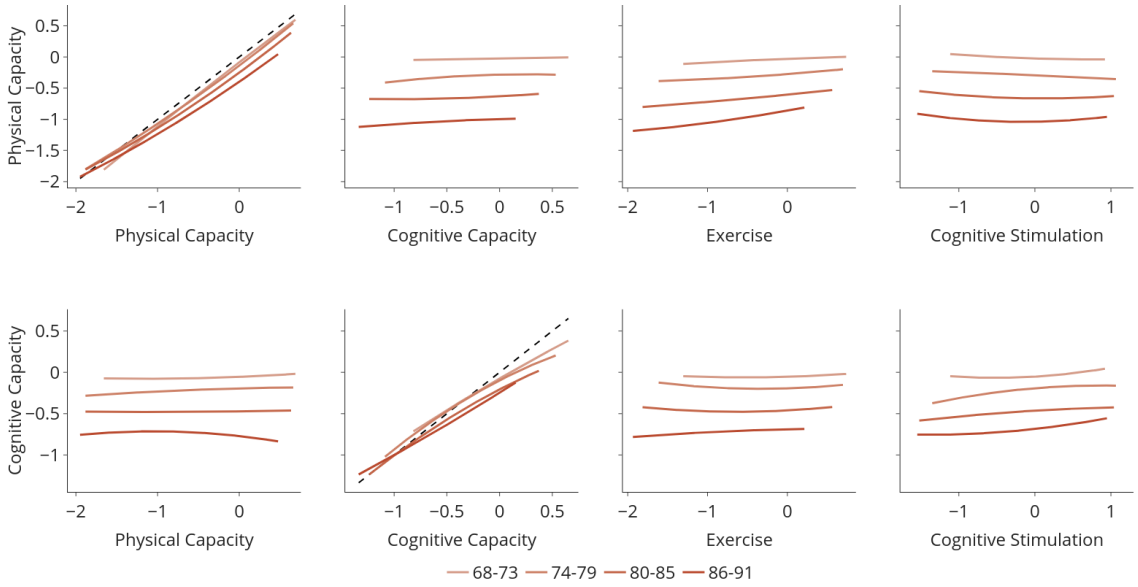
Increased cognitive capacity is associated with a slightly more favorable evolution of physical capacity. For example, changing cognitive capacity from its first quartile (−0.63) to its third quartile (−0.17) at age 80 is associated with an increase of age-82 physical capacity of 0.02 units or just under 2 percentiles. The corresponding effects of increased exercise are positive as well and tend to be larger. The same interquartile move for exercise at age 80 (from −0.81 to −0.06) leads to an increase of physical capacity by 0.16 units, which corresponds to almost 5 percentiles. The effects of cognitive stimulation on the dynamics of physical capacity are often slightly negative at median levels of physical capacity, cognitive capacity, and exercise.

The second row of Figure 1.4.1a shows the corresponding effects for the evolution of cognitive capacity. We start with the second panel, which contains the own-effects. They are much less persistent than the own-effects for physical capacity as evident by the flatter slopes at all ages. The four lines are also further apart except at the very bottom of the distribution of cognitive capacity. This means that at almost any level of cognitive capacity, the dynamics are worse for higher ages, provided all other factors are at their median.

The first panel in the second row of Figure 1.4.1a displays modestly positive effects of physical capacity in the lower age groups; these become zero for higher ages and, in the highest age group, turn out to be negative at very low levels of physical capacity. Exercise has mostly positive effects



(a) Transitions, females



(b) Transitions, males

Figure 1.4.1. Next period states as a function of current states, other factors evaluated at the median

on the evolution of cognitive capacity at median values of other states with an exception being in the lower half of the exercise distribution during women's upper seventies. Finally, cognitive stimulation has positive effects almost everywhere. Note that the lines are visually misleading to some extent because of the long left tail of cognitive stimulation. For example, the first quartile

at age 80 is −0.66, the slope is steepest to the left of it. Moving cognitive stimulation to its third quartile at 0.14 has hardly any effect.

Figure 1.4.1b shows the same set of transition functions for men. Again, the broad patterns are fairly similar to women, but there are some important differences. For example, physical capacity is deteriorating more quickly for ages 74 and beyond across the entire distribution of current physical capacity; only at the very bottom of the distribution there is some sign of mean reversion. For the own-effects of physical capacity, there is a similar pattern to what we noted for the own effects of cognitive capacity among women: At almost any level of physical capacity, the dynamics are worse for higher ages, provided all other factors are at their median. In contrast, for cognitive capacity, the same effect is somewhat less pronounced then for women; the lower two age group and the upper two age groups look much more similar to each other there. The signs and magnitudes we noted for the off-diagonal elements generally hold up, although some curvatures appear markedly different. These mostly concern the tails of the distributions, however.

1.4.3 Dynamic Effects Over Several Periods

A major benefit of our dynamic model over multiple periods is that it can be used to evaluate the dynamic effects of interventions through various channels. For example, a positive relation between exercise and cognitive capacity in the cross-section does not mean much because it is not the snapshot that matters, but the history of processes that has led there. In this section, we highlight a few examples of how the distributions of factors change in several years' time when we exogenously manipulate the factors measuring investments, i.e., exercise or cognitive stimulation.

Tables 1.4.2 and 1.4.3 contain the effects of one possible set of such exercises for women and men, respectively. In the baseline scenario, we fix all factors at their age-80 medians. The next row shows the age-86 quantiles the factors are expected to end up at. We then change exercise to its first quartile at age 80, leaving all other factors at their median and letting all of them evolve according to the estimated transition equations until age 86. We repeat this exercise for setting the exercise factor to its third quartile at age 80. The last two panels do the same for cognitive stimulation. We do not take into account that mortality might be affected by the experiment – all effects are conditional on the corresponding individual in the data still being alive at age 86.

The main takeaway from the baseline exercise is that even when fixing everything at the median, there can be large expected changes just a few years down the road. For women, physical capacity is expected to be at the 45^{th} percentile at age 86, whereas cognitive capacity would be expected at its 64^{th} percentile. Hardly any change would be expected in the quantiles of exercise or cognitive stimulation. In stark contrast, for men there would be large drops in the expected quantiles of physical capacity, exercise, and cognitive stimulation along with a tiny drop in the quantile of cognitive capacity.

Due to the high persistence in exercise (for women, see Table 1.D.13 or Figure 1.D.13 in the Appendix; the numbers for men follow directly after those), changing at age 80 essentially means setting it to the same quantile all three periods. Doing so has a large effect on physical capacity (drops by 6-7 percentiles); cognitive capacity and cognitive stimulation barely change. The effect of increasing exercise to its third quartile is almost symmetric for women; the improvement is

Table 1.4.2. 6-year-ahead effects of changing exercise or cognitive stimulation, females

Scenario	Age	Physical Capacity	Cognitive Capacity	Exercise	Cognitive Stimulation
Baseline	80	0.50	0.50	0.50	0.50
	86	0.45	0.64	0.49	0.51
Exercise low	80	0.50	0.50	0.25	0.50
	86	0.39	0.63	0.28	0.51
Exercise high	80	0.50	0.50	0.75	0.50
	86	0.50	0.65	0.67	0.52
Cognitive Stimulation low	80	0.50	0.50	0.50	0.25
	86	0.46	0.52	0.48	0.29
Cognitive Stimulation high	80	0.50	0.50	0.50	0.75
	86	0.40	0.71	0.49	0.72

Table 1.4.3. 6-year-ahead effects of changing exercise or cognitive stimulation, males

Scenario	Age	Physical Capacity	Cognitive Capacity	Exercise	Cognitive Stimulation
Baseline	80	0.50	0.50	0.50	0.50
	86	0.34	0.48	0.39	0.38
Exercise low	80	0.50	0.50	0.25	0.50
	86	0.27	0.49	0.23	0.37
Exercise high	80	0.50	0.50	0.75	0.50
	86	0.37	0.50	0.54	0.39
Cognitive Stimulation low	80	0.50	0.50	0.50	0.25
	86	0.35	0.43	0.39	0.19
Cognitive Stimulation high	80	0.50	0.50	0.50	0.75
	86	0.33	0.53	0.40	0.60

only three percentiles for men. Note that, by age 86, exercise has reverted to its 54^{th} percentile, so less of an effect might be expected, too.

Fixing cognitive stimulation at its first quartile reduces cognitive capacity by 12 percentiles for women and by 5 percentiles for men. Interestingly, the larger effect for women occurs despite the fact that at age 86, cognitive stimulation is expected to be at its 29^{th} percentile for women compared to the 19^{th} percentile for men. Conversely, increasing cognitive stimulation substantially improves cognition for both genders. It also has a detrimental effect on women's health whereas there is no effect for men.

1.5 Conclusions and Outlook

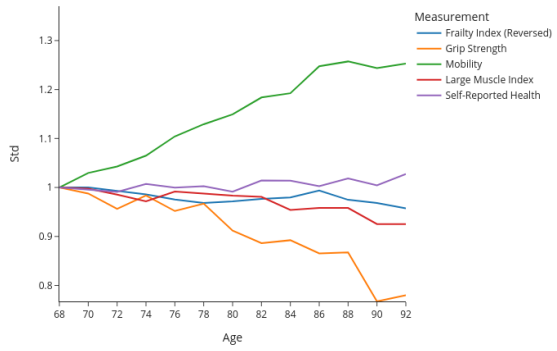
We adapt a nonlinear dynamic latent factor framework that was developed for skill formation of children to study the physical and cognitive decline between ages 68 and 93. To this end, we incorporate mortality into the model. The model is estimated with a rich set of measures from the Health and Retirement Study.

We document a large amount of measurement error in all observed variables. While most measurements have a high correlation with the latent factor they measure, no single measurement is a good enough proxy to use in isolation. A dynamic latent factor model is therefore a good fit for this setting. Having a rich set of time invariant measurements for each latent factor, lets us overcome some of the challenges related to the interpretability of latent factors. To make our results even more interpretable we also present them in terms of population ranks and use survival probabilities to anchor physical capacity.

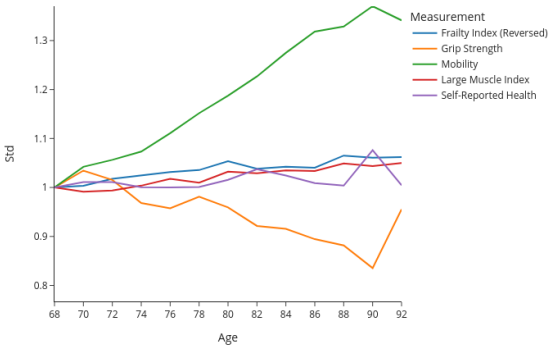
We find that, despite a strong decline in means for physical and cognitive capacity, the rank order of these latent factors is remarkably stable over periods. Nevertheless, physical and cognitive capacity can be influenced by investments until very high ages. Cognitive stimulation is a specific investment into cognitive capacity. Physical exercise has a larger effect on physical capacity and a small effect on cognitive capacity.

We leave a few extensions of our approach for future work. Besides expanding the sampling period by another wave, we want to add mental health as a separate latent factor that is different from cognitive and physical capacity but can influence both. As a robustness check we want to replace the linear probability model of mortality by a probit model. This requires the addition of nonlinear measurement equations to the model. To address any concerns of endogeneity of investments, we will use a control function approach similar to recent skill formation papers (Agostinelli and Wiswall, 2016a; Attanasio, Meghir, and Nix, 2020) as endogeneity correction. Finally we will use the model to simulate the effect of different investment policies and use Shapley decompositions to attribute the dynamic of investments over multiple periods to different channels of transmission.

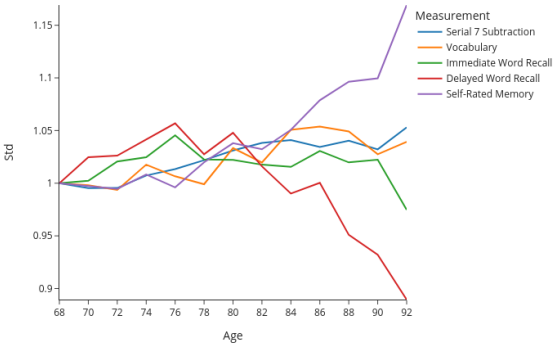
Appendix 1.A Additional Background on the Data and Measurements



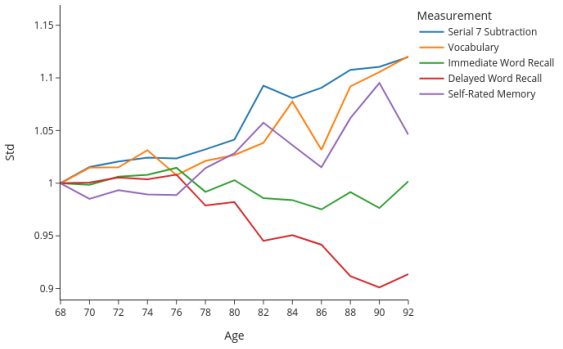
(a) Physical capacity, females



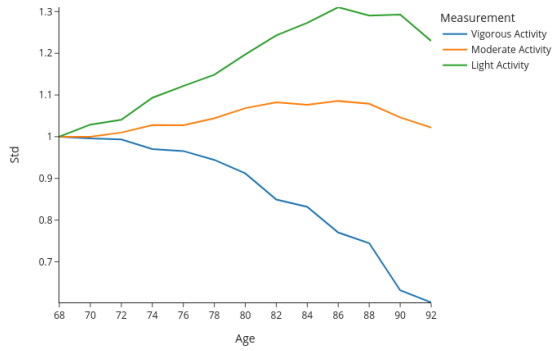
(b) Physical capacity, males



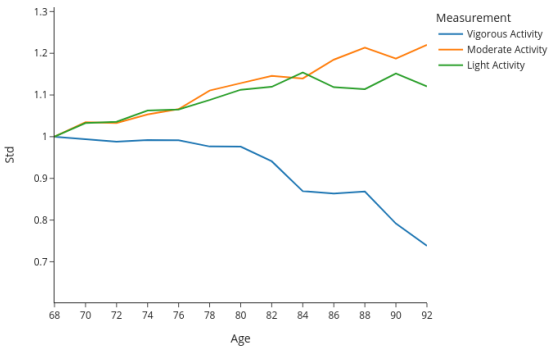
(c) Cognitive capacity, females



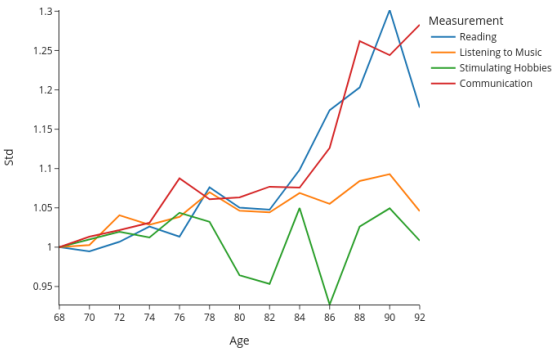
(d) Cognitive capacity, males



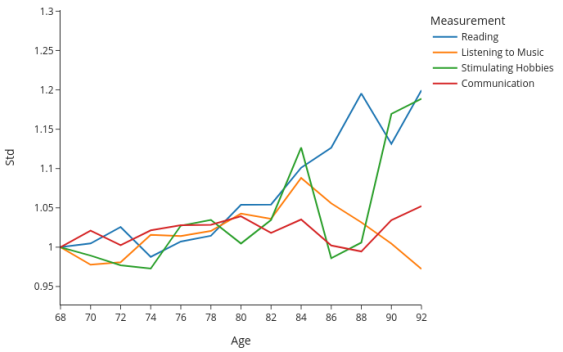
(e) Exercise, females



(f) Exercise, males



(g) Cognitive Stimulation, females



(h) Cognitive Stimulation, males

Figure 1.A.1. Standard deviation of measurements by age

Appendix 1.B The Maximum Likelihood Estimator

1.B.1 State Estimation

1.B.1.1 Preliminaries

To discuss the econometric approach used in this paper and potential alternatives it is convenient to express the model in state space notation.

To do so, let $\boldsymbol{x}_t \in \mathscr{R}^N$ denote the vector of latent factors (i.e. physical capacity, cognitive capacity, physical exercise and cognitive stimulation) in period t.

Similarly, let $\boldsymbol{y}_t \in \mathscr{R}^{L_t}$ denote the vector of all observable measurements in period t.

Then the transition function of the latent factors can be written as:

$$\boldsymbol{x}_{t+1} = F_t(\boldsymbol{x}_t) + \eta_t \tag{1.B.1}$$

where η_t is a vector of error terms with η_t^j on the j^{th} position. Let $\boldsymbol{Q}_t$ denote the covariance matrix of η_t

The linear measurement system can be written as:

$$\boldsymbol{y}_t = \boldsymbol{H_t}\boldsymbol{x}_t + \boldsymbol{\epsilon}_t \tag{1.B.2}$$

where $\boldsymbol{H_t}$ is a matrix of coefficients known as factor loadings and $\boldsymbol{\epsilon}_t$ is a vector of measurement errors with $\epsilon_{t,l}$ on the l^{th} position. Let $\boldsymbol{R}_t$ denote the covariance matrix of $\boldsymbol{\epsilon}_t$.

Equations 1.B.1 and 1.B.2 define a state space model. Equation 1.B.1 is called transition equation. Equation 1.B.2 is called measurement equation. The vector $\boldsymbol{x}_t$ is called the state of the system. The matrices $\boldsymbol{Q}_t$ and $\boldsymbol{R}_t$ are called process noise and measurement noise, respectively.

To see why it was handy to rewrite the technology of skill formation in state form, assume for a moment that the transition function F_t (including parameters) as well as the matrices $\boldsymbol{H_t}$, $\boldsymbol{Q}_t$ and $\boldsymbol{R}_t$ are known for all $t \in T$ but the state vectors $\boldsymbol{x}_t$ are unknown and have to be estimated from measurements $\boldsymbol{y}_t$. This problem is known as optimal state estimation, which is a well researched topic in physics and engineering.

To efficiently estimate the state vector in period t, an estimator should not only use measurements from this period, but also take the information from all previous measurements into account. For linear systems, Kalman filters are the method of choice for state estimation (Kalman, 1960). For nonlinear systems, several nonlinear variants of the Kalman filter have been developed. Kalman filters treat the state of a system itself as random vector. Therefore, they are sometimes classified as Bayesian filters.

Kalman filters consist of a predict and an update step. They are initialised with an initial estimate for the mean $\bar{\boldsymbol{x}}_0$ and covariance matrix $\boldsymbol{P}_0$ of the distribution of the state vector. Then, in each period, the new measurements are incorporated to update the mean and covariance matrix of the state vector. After that, the transition equation is used to predict the mean and covariance matrix of the state vector in the next period. This predicted state vector can then again be updated with measurements.

For the application of Kalman filters, the following assumptions must hold:

1. $\eta_t \sim \mathcal{N}(\mathbf{0}_N, \boldsymbol{Q}_t)$ where $\mathbf{0}_N$ denotes a vector of zeros of length N, $\boldsymbol{Q}_t$ is a diagonal matrix.
2. The η_t^j are serially independent over all t.
3. $\epsilon_t \sim \mathcal{N}(\mathbf{0}_{L_t}, \boldsymbol{R}_t)$ where $\boldsymbol{R}_t$ is a diagonal matrix.
4. The $\epsilon_{t,l}$ are serially independent over all t.
5. $\epsilon_{t,l}$ and η_t^j are independent of $\boldsymbol{x}_t$ for all $t = 1, \ldots, T$, $l = 1, \ldots, L$ and each factor j.
6. The distribution of the state vector $p(\boldsymbol{x}_t)$ can be approximated by a mixture of normal distributions for all $t = 1, \ldots, T$

Due to the assumption of a linear measurement system, the state vector can be estimated by combining the update step of a linear Kalman filter with the predict step of a nonlinear Kalman filter. For computational reasons, it will be convenient not to incorporate all measurements at once but to perform a separate update step for each measurement.

1.B.1.2 The Update Step of the Kalman Filter

The aim of the Kalman update is to efficiently combine information from measurements in the current period with previous measurements. To do so, the measurement function is used to convert the pre-update state vector into predicted measurements for the current period (equation 1.B.3). The difference between the predicted and actual measurements is called residual (equation 1.B.4). This residual, scaled by the so called Kalman gain, is then added to the pre-update state vector (equation 1.B.8). The Kalman gain is smaller if the variance of the measurement (calculated by equation 1.B.6) is large. This has the intuitive consequence that noisy measurements receive a low weight. The Kalman gain becomes larger if the pre-update covariance matrix has large diagonal entries (equation 1.B.5 and 1.B.7). Thus, measurements receive more weight if the pre-update state is known imprecisely due to bad initial values or a high process noise, for example. After the incorporation of the measurements, the state is always known with the same or more precision than before. This is reflected by subtracting a positive semi-definite matrix from the pre-update covariance matrix (equation 1.B.9).

Let $\bar{\boldsymbol{x}}_{t|y_{t,l}^-}$ denote the mean of the conditional distribution of the state vector given all measurements up to but not including the l^{th} measurement in period t. Let $\boldsymbol{P}_{t|y_{t,l}^-}$ denote the covariance matrix of this distribution. Let $\boldsymbol{h}_{t,l}$ denote the l^{th} row of $\boldsymbol{H}_t$. Let $r_{t,l,l}$ be the l^{th} diagonal element of $\boldsymbol{R}_t$. The update step that incorporates the l^{th} measurement into the estimate is given by the following equations:

$$\bar{y}_{t,l|y_{t,l}^-} = \boldsymbol{h}_{t,l}\bar{\boldsymbol{x}}_{t|y_{t,l}^-} \qquad \bar{y}_{t,l|y_{t,l}^-} = E(y_{t,l}|y_{t,l}^-) \tag{1.B.3}$$

$$\delta_{t,l} = y_{t,l} - \bar{y}_{t,l|y_{t,l}^-} \qquad \delta_{t,l} \text{ can be interpreted as residual} \tag{1.B.4}$$

$$\boldsymbol{f}_{t,l} = \boldsymbol{P}_{t|y_{t,l}^-}\boldsymbol{h}_{t,l}^T \qquad \boldsymbol{f}_{t,l} \text{ is an intermediate result} \tag{1.B.5}$$

$$\sigma_{t,l} = \boldsymbol{h}_{t,l}\boldsymbol{f}_{t,l} + r_{t,l,l} \qquad \sigma_{t,l} \text{ is the variance of } y_{t,l} \tag{1.B.6}$$

$$\boldsymbol{k}_{t,l} = \frac{1}{\sigma_{t,l}}\boldsymbol{f}_{t,l} \qquad \boldsymbol{k}_{t,l} \text{ is the (scaled) Kalman gain} \tag{1.B.7}$$

$$\bar{\boldsymbol{x}}_{t|y_{t,l}} = \bar{\boldsymbol{x}}_{t|y_{t,l}^-} + \boldsymbol{k}_{t,l}\delta_{t,l} \qquad \bar{\boldsymbol{x}}_{t|y_{t,l}} \text{ is the updated mean} \tag{1.B.8}$$

$$\boldsymbol{P}_{t|y_{t,l}} = \boldsymbol{P}_{t|y_{t,l}^-} - \frac{1}{\sigma_{t,l}}\boldsymbol{f}_{t,l}\boldsymbol{f}_{t,l}^T \qquad P_{t|y_{t,l}} \text{ is the updated covariance matrix} \tag{1.B.9}$$

1.B.1.3 The Predict Step of the Kalman Filter

In linear systems, the mean and covariance matrix of the system can be propagated to the next period by simply applying the linear transition equation. With a nonlinear transition function, however, this is not possible, as $E(f(X) \neq f(E(X)$ in general. For the nonlinear predict step, two basic options exist: The *extended Kalman filter* and the *unscented Kalman filter*. Cunha, Heckman and Schennach choose the unscented Kalman filter because it has been shown to be more reliable in a wide range of settings (Van Der Merwe, 2004).

The intuition of the predict step of the unscented Kalman filter is relatively simple: firstly, a deterministic sample of points in the state space, called sigma points (equation 1.B.10), and accompanying weights are chosen (equation 1.B.11). Usually these are $2N + 1$ points and weights, where N is the length of the state vector. Secondly, these sigma points are transformed using the true nonlinear transition equation. Thirdly, the weighted sample mean is used as estimate for the next period mean of the state vector (equation 1.B.12). Fourthly, the sum of the covariance matrix of the process noise and the weighted sample covariance of the transformed sigma points is used as estimate of the covariance matrix of the state vector (equation 1.B.13). Intuitively, the addition of the process noise accounts for the fact that the prediction always adds some uncertainty about the state of the system.

For the choice of sigma points and sigma weights, many different algorithms exist. All have in common that some form of matrix square root of the covariance matrix of the state vector is taken. Two definitions of matrix square root exist: 1) $\boldsymbol{A}$ is a matrix square root of $\boldsymbol{P}$ if $\boldsymbol{P} = \boldsymbol{A}\boldsymbol{A}$. 2) $\boldsymbol{A}$ is a matrix square root of $\boldsymbol{P}$ if $\boldsymbol{P} = \boldsymbol{A}\boldsymbol{A}^T$. The matrix square root is not unique in general and some matrices do not have a square root. However, all symmetric positive semi-definite matrices, i.e. all valid covariance matrices, can be decomposed into $\boldsymbol{P} = \boldsymbol{L}\boldsymbol{L}^T$ where $\boldsymbol{L}$ is lower triangular (Zhang, 1999). For the unscented Kalman filter, both definitions of matrix square root work. Below, the sigma point algorithm proposed by Julier and Uhlmann (1997), is presented without reference to a particular type of matrix square root:

Let $\kappa \in \mathbb{R}$ be a scaling parameter. Usually, κ is set to 2 if the distribution of the state vector is assumed to be normal. Let $\boldsymbol{P}_{t|t}$ denote the covariance matrix of the state vector, conditional on all

measurements up to and including period t. Define $\boldsymbol{S}_{t|t} \equiv \sqrt{\boldsymbol{P}_{t|t}}$ as the matrix square root of $\boldsymbol{P}_{t|t}$ and let $\boldsymbol{s}_{t,n}$ denote its n^{th} column.

Sigma points are calculated according to the following equations:

$$\begin{aligned} \chi_{t,n} &= \bar{\boldsymbol{x}}_{t|t} && \text{for } n = 0 \\ \chi_{t,n} &= \bar{\boldsymbol{x}}_{t|t} + \sqrt{N+\kappa}\ \boldsymbol{s}_{t,n} && \text{for } n = 1,\dots N \\ \chi_{t,n} &= \bar{\boldsymbol{x}}_{t|t} - \sqrt{N+\kappa}\ \boldsymbol{s}_{t,n} && \text{for } n = N+1,\dots 2N \end{aligned} \tag{1.B.10}$$

where $\chi_{t,n}$ is the n^{th} sigma point at period t that is calculated after incorporating all measurements of that period. The corresponding sigma weights are calculated as follows:

$$\begin{aligned} w_{t,n} &= \frac{\kappa}{N+\kappa} && \text{for } n = 0 \\ w_{t,n} &= \frac{1}{2(N+\kappa)} && \text{for } n = 1,\dots,2N \end{aligned} \tag{1.B.11}$$

where $w_{t,n}$ is the n^{th} sigma weight. Define $\widetilde{\chi}_{t,n} \equiv F_t(\chi_{t,n})$ where $F_t(\cdot)$ is defined as in equation 1.B.1. Then the predict step of the unscented Kalman filter is given by:

$$\bar{\boldsymbol{x}}_{t+1|t} = \sum_{n=0}^{2N} w_{t,n}\widetilde{\chi}_{t,n} \tag{1.B.12}$$

$$\boldsymbol{P}_{t+1|t} = \left[\sum_{n=0}^{2N} w_{t,n}(\widetilde{\chi}_{t,n} - \bar{\boldsymbol{x}}_{t+1|t})(\widetilde{\chi}_{t,n} - \bar{\boldsymbol{x}}_{t+1|t})^T\right] + \boldsymbol{Q_t} \tag{1.B.13}$$

1.B.2 The Likelihood Interpretation of the Kalman Filter

Of course, the parameters of the function F_t and the matrices $\boldsymbol{H}_t$, $\boldsymbol{Q}_t$ and $\boldsymbol{R}_t$ are unknown in reality. However, they can be estimated by maximum likelihood. The direct maximization of the likelihood function would involve the evaluation of high dimensional integrals which is computationally very expensive (Cunha, Heckman, and Schennach, 2010). Instead, Kalman filters can be used to reduce the sumber of computations required for each evaluation of the likelihood function dramatically.

To see how, define $\boldsymbol{\theta}$ as the vector with all estimated parameters of the model. Then, the likelihood contribution of individual i is given by:

$$\mathscr{L}(\boldsymbol{\theta}|\boldsymbol{y}_1,\dots,\boldsymbol{y}_T) \equiv p_\theta(\boldsymbol{y}_1,\dots,\boldsymbol{y}_T) = \prod_{t=1}^{T}\prod_{l=1}^{L_t} p_\theta(y_{t,l}|y_{t,l}^{-}) \tag{1.B.14}$$

where $p_\theta(\boldsymbol{y}_1,\dots,\boldsymbol{y}_T)$ denotes the joint density of all measurements for individual i, conditional on the parameter vector θ and $p_\theta(y_{t,l}|y_{t,l}^{-})$ is the density of the l^{th} measurement in period t, given

all measurements up to but not including this measurement. The subscript i is again omitted for readability.

To see how this relates to the Kalman filter, recall that for each $t = 1, \ldots, T$ and each $l = 1, \ldots, L_t$, equation 1.B.3 calculates $\bar{y}_{t,l|y_{t,l}^-}$, i.e the expected value of the l^{th} measurement in period t, conditional on all previous measurements. In addition, due to the normality and independence assumptions on the error terms and the factor distribution, $y_{t,l}$ is normally distributed around $\bar{y}_{t,l|y_{t,l}^-}$. Equation 1.B.6 can be used to calculate the variance $\sigma_{t,l}$ of this distribution. Thus, $p_\theta(y_{t,l}|y_{t,l}^-) = \phi_{\bar{y}_{t,l|y_{t,l}^-},\sigma_{t,l}}(y_{t,l})$ where $\phi_{\mu,\sigma}(\cdot)$ is the density of a normal random variable with mean μ and variance σ.

A nice feature of the estimator based on this factorization of the likelihood function is that it can deal very well with missing observations. If measurement $y_{t,l}$ is missing for inidividual i, the corresponding update of the state vector is just skipped. More formally, this means that the missing measurement is integrated out from the likelihood function.

1.B.3 Numerical Stability

1.B.3.1 Numerical Challenges

While the Kalman filter based maximum likelihood estimator is statistically and computationally efficient, it is numerically unstable. The numerical instability caused by floating point imprecision is inherent to Kalman filters and has been discovered soon after Kalman published his original article. Since then, the precision of computers has increased enormously such that nowadays numerical problems are not a big issue for well specified Kalman filters. However, during the maximization of the likelihood function the optimizer might pick parameter combinations that are far from leading to a well specified filter.

The numerical problems manifest themselves in two places:

1. In the update step, the subtraction in equation 1.B.9 can lead to negative diagonal elements in the updated covariance matrix of the state vector. While this is mathematically impossible in a well specified Kalman filter, numerical imprecisions and badly specified Kalman filters during the maximization process make it possible.
2. Even if the covariance matrix of the state vector has nonnegative diagonal entries, numerical imprecisions might render it not positive semi-definite. With this the existence of a matrix square root is not guaranteed, which can make the calculation of sigma points impossible.

Cunha, Heckman and Schennach mention the numerical problems in their supplementary material. To solve the first problem, they recommend to find good initial values for the maximization by first constraining some parameters and letting the code find good initial values for the others. For the second problem, they propose to set all off-diagonal elements of $\boldsymbol{P}$ to zero before taking the square root, which then corresponds to taking the element wise square root of the diagonal elements. While this prevents the estimator from crashing, it is not standard practice in Kalman filtering and it is not guaranteed that an estimator based on this type of matrix square root produces reliable results.

1.B.3.2 Outline of the Solution

A better approach is to use a square root implementation of the Kalman filter. Many different square root Kalman filters exist. They are mathematically equivalent to normal Kalman filters but numerically more stable.

Instead of propagating the full covariance matrix of the state vector, square root Kalman filters propagate the square root of this matrix. This has three advantages:

1. It avoids overflow errors due to numbers with very small or large absolute values, as taking the square root makes large numbers smaller and small numbers larger.
2. By using a matrix square root $\boldsymbol{A}$ of the type $\boldsymbol{P} = \boldsymbol{A}\boldsymbol{A}^T$, the problematic covariance matrix is guaranteed to be positive semi-definite (Zhang, 1999), i.e. a valid covariance matrix. In particular, its diagonal entries are sums of squared terms and, consequently, guaranteed to be nonnegative. This solves the first problem.
3. By choosing an appropriate pair of square root update and predict algorithms, taking matrix square roots can be completely avoided. This eliminates the second problem.

The computational requirements of square root filters are comparable to those of normal Kalman filters. In the nonlinear case, they are even lower. For a maximally robust estimator, we use a pair of square root update and predict algorithms that completely avoid taking matrix square roots. The algorithm for the update was developed by Prvan and Osborne (Prvan and Osborne, 1988). The unscented square root predict step was proposed by Van Der Merwe and Wan (van der Merwe and Wan, 2001). Both propagate the transpose of a lower triangular matrix square root of the state covariance matrix.

1.B.3.3 The QR Decomposition of a Matrix

Both square root algorithms rely on a matrix factorization called QR decomposition. Note that in this subsection, $\boldsymbol{Q}$ and $\boldsymbol{R}$ do not denote the covariance matrices of the process and measurement noise but factors into which a matrix is decomposed.

$\boldsymbol{Q}\boldsymbol{R}$ is called QR decomposition of an $m \times n$ matrix $\boldsymbol{A}$ with $m \geq n$ if:

1. $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$
2. $\boldsymbol{Q}$ is an orthogonal $m \times m$ matrix
3. $\boldsymbol{R}$ is an $m \times n$ matrix and the first n rows of $\boldsymbol{R}$ form a upper triangular matrix and its remaining rows only contain zeros

The QR decomposition of a matrix always exists but is not unique. A useful property of the QR decomposition is that:

$$\boldsymbol{A}^T\boldsymbol{A} \;=\; (\boldsymbol{Q}\boldsymbol{R})^T\boldsymbol{Q}\boldsymbol{R} \;=\; \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{R} \;=\; \boldsymbol{R}^T\boldsymbol{R} \tag{1.B.15}$$

where the last equality comes from the defining property of orthogonal matrices that $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}$, where $\boldsymbol{I}$ denotes the identity matrix. Thus, the upper triangular part of $\boldsymbol{R}$ is the transpose of a lower triangular matrix square root of $\boldsymbol{A}^T\boldsymbol{A}$. For convenience, let $qr(\boldsymbol{A})$ denote the QR decomposition of $\boldsymbol{A}$ that only returns the upper triangular part of the matrix $\boldsymbol{R}$.

1.B.3.4 The Update Step of the Square-Root Kalman Filter

Let $\boldsymbol{S}_{t|y_{l,t}^-}$ be a lower triangular matrix square root of $\boldsymbol{P}_{t|y_{l,t}^-}$ and keep the rest of the notation as in section 1.B.1. Then, the square root update that incorporates the l^{th} measurement in period t is given by the following equations:

$\bar{y}_{t,l|y_{t,l}^-}$ and $\delta_{t,l}$ are calculated as in equation 1.B.3 and 1.B.4 respectively. Then the following intermediate results are calculated.

$$\boldsymbol{f}_{t,l}^* = \boldsymbol{S}_{t|y_{t,l}^-}^T \boldsymbol{h}_{t,l}^T \tag{1.B.16}$$

$$\boldsymbol{M}_{t,l} = \begin{bmatrix} \sqrt{r_{t,l,l}} & \boldsymbol{0}_N^T \\ \boldsymbol{f}_{t,l}^* & \boldsymbol{S}_{t|y_{t,l}^-}^T \end{bmatrix} \tag{1.B.17}$$

It can be shown that:

$$qr(\boldsymbol{M}_{t,l}) = \begin{bmatrix} \sqrt{\sigma_{t,l}} & \frac{1}{\sqrt{\sigma_{t,l}}}\boldsymbol{f}_{t,l}^T \\ \boldsymbol{0}_N & \boldsymbol{S}_{t|y_{t,l}}^T \end{bmatrix} \tag{1.B.18}$$

where $\boldsymbol{S}_{t|y_{t,l}}^T$ is the transpose of a lower triangular square root of the updated covariance matrix and $\boldsymbol{0}_N$ denotes a column vector of length N that is filled with zeros.

The matrix in equation 1.B.18 also contains $\boldsymbol{f}_{t,l}$ and $\sigma_{t,l}$ such that the Kalman gain can be calculated as in equation 1.B.7 and the mean of the state vector can be updated as in equation 1.B.8.

To see why equation 1.B.18 holds, define $\boldsymbol{U}_{t,l} \equiv qr(\boldsymbol{M}_{t,l})$ and partition it as follows:

$$\boldsymbol{U}_{t,l} = \begin{bmatrix} \boldsymbol{U}_{1,1} & \boldsymbol{U}_{1,2} \\ \boldsymbol{0} & \boldsymbol{U}_{2,2} \end{bmatrix} \tag{1.B.19}$$

where $\boldsymbol{U}_{1,1}$ is a scalar, $\boldsymbol{U}_{1,2}$ a row vector of length N, $\boldsymbol{0}$ a column vector of length N filled with zeros and $\boldsymbol{U}_{2,2}$ an upper triangular $N \times N$ matrix. Recall from the definition of $\boldsymbol{U}_{t,l}$ and equation 1.B.15 that $\boldsymbol{U}_{t,l}^T\boldsymbol{U}_{t,l} = \boldsymbol{M}_{t,l}^T\boldsymbol{M}_{t,l}$. Multiplying out both sides of this equality yields:

$$\begin{bmatrix} r_{t,l,l} + \boldsymbol{f}_{t,l}^{*T}\boldsymbol{f}_{t,l}^* & \boldsymbol{f}_{t,l}^{*T}\boldsymbol{S}_{t|y_{t,l}^-}^T \\ \boldsymbol{S}_{t|y_{t,l}^-}\boldsymbol{f}_{t,l}^* & \boldsymbol{S}_{t|y_{t,l}^-}\boldsymbol{S}_{t|y_{t,l}^-}^T \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}_{1,1}^2 & \boldsymbol{U}_{1,1}\boldsymbol{U}_{1,2} \\ \boldsymbol{U}_{1,2}^T\boldsymbol{U}_{1,1} & \boldsymbol{U}_{1,2}^T\boldsymbol{U}_{1,2} + \boldsymbol{U}_{2,2}^T\boldsymbol{U}_{2,2} \end{bmatrix} \tag{1.B.20}$$

It is obvious from equation 1.B.6 and 1.B.16 that $\boldsymbol{U}_{1,1} = \sqrt{\sigma_{t,l}}$. Using this and noting that $\boldsymbol{f}_{t,l}^{*T}\boldsymbol{S}_{t|y_{t,l}^-}^T = \boldsymbol{f}_{t,l}^T$, where $\boldsymbol{f}_{t,l}$ is defined as in equation 1.B.5, one obtains that:

$$\boldsymbol{U}_{1,2} \;=\; \frac{\boldsymbol{f}_{t,l}^{T}}{\sqrt{\sigma_{t,l}}} \tag{1.B.21}$$

It remains to show that $\boldsymbol{U}_{2,2} = \boldsymbol{S}_{t|y_{t,l}}^{T}$. By noting that the bottom right element of the left hand side of equation 1.B.20 is, by definition, equal to the pre-update covariance matrix $\boldsymbol{P}_{t|y_{t,l}^{-}}$ and plugging in the value for $\boldsymbol{U}_{1,2}$, one obtains that:

$$\boldsymbol{U}_{2,2}^{T}\boldsymbol{U}_{2,2} \;=\; \boldsymbol{P}_{t|y_{t,l}^{-}} - \frac{1}{\sigma_{t,l}}\boldsymbol{f}_{t,l}\boldsymbol{f}_{t,l}^{T} \;=\; \boldsymbol{P}_{t|y_{t,l}} \tag{1.B.22}$$

where the last equality comes comes from equation 1.B.9. Thus $\boldsymbol{U}_{2,2}^{T}$ is a matrix square root of $\boldsymbol{P}_{t|y_{t,l}}$ and by the definition of the QR decomposition it is lower triangular, which completes the proof. Importantly, no part of the proof requires the lower triangular square roots of $\boldsymbol{P}_{t|y_{t,l}^{-}}$ or $\boldsymbol{P}_{t|y_{t,l}}$ to be unique or makes reference to a specific type of matrix square root.

1.B.3.5 The Predict Step of the Square-Root Kalman Filter

For the square root implementation of the unscented predict step in period t, firstly the sigma points are calculated as in equation 1.B.10, where this time $\boldsymbol{S}_{t|t}$ is required to be a lower triangular matrix square root of $\boldsymbol{P}_{t|t}$. Again, $\tilde{\mathcal{X}}_t$ denotes the $(2N+1)\times N$ matrix of the transformed sigma points. The calculation of the predicted mean of the state vector remains the same as before (equation 1.B.12).

Define $\boldsymbol{A}_t$ as stacked matrix of weighted deviations of the sigma points from the predicted mean and the covariance matrix of the transition shocks:

$$\boldsymbol{A}_t \;\equiv\; \begin{bmatrix} \sqrt{w_{t,0}}(\tilde{\mathcal{X}}_{t,0} - \bar{\boldsymbol{x}}_{t+1|t})^{T} \\ \cdots \\ \sqrt{w_{t,2n}}(\tilde{\mathcal{X}}_{t,2n} - \bar{\boldsymbol{x}}_{t+1|t})^{T} \\ \sqrt{\boldsymbol{Q}_t} \end{bmatrix} \tag{1.B.23}$$

Then equation 1.B.13 can be rewritten as:

$$\boldsymbol{P}_{t+1|t} \;=\; \boldsymbol{A}_t^{T}\boldsymbol{A}_t \tag{1.B.24}$$

and by the relation of the QR decomposition and and the lower triangular matrix square root (equation 1.B.15) a lower triangular matrix square root of $\boldsymbol{P}_{t+1|t}$ is given by $qr(\boldsymbol{A}_t)^{T}$.

Appendix 1.C Detailed Model Setup

1.C.1 Background on Identification

Cunha, Heckman, and Schennach (2010) provide very general nonparametric Identification result for nonlinear dynamic latent factor models. The exact conditions for identification depend on the assumptions one is willing to put on the measurement error. However, having at least two

dedicated measurements for each latent factor in each period is sufficient to identify an arbitrary production function under mild conditions. Since latent factors do not have a natural unit of measurement, the identification requires normalizations of location and scale. Thus, Cunha, Heckman, and Schennach (2010) normalize one loading of each factor in each period to 1 and one intercept of each factor in each period to 0. While the identification result works for arbitrary production functions, they use a parametric CES function in their empirical application.

Agostinelli and Wiswall (2016b) criticize the identification result by Cunha, Heckman, and Schennach (2010) to be flawed. They point out that the CES production function already puts a restriction on the scale and location of its output. Thus, normalization of scale and location are only required in the first period and re-normalizations in each period are actually not normalizations but testable assumptions. Moreover, they show that under the implicit restrictions imposed by the CES production function, identification under a linear measurement system can be achieved with as little as one measurement per latent factor and period as long as there are at least two measurements in the first period.

Freyberger (2021) shows that the CES production function also imposes implicit restrictions on the relative scale of the latent factors and thus identification can be achieved if only the location and scale of a single factor are normalized in the first period.

While the critique by Agostinelli and Wiswall (2016b) that over-normalizations are detrimental is correct, it mostly applies to the empirical application and not the general identification result in Cunha, Heckman, and Schennach (2010) nor the maximum likelihood estimator used in the paper. The identification result states that latent factors have no natural scale and location that could be be identified from data and thus their location and scale has to be fixed by restrictions imposed by the econometrician. Cunha, Heckman, and Schennach (2010) restrict factor loadings and intercepts but mention, that instead of factor loadings, the variances of measurement errors could be restricted. Of course, these restrictions are mutually exclusive and it would not be valid to restrict factor loadings and variances of measurement error at the same time. The main contribution of Agostinelli and Wiswall (2016b) is to point out that using restrictive functional forms for the production function is yet another way of fixing the location and scale of the latent factors.

Appendix 1.D Additional Tables and Figures for the Main Specification

1.D.1 Complete Set of Parameters of the Measurement System

Table 1.D.1. Intercepts, Loadings, and Measurement Standard Deviations for Physical Capacity, Females

		Intercept	Loading	Meas. Std.
All	Frailty Index (Reversed)	0.000	1.000	0.707***
				(0.001)
	Mobility	−0.113***	1.228***	0.766***
		(0.003)	(0.005)	(0.003)
	Large Muscle Index	0.005*	0.929***	0.750***
		(0.003)	(0.005)	(0.002)
	Self-Reported Health	−0.048***	0.950***	0.765***
		(0.003)	(0.004)	(0.002)
70	Alive	0.897***	0.042***	0.303***
		(0.103)	(0.011)	(0.039)
	Grip Strength	−0.126***	0.489***	0.933***
		(0.027)	(0.042)	(0.015)
72	Alive	0.909***	0.045***	0.288***
		(0.107)	(0.011)	(0.038)
	Grip Strength	−0.241***	0.396***	0.922***
		(0.028)	(0.042)	(0.016)
74	Alive	0.902***	0.060***	0.301***
		(0.097)	(0.013)	(0.036)
	Grip Strength	−0.292***	0.466***	0.935***
		(0.030)	(0.043)	(0.018)
76	Alive	0.885***	0.073***	0.327***
		(0.101)	(0.018)	(0.043)
	Grip Strength	−0.471***	0.368***	0.924***
		(0.030)	(0.049)	(0.012)
78	Alive	0.879***	0.075***	0.339***
		(0.103)	(0.019)	(0.046)
	Grip Strength	−0.540***	0.447***	0.924***
		(0.033)	(0.048)	(0.019)
80	Alive	0.871***	0.091***	0.353***
		(0.097)	(0.023)	(0.047)
	Grip Strength	−0.758***	0.367***	0.882***
		(0.034)	(0.052)	(0.021)
82	Alive	0.870***	0.089***	0.359***
		(0.112)	(0.026)	(0.054)
	Grip Strength	−0.789***	0.336***	0.861***
		(0.037)	(0.055)	(0.020)
84	Alive	0.869***	0.110***	0.371***
		(0.105)	(0.030)	(0.053)
	Grip Strength	−0.980***	0.334***	0.866***
		(0.042)	(0.061)	(0.026)
86	Alive	0.856***	0.123***	0.391***
		(0.122)	(0.040)	(0.067)
	Grip Strength	−0.997***	0.337***	0.839***
		(0.046)	(0.071)	(0.028)
88	Alive	0.846***	0.129**	0.406***
		(0.146)	(0.053)	(0.086)
	Grip Strength	−1.191***	0.413***	0.827***
		(0.060)	(0.084)	(0.035)
90	Alive	0.828***	0.137*	0.425***
		(0.203)	(0.081)	(0.133)
	Grip Strength	−1.105***	0.357***	0.736***
		(0.062)	(0.099)	(0.032)
92	Alive	0.819***	0.168	0.443***
		(0.215)	(0.116)	(0.148)
	Grip Strength	−1.362***	0.350***	0.746***
		(0.083)	(0.116)	(0.048)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.D.2. Intercepts, Loadings, and Measurement Standard Deviations for Physical Capacity, Males

		Intercept	Loading	Meas. Std.
All	Frailty Index (Reversed)	0.000	1.000	0.796***
				(0.002)
	Mobility	−0.015***	1.331***	0.750***
		(0.005)	(0.007)	(0.003)
	Large Muscle Index	0.042***	1.032***	0.761***
		(0.004)	(0.006)	(0.003)
	Self-Reported Health	0.026***	0.963***	0.793***
		(0.003)	(0.006)	(0.003)
70	Alive	0.901***	0.058***	0.303***
		(0.092)	(0.013)	(0.035)
	Grip Strength	−0.056	0.578***	0.978***
		(0.034)	(0.053)	(0.020)
72	Alive	0.907***	0.075***	0.298***
		(0.083)	(0.015)	(0.030)
	Grip Strength	−0.294***	0.550***	0.959***
		(0.034)	(0.053)	(0.020)
74	Alive	0.900***	0.061***	0.310***
		(0.119)	(0.017)	(0.046)
	Grip Strength	−0.318***	0.497***	0.922***
		(0.035)	(0.057)	(0.021)
76	Alive	0.876***	0.073***	0.344***
		(0.129)	(0.024)	(0.059)
	Grip Strength	−0.506***	0.559***	0.898***
		(0.037)	(0.057)	(0.020)
78	Alive	0.872***	0.081***	0.355***
		(0.130)	(0.027)	(0.062)
	Grip Strength	−0.560***	0.553***	0.920***
		(0.041)	(0.059)	(0.023)
80	Alive	0.866***	0.089***	0.367***
		(0.135)	(0.031)	(0.068)
	Grip Strength	−0.737***	0.571***	0.891***
		(0.043)	(0.061)	(0.023)
82	Alive	0.852***	0.136***	0.394***
		(0.117)	(0.043)	(0.066)
	Grip Strength	−0.959***	0.468***	0.872***
		(0.046)	(0.065)	(0.025)
84	Alive	0.868***	0.139***	0.387***
		(0.130)	(0.047)	(0.068)
	Grip Strength	−1.042***	0.557***	0.842***
		(0.052)	(0.069)	(0.027)
86	Alive	0.849***	0.140**	0.408***
		(0.157)	(0.062)	(0.092)
	Grip Strength	−1.238***	0.488***	0.841***
		(0.064)	(0.085)	(0.034)
88	Alive	0.856***	0.176**	0.418***
		(0.154)	(0.074)	(0.090)
	Grip Strength	−1.283***	0.471***	0.826***
		(0.071)	(0.109)	(0.045)
90	Alive	0.850***	0.204*	0.430***
		(0.212)	(0.121)	(0.128)
	Grip Strength	−1.358***	0.508***	0.766***
		(0.102)	(0.120)	(0.055)
92	Alive	0.765**	0.183	0.464*
		(0.312)	(0.218)	(0.268)
	Grip Strength	−1.493***	0.684***	0.816***
		(0.123)	(0.166)	(0.077)

Note: ***p<0.01;**p<0.05;*p<0.1

Table 1.D.3. Intercepts, Loadings, and Measurement Standard Deviations for Cognitive Capacity, Females

		Intercept	Loading	Meas. Std.
All	Serial 7 Subtraction	0.000	1.000	0.890***
				(0.003)
	Vocabulary	0.043***	0.839***	0.923***
		(0.006)	(0.013)	(0.004)
	Immediate Word Recall	−0.161***	1.801***	0.583***
		(0.006)	(0.015)	(0.003)
	Delayed Word Recall	−0.189***	1.805***	0.595***
		(0.006)	(0.014)	(0.002)
70	Self-Rated Memory	0.005	0.576***	0.961***
		(0.014)	(0.031)	(0.009)
72	Self-Rated Memory	0.029**	0.593***	0.955***
		(0.015)	(0.030)	(0.009)
74	Self-Rated Memory	0.016	0.555***	0.973***
		(0.015)	(0.030)	(0.009)
76	Self-Rated Memory	0.028*	0.497***	0.968***
		(0.017)	(0.033)	(0.010)
78	Self-Rated Memory	0.045**	0.501***	0.992***
		(0.019)	(0.035)	(0.011)
80	Self-Rated Memory	0.052**	0.470***	1.013***
		(0.022)	(0.038)	(0.012)
82	Self-Rated Memory	0.069**	0.460***	1.010***
		(0.027)	(0.043)	(0.013)
84	Self-Rated Memory	0.083**	0.398***	1.035***
		(0.032)	(0.050)	(0.015)
86	Self-Rated Memory	0.079*	0.393***	1.063***
		(0.041)	(0.058)	(0.018)
88	Self-Rated Memory	0.261***	0.549***	1.069***
		(0.055)	(0.075)	(0.021)
90	Self-Rated Memory	0.210***	0.459***	1.081***
		(0.074)	(0.097)	(0.026)
92	Self-Rated Memory	0.218**	0.538***	1.145***
		(0.110)	(0.133)	(0.040)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.D.4. Intercepts, Loadings, and Measurement Standard Deviations for Cognitive Capacity, Males

		Intercept	Loading	Meas. Std.
All	Serial 7 Subtraction	0.000	1.000	0.907***
				(0.004)
	Vocabulary	0.048***	0.960***	0.900***
		(0.008)	(0.016)	(0.004)
	Immediate Word Recall	−0.183***	1.684***	0.599***
		(0.008)	(0.016)	(0.003)
	Delayed Word Recall	−0.200***	1.648***	0.605***
		(0.008)	(0.015)	(0.003)
70	Self-Rated Memory	−0.041**	0.626***	0.937***
		(0.017)	(0.035)	(0.011)
72	Self-Rated Memory	−0.052***	0.560***	0.955***
		(0.017)	(0.034)	(0.011)
74	Self-Rated Memory	−0.043**	0.573***	0.949***
		(0.017)	(0.035)	(0.011)
76	Self-Rated Memory	−0.039**	0.527***	0.955***
		(0.020)	(0.040)	(0.012)
78	Self-Rated Memory	−0.051**	0.607***	0.972***
		(0.022)	(0.043)	(0.013)
80	Self-Rated Memory	−0.002	0.589***	0.988***
		(0.026)	(0.048)	(0.015)
82	Self-Rated Memory	−0.019	0.479***	1.033***
		(0.034)	(0.057)	(0.018)
84	Self-Rated Memory	−0.019	0.520***	1.007***
		(0.040)	(0.063)	(0.020)
86	Self-Rated Memory	−0.019	0.464***	0.992***
		(0.046)	(0.071)	(0.022)
88	Self-Rated Memory	0.007	0.509***	1.035***
		(0.065)	(0.091)	(0.028)
90	Self-Rated Memory	0.011	0.386***	1.080***
		(0.089)	(0.120)	(0.038)
92	Self-Rated Memory	0.003	0.599***	1.011***
		(0.125)	(0.182)	(0.049)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.D.5. Intercepts, Loadings, and Measurement Standard Deviations for Exercise, Females

		Intercept	Loading	Meas. Std.
All	Vigorous Activity	−0.009	0.682***	0.809***
		(0.006)	(0.010)	(0.004)
	Moderate Activity	0.000	1.000	0.794***
				(0.004)
	Light Activity	−0.127***	1.076***	0.933***
		(0.007)	(0.012)	(0.004)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.D.6. Intercepts, Loadings, and Measurement Standard Deviations for Exercise, Males

		Intercept	Loading	Meas. Std.
All	Vigorous Activity	−0.012**	0.741***	0.814***
		(0.006)	(0.012)	(0.005)
	Moderate Activity	0.000	1.000	0.816***
				(0.004)
	Light Activity	−0.077***	0.927***	0.861***
		(0.007)	(0.013)	(0.004)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.D.7. Intercepts, Loadings, and Measurement Standard Deviations for Cognitive Stimulation, Females

		Intercept	Loading	Meas. Std.
All	Reading	0.000	1.000	0.780***
				(0.006)
	Listening to Music	−0.168***	0.512***	0.980***
		(0.006)	(0.010)	(0.006)
	Stimulating Hobbies	−0.069***	0.578***	0.925***
		(0.007)	(0.011)	(0.005)
	Communication	−0.062***	0.523***	0.999***
		(0.006)	(0.010)	(0.005)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.D.8. Intercepts, Loadings, and Measurement Standard Deviations for Cognitive Stimulation, Males

		Intercept	Loading	Meas. Std.
All	Reading	0.000	1.000	0.683***
				(0.007)
	Listening to Music	−0.175***	0.229***	1.004***
		(0.007)	(0.010)	(0.007)
	Stimulating Hobbies	−0.012	0.375***	0.969***
		(0.009)	(0.012)	(0.005)
	Communication	−0.083***	0.325***	0.989***
		(0.007)	(0.011)	(0.006)
Note:				***p<0.01;**p<0.05;*p<0.1

1.D.2 Correlations Between Measurements and Factors

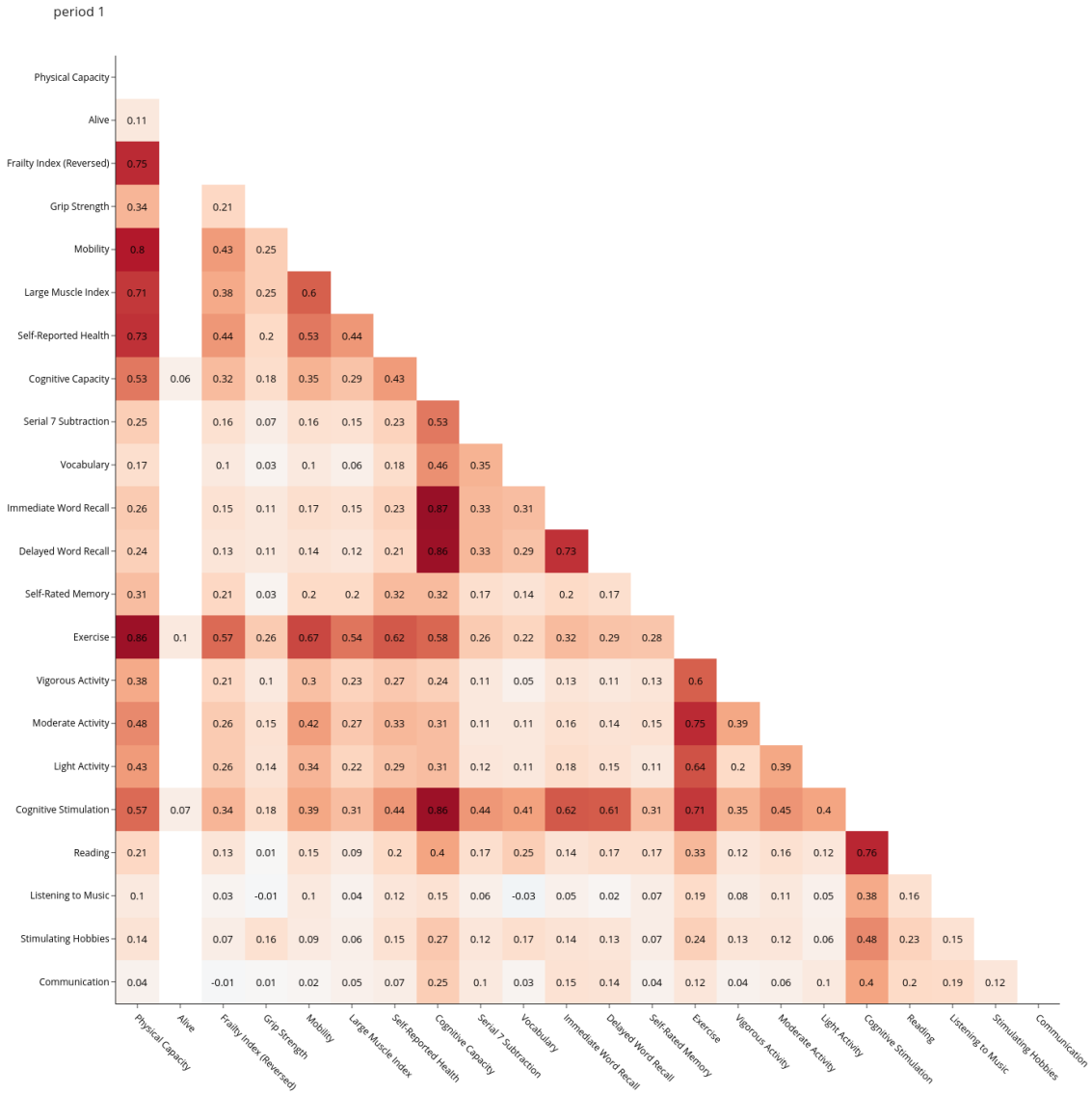


Figure 1.D.1. Correlations across implied factors and measurement correlations – females aged 70

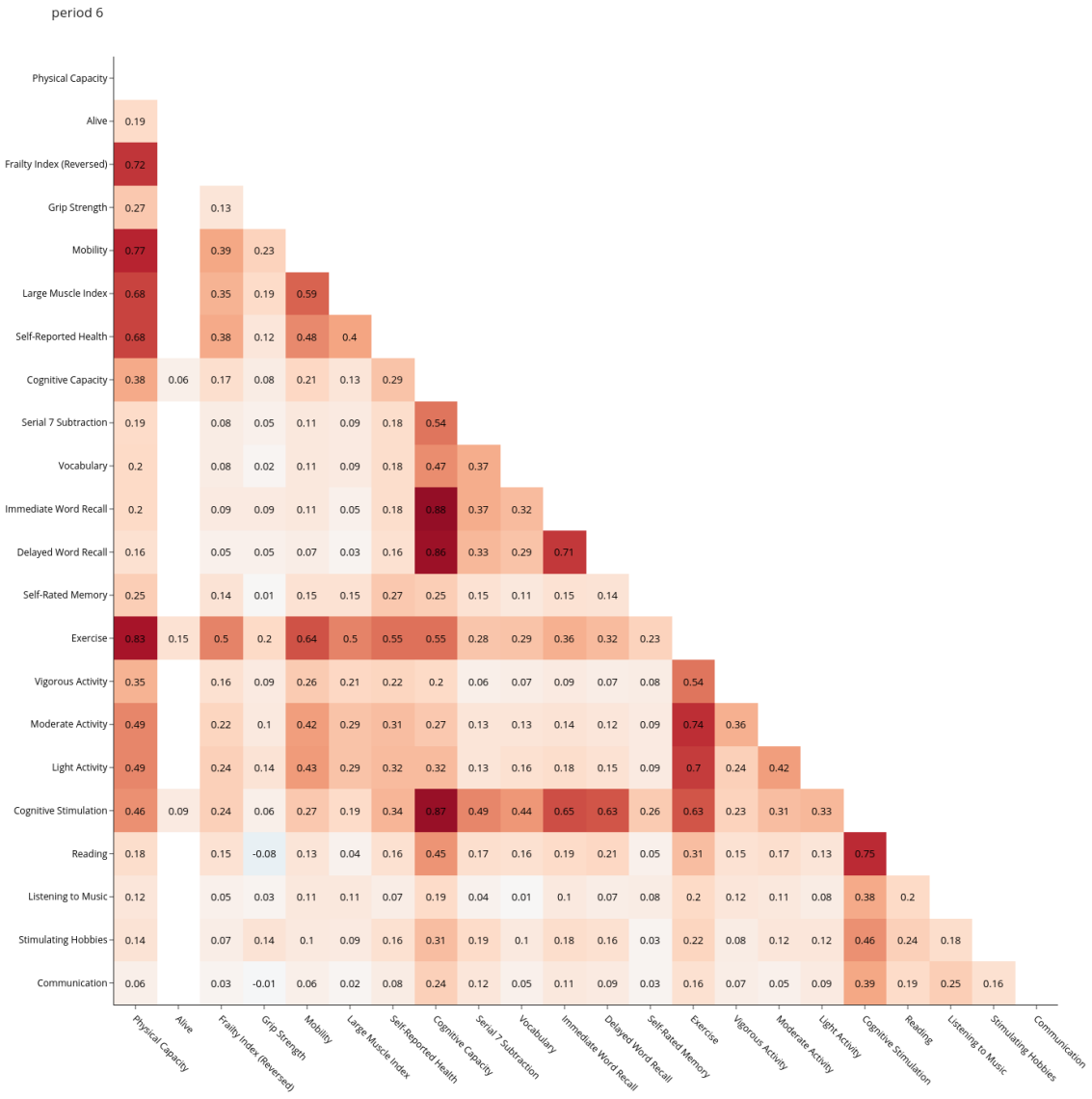


Figure 1.D.2. Correlations across implied factors and measurement correlations – females aged 80

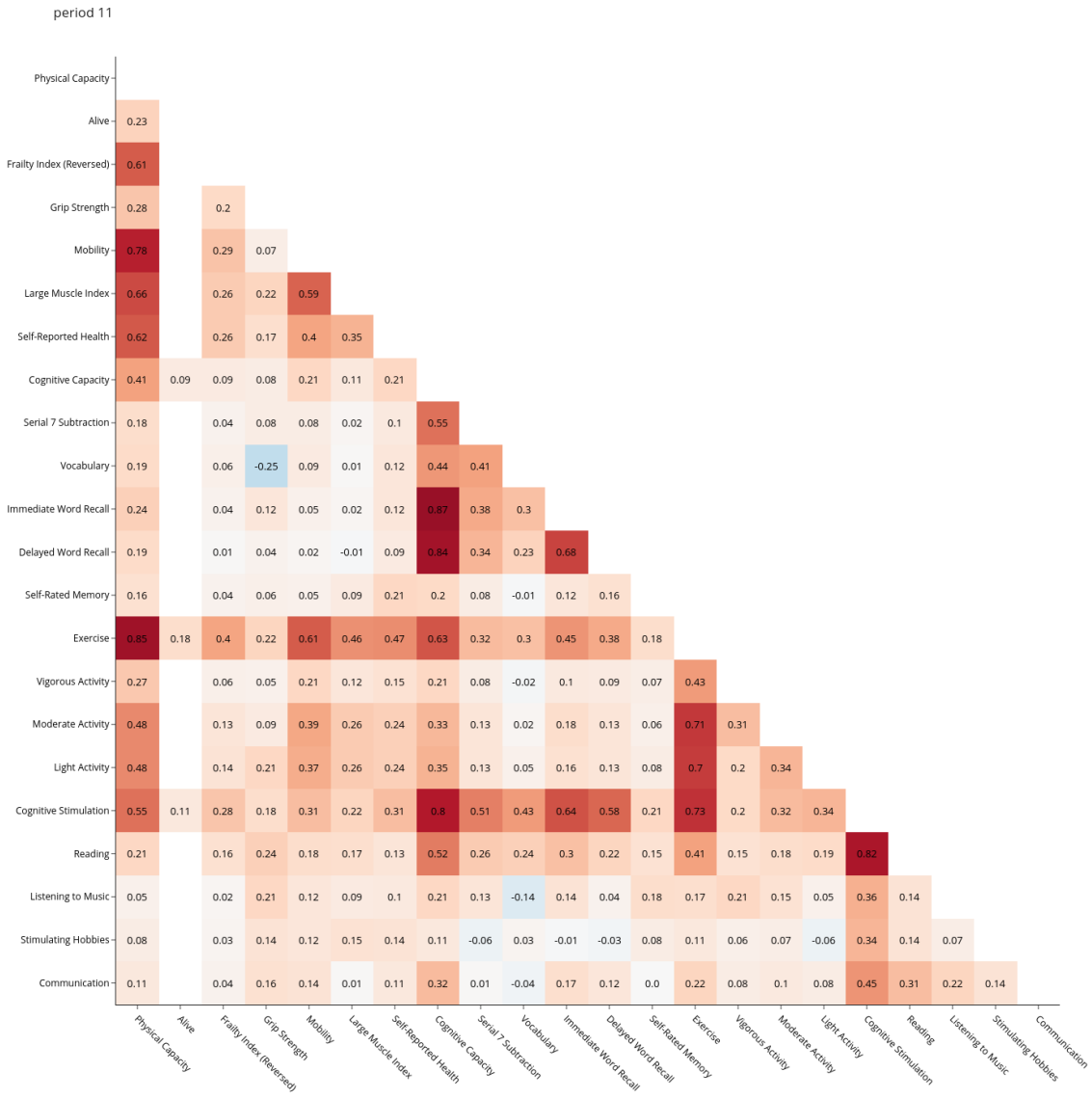


Figure 1.D.3. Correlations across implied factors and measurement correlations – females aged 90

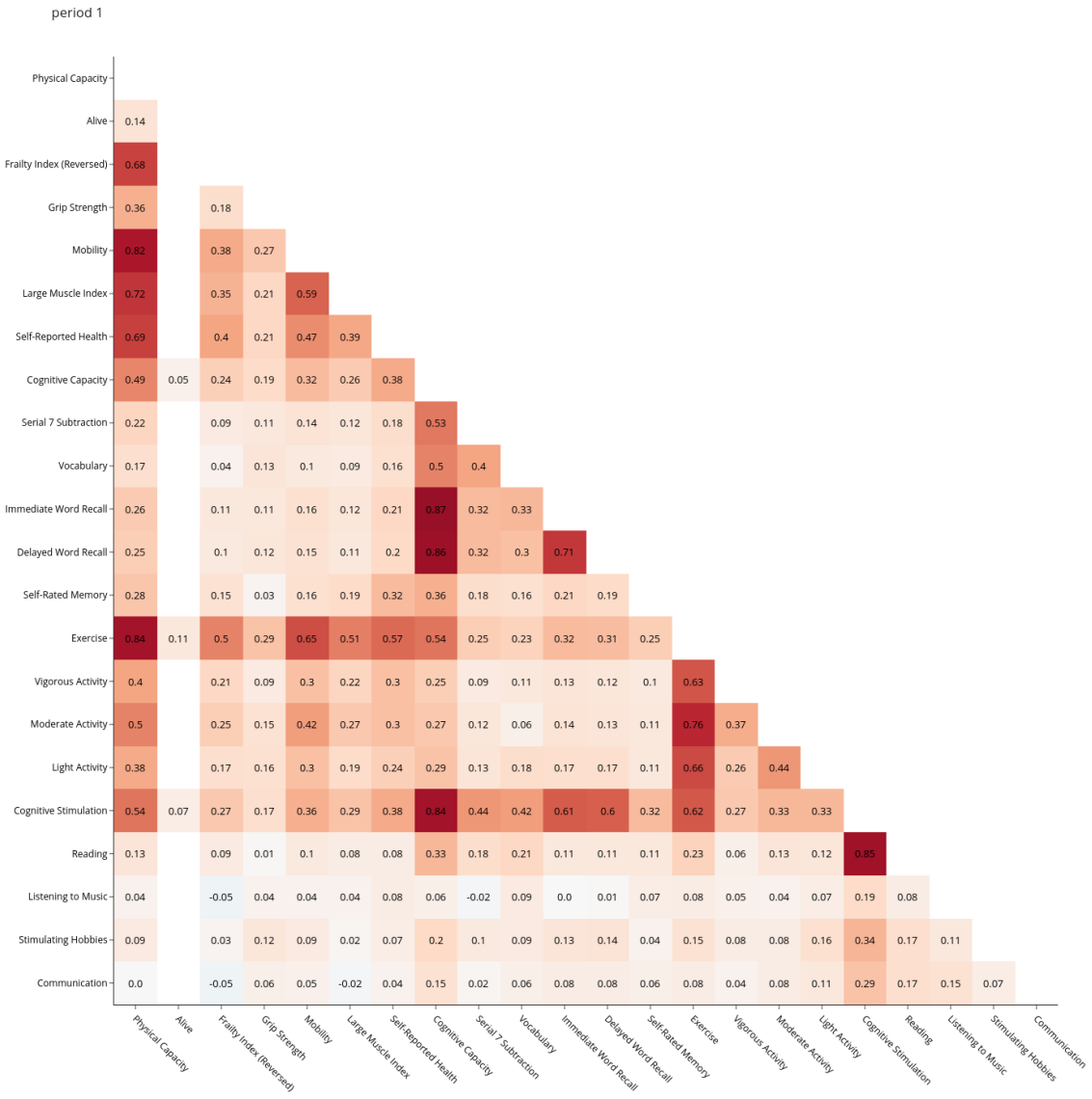


Figure 1.D.4. Correlations across implied factors and measurement correlations – males aged 70

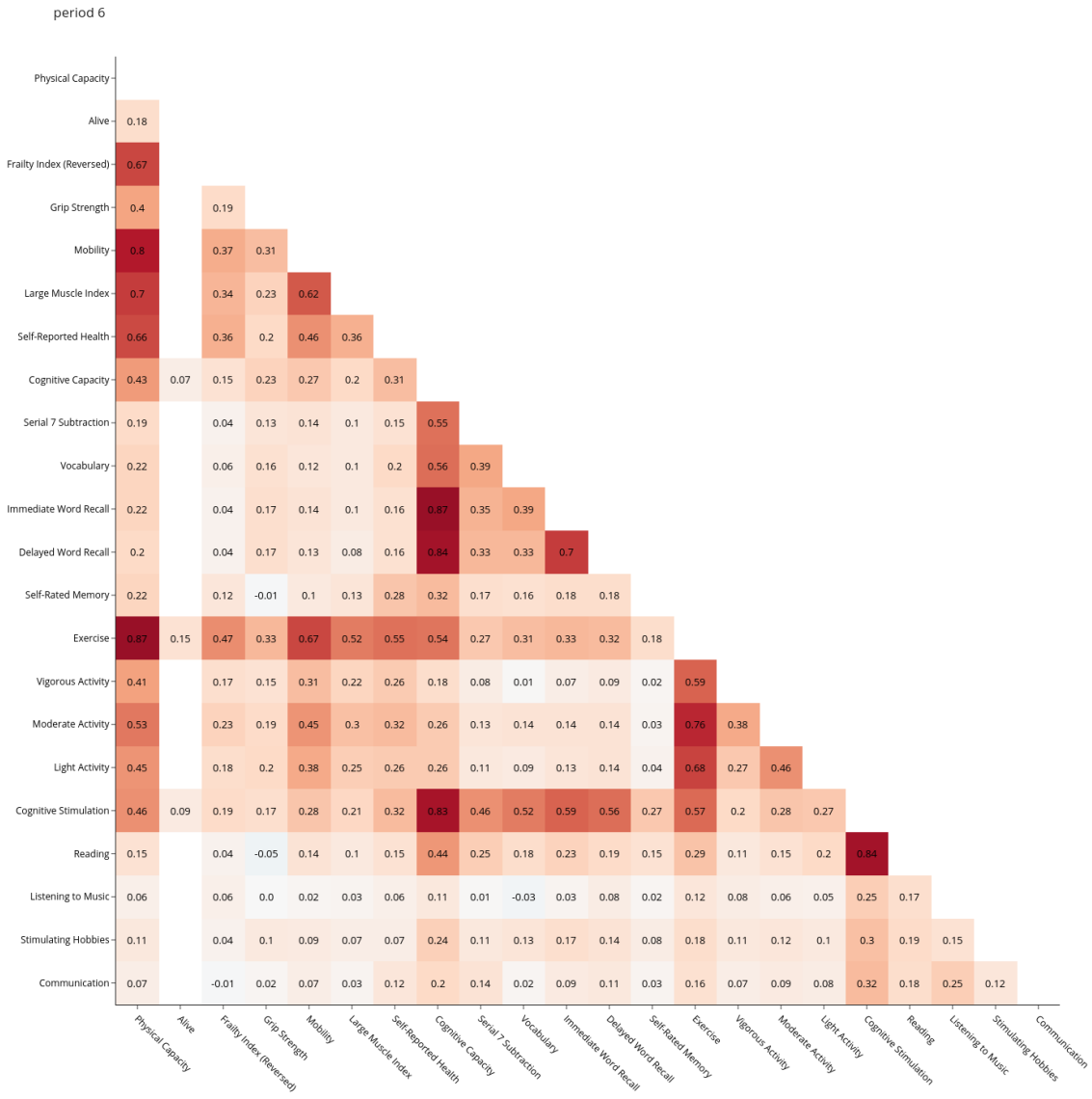


Figure 1.D.5. Correlations across implied factors and measurement correlations – males aged 80

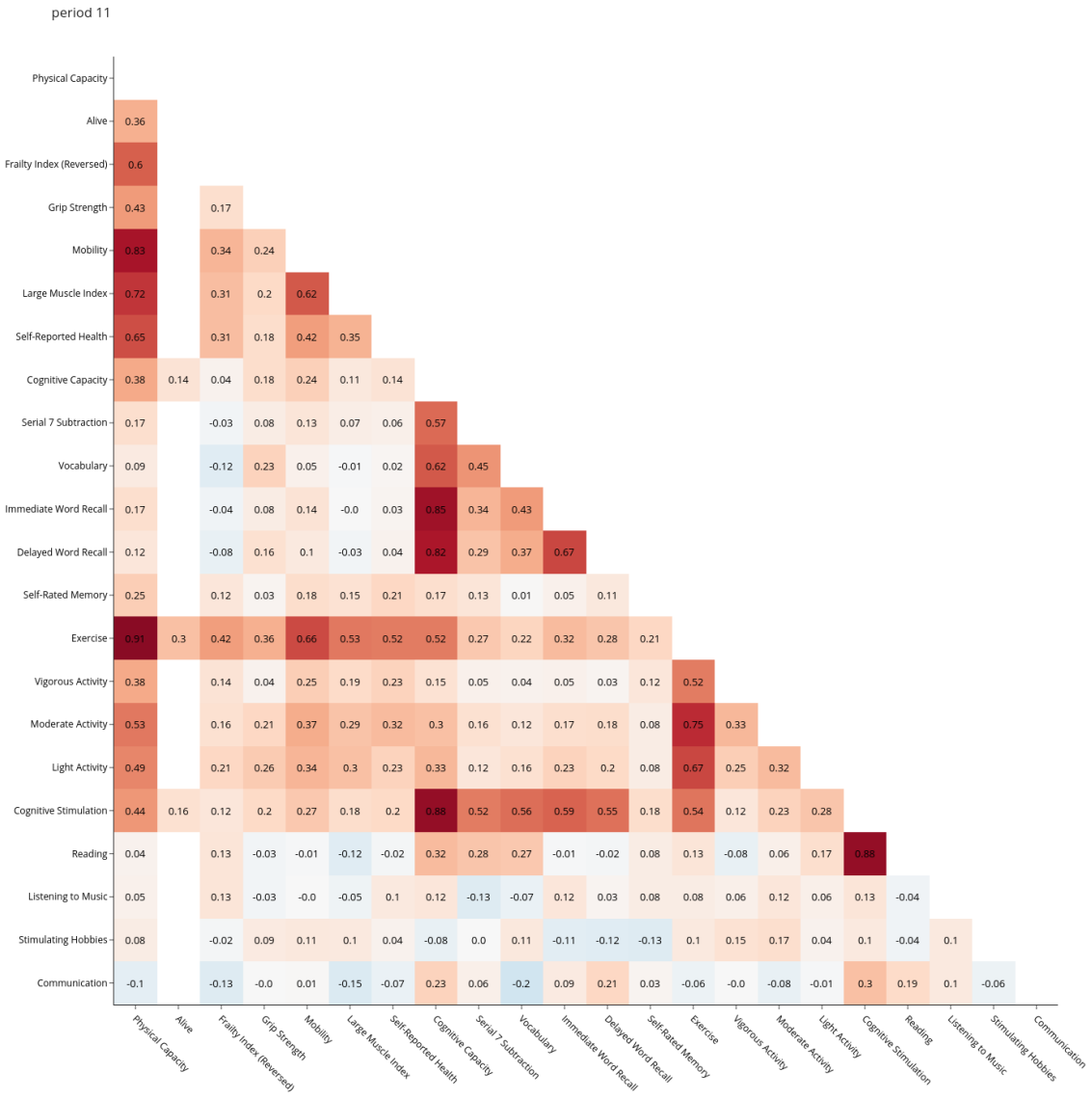


Figure 1.D.6. Correlations across implied factors and measurement correlations – males aged 90

1.D.3 Factor Distributions

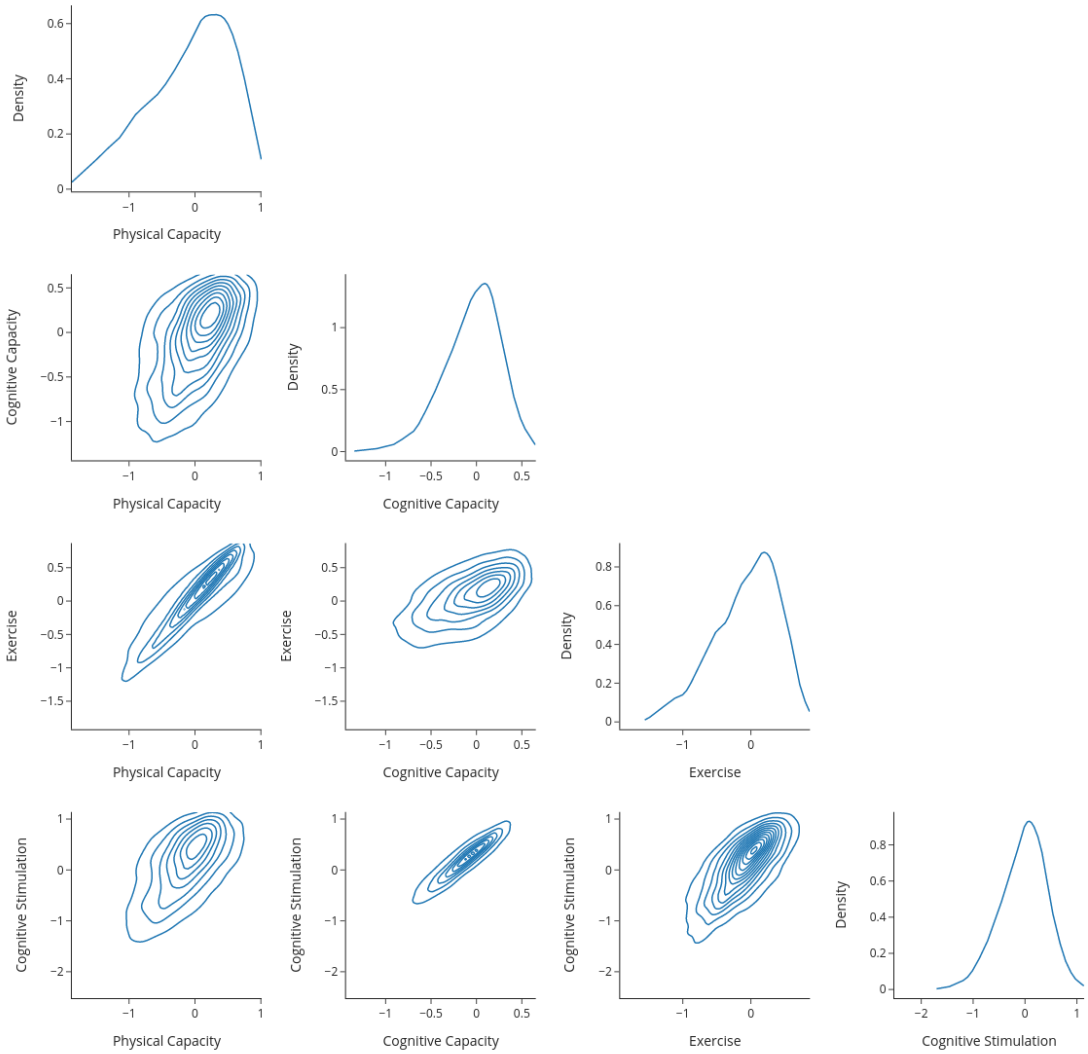


Figure 1.D.7. Factor distributions – females aged 70

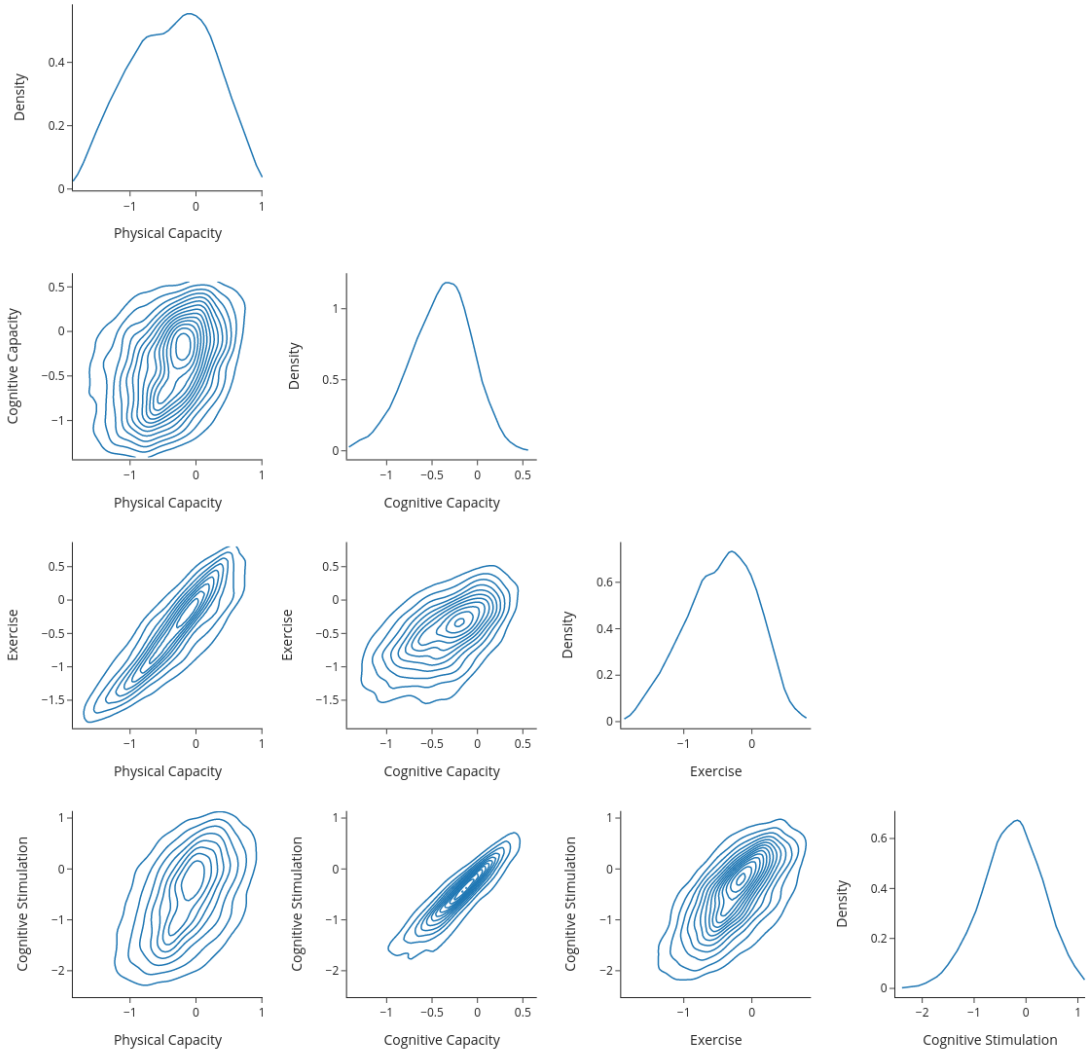


Figure 1.D.8. Factor distributions – females aged 80

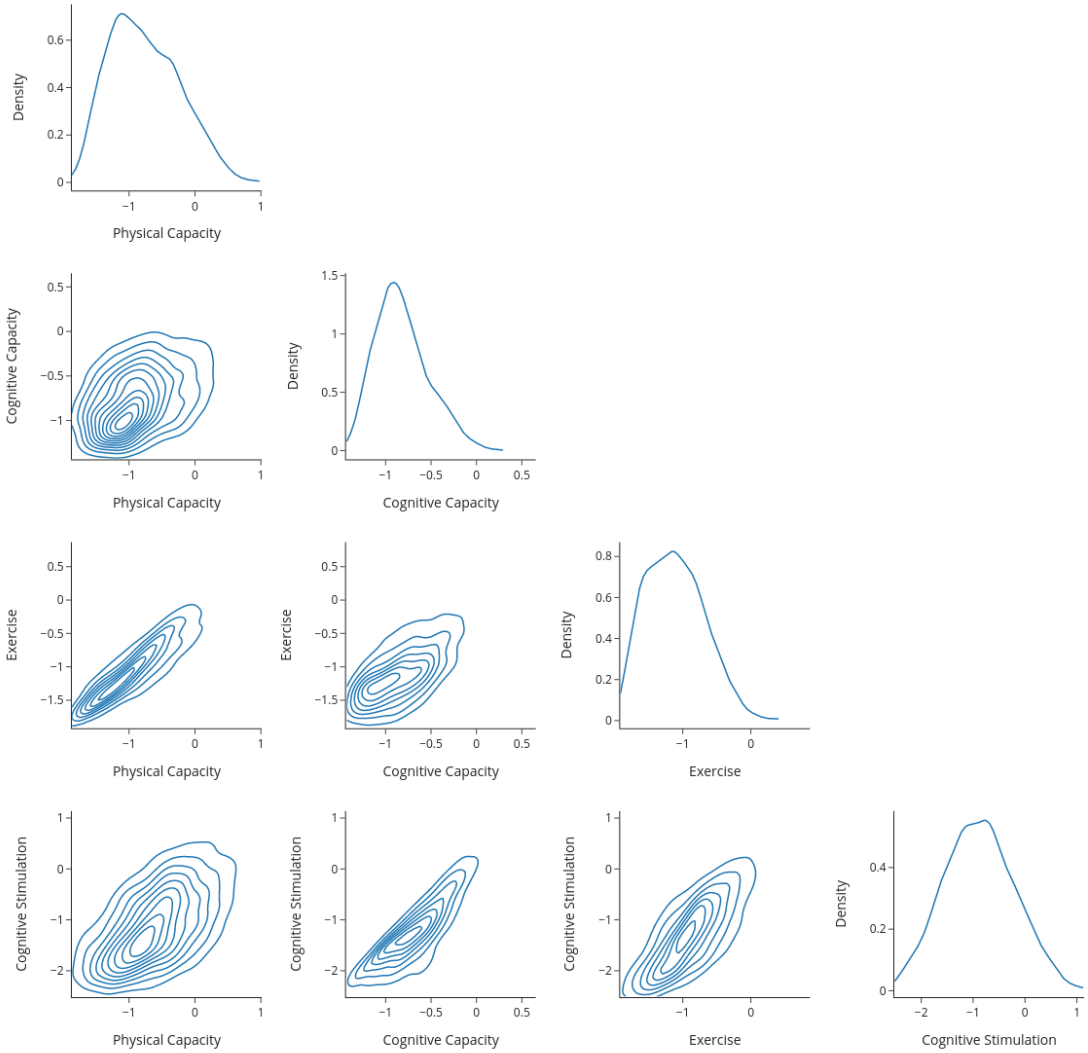


Figure 1.D.9. Factor distributions – females aged 90

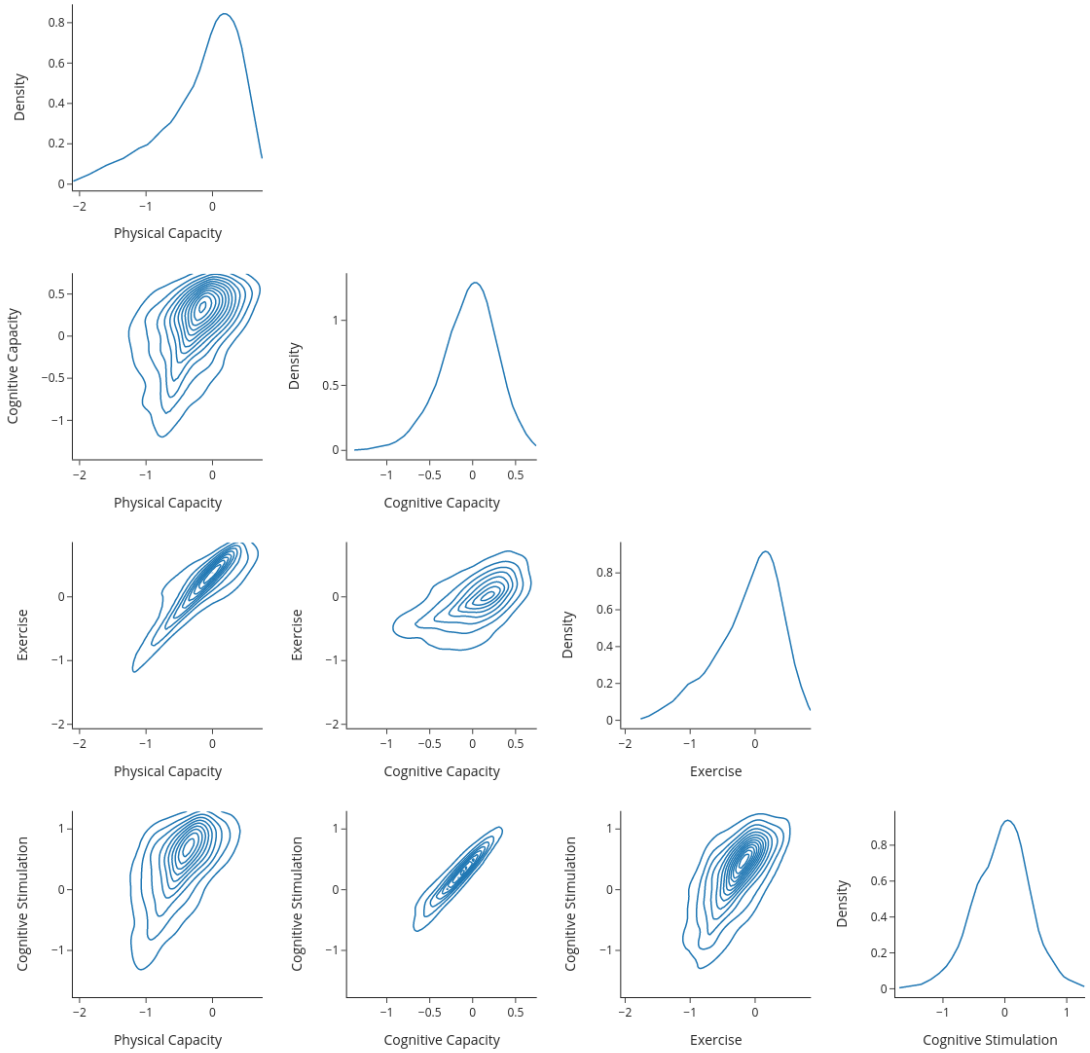


Figure 1.D.10. Factor distributions – males aged 70

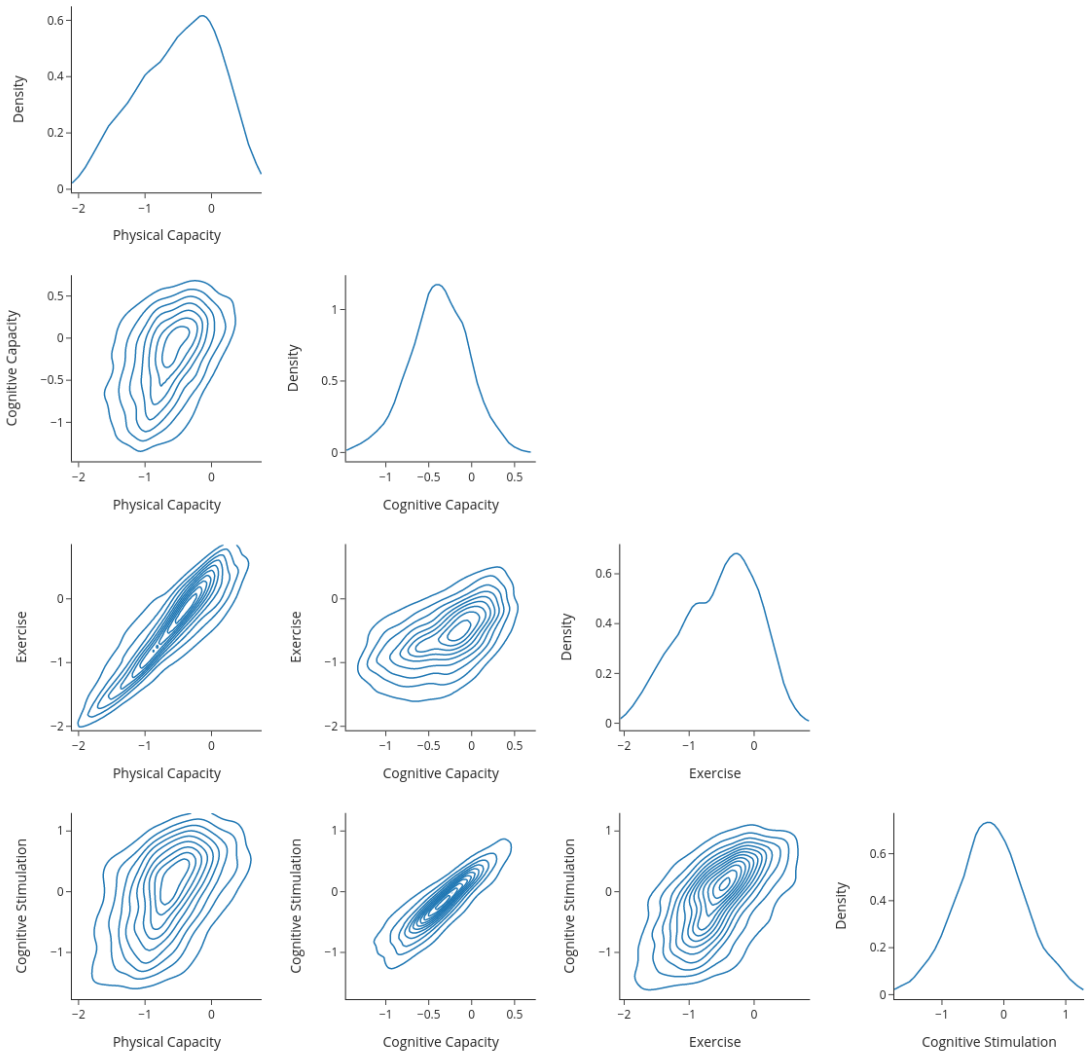


Figure 1.D.11. Factor distributions – males aged 80

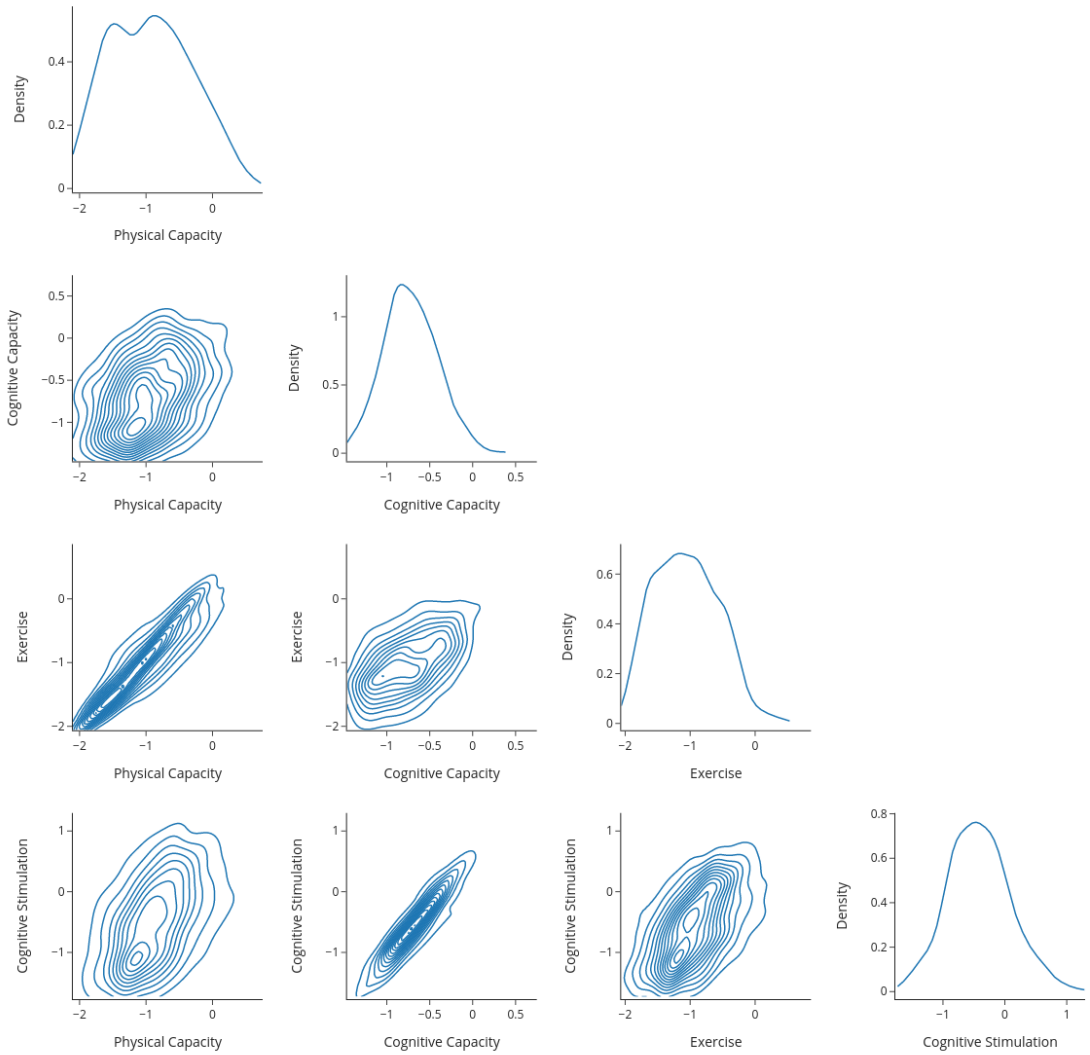


Figure 1.D.12. Factor distributions – males aged 90

1.D.4 Transition Equations

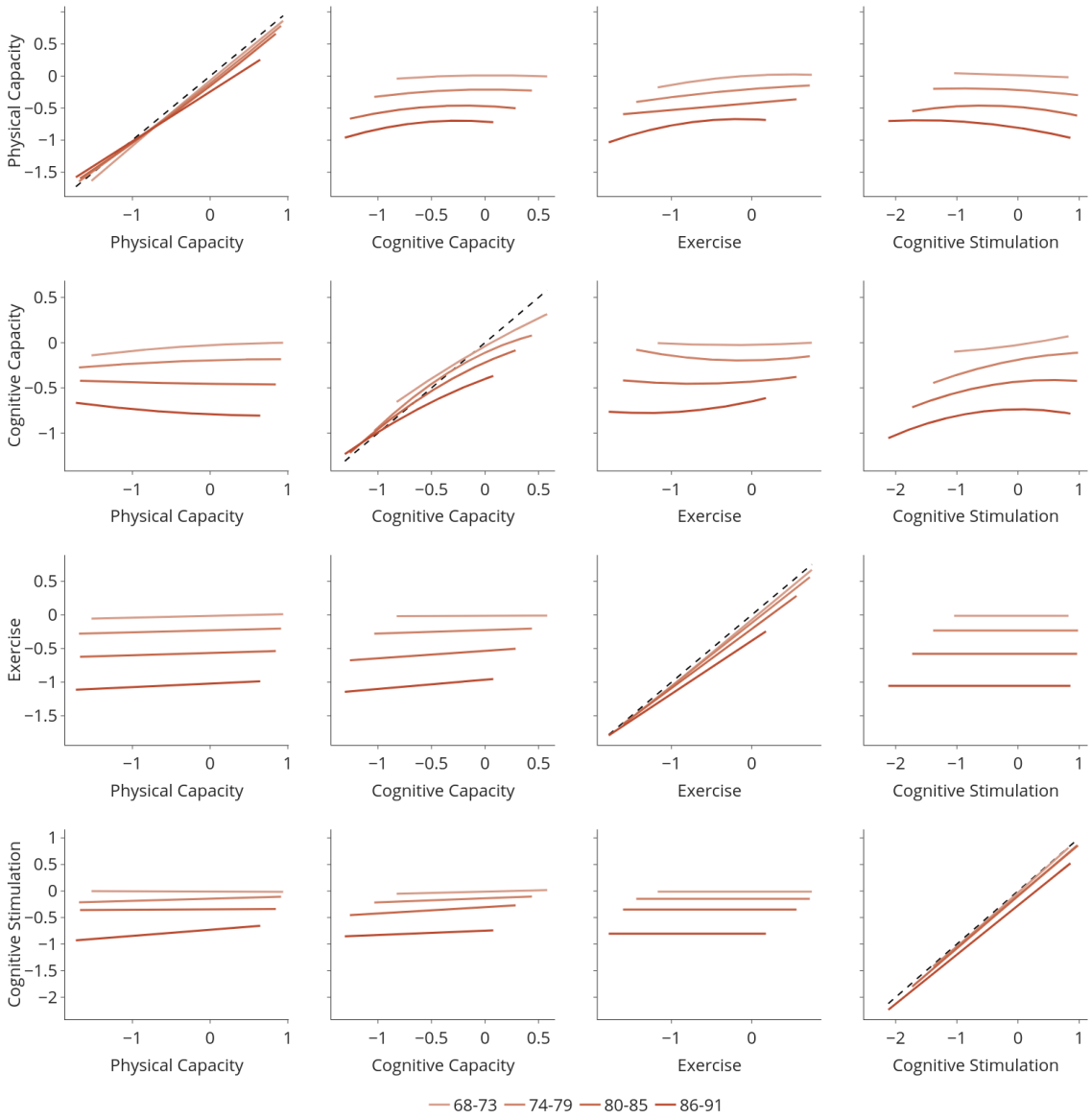


Figure 1.D.13. Transition equations for all factors (other factors evaluated at the median), females

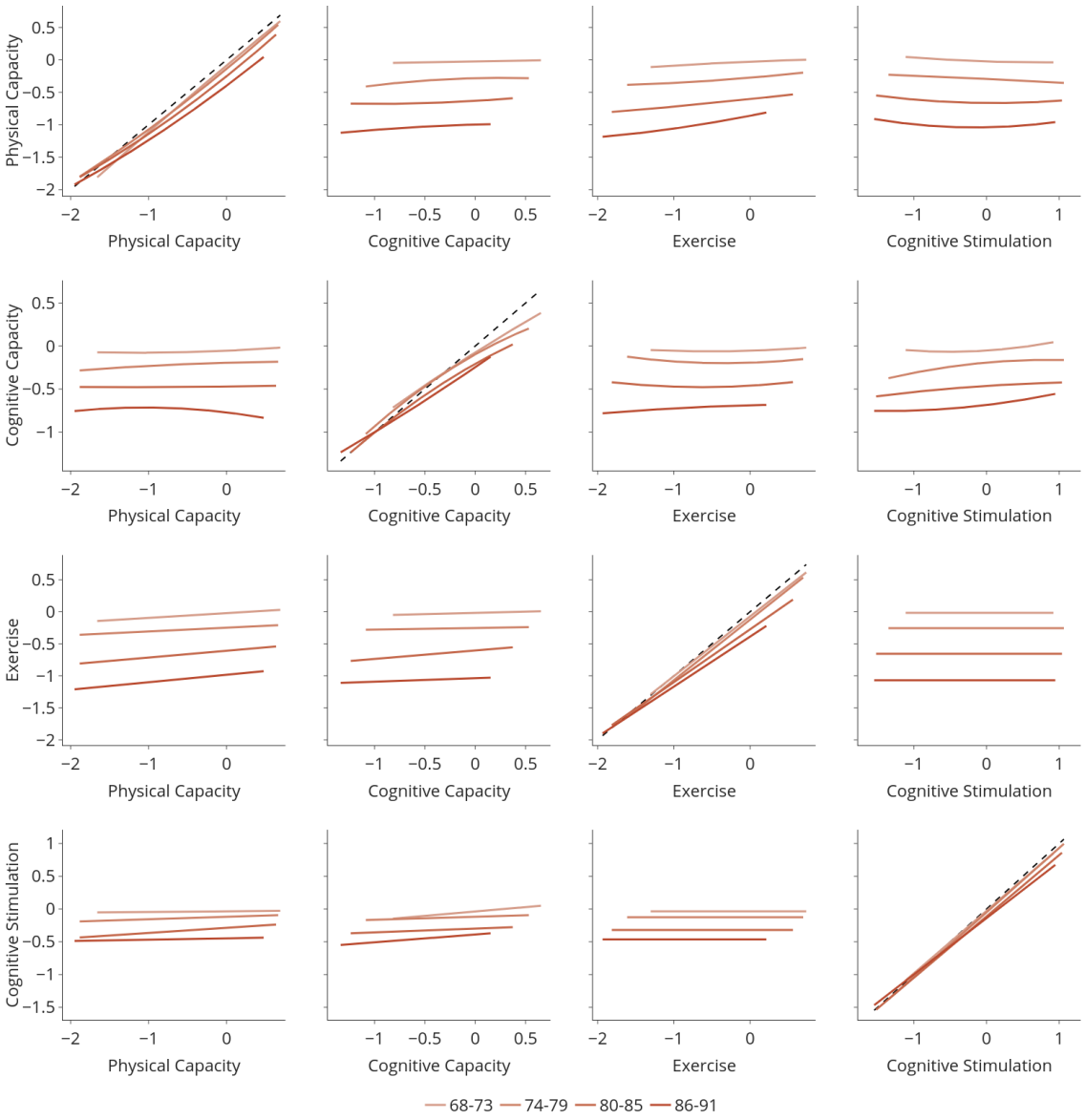


Figure 1.D.14. Transition equations for all factors (other factors evaluated at the median), males

Table 1.D.9. Transition Parameters for Physical Capacity, Females

	68-73	74-79	80-85	86-91
Physical Capacity	0.999***	0.972***	0.969***	0.994***
	(0.006)	(0.008)	(0.014)	(0.035)
Cognitive Capacity	0.007	0.003	−0.086***	−0.093
	(0.008)	(0.011)	(0.023)	(0.069)
Exercise	0.066***	0.074***	0.054***	−0.022
	(0.007)	(0.009)	(0.015)	(0.043)
Cognitive Stimulation	−0.035***	−0.025***	0.039***	0.021
	(0.006)	(0.008)	(0.015)	(0.034)
Physical Capacity Squared	−0.015	0.030***	0.041***	0.002
	(0.009)	(0.012)	(0.015)	(0.030)
Cognitive Capacity Squared	−0.066***	−0.097***	−0.180***	−0.235***
	(0.015)	(0.019)	(0.031)	(0.067)
Exercise Squared	−0.074***	−0.031*	−0.001	−0.136***
	(0.016)	(0.016)	(0.021)	(0.040)
Cognitive Stimulation Squared	−0.004	−0.029**	−0.066***	−0.045**
	(0.013)	(0.012)	(0.015)	(0.021)
Physical Capacity × Cognitive Capacity	0.069***	0.099***	0.142***	0.221***
	(0.018)	(0.021)	(0.032)	(0.065)
Physical Capacity × Exercise	0.075***	0.024	0.014	0.161***
	(0.019)	(0.021)	(0.027)	(0.054)
Physical Capacity × Cognitive Stimulation	−0.074***	−0.036**	−0.080***	−0.115***
	(0.015)	(0.016)	(0.021)	(0.036)
Cognitive Capacity × Exercise	−0.099***	−0.211***	−0.267***	−0.254***
	(0.024)	(0.026)	(0.040)	(0.071)
Cognitive Capacity × Cognitive Stimulation	0.077***	0.148***	0.187***	0.170***
	(0.022)	(0.025)	(0.037)	(0.060)
Exercise × Cognitive Stimulation	0.085***	0.094***	0.170***	0.148***
	(0.024)	(0.020)	(0.027)	(0.043)
Constant	−0.072***	−0.104***	−0.120***	−0.089***
	(0.006)	(0.007)	(0.011)	(0.025)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.D.10. Transition Parameters for Physical Capacity, Males

	68-73	74-79	80-85	86-91
Physical Capacity	1.010***	0.989***	0.997***	0.939***
	(0.008)	(0.011)	(0.023)	(0.062)
Cognitive Capacity	0.034***	0.036***	0.034	−0.031
	(0.009)	(0.013)	(0.035)	(0.113)
Exercise	0.051***	0.082***	0.070***	0.133*
	(0.006)	(0.009)	(0.020)	(0.070)
Cognitive Stimulation	−0.039***	−0.037***	−0.029	−0.048
	(0.007)	(0.009)	(0.022)	(0.065)
Physical Capacity Squared	−0.020*	0.057***	0.073***	0.062
	(0.011)	(0.013)	(0.024)	(0.054)
Cognitive Capacity Squared	0.004	−0.074***	0.047	−0.037
	(0.018)	(0.021)	(0.046)	(0.112)
Exercise Squared	−0.012	0.019	0.009	0.032
	(0.012)	(0.014)	(0.025)	(0.062)
Cognitive Stimulation Squared	0.024***	−0.005	0.044**	0.067*
	(0.009)	(0.011)	(0.021)	(0.035)
Physical Capacity × Cognitive Capacity	−0.010	0.024	0.131***	0.120
	(0.019)	(0.024)	(0.047)	(0.114)
Physical Capacity × Exercise	−0.006	−0.045**	0.002	−0.019
	(0.017)	(0.020)	(0.039)	(0.088)
Physical Capacity × Cognitive Stimulation	0.006	0.026	−0.074**	−0.052
	(0.015)	(0.017)	(0.032)	(0.071)
Cognitive Capacity × Exercise	−0.078***	−0.073***	−0.207***	−0.152
	(0.019)	(0.025)	(0.050)	(0.119)
Cognitive Capacity × Cognitive Stimulation	−0.026	0.076***	−0.088*	−0.088
	(0.021)	(0.024)	(0.050)	(0.105)
Exercise × Cognitive Stimulation	0.047***	0.011	0.101***	0.027
	(0.016)	(0.017)	(0.034)	(0.080)
Constant	−0.090***	−0.124***	−0.191***	−0.246***
	(0.007)	(0.008)	(0.016)	(0.044)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.D.11. Transition Parameters for Cognitive Capacity, Females

	68-73	74-79	80-85	86-91
Physical Capacity	0.042***	0.046***	0.011	0.016
	(0.008)	(0.009)	(0.013)	(0.028)
Cognitive Capacity	0.664***	0.568***	0.600***	0.521***
	(0.010)	(0.012)	(0.019)	(0.052)
Exercise	0.013	−0.011	0.038**	0.039
	(0.011)	(0.011)	(0.016)	(0.037)
Cognitive Stimulation	0.100***	0.151***	0.129***	0.194***
	(0.010)	(0.010)	(0.014)	(0.027)
Physical Capacity Squared	−0.019	−0.014	0.005	0.022
	(0.013)	(0.014)	(0.016)	(0.024)
Cognitive Capacity Squared	−0.096***	−0.274***	−0.210***	−0.150***
	(0.018)	(0.023)	(0.027)	(0.058)
Exercise Squared	0.025	0.067***	0.047**	0.072**
	(0.026)	(0.022)	(0.024)	(0.036)
Cognitive Stimulation Squared	0.025	−0.045***	−0.058***	−0.069***
	(0.019)	(0.016)	(0.015)	(0.019)
Physical Capacity × Cognitive Capacity	0.064***	0.203***	0.132***	0.218***
	(0.021)	(0.024)	(0.033)	(0.058)
Physical Capacity × Exercise	0.030	−0.022	−0.034	−0.056
	(0.029)	(0.026)	(0.030)	(0.046)
Physical Capacity × Cognitive Stimulation	−0.002	−0.070***	−0.051**	−0.081**
	(0.022)	(0.021)	(0.022)	(0.033)
Cognitive Capacity × Exercise	−0.044	−0.178***	−0.056	−0.243***
	(0.029)	(0.030)	(0.038)	(0.072)
Cognitive Capacity × Cognitive Stimulation	0.015	0.227***	0.199***	0.262***
	(0.027)	(0.031)	(0.033)	(0.054)
Exercise × Cognitive Stimulation	−0.070**	0.029	0.016	0.080*
	(0.034)	(0.028)	(0.029)	(0.042)
Constant	−0.045***	−0.098***	−0.163***	−0.263***
	(0.009)	(0.010)	(0.011)	(0.020)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.D.12. Transition Parameters for Cognitive Capacity, Males

	68-73	74-79	80-85	86-91
Physical Capacity	0.040***	0.030**	0.027	−0.041
	(0.010)	(0.012)	(0.017)	(0.045)
Cognitive Capacity	0.738***	0.668***	0.699***	0.760***
	(0.011)	(0.013)	(0.026)	(0.075)
Exercise	0.029***	0.019*	0.025	0.090*
	(0.010)	(0.011)	(0.016)	(0.053)
Cognitive Stimulation	0.059***	0.099***	0.099***	0.020
	(0.010)	(0.010)	(0.016)	(0.039)
Physical Capacity Squared	0.018	−0.012	0.006	−0.052
	(0.017)	(0.018)	(0.023)	(0.044)
Cognitive Capacity Squared	−0.054***	−0.177***	−0.141***	0.044
	(0.018)	(0.024)	(0.040)	(0.088)
Exercise Squared	0.026	0.046**	0.040*	−0.013
	(0.021)	(0.020)	(0.023)	(0.057)
Cognitive Stimulation Squared	0.054***	−0.044***	−0.017	0.037
	(0.014)	(0.014)	(0.017)	(0.025)
Physical Capacity × Cognitive Capacity	0.043*	0.062**	0.071*	−0.023
	(0.025)	(0.031)	(0.043)	(0.087)
Physical Capacity × Exercise	−0.001	−0.008	−0.030	0.092
	(0.030)	(0.031)	(0.034)	(0.081)
Physical Capacity × Cognitive Stimulation	−0.000	−0.028	0.006	0.006
	(0.022)	(0.023)	(0.028)	(0.055)
Cognitive Capacity × Exercise	−0.039	−0.083***	−0.033	0.016
	(0.028)	(0.031)	(0.042)	(0.092)
Cognitive Capacity × Cognitive Stimulation	−0.068***	0.162***	0.114***	−0.098
	(0.025)	(0.027)	(0.042)	(0.076)
Exercise × Cognitive Stimulation	−0.023	0.031	−0.003	−0.031
	(0.027)	(0.025)	(0.029)	(0.059)
Constant	−0.079***	−0.082***	−0.164***	−0.209***
	(0.010)	(0.011)	(0.014)	(0.028)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.D.13. Transition Parameters for Exercise, Females

	68-73	74-79	80-85	86-91
Physical Capacity	0.026***	0.029***	0.035**	0.053**
	(0.010)	(0.011)	(0.014)	(0.021)
Cognitive Capacity	0.006	0.050***	0.110***	0.138***
	(0.011)	(0.011)	(0.015)	(0.027)
Exercise	0.990***	0.941***	0.880***	0.790***
	(0.014)	(0.014)	(0.018)	(0.027)
Constant	−0.074***	−0.109***	−0.155***	−0.258***
	(0.004)	(0.005)	(0.008)	(0.017)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.D.14. Transition Parameters for Exercise, Males

	68-73	74-79	80-85	86-91
Physical Capacity	0.075***	0.059***	0.106***	0.117***
	(0.012)	(0.013)	(0.021)	(0.038)
Cognitive Capacity	0.038***	0.025*	0.132***	0.056
	(0.013)	(0.014)	(0.022)	(0.044)
Exercise	0.933***	0.945***	0.825***	0.782***
	(0.015)	(0.015)	(0.022)	(0.047)
Constant	−0.078***	−0.111***	−0.178***	−0.266***
	(0.005)	(0.006)	(0.011)	(0.025)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.D.15. Transition Parameters for Cognitive Stimulation, Females

	68-73	74-79	80-85	86-91
Physical Capacity	−0.006	0.041***	0.008	0.116***
	(0.013)	(0.014)	(0.020)	(0.041)
Cognitive Capacity	0.050**	0.076***	0.120***	0.081
	(0.022)	(0.023)	(0.035)	(0.072)
Cognitive Stimulation	1.020***	0.962***	0.985***	0.927***
	(0.018)	(0.017)	(0.024)	(0.046)
Constant	−0.033***	−0.071***	−0.046***	−0.144***
	(0.007)	(0.009)	(0.014)	(0.035)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.D.16. Transition Parameters for Cognitive Stimulation, Males

	68-73	74-79	80-85	86-91
Physical Capacity	0.011	0.037**	0.079**	0.020
	(0.018)	(0.019)	(0.032)	(0.080)
Cognitive Capacity	0.134***	0.046	0.057	0.119
	(0.026)	(0.031)	(0.050)	(0.141)
Cognitive Stimulation	0.953***	0.982***	0.936***	0.858***
	(0.018)	(0.019)	(0.030)	(0.067)
Constant	−0.033***	−0.039***	−0.056**	−0.051
	(0.009)	(0.011)	(0.022)	(0.070)
Note:				***p<0.01;**p<0.05;*p<0.1

1.D.5 Distributions of Initial Factors and of Shocks to Factors

Table 1.D.17. Distribution of the initial states, females

			Correlation with			
Factor	Mean	Standard Deviation	Physical Capacity	Cognitive Capacity	Exercise	Cognitive Stimulation
Physical Capacity	0.19	0.68	1.00	0.35	0.66	0.36
Cognitive Capacity	0.11	0.46	0.35	1.00	0.32	0.52
Exercise	0.15	0.59	0.66	0.32	1.00	0.51
Cognitive Stimulation	0.08	0.68	0.36	0.52	0.51	1.00

Table 1.D.18. Distribution of the intitial states, males

			Correlation with			
Factor	Mean	Standard Deviation	Physical Capacity	Cognitive Capacity	Exercise	Cognitive Stimulation
Physical Capacity	0.10	0.61	1.00	0.30	0.58	0.30
Cognitive Capacity	0.11	0.49	0.30	1.00	0.27	0.42
Exercise	0.12	0.64	0.58	0.27	1.00	0.33
Cognitive Stimulation	0.04	0.79	0.30	0.42	0.33	1.00

Table 1.D.19. Standard deviations of shocks

		Male	Female
68-73	Physical Capacity	0.094***	0.005
		(0.008)	(0.103)
	Cognitive Capacity	0.292***	0.308***
		(0.005)	(0.004)
	Exercise	0.236***	0.164***
		(0.012)	(0.011)
	Cognitive Stimulation	0.155***	0.001
		(0.028)	(2.725)
74-79	Physical Capacity	0.161***	0.159***
		(0.006)	(0.005)
	Cognitive Capacity	0.283***	0.302***
		(0.005)	(0.004)
	Exercise	0.261***	0.240***
		(0.012)	(0.009)
	Cognitive Stimulation	0.190***	0.185***
		(0.026)	(0.017)
80-85	Physical Capacity	0.231***	0.188***
		(0.009)	(0.008)
	Cognitive Capacity	0.240***	0.274***
		(0.006)	(0.005)
	Exercise	0.326***	0.275***
		(0.015)	(0.012)
	Cognitive Stimulation	0.319***	0.225***
		(0.031)	(0.024)
86-91	Physical Capacity	0.315***	0.227***
		(0.020)	(0.012)
	Cognitive Capacity	0.250***	0.238***
		(0.013)	(0.010)
	Exercise	0.372***	0.304***
		(0.028)	(0.017)
	Cognitive Stimulation	0.471***	0.309***
		(0.059)	(0.047)
Note:		***p<0.01;**p<0.05;*p<0.1	

Appendix 1.E Results for a Linearized Model

1.E.1 Measurement System

Table 1.E.1. Loadings and Measurement Standard Deviations for Physical Capacity, Females

		Intercept	Loading	Meas. Std.
All	Frailty Index (Reversed)	0.000	1.000	0.705***
				(0.001)
	Mobility	−0.114***	1.222***	0.768***
		(0.003)	(0.005)	(0.002)
	Large Muscle Index	0.005*	0.926***	0.750***
		(0.003)	(0.005)	(0.002)
	Self-Reported Health	−0.048***	0.947***	0.765***
		(0.003)	(0.004)	(0.002)
70	Alive	0.897***	0.042***	0.303***
		(0.101)	(0.011)	(0.038)
	Grip Strength	−0.125***	0.488***	0.933***
		(0.027)	(0.042)	(0.015)
72	Alive	0.910***	0.045***	0.288***
		(0.106)	(0.011)	(0.037)
	Grip Strength	−0.240***	0.395***	0.922***
		(0.028)	(0.042)	(0.016)
74	Alive	0.902***	0.060***	0.301***
		(0.096)	(0.013)	(0.036)
	Grip Strength	−0.291***	0.464***	0.936***
		(0.030)	(0.042)	(0.018)
76	Alive	0.886***	0.073***	0.327***
		(0.099)	(0.018)	(0.042)
	Grip Strength	−0.470***	0.367***	0.924***
		(0.030)	(0.048)	(0.012)
78	Alive	0.879***	0.075***	0.339***
		(0.101)	(0.019)	(0.045)
	Grip Strength	−0.540***	0.445***	0.924***
		(0.033)	(0.048)	(0.019)
80	Alive	0.870***	0.091***	0.353***
		(0.097)	(0.022)	(0.046)
	Grip Strength	−0.758***	0.365***	0.882***
		(0.034)	(0.052)	(0.021)
82	Alive	0.871***	0.090***	0.359***
		(0.109)	(0.026)	(0.053)
	Grip Strength	−0.789***	0.339***	0.860***
		(0.036)	(0.054)	(0.020)
84	Alive	0.869***	0.110***	0.371***
		(0.103)	(0.030)	(0.052)
	Grip Strength	−0.979***	0.336***	0.866***
		(0.041)	(0.060)	(0.025)
86	Alive	0.855***	0.120***	0.391***
		(0.124)	(0.040)	(0.069)
	Grip Strength	−0.999***	0.332***	0.840***
		(0.046)	(0.070)	(0.028)
88	Alive	0.845***	0.128**	0.406***
		(0.142)	(0.051)	(0.084)
	Grip Strength	−1.190***	0.415***	0.826***
		(0.059)	(0.082)	(0.035)
90	Alive	0.826***	0.133*	0.425***
		(0.204)	(0.080)	(0.135)
	Grip Strength	−1.099***	0.371***	0.734***
		(0.061)	(0.097)	(0.031)
92	Alive	0.816***	0.164	0.444***
		(0.228)	(0.120)	(0.159)
	Grip Strength	−1.357***	0.356***	0.745***
		(0.083)	(0.115)	(0.047)

Note: ***p<0.01;**p<0.05;*p<0.1

Table 1.E.2. Loadings and Measurement Standard Deviations for Physical Capacity, Males

		Intercept	Loading	Meas. Std.
All	Frailty Index (Reversed)	0.000	1.000	0.796***
				(0.002)
	Mobility	−0.015***	1.330***	0.751***
		(0.005)	(0.007)	(0.003)
	Large Muscle Index	0.043***	1.033***	0.761***
		(0.004)	(0.006)	(0.003)
	Self-Reported Health	0.027***	0.964***	0.792***
		(0.003)	(0.006)	(0.003)
70	Alive	0.901***	0.057***	0.303***
		(0.093)	(0.013)	(0.035)
	Grip Strength	−0.055	0.578***	0.977***
		(0.034)	(0.053)	(0.020)
72	Alive	0.907***	0.074***	0.298***
		(0.083)	(0.015)	(0.030)
	Grip Strength	−0.294***	0.549***	0.959***
		(0.034)	(0.053)	(0.020)
74	Alive	0.900***	0.061***	0.310***
		(0.119)	(0.017)	(0.046)
	Grip Strength	−0.317***	0.499***	0.922***
		(0.035)	(0.057)	(0.021)
76	Alive	0.876***	0.073***	0.344***
		(0.129)	(0.024)	(0.059)
	Grip Strength	−0.505***	0.557***	0.898***
		(0.036)	(0.056)	(0.020)
78	Alive	0.872***	0.081***	0.355***
		(0.128)	(0.026)	(0.061)
	Grip Strength	−0.559***	0.552***	0.920***
		(0.040)	(0.058)	(0.022)
80	Alive	0.866***	0.089***	0.367***
		(0.132)	(0.031)	(0.066)
	Grip Strength	−0.736***	0.573***	0.891***
		(0.042)	(0.062)	(0.023)
82	Alive	0.853***	0.136***	0.393***
		(0.114)	(0.042)	(0.064)
	Grip Strength	−0.959***	0.466***	0.873***
		(0.046)	(0.064)	(0.025)
84	Alive	0.869***	0.138***	0.387***
		(0.127)	(0.046)	(0.067)
	Grip Strength	−1.040***	0.561***	0.842***
		(0.052)	(0.068)	(0.027)
86	Alive	0.847***	0.137**	0.408***
		(0.158)	(0.061)	(0.092)
	Grip Strength	−1.237***	0.491***	0.841***
		(0.063)	(0.083)	(0.033)
88	Alive	0.858***	0.179**	0.416***
		(0.145)	(0.071)	(0.083)
	Grip Strength	−1.280***	0.480***	0.824***
		(0.069)	(0.107)	(0.044)
90	Alive	0.851***	0.204*	0.429***
		(0.201)	(0.116)	(0.120)
	Grip Strength	−1.361***	0.503***	0.767***
		(0.097)	(0.114)	(0.053)
92	Alive	0.765**	0.183	0.464*
		(0.317)	(0.220)	(0.271)
	Grip Strength	−1.487***	0.683***	0.817***
		(0.120)	(0.162)	(0.076)

Note: ***p<0.01;**p<0.05;*p<0.1

Table 1.E.3. Loadings and Measurement Standard Deviations for Cognitive Capacity, Females

		Intercept	Loading	Meas. Std.
All	Serial 7 Subtraction	0.000	1.000	0.890***
				(0.003)
	Vocabulary	0.044***	0.840***	0.923***
		(0.006)	(0.013)	(0.004)
	Immediate Word Recall	−0.161***	1.799***	0.584***
		(0.006)	(0.014)	(0.003)
	Delayed Word Recall	−0.189***	1.803***	0.595***
		(0.006)	(0.014)	(0.002)
70	Self-Rated Memory	0.005	0.577***	0.961***
		(0.014)	(0.031)	(0.009)
72	Self-Rated Memory	0.029**	0.595***	0.954***
		(0.014)	(0.030)	(0.009)
74	Self-Rated Memory	0.016	0.562***	0.972***
		(0.015)	(0.030)	(0.009)
76	Self-Rated Memory	0.028*	0.496***	0.968***
		(0.017)	(0.032)	(0.010)
78	Self-Rated Memory	0.046**	0.501***	0.992***
		(0.019)	(0.035)	(0.011)
80	Self-Rated Memory	0.054**	0.480***	1.012***
		(0.022)	(0.038)	(0.012)
82	Self-Rated Memory	0.069**	0.460***	1.009***
		(0.027)	(0.043)	(0.013)
84	Self-Rated Memory	0.082**	0.396***	1.035***
		(0.032)	(0.050)	(0.015)
86	Self-Rated Memory	0.079**	0.393***	1.063***
		(0.040)	(0.058)	(0.018)
88	Self-Rated Memory	0.261***	0.549***	1.069***
		(0.054)	(0.074)	(0.021)
90	Self-Rated Memory	0.213***	0.463***	1.080***
		(0.074)	(0.096)	(0.026)
92	Self-Rated Memory	0.215**	0.532***	1.146***
		(0.108)	(0.131)	(0.040)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.E.4. Loadings and Measurement Standard Deviations for Cognitive Capacity, Males

		Intercept	Loading	Meas. Std.
All	Serial 7 Subtraction	0.000	1.000	0.907***
				(0.004)
	Vocabulary	0.048***	0.962***	0.900***
		(0.007)	(0.015)	(0.004)
	Immediate Word Recall	−0.184***	1.683***	0.600***
		(0.008)	(0.015)	(0.003)
	Delayed Word Recall	−0.201***	1.647***	0.607***
		(0.008)	(0.015)	(0.003)
70	Self-Rated Memory	−0.041**	0.627***	0.937***
		(0.016)	(0.035)	(0.011)
72	Self-Rated Memory	−0.052***	0.563***	0.955***
		(0.017)	(0.034)	(0.011)
74	Self-Rated Memory	−0.043**	0.579***	0.948***
		(0.017)	(0.035)	(0.011)
76	Self-Rated Memory	−0.039**	0.528***	0.955***
		(0.019)	(0.039)	(0.012)
78	Self-Rated Memory	−0.050**	0.610***	0.971***
		(0.022)	(0.043)	(0.013)
80	Self-Rated Memory	−0.001	0.596***	0.988***
		(0.026)	(0.048)	(0.015)
82	Self-Rated Memory	−0.019	0.478***	1.033***
		(0.034)	(0.056)	(0.018)
84	Self-Rated Memory	−0.018	0.520***	1.007***
		(0.039)	(0.062)	(0.020)
86	Self-Rated Memory	−0.018	0.465***	0.992***
		(0.045)	(0.069)	(0.021)
88	Self-Rated Memory	0.007	0.511***	1.035***
		(0.063)	(0.088)	(0.027)
90	Self-Rated Memory	0.013	0.391***	1.080***
		(0.086)	(0.117)	(0.037)
92	Self-Rated Memory	0.006	0.602***	1.011***
		(0.124)	(0.180)	(0.048)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.E.5. Loadings and Measurement Standard Deviations for Exercise, Females

		Intercept	Loading	Meas. Std.
All	Vigorous Activity	−0.009	0.683***	0.808***
		(0.006)	(0.010)	(0.004)
	Moderate Activity	0.000	1.000	0.794***
				(0.004)
	Light Activity	−0.127***	1.077***	0.933***
		(0.007)	(0.012)	(0.004)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.E.6. Loadings and Measurement Standard Deviations for Exercise, Males

		Intercept	Loading	Meas. Std.
All	Vigorous Activity	−0.012**	0.742***	0.813***
		(0.006)	(0.012)	(0.005)
	Moderate Activity	0.000	1.000	0.816***
				(0.004)
	Light Activity	−0.078***	0.927***	0.861***
		(0.007)	(0.013)	(0.004)
Note:				***p<0.01;**p<0.05;*p<0.1

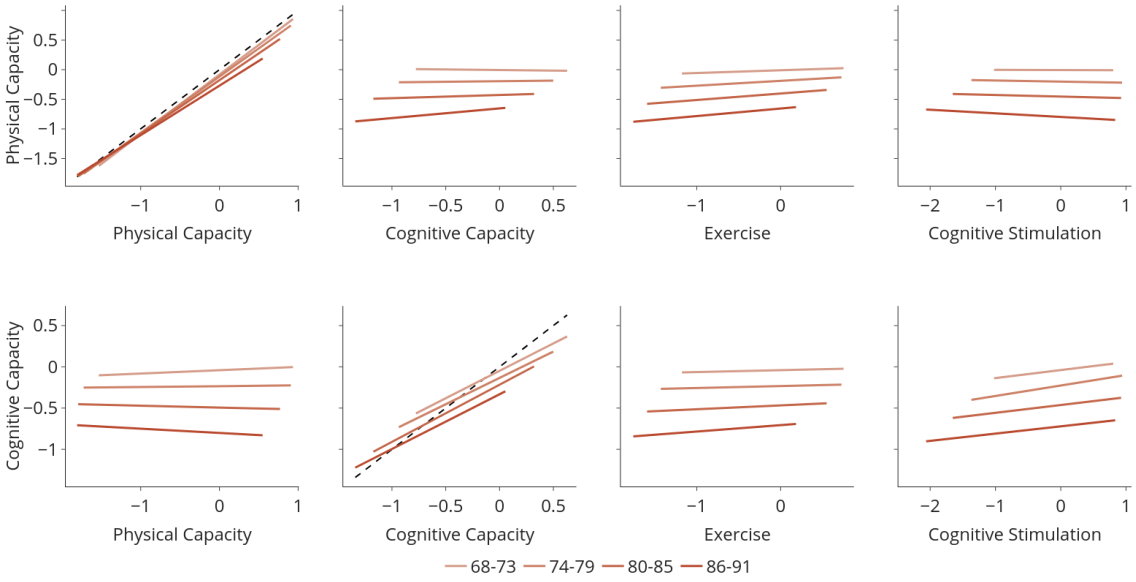
Table 1.E.7. Loadings and Measurement Standard Deviations for Cognitive Stimulation, Females

		Intercept	Loading	Meas. Std.
All	Reading	0.000	1.000	0.769***
				(0.006)
	Listening to Music	−0.168***	0.498***	0.981***
		(0.006)	(0.010)	(0.006)
	Stimulating Hobbies	−0.068***	0.564***	0.926***
		(0.007)	(0.011)	(0.005)
	Communication	−0.062***	0.513***	0.999***
		(0.006)	(0.010)	(0.005)
Note:				***p<0.01;**p<0.05;*p<0.1

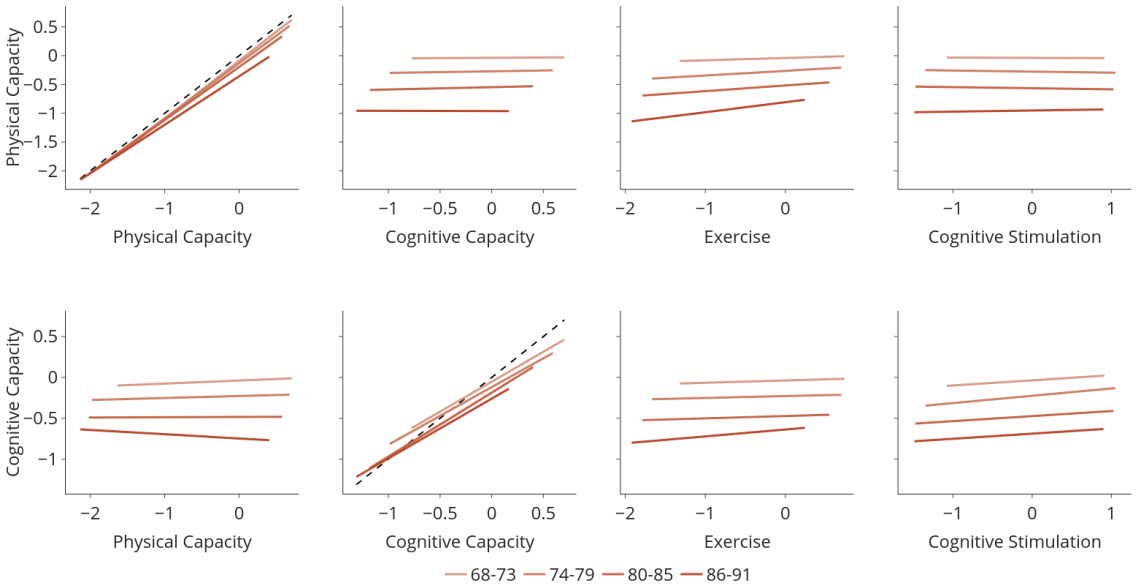
Table 1.E.8. Loadings and Measurement Standard Deviations for Cognitive Stimulation, Males

		Intercept	Loading	Meas. Std.
All	Reading	0.000	1.000	0.674***
				(0.007)
	Listening to Music	−0.175***	0.223***	1.005***
		(0.007)	(0.010)	(0.007)
	Stimulating Hobbies	−0.011	0.369***	0.970***
		(0.009)	(0.011)	(0.005)
	Communication	−0.082***	0.320***	0.990***
		(0.007)	(0.010)	(0.006)
Note:				***p<0.01;**p<0.05;*p<0.1

1.E.2 Transition Equations



(a) Transitions, females



(b) Transitions, males

Figure 1.E.1. Transition equations (other factors evaluated at the median)

Table 1.E.9. Transition Parameters for Physical Capacity, Females

	68-73	74-79	80-85	86-91
Physical Capacity	1.000***	0.950***	0.905***	0.835***
	(0.006)	(0.007)	(0.009)	(0.017)
Cognitive Capacity	−0.020***	0.020**	0.053***	0.164***
	(0.007)	(0.009)	(0.015)	(0.029)
Exercise	0.046***	0.081***	0.109***	0.128***
	(0.007)	(0.008)	(0.010)	(0.019)
Cognitive Stimulation	−0.003	−0.019**	−0.026**	−0.061***
	(0.007)	(0.008)	(0.011)	(0.017)
Constant	−0.086***	−0.105***	−0.117***	−0.090***
	(0.002)	(0.004)	(0.007)	(0.018)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.E.10. Transition Parameters for Physical Capacity, Males

	68-73	74-79	80-85	86-91
Physical Capacity	1.010***	0.957***	0.922***	0.842***
	(0.007)	(0.008)	(0.013)	(0.030)
Cognitive Capacity	0.009	0.028**	0.038**	−0.003
	(0.009)	(0.011)	(0.019)	(0.042)
Exercise	0.039***	0.081***	0.098***	0.173***
	(0.006)	(0.007)	(0.013)	(0.030)
Cognitive Stimulation	−0.004	−0.018**	−0.020	0.020
	(0.007)	(0.008)	(0.014)	(0.029)
Constant	−0.090***	−0.115***	−0.139***	−0.200***
	(0.003)	(0.005)	(0.010)	(0.029)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.E.11. Transition Parameters for Cognitive Capacity, Females

	68-73	74-79	80-85	86-91
Physical Capacity	0.040***	0.011	−0.023**	−0.052***
	(0.007)	(0.008)	(0.010)	(0.015)
Cognitive Capacity	0.664***	0.639***	0.693***	0.664***
	(0.010)	(0.011)	(0.014)	(0.026)
Exercise	0.023**	0.024**	0.047***	0.078***
	(0.010)	(0.011)	(0.012)	(0.018)
Cognitive Stimulation	0.097***	0.126***	0.095***	0.087***
	(0.010)	(0.010)	(0.011)	(0.016)
Constant	−0.054***	−0.122***	−0.180***	−0.248***
	(0.003)	(0.004)	(0.006)	(0.013)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.E.12. Transition Parameters for Cognitive Capacity, Males

	68-73	74-79	80-85	86-91
Physical Capacity	0.037***	0.024**	0.003	−0.052**
	(0.009)	(0.010)	(0.012)	(0.024)
Cognitive Capacity	0.733***	0.704***	0.782***	0.730***
	(0.011)	(0.013)	(0.017)	(0.033)
Exercise	0.028***	0.023**	0.028**	0.084***
	(0.010)	(0.010)	(0.012)	(0.026)
Cognitive Stimulation	0.063***	0.089***	0.061***	0.063***
	(0.010)	(0.010)	(0.013)	(0.023)
Constant	−0.056***	−0.105***	−0.160***	−0.203***
	(0.004)	(0.004)	(0.007)	(0.017)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.E.13. Transition Parameters for Exercise, Females

	68-73	74-79	80-85	86-91
Physical Capacity	0.027***	0.030***	0.027**	0.052**
	(0.010)	(0.011)	(0.014)	(0.021)
Cognitive Capacity	0.005	0.040***	0.101***	0.115***
	(0.010)	(0.011)	(0.016)	(0.027)
Exercise	0.991***	0.941***	0.886***	0.802***
	(0.014)	(0.014)	(0.018)	(0.026)
Constant	−0.073***	−0.111***	−0.158***	−0.261***
	(0.004)	(0.005)	(0.008)	(0.016)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.E.14. Transition Parameters for Exercise, Males

	68-73	74-79	80-85	86-91
Physical Capacity	0.073***	0.063***	0.106***	0.117***
	(0.012)	(0.013)	(0.020)	(0.037)
Cognitive Capacity	0.038***	0.022	0.133***	0.045
	(0.013)	(0.014)	(0.022)	(0.042)
Exercise	0.934***	0.942***	0.820***	0.786***
	(0.015)	(0.015)	(0.021)	(0.046)
Constant	−0.078***	−0.111***	−0.180***	−0.268***
	(0.005)	(0.006)	(0.011)	(0.024)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.E.15. Transition Parameters for Cognitive Stimulation, Females

	68-73	74-79	80-85	86-91
Physical Capacity	−0.010	0.024*	−0.000	0.101**
	(0.013)	(0.014)	(0.021)	(0.043)
Cognitive Capacity	0.041*	0.097***	0.115***	0.109
	(0.023)	(0.024)	(0.038)	(0.076)
Cognitive Stimulation	1.030***	0.960***	0.980***	0.922***
	(0.018)	(0.017)	(0.025)	(0.049)
Constant	−0.037***	−0.064***	−0.057***	−0.141***
	(0.007)	(0.009)	(0.015)	(0.034)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.E.16. Transition Parameters for Cognitive Stimulation, Males

	68-73	74-79	80-85	86-91
Physical Capacity	−0.007	0.033*	0.067**	0.025
	(0.018)	(0.019)	(0.032)	(0.080)
Cognitive Capacity	0.143***	0.059*	0.048	0.110
	(0.026)	(0.030)	(0.049)	(0.138)
Cognitive Stimulation	0.950***	0.974***	0.939***	0.843***
	(0.018)	(0.019)	(0.032)	(0.067)
Constant	−0.034***	−0.036***	−0.062***	−0.059
	(0.009)	(0.011)	(0.023)	(0.069)
Note:				***p<0.01;**p<0.05;*p<0.1

Table 1.E.17. Standard deviations of shocks

		Male All	Male All Linear	Female All	Female All Linear
68-73	Physical Capacity	0.094***	0.111***	0.005	0.038***
		(0.008)	(0.006)	(0.103)	(0.014)
	Cognitive Capacity	0.292***	0.294***	0.308***	0.308***
		(0.005)	(0.004)	(0.004)	(0.004)
	Exercise	0.236***	0.235***	0.164***	0.158***
		(0.012)	(0.012)	(0.011)	(0.011)
	Cognitive Stimulation	0.155***	0.171***	0.001	0.001
		(0.028)	(0.026)	(2.725)	(2.840)
74-79	Physical Capacity	0.161***	0.181***	0.159***	0.183***
		(0.006)	(0.005)	(0.005)	(0.004)
	Cognitive Capacity	0.283***	0.287***	0.302***	0.309***
		(0.005)	(0.005)	(0.004)	(0.004)
	Exercise	0.261***	0.264***	0.240***	0.243***
		(0.012)	(0.012)	(0.009)	(0.009)
	Cognitive Stimulation	0.190***	0.198***	0.185***	0.192***
		(0.026)	(0.025)	(0.017)	(0.017)
80-85	Physical Capacity	0.231***	0.252***	0.188***	0.219***
		(0.009)	(0.007)	(0.008)	(0.006)
	Cognitive Capacity	0.240***	0.243***	0.274***	0.279***
		(0.006)	(0.006)	(0.005)	(0.004)
	Exercise	0.326***	0.328***	0.275***	0.273***
		(0.015)	(0.015)	(0.012)	(0.012)
	Cognitive Stimulation	0.319***	0.326***	0.225***	0.231***
		(0.031)	(0.031)	(0.024)	(0.024)
86-91	Physical Capacity	0.315***	0.341***	0.227***	0.268***
		(0.020)	(0.014)	(0.012)	(0.010)
	Cognitive Capacity	0.250***	0.254***	0.238***	0.258***
		(0.013)	(0.009)	(0.010)	(0.007)
	Exercise	0.372***	0.372***	0.304***	0.295***
		(0.028)	(0.027)	(0.017)	(0.018)
	Cognitive Stimulation	0.471***	0.487***	0.309***	0.313***
		(0.059)	(0.056)	(0.047)	(0.049)

Note: ***p<0.01;**p<0.05;*p<0.1

References

Agostinelli, Francesco, and Matthew Wiswall. 2016a. "Estimating the Technology of Children's Skill Formation." Working Paper 22442. National Bureau of Economic Research. [20, 22, 31]

Agostinelli, Francesco, and Matthew Wiswall. 2016b. "Identification of Dynamic Latent Factor Models: The Implications of Re-Normalization in a Model of Child Development." Working Paper 22441. National Bureau of Economic Research. [21, 41]

Amengual, Dante, Jesús Bueren, and Julio A Crego. 2021. "Endogenous health groups and heterogeneous dynamics of the elderly." *Journal of Applied Econometrics* 36 (7): 878–97. [9]

Attanasio, Orazio, Flávio Cunha, and Pamela Jervis. 2019. "Subjective Parental Beliefs. Their Measurement and Role." NBER Working Papers 26516. National Bureau of Economic Research. [22]

Attanasio, Orazio, Costas Meghir, and Emily Nix. 2020. "Human Capital Development and Parental Investment in India." *Review of Economic Studies* 87 (6): 2511–41. eprint: https://academic.oup.com/restud/article-pdf/87/6/2511/34133061/rdaa026.pdf. [21, 22, 31]

Baker, Michael, Mark Stabile, and Catherine Deri. 2004. "What do self-reported, objective, measures of health measure?" *Journal of human Resources* 39 (4): 1067–93. [6, 8]

Ball, Karlene, Daniel Berch, Karin Helmers, Jared Jobe, Mary Leveck, Michael Marsiske, John Morris, George Rebok, David Smith, Sharon Tennstedt, Frederick Unverzagt, and Sherry Willis. 2002. "Effects of Cognitive Training Interventions With Older Adults: A Randomized Controlled Trial." *JAMA : the journal of the American Medical Association* 288 (12): 2271–81. [6]

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement error in survey data." In *Handbook of econometrics*. Vol. 5, Elsevier, 3705–843. [6]

Bradbury, James, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. *JAX: composable transformations of Python+NumPy programs*. Version 0.3.13. [23]

Chang, Yu-Hung, I-Chien Wu, and Chao A. Hsiung. 2021. "Reading activity prevents long-term decline in cognitive function in older people: evidence from a 14-year longitudinal study." *International Psychogeriatrics* 33 (1): 63–74. [14]

Clouston, Sean A. P., Paul Brewster, Diana Kuh, Marcus Richards, Rachel Cooper, Rebecca Hardy, Marcie S. Rubin, and Scott M. Hofer. 2013. "The Dynamic Relationship Between Physical Function and Cognition in Longitudinal Aging Cohorts." *Epidemiologic Reviews* 35 (1): 33–50. eprint: https://academic.oup.com/epirev/article-pdf/35/1/33/7287926/mxs004.pdf. [6]

Contoyannis, Paul, Andrew M. Jones, and Nigel Rice. 2004. "The dynamics of health in the British Household Panel Survey." *Journal of Applied Econometrics* 19 (4): 473–503. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jae.755. [8]

Crimmins, Eileen M. 2020. "Social hallmarks of aging: Suggestions for geroscience research." *Ageing research reviews* 63: 101136. [6, 9]

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97 (2): 31–47. [19]

Cunha, Flavio, and James J. Heckman. 2008. "Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation." *Journal of human resources* 43 (4): 738–82. [21]

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica* 78 (3): 883–931. [6, 7, 9, 19–23, 26, 36, 40, 41]

Cutler, David M, and Elizabeth Richardson. 1997. "Measuring the health of the US population." *Brookings papers on economic activity. Microeconomics* 1997: 217–82. [8]

Del Bono, Emilia, Josh Kinsler, and Ronni Pavan. 2020. "A Note on the Importance of Normalizations in Dynamic Latent Factor Models of Skill Formation." IZA Discussion Papers 13714. Institute of Labor Economics (IZA). [21]

Dixon, Roger A., and Cindy M. de Frias. 2014. "Cognitively elite, cognitively normal, and cognitively impaired aging: Neurocognitive status and stability moderate memory performance." *Journal of Clinical and Experimental Neuropsychology* 36 (4): 418–30. [13]

Dowd, Jennifer Beam, and Anna Zajacova. 2007. "Does the predictive power of self-rated health for subsequent mortality risk vary by socioeconomic status in the US?" *International journal of epidemiology* 36 (6): 1214–21. [8]

Dowd, Jennifer Beam, and Anna Zajacova. 2010. "Does self-rated health mean the same thing across socioeconomic groups? Evidence from biomarker data." *Annals of epidemiology* 20 (10): 743–49. [8]

Fiatarone, Maria A., Evelyn F. O'Neill, Nancy Doyle Ryan, Karen M. Clements, Guido R. Solares, Miriam E. Nelson, Susan B. Roberts, Joseph J. Kehayias, Lewis A. Lipsitz, and William J. Evans. 1994. "Exercise Training and Nutritional Supplementation for Physical Frailty in Very Elderly People." *New England Journal of Medicine* 330 (25): 1769–75. eprint: https://doi.org/10.1056/NEJM199406233302501. PMID: 8190152. [6]

Freyberger, Joachim. 2021. "Normalizations and misspecification in skill formation models." Working Paper. [21, 41]

Gabler, Janos. 2022. *A Python Library to Estimate Nonlinear Dynamic Latent Factor Models*. [23]

Gabler, Janoś, Tobias Raabe, Klara Röhrl, and Hans-Martin von Gaudecker. 2022. "The effectiveness of testing, vaccinations and contact restrictions for containing the CoViD-19 pandemic." en. *Sci. Rep.* 12 (1): 8048. [23]

Gaudecker, Hans-Martin von. 2019. "Templates for Reproducible Research Projects in Economics." [23]

Gill, Thomas M., Evelyne A. Gahbauer, Ling Han, and Heather G. Allore. 2010. "Trajectories of Disability in the Last Year of Life." *New England Journal of Medicine* 362 (13): 1173–80. eprint: https://doi.org/10.1056/NEJMoa0909087. PMID: 20357280. [25]

Health and Retirement Study. 2022a. "Cross-Wave Imputation of Cognitive Functioning Measures 1992-2018." [13]

Health and Retirement Study. 2022b. "HRS Public Survey Data." [10]

Health and Retirement Study. 2022c. "RAND HRS Products." [10]

Heckman, James J, and Stefano Mosso. 2014. "The Economics of Human Development and Social Mobility." *Annu. Rev. Econ.* 6 (1): 689–733. [5]

Heiss, Florian. 2011. "Dynamics of self-rated health and selective mortality." *Empirical Economics* 40 (1): 119–40. [8]

Hosseini, Roozbeh, Karen A. Kopecky, and Kai Zhao. 2022. "The evolution of health over the life cycle." *Review of Economic Dynamics* 45: 237–63. [5, 6, 8, 11]

Huang, Zhiyong, and Jürgen Maurer. 2019. "Validity of Self-Rated Memory Among Middle-Aged and Older Chinese Adults: Results From the China Health and Retirement Longitudinal Study (CHARLS)." *Assessment* 26 (8): 1582–93. eprint: https://doi.org/10.1177/1073191117741188. PMID: 29126348. [25]

Idler, Ellen L., and Yael Benyamini. 1997. "Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies." *Journal of Health and Social Behavior* 38 (1): 21–37. [8]

Jenicek, Milos, Robert Cleroux, and Michel Lamoureux. 1979. "Principal component analysis of four health indicators and construction of a global health index in the aged." *American Journal of Epidemiology* 110 (3): 343–49. [8]

Julier, Simon J., and Jeffrey K. Uhlmann, editors. 1997. *New extension of the Kalman filter to nonlinear systems*. International Society for Optics and Photonics. [35]

Jürges, Hendrik. 2007. "True health vs response styles: exploring cross-country differences in self-reported health." *Health Economics* 16 (2): 163–78. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hec.1134. [8]

Kalman, Rudolph Emil. 1960. "A new approach to linear filtering and prediction problems." *0021-9223*, [33]

Kapteyn, Arie, James Banks, Mark Hamer, James P Smith, Andrew Steptoe, Arthur Van Soest, Annemarie Koster, and Saw Htay Wah. 2018. "What they say and what they do: comparing physical activity across the USA, England and the Netherlands." *J Epidemiol Community Health* 72 (6): 471–76. [6, 7]

Karzmark, Peter. 2000. "8." *Internation journal of geriatric psychiatry* Validity of the serial seven procedure (15): [13]

Kasper, Judith D, Kitty S Chan, and Vicki A Freedman. 2017. "Measuring physical capacity: an assessment of a composite measure using self-report and performance-based items." *Journal of aging and health* 29 (2): 289–309. [6]

Latham, Kenzie, and Chuck W. Peek. 2012. "Self-Rated Health and Morbidity Onset Among Late Midlife U.S. Adults." *Journals of Gerontology: Series B* 68 (1): 107–16. eprint: https://academic.oup.com/psychsocgerontology/article-pdf/68/1/107/1694524/gbs104.pdf. [8]

Lindeboom, Maarten, and Eddy Van Doorslaer. 2004. "Cut-point shift and index shift in self-reported health." *Journal of health economics* 23 (6): 1083–99. [8]

Maurer, Jürgen, Roger Klein, and Francis Vella. 2011. "Subjective health assessments and active labor market participation of older men: evidence from a semiparametric binary choice model with nonadditive correlated individual-specific effects." *Review of Economics and Statistics* 93 (3): 764–74. [8]

McCammon, Ryan J., Gwenith G. Fisher, Halimah Hassan, Jessica D. Faul, Willard L. Rodgers, and David R. Weir. 2022. "Health and Retirement Study – Imputation of Cognitive Functioning Measures: 1992-2018 Data Description." Working paper. Version 7.0. Survey Research Center University of Michigan. [13]

McFadden, Daniel. 2008. "Human capital accumulation and depreciation." *Applied Economic Perspectives and Policy* 30 (3): 379–85. [5]

Montero-Odasso, Manuel, Joe Verghese, Olivier Beauchet, and Jeffrey M Hausdorff. 2012. "Gait and cognition: a complementary approach to understanding brain function and the risk of falling." *Journal of the American Geriatrics Society* 60 (11): 2127–36. [6]

Nakazato, Yuichi, Tomoko Sugiyama, Rena Ohno, Hirofumi Shimoyama, Diana L Leung, Alan A Cohen, Riichi Kurane, Satoru Hirose, Akihisa Watanabe, and Hiromi Shimoyama. 2020. "Estimation of homeostatic dysreg-

ulation and frailty using biomarker variability: a principal component analysis of hemodialysis patients." *Scientific Reports* 10 (1): 1–12. [8]

Niccoli, Teresa, and Linda Partridge. 2012. "Ageing as a Risk Factor for Disease." *Current Biology* 22 (17): R741–R752. [6]

Poterba, James M, Steven F Venti, and David A Wise. 2017. "The asset cost of poor health." *Journal of the Economics of Ageing* 9: 172–84. [8]

Prvan, Tania, and M. R. Osborne. 1988. "A Square-Root Fixed-Interval Discrete-Time Smoother." *Journal of the Australian Mathematical Society. Series B. Applied Mathematics* 30 (1): 57–68. [23, 38]

Raabe, Tobias. 2020. "A Python tool for managing scientific workflows." [23]

Rockwood, Kenneth, and Arnold Mitnitski. 2007. "Frailty in relation to the accumulation of deficits." *Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 62 (7): 722–27. [8]

Runge, Shannon K. 2015. "Word Recall: Cognitive Performance Within Internet Surveys." *JMIR Mental Health* 2 (2): [13]

Salthouse, Timothy. 2010. "Selective review of aging." *Journal of the International Neuropsychological Society : JINS* 16 (09): 754–60. [7]

Salthouse, Timothy. 2012. "Consequences of age-related cognitive declines." *Annual review of psychology* 63: 201. [7]

Särkämö, Teppo, and David Soto. 2012. "Music listening after stroke: beneficial effects and potential neural mechanisms." *Annals of the New York Academy of Sciences* 1252 (1): 266–81. eprint: https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.2011.06405.x. [14]

Särkämö, Teppo, Mari Tervaniemi, Sari Laitinen, Anita Forsblom, Seppo Soinila, Mikko Mikkonen, Taina Autti, Heli M. Silvennoinen, Jaakko Erkkilä, Matti Laine, Isabelle Peretz, and Marja Hietanen. 2008. "Music listening enhances cognitive recovery and mood after middle cerebral artery stroke." *Brain* 131 (3): 866–76. eprint: https://academic.oup.com/brain/article-pdf/131/3/866/907374/awn013.pdf. [14]

Schiele, Valentin, and Hendrik Schmitz. 2021. "Understanding cognitivedecline in older ages: The role of health shocks." *Ruhr Economic Papers*, (919): [9]

Steiber, Nadia. 2016. "Strong or Weak Handgrip? Normative Reference Values for the German Population across the Life Course Stratified by Sex, Age, and Body Height." *PLOS One* 10 (11): [11]

Van Der Merwe, Rudolph. 2004. "Sigma-Point Kalman Filters for Probabilistic Inference in Dynamic State-Space Models." Doctoral dissertation. OGI School of Science & Engineering at Oregon Health & Science University. [35]

van der Merwe, Rudolph, and Eric A. Wan, editors. 2001. *The square-root unscented Kalman filter for state and parameter-estimation*. IEEE. [23, 38]

World Bank. 2018. *World development report 2019: The changing nature of work*. The World Bank. [5]

Zhang, Fuzhen. 1999. *Matrix Theory: Basic Results and Techniques*. Universitext (Berlin. Print). Springer. [35, 38]

Chapter 2

Intrinsic and External Determinants of Age-Related Decline in Functioning

Joint with Jürgen Maurer

2.1 Introduction

Aging is an inevitable biological process that affects all living organisms. At the biological level, the gradual accumulation of molecular and cellular damage associated with aging (World Health Organization, 2015) can lead to a broad spectrum of impairments that limit an individual's ability to perform daily activities and maintain independence. Yet, age-related functional decline and the associated risk of disability is neither a deterministic nor a linear function of the biological age (World Health Organization, 2015), but is influenced by a complex interplay between *intrinsic* and *environmental* factors.

Deterioration of intrinsic capacities manifested mainly through accumulating multiple chronic conditions in older adults can strain their physical and mental resources, leading to reduced mobility, fatigue, and cognitive impairment. This cumulative burden of chronic conditions can disrupt daily routines and activities of daily living, such as bathing, dressing, and eating, leading to social isolation, increased dependence on others, and a decline in overall quality of life.

At the same time, environmental factors, including social support, access to healthcare, and living conditions, can either exacerbate or mitigate the effects of intrinsic factors.

Thus, a comprehensive assessment of functioning in older age is inevitably tied not only to an individual's physical and mental limitations but also to the social environment and norms around them. Ultimately, these intrinsic and broader environmental factors interact in non-trivial ways to determine old-age outcomes related to functioning and wellbeing, as captured by an individual's happiness, satisfaction, and fulfillment.

*

Understanding the intricate relationship between intrinsic and environmental factors is crucial for developing effective interventions to promote healthy aging and prevent functional decline. By identifying modifiable environmental factors, we can design strategies to optimize individuals' environments and enhance their ability to maintain independence and quality of life as they age.

We derive the motivation for our study primarily from the World Health Organization's (WHO) 2015 report on healthy aging (World Health Organization, 2015). The report defines healthy ageing as a multifaceted process encompassing the development and maintenance of *functional ability*, enabling individuals to experience wellbeing and live fulfilling lives in their later years. As defined in the report, functional abilities encompass the health-related attributes enabling individuals to perform activities they value. They are determined by an individual's intrinsic capabilities, the external environment, and the interactions between these factors (World Health Organization, 2015). Uncovering the implications of the environmental factors for age-related functional decline bears importance for the policies aimed at promoting healthy aging and reducing disparities in the age trajectories of disability.

A valuable framework for understanding the interplay between an individual's intrinsic capacities and external environment in shaping disability is developed in the seminal paper Verbrugge and Jette (1994). The paper proposes a sociomedical model of disability, emphasizing the dynamic and interactive nature of the disablement process. Intrinsic factors, such as age, genetics, and comorbidities, contribute to the underlying impairments and functional limitations. Extrinsic factors, including social structures, environmental barriers, and personal resources, influence how individuals with impairments function and participate in society. Disability, the paper suggests, is then determined by the gap between the demand for functional abilities imposed by one's environment and the actual functional abilities arising from one's intrinsic capacities.

Various empirical studies have also been conducted into intrinsic and extrinsic predictors of disabilities among older adults. In most studies, disability was defined in terms of difficulties in performing activities of daily living (ADL) and instrumental activities of daily living (IADL)[1]. As a summary index of health status in older adults, frailty, defined as the presence of multiple health risk factors, has been found to be a major factor contributing to increased risk of disability in older people (see, e.g., Vermeulen, Neyens, Rossum, Spreeuwenberg, and Witte, 2011; Makizako, Shimada, Doi, Tsutsumimoto, and Suzuki, 2015; Cunha, Veronese, Melo Borges, and Ricci, 2019). On a more disaggregated level, pertaining to more specific physical measurements, hand grip strength has been found to be negatively associated with increased risk of functional disability (see e.g., Taekema, Gussekloo, Maier, Westendorp, and Craen, 2010; McGrath, Vincent, Jurivich, Hackney, Tomkinson, et al., 2020).

Cognitive health has similarly been found to have significant implications for functional decline in older adults. In studies of aging and dementia, several papers have found finds that cognitive decline is either a precursor to functional decline or coexists with it (see e.g., Auyeung, Kwok,

1. ADL include basic self-care tasks, such as bathing, dressing, and getting in and out of bed, while IADL encompass complex activities that assist individuals in living independently in the community, including meal preparation, shopping, housekeeping, etc.

Lee, Leung, Leung, et al., 2008; Burton, Strauss, Bunce, Hunter, and Hultsch, 2009; Zahodne, Manly, MacKay-Brandt, and Stern, 2013).

A wide range of environmental factors have also been studied as potential predictors for healthy aging in general and sustentation of functional abilities in particular, in later stages of life. Unsurprisingly, built environments with lower accessibility of dwellings and transportation contribute to increased risk of disability among older adults with functional limitations (see e.g., Keysor, Jette, LaValley, Lewis, Torner, et al., 2009; Lien, Guo, Chang, Lin, and Kuan, 2014; Satariano, Kealey, Hubbard, Kurtovich, Ivey, et al., 2014). Apart from physical barriers to mobility, socioeconomic and psychosocial factors have also been found to be linked to functional disability in older adults. Zhong, Wang, and Nicholas (2017), for instance, finds that both childhood and adult socioeconomic status were associated with functional disability. One should, however, be wary of the apparent endogenous nature of economic status (i.e., income) relative to the disability status. A number of studies in gerontology have found that promoting social participation among older adults is beneficial in terms of reduced risk of disability and overall functional health (see e.g., Mendes de Leon, 2003; Kanamori, Kai, Aida, Kondo, Kawachi, et al., 2014; Gao, Sa, Li, Zhang, Tian, et al., 2018).

While the research on both intrinsic capacities and external factors as possible determinants of old-age disability is rich, and the importance of viewing disability in the context of one's environment has largely been established, to the best of our knowledge, in the existing literature, the prevalent econometric approach is the modeling of the relevant variables in a somewhat simplified, linear manner.

The main contribution of our study to the existing literature is, thus, the estimation of *nonlinear* interaction terms between intrinsic and extrinsic factors and their impact on disability rate. To this end, we use the semi-parametric double index binary choice estimator developed in Klein and Vella (2009). In this econometric model, we identify two indices, *intrinsic* and *extrinsic*, which are summary quantification of intrinsic and environmental factors, respectively. Within this framework, we are able to abstain from parametric assumptions regarding the functional form of the link function between the two indices to obtain the predicted probability of being disabled.

Our main results are aligned with the prevailing view in the gerontology literature. Specifically, we find that the environmental index has a nontrivial impact on the predicted probability of being disabled. We also find considerable nonlinear interaction effects between intrinsic and extrinsic indices. In particular, we find the intrinsic gradient of the predicted probability of disability to be steeper at lower quantiles of the environmental index.

The rest of the paper is organized as follows: Section 2.2 describes the data source we use for our estimation and discusses the specific variables used in the econometric specification. Section 2.3 presents the estimation framework of Klein and Vella (2009) and discusses specific estimators of interest. Section 2.4 discusses our results, and Section 3.6 concludes.

2.2 Data and model specification

In our empirical application, we use data from the first wave[2] of the WHO Study on global AGEing and Adult Health (SAGE)(World Health Organization, 2022), pertaining four lower- to upper-middle-income countries (China, India, Russian Federation, and Ghana)[3]. SAGE collects data on a wide range of individual and household-level characteristics of individuals aged 50 and above, its main target population. Sample sizes vary significantly from country to country (see Table 2.2.1). The Chinese sample has the largest number of observations (more than 11 000), and the Russian sample has the lowest number of observations (less than 2 500). However, the sampling was conducted for each country to ensure that the data are representative on the national level.

Extensive coverage of individuals' health status, health-related impairments, and social networks make SAGE data especially suitable for our study. We extract our outcome and most of the independent variables from the individual-level questionnaire of SAGE. The individual questionnaire contains various variables covering socio-demographic characteristics, difficulties due to health conditions, and measures (objective and self-reported) of physical and mental/cognitive health. We use the information on the estimated household income and the number of household members from the household-level data sets.

We proceed by describing our outcome variable and variables included in each index in the following subsections.

2.2.1 Outcome variable

We base our outcome variable, the binary disability indicator, on the WHO Disability Assessment Schedule 2.0 (WHODAS-2.0). WHODAS-2.0 is a 12-item, self-reported questionnaire that asks how much health-state related difficulty a person has had in performing physical or cognitive/mental activities in six domains: 1) cognition (understanding and communicating), 2) mobility (moving and getting around), 3) self-care (hygiene, dressing, eating and staying alone), 4) getting along (interacting with others), 5) life activities (domestic responsibilities, leisure, work, and school), and 6) participation (participation in community activities and society) (T.B. Ustun, N. Kostanjesek, S. Chatterji, J. Rehm, 2010). Possible scores assigned to each question range from 0 (no difficulty at all) to 5 (extreme difficulty or impossible to perform the task.). We calculate the overall score as the sum of scores across the 12 questions. Finally, in our primary analysis, to define our outcome disability variable, we choose the cut-off of the upper 20th percentile of the disability score distribution. The average rate of disability, perhaps somewhat mechanically, is around 20% in most of the country samples (see Table 2.2.1).

We consider a stricter (10th percentile) and a looser (30th percentile) cut-off in additional results reported in Appendix 2.A and Appendix 2.B.

2. Although WHO has implemented three waves of the survey (Wave 1: 2007-2010, Wave 2: 2014-2015, and Wave 3: 2018-2019), the first wave is to date the only available one.

3. In our application, we do not use the data from Mexico and South Africa given a large number of missing values in critical variables of our analysis.

Table 2.2.1. Summary statistics, outcome and intrinsic variables

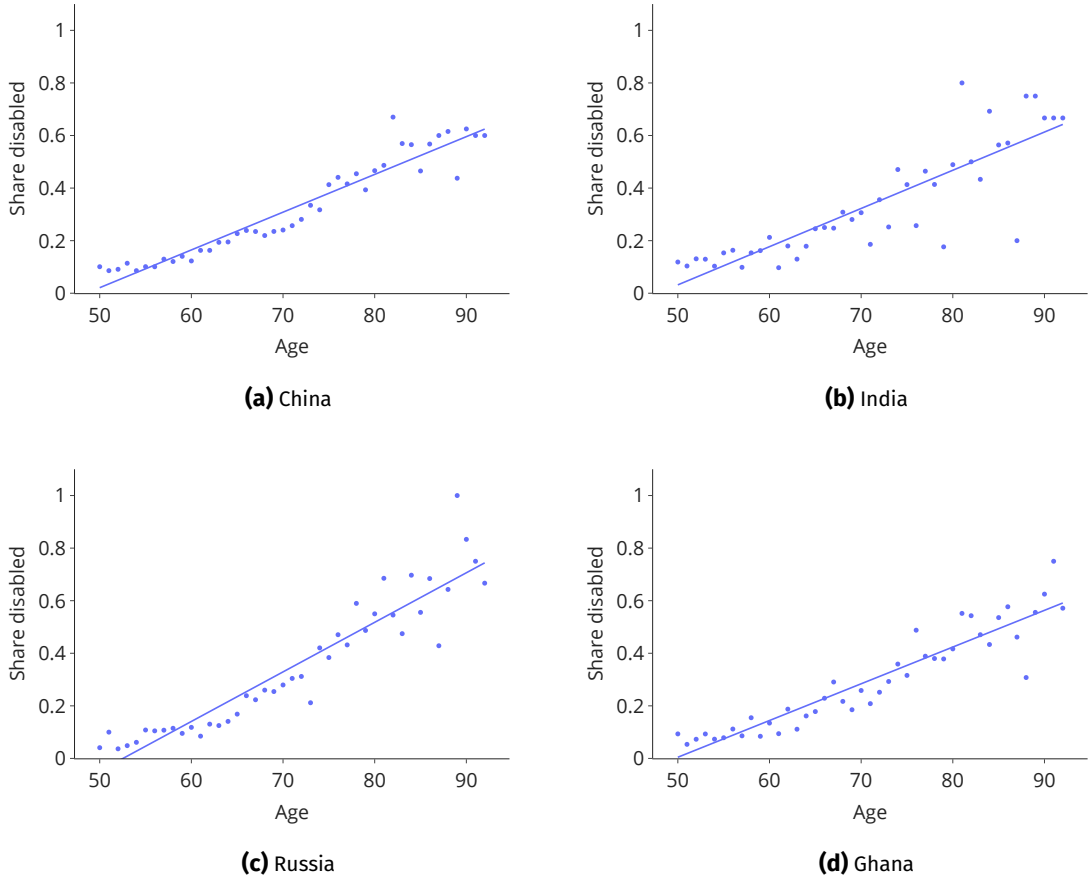
	China		India		Russia		Ghana	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Disability	0.20	0.40	0.19	0.40	0.18	0.38	0.19	0.39
Age	62.78	9.12	61.28	8.56	63.38	9.61	63.75	10.07
Male	0.47	0.50	0.53	0.50	0.39	0.49	0.52	0.50
Grip Strength	0.28	0.12	0.21	0.09	0.30	0.13	0.26	0.12
Cognition Score	0.51	0.13	0.45	0.11	0.55	0.14	0.50	0.11
Diabetes	0.06	0.24	0.07	0.26	0.08	0.28	0.04	0.19
Lung	0.09	0.28	0.04	0.20	0.17	0.38	0.01	0.07
Arthritis	0.27	0.44	0.28	0.45	0.38	0.48	0.26	0.44
Angina	0.11	0.31	0.19	0.39	0.38	0.49	0.16	0.36
Stroke	0.03	0.18	0.02	0.13	0.04	0.20	0.02	0.15
Asthma	0.05	0.22	0.12	0.32	0.07	0.25	0.05	0.22
Hypertension	0.55	0.50	0.30	0.46	0.58	0.49	0.56	0.50
Depression	0.02	0.13	0.13	0.34	0.08	0.27	0.08	0.27
N. Obs	11062		5458		2417		3652	

2.2.2 Intrinsic variables

As discussed previously, the intrinsic index summarizes the individual characteristics of aging adults. In the intrinsic index, we thus include the age and variables measuring physical and cognitive capacities, frailty, age, and gender.

The first variable in the intrinsic index is the **age** variable. By convention, setting age as the first variable in the intrinsic index implies that coefficient estimates of the variables in the index are to be interpreted in units (a year) of age. To be more precise, in our estimation, we will normalize the coefficient of age variable to -1. Using age as the normalizing variable seems like the natural choice given first, the context of our analysis, that is, age-related decline in functioning, and second, the strong and positive association between age and the average disability rate, as evidenced by Figure 2.2.1. For all countries, we see a smooth increase (albeit at various rates) in average disability along the age axis. To control for different trajectories (from a biological perspective) in functional decline between men and women, we include **gender** in the index of intrinsic variables[4]. As we can see in 2.2.1, in terms of gender representation, samples from China, India, and Ghana are relatively balanced; the Russian sample has a slightly higher rate of female participation.

4. Liang, Bennett, Shaw, Quiñones, Ye, et al. (2008) shows, for the US population, that among older adults, women have lower baseline level, as well as a higher rate of worsening of functional status.



(a) China **(b)** India

(c) Russia **(d)** Ghana

Figure 2.2.1. Disability rate by age

The next variable in the intrinsic index, the **grip strength**, has been used as a biomarker for identifying adults at risk of cognitive and functional impairments (see Bohannon (2019) and Rantanen, Guralnik, Foley, Masaki, Leveille, et al. (1999)). In the SAGE survey, two tests were conducted for each hand, using a specialized tool to measure grip strength in kilograms. Our variable of interest is constructed using the average measurement for the dominant hand, converted to a score ranging from zero to one.

We include the variable **cognition score** to measure cognitive functioning. Cognition score has been calculated as the composite of scores from the following tests administered during the survey interviews:

Immediate word recall test reads a list of ten words to the interviewees and asks them to recall them (in any order) in three subsequent trials. The summary score is the average of correctly recalled words across the three trials. In the *delayed word recall* test, participants are asked to recall the same list of words after a delay of ten minutes.

In *backward and forward counting* tests, respondents are asked to recall (in the correct order, backward and forward, respectively) lists of numbers of different lengths in two subsequent trials. The length of the most extended list determines the final score recited correctly in either of the trials.

To account for health-related risk factors, we include indicator variables for the following diseases: diabetes, lung disease, arthritis, angina, stroke, asthma, hypertension, and depression. The binary variables have been constructed based on the combination of (self-reported) diagnoses and symptomatic conditions. An exception is the indicator for hypertension, which is based on the blood pressure measurements conducted during the at-home interviews.

As we can see in Table 2.2.1, the prevalence rates vary quite significantly across the diseases. In particular, in all the country samples, we observe higher prevalence rates for arthritis and hypertension. Depression and stroke, on the other hand, have the lowest prevalence rate in all the country samples.

2.2.3 Environmental variables

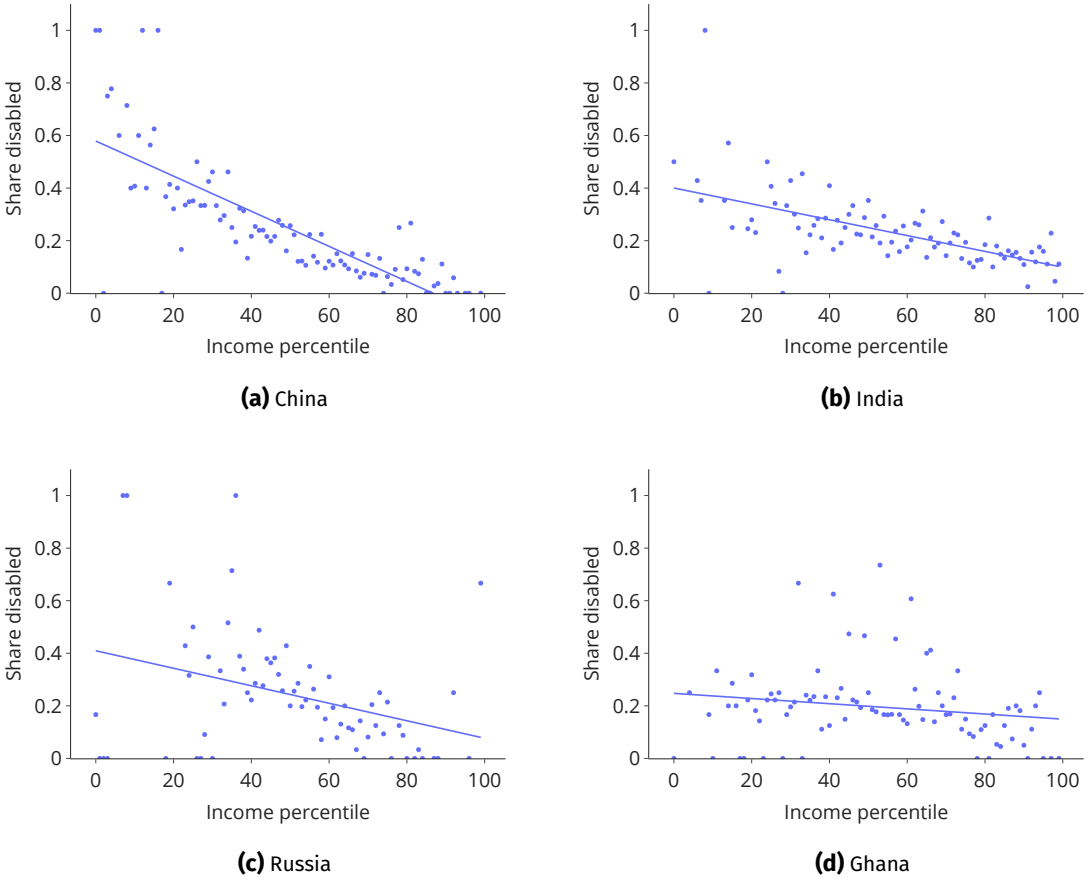
To capture the environmental factors impacting individuals' functional abilities, we consider several variables from the socioeconomic domain, as well as social cohesion variables, which describe an individual's connectedness to their community.

The first variable we include in the environmental index, which is also the normalizing variable, is the **income percentile**. We construct the variable based on the (per country) sample distribution of the permanent income score. The SAGE data estimates households' permanent income based on current income and household assets. An obvious concern with including income in the environmental index is its endogenous nature, as current income can strongly depend on disability status. A potentially alleviating factor for the endogeneity concern is the inclusion of household assets, reflecting long-term wealth, in estimating the permanent income score.

Fig 2.2.2 offers a first look at how disability rates vary along the income distribution. We mainly observe gradually declining disability rates as we move up along the income distributions. The income gradient of the disability rate is especially pronounced in the Chinese sample.

Table 2.2.2. Summary statistics, environmental variables

	China		India		Russia		Ghana	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Income Percentile	48.71	16.41	60.99	21.13	54.89	12.88	53.82	15.96
Education Years	5.48	4.42	3.80	4.77	11.32	3.62	4.19	5.32
Num. HH Memberds	1.76	1.32	5.45	3.68	1.55	1.58	4.52	3.33
Cohabiting	0.83	0.37	0.76	0.43	0.62	0.49	0.57	0.49
Comm Loc. Affairs	1.18	0.48	1.38	0.67	1.44	0.74	1.75	0.86
Comm Community Leader	1.10	0.40	1.43	0.77	1.25	0.60	2.42	1.25
Comm Soc. Clubs	1.24	0.61	1.44	0.75	1.35	0.67	2.28	1.18
Comm Work Nbhd.	1.62	0.84	1.72	0.95	1.53	0.70	2.03	1.10
Comm Friends Over	2.12	0.91	2.79	1.21	2.48	0.87	3.55	1.31
Comm Diff. Nbhd.	1.94	0.91	2.80	1.20	2.06	0.85	3.16	1.33
Comm Got Out	2.08	0.74	2.09	0.82	2.29	1.18	3.14	1.26
Trust Nbhd	3.89	0.70	3.30	1.00	2.94	1.03	3.38	1.10
Trust Work	3.84	0.69	3.08	1.04	3.02	1.01	3.13	1.13
Trust Stangers	1.57	0.79	2.23	1.09	1.82	0.90	2.44	1.16
Trust Gen	0.89	0.31	0.55	0.50	0.32	0.47	0.61	0.49
Trust Someone	0.98	0.14	0.81	0.40	0.80	0.40	0.78	0.41
Gov Impact	1.80	0.98	2.06	1.20	2.12	0.96	2.71	1.25
Gov Freedom	3.73	0.85	2.63	1.43	3.08	1.19	3.57	1.09
N. Obs	11062		5458		2417		3652	



(a) China

(b) India

(c) Russia

(d) Ghana

Figure 2.2.2. Disability rate by income percentile

The second variable we include in the environmental index is **education years**. As part of a larger set of socio-economic variables, education has been shown to be negatively associated with functional decline (see, e.g., Beydoun and Popkin, 2005; Larnyo, Dai, Nutakor, Ampon-Wireko, Larnyo, et al., 2022). In terms of sample means, Russia stands out among the four countries with average education years of more than ten years, compared to the 4-5 years for the other countries (Table 2.2.2). This stark difference can arise either from specifics of the Soviet educational system (when our survey participants would have attained their education), namely compulsory secondary education and state-sponsored free higher education, or from oversampling of the urban population in the Russian data (80% compared to 25% in India, 40% in Ghana, and 50% in China).

A number of studies have shown that a more supportive social environment can be a mitigating factor in age-related functional decline and psychological distress associated with it (see, e.g., Backe, Patil, Nes, and Clench-Aas, 2018; Hajek, Brettschneider, Eisele, Mallon, Oey, et al., 2022). **number of household members** is the first variable from this domain of variables that we include in the environmental index. Table 2.2.2 shows that the largest average household size is observed in India (5.5), followed by Ghana (4.51), China (1.75), and Russia (1.5). We also include an indicator variable for the marital status (variable **cohabiting** in Table 2.2.2), which is coded as one if an individual is married or cohabits with another person.

We further include several variables associated with social cohesion. The first set of these questions asks about the frequency of community involvement during the period of 12 months prior to the interview. An example question is how often an individual has attended group, society, or union meetings (**Comm Soc Clubs**), with possible answers ranging from never (score of 1) to daily (score of 5). The second set of variables concerns trust for different groups of people. Two binary answer questions ask whether respondents think people can be trusted in general (**Trust Gen**) and whether they have someone they can trust (**Trust Someone**). The rest of the questions in this set ask to which extent respondents trust strangers (**Trust Strangers**) and people at work (**Trust Work**) and neighborhood (**Trust Nbhd**). Possible answers range from "to a very great extent" (score of 1, but coded as 5 in our analysis) to "to a very small extent" (score of 5, but coded as 1 in our analysis). Finally, **Gov Impact** and **Gov Freedom** ask respondents, on a scale of from 1 to 5, how much influence they think they have in getting the government to address issues relevant to them and how much freedom they have in expressing political opinions without fear of reprisal, respectively.

2.3 Estimator and quantities of interest

2.3.1 Estimator

We formalize and estimate our empirical model using the binary response double index framework developed in Klein and Vella (2009). Namely, we model the conditional probability of being disabled in the following form:

$$Pr(Disa = 1|I,E) = h(I,E) \tag{2.3.1}$$

$$\Leftrightarrow Pr(Disa = 1|X^I,X^E) = h(X^I\beta^I,X^E\beta^E) \tag{2.3.2}$$

The empirical formulation of equation 2.3.1 implies that an individual's intrinsic and environmental characteristics enter the respective indices in linear form. Furthermore, the probability link function h is determined nonparametrically, based on the observed data, and allows for nonlinear interaction between the indices.

Our identification strategy is analogous to that of Drerup, Enke, and von Gaudecker (2017). Specifically, in our econometric specification, the continuous age and income percentile variables are included only in the intrinsic and environmental indices, respectively. We further normalize the coefficients of those variables to -1 and 1.

Utilizing Bayes rule, the probability link function can be expressed as a function of the joint densities of the two indices and unconditional probability of being disabled:

$$Pr(Disa = 1|X^I,X^E) = h(X^I\beta^I,X^E\beta^E) = \frac{f_{Disa=1}(X^I\beta^I,X^E\beta^E)Pr(Disa = 1)}{f(X^I\beta^I,X^E\beta^E)} \tag{2.3.3}$$

where $f(\cdot,\cdot)$ and $f_{Disa=1}(\cdot,\cdot)$ denote the unconditional and conditional (on being disabled) joint densities of the intrinsic and environmental indices, respectively, and $Pr(Disa = 1)$ is the unconditional probability of being disabled. Empirical estimates of the quantities involved in 2.3.3 are obtained through the multistage kernel density estimation procedure of Klein and Vella (2009). The index parameter vectors β^I and β^E then can then be estimated via maximum likelihood estimation:

$$(\hat{\beta}^I,\hat{\beta}^E) = \arg\max_{\beta^I,\beta^E} \sum_{i=1}^{N} \tau_i[Disa_i \ln(\hat{P}_i(\beta^I,\beta^E)) + (1 - Disa_i)\ln(1 - \ln\hat{P}_i)] \tag{2.3.4}$$

where τ_i is a trimming function preventing densities from diminishing, and $\hat{P}_i(\cdot,\cdot)$ is the empirical estimate of the conditional probability of being disabled.

2.3.2 Parameters of interest

As a summary version of the link function h, we estimate the average structural function (ASF) proposed in Blundell and Powell (2003) and Blundell and Powell (2004). ASF summarizes the dependence of the binary response variable (disability rate in our case) on either of the structural indices by taking the average over the marginal distribution of the respective other index. In formal terms:

$$ASF_I(I^0) = \int h(I^0,E)dF_E \tag{2.3.5}$$

$$ASF_E(E^0) = \int h(I,E^0)dF_I \tag{2.3.6}$$

where the marginal distriubtions of the indices, F_I and F_E can be estimated, given the emprical estimates of the index levels, $\hat{I} = X^I\hat{\beta}^I$ and $\hat{E} = X^E\hat{\beta}^E$, respectively. While the parameters estimated via

the maximum likelihood procedure describe the contribution of each microvariable to the structural indices, we summarize the effects of the individual environmental and intrinsic variables on the predicted probability of being disabled by means of "average partial effects" (APE) of Klein and Vella (2009),Maurer (2009) and Maurer, Klein, and Vella (2011). The average partial effect of changing the value of a micro variable x_j from x_j^0 to x_j^1 can be calculated as the difference of the ASF of the corresponding index when the value of x_j is fixed at x_j^0 and x_j^1 throughout. Formally, the APE of an x_j variable through the intrinsic index is calculated as:

$$APE_I(x_j^0, x_j^1) = \int h(I(x_j^1), E) dF_{I(x_j^1),E} - \int h(I(x_j^0), E) dF_{I(x_j^0),E} \tag{2.3.7}$$

with analogous calculations of the effects through the environmental index.

2.4 Main results

We will present our estimation results in several steps. We first discuss parameter estimates of the linear coefficients of the intrinsic and environmental indices. We next present the estimated semiparametric probability function, which illustrates the joint effect of the two indices on the predicted probability of being disabled. Further, we discuss the marginal effects of the intrinsic and environmental indices on the predicted probability of being disabled, wherein the effects of the environmental and intrinsic indices, respectively, are integrated out. Finally, we discuss the effects of individual covariates in each index on the predicted probability of being disabled.

2.4.1 Coefficient estimates

Tables 2.4.1 and 2.4.2 illustrate the estimated model parameters in the intrinsic and environmental indices, respectively. Recall that, for identification purposes, we normalize the constant terms in both indices to zeros. Further, in the intrinsic index, we normalize the coefficient of the age variable to -1, and in the environmental index, we normalize the coefficient of income percentile to 1. This normalization setting implies that we should interpret the coefficients in each index in units of the corresponding normalizing variables.

The negative association of age with the intrinsic index imposes certain expectations regarding the remaining variables. In particular, we would expect the risk factors (variables indicating various diagnoses) to enter the index with a negative coefficient, whereas the cognition score and the grip strength with a positive one.

The estimate of the coefficient of the gender variable is positive and statistically significant in the Chinese, Indian, and Ghanaian samples. The signs of the coefficient estimates imply that being male is associated with a lower risk of disability, which is in line with results from several studies in gerontology literature (see, e.g., Darkwah, Iddi, Nonvignon, and Aikins, 2022; Malik, 2022). If we were to interpret the sizes of the estimates, then, in the Ghanaian sample, for instance, being male has the same contribution to the intrinsic index as reducing age by almost 4 years. In any kind of interpretation, however, we should keep in mind that these coefficients are considered

Table 2.4.1. Estimated parameters, intrinsic index

	China	India	Russia	Ghana
Age	−1.000	−1.000	−1.000	−1.000
Male	2.965**	3.353**	−4.166	3.808**
	(1.315)	(1.690)	(2.800)	(1.828)
Grip Strength	36.389***	24.529**	60.752***	−19.570**
	(6.449)	(10.870)	(13.409)	(8.051)
Cognition Score	52.770***	54.716***	53.268***	64.285***
	(6.602)	(9.954)	(13.411)	(9.614)
Diabetes	−2.671	−1.126	−6.455*	−10.050**
	(2.344)	(3.031)	(3.550)	(3.934)
Lung	−0.165	−3.658	−4.190	−4.345
	(1.916)	(4.050)	(2.818)	(8.736)
Arthritis	−9.703***	−6.986***	−10.979***	−8.169***
	(1.327)	(1.659)	(2.572)	(1.817)
Angina	−14.350***	−12.298***	−16.027***	−8.201***
	(2.020)	(1.952)	(2.967)	(2.024)
Stroke	−28.424***	−8.959	−7.595	−24.583***
	(4.453)	(5.886)	(4.907)	(7.343)
Asthma	−9.933***	−13.628***	−8.707**	2.450
	(2.682)	(2.610)	(4.173)	(3.441)
Hypertension	−1.671	−4.653***	1.322	−2.831*
	(1.184)	(1.629)	(2.362)	(1.566)
Depression	−26.617***	−13.139***	−10.355***	1.770
	(5.356)	(2.301)	(3.847)	(2.808)
Observations	11062	5458	2417	3652
Note:				***p<0.01;**p<0.05;*p<0.1

nuisance parameters in the Klein and Vella (2009) estimator, which is primarily concerned with identifying the structural indices.

Further, as we see in Table 2.4.1, in all country samples, except for Ghana, the estimated coefficient of the grip strength variable has the expected sign and is statistically significant. Next, the estimated coefficients of the cognition score variable are positive and significant in all the samples and are of the same order of magnitude as those of the grip strength score variable. The normalization in our model and the standardization of cognition and grip strength variables render it challenging to interpret the estimated coefficients by more than to say that in all the country samples (with the noted exception), we observe statistically significant estimates and that the effects of higher scores of grip strength and cognitive abilities on the intrinsic index go in the same direction as that of lower age.

Turning to the risk factors, the estimated coefficients mainly have the expected negative sign. Many coefficient estimates are also statistically significant.

Overall, while the precision of the coefficient estimates in the intrinsic index varies across the country samples and for different risk factors, the results are rather satisfactory, with many of the estimates being statistically significant and most of the coefficients having the correct expected sign.

Table 2.4.2 presents the estimates of coefficients for the variables in the environmental index. Given the assumed positive relationship between income and the environmental index, we would expect all the variables in the environmental index to have positive estimated coefficients, as none of them reflects a risk factor but rather is indicative of a better environment in terms of social network and support. In the case of education years, the coefficient signs are aligned with our expectations in all the country samples. Considering the Chinese sample, for example, an additional year of education has the same effect on the environmental index as moving up the income distribution by 1.8 percentage points.

Results are somewhat disappointing for the other variables in the environmental index, both in terms of coefficient signs and estimation precisions. In particular, we would expect the household size to contribute to the environmental index in the same direction as income. However, the estimated coefficients are negative in most of the country samples. Further, the coefficients of social cohesion variables appear not to have a consistent sign either across country samples or the types of social cohesion. Within the community involvement subgroup, for example, attending meetings where local affairs are discussed (Comm. Loc. Affairs) contributes to the environmental index negatively (in terms of income percentile) in the Chinese sample but positively in the other country samples.

Table 2.4.2. Estimated parameters, environmental index

	China	India	Russia	Ghana
Income Percentile	1.000	1.000	1.000	1.000
Education Years	1.787***	1.612	1.926	−0.949
	(0.383)	(1.008)	(1.304)	(0.846)
Num. HH Memberds	−0.309	−3.837***	−7.702**	−3.523***
	(0.904)	(0.912)	(3.161)	(1.226)
Cohabiting	5.685*	17.860**	−4.744	−8.612
	(3.172)	(7.646)	(8.491)	(8.773)
Comm Loc. Affairs	−2.518	22.371***	15.633	58.000***
	(2.810)	(7.585)	(10.388)	(15.265)
Comm Community Leader	−17.981***	23.498***	2.705	−12.573**
	(3.029)	(6.380)	(9.367)	(5.328)
Comm Soc. Clubs	3.184	−0.462	−0.361	7.551*
	(2.371)	(5.265)	(10.098)	(4.545)
Comm Work Nbhd.	2.332	−4.052	19.562**	26.330***
	(1.627)	(4.050)	(8.733)	(7.528)
Comm Friends Over	0.127	−11.582***	−3.040	−16.860***
	(1.451)	(3.568)	(5.450)	(5.272)
Comm Diff. Nbhd.	11.877***	9.780***	9.209	6.258
	(1.722)	(3.469)	(6.305)	(4.052)
Comm Got Out	13.415***	1.700	4.200	1.221
	(2.163)	(4.252)	(3.949)	(3.465)
Trust Nbhd	−3.566	13.420***	7.356	−7.066
	(2.772)	(4.360)	(5.307)	(5.651)
Trust Work	5.608*	−8.561**	9.380*	28.070***
	(2.865)	(3.944)	(5.553)	(8.593)
Trust Stangers	−3.907**	1.986	3.424	−16.981***
	(1.532)	(3.225)	(5.482)	(5.826)
Trust Gen	−0.003	−11.490	0.797	−4.809
	(3.977)	(7.008)	(9.635)	(8.237)
Trust Someone	3.782	18.172**	−3.852	−8.789
	(8.655)	(9.029)	(9.208)	(9.665)
Gov Impact	4.148***	−5.592	−8.819*	−5.137
	(1.356)	(3.428)	(4.884)	(3.853)
Gov Freedom	5.698***	7.553**	12.350**	−4.922
	(1.575)	(2.984)	(5.416)	(3.971)
Observations	11062	5458	2417	3652
Note:				***p<0.01;**p<0.05;*p<0.1

On the other hand, having friends over contributes to the index negatively in all the country samples except for the Chinese one. In sum, we don't see a pattern emerge in terms of the relative contributions of the social cohesion variables to the environmental index, and many of the estimates lack statistical significance.

2.4.2 Average partial effects of individual covariates

To gauge how each microvariable, on average, affects the probability of being disabled, we look at the estimated partial effects in Table 2.4.3 and Table 2.4.4.

For the binary variables, the average partial effects are calculated as the difference between values of the non-parametric probability link function estimated ate index values when setting the control variable's value to one and zero for all observations, respectively. For the age, the number of household members, and education years, we calculate the change in the link function when increasing the values of the variable observations by one. We calculate the effect of a five percentage points increase for the income percentile. We consider the effect of one standard deviation increase in the observed values for the remaining variables.

Table 2.4.3. APEs, intrinsic index

	China	India	Russia	Ghana
Age	0.005	0.006	0.005	0.007
Male	−0.014	−0.019	0.020	−0.025
Grip Strength	−0.019	−0.012	−0.035	0.016
Cognition Score	−0.028	−0.029	−0.033	−0.040
Diabetes	0.013	0.006	0.033	0.073
Lung	0.001	0.021	0.021	0.030
Arthritis	0.052	0.042	0.060	0.057
Angina	0.083	0.081	0.095	0.058
Stroke	0.168	0.056	0.039	0.193
Asthma	0.055	0.091	0.045	−0.015
Hypertension	0.008	0.027	−0.006	0.018
Depression	0.156	0.087	0.054	−0.011

We start by looking at the results for the variables in the intrinsic index. As we could expect, the average rate of disability increases with age. Furthermore, the estimated partial effects are of the same order of magnitude across the country samples. Adding one year to age increases the probability of being disabled by 0.5 percentage points in the Chinese and Russian samples, by 0.6 percentage points in the Indian sample, and by 0.7 percentage points in the Ghanaian sample.

Further, given that in our model specification we do not have a variable appear in both indices, the relative signs and magnitudes of the coefficient estimates (Table 2.4.1) hint at signs of the average partial effects. In particular, in line with the coefficient estimates, being male is associated

with a decreased probability of being disabled in China (by 1.4% points), India (by 1.9% points), and Ghana (by 2.5% points), and an increased probability of being disabled in Russian (by 2.0% points). The impact of higher grip strength on the disability rate is negative across all countries except Ghana. In particular, increasing the grip strength score by one standard deviation leads to a decrease in disability rate by 1.9% points in the Chinese sample, 1.2% points in the Indian sample, and 3.5% points in the Russian sample. Similarly, increasing the cognition score by one standard deviation decreases the predicted probability of being disabled in all the country samples, including Ghana. The estimated average partial effects of the cognition score variable are of the same magnitude as those of the grip strength variable.

Turning to the risk factors, we observe that in most cases, their presence is associated with a higher probability of being disabled. The exceptions are the presence of symptoms or diagnosis of asthma and depression in the Ghanaian sample and hypertension in the Russian sample, where the estimated average partial effects are negative. Compared to the other health-related risk factors, the effect of stroke on the predicted probability of being disabled is the largest in the Chinese (16.8% points) and the Ghanaian (19.3% points) samples. Suffering from depression is also among the top risk factors in the Chinese, as well as the Indian samples, with estimated average partial effects of 15.6% (second to stroke) and 8.7% (second to asthma) points, respectively.

Table 2.4.4. APEs, environmental index

	China	India	Russia	Ghana
Income Percentile	−0.011	−0.005	−0.005	−0.005
Education Years	−0.004	−0.001	−0.002	0.001
Num. HH Memberds	0.001	0.004	0.008	0.004
Cohabiting	−0.013	−0.017	0.005	0.009
Comm Loc. Affairs	0.003	−0.014	−0.012	−0.045
Comm Community Leader	0.017	−0.017	−0.001	0.016
Comm Soc. Clubs	−0.004	0.000	0.001	−0.009
Comm Work Nbhd.	−0.004	0.004	−0.014	−0.029
Comm Friends Over	−0.000	0.013	0.003	0.022
Comm Diff. Nbhd.	−0.023	−0.011	−0.008	−0.009
Comm Got Out	−0.021	−0.001	−0.005	−0.001
Trust Nbhd	0.006	−0.013	−0.007	0.008
Trust Work	−0.009	0.008	−0.009	−0.032
Trust Stangers	0.007	−0.002	−0.003	0.020
Trust Gen	0.000	0.011	−0.001	0.005
Trust Someone	−0.009	−0.017	0.004	0.009
Gov Impact	−0.009	0.006	0.009	0.007
Gov Freedom	−0.011	−0.010	−0.015	0.006

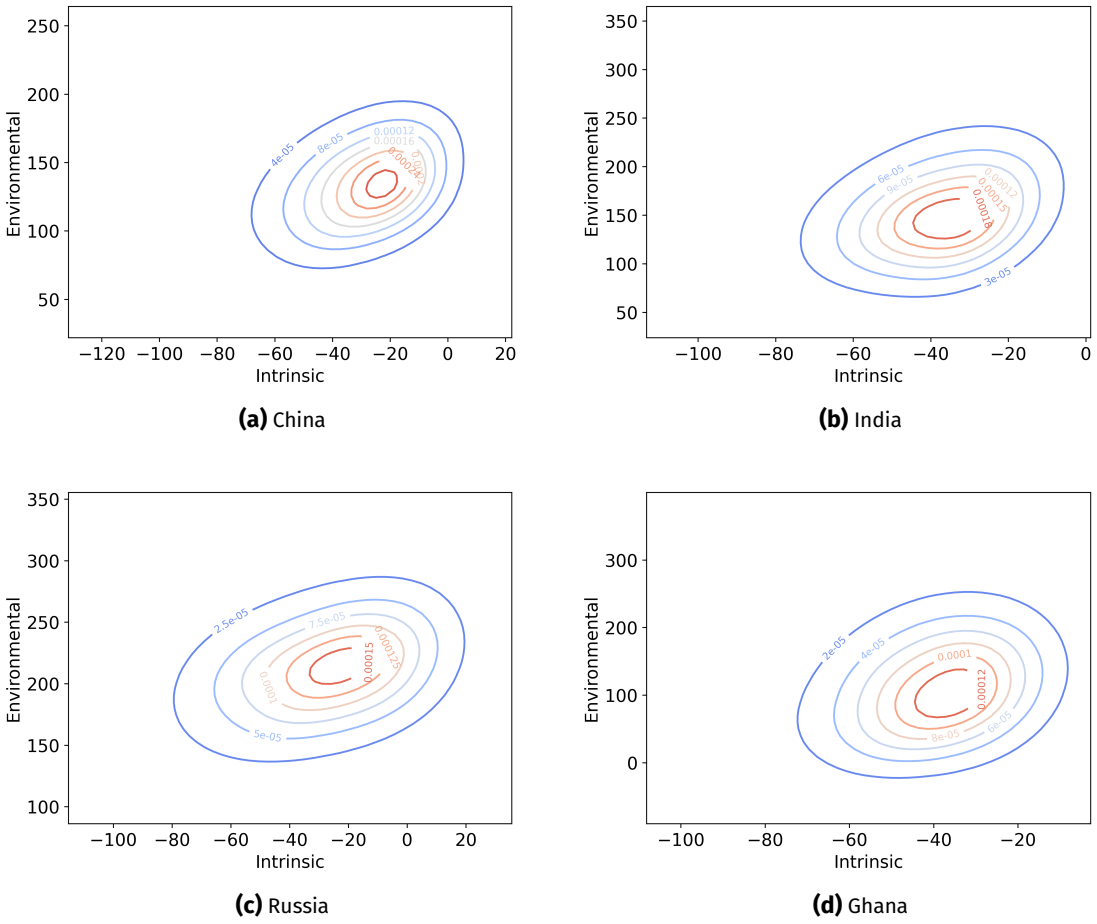


Figure 2.4.1. Bivariate density contour of the two indices

Table 2.4.4 reports the average partial effects for the variables in the environmental index. Our estimates indicate that being higher up in the income distribution is associated with a lower disability rate. Increasing the income percentile by 5 percentage points is associated with a decrease in the probability of being disabled by 1.1 percentage points in the Chinese sample, and by 0.5 percentage points in the other samples. An additional year of education has a negative partial effect on the predicted probability of being disabled in all the country samples, except for the Ghanaian one, in line with the coefficient estimates (see Table 2.4.2).

Contrary to our expectations, household size has a positive impact on the disability rate. The largest effect is observed in the Russian sample, with an estimated average partial effect of 0.8 percentage points.

Turning to the average partial effects of the variables of social cohesion, we observe mixed directions of impacts of these variables on the predicted probability of being disabled, both across countries and within the variable groups. Among the community involvement variables, in particular, more involvement in local affairs (a one standard deviation increase in the corresponding

score) is associated with an increased disability rate in the Chinese sample (0.3% points) and with a decreased disability rate in the Ghanaian (4.5% points), Russian (1.2% points), and Indian (1.4% points) samples. Having friends over, on the other hand, is associated with an increase in the disability rate in all the country samples, except for the Chinese one, where the estimated partial effect is nill. Working with other people in the neighborhood leads to a decrease in the predicted probability of being disabled in all the country samples, except for India, with the largest effect of -2.9% points observed in the Ghanaian sample.

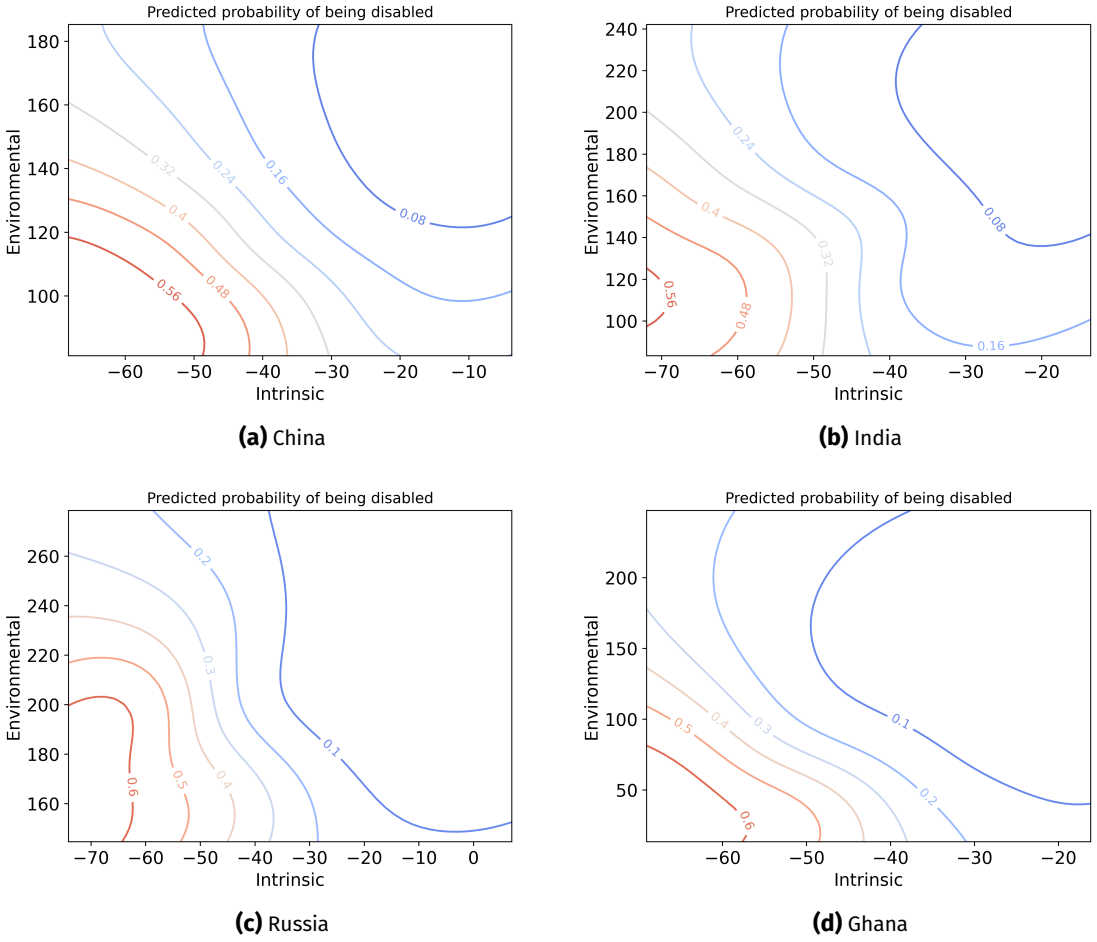
The average partial effects of variables associated with trust have mixed signs both across and within countries. A higher degree of trust in the neighborhood, for example, is associated with a higher disability rate in the Chinese (0.6 percentage points) and Ghanaian (0.8 percentage points) samples and with a lower disability rate in Indian (-1.3 percentage points) and Russian (-0.7 percentage points) samples. Trusting people at work, on the other hand, is associated with a decrease in the predicted probability of being disabled in all the country samples except for the Indian sample.

To summarize, we have mixed results regarding how each of the covariates in the two indices affects the predicted probability of being disabled. For the variables in the intrinsic index, we mostly find the signs of the estimated effect to conform to prior expectations stemming from existing literature. In the case of the variables in the environmental index, however, we have more inconsistent results both across and within countries. Unfortunately, in some cases, the estimated effects contradicted our expectations based on health and gerontology literature (see e.g., Noguchi, Kondo, Saito, Nakagawa-Senda, and Suzuki, 2019; Fujihara, Miyaguni, Tsuji, and Kondo, 2022).

2.4.3 The interaction terms

This section discusses how the estimated index levels interact non-parametrically to determine the predicted probability of being disabled. Since the model does not allow for predictions outside the support of the indices, we begin by looking at the joint distribution of the two indices.

Figure 2.4.1 plots the bivariate density contours of the intrinsic and environmental indices. We can see that, to varying degrees, a positive correlation exists between the two structural indices in all the country samples. Importantly, we observe these correlation structures even though we do not have any common variables influencing both the intrinsic and the environmental index. In what follows, we trim the index values to lie between the 5th and 95th percentiles of their marginal distributions, where the masses are concentrated.



(a) China **(b)** India **(c)** Russia **(d)** Ghana

Figure 2.4.2. Predicted probability of being disabled: Both Indices

Figure 2.4.2 depicts the predicted probability of being disabled as a function of the intrinsic and environmental indices. We can see that in all the country samples, the disability rate is decreasing in both indices. Panel (d) of Figure 2.4.2, for example, illustrates that in the Russian sample, the disability rate goes from 60 percentage points at the lowest values of the indices to 10 percentage points at the highest values. We can also detect non-linear interaction effects between the indices. In particular, the intrinsic gradient of the predicted probability of being disabled is more prominent (in absolute terms) at lower levels of the environmental index.

To get a more concrete idea on the magnitude of the non-linearities, figures 2.4.3 and 2.4.4 depict slices from the bivariate probability function.

In Figure 2.4.3 specifically, we can see the dependence of the predicted probability of being disabled on the intrinsic index when fixing the environmental index at its 5th (dashed line) and 95th (solid line) percentiles. The most striking difference is observed in the Ghanaian sample, where the predicted probability of being disabled changes only by 18% points when moving from

the 5th to the 95th percentile of the intrinsic index at the 95th percentile of the environmental index, whereas the difference is 54% points at the 5th percentile of the environmental index.

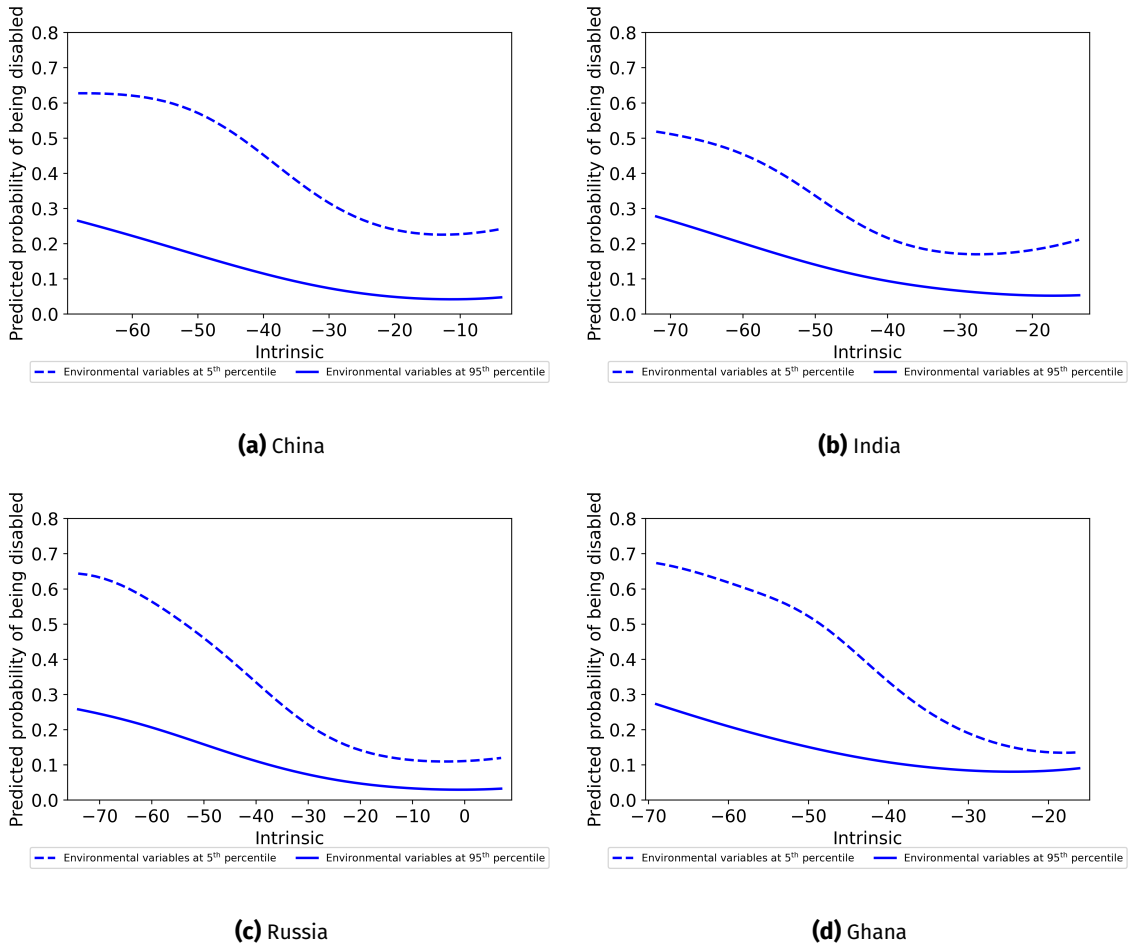


Figure 2.4.3. Predicted probability of being disabled: Intrinsic Index. Depicts how the disability rate depends on the intrinsic index at different quantiles of the environmental index

Figure 2.4.4 presents analogous plots for the intrinsic index's non-linear effects on the probability function's environmental gradient. In this case, we observe even more pronounced non-linear effects. In particular, continuing with the Ghanaian sample, the increase in the predicted probability of being disabled along the environmental index is eight times as large at lower values of the intrinsic index (40% points), as it is at higher values of the intrinsic index (5% points).

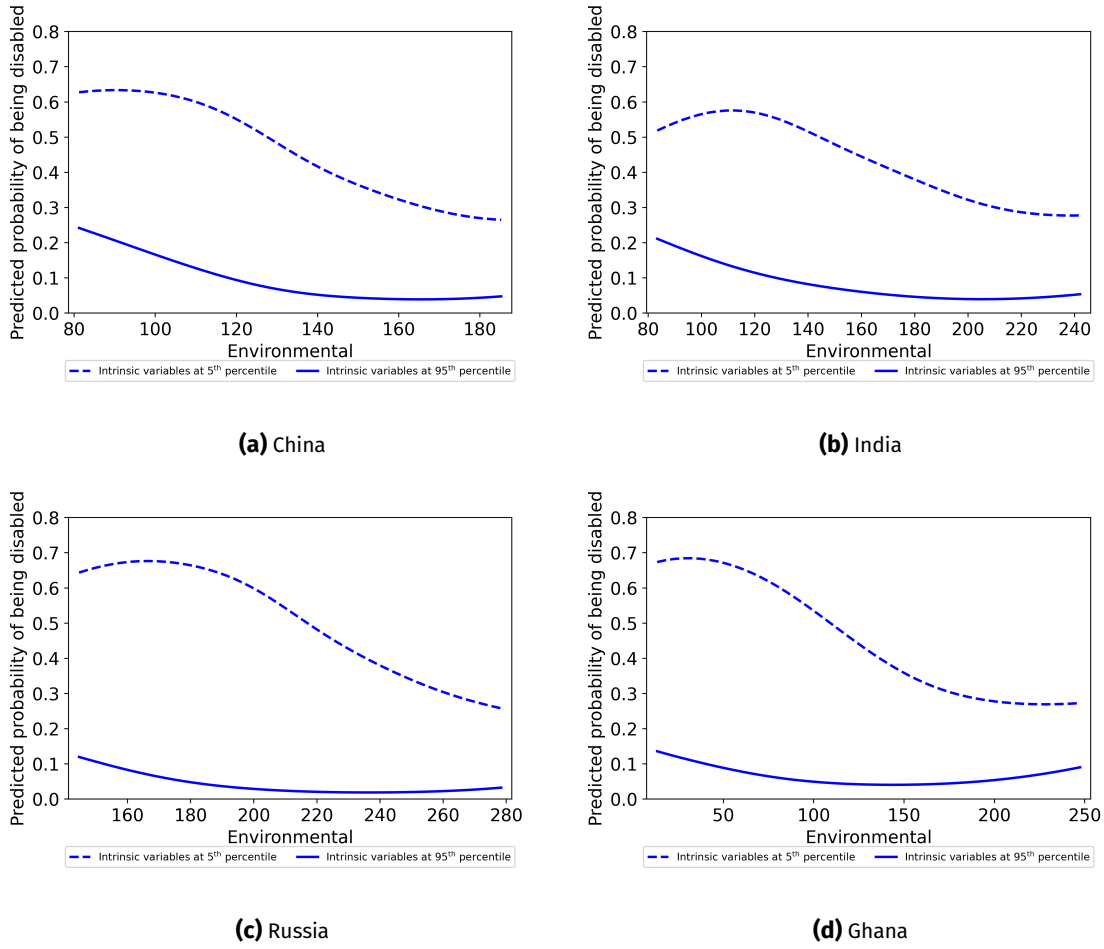


Figure 2.4.4. Predicted probability of being disabled: Environmental Index. Depicts how the disability rate depends on the environmental index at different quantiles of the intrinsic index

The difference in the non-linear effects speaks to the intrinsic index's relative importance for predicting disability. However, as we saw in Figure 2.4.3, the environmental index non-trivially affects how the disability rate reacts to changes in the intrinsic index.

In our last set of results, we discuss the structural dependency of the predicted probability of being disabled on the intrinsic and environmental indices through the means of respective average structural functions (ASF). As discussed earlier, the ASFs represent how the disability rate depends on the intrinsic (environmental) index when the environmental (intrinsic) effects have been integrated out using their marginal distribution. Figure 2.4.5 illustrates the estimated ASFs of the intrinsic index (solid lines), as well as the corresponding 95% confidence interval (the area between the dashed lines). In all the country samples, ASF is a decreasing function of the intrinsic index. The ASF values are of similar magnitudes across the country samples and lie between just under 50% points (highest value observed in the Russian sample, panel (d) of Figure 2.4.5) for the initial values of the intrinsic index and goes down to under 10% points. This

negative relationship is especially strong for the index values in the middle part of its distribution. Closer to the right tail of the support of the intrinsic index, the ASF either plateaus (in the Russian and Ghanaian samples) or displays an ever-so-slight upward slope (in the Indian and Chinese samples). In addition, the ASFs are rather precisely estimated, as evidenced by the narrow error bounds, in most of the country samples.

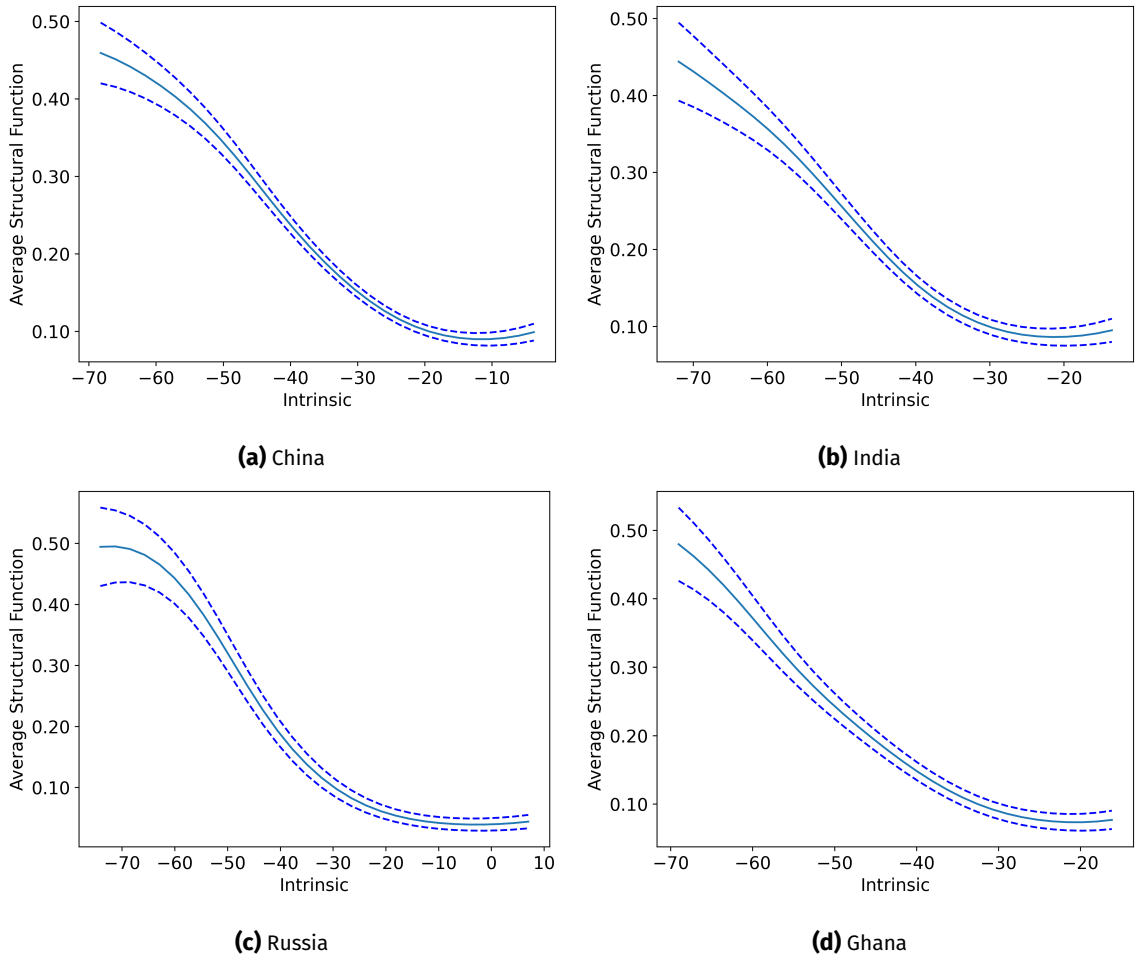


Figure 2.4.5. Average structural function, intrinsic index

The estimates of the ASF for the environmental index are presented in Figure 2.4.6. As in the case of the intrinsic index, the ASF of the environmental index has a negative slope across all country samples. Considering the Chinese sample (panel (a) of Figure 2.4.6), the ASF of the environmental index is initially around 37.5% and gradually declines until around 10% points for higher values of the environmental index. We also observe larger standard errors of the estimates, with significantly wider confidence intervals compared to Figure 2.4.5.

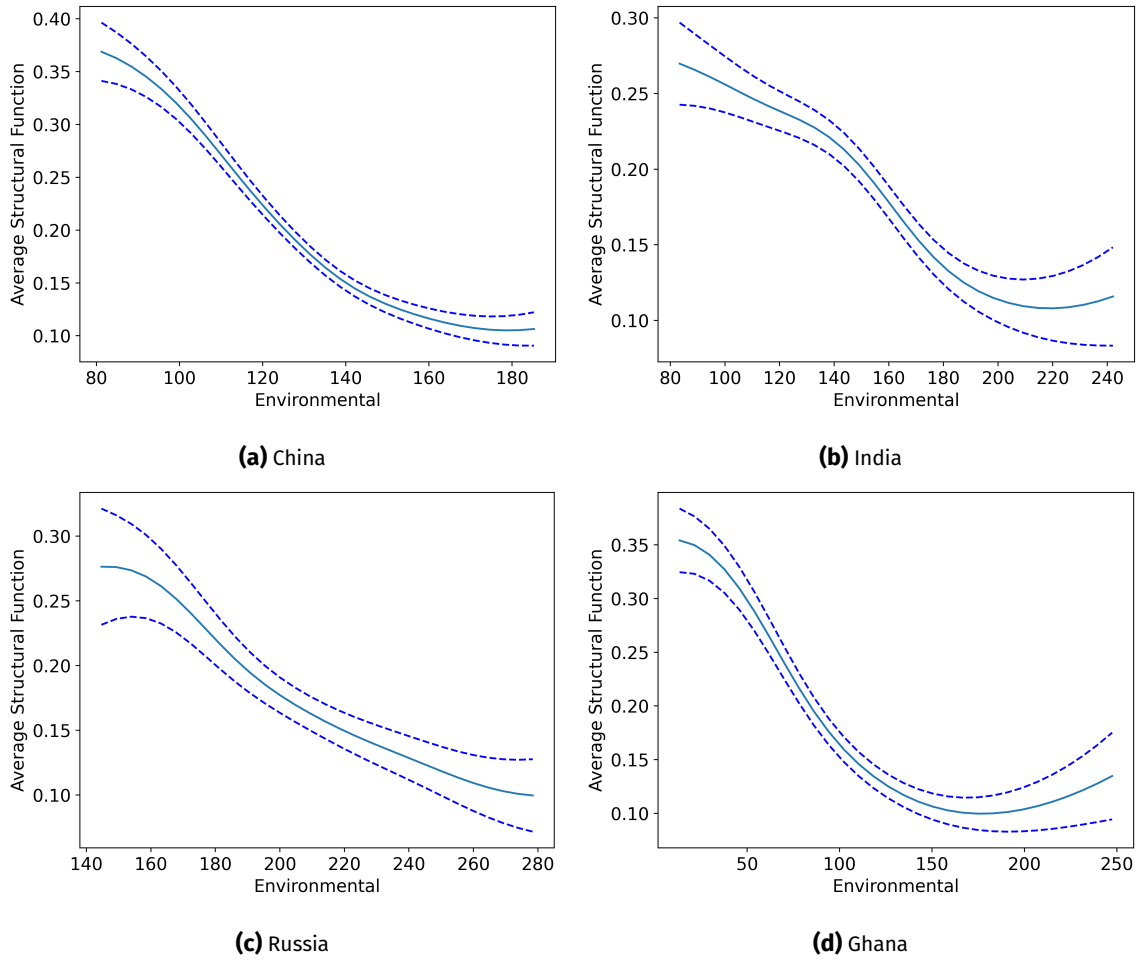


Figure 2.4.6. Average structural function, environmental index

To summarize our results, due to limitations of the data at hand, we were not able to estimate the environmental index as well as the intrinsic one. Nevertheless, we were able to identify considerable non-linear interaction effects between the two indices. Importantly, these non-linearities arise without any mechanical effects that could have potentially be driven by common index variables.

Our main results, in particular the shapes of the ASF functions and the non-linear interaction terms between the structural indices persist through considerations of looser (30th percentile) and stricter (10th percentile) cut-offs. For the full set of tables and figures of the results we refer to appendices 2.A and 2.B.

2.5 Conclusion

We study functional abilities in older age as an outcome of intrinsic capacities, which constitute biological age, gender and accumulated frailties, and environmental factors characterized by socio-

economic indicators and social cohesion. We use the data from the first wave of the World Health Organization's Study on global AGEing (World Health Organization, 2022) to estimate a semi-parametric double index model for the predicted probability of being disabled. This estimation procedure allows us to model the link function of the intrinsic and environmental indices in a non-parametric fashion, thus allowing for complex, non-linear interaction terms between the two.

This framework reveals significant interaction effects between the environmental and intrinsic indices. Importantly, we find that the curvature of the intrinsic gradient of the probability link function is affected by the level of the environmental index in a non-negligible way. In particular, the change in the predicted probability of being disabled moving from the top 5th percentile of the intrinsic index to the bottom 5th percentile can be up to two times as large at the low levels of the environmental index as the change at the high levels of the latter index (2.4.3). We obtain these results despite the relatively simple model specification and no common variables shared among the two indices, meaning that the correlation between the indices does not arise mechanically.

Our results are an important first step in estimating how environmental factors and intrinsic capacities interact to determine functional abilities in older people. At the same time, we acknowledge the limitations of our study. In particular, the variables in the environmental index are poorly identified compared to their intrinsic counterparts. Additionally, the normalizing variable in the environmental index potentially suffers from endogeneity issues. Most of the issues in our study stem from the data at hand. In future work, we intend to tackle the research question with datasets with a more granular and wider set of variables for the environmental index.

We leave a few extensions of the estimation procedure employed in this study for future work. We plan to: 1) Implement a multiple choice double index model to differentiate between different levels of disability. 2) Extend the estimator to apply it to a continuous outcome variable to estimate the impact of the intrinsic capacities and environmental factors on subjective well-being, retrieved as a continuous score variable from the SAGE data.

Appendix 2.A Additional results: 10th percentile cutoff for disability

2.A.1 Coefficient estimates

Table 2.A.1. Estimated parameters, intrinsic index

	China	India	Russia	Ghana
Age	−1.000	−1.000	−1.000	−1.000
Male	3.226**	0.585	−4.190	3.055
	(1.462)	(1.795)	(4.096)	(2.424)
Grip Strength	32.990***	24.412**	25.901	−10.647
	(6.742)	(10.503)	(16.062)	(11.345)
Cognition Score	51.101***	37.189***	86.296***	73.291***
	(6.937)	(9.021)	(23.226)	(13.525)
Diabetes	−2.000	−2.320	−15.907***	−11.761**
	(2.505)	(2.756)	(5.455)	(5.272)
Lung	−0.030	−6.185	−10.223**	−17.577
	(2.206)	(4.216)	(5.082)	(11.967)
Arthritis	−3.713***	−5.214***	−11.341***	−7.930***
	(1.392)	(1.651)	(3.949)	(2.337)
Angina	−10.135***	−12.743***	−16.267***	−10.079***
	(1.908)	(1.773)	(4.600)	(2.623)
Stroke	−19.760***	−4.817	−9.862	−30.211***
	(3.483)	(5.105)	(7.254)	(8.256)
Asthma	−11.667***	−10.168***	−14.173**	4.915
	(2.901)	(2.322)	(5.900)	(4.622)
Hypertension	1.338	−3.575**	5.985*	−3.904*
	(1.293)	(1.628)	(3.484)	(2.210)
Depression	−29.373***	−5.553***	−17.541***	10.153**
	(5.529)	(2.026)	(5.726)	(4.495)
Observations	11062	5458	2417	3652
Note:				***p<0.01;**p<0.05;*p<0.1

Table 2.A.2. Estimated parameters, environmental index

	China	India	Russia	Ghana
Income Percentile	1.000	1.000	1.000	1.000
Education Years	2.063***	−0.050	10.601	−0.022
	(0.561)	(1.119)	(8.798)	(1.745)
Num. HH Memberds	−1.289	−3.602***	5.849	−3.170
	(1.275)	(0.900)	(10.504)	(2.126)
Cohabiting	−6.995*	11.984	10.112	8.648
	(4.178)	(7.462)	(25.293)	(16.064)
Comm Loc. Affairs	−5.667	15.019	61.959	48.670*
	(3.824)	(9.209)	(52.747)	(25.170)
Comm Community Leader	−23.507***	19.411**	0.025	−13.942
	(4.004)	(8.291)	(27.633)	(9.396)
Comm Soc. Clubs	1.617	5.234	49.745	23.421
	(3.485)	(7.204)	(51.202)	(16.074)
Comm Work Nbhd.	7.002***	13.675***	43.403	22.659*
	(2.567)	(5.242)	(38.418)	(13.666)
Comm Friends Over	2.249	−5.997*	9.662	−4.415
	(2.117)	(3.483)	(17.069)	(7.149)
Comm Diff. Nbhd.	12.378***	−7.019*	5.099	4.488
	(2.440)	(3.877)	(15.643)	(6.620)
Comm Got Out	12.212***	10.054**	30.316	26.643**
	(2.786)	(5.097)	(25.995)	(12.519)
Trust Nbhd	−4.042	15.427***	−9.385	−7.833
	(3.778)	(5.353)	(13.649)	(11.138)
Trust Work	4.962	−14.282***	56.122	22.149
	(3.664)	(4.795)	(46.415)	(14.533)
Trust Stangers	−5.392***	2.864	4.675	−16.569
	(2.065)	(3.418)	(13.504)	(11.451)
Trust Gen	−9.157	−4.194	13.035	−2.455
	(5.737)	(7.023)	(26.519)	(15.406)
Trust Someone	12.698	17.401**	−60.617	−39.351
	(11.603)	(8.806)	(57.965)	(24.353)
Gov Impact	4.240**	−26.372***	−18.992	1.084
	(1.992)	(6.556)	(15.941)	(6.743)
Gov Freedom	6.852***	22.310***	5.068	−15.447
	(2.152)	(5.649)	(9.976)	(9.600)
Observations	11062	5458	2417	3652
Note:				***p<0.01;**p<0.05;*p<0.1

2.A.2 Averge partial effects

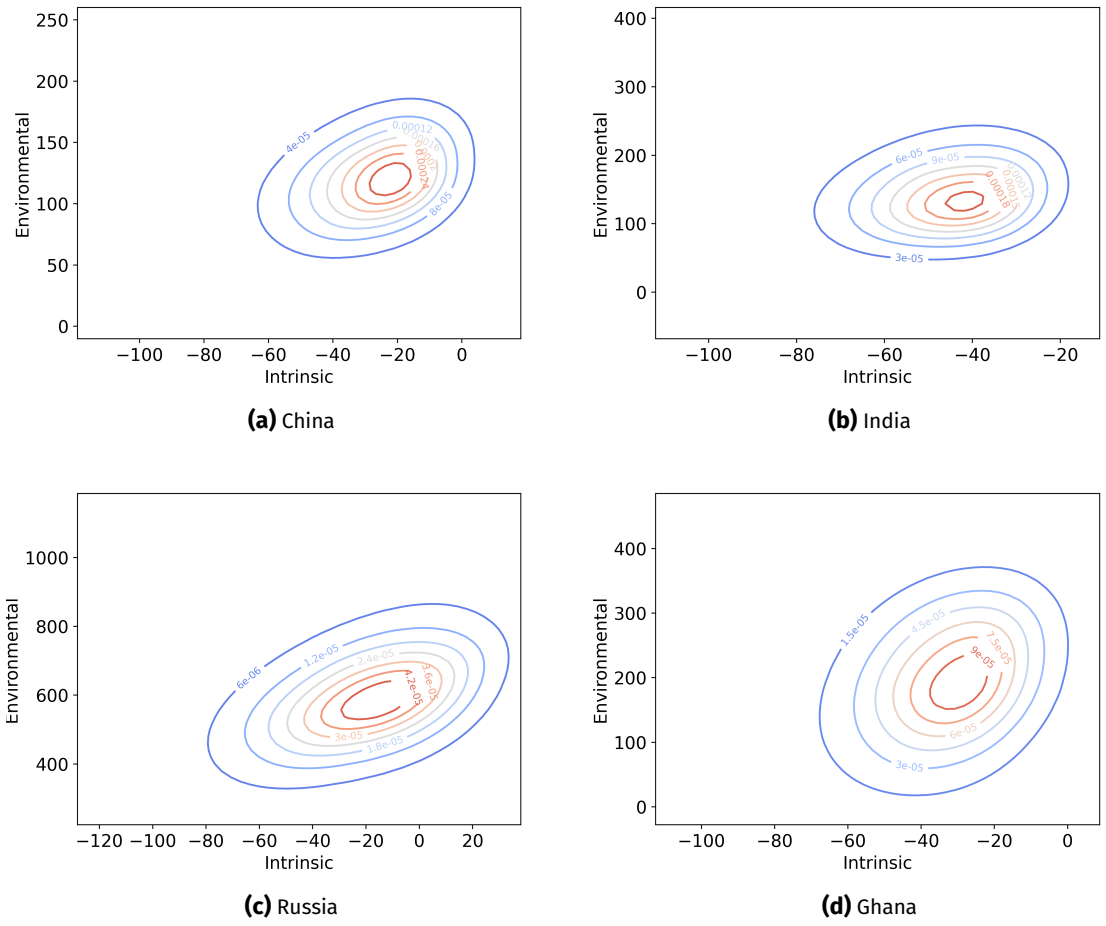
Table 2.A.3. Average partial effects, intrinsic index

	China	India	Russia	Ghana
Age	0.004	0.004	0.003	0.004
Male	−0.011	−0.002	0.011	−0.011
Grip Strength	−0.012	−0.008	−0.008	0.005
Cognition Score	−0.019	−0.015	−0.029	−0.024
Diabetes	0.007	0.010	0.045	0.051
Lung	0.000	0.028	0.028	0.082
Arthritis	0.013	0.023	0.033	0.031
Angina	0.042	0.066	0.051	0.041
Stroke	0.092	0.022	0.027	0.155
Asthma	0.050	0.051	0.040	−0.016
Hypertension	−0.005	0.015	−0.016	0.014
Depression	0.143	0.025	0.049	−0.029

Table 2.A.4. Average partial effects, environmental index

	China	India	Russia	Ghana
Income Percentile	−0.006	−0.003	−0.001	−0.002
Education Years	−0.002	0.000	−0.003	0.000
Num. HH Memberds	0.002	0.003	−0.001	0.002
Cohabiting	0.008	−0.008	−0.003	−0.004
Comm Loc. Affairs	0.003	−0.007	−0.013	−0.016
Comm Community Leader	0.012	−0.010	0.000	0.008
Comm Soc. Clubs	−0.001	−0.003	−0.009	−0.011
Comm Work Nbhd.	−0.007	−0.009	−0.009	−0.010
Comm Friends Over	−0.002	0.005	−0.002	0.003
Comm Diff. Nbhd.	−0.012	0.006	−0.001	−0.002
Comm Got Out	−0.010	−0.005	−0.010	−0.014
Trust Nbhd	0.004	−0.010	0.003	0.004
Trust Work	−0.004	0.009	−0.017	−0.010
Trust Stangers	0.005	−0.002	−0.001	0.009
Trust Gen	0.011	0.003	−0.004	0.001
Trust Someone	−0.016	−0.012	0.017	0.017
Gov Impact	−0.005	0.018	0.005	−0.000
Gov Freedom	−0.007	−0.021	−0.001	0.008

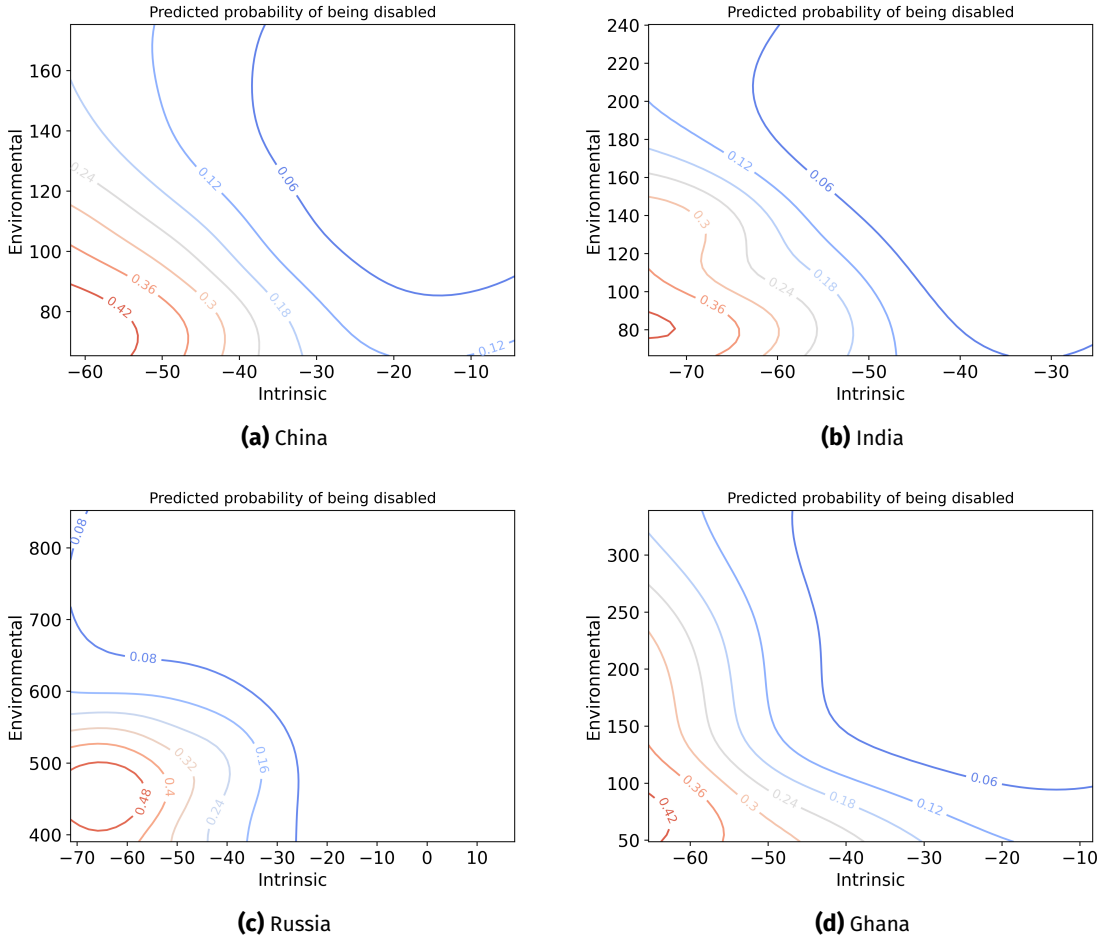
2.A.3 Interaction terms



(a) China **(b)** India

(c) Russia **(d)** Ghana

Figure 2.A.1. Bivariate density contour of the two indices



(a) China **(b)** India **(c)** Russia **(d)** Ghana

Figure 2.A.2. Predicted probability of being disabled

Table 2.A.5. The effect of intrinsic index on disability rate at different values of the environmental index

Environmental Index	China	India	Russia	Ghana
Low	0.31	0.32	0.40	0.32
High	0.15	0.09	0.08	0.15

Table 2.A.6. The effect of the environmental index on disability rate at different values of the intrinsic index

Intrinsic Index	China	India	Russia	Ghana
Low	0.27	0.27	0.35	0.26
High	0.11	0.04	0.02	0.09

2.A.4 Average structural functions

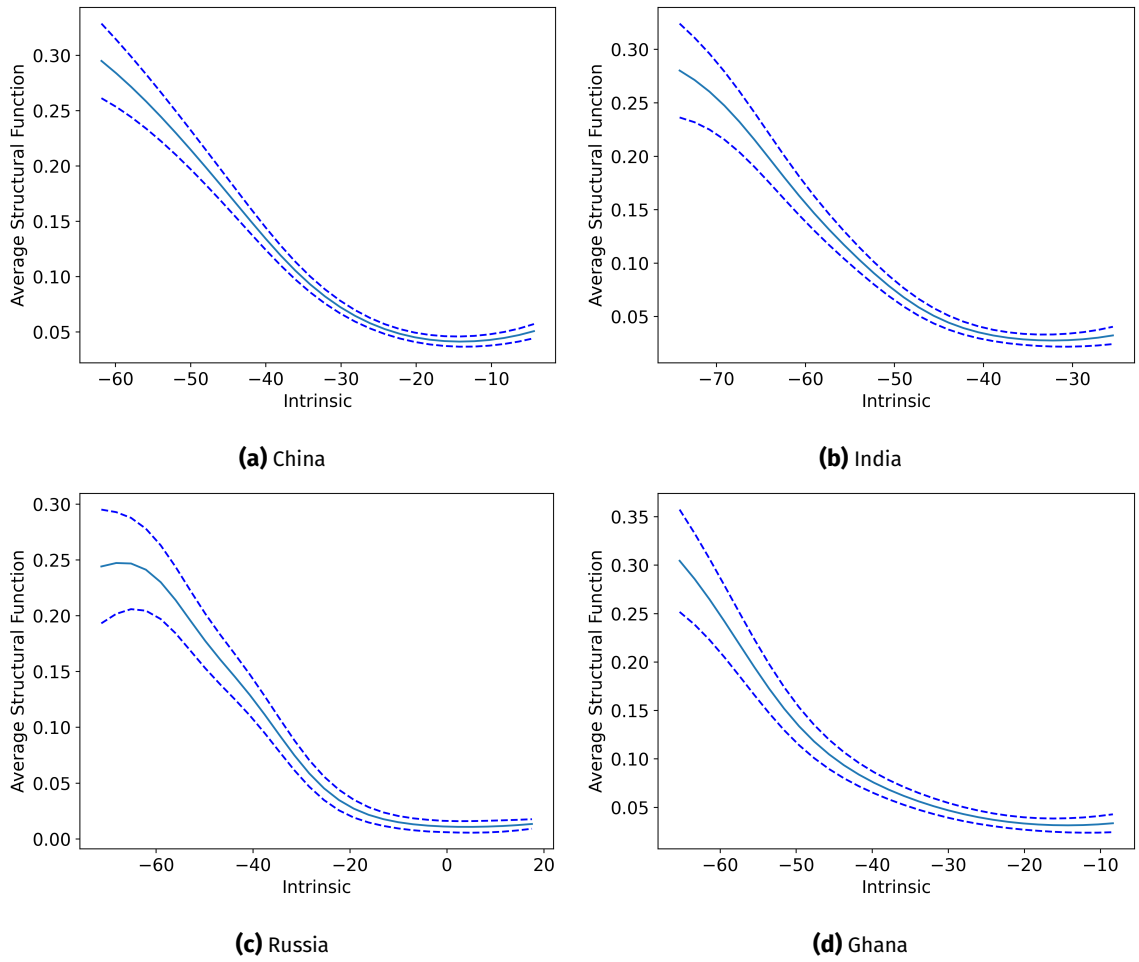
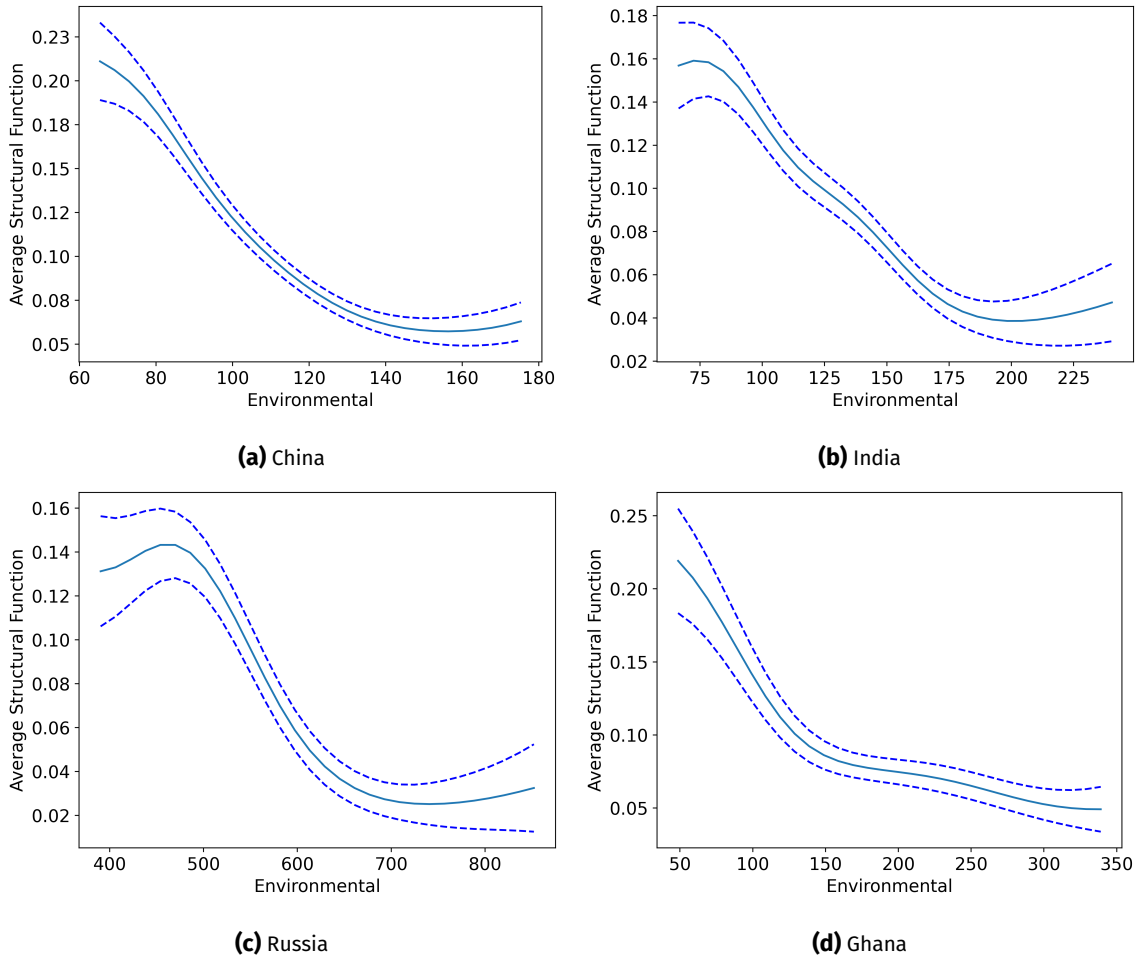


Figure 2.A.3. Average structural function, intrinsic index



(a) China **(b)** India

(c) Russia **(d)** Ghana

Figure 2.A.4. Average structural function, environmental index

Appendix 2.B Additional results: 30th percentile cutoff for disability

2.B.1 Coefficient estimates

Table 2.B.1. Estimated parameters, intrinsic index

	China	India	Russia	Ghana
Age	−1.000	−1.000	−1.000	−1.000
Male	2.231	10.620***	−4.896	5.242***
	(1.382)	(1.581)	(3.657)	(1.535)
Grip Strength	58.105***	18.152*	90.173***	−24.196***
	(7.745)	(10.071)	(22.576)	(6.943)
Cognition Score	74.305***	38.636***	67.202***	56.112***
	(8.032)	(8.872)	(19.084)	(7.958)
Diabetes	−3.857*	−0.940	−17.675***	−10.480***
	(2.330)	(2.835)	(5.750)	(3.713)
Lung	−1.377	−4.094	−17.314***	−0.364
	(2.116)	(3.672)	(4.610)	(8.429)
Arthritis	−11.786***	−7.418***	−14.564***	−8.669***
	(1.492)	(1.563)	(3.726)	(1.600)
Angina	−19.287***	−14.291***	−19.864***	−8.945***
	(2.383)	(2.052)	(4.530)	(1.792)
Stroke	−23.394***	−7.226	−29.354***	−23.008***
	(4.269)	(5.158)	(10.438)	(7.184)
Asthma	−11.642***	−13.010***	0.849	0.191
	(2.895)	(2.497)	(5.574)	(3.201)
Hypertension	−2.460**	−3.033**	−9.326***	−1.819
	(1.232)	(1.543)	(3.344)	(1.363)
Depression	−34.240***	−14.052***	−15.992***	−0.887
	(6.477)	(2.348)	(5.681)	(2.535)
Observations	11062	5458	2417	3652
Note:				***p<0.01;**p<0.05;*p<0.1

Table 2.B.2. Estimated parameters, environmental index

	China	India	Russia	Ghana
Income Percentile	1.000	1.000	1.000	1.000
Education Years	0.807***	1.339**	12.149	−0.314
	(0.273)	(0.633)	(7.896)	(0.722)
Num. HH Memberds	−1.946***	−3.184***	19.920	−2.211**
	(0.717)	(0.658)	(14.067)	(0.948)
Cohabiting	2.882	5.102	−23.653	−7.003
	(2.641)	(5.202)	(20.515)	(7.261)
Comm Loc. Affairs	−7.964***	11.952**	44.218	35.565***
	(2.306)	(4.669)	(32.843)	(8.885)
Comm Community Leader	24.724***	−6.522*	−46.322	−16.240***
	(3.393)	(3.945)	(32.521)	(4.986)
Comm Soc. Clubs	0.309	11.448***	12.221	5.031
	(1.776)	(4.181)	(17.578)	(3.431)
Comm Work Nbhd.	−0.861	−1.235	47.999	35.047***
	(1.289)	(2.806)	(31.553)	(8.490)
Comm Friends Over	1.768	−7.783***	−39.863	−16.046***
	(1.208)	(2.291)	(26.287)	(4.612)
Comm Diff. Nbhd.	5.584***	7.460***	46.178	1.173
	(1.213)	(2.302)	(29.717)	(3.262)
Comm Got Out	13.074***	−2.173	0.592	7.174**
	(1.733)	(2.925)	(6.720)	(3.306)
Trust Nbhd	−2.250	0.833	27.357	2.931
	(2.101)	(2.750)	(18.226)	(4.510)
Trust Work	2.764	−3.953	−3.349	21.331***
	(2.169)	(2.679)	(8.919)	(6.309)
Trust Stangers	−0.940	0.440	17.982	−15.588***
	(1.204)	(2.212)	(14.583)	(4.776)
Trust Gen	7.192**	8.090*	−25.141	1.475
	(3.101)	(4.494)	(22.625)	(7.005)
Trust Someone	−0.197	18.090***	−25.112	−6.222
	(7.031)	(6.747)	(23.292)	(8.054)
Gov Impact	0.609	2.064	−18.529	1.386
	(0.989)	(2.249)	(13.539)	(3.123)
Gov Freedom	3.564***	1.406	11.647	−4.419
	(1.211)	(1.846)	(9.340)	(3.400)
Observations	11062	5458	2417	3652
Note:				***p<0.01;**p<0.05;*p<0.1

2.B.2 Averge partial effects

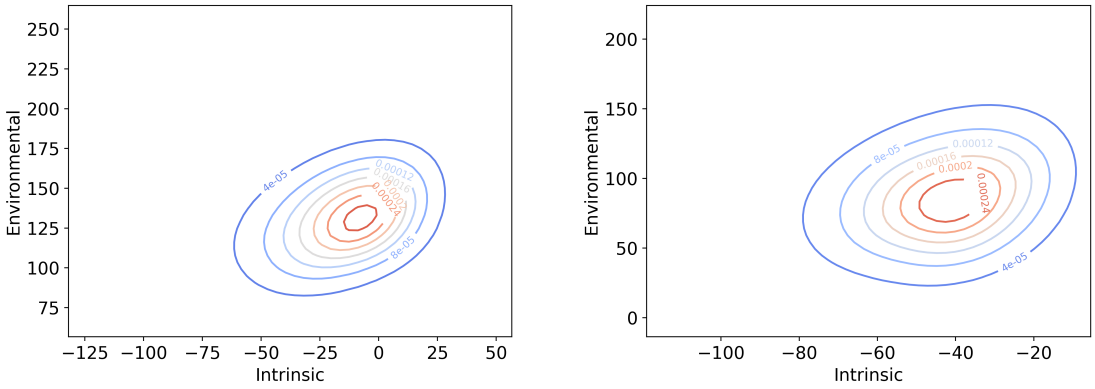
Table 2.B.3. Average partial effects, intrinsic index

	China	India	Russia	Ghana
Age	0.005	0.007	0.005	0.009
Male	−0.012	−0.076	0.022	−0.045
Grip Strength	−0.034	−0.011	−0.049	0.026
Cognition Score	−0.045	−0.027	−0.039	−0.048
Diabetes	0.022	0.006	0.086	0.095
Lung	0.008	0.029	0.083	0.003
Arthritis	0.070	0.054	0.071	0.079
Angina	0.122	0.113	0.105	0.082
Stroke	0.146	0.053	0.141	0.206
Asthma	0.070	0.100	−0.004	−0.002
Hypertension	0.013	0.021	0.044	0.015
Depression	0.210	0.110	0.077	0.008

Table 2.B.4. Average partial effects, environmental index

	China	India	Russia	Ghana
Income Percentile	−0.016	−0.009	−0.003	−0.008
Education Years	−0.003	−0.002	−0.008	0.001
Num. HH Memberds	0.006	0.006	−0.013	0.004
Cohabiting	−0.009	−0.009	0.016	0.011
Comm Loc. Affairs	0.012	−0.014	−0.022	−0.044
Comm Community Leader	−0.030	0.009	0.021	0.031
Comm Soc. Clubs	−0.001	−0.015	−0.006	−0.009
Comm Work Nbhd.	0.002	0.002	−0.022	−0.053
Comm Friends Over	−0.005	0.016	0.026	0.032
Comm Diff. Nbhd.	−0.016	−0.016	−0.025	−0.002
Comm Got Out	−0.029	0.003	−0.000	−0.014
Trust Nbhd	0.005	−0.001	−0.019	−0.005
Trust Work	−0.006	0.007	0.003	−0.035
Trust Stangers	0.002	−0.001	−0.011	0.028
Trust Gen	−0.024	−0.014	0.018	−0.002
Trust Someone	0.001	−0.032	0.017	0.009
Gov Impact	−0.002	−0.004	0.013	−0.003
Gov Freedom	−0.010	−0.003	−0.009	0.008

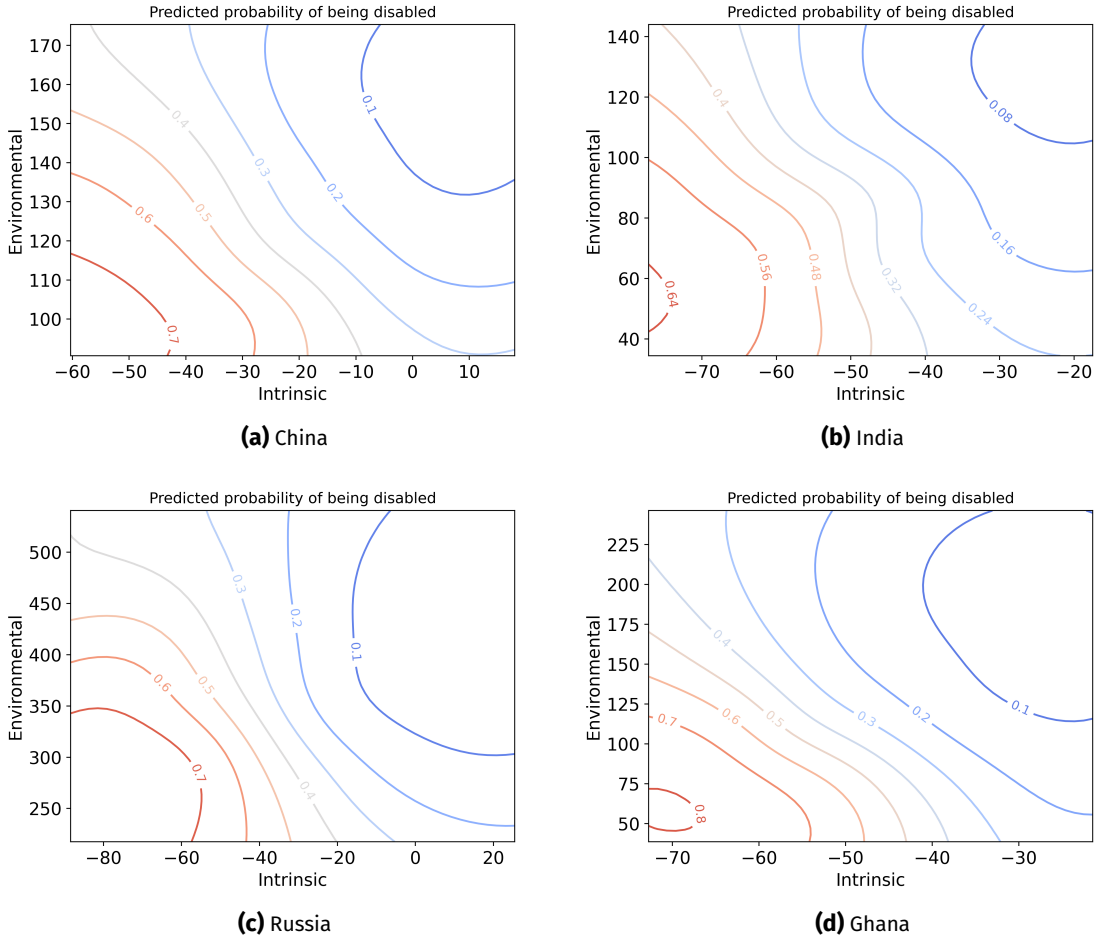
2.B.3 Interaction terms



(a) China **(b)** India

(c) Russia **(d)** Ghana

Figure 2.B.1. Bivariate density contour of the two indices



(a) China **(b)** India **(c)** Russia **(d)** Ghana

Figure 2.B.2. Predicted probability of being disabled

Table 2.B.5. The effect of intrinsic index on disability rate at different values of the environmental index

Environmental Index	China	India	Russia	Ghana
Low	0.42	0.38	0.56	0.55
High	0.34	0.34	0.33	0.27

Table 2.B.6. The effect of the environmental index on disability rate at different values of the intrinsic index

Intrinsic Index	China	India	Russia	Ghana
Low	0.32	0.21	0.39	0.41
High	0.23	0.17	0.16	0.13

2.B.4 Average structural functions

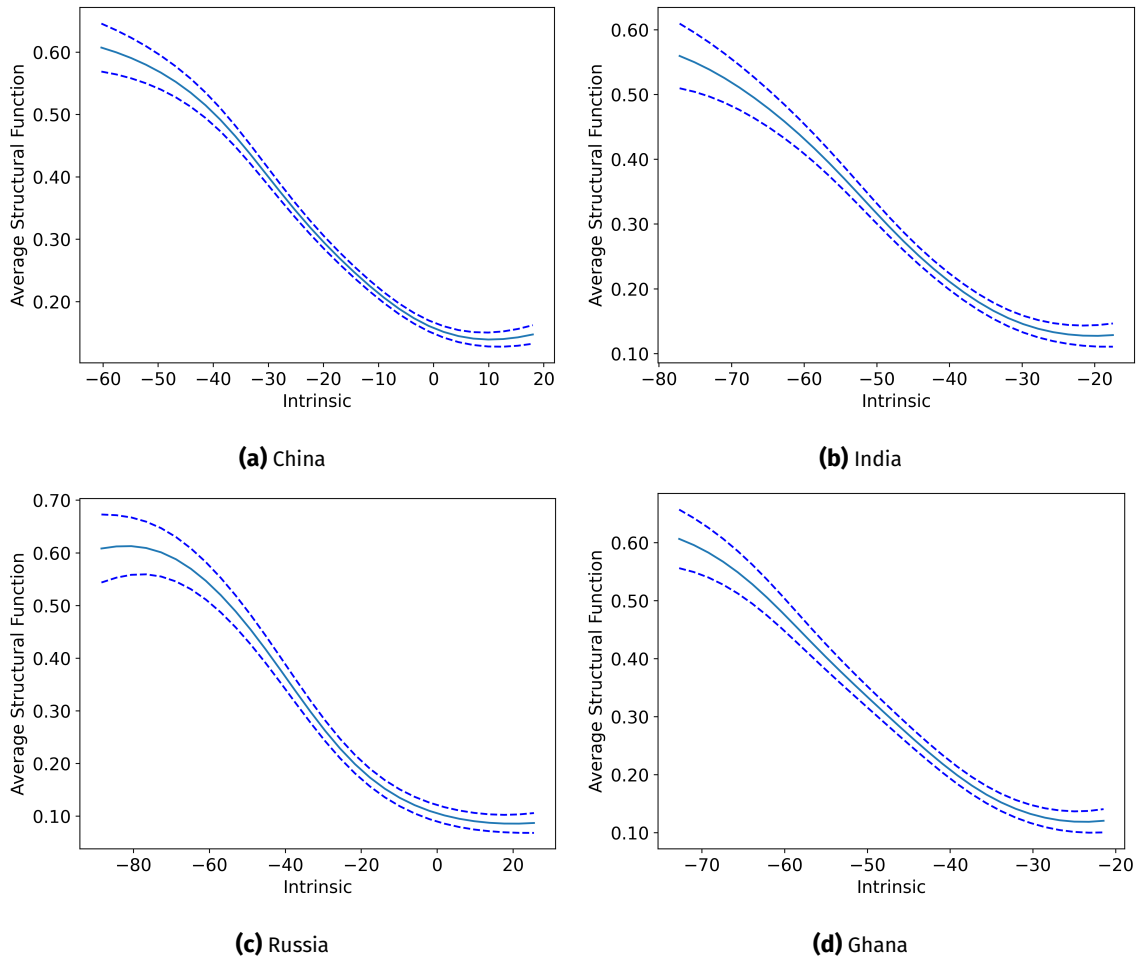
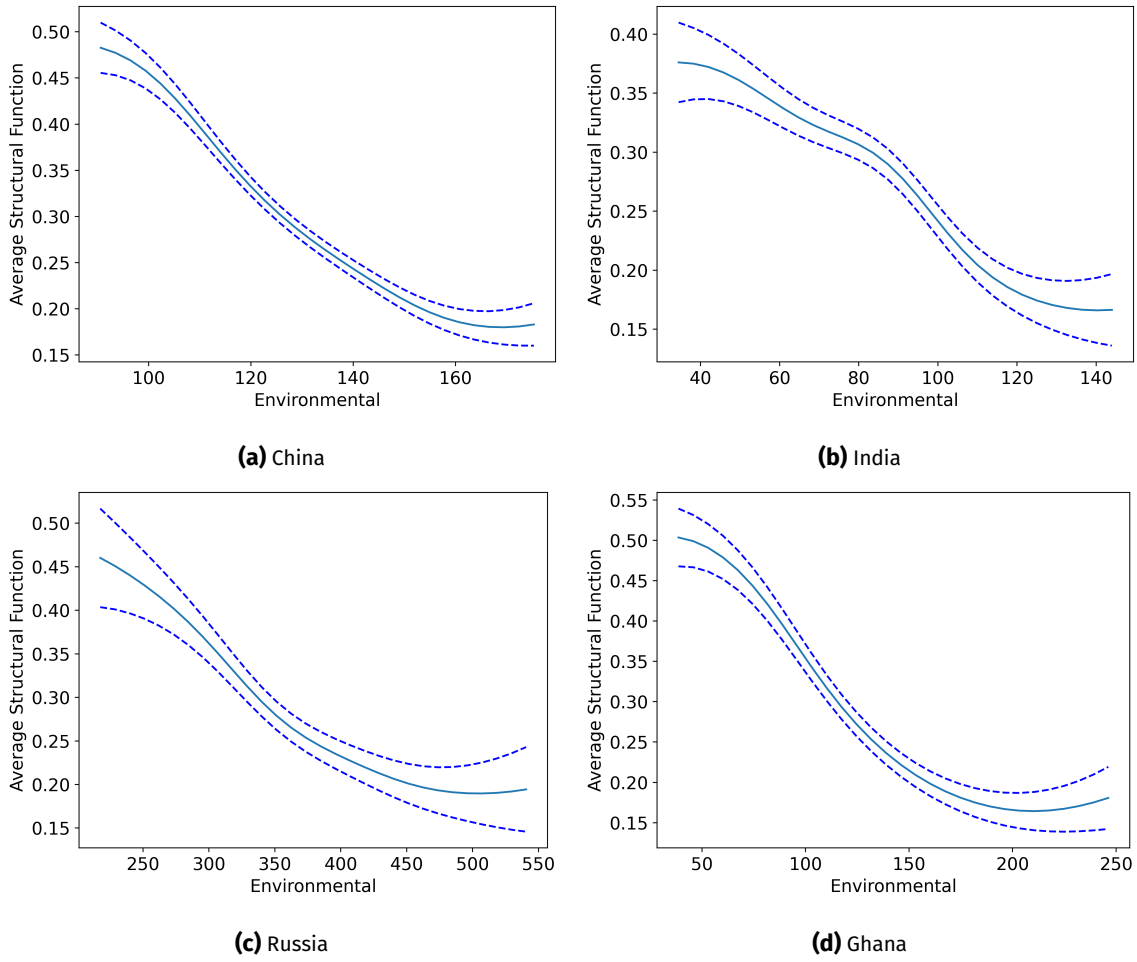


Figure 2.B.3. Average structural function, intrinsic index



(a) China **(b)** India

(c) Russia **(d)** Ghana

Figure 2.B.4. Average structural function, environmental index

References

Auyeung, Tung Wai, Timothy Kwok, Jenny Lee, Ping Chung Leung, Jason Leung, and Jean Woo. 2008. "Functional Decline in Cognitive Impairment – The Relationship between Physical and Cognitive Function." *Neuroepidemiology* 31 (3): 167–73. [88]

Backe, Ingeborg Flåten, Grete Grindal Patil, Ragnhild Bang Nes, and Jocelyne Clench-Aas. 2018. "The relationship between physical functional limitations, and psychological distress: Considering a possible mediating role of pain, social support and sense of mastery." *SSM - Population Health* 4: 153–63. [96]

Beydoun, May A., and Barry M. Popkin. 2005. "The impact of socio-economic factors on functional status decline among community-dwelling older adults in China." *Social Science & Medicine* 60 (9): 2045–57. [96]

Blundell, Richard W., and James L. Powell. 2003. "Endogeneity in Nonparametric and Semiparametric Regression Models." In *Advances in Economics and Econometrics: Theory and Applications*. Edited by Lars Peter Hansen Mathias Dewatripont and Stephen J. Turnovsky. 1st edition. Vol. 2, Cambridge University Press. Chapter 8, 312–57. [97]

Blundell, Richard W., and James L. Powell. 2004. "Endogeneity in Semiparametric Binary Response Models." *Review of Economic Studies* 71 (3): 655–79. [97]

Bohannon, Richar W. 2019. "Grip Strength: An Indispensable Bomarker For Older Adults." *Clinical Interventions in Aging*, (14): [93]

Burton, Catherine L., Esther Strauss, David Bunce, Michael A. Hunter, and David F. Hultsch. 2009. "Functional Abilities in Older Adults with Mild Cognitive Impairment." *Gerontology* 55 (5): 570–81. [89]

Cunha, Ana Izabel Lopes, Nicola Veronese, Sheila de Melo Borges, and Natalia Aquaroni Ricci. 2019. "Frailty as a predictor of adverse outcomes in hospitalized older adults: A systematic review and meta-analysis." *Ageing Research Reviews* 56 (12): 100960. [88]

Darkwah, Kwasi Adjepong, Samuel Iddi, Justice Nonvignon, and Moses Aikins. 2022. "Characterization of functional disability among older adults in Ghana: A multi-level analysis of the study on global ageing and adult health (SAGE) Wave II." *PLOS ONE* 17 (11): edited by Alexandra J. Mayhew, e0277125. [98]

Drerup, Tilman, Benjamin Enke, and Hans-Martin von Gaudecker. 2017. "The precision of subjective data and the explanatory power of economic models." *Journal of Econometrics* 200 (2): 378–89. Measurement Error Models. [97]

Fujihara, Satoko, Yasuhiro Miyaguni, Taishi Tsuji, and Katsunori Kondo. 2022. "Community-level social participation and functional disability among older adults: A JAGES multilevel longitudinal study." *Archives of Gerontology and Geriatrics* 100: 104632. [105]

Gao, Min, Zhihong Sa, Yanyu Li, Weijun Zhang, Donghua Tian, Shengfa Zhang, and Linni Gu. 2018. "Does social participation reduce the risk of functional disability among older adults in China? A survival analysis using the 2005–2011 waves of the CLHLS data." *BMC Geriatrics* 18 (1): [89]

Hajek, André, Christian Brettschneider, Marion Eisele, Tina Mallon, Anke Oey, Birgitt Wiese, Siegfried Weyerer, Jochen Werle, Angela Fuchs, Michael Pentzek, Uta Gühne, Susanne Röhr, Dagmar Weeg, Horst Bickel, Luca Kleineidam, Michael Wagner, Martin Scherer, Wolfgang Maier, Steffi G. Riedel-Heller, and Hans-Helmut König. 2022. "Social Support and Functional Decline in the Oldest Old." *Gerontology* 68 (2): 200–8. [96]

Kanamori, Satoru, Yuko Kai, Jun Aida, Katsunori Kondo, Ichiro Kawachi, Hiroshi Hirai, Kokoro Shirai, Yoshiki Ishikawa, and Kayo Suzuki. 2014. "Social Participation and the Prevention of Functional Disability in Older Japanese: The JAGES Cohort Study." *PLoS ONE* 9 (6): edited by Jerson Laks, e99638. [89]

Keysor, J. J., A. M. Jette, M. P. LaValley, C. E. Lewis, J. C. Torner, M. C. Nevitt, and D. T. Felson. 2009. "Community Environmental Factors Are Associated With Disability in Older Adults With Functional Limitations: The MOST Study." *Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 65A (4): 393–99. [89]

Klein, Roger, and Francis Vella. 2009. "A semiparametric model for binary response and continuous outcomes under index heteroscedasticity." *Journal of Applied Econometrics* 24 (5): 735–62. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jae.1064. [89, 96–98, 100]

Larnyo, Ebenezer, Baozhen Dai, Jonathan Aseye Nutakor, Sabina Ampon-Wireko, Abigail Larnyo, and Ruth Appiah. 2022. "Examining the impact of socioeconomic status, demographic characteristics, lifestyle and other risk factors on adults' cognitive functioning in developing countries: an analysis of five selected WHO SAGE Wave 1 Countries." *International Journal for Equity in Health* 21 (1): 31. [96]

Liang, Jersey, Joan Bennett, Benjamin Shaw, Ana Quiñones, Wen Ye, Xiao xu, and Mary Beth Ofstedal. 2008. "Gender Differences in Functional Status in Middle and Older Age: Are There Any Age Variations?" *journals of gerontology. Series B, Psychological sciences and social sciences* 63 (10): S282–92. [91]

Lien, Wei-Chih, Nai-Wen Guo, Jer-Hao Chang, Yu-Ching Lin, and Ta-Shen Kuan. 2014. "Relationship of perceived environmental barriers and disability in community-dwelling elderly in Taiwan – a population-based study." *BMC Geriatrics* 14 (1): [89]

Makizako, Hyuma, Hiroyuki Shimada, Takehiko Doi, Kota Tsutsumimoto, and Takao Suzuki. 2015. "Impact of physical frailty on disability in community-dwelling older adults: a prospective cohort study." *BMJ Open* 5 (9): e008462. [88]

Malik, Manzoor Ahmad. 2022. "Functional disability among older adults in India; a gender perspective." *PLOS ONE* 17 (9): edited by Farzad Taghizadeh-Hesary, e0273659. [98]

Maurer, Jürgen. 2009. "Who has a clue to preventing the flu? Unravelling supply and demand effects on the take-up of influenza vaccinations." *Journal of Health Economics* 28 (3): 704–17. [98]

Maurer, Jürgen, Roger Klein, and Francis Vella. 2011. "SUBJECTIVE HEALTH ASSESSMENTS AND ACTIVE LABOR MARKET PARTICIPATION OF OLDER MEN: EVIDENCE FROM A SEMIPARAMETRIC BINARY CHOICE MODEL WITH NONADDITIVE CORRELATED INDIVIDUAL-SPECIFIC EFFECTS." *Review of Economics and Statistics* 93 (3): 764–74. [98]

McGrath, Ryan, Brenda M Vincent, Donald A Jurivich, Kyle J Hackney, Grant R Tomkinson, Lindsey J Dahl, and Brian C Clark. 2020. "Handgrip Strength Asymmetry and Weakness Together Are Associated With Functional Disability in Aging Americans." *Journals of Gerontology: Series A* 76 (2): edited by Anne Newman, 291–96. [88]

Mendes de Leon, C. F. 2003. "Social Engagement and Disability in a Community Population of Older Adults: The New Haven EPESE." *American Journal of Epidemiology* 157 (7): 633–42. [89]

Noguchi, Taiji, Katsunori Kondo, Masashige Saito, Hiroko Nakagawa-Senda, and Sadao Suzuki. 2019. "Community social capital and the onset of functional disability among older adults in Japan: a multilevel longitudinal study using Japan Gerontological Evaluation Study (JAGES) data." *BMJ Open* 9 (10): e029279. [105]

Rantanen, Taina, Jack M. Guralnik, Dan Foley, Kamal Masaki, Suzanne Leveille, J. David Curb, and Lon White. 1999. "Midlife Hand Grip Strength as a Predictor of Old Age Disability." *JAMA* 281 (6): 558–60. eprint: https://jamanetwork.com/journals/jama/articlepdf/188748/jbr80447.pdf. [93]

Satariano, William A., Melissa Kealey, Alan Hubbard, Elaine Kurtovich, Susan L. Ivey, Constance M. Bayles, Rebecca H. Hunter, and Thomas R. Prohaska. 2014. "Mobility Disability in Older Adults: At the Intersection of People and Places." *Gerontologist* 56 (3): 525–34. [89]

T.B. Ustun, N. Kostanjesek, S. Chatterji, J. Rehm. 2010. *Measuring Health and Disability:Manual for WHO Disability Assessment Schedule (WHODAS 2.0)*. World Health Organization. [90]

Taekema, D. G., J. Gussekloo, A. B. Maier, R. G. J. Westendorp, and A. J. M. de Craen. 2010. "Handgrip strength as a predictor of functional, psychological and social health. A prospective population-based study among the oldest old." *Age and Ageing* 39 (3): 331–37. [88]

Verbrugge, Lois M., and Alan M. Jette. 1994. "The disablement process." *Social Science & Medicine* 38 (1): 1–14. [88]

Vermeulen, Joan, Jacques CL Neyens, Erik van Rossum, Marieke D Spreeuwenberg, and Luc P de Witte. 2011. "Predicting ADL disability in community-dwelling elderly people using physical frailty indicators: a systematic review." *BMC Geriatrics* 11 (1): [88]

World Health Organization. 2015. *World report on ageing and health*. World Health Organization, 246 p. [87, 88]

World Health Organization. 2022. "Study on global AGEing and adult health (SAGE)." [90, 111]

Zahodne, Laura B., Jennifer J. Manly, Anna MacKay-Brandt, and Yaakov Stern. 2013. "Cognitive Declines Precede and Predict Functional Declines in Aging and Alzheimers Disease." *PLoS ONE* 8 (9): edited by Kenji Hashimoto. [89]

Zhong, Yaqin, Jian Wang, and Stephen Nicholas. 2017. "Gender, childhood and adult socioeconomic inequalities in functional disability among Chinese older adults." *International Journal for Equity in Health* 16 (1): [89]

Chapter 3

Tranquilo

Joint with Janoś Gabler, Tim Mensinger, and Sebastian Gsell

3.1 Introduction

Economists frequently encounter "hard" optimization problems when fitting structural models to empirical data. By "hard" we mean that solving the optimization problem requires a significant amount of computation time, often hours or days; that manual intervention like tuning start values or adjusting algorithm parameters is required to obtain a solution; and that solving the optimization problems takes up a significant portion of the researcher's time. A prime example where such problems arise is the estimation of discrete choice models via the method of simulated moments (MSM).

Despite the prevalence of MSM estimation in structural papers (see Eisenhauer, Heckman, and Mosso (2015) for a review) and widely available anecdotal evidence that structural researchers would love to spend less time on solving optimization problems, there are no specialized optimization algorithms that are tailored to the characteristics of MSM estimation problems.

The goal of our paper is to close this gap by proposing the *tranquilo* (**T**rust**R**egion **A**daptive **N**oise robust **QU**adrat**I**c or **L**inear approximation **O**ptimizer) algorithm – an optimizer that helps researchers solve hard optimization problems, as they arise during MSM estimation, faster and with less need for manual intervention. *tranquilo* is designed to take three main characteristics of MSM estimation problems into account:

First, MSM estimation problems are *nonlinear least-squares* problems. Least-squares optimization problems lie within the general class of *blackbox* optimization problems

$$\min_{l \leq x \leq u} f(x) \tag{DET}$$

where $f : \mathbb{R}^p \to \mathbb{R}$. In the least-squares case, it is assumed that the objective function f has the structure $f(x) = \sum_{i=1}^{k} r_i(x)^2$ and thus the optimization problem can be written as

$$\min_{l \leq x \leq u} \sum_{i=1}^{k} r_i(x)^2 \;=\; \min_{l \leq x \leq u} \|r(x)\|^2 \tag{DET-LS}$$

where $r(x) \equiv [r_1(x), \ldots, r_k(x)]^T : \mathbb{R}^p \to \mathbb{R}^k$ and $\|\cdot\|$ is the 2-norm of a vector. $r_i(x)$ is called a least-squares residual and $r(x)$ is called a residual vector.

It is easy to see that MSM problems are nonlinear least-squares problems. The objective function of an MSM problem is given by

$$f(x) \;=\; (m(x) - \hat{m})^T W (m(x) - \hat{m})$$

where x are the parameters to be estimated, $m(x)$ is a vector of simulated moments from the economic model, and $\hat{m}$ is a vector of empirical moments. W is a positive definite weighting matrix.

By defining $L = chol(W)$ as the lower triangular Cholesky factor of W, we can rewrite the objective function as

$$f(x) \;=\; (m(x) - \hat{m})^T L L^T (m(x) - \hat{m})$$

By defining $r(x) = L^T(m(x) - \hat{m})$, we can rewrite the objective function as

$$f(x) \;=\; r(x)^T r(x) \;=\; \sum_{i=1}^{k} r_i(x)^2 \;=\; \|r(x)\|^2$$

Which shows the least-squares structure of the MSM objective function.

There is a class of optimization algorithms that exploit the least-squares structure of the objective function, and it is a robust result that they outperform general-purpose algorithms when applicable (Levenberg, 1944; Marquardt, 1963; Wild, 2017; Cartis, Fiala, Marteau, and Roberts, 2019). A review of existing algorithms can be found in Section 3.2. *Tranquilo* builds on the class of derivative-free least-squares optimizers and extends them to meet the requirements of an efficient optimizer for MSM problems.

Second, MSM estimation problems as they arise in economics have an expensive objective function that is hard to parallelize.

In structural economic models, each evaluation of the MSM objective function first requires solving the model and then simulating data based on the solution. Solving a model can take anywhere from a few seconds to a few hours but is never in the range of milliseconds. Simulating data from a model is usually much faster but still adds some computation time.

An optimization algorithm that is designed for MSM estimation problems can thus assume that the evaluation of the objective function is a runtime bottleneck. Whenever there is a trade-off between reducing calculations done by the optimizer (e.g., doing linear algebra to calculate candidate points) vs. saving a few function evaluations, the optimizer should always prioritize saving function evaluations. This is in stark contrast to traditional objectives of optimization algorithms where often a significant amount of effort is spent on optimizing calculations done by the optimizer so that they can solve benchmarks consisting of very fast objective functions as fast as possible.

Nowadays, most researchers have access to parallel hardware: 8 to 16 cores in a laptop; 16 to 64 cores in desktop computers and small servers, up to hundreds of cores in clusters. However, in most codebases we are aware of, the objective function is not parallelized. This may partially be due to the fact that economists are not trained in parallel programming, but there are also inherent difficulties in parallelizing the solution of economic models. In any case, parallelizing the objective function would require a large time investment of the researcher.

This implies that an optimization algorithm should parallelize the evaluation of the objective function and thus shift the burden of parallel programming from researchers to algorithm developers. Algorithms that parallelize on the level of function evaluations exist (Lee and Wiswall, 2007), but to the best of our knowledge, none of them exploits the least-squares structure of the objective function.

A primary goal for *tranquilo* was to develop an algorithm that evaluates the objective function in batches and, instead of trying to minimize the number of function evaluations, tries to minimize the number of batches. The batch size corresponds to the number of available cores in the system. This design choice stems from the understanding that researchers typically do not benefit from idle cores on their computers. Instead, their priority is often to minimize the time it takes to solve the optimization problem.

Third, MSM estimation problems have a noisy objective function. Noise in the objective function means that we only observe noisy evaluations of the true objective function, but we are interested in finding the minimum or minimizer of the true objective function.

Thus, the problem to be solved is not given by equations DET or DET-LS but by their stochastic counterparts

$$\min_{l \leq x \leq u} \mathbb{E} f(x, \xi) \tag{STOCH}$$

$$\min_{l \leq x \leq u} \mathbb{E} \|r(x, \xi)\|^2 \;=\; \min_{l \leq x \leq u} \mathbb{E} \sum_{i=1}^{k} r_i(x, \xi_i)^2 \tag{STOCH-LS}$$

In the method of simulated moments, the noise in the objective function comes from the fact that we are simulating data of a stochastic model. As researchers, we can influence the amount of noise in the objective function by increasing the number of simulation draws. However, this comes at the cost of increased computation time, and reducing the amount of noise to a level that is acceptable for standard optimizers is usually prohibitively expensive. This is especially the case in dynamic discrete models where initially small random influences propagate over time and can lead to large differences in the simulated data of final periods.

A common approach to deal with noisy objective functions inolves fixing the seed of the random number generator and using the same random draws in each iteration of the optimizer. In general, this approach is not suitable for solving STOCH or STOCH-LS. The reason is that the optimizer will be influenced by lucky draws under the chosen seed and has no chance to optimize the true objective function, i.e., the expected value of the observable function.

Moreover, in dynamic discrete models, this approach of fixing the seed also fails to produce a well-behaved objective function. While it renders the objective function deterministic, it can introduce discontinuities and local optima even if the underlying true objective function is smooth.

Any optimizer that aims to solve STOCH or STOCH-LS has to evaluate the objective function more often than an equivalent optimizer for deterministic problems so the effect of noise can be averaged out. Typically, such optimizers ask a user to specify the number of function evaluations, either as a fixed sequence or a function that depends on the iteration counter and other internal variables of the optimizer. This does not only require knowledge about the inner workings of the optimizer but is very hard to do in practice, as the ideal number of function evaluations depends on problem properties that are not known ex-ante. We illustrate this in Section 3.5.1. Usually, the ideal sequence is increasing in the iteration counter. Choosing too many evaluations slows down the progress. Choosing too few can lead to a catastrophic failure of the optimizer.

A primary goal of our optimizer is, therefore, to determine the optimal number of function evaluations in an adaptive fashion without requiring any user-provided information on the amount or type of noise in the objective function.

While *tranquilo* is tailored to MSM estimation problems as they arise in economics, it is not limited to these. Problems with the same characteristics are also encountered in other fields. Prime examples are design optimization in engineering or calibrating epidemiological models to empirical data. In fact, one of the main motivations for developing *tranquilo* comes from an epidemiological model (Gabler, Raabe, Röhrl, and Gaudecker, 2022).

To summarize the contributions on a technical level, some familiarity with derivative-free trust-region optimizers is required. We describe the basic intuition of derivative-free trust-region optimization in Section 3.2 and refer the reader to Conn, Gould, and Toint (2000) for a more detailed introduction. The technical contributions are as follows.

First, We take a fairly standard trust-region framework for nonlinear least-squares optimizers (see for example, Conn, Gould, and Toint (2000)) and reformulate it in a modular fashion that allows us to replace individual components of the algorithm in order to customize it to the characteristics of MSM estimation problems. Besides the obvious benefit of easing the implementation, this modularization generates important insights. For example, we show that the fundamental difference between a scalar and least-squares version of *tranquilo* is concentrated in one single step (see Section 15), which is not obvious when looking at other codebases that implement scalar and least-squares versions of an algorithm (e.g., Cartis, Fiala, et al. (2019) implement the scalar *PY-BOBYQA* and least-squares *DFO-LS* in two separate codebases even though they highlight the similarity of both algorithms in their paper).

Second, we add parallelization capabilities to the trust-region framework. Some parts of derivative-free trust-region algorithms, such as the evaluation of the objective function on an initial set of points, are embarrassingly parallel and have been parallelized in other algorithms (e.g., Cartis, Fiala, et al. (2019)). We add two new ideas for more efficient parallelization: The first is a parallel line search that tries out multiple step lengths in the search direction obtained by solving the trust-region subproblem. The second is speculative sampling: While doing the function evaluation(s) needed to decide whether a candidate point is accepted, we already sample points that would be helpful in the next iteration if the candidate point is accepted, and evaluate

the objective function on those points. Both strategies have diminishing returns if many cores are available. Therefore, we find that a combination of both approaches works best.

Third, we propose novel ways of adaptively determining how many function evaluations are needed to average out the noise just enough so that the optimizer can make progress. We distinguish two different situations within each iteration.

In the *model building phase*, we need to determine how often the objective function should be evaluated on each model point. The goal here is to build a model that is as cheap as possible but good enough to send us in the right direction. We treat the error that derives from noise in a similar form as the error that derives from approximating a general nonlinear function over the trust-region with a low-order polynomial. To this end, we introduce a new measure of model quality ρ^{noise} that measures how strongly random error impedes the surrogate model's ability to produce good candidate points. This measure is then used to adjust the number of repeated function evaluations at each model point. In this sense, it is similar to the traditional measure of model quality ρ that is used to adjust the trust-region radius. The calculation of ρ^{noise} is based on a simulation approach that is computationally costly compared to an iteration of a normal trust-region algorithm but small compared to a single evaluation of the objective function in typical applications.

In the *acceptance phase*, we need to determine how many function evaluations are needed to decide whether the candidate point is actually an improvement over the currently accepted point. We use power analysis to determine the minimal number of additional function evaluations on both the candidate point and the currently accepted point that are needed to achieve a certain power in deciding which point is better. The approach takes the existing number of evaluations on both points as well as the expected improvement – a side-product of the trust-region step – into account.

Both approaches require an estimate of the variance of the noise in the objective function. We estimate this variance from existing function evaluations on points in a neighborhood of the current trust-region. By doing multiple function evaluations on the start parameters, we can guarantee that a sufficient number of function evaluations is available in all iterations and no extra function evaluations are needed for the noise estimation. This approach treats the noise variance as locally constant over the trust-region but otherwise accommodates both additive and multiplicative noise as well as mixtures thereof and does not require the user to specify which type of noise is present.

Fourth, we make *tranquilo* (Gabler, Gsell, Mensinger, and Petrosyan, 2024) available as an open-source Python package that can be used in isolation or via the estimagic package (Gabler, 2022).

Tranquilo builds heavily on previous algorithms and the literature on derivative-free optimization (Conn, Gould, and Toint, 2000; Powell, 2009; Wild, 2017; Cartis, Fiala, et al., 2019). While this literature uses terminology that is not commonly familiar to economists, it is surprising that large parts of *tranquilo* can be understood in terms of concepts that are familiar to economists: Power analysis is routinely used to determine sample sizes in empirical work; in *tranquilo*, it is used to determine the number of function evaluations for accepting a candidate point. Using model-based simulations to determine optimal policies is the core business of struc-

tural economists; in *tranquilo* we use them to determine an optimal policy for the number of function evaluations used in the model building phase. Finally, ordinary least-squares regression is the workhorse method for every empirical economists; in *tranquilo* we use it to fit linear or quadratic approximations to a general nonlinear function.

We benchmark *tranquilo* against existing solvers on the Moré-Wild benchmark set Moré and Wild (2009), which is the standard benchmark set for derivative-free least-squares solvers. To assess the performance of *tranquilo* on noisy problems, we add artificial noise to the objective functions of the benchmark set.

In a baseline setting without noise and parallelism, the least-squares and scalar versions of *tranquilo* are competitive with comparable existing solvers. The least-squares version is slightly slower than the best existing least-squares solver *DFO-LS*, but faster than *POUNDERS*. The scalar version is slightly slower than the NlOpt implementation of *BOBYQA* but beats the scipy and NlOpt implementations of Nelder-Mead as well as the NAG implementation of *BOBYQA*. While this is not the primary use case for *tranquilo*, it is reassuring that *tranquilo* is competitive with existing solvers in the baseline setting. The full results and details of the benchmarking procedure for this setting can be found in Section 3.3.3.

To assess parallel performance, we compare *tranquilo* versions with 1, 2, 4, and 8 cores against each other. We also include *DFO-LS* as a reference. Importantly, this time, the goal is not to minimize the number of function evaluations but the number of batches, where each batch is a set of function evaluations that can be run in parallel. As before, we find that *DFO-LS* is faster than the serial version of *tranquilo*, but with two cores, *tranquilo* is already considerably faster than *DFO-LS*. Adding more cores keeps improving the performance of *tranquilo*, and the 8-core version is the fastest solver for more than 80% of the problems in the benchmark set. The full results and details of the benchmarking procedure for this setting can be found in Section 3.4.2.

In a noisy setting, we compare *tranquilo* against *DFO-LS* – the only other derivative-free least-squares solver that is designed to handle noisy objective functions. Since *DFO-LS* requires the user to specify the number of function evaluations at each parameter vector, we compare *tranquilo* against multiple variants of *DFO-LS*. We restrict our attention to a fixed number of function evaluations because correctly guessing sequences that vary in each iteration is very hard to do in practice. Compared to the noisy benchmarks in Cartis, Fiala, et al. (2019), we use a much larger amount of noise.

We find that *tranquilo* outperforms all configurations of *DFO-LS* in the noisy setting. While *DFO-LS* configurations with few function evaluations per parameter vector solve some problems very quickly, they fail to solve others. Configurations with many function evaluations per parameter vector solve more problems but are very slow. Due to its adaptive nature, *tranquilo* is able to solve problems quickly while still being robust to solve many problems. The full results and details of the benchmarking procedure for this setting can be found in Section 3.5.4.

The remainder of the paper is structured as follows: Section 3.2 reviews core concepts and terminology of derivative-free optimization and discusses how existing algorithms relate to *tranquilo*. Section 3.3 describes the modular formulation of our general trust-region framework and discusses the implementation of each component for the baseline case without noise and parallelization. It also shows the results of benchmarking *tranquilo* against existing solvers in this

setting. Section 3.4 explains our two ideas for improving the parallelization of derivative-free trust-region optimizers and shows the speed-up we achieve via parallelization. Section 3.5 describes our approaches for noise handling as well as the corresponding benchmarks. Section 3.6 concludes.

3.2 Literature review

The literature review is split into two parts. The first reviews important concepts of derivative-free optimization and is dedicated to readers with little or no background in optimization. We introduce all essential concepts needed to understand the rest of the paper, as well as the technical description of contributions in the introduction. The second part reviews related algorithms and identifies gaps in the literature that are filled by *tranquilo*.

3.2.1 Concepts of derivative-free optimization

Local and global optimization. In economics and statistics, it is often the goal to find a global minimum of a scalar objective function defined on $\mathbb{R}^p$. Without further assumptions, this is an impossible task, as the only way to guarantee that a global optimum was found is to evaluate the objective function at all points in $\mathbb{R}^p$.

There are two ways to solve global optimization problems in practice: Global optimizers or local optimizers in a multistart framework.

Global optimizers require finite bounds for all parameters and sample the parameter space. The simplest algorithms are random search and grid search; other algorithms sample candidate points in more sophisticated ways. Global algorithms often yield relatively imprecise solutions that must be refined with a local optimizer. Moreover, they suffer from the curse of dimensionality, i.e., they become extremely expensive as soon as there are more than a handful of parameters. A big drawback of global optimizers is that they typically do not exploit any known properties of the objective function. For example, we are not aware of global optimizers that exploit the least-squares structure. Without further precautions, global optimizers are also not robust to noise in the objective function. A simple example of this is random search. While random search is a very robust global optimizer for deterministic functions, it breaks down if there is considerable noise in the objective function, as it might select a point that just had a lucky draw.

Multistart frameworks run local optimizers from multiple starting points. While any single optimization run might get stuck in a local minimum, the hope is that the best local minimum is also the global minimum. As with global optimizers, multistart frameworks come without guarantees that the global minimum was found. Their biggest advantage is that they work with any local optimizers and, thus, can exploit known properties of the objective function, such as the least-squares structure. As long as the local optimizer is robust to noise, multistart frameworks are also robust to noise.

Given these trade-offs, we decided to develop a local optimizer. If a global optimum is required, we recommend to run *tranquilo* in an efficient multistart framework such as tiktak (Arnoud, Guvenen, and Kleineberg, 2019).

Derivative free optimization. Local optimizers move iteratively through the parameter space of an optimization problem to find a parameter vector that minimizes the objective function. Thus, in each iteration, the algorithm needs to decide on a search direction and a step size. In derivative-based optimization, the search direction is usually based on the gradient of the objective function, and the step size is chosen based on its Hessian (see Nocedal and Wright (2006) for examples).

While this approach is very successful, it requires a means to evaluate the gradient and, potentially, the Hessian of the objective function. Whenever one has access to the objective function itself, a way to get at its derivatives is to use finite differences. However, this approach is very expensive. If there are p parameters, calculating a gradient via finite differences takes at least p additional evaluations of the objective function, and second derivatives are even more expensive.

Gradient-free optimizers do not make direct use of the derivatives of the objective function. By not using derivatives, their goal is to be faster than a gradient-based optimizer employing finite differences. There are different classes of gradient-free optimizers. Each of them uses a different approach to finding a search direction and a step size without using the derivatives of the objective function. We restrict our attention to the class of derivative-free trust region optimizers. For an overview of other approaches, see Larson, Menickelly, and Wild (2019).

Importantly, many derivative-free optimizers assume that the derivatives of the objective function exist. They simply do not use them because they are too expensive to evaluate. The existence of derivatives is needed for convergence proofs. In practice, some derivative-free optimizers work even if the derivatives do not exist.

The basic idea of trust-region optimization. An important class of derivative-free optimizers are trust-region methods (Conn, Gould, and Toint, 2000; Nocedal and Wright, 2006). One iteration of a prototypical trust-region algorithm looks as follows

1. Given a current parameter vector x_t as trust-region center and a radius Δ_t, form a surrogate model M_t that approximates the objective function inside the trust-region. The surrogate model is usually a quadratic model or some other low-order polynomial.
2. Find the minimizer of the surrogate model using a specialized optimizer that is tailored to the functional form of the surrogate model. This minimizer becomes a candidate step.
3. Evaluate the objective function at the candidate point and accept or reject the candidate point.
4. Adjust the trust-region radius for the next iteration based on a measure of progress.

The basic idea of a trust-region optimizer is to iteratively replace an expensive objective function that is hard to optimize with a local surrogate model that can be optimized very cheaply. The acceptance decision and trust-region management play an important role in ensuring that the model approximates the function well enough.

Surrogate models. Different trust-region optimizers form the surrogate models in different ways. Derivative-based methods use the gradient and Hessian of the objective function at x_t to form a second-order Taylor expansion that serves as the surrogate model. To save costly evaluations of the Hessian, some optimizers use approximations to the Hessian. A robust result in the literature is that (underdetermined) quadratic surrogate models work best (Conn, Gould, and Toint, 2000). Linear surrogate models have no internal minimum and can thus only suggest candidate points on the boundary of the trust-region, which makes them unsuitable for choosing good step lengths. Higher-order polynomials are not just too expensive to form but also too hard to optimize.

Derivative-free optimizers form surrogate models by evaluating the objective function on a sample of points and forming a quadratic model by interpolation or regression. The points are chosen carefully based on geometric considerations to maximize the model's approximation accuracy. To save function evaluations, only a few points in the sample are replaced in each iteration. Fully determined quadratic interpolation models require the function to be evaluated at $\frac{(p+1)(p+2)}{2}$ points. Since this number grows quickly in p, many algorithms use underdetermined interpolation models based on $2p+1$ function evaluations. The remaining degrees of freedom are resolved by choosing a solution to the interpolation conditions that minimize the Frobenius norm of the model Hessian or the Frobenius norm of the change in the model Hessian between iterations. This idea was first popularized by Powell in the *NEWUOA* and *BOBYQA* algorithms (Powell, 2006; Powell, 2009) and has since been used by many others (see Larson, Menickelly, and Wild (2019) for a review).

A special case are derivative-free trust-region methods for least-squares problems. Instead of forming just one surrogate model for the function value, they form a surrogate model for each least-squares residual. The surrogate models for the residuals are then aggregated into a surrogate model for the actual function value. While there are no proofs that this approach works better than forming scalar surrogate models directly, a vast amount of benchmarks shows that as few as $p+1$ function evaluations can be enough to create useful surrogate models using this principle (see for example Cartis, Fiala, et al. (2019) and Cartis and Roberts (2019)).

Trustregion radius management. The surrogate models in trust-region optimizers only approximate the objective function locally. Using simple models like quadratic ones can, therefore, be justified by Taylor's theorem. This shows that the trust-region radius plays a central role in governing the approximation quality. If the radius is large, the optimizer can make large steps, but the model might be a poor approximation to the objective function. Making the radius smaller increases the model accuracy at the cost of slower progress.

It is very important that the model only has to be good enough to make progress, and it is not an explicit goal to minimize the overall approximation error on the trust-region. The radius adjustment is, therefore, based on a measure of model quality that specifically takes into account how well the model predicts good candidate points

$$\rho_t \equiv \frac{f(x_t^*) - f(x_t^* + s_t)}{M_t^s(x_t^*) - M_t^s(x_t^* + s_t)} = \frac{\text{Actual Improvement}}{\text{Expected Improvement}} \tag{3.2.1}$$

The basic idea is then as follows: If ρ is large, the model worked well in predicting a descent direction, and the radius can be increased or kept constant. If ρ is small, the radius has to be reduced in order to improve the model quality in the next iteration. If suitable surrogate models are used and regularity conditions are fulfilled, Taylor-like error bounds guarantee that a good approximation quality can be achieved by making the radius small enough. The actual radius adjustment is slightly more complex and depends on additional quantities and conditions. Several methods are discussed in Conn, Gould, and Toint (2000).

Convergence. The word *convergence* is used for two very different things: In the theoretical literature, a convergence proof means that a mathematical algorithm is guaranteed to find a local optimum or stationary point if run long enough. Among practitioners, convergence means that an algorithm stopped the optimization process because a condition was achieved. Since those conditions can usually be set by a user, reaching them is not a strong guarantee that an optimum has been found, and practitioners should always verify that convergence was not spuriously induced by weak convergence criteria.

Tranquilo is loosely based on a trust-region framework for which a convergence proof exists (Conn, Gould, and Toint, 2000), and the components that play a central role in the convergence proof (e.g., solvers for the surrogate problem and trust-region radius handling) are fairly standard. However, *tranquilo* is meant as an algorithm for practitioners, and we do not make an attempt at extending the convergence proof to cover the modifications we propose in *tranquilo*. Instead, we rely on extensive benchmarks to show the practical performance of *tranquilo*.

3.2.2 Related algorithms

We restrict our attention to derivative-free trust-region methods for bound-constrained optimization. A more comprehensive overview discussing other methods can be found in Larson, Menickelly, and Wild (2019).

While derivative-free trust-region optimizers based on quadratic models have been used since the early 1970s (Winfield, 1973), the interest in these methods has been revitalized by the influential work of Powell. An important contribution of Powell was the introduction of underdetermined interpolation for the construction of quadratic surrogate models –first introduced in the *NEWUOA* and *BOBYQA* algorithms– which drastically improve the efficiency for higher dimensional problems (Powell, 2006; Powell, 2009). As an algorithm that supports bound constraints, *BOBYQA* can be seen as the direct predecessor of most algorithms that we discuss in this section.

The *BOBYQA* algorithm (Powell, 2009) maintains a sample of $2p + 1$ points that are used to form a quadratic surrogate model. The model is fit using underdetermined interpolation. The remaining degrees of freedom are resolved by choosing the solution to the interpolation conditions that minimize the Frobenius norm of the change in the model Hessian between two iterations. Between two iterations, at most one model point is replaced. The replacement point is chosen to maximize the stability of the model. Several variants of the *BOBYQA* algorithm are available as open-source software and compare very favorably against other derivative-free optimizers like the Nelder-Mead algorithm (see Section 3.3.3).

The *DFBOLS* (Zhang, Conn, and Scheinberg, 2010) and *POUNDERS* algorithm (Wild, 2017) can be seen as a translation of *BOBYQA* to least-squares problems. Both algorithms use $2p + 1$ interpolation points and the same underdetermined interpolation method as *BOBYQA*. The main difference is that they construct one quadratic surrogate model for each least-squares residual and aggregate those models into a quadratic model for the objective function. The aggregation method differs between the optimizers: *POUNDERS*' aggregation method can be described as a Full-Newton approach whereas *DFBOLS* incorporates elements from a Gauss-Newton approach. The *POUNDERS* algorithm is available as a pure Python implementation in the estimagic library. A C implementation of *POUNDERS* is available in the toolkit for advanced optimization (TAO) (Dener, Denchfield, Suh, Munson, Sarich, et al., 2021). *DFBOLS* is available as Fortran code. The performance of *DFBOLS* and *POUNDERS* is expected to be very similar (Wild, 2017). Due to the lack of a *DFBOLS* implementation with Python bindings, we only compare *tranquilo* to *POUNDERS*.

DFO-LS (Cartis, Fiala, et al., 2019) is another derivative-free trust-region method for least-squares problems. The key difference is that *DFO-LS* uses only $p + 1$ interpolation points and fits fully determined linear surrogate models for each residual. Those linear models are then aggregated into a quadratic model for the objective function. This change drastically improves *DFO-LS*'s performance for larger problems. We use the same approach in *tranquilo*. On top of this change, *DFO-LS* introduces several new features: First, a fast start option tries to make progress before there are enough function evaluations to fit an initial model. Second, there is a heuristic that detects whether the trust-region radius collapsed due to the presence of noise and if so, the algorithm is automatically restarted. Third, the user can specify sequences that control how often a noisy objective function should be evaluated. The sequence can depend on several quantities, among them the iteration counter and a restart counter. The same new features are also available in *Py-BOBYQA*, which is developed in the same paper and works for scalar objective functions. The performance of *DFO-LS* is excellent (see Section 3.3.3) and we use it as the main benchmark for *tranquilo*. *Py-BOBYQA* is also included in the benchmarks but performs slightly worse than other *BOBYQA* implementations. Both algorithms are available as standalone Python packages.

The main problem of derivative-free trust-region optimizers applied to noisy objective functions is that the trust-region radius collapses to zero. This is caused by bad candidate points from noise-affected surrogate models and spurious rejections due to unlucky draws in the acceptance evaluation. While *DFO-LS* is the only least-squares optimizers for noisy objective functions that we are aware of, there are several optimizers for scalar objective functions that employ strategies to avoid the collapsing of the radius.

SNOWPACK (Augustin and Marzouk, 2017) ties the trust-region radius management to an estimate of the noise in function evaluations. Moreover, it uses Gaussian process models instead of quadratic interpolation models to reduce the effect of noise.

Shashaani, Hashemi, and Pasupathy (2018) recognize that user-specified sequences for the number of function evaluations needed to average out noise are impractical and propose the *ASTRO-DF* algorithm that uses adaptive sampling: The number of evaluations is increased until an estimated standard error falls under a threshold. The threshold is a fixed factor of the squared trust-region radius. This incorporates the idea that smaller trust-region radii require more precise models. Moreover, it prevents the radius from shrinking too much before a good model quality

has been achieved. The adaptive sampling in *ASTRO-DF* is, however, not based on the magnitude of the function evaluations. *ASTRO-DF* is available as part of the simopt library, where it can be benchmarked against other simopt optimizers. We currently exclude *ASTRO-DF* from our benchmarks because we could not get it to solve our benchmark problems precisely enough, but we want to exclude all errors that might be caused on our side before drawing any conclusions.

Parallelization on the algorithm level has not been a focus of the literature on derivative-free trust-region optimizers or derivative-free least-squares optimizers. However, there are parallel direct search algorithms for scalar problems.

Lee and Wiswall (2007) introduce a parallel version of the *Nelder-Mead* simplex algorithm. The classical *Nelder-Mead* algorithm maintains a set of $p + 1$ points that are used to form a simplex in parameter space. In each iteration, the worst point is replaced by a new point. There are different strategies for calculating the new point, which are selected based on the function values. The parallel version replaces more than one point in each iteration and evaluates the objective functions on all new points in parallel. Depending on the function value, an initial candidate for a new point might be rejected, and a second function evaluation is necessary before a new point is accepted. The empirical results in Lee and Wiswall (2007) show strong gains in efficiency, which are sometimes substantially larger than the number of processors. They explain this by the fact that the parallel version might sometimes create better search directions than the serial one. An implementation of the parallel *Nelder-Mead* algorithm is available in the estimagic library. We are currently working on incorporating it into our benchmarks.

3.3 Tranquilo core algorithm

In this section, we describe a core version of the *tranquilo* algorithm that is suitable for solving the deterministic nonlinear least-squares problem DET-LS as well as the deterministic scalar problem DET without using parallelization. The extension to the parallel case is described in Section 3.4. The extension to the stochastic case is described in Section 3.5.

The structure is as follows: In Section 3.3.1, we describe our modular formulation of a general trust-region algorithm that formalizes the interface of components in the algorithm. Most components are mathematical functions that have a one-to-one correspondence in the Python implementation of the algorithm. At this stage, we only describe the inputs and outputs of functions and are agnostic about their inner workings. In Section 3.3.2, we change our focus and describe the algorithmic implementation of each component. We focus on the deterministic and serial case, and draw ample comparisons to existing algorithms. In Section 3.3.3, we describe how we benchmark optimizers and show how *tranquilo* compares to other algorithms.

3.3.1 The trust region framework

In this section, we review the general trust-region framework of the *tranquilo* algorithm in a modular fashion with a high level of abstraction. Doing so allows us to describe the concrete implementation of our baseline algorithm as well as its extensions to the parallel and noisy case clearly

and without repeating what stays unchanged. The full algorithm is described in Algorithm 1. A lookup table for our notation can be found in Appendix 3.A.

Tranquilo is flexible because it is made up of *replaceable components*. By replaceable component, we mean a function that takes a specified set of inputs and produces a specified set of outputs. A simple example of a replaceable component is a *Sampler*, which takes existing points, a trust-region, and a target sample size as inputs and produces a set of new points as output. How the new points are created is not specified and varies across different samplers. Importantly, all other parts of *tranquilo* will work with any sampler that conforms to the specified set of inputs and outputs. This has, of course, a clear mapping to the Python implementation of *tranquilo*: For every replaceable component, we implement several different versions that a user of the algorithm can select by providing the name of that version. Advanced users can go beyond what we offer and implement their own versions of components.

A full list of replaceable components and a definition of their interfaces can be found in Table 3.A.4. The implemented versions of each component are described in Sections 3.3.2, 3.4.1, and 3.5. Before looking at these implementations, we first describe how the different components interact to create the *tranquilo* algorithm.

Algorithm 1: *Tranquilo* algorithm

Input: Starting point x_0^*, initial trust-region radius Δ_0^{region}, target sample size n^{target}, search factor γ^{search}, minimum step size s^{min}, sample increment n_{stag}^{drop}, maximum number of iterations t^{max}, maximum number of trials to avoid stagnation n_{stag}^{max}, lower and upper bounds l and u.

Initialize history with $\mathcal{H}_0 = \{(x_0^*, r(x_0^*))\}$
Initialize vector model M_0^v with intercept terms at $r(x_0^*)$ and all other coefficients set to zero
for $t=0,1,\ldots,t^{max}$ **do**
    Calculate the search radius $\Delta_t^{search} = \gamma^{search} \Delta_t^{region}$
    Calculate the effective trust-region R_t based on x_t^*, Δ_t^{region}, l and u
    Scan the history for existing points $\mathcal{X}_t^{existing} = \{x \in \mathcal{H}_t : \|x_t^* - x\| \leq \Delta_t^{search}\}$
    Filter existing points: $\mathcal{X}_t^{filtered} = Filter(\mathcal{X}_t^{existing})$
    if $|\mathcal{X}_t^{filtered}| < n^{target}$ **then**
        Sample $n^{target} - |\mathcal{X}_t^{filtered}|$ new points in the trust-region: $\mathcal{X}_t^{new} = Sample(\mathcal{X}_t^{filtered}, R_t, n^{target})$
        $\mathcal{X}_t^{model} = \mathcal{X}_t^{filtered} \cup \mathcal{X}_t^{new}$
    else
        $\mathcal{X}_t^{model} = \mathcal{X}_t^{filtered}$
    end
    Build a vector model $M_t^v = Fit(\mathcal{X}_t^{model}, \mathcal{R}_t^{model}, M_{t-1}^v, R_t)$
    Aggregate the vector model: $M_t^s = Aggregate(M_t^v)$
    Solve the surrogate problem: $s_t = Subsolve(M_t^s, R_t)$
    while $|\mathcal{X}_t^{model}| > n^{target}$ *and* $\|s_t\| \leq s^{min}$ **do**
        Reduce the sample: $\mathcal{X}_t^{reduced} = Drop(\mathcal{X}_t^{model}, n_{stage}^{drop}, \Delta_t^{region})$ and set $\mathcal{X}_t^{model} = \mathcal{X}_t^{reduced}$
        Build a vector model $M_t^v = Fit(\mathcal{X}_t^{model}, \mathcal{R}_t^{model}, M_{t-1}^v, R_t)$
        Aggregate the vector model: $M_t^s = Aggregate(M_t^v)$
        Solve the surrogate problem: $s_t = Subsolve(M_t^s, R_t)$
    end
    $n_{stag} = 0$
    while $\|s_t\| \leq s^{min}$ *and* $n_{stag} \leq n_{stag}^{max}$ **do**
        Reduce the sample: $\mathcal{X}_t^{reduced} = Drop(\mathcal{X}_t^{model}, n_{stag}^{drop}, \Delta_t^{region})$
        Sample new points in the trust-region: $\mathcal{X}_t^{new} = Sample(\mathcal{X}_t^{reduced}, R_t, n^{target})$ and set $\mathcal{X}_t^{model} = \mathcal{X}_t^{reduced} \cup \mathcal{X}_t^{new}$
        Build a vector model $M_t^v = Fit(\mathcal{X}_t^{model}, \mathcal{R}_t^{model}, M_{t-1}^v, R_t)$
        Aggregate the vector model: $M_t^s = Aggregate(M_t^v)$
        Solve the surrogate problem: $s_t = Subsolve(M_t^s, R_t)$
        $n_{stag} = n_{stag} + 1$
    end
    Calculate $\Delta M_t^s = M_t^s(x_t^*) - M_t^s(x_t^* + s_t)$
    Accept or reject the step and calculate a measure of progress $(x_{t+1}^*, \rho_t) = Accept(x_t^*, s_t, \Delta M_t^s)$
    Adjust the trust-region radius: $\Delta_{t+1}^{region} = AdjustRadius(\Delta_t^{region}, \rho_t, s_t)$
    if $x_{t+1}^* \neq x_t^*$ *and* $Converged(\mathcal{H}_t, M_t^s, x_t^*, x_{t+1}^*)$ **then**
        break
    end
end

At the beginning of *tranquilo*, we are equipped with a starting point $x_0^* \in \mathbb{R}^p$, an initial radius $\Delta_0^{region} > 0$, as well as the lower and upper bounds of the optimization problem $l, u \in \mathbb{R}^p$. Together, those quantities define the initial trust-region. Moreover, we have several algorithm constants like the target sample size n^{target}, the search factor γ^{search}, the minimum step size s^{min}, the sample increment n_{stag}^{drop}, the maximum number of iterations t^{max}, and the maximum number of trials to avoid stagnation n_{stag}^{max}. For now, we abstract from constants that are only used by the specific implementation of components.

The algorithm starts by evaluating the objective function at the starting point and initializing the history of function evaluations with $\mathscr{H}_0 = \{(x_0^*, r(x_0^*))\}$ —If it is clear from the context, we sometimes write $x \in \mathscr{H}$ instead of $(x, r(x)) \in \mathscr{H}$. The history of function evaluations is scanned at the beginning of each iteration to find points on which the objective function has previously been evaluated and which are inside or near the current trust-region. Moreover, we initialize a surrogate model for the least-squares residuals M_0^v to equal the constant $r(x_0^*)$ for all points inside the trust-region. We call this a vector model to distinguish it from the aggregated scalar model that approximates the objective function instead of the residuals.

Before the first trust-region iteration, we calculate the effective trust-region R_t which is the subset of the parameter space in which new points can be sampled and to which the solution of the trust-region subproblem will be constrained. If no bounds are binding, the effective trust-region is just the trust-region, i.e., a ball with center x_t^* and radius Δ_t^{region} in Euclidean norm. If bounds are binding, we switch to a hypercube trust-region with the same volume as a ball of radius Δ_t^{region}. The hypercube is also centered at x_t^* and clipped to comply with the bounds of the optimization problem. Note that a hypercube can be viewed as a ball under the maximum-norm. To avoid confusion, we stick to saying ball for spherical regions and hypercube for cubical regions. Other trust-region algorithms that allow for bound constraints (e.g., Wild (2017)) work with a radius in maximum-norm from the beginning. However, we found that switching between the two shapes yields a better performance in benchmarks.

At the beginning of each iteration, we scan the history of function evaluations for points that lie within the search radius of the current trust-region center. These points can be re-used when building the surrogate model. Next, the set of points is filtered. The filtering step is the first replaceable component of the *tranquilo* algorithm. Having a filtering step is a design choice inspired by the following seemingly counter-intuitive observation: A sample size that is *too large* can actually make surrogate models worse (Powell, 2009; Larson, Menickelly, and Wild, 2019). The filtering step provides the option to discard points that are too close to each other or too close to the trust-region center. In our practical experiments, however, we could not confirm this observa-

tion and use the identity function as a filter. Other filters we tried and implemented are described in Section 3.3.2.1.

The scanning and filtering approach differs from other trust-region algorithms that do not maintain a full history and only store a fixed-size set of model points (Powell, 2009; Wild, 2017; Cartis, Fiala, et al., 2019). To add a new point, old ones have to be discarded. We find the scanning and filtering approach appealing because it allows the user to warm-start the algorithm with a database of previous function evaluations and for costly objective functions, the memory overhead of storing the full history is not a concern.

If the number of filtered points is smaller than the target sample size n^{target}, we sample new points in the current trust-region until we reach the target sample size. The sampling step is another replaceable component of the *tranquilo* algorithm. The sampling can be based on the geometry of the existing points, and all sampled points must lie inside the effective trust-region. We discuss the sampling strategies we implemented in Section 3.3.2.2.

Scanning, filtering, and sampling leave us with a set of model points $\mathcal{X}_t^{model}$. After evaluating the objective function on the newly sampled points, we can also construct a corresponding set of least-squares residuals $\mathcal{R}_t^{model}$. These can be used to fit a vector model M_t^v. Fitting is another replaceable component of the *tranquilo* algorithm that allows us to nest the fitting strategies of different algorithms in a simple way. In addition to $\mathcal{X}_t^{model}$ and $\mathcal{R}_t^{model}$, the fitting method needs two more ingredients: First, the previous vector model M_{t-1}^v, which can, for example, be used to penalize changes in the model Hessian and second, the effective trust-region, which is used to scale the model to a unit-ball or unit-hypercube –depending on the shape of the trust-region– for numerical stability.

Fitting methods can differ by the type of vector model they fit (e.g., linear or quadratic), by the way they resolve degrees of freedom in the case of underdetermined interpolation (e.g., penalize Hessian terms or changes in Hessian terms), and by the way they do the actual fitting (e.g., ordinary least squares, least absolute deviation, lasso or ridge regression). To ensure that the model fitting is well-defined, all fitting methods must work for underdetermined, just-determined, and over-determined fitting problems. We discuss the fitting methods we implemented in Section 3.3.2.3.

The next step is to aggregate the vector model M_t^v (the surrogate model that approximates the residual function $r(x)$) into a scalar model M_t^s (the surrogate model that approximates the objective function $f(x)$). The minimizer of the scalar surrogate model will become the next candidate point. The aggregation step is another replaceable component of the *tranquilo* algorithm, but it is important that the fitting step, which decides whether linear or quadratic residual models

are built, and the aggregation step are compatible and produce a well-defined quadratic scalar model.

By choosing appropriate pairs of fitting and aggregation methods, we can nest the fitting and aggregation strategies of different algorithms; such as fitting linear residual models and aggregating them into a quadratic scalar model (Cartis, Fiala, et al., 2019) or fitting quadratic residual models and aggregating them into a scalar model (Zhang, Conn, and Scheinberg, 2010; Wild, 2017).

By treating scalar optimization problems as outputting a vector of size one and using an identity function as the aggregation method, we can even nest the fitting approach of scalar algorithms like *BOBYQA* (Powell, 2009). Even though our primary focus is developing a least-squares optimizer, we show that the resulting algorithm is competitive with other derivative-free scalar optimizers (see Section 3.3.3). Another possible extension would be a dedicated optimizer for likelihood functions that leverages the information matrix equality to construct a quadratic scalar model from linear surrogate models. This would be a derivative-free analog of the popular *BHHH* algorithm (Berndt, Hall, Hall, and Hausman, 1974).

Using the scalar model M_t^s, we solve the trust-region subproblem to obtain a candidate step length s_t. The subsolver is again a replaceable component of the *tranquilo* algorithm. After extensive experimentation, we found that two common methods work best: If bounds are binding, we use the *BNTR* algorithm, otherwise we use the *GGTPAR* algorithm. Both solvers are also available in the *POUNDERS* algorithm (Wild, 2017). However, there, the user has to decide before the optimization which one should be used, whereas we switch dynamically between the two. Both algorithms solve the quadratic problem (almost) exactly, which is a good choice for our setting with expensive objective functions. The details of the *BNTR* and *GGTPAR* algorithms are described in Section 3.3.2.5.

If the candidate step s_t is *large enough*, we directly move to the acceptance step. Here, whether a step is large enough is determined based on a cutoff that is relative to the trust-region radius Δ_t^{region}. If the candidate step is too small, we take extra measures to avoid stagnation. If the sample size of the model points is larger than the target sample size n^{target}, we drop n_{stag}^{drop} points, re-fit a vector model, aggregate it, and solve the new trust-region subproblem to get another step size. This process is repeated until the step size becomes large enough or the sample size equals n_{target}. Which points are dropped is again determined by a replaceable component described in Section 3.3.2.6.

If this is not enough to produce a large enough step, we keep dropping n_{stag}^{drop} points, but this time, we replace them with new points. This process is repeated up to n_{stag}^{max} times.

Two things are important to note here: First, solving the trust-region subproblem many times adds some overhead, but this cost is negligible compared to the cost of a single evaluation of the objective function. Second, it seems counterintuitive to drop points instead of simply adding new ones. However, we found that this approach works better in practice. If only new points are added, they have a rather small impact on the model and can, therefore, not avoid stagnation. In our experiments, we found that dropping one point at a time, i.e., $n_{stag}^{drop} = 1$, works best in a serial algorithm.

Once a sufficiently large candidate step s_t has been found or the maximum number of trials for avoiding stagnation has been reached, we move on to the acceptance step. The acceptance step is a replaceable component of *tranquilo* that contributes strongly to the flexibility of our framework.

On an abstract level, the acceptance step looks as follows

$$(x_{t+1}^*, \rho_t) \;=\; Accept(x_t^*, s_t, \Delta M_t^s)$$

where x_{t+1}^* is the candidate point for the next iteration, ρ_t is a measure of progress or model quality, and ΔM_t^s is the expected improvement from taking step s_t. x_{t+1}^* can be equal to $x_t^* + s_t$, x_t^* or an entirely different point. ρ_t can either be calculated as in Equation 3.2.1 or in a different way. Any evaluation of the objective function that happens in the acceptance step will be added to the history and can be used in the next iteration.

This approach nests the classical case where the acceptance step consists of evaluating the objective function at $x_t^* + s_t$ and accepting the step if the improvement is large enough, which often just means larger than zero. Then ρ_t is simply calculated as in Equation 3.2.1. The implementation of such a simple acceptance step is described in Section 3.3.2.7. However, our approach can also express entirely different methods. Examples are a parallel line-search and speculative sampling, which we will describe in Section 3.4.1.

Given ρ_t and the step size of s_t, we can adjust the trust-region radius for the next iteration. Again, we make the radius adjustment a replaceable component. For our empirical results, we use the radius adjustment rules of the *POUNDERS* algorithm (Wild, 2017), which is described further in Section 3.3.2.8.

If $x_{t+1}^* \neq x_t^*$, we check for convergence of the algorithm. The convergence check is based on the history of function evaluations $\mathcal{H}_t$ as well as the current scalar model M_t^s. This allows for all common convergence criteria, which are either based on absolute or relative improvements in the objective function, absolute or relative step sizes, or the gradient terms of the scalar surrogate model. The exact implementation of the convergence check is again replaceable and described in Section 3.3.2.9

The flexible nature of the *tranquilo* framework allows for quick experimentation and benchmarking of different components that are commonly used in trust-region algorithms. For example, we can easily compare the performance of different model fitting and aggregation strategies, leaving everything else equal. In the traditional literature, these changes would be considered large enough to warrant a new algorithm name (compare, e.g., *DFBOLS* and *POUNDERS*). This flexibility will be extremely useful when looking at extensions for parallelization and noise handling. The basic algorithm as described in this section, will stay virtually unchanged, and only components will be swapped out. The few changes needed in the algorithm itself are the introduction of a few new quantities (e.g., a batch size for parallelization or a noise estimate for noisy problems) that were omitted in the baseline version for simplicity. Moreover, the parallel version of *tranquilo* (see Section 3.4) and the noisy version of *tranquilo* (see Algorithm 3) nest the baseline algorithm.

3.3.2 Implementation of the components

In this section, we provide a detailed description of each component and their different implementations in *tranquilo*. We begin with a discussion of the components of the noise-free serial optimization problem, deferring the discussion of components of the noisy and parallel optimization problem to Section 3.5.3 and 3.4.1.

3.3.2.1 Filtering

At the beginning of each iteration, we get a set of existing points $\mathcal{X}_t^{existing}$ in the neighborhood of the trust-region center x_t^*. These points and their corresponding function evaluations can be used to construct a surrogate model. They are "free" from a computational budget perspective, in the sense that using them for the surrogate model does not incur any additional evaluations of the objective function. However, there can be reasons why not all of those points should be used. First, some of them might be far away from the current trust-region, which might hinder the model from approximating the objective function well locally. Second, some of them might be very close to each other, which can lead the model to overfit certain areas of the trust-region. Third, there might simply be so many points that any newly added point has only a small impact on the model, and therefore, the next candidate point will be very close to the old one.

Filters can address these issues by discarding some of the existing points. More formally, they take the following form

$$\mathcal{X}_t^{filtered} = Filter(\mathcal{X}_t^{existing})$$

Since existing algorithms do not maintain a full history of function evaluations, filtering has no counterpart in the literature. However, the filtering methods we implement are inspired by the traditional goal of producing a well-poised set of model points, i.e., model points with geometric properties that lead to surrogate models with tight error bounds. This topic is discussed in more detail in Section 3.3.2.2.

We implement the following filters:

Keep all. As the name suggests, this filter does not discard any points. We use this filter in our benchmarks for the noise-free and serial case as it yields the best performance in this setting.

Discard all. This filter discards all existing points. Using this filter makes the optimizer slower but, in some cases, more robust, as it uses a new high-quality sample of points in each iteration and is therefore not prone to stagnation. The slowdown is not as severe as one might expect: While each model is now more costly to build, the model quality is also higher, and fewer iterations are needed.

Drop excess. This filter only drops points if more than n^{filter} points are available. If so, we first discard excessive points that are outside the trust-region. We begin with the point farthest from the trust-region center. If all remaining points are inside the trust-region, we look for the two points that are closest to each other and discard the one that is closer to the trust-region center, unless one of the points is the center itself. We repeat this process until only n^{filter} points remain. The idea behind this filter is that we want to have points as far out as possible as long as they are inside the trust-region. This filter is used in our benchmarks for the parallel case with $n^{filter} = 3n^{target}$.

3.3.2.2 Sampling

Sampling refers to the process of creating new model points $\mathcal{X}^{model}$ at which the objective function is evaluated in order to construct a surrogate model. In the first iteration, a full sample is created from scratch, and the sample size is set to n^{target}. In most other iterations, sampling only complements a set of existing points, and the sample size might be larger than n^{target}. In all cases, the goal is to create a set of model points with geometric properties that lead to tight error bounds of the surrogate model. Which points are optimal depends on the type of surrogate model used (e.g., linear or quadratic).

We restrict our attention to simple polynomial models that are linear in the parameters. Let n_t denote the number of model points in iteration t, and d the number of coefficients of the

model. In this case, the quality of the sample is not directly assessed on the model points $\mathscr{X}^{model} = [x_1, \ldots, x_{n_t}]^T$ but on a design matrix $X \in \mathbb{R}^{n_t \times d}$ that is constructed from the model points given the model class. The design matrix is also known as the matrix of regressors. Note that for all $i = 1, \ldots, n_t$ the model points must be inside the effective trust-region, i.e., $x_i \in R_t \subset \mathbb{R}^p$. Abusing notation slightly, we write $\mathscr{X}^{model} \in R_t^{n_t}$.

In the case of a linear model, we have $d = p + 1$ coefficients, and the construction of the design matrix is as follows

$$D^l(\mathscr{X}^{model}) \equiv X^l = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_t,1} & \ldots & x_{n_t,p} \end{pmatrix}$$

In the case of a quadratic model, there are additional columns for the cross products and square terms, so we have $d = (p+1)(p+2)/2$ coefficients. The design matrix is constructed as follows

$$D^q(\mathscr{X}^{model}) \equiv X^q = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,p} & x_{1,1}^2 & x_{1,1}x_{1,2} & \ldots & x_{1,p}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_t,1} & \ldots & x_{n_t,p} & x_{n_t,1}^2 & x_{n_t,1}x_{n_t,2} & \ldots & x_{n_t,p}^2 \end{pmatrix}$$

There are multiple strands of literature that discuss the optimal sampling of model points. The one that is closest to economics is the one on optimal design (see Pukelsheim (2006) for a comprehensive overview). Optimal design asks the question of how to choose a set of points in the space of potential experiments (which, in our case, is the parameter space) that leads to the most informative data, i.e., data that allows us to estimate parameters of interest with the highest precision. Depending on the goal of the experimenter, different statistical measures of precision are maximized. In the case of a regression model, a frequently used measure is D-optimality. The D-optimal sample is the one whose design matrix minimizes the determinant of the inverse Fisher information matrix, i.e.,

$$\mathscr{X}^{d*} = \underset{\mathscr{X} \in R_t^{n_t}}{\arg\min} \det\left(\left[D(\mathscr{X})^T D(\mathscr{X})\right]^{-1}\right)$$

A closely related strand of literature is the one on function approximation. The main difference is that optimal design typically looks at cases where $n_t \geq d$, whereas function approximation looks at the case where $n_t = d$. In this case, the design matrix is square. In the function approximation literature, the sample of choice is called Fekete points. Fekete points are the set of points that

maximize the determinant of the design matrix (see for examples Briani, Sommariva, and Vianello (2012)), i.e.,

$$\mathcal{X}^{f*} = \underset{\mathcal{X} \in R_t^{n_t}}{\arg\max} \det\left(D(\mathcal{X})\right)$$

In the case of a square design matrix, the Fekete points are equivalent to the D-optimal sample. To see this, note that

$$\det\left(\left[D(\mathcal{X})^T D(\mathcal{X})\right]^{-1}\right) = \det\left(D(\mathcal{X})^T D(\mathcal{X})\right)^{-1} = \det\left(D(\mathcal{X})\right)^{-2}$$

While familiar to economists, neither of the previous approaches extends to the case where $n < d$. Therefore, the literature on trust-region optimizers introduces the concept of Λ-poisedness to measure the quality of samples. This concept is based on Lagrange polynomials and works for over-determined, just-determined, and underdetermined interpolation problems. For a definition and comprehensive treatment of Λ-poisedness, see Conn, Gould, and Toint (2000). While the definition of the measure Λ relies on concepts that are not typically familiar to economists, it has a simple interpretation: Λ^{-1} can be interpreted as the distance a set of model points has to linear dependence, i.e., the smaller Λ is, the more linearly independent the model points are. In the case of just-determined and over-determined interpolation problems, the optimal sample according to Λ-poisedness is equivalent to Fekete points or D-optimal points, respectively.

It is very instructive to look at the optimal samples for linear and quadratic models in the case of a spherical trust-region in two dimensions. These samples are shown in Figure 3.3.1.

While one might intuitively think that the optimal sample fills the space uniformly, this is not the case. The optimal sample for a linear model consists of points that are uniformly spaced on the sphere, i.e., the boundary of the ball. The same holds for the optimal samples of a quadratic model, except that here, there is one additional point in the center. This pattern carries over to higher dimensions and larger samples. In the case of a cubical-shaped trust-region, the optimal sample also consists of points on the boundary of the cube (and one point in the center for quadratic models), but the spacing is less regular. When moving to higher-order polynomial models, the optimal sample consists of concentric rings (see Briani, Sommariva, and Vianello (2012)).

Calculating optimal samples directly based on Λ-poisedness, D-optimality or the Fekete criterion is expensive. We, therefore, exploit the pattern described above in several of our samplers.

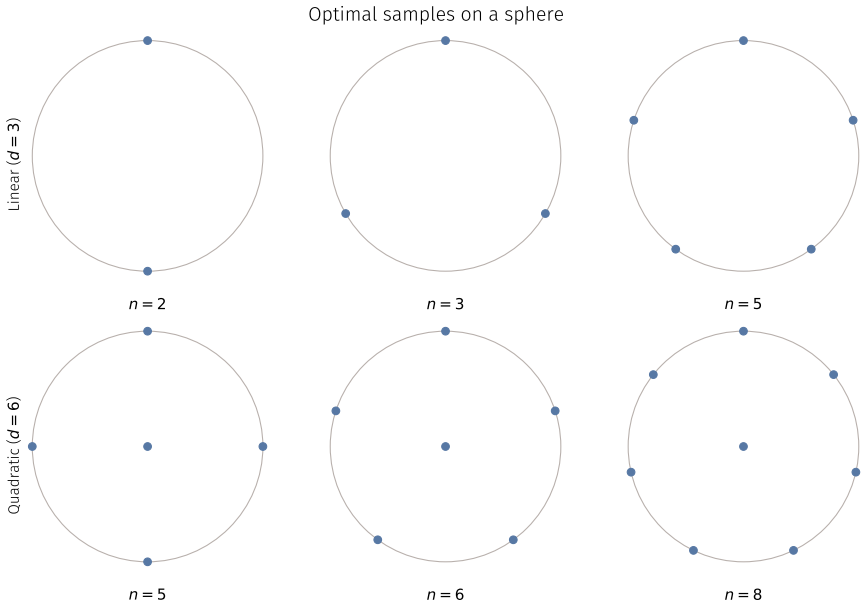


Figure 3.3.1. Optimal samples for linear and quadratic models on a ball. The first row shows the optimal samples for linear models, the second row shows the optimal samples for quadratic models. The three columns look at the under-determined, just-determined, and over-determined case. Optimal samples are not space-filling. For linear models, all points lie on the boundary of the ball. For quadratic models, there is one additional point in the center.

Random hull sampling. This sampler draws points uniformly on the boundary of a spherical or cubic trust-region. When used to complement an existing sample, it does not take the position of existing points into account. The sampler is very fast and provides a good baseline for testing against optimal samplers. Note that even for quadratic models, it is not necessary to sample a point in the center of the trust-region, as the acceptance step from the previous iteration already evaluated the objective function at that point.

Optimal hull sampling. This sampler uses the Random hull sampler to create an initial set of points and refines these points by maximizing the minimal distance between points, i.e.,

$$\underset{\mathcal{X} \in R_t^{n_t}}{\arg\max} \left\{ \min \left\{ \|x_i - x_j\| \ : \ i,j \ = \ 1,\ldots,n_t, i \ \neq \ j \right\} \right\}$$

To make the problem differentiable, we approximate the minimal distance by a smooth minimum. A smooth minimum of a vector $z = (z_1, \ldots, z_n)$ can be constructed using various approaches. We choose the log-sum-exp function, modified for the minimum-case

$$\text{SmoothMin}(z) \;=\; -\frac{1}{h}\ln\left(\sum_{i=1}^{n}\exp\left(-hz_i\right)\right)$$

Where h determines the hardness of the smooth minimum. As $h \to \infty$, the smooth minimum approaches the true minimum. If used to complement existing points, the optimal hull sampler takes the position of all points into account and positions the new points away from the existing points. This sampler is used by default in *tranquilo* and was used in all benchmarks. It provides a good compromise between sample quality and speed.

Determinant sampler. This sampler creates a D-optimal sample by minimizing the D-optimality criterion using a local optimizer. It is much slower than the optimal hull sampler and can produce lower-quality samples if the optimizer gets stuck in a local optimum. Therefore, this sampler is not used in any of our benchmarks.

3.3.2.3 Fitting

Fitting refers to the process of taking a set of model points $\mathcal{X}^{model}$ and corresponding evaluations of the residual function $\mathcal{R}^{model}$ and constructing a surrogate model M^v that approximates the residual function. As is common in the literature, the model points are scaled to the trust-region before fitting. This means that the solution of the trust-region subproblem is always performed over a unit-ball or unit-hypercube. Moreover, scaling increases the numerical stability of the model fitting. We emphasize this by using s instead of x to denote a scaled model point.

We restrict our attention to linear or quadratic models. A linear model for residual i takes the following form

$$M_t^i(s) \;=\; c_t^i + s^T g_t^i \tag{3.3.1}$$

Where $c_t^i \in \mathbb{R}$ is a scalar intercept term and $g_t^i \in \mathbb{R}^p$ is a vector of slope coefficients. g_t^i is also known as the model gradient. Thus, in total, a linear model has $p+1$ coefficients per residual and we require $p+1$ model points for a just-determined interpolation model.

A quadratic model includes additional terms

$$M_t^i(s) \;=\; c_t^i + s^T g_t^i + \frac{1}{2}s^T H_t^i s \tag{3.3.2}$$

Here H_t^i is a symmetric matrix of second-order coefficients, which is also known as the model Hessian. Due to the symmetry of H_t^i, the total number of coefficients is $(p+1)(p+2)/2$ and we require $(p+1)(p+2)/2$ model points for a just-determined interpolation model.

If we have more model points than coefficients, the model is over-determined, and instead of solving the interpolation conditions exactly, a least-squares solution is used. If fewer model points than coefficients are available, the model is under-determined, and the remaining degrees of freedom need to be resolved by an additional criterion. Additional criteria are usually based on the absolute values or norm of the model coefficients. Methods differ according to two main dimensions: First, whether they penalize all coefficients or only the coefficients in H_t^i. And second, whether they penalize the magnitude of the coefficients or the magnitude of the change in coefficients between two iterations. To capture the first dimension, we implement different fitting methods. To capture the second one, we use a residualization approach that can be used in combination with any fitting method.

Residualization means that on the left-hand side of the interpolation conditions, we do not use $\mathscr{R}_t^{model}$ directly. Instead, we subtract the predicted residuals of the previous model M_{t-1}^v, evaluated at the current model points $\mathscr{X}_t^{model}$ from the residuals. Thus, the coefficients of the fitted model only capture the difference between the previous and the current model, and any penalization that might be done by the fitting method only penalizes the change in coefficients. After fitting the model, the coefficients of the previous model can be added back to produce a model that approximates the current residual function.

OLS fitting. OLS fitting solves overdetermined models by minimizing the squared norm of the residuals. For just-determined models, this is equivalent to solving the interpolation conditions exactly. For underdetermined models, there are multiple solutions to the interpolation conditions. From those solutions, the solution where the Euclidean norm of all coefficients is smallest is chosen. If used in combination with residualization, the solution with the smallest change in coefficients (in Euclidean norm) is chosen.

Hessian-norm fitting. This fitting method is equivalent to OLS fitting for over and just-determined models. For underdetermined models, it penalizes the Frobenius norm of the Hessian coefficients. This approach is used by Wild (2008) and motivated by theoretical results that guarantee an approximation quality of the quadratic model (Larson, Menickelly, and Wild, 2019). If used in combination with residualization, the penalty is only applied to the change in Hessian coefficients between two iterations. This approach was first introduced by Powell (see Powell (2006) and Powell (2009)) and is also used in *POUNDERS* (Wild, 2017) and several other algorithms (Larson, Menickelly, and Wild, 2019).

Mixed fitting. Motivated by the comparable success of OLS-fitting and Hessian-norm-fitting in our benchmarks, we also implement a fitting method that combines the two approaches smoothly. Instead of only penalizing Hessian coefficients or penalizing all coefficients equally, we introduce the possibility of weighted penalization that differs across the intercept, gradient terms, and Hessian terms. We use this fitting method by default for underdetermined fitting problems and find that we get the best performance if we put a weak penalty on the intercept, a medium penalty on the gradient terms, and a strong penalty on the Hessian terms. The exact weights are not interpretable and were set empirically by tuning the algorithm against a benchmark set. The fitting method uses standard OLS-fitting after scaling the columns of the design matrix. After the fitting, we rescale the coefficients to undo the effects of rescaling the data. For over- and just-determined problems, this fitting method is equivalent to standard OLS-fitting.

Ridge fitting. An alternative approach to the above three fitting methods is ridge regression. Ridge regression performs an ℓ_2-regularization of the coefficients by adding a penalty to the objective function of a least-squares regression

$$\min_{\Theta \in \mathbb{R}^{k \times d}} \sum_{i=0}^{n_t} \|M^v(\tilde{x}_i; \Theta) - r(x_i)\|^2 + \lambda \|\Theta\|^2 \tag{3.3.3}$$

where Θ are the coefficients of the regression problem, and the constant λ is a penalty term that controls the shrinkage of coefficients, which leads to relatively smaller estimates for coefficients with low explanatory power. The big difference between ridge regression and the fitting methods discussed above is that the penalty has an effect even for over-determined models. While this could be attractive in the presence of noise, it introduces a practical problem: The penalty parameter λ has to be set, for which we did not find an adequate solution that performed well across all benchmarks.

3.3.2.4 Aggregation

Aggregation refers to the process of converting a vector model M_t^v, that approximates the residual function $r(x)$, into a scalar model M_t^s, that approximates the objective function $f(x)$. The scalar model is used to solve the trust-region subproblem to produce a candidate point s_t. While vector models might be linear, we only consider aggregation methods that result in quadratic scalar models. This is because linear scalar models are not capable of having internal optima, which makes them unsuitable for trust-region optimization. The choice of aggregation method

depends on the type of residual model (linear or quadratic) and the type of objective function (scalar or least-squares). We implement aggregation methods for three cases:

Least-squares objective, linear residual models. With linear vector model M_t^v, we follow *DF-OLS* by building the scalar model M_t^s through substitution of r^i by M_t^i (Equation 3.3.1) in the definition of the full objective function DET-LS

$$M_t^s(s) \equiv \sum_{i=1}^{k} M_t^i(s)^2 = c_t^s + s^T g_t^s + \frac{1}{2} s^T H_t^s s$$

Where

$$c_t^s \equiv \sum_{i=1}^{k} (c_t^i) \in \mathbb{R} \tag{3.3.4}$$

$$g_t^s \equiv 2 \sum_{i=1}^{k} c_t^i g_t^i \in \mathbb{R}^p \tag{3.3.5}$$

$$H_t^s \equiv 2 \sum_{i=1}^{k} (g_t^i)(g_t^i)^T \in \mathbb{R}^{p \times p} \tag{3.3.6}$$

We define the gradient of M_t^s as $g_t^M = \frac{d}{ds} M_t^s$, which can be derived as

$$g_t^M(s) = g_t^s + H_t^s s \tag{3.3.7}$$

Least-squares objective, quadratic residual models. With quadratic residual models, *POUNDERS* obtains an aggregate model using a "full Newton" approach. The full Newton model approximates the scalar model obtained by direct substitution of the residual functions by a second-order Taylor expansion around the current candidate point x_t^*

$$M_t^s(s) \equiv \sum_{i=1}^{k} (M_t^i(s))^2 \approx c_t^s + g_t^s s^T + \frac{1}{2} s^T H_t^s s \tag{3.3.8}$$

where c_t^s and g_t^s are defined as in 3.3.4 and 3.3.5, respectively, and

$$H_t^s \equiv 2 \sum_{i=1}^{k} \left((g_t^i)(g_t^i)^T + H_t^i c_t^i \right)$$

An alternative approach is implemented in *DFBOLS* (Zhang, Conn, and Scheinberg, 2010), where the second-order term of the scalar model is regularized based on cut-offs on the intercept and the linear terms. The regularization is designed to provide fast local convergence for problems with sparse residuals (Zhang, Conn, and Scheinberg, 2010). This approach could be implemented in *tranquilo* in the future.

Scalar objective, quadratic "residual" models. In the case of a scalar objective function, the residual function is simply the objective function, and the aggregation method is the identity function. Using the term aggregation here is simply an abstraction that allows us to use the same algorithmic framework for both cases.

3.3.2.5 Subsolvers

After obtaining a scalar model M_t^s, we solve the trust-region subproblem to obtain a candidate step s_t. The model is already scaled such that the subproblem is always solved over the same space, which is either a unit-ball or a unit-hypercube. More formally, we solve one of the following problems

$$\min_{\tilde{s}\in\mathbb{R}^p} M_t^s(\tilde{s}) \text{ s.t. } \|\tilde{s}\| \leq 1 \tag{SP-Ball}$$

$$\min_{\tilde{s}\in\mathbb{R}^p} M_t^s(\tilde{s}) \text{ s.t. } \tilde{s} \in [-1,1]^p \tag{SP-Cube}$$

After solving the subproblem, the resulting vector $\tilde{s}$ is rescaled with the radius of the effective trust-region to obtain the candidate step s_t.

The literature on subproblem optimizers is extremely well-developed, and we do not innovate in this area. Traditionally, subproblem solvers only look for approximate solutions in order to save computational resources. However, in a setting with expensive objective functions, solving the subproblem precisely incurs only a negligible overhead that is outweighed by the benefits of a precise solution. For solving the Problem SP-Ball, we use the *GQTPAR* algorithm. For solving the Problem SP-Cube, we use *BNTR*. Both algorithms are also used by *POUNDERS* (Wild, 2017). We provide numba-accelerated Python reimplementations of both algorithms. Our implementations are described further in Appendix 3.C.

3.3.2.6 Dropping Points

To avoid stagnation, there are two situations in which we drop points: In the while loop starting in Line 22, we are in a situation where the sample is larger than the target sample size and drop points without replacing them. In the while loop starting in Line 31, we have reached the target sample size and replace each dropped point with a new one. In both cases, we use the same dropping algorithm which is also used by the drop-excess filter described in Section 3.3.2.1.

3.3.2.7 Acceptance decision

The way we formalize the acceptance step in *tranquilo* plays a key role in making *tranquilo* a flexible algorithmic framework for trust-region optimization. Formally, the acceptance step looks as follows

$$(x^*_{t+1}, \rho_t) \;=\; Accept(x^*_t, s_t, \Delta M^s_t) \tag{3.3.9}$$

where x^*_{t+1} is the candidate point for the next iteration, ρ_t is a measure of progress or model quality, and ΔM^s_t is the expected improvement from taking step s_t.

Within these boundaries, many different implementations of acceptance steps are possible. Traditionally, ρ_t is calculated as the ratio of actual and expected improvement (see Equation 3.2.1), and x^*_{t+1} is either the candidate point $x^*_t + s_t$ or the current point x^*_t. In some algorithms, x^*_{t+1} can also be a model point if it yields an improvement over both the candidate point and the current point (see, for example, Cartis, Fiala, et al. (2019)). Typically, the acceptance step comprises only one new objective function evaluation at $x^*_t + s_t$.

In *tranquilo*, the acceptance step can calculate ρ_t in any way that is useful for the radius management and can use any number of objective function evaluations to create x^*_{t+1}. While we stick to traditional approaches for the serial and noise-free case, our extensions to parallel and noisy settings mainly consist of modifications to the acceptance step. Those extensions are described in Section 3.4.1.2 and 3.5.3.5.

Accept classic. In this acceptance step, ρ_t is calculated as in Equation 3.2.1 and x^*_{t+1} is either $x^*_t + s_t$ or x^*_t. The candidate point is accepted if it yields any improvement over the current point.

3.3.2.8 Trustregion radius adjustment

We base the implementation of the trust-region radius adjustment step on the radius adjustment rules of Wild (2017), which is given by

$$\Delta_{t+1}^{region} = \begin{cases} \min\{\gamma^{inc}\Delta_t^{region}, \Delta^{max}\} & \text{if } \rho_t \geq \rho^{inc} \text{ and } s_t \geq c^{ls}\Delta_t^{region} \\ \gamma^{dec}\Delta_t & \text{if } \rho_t < \rho^{dec} \\ \Delta_t & \text{otherwise} \end{cases} \tag{3.3.10}$$

As we can see from Equation 3.3.10, the updates to the trust-region radius depend on model performance, measured by ρ_t, and the length of the step s_t. Only if both are large, the trust-region radius is increased by a factor γ^{inc}. Here, a high ρ_t indicates that the model is good enough that we can afford a larger radius. A large step length indicates that the solution lies outside the current trust-region, and we would actually benefit from a larger trust-region. What counts as a large enough ρ_t is determined by a constant cutoff ρ^{inc}. As in *POUNDERS* (Wild, 2017), we bound the trust-region radius by a constant Δ^{max}.

On the other hand, if the ratio of actual to expected improvement falls below a threshold $\rho^{dec} \leq \rho^{inc}$, we shrink the trust-region radius by a factor $\gamma^{dec} \leq \gamma^{inc}$.

For the values of ρ_t between cut-off values ρ^{dec} and ρ^{inc}, we leave the trust-region radius unchanged. Similarly, if $\rho_t > \rho^{inc}$ but the step-length is small $s_t < c^{ls}\Delta_t^{region}$, we also leave the trust-region radius unchanged.

In *tranquilo*, we use the values $\rho^{dec} = \rho^{inc} = 0.1$ for the cut-offs on the ratio ρ_t. For the expansion and shrinkage factors of the trust-region radius, we use the values $\gamma^{inc} = 2$ and $\gamma = 0.5$. To identify large candidate steps, we use the value of $c^{ls} = 0.5$ for the relative step length. Finally, for Δ^{max}, we use the value of 10^6. All of these values are taken from the TAO implementation of *POUNDERS* (Dener et al., 2021).

3.3.2.9 Convergence and stopping criteria

We use common convergence criteria in Line 35 of Algorithm 1 based either on absolute or relative improvements in the objective function, absolute or relative step sizes, or the linear terms of the scalar surrogate model. Specifically, the algorithm stops at iteration t if $Converged(\mathcal{H}_t, M_t^s, x_t^*, x_{t+1}^*)$ in Line 35 evaluates to *True*. This happens if any of the following conditions are satisfied

$$
\begin{aligned}
|f(x_t^*) - f(x_{t+1}^*)| &\leq \epsilon^{fatol} \\
|f(x_t^*) - f(x_{t+1}^*)|/|f(x_t^*)| &\leq \epsilon^{frtol} \\
\|g_t^M(x_{t+1}^*)\| &\leq \epsilon^{gatol} \\
\|g_t^M(x_{t+1}^*)\|/|f(x_{t+1}^*)| &\leq \epsilon^{grtol} \\
\|x_{t+1}^* - x_t^*\| &\leq \epsilon^{xatol} \\
\|x_{t+1}^* - x_t^*\|/\|x_t^*\| &\leq \epsilon^{xrtol}
\end{aligned}
$$

These convergence criteria are taken from the *POUNDERS* implementation, described in Dener et al. (2021). Note that g_t^M is the gradient of the scalar model M_t^s, as defined in Equation 3.3.7.

3.3.3 Benchmarking

The ideal way to evaluate the performance of an optimization algorithm would be to run it on a large set of real-world problems and compare its performance to other optimizers. However, this approach is not feasible for several reasons. First, for interesting real-world problems, the exact solution is typically unknown. Second, the real-world problems we are interested in are too costly to be used in a benchmark. Third, there are no standard sets of real-world problems, so our results would not be comparable to other results in the literature.

For these reasons, it is common to evaluate optimization algorithms on standardized sets of benchmark problems with known solutions. These problems are designed to be representative of real-world problems and to include features that are challenging for optimization algorithms. However, they are fast to evaluate, so the benchmark can be run in minutes or hours instead of days or weeks. A complete benchmark is defined in terms of a set of problems, a set of solvers, and a convergence test (Dolan and Moré (2002)).

Throughout this paper, we use modified versions of the Moré-Wild benchmark set (Moré and Wild (2009)) to evaluate the performance of our algorithms. This benchmark set contains 53 non-linear least squares problems with known solutions. These test cases are constructed based on 22 functions originally derived from the CUTEr Problems (Gould, Orban, and Toint (2003)). The objective functions are twice continuously differentiable, but we do not make use of the derivatives in our benchmarks. The parameter dimensions p vary between 2 and 12, where the median dimension is 7. The dimension of the least squares residuals k is between 2 and 65. Only three of the 53 problems have local minimizers that are not global minimizers. These are based on the Freudenstein and Roth function and the Brown almost-linear function. The remaining 50 problems have a unique minimizer.

The Moré-Wild benchmark set is standard in the recent literature on derivative-free optimization. Among others, it has been used to benchmark *POUNDERS* (Wild (2017)), *DFOGN* (Cartis and Roberts (2019)), and *DFOLS* (Cartis, Fiala, et al. (2019)).

The benchmark set plays an important role not only in measuring the final performance of the algorithm but also in tuning the algorithm's hyperparameters during development. In order to avoid overfitting the tuning parameters to the benchmark set and to improve the robustness of our conclusions, we extend the benchmark set with randomly generated problems. For each of the 53 problems, we generate four additional problems by drawing a new vector of start parameters in the neighborhood of the original start parameters. The neighborhood is defined by multiplying the original start parameters with 0.9 or 1.1. In the case of parameter values smaller than 1, we switch to additive perturbations by adding and subtracting 0.1. The new start parameters are drawn uniformly from the neighborhood. If the objective function is undefined at the new starting values, we tighten the neighborhood until we find a valid starting point.

To measure the performance of different algorithms, we need a convergence test. Importantly, a onvergence test is only based on the history of function values of each optimizer and the known solution of the problem. It is independent of the algorithm's internal convergence criteria. We use the following convergence test, as proposed by Moré and Wild (Moré and Wild (2009)), to test whether algorithm j solved problem i

$$\frac{f_i(x_{ij}^*) - f_i^*}{f_i(x_{i0}) - f_i^*} \leq \tau \tag{3.3.11}$$

where $\tau > 0$ is a tolerance level, x_{i0} is the vector of start parameters, f_i^* is the known minimum of the objective function, and $f_i(x_{ij}^*)$ is the lowest objective function value obtained by the optimizer. Note that x_{ij}^* can be any point that has been tried out by algorithm j. For noise-free problems, we set $\tau = 10^{-3}$.

Once we have the convergence test to decide whether an algorithm solved a problem, we need a way to measure the computational budget the algorithm needed until it found a solution. The computational budget can also be interpreted as the runtime until solution. Since we are interested in applications where the objective function is expensive, meaning that, by assumption, the algorithm will spent most of its runtime on evaluating the objective function, we use the number of function evaluations as the measure of the computational budget. Using walltime instead would mostly measure how much work is done inside the algorithm itself because all objective functions in the benchmark set are very fast to evaluate. Using the number of function evaluation is common practice in the literature on derivative-free least-squares optimization (see for example Wild (2017) and Cartis, Fiala, et al. (2019)).

The standard way of visualizing the performance of a set of solvers on a benchmark set are *performance profiles* (Moré and Wild (2009)), which are also known as *profile plots*. Performance profiles show the share of solved problems on the y-axis. On the x-axis, they show a normalized measure of the computational budget. Normalized here means that the number of function evaluations each algorithm needed to solve a given problem is divided by the runtime that the fastest algorithm needed to solve the problem. This makes performance profiles useful even for benchmark sets that contain problems with very different difficulty levels. Without normalization, the performance profile would be dominated by the hardest problems. The x-axis of performance profiles starts at 1. The y-value each algorithm achieves at 1 is the share of problems for which this algorithm was the fastest.

Figure 3.3.2 shows the performance profiles for the least-squares version of *tranquilo* and compares it against *DFO-LS* and *POUNDERS*.

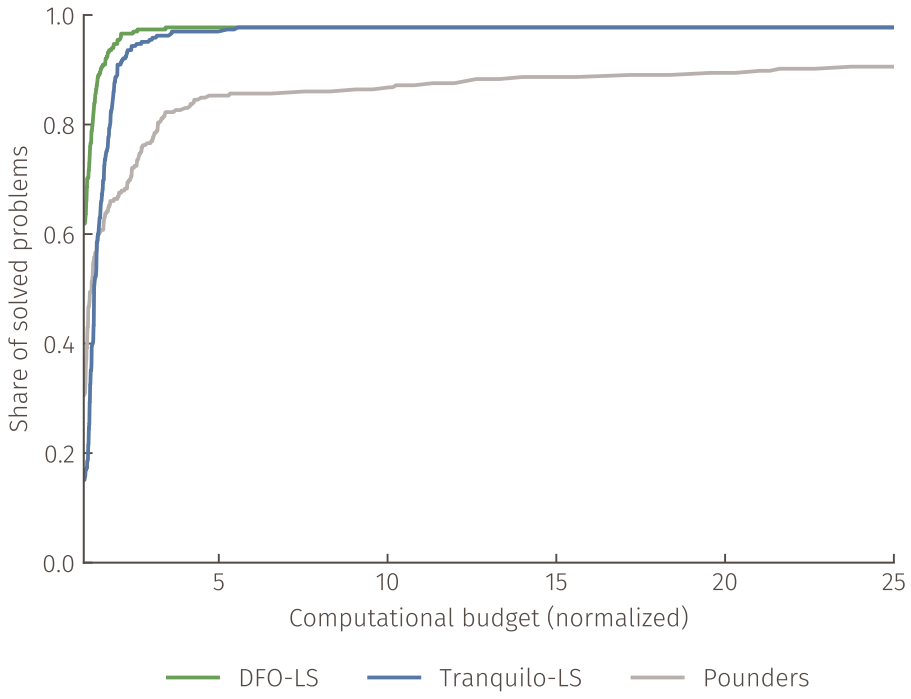


Figure 3.3.2. Comparison of least-squares optimizers on an augmented Moré-Wild benchmark set. The y-axis shows the share of problems solved. The x-axis shows the normalized computational budget. The computational budget is measured in terms of objective function evaluations needed by the optimizers. Normalized means that the number of function evaluations each algorithm needed to solve a given problem is divided by the number of function evaluations the fastest algorithm needed to solve that problem. Both *DFO-LS* and *tranquilo* solve the same number of problems. In most problems, *DFO-LS* is slightly faster than *tranquilo*. *POUNDERS* is slower than the other two on most problems. Moreover, it fails to solve some problems to the required level of precision.

Both *DFO-LS* and *tranquilo* solve the same number of problems. In most problems, *DFO-LS* is slightly faster than *tranquilo*. *POUNDERS* is slower than the other two on most problems. Moreover, it fails to solve some problems to the required level of precision. This is in line with results by Cartis, Fiala, et al. (2019) who suspect that the lack of precision is related to the minimal trust-region radius *POUNDERS* uses.

Figure 3.3.3 shows the performance profiles for the scalar version of tranquilo and compares its performance against *BOBYQA* implementations from NlOpt and the Numerical Algorithms Group (NAG) as well as *Nelder-Mead* implementations from NlOpt and SciPy.

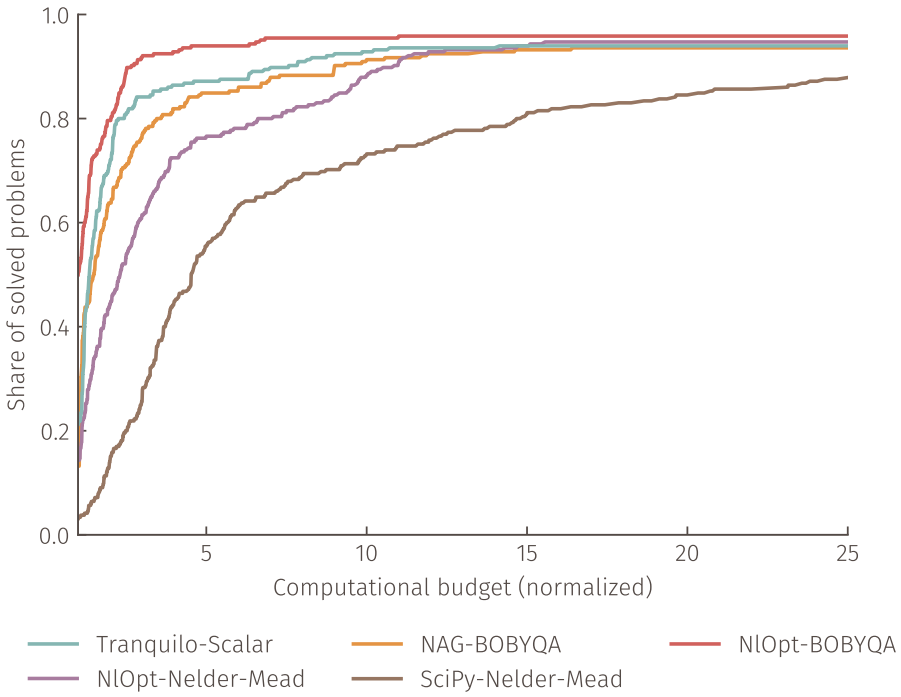


Figure 3.3.3. Comparison of scalar optimizers on an augmented Moré-Wild benchmark set. The *y*-axis shows the share of problems solved. The *x*-axis shows the normalized computational budget. The computational budget is measured in terms of objective function evaluations needed by the optimizers. Normalized means that the number of function evaluations each algorithm needed to solve a given problem is divided by the number of function evaluations the fastest algorithm needed to solve that problem. The fastest and most robust optimizer is the NlOpt implementation of *BOBYQA*. The slowest and least robust optimizer is the SciPy implementation of *Nelder-Mead*. All other algorithms solve slightly fewer problems than the NlOpt implementation of *BOBYQA*. Among them, *tranquilo* is the fastest, followed by the NAG implementation of *BOBYQA* and the NlOpt implementation of *Nelder-Mead*.

The fastest and most robust optimizer is the NlOpt implementation of *BOBYQA*. The slowest and least robust optimizer is the SciPy implementation of *Nelder-Mead*. All other algorithms solve

slightly fewer problems than the NlOpt implementation of *BOBYQA*. Among them, *tranquilo* is the fastest, followed by the NAG implementation of *BOBYQA* and the NlOpt implementation of *Nelder-Mead*. Generally, the derivative free trust-region optimizers seem faster and more robust than the direct search methods.

Figure 3.3.4 combines the two cases and compares scalar and least-squares algorithm in a single plot. The main purpose of this plot is to show that the least-squares algorithms indeed outperform similar scalar algorithms when applicable.

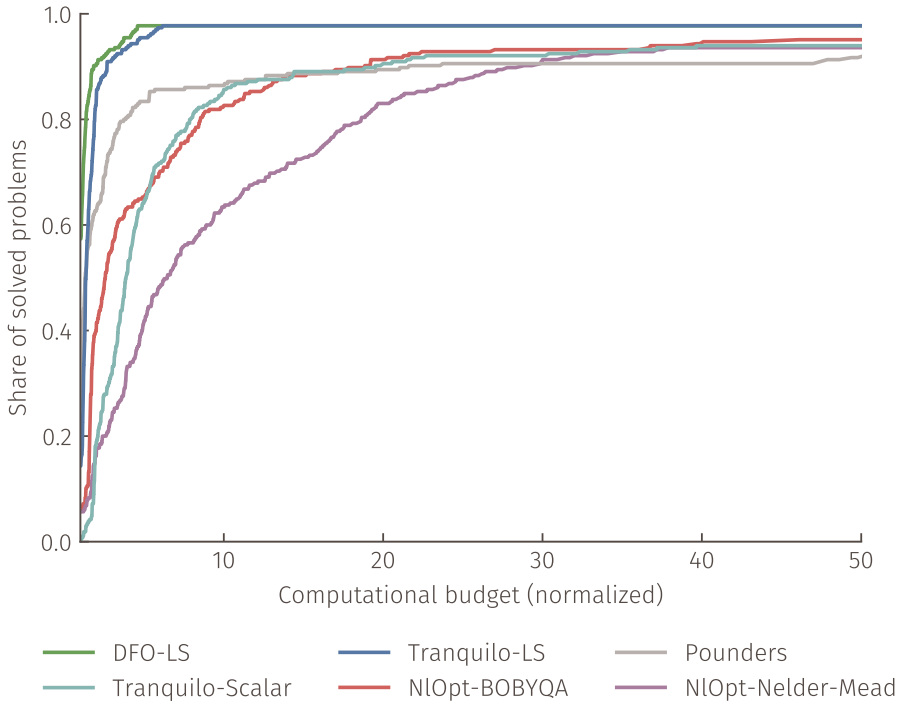


Figure 3.3.4. Comparison of scalar and least-squares optimizers on an augmented Moré-Wild benchmark set. The *y*-axis shows the share of problems solved. The *x*-axis shows the normalized computational budget. The computational budget is measured in terms of objective function evaluations needed by the optimizers. Normalized means that the number of function evaluations each algorithm needed to solve a given problem is divided by the number of function evaluations the fastest algorithm needed to solve that problem. The plot shows that least-squares algorithms are generally faster and more robust than their scalar counterparts.

The combined plot shows a clear separation of several groups of algorithms: Least-squares algorithms that use linear residual models and aggregate them into quadratic scalar models are clearly faster than all other algorithms. *POUNDERS*, as the only least-squares algorithm that uses a quadratic residual model, is faster than scalar optimizers but cannot solve all problems to the required level of precision. The scalar optimizers can again be split into two groups: The two

BOBYQA implementations (i.e., model-based trust-region optimizers) are faster than the NlOpt implementation of *Nelder-Mead* (i.e., a direct search method). The SciPy implementation of *Nelder-Mead* is omitted because it is much slower than the other algorithms (see Figure 3.3.3).

3.4 Parallelization

Tranquilo is designed for objective functions f that cannot easily be parallelized. This means that if parallel hardware is available, the parallelization should be done on the algorithm level.

When designing a parallel algorithm, the focus shifts from minimizing the number of objective function evaluations to minimizing the number of *batches*, where each batch consists of n^{batch} objective function evaluations that can be done in parallel.

For instance, assume that n^{batch} is equal to four and we require seven function evaluations. This example is illustrated in Figure 3.4.1. If all seven function evaluations are independent, we can perform them in two batches. Further, the second batch is not full. If we do not derive utility from idle resources, we could do one "free" evaluation in the second batch. The goal of our parallelization approaches in tranquilo is therefore to do as many function evaluations as possible in parallel and to find good ways to use up any "free" evaluations.

$$\left\{\left[\underbrace{f(x_1), f(x_2), f(x_3), f(x_4)}_{\text{Batch 1}}\right], \left[\underbrace{f(x_5), f(x_6), f(x_7)}_{\text{Batch 2}}\right]\right\}$$

Figure 3.4.1. Illustration of batched evaluations. We assume that the batch-size is four and therefore 7 independent function evaluations can be done in two batches. The second batch is not full and therefore contains one "free" function evaluation.

In most cases, the batch size (n^{batch}) is equal to the number of cores that are available to the optimizer. However, we allow for the batch size to be larger than the number of cores. This allows us to simulate the behavior of heavily parallelized algorithms on machines with fewer cores.

The basic version of *tranquilo* as described in Algorithm 1 already creates some situations in which the objective function needs to be evaluated on multiple points and there are no dependencies between these evaluations: In the first iteration, the objective function is evaluated at every point in the initial sample which contains at least $p + 1$ points. In all subsequent iterations, the objective function is evaluated at all newly sampled points. Of course, the parallel version of tranquilo exploits these situations. However, in most of these cases, the number of function evaluations required are not multiples of the batch size and therefore, "free" evaluations are left on the table. Moreover, there are several situations in which just one function evaluation is required.

This is for example the case when points are replaced to avoid stagnation (Line 22 of Algorithm 1) and in the acceptance step. The parallel version of tranquilo exploits most of these situations and fills up the "free" function evaluations with different strategies.

3.4.1 Adding parallelization to tranquilo

Almost all of the changes required to add parallelization to tranquilo are done by switching out components. The only exception is the sampling step in each iteration. Here, instead of passing n^{target} as target sample-size into the samplers (see Line 13), we pass in a target sample size that makes sure that the number of newly sampled points is a multiple of n^{batch}. Moreover, the acceptance step now depends on the batch size.

3.4.1.1 Filtering

As the usage of these "free" evaluations potentially leads to many more available points in the history, we observed that using no filter leads to worse benchmark results compared to using the *Drop excess* filter. The *Drop excess* filter is described in Section 3.3.2.1 and activated by default when $n^{batch} > 1$. In the parallel benchmarks we set $n^{filter} = 3n^{target}$; see Figure 3.4.4 for reference.

3.4.1.2 Acceptance Step

During the acceptance step, we determine the new candidate point x^*_{t+1} and a measure of model quality ρ_t. A typical serial acceptance step requires a single function evaluation at the candidate point $x^*_t + s_t$ (see Section 3.3.2.7). In the parallel case, this leaves us with many "free" evaluations, which we can use to improve efficiency.

In the parallel acceptance step, we invest one evaluation in the candidate point, which leaves us with $n^{batch} - 1$ available evaluations to fill up the batch. We use these in two ways. First, we check whether the candidate point lies at the trust-region border. If so, we interpret it as a signal that the direction of the step is good, but the step size might be too small. We thus sample points on the line that goes through the current best point and the candidate point. We call this a line search. Second, we try to predict at which points the objective function needs to be evaluated in the next iteration and perform the evaluations. We call this speculative sampling.

Line Search. Consider the illustration in Figure 3.4.2. For the line search, we sample points on a line starting at the current best point (the center point of the circle in the illustration) and going

through the candidate point (the red point in the illustration). Formally, the line is given by

$$Line(\alpha) = x_t^* + \alpha s_t$$

where $Line(1)$ equals the candidate point. For values $\alpha < 1$, $Line(\alpha)$ represents points inside the trust-region, while for $\alpha > 1$, $Line(\alpha)$ represents points outside of the trust-region. Since we believe the step was too small, we want to sample outside of the trust-region. In one iteration, the trust-region radius can increase by a maximum of 2. To simulate a continuous maximal increase of the radius, we sample a maximum of three points on the line with $\alpha = 2, 4, 8$, respectively, depending on the number of "free" evaluations (the blue points in the illustration).

If the search direction of the candidate step is good, a line search can dramatically increase the speed of the algorithm, allowing us to go as far as 2^3 times the initial trust-region radius, which translates to jumping ahead three iterations of the algorithm. If any of the line search points is better than the current best point, we accept it.

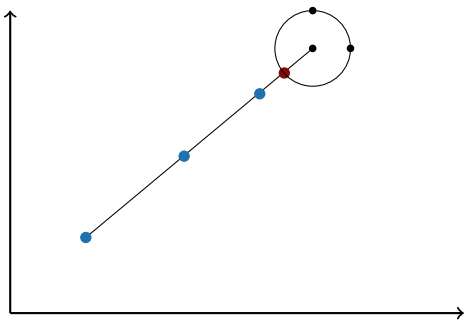


Figure 3.4.2. Illustration of the line search. The candidate point is shown in red. The black dots show current model points. The blue dots show the line search points. The line search points are all on a line that goes through the current best point and the candidate point. The spacing is at 2, 4, and 8 times the current current trust-region radius.

Speculative Sampling. Consider the illustration in Figure 3.4.3. For the speculative sampling, we assume that the candidate point will be accepted (the red point in the illustration). In this case, we can use any "free" evaluations to sample points around the candidate, as these points will be required in the next iteration. We do not know, however, how the radius will change. After empirical testing, we found that setting the radius of the region from which we draw the speculative sampling to 0.75 times the current trust-region radius results in the best benchmark performance. The speculative sample points are shown in blue in the illustration.

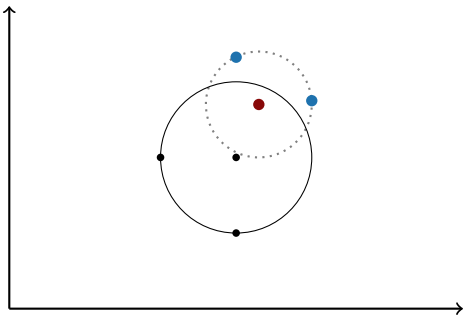


Figure 3.4.3. Illustration of the speculative sampling. The candidate point is shown in red. The black dots show a hypothetical sample of existing points that would be available in the next iteration if the candidate point was accepted. The blue dots show the speculative sample. The points are sampled in the same way they would be sampled in the next iteration if the candidate point was accepted and the radius was 0.75 times the current trust-region radius.

Implementation of the Parallel Acceptance Step. If enough "free" function evaluations are available, we combine both, the line search and speculative sampling in our parallel acceptance step. For an efficient combination, we first calculate the line-search points and already take them into account as existing points when creating the speculative sampling. Of course, the function evaluations on both, the line-search points and the speculative sample are done in parallel, together with the function evaluation at the candidate point. The parallel acceptance step is shown in Algorithm 2.

Algorithm 2: Parallel acceptance step

Input: The current parameter vector x_t^*, the candidate step s_t, the expected improvement ΔM_t^s, the effective trust-region R_t, the history $\mathcal{H}_t$, and the batch size n^{batch}.

if $n^{batch} = 1$ **then**
    return $AcceptClassic(x_t^*, s_t, \Delta M_t^s)$
else
end
if $x_t^* + s_t$ *is on the border of* R_t **then**
    Calculate the number of available line search points: $n^{ls} = \min\{n^{batch} - 1, 3\}$
    Sample n^{ls} points on a line: $\mathcal{X}_t^{ls} = \{x_t^* + 2^i s_t : i = 1, \ldots, n^{ls}\}$
else
    $n^{ls} = 0$ and $\mathcal{X}_t^{ls} = \{\}$
end
Calculate number of speculative sampling points: $n^{speculative} = n^{batch} - 1 - n^{ls}$
Define a speculative sampling region: $R_t^{speculative}$ with the same center as R_t, and a radius of 0.75 times that of R_t
Scan the history for existing points $\mathcal{X}_t^{existing} = \{x \in \mathcal{H}_t : x \in R_t^{speculative}\}$
Sample speculative points: $\mathcal{X}_t^{speculative} = Sample(\mathcal{X}_t^{ls} \cup \mathcal{X}^{existing}, R_t^{speculative}, n^{speculative})$
Compute $\rho_t = (f(x_t^*) - f(x_t^* + s_t))/\Delta M_t^s$
Compute $x_{t+1}^* = \arg\min\{f(x) : x \in \{x_t^* + s_t\} \cup \mathcal{X}_t^{ls} \cup \mathcal{X}_t^{speculative}\}$
return (x_{t+1}^*, ρ_t)

3.4.2 Benchmarking

The performance-profiles have to be adjusted for the parallel case, as the number of objective evaluations does not provide a good measure of runtime anymore. Instead, we use the number of batch evaluations to measure the computational budget. This reflects our assumption that in the parallel case the evaluation of a batch takes as much time as the evaluation at a single point.

Figure 3.4.4 shows the benchmark results of our parallel algorithm on the Moré and Wild (2009) benchmark set; see Section 3.3.3 for details on benchmarking. We compare the parallel least-squares version of *tranquilo*, for batch sizes 2, 4, and 8, to the serial version of *tranquilo* and the *DF-OLS* algorithm.

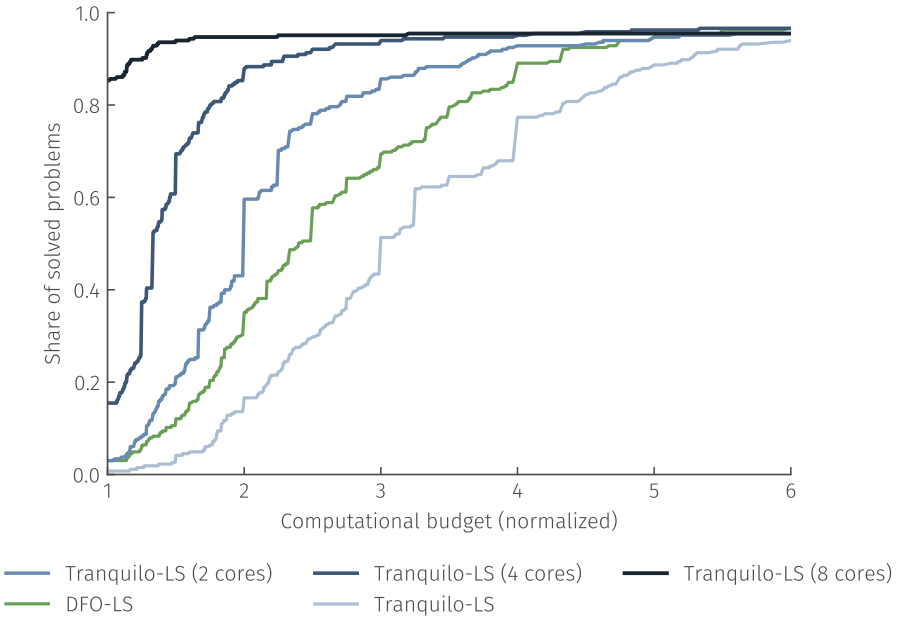


Figure 3.4.4. Comparison of parallel and serial least-squares optimizers on an augmented Moré-Wild benchmark set. The *y*-axis shows the share of problems solved. The *x*-axis shows the normalized computational budget. The computational budget is measured in terms of batches of objective function evaluations needed by the optimizers. Normalized means that the number of batches each algorithm needed to solve a given problem is divided by the number of batches the fastest algorithm needed to solve that problem. The plot shows that *tranquilo* strongly benefits from having more cores available. The 8-core version is the fastest algorithm for roughly 85% of the problems.

As in Section 3.3.3, the *y*-axis denotes the share of solved problems, while the *x*-axis denotes the multiple of the minimal number of *batches* needed to solve the problem. This implies that the intercept can be interpreted as the share of problems that the respective algorithm was able to solve the fastest.

The serial version of *tranquilo* (lightest blue) is slower than the *DF-OLS* algorithm (green), as was also seen in the least-square benchmark (see Figure 3.3.2). The parallel versions of *tranquilo*, however, dominate the *DF-OLS* algorithm for given batch sizes. In particular, when using a batch size of 8, *tranquilo* is the fastest algorithm for roughly 85% of the problems. In some sense, this comes as no surprise, as there are multiple effects playing a role when using a larger batch size. First, the sample sizes will be larger, and second, we can fully utilize the combination of line search and speculative sampling.

3.5 Noisy optimization

As described in Section 3.2.2, the main challenge for trust-region optimizers in the case of noisy objective functions is to determine how often the objective function should be evaluated at each point. Multiple evaluations at the same points are necessary in order to average out the noise to a level that allows the optimizer to make progress. *DFO-LS* puts this burden on the user. *ASTRO-DF* determines the sample size adaptively by adding evaluations until an estimated standard error falls under a fixed factor of the squared trust-region radius.

Tranquilo takes an entirely different approach that recognizes that the effects of noise are similar to the effect of approximation error in the surrogate model –which is handled very well by trust-region optimizers. *Tranquilo* therefore introduces a new measure of model quality, ρ^{noise}, that can be used to adjust the number of function evaluations used to construct surrogate models. Moreover, we borrow ideas from statistical power analysis to determine the number of function evaluations required to make an acceptance decision.

The structure of this section is as follows: Section 3.5.1 discusses the effects of setting sub-optimal sample sizes and why determining optimal sample sizes ex-ante is hard. Section 3.5.3 describes the changes to the core algorithm framework that are necessary to make *tranquilo* robust to noise, as well as the implementation of new components for noisy optimization. Section 3.5.4 compares the noise-robust version of *tranquilo* against different configurations of *DFO-LS* on a noisy version of the Moré-Wild benchmark set.

3.5.1 The importance of sample sizes

To make things precise, we use the following notation: m_t^{model} denotes the number of repeated function evaluations at each model point in iteration t. m_{t1}^{accept} and m_{t2}^{accept} are the number of repeated function evaluations at the current x and the candidate point in the acceptance step of iteration t. For convenience, most algorithms for noisy optimization set all three numbers equal, even though they are conceptually quite different. In *tranquilo*, we therefore keep the distinction, and since each number of repetitions is set adaptively, they do not generally coincide. Although the number of repetitions (symbolized by the letter m) will often be called sample size in the following sections, it is not to be confused with the sample size as used in the earlier sections of this paper (symbolized by the letter n), which measures the number of distinct x-vectors used for building a surrogate model. If the distinction is not clear from the context, we will use the term number of repetitions.

If m_t^{model} is too small, the surrogate model will be estimated imprecisely, even if the trust-region radius is chosen appropriately and a quadratic model can approximate the true objective function

well. This means that the candidate points obtained from minimizing the surrogate model have low quality, and therefore, the measure of model quality ρ will be low in many iterations. If no further measures are taken, the radius is decreased until it collapses to zero, and the algorithm stagnates. On the other hand, if the sample size is too large, the algorithm will become prohibitively expensive.

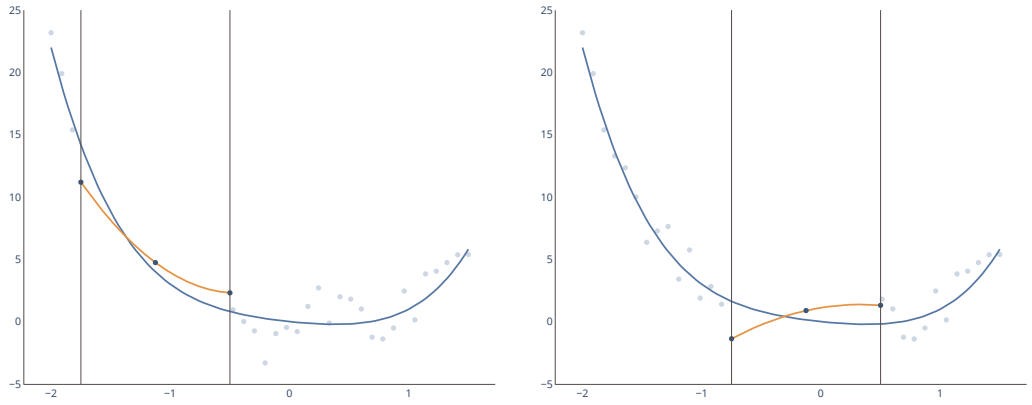
A similar effect occurs in the acceptance step: If the sample size is too small, the acceptance decision will be based on noisy information. It becomes possible that candidates that are worse than the current point in expectation are accepted due to lucky draws and, conversely, it can happen that very good candidates are rejected. Moreover, ρ becomes noisy and the radius adjustment erratic.

To talk about noisy and noise-free function evaluations and residuals, respectively, we use the following conventions: $f(x, \xi_j)$ is the j-th noisy realization of the objective function at x and $\mathbb{E}f(x, \xi)$ is the expectation of the objective function at x. $\overline{\mathscr{F}}_t^{model}$ is used to denote the average of all function evaluations at the model points. Analogous notation is used for the residual function r.

Figure 3.5.1 illustrates why it is hard to pick optimal sample sizes. We focus on m^{model}, but very similar arguments apply to m^{accept}. The left and right plot show the same objective function. The vertical lines mark trust-regions bounds. In both plots, the trust-region radius is the same, but the centers differ. In each trust-region, we plot a D-optimal sample (x_1, x_2, x_3). At each point, we observe one noisy function evaluation $f(x_1, \xi_1), f(x_2, \xi_1)$ and $f(x_3, \xi_1)$. The realizations of the random term ξ are the same in both plots.

In the left plot, the trust-region is in a steep area of the objective function. While the effect of noise makes the approximation quality of the surrogate model worse compared to a noise-free case, the differences in observed function evaluations f is dominated by differences in $\mathbb{E}f$. Therefore, the surrogate model still points into the right direction.

In the right plot, the trust-region is in a flat area of the objective function. Therefore, the differences in observed function evaluations f are dominated by differences in the realized noise and do not reflect differences in $\mathbb{E}f$. As a result, the surrogate model points in the wrong direction.



(a) Noise with a steep function

(b) Noise with a flat function

Figure 3.5.1. Effect of noise on a surrogate model

This simple illustration shows that setting the sample size based on the variance of the error term and the trust-region radius alone is not sufficient to ensure that sample sizes are close to optimal. Since objective functions are typically steeper in the beginning and flatter as we get closer to the optimum, the optimal sample size will typically be increasing in the iteration. However, it is very hard to pick an optimal schedule for this. Therefore, approaches that require the user to set a sequence of sample sizes as a function of the iteration counter require a lot of trial and error in practice. This motivates us to develop a fully adaptive approach to selecting the sample sizes in *tranquilo*.

3.5.2 Core ideas for noise handling

3.5.2.1 Determining m^{model}

Our approach is based on the observation that the effects of noise on model quality are similar to the effects of approximation error and can, therefore, be handled in a similar way.

Approximation error is introduced by the fact that a quadratic surrogate model is not able to capture the shape of the objective function exactly. The tuning parameter to govern the size of the approximation error is the trust-region radius. A larger radius means a larger approximation error. Talyor-like error bounds ensure that by making the radius small enough, the approximation error can be made arbitrarily small. Of course, a small radius comes with the cost of smaller step sizes and slower progress. Therefore, it is not a goal to make the approximation error as small as

possible but to make the radius as large as possible while ensuring that the model is good enough to produce suitable candidate points.

To achieve this balance, trust-region optimizers use ρ, the ratio of actual to predicted improvement, as a measure of model quality. If the actual improvement is larger or similar to the predicted improvement, the model was good enough to find suitable candidates, and, therefore, the radius can be increased or kept constant. Otherwise, the radius can be decreased.

Random error is introduced by the fact that observations of the objective function are noisy. The tuning parameter to govern the size of the random error is the number of repeated function evaluations. Under regularity conditions, a law of large numbers ensures that making m^{model} large enough will make the random error arbitrarily small. It is worth emphasizing that making the error small comes at the cost of more function evaluations and it is, therefore, not a goal to make the error as small as possible but to make it just small enough to ensure that the model is good enough to find suitable candidates.

The presence of random error does, of course, not alleviate the approximation error. If ρ is calculated as usual it reflects both kinds of errors:

$$\rho = \frac{f(x^*, \xi) - f(x^* + s, \xi)}{M^s(x^*) - M^s(x^* + s)} = \frac{\text{Actual Improvement}}{\text{Expected Improvement}}$$

To obtain a measure that mostly reflects approximation error, we could, of course, replace the noisy function evaluations with averages over multiple repetitions. However, this again requires determining a sample size. We, therefore, first focus on finding a measure ρ^{noise} that only reflects the effect of random error on the model's ability to produce good candidate points.

We start by noting that the denominator of ρ is made up of quadratic models and that any alternative measure that puts a similar focus on the model's ability to produce good candidate points will likely have quadratic models in the denominator as well. If ρ^{noise} should not be affected by approximation error, we have to replace all occurrences of the objective function f in the numerator with a quadratic model that approximates f. Of course, the best quadratic approximation of f we have available is the surrogate model M^s.

We therefore construct ρ^{noise} with a simulation approach in which we use the current surrogate models M^v and M^s as a stand-in for the residual and objective functions. Using M^v we can create a sample of "true" residuals for each point in the current set of model points $\mathcal{X}_t^{model}$. Using an estimate of the noise variance, we can then simulate noisy function evaluations. On the simulated function evaluations, we can fit simulated vector models $M^{v,sim}$, aggregate them into simulated scalar models $M^{s,sim}$, and solve the simulated trust-region subproblem. This yields a candidate step s^{sim}. Given these ingredients, we can calculated ρ^{noise} as follows:

$$\rho^{noise} = \frac{M^s(x^*) - M^s(x^* + s^{sim})}{M^{s,sim}(x^*) - M^{s,sim}(x^* + s^{sim})} \tag{3.5.1}$$

All steps in the calculation of the simulated candidate step s^{sim} completely mirror the steps done to calculate the normal candidate step s in tranquilo. The only difference is that the true objective function f is replaced by the surrogate model M^s and that noisy function evaluations are replaced by their simulated counterparts. This means that ρ^{noise} is a pure measure of the effects of random error on the model's ability to produce good candidate points. As both the numerator and denominator are made up of quadratic models, it does not contain any approximation error.

In our practical implementation, we repeat the simulation b times to create a vector of ρ^{noise} values. This vector can then be used to adjust m_t^{model}. We describe the details of this implementation in Section 3.5.3.3.

3.5.2.2 Determining m^{accept}

In the absence of noise, accepting or rejecting a candidate step s boils down to the simple question of whether $f(x_t^* + s)$ is smaller than $f(x_t^*)$. In the presence of noise, this turns into a question of expected values: Is $\mathbb{E}f(x_t^* + s, \xi)$ smaller than $\mathbb{E}f(x_t^*, \xi)$? This is a hypothesis test.

Empirical economists who collect data frequently have to make decisions about sample sizes. Collecting data is expensive, but collecting too little data might render non-zero effects statistically insignificant. The method of choice for determining sample sizes to alleviate this problem is power analysis.

For simplicity, we assume that our test statistic of interest –the difference in means between function evaluations at the current and candidate point– is normally distributed. This can either be seen as a finite sample assumption or justified with asymptotic arguments. Given this assumption, we need to fix three ingredients for a power analysis: First, the significance level of the hypothesis test that is going to be performed. Second, the desired power for detecting an effect. Third, the minimal detectable effect size.

We treat the first two as tuning parameters of the algorithm and set them to maximize performance in benchmarks. For the minimal detectable effect size, this cannot be done because it depends on the scale of the objective function. However, the solution of the trust-region subproblem generates an expected improvement as a by-product. We use this expected improvement as a minimal detectable effect size.

The results of the power analysis, together with the existing number of function evaluations at x_t^* and $x_t^* + s$, can then be used to calculate optimal values for m_{t1}^{accept} and m_{t2}^{accept} that minimize the number of additionally required function evaluations while achieving the desired power. In our practical implementation we also keep the number of function evaluations used in the acceptance step between a lower and upper bound that can be set by the user. The details of this implementation are described in Section 3.5.3.5.

3.5.3 Adding noise handling to tranquilo

3.5.3.1 Noisy-trustregion-framework

In this section, we describe the changes to the core algorithm framework as depicted in Algorithm 1 that are necessary to make *tranquilo* robust to noise. The modified algorithm is shown in Algorithm 3. The changes are highlighted in green.

The noisy version of *tranquilo* contains two additional inputs: m_0, which determines how often the objective function is evaluated at the start parameters, and m_0^{model}, which is the initial value for the adaptively chosen number of repeated function evaluations. m_0 needs to be larger or equal to two, such that we cannot just get an estimate of the function value at the start parameters but also an estimate of the noise variance. In our benchmarks, we set m_0 to 5 and m_0^{model} to 1.

The first thing that changes in the algorithm is that the History $\mathscr{H}_t$ is now initialized with a set of function evaluations at the start parameters instead of just one function evaluation. In general, each parameter vector in the history can now be associated with several observed function evaluations, and the number of function evaluations varies over time. As a consequence, $\mathscr{F}_t^{model}$ and $\mathscr{R}_t^{model}$ are now replaced by $\overline{\mathscr{F}}_t^{model}$ and $\overline{\mathscr{R}}_t^{model}$ which contain averages over multiple function or residual evaluations at each model point. Similarly, the initial vector model M_0^v is now initialized with the average of the function evaluations as intercept terms. All other coefficients stay at zero.

The main loop of *tranquilo* proceeds as before through the process of scanning the history, filtering existing points, sampling new points, fitting and aggregating vector models, and solving the trust-region subproblem. The two while loops for stagnation handling are also unchanged.

Algorithm 3: *Tranquilo* algorithm (noisy case)

Input: Starting point x_0^*, initial trust-region radius Δ_0^{region}, target sample size n^{target}, search factor γ^{search}, minimum step size s^{min}, sample increment n_{stag}^{drop}, maximum number of iterations t^{max}, maximum number of trials to avoid stagnation n_{stag}^{max}, lower and upper bounds l and u, the number of function evaluations at the start parameters m_0, and the initial value for the number of repeated function evaluations m_0^{model}.

Initialize history with $\mathscr{H}_0 = \left\{\left(x_0^*, \{r_j(x_0^*) : j = 1, ..., m_0\}\right)\right\}$

Initialize vector model M_0^v with intercept terms at $\frac{1}{m_0}\sum_{j=1}^{m_0} r_j(x_0^*)$ and all other coefficients set to 0

for t=0,1,...,t^{max} **do**

  Calculate the search radius $\Delta_t^{search} = \gamma^{search}\Delta_t^{region}$

  Calculate the effective trust-region R_t based on x_t^*, Δ_t^{region}, l and u

  Scan the history for existing points $\mathscr{X}_t^{existing} = \{x \in \mathscr{H}_t : \|x_t^* - x\| \leq \Delta_t^{search}\}$

  Filter existing points: $\mathscr{X}_t^{filtered} = Filter(\mathscr{X}_t^{existing})$

  if $|\mathscr{X}_t^{filtered}| < n^{target}$ **then**

    Sample $n^{target} - |\mathscr{X}_t^{filtered}|$ new points in the trust-region: $\mathscr{X}_t^{new} = Sample(\mathscr{X}_t^{filtered}, R_t, n^{target})$

    $\mathscr{X}_t^{model} = \mathscr{X}_t^{filtered} \cup \mathscr{X}_t^{new}$

  else

    $\mathscr{X}_t^{model} = \mathscr{X}_t^{filtered}$

  end

  Build a vector model $M_t^v = Fit(\mathscr{X}_t^{model}, \overline{\mathscr{R}}_t^{model}, M_{t-1}^v, R_t)$

  Aggregate the vector model: $M_t^s = Aggregate(M_t^v)$

  Solve the surrogate problem: $s_t = Subsolve(M_t^s, R_t)$

  while $|\mathscr{X}_t^{model}| > n^{target}$ *and* $\|s_t\| \leq s^{min}$ **do**

    Reduce the sample: $\mathscr{X}_t^{reduced} = Drop(\mathscr{X}_t^{model}, n_{stage}^{drop}, \Delta_t^{region})$ and set $\mathscr{X}_t^{model} = \mathscr{X}_t^{reduced}$

    Build a vector model $M_t^v = Fit(\mathscr{X}_t^{model}, \overline{\mathscr{R}}_t^{model}, M_{t-1}^v, R_t)$

    Aggregate the vector model: $M_t^s = Aggregate(M_t^v)$

    Solve the surrogate problem: $s_t = Subsolve(M_t^s, R_t)$

  end

  $n_{stag} = 0$

  while $\|s_t\| \leq s^{min}$ *and* $n_{stag} \leq n_{stag}^{max}$ **do**

    Reduce the sample: $\mathscr{X}_t^{reduced} = Drop(\mathscr{X}_t^{model}, n_{stag}^{drop}, \Delta_t^{region})$

    Sample new points in the trust-region: $\mathscr{X}_t^{new} = Sample(\mathscr{X}_t^{reduced}, R_t, n^{target})$ and set $\mathscr{X}_t^{model} = \mathscr{X}_t^{reduced} \cup \mathscr{X}_t^{new}$

    Build a vector model $M_t^v = Fit(\mathscr{X}_t^{model}, \overline{\mathscr{R}}_t^{model}, M_{t-1}^v, R_t)$

    Aggregate the vector model: $M_t^s = Aggregate(M_t^v)$

    Solve the surrogate problem: $s_t = Subsolve(M_t^s, R_t)$

    $n_{stag} = n_{stag} + 1$

  end

  Estimate the noise variance of the objective and residual functions $\sigma_t, \Sigma_t = Varest(\mathscr{H}_t, R_t)$

  Calculate $\Delta M_t^s = M_t^s(x_t^*) - M_t^s(x_t^* + s_t)$

  Accept or reject the step and calculate a measure of progress $(x_{t+1}^*, \rho_t) = Accept(x_t^*, s_t, \Delta M_t^s, \sigma_t)$

  Simulate the effect of noise on ρ:
  $\rho^{noise} = \{\rho_1^{noise}, \dots \rho_b^{noise}\} = SimNoise(\mathscr{X}^{model}, \mathscr{R}^{model}, M_{t-1}^v, M_t^v, R_t, \Sigma_t)$

  Adjust the number of repeated function evaluations: $m_{t+1}^{model} = AdjustRep(\rho^{noise}, \rho_t, m_t^{model})$

  Adjust the trust-region radius: $\Delta_{t+1}^{region} = AdjustRadius(\Delta_t^{region}, \rho_t, s_t, m_t^{model}, m_{t+1}^{model})$

  if $x_{t+1}^* \neq x_t^*$ *and* $Converged(\mathscr{H}_t, M_t^s, x_t^*, x_{t+1}^*)$ **then**

    break

  end

end

The major changes appear when the original algorithm would have proceeded with the acceptance step. In the noisy case, we first estimate the variance of the noise term in the objective function (σ_t) as well as the variance-covariance matrix of the noise terms in the residual function (Σ_t). The actual implementation of the noise estimation is again a replaceable component, which is further described in Section 3.5.3.2.

While the expected improvement is calculated as before, the acceptance step now takes the estimated noise variance σ_t into account. Implementations of noise robust acceptance steps are described in Section 3.5.3.5.

After the acceptance step, two new steps are introduced. The first is to simulate our noisy measure of model quality ρ^{noise}, and the second is to adjust the number of repeated function evaluations m_t^{model} based on the simulated values. Both are replaceable components, which are further described in Section 3.5.3.3 and Section 3.5.3.4. Finally, the trust-region radius is adjusted as before. The only difference is that it now takes two additional arguments m_t^{model} and m_{t+1}^{model}. This can be used to skip radius decreases in situations where the number of repetitions was increased.

3.5.3.2 Estimation of noise variance

The literature on noisy optimization generally distinguishes between two types of noise: Additive noise is a noise term with fixed variance over the entire parameter space that is added to the objective function. Multiplicative noise is a noise term that enters the objective function as a multiplicative factor, and therefore, the effective variance of the noise varies with the value of the objective function. In least-squares optimization, the noise term is added to the residual function, and therefore, even additive noise leads to a noise term whose variance varies over the parameter space.

In *tranquilo*, we treat the noise term as constant over the current trust-region. This can be seen as a locally constant approximation to more general noise terms. We do not make any assumptions about how the noise term varies between trust-regions. We distinguish between Σ_t, the variance-covariance matrix of the noise terms in the residual function, and σ_t, the variance of the noise term in the objective function.

While we implement the noise estimation as a replaceable component, we provide just one implementation: We first calculate a search radius $\Delta_t^{varest} = \gamma^{varest} \Delta_t^{region}$ and scan the history for function evaluations within this radius of the current point. Out of these points, we only keep the ones at which the objective function was evaluated at least m^{varest} times. Next, we de-mean the function and residual evaluations at each point. Finally, we calculate σ_t as the variance of

the de-meaned function evaluations and Σ_t as the variance-covariance matrix of the de-meaned residual evaluations.

To make sure that at least one point exists in the current trust-region at which the function has been evaluated often enough to get a variance estimate, the minimal number of repeated function evaluations in the acceptance step m_{min}^{accept} needs to be set larger or equal to m_{min}^{varest}. In our benchmarks, we set $m_{min}^{varest} = 3$ and $m_{min}^{accept} = 4$.

3.5.3.3 Simulating ρ^{noise}

The goal of this step is to simulate multiple instances of measures of model quality $\rho^{noise} = \{\rho_1^{noise}, \ldots, \rho_b^{noise}\}$ that can be used to adjust m_t^{model}. In principle, there are many possibilities for doing this, which can range from heuristics to computationally costly simulation approaches. Currently, we implement just one approach, which is based on simulations.

The approach for generating ρ^{noise} is described in Algorithm 4. The inputs of the algorithm are the current set of model points $\mathcal{X}_t^{model}$, the current and previous vector models M_{t-1}^v and M_t^v, the current effective trust-region R_t, the estimated variance-covariance matrix of the noise terms in the residual function Σ_t, the current parameter vector x_t^*, the number of simulations b, and the number of repeated function evaluations m_t^{model}. The first few inputs provide almost all ingredients for a standard fitting step in *tranquilo*. The only thing that is missing are the residuals at the model points $\mathcal{R}_t^{model}$, because those will be replaced by simulated counterparts.

The simulation starts by calculating the "true" and noise-free residuals at the model points. They are denoted by $\mathcal{R}_{sim,true}^{model}$ and calculated by evaluating the current vector model M_t^v at the model points. These "true" residuals play the role of the unobservable $\mathbb{E}r(x, \xi)$ during the simulation.

For each $\ell = 1, \ldots, b$ simulation draw, we start by creating averages of simulated noisy residuals, denoted by $\overline{\mathcal{R}}_{sim,\ell}^{model}$. These play the role of the average observed residuals $\overline{\mathcal{R}}^{model}$ in the simulation, i.e., they will be used to fit vector models $M_\ell^{v,sim}$. To capture residualized model fitting, the previous vector model M_{t-1}^v is used inside each simulated fitting step. The simulated vector models are then aggregated into simulated scalar models $M_\ell^{s,sim}$ and used to solve the simulated trust-region subproblem. This yields a candidate step s_ℓ^{sim}.

Given these ingredients, we can calculate the simulated measure of model quality ρ_ℓ^{noise} as in Equation 3.5.3.3. Since the simulated ρ_ℓ^{noise} mimics all steps of the actual algorithm, it is a pure measure of the effect of random error on the model's ability to produce good candidate points. All other errors, such as approximation error, as well as imperfect solutions of the trust-region

subproblem or numerical imprecisions in the fitting process are reflected in the standard ρ but not in ρ^{noise}.

Algorithm 4: Simulating ρ^{noise}

Input: Model points $\mathcal{X}_t^{model}$, current and previous vector models M_{t-1}^v and M_t^v, the current effective trust-region R_t, The variance-covariance matrix of the noise terms in the residual function Σ_t, the current parameter vector x_t^*, the number of simulations b and the number of repeated function evaluations m_t^{model}.

1 Calculate "true" residuals $\mathcal{R}_{sim,true}^{model} = \{M_t^v(x) : x \in \mathcal{X}_t^{model}\}$
2 **for** $\ell = 1, \ldots, b$ **do**
3 Simulate average noisy residuals $\overline{\mathcal{R}}_{sim,\ell}^{model}$ over m_t^{model} simulated noisy residuals that are created by adding noise draws from $N(0, \Sigma_t)$ to "true" residuals in $\mathcal{R}_{sim,true}^{model}$
4 Fit a simulated vector model: $M_\ell^{v,sim} = Fit(\mathcal{X}_t^{model}, \overline{\mathcal{R}}_{sim,\ell}^{model}, M_{t-1}^v, R_t)$
5 Aggregate the simulated vector model: $M_\ell^{s,sim} = Aggregate(M_\ell^{v,sim})$
6 Solve the simulated trust-region subproblem: $s_\ell^{sim} = Subsolve(M_\ell^{s,sim}, R_t)$
7 Calculate the simulated measure of model quality: $\rho_\ell^{noise} = \frac{M_t^s(x_t^*) - M_t^s(x_t^* + s_\ell^{sim})}{M_\ell^{s,sim}(x_t^*) - M_\ell^{s,sim}(x_t^* + s_\ell^{sim})}$
8 **end**

In our practical implementation, we set $b = 100$. This means that simulating ρ^{noise} incurs the computational overhead of fitting, aggregating and minimizing 100 surrogate models. While this overhead is much larger than an iteration of a typical trust-region algorithm, it is justified in a setting with an expensive objective function.

3.5.3.4 Adjusting m^{model}

The goal of this step is to adjust the number of repeated function evaluations m_t^{model} based on the simulated ρ_ℓ^{noise} values. A simple possibility would be to calculate the average of the simulated ρ_ℓ^{noise} values and then adjust m_t^{model} in a very similar way to the adjustment of the trust-region radius. A drawback of this approach is that the denominator of the simulated ρ_ℓ^{noise} can be very small and therefore, a non-robust statistic like the average is strongly affected by a few outliers.

To avoid this problem, we choose an approach that is not based on the average but on the share of simulated ρ_ℓ^{noise} values below and above certain cutoffs. In particular, we use the following approach: ρ_{high}^{noise} and ρ_{low}^{noise} are cutoffs thate determine whether a simulated ρ_ℓ^{noise} is considered high

or low. If more than π^{high} of the simulated ρ_ℓ^{noise} are high, we conclude that m_t^{model} is unnecessarily large and decrease it by one in order to save costly function evaluations in the next iteration. If this is not the case but more than π^{low} of the simulated ρ_ℓ^{noise} are high or the overall ρ_t is larger than ρ_{keep}, we conclude that m_t^{model} is just right and leave it unchanged. Otherwise, we increase m_t^{model} by one.

To improve robustness, we also keep the number of repeated function evaluations between a lower and upper bound. m_{min}^{model} is the minimal number of function evaluations, which we set to 1 in our benchmarks. m_{max}^{model} is the maximal number of function evaluations, which we set to 30.

3.5.3.5 Noisy acceptance steps

The noisy acceptance step requires an additional input σ_t, over the noise-free acceptance step (Equation 3.3.9). It is a replaceable component, but we only provide one implementation based on the power analysis ideas described in Section 3.5.2.2.

$$(x_{t+1}^*, \rho_t) \;=\; Accept(x_t^*, s_t, \Delta M_t^s, \sigma_t)$$

As in the noise-free case, x_{t+1}^* is the candidate point for the next iteration, ρ_t is a measure of progress or model quality, ΔM_t^s is the expected improvement from taking step s_t, and now additionally, σ_t is the estimated variance of the noise term in the objective function.

In the noise-free case, we would, generally, accept the candidate step s_t if it yields any improvement over the current point. That is, if $f(x_t^* + s_t) < f(x_t^*)$. In the noisy case, we are ultimately interested in minimizing the $\mathbb{E}f$, and thus we would like to make the comparison $\mathbb{E}f(x_t^* + s_t, \xi) < \mathbb{E}f(x_t^*, \xi)$. However, this is, of course, not observed. Instead, we can try to reduce the effect of the noise by averaging multiple function evaluations at each point. Define the averages of the objective function at the candidate and current point as:

$$\bar{f}(x_t^* + s_t) \;=\; \frac{1}{m_{t2}^{accept}} \sum_{i=1}^{m_{t2}^{accept}} f(x_t^* + s_t, \xi_i) \quad \text{and} \quad \bar{f}(x_t^*) \;=\; \frac{1}{m_{t1}^{accept}} \sum_{j=1}^{m_{t1}^{accept}} f(x_t^*, \xi_j)$$

The noisy acceptance decision is then based on the following condition:

$$\text{Accept } s_t \quad \Longleftrightarrow \quad \bar{f}(x_t^* + s_t) \;<\; \bar{f}(x_t^*) \tag{3.5.2}$$

Similarly to the noise-free case in Equation 3.2.1, we can compute ρ_t by replacing f with $\bar{f}$:

$$\rho_t = \frac{\bar{f}(x_t^*) - \bar{f}(x_t^* + s_t)}{M_t^s(x_t^*) - M_t^s(x_t^* + s_t)}$$

While the mechanics of the noisy acceptance step are straightforward, as alluded to in the previous sections, the difficulty stems from determining m_{t1}^{accept} and m_{t2}^{accept} such that the decision based on the averages $\bar{f}$ is a good proxy for the decision based on expected values $\mathbb{E}f$.

One way to solve this problem is to use a two-sample power analysis. For this we need to assume that

1. $f(x_t^* + s_t, \xi_i)$ is independent of $f(x_t^*, \xi_j)$ for all i, j
2. $\frac{1}{m_{t1}^{accept}} \sum_{j=1}^{m_{t1}^{accept}} f(x_t^*, \xi_j) \approx N\left(\mathbb{E}f(x_t^*), \sigma^2 / m_{t1}^{accept}\right)$
3. $\frac{1}{m_{t2}^{accept}} \sum_{i=1}^{m_{t2}^{accept}} f(x_t^* + s_t, \xi_i) \approx N\left(\mathbb{E}f(x_t^* + s_t), \sigma^2 / m_{t2}^{accept}\right)$
4. σ_t is a reasonable estimate of σ

Given a significance level $\alpha \in (0, 1)$, a power level $1 - \beta \in (0, 1)$, and a minimal detectable effect size $M_t^s(x_t^*) - M_t^s(x_t^* + s_t)$, by choosing m_{t1}^{accept} and m_{t2}^{accept} under following condition, we can guarantee that the noisy acceptance condition (Equation 3.5.2) is done with a significance level of α and a power level of $1 - \beta$:

$$\frac{m_{t1}^{accept} m_{t2}^{accept}}{m_{t1}^{accept} + m_{t2}^{accept}} \geq \left[\frac{\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)}{(M_t^s(x_t^*) - M_t^s(x_t^* + s_t))/\sigma_t} \right]^2 \tag{3.5.3}$$

Since we still want to minimize the number of function evaluations, the actual choice of m_{t1}^{accept} and m_{t2}^{accept} is based on the following problem:

$$\underset{m_{t1}^{accept}, m_{t2}^{accept} \in \mathbb{N}}{\text{minimize}} \; m_{t1}^{accept} + m_{t2}^{accept} \text{ s.t. Equation 3.5.3 holds}$$

A detailed derivation of Equation 3.5.3 is provided in Appendix 3.B.

3.5.3.6 Noisy radius adjustment

The noise robust radius adjustment is identical to the noise-free version, except that radius decreases are skipped if m_{t+1}^{model} is larger than m_t^{model}. This successfully prevents the trust-region radius from collapsing to zero before a suitable value for m^{model} is found.

3.5.4 Benchmarking

To test the performance of *tranquilo* in a noisy setting, we use the bootstrapped version of the Moré-Wild benchmark set. Following Cartis, Fiala, et al. (2019), we add identical and independently normal-distributed noise terms to each residual. We choose a large standard deviation of 1.2 to create a challenging benchmark set (compared to 0.01 in Cartis, Fiala, et al. (2019)). Note that the scale of the residuals in the Moré-Wild benchmark varies drastically across problems. This means that even though we add the same amount of noise to each residual, we obtain problems with very different difficulties and with very different optimal sequences of sample sizes.

Since *tranquilo* is fully adaptive, we only run it in one configuration. For *DFO-LS*, we choose configurations with three different sample sizes. Note that in *DFO-LS*, $m^{model} = m^{accept}$. Since it is very hard to pick optimally increasing sequences of sample sizes in practice, we restrict ourselves to fixed sequences of 3, 5, and 10 function evaluations at each point.

Since we are interested in minimizing the expected value of our objective functions, the convergence test is based on evaluating the noise-free objective function at the parameter vectors generated by the algorithm. Then the convergence test is as before, but the tolerance level τ is relaxed to 0.1 to reflect that we cannot expect the same precision for noisy and noise-free optimization problems. The results of the benchmark are shown in Figure 3.5.2.

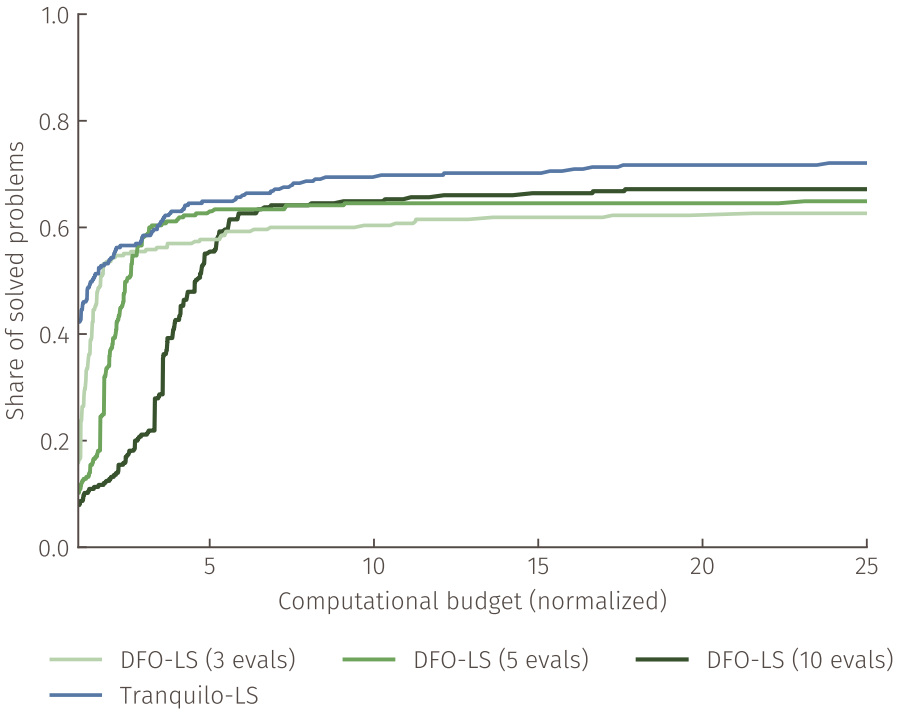


Figure 3.5.2. Comparison of least-squares optimizers on an augmented Moré-Wild benchmark set with added noise. The noise is normally distributed with a standard deviation of 1.2. The *x*-axis shows the normalized computational budget. The computational budget is measured in terms of objective function evaluations needed by the optimizers. Normalized means that the number of function evaluations each algorithm needed to solve a given problem is divided by the number of function evaluations the fastest algorithm needed to solve that problem. The different *DFO-LS* configurations vary in the number of repeated function evaluations at each point. *tranquilo* is fully adaptive and therefore does not need multiple configurations. The plot shows that *tranquilo* outperforms the *DFO-LS* configurations in speed and robustness.

We see that *DFO-LS* with three evaluations solves some problems very quickly but then stagnates abruptly. Using 5 evaluations at each point makes the algorithm slower but helps to solve more problems. The pattern repeats for 10 evaluations, even though only a few additional problems are solved by switching from 5 to 10 evaluations. This shows that it is very hard to pick a sample size that works well for several problems, and in fact, the sample sizes 3, 5, and 10 are already the result of some trial and error in which the whole benchmark set was solved multiple times.

Tranquilo starts at a low sample size and can, therefore, solve easy problems very quickly. If necessary, the sample size is increased, and therefore, *tranquilo* solves more problems than any configuration of *DFO-LS*. In total, *tranquilo* is the fastest algorithm for more than 40 % of the

problems. Moreover, its performance-profile is consistently above the performance profiles of all *DFO-LS* configurations.

3.6 conclusion

This paper presents the *tranquilo* algorithm, an optimizer for noisy nonlinear least-squares problems with expensive objective functions. A typical situation in which such problems arise is the estimation of econometric models using the method of simulated moments (MSM). *Tranquilo* improves over existing least-squares optimizers in two important ways: By introducing a line-search and speculative sampling approach, the algorithm becomes more parallelizable and the solution can be accelerated if multi-core machines are available. By introducing novel approaches for adaptive noise handling, the algorithm can solve noisy optimization problems without requiring the user to set any advanced algorithm parameters.

We show that in a noise-free and serial setting, tranquilo is roughly competitive with other state-of-the-art optimizers. The parallel version of *tranquilo* is much faster that the serial version. For noisy objective functions, tranquilo outperforms existing optimizers.

Appendix 3.A Notation

Table 3.A.1. Algorithm constants

Symbol	Description
$p \in \mathbb{N}$	Number of parameters in the optimization problem
$k \in \mathbb{N}$	Number of least-squares residuals. 1 for scalar problems
$n^{target} \in \mathbb{N}$	Target for the number of points used to construct surrogate models. Independent of the number of evaluations at each point in the noisy case. Usually $p + 1$ for least-squares optimizers
$n^{filter} \in \mathbb{N}$	Maximum number of points that remain after filtering. Typically larger than n^{target}
$l \in \mathbb{R}^p \cup -\infty$	Lower bounds for parameters
$u \in \mathbb{R}^p \cup \infty$	Upper bounds for parameters
$\gamma^{search} \in \mathbb{R}^+$	Search radius factor, usually ≥ 1
$s^{min} \in \mathbb{R}^+$	Minimum step size
$n_{stag}^{max} \in \mathbb{N}$	Maximum number of trials to avoid stagnation
$n_{stag}^{drop} \in \mathbb{N}$	Sample increment
$t^{max} \in \mathbb{N}$	Maximum number of iterations
$d^s \in \mathbb{N}$	Number of free coefficients of a scalar surrogate model
$d^v \in \mathbb{N}$	Number of free coefficients of each individual model in a vector surrogate model
$n^{cores} \in \mathbb{N}$	Number of available cores
$n^{batch} \in \mathbb{N}$	Batch size

Appendix 3.B Power Analysis

In this section, we derive the optimization problem we solve in the noisy acceptance step to determine the optimal number of objective function evaluations. For a more detailed discussion of power analysis, we refer to Montgomery (2008) or Cohen (1988).

3.B.1 Statistical Motivation

Suppose we observe two samples $\{y_1^{(1)}, \ldots, y_{n_1}^{(1)}\}$ and $\{y_1^{(2)}, \ldots, y_{n_2}^{(2)}\}$. The first sample has a mean of $\mathbb{E}[y_i^{(1)}] = \mu_1$ and the second a mean of $\mathbb{E}[y_i^{(2)}] = \mu_2$. We assume that both samples are independent of another and that the variance of both samples is σ^2. Assume further that the sample averages can be approximated by normal distributions, i.e.,

$$\frac{1}{n_1}\sum_{i=1}^{n_1} y_i^{(1)} \approx N(\mu_1, \sigma^2/n_1) \quad \text{and} \quad \frac{1}{n_2}\sum_{i=1}^{n_2} y_i^{(2)} \approx N(\mu_2, \sigma^2/n_2)$$

Note that this can be justified by distribution assumptions on the $y_i^{(k)}$ or by asymptotic arguments. In particular, we do not want to assume that the variables in a sample are independent, i.e., $y_i^{(k)}$ and $y_j^{(k)}$ may be dependent.

Our target parameter is $\Delta := \mu_1 - \mu_2$, and our goal is to test whether this parameter is greater than zero: $\Delta > 0$. The key assumption underlying power analysis is that we can choose the values of n_1 and n_2.

For this, we first need to select a statistical significance level $\alpha \in (0,1)$, a power level $1-\beta \in (0,1)$, and a minimal detectable effect size Δ_{min}. The formal test is then

$$H_0 \ : \ \Delta \ = \ 0 \text{ v.s. } H_1 \ : \ \Delta \ = \ \Delta_{min}$$

Define the estimators $\hat{\mu}_1 := \frac{1}{n_1}\sum_{i=1}^{n_1} y_i^{(1)}$ and $\hat{\mu}_2 := \frac{1}{n_2}\sum_{i=1}^{n_2} y_i^{(2)}$ of μ_1 and μ_2, respectively, and the estimator $\hat{\Delta} = \hat{\Delta}(n_1, n_2) := \hat{\mu}_1 - \hat{\mu}_2$. Under the normality assumption stated above, we get

$$\hat{\Delta}(n_1, n_2) \ \approx \ N(\Delta, \sigma_\Delta^2)$$

With $\sigma_\Delta^2 = \sigma^2 \frac{n_1+n_2}{n_1 n_2}$. Define the t-test statistic $\hat{t} = \hat{\Delta}/\sigma_\Delta$.

Under the null hypothesis H_0, we find $\hat{t} \approx N(0,1)$, so that we can choose the critical value, i.e., the value such that the Type-1 error is α, as $\Phi^{-1}(1-\alpha)$. Note furthter that under the alternative hypothesis H_1, we have $\hat{t} - \Delta_{min}/\sigma_\Delta \approx N(0,1)$.

We also require that the Type-2 error is at most β, i.e., we want that the probability of accepting H_0, when H_1 is true, is at most β. More formally,

$$\begin{aligned}
\beta \ &\geq \ \mathbb{P}[\hat{t} \ \leq \ z_{1-\alpha} | H_2] \\
&= \ \mathbb{P}[\hat{t} - \Delta_{min}/\sigma_\Delta \ \leq \ \Phi^{-1}(1-\alpha) - \Delta_{min}/\sigma_\Delta | H_2] \\
&\approx \ \Phi(\Phi^{-1}(1-\alpha) - \Delta_{min}/\sigma_\Delta)
\end{aligned}$$

And so,

$$\begin{aligned}
\Phi^{-1}(\beta) \ &\gtrsim \ \Phi^{-1}(1-\alpha) - \Delta_{min}/\sigma_\Delta \\
&= \ \Phi^{-1}(1-\alpha) - \Delta_{min}/\sigma \sqrt{\frac{n_1 n_2}{n_1 + n_2}}
\end{aligned}$$

Rearranging the previous equation then gives

$$\frac{n_1 n_2}{n_1 + n_2} \geq \left[\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)}{\Delta_{min}/\sigma} \right]^2 \tag{3.B.1}$$

3.B.2 Optimal sample sizes

Given the condition in Equation 3.B.1, we want to minimize the total number of samples $n_1 + n_2$. However, the exact problem we face in the noisy acceptance step (Section 3.5.3.5) in *tranquilo* may slightly differ from Equation 3.B.1, as there may already exist previous samples n_1^{exist} and n_2^{exist}.

In this case, we solve the following problem

$$\underset{n_1, n_2 \in \mathbb{N}}{\text{minimize}} \, n_1 + n_2 \text{ s.t } \frac{(n_1 + n_1^{exist})(n_2 + n_2^{exist})}{n_1 + n_1^{exist} + n_2 + n_2^{exist}} \geq \left[\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)}{\Delta_{min}/\sigma} \right]^2$$

Appendix 3.C Subsolvers

3.C.1 GQTPAR

In the SP-Ball case, *GQTPAR* finds an exact solution to the trust-region subproblem (Moré and Sorensen (1983)), which satisfies

$$(H + \lambda I)\, s^* = -g \tag{3.C.1}$$

where g is the model gradient, H denotes the model Hessian, and I is the identity matrix. *GQTPAR* determines the Lagrange multiplier $\lambda \geq 0$ such that the matrix $(H + \lambda I)$ is positive definite and $\lambda(1 - \|s^*\|) = 0$. The latter is a complementary slackness condition which states that at least one of the quantities λ and $(1 - \|s^*\|)$ must be zero at the optimum s^*. Recall that in the problem SP-Ball, the subspace B is a ball with a center of 0 and a radius of 1, defined as $B := \{s \in \mathbb{R}^p : \|s\| \leq 1\}$. When the solution s^* is interior to B, i.e. $\|s^*\| < 1$, then $\lambda = 0$ and *GQTPAR* terminates immediately. Otherwise, when s^* lies on the boundary of B, i.e. $\|s^*\| = 1$, $\lambda > 0$, and Newton's method is applied to find the value of λ such that $\|s^*\| = 1$ is satisfied. Rearranging

Equation 3.C.1, Moré and Sorensen (1983) show that the unique solution s^*, which depends on λ, is defined as

$$s^*(\lambda) \;=\; -(H + \lambda I)^{-1}\, g \tag{3.C.2}$$

for $\lambda > 0$ sufficiently large so that $(H + \lambda I)$ is positive definite and $\|s(\lambda)\| = 1$. To obtain s^*, one first needs an expression for λ. Starting with an initial guess, *GQTPAR* updates λ in each iteration ℓ of Algorithm 5 via

$$\lambda^{(\ell+1)} \;=\; \lambda^{(\ell)} - \frac{\phi(\lambda^{(\ell)})}{\phi'(\lambda^{(\ell)})} \tag{3.C.3}$$

where the function $\phi(\lambda^{(\ell)})$ is defined as:

$$\phi(\lambda^{(\ell)}) \;=\; \frac{1}{\|s(\lambda^{(\ell)})\| - 1} \tag{3.C.4}$$

and $\phi'(\lambda^{(\ell)})$ is the first derivative of $\phi(\lambda^{(\ell)})$ with respect to $\lambda^{(\ell)}$. *GQTPAR* finds the optimal $\lambda^{(\ell)}$ by applying Newton's root finding method to the function $\phi(\lambda^{(\ell)})$ in Equation 3.C.4, which is almost linear around the optimal $\lambda^{(\ell)}$ (Nocedal and Wright (2006)). Before updating $\lambda^{(\ell)}$, however, an expression for $s(\lambda^{(\ell)})$ satisfying Equation 3.C.2 is needed. *GQTPAR* obtains a candidate $s(\lambda^{(\ell)})$ by factorizing the model Hessian H via Cholesky factorization and solving the resulting linear system (see lines 4 and 5 of Algorithm 5). Conn, Gould, and Toint (2000) show that Equation 3.C.3 can be simplified to the updating formula in line 11 of Algorithm 5, which makes the dependence on s_ℓ apparent. The vector q_ℓ is the solution to the linear system in line 10 of Algorithm 5, where R denotes the upper triangular matrix of H. The details are omitted here for brevity but are available in Conn, Gould, and Toint (2000). With expressions for s_ℓ and q_ℓ in hand, *GQTPAR* calculates their respective norms $\|s_\ell\|$ and $\|q_\ell\|$, and updates $\lambda^{(\ell+1)}$ in line 11 of Algorithm 5.

Note that in line 2 of Algorithm 5, $\lambda^{(\ell)}$ is safeguarded. This is necessary to ensure that the matrix $(H + \lambda^{(\ell)})I$ is positive definite and its Cholesky factorization exists. For details on the safeguarding procedure, see Moré and Sorensen (1983) or Nocedal and Wright (2006). This concludes Algorithm 5 for the "easy case".

There may be situations, where $(H + \lambda^{(\ell)}I)$ is positive definite but the solution s^* to Equation 3.C.1 is not unique. This is what Moré and Sorensen (1983) call the "hard case", which is not

Algorithm 5: GQTPAR algorithm - The "easy case"

Input: Initial guess s_0, $\lambda^{(0)}$, $\lambda_L^{(0)}$, $\lambda_U^{(0)}$

for ℓ=0,1,2,... **do**
  Safeguard $\lambda^{(l)}$ to obtain $\lambda_S^{(\ell)}$
  if $H + \lambda^{(\ell)} I$ *is positive definite* **then**
    Factor $H + \lambda^{(l)} I = R^T R$
    Solve $s_\ell = -(R^T R)^{-1} g$
  end
  Update $\lambda_L^{(\ell)}$, $\lambda_U^{(\ell)}$, $\lambda_S^{(\ell)}$
  Check convergence criteria
  if $H + \lambda^{(l)} I$ *is positive definite and* $g \neq 0$ **then**
    Solve $q_\ell = (R^T)^{-1} s_\ell$
    Set $\lambda^{(\ell+1)} = \lambda^{(\ell)} + \left(\frac{\|s_\ell\|}{\|q_\ell\|}\right)^2 \left(\|s_\ell\| - 1\right)$
  else
    Set $\lambda^{(\ell+1)} = \lambda_S^{(\ell)}$
end

described in Algorithm 5. We refer the interested reader to Moré and Sorensen (1983) and Conn, Gould, and Toint (2000) for details. In short, in the "hard case", Equation 3.C.1 is replaced by

$$\left(H - \lambda_1 I\right)(s^* + \tau z) \;=\; -g \tag{3.C.5}$$

where z is the eigenvector of the model Hessian H corresponding to the first eigenvalue λ_1 of H. Moreover, z is such that $\|s + \tau z\,(\lambda)\| = 1$ for some τ.

3.C.2 BNTR

BNTR stands for "Bounded Newton Trust-Region" algorithm and was developed for the Toolkit of Advanced Optimization (Dener et al. (2021)). It employs a trust-region-like approach combined with an active set method to solve the bound-constrained problem SP-Cube. A search direction of the candidate step is considered "active" if it lies at the boundary of the trust-region. The active and inactive sets are defined as follows (Bertsekas (1982))

$$
\begin{aligned}
\text{lower bounded: } \mathscr{L}(s) &= \{i : s_i \leq l_i \wedge g(s)_i > 0\},\\
\text{upper bounded: } \mathscr{U}(s) &= \{i : s_i \geq u_i \wedge g(s)_i < 0\},\\
\text{fixed: } \mathscr{F}(s) &= \{i : l_i = u_i\},\\
\text{active-set: } \mathscr{A}(s) &= \mathscr{L}(s) \cup \mathscr{U}(s) \cup \mathscr{F}(s),\\
\text{inactive-set: } \mathscr{I}(s) &= \{1, 2, \ldots, n\} \setminus \mathscr{A}(s).
\end{aligned}
$$

where l_i and u_i are the lower and upper bound on the ith search direction in s, respectively. Instead of fitting a full surrogate model, *BNTR* uses a simplified quadratic model in the surrogate step in line 5 of Algorithm 6. In particular, it solves for the minimizer r_ℓ of the reduced quadratic model $M_\ell^r(r)$ for the unconstrained (i.e. inactive) search directions only via a conjugate gradient method. The available methods are *Conjugate Gradient*, *Steihaug-Toint*, and *TRSBOX*. The reduced model is formed using the reduced model gradient g^r and the reduced model Hessian H^r based on the inactive set $\mathscr{I}(s)$, i.e. the unbounded search directions.

With the reduced conjugate gradient step r_ℓ in hand, *BNTR* constructs a new candidate step p_ℓ (line 6 of Algorithm 6). It does so by projecting r_ℓ onto the lower and upper bounds of the active set $\mathscr{A}(s)$

$$
p = \begin{cases} l_i & \text{if } s_i < l_i \\ u_i & \text{if } s_i > u_i \\ r_i & \text{otherwise} \end{cases} \qquad \text{(bound-projection)}
$$

where the subscript ℓ is omitted for readability. Similar to other trust-region optimizers, acceptance of the candidate step p_ℓ is determined based on the ratio of the actual over expected improvement of the surrogate model (see line 7 of Algorithm 6). The actual improvement is defined as the difference between the surrogate model evaluated at the candidate $s_\ell + p_\ell$ and the surrogate model evaluated at the current iterate s_ℓ. The expected improvement is defined as the value of the reduced surrogate model evaluated at r_ℓ. If the ratio is larger than a threshold, the candidate step p_ℓ is accepted and the trust-region radius is increased. Else, the candidate is rejected and the trust-region radius is decreased. The process is repeated until convergence criteria are met.

Algorithm 6: BNTR algorithm

Input: Initial guess s_0, $\Delta_0^{sub} > 0$

Take a finite number of gradient descent steps and update s_0, Δ_0^{sub}

for $\ell = 0,1,2,\ldots$ **do**

  Create active set of bounds $\mathcal{A}(s)$ and set of inactive directions $\mathcal{I}(s)$

  Construct reduced model gradient and reduced model Hessian based on $\mathcal{I}(s)$

  Solve for the optimum of the reduced model M_ℓ^r: $r \approx arg\min_r M_\ell^r(r)$ s.t. $r \leq \Delta_\ell^{sub}$

  Construct new candidate step p_ℓ by projecting r_ℓ onto $\mathcal{A}(s)$

  Calculate $\kappa_\ell = \frac{M_\ell^s(s_\ell + p_\ell) - M_\ell^s(s_\ell)}{M_\ell^r(r_\ell)}$, the ratio of actual over expected improvement

  if κ_ℓ *is larger than a threshold* **then**

    Accept step $s_{\ell+1} = s_\ell + p_\ell$

    Expand trust-region radius: $\Delta_{\ell+1}^{sub} = \alpha^{inc}\Delta_\ell^{sub}$

  else

    Reject step $s_{\ell+1} = s_\ell$

    Shrink trust-region radius $\Delta_{\ell+1}^{sub} = \alpha^{dec}\Delta_\ell^{sub}$

  Check convergence criteria

end

References

Arnoud, Antoine, Fatih Guvenen, and Tatjana Kleineberg. 2019. "Benchmarking Global Optimizers." NBER Working Papers 26340. National Bureau of Economic Research, Inc. Tiktak. [137]

Augustin, F., and Youssef Marzouk. 2017. "A trust-region method for derivative-free nonlinear constrained stochastic optimization." (03): [141]

Berndt, Ernst R., Bronwyn Hall, Robert Hall, and Jerry Hausman. 1974. "Estimation and Inference in Nonlinear Structural Models." In *Annals of Economic and Social Measurement, Volume 3, number 4*. National Bureau of Economic Research, Inc, 653–65. [147]

Bertsekas, Dimitri P. 1982. "Projected Newton Methods for Optimization Problems with Simple Constraints." *SIAM Journal on Control and Optimization* 20 (2): 221–46. [191]

Briani, Matteo, Alvise Sommariva, and Marco Vianello. 2012. "Computing Fekete and Lebesgue points: Simplex, square, disk." *Journal of Computational and Applied Mathematics* 236 (9): 2477–86. [152]

Cartis, Coralia, Jan Fiala, Benjamin Marteau, and Lindon Roberts. 2019. "Improving the Flexibility and Robustness of Model-Based Derivative-Free Optimization Solvers." *ACM Trans. Math. Softw.* 45 (3): [132, 134–136, 139, 141, 146, 147, 159, 162, 164, 184]

Cartis, Coralia, and Lindon Roberts. 2019. "A derivative-free Gauss–Newton method." *Mathematical Programming Computation* 11 (4): 631–74. [139, 162]

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates. [187]

Conn, Andrew R., Nicholas I. M. Gould, and Philippe L. Toint. 2000. *Trust Region Methods*. Society for Industrial, and Applied Mathematics. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9780898719857. [134, 135, 138–140, 152, 190]

Dener, Alp, Adam Denchfield, Hansol Suh, Todd Munson, Jason Sarich, Stefan Wild, Steven Benson, and Lois Curfman McInnes. 2021. "TAO Users Manual (Rev. 3.15)." (3): [141, 160, 161, 191]

Dolan, Elizabeth D., and Jorge J. Moré. 2002. "Benchmarking optimization software with performance profiles." *Mathematical Programming* 91: 201–13. [161]

Eisenhauer, Philipp, James J. Heckman, and Stefano Mosso. 2015. "ESTIMATION OF DYNAMIC DISCRETE CHOICE MODELS BY MAXIMUM LIKELIHOOD AND THE SIMULATED METHOD OF MOMENTS." *International Economic Review* 56 (2): 331–57. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/iere.12107. [131]

Gabler, Janoś. 2022. "A Python Tool for the Estimation of large scale scientific models." [135]

Gabler, Janoś, Sebastian Gsell, Tim Mensinger, and Mariam Petrosyan. 2024. "Tranquilo." [135]

Gabler, Janoś, Tobias Raabe, Klara Röhrl, and Hans-Martin von Gaudecker. 2022. "The effectiveness of testing, vaccinations and contact restrictions for containing the CoViD-19 pandemic." en. *Sci. Rep.* 12 (1): 8048. [134]

Gould, Nicholas I. M., Dominique Orban, and Philippe L. Toint. 2003. "CUTEr and SifDec: A Constrained and Unconstrained Testing Environment, Revisited." *ACM Trans. Math. Softw.* 29 (4): 373–94. [161]

Larson, Jeffrey, Matt Menickelly, and Stefan M. Wild. 2019. "Derivative-free optimization methods." *Acta Numerica* 28 (5): 287–404. [138–140, 145, 155]

Lee, Donghoon, and Matthew Wiswall. 2007. "A Parallel Implementation of the Simplex Function Minimization Routine." *Computational Economics* 30 (02): 171–87. [133, 142]

Levenberg, Kenneth. 1944. "A Method for the Solution of Certain Non-Linear Problems in Least Squares." *Quarterly of Applied Mathematics* 2 (2): 164–68. [132]

Marquardt, Donald W. 1963. "An Algorithm for Least-Squares Estimation of Nonlinear Parameters." *Journal of the Society for Industrial and Applied Mathematics* 11 (2): 431–41. [132]

Montgomery, D.C. 2008. *Design and Analysis of Experiments*. Student solutions manual. John Wiley & Sons. [187]

Moré, Jorge J., and Danny C. Sorensen. 1983. "Computing a Trust Region Step." *Siam Journal on Scientific and Statistical Computing* 4: 553–72. [189, 190]

Moré, Jorge J., and Stefan M. Wild. 2009. "Benchmarking Derivative-Free Optimization Algorithms." *SIAM Journal on Optimization* 20 (1): 172–91. [136, 161–163, 170]

Nocedal, Jorge, and Stephen Wright. 2006. *Numerical optimization*. Springer Science & Business Media. [138, 190]

Powell, M. 2009. "The BOBYQA Algorithm for Bound Constrained Optimization without Derivatives." Working paper DAMTP 2009/NA06. Centre for Mathematical Sciences, University of Cambridge. [135, 139, 140, 145–147, 155]

Powell, M. J. D. 2006. "The NEWUOA software for unconstrained optimization without derivatives." In. *Large-Scale Nonlinear Optimization*. Edited by G. Di Pillo and M. Roma. Boston, MA: Springer US, 255–97. [139, 140, 155]

Pukelsheim, Friedrich. 2006. *Optimal Design of Experiments (Classics in Applied Mathematics) (Classics in Applied Mathematics, 50)*. USA: Society for Industrial, and Applied Mathematics. [151]

Shashaani, Sara, Fatemeh S. Hashemi, and Raghu Pasupathy. 2018. "ASTRO-DF: A Class of Adaptive Sampling Trust-Region Algorithms for Derivative-Free Stochastic Optimization." *SIAM Journal on Optimization* 28 (4): 3145–76. eprint: https://doi.org/10.1137/15M1042425. [141]

Wild, Stefan M. 2008. "MNH: A Derivative-Free Optimization Algorithm Using Minimal Norm Hessians." In *Tenth Copper Mountain Conference on Iterative Methods*. [155]

Wild, Stefan M. 2017. "Solving Derivative-Free Nonlinear Least Squares Problems with POUNDERS." In *Advances and Trends in Optimization with Engineering Applications*. Edited by Tamas Terlaky, Miguel F. Anjos, and Shabbir Ahmed. SIAM, 529–40. [132, 135, 140, 141, 145–148, 155, 158–160, 162]

Winfield, D. 1973. "Function Minimization by Interpolation in a Data Table." *IMA Journal of Applied Mathematics* 12 (3): 339–47. eprint: https://academic.oup.com/imamat/article-pdf/12/3/339/1918553/12-3-339.pdf. [140]

Zhang, Hongchao, Andrew R. Conn, and Katya Scheinberg. 2010. "A Derivative-Free Algorithm for Least-Squares Minimization." *SIAM Journal on Optimization* 20 (6): 3555–76. [140, 147, 157]

Table 3.A.2. Component specific constants

Symbol	Description
AdjustRadius	
$\rho^{inc} \in \mathbb{R}^+$	Radius shrinking cutoff
$\rho^{dec} \in \mathbb{R}^+$	Radius expansion cutoff
$\gamma^{inc} \in \mathbb{R}^+$	Radius expansion factor
$\gamma^{dec} \in \mathbb{R}^+$	Radius shrinking factor
$\Delta^{max} \in \mathbb{R}^+$	Radius bound
$c^{ls} \in \mathbb{R}^+$	Large radius cut-off
Converge	
$\varepsilon^{fatol} \in \mathbb{R}^+$	Convergence absolute tolerance objective function
$\varepsilon^{frtol} \in \mathbb{R}^+$	Convergence relative tolerance objective function
$\varepsilon^{gatol} \in \mathbb{R}^+$	Convergence absolute tolerance surrogate model gradient
$\varepsilon^{grtol} \in \mathbb{R}^+$	Convergence relative tolerance surrogate model gradient
$\varepsilon^{xatol} \in \mathbb{R}^+$	Convergence absolute tolerance parameters
$\varepsilon^{xrtol} \in \mathbb{R}^+$	Convergence relative tolerance parameters
Varest	
γ^{varest}	Factor to calculate a search radius for points used for noise variance estimation
m_{min}^{varest}	Minimal number of function evaluations required to use a point for variance estimation
Accept	
m_{min}^{accept}	Minimal number of repeated function evaluations for acceptance steps
m_{max}^{accept}	Maximal number of repeated function evaluations for acceptance steps
SimulateNoise	
b	Number of simulation runs for the calculation of ρ^{noise}
AdjustRep	
m_0	Number of repeated function evaluations at start parameters
ρ_{high}^{noise}	Threhshold for a simulated ρ^{noise} to be considered high
ρ_{low}^{noise}	Threhshold for a simulated ρ^{noise} to be considered low
π^{high}	Minimal share of high ρ^{noise}-estimates required to decrease m_t^{model}
π^{low}	Minimal share of low ρ^{noise}-estimates required to increase m_t^{model}
ρ_{keep}	Threshold for ρ_t to be considered good enough that m does not have to be increased. This refers to the overall ρ, not ρ^{noise}
m_{min}^{model}	Minimal number of repeated function evaluations for surrogate model construction
m_{max}^{model}	Maximal number of repeated function evaluations for surrogate model construction

Table 3.A.3. Internal algorithm variables

Symbol	Description
$x_t^* \in \mathbb{R}^p$	Accepted parameter vector at the beginning of iteration t. Also serves as trustregion center
Δ_t^{region}	Trust region radius in iteration t
Δ_t^{search}	Search radius, defined as $\gamma^{search}\Delta_t^{region}$
R_t	Effective trustregion in iteration t. If no bounds are binding, R_t is defined as a ball with center x_t^* and radius Δ_t^{region}. If bounds are binding, R_t is defined as a hypercube with the same volume as a ball with radius Δ_t^{region} that contains x_t^* and respects bound constraints
$\mathcal{H}_t$	History of function evaluations and parameter vectors up to period t
$\mathcal{X}_t^{existing} \subset \mathbb{R}^p$	Points inside the search radius for which the function has previously been evaluated
$\mathcal{X}_t^{filtered} \subset \mathbb{R}^p$	Filtered existing points
$\mathcal{X}_t^{new}$	Newly sampled points in iteration t. Defined as $\mathcal{X}^{filtered} \cup \mathcal{X}_t^{new}$
$\mathcal{X}_t^{model}$	Model points
$\mathcal{R}_t^{existing}, \mathcal{R}_t^{filtered}, \ldots$	Least-squares residuals evaluated on the corresponding set of points
$\mathcal{F}_t^{existing}, \mathcal{F}_t^{filtered}, \ldots$	Objective function evaluations on the corresponding set of points
$M_t^s \in \mathcal{M} = \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^{p \times p}$	Scalar quadratic model defined by an intercept, gradient-terms and hessian-terms
$M_t^v \in \mathcal{M}^k$	Vector model consisting of one scalar model per least-squares residual
$\Delta f_t \equiv f(x_t^*) - f(x_t^* + s_t)$	Actual improvement through step s_t
$\Delta M_t^s \equiv M_t^s(x_t^*) - M_t^s(x_t^* + s_t)$	Expected improvement through step s_t
$\overline{\mathcal{R}}_t^{existing}, \overline{\mathcal{R}}_t^{filtered}, \ldots$	Averaged Least-squares residuals evaluated on the corresponding set of points
$\overline{\mathcal{F}}_t^{existing}, \overline{\mathcal{F}}_t^{filtered}, \ldots$	Averaged function evaluations on the corresponding set of points
σ_t	Estimate of the noise variance in the objective function in iteration t
Σ_t	Estimate of the noise variance-covariance matrix in the residual function in iteration t
m_t^{model}	Number of repeated function evaluations for surrogate model construction in iteration t
m_{t1}^{accept} and m_{t2}^{accept}	Number of repeated function evaluations at x_t^* and $x_t^* + s_t$ for the acceptance step in iteration t. These are actually component variables of acceptance steps but listed here because all noise-robust acceptance steps use these variables

Table 3.A.4. Component functions

Symbol	Description
$\mathcal{X}_t^{filtered} = Filter(\mathcal{X}_t^{existing})$	Filter applied to sample of existing points
$\mathcal{X}_t^{new} = Sample(\mathcal{X}_t^{filtered}, R_t, n^{target})$	Sample new points inside the trustregion
$M_t^v = Fit(\mathcal{X}_t^{model}, \mathcal{R}_t^{model}, M_{t-1}^v, R_t)$	Fit a quadratic model scaled to the trustregion
$M_t^s = Aggregate(M_t^v)$	Aggregate vector model into scalar model
$s_t = Subsolve(M_t^s, R_t)$	Solve the trustregion subproblem
$\mathcal{X}_t^{reduced} = Drop(\mathcal{X}_t^{model}, n^{drop}, \Delta_t^{region})$	Drop the n^{drop} worst points
$(x_{t+1}^*, \rho_t) = Accept(x_t^*, s_t, \Delta M_t^s, \sigma_t)$	Accept or reject the proposed step and calculate a measure of progress. The argument σ_t is only used in the noisy case
$\Delta_{t+1}^{region} = AdjustRadius(\Delta_t^{region}, \rho_t, s_t)$	Adjust the trustregion radius
$Converged(\mathcal{H}_t, M_t^s, x_t^*, x_{t+1}^*)$	Check for convergence
$\sigma_t, \Sigma_t = Varest(\mathcal{H}_t, R_t)$	Estimate the noise variance
$\rho^{noise} = SimNoise(\mathcal{X}^{model}, \mathcal{R}^{model}, M_{t-1}^v, M_t^v, R_t, \Sigma_t)$	Simulate ρ that would obtain in the absence of Approximation error due to noise. $\rho^{noise} = (\rho_1^{noise}, \ldots, \rho_b^{noise})$ is a vector of simulated ρ's.
$m_{t+1}^{model} = AdjustRep(\rho^{noise}, \rho_t, m_t^{model})$	Adjust the number of repeated function evaluations for surrogate model construction

Table 3.A.5. Mathematical symbols

Symbol	Description
$\|\|x\|\|$	The Euclidean norm of a vector x
$N(\mu, \Sigma)$	A (multivariate) Normal distribution with mean μ and variance-covariance (matrix) Σ
$\Phi(x)$	The cumulative distribution function of the standard Normal distribution evaluated at x
$\lceil x \rceil$	The smallest integer greater than or equal to x