# Essays in Microeconometrics

Dissertation

zur Erlangung des akademischen Grades

doctorum rerum politicarum

(Doktor der Wirtschaftswissenschaft)

eingereicht an der

Wirtschaftswissenschaftlichen Fakultät

der Humboldt-Universität zu Berlin

von

**M.Sc. Stephan Martin**

Präsidentin der Humboldt Universität zu Berlin:
Prof. Dr. Julia von Blumenthal

Dekan der Wirtschaftswissenschaftlichen Fakultät:
Prof. Dr. Daniel Klapper

Gutachterinnen/Gutachter:
Prof. Dr. Christoph Breunig
Prof. Dr. Sonja Greven

Tag des Kolloquiums: 29.06.2023

# Abstract

This dissertation comprises three individual papers on various topics in microeconometrics, which is the study of econometric theory in the context of problems arising from e.g. the analysis of cross-sectional data.

In the first chapter, which is joint work with Christoph Breunig, we study a semi-/nonparametric regression model with a general form of nonclassical measurement error in the outcome variable. We show equivalence of this model to a generalized regression model and provide conditions under which the regression function is identifiable under appropriate normalizations. We propose a novel sieve rank estimator for the regression function and establish its rate of convergence. We find that our estimator corrects for biases induced by nonclassical measurement error in Monte Carlo simulations and an empirical application on belief formation of stock market expectations with survey data from the German Socio-Economic Panel (SOEP).

The second chapter deals with the estimation of conditional random coefficient models. Here I propose a two-stage sieve estimation procedure. First, a closed-form sieve approximation of the conditional RC density is derived where each sieve coefficient can be expressed as conditional expectation function varying with controls. Second, sieve coefficients are estimated with generic machine learning procedures and under appropriate sample splitting rules. I derive the $L_2$-convergence rate of the conditional RC-density estimator and also provide a result on pointwise asymptotic normality. The performance and applicability of the estimator is illustrated using random forest algorithms over a range of Monte Carlo simulations and in an empirical application studying behavioral heterogeneity in an economic experiment on portfolio choice.

The third chapter presents a novel and simple approach to estimating a class of semi(non)parametric discrete choice models imposing shape constraints on the infinite-dimensional and unknown link function parameter. I study multiple-index discrete choice models where the link function is known to be bounded between zero and one and is (partly) monotonic. In the paper I present an easy to implement and computationally efficient sieve GLS estimation approach using a sieve space of constrained I- and B-spline basis functions. The estimator is shown to be consistent and that imposing shape constraints speeds up the convergence rate of the estimator in a weak Fisher-like norm. The asymptotic normality of relevant smooth functionals of model parameters is derived and I illustrate that necessary assumptions are milder if shape constraints are imposed. A Monte Carlo Simulation study shows the finite-sample properties of the estimator and gains of imposing shape constraints in finite samples.

# Zusammenfassung

Diese Dissertation umfasst drei Aufsätze zu verschiedenen Themen aus dem Bereich der Mikroökonometrie, einem Teilgebiet der theoretischen Ökonometrie, das sich insbesondere mit Problemen bei der Analyse von Querschnittsdaten befasst. Das erste Kapitel ist eine gemeinsame Arbeit mit Christoph Breunig und umfasst semi/nichtparametrische Regressionsmodelle, in denen die abhängige Variable einen nicht-klassischen Messfehler aufweist. Zunächst werden Bedingungen erarbeitet, unter denen die Regressionsfunktion bis auf eine Normalisierung identifiziert werden kann. Zur Schätzung wird ein neuer Schätzer entwickelt, bei dem eine Rang-basierte Kriteriumsfunktion über einen sieve-Raum optimiert wird und dessen Konvergenzrate hergeleitet. Im Rahmen einer Monte Carlo Simulationsstudie wird gezeigt, dass der Schätzer Verzerrungen durch den nichtklassischen Messfehler korrigieren kann. Dies wird in einer Anwendung zur Entstehung von Erwartungen am Aktienmarkt auch empirisch illustriert.

Das zweite Kapitel beschäftigt sich mit der Schätzung von bedingten Dichtefunktionen von zufälligen Koeffizienten in linearen Regressionsmodellen. Es wird ein zweistufiges Schätzverfahren entwickelt, in dem zunčhst eine Approximation der bedingten Dichte der Regressions-Koeffizienten hergeleitet wird, die durch Funktionen von Kontrollvariablen parametrisiert ist. In einem weiteren Schritt können diese Funktionen mit generischen Methoden des maschinellen Lernens geschätzt werden. Des Weiteren wird auch die Konvergenzrate des Schätzers in der $L_2$-Norm hergeleitet sowie dessen punktweise, asymptotische Normalität. Der Schätzer wird mittels random forest Algorithmen angewandt und seine Eigenschaften in Monte Carlo Simulationen untersucht. Zudem wird mit der Methode die Heterogenität von Verhalten in einem Labor-Experiment zur Portfolio Selektion analysiert.

Im dritten Kapitel wird ein neuer und einfach umsetzbarer Ansatz zur Schätzung semi(nicht)parametrischer diskreter Entscheidungsmodelle, unter Berücksichtigung von Restriktionen auf die funktionalen Parameter des Modells, vorgestellt. Die untersuchten Modelle weisen funktionale Parameter auf, welche allgemein durch null und eins begrenzt, sowie monoton steigend in einigen Argumenten sind. Zentraler Teil der Arbeit ist die Entwicklung eines GLS-Schätzers über einen geeigneten sieve-Raum, der aus I- und B-Spline Basisfunktionen unter geeigneten Restriktionen basiert. Es wird gezeigt, dass sich die Berücksichtigung der Restriktionen auf die funktionale Form positiv auf die Konvergenzrate des Schätzers in einer schwachen Norm auswirkt und so notwendige Bedingungen für die asymptotische Normalität semiparametrischer Schätzer einfacher erreichen lässt. Eine Monte Carlo Studie stellt die Eigenschaften des Schätzers in endlichen Stichproben dar.

# Acknowledgements

This work is the culmination of a very eventful and enriching period of my life. For the academic aspects of this work I like to thank my advisor Christoph Breunig for introducing me to the world of theoretical econometrics and for his support and understanding along the way. Sonja Greven for agreeing to review this work despite thematic differences. The CRC 190, especially, Georg Weizsäcker for providing financial support and a superb environment for doing research. Many individuals along the way had an impact on some level or another on this work so I like to thank all members of the Chair of Econometrics at Humboldt University, all fellow students from my cohort at BDPEMS/BSE, members of the CRC 190, all peers from the PhD cohorts at the universities of Konstanz and Frankfurt who were kind to host me and ultimately also my colleagues from the Deutsche Bundesbank. It has been a long and enlightening ride.

The largest thanks is reserved to my wife Anika, who is intricably intertwined with the process of how this work came to be and for her love and patience along the way. I thank my parents who unconciously put me on this track of life for which I am eternally grateful and for my entire family who supported me especially in the last stages of this work. The work is dedicated to our daughter Victoria, who has never been helpful in finishing this work but turned every day brighter.

# Contents

# Chapter 1

# Nonclassical Measurement Error in the Outcome Variable

## 1.1 Introduction

In empirical research, measurement error is a recurring issue. In recent years, much attention has been given to various forms of measurement error in the covariates of econometric models, whereas measurement error of the dependent variable is mostly ignored. In many economic environments, measurement error of the dependent variable may be driven (in a nonlinear fashion) by the underlying variable. This nonclassical measurement error implies biased estimation results if not accounted for.

This paper is concerned with semi-/nonparametric regression models where the dependent variable of interest $Y^*$ is generally not observed and only a possibly error-contaminated measurement $Y$ is observable. Specifically, $Y^*$ satisfies

$$Y^* = g(X) + U, \tag{1.1}$$

where the unknown function $g$ is of interest given observed covariates $X$ and unobservables $U$. We study the *nonclassical* measurement error case where $\mathbf{E}[Y|Y^*, X] \neq Y^*$. Hence, the regression function $g$ does in general not coincide with conditional expectations of observable variables and we cannot impose $g(x) = \mathbf{E}[Y|X = x]$.

Nonparametric identification of our model relies on the availability of covariates which do not affect the measurement error directly. We impose such type of exclusion restriction on a subset $Z$ of the vector $X = (Z, W)$, where $W$ are additional controls. Under a monotonicity condition on the measurement error mechanism, we show in this paper that model (1.1) can be reformulated as a generalized regression model

of the form

$$\mathbf{E}[Y|X = x] = H(g(x), w),$$

where $H(\cdot, w)$ is a nonlinear, monotonic function for $w$ in the support of $W$. Identification of the function $g$, up to strictly monotonic transformations, immediately follows, which allows us to infer on economically relevant quantities such as the direction and shape of partial effects.

Under scale and location normalization of the unknown link function $H$, nonparametric identification of the regression function $g$ is obtained. We highlight that normalization of the link function $H$ is equivalent to imposing mild shape restrictions on the measurement error mechanism. Additionally, our normalization conditions on the link function do not only naturally extend the classical measurement case but are also satisfied if there is a range of $Y^*$ where measurement error is classical. Our nonparametric identification results build thus on intuitive assumptions without relying on high-level assumptions such as completeness, see Hu and Schennach (2008).

We consider a sieve, rank-based minimum distance estimator and establish its asymptotic properties. We derive the rate of convergence in $L^2$ sense of our estimator. We find that the sieve rank estimator generally suffers from ill-posedness in the convergence rate as the rank-based criterion function is not continuous in the usual $L^2$-norm. We develop the theory for the case where $W$ is discrete and provide an extension to allow continuous controls $W$ using kernel weights in the appendix of this paper.

We analyze the performance of the estimator in a Monte Carlo simulation study and in an empirical application using survey data. We apply our estimator to study belief formation with subjective belief data from the German Socio-Economic Panel innovation sample (SOEP-IS). Subjective belief data is known to be plagued by substantial measurement error and it is in general hard to justify that the measurement error is classical and thus not sensitive to the underlying true individual belief. We study the impact of an exogenous display of historic stock market returns provided to survey respondents prior to eliciting their belief on future returns. Applying our method, we find a monotonic and concave relationship between the historic information and stated beliefs indicating that individuals acknowledge the given information conservatively.

**Literature** Our work ties into the literature on measurement error in observable variables of econometric models. The literature on measurement error in covariates

is extensive, whereas measurement error in the outcome variable has received much less attention. For a review of models with errors in covariates, see e.g. Chen et al. (2011) and Schennach (2013). Chen et al. (2005) develop a general way of accounting for measurement error in any variable of a class of semiparametric models once auxiliary data, e.g. from validation samples is available. However, this is hardly the case in most practical applications. Models focusing on nonclassical measurement error in the outcome side are rare. Chapter 3 of Abrevaya and Hausman (1999) considers a semiparametric model with a more simplistic measurement error mechanism. Hoderlein and Winter (2010) and Hoderlein et al. (2015) develop structural models of response error in surveys due to imperfect recall and derive testable implications for econometric analyses. The latter paper focuses on the role of rounding in individual reporting behavior which is also a more specific form of nonclassical measurement error.

de Nadai and Lewbel (2016) allows for classical measurement error in the outcome variable that is correlated with an error in covariates. Abrevaya and Hausman (2004) consider classical measurement error of the dependent variable in a transformation model. Given we have a precise idea on the form of measurement error, a sizeable literature is usually available providing different strategies for identification. For instance a special case of nonclassical measurement error is selective non-response in the outcome variable, see e.g. D'Haultfoeuille (2010) or Breunig et al. (2018) and references therein. A non-nested form of nonclassical measurement error are Berkson-type errors, see Berkson (1950) and Schennach (2013, section 6.3).

Our identifying assumptions lead us to the literature on generalized regression models as introduced in Han (1987) or the class of nonlinear index models in Matzkin (2007). See also the model studied in Jacho-Chavez et al. (2010). Estimation of such models often proceeds by rank-based estimation strategies, see Han (1987), Cavanagh and Sherman (1998), Khan (2001), Shin (2010) and Abrevaya and Shin (2011) which all consider parametric regression models with the exception of Matzkin (1991b) who studies a nonparametric model with additional shape restrictions on the link function. A recent contribution studying rank estimators in a high-dimensional setting is Fan et al. (2020). To the best of our knowledge, we are the first to study nonparametric M-estimation with rank-based criterion functions and to point out and illustrate the ill-posedness of the estimation problem. Jureckova et al. (2016) study a different class of rank estimators in the context of a parametric model with measurement error in both regressors and outcome. Their the outcome error may not be nonclassical as in our general notion but can at most depend on observable regressors.

The remainder of the paper is organized as follows. In section 1.2 we present our model setup and give a nonparametric identification result for features of the mean regression function when there is a form of nonclassical measurement error in the outcome variable. In section 1.3 we introduce a sieve estimator with a rank based criterion function and establish its convergence. In section 1.4 we analyze finite sample properties of the estimator in a Monte Carlo simulation study. Section 1.5 contains an application of our method to belief formation of stock market expectations. Appendix 1.7.1 provides an extension to weighted sieve rank estimation, when control variables are continuous. All proofs are postponed to the Appendix 1.7.3.

## 1.2 Model Setup and Identification

We consider a nonparametric econometric model with measurement error in the outcome variable. The model we study is

$$Y^* = g(X) + U, \tag{1.2}$$

where $Y^*$ is the scalar, outcome variable, $X$ is a $d_x$-dimensional vector of exogenous covariates, $U$ is a scalar error term, and $g$ a nonparametric function of interest. The outcome variable $Y^*$ is not observed by the researcher; only an error contaminated measurement $Y$ is available. We are primarily interested in the case where the error satisfies $\mathbf{E}[U|X] = 0$ and thus $g$ is the unknown conditional expectation function of $Y^*$ given $X$. Our identification approach can be readily extended to more general nonseparable models of the form $Y^* = m(g(X), U)$ where $m$ is strictly monotonic in its first argument.

Throughout the paper, we assume that the regressors $X$ can be decomposed such that $X = (Z', W')'$, where $Z$ has no direct effect on the measurement error and $W$ are control variables. Also we introduce the notation $g_w(\cdot) \equiv g(\cdot, w)$ for the regression function evaluated at a fixed $w$ in the support of $W$. We now provide conditions, which allow for nonparametric identification of $g_w$ up a strictly monotonic transformation.

**Assumption 1** (Exclusion Restriction)**.** *The observed outcome $Y$ is conditionally mean independent of $Z$ given $Y^*$ and $W$, i.e., $\mathbf{E}[Y|Y^*, Z, W] = \mathbf{E}[Y|Y^*, W]$.*

Assumption 1 rules out that $Z$ has a direct effect on the measurement $Y$ in conditional expectations. Assumption 1 is generally weaker than assuming that the conditional distribution of $Y$ given $(Y^*, Z, W)$ does not depend on $Z$, which restricts

$Z$ to have no information on $Y$ that is not captured by $(Y^*, W)$. Analogue exclusion restrictions are commonly imposed in the literature on nonclassical measurement error in covariates. In Hu and Schennach (2008, Assumption 2 (ii)), the distribution of the error-contaminated regressor is independent of instruments conditional on the latent regressor (see also Schennach (2013, section 4.3)). Assumption 1 is less restrictive than other exclusion restrictions found in the measurement error literature, see Ben-Moshe et al. (2017, Assumption 2.1 (iii)).

Conditions similar to Assumption 1 can also be found in the literature on selective non-response, which is a special case of nonclassical measurement error in the outcome. Individuals either report the outcome truthfully (response indicator $D = 1$) or not at all ($D = 0$) so the observed outcome in this case is $Y = DY^*$. See also Remark 1.2 below. An identifying assumption in D'Haultfoeuille (2010) and Breunig et al. (2018) is that $D \perp\!\!\!\perp X \mid (Y^*, W)$, which is related to Assumption 1.

In the following, we make use of the notation $h(Y^*, W) = \mathbf{E}[Y|Y^*, W]$. Assumption 1 implies the measurement error model

$$Y = h(Y^*, W) + V,$$

where $\mathbf{E}[V|Y^*, W] = 0$. Consequently, Assumption 1 implies conditional mean independence of the measurement error $V$ given the regression error $U$, that is, $\mathbf{E}[V|U] = 0$.

**Assumption 2** (Monotonicity)**.** *For any $w \in supp(W)$, the function $h(\cdot, w)$ is weakly monotonic and non-constant over the support of $Y^*$.*

Assumption 2 imposes that the expected observed outcome $Y$ is monotonic in the latent outcome $Y^*$ given $W$. This is trivially satisfied when the measurement error is classical, i.e., when $h$ does not depend on $W$ and is the identity. A similar monotonicity condition has also been imposed in the measurement error model in Abrevaya and Hausman (1999, Example 3).[1] Note that $h$ does not need to be strictly monotonic which allows to consider models with rounding error in the outcome, see Hoderlein et al. (2015).A simple example is rounding to the nearest integer, where e.g. $\mathbf{E}[Y|Y^* = 1.3, W = w] = 1$ and $\mathbf{E}[Y|Y^* = 1.5, W = w] = 2$ which is in line with the above weak monotonicity assumption. Settings where the assumption does not hold can be constructed but appear in general not very plausible in applications. [2]

[1] In our notation Abrevaya and Hausman (1999) consider the error mechanism $Y = h(Y^*, V)$, with $\partial_y h(Y^*, V) > 0$, $\partial_v h(Y^*, V) > 0$ and $V \perp\!\!\!\perp (X, U)$. As we allow for heteroscedasticity in the measurement error model, condition $\partial_y h(Y^*, V) > 0$ may lead to one sided error restrictions.

[2] Let $Y^*$ be the latent income of an individual and $Y$ the income reported to the tax authority. Assume there is a threshold of 1000$ below which the income is tax-free. An individual with an

We furhter discuss the plausiblity of Assumption 2 in the context of the application in section 1.5 in a setting with survey data.

**Assumption 3** (Conditional Exogeneity)**.** *The conditional independence restriction* $Z \perp\!\!\!\perp U \mid W$ *holds.*

Assumption 3 imposes a conditional independence restriction of $Z$ and the regression error $U$. This condition is also known as conditional exogeneity assumption following White and Chalak (2010). Independence assumptions can be restrictive, but are often required in the measurement error literature (see, e.g. Hausman et al. (1991), Schennach (2007), Ben-Moshe et al. (2017, Assumption 2.2)), or when accounting for endogeneity using control functions (see, e.g. Newey et al. (1999)). We relax such restrictions by imposing independence to hold only conditional on control variables $W$. Similar conditions are often employed for identification in the econometrics literature, see e.g. Chiappori et al. (2015) for nonparametric identification in a transformation model. Assumption 3 is also closely related to the special regressor assumption, see Lewbel (2014) for a review.

Next, we need the following set of regularity conditions. We introduce the notation $supp(V)$ for the support of a random vector $V$.

**Assumption 4.** *For any* $w \in supp(W)$*: (i) the function* $g_w$ *is continuous; (ii) and any* $z_1, z_2 \in supp(Z)$ *such that* $g_w(z_1) < g_w(z_2)$ *there exists* $u \in supp(U)$ *satisfying* $h(g_w(z_1)+u, w) < h(g_w(z_2)+u, w)$*; (iii) there is at least one variable* $Z_{(1)}$ *such that* $Z = (Z_{(1)}, Z_{(-1)})$ *with* $f_{Z_{(1)}|Z_{(-1)},W}(z_1|z_{-1}, w) > 0$ *for all* $(z_1, z_{-1}) \in supp(Z)$.

Assumption 4 (ii) is a mild support condition on $U$ conditional on $W = w$. The unobservable $U$ must vary sufficiently to shift $g_w(Z)$ out of a flat region of $h$. The assumption is not required if $h$ is strictly monotonic in its first argument. Assumption 4 (iii) requires $Z$ to contain at least one continuously distributed variable with sufficient variation. If $Z$ is scalar then Assumption 4 (iii) may be replaced by $f_{Z|W}(z|w) > 0$ for all $z \in supp(Z)$. This rules out the case of $Z$ being a discrete scalar variable. For a nonseparable model $Y^* = m(g(X), U)$, Assumption 4 (ii) would need to be reformulated to satisfying $h(m(g_w(z_1), u), w) < h(m(g_w(z_2), u), w)$. This does not strengthen the Assumption as long as $m$ is strictly monotonic in its second

income of 999$ has no incentive to misreport, yet individuals with an income of 1000$ may profit from underreporting at the risk of sanctions. Assume there is a share $p$ of types in the population willing to misreport and $1-p$ who is not. Assume $\mathbf{E}[Y|Y^* = 1000] = p \cdot 990 + (1-p) \cdot 1000$, then Assumption 2 is violated if $p > 10\%$. Here the group of misreporters must be sizeable and their average underreporting sufficiently far away from the threshold to break the assumption, making the example appear rather constructed, as without strong incentives misreporters may report only slightly below the threshold.

argument. For nonseparable models without monotonicity of $m$ in the unobserved $U$ this is a more high-level assumption which can nevertheless be satisfied if $U$ has large support and a sufficient impact on $Y^*$ via the function $m$.

Under the stated assumptions, now provide establish equivalence to the regression model (1.2) to a generalized regression model specified by the link function $H(g_w(z), w) = \mathbf{E}[h(g(z, W) + U, W) \mid W = w]$. Below, $\mathbb{1}\{\cdot\}$ denotes the indicator function.

**Theorem 1.1.** *Let Assumptions 1–4 be satisfied, then for any $w \in supp(W)$ it holds*

$$\mathbf{E}[Y|X = x] = H(g_w(z), w), \tag{1.3}$$

*where $H(\cdot, w)$ is strictly monotonically increasing and $g_w(z)$ maximizes the function*

$$\mathcal{Q}(\phi, w) = \mathbf{E}[Y_1 \mathbb{1}\{\phi(X_1) > \phi(X_2)\} \mid W_1 = W_2 = w]. \tag{1.4}$$

*with respect to $\phi$ which is a generic function sharing the properties of $g_w$. In particular, the function $g_w(\cdot)$ is identified up to strictly increasing transformations.*

The model (1.3) falls into the class of generalized regression models studied by Han (1987), Matzkin (1991b), and Cavanagh and Sherman (1998). Further note that nonclassical measurement error implies heterogeneous biases for the marginal effects. When $\partial_z H(g_w(z), w) < 1$ we obtain an *attenuation bias* for the marginal effect $\partial_z g_w(z)$ and when $\partial_z H(g_w(z), w) > 1$ we get an *augmentation bias* for $\partial_z g_w(z)$.

Theorem 1.1 implies identification of features of $g_w$ that are preserved under monotonic transformations. This includes the sign of partial effects, the ratio of two partial effects[3] and properties such as quasi-concavity (-convexity) of the function. For the remainder of the paper we consider identification and estimation of $g_w$ in the point identified case.

We impose the following restriction on the model and the measurement error mechanism described by the function $H$.

**Assumption 5.** *(i) The function $g_w$ is additively separable such that there exists a decomposition $Z = (Z_1, Z_{-1})$ such that $g_w(Z) = m_w(Z_1) + l_w(Z_{-1})$ for some functions $m_w, l_w$. (ii) There exists $\{z_1, z_2\} \subset supp(Z)$ with $g_w(z_1) \neq g_w(z_2)$ and $\mathbf{E}[Y|Z = z, W = w] = \mathbf{E}[Y^*|Z = z, W = w]$ for $z \in \{z_1, z_2\}$.*

Assumption 5 (i) imposes an additive separable structure on the regression function $g_w$. Following the identification statement in Theorem 1.1, mere location and

[3]Note that for $g(z_1, z_2)$ it holds that $\frac{\partial g}{\partial z_1} / \frac{\partial g}{\partial z_2} = \frac{\partial H(g)}{\partial z_1} / \frac{\partial H(g)}{\partial z_2}$ whenever these quantities and ratios are well-defined.

scale normalizations are not sufficient to point identify $g_w$. However, for any additive separable model this is the case, see also Jacho-Chavez et al. (2010). Assumption 5 (ii) restricts the measurement error for at least to realizations of $Z$. For instance, one can think of pension information to account for nonclassical measurement error in labor income survey questions (see Breunig and Haan (2018)). Here, for certain ranges of labor income (e.g. close to the median) we may assume that the measurement error is of classical form. Assumption 5 (ii) is also in line with normalization requirements for identification under nonclassical measurement error. For instance, Assumption 5 of Hu and Schennach (2008) requires some functional of the distribution of the measurement error conditional on the value of the true variable to be equal to the true variable itself, such as some quantile of $Y|Y^* = y^*$ to correspond to $y^*$.

Economic restrictions on the model can also be employed to sufficiently restrict the function space. We refer to the discussion in sections 3.4 and 4.4 in Matzkin (2007) where several possible function spaces are discussed that can replace Assumption 5(i). This includes the spaces of functions that are homogeneous of degree one or so called "least-concave" functions, see also Matzkin (1994). Matzkin (2007) shows that imposing homogeneity of degree 1 and a location normalization is sufficient for Assumption 5. Homogeneous functions are frequently encountered in microeconomics. Thus, in applications where the function $g$ has the structural interpretation of a production or cost function, homogeneity can be a reasonable restriction on the parameter space.

**Corollary 1.1.** *Let Assumptions 1– 5 (i) be satisfied, then the function $g_w$ is identified up to a location and scale normalization. If 5 (ii) is additionally satisfied then the function $g_w$ is point identified.*

Corollary 1.1 establishes identification of the regression function under normalization imposed in Assumption 5. The shape restrictions imposed in Assumption 5 imply a normalization of the unknown, nonparametric link function $H$, in contrast to nonparametric generalized regression models, where normalization is typically imposed on the unknown function of interest.

We neither restrict the support of the observed outcome $Y$, nor require continuity in the function $h(\cdot, w)$. Thus, we can also cover cases where the observed outcome is categorical or has mass points. This likely occurs in survey data as respondents tend to provide rounded values. The following examples consider a generalization and special case of model (1.2).

**Example 1.1** (Control function approach)**.** *We can also motivate the presence of $W$ in Assumption 3 as a control function. To this end we deviate for a moment*

*from our previous notation and introduce the following triangular model*

$$\begin{aligned} Y^* =& g(X) + U \\ X =& m(Z, \eta) \end{aligned}$$

*where for simplicity $X$ is a one-dimensional endogenous covariate that may correlate with the model error $U$. The function $m$ is strictly monotonic in $\eta$ and $Z$ is an instrumental variable satisfying $Z \perp\!\!\!\perp (U, \eta)$. Under additional regularity conditions, following Imbens and Newey (2009, Theorem 1) it holds that*

$$\begin{aligned} &X \perp\!\!\!\perp U \mid W \quad \textit{with} \\ &W = F_{X|Z}(X, Z) = F_\eta(\eta), \end{aligned}$$

*where $F_V$ denotes the cummulative distribution function of a random variable $V$. As in Assumption 1 we impose $\mathbf{E}[Y|Y^*, Z, W] = \mathbf{E}[Y|Y^*, W]$. Thus, following Theorem 1.1, we obtain identification of the structural function $g$ up to a strictly monotonic transformation.*

**Example 1.2** (Selective Nonresponse)**.** *Consider a nonresponse model*

$$\begin{aligned} Y &= DY^* \\ D &= \phi(Y^*, W, V), \end{aligned}$$

*for some unknown function $\phi$, where the response indicator $D \in \{0, 1\}$ is always observed and $Y^*$ is only observed if $D = 1$. This framework, where the response mechanism is mainly driven by the latent outcome $Y^*$ has been studied by D'Haultfoeuille (2010) and Breunig et al. (2018). As long as the conditional mean function $h(Y^*, W) = P(D = 1|Y^*, W)Y^*$ is monotonic in its first argument, the model is in accordance to Assumption 2. This holds e.g. when the conditional response probability function is monotonic and the support of $Y^*$ is bounded below*[4]*. In this case, a completeness condition for nonparametric identification of the conditional selection probability $P(D = 1|Y^*, W)$ (see D'Haultfoeuille (2010) and Breunig et al. (2018)) via conditional moment restrictions is not required.*

[4]If $Y^*$ is bounded below, then $Y^*$ can be redefined such that without loss of generality $Y^* \geq 0$ and monotonicity of $h(Y^*, W) = P(D = 1|Y^*, W)Y^*$ follows from taking the derivative.

## 1.3 Estimation and Asymptotic Properties

In this section, we introduce a nonparametric sieve M-estimator with a simple, rank-based criterion function. For simplicity, we consider only the case where $W$ consists of discrete variables and defer the estimation with continuous $W$ to Appendix 1.7.1.

### 1.3.1 The Sieve Rank Estimator

Our identification result builds on shape restrictions imposed on the measurement error mechanism, which imply identified moment conditions. Specifically, for a given $w$ we have from the identification statement in Theorem 1.1 that the true $g_w$ maximizes the function

$$\mathcal{Q}(\phi, w) = \mathbf{E}[Y_1 \mathbb{1}\{\phi(X_1) > \phi(X_2)\} \mid W_1 = W_2 = w].$$

Based on this population criterion, we now consider a sieve rank estimator, which implicitly accounts for imposed shape restrictions required for identification.

We propose the following sieve rank estimator

$$\begin{aligned} \widehat{g}_w &= \arg\max_{\phi \in \mathcal{G}_K} \mathcal{Q}_n(\phi, w) \text{ where} \\ \mathcal{Q}_n(\phi, w) &:= \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} Y_i \mathbb{1}\{W_i = W_j = w\} \mathbb{1}\{\phi(Z_i) > \phi(Z_j)\}, \end{aligned} \tag{1.5}$$

for some $K = K(n)$ dimensional sieve space $\mathcal{G}_K$. Here, the dimension parameter $K$ grows slowly with sample size $n$. For the special case where $W$ is absent, the criterion reduces to

$$\mathcal{Q}_n(\phi) = \frac{1}{n(n-1)} \sum_{i=1}^{n} Y_i \operatorname{Rank}(\phi(Z_i)), \tag{1.6}$$

where the rank function is defined as $\operatorname{Rank}(\phi(Z_i)) = \sum_{j \neq i}^{n} \mathbb{1}\{\phi(Z_i) > \phi(Z_j)\}$. This is a nonparametric version of the criterion of Cavanagh and Sherman (1998).

The specific choice of $\mathcal{G}_K$ hinges on the chosen normalization. Under a normalization of the link function $H$, see Corollary 1.1, we may consider a linear sieve space $\mathcal{G}_K = \{\phi : \phi(z) = \gamma_w' p^K(z)\}$. Let $p^K = (p_1, \ldots, p_K)$ be a $K$- dimensional vector of known basis functions such as polynomials, splines or similar. We can in principal also apply the general sieve estimation technique of Chen (2007) based on the conditional moment restriction $\mathbf{E}[Y|X = x] = H(g_w(z), w)$. This would require to estimate $H$ along with $g_w$ and nesting of two sieve spaces. Our estimation strategy

constructively arises from the identification argument and provides a simple direct estimate of $g_w$. We also directly leverage the monotonicity condition on $H$ in the estimation so there is no need to introduce additional shape-constraints.

### 1.3.2 Convergence Rate

In this section, we derive a rate of convergence of the sieve rank estimator $\widehat{g}_w$ given in (1.5). To keep notation simple, we omit the controls $W$ entirely from the following analysis. In this case, estimation amounts to maximizing the criterion in (1.6) from the previous section over a suitable sieve space.

For the remainder of the paper we consider the centered criterion function

$$\mathcal{Q}(\phi) = \mathbf{E}\left[Y_i\big(\mathbb{1}\{\phi(Z_i) > \phi(Z_j)\} - \mathbb{1}\{g(Z_i) > g(Z_j)\}\big)\right] \tag{1.7}$$

where $g$ is the regression function satisfying the model equation (1.2). Centering does not change the maximizer in the optimization problem and is thus without loss of generality.

Our analysis builds on a linearization of the nonlinear criterion function $\mathcal{Q}(\cdot)$. The first directional derivative of $\mathcal{Q}$ is equal to zero for any arbitrary direction and hence, we consider the second directional derivative which can be viewed as a quadratic approximation to the criterion function $\mathcal{Q}(\cdot)$. Specifically, we introduce

$$Q(\phi - g) := \frac{\partial^2}{\partial \tau^2} \mathcal{Q}(g + \tau(\phi - g)) \Big|_{\tau=0}$$

denote the second directional derivative of the non-linear functional $\mathcal{Q}$ in the direction $\phi - g$. We assume that the functional $Q(\cdot)$ is bi-linear and continuous. Below, we denote $L^2(Z) = \{\phi : \|\phi\|_{L^2(Z)} < \infty\}$ where $\|\phi\|_{L^2(Z)} := \sqrt{\mathbf{E}\,\phi^2(Z)}$.

To account for the potential instability of the estimation problem, we introduce the sieve measure of ill-posedness

$$\tau_K = \sup_{\phi \in \mathcal{G}_K} \frac{\|\phi - \Pi_K g\|_{L^2(Z)}}{Q(\phi - \Pi_K g)}$$

to account for the fact that the criterion function and the $L^2$-norm are generally not (locally) equivalent. If $\tau_K \to \infty$ as $K \to \infty$ the problem of estimating $g$ is ill-posed in rate and additional regularization slows down convergence in the strong $L^2$- norm. In contrast to Chen and Pouzo (2012), we rely on the second directional derivative in the denominator.

For the following assumption we introduce a local neighborhood of $g$ and define

the space $\mathcal{G}_K^\delta = \{\phi \in \mathcal{G}_K : \|\phi - g\|_{L^2(Z)} < \delta\}$ with $\delta > 0$.

**Assumption 6.** *(i) A random sample $\{(Y_i, Z_i)\}_{i=1}^n$ of $(Y, Z)$ is observed; (ii) there exists $\Pi_K g \in \mathcal{G}_K$ such that $\|\Pi_K g - g\|_{L^2(Z)} = O(K^{-\alpha/d_z})$; (iii) $\mathbf{E}[U^2] < \infty$ and $g \in L^2(Z)$; (iv) for any $\phi$ in $\mathcal{G}_K^\delta$ there exists a constant $0 < \eta < 1$ such that $|\mathcal{Q}(\phi) - Q(\phi - g)| \leq \eta \cdot Q(\phi - g)$; (v) the cdf of $g(Z)$ is Lipschitz continuous, i.e., $|F_{g(Z)}(a) - F_{g(Z)}(b)| \leq C|a - b|$ for some constant $C$ and any $a, b$; and (vi) $\tau_K \sqrt{K/n} = o(1)$.*

Assumption 6 (ii) imposes regularity on the regression function $g$ via a sieve approximation error, see also Chen (2007) for examples. Assumption 6 (iv) is also known as the tangential cone condition and implies that $\mathcal{Q}(\phi)$ is locally equivalent to $Q(\phi - g)$ which is a typical condition required to derive the convergence rate for sieve estimators; see Chen and Pouzo (2012, Assumption 4.1(ii)) and also Dunker et al. (2011). Assumption 6 (v) amounts to a local continuity assumption for the kernel of an empirical process, see e.g. Chen (2007, Condition 3.8). Assumption 6 (vi) restricts the growth of $K$ relative to the sieve measure of ill-posedness $\tau_K$ and is required for consistency, see Lemma 1.3.

**Remark 1.1** (Illustration of Ill-Posedness)**.** *To give an insight on the source of ill-posedness, note that*

$$\mathcal{Q}(\phi) = \mathbf{E}\left[Y_i \left(F_{g(Z_i)|Y_i}(g(Z_j)) - F_{\phi(Z_i)|Y_i}(\phi(Z_j))\right)\right]$$

*which shows that if there is little variation in the distribution of $F_{g(Z)|Y}$ for variations of $g$ then the ill-posed inverse problem becomes more severe. This is further illustrated by the following lemma where we study a special case for which we can derive $Q$ analytically and give sufficient conditions for Assumption 6 (iv).*

**Lemma 1.1.** *Consider the additive separable model $g(Z) = Z_1 + \widetilde{g}(Z_2)$ with bivariate $Z = (Z_1, Z_2)$. Then Assumption 6 (iv) is satisfied if $f'_{Z_1|Z_2}$ is uniformly bounded away from zero and $f''_{Z_1|Z_2}$ is uniformly bounded above.*

*The special case outlined in Lemma 1.1 illustrates the behavior of $\tau_K$. If the density $f_{Z_{21}|Z_{22}}$, that is the conditional density of the separable covariate, is flat in the relevant support, we may encounter the case that the criterion $\mathcal{Q}$ is close to zero for candidate functions that are arbitrarily far away from the true function in the $L^2$- sense.*

We further illustrate this issue in a Monte Carlo simulation study in section 1.4, where we show that the estimation problem becomes more difficult as $f_{Z_{21}|Z_{22}}$

becomes more flat. We are now in a position to provide a general rate of convergence of our sieve rank estimator $\widehat{g}$.

**Theorem 1.2.** *Let Assumptions 1-6 be satisfied. It holds that*

$$\|\widehat{g} - g\|_{L^2(Z)} = O_p\Big( \max\Big\{\tau_K \sqrt{\frac{K}{n}},\ K^{-\alpha/d_z}\Big\}\Big)$$

The proof of Theorem (1.2) makes use of a representation of second-order U-processes as empirical processes following Clemencon et al. (2008). To the best of our knowledge, this is the first convergence rate result for nonparametric M-estimators with a rank-based criterion function in the presence of ill-posedness.

The next corollary provides concrete rates of testing when the dimension parameter $K$ is chosen to level variance and square bias under classical smoothness conditions. We call our model *mildly ill-posed* if: $\tau_k \sim k^{\gamma/d_z}$ with $\gamma > 0$ and *severely ill-posed* if: $\tau_k \sim \exp(k^{\gamma/d})$, with $\gamma > 0$.[5]

**Corollary 1.2.** *Let Assumptions 1-6 be satisfied.*

1. *Mildly ill-posed case: setting* $K \sim n^{d_z/d_z+2\gamma+2\alpha}$ *yields*

$$\|\widehat{g} - g\|_{L^2(Z)} = O_p(n^{-\alpha/(2\alpha+2\gamma+d_z)}).$$

2. *Severely ill-posed case: setting* $K \sim \log(n)^{d/\gamma}$ *yields*

$$\|\widehat{g} - g\|_{L^2(Z)} = O_p(\log(n)^{-\alpha/\gamma}).$$

Both convergence rates are the optimal rates for ill-posed problems. As outlined in the discussion following Lemma 1.1, the severity of the ill-posedness will generally depend on the chosen normalization and features of the data.

## 1.4 Monte Carlo Simulation Study

This section demonstrates how nonclassical measurement errors in the outcome alters mean regression results in finite samples and shows the usefulness of our approach to correct for such biases. We compare regression function estimates obtained from simply ignoring the measurement error with our estimator, which accounts for

[5] If $\{a_n\}$ and $\{b_n\}$ are sequences of positive numbers, we use the notation $a_n \lesssim b_n$ if $\limsup_{n\to\infty} a_n/b_n < \infty$ and $a_n \sim b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

the presence of the error. Throughout this section, simulation results are based on a sample of size of $n = 1000$ and 1000 Monte Carlo iterations.

We consider the following data generating process

$$\begin{aligned} Y^* &= Z_1 + g(Z_2) + U \\ Y &= h(Y^*) + V, \end{aligned}$$

where $Z_1 \sim \mathcal{N}(1, \sigma^2)$, $Z_2 \sim \mathcal{U}[-3, 3]$ independent of each other, $g(\cdot) = \sin(\cdot)$ and the error terms $(U, V) \sim \mathcal{N}(0, I_2)$. Here, $I_2$ is the 2-dimensional identity matrix and for the standard deviation of $Z_1$ we choose $\sigma = 1$, which will be varied later. In the above model, $g$ is identified up to a location normalization. Analogously we could specify a linear or nonlinear function on $Z_1$ and impose an additional scale normalization on $g$. The function $h$ in the measurement error equation is chosen as

$$h(Y^*) = \begin{cases} q_{0.7} + b(Y^* - q_{0.7}) & \text{if } Y^* > q_{0.7} \\ Y^*, & \text{if } q_{0.3} \leq Y^* \leq q_{0.7} \\ q_{0.3} - a(q_{0.3} - Y^*) & \text{otherwise} \end{cases}$$

where $q_{0.3}, q_{0.7}$ denote the 30%- and 70%-quantile of $Y^*$ (determined via numerical approximation). The setup is analogous to a typical survey data setting with over- or underreporting in the tails of $Y^*$, whereas the center of the distribution is not affected. The scalars $a, b$ are chosen to vary the magnitude of measurement error.

Figure 1.1 illustrates the effects of the measurement error for the case $a = b = 0.5$. We show the realizations of $Y$ and $Y^*$ for a specific draw of the data generating process and plots the function $h$. We compare the measurement error function $h$ (depicted as red solid line) with the setup of classical measurement error, which is captured by the 45° line (depicted as black dashed line).

We implement the sieve rank estimator $\widehat{g}$ given in (1.5) using a linear sieve space with B-spline basis functions of order 3 with 2 interior knots that are placed according to quantiles of the empirical distribution. Thus we have $K = 4$. The elements of the sieve space are normalized at the point $(0, 0)$ which is the correct value of the true function $\sin(\cdot)$ at 0. This normalization can also be rationalized as utilizing prior knowledge on the measurement error mechanism in the sense of Assumption 5 (ii). For instance, we can expect that ignoring the measurement error results in estimates that are close to the true function $g$ in the center of the distribution of $Z_2$. Figure 1.2 shows the sieve rank estimates $\widehat{g}$ and compares them to a nonparametric series regression that does not account for nonclassical measurement

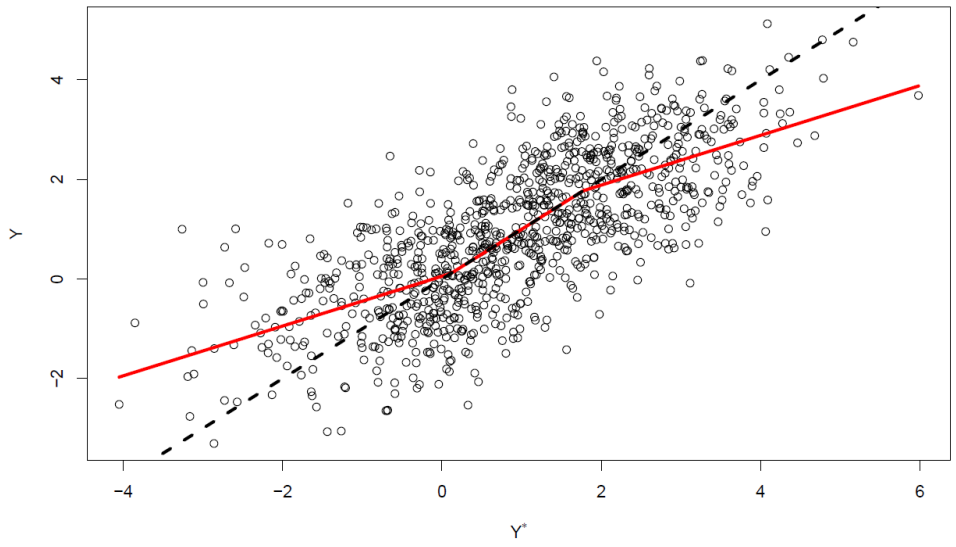


Figure 1.1: Realizations of $Y^*, Y$ when $a = b = 0.5$ based on a random draw of size $n = 1000$. The red solid line depict the function $h$ and the black dashed line the 45° line.

error in the outcome using the same order and the same knot placement as for $\widehat{g}$. For the latter estimator the same choice of basis functions and tuning parameters is adopted. We study different values for $a, b$ amongst which is the severe case $a = b = 0$ which essentially implies that at some point the measurements $Y$ are merely random fluctuations around a constant value[6]. We observe from the results in Figure 1.2 that our estimation strategy results in an accurate estimate of $g$ in any of the cases, whereas ignoring the measurement error yields estimates with a sizeable bias in the tails of $Z_2$. In the severe setting depicted in the right panel, ignoring measurement error results in a rather flat estimate which is significantly different from the sieve rank estimator.

The data generating process chosen here is in line with the model in Lemma 1.1 and thus allows us to study the degree of ill-posedness in the convergence rate of the estimator. As pointed out in the discussion following Lemma 1.1, the behavior of the sieve measure of ill-posedness $\tau_K$ is governed by the conditional density $f_{Z_1|Z_2}$. If the density $f_{Z_1|Z_2}$ is flat over the relevant support, $\tau_K$ diverges faster and the ill-posedness is more severe.

[6]Additionally we perform Kolmogorov-Smirnov tests to test the null hypothesis that $Y$ and $Y^*$ follow the same probability distribution on every drawn sample of the MC study. In the $a = b = 0.5$ setting we reject the null on a 5% - level only once in 1000 samples and in the $a = b = 0$ case we reject the null in 966 cases. Thus in the strong ME setting, $Y$ and $Y^*$ have different marginal distributions in contrast to the mild ME setting, where differences are virtually undetectable.

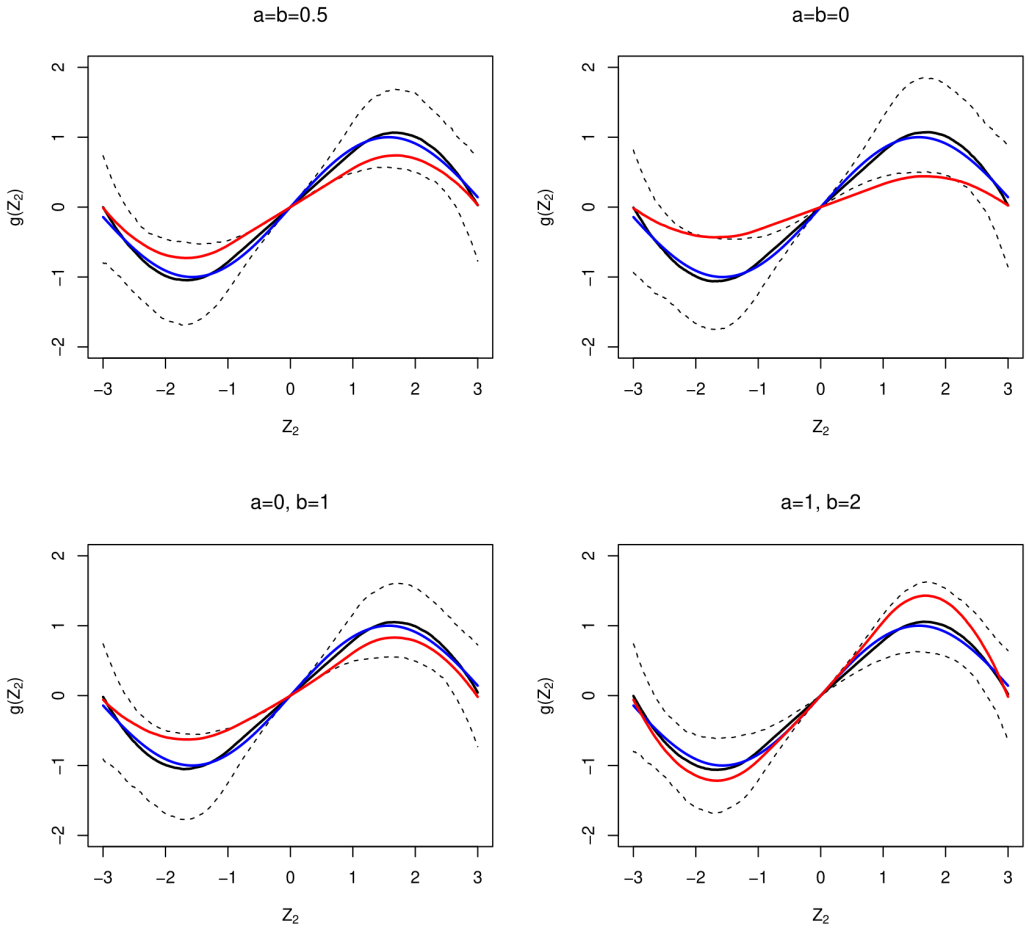


Figure 1.2: Estimation results normalized to go through the coordinate $(0, 0)$: Solid black line is the median of our sieve rank estimator $\widehat{g}$, solid red line is the median of a series estimator with same B-splines specification, solid blue line shows true $g(\cdot)$ function, and dashed black lines are the 0.95 and 0.05 quantiles over all Monte Carlo rounds.

Table 1.1 below shows mean squared errors of function estimates across different standard deviations of the separable covariate $Z_1$ which affects the slope of the density $f_{Z_1|Z_2}$. For small standard deviations, the conditional density $f_{Z_1|Z_2}$, i.e., here $f_{Z_1}$ by full independence, will be rather flat over most of the support. For small standard deviations of $Z_1$, the MSE increases more severely with $K$ as compared to large standard deviations. This illustrates that the degree of ill-posedness of the estimation problem is more severe whenever the slope of the density $f_{Z_1|Z_2}$ is small. Additionally we see that this is not the case when the distribution of $Z_1$ is fixed and the dispersion of $Z_2$ is varied. This confirms that the ill-posedness in this setting is not driven by the distribution of $Z_2$ in this setting.

| St. Dev. of $Z_1$ | $Z_2 \sim \mathcal{U}[-c, c]$ | MSE($\widehat{g}$) for sieve dim. | | | |
|---|---|---|---|---|---|
| $\sigma$ | c | $K = 3$ | $K = 4$ | $K = 5$ | $K = 6$ |
| 0.5 | 1 | 0.02209 | 0.06843 | 0.17294 | 0.52289 |
| | 3 | 0.02389 | 0.05982 | 0.17068 | 0.62054 |
| 1 | 1 | 0.01579 | 0.04293 | 0.09087 | 0.20775 |
| | 3 | 0.01807 | 0.04783 | 0.08118 | 0.19650 |
| 2 | 1 | 0.01489 | 0.04316 | 0.09514 | 0.20622 |
| | 3 | 0.01640 | 0.04580 | 0.08593 | 0.19877 |

Table 1.1: Results for the MSE($\widehat{g}$) for varying values of the standard deviation $\sigma$ of $Z_1$ and the range $c$ of $Z_2$.

## 1.5 Application: Beliefs on Stock Market Returns

Subjective beliefs on stock market returns are an important determinant in economic models that seek to explain stock market participation and portfolio choice, see e.g. Breunig et al. (2019) and the references therein. Subjective belief data, however, is known to be prone to a large degree of measurement error, see the discussion and references in Drerup et al. (2017).

We study the impact of historic return information on subjective beliefs of future stock market returns. We account for nonclassical measurement error in the outcome variable by applying our sieve rank method and contrast the results to a model where we simply ignore measurement error in the outcome.

We use novel data from the innovation sample of the 2017 wave of the German Socio Economic Panel (SOEP-IS), which contains survey questions on individual beliefs on future stock market returns. In the interviews, respondents are asked their expectations on the DAX, Germany's prime blue chip stock market index, in one, two, ten and thirty years with respect to the current level. They are asked to provide a direction of the change (increase or decrease) as well as a percentage change.

Prior to elicitation of their beliefs, individuals obtain information about historical DAX returns. Two observations of the time series of yearly DAX returns from 1951 to 2016 are randomly drawn and presented to the respondent. Afterwards they are asked to report their beliefs on how the DAX changes in the next year (in percentage points).

In this application, we are interested in the effect of the historical DAX information on the individuals expected DAX return in one year. Let $Y^*$ denote the individual true belief on the DAX return in one year and let $Z_1, Z_2$ be the two treatment variables, i.e., the randomly drawn historical returns. The reported belief is

| | Min. | 1. Quant | Median | Mean | 3. Quant. | Max. |
|---|---|---|---|---|---|---|
| $Y$ | -50.00 | 1.00 | 4.00 | 3.55 | 7.00 | 130.00 |
| $Z_1$ | -43.94 | -6.08 | 11.36 | 14.77 | 29.06 | 116.06 |
| $Z_2$ | -43.94 | -6.08 | 13.99 | 17.13 | 34.97 | 116.06 |

Table 1.2: Summary Statistics (all units are percentage points)

denoted by $Y$. We consider the following flexible additively separable model

$$Y^* = g_1(Z_1) + g_2(Z_2) + g_3(Z_1 \cdot Z_2) + U, \quad \text{where } Z \perp\!\!\!\perp U. \tag{1.8}$$

It is difficult to rationalize a classical measurement error assumption a priori. Various forms of nonclassical measurement error may occur in this setting: (i) Respondents may tend to provide rounded values instead of precise beliefs, (ii) respondents may systematically over- or underreport their beliefs, e.g., individuals with extreme beliefs may resort to reporting more modest values, or (iii) the reporting may additionally depend on variables $W$ such as certain cognitive skills or personality traits like patience or perseverance. Note that by the experimental design $Z_1, Z_2$ and $W$ are credibly fully independent so we subsume any effect of $W$ on $Y^*$ in $U$ and consider an unweighted version of the sieve rank estimator

We now discuss the plausibility of Assumptions 1-3 required for identification. Assumptions 1 posits that given true beliefs $Y^*$ and relevant individual characteristics $W$, the historic return information $Z_1, Z_2$ have no impact on the mean reported belief. Assumption 2 imposes a mild restriction on the measurement error mechanism in that it requires monotonicty in the reporting of beliefs (in the conditional mean). Assumption 3 is satisfied as $Z_1, Z_2$ are by the experimental setup credibly fully independent of unobservables $U$. The data consists of 1084 interviewed persons but 306 people do not respond to the question on beliefs. We removed missing values and report the summary statistics in Table 1.2.

We estimate functions $g_1$, $g_2$, $g_3$ with our method outlined in (1.6) and contrast the results to estimates obtained from assuming classical measurement error, i.e., from a standard additive-separable, nonparametric regression of $Y$ on $Z_1$ and $Z_2$ with the respective interaction term. We choose a B-Spline basis of degree two without interior knots for each function estimate. This choice is motivated by a 10-fold cross-validation on the model ignoring the measurement error.

The results are presented in Figure 1.3. Accounting for the measurement error leads to a concave, symmetric effect of both treatments on the individual beliefs. When ignoring the possibility of measurement error, results are much more asym-

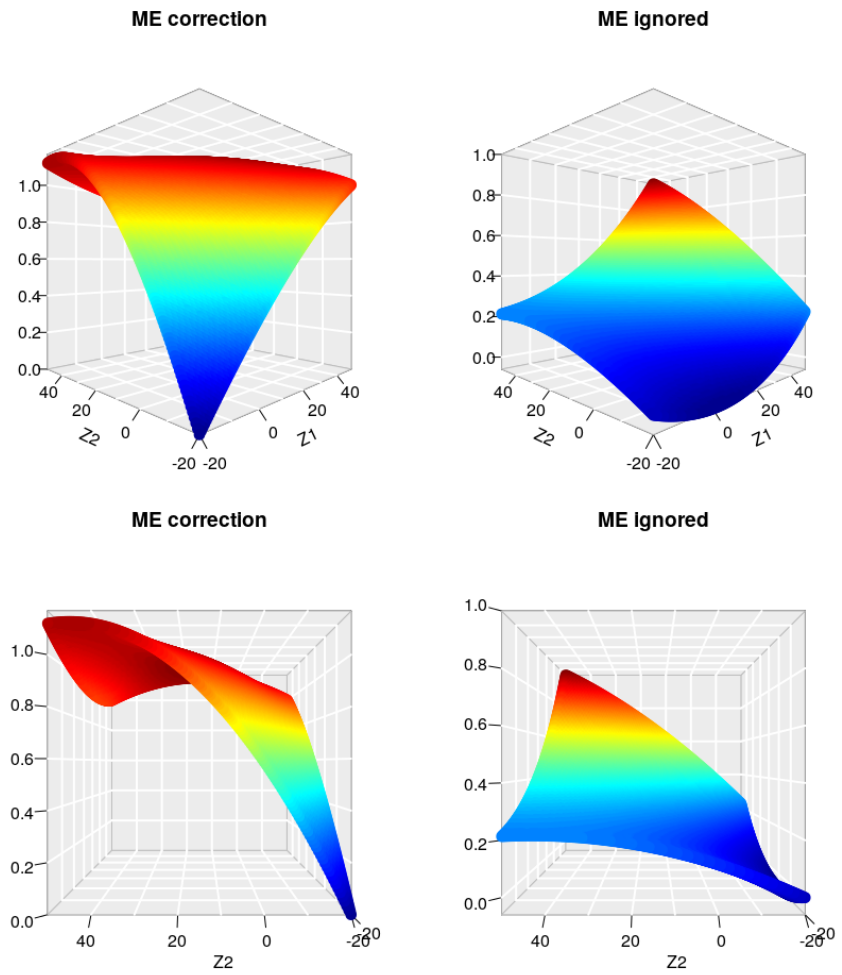


Figure 1.3: Nonparametric estimates of $g(Z_1, Z_2) = g_1(Z_1) + g_2(Z_2) + g_3(Z_1, Z_2)$. The first column contains the estimate from our sieve rank estimator and the second column the estimate from ignoring measurement error.

metric, including convex marginals for the first treatment and flat parts in the surface. In contrast, our method yields that individuals learn conservatively from both treatments which is in line with the a priori economic intuition. Note that on the z-axis that estimates in both columns have been normalized to move through coordinates (-20,-20,0) and (50,50,1). Functions are evaluated on a grid ranging from -20 to 50 which corresponds to the 10%- and 90%-quantile of the marginal distributions of the treatment variables. Summarizing, accounting for possible non-classical measurement error in the outcome variable delivers function estimates of belief formation that are more in line with economic intuition.

## 1.6 Conclusion

This paper provides new insights on the analysis of regression models with nonclassical measurement error in the outcome variable. Our nonparametric identification result is based on intuitive assumptions involving shape restrictions on measurement error functions. This novel result builds on the equivalence of nonclassical measurement models and generalized regression models. We consider a sieve rank estimator which constructively arises from our identification result and implicitly accounts for the required shape restrictions. We establish the rate of convergence of the sieve rank estimator which is affected by a potentially ill-posed inverse problem. The proposed estimation method is easy to implement and provides numerically stable results as demonstrated in a finite sample analysis. Finally, we demonstrate the usefulness of our method in an empirical application on belief elicitation, where we find measurement error in subjective belief data to be of a nonclassical form.

## 1.7 Supplemental Material

This Supplemental Material consists of an extension of our sieve rank estimator to continuous control variables and proofs of our theoretical results. Appendix 1.7.1 provides an extension to weighted sieve rank estimation, when control variables are continuous. All proofs are postponed to the Appendix 1.7.3.

### 1.7.1 Extension: Estimation with Continuous $W$

When $W$ does contain continuous variables, we can simply replace the indicator in (1.5) with a kernel function to account for the fact that $W_i = W_j = w$ is a null event. Then estimation can proceed with

$$\widehat{g}_w = \arg\max_{\phi \in \mathcal{G}_K} \mathcal{Q}_n(\phi, w) \text{ where} \tag{1.9}$$
$$\mathcal{Q}_n(\phi, w) := \sum_{1 \le i < j \le n} Y_i \mathcal{K}_s(W_i - w)\mathcal{K}_s(W_j - w)\mathbb{1}\{\phi(Z_i) > \phi(Z_j)\}$$

where $K_h$ is defined as

$$\mathcal{K}_s(W_i - w) = \prod_{l=1}^{d_w} \mathcal{K}\left(\frac{W_{l,i} - w}{s_l}\right)$$

and $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ is some kernel function and $s \in \mathbb{R}^{d_w}$ a vector of bandwidths.

As we move from the original criterion of Cavanagh and Sherman (1998) to

the conditional version with continuous $W$ the computational complexity of the maximization problem increases. Ranking is an $O(n\log(n))$ operation whereas the weighted ranking is performed in $O(n^2)$ time. This implies that the conditional estimation method is not scalable to large data sets and computation time increases heavily with the sample size.

The following criterion can be used to deal with continuous W and computation time scales in $n$.

$$\begin{aligned}\mathcal{Q}_n(\phi, w) &= \sum_{1\leq i<j\leq n} \mathcal{K}_s^U(W_i - w) Y_i \mathcal{K}_s^U(W_j - w) \mathbb{1}\{\phi(Z_i) > \phi(Z_j)\} \\ &= \sum_{i:\ w-s<W_i<w+s} Y_i \text{Rank}_s(\phi(Z_i)) \end{aligned} \tag{1.10}$$

with uniform kernel

$$\mathcal{K}_s^U(W_i - w) := \mathbb{1}\{w - s < W_i < w + s\}$$

which is again equivalent to applying the sieve rank estimator over a subsample of the data obtained by considering a window of size $2s$ around $w$. Weighted rank estimation is studied in Shin (2010) and Abrevaya and Shin (2011) for semiparametric and additively separable models. An important special case is again the setting where the function $g(\cdot, w)$ does not vary with $w$ which is the case of $g$ is additvely separable in a function of $Z$ and $W$.

**Remark 1.2.** *Assume the function $g(Z)$ does not depend on $W$. We can consider the following estimator*

$$\begin{aligned}\widehat{g} &= \arg\max_{\phi\in\mathcal{G}_K} Q_n(\phi) \quad \textit{where} \\ Q_n(\phi) &:= \sum_{1\leq i<j\leq n} Y_i \mathcal{K}_h(W_i - W_j) \mathbb{1}\{\phi(Z_i) > \phi(Z_j)\}\end{aligned}$$

*In contrast to before we consider only those observations in a neighborhood around a fixed value $w$ but we choose the weights according to which distance any pair $(W_i, W_j)$ has to each other. Similar to the approach in (3.5) this is associated with increasing computational complexity as the computation time does not scale with the sample size.*

*We thus suggest the following strategy:*

*First use the criterion in (1.10) to obtain estimates $\widehat{g}_w$ across different values of $w \in supp(W)$. Each is an estimate of $g$ as $g$ does not depend in theory on $w$, but estimation results may nevertheless vary for different $w$. Second, aggregate*

*the different estimates $\widehat{g}_w$ to one final estimator for $g$. To this end, we can follow Chiappori et al. (2015) which discuss the following two 'aggregation' procedures.*

$$\widehat{g}_{LS}(z) = \arg\min_{q\in\mathbb{R}} \int_{supp(W)} \nu(w)[\widehat{g}(z,w) - q]^2 dw$$
$$\widehat{g}_{LAD}(z) = \arg\min_{q\in\mathbb{R}} \int_{supp(W)} \nu(w)|\widehat{g}(z,w) - q| dw$$

*where $\nu$ is some weighting function with $\int_{supp(W)} \nu(w)dw = 1$.*

*The implementation is simple. Random draws from $\{W_i\}_{i=1}^N$ yields a set of different realizations $w$ on which to evaluate the local estimators $\widehat{g}_w$. The LS criterion takes the average of the local estimators, the LAD criterion takes the empirical median to aggregate to a final estimator for $g$. In simulations Chiappori et al. (2015) find that the latter estimator performs best as for $w$ in the tails of the distribution of $W$ we may get erratically behaving $\widehat{g}_w$.*

### 1.7.2 Weighted Rank Estimation

In this section we assess the performance of a weighted rank estimator for a setting as described in Remark 1.2. We consider the following data generating process similar to Section 1.4,

$$Y^* = Z_1 + g(Z_2) + m(W) + U \cdot W^2$$
$$Y = h(Y^* + W) + V \cdot |W|$$

where $g(\cdot) = \sin(\cdot)$, $m(\cdot) = cos(\cdot)$, $W = 0.5 \cdot Z_2 + 0.5 \cdot U$ and the remaining variables as in Section 1.4 with $h$ parameterized by $a = b = 0$. In this setting there is correlation between $Z_2$ and $W$. Further the measurement is additionally affected by the variable $W$. This setting is in line with Remark 1.2 as $g$ does not vary with $W$, and we implement the procedure outlined at the end of this remark with the LAD-criterion as aggregating procedure.

In order to calculate an estimate of $g$ for each Monte Carlo sample, we first take 50 random draws of the variable $W$, calculate $\widehat{g}_w$ by maximizing (1.10) for each of the 50 different realizations $w$. Finally, we aggregate the results to a final estimate by taking the sample median over the local estimates $\widehat{g}_w$. We vary the bandwidth parameter $\widetilde{s}$ across different experiments. The sample size is $n = 1000$ and 500 Monte Carlo replications are considered. The following Figure 1.4 shows the results.

If we choose $s$ reasonably small, our estimation procedure is quite close to the truth and outperforms the standard nonparametric estimator that simply ignores

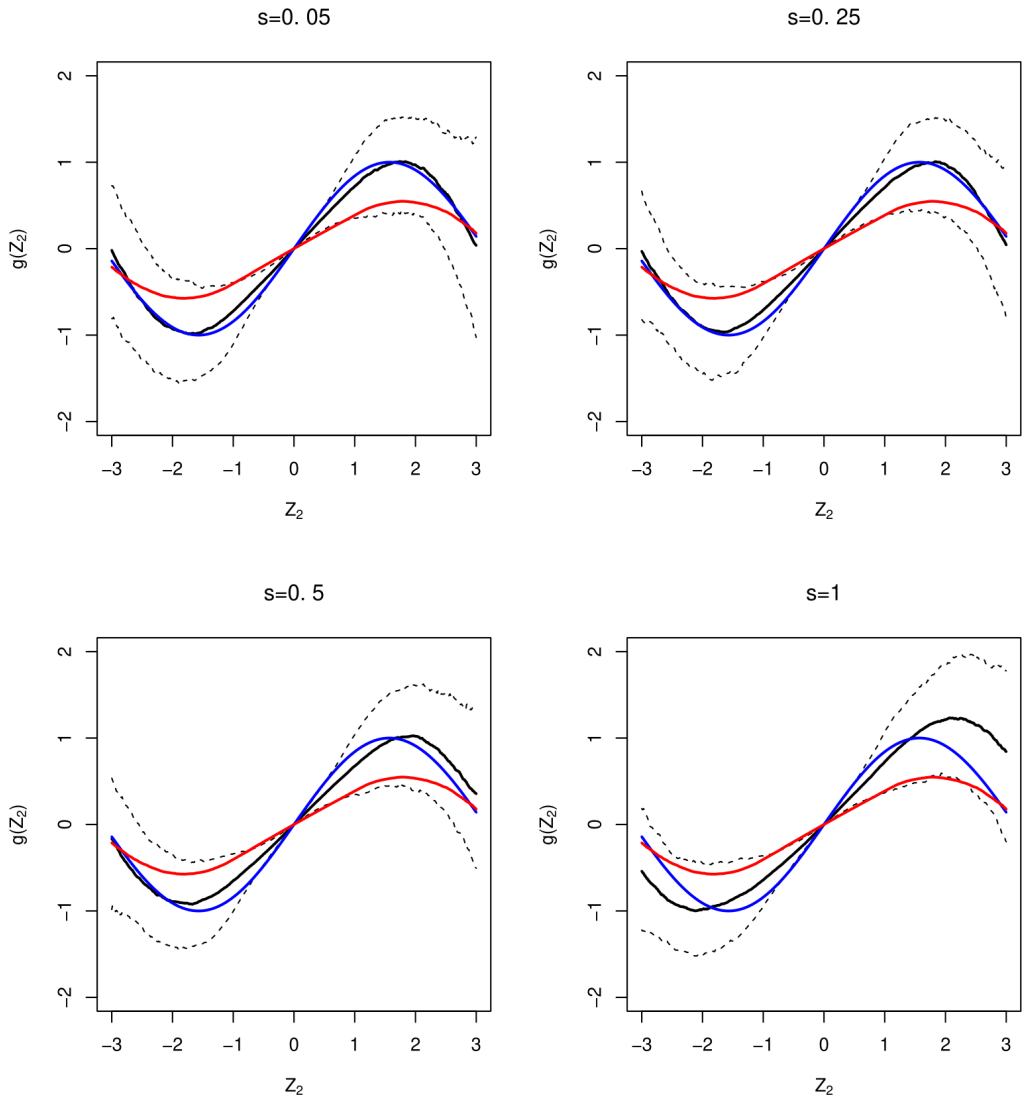


Figure 1.4: The blue line is the $g(\cdot) = \sin(\cdot)$ function, the solid black line denotes the median and the dotted lines the respective 0.95 and 0.05 quantiles of the weighted sieve rank estimator over the Monte Carlo experiments. The red line is the median of series estimates of $g$ in the model $Y = Z_1 + g(Z_2) + m(W) + U$. Basis functions are set as in Section 1.4 with $K = 4$.

the measurement error. Increasing the bandwidth $s$ leads to smaller confidence bands, but considerably increases the bias of the estimate. However in this strong measurement error setting, the weighted sieve rank estimator still outperforms the estimate from ignoring the measurement error.

### 1.7.3 Proofs and Technical Results

First, recall that $X = (Z, W)$ and that $g_w = g(\cdot, w)$.

Proof of Theorem 1.1. Proof of (1.3). The exclusion restriction captured in Assumption 1 implies

$$\begin{aligned}\mathbf{E}[Y|X = x] &= \mathbf{E}[h(Y^*, W) \mid Z = z, W = w] \\ &= \mathbf{E}[h(g(Z, W) + U, W) \mid Z = z, W = w] \\ &= \mathbf{E}[h(g(z, W) + U, W) \mid W = w], \end{aligned} \tag{1.11}$$

where the last equation is due to the conditional exogeneity imposed in Assumption 3. The results follows from strict montonicity of $H(g_w(z), w) = \mathbf{E}[h(g(z, W) + U, W) \mid W = w]$ in its first argument, which is due to Assumption 2 and Assumption 4 (ii).

Proof of (1.4). By the law of iterated expectations, the criterion function $Q(\phi, w) = \mathbf{E}[Y_1 \mathbb{1}\{\phi(X_1) > \phi(X_2)\}|W_1 = W_2 = w]$ can be rewritten as

$$\begin{aligned}Q(\phi, w) =& \frac{1}{2}\mathbf{E}[H(g(X_1), W_1)\mathbb{1}\{\phi(X_1) > \phi(X_2)\} \\ &+ H(g(X_2), W_2)\mathbb{1}\{\phi(X_1) < \phi(X_2)\} \mid W_1 = W_2 = w],\end{aligned}$$

using $\mathbf{E}[Y|X] = H(g(X), W)$ by equation (1.3). Under Assumption 2, we may consider the case that holds with $h(\cdot, w)$ weakly monotonically increasing, without loss of generality. Now the function $g_w$ is a maximizer of $Q(\cdot, w)$, which follows by

$$Q(g_w, w) = \frac{1}{2}\mathbf{E}[\max\{H(g(X_1), W_1), H(g(X_2), W_2)\} \mid W_1 = W_2 = w]$$

and using monotonicity of $H$ in its first argument. In particular, $m \circ g_w$ is a maximizer of $Q(\cdot, w)$ for any strictly increasing function $m$ (here $\circ$ denotes function composition).

It remains to show that $g_w$ is a unique maximizer up to strictly increasing transformations. Specifically, we show that for any function $\widetilde{g}_w \neq m \circ g_w$ for an arbitrary strictly monotonic transformation $m$ we have that $Q(\widetilde{g}_w, w) < Q(g_w, w)$. To do so, consider some arbitrary function $\phi \in \mathcal{G}$ that is not a strictly monotonic transformation of $g_w$. Therefore, there exist $z', z'' \in supp(Z)$ such that $g_w(z') < g_w(z'')$ and $\phi(z') > \phi(z'')$. By (1.3), $H(\cdot, w)$ is strictly monotonic and it holds for every $w$ that

$$H(g_w(z'), w) < H(g_w(z''), w).$$

By continuity of the functions following Assumption 4 (i) the above inequalities hold in neighborhoods $B_1$ around $z'$ and $B_2$ around $z''$, respectively. By Assumption 4 (iii) these neighborhoods have a strictly positive probability measure. This implies

$$\begin{aligned} &Q(g_w, w) - Q(\phi, w) \\ \geq &\frac{1}{2}\,\mathbf{E}[H(g_w(Z_1), W_1) - H(g_w(Z_2), W_2) | Z_1, Z_2 \in B_1 \times B_2, W_1 = W_2 = w] \\ &\times \mathbb{P}(Z_1, Z_2 \in B_1 \times B_2 \mid W_1 = W_2 = w) > 0. \end{aligned}$$

Thus, $Q(\cdot, w)$ is only maximized by $g_w$ and strictly monotonic transformations of it. Hence, $g_w$ is identified up to a strictly monotonic transformation. □

Proof of Corollary 1.1. Under Assumption 5 (i) any candidate regression function $\widetilde{g}_w(Z) = \widetilde{m}_w(Z_1) + \widetilde{l}_w(Z_{-1})$ must satisfy

$$\widetilde{g}_w(Z) = M_w(g_w(Z)) = M_w(m_w(Z_1) + l_w(Z_{-1}) = \widetilde{m}_w(Z_1) + \widetilde{l}_w(Z_{-1})$$

for a strictly monotonic function $M_w$. Thus $M_w$ must be linear and $g_w$ is identified up to location and scale transformation. Indeed, given linear and strictly monotonic transformations, $g_w$ is the only maximizer of $Q(\cdot, w)$. Under Assumption 5 (ii) we have that $g_w(z_1) = \mathbf{E}[Y|Z = z_1, W = w]$ and $g_w(z_2) = \mathbf{E}[Y|Z = z_2, W = w]$ and fixing the parameter space to move through both points leads to $g_w$ being the unique maximizer of $Q(\cdot, w)$ over $\mathcal{G}$ and thus $g_w$ is point identified. □

Proof of Lemma 1.1. Let $Z_1, Z_2$ be independent copies of $Z$. Consider the additive separable case $g(Z_1) = Z_{11} + \widetilde{g}(Z_{12})$ with bivariate $Z_1 = (Z_{11}, Z_{12})$. Analogously we denote $\phi(Z_1) = Z_{11} + \widetilde{\phi}(Z_{12})$. The following holds for the criterion $\mathcal{Q}$

$$\begin{aligned} |\mathcal{Q}(\phi)| =& \mathbf{E}[Y_1(\mathbb{1}\{Z_{11} + \widetilde{g}(Z_{12}) > g(Z_2)\} - \mathbf{E}[Y_1 \mathbb{1}\{Z_{11} + \widetilde{\phi}(Z_{12}) > \phi(Z_2)\}] \\ =& \mathbf{E}[Y_1(F_{Z_{21}|Z_{22}}(\phi(Z_1) - \widetilde{\phi}(Z_{21})) - F_{Z_{21}|Z_{22}}(g(Z_1) - \widetilde{g}(Z_{21})))], \end{aligned}$$

as $g$ is the maximizer of $\mathcal{Q}$ and with the second equation due to the law of iterated expectation. Using a second-order Taylor decomposition with directional derivatives yields for all $\phi$ in a neighborhood around $g$

$$|\mathcal{Q}(\phi)| = Q_g(\phi - g) + \underbrace{\mathbf{E}[Y_1 f''_{Z_{21}|Z_{22}}(\xi)(\widetilde{\phi}(Z_{12}) - \widetilde{g}(Z_{12}) + \widetilde{g}(Z_{22}) - \widetilde{\phi}(Z_{22}))^3]}_{=R},$$

where $\xi$ is some intermediate variable[7] and $Q_g$ denotes the directional derivative of $\mathcal{Q}$ at $g$ which is given by

$$Q_g(\phi - g) = \mathbf{E}[Y_1 f'_{Z_{21}|Z_{22}}(g(Z_1) - \widetilde{g}(Z_{21}))(\widetilde{\phi}(Z_{12}) - \widetilde{g}(Z_{12}) + \widetilde{g}(Z_{22}) - \widetilde{\phi}(Z_{22}))^2].$$

Applying the Cauchy-Schwarz inequality to $Q_g(\phi - g)$ shows that $Q_g$ is weaker than the $L^2$-norm. Further, the remainder term $R$ satisfies

$$|R| \leq \mathbf{E}\left[\left|\frac{f''_{Z_{21}|Z_{22}}(\xi)}{f'_{Z_{21}|Z_{22}}(g(Z_1) - \widetilde{g}(Z_{21}))}(\widetilde{\phi}(Z_{12}) - \widetilde{g}(Z_{12}) + \widetilde{g}(Z_{22}) - \widetilde{\phi}(Z_{22}))\right|\right] \cdot Q_g(\phi - g)$$

and thus the tangential cone condition in Assumption 6 (iv) is satisfied if the first factor on the right hand side is bounded between 0 and 1. The lower bound holds directly and the upper bound is easily satisfied if the $\delta-$ neighborhood around $g$ is chosen sufficiently small and derivatives of the density are bounded away from zero and infinity, as is condition. □

For the proof of the next results, we require some additional notation to deal with the Hoeffding decomposition of U-statistics, specific function spaces and their respective envelope functions.

We introduce the empirical criterion $\mathcal{Q}_n(\phi)$ that can be denoted as

$$\mathcal{Q}_n(\phi) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \Gamma(S_i, S_j, \phi)$$

where $S_i = (Y_i, Z_i)$ and which is a second order U-statistic with kernel

$$\Gamma(S_i, S_j, \phi) = Y_i\big(\mathbb{1}\{\phi(Z_i) > \phi(Z_j)\} - \mathbb{1}\{g(Z_i) > g(Z_j)\}\big)$$

indexed by $\phi \in \mathcal{G}_K$ making it a second-order U-process. Note that $\mathcal{Q}_n$ is centered here which does not affect the optimization. Using the kernel notation, the criterion function $\mathcal{Q}$ given in (1.7) satisfies $\mathcal{Q}(\phi) = \mathbf{E}[\Gamma(S_i, S_j, \phi)]$.
For the asymptotic analysis we make use of the Hoeffding decomposition of a U-statistic (see e.g. van der Vaart (1998))

$$\mathcal{Q}_n(\phi) = \mathcal{Q}(\phi) + \nu_n(\phi) + \xi_n(\phi) \tag{1.12}$$

[7] More precisely $\xi = g(Z_1) - \widetilde{g}(Z_{22}) + s[\phi(Z_1) - \widetilde{\phi}(Z_{21}) + \widetilde{g}(Z_{21}) - g(Z_1)]$ for some $s \in (0, 1)$.

with short hand notations

$$\begin{aligned}
\nu_n(\phi) &:= \frac{1}{n}\sum_{i=1}^{n}\nu(S_i,\phi),\\
\nu(S_i,\phi) &:= \mathbf{E}[\Gamma(S_i,S_j,\phi)|S_i] + \mathbf{E}[\Gamma(S_j,S_i,\phi)|S_i] - 2\,\mathbf{E}[\Gamma(S_i,S_j,\phi)],\\
\xi_n(\phi) &:= \frac{2}{n(n-1)}\sum_{1\le i<j\le n}\xi(S_i,S_j,\phi),\\
\xi(S_i,S_j,\phi) &:= \Gamma(S_i,S_j,\phi) - \mathbf{E}[\Gamma(S_i,S_j,\phi)|S_i] - \mathbf{E}[\Gamma(S_i,S_j,\phi)|S_j] + \mathbf{E}[\Gamma(S_i,S_j,\phi)].
\end{aligned}$$

This decomposition is frequently deviced in the rank estimation literature to obtain asymptotic results, see e.g. Sherman (1993). The first summand in the decomposition is a smooth function of the parameter $\phi$, $\nu_n$ is an empirical process and $\xi_n$ a degenerate U-process, both indexed by the function space $\mathcal{G}_K$. Further, we define the function classes $\mathcal{F}_{\nu,K} = \{\nu(\cdot,\phi) : \phi \in \mathcal{G}_K^\delta\}$ and $\mathcal{F}_{\xi,K} = \{\xi(\cdot,\cdot,\phi) : \phi \in \mathcal{G}_K^\delta\}$. Let $\overline{F}_\nu$ and $\overline{F}_\xi$ denote respective envelope functions . The envelope function is defined as any function satisfying $|\nu(\cdot,\phi)| \le \overline{F}_\nu(\cdot)$.

In this setting, $\overline{F}_\nu(S_i) = |Y_i| + 3\,\mathbf{E}[|Y_i|]$, since

$$\begin{aligned}
|\nu(S_i,\phi)| =& |Y_i\,\mathbf{E}[\mathbb{1}\{\phi(Z_i) > \phi(Z_j)\} - \mathbb{1}\{g(Z_i) > g(Z_j)\}|Z_i]\\
&+ \mathbf{E}[Y_j\,(\mathbb{1}\{\phi(Z_j) > \phi(Z_i)\} - \mathbb{1}\{g(Z_j) > g(Z_i)\})\,|Z_i]\\
&- 2\,\mathbf{E}[Y_i\,(\mathbb{1}\{\phi(Z_i) > \phi(Z_j)\} - \mathbb{1}\{g(Z_i) > g(Z_j)\})]|\\
\le& |Y_i| + 3\,\mathbf{E}[|Y_i|],
\end{aligned}$$

where $\left\|\overline{F}_\nu\right\|_{L^2(S)} \le \sqrt{4\,\mathbf{E}[Y^2]} =: C_\nu$. In addition we have $\overline{F}_\xi(S_i,S_j) = 2|Y_i| + 2\,\mathbf{E}[|Y_i|]$ as

$$\begin{aligned}
|\xi(S_i,S_j,\phi)| =& |Y_i(\mathbb{1}\{\phi(Z_i) > \phi(Z_j)\} - \mathbb{1}\{g(Z_i) > g(Z_j)\})\\
&- Y_i\,\mathbf{E}[\mathbb{1}\{\phi(Z_i) > \phi(Z_j)\} - \mathbb{1}\{g(Z_i) > g(Z_j)\}|Z_i]\\
&- \mathbf{E}[Y_i(\mathbb{1}\{\phi(Z_i) > \phi(Z_j)\} - \mathbb{1}\{g(Z_i) > g(Z_j)\})|Z_j]\\
&+ \mathbf{E}[Y_i(\mathbb{1}\{\phi(Z_i) > \phi(Z_j)\} - \mathbb{1}\{g(Z_i) > g(Z_j)\})]|
\end{aligned}$$

and $\left\|\overline{F}\right\|_{L^2(S)} \le \sqrt{12\,\mathbf{E}[Y^2]} =: C_\eta$. By Assumption 6 (iii) we have $C_\nu, C_\eta < \infty$. Ultimately, we define the bracketing integral $J_{[]}$ of the space $\mathcal{F}_{\nu,K}$

$$J_{[]}(1,\mathcal{F}_{\nu,K},L^2(S)) = \int_0^1 \sqrt{1 + \log N_{[]}(\epsilon\cdot\left\|\overline{F}_\nu\right\|_{L^2(S)},\mathcal{F}_{\nu,K},L^2(S))}d\epsilon.$$

and analogously for $\mathcal{F}_{\xi,K}$.

PROOF OF THEOREM 1.2. We begin by noting that consistency of $\widehat{g}$ in the $L^2$-norm follows from Lemma 1.3. Due to the consistency result in Lemma 1.3, we may restrict the function spaces to a local neighborhood around $g$, i.e. we define the space $\mathcal{G}_K^\delta = \{\phi \in \mathcal{G}_K : \|\phi - g\|_{L^2(Z)} < \delta\}$ and assume that $\widehat{g} \in \mathcal{G}_K^\delta$. Further we introduce the space $\mathcal{G}_K^{\delta,r_n} = \{\phi \in \mathcal{G}_K^\delta : Q_g(\phi - g) > Mr_n\}$ where $M > 0$. It holds that

$$\begin{aligned}\mathbb{P}\big(Q_g(\widehat{g} - g) \geq Mr_n\big) &\leq \mathbb{P}\left(\sup_{\phi \in \mathcal{G}_K^{\delta,r_n}} \mathcal{Q}_n(\phi) \geq \mathcal{Q}_n(\Pi_K g)\right) \\ &\leq \mathbb{P}\left(\sup_{\phi \in \mathcal{G}_K^{\delta,r_n}} \mathcal{Q}(\phi) + \nu_n(\phi) + \xi_n(\phi) \ \geq \mathcal{Q}(\Pi_K g) + \nu_n(\Pi_K g) + \xi_n(\Pi_K g)\right),\end{aligned}$$

by applying the Hoeffding decomposition (1.12). Due to Assumption 6 (iv) we have local equivalence of $|\mathcal{Q}(\cdot)|$ and $Q_g(\cdot)$. Since $\mathcal{Q}(\cdot)$ is negative and thus $|\mathcal{Q}(\cdot)| = -\mathcal{Q}(\cdot)$ it follows that

$$\begin{aligned}&\mathbb{P}\big(Q_g(\widehat{g} - g) \geq Mr_n\big) \\ \leq &\mathbb{P}\left(\sup_{\phi \in \mathcal{G}_K^{\delta,r_n}} \Big(\mathcal{Q}(\phi) + \nu_n(\phi) - \nu_n(\Pi_K g) + \xi_n(\phi) - \xi_n(\Pi_K g))\Big) \geq -\eta Q_g(\Pi_K g - g)\right) \\ \leq &\mathbb{P}\left(\sup_{\phi \in \mathcal{G}_K^{\delta,r_n}} \Big(\nu_n(\phi) - \nu_n(\Pi_K g) + \xi_n(\phi) - \xi_n(\Pi_K g) + \eta Q_g(\Pi_K g - g)\Big) \geq \inf_{\phi \in \mathcal{G}_K^{\delta,r_n}} |\mathcal{Q}(\phi)|\right) \\ \leq &\mathbb{P}\left(\sup_{\phi \in \mathcal{G}_K^{\delta}} \nu_n(\phi) - \nu_n(\Pi_K g) + \sup_{\phi \in \mathcal{G}_K^{\delta}} \xi_n(\phi) - \xi_n(\Pi_K g) + \eta Q_g(\Pi_K g - g) \geq C_2 M r_n\right),\end{aligned}$$

where it remains to study the asymptotic behavior of each summand in the last line separately. Note that both summands on the left hand-side are positive, hence if $\sup_{\mathcal{G}_K^\delta} \nu_n(\phi)$ is bounded in probability so is $\nu_n(\Pi_K g)$ and similarly for $\xi_n$.

First we study the asymptotic behavior of the empirical process part $\sup_{\phi \in \mathcal{G}_K^\delta} \nu_n(\phi)$. Recall the definition $\mathcal{F}_{\nu,K} = \{\nu(\cdot, \phi) : \phi \in \mathcal{G}_K^\delta\}$ with envelope $\overline{F}_\nu$. By applying the last display of Theorem 2.14.2 of van der Vaart and Wellner (2000) we can conclude that

$$\mathbf{E}\Big|\sup_{\phi \in \mathcal{G}_K^\delta} \nu_n(\phi)\Big| = \mathbf{E}\Big|\sup_{\nu \in \mathcal{F}_{\nu,K}} \frac{1}{n}\sum_{i=1}^n \nu(S_i)\Big| \leq J_{[]}(1, \mathcal{F}_{\nu,K}, L^2(S)) \cdot \big\|\overline{F}_\nu\big\|_{L^2(S)} \cdot n^{-1/2}$$

where $\big\|\overline{F}_\nu\big\|_{L^2(S)} \leq \big\|\overline{F}_\nu\big\|_{L_\infty(S)} \leq C_\nu < \infty$. By Lemma 1.2 (i) and (ii) we have

$$\log N_{[]}(\epsilon \cdot \big\|\overline{F}_\nu\big\|_{L_\infty(S)}, \mathcal{F}_{\nu,K}, L_\infty(S)) \leq c_0 K \log(1/\epsilon \cdot C_\nu^{-1})$$

and ultimately we obtain $J_{[]}(1, \mathcal{F}_{\nu,K}, L_\infty(S)) = O(\sqrt{K})$ and by Markov's inequality $\sup_{\phi \in \mathcal{G}_K^\delta} \nu_n(\phi) = O_p(\sqrt{K/n})$.

It remains to analyze the convergence rate of the degenerate U-process $\sup_{\phi \in \mathcal{G}_K^\delta} \xi_n(\phi)$. Similar to Lemma A.1 in Clemencon et al. (2008) we can make use of the following equality for second-order U-statistics

$$\frac{1}{n(n-1)} \sum_{i \neq j} \xi(S_i, S_j, \phi) = \frac{1}{n!} \sum_{\pi} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \xi(S_i, S_{\lfloor n/2 \rfloor + i}, \phi) \tag{1.13}$$

where $\pi$ is short-hand for all permutations of $\{1, \ldots, n\}$. Then applying the triangle inequality to (1.13) leads to

$$\mathbf{E}\left[ \Big| \sup_{\phi \in \mathcal{G}_K^\delta} \frac{1}{n(n-1)} \sum_{i \neq j} \xi(S_i, S_j, \phi) \Big| \right] \leq \mathbf{E}\left[ \Big| \sup_{\phi \in \mathcal{G}_K^\delta} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \xi(S_i, S_{\lfloor n/2 \rfloor + i}, \phi) \Big| \right] \tag{1.14}$$

from which we can conclude that for obtaining the convergence rate of the degenerate U-process on the left-hand side of (1.14) it is sufficient to analyze the convergence rate of an empirical process with kernel $\xi$ indexed by the function $\mathcal{G}_K^\delta$.

The kernel $\xi$ contains non-smooth indicator functions so we cannot apply the exact same reasoning we used earlier to derive a bound for $\nu_n$, as $\xi(S_i, S_j, \phi)$ is not continuous in $\phi$. However we can use the fact that $\xi(\cdot, \cdot, \phi)$ belongs to a VC-subgraph family and we can thus derive the complexity bound in Lemma 1.2 (iii).

Recall the definition $\mathcal{F}_{\xi,K} = \{\xi(\cdot, \cdot, \phi) : \phi \in \mathcal{G}_K^\delta\}$ and the associated envelope function $\overline{F}_\xi$. Now we apply Theorem 2.14.1 of van der Vaart and Wellner (2000)

$$\mathbf{E}\left[ \Big| \sup_{\phi \in \mathcal{G}_K^\delta} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \xi(S_i, S_{\lfloor n/2 \rfloor + i}, \phi) \Big| \right] \leq J_{[]}(1, \mathcal{F}_{\xi,K}, L^2(S)) \big\| \overline{F}_\xi \big\|_{L^2(S)} \lfloor (n/2) \rfloor^{-1/2}$$

Applying Lemma 1.2 (iii) we obtain the bound

$$J_{[]}(1, \mathcal{F}_{\xi,K}, L^2(S)) \leq \int_0^1 \sqrt{1 + c_1 + c_2 K \log(1/\epsilon)} d\epsilon = O(\sqrt{K})$$

and by Markov's inequality that $\sup_{\phi \in \mathcal{G}_K^\delta} \xi_n(\phi) = O_p(\sqrt{K/n})$. Finally, we can conclude that

$$\mathbb{P}\left(Q_g(\widehat{g} - g) \geq M r_n\right) \leq \mathbb{P}\left( \sup_{\phi \in \mathcal{G}_K^\delta} \nu_n(\phi) + \sup_{\phi \in \mathcal{G}_K^\delta} \xi_n(\phi) + Q_g(\Pi_K g - g) \geq C_2 M r_n \right).$$

with $\sup_{\phi\in\mathcal{G}_K^\delta}\nu_n(\phi) = O_p(\sqrt{K/n})$ and $\sup_{\phi\in\mathcal{G}_K^\delta}\xi_n(\phi) = O_p(\sqrt{K/n})$. Consequently, choosing $r_n = \max\{\sqrt{K/n}, Q_g(\Pi_K g - g)\}$ we see that the right hand side probability converges to zero as $M \to \infty$. Thus $Q_g(\widehat{g} - g) = O_p(r_n)$. By the definition of the sieve measure of ill-posedness $\tau_K$ we obtain

$$\begin{aligned}\|\widehat{g} - g\|_{L^2(Z)} \leq \tau_K Q_g(\widehat{g} - g) \leq& \tau_K O_p\left(\max\{\sqrt{K/n}, Q_g(\Pi_K g - g)\}\right)\\ =& O_p\left(\tau_K\sqrt{K/n}, \|\Pi_K g - g\|_{L^2(Z)}\right)\end{aligned}$$

which concludes the proof. □

**Lemma 1.2.** *Under Assumption 6 it holds that*
*(i)* $\sup_{\|\phi-g\|_\infty\leq\delta}|\nu(S_i,\phi)| \leq M_1(S_i)\cdot\delta$ *with* $\mathbf{E}[M_1(S_i)] < \infty$,
*(ii)* $\log N_{[]}(\epsilon, \mathcal{F}_{\nu,K}, L_\infty(S)) \leq c_0 K\log(1/\epsilon)$ *for some positive constant* $c_0$,
*(iii)* $\log N(\epsilon, \mathcal{F}_{\xi,K}, L^2(S)) \leq c_1 + c_2 K\log(1/\epsilon)$, *for positive constants* $c_1, c_2$.

Proof of Lemma 1.2. Proof of part (i). It holds that

$$\begin{aligned}\nu(S_i,\phi) =& Y_i\,\mathbf{E}[\mathbb{1}\{\phi(Z_i) > \phi(Z_j)\} - \mathbb{1}\{g(Z_i) > g(Z_j)\}|Z_i]\\ &+ \mathbf{E}[Y_j\,(\mathbb{1}\{\phi(Z_j) > \phi(Z_i)\} - \mathbb{1}\{g(Z_j) > g(Z_i)\})\,|Z_i]\\ &- 2\,\mathbf{E}[Y_i\,(\mathbb{1}\{\phi(Z_i) > \phi(Z_j)\} - \mathbb{1}\{g(Z_i) > g(Z_j)\})]\end{aligned}$$

We make use of the fact that as $\|\phi - g\|_\infty \leq \delta$ and thus $g(z) - \delta \leq \phi(z) \leq g(z) + \delta$ for any $z$ in the support of $Z$. Following Chen et al. (2003) (p. 1599-1600) we have that

$$\begin{aligned}&\sup_{\|\phi-g\|_\infty\leq\delta}|\mathbb{1}\{\phi(Z_j) < \phi(Z_i)\} - \mathbb{1}\{g(Z_j) < g(Z_i)\}|\\ \leq&|\mathbb{1}\{g(Z_j) < \phi(Z_i) + \delta\} - \mathbb{1}\{g(Z_j) < g(Z_i) - \delta\}|\end{aligned}$$

and thus

$$\begin{aligned}|\nu(S_i,\phi)| \leq& |Y_i|\cdot|F_{g(Z)}(\phi(Z_i)+\delta) - F_{g(Z)}(g(Z_i)-\delta)|\\ &+ |\,\mathbf{E}[Y_i|Z_i]|\cdot|F_{g(Z)}(\phi(Z_i)+\delta) - F_{g(Z)}(g(Z_i)-\delta)|\\ &+ |\,\mathbf{E}[Y_i]|\cdot\mathbf{E}[|F_{g(Z)}(\phi(Z_i)+\delta) - F_{g(Z)}(g(Z_i)-\delta)|]\\ \leq&(|Y_i| + |\,\mathbf{E}[Y_i|Z_i]| + |\,\mathbf{E}[Y_i]|)\cdot 3\delta\end{aligned}$$

where the last inequality follows from Assumption 6 (v), the Lipschitz continuity for the cdf of $g(Z)$. Define $M_1(S_i) = |Y_i| + |\,\mathbf{E}[Y_i|Z_i]| + |\,\mathbf{E}[Y_i]|$. From Assumption 6 (iii) follows that $\mathbf{E}[M_1(S_i)] < \infty$ which concludes the argument.

We continue with the proof of part (ii). By Lemma 1.2 (i) we have

$$\log N_{[]}(\epsilon, \mathcal{F}_{\nu,K}, L_\infty(S)) \leq \log N_{[]}(\epsilon, \mathcal{G}_K, L_\infty(Z)) \leq cK \log(1/\epsilon)$$

where both inequalities are due to Chen (2007) (pp. 5595 and 5601).

We conclude with the proof of part (iii). We make use of the decomposition $\xi(S_i, S_j, \phi) = \xi_1(S_i, S_j, \phi) + \xi_2(S_i, S_j, \phi)$ where $\xi_1(S_i, S_j, \phi) = \Gamma(S_i, S_j, \phi)$ and

$$\xi_2(S_i, S_j, \phi) = -\mathbf{E}[\Gamma(S_i, S_j, \phi)|S_i] - \mathbf{E}[\Gamma(S_i, S_j, \phi)|S_j] + \mathbf{E}[\Gamma(S_i, S_j, \phi)].$$

Following for instance Nolan and Pollard (1987, Lemma 16) we conclude

$$\log N(\epsilon, \mathcal{F}_{\xi,K}, L^2(S)) \leq \log N(\epsilon, \mathcal{F}_{\xi_1,K}, L^2(S)) + \log N(\epsilon, \mathcal{F}_{\xi_2,K}, L^2(S)).$$

Similar to the proof of part (ii) of Lemma 1.2 we obtain $\log N(\epsilon, \mathcal{F}_{\xi_2,K}, L^2(S)) \leq cK \log(1/\epsilon)$ for some constant $c$. Below, we follow Chapter 5 of Sherman (1993) to establish that $\mathcal{F}_{\xi_1,K}$ belongs to a VC-subgraph class. To this end define the subgraph

$$\begin{aligned}
&\text{subgraph}\,(\xi_1(\cdot,\cdot,\phi)) \\
=&\{(s_i, s_j, t) \in \mathit{supp}(S)^2 \times \mathbb{R} : 0 < t < y_i[\mathbb{1}\{\phi(z_i) > \phi(z_j)\} - \mathbb{1}\{g(z_i) < g(z_j)\}]\} \\
=&\{y_i > 0\}\{\phi(z_i) - \phi(z_j) > 0\}\{t > 0\}\{t < \overline{F}_{\xi_1}(z_i, z_j)\}\{g(z_i) - g(z_j) < 0\} \\
&\cup \{y_i < 0\}\{\phi(z_i) - \phi(z_j) < 0\}\{t > 0\}\{t < \overline{F}_{\xi,1}(z_i, z_j)\}\{g(z_i) - g(z_j) > 0\}
\end{aligned}$$

and introduce the function

$$m(t, s_1, s_2; \gamma_1, \gamma_2, \pi_1, \pi_2) := \gamma_1 t + \gamma_2 y_1 + (g(z_1), p^K(z_2))'\pi_1 + (g(z_2), p^K(z_2))'\pi_2$$

with the associated function space

$$\mathcal{M} = \{m(\cdot,\cdot,\cdot;\gamma_1,\gamma_2,\pi_1,\pi_2) : \gamma_1 \in \mathbb{R}, \gamma_2 \in \mathbb{R}, \pi_1 \in \mathbb{R}^{K+1}, \pi_2 \in \mathbb{R}^{K+1}\}.$$

Note that $\mathcal{M}$ is a finite vector space of dimension $2(K+2)$ and the subgraph can be written as

$$\text{subgraph}\,(\xi_1(\cdot,\cdot,\phi)) = \bigcup_{i=1}^{10}\{m_i > 0\} \tag{1.15}$$

with functions $m_i \in \mathcal{M}$ for any $i = 1, \ldots, 10$. Following e.g. Lemma 2.4 and 2.5 in Pakes and Pollard (1989) it can be established that subgraph $(\xi_1(\cdot,\cdot,\phi))$ belongs to a VC-class of sets and thus the space $\mathcal{F}_{\xi_1}$ is a VC-class of functions. To bound the

complexity of the space we require the VC-index of $\mathcal{F}_{\xi_1}$ which we denote as $V(\mathcal{F}_{\xi_1}) = V(\text{subgraph}(\xi_1))$. From Pollard (1984, Lemma 18) it follows that $V(\{m_i > 0\}) \leq 2(K+2)$. Applying in van der Vaart and Wellner (2009, Theorem 1.1) to (1.15) then leads to $V(\text{subgraph}(\xi_1)) \lesssim 2(K+2)$, so the VC-index of the space $\mathcal{F}_{\xi_1}$ increases with the same order as the sieve dimension $K$. Now applying van der Vaart (1998, Theorem 2.6.7) yields

$$\begin{aligned}\log N(\epsilon, \mathcal{F}_{\xi_1,K}, L^2(S)) &\leq \log(C \cdot V(\mathcal{F}_{\xi_1})(16e)^{V(\mathcal{F}_{\xi_1})}(1/\epsilon)^{2V(\mathcal{F}_{\xi_1})-2}) \\ &= \log(C) + \log(2(K+2)) \\ &\quad + 2(K+2)\log(16e) + 2(K+2)\log(1/\epsilon)\end{aligned}$$

and together with $\log N(\epsilon, \mathcal{F}_{\xi_2,K}, L^2(S)) \leq cK\log(1/\epsilon)$ the stated result follows. □

**Lemma 1.3.** *Under Assumptions 1–6 it holds that* $\|\widehat{g} - g\|_{L^2(Z)} = o_p(1)$.

Proof of Lemma 1.3. We need to check the conditions in Lemma A.2 of Chen and Pouzo (2012). In their notation $\overline{Q}_n = \mathcal{Q}$ and

$$g_0(k, n, \epsilon) = \inf_{\phi \in \mathcal{G}_K : \|\phi - g\|_{L^2(Z)} \geq \epsilon} |\mathcal{Q}(\phi)|$$

Their condition a is thus satisfied and $g_0(n, k, \epsilon) > 0$ by the identification result in Theorem 1.1. Condition b holds by Assumption 6 (ii) and the fact that for large enough $K$ the following holds

$$|\mathcal{Q}(\Pi_K g) - \mathcal{Q}(g)| \lesssim Q_g(\Pi_K g - g) \lesssim \tau_K^{-1}\|\Pi_K g - g\|_{L^2(Z)},$$

and thus $\mathcal{Q}(\Pi_K g) - \mathcal{Q}(g) = o(1)$. Next, Condition c is implicitly assumed to hold and it remains to check condition d which translates as

$$\frac{\max\{|\mathcal{Q}(\Pi_K g) - \mathcal{Q}(g)|, \sup_{\phi \in \mathcal{G}_K} |\mathcal{Q}_n(\phi) - \mathcal{Q}(\phi)|\}}{g_0(n, k, \epsilon)} = o(1).$$

Analogous to the empirical process result from (1.13) and (1.14) and the subsequent proceedings, it holds that $\sup_{\phi \in \mathcal{G}_K} |\mathcal{Q}_n(\phi) - \mathcal{Q}(\phi)| \lesssim \sqrt{K/n}$. Then ultimately consider that for any $\epsilon > 0$ there is some $\epsilon^* > 0$ that is sufficiently small such that the

local equivalence relation in Assumption 6 (iv) is valid and we can conclude

$$\begin{aligned} g_0(k,n,\epsilon) = \inf_{\mathcal{G}_K:\|\phi-g\|_{L^2(Z)}\geq\epsilon} |\mathcal{Q}(\phi)| &\geq \inf_{\mathcal{G}_K:\|\phi-g\|_{L^2(Z)}\geq\epsilon^*} Q_g(\phi-g) \\ &\geq \inf_{\mathcal{G}_K:\|\phi-g\|_{L^2(Z)}\geq\epsilon^*} \tau_K^{-1}\|\phi-g\|_{L^2(Z)} \\ &\geq \tau_K^{-1}\epsilon^*. \end{aligned}$$

In summary we require that

$$\begin{aligned} &\max\{|\mathcal{Q}(\Pi_K g)-\mathcal{Q}(g)|, \sup_{\phi\in\mathcal{G}_K}|\mathcal{Q}_n(\phi)-\mathcal{Q}(\phi)|\}\big/ g_0(n,k,\epsilon) \\ \lesssim &\tau_K \max\{\sqrt{K/n}, \tau_K^{-1}\|\phi-g\|_{L^2(Z)}\} = o(1), \end{aligned}$$

which follows from the rate restriction in Assumption 6 (vi). □

# Chapter 2

# Estimation of Conditional Random Coefficient Models using Machine Learning Techniques

## 2.1 Introduction

In recent years microeconometric models aimed at capturing complex heterogeneity in individual behavior. In particular, it has become relevant to include observed and/or unobserved heterogeneity when modeling (average) partial effects in regression models.

An important model that accounts very flexibly for such heterogeneity is the nonparametric random coefficient model

$$Y = B_0 + B_1 W,$$

where $(B_0, B_1)$ is a vector of $p + 1$ random variables and $W \in \mathbb{R}^p$ is a vector of regressors. Let $X$ be a set of additional control variables that may be related to $(W, B_0, B_1)$. $B_1$ is the individual effect of a change in $W$ on $Y$ and the distribution of $B_1$ reflects the heterogeneity of individual effects in the population.

The model is nonparametric in that it does not impose distributional assumptions on partial effects $B_1$ and the nuisance $B_0$. The primary goal for this class of models is to identify and estimate the entire distribution of the vector of random coefficients $(B_0, B_1)$ such as the joint density function. The marginal density of the random slope parameter $B_1$ is of special interest in economic applications as $B_1$ can be interpreted as an average partial effect of a change in $W$ on the outcome. Its distribution reflects the heterogeneity of this effect in the underlying population. If $W$ is an exogenous

treatment then the expected value of $B_1$ corresponds to an average treatment effect, its conditional expectation to a conditional average treatment effect and so on.

A crucial identifying assumption in the literature is full independence of covariates $W$ and random coefficients $(B_0, B_1)$. This is satisfied if $W$ is randomly assigned, such as in experimental data settings, and the marginal slope density captures the entire heterogeneity of a partial effect in the population. This knowledge does not allow to link the heterogeneity to any observable characteristics. For instance, the shape of the RC-density may vary across observable individual characteristics. Learning the RC-density conditional on a set of control variables $X$ provides additional insight on how the shape of the heterogeneity varies across subpopulations with differing observable characteristics. This allows to disentangle heterogeneity in *observable* and *unobservable* heterogeneity.

Furthermore, when dealing with observational data there is always room for a potential dependence between $W$ and control variables $X$. The random intercept $B_0$ subsumes the effects of $X$ on $Y$ which violates full independence between $W$ and random coefficients. In this work the identifying restriction can be weakened to allow for conditional independence, i.e. $B \perp\!\!\!\perp W|X$, which corresponds to a selection-on-observables assumption. In most economic applications the set of control variables $X$ is of considerable size or even high-dimensional and there is generally no prior knowledge about which variables in $X$ drive heterogeneous partial effects. Modern Machine Learning (ML) methods allow to deal with large dimensional set of controls and perform some form of variable selection to identify those elements of $X$ inducing different shapes of heterogeneity. Recently, ML-techniques have proven useful in relating features of the distribution of partial effects $B_1$ to additional observable characteristics, such as in the estimation of conditional average treatment effects, also referred to as heterogeneous treatment effects. In this work, I link the entire distribution of random coefficients to observable characteristics by studying a *conditional random coefficient model*. This allows to uncover (i) which variables generally drive heterogeneity in partial effects and (ii) how the distribution of partial effects varies across subpopulations with different observable characteristics.

First I begin by providing an identification statement for the conditional RC-density. Deriving from this identification statement I can formulate a sieve approximation to the RC-density conditional on a fixed value of $X$. This sieve approximation has a closed form expression and each sieve coefficient can be expressed as a conditional expectation function of some nonlinear transformation of $Y$ and $W$ varying with controls $X$.

Generic ML-methods can be used to estimate this set of conditional expectation functions. Various practical considerations to be outlined later require to orthogonalize both the outcome $Y$ and treatment $W$ using ML technqiues before estimating the sieve coefficients itself. This requires the use of iterated sample splitting to deal with nested ML-steps.

Following the outline of the estimation strategy I derive the $L_2$-convergence rate of the final conditional RC-density estimator. This convergence rate is crucially determined by the asymptotic properties of the ML-methods employed. Assuming that the slowest ML-estimator used converges at a polynomial rate the $L_2$-convergence rate of the RC-density estimator is slower by some factor than that of the slowest ML-estimator employed. The factor hinges on the degree of ill-posedness of the underlying inverse problem and the overall smoothness of the density.

In addition to the asymptotic properties of the estimator, I introduce a cross-validation strategy to inform the choice of tuning parameters.

Finally, I apply the estimator to study behavioral heterogeneity in an economic experiment on portfolio choice. Survey respondents of the german socio-economic panel (SOEP) are asked to invest an hypothetical monetary amount into a riskfree asset or into a risky asset with payoff depending on the return of a stock market index. I study the effect of stock market beliefs on the investment decision with a random coefficient model. I find a bi-modal RC-density that reflects the presence of two types in the population. One type complies with economic theory and stock market beliefs have a positive impact on the amount invested in the risky asset. For a second type this is not the case and the effect is centered around zero. The division into types prevails when varying the values of controls. This suggests that the assignment to types is largely based on unobservable characteristics not in the data. An exemption here is age, as for a subpopulation of elder respondents the share of non-compliers is substantially larger.

**Related literature** Identification and estimation of the nonparametric random coefficient model is studied in Beran and Hall (1992), Beran and Millar (1994), Beran et al. (1996) and Hoderlein et al. (2010). See also Masten (2017) for a refined identification result. All of these works operate under the assumption that random coefficients and regressors are fully independent. Breunig (2021) considers a so called *varying random coefficient model*, where each random coefficient is made up of a nonparametric function of control variables and an additively separable random component which is fully independent of controls. Further, the number of control variables affecting random coefficients is effectively smaller than the number

of random coefficients itself whereas in the present setting the number of controls is allowed to be larger. Sieve estimation for random coefficient models is used for the testing procedure in Breunig and Hoderlein (2018) and in Breunig (2021).

Machine Learning estimation and econometric models have been paired frequently in recent years. Chernozhukov et al. (2015), Chernozhukov et al. (2017) and Chernozhukov et al. (2020) study estimation and inference on parameters and linear functionals of parameters in high-dimensional linear models. Thereby highlighting the importance of iterated ML estimation and sample splitting for achieving consistency and asymptotic normality of parameter estimates.

Of particular importance for this work is the estimation of heterogeneous treatment effects using machine learning methods as considered in Wager and Athey (2018) and Athey et al. (2019), see also the references therein. This is due to the fact that a heterogeneous treatment effect can be viewed as the conditional expectation function of a random slope coefficient. In contrast, the conditional RC-density studied here is informative about the entire distribution of a treatment effect in a given (sub-)population. This also includes learning the form of *unobservable* heterogeneity which remains otherwise unknown when only the mean of a random coefficient is studied. Chernozhukov et al. (2019) considers identification and estimation of particular subfeatures of the conditional expectation function of a random slope.

The theory of the conditional RC density estimate developed here holds for generic machine learning techniques. However I make use of the causal forest algorithms of Athey et al. (2019) and the popular ML-tool of random forests introduced by Breiman (2001) in the implementation of the estimator and in the asymptotic theory. The asymptotic theory of random forest estimators has been studied in Scornet et al. (2015), Wager and Walther (2016), Wager and Athey (2018) and Athey et al. (2019).

The remainder of this paper is organized as follows. Section 2.2 introduces the main model and discusses identification of conditional RC-densities. Section 2.3 outlines the estimation strategy. Section 2.4 presents the asymptotic properties of the estimator and section 2.5 asymptotic inference on the conditional RC density estimates. Section 2.6 addresses auxiliary topics, i.e. marginal density estimation, variable importance measures and the choice of tuning parameters. Section 2.7 contains a Monte Carlo simulation study. Section 2.8 contains an empirical application of the estimation procedure using survey data. Section 2.9 concludes.

## 2.2 Model Setup and Identification

This paper considers the following random coefficient model

$$Y = B_0 + B_1 \cdot W \tag{2.1}$$

where $B = (B_0, B_1)$ consists of two scalar random variables and $W$ is a scalar regressor of interest to the researcher. Here $B_1$ is the average partial effect of a change in $W$ on the outcome $Y$. Within the model framework there exists a vector of additional covariates $X \subseteq \mathbb{R}^d$ that may affect both the random coefficients as well as the regressor $W$. The goal of this section is to give conditions under which we can identify the conditional random coefficient density $f_{B|X=x}$. That is the random coefficient density for a given subpopulation with characteristics $X = x$. Note that the results of this paper can be readily extended to cover the case where $W$ is multivariate, though for conciseness, I focus on the scalar $W$ case which is of the most practical relevance.

The model (2.1) can be interpreted as a reduced form of a more general multivariate random coefficient model with the random intercept absorbing all the (heterogenous) effects of other covariates on the outcome. Without loss of generality, the random coefficients satisfy

$$B_j = g_j(X) + A_j, \quad \text{where} \quad \mathbf{E}[A_j \mid X] = 0, \quad j = 0, 1 \ ,$$

which illustrates that (2.1) nests many popular mean regression models. For instance it can be viewed as an extension of Robinson (1988) with a random coefficient instead of a deterministic one.
For obtaining identification the following assumptions are imposed.

**Assumption 1.** *(i) $B \perp\!\!\!\perp W \mid X$ (ii) for every $x$ in the support of $X$ the random variable $W \mid X = x$ has full support $\mathbb{R}$.*

Assumption 1 (i) requires that the regressor of interest $W$ is independent of random coefficients conditional on controls $X$. This restriction allows for some dependence between $B$ and $W$ and is thereby weaker then the full independence assumption typically encountered in random coefficient models, see Beran and Hall (1992), Hoderlein et al. (2010) and Masten (2017). It can also be interpreted as an exogeneity condition on the regressor $W$. If $W$ is a (quasi-)experimental intervention then (i), or more precisely the part $B_0 \perp\!\!\!\perp W \mid X$, corresponds to a selection on observables assumption. It is also one of the main assumptions for identifying heterogenous treatment effects as in the RC model in section 6 of Athey et al. (2019).

Assumption 1 (ii) strengthens the common large support restriction from the random coefficients literature, see e.g. Hoderlein et al. (2010) to also hold conditional on $X = x$. The assumption rules out the case where $W$ is a deterministic function of $X$ and may be problematic if otherwise certain realizations $x$ provide strong information about $W$. A workaround for this assumptions is provided by the varying RC model of Breunig (2021). There, however, functional form restrictions on the random coefficients need to be made, which outlines a tradeoff between the above full support assumption and further restrictions on the random coefficient model. If this assumption is violated for some realizations of $X$ then, nevertheless, identification for different $x$-values which satisfy (ii) can be established. Masten (2017) discusses identification in the case of bounded support of the regressor $W$. Taking this results into account, the following identification result can be formulated.

**Lemma 2.1.** *If Assumption 1 holds, then for every $x$ in the support of $X$ the density function $f_{B|X=x}$ is identified. If $W \mid X = x$ has instead only compact support then $f_{B|X=x}$ is point identified if and only if the distribution of $B \mid X = x$ is determined solely by its moments and all absolute moments are finite.*

The proof of Lemma 2.1 immediately follows from extending classical identification results for random coefficient models as synthesized in Masten (2017) to the conditional case.

## 2.3 Estimation of Conditional Random Coefficient Densities

This section introduces the estimation strategy for conditional RC-densities. Subsection 2.3.1 outlines the principal idea and introduces the main notation whereas Subsection 2.3.2 is of most practical relevance. It discusses a demeaned random coefficient model and provides further details on machine learning estimators and sample splitting rules. Before moving forward the following general notation needs to be introduced. Let

$$\phi_{Y|X}(t|x) = \mathbf{E}[\exp(itY) \mid X = x]$$

denote the characteristic function of $Y$ conditional on $X = x$. The Fourier transformation $\mathcal{F}$ and the inverse Fourier transformation $\mathcal{F}^{-1}$ are defined as

$$(\mathcal{F}f)(t) = \int_{\mathbb{R}^d} \exp(it'a) f(a) da$$
$$(\mathcal{F}^{-1}g)(a) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \exp(-ia't) g(t) dt$$

for some functions $f, g : \mathbb{R}^d \to \mathbb{R}$. The operators $\mathcal{F} : \mathbb{R}^d \to \mathbb{C}^d$ and $\mathcal{F}^{-1} : \mathbb{C}^d \to \mathbb{R}^d$ relate the characteristic function of a random variable to its probability density function, given the latter exists. For some random variable $A$ with density $f_A$ it holds that $\phi_A(t) = (\mathcal{F}f_A)(t)$ and vice versa $(\mathcal{F}^{-1}\phi_A)(a) = f_A(a)$.

### 2.3.1 A Two-Stage Sieve Estimation Approach

The essential implication of Assumption 1 that will be leveraged for estimation is the identity

$$(\mathcal{F}f_{B|X=x})(t, tw) = \phi_{Y|X,W}(t|x, w),$$

which holds for every $x$ in the support of $X$ and $t, w \in \mathbb{R}$. The identity in turn implies the following $L_2$-condition

$$\int_{\mathbb{R}^2} \left| (\mathcal{F}f_{B|X=x})(t, tw) - \phi_{Y|X,W}(t|x, w) \right|^2 d\nu(t) d\mu(w) = 0, \tag{2.2}$$

where $v, \mu$ are arbitrary probability measures on $\mathbb{R}$ that are discussed later in more detail. Following Breunig (2021) the $L_2$-criterion in (2.2) can be used to construct a sieve estimator of the density $f_{B|X=x}$.

To this end, let $q^K = (q_1, \ldots, q_K)$ denote a $K = K(n)$-dimensional vector of known basis functions that span the linear sieve space $\mathcal{B}_K = \{\phi(\cdot) = q^K(\cdot)'\pi\}$. As $f_{B|X=x}$ is bivariate, $q^K$ typically is a tensor product of univariate basis functions, i.e. $q^K(b_0, b_1) = q^{K_1}(b_0) \otimes q^{K_2}(b_1)$ with $K = K_1 \cdot K_2$.

If the characteristic function $\phi_{Y|X,W}(\cdot|x, w)$ were known, then a sieve estimator of the conditional random coefficient density is

$$\widetilde{f}_{B|X}(\cdot|x) = \arg\min_{\phi \in \mathcal{B}_K} \int_{\mathbb{R}^2} \left| (\mathcal{F}\phi)(t, tw) - \phi_{Y|X,W}(t|x, w) \right|^2 d\nu(t) d\mu(w),$$

which has the following closed form expression

$$\widetilde{f}_{B|X}(b|x) = q^K(b)'Q^{-1}\int_{\mathbb{R}^2}(\mathcal{F}q^K)(-t,-tw)\phi_{Y|X,W}(t|x,w)d\nu(t)d\mu(w) \tag{2.3}$$

and where

$$Q = \int_{\mathbb{R}^2}(\mathcal{F}q^K)(t,tw)(\mathcal{F}q^K)'(-t,-tw)d\nu(t)d\mu(w). \tag{2.4}$$

Note that the estimator in (2.3) is not feasible as $\phi_{Y|X,W}(t|x,w)$ is not known. Breunig (2021) proceeds by replacing the unknown characteristic function with a nonparametric plug-in estimate. A general problem in this setting is the presence of the possibly large-dimensional set of controls $X$ that cannot be reduced a priori in most practical applications. Breunig (2021) studies a varying random coefficient model which puts additional structure on the relationship of random coefficients and controls and where $X$ is low-dimensional.

Modern Machine Learning (ML-)estimators are well suited for the estimation of conditional expectation functions like $\phi_{Y|X,W}(t|x,w) = \mathbf{E}[\exp(itY)|X=x,W=w]$ in the presence of possibly high-dimensional controls $X$. However, an additional issue arises in that we would require to perform different Machine Learning steps over a continuum of values for $t$.

In order to enable the use of Machine Learning techniques, a different strategy is needed. Further rearranging of (2.3) yields

$$\begin{aligned}
&\widetilde{f}_{B|X}(b|x)\\
=&q^K(b)'Q^{-1}\int_{\mathbb{R}}\int_{\mathbb{R}}\mathbf{E}[(\mathcal{F}q^K)(-t,-tW)\exp(itY)|X=x,W=w]d\nu(t)d\mu(w)\\
=&q^K(b)'Q^{-1}\int_{\mathbb{R}}\mathbf{E}[\int_{\mathbb{R}}(\mathcal{F}q^K)(-t,-tW)\exp(itY)d\nu(t)|X=x,W=w]d\mu(w)
\end{aligned}$$

$$=q^K(b)'Q^{-1}\int_{\mathbb{R}}\mathbf{E}[T(W,Y)|X=x,W=w]d\mu(w), \tag{2.5}$$

where the operator $T(w,y) := \int_{\mathbb{R}}(\mathcal{F}q^K)(-t,-tw)\exp(ity)d\nu(t)$ is short-hand for the nonlinear mapping $T : \mathcal{W}\times\mathcal{Y}\to\mathbb{R}^K$. The operator $T$ can be computed via numeric integration and thus, I consider it deterministic for the remainder of this paper. Also for clarification, let $T(W,Y) = (T_1(W,Y),\dots,T_K(W,Y))$ with functions $T_k : \mathcal{W}\times\mathcal{Y}\to\mathbb{R}$, for $k=1,\dots,K$. Further, define the functions

$$\pi_k(x,w) = \mathbf{E}[T_k(W,Y)|X=x,W=w]$$

for $k=1,\dots,K$ with $\pi(x,w) = (\pi_1(x,w),\dots,\pi_K(x,w)$. A distinguishing feature of the sieve approximation in (2.5) is that sieve coefficients are the relevant quantity that varies in $X$.

Finally, I construct a feasible estimator by replacing $Q$ with a sample mean and $\pi(x,w)$ with a vector of Machine Learning estimates. The resulting RC-density estimator is

$$\widehat{f}_{B|X}(b|x) = q^K(b)'\widehat{Q}^{-1}\int_{\mathbb{R}}\widehat{\pi}(x,w)d\mu(w), \tag{2.6}$$

where $\widehat{\pi}$ is a generic ML-estimate of the unknown function $\pi$ and

$$\widehat{Q} = \frac{1}{n}\sum_{i=1}^{n}(\mathcal{F}q^K)(N_i,N_iM_i)(\mathcal{F}q^K)'(-N_i,-N_iM_i),$$

where the $(M_i,N_i)$'s are $n$ Monte Carlo draws from the probability distributions $\mu$ and $\nu$ that are specified by the researcher. In general $Q$ can be calculated directly via numerical methods for most measures $\mu,\nu$, however this representation is introduced here, as we will later consider the case where $\mu$ is the distribution of $W$ and use sample realizations $W_i$ instead of the generated $M_i$.

Notice that (2.6) is a direct estimate of the closed-form sieve projection in (2.3). The sieve coefficients can be expressed in terms of different conditional expectation functions which can in turn be conveniently estimated by generic machine learning routines even if the set of controls $X$ is high-dimensional.

**Choice of weighting measures** The choice of the measure $\mu$ leaves room for further simplification of the estimator in (2.6). If we choose $f_{W|X=x}$ as the density of the weighting measure $\mu$, then it holds that $\int_{\mathbb{R}}\pi(x,w)d\mu(w) = \mathbf{E}[T(W,Y)|X=x]$ and we only need to consider ML-estimation of a conditional expectation function

and there is no need for additional weighting of this ML-estimator. This does slightly ease computation and simplifies the asymptotic analysis as asymptotic properties of ML-estimators of conditional expectation functions are readily available. The properties of the transformed ML-estimator $\int_{\mathbb{R}} \widehat{\pi}(x,w)d\mu(w)$ used to calculate the estimator in (2.6) have not been studied explicitly.

A caveat of choosing $f_{W|X=x}$ is that the matrix $Q$ will vary in $x$, which is problematic from both the computational as well as the theoretical stance[1]. A possible workaround is to orthogonalize $W$, which I will elaborate on in detail in the next section. However, this workaround will slow down the convergence of the RC-density estimator, as will be shown in section 2.4. Hence, choosing $d\mu/dw = f_{W|X=x}$ as weighting measure is problematic.

In any case, choosing a $\mu$ that is related to the distribution of $W$ is appropriate. ML-estimators like $\widehat{\Pi}(x,w)$ perform best for points from the center of the distribution and will be less accurate in the tails. Moving forward, I will focus on the case where $d\mu(w)/dw = f_W(w)$, which automatically weighs down areas where the ML-estimates may be less accurate. This respects the finite sample behavior of each machine learned sieve coefficient and reduces the problem of estimating $Q$ to a simple sample mean. The additional weighting of the ML-estimate $\widehat{\pi}(x,w)$ is a minor issue compared to the difficulties arising from alternative choices for $\mu$.

Following Breunig (2021), I choose $\nu$ to follow a lognormal$(0,\sigma_t)$ distribution where $\sigma_t > 0$ is then the second tuning parameter to be chosen by the researcher, along with $K$. This particular choice of weighting measure works well in the settings of Breunig (2021) and also in the simulations and applications in this work. Theoretical justifications are given in Breunig and Hoderlein (2018), Breunig (2021) and in section 2.3, but these do not preclude other choices of weighting distributions.

**Choice of sieve basis functions** Throughout this paper, I again follow Breunig (2021) and choose Hermite functions as sieve basis $q^K$.

Hermite functions are a $L_2$-basis and have appealing theoretical properties. They are eigenfunctions of the Fourier transformation and satisfy $\mathcal{F}q_k(a,b) = \sqrt{2\pi}i^{k-1}q_k(a,b)$. This property simplifies the computation of the estimator considerably. A drawback of Hermite functions is that most of the support concentrates around zero even if $K$ is moderately large. Thus, any moderately sized sieve approximation will fail to be a good approximation of a density function that is centered away from zero and/or has a particularly large support. This is a major motivation for considering

[1] Each of the $K^2$ elements of $Q$ would need to be estimated by a ML-step. Further the asymptotic properties of a machine-learned matrix whose dimensions increase with the sample size remain unclear.

a demeaned random coefficient model in the next subsection.

### 2.3.2 Demeaning of Random Coefficients

This section discusses estimation of a demeaned version of the random coefficient model in (2.1). The reason is that if marginal densities of $B_0$ and $B_1$ are centered away from zero, estimation of the bivariate density function with a Hermite function sieve will require a possibly large choice of $K$ and thus prohibitively many ML-steps. The computational cost associated with each ML-step and the general ill-posedness of the RC-density estimation problem lead to a strong preference for a coarse choice of $K$.

Second, as the random slope density is of particular interest in economics, specific ML-routines which provide high quality estimates for the conditional expectation $\mathbf{E}[B_1|X]$ have already been developed, see Wager and Athey (2018) and Athey et al. (2019). By demeaning, we can separate estimation of the conditional expectation from the remaining conditional shape of the RC-density. This enables the use of ML-tools that are tailored to the specific predictive tasks such as the estimators in Athey et al. (2019) for the conditional expectation function of $B_1$. Further, these direct estimators of the conditional expectation will perform better than those one can infer from an indirect estimate via integrating the entire conditional density function.

We can reformulate the original RC-model (2.1),

$$\begin{aligned} Y - \mathbf{E}[Y|X,W] &= A_0 + A_1 \cdot W, \\ \text{where } \; A_0 &= B_0 - \mathbf{E}[B_0|X] \;\text{ and }\; A_1 = B_1 - \mathbf{E}[B_1|X] \end{aligned}$$

and estimate the joint density of the demeaned random coefficients $f_{A|X=x}$ with the procedure outlined in the previous section. To this end, let $\beta(x) = \mathbf{E}[B|X = x]$ with $\beta(x) = (\beta_0(x), \beta_1(x))$ denoting the conditional expectation of the intercept and slope, respectively. Further, let $m(x,w) = \mathbf{E}[Y|X = x, W = w]$. Then, the closed form of the sieve approximation analogous to the previous section is

$$\begin{aligned} \widetilde{f}_{B|X}(b|x) &= \widetilde{f}_{A|X}(b - \beta(x)|x) \\ &= q^K(b - \beta(x))' Q^{-1} \int_{\mathbb{R}} \mathbf{E}[T(W, Y - m(X,W))|X = x, W = w] d\mu(w) \end{aligned} \tag{2.7}$$

with

$$Q = \int_{\mathbb{R}^2} (\mathcal{F}q^K)(t, tw)(\mathcal{F}q^K)'(-t, -tw)d\nu(t)d\mu(w).$$

Further, define

$$\begin{aligned}\Pi(x) &= \int_{\mathbb{R}} \mathbf{E}[T(W, Y - m(X, W))|X = x, W = w]d\mu(w) \\ \Pi_{dm}(x) &= \int_{\mathbb{R}} \mathbf{E}[T(W, Y - \widehat{m}(X, W))|X = x, W = w]d\mu(w)\end{aligned}$$

with $\widehat{m}$ denoting a generic (ML)-estimator for the unknown function $m$. By choosing $d\mu(w)/dw = f_W(w)$, an estimator is

$$\begin{aligned}\widehat{f}_{B|X}(b|x) &= \widehat{f}_{A|X}(b - \widehat{\beta}(x)|x) \\ &= q^K(b - \widehat{\beta}(x))'\widehat{Q}^{-1}\widehat{\Pi}_{dm}(x) \qquad (2.8)\end{aligned}$$

where

$$\widehat{Q} = \frac{1}{n}\sum_{i=1}^{n}(\mathcal{F}q^K)(t, t \cdot W_i)(\mathcal{F}q^K)'(-t, -t \cdot W_i)d\nu(t) \qquad (2.9)$$

and $\widehat{\Pi}_{dm}(x)$ is an ML-estimate of $\Pi_{dm}(x)$. The conditional expectation $\mathbf{E}[T(W, Y - \widehat{m}(X, W))|X = x, W = w]$ can be conveniently estimated with ML methods, but it remains to construct an estimate for the quantity $\int_{\mathbb{R}} \mathbf{E}[T(W, Y - \widehat{m}(X, W))|X = x, W = w]d\mu(w)$. I suggest to use

$$\widehat{\Pi}_{dm}(x) = \frac{1}{R}\sum_{r=1}^{R}\widehat{\mathbf{E}}[T(W, Y - \widehat{m}(X, W))|X = x, W = W_r], \qquad (2.10)$$

where $\widehat{\mathbf{E}}[T(W, Y - \widehat{m}(X, W))|X = x, W = w]$ is an ML-estimator of the respective conditional expectation function and thus $\widehat{\Pi}_{dm}(x)$ is a sample average of different predictions from the ML-estimators over a hold-out sample of size $R$, which has not been left out of the estimation before. Another possibility that does not rely on a hold-out sample is to use a leave-one-out ML-estimator. In the applications and simulation studies I simply calculate (2.10) on the entire sample observations for $W$. This is theoretically not valid, yet in simulations there is practically no difference between using (2.10) on the entire sample for $W$ or an equally-sized hold out sample of $W$.

Therefore, I assume for the remainder of the paper that

$$\widehat{\Pi}_{dm}(x) = \int_{\mathbb{R}} \widehat{\mathbf{E}}[T(W, Y - \widehat{m}(X, W))|X = x, W = w]d\mu(w). \tag{2.11}$$

When studying the asymptotics of (2.11) in the next section, it is implicitly assumed that the rather slow ML-estimators dominate the asymptotic behavior of $\widehat{\Pi}_{dm}$, i.e. convergence of the sample mean to the integral is negligible compared to the convergence of ML-estimators.

The estimator (2.8) nests several ML-estimates and it is therefore apparent that sample splitting is required to achieve consistency.

A particular requirement is that $\widehat{m}$ is calculated on a different sample than $\widehat{\Pi}_{dm}$ which takes $\widehat{m}$ as input.
The use of sample splitting is somewhat mandatory for nested ML-estimators, see Chernozhukov et al. (2015).

The subsequent paragraph outlines the precise use of sample splitting along with a concise summary of the estimation procedure.

**Estimation Procedure**

The sample is $(X_i, W_i, Y_i)$ with $i = 1, \ldots, n$. Set tuning parameters $K$ and $\nu$.

Step 1: Calculate $\widehat{\beta}(x)$ and $\widehat{Q}$ on the full sample.

Step 2: Randomly split the sample in two parts of equal size $n/2$. The two samples are referred to as sample $\mathcal{D}$ and sample $\mathcal{R}$.

Step 3: Use sample $\mathcal{D}$ as training sample to learn $\widehat{m}$ with some ML method.

Step 4: Taking $\widehat{m}$ as given, use sample $\mathcal{R}$ to learn $\widehat{\Pi}_{dm}$ with some ML method.

Then perform cross-fitting, i.e. iterate steps 2-4 a number of $M$ times to obtain $M$ different estimates $\widehat{\Pi}_{dm,m}(x)$ for $m = 1, \ldots, M$. Then aggregate these to a final conditional RC-density estimate

$$\widehat{f}_{B|X}(b|x) = \frac{1}{M} \sum_{m=1}^{M} q^K(b - \widehat{\beta}(x))' \widehat{Q}^{-1} \widehat{\Pi}_{dm,m}(x). \tag{2.12}$$

The cross-fitting procedure is optional, yet highly recommended as it stabilizes the estimates considerably.

In many works linking causal inference with machine learning methods, orthogonalization of treatments $W$ is often mandatory to achieve consistent estimation of causal effects, see e.g. Chernozhukov et al. (2015) or at least desirable for the performance of machine learning methods, see section 6.1.1. in Athey et al. (2019). Therefore, this section ends with a brief discussion on how to handle the case of a demeaned $W$ in the estimation. In the next section, it is shown that orthogonalization of the treatment $W$ is not innocuous in the RC model, as it slows down the convergence rate of the RC-density estimator.

**Remark 2.1.** *Additional orthogonalization of $W$ leads to a random coefficient model*

$$\begin{aligned} Y - \mathbf{E}[Y|X,W] &= A_0 + A_1 \cdot (W - E[W|X]), \\ &\text{where } A_0 = B_0 - \mathbf{E}[B_0|X] + A_1\,\mathbf{E}[W|X] \text{ and } A_1 = B_1 - \mathbf{E}[B_1|X] \end{aligned}$$

*which does not change the interpretation of the random slope. Define $\mathbf{E}[W|X = x] = g(x)$ and the variable $\overline{W} = W - g(X)$, then the estimation of $g$ needs to be taken into account and the following quantities reformulated to*

$$\begin{aligned} &\widetilde{f}_{A|X}(b - \beta(x)|x) \\ =&q^K(b-\beta(x))'Q^{-1}\int_{\mathbb{R}} \mathbf{E}[T(\overline{W}, Y - m(X,W))|X = x, \overline{W} = \overline{w}]d\mu(\overline{w}) \end{aligned}$$

*with*

$$Q = \int_{\mathbb{R}^2} (\mathcal{F}q^K)(t, t\overline{w})(\mathcal{F}q^K)'(-t, -t\overline{w})d\nu(t)d\mu(\overline{w}).$$

*Then the estimation procedure needs to be amended. In Step 3, the function $g$ is estimated additionally by an ML- estimator $\widehat{g}$. In Step 4, we use sample $\mathcal{R}$ to estimate $Q$ as well, in particular*

$$\widehat{Q} = \frac{1}{|\mathcal{R}|}\sum_{i=1}^{|\mathcal{R}|}(\mathcal{F}q^K)(t, t\cdot(W_i - \widehat{g}(X_i)))(\mathcal{F}q^K)'(-t, -t\cdot(W_i - \widehat{g}(X_i)))d\nu(t)$$

## 2.4 Asymptotic Analysis

The following section develops the asymptotic theory of the estimator in (2.12) with its composite parts in (2.9) and (2.11). Estimation in the orthogonalized $W$ case outlined in Remark 2.1 is also considered. The following notation is required. Let $\lambda_{min}(\Omega), \lambda_{\max}(\Omega)$ denote the smallest and largest eigenvalues of a matrix $\Omega$. Further, define the $L_2$-norm $\|g\| = \int |g(a)|^2 da$ and the weighted $L_2$-

norm $\|g\|_{\nu,\mu} = \int |g(t,x)|^2 d\nu(t) d\mu(x)$ for a generic, possibly complex-valued function $g$. To avoid confusion at some points, $\|\cdot\|_E$ denotes the euclidean norm of a (complex) vector. $P_K g$ denotes the $L_2$-projection on a linear sieve space $\mathcal{B}_K$, i.e. $P_K g = \arg\min_{f \in \mathcal{B}_K} \|g - f\|$. The relation $a_n \lesssim b_n$ is shorthand for $a_n \leq C \cdot b_n$ for some constant $C > 0$ and sequences $a_n, b_n$.

The following set of assumptions is necessary.

**Assumption 2.** *(i)* $\sup_{b \in \mathbb{R}^2} \|q^K(b)\| \lesssim \sqrt{K}$ *(ii) the smallest eigenvalue of* $Q$ *satisfies* $\lambda_{\min}(Q) = O(\tau_K)$ *with* $\tau_K \geq 0$ *and* $\tau_K$ *decreasing to zero and* $\lambda_{\max}(\int_{\mathbb{R}^2} q^K(b) q^K(b)' db) = O(1)$ *(iii) for any* $x$ *in the support of* $X$ *we have that* $\|P_K f_{A|X=x} - f_{A|X=x}\| = O(K^{-\alpha})$ *for some* $\alpha > 0$ *and* $\|\mathcal{F} f_{A|X=x} - \mathcal{F} P_K f_{A|X=x}\|_{v,\mu} = O(\tau_K \|P_K f_{A|X=x} - f_{A|X=x}\|)$ *(iv)* $\int_{\mathbb{R}} t^2 d\nu(t) < \infty$ *and* $\sup_{x \in \mathcal{X}} |\Pi_k(x)| \leq C_1$ *for any* $k$ *and some* $C_1 > 0$ *(v) for any* $x$ *in the support of* $X$ *it holds* $\int_{\mathbb{R}^2} \|\nabla f_{A|X}(a|x) da\| \leq C_2$ *for some constant* $C_2 > 0$.

Assumption (2) (i) is satisfied for the most commonly employed sieve bases such as splines, fourier series or wavelets, see e.g. Belloni et al. (2015) as well as for Hermite functions. Sufficient conditions for Assumptions (ii) and (iii) are given in Breunig (2021) in the absence of the measure $\mu$ in (2.2). With the additional measure the eigenvalue decay $\tau_K$ will generally depend on $\mu$, i.e. in my preferred specification on the distribution of $W$. Simulations show that the decay is faster the more light-tailed the distribution of $W$ and the smaller its support. Condition (iii) is a typical assumption on the approximating properties of the basis functions and the parameter $\alpha$ is solely related to the smoothness of the density functions $f_{A|X=x}$ as the dimension of the random coefficient vector is not of interest in this analysis. See e.g. Chen (2007) for a review of approximation properties of various sieve bases across different smoothness classes. The remaining parts (iv) and (v) are standard regularity conditions on the density $f_{A|X=x}$. In particular (iv) imposes that for any $x$ the $L_2$-projection of $f_{A|X=x}$ has bounded coefficients.

**Assumption 3.** *(i) For any* $k = 1, \ldots, K$ *and fixed* $x, w$ *assume that*

$$\max \Big\{ (\widehat{g}(x) - g(x))^2, (\widehat{m}(x,w) - m(x,w))^2, (\widehat{\beta}_0(x) - \beta_0(x))^2, (\widehat{\beta}_1(x) - \beta_1(x))^2, \\ (\widehat{\Pi}_{dm,k}(x) - \Pi_{dm,k}(x))^2 \Big\} = O_p\left(n^{-2\varphi}\right)$$

*(ii)* $K \tau_K^{-1} \log(K) = o(n)$ *(iii)* $K \tau_K \log(K) = o(n^{1-2\varphi})$

Assumption 3 (i) states an abstract upper bound for the pointwise convergence rates of various ML-estimates. Thereby, one can abstract from considering different convergence rates for each ML-estimator by simply focusing on the slowest rate among those ML-estimators employed. Typically, for any ML-method $\varphi < 1/2$.

Part (ii) is a common rate restriction in the series estimation literature to achieve that $\|\widehat{Q}^{-1} - Q^{-1}\| \to 0$, see Belloni et al. (2015). The last part (iii) is an additional rate restriction that is trivially satisfied if $\tau_K^{-1} = K^\gamma$ with $\gamma > 1$, i.e. if the eigenvalue decay $\tau_K$ is sufficiently fast. It holds more generally if $\varphi$ is sufficiently small.

In general, the particular convergence rates depend on factors such as the effective dimension and the smoothness of the conditional expectation function that is to be estimated. An additional aspect is the proper choice of tuning parameters for any ML-method that is applied and the precise notion of high-dimensionality, i.e. the rates at which $\dim(X)$ may go to infinity relative to the sample size. In order to derive these convergence rates, additional restrictions will be required.

Thus, Assumption 3 abstracts from many theoretical and practical details of the ML-techniques employed. However, the complexity of any given ML-method makes the joint parameter choice of our model tuning parameters $K$ and $\sigma_t$ along with other parameters of the ML-routines impractical. Therefore, I suppose that any ML-estimator used has been properly tuned by e.g. data-driven methods to the prediction task at hand. Thus, the rate in Assumption 3 can be viewed as the best rate achievable given a set of ML-estimators that are properly tuned to their respective estimation problem. This abstraction is in line with other works using generic machine learning techniques such as Chernozhukov et al. (2015) or Chernozhukov et al. (2019).

Below, I discuss Assumption 3 in the context of regression forests which will be used in the applied segments of this paper.

**Remark 2.2.** *Suppose estimates for the various functions summarized in Assumption 3 are obtained from applying random forest algorithms. Originally devised by Breiman (2001), random forests are a popular ML-tool among practitioners and several works have since considered the asymptotic properties of random forests. Some theoretical properties like consistency have been established for various tree-growing schemes, see e.g. Biau (2012), Scornet et al. (2015) and Wager and Walther (2016), but the development of theory is ongoing.*

*For obtaining pointwise results as in Assumption 3 (i) we can invoke Theorem 3.1 in Wager and Athey (2018). From that, it follows for any estimate of a conditional expectation function that*

$$n^\varphi = n^{1-b} \cdot \log(n^b)^d,$$

*where $b$ satisfies*

$$b_{\min} := 1 - \left(1 + \frac{d}{\pi} \frac{\log(\omega)^{-1}}{\log((1-\omega)^{-1})}\right)^{-1} < b < 1$$

*and where $\omega, \pi$ are hyperparameters of the forest algorithm, i.e. the regularity parameter and splitting probability. For additional assumptions to obtain this result, see Theorem 3.1 of Wager and Athey (2018). Smoothness assumptions on the underlying conditional expectation function and other regularity conditions are required. The result provides a worst-case convergence rate and obtaining optimal rates in high-dimensional settings remains an open question. The generalized random forest algorithm of Athey et al. (2019) yields a similar result.*

*An additional example that will be referred to later is Wager and Walther (2016), which derive $L_2$-rates under additional sparsity assumptions, for a different class of random forest algorithms. In their Theorem 4, Wager and Walther (2016) establish that for any estimate $\widehat{\tau}(x)$ of a conditional expectation function $\tau(x)$ it holds that*

$$\mathbf{E}[(\widehat{\tau}(X) - \tau(X))^2] = O(n^{\log(\xi)/\log(2\xi)}),$$

*where $\xi = 1/(1 - 3/(4q))$ and $q$ is the effective dimension of the true conditional expectation function. See their Theorem for more details. They admit a high-dimensional setting where the number of covariates may grow with the sample size[2] but need additional restrictions on minimum effect sizes of some covariates, see in particular their Assumptions 3 and 4.*

The remark above gives convergence rates for random forest estimators of conditional expectation functions. Further, Assumption 3 imposes a convergence rate on $\widehat{\Pi}_{dm,k}(x)$ which is by definition (2.11) itself an average of different random forest estimates by averaging over $W$. Thus its convergence rate can be expected to be faster then the rate of the random forest estimate $\widehat{\mathbf{E}}[T(W, Y - \widehat{m}(X,W))|X = x, W = w]$ for any fixed $w$.

Finally the following convergence rate result holds.

**Theorem 2.1.** *Under Assumptions 1- 3 it holds that*

$$\int \left[\widehat{f}_{B|X}(b|x) - f_{B|X}(b|x)\right]^2 db = O_p(\tau_K^{-2} \frac{K}{n^{2\varphi}} + K^{-\alpha})$$

[2]More precisely $\liminf d/n > 0$

Here the decay of eigenvalues $\tau_K$ of the matrix $Q$ serves as a measure of ill-posedness. The estimation of the RC density is known to be an ill-posed inverse problem which implies a slower convergence rate of estimators, see Hoderlein et al. (2010) or Breunig (2021). If $\tau_K$ decays polynomially, e.g. $\tau_K \sim K^{-\gamma/2}$ then choosing $K \sim n^{\frac{2\varphi}{1+\gamma+\alpha}}$ balances bias and variance and results in the convergence rate

$$\int \left[\widehat{f}_{B|X}(b|x) - f_{B|X}(b|x)\right]^2 db = O_p(n^{-\frac{\alpha}{1+\alpha+\gamma}2\varphi}).$$

This shows that the convergence rate $\varphi$ of the generic machine learning estimators is slowed down by a factor $\alpha/(1+\alpha+\gamma) < 1$. This loss of speed is increasing in the eigenvalue decay parameter $\gamma$ and decreasing in the smoothness parameter $\alpha$. Note that the density considered here is bivariate and thus there is no explicit parameter for the number of random coefficients. If $W$ is multidimensional, its dimension enters the factor and further slows down convergence.

This rate result is not sharp in that the convergence rate can be improved for a given ML-technique. As $\varphi$ depends on tuning parameters specific to the chosen ML- technique, a joint choice of $K$ and the tuning parameters subsumed in $\varphi$ may improve the rate of convergence. However, calculating these exact rates may prove difficult and in practice, joint tuning of $K$ along with the parameters of the specific ML-techniques leads to excessive computational costs. Further note that using the result from Wager and Walther (2016) in Remark 2.2 a rate for $\mathbf{E}[\int \left[\widehat{f}_{B|X}(b|X) - f_{B|X}(b|X)\right]^2 db]$ can be derived analogously.

In Remark 2.1, estimation with orthogonalized treatment $W$ is considered. This is important for some ML estimators applied in causal inference. For the estimation of $\beta_1(x) = \mathbf{E}[B_1|X = x]$, Athey et al. (2019) suggest orthogonalization of both the outcome $Y$ and treatment $W$ to improve the performance of the random forest routines involved, see section 6.1.1 in Athey et al. (2019). The subsequent Corollary shows that in the context of random coefficient models orthogonalization of $W$ is not innocuous, as it slows down the convergence rate of the RC-density estimator. This is in contrast to Theorem 2.1, where sole orthogonalization of the outcome $Y$ does not result in a slower rate.

**Corollary 2.1.** *Let Assumptions 1- 3 hold. Consider the orthogonalized $W$ case outlined in Remark 2.1. Assume additionally that $K = o(\tau_K^{-1})$, then*

$$\int \left[\widehat{f}_{B|X}(b|x) - f_{B|X}(b|x)\right]^2 db = O_p(\tau_K^{-2}\frac{K^2}{n^{2\varphi}} + K^{-\alpha})$$

This slower convergence rate is due to the fact, that the "generated regressor" $W - \widehat{g}(X)$ appears within Hermite functions $q^K$. As is shown in the proof, the derivatives of Hermite functions $q^K$ diverge in $K$ and thus, an additional $K$-term appears in the derivation of the convergence rate. This is a general issue that does not seem to have been noticed so far. The convergence rate of any Hermite function sieve estimator of a RC density is slower when a generated regressor enters the sieve basis functions.

## 2.5 Inference

This section discusses inference for the conditional RC-density estimator, in particular, pointwise inference of the conditional RC density function estimate. Asymptotic normality of the RC-density estimator follows from asymptotic normality of ML-estimators linked with theory from the series estimation literature. For random forests, such asymptotic normality results have been recently provided by Athey et al. (2019) and Wager and Athey (2018). The main issue is how to establish the asymptotic covariance of the $K$ different ML-estimators, i.e. to obtain estimates for $\mathbf{E}[\widehat{\Pi}_{dm,j}(x) \cdot \widehat{\Pi}_{dm,l}(x)]$ for any pair $1 \leq j, l \leq K$. This issue can be overcome by introducing an additional layer of sample splitting. If we split the sample $\mathcal{R}$ on which sieve coefficients are learned into $K$ equally sized subsamples and use a different subsample for estimating each sieve coefficient, then naturally these estimators are stochastically independent. In that case, the asymptotic variance of each sieve coefficient can be obtained by resigning to established variance estimators for the respective ML-method. This approach requires the block size $|\mathcal{R}|/K$ to be of meaningful size in practice, yet, again cross-fitting, i.e. iterated sample splitting and averaging of estimates stabilizes the results and reduces any losses in efficiency.

More precisely, we need to amend Step 4. of the estimation procedure at the end of section 2.3.1.

Step 4: Additionally, split $\mathcal{R}$ into $K$ subsamples $\mathcal{R}_1, \ldots, \mathcal{R}_K$ of size $\lfloor |\mathcal{R}|/K \rfloor$ and calculate each $\widehat{\Pi}_{dm,j}$ using subsample $\mathcal{R}_j$.

Introduce the notation $r_p(n) = n^{\varphi}$. The resulting estimator does not coincide with the one in the previous section. The convergence rate is similar with the term $r_p(n/K)$ appearing in the convergence rate rather then $r_p(n)$ which is due to the fact that each sieve coefficient is learned on a sample of size $n/K$. Nevertheless in implementations the finite sample performance is comparable to the estimator in the previous section.

**Assumption 4.** *(i) For each $k = 1, \dots, K$ there exists a non-increasing sequence $\sigma_{dm,k}(x)$ such that*

$$\frac{\widehat{\Pi}_{dm,k}(x) - \Pi_{dm,k}(x)}{\sigma_{dm,k}(x)} \xrightarrow{d} N(0,1),$$

*where $\sigma_{dm,k}(x) \propto r_p(n/K)^{-1}$ and $\min_k \sigma_{dm,k} > 0$*
*(ii) $\max_k \sigma_{dm,k} / \min_k \sigma_{dm,k} = O(1)$*
*(iii) Define*

$$\begin{aligned}\Sigma_n(x) :=& diag\left(\sigma_{dm,1}(x), \dots, \sigma_{dm,K}(x)\right)\\ v_n(b,x) :=& \|q^K(b)'Q^{-1}\Sigma_n(x)\|_E = \sqrt{q^K(b)'Q^{-1}\Sigma_n^2(x)Q^{-1}q^K(b)}\end{aligned}$$

*and it holds that*

$$\frac{q^K(b)Q^{-1}\left(\widehat{\Pi}_{dm}(x) - \Pi_{dm}(x)\right)}{v_n(b,x)} \xrightarrow{d} N(0,1)$$

*with $v_n(b,x)$ bounded away from zero (iv) $\sqrt{K/r_p(n/K)}\tau_K^{-1} = o(1)$.*

Assumption 4 (i) establishes asymptotic normality of various ML-estimators. Such asymptotic normality results are standard for many ML-methods. For (honest) random forests such results have been established by Wager and Athey (2018). Asymptotic normality results for different tree-based algorithms are presented in the references therein and also in Athey et al. (2019). $\sigma_{dm,k}$ is the individual standard error of the $k$-th ML step and subsumes both the convergence rate and the residual standard deviation. Part (ii) of the above assumption imposes that the residual standard deviation in any of the $K$ ML-regressions is bounded away from zero and infinity. This assumption can be weakened to allow for standard deviations to diverge as $K$ grows at the cost of introducing an additional rate parameter that would require to further strengthen rate restrictions. In Assumption 4 (iii), $v_n(b,x)$ serves as the standard error of the conditional RC-density estimate. It holds that

$$v_n(b,x) \gtrsim \sqrt{K}\tau_K^{-1} \min_k \sigma_{dm,k},$$

which is bounded away from zero and approaching zero asymptotically under the rate restriction in (iv) which is required for consistency of the RC-density estimate, see Theorem 2.1.

Part (i) and (iii) are the most intricate conditions in Assumption 4 and typically involve additional regularity conditions and restrictions on the growth of $K$. The

following Lemma gives conditions such that Assumption 4 (i) and (iii) are satisfied for honest regression forests.

**Lemma 2.2.** *Assume the following conditions hold: (i) The density $f_X$ is bounded away from zero and infinity (ii) for any $k$ in $1, \dots, K$ the function $\Pi_{dm,k}(x)$ is Lipschitz continuous and also $\mathbf{E}[T_k(W,Y)^2|X=x]$ is Lipschitz continuous. (iii) for any $k$ and uniformly in $x$ it holds $Var(T_k(W,Y)|X=x) > 0$ and $\mathbf{E}[|T_k(W,Y) - \mathbf{E}[T_k(W,Y)|X=x]|^{2+\delta}|X=x] < M$ for some constants $\delta, M > 0$. (iv) $K = o(r_p(n/K)^c)$ with $c = \min\{\delta/2, 1, -\beta^*/b\}$ and $\beta^* = 1 + \epsilon - b/\beta_{\min} < 1$. Then, if $\widehat{\Pi}_{dm,k}(x)$ is an honest random forest estimator in the sense of Theorem 3.1. of Wager and Athey (2018), then Assumption 4 (i) is satisfied under conditions (i)-(iii). If additionally condition (iv) and Assumption 4 (ii) are satisfied, then Assumption 4 (iii) holds .*

Assumptions (i) to (iii) in the Lemma above are as in Theorem 3.1. of Wager and Athey (2018) that establishes asymptotic normality of a single (honest) random forest estimator of a mean regression function. Assumption (iv) is an additional rate restriction that is needed to achieve asymptotic normality of the sieve coefficient estimates. Additional rate restrictions are common in the series estimation literature, see e.g. Theorem 4.2. (iii) in Belloni et al. (2015) and also appear in Assumption 5 (ii) of Breunig (2021) for an RC-density estimate. Here, the rate restriction is milder then the one in 4 (ii) if for instance $\delta > 2$ and the convergence rate $b$ is sufficiently fast such that $-\beta^*/\beta_{\min} > 1$. In this case, the rate restriction in Assumption 4 (iv) that guarantees consistency of the RC-density estimate is already sufficient for Assumption 4 (iii). If however $\delta$ is rather small and the convergence rate $b$ close to the worst-case $\beta_{\min}$, there can be cases, where the rate restriction of Lemma 2.2 is stronger compared to the one in 4 (ii), especially if the decay of $\tau_K$ is slow.

The following additional assumption is required.

**Assumption 5.** *For any $x$ in the support of $X$ and for any $a \in \mathbb{R}$ it holds $P_K f_{A|X}(a|x) - f_{A|X}(a|x) = o(v_n(a,x))$.*

Assumption 5 is an undersmoothing condition that is standard for pointwise inference of a series estimator, see Belloni et al. (2015) (4.18). Note that similar rate restrictions are not needed for $\widehat{\beta}(x), \widehat{g}, \widehat{m}$, as these are calculated on a sample proportional to $n$ and thus, converge at rate $r_p(n)$ which is always faster then the standard error rate $v_n(b,w)$. An estimator for $v_n(b,x)$ is

$$\widehat{v}_n(b,x) = \|q^K(b - \widehat{\beta}(x))\widehat{Q}^{-1}\widehat{\Sigma}_n(x)\|_E,$$

where $\widehat{\Sigma}_n(x) = diag\left(\widehat{\sigma}_{dm,1}(x), \ldots, \widehat{\sigma}_{dm,K}(x)\right)$ and the individual standard error estimates $\widehat{\sigma}_{dm,k}(x)$ are specific to the employed ML-method. For random forests these can be obtained from applying the infinitisimal jackknife procedure of Efron (2014), see also the discussion in Wager and Athey (2018). Additional rate restrictions required for consistent estimation of the standard error $v_n(b,x)$ are not required. In the proof of the subsequent Theorem 2.2 it is shown that $\widehat{v}_n(b,x)$ is consistent for $v_n(b,x)$ under the assumptions given so far. The following pointwise asymptotic normality result holds.

**Theorem 2.2.** *If Assumptions 1-5 are satisfied then,*

$$\frac{\widehat{f}_{B|X}(b|x) - f_{B|X}(b|x)}{v_n(b,x)} \xrightarrow{d} N(0,1)$$

*and further*

$$\frac{\widehat{f}_{B|X}(b|x) - f_{B|X}(b|x)}{\widehat{v}_n(b,x)} \xrightarrow{d} N(0,1).$$

This determines the asymptotic normality of the estimator conditional on a given sample split, i.e. the case $M = 1$. To handle the cross-fitting case $M > 1$ and additional uncertainty due to sample splitting, we can follow the variational inference approach of Chernozhukov et al. (2019). The idea summarizes as follows.

Suppose there are $M$ different estimates $\widehat{f}^l_{B|X}(b|x)$ for $l = 1, \ldots, M$. For each estimate it is possible to construct a (1-$\alpha$)-confidence interval $[L_{1-\alpha,l}, U_{1-\alpha,l}]$ from Theorem 2.2 with $L_l = \widehat{f}^l_{B|X}(b|x) - c_{1-\alpha} \cdot \widehat{v}_n(b,w)$ and $U_l = \widehat{f}^l_{B|X}(b|x) + c_{1-\alpha} \cdot \widehat{v}_n(b,w)$ and $c_{1-\alpha}$ denoting the respective $1-\alpha$ quantile of the standard normal distribution.

To construct an asymptotically valid $1-\alpha$- confidence intervals for $f_{B|X}(b|x)$, Chernozhukov et al. (2019) propose $[\underline{Med}(\{L_{1-\alpha/2,l}\}_{l=1}^M), \overline{Med}(\{U_{1-\alpha/2,l}\}_{l=1}^M)]$ with $\underline{Med}$ denoting the lower median and $\overline{Med}$ the upper median. The confidence level of each single interval needs to be discounted to $1-\alpha/2$. Chernozhukov et al. (2019) provide a similar reasoning for constructing adjusted p-values.

## 2.6 Marginal Densities, Variable Importance Measures and Cross-Validation

This section addresses additional important aspects for the practical application of the estimation procedure. First, I discuss how to construct estimates of the marginal random coefficient density $f_B$. Second, I present a measure of variable importance

that assigns an importance score to every variable in $X$. This is an important descriptive tool for uncovering which variables in $X$ drive the heterogeneity in conditional RC densities. Lastly, I discuss a cross-validation procedure for a data-driven choice of tuning parameters.

**Estimating marginal RC densities** There are various direct estimators for marginal RC densities in the literature such as the Radon transform estimator of Hoderlein et al. (2010) or an adaptation of the sieve estimation strategy from Breunig and Hoderlein (2018) and Breunig (2021). The common identifying restriction is, however, full independence between $B$ and $W$, which is difficult to maintain in non-experimental data settings.

Maintaining the weaker conditional independence condition in Assumption 1 (i) estimates of the marginal density can be readily constructed by averaging over leave-one-out estimates of conditional density estimates $\widehat{f}_{-i,B|X}$. Here, the estimate is calculated without using the $i$-th datapoint $(Y_i, W_i.X_i)$. A consistent estimator for the marginal density is

$$\widehat{f}_B(b) = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{-i,B|X}(b|X_i).$$

The estimator $\widehat{f}_B$ will inherit its asymptotic properties from the conditional estimate $\widehat{f}_{B|X}$ which is discussed in the previous section. Thus, the convergence rate is slower compared to direct marginal RC density estimators making use of full independence between random coefficients and covariates. To the best of my knowledge there are, however, currently no alternative estimators for the marginal random coefficient density that operate under Assumption 1 (i).

**Variable Importance Measures** The estimation procedure outlined so far yields consistent estimates of $f_{B|X=x}$ for any given point $x$. An important question in applications is to identify those variables in the set of controls $X$ that drive heterogeneity in conditional RC densities, i.e. a criterion to guide the choice of interesting points $x$ on which to evaluate the estimate $\widehat{f}_{B|X}(b|x)$.

We focus here on our running example that makes use of regression forests. Note that for regression forests generally no post-selection inference problems arise as variable selection is done within in the various ML-steps of the estimation procedure. The goal is to find points $x$ that reveal interesting heterogeneities to the researcher. This is analogous to the role of variable importance measures for the causal forests of Athey et al. (2019).

For each of the ML-estimators used we can calculate a measure of variable importance that assigns an importance score to each covariate that is normalized to sum to one. This score is informative on how often a specific variable has been used for placing splits in the growing of the forest.

First, I focus on the ML-estimates $\widehat{\Pi}_{dm}$ which constitute the sieve coefficients and thus, determine the shape of the RC density. For each $k = 1, \ldots, K$ let $VI_k(X_l)$ denote an importance score assigned to covariate $X_l \in X$ by the regression forest estimator $\widehat{\Pi}_{dm}$.

To obtain a global measure of variable importance for the shape of the function $f_{A|X=x}$ we can simply average over $K$. Thus, define the variable importance of $X_l$ for the shape of the density as

$$VI_{shape}(X_l) = \frac{1}{K} \sum_{k=1}^{K} VI_k(X_l).$$

A measure of variable importance for the conditional expectation of random coefficients $\beta(x)$ is directly available by considering the variable importance measure for causal forests as implemented in the Athey et al. (2019)-package. In contrast to $VI_{shape}$, this measure of variable importance for the center of the density will be henceforth referred to as $VI_{mean}$.

**Parameter Tuning** In this paragraph I propose a cross-validation procedure for the choice of tuning parameters. Analogous to classical density estimation tuning parameters are chosen by minimization of the integrated squared error

$$\arg\min_{K,\sigma_t} ISE(K, \sigma_t) := \int_{\mathbb{R}^2} \left( \widehat{f}_{B|X}(b|x, K, \sigma_t) - f_{B|X}(b|x) \right)^2 db,$$

which is equivalent to minimizing the criterion

$$J(K, \sigma_t) := \int_{\mathbb{R}^2} \widehat{f}_{B|X}(b|x, K, \sigma_t)^2 db - 2 \int_{\mathbb{R}^2} \widehat{f}_{B|X}(b|x, K, \sigma_t) f_{B|X}(b|x) db.$$

The first part is simply the integrated squared RC density estimate. The second term is typically estimated via cross-validation. However, it is not possible to observe realizations of random coefficients. The following Lemma links the second part to an expression that can be estimated via leave-one-out cross validation.

**Lemma 2.3.** *Let Assumption 1 hold, then the following identity holds*

$$\int_{\mathbb{R}^2} \widehat{f}_{B|X}(b|x) f_{B|X}(b|x) db = \int_{\mathbb{R}} \mathbf{E}[V(Y,W)' \widehat{Q}^{-1} \widehat{\Pi}_{dm}(X) | X = x, W = w] dw,$$

*where* $V(y,w) = (V_1(y,w), \ldots, V_K(y,w))$ *with*

$$V_k(y,w) = \frac{1}{2\pi^2} \int q_k(b) \cdot |t| \cdot \exp[it(y - b'(1,w))] dt db.$$

Again, a weighting for $w$ should be considered for practical reasons. Defining

$$V_k(y,w) = \frac{1}{2\pi^2} \int q_k(b) \cdot |t| \cdot \exp[it(y - b'(1,w))] / f_W(w) dt db,$$

it is equivalent to consider the integral $\int_{\mathbb{R}} \mathbf{E}[V(Y,W)' \widehat{Q}^{-1} \widehat{\Pi}_{dm}(X) | X = x, W = w] f_W(w) dw$ in the first equality of Lemma 2.3. Using a plug-in estimate for the unknown density $f_W$, the function $V$ can be computed and cross-validation used to estimate the conditional expectation with a machine learning estimator. Here, either a subsample of observations that has not been used for calculating $\widehat{f}_{B|X}$ can be used for the prediction task or a leave-one-out estimator for $\widehat{f}_{B|X}$. Standard practices of cross-validation apply.

## 2.7 Monte Carlo Simulations

This section evaluates the finite sample performance of the RC-density estimator outlined in the earlier sections. The following data generating process is studied first,

$$\begin{aligned} Y &= B_0 + B_1 \cdot W, \quad \text{with} \\ B_0 &= \sin(X_1) + A_0, \\ B_1 &= X_2 + 0.5 \cdot X_3 + 0.25 \cdot X_2 \cdot X_3 + A_1, \\ W &= 1 + X_3 + (1 + X_3^2) \cdot V \end{aligned} \tag{2.13}$$

where $A_0, V$ are standard normal random variables and $A_1$ is a mixture of a $N(-1.5, 1)$ and a $N(1.5, \sqrt{1/2})$ random variable with weights $1/2$. In this setting, the density of the random slope $B_1$ is bi-modal and any testpoint $X = x$ solely determines the center of the density function. The controls $X$ are a $p$-dimensional vector of iid standard normal variables. Here, I set $p = 10$, but as we see from the setup above, only variables $X_1, X_2, X_3$ are of importance in this toy model. This reflects

the common practical problem, that there is a large set of control variables but only some of them drive the heterogeneity in $B_1$ or may otherwise affect the outcome $Y$. Further, I introduce a form of heteroskedasticity in the equation for $W$, such that we do not only consider the clean case where orthogonalization removes all dependence between $W$ and $X$. There is also some form of dependence between the regressor $W$ and the random slope $B_1$ as both depend on the regressor $X_3$.

The goal is to estimate the density of the random slope $B_1$ conditional on some testpoint $X = x$. Here, I implement the estimator in (2.6) with the algorithm outlined at the end of section 2.3. In this setting $W$, does not need to be orthogonalized.

The other parameters of the estimation problem are chosen as follows. I set $K_1 = K_2 = 3$ and thus, there are a total number of $K = 9$ basis functions. Hermite polynomials are used as sieve basis $q^K$ and the weighting measure follows a log-normal law, i.e. $\mu \sim lognormal(0, \sigma_t)$ with $\sigma_t = 1$. In practice, when only the slope parameter is of interest, $K_1$ should be fixed and cross-validation performed to guide the choice of $K_2$ and $\sigma_t$. Simulations show that $K_2$ is the more relevant parameter for estimates compared to $\sigma_t$, so sole cross-validation of $K_2$ may be sufficient if computation time is a concern. To reduce computational effort, parameters in this simulation study are not chosen via cross-validation and there is no cross-fitting as well. Therefore $M = 1$ and the sample is split only once in equally sized parts $\mathcal{R}$,$\mathcal{D}$ of size $n/2$ and RC-density estimates are computed only once per Monte Carlo iteration. The testpoint is chosen as $x = (0, 0.3, 0, \ldots, 0)$, so the correct density is centered around 0.3.

All ML estimates are obtained from using honest regression forests, respectively causal forests for the quantity $\beta_1(x)$, see Wager and Athey (2018) and Athey et al. (2019), with the implementation taken from the **grf**-package in *R*. Each random forest is tuned using implemented data-driven routines, the number of trees in each forest is set to 2000, which is the packages default setting. In general, I find that tuning of internal forest parameters does improve the quality of estimates but is only of secondary importance for the overall shape of the density estimate.

The sample size is $n = 1000$ and 100 Monte Carlo draws of the model in (2.13) are performed. The simulation results for $\widehat{f}_{B_1|X=x}$ are presented in Figure 2.1. Figure 2.1 shows a favorable performance of the estimator even for a moderate sample size and for a coarse choice of $K_2$.

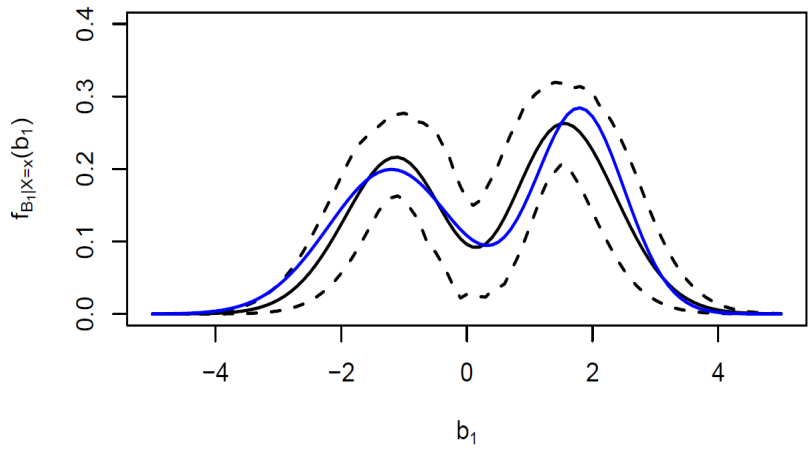


Figure 2.1: The solid black line denotes the median of the Monte Carlo estimates. The dotted lines the 95%- and 5%-quantiles. The solid blue line is the correct density. Key parameters: $K_2 = 3$ and $\sigma_t = 1$

The second data generating process is,

$$\begin{aligned} Y &= B_0 + B_1 \cdot W, \quad \text{with} \\ B_0 &= \sin(X_1) + A_0, \\ W &= 1 + X_3 + V \cdot (1 + X_3^2), \end{aligned} \tag{2.14}$$

where all random variables are chosen as before and $B_1$ is a mixture distribution like $A_1$ in the first setting, but now with weights $\Phi(X_2), 1-\Phi(X_2)$. In this setting, $X$ determines the entire shape of the density function as opposed to the first setting, where $X$ only determines the center of the density. For the testpoint $x = (0, 0.3, 0, \ldots, 0)$, the conditional density is again bi-modal, but now the mode on the negative part of the domain is more pronounced. All parameters and hyperparameters of the ML-procedures are as before, but now $K_2 = 7$ to illustrate the performance of the estimator for a more complex model. As the density of $B_1$ is more dispersed compared to the first setting, this increase in complexity can be rationalized. Note that the support of each of the Hermite basis functions increases with $K$. This leads to the suggestion to increase $K$ for highly dispersed densities or to otherwise scale down $Y$ and $W$ accordingly to control the maximal dispersion of the density. The simulation results are presented in Figure 2.2. Through a larger number of basis functions the bias is comparably lower then in Figure 2.3 at the expense of increased confidence intervals. As there are no shape constraints, we see that density estimates

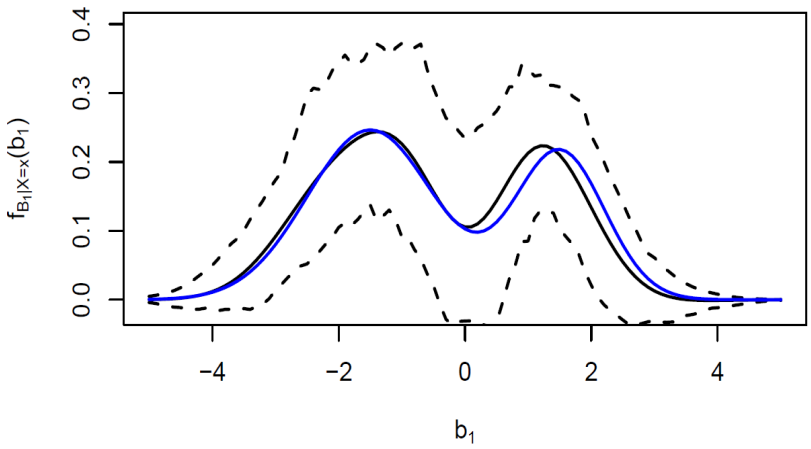


Figure 2.2: The solid black line denotes the median of the Monte Carlo estimates. The dotted lines the 95%- and 5%-quantiles. The solid blue line is the correct density. Key parameters: $K_2 = 7$ and $\sigma_t = 1$.

can in principal have negative parts. Yet, the method can detect the conditional density reliably even when the entire shape of the conditional density varies with $X$.

## 2.8 Empirical Application

In this section I apply the estimation strategy outlined before to study heterogeneous effects of stock market expectations on portfolio choice. I make use of the innovation sample of the german socio-economic panel (SOEP-IS). Therein, survey respondents were supplied with a hypothetical amount of $50,000$ Euros and asked to split their investment among one risk-free and one risky asset with returns paid out one year later. The risk-free asset is a state claim with a fixed annual interest rate of 4% whereas the risky asset's return hinges on the return of the german stock market index (DAX) within the next year.

This experiment has been previously analyzed in Huck et al. (2015). Economic theory suggests that stock market expectations and risk preferences are the main determinants of the portfolio choice task at hand.

The goal is to study the effect of stock market expectations on the investment in the risky asset. Formulated as a random coefficient model, I study the following

econometric model,

$$Y_i = B_{0,i} + B_{1,i} \cdot W_i,$$

where $Y_i$ denotes the individual investment in the risky asset, $W_i$ is the individual belief on the development of the DAX for the next year and $B_{1,i}$ is the individual effect of interest. The random intercept $B_{0,i}$ subsumes the effects of other controls $X$ and further unobservable characteristics on the outcome $Y_i$. The set of controls

| | Min. | 1. Quant | Median | Mean | 3. Quant. | Max. |
|---|---|---|---|---|---|---|
| $Y$ (in Euro) | 1000 | 15000 | 25000 | 24029 | 30000 | 50000 |
| $W$ (in %-points) | -50 | 2 | 5 | 4.90 | 8 | 130 |

Table 2.1: Summary Statistics

is quite rich and contains 75 variables including information on socio-demographics such as gender, age or tertiary degrees as well as self-assessed measures of risk aversion, personality traits or skills in mathematical calculations. Summary statistics for the main variables are provided in Table 2.1.

In order to apply the estimation method, we need to assume that the conditional independence restriction of Assumption 1 (i) holds. Applied to the present setting, this implies that stock market beliefs are exogenous conditional on the set of controls $X$. The data is observational and beliefs are self-reported, so we cannot rule out relations between beliefs and other controls which rules out considering the standard, unconditional RC model that relies on full independence of random coefficients and controls.

Further, beliefs $W$ must vary sufficiently in the population to plausibly fulfill the support restrictions in Assumption 1 (ii), which is the case in this setting.

The analysis begins by choosing the tuning parameters $K$ and $\sigma_t$. As the random slope is of main interest, I fix $K_1 = 3$ and $\sigma_t = 1$ and vary the choice of $K_2$. Figure 2.6 presents various estimates for different choices of $K_2$ and a suitable choice of $K_2$ can be eyeballed. For most choices of $K_2$, the two modes of the density are centered as for the case $K_2 = 5$ which thus appears to be a reasonable and coarse choice for the remainder of this analysis.

Next, I set a testpoint $x$ that corresponds to the medians of the variables in $X$. For those individuals with "median" characteristics $X = x$, an estimate of the random slope density $f_{B_1|X=x}$ is presented in 2.3 below. Most notable is the bi-modal shape of the density with one mode centered around zero and another around 2. Note that variables $Y$ and $W$ have been rescaled such that a value of 2

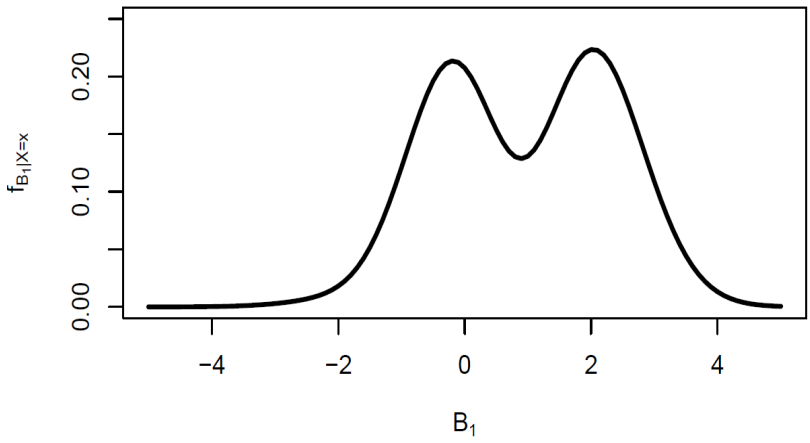


Figure 2.3: Estimate of the conditional density $f_{B_1|X=x}$. The tuning parameters have been chosen as $K_1 = 3$, $K_2 = 5$ (in total $K = 15$) and $\sigma_t = 1$. $M = 100$ sample splits are performed. The testpoint $x$ is chosen as the medians of the variables in $X$.

can be interpreted in the following way. A one percentage point increase in beliefs is associated with investing 1000 Euro (that is 2% of available funds) more into the risky asset.

Such a bi-modal density corresponds to the existence of two types in the population. For one part of the population, stock market expectations are actually linked to investment in the stock index as predicted by economic theory. Higher expectations also lead to a larger investment in the stock index. This does not seem to be true for a second group in the population, where the marginal effect centers around zero. This part of the population may follow different, e.g. heuristic decision rules in their portfolio choice and their stated beliefs do not appear to be a relevant constituent of the investment decision.

This appearance of types is in line with other results from the portfolio choice literature such as Drerup et al. (2017). They establish a link between the precision of subjective beliefs and the predictive power of economic models. Whenever beliefs are rather crude and imprecise they are likely not determinants of a rational portfolio choice. For individuals with such beliefs, economic theory has a rather low power in predicting their stock market participation. This is in line with my finding that for some part of the population their stated, subjective beliefs do not seem to influence their investment decision.

So far this finding indicates the existence of two, equally-large groups in the population. One group for which beliefs seem to have an impact on investment choice and one where it does not. Next, I study heterogeneity of the random slope densities, i.e. consider estimates of $f_{B_1|X=x}$ evaluated at different testpoints $x$. This is interesting, because the type distribution may vary across subpopulations with different observable characteristics $X$.

To get an idea of which variables may drive the heterogeneity I report a variable importance measure for the density's shape and center, as outlined in section 2.6. The largest variable importance scores among the 75 control variables are reported in Table 2.2.

| | "age" | "daxnetto1" | "daxnetto2" | "prisk" | "isb011" |
|---|---|---|---|---|---|
| $VI_{shape}$ | 0.052 | 0.043 | 0.046 | 0.035 | 0.031 |
| $VI_{mean}$ | 0.032 | 0.050 | 0.049 | 0.025 | 0.025 |

Table 2.2: Variable importance measures for those variables in $X$ with largest scores.

There does not appear to be much variation in densities across different controls. Most importance is given to the age variable followed by "daxnetto1" and "daxnetto2", which are randomly selected information on past annual DAX returns that were presented to the survey respondents before the investment game. The other two are a measure of risk-aversion and a measure of self-assessed patience.

Taking these results allows to investigate heterogeneity with respect to age and the historic information.

Figure 2.4 shows the heterogeneity in random slope densities for different age groups. Therefore, the conditional density estimate is evaluated at three different testpoints. The variable age is varied but all other points are set to the respective sample median value of the variables. The value $x$ is as in Figure 2.3 except that age is varied. The most prominent descriptive fact is that the type composure in the population seems to vary with age. For the young and medium aged subpopulations both types are equal in size. For the elder subpopulation, fewer people behave in accordance to economic theory.

Next, I also consider heterogeneity with respect to the historic information that has been displayed to the respondents. Again, there are three testpoints. There is one testpoint where both historic informations have been very positive (return of 35%), one where both informations are negative (return of -5%) and one mixed with a positive first and a negative second information. The results are displayed in Figure 2.5. Here, there is no apparent or robust heterogeneity with respect to the historic information. Therefore, we conclude the analysis and do not vary according

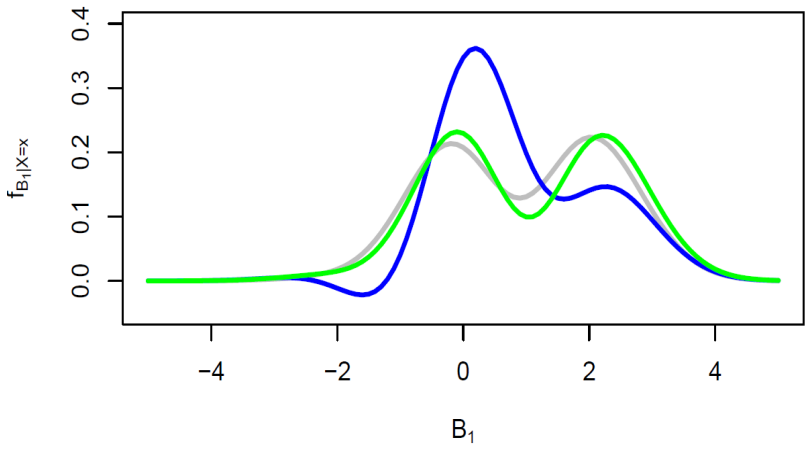


Figure 2.4: Estimate of the conditional density of $B_1|X = x$ for three different testpoints for $x$. The green line denotes the density for $age = 30$, the grey line for $age = 49$ and the blue line for $age = 70$. The tuning parameters have been chosen as in Figure 2.3.

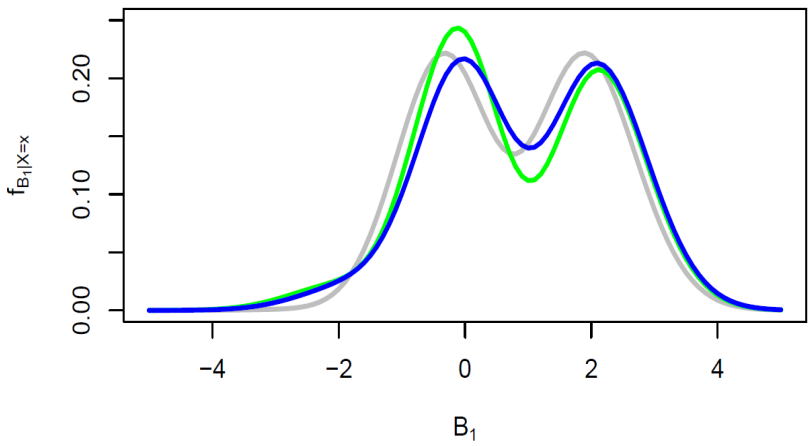


Figure 2.5: Estimate of the conditional density of $B_1|X = x$ for three different testpoints for $x$. The blue line denotes the density for $x$ with $daxnetto1 = daxnetto2 = 35$, the grey line for $daxnetto1 = daxnetto2 = -5$ and the green line for $daxnetto1 = 35$ and $daxnetto2 = -5$. The tuning parameters have been chosen as in Figure 2.3.

to variables with a lower variable importance score than that of *daxnetto*2.

The random coefficient analysis suggests the presence of two, roughly equally-sized types in the population. One group of individuals complies with economic theory in that their stock market expectation also explains their investment in a risky asset. A second group seems to follow different decision rules, their beliefs have no impact on their investment decision. Due to a possible correlation of beliefs and random coefficients, this finding cannot be inferred from estimating standard, marginal random coefficient models. Regarding heterogeneity, I find that the mixture of types in the populuation may depend on age but fail to uncover more interesting heterogeneity with the given data. It appears that the main determinants of type membership are unobservables that are not captured in the given data set.

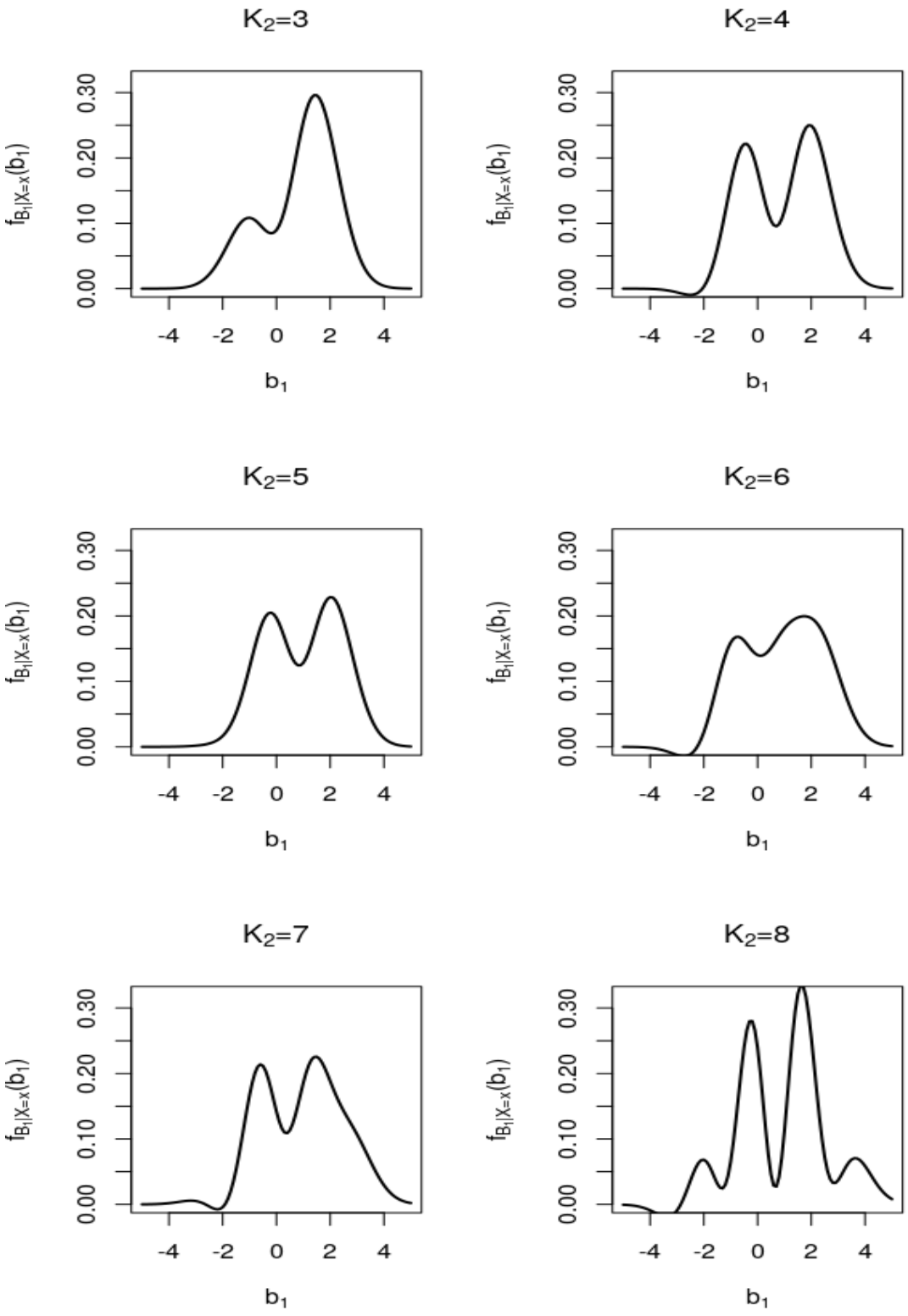


Figure 2.6: Estimates of the RC-density for various choices of $K_2$.

## 2.9 Conclusion

This paper discusses the estimation of conditional random coefficient densities when the set of conditioning variables is large. The very general conditional RC model has rarely been studied in both theory and application. This paper provides a general sieve estimation strategy for estimating conditional RC densities. The approach enables the use of generic machine learning methods to estimate sieve coefficients in the presence of a large dimensional set of control variables. Therefore, the estimator is applicable in many economic settings in which a continuous treatment variable is available. Theoretical results of the paper include convergence rate and inference results for the conditional sieve RC density estimator which combine asymptotic theories of sieve estimators and machine learning methods, in particular, applying results on (honest) random forests.

The finite sample properties of the estimator are illustrated in a Monte Carlo simulation study and an empirical application. The application reveals behavioral heterogeneity in an experimental portfolio choice task which is in line with recent empirical findings in the literature.

## 2.10 Appendix

Proof of Lemma 2.1. Let $\phi_{Y|X}(t|x) = \mathbf{E}[\exp(itY) \mid X = x]$ denote the conditional characteristic function of $Y$ given $X = x$. The following holds

$$\begin{aligned}
\phi_{Y|X,W}(t|x,w) &= \mathbf{E}[\exp(itY) \mid X = x, W = w] \\
&= \mathbf{E}[\exp(it(B_0 + B_1 W)) \mid X = x, W = w] \\
&= \mathbf{E}[\exp(i(t, tw)'(B_0, B_1)) \mid X = x, W = w] \\
&= \mathbf{E}[\exp(i(t, tw)'(B_0, B_1)) \mid X = x] \\
&= \phi_{B_0,B_1|X}(t, tw|x),
\end{aligned}$$

which is in fact already enough to point identify the probability distribution of $B \mid X = x$. By varying both $t$ and $w$ it is possible to evaluate the characteristic function of $B \mid X = x$ at any point in $\mathbb{R}^2$. Here, Assumption 1 is required in that the support of $W \mid X = x$ is the entire real line $\mathbb{R}$. See the proof of Lemma 1 in Masten (2017) and the references therein for details.

The main interest in practical applications is in identifying the density function $f_{B|X=x}$, which follows from applying the inverse Fourier transform to $\phi_{B_0,B_1|X=x}(t, tw)$. The Fourier transformation $\mathcal{F}$ and the inverse Fourier transformation $\mathcal{F}^{-1}$ are de-

fined as

$$(\mathcal{F}f)(t) = \int_{\mathbb{R}^d} \exp(it'a) f(a) da$$
$$(\mathcal{F}^{-1}g)(a) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \exp(-ia't) g(t) dt$$

for some functions $f, g : \mathbb{R}^d \to \mathbb{R}$ and $\mathcal{F} : \mathbb{R}^d \to \mathbb{C}^d$ and $\mathcal{F}^{-1} : \mathbb{C}^d \to \mathbb{R}^d$. The Fourier transform generally links the characteristic function of a random variable to its density function, in particular $\phi_{B_0,B_1|X}(t, tw|x) = (\mathcal{F}f_{B_0,B_1|X=x})(t, tw)$.

From this, we can infer the following

$$\begin{aligned}
f_{B|X}(b|x) &= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \exp(-ib's)(\mathcal{F}f_{B|X=x})(s) ds \\
&= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} |t| \exp(-ib'(t, tw))(\mathcal{F}f_{B|X=x})(t, tw) dt dw \\
&= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} |t| \exp(-ib'(t, tw))(\mathcal{F}f_{B|X=x})(t, tw) dt dw \\
&= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} |t| \exp(-ib'(t, tw))\phi_{B_0,B_1|X}(t, tw|x) dt dw \\
&= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} |t| \exp(-ib'(t, tw))\phi_{Y|X,W}(t|x, w) dt dw,
\end{aligned}$$

which establishes identification of the density function $f_{B|X=x}$. □

Proof of Theorem 2.1. Let $\|\cdot\|_E$ denote the euclidean norm of a (complex-) vector and define $\|f - g\| = \int_{\mathbb{R}^2} |f(a) - g(a)|^2 da$ and $\|f - g\|_{\nu,\mu} = \int_{\mathbb{R}^2} |\mathcal{F}f(t, w) - \mathcal{F}g(t, w)|^2 d\nu(t) d\mu(w)$ for arbitrary functions $f, g : \mathbb{C}^2 \to \mathbb{R}$. Consider the following decomposition

$$\begin{aligned}
&\|\widehat{f}_{B|X=x} - f_{B|X=x}\|^2 \\
\leq &\|\widehat{f}_{A|X=x} - f_{A|X=x}\|^2 + \|f_{A|X=x}(\cdot - \widehat{\beta}(x)) - f_{A|X=x}(\cdot - \beta(x))\|^2 \\
\leq &\|\widehat{f}_{A|X=x} - \widetilde{f}_{A|X=x}\|^2 + \|\widetilde{f}_{A|X=x} - f_{A|X=x}\|^2 \\
&\quad + \|f_{A|X=x}(\cdot - \widehat{\beta}(x)) - f_{A|X=x}(\cdot - \beta(x))\|^2 \\
= &A + B + C.
\end{aligned}$$

The proof begins by examining summand $B$. To this end recall that

$$P_K f_{A|X=x} = \arg\min_{\phi \in \mathcal{B}_K} \|\phi - f_{A|X=x}\|,$$

which is the $L_2$-projection of $f_{A|X=x}$ on the sieve space $\mathcal{B}_K$. It further holds for every

$x \in \mathcal{X}$,

$$
\begin{aligned}
&\|\widetilde{f}_{A|X=x} - f_{A|X=x}\|^2 \\
\leq &\|\widetilde{f}_{A|X=x} - P_K f_{A|X=x}\|^2 + \|P_K f_{A|X=x} - f_{A|X=x}\|^2 \\
\leq &{\tau_K}^{-1}\|\mathcal{F}\widetilde{f}_{A|X=x} - \mathcal{F}P_K f_{A|X=x}\|_{v,\mu}^2 + O(K^{-\alpha}) \\
\leq &{\tau_K}^{-1}\Big[\|\mathcal{F}\widetilde{f}_{A|X=x} - \mathcal{F}f_{A|X=x}\|_{v,\mu}^2 + \|\mathcal{F}f_{A|X=x} - \mathcal{F}P_K f_{A|X=x}\|_{v,\mu}^2\Big] + O(K^{-\alpha}) \\
\leq &{\tau_K}^{-1}\|\mathcal{F}f_{A|X=x} - \mathcal{F}P_K f_{A|X=x}\|_{v,\mu}^2 + O(K^{-\alpha}) \\
= &O(K^{-\alpha}),
\end{aligned}
$$

where we have used the link condition $\|\mathcal{F}f_{A|X=x} - \mathcal{F}\Pi_K f_{A|X=x}\|_{v,\mu}^2 = O(\tau_K\|\Pi_K f_{A|X=x} - f_{A|X=x}\|^2)$ and the fact that

$$
\widetilde{f}_{A|X=x} = \arg\min_{\phi \in \mathcal{B}_K} \|\mathcal{F}\phi - \mathcal{F}f_{A|X=x}\|_{v,\mu}^2.
$$

Next, consider the first summand $A$. It holds that

$$
\begin{aligned}
&\|\widehat{f}_{A|X=x} - \widetilde{f}_{A|X=x}\|^2 \\
= &\|q^K(\cdot)'\widehat{Q}^{-1}\widehat{\Pi}_{dm}(x) - q^K(\cdot)'Q^{-1}\Pi(x)\|^2 \\
\leq &\|q^K(\cdot)'(\widehat{Q}^{-1} - Q^{-1})\Pi(x)\|^2 + \|q^K(\cdot)'(\widehat{Q}^{-1} - Q^{-1})(\widehat{\Pi}_{dm}(x) - \Pi(x))\|^2 \\
&+ \|q^K(\cdot)'\widehat{Q}^{-1}(\widehat{\Pi}_{dm}(x) - \Pi(x))\|^2 \\
= &: I + II + III,
\end{aligned}
\tag{2.15}
$$

where in the following we consider each term separately. We begin with term III, where we have

$$
\begin{aligned}
&\|q^K(\cdot)'Q^{-1}(\widehat{\Pi}_{dm}(x) - \Pi(x))\|^2 \\
= &[\widehat{\Pi}_{dm}(x) - \Pi(x)]'Q^{-1}\left(\int_{\mathbb{R}^2} q^K(b)q^K(b)'db\right)Q^{-1}[\widehat{\Pi}_{dm}(x) - \Pi(x)] \\
\lesssim &\|\widehat{\Pi}_{dm}(x) - \Pi(x)\|_E^2\|Q^{-1}\|^2 \\
\lesssim &\tau_K^{-2}\Big[\|\widehat{\Pi}_{dm}(x) - \Pi_{dm}(x)\|_E^2 + \|\Pi_{dm}(x) - \Pi(x)\|_E^2\Big] \\
\lesssim &\tau_K^{-2}\Big[K \cdot O_p(n^{-2\varphi}) + O_p(K \cdot n^{-2\varphi})\Big] \\
= &O_p(\tau_K^{-2}\frac{K}{n^{2\varphi}}).
\end{aligned}
$$

and made use of the sample splitting rule and the convergence rate of ML-estimates in Assumption 3 (i). Without sample splitting, the behavior of $\|\widehat{\Pi}_{dm}(x) - \Pi_{dm}(x)\|_E$

cannot be established. The behavior of $\|\Pi_{dm}(x) - \Pi(x)\|_E^2$ follows from Lemma 2.4 (ii). Next, consider the term $I$. We have

$$\begin{aligned}
&\|q^K(\cdot)'(\widehat{Q}^{-1} - Q^{-1})\Pi(x)\|^2 \\
=&\Pi(X)'(\widehat{Q}^{-1} - Q^{-1}) \left( \int_{\mathbb{R}^2} q^K(b) q^K(b)' db \right) (\widehat{Q}^{-1} - Q^{-1})\Pi(x) \\
\lesssim&\|\Pi(x)\|_E^2 \cdot \|\widehat{Q}^{-1} - Q^{-1}\|^2 \\
\lesssim&K \cdot O_p\left( \frac{K\tau_K^{-1}\log(K)}{n} \right) \\
\lesssim&o_p(\tau_K^{-2}\frac{K}{n^{2\varphi}}),
\end{aligned}$$

which holds by Assumption 2 (iv) and applying Rudelsons LLN, as in the second part of Lemma 6.2. in Belloni et al. (2015) to $\|\widehat{Q}^{-1} - Q^{-1}\|$. The last inequality holds by the rate restriction in Assumption 3 (iii).

It remains to analyze II. Under the same reasoning as above, we obtain

$$\begin{aligned}
&\|q^K(\cdot)'(\widehat{Q}^{-1} - Q^{-1})(\widehat{\Pi}_{dm}(X) - \Pi_{dm}(X))\| \\
\lesssim& \|\widehat{\Pi}_{dm}(X) - \Pi(X)\|_E^2 \|\widehat{Q}^{-1} - Q^{-1}\|^2 \\
\lesssim& O_p\left( \frac{K}{n^{2\varphi}} \right) \cdot O_p\left( \frac{K\tau_K^{-1}\log(K)}{n} \right) \\
\lesssim& o_p\left( \tau_K^{-2}\frac{K}{n^{2\varphi}} \right)
\end{aligned}$$

and collecting terms we can conclude that $A = O_p(\tau_K^{-2}K/n^{2\varphi})$. Finally, it remains to consider $C$. There exists some $\tau \in (0,1)$ such that the following holds

$$\begin{aligned}
\|f_{A|X}(\cdot - \widehat{\beta}(X)) - f_{A|X}(\cdot - \beta(X))\| \leq& \int \|\nabla f_{A|X}(b - \xi)\| db \cdot \|\widehat{\beta}(X) - \beta(X)\|_E^2 \\
=& O_p(1) \cdot O_p(n^{-2\varphi}) = o_p(\tau_K^{-2}\frac{K}{n^{2\varphi}}),
\end{aligned}$$

where $\xi = \beta(X)(1-\tau) + \tau\widehat{\beta}(X)$ and the last bound following from Assumption 2 (v) and 3. This establishes the final result of Theorem 2.1. □

Proof of Corollary 2.1. In the case outlined in the corollary, we have

$$
\begin{aligned}
Q &= \mathbf{E}\left[\int_{\mathbb{R}} \mathcal{F}q^K(-t,-t(W-g(X)))\mathcal{F}q^K(-t,-t(W-g(X)))' d\nu(t)\right] \\
\widehat{Q} &= \sum_{i=1}^{\mathcal{R}} \int_{\mathbb{R}} \mathcal{F}q^K(-t,-t(W_i-\widehat{g}(X_i))\mathcal{F}q^K(-t,-t(W_i-\widehat{g}(X_i)))' d\nu(t) \\
\widetilde{Q} &= \mathbf{E}\left[\int_{\mathbb{R}} \mathcal{F}q^K(-t,-t(W-\widehat{g}(X)))\mathcal{F}q^K(-t,-t(W-\widehat{g}(X)))' d\nu(t)\right].
\end{aligned}
$$

The main part is to consider the behavior of $\|\widehat{Q}-Q\|$. It holds that

$$
\begin{aligned}
\|\widehat{Q}-Q\| &= \|\widehat{Q}-\widetilde{Q}\| + \|\widetilde{Q}-Q\| \\
&= O_p\left(\sqrt{\frac{K\tau_K^{-1}\log(K)}{n}}\right) + \|\widetilde{Q}-Q\|
\end{aligned}
$$

again by Rudelson's LLN. For the second part $\|\widetilde{Q}-Q\|$, it suffices to check the quantity

$$
\begin{aligned}
&||\mathcal{F}q^K(-t,-t(W-\widehat{g}(X)))\mathcal{F}q^K(-t,-t(W-\widehat{g}(X))) \\
&-\mathcal{F}q^K(-t,-t(W-g(X)))\mathcal{F}q^K(-t,-t(W-g(X)))||.
\end{aligned} \tag{2.16}
$$

For arbitrary complex vectors $a, b$, it holds that

$$
\begin{aligned}
\|aa'-bb'\| &= \|(a-b)(a-b)' + (a-b)b' + b(a-b)'\| \\
&\leq 2\cdot\|a-b\| + 2\cdot\|b\|\cdot\|a-b\|
\end{aligned}
$$

and applying this to (2.16) leads to

$$
\begin{aligned}
(2.16) \leq & 2\cdot(1+\|\mathcal{F}q^K(-t,-t(W-g(X)))\|) \\
& \cdot\|\mathcal{F}q^K(-t,-t(W-\widehat{g}(X)) - \mathcal{F}q^K(-t,-t(W-g(X))\| \\
\leq & 2\cdot(1+\|\mathcal{F}q^K(-t,-t(W-g(X)))\|)\cdot\|D\mathcal{F}q^K(\xi)\|\cdot\|\widehat{g}(X)-g(X)\|,
\end{aligned}
$$

which implies that

$$
\|\widetilde{Q}-Q\| \lesssim \sqrt{K}\cdot K\cdot O_p(n^{-2\varphi})
$$

and thus,

$$\|\widehat{Q} - Q\| = O_p\left(\sqrt{\frac{K\tau_K^{-1}\log(K)}{n}}\right) + O_p\left(\frac{K^{3/2}}{n^{2\varphi}}\right).$$

The difference to the proof of Theorem 2.1 is only in checking terms I, II and III. By applying Lemma 2.4 (i), it holds that

$$III = O_p(\tau_K^{-2}\frac{K^2}{n^{2\varphi}})$$

and further from the rate of $\|\widehat{Q} - Q\|$ above and the rate restriction stated in the Corollary that

$$I = o_p(\tau_K^{-2}\frac{K^2}{n^{2\varphi}}).$$

From III and I it is apparent that II is asymptotically negligible, which leads to the stated result. □

Proof of Lemma 2.2. For any single (honest) random forest predictor $\widehat{\Pi}_{dm,k}(x)$ Theorem 1 of Wager and Athey (2018) establishes its asymptotic normality under the assumptions stated in the Theorem itself. For the proof of Lemma 2.2 it suffices to adapt their steps to the case $q^K(b)'Q^{-1}(\widehat{\Pi}_{dm,1}(x), \ldots, \widehat{\Pi}_{dm,K}(x)$. To simplify notation for the remainder of the proof, $q^K = q^K(b)$ and $\widehat{\Pi}_{dm,k} = \widehat{\Pi}_{dm,k}(x)$. The proof proceeds treating $\widehat{\Pi}_{dm,k}$ as a pure forest estimator. Applying the integration to obtain the true $\widehat{\Pi}_{dm,k}$ of (2.11) does not change the derivations, as integration is a linear, monotonic operator and the theory in Wager and Athey (2018) goes through.

Let $\overset{\circ}{\widehat{\Pi}}_{dm,k}$ denote the Hajek projection of the forest predictor and under a slight abuse of notation let $\overset{\circ}{\widehat{\Pi}}_{dm} = (\overset{\circ}{\widehat{\Pi}}_{dm,1}, \ldots, \overset{\circ}{\widehat{\Pi}}_{dm,K})'$. In broad steps the proof of Wager and Athey (2018) proceeds by checking that for a given forest predictor $\widehat{\Pi}_{dm,k}$ it holds that

$$\frac{\overset{\circ}{\widehat{\Pi}}_{dm,k} - \mathbf{E}[\overset{\circ}{\widehat{\Pi}}_{dm,k}]}{\sigma_{dm,k}} \xrightarrow{d} N(0,1) \tag{2.17}$$

$$\mathbf{E}\left[\left(\widehat{\Pi}_{dm,k} - \overset{\circ}{\widehat{\Pi}}_{dm,k}\right)^2\right] / \sigma^2_{dm,k} \to 0 \tag{2.18}$$

$$\frac{\mathbf{E}[\widehat{\Pi}_{dm,k}] - \Pi_{dm,k}}{\sigma_{dm,K}} \to 0, \tag{2.19}$$

where (A.1) and (A.2) are shown in the proof of Theorem 8 and (A.3) in the proof of Theorem 1 of Wager and Athey (2018). The quantity $\sigma_{dm,k}$ is in fact the standard deviation of the Hajek projection.

I follow along their steps and show that the following holds under the Assumptions stated in Lemma 2.2:

$$
\begin{aligned}
I :=& \frac{q^K Q^{-1}(\mathring{\widehat{\Pi}}_{dm} - \mathbf{E}[\mathring{\widehat{\Pi}}_{dm}])}{\|q^K Q^{-1}\Sigma_n\|} \xrightarrow{d} N(0,1) \\
II :=& \mathbf{E}\left[\left(q^K Q^{-1}\left(\widehat{\Pi}_{dm} - \mathring{\widehat{\Pi}}_{dm}\right)\right)^2\right] / \|q^K Q^{-1}\Sigma_n\|^2 \to 0 \\
III :=& \frac{q^K Q^{-1}(\mathbf{E}[\widehat{\Pi}_{dm}] - \Pi_{dm})}{\|q^K Q^{-1}\Sigma_n\|} \to 0
\end{aligned}
$$

which taken together implies that $q^K Q^{-1}(\widehat{\Pi}_{dm} - \Pi_{dm})/\|q^K Q^{-1}\Sigma_n\|$ is asymptotically normal.

We need to introduce and adapt some of the notation from the proofs of Wager and Athey (2018). Let $s$ denote the subsample size used to construct the random forest from single tree predictors $\widehat{T} = (\widehat{T}_1, \ldots, \widehat{T}_K)$, where $\widehat{T}_k = \widehat{T}_k(x; \mathcal{R}_k)$ is a single tree predictor for the conditional expectation $\mathbf{E}[T_k(W - \widehat{g}(X), Y - \widehat{m}(X,W))|X = x]$ making use of the data points in the respective sample $\mathcal{R}_k$.

We begin with part I. Plugging in the expression for the Hajek projection of the random forest on page 53 of the supplemental material of Wager and Athey (2018), we obtain the identity

$$
q^K Q^{-1}(\mathring{\widehat{\Pi}}_{dm} - \mathbf{E}[\mathring{\widehat{\Pi}}_{dm}]) = \frac{s \cdot K}{n} \sum_{i=1}^{n/K} q^K Q^{-1}(\mathbf{E}[\widehat{T}|\mathcal{R}_i] - \mathbf{E}[\widehat{T}])
$$

where $\mathbf{E}[\widehat{T}|\mathcal{R}_i] = (\mathbf{E}[\widehat{T}_1|\mathcal{R}_{1,i}], \ldots, \mathbf{E}[\widehat{T}_K|\mathcal{R}_{K,i}])'$ and $\mathcal{R}_{k,i}$ is the $i$-th observation in sample $\mathcal{R}_k$. Further,

$$
\|q^K Q^{-1}\Sigma_n\| = \frac{s \cdot K}{n} \sqrt{\sum_{i=1}^{n/K} q^K Q^{-1} Var(\widehat{T}) Q^{-1} q^K}
$$

where $Var(\widehat{T}) = diag(Var(\widehat{T}_1), \ldots, Var(\widehat{T}_K))$ and which holds from applying the identity on the last line of page 52 and thus, we can write

$$
I = \frac{\sum_{i=1}^{n/K} q^K Q^{-1}(\mathbf{E}[\widehat{T}|\mathcal{R}_i] - \mathbf{E}[\widehat{T}])}{\sqrt{\sum_{i=1}^{n/K} q^K Q^{-1} Var(\widehat{T}) Q^{-1} q^K}}
$$

and establish the asymptotic normality of I by checking Lyapunov's condition

$$\frac{\sum_{i=1}^{n/K} \mathbf{E}[|q^K Q^{-1}(\mathbf{E}[\widehat{T}|\mathcal{R}_i] - \mathbf{E}[\widehat{T}])|^{2+\delta}]}{\left(\sum_{i=1}^{n/K} q^K Q^{-1} Var(\widehat{T}) Q^{-1} q^K\right)^{1+\delta/2}} \to 0, \tag{2.20}$$

For the numerator, we have by Cauchy-Schwarz

$$\begin{aligned}
&\sum_{i=1}^{n/K} \mathbf{E}[|q^K Q^{-1}(\mathbf{E}[\widehat{T}|\mathcal{R}_i] - \mathbf{E}[\widehat{T}])|^{2+\delta}] \\
\leq &\|q^K Q^{-1}\|^{2+\delta} \cdot \sum_{k=1}^{K} \sum_{i=1}^{n/K} \mathbf{E}[|\, \mathbf{E}[\widehat{T}_k|\mathcal{R}_{k,i}] - \mathbf{E}[\widehat{T}_k]|^{2+\delta}] \\
\leq &\|q^K Q^{-1}\|^{2+\delta} \cdot K \cdot \sum_{i=1}^{n/K} \max_k \mathbf{E}[|\, \mathbf{E}[\widehat{T}_k|\mathcal{R}_{k,i}] - \mathbf{E}[\widehat{T}_k]|^{2+\delta}],
\end{aligned}$$

where the last inequality is due to the last display in the proof of Theorem 8. The denominator satisfies

$$\left(\sum_{i=1}^{n/K} q^K Q^{-1} Var(\widehat{T}) Q^{-1} q^K\right)^{1+\delta/2} \geq \|q^K Q^{-1}\|^{2+\delta} \cdot \left(\sum_{i=1}^{n/K} \min_k Var(\widehat{T}_k)\right)^{1+\delta/2},$$

which follows from the last steps of the proof on page 54. Define

$$\begin{aligned}
k^* &:= \arg\max_{k \in K} \mathbf{E}[|\, \mathbf{E}[\widehat{T}_k|\mathcal{R}_{k,i}] - \mathbf{E}[\widehat{T}_k]|^{2+\delta}] \\
\overline{k} &:= \arg\max_{k \in K} Var(\widehat{T}_k) \\
\underline{k} &:= \arg\min_{k \in K} Var(\widehat{T}_k),
\end{aligned}$$

then following the proof of Theorem 8, one can conclude that

$$\begin{aligned}
(2.20) \leq &\frac{K \cdot \sum_{i=1}^{n/K} \mathbf{E}[|\, \mathbf{E}[\widehat{T}_{k^*}|\mathcal{R}_{k^*,i}] - \mathbf{E}[\widehat{T}_{k^*}]|^{2+\delta}]}{\left(\sum_{i=1}^{n/K} Var(\widehat{T}_{k^*})\right)^{1+\delta/2}} \cdot \frac{\sum_{i=1}^{n/K} Var\left(\widehat{T}_{\overline{k}}\right)^{1+\delta/2}}{\sum_{i=1}^{n/K} Var\left(\widehat{T}_{\underline{k}}\right)^{1+\delta/2}} \\
\leq &K \cdot r_p (n/K)^{-\delta/2}
\end{aligned}$$

which holds by the last display in the proof of Theorem 8 in the supplemental

material of Wager and Athey (2018) and Assumption 4 (ii), which further implies

$$\sum_{i=1}^{n/K} Var\left(\widehat{T}_{\overline{k}}\right)^{1+\delta/2} / \sum_{i=1}^{n/K} Var\left(\widehat{T}_{\underline{k}}\right)^{1+\delta/2} = O(1).$$

Then, (2.20) tends to zero by the rate restriction stated in Lemma 2.2, which establishes asymptotic normality of I. Now, consider II. By applying Cauchy Schwarz and the lower bound for sieve variance we obtain

$$\begin{aligned} II \leq & \frac{\|q^K Q^{-1}\|^2}{\|q^K Q^{-1}\Sigma_n\|^2} \mathbf{E}[\|\widehat{\Pi}_{dm} - \overset{\circ}{\widehat{\Pi}}_{dm}\|^2] \\ \leq & \frac{1}{\min_k \sigma^2_{dm,k}} \sum_{k=1}^{K} \mathbf{E}[(\widehat{\Pi}_{dm,k} - \overset{\circ}{\widehat{\Pi}}_{dm,k})^2] \\ \leq & K \cdot r_p(n/K)^{-1} \to 0, \end{aligned}$$

which holds by the same reasoning as in the beginning of the proof of Theorem 8 in the supplement of Wager and Athey (2018) and by the rate restriction stated in Lemma 2.2. It remains to consider III. By the same reasoning as before we obtain

$$\begin{aligned} III \leq & \frac{\|\mathbf{E}[\widehat{\Pi}_{dm}] - \Pi_{dm}\|}{\min_k \sigma_{dm,k}} \\ \leq & \frac{\sqrt{K} \cdot \max_k \mathbf{E}[\widehat{\Pi}_{dm,k}] - \Pi_{dm,k}}{\min_k \sigma_{dm,k}} \\ \lesssim & \sqrt{K} \cdot \left(\frac{n}{K}\right)^{\frac{1}{2}\beta^*} = O\left(\sqrt{\frac{K}{r_p(n/K)^{-\beta^*/b}}}\right), \end{aligned}$$

where $\beta^* := 1 + \epsilon - b/\beta_{\min}$. The last inequality follows from the Proof of Theorem 1 on page 40 of the supplement of Wager and Athey (2018) and under these conditions $\beta^* < 0$ and $b/\beta_{\min} > 0$. Under the rate restrictions in the Lemma the right hand-side above converges to zero which concludes the proof. □

Proof of Theorem 2.2. The proof begins with the following decomposition

$$\begin{aligned}
&\widehat{f}_{B|X}(b|x) - f_{B|X}(b|x) \\
=&\underbrace{\widehat{f}_{A|X}(b-\widehat{\beta}(x)|x) - \widehat{f}_{A|X}(b-\beta(x)|x)}_{I} \\
+&\underbrace{\widehat{f}_{A|X}(b-\beta(x)|x) - q^K(b-\beta(x))'Q^{-1}\Pi_{dm}(x)}_{II} \\
+&\underbrace{q^K(b-\beta(x))'Q^{-1}\Pi_{dm}(x) - q^K(b-\beta(x))'Q^{-1}\Pi(x)}_{III} \\
+&\underbrace{q^K(b-\beta(x))'Q^{-1}\Pi(x) - f_{A|X}(b-\beta(x)|x)}_{IV}
\end{aligned}$$

and proceeds by checking the individual terms separately.

For I it holds that

$$\begin{aligned}
I \leq& |\widehat{f}_{A|X}(b-\widehat{\beta}(x)|x) - \widehat{f}_{A|X}(b-\beta(x)|x)| \\
\leq& \|D\widehat{f}_{A|X}(b-\beta(x)-(1-\tau)(\widehat{\beta}(x)-\beta(x))\| \cdot \|\widehat{\beta}(x)-\beta(x)\| \\
=& o_p(v_n(b,w)),
\end{aligned}$$

for some $\tau \in (0,1)$ by consistency of $\widehat{f}_{A|X}$ and Assumption 2 (v). This is due to the fact that $\|\widehat{\beta}(x)-\beta(x)\| = r_p(n)^{-1}$ as $\widehat{\beta}$ is calculated on a sample proportional to $n$ and thus, by the rate of $v_n(b,w)$ it always holds that $I = o_p(v_n(b,w))$. For II, it holds by Assumption 4 (i) that

$$II/v_n(b,x) \xrightarrow{d} N(0,1),$$

For III, we have from Lemma 2.4 (ii) that

$$III \lesssim_P \tau_K^{-1}\sqrt{K} \cdot \sqrt{K/n^{2\varphi}}$$

and thus,

$$III/v_n(b,w) \lesssim_P \frac{\tau_K^{-1}\sqrt{K}\sqrt{K/n^{2\varphi}}}{\tau_K^{-1}\sqrt{K}r_p(n/K)^{-1}} = \frac{K}{n} = o(1)$$

and then $III/v_n(b,x) = o_p(1)$. Finally, for IV

$$\left(P_K f_{A|X}(b-\beta(x)|x) - f_{A|X}(b-\beta(x)|x)\right) = o(v_n(b,w))$$

by Assumption 5 (i) the approximation error is negligible compared to $v_n$.

For the final part of the statement it remains to show

$$\left|\frac{\widehat{v}_n(b,w)}{v_n(b,w)} - 1\right| = o_p(1),$$

Let $s(b,x)' = q^K(b-\beta(x))'Q^{-1}$ and $\widehat{s}(b,x)' = q^K(b-\widehat{\beta}(x))'\widehat{Q}^{-1}$, it holds that

$$\begin{aligned}
|\widehat{v}_n(b,w) - v_n(b,w)| \leq& \left| \|\widehat{s}(b,x)'\widehat{\Sigma}_n(x)\| - \|s(b,x)'\Sigma_n(x)\| \right| \\
\leq& \|\widehat{s}(b,x)'\widehat{\Sigma}_n(x) - s(b,x)'\Sigma_n(x)\| \\
\leq& \|[\widehat{s}(b,x) - s(b,x)]'\widehat{\Sigma}_n(x)\| + \|s(b,x)'(\widehat{\Sigma}_n(x) - \Sigma_n(x))\| \\
\leq& \max_k \widehat{\sigma}_{dm,k} \cdot \|\widehat{s}(b,x) - s(b,x)\| \\
&+ \max_k |\widehat{\sigma}_{dm,k} - \sigma_{dm,k}| \cdot \|s(b,x)\|
\end{aligned}$$

by the triangle inequality and the fact that $\widehat{\Sigma}_n(x), \Sigma_n(x)$ is a diagonal matrix. Further,

$$\begin{aligned}
&\|\widehat{s}(b,x) - s(b,x)\| \\
=&\|q^K(b-\widehat{\beta}(x))'\widehat{Q}^{-1} - q^K(b-\beta(x))'Q^{-1}\| \\
\leq&\|[q^K(b-\widehat{\beta}(x)) - q^K(b-\beta(x))]'\widehat{Q}^{-1} + q^K(b-\beta(x))'(\widehat{Q}^{-1} - Q^{-1})\| \\
\leq&\|Dq^K(b-\beta(x) - (1-\tau)(\widehat{\beta}(x)-\beta(x)))\| \cdot \|\widehat{Q}^{-1}\| \cdot \|\widehat{\beta}(x) - \beta(x)\| \\
&+ \|q^K(\cdot - \beta(x))\| \cdot \|\widehat{Q}^{-1} - Q^{-1}\| \\
\lesssim_P& \sqrt{K}\tau_K^{-1} \cdot r_p(n)^{-1} + \sqrt{K} \cdot \sqrt{K\tau_K^{-1}\log(K)/n},
\end{aligned}$$

Summarizing, by the properties of $v_n(b,x)$, we have

$$\begin{aligned}
&\left|\frac{\widehat{v}_n(b,w)}{v_n(b,w)} - 1\right| \\
\lesssim& \frac{\max_k \widehat{\sigma}_{dm,k}}{\max_k \sigma_{dm,k}} \cdot \frac{\max_k \sigma_{dm,k}}{\min_k \sigma_{dm,k}} \cdot \frac{\|\widehat{s}(b,x) - s(b,x)\|}{\sqrt{K}\tau_K^{-1}} \\
&+ \frac{\max_k |\widehat{\sigma}_{dm,k} - \sigma_{dm,k}|}{\sigma_{dm,k^*}} \cdot \frac{\sigma_{dm,k^*}}{\min_k \sigma_{dm,k}} \cdot \frac{\|s(b,x)\|}{\sqrt{K}\tau_K^{-1}} \\
=& (1+o_p(1)) \cdot O(1) \cdot O_p\left(r_p(n)^{-1} + \sqrt{K\tau_K\log(K)/n}\right) + o_p(1) \cdot O(1) \\
=& o_p(1)
\end{aligned}$$

with the right hand side converging to zero by consistency of $\widehat{\sigma}_{dm,k}$, the fact that $\max_k \sigma_{dm,k} / \min_k \sigma_{dm,k} = O(1)$ and the rate restriction in Assumption 4 (ii). □

PROOF OF LEMMA 2.3. By definition of $\widehat{f}_{B|X}(b|x)$ and the last display in the proof of Lemma 2.1, it holds that

$$\begin{aligned}
&\int_{\mathbb{R}^2} \widehat{f}_{B|X}(b|x) f_{B|X}(b|x) db \\
&= \int_{\mathbb{R}^2} q^K(b - \widehat{\beta}(x))' \widehat{Q}^{-1} \widehat{\Pi}_{dm}(x) \\
&\quad \cdot \left[ \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} |t| \exp(-ib'(t, tw)) \phi_{Y|X,W}(t|x, w) dt dw \right] db \\
&= \int_{\mathbb{R}^2} q^K(b - \widehat{\beta}(x))' \widehat{Q}^{-1} \widehat{\Pi}_{dm}(x) \\
&\quad \cdot \left[ \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} |t| \exp[-it(y - b'(1, w))] f_{Y|X,W}(y|x, w) dy dt dw \right] db
\end{aligned}$$

with the last equality following from plugging in the definition for $\phi_{Y|X,W}(t|x, w)$. Rearranging and using the definition of $V(Y, W)$ yields,

$$\begin{aligned}
&\int_{\mathbb{R}^2} \widehat{f}_{B|X}(b|x) f_{B|X}(b|x) db \\
&= \int_{\mathbb{R}} \mathbf{E}[V(Y, W)' \widehat{Q}^{-1} \widehat{\Pi}_{dm}(X)|X = x, W = w] dw,
\end{aligned}$$

which is the statement of the Lemma. □

**Lemma 2.4.** *Let Assumptions 2- 3 be satisfied and $q^K$ be the Hermite function basis. The following conditions hold.*

*(i) If $\Pi_{dm}(x) = \mathbf{E}[T_k(W - \widehat{g}(X), Y - \widehat{m}(X, W))|X = x]$, it holds that*

$$\|\Pi_{dm}(x) - \Pi(x)\|^2 = O(K^2 \cdot n^{-2\varphi}).$$

*(ii) If $\Pi_{dm}(x) = \mathbf{E}[T_k(W, Y - \widehat{m}(X, W))|X = x]$, then*

$$\|\Pi_{dm}(x) - \Pi(x)\|^2 = O(K \cdot n^{-2\varphi}).$$

PROOF OF LEMMA 2.4. The proof begins with case (i). By definitions made earlier, it holds

$$\begin{aligned}
&\|\Pi_{dm}(x) - \Pi(x)\|^2 \\
&\leq \sum_{k=1}^{K} \mathbf{E}\left[ |T_k(W - \widehat{g}(X), Y - \widehat{m}(X, W)) - T_k(W - g(X), Y - m(X, W))|^2 \, |X = x \right].
\end{aligned}$$

Then, by the properties of complex numbers and a mean value argument

$$\begin{aligned}
&|T_k(W-\widehat{g}(X),Y-\widehat{m}(X,W))-T_k(W-g(X),Y-m(X,W))|^2\\
=&[\mathrm{Re}(T_k(W-\widehat{g}(X),Y-\widehat{m}(X,W)))-\mathrm{Re}(T_k(W-g(X),Y-m(X,W)))]^2\\
&+[\mathrm{Im}(T_k(W-\widehat{g}(X),Y-\widehat{m}(X,W)))-\mathrm{Im}(T_k(W-g(X),Y-m(X,W)))]^2\\
=&\nabla\mathrm{Re}(T_k)(\xi_1)'(\widehat{g}(X)-g(X),\widehat{m}(X,W)-m(X,W))^2\\
&+\nabla\mathrm{Im}(T_k)(\xi_2)'(\widehat{g}(X)-g(X),\widehat{m}(X,W)-m(X,W))^2
\end{aligned}$$

for $\xi_j=(W,Y)-\tau_j\cdot(g(X),m(X,W))+(1-\tau_j)\cdot(\widehat{g}(X),\widehat{m}(X,W))$, where $\tau_j\in(0,1)$. Hence,

$$\begin{aligned}
&\sum_{k=1}^{K}\|\Pi_{dm}(x)-\Pi(x)\|^2\\
\leq&\sum_{k=1}^{K}\mathbf{E}\left[\|\nabla\mathrm{Re}(T_k)(\xi_1)\|^2+\|\nabla\mathrm{Im}(T_k)(\xi_1)\|^2|X=x\right]\\
&\cdot\|(\widehat{g}(x)-g(x),\mathbf{E}[\widehat{m}(X,W)-m(X,W)|X=x])\|^2.
\end{aligned}$$

Using the definition of $T_k$ along with the eigenfunction property of Hermite functions that

$$\mathcal{F}q^K(-t,-tw)=\sqrt{2\pi}i^{k-1}q^K(-t,-tw),$$

we obtain

$$\begin{aligned}
\mathrm{Re}(T_k)(y,w)&=\int_{\mathbb{R}}q^K(-t,-tw)\cdot\cos(\frac{\pi\cdot K}{2}+ty)d\nu(t),\\
\mathrm{Im}(T_k)(y,w)&=\int_{\mathbb{R}}q^K(-t,-tw)\cdot\sin(\frac{\pi\cdot K}{2}+ty)d\nu(t).
\end{aligned}$$

Let $\xi_1=(\xi_w,\xi_y)$, then we have for the real part

$$\begin{aligned}
\|\nabla\mathrm{Re}(T_k)(\xi_1)\|^2=&\frac{\partial\mathrm{Re}(T_k)(\xi_1)}{\partial w}^2+\frac{\partial\mathrm{Re}(T_k)(\xi_1)}{\partial y}^2\\
\leq&\int_{\mathbb{R}}\frac{\partial q^K(-t,-t\xi_w)}{\partial w}^2\cos(\frac{\pi k}{2}+t\xi_y)^2t^2d\nu(t)\\
&+\int_{\mathbb{R}}q^K(-t,-t\xi_w)^2\sin(\frac{\pi k}{2}+t\xi_y)^2t^2d\nu(t)
\end{aligned}\tag{2.21}$$

$$\begin{aligned}
&\leq \sup_{b\in\mathbb{R}^2} \frac{\partial q^K(b_1,b_2)}{\partial b_2}^2 \cdot \int_{\mathbb{R}} t^2 d\nu(t) + \sup_{b\in\mathbb{R}^2} q^K(b_1,b_2)^2 \cdot \int_{\mathbb{R}} t^2 d\nu(t) \\
&\lesssim K,
\end{aligned}$$

which is due to the fact that the distribution $\nu$ has finite second moments, the boundedness of Hermite functions and the following property of the derivative of Hermite functions,

$$\partial q^k(b)/\partial b = \sqrt{\frac{k}{2}} q^{k-1}(b) - \sqrt{\frac{k+1}{2}} q^{k+1}(b).$$

The argument is analogous for the imaginary part and summarizing

$$\begin{aligned}
&\|\Pi_{dm}(x) - \Pi(x)\|^2 \\
\lesssim &\sum_{k=1}^{K} 2 \cdot K \cdot \|\widehat{g}(x) - g(x), \mathbf{E}[\widehat{m}(X,W) - m(X,W)|X=x]\|^2 \\
\lesssim &K^2 \cdot O_p(n^{-2\varphi}).
\end{aligned}$$

The proof for part (ii) is analogous. Here, (2.21) is only the derivative with respect to y and it immediately follows from (2.21) that

$$\|\nabla \mathrm{Re}(T_k)(\xi_1)\|^2 \lesssim 1$$

and thus,

$$\|\Pi_{dm}(x) - \Pi(x)\|^2 = O(K \cdot n^{-2\varphi}),$$

which concludes the proof. □

# Chapter 3

# A Simple Shape-Constrained Estimator for Semi(non)parametric Discrete Choice Models

## 3.1 Introduction

Parametric discrete choice models like Logit and Probit are important workhorses in applied econometrics as well as in the statistical literature on classification.

Distributional assumptions on nuisance parameters and functional form restrictions on structural parameters spawned the econometric literature on semiparametric discrete choice models. The main branch of this literature focuses on relaxing distributional assumptions on unobservable error terms, while retaining the standard linear index formulation ($X'\beta$) of the structural part in the model setup.

In this paper, I study a class of discrete choice models, where there is a set of $L$ mutually exclusive choice alternatives and choice probability functions follow the form $P[d_l = 1|X] = G_{0,l}(\phi_{l,1}(X, \beta_0), \ldots, \phi_{l,M}(X, \beta_0)$, where $d_l$ is a binary indicator coding whether alternative $l$ is chosen or not and any element $\phi_{l,j}(X, \beta_0)$ characterizes an "index" of the model. The functional form $\phi_l$ of the index is considered to be known, whereas the finite-dimensional parameter $\beta_0$ is unknown. In such multiple index models, finite-dimensional parameters $\beta_0$ are often the main parameter of interest and the unknown function $G_{0,l}$ is considered a nuisance rendering the model a semiparametric one. As $G_{0,l}$ is required for identifying partial effects or to

perform predictions, I do not consider it a nuisance and henceforth adopt the term semi(non)parametric for describing this class of models.

The model setup implies a set of $L-1$ moment conditions allowing identification and estimation of the model parameters. For $L > 2$, the model nests the classical multinomial choice model for which semiparametric estimation has been considered by Lee (1995). For $L = 2$, we have the popular binary choice model results which has been frequently studied e.g by Klein and Spady (1993). For the $L = 2$ case, there are other more general models with more then one index ($M > 1$) that are covered by the model setup, like various double-index models, which extend the binary model to allow for heteroskedasticity [Klein and Vella (2009)], endogeneity [Blundell and Powell (2004), Rothe (2009)] or sample selection [Klein et al. (2015), Escanciano et al. (2016)].

The typical procedure in the semiparametric estimation of such discrete choice models is to estimate $G_{0,l}$ nonparametrically and then to solve for $\beta_0$ over an empirical criterion function (in many cases maximum likelihood). However, these approaches generally disregard prior information on the functional form of $G_{0,l}$. In any case, the choice model setup implies by definition $0 \leq G_{0,l} \leq 1$ uniformly and, depending on the underlying choice model, $G_{0,l}$ is monotone increasing in some of its indices. For the extended binary model case ($L = 2$, $M > 1$), $G_{0,l}$ is monotone in the first index which determines the mean response, however, not necessarily in the second index, which is required to control for deviations from the full independence assumption of regressors and unobservable error terms. For the multinomial choice model, $G_{0,l}$ is monotone increasing in all indices, provided we assume the underlying choice model is the popular random utility model with full independence between regressors and errors, see e.g. Train (2009). Yet, also many other stochastic choice models usually impose monotonicity in some of the indices, see e.g. Breitmoser (2018) for different stochastic choice models other than random utility.

This work proposes a computationally simple estimator for both parameters $G_{0,l}$ and $\beta_0$, imposing shape constraints in the form of boundedness and monotonicity on $G_{0,l}$. The estimation follows a sieve GLS approach and shape constraints are imposed by considering a constrained I-Spline and B-Spline tensor sieve space. I-Spline basis functions have the same form as cumulative distribution functions and are therefore natural candidates for approximating the unknown $G_{0,l}$. Though these functions have been suggested in the statistical literature, there appears to be no econometric work making use of these basis functions yet. This work introduces how to incorporate these functions in a sieve estimation framework and presents results on approximation properties and sieve complexity. Imposing both boundedness

and monotonicity constraints has also an effect on the asymptotic properties of the estimator. It can be shown that imposing the full set of shape constraints allows for a smaller bound to the $L_2$-complexity of the sieve space compared to the standard unconstrained sieve case. This affects the convergence rate of the estimator in a weak "Fisher-like" norm that was originally introduced by Ai and Chen (2003). There, it has been established that a sufficiently fast convergence rate in this weak norm is required to obtain asymptotic normality of the estimates for $\beta_0$. For the weak norm the effects of imposing our set of shape constraints are comparable to a dimension-reduction on the functions $G_{0,l}$ as the convergence rate only hinges on the dimensionality of the non-monotonic arguments of $G_{0,l}$. In the case where $G_{0,l}$ is monotonically increasing in each of its arguments, like in the binary choice or a random utility multinomial choice model, the convergence rate in the weak norm can even reach the parametric rate of $\sqrt{n}$ provided the number of choices is sufficiently limited.

Imposing shape-constraints speeds up the optimal convergence rate in the weak norm and thereby weakens convergence rate restrictions on the sieve estimator in the weak norm that are generally required to obtain asymptotic normality of the estimate of $\beta_0$ or other smooth functionals of the parameter set. These results are novel in the context of semiparametric discrete choice models, but relate to results from the econometric literature that also show how imposing shape constraints affects the convergence rate in some weak norm, see Chetverikov and Wilhelm (2017) and establishes how shape constraints can have an effect on semiparametric estimators apart from mere improvements in finite samples. The resulting estimator is computationally attractive, as imposing the shape constraints results in a convenient quadratic programming problem. The estimator is thus easy to implement and computationally cheap. A Monte Carlo simulation study reveals a benefit from imposing shape constraints in finite samples and in comparison to a benchmark estimator from the semiparametric discrete choice literature.

**Literature** This work relates to the literature on semiparametric discrete choice models, where main attention has been given to the binary choice case and a quite extensive list of different estimators has accumulated to this day. Prominent examples are the maximum-score and smoothed maximum score estimators of Manski (1985) and Horowitz (1992), the single index estimators by (Ichimura, 1993) and Klein and Spady (1993), as well as the special regressor approach of Lewbel (2014). In comparison, the case of multiple and ordered choices has been much less explored. Maximum score estimators for the general multiple case have beeen studied

by Manski (1975), Fox (2007) and Yan (2014). The approach of Lewbel (2000) is also applicable to the multinomial case. General semiparametric multiple index models have been analyzed in Ichimura and Lee (1991) and Lee (1995).

Further, this paper relates to the statistical literature on shape-constrained estimation. A review on the role of shape constraints in econometrics is provided by Matzkin (1994) and Chetverikov et al. (2018). See also Horowitz and Lee (2017), Blundell et al. (2017), Freyberger and Reeves (2019), Compiani (2021) and the references therein for nonparametric estimation by directly imposing shape restrictions. Breunig and Chen (2023) provide a data-driven methods to conduct optimal testing of shape constraints. Estimation under shape constraints is discussed in Chen (2007), where various shape-preserving sieve spaces are presented. However, none of the sieve spaces presented there imposes the set of shape constraints required in this application, and I-Spline basis functions are not part of the discussion either. Additionally, many of the shape-preserving sieves, like cardinal B-Spline wavelet sieves, may be hard to implement practically and have not been applied in practical economic applications yet.

There is a large literature on nonparametric estimation under monotonicity constraints. For a review see the section in Chetverikov and Wilhelm (2017) and e.g. Delecroix and Thomas-Agnan (2000). The case of semiparametric estimation has received much less attention, see e.g. Wu and Sickles (2018). Mammen (1991) shows that the convergence rate of a constrained monotonic estimator, in a strong norm, is the same as for the unconstrained estimator. Chetverikov and Wilhelm (2017) consider estimation of a monotonic regression function in the context of a nonparametric IV model and show that imposing a monotonicity constraint provides a faster convergence rate in a weaker, truncated $L_2$-norm. Further, they find that imposing monotonicity is most beneficial in terms of a non-asymptotic error bound if the function has flat parts, which is the case in our setting where the tails of the function necessarily tend to 0 and 1, implying flatness over a possibly large region. For the semiparametric binary choice model Banerjee et al. (2009) have considered monotonicity constrained estimation of a baseline choice probability function and how to perform inference on structural parameters without the need to estimate nuisance parameters.

This paper is organized as follows. Section 3.2 introduces the model framework and gives examples for discrete choice models from the literature that fit into the framework. Section 3.3 discusses identification of parameters and outlines the estimation strategy. Therein, subsection 3.3.2 provides an overview of I-Spline basis

functions and their usage for approximating bounded, monotonic functions. Section 3.4 deals with the asymptotic properties of the sieve GLS procedure. The first subsection presents properties of the constrained I-Spline sieve space. The subsequent subsections derive consistency of the estimator in a strong norm, the convergence rate in a weak norm and asymptotic normality of certain smooth functionals of the parameter space. In section 3.5 the finite sample performance of the estimator is analyzed in a Monte Carlo simulation study and section 3.6 concludes.

## 3.2 Model Framework

Given a finite set of $L$ mutually exclusive choice alternatives, let $d = (d_1, \ldots, d_L)$ be a vector of binary indicator variables, where $d_l = 1$ if a decision-maker chooses alternative $l$ and $d_l = 0$ else.

In the most general sense, I analyze shape-constrained estimation in a class of choice models, where the choice probabilities can be expressed as a multiple-index model characterized by the following set of moment conditions:

$$P[d_l = 1|X] = G_{0,l}(\phi_{1,l}(X, \beta_0), \ldots, \phi_{M,l}(X, \beta_0)), \quad l = 1, \ldots, L-1. \tag{3.1}$$

All other observable variables entering the model are summarized in $X \in \mathbb{R}^{d_X}$. This typically constitutes the set of exogenous regressors, but may also include auxiliary variables such as control functions. $\beta_0 \in B \subseteq \mathbb{R}^{d_\beta}$ is a finite dimensional coefficent vector. Let $\theta_{0,l} = (G_{0,l}, \beta_0)$ denote the entire set of parameters associated with (3.1). The mapping $\phi_{l,m}(X, \beta_0) : \mathbb{R}^{d_X} \times B \to \mathbb{R}$ is a short-hand notation for the $m$-th linear index entering the $l$-th moment condition of the model. This notation is chosen to clarify that observables and parameters may be different for each index and each moment condition depending on the underlying discrete choice model. Since choice probabilities have to sum up to one, we implicitly impose the normalization $G_{0,L}(\cdot) = 1 - \sum_{l=1}^{L-1} G_{0,l}(\cdot)$ and thus the set of $L-1$ conditional moment restrictions is sufficient to characterize the model.

Note that any $\phi_{l,m}$ can also encompass known non-linear transformations of observables and parameters. I do not consider the case where $\phi_{l,m}$ includes additional nonparametric parameters, though the estimation strategy may be extendable to more general types of discrete choice models. See Matzkin (1991a) and Matzkin (1993) for nonparametric identification of the indices in discrete choice models.

Dealing with choice probabilities, it is clear that each $G_{0,l}(\cdot)$ is bounded between zero and one. Furthermore, it will often be the case that the function is monotone increasing in some or all of its arguments. The following examples illustrate some

of the most important models that are covered by the above model setup.

**Example 3.1. Multinomial Choice Model**

Consider the well known random utility model, see e.g. Train (2009).

$$Y_{il}^* = X_{il}'\beta_0 + \epsilon_{il}, \quad l = 1, ..., L$$
$$Y_i = \arg\max_{l=1,...,L} X_{il}'\beta_0 + \epsilon_{il},$$

where $X_{il} \in \mathbb{R}^q$ denotes the vector of covariates for alternative $l$, the random utility $Y_{il}^*$ for alternative $l$ is latent, $Y_i$ is a multinomial variable indicating an individuals observed choice and $\epsilon_{il}$ the scalar error term that is generally assumed to be fully independent of each regressor. Let $d_{il}$ be the dummy variable indicating whether individual $i$ chooses alternative $l$. The following equalities hold

$$\begin{aligned} P(d_{il} = 1|X_i) &= P(\epsilon_{ij} - \epsilon_{il} \leq (X_{il} - X_{ij})'\beta_0, \ \forall j = 1, \ldots, l-1, l+1, \ldots, L|X_i) \\ &= G_{0,l}((X_{il} - X_{i1})'\beta_0, ..., (X_{il} - X_{il-1})'\beta_0, (X_{il} - X_{il+1})'\beta_0, \\ &\quad ..., (X_{il} - X_{iL})'\beta_0) \end{aligned}$$

and imply an $L-1$-index model, where the same parameter $\beta_0$ enters each index, but where covariates of each index differ across choices. Analogously, it is possible to include covariates that do not vary across choices, but then the associated finite-dimensional parameters need to vary across choices.

In the random utility model $G_{0,l}$, is a joint cdf and thus, the function is monotone increasing in every argument. Parametric multinomial choice models typically impose the restriction $\epsilon_{ij} \perp X_{ik}$ for any pair $(j, k)$ from the choice set, but this is not required in a general semi(non)parametric model, and error terms may correlate across choices.

**Example 3.2. Binary Choice with Index Heteroskedasticity**

Klein and Vella (2009) study the model

$$d = \mathbb{1}\{X_1'\beta_0 + S(X_2'\gamma_0)\epsilon > 0\},$$

with $S(\cdot) : \mathbb{R} \to \mathbb{R}_+$, some unknown positive scale function. Variables in $X_1$ and $X_2$ can overlap, but should contain at least one continuous variable not included in the other one. It is immediate to see that $P(d = 1|X_1, X_2) = G_0(X_1'\beta, X_2'\gamma)$. In this case, $G_0$ is still monotone increasing in the first, however, generally not in the second argument.

**Example 3.3. Binary Choice with Endogeneity** The following triangular model is a special case of the models considered by Blundell and Powell (2004) and Rothe (2009):

$$\begin{aligned} d &= \mathbb{1}\{X'\beta_0 + \epsilon > 0\} \\ X^e &= Z'\pi + V, \end{aligned}$$

with $X = (X^e, X^{-e})$ and $Z = (Z^e, X^{-e})$. $X^e$ are endogenous variables in the outcome equation in the sense that the condition $X^e \perp \epsilon$ is violated. Let $V = X^e - Z'\pi$. In some applications, the triangular model may allow to use $V$ as a control function to achieve identification. In these settings, the following distributional exclusion restrictions hold

$$\epsilon\,|X, Z \;\sim\; \epsilon\,|X, V \;\sim\; \epsilon\,|V,$$

with $\sim$ denoting equality in distribution. This implies

$$\begin{aligned} P(d = 1|X, V) &= P(-\epsilon < X'\beta_0|X, V) \\ &= P(-\epsilon < X'\beta_0|V) \\ &= G_0(X'\beta_0, V) \\ &= G_0(X'\beta_0, X^e - Z'\pi), \end{aligned}$$

and by way of the distributional exclusion restrictions the following moment condition holds

$$E(d|X, Z) = G_0(X'\beta_0, X^e - Z'\pi),$$

which is in turn a model covered by our framework. Generally, we have only monotonicity in the first argument of $G_0$ but not in the second. Blundell and Powell (2004) and Rothe (2009) consider the case of nonparametrically generated regressors i.e $E[X^e|Z] = g(Z)$. The latter paper is explicitly concerned with how a first step nonparametric regression estimate of the above function affects properties of estimators of $\beta_0$. We do not consider nonparametric parameters entering the indices of the link function or for that matter first stage estimates entering the moment condition (so called nonparametrically generated regressors), nevertheless, relevant aspects of our estimation strategy may be extendable to these settings.

The list of models put forward here is not exhaustive. For instance our estimation approach can also be used in the binary model with sample selection studied by Klein

et al. (2015). Here, it can be used for estimating the probability function $P(d = 1|\delta = 1, \phi_1, \phi_2)$, where $\delta$ is the selection indicator and $\phi$'s are some linear indices. The structural choice probability function ($P(d = 1|\phi_1, \phi_2)$, so not conditional on selection $\delta$) can then be recovered using identification at infinity arguments, see Klein et al. (2015) for details. Furthermore, some of the models discussed in Escanciano et al. (2016) fit into the model setup if the set of shape constraints we study here does apply.

## 3.3 Identification and Estimation Strategy

In this section, I present a sieve GLS estimation strategy that conveniently allows to impose shape constraints on the nonparametric components of the parameters entering (3.1). The first subsection briefly discusses identification of the parameter $\theta_0$ from the moment condition (3.1). The second subsection introduces so called I-Spline basis functions and illustrates their usefulness for imposing all desired shape constraints on the sieve space. The last two subsections then introduce the estimation strategy, first, from a theoretical and second, from the point of view of practical implementation.

### 3.3.1 Identification

Throughout this paper the parameter $\theta_0 = (\beta_0, G_{0,1}, \ldots, G_{0,L-1})$ is assumed to be identified from the set of conditional moment restrictions $E[\rho(d, X, \theta_0)|X] = 0$, where $\rho$ is an $L - 1$ vector of moments with $l$-element

$$\rho_l(d, X, \theta_0) \equiv d_l - G_{0,l}(\phi_{l,1}(X, \beta_0), \ldots, \phi_{l,M}(X, \beta_0)).$$

The parameter space is $\Theta = B \times \prod_{j=1}^{L-1} \mathcal{G}_j$, where the infinite-dimensional parameter spaces $\mathcal{G}_j$ are to be defined later and $B$ is the space of the finite dimensional parameter.

We formulate the following abstract identification condition:

**Assumption 1.** *The parameter $\theta_0$ is uniquely identified from the conditional moment restrictions $E[\rho(d, X, \theta_0)|X] = 0$, i.e. $E[\rho(d, X, \theta)|X] = 0$ if and only if $\theta = \theta_0$.*

In general, identification conditions vary across discrete choice models. Since I summarize a class of models that fit into the multiple index model framework implied by (3.1), identification generally needs to be considered individually, e.g. separately

for each of the models in section 3.2. There are however general identification arguments for the class of linear multiple index models in Lemma 3 of Ichimura and Lee (1991). For sufficient conditions for identification see also the discussion in section 2.3.1 of Horowitz (1998). Identification conditions typically include that every linear index $\phi_{l,m}(X, \beta_0)$ contains a continuous explanatory variable with nonzero coefficient, which is not contained in any other index $\phi_{l,k}(X, \beta_0)$, $k \neq m$. This non-zero coefficient is normalized to one. Further, each true choice probability function $G_{0,l}(\cdot)$ needs to be differentiable in each argument and derivatives are not allowed to be almost surely linearly dependent with the constant function 1. Further, the matrix of regressors $X$ is assumed to have full rank to rule cases of perfect multicollinearity.

### 3.3.2 Estimation under monotonicity and boundedness constraints

Shape constrained estimation has been substantively studied in the nonparametric estimation literature. The case of monotonicity constrained estimation has received much attention, see the isotonic regression literature like Mammen (1991) in statistics, and also recent contributions from the econometrics literature, such as Horowitz and Lee (2017) and Chetverikov and Wilhelm (2017). Typically, monotonicity constrained estimation proceeds by introducing a set of linear constraints on the derivative of the estimated function over a finite grid of observations. Asymptotic theory requires the grid to grow finer with increasing sample size. See Horowitz and Lee (2017) or also Beresteanu (2004). An alternative is to rearrange the function estimate ex-post to enforce monotonicity, like in Chernozhukov et al. (2009).

Positivity constrained estimation has received less attention, although the applications, such as density estimation, are straightforward. In this case, specific transformations are employed before the estimation to ensure positivity of the estimates. In the context of sieve estimation, Chen (2007) discusses shape-preserving sieve spaces. Here, examples are given how to impose monotonicity by using monotonic spline wavelet sieves. There are, however, no applications of such a procedure in the applied literature probably due to the complexity of wavelet basis functions.

In the following, I present a computationally simple way to impose three shape constraints, i.e. monotonicity, positivity and boundedness from above by one, in a single estimation step. The procedure does not rely on any finite dimensional grid on which we enforce the constraints and thus, do not need the additional asymptotic theory on the grid size as in Horowitz and Lee (2017).

The approach makes use of so called I-Spline basis functions which were put forward by Curry and Schoenberg (1966) and Ramsay (1988). I-Splines did thusfar

not appear in econometric contexts but have been applied in the statistics literature to e.g. estimate bivariate distribution functions, see Wu and Zhang (2012).

We begin by defining M- and I-Spline basis functions and their properties:

**Definition 3.1.** *The following definitions are taken from Wu and Zhang (2012), which built up on Ramsay (1988). Consider a knot sequence* $t = (t_1, \dots, t_{n+k})$ *on the compact, real interval* $[U, L]$*:*

$$\begin{aligned} U = t_1 = \cdots = t_k \\ t_{n+1} = \cdots = t_{n+k} = L, \\ t_i < t_{i+k}, \quad \forall i. \end{aligned}$$

*Given the knot sequence* $t$ *the* $i$*-th M-Spline basis function of order* $k$ *is defined by the following recursion:*

*For* $l = 1$*:*

$$M_i(x|1,t) = \begin{cases} \frac{1}{t_{i+1}-t_i}, & \text{if } t_i \leq x \leq t_{i+1} \\ 0 & \text{else.} \end{cases}$$

*For* $l > 1$*:*

$$M_i(x|l,t) = \frac{l[(x-t_i)M_i(x|l-1,t) + (t_{i+k}-x)M_{i+1}(x|l-1,t)]}{(l-1)(t_{i+l}-t_i)}.$$

*M-Spline basis functions have the same properties as probability density functions and thus, are always non-negative and integrate to one.*

*I-Spline basis functions are then defined as:*

$$I_i(x|l,t) = \int_L^x M_i(u|l,t)du.$$

*Since I-splines are integrated M-Splines, they share properties of cumulative distribution functions. In particular, they are monotone increasing and bounded between zero and one and thus, natural candidates for approximating monotone choice probability functions.*

For expositional purpose, assume there is i.i.d data $(Y_i, X_i)$ with $X_i \in \mathcal{X}^d$ and $\mathcal{X}$ some compact interval on $\mathbb{R}$. Consider a standard nonparametric mean regression

model

$$Y = g(X) + U, \quad E[U \mid X] = 0, \tag{3.2}$$

where $g(\cdot)$ is known to be monotone increasing and bounded between zero and one. We first consider the univariate case $d = 1$ and use the short hand notation $I_i^l(x) \equiv I_i(x|l,t)$. Since I-Spline basis functions satisfy all our desired shape constraints, it is intuitive to approximate $g(\cdot)$ with a weighted sum of I-Spline basis functions. By constraining the weights to be positive and to sum up to at most one, it can be ensured that the fitted object obeys all shape constraints. The following constrained series estimator for $g(\cdot)$ can be formulated

$$\min_{\pi} \sum_{i=1}^{N}(y_i - \sum_{j=1}^{K} I_j^l(x_i)\pi_j)^2 \quad s.t. \ \forall j \ \pi_j \geq 0, \ \sum_{j=1}^{K} \pi_j \leq 1. \tag{3.3}$$

Abstracting for now from the properties of such an estimator, note that it is a computationally cheap way to impose the shape constraints since this is a quadratic programming problem, which can be solved very fast by standard software routines. Further, note that with estimates of $\pi_j$, we immediately have estimates for the derivative of the estimated function, making use of the link between I- and M-Spline functions shown earlier.

Furthermore, there is a link between I- and B-Spline basis functions that allows to view the constrained series estimator over an I-Spline basis as constrained series estimator with B-Spline basis functions. As pointed out by Wu and Zhang (2012), any I-Spline basis function of order $l$ can be expressed as sum of B-Spline basis functions of order $l + 1$. Specifically,

$$I_i^l(x) = \sum_{m=i}^{K} B_m^{l+1}(x).$$

Thus, the following constrained least squares problems yields the same fitted function as (3.3):

$$\min_{\alpha} \sum_{i=1}^{N}(y_i - \sum_{j=1}^{K} B_j^{l+1}(x_i)'\alpha_j)^2 \quad s.t. \ 0 \leq \alpha_1 \leq \cdots \leq a_K \leq 1. \tag{3.4}$$

Consequently, we can thus express any constrained series estimator with I-Spline basis functions as constrained series estimator with B-Spline basis functions under a different set of constraints. Theoretical properties of the latter are very well

understood and as presented in section 3.4, approximation properties of B-Spline sieves carry over to I-Spline sieves. In subsection 3.3.3, we see that in the practical setting of this estimation problem it is more convenient to enforce the coefficient constraints on I-Spline sieves compared to the analogous B-Spline sieve formulation. In the B-Spline case, the optimization problem in (3.4) is no quadratic programming problem and thus, more difficult to compute.

To estimate general multivariate functions, which are monotone increasing in each argument, it is straightforward to use the tensor product of I-Spline basis functions with above constraints on coefficients, see e.g. Chen (2007) for the use of tensor product sieve spaces for the estimation of multivariate functions.

This exposition ends by showing how to approximate a bivariate function $g(x, z)$ that is bounded between zero and one, but only monotonically increasing in its first argument by using basis functions obtained from a tensor product of I- and B-Spline basis functions. I choose I-Spline functions over the support of $x$ and B-Splines over $z$. Then solving

$$\min_{\pi} \sum_{i=1}^{N}(y_i - \sum_{j=1}^{K_1}\sum_{k=1}^{K_2} I_j^l(x_i)B_k^{l+1}(z)\pi_{jk})^2 \quad s.t. \;\; \forall j,k \;\; \pi_{j,k} \geq 0, \; \sum_{j=1}^{K_1}\sum_{k=1}^{K_2}\pi_{j,k} \leq 1 \tag{3.5}$$

yields a fitted function $\widehat{g}(x, z) = \sum_{j=1}^{K_1}\sum_{k=1}^{K_2} I_j^l(x_i)B_k^{l+1}(z)\widehat{\pi}_{jk}$ that satisfies the shape constraints of $g(x, z)$. Since all coefficients are positive and sum to at most one and B-Splines are bounded between zero and one, it is clear that the fitted $\widehat{g}(x, y)$ is between zero and one. Considering the derivative of $\widehat{g}(x, y)$ with respect to the first argument, we can see that the function again is monotone in $x$ by positivity of B-Splines. Taking the derivative with respect to $y$, we see that the sign is determined by the derivatives of the B-Spline basis functions which are left unrestricted.

The above problem can again be formulated as constrained series estimation with B-Spline basis. In particular, solving

$$\min_{\alpha} \sum_{i=1}^{N}(y_i - \sum_{j=1}^{K_1}\sum_{k=1}^{K_2} B_j^{l+1}(x_i)B_k^{l+1}(z)\alpha_{jk})^2 \quad s.t. \;\; \forall j,k \;\; 0 \leq \alpha_{j,k} \leq 1, \tag{3.6}$$

$$\alpha_{j+1,k} - \alpha_{j,k} \geq 0$$

provides us with the same minimizer as (3.5).

In the following subsection, I formulate a sieve GLS estimator with an I-Spline or mixed I-spline/ B-Spline tensor sieve, where shape constraints can be respected

by imposing linear constraints on the sieve coefficients. The GLS criterion, allows to implement the estimator as a stepwise (so called profile sieve) procedure, which makes use of the computational appeal of the constrained series estimation step introduced in (3.3) and (3.5). This is in contrast to more natural candidate criterion functions such as sieve maximum likelihood.

### 3.3.3 Sieve GLS Estimation

We propose the following sieve GLS estimator of $\theta_0$:

$$\widehat{\theta} = \arg\min_{\theta \in \Theta_K^c} \sum_{i=1}^{N} \rho(d_i, X_i, \theta)' \, \Sigma(X)^{-1} \rho(d_i, X_i, \theta) \tag{3.7}$$

with sieve space $\Theta_K^c = B \times \prod_{j=1}^{L} \mathcal{G}_{K,j}^c$ and generalized residuals

$$\begin{aligned}
\rho(d_i, X_i, \theta) &= (\rho_1(d_i, X_i, \theta), \ldots, \rho_{L-1}(d_i, X_i, \theta)), \\
\rho_l(d_i, X_i, \theta) &= d_{il} - G_l(\phi_{l,1}(X_i, \beta), \ldots, \phi_{l,M}(X_i, \beta)), \\
m(X_i, \theta) &= \mathbf{E}[\rho(d_i, X_i, \theta) | X_i],
\end{aligned}$$

where for now $\Sigma(X)$ is a positive-definite weighting matrix, which is considered to be known. Further, the following short-hand notation for the indices is applied $\phi = (\phi_1, \ldots, \phi_M)$ with $\phi_j = \phi_j(X, \beta)$. Here, the dependency of $\phi$ on $l$ is implicitly suppressed, i.e. the fact that any index-function may be specific to the moment condition (or rather choice probability function) at hand, as it is not relevant for the formulation of sieve spaces.

It remains to characterize the shape constrained sieve space $\Theta_K^c$. Typically in the sieve estimation literature, $G_0$ is assumed to be an element of a smoothness class $\mathcal{G}^u$ such as a Sobolev or Hölder space. Then an appropriate sieve space $\mathcal{G}_K^u$ of dimension $K$ is determined, which can approximate elements of $\mathcal{G}^u$ well. The most convenient choices are finite-dimensional linear sieves consising of polynomial or spline basis functions.

In this setting, we have additional knowledge on the shape of $G_0$, atop of mere smoothness properties. $G_0$ is bounded between zero and one and monotonically increasing in some of its arguments. Thus, we know that $G_0$ belongs to a space $\mathcal{G}^c \subseteq \mathcal{G}^u$. This a priori information on the shape of $G_0$ can be imposed on every sieve space, so we consider a constrained sieve $\mathcal{G}_K^c \subseteq \mathcal{G}_K^u$. Here, an unconstrained sieve is a standard B-Spline sieve, whereas the constrained sieve is a B-Spline sieve with constraints on the sieve coefficients or by equivalence a (mixed) I-Spline sieve

with constraints.

In particular, we define

$$\mathcal{G}^c = \{G \in \mathcal{G}^u : 0 \leq G \leq 1, \;\; D^\alpha G \geq 0 \;\; \forall \alpha \in \mathcal{M}\},$$

where $\mathcal{G}^u$ is a smoothness class that will be further specified in the next section and also the differential operator $D^\alpha$ is introduced there. The set $\mathcal{M}$ indicates the arguments of the function $G$ on which it is monotonically increasing. Specifically define,

$$\mathcal{M} = \{\alpha \in \{0,1\}^M : \; [\alpha] = 1, \alpha^{(i)} = 1 \\ \text{iff G is non-decreasing in its i-th argument}\}.$$

The shape constrained sieve space is then

$$\mathcal{G}_K^c = \{G : G = \sum_{j=1}^{K} p_j(\phi)\pi_j, \; \phi = (\phi_1, \ldots, \phi_M), \; \forall j \; \pi_j \geq 0, \; \sum_{j=1}^{K} \pi_j \leq 1\},$$

where the vector of basis functions $p$ is the tensor product of I-Spline basis functions on those elements of $\phi$ where the function is monotonic and of B-Spline basis functions for the remaining arguments. Without loss of generality assume that the first $m$ arguments of the function are supposed to be monotonic i.e.

$$\sum_{j=1}^{K} p_j(\phi)\pi_j = \sum_{j_1=1}^{K_1} \cdots \sum_{j_M=1}^{K_M} I_{j_1}^l(\phi_1) \ldots I_{j_m}^l(\phi_m) B_{j_{m+1}}^{l+1}(\phi_{m+1}) \ldots B_{j_M}^{l+1}(\phi_M) \pi_{j_1,\ldots,j_M},$$

where $K = \prod_{l=1}^{M} K_l$ and where we denote $K_{(m)} = \prod_{l=1}^{m} K_l$ as the sieve dimension spanning the monotonic parts of the function and $K_{(\neg m)} = \prod_{l=m+1}^{M} K_l$ spanning the remaining arguments. We introduce the constrained, univariate I-Spline sieve of dimension $K$

$$\mathcal{I}_K^c = \{G : \sum_{j=1}^{K} I_j^l(\phi)\pi_j, \; \forall j \; \pi_j \geq 0, \;\; \sum_{j=1}^{K} \pi_j \leq 1\}$$

and the univariate B-Spline sieve with positivity and boundedness constraints of dimension $K$

$$\mathcal{B}_K^c = \{G : \sum_{j=1}^{K} B_j^{l+1}(\phi)\alpha_j, \;\; 0 \leq \alpha_j \leq 1\}.$$

Then, following the definition of tensor sieve spaces of Chen (2007) p. 5573, we have that

$$\mathcal{G}_K^c = \mathcal{I}_{K_1}^c \otimes \cdots \otimes \mathcal{I}_{K_{(m)}}^c \otimes \mathcal{B}_{K_{(m)}+1}^c \otimes \cdots \otimes \mathcal{B}_{K_M}^c = \left( \bigotimes_{l=1}^{m} \mathcal{I}_{K_l}^c \right) \otimes \left( \bigotimes_{l=m+1}^{M} \mathcal{B}_{K_l}^c \right).$$

For the sake of completeness, we also introduce the following constrained B-Spline sieve space $\widetilde{\mathcal{G}}_K^c$ that is equivalent to $\mathcal{G}_K^c$ is denoted as

$$\widetilde{\mathcal{G}}_K^c = \{G : \sum_{j_1=1}^{K_1} \cdots \sum_{j_M=1}^{K_M} B_{j_1}^{l+1}(\phi_1) \ldots B_{j_M}^{l+1}(\phi_M)\gamma_{j_1,\ldots j_M}, \\ 0 \leq \gamma_{j_1,\ldots j_M} \leq 1 \text{ and } \gamma_{j_1,\ldots,j_l+1,\ldots,j_M} \geq \gamma_{j_1,\ldots,j_l,\ldots,j_M} \text{ for } l \leq m\}.$$

The following subsection discusses the implementation of the sieve GLS estimator in a stepwise fashion.

### 3.3.4 Profile Sieve Procedure

For practical use, I propose the following profile sieve GLS procedure. $\Sigma(X)$ is a known weighting matrix. The optimally weighted case and the case, where an estimate from the data $\widehat{\Sigma}(X)$ is considered, is omitted for now, but can be readily incorporated in the procedure, see Ai and Chen (2003) or section 4.2, 4.3 of Chen (2007). Though the theoretical analysis considers joint estimation of the components of $\theta$ as in (3.7), the implementation of the estimator I suggest here, follows an implicit two-stage approach.

**Step 1** For each $\beta \in B$ calculate the optimal function:

$$\widetilde{G}(\beta) = \arg\min_{G \in \mathcal{G}_K^c} \sum_{i=1}^{N} \rho(d_i, X_i, G, \beta)' \Sigma(X)^{-1} \rho(d_i, X_i, G, \beta). \tag{3.8}$$

**Step 2** Calculate the optimal $\beta$ given the first stage estimates:

$$\widehat{\beta} = \arg\min_{\beta \in B} \sum_{i=1}^{N} \rho(d_i, X_i, \widetilde{G}(\beta), \beta)' \Sigma(X)^{-1} \rho(d_i, X_i, \widetilde{G}(\beta), \beta). \tag{3.9}$$

The optimal $(\widehat{G}, \widehat{\beta})$ solving equation (3.7) is then $(\widetilde{G}(\widehat{\beta}), \widehat{\beta})$.

This practical implementation rationalizes the use of a GLS procedure and of the I-Spline formulation of sieve spaces instead of sieve maximum likelihood over

constrained B-Spline sieves, which appears to be the most natural candidate criterion. For each $\beta$ in Step 1, the optimization problem is a quadratic programming problem, which can be solved fast by standard software routines. This makes the evaluation of the criterion function in Step 2 cheap and global search algorithms like e.g. differential evolution can perform Step 2 effectively. By avoiding the usage of a log-likelihood procedure, we circumvent additional numerical problems from performing divisions with nonparametric estimates, which are typically corrected by introducing some sort of trimming, see e.g. the approach of Klein and Spady (1993) as an example for such maximum-likelihood routines.

Note, that a GLS procedure can nevertheless also provide us with semiparametrically efficient estimates of $\beta$, see Chen (2007), so this feature is not exclusive to maximum likelihood.

## 3.4 Asymptotic properties

First, some notation needs to be introduced to specify the smoothness class $\mathcal{G}^u$ containing $G_0$. Let $\Phi = \Phi_1 \times \cdots \times \Phi_M$ be the cartesian product of compact, real intervals $\Phi_1, \ldots, \Phi_M$. Suppose $\alpha = (\alpha_1, \ldots, \alpha_M)$ is a $M$-tupel of nonnegative integers and $[\alpha] = \alpha_1 + \cdots + \alpha_M$. Define the differential operator:

$$D^\alpha = \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1} \ldots \partial x_M^{\alpha_M}}$$

such that $D^\alpha G$ describes the $\alpha$-partial derivative of $G(\phi_1, \ldots, \phi_M)$.

Let $\eta$ be a nonnegative integer, $\gamma \in (0, 1]$ and $p = \eta + \gamma$. A function $G$ on $\Phi$ is called $p$-smooth if it is $\eta$-times continuously differentiable ($G \in C^\eta(\Phi)$) and $D^\alpha G$ satisfies a Hölder condition with exponent $\gamma$ for all $\alpha$ with $[\alpha] = \eta$, i.e. there exists a constant $C$ such that $|D^\alpha G(a) - D^\alpha G(b)| \leq C\|a - b\|^\gamma$ for $a, b \in \Phi$.

The space of all p-smooth functions on $\Phi$ is denoted by $\Lambda^p(\Phi)$ . Define the Hölder space with smoothness $p$:

$$\Lambda_\infty^p(\Phi) = \{G \in C^\eta(\Phi) : \sup_{[\alpha]\leq\eta} \sup_{\phi\in\Phi} |D^\alpha G(\phi)| < \infty, \sup_{[\alpha]=\eta} \sup_{a,b\in\Phi, a\neq b} \frac{|D^\alpha G(a) - D^\alpha G(b)|}{\|a - b\|^\gamma} < \infty\}$$

Analogously it can be defined as

$$\Lambda_\infty^p(\Phi) = \{G \in C^\eta(\Phi) : \|G\|_{\eta,\gamma} < \infty\}$$

with Hölder norm

$$\|G\|_{\eta,\gamma} = \sup_{[\alpha]\leq\eta}\sup_{\phi\in\Phi}|D^\alpha G(\phi)| + \sup_{[\alpha]=m}\sup_{a,b\in\Phi,a\neq b}\frac{|D^\alpha G(a) - D^\alpha G(b)|}{\|a-b\|^\gamma}.$$

Hölder spaces can be well approximated by various linear sieve spaces, in particular, $\Lambda^p_\infty(\Phi)$ on a bounded domain is compact with respect to the sup-norm, see Freyberger and Masten (2015), also for a much more general discussion on the compactness of prominent smoothness classes based on norm bounds.

In the following, define $\mathcal{G}^u = \Lambda^p_\infty(\Phi)$ as the relevant smoothness class. The parameter space $\mathcal{G}^c$ and the corresponding sieve space $\mathcal{G}^c_K$ are defined as in the preceding section. Consider an arbitrary space of function $\mathcal{F}$, where each $f \in \mathcal{F}$ is $f : \phi \mapsto f(\phi)$. The envelope function is defined as any function such that for any $f \in \mathcal{F}$ and $\phi$ in the support, $|f(\phi)| \leq F(\phi)$. Let $N_{[\,]}$ denote the bracketing number of the space $\mathcal{F}$, which is the minimum number of $\delta$-brackets required to cover $\mathcal{F}$ , see e.g. Chen (2007) p. 5594 for more details. Next, we present some properties of the constrained sieve space $\Theta^c_K = \mathcal{G}^c_K \times B$. To this end, let $\|g\|_\infty = \sup_{\phi\in\Phi}|g(\phi)|$ denote the sup-norm of some function.

**Assumption 2.** *(i) For any $j = 1,\dots,M$, the support $\Phi_j$ is a compact interval on $\mathbb{R}$ (ii) $G_{0,l} \in \Lambda^p_\infty(\Phi)$ for any $l = 1,\dots,L-1$ and (iii) the spline order satisfies $l \geq \eta$.*

**Lemma 3.1.** *Under Assumption 2, it holds that (i) $\Theta^c_k \subseteq \Theta^c_{k+1} \subseteq \Theta^c$ for any $k = 1,2,\dots$ and (ii) for each $\theta \in \Theta^c$, there exists $\Pi_K\theta \in \Theta^c_K$ such that $\|\Pi_K\theta - \theta\|_\infty = O(K^{-p/d})$ (iii) if $K_{(\neg m)} \geq 1$, then $N_{[\,]}(\epsilon,\Theta^c_K,L^r(\Phi)) = O(K_{(\neg m)})$, if $K_{(\neg m)} = 0$, then $N_{[\,]}(\epsilon,\Theta^c_K,L^r(\Phi)) = O(1)$ for $r \geq 1$.*

Properties (i) and (ii) are standard approximating properties of a sieve space that are required to establish consistency of any sieve estimator, see e.g. Condition 3.2 in Chen (2007). The approximating property (ii) for I-Spline sieves follows from arguments in Wu and Zhang (2012). Condition (iii) shows that the complexity of the sieve space, measured by the bracketing number, only increases in the number of basis functions that span the non-monotonic parts of the function. In the special case where the function is univariate and non-decreasing, the complexity of the sieve space is constant and does not increase in the number of basis functions. An example for that is the case of the binary choice model. This bound is an important property that impacts the asymptotic results following in this section. We will see that this property results in a faster convergence rate in a certain weak norm compared to the case, where no shape constraints are imposed on the sieve space.

The assumptions required for Lemma 3.1 are standard. The support of any index needs to be compact, which is the case for any linear index if the support of $X$ and

the parameter space $B$ are compact. Dealing with indices on a non-compact domain (e.g. via covariates supported on the entire real line) is possible by incorporating a known strictly monotonic transformation into the formulation of any $\phi$ that maps from the real line into a compact set such as e.g. the arctan function. For the mere result of Lemma 2, part (ii) of Assumption 2 can be weakened as we only need that the function is $\eta$-times continuously differentiable.

There are alternative shape-preserving sieves that are appropriate for approximating nondecreasing bounded functions. Examples are the cardinal B-Spline wavelet sieves with nondecreasing coefficient sequence in section 2.3.5 in Chen (2007). However, these sieves are fairly harder to implement and apply in practice and we cannot easily impose the specific boundedness constraints, which we require in the discrete choice context.

For deriving the asymptotic properties of the estimates I proceed with the following steps in the next subsections. First, we prove consistency in a strong norm, i.e. $\|\widehat{\theta} - \theta_0\|_c = \|\widehat{\beta} - \beta_0\|_E + \sum_{l=1}^{L-1} \|\widehat{G}_l - G_{l0}\|_\infty = o_p(1)$. Building on this result, we can derive the convergence rate in a certain "Fisher-like" weak norm in the second subsection. Here, the effect of imposing the shape constraints materializes in the asymptotic analysis. The next subsection shows that under a fast enough convergence rate in the weak norm, we can obtain asymptotic normality of the estimated $\widehat{\beta}$ and other smooth functionals of the parameter $\widehat{\theta}$ such as average partial effects. For this, we can rely on established theories of Shen (1997), Ai and Chen (2003) or Chen (2007).

### 3.4.1 Consistency

We derive consistency of the estimator in the norm

$$\|\widehat{\theta} - \theta_0\|_c = \|\widehat{\beta} - \beta_0\|_E + \sum_{l=1}^{L-1} \|\widehat{G}_l - G_{0,l}\|_\infty,$$

with $\|\cdot\|_E$ denoting the euclidean distance of the finite-dimensional parameter vector and $\|G_l - G_{0,l}\|_\infty = \sup_{x \in \mathcal{X}} \sup_{\beta \in B} |G_l(\phi_l(x, \beta)) - G_{0,l}(\phi_l(x, \beta))|$ for any $l$. The following set of assumptions is imposed to obtain consistency of the estimator.

**Assumption 3.**
*(i) The data $Z_i = (d_i, X_i)$ is i.i.d. across $i = 1, \ldots, N$.*
*(ii) Any weighting matrix $\Sigma(X)$ is real and its largest and smallest eigenvalues are bounded and bounded away from zero, uniformly for all $X$.*

*(iii) for any* $l = 1, \ldots, L-1$ *and* $G \in \Theta_K^c$, $\sup_{x \in \mathcal{X}} \sup_{\beta \in B} \|\nabla G_l(\phi(X, \beta))\|_E$ *is bounded and bounded away from zero and* $\|\phi_l(X, \beta_1) - \phi_l(X, \beta_2)\|_E \leq U(X)\|\beta_1 - \beta_2\|_E$ *for some random variable* $U(X)$ *with* $\mathbf{E}[U(X)] < \infty$.
*(iv)* $K \to \infty$ *and* $K = o(n)$.

**Theorem 3.1.** *Under Assumptions 1-3, it holds that* $\|\widehat{\theta} - \theta_0\|_c = o_p(1)$.

Assumptions (i), (ii), (iv), along with Assumption 2 and the results in Lemma 3.1 (i), (ii) are standard assumptions for obtaining consistency of sieve M- or minimum distance estimators, see e.g. Remark 3.3 and 3.4 in Chen (2007). Part (iii) of Assumption 3 implies that the criterion function is Lipschitz over $\Theta_K^c$ with respect to the metric $\|\cdot\|_c$, which is a condition that is also standard, see Condition 3.5M (ii) and 3.5MD (ii) in Chen (2007). The latter part of the assumption depends on the underlying discrete choice model and is satisfied in the linear index case, when regressors have finite expectation. $\phi_l$ may include a known nonlinear transformation if it is also Lipschitz. The rate requirement in (iv) cannot be relaxed in our context. This is due to the fact that the result on the bracketing entropy of the constrained sieve space $\Theta_K^c$ in Lemma 3.1 (iii) does not generalize to the case of the consistency metric $\|\cdot\|_c$ or other metrics that are uniform over the finite dimensional parameter space $B$.

### 3.4.2 Convergence Rate in a Weak Metric

In order to derive asymptotic normality of the parametric components of $\theta$, it is typically required that the convergence rate of $\theta$ is sufficiently fast in some metric that is locally equivalent to the GLS criterion function. This metric may be weaker than the consistency metric $\|\cdot\|_c$. Ai and Chen (2003) have established that this is the case if the convergence rate in some so called Fisher-like norm is faster than $n^{-1/4}$. The Fisher-like norm being weaker than the usual $L_2$ and sup-norms.

I follow their approach and derive the convergence rate in the Fisher-like norm. The matter is simplified by the fact that we do not have endogenous variables entering the difference of moment conditions $\rho(Z, \theta_1) - \rho(Z, \theta_2)$ in our setting and thus there is no need for first stage estimates of conditional expectation functions. Therefore, we are able to rely on general convergence rate results for sieve M-estimators. However, when deriving the rate, we can make use of the result on constrained sieve space complexity in Lemma 3.1. Imposing shape constraints reduces the complexity of the underlying sieve space and if the discrete choice model exhibits monotonically increasing choice probability functions, this will speed up the rate of convergence in the Fisher-like norm.

To begin the exposition, we first need to introduce some necessary terminology. This is standard and analogous to Ai and Chen (2003) or Chen (2007). Assume that for any $\theta_1, \theta_2 \in \Theta$ there exists a continuous path $\theta(\tau) = \theta_1 + \tau(\theta_2 - \theta_1)$ with $\theta(0) = \theta_1$ and $\theta(1) = \theta_2$ such that $\{\theta(\tau) : \tau \in [0,1]\} \subset \Theta$.
The directional derivative in direction $[\theta_2 - \theta_1]$ evaluated at $\theta_1$ is defined as

$$\frac{d\rho(d, X, \theta_1)}{d\theta}[\theta_2 - \theta_1] \equiv \frac{d\rho(d, X, \theta_1 + \tau(\theta_2 - \theta_1)}{d\tau}\Big|_{\tau=0}$$

and analogously

$$\frac{d\rho(d, X, \tilde{\theta})}{d\theta}[\theta_2 - \theta_1] \equiv \frac{d\rho(d, X, \theta_1 + \tau(\theta_2 - \theta_1)}{d\tau}\Big|_{\tau=\tilde{\tau}}$$

where $\tilde{\theta} = \theta(\tilde{\tau})$ i.e. the directional derivative is evaluated at some $\tilde{\theta}$ between $\theta_1$ and $\theta_2$.

The following norm will be from now on referred to as (weighted) Fisher-like norm:

$$\|\theta - \theta_0\| := \sqrt{E\left[\left|\left|\frac{d\rho(d, X, \theta_0)}{d\theta}[\theta - \theta_0]\right|\right|^2_{\Sigma(X)^{-1}}\right]},$$

where we make use of the short-hand notation

$$\left|\left|\frac{d\rho(d, X, \theta_0)}{d\theta}[\theta - \theta_0]\right|\right|^2_{\Sigma(X)^{-1}} = \frac{d\rho(d, X, \theta_0)}{d\theta}[\theta - \theta_0]'\Sigma(X)^{-1}\frac{d\rho(d, X, \theta_0)}{d\theta}[\theta - \theta_0].$$

Due to the fact that the difference $\rho(d, X, \theta_0) - \rho(d, X, \theta)$ does not depend on the endogenous variable $d$ we likewise have

$$\|\theta - \theta_0\| = \sqrt{E\left[\left|\left|\frac{dm(X, \theta_0)}{d\theta}[\theta - \theta_0]\right|\right|^2_{\Sigma(X)^{-1}}\right]},$$

which will be the preferred notation going onwards.

Furthermore, the following quantities that reflect the least-favorable directions of the semiparametric estimation problem need to be introduced. Let $\overline{\boldsymbol{V}}$ denote the closure of the linear span of the space $\Theta^c - \theta_0$ with respect to $\|\cdot\|$ where $\overline{\boldsymbol{V}} = \mathbb{R}^{d_\beta} \times \overline{\mathcal{G}}^c$.

Let

$$\begin{aligned} D_{w_j}(X) =& \frac{dm(X,\beta,h_0(\cdot))}{d\beta_j}\Big|_{\beta=\beta_0} - \frac{dm(X,\beta_0,h_0(\cdot)+\tau w_j(\cdot))}{d\tau}\Big|_{\tau=0} \\ =& \frac{dm(X,\theta_0)}{d\beta_j} - \frac{dm(X,\theta_0)}{dh}[w_j] \end{aligned}$$

and $D_w = (D_{w_1}, \ldots, D_{w_{d_\beta}})$ with $w = (w_1, \ldots, w_{d_\beta})$. For any $j = 1, \ldots, d_\beta$, let $w_j^*$ be the argument that satisfies

$$\inf_{w_j \in \mathcal{W}} \mathbf{E}[D_{w_j}(X)'\Sigma(X)^{-1}D_{w_j}(X)]$$

and

$$\mathbf{E}[D_{w^*}(X)'\Sigma(X)^{-1}D_{w^*}(X)] = \left(\frac{dm(X,\theta_0)}{d\beta} - \frac{dm(X,\theta_0)}{dh}[w*]\right)$$

with

$$\frac{dm(X,\theta_0)}{dh}[w^*] = \left(\frac{dm(X,\theta_0)}{dh}[w_1^*], \ldots, \frac{dm(X,\theta_0)}{dh}[w_{d_\beta}^*]\right).$$

In order to obtain the convergence rate in $\|\cdot\|$, we require the following additional assumptions.

**Assumption 4.** *(i)* $\|\nabla G_{0,l}(\phi_l(X,\beta_0))'J_{\phi_l}(X,\beta_0)\|_{L_2(X)} < \infty$ *for any* $l = 1, \ldots, L-1$. *(ii)* $\mathbf{E}[D_{w^*}(X)'D_{w^*}(X)]$ *is finite positive definite and* $tr\left(\mathbf{E}\left[\frac{dm(X,\theta_0)}{dh}[w^*]'\frac{dm(X,\theta_0)}{dh}[w^*]\right]\right)$ *is finite. (iii)* $\sup_{x\in\mathcal{X}}\sup_{\beta\in B}\|\nabla G_{0,l}(\phi(X,\beta))\|_E$ *is bounded and bounded away from zero*

Here, $J_{\phi_l}(X,\beta_0)$ denotes the Jacobian of $\phi_l$ with respect to $\beta$ and the operator $tr(\cdot)$ the trace of a matrix. Assumption 4 are mere smoothness and regularity conditions on the choice probability functions that are easily satisfied for the cdf-like functions in the discrete choice setup. Condition (ii) is a standard regularity condition in semiparametric estimation, see Condition 4.1(ii) in Chen (2007) and will appear in the asymptotic variance-covariance matrix of $\widehat{\beta}$.

Then, the following result for the convergence rate holds.

**Theorem 3.2.** *Under Assumptions 1-4 it follows that*

$$\|\widehat{\theta} - \theta_0\| = O_p\left(\max\left\{K^{-p/M}, \sqrt{\frac{K_{(\neg m)}}{n}}\right\}\right)$$

The result follows from applying the general convergence rate result for sieve M-estimators, Theorem 3.2. in Chen (2007), which itself builds on older results like Chen and Shen (1998). Imposing the proposed set of shape constraints reduces the asymptotic variance term from $K$ to $K_{(\neg m)}$. In order to obtain consistency in the first place, we also require by Assumption 3 (iv) that $K_{(m)} = o(n)$. Now, we can choose $K = K_{(\neg m)}^{M/(M-m)}$. Then, balancing bias and variance, we obtain $\|\theta - \theta_0\| = O_p(n^{-p/(2p+M-m)})$ as the optimal convergence rate. Therefore, the optimal rate in the Fisher-like norm depends only on the dimension of the non-monotonic parts of the function. If the function has only monotonically increasing arguments, the asymptotic variance part is of order $n^{-1/2}$ and the optimal convergence rate may be the parametric rate. By the rate restrictions given in Assumption 3 (iv) however the parametric rate will only hold if $M < 2p$. In a multinomial choice model with $L$ choices, $M = L - 1$ and e.g. if $\lfloor p \rfloor = 2$, then the parametric rate is only possible if $L \leq 5$.

Generally, shape-constrained estimation does not improve the convergence in rate in strong norms like sup or $L_2$ norms, see e.g. Mammen (1991) or Chetverikov et al. (2018). However, for a weak norm like the Fisher-like norm this may be possible, as illustrated in the nonparametric IV study by Chetverikov and Wilhelm (2017).

### 3.4.3 Asymptotic Normality of Smooth Functionals of $\theta$

A main interest in semiparametric estimation is inference on the parametric components of $\theta$ or other economically relevant smooth functionals of $\theta$ such as average partial effects. Ai and Chen (2003) have established asymptotic normality of sieve minimum distance estimates of $\beta$ and generalizations to smooth functionals of $\theta$ can be found in e.g. Chen (2007). This section begins with establishing asymptotic normality of $\widehat{\beta}$ for which we can directly apply a result for sieve GLS estimators from section 4.2.2. in Chen (2007). To this end, introduce for some $\lambda \in \mathbb{R}^{d_\beta}$, $\|\lambda\|_E = 1$,

$$v_\beta^* = (\mathbf{E}[D_{w_j^*}(X)'\Sigma(X)^{-1}D_{w_j^*}(X)])^{-1}\lambda$$

and $v_h^* = -w^* v_\beta^*$ and $v^* = (v_\beta^*, v_h^*)$. The following additional assumptions are required.

**Assumption 5.** *(i)* $\beta_0 \in int(B)$ *(ii)* $\mathbf{E}[D_{w_j^*}(X)'\Sigma(X)^{-1}D_{w_j^*}(X)]$ *is positive definite (iii) there is* $\Pi_K v^* \in \Theta_K^c$ *such that* $\|\Pi_K v^* - v^*\| \cdot \|\theta - \theta_0\| = o(n^{-1/2})$ *(iv)* $\Sigma_0(X) = Var[\rho(Z, \theta_0)]$ *is positive definite and uniformly bounded over* $X$ *(v)* $\rho(Z, \theta)$ *is twice continuously pathwise differentiable with respect to* $\theta \in \Theta^c$ *for all* $\|\theta - \theta_0\| = o(1)$ *(vi)* $\|\widehat{\theta} - \theta\| = o_p(n^{-1/4})$

These assumptions are analogous to those in Ai and Chen (2003) for the general sieve minimum distance estimator. The rate restriction in (iii) is easily satisfiable if we choose $\|\Pi_K v^* - v^*\| = O(K^{-p/d})$, i.e. the same dimension parameter as for our constrained sieve space. The rate (vi) is a necessary condition in Ai and Chen (2003) (see their Theorem 3.1) and related literature on semiparametric estimation (see the references therein) to derive asymptotic normality of $\widehat{\beta}$. By our Theorem 3.2, however, this rate is more easily obtainable if one imposes our set of shape constraints as this will result in a generally faster rate. In many cases, like for the standard binary choice model, these rate restrictions can even be neglected as the rate in the Fisher-like norm is already of order $\sqrt{n}$. The following result can be established.

**Theorem 3.3.** *Under Assumptions 1-5 it holds that*

$$\sqrt{n}(\widehat{\beta} - \beta_0) \to N(0, V_1^{-1} V_2 V_1^{-1})$$

*with*

$$V_1 = \mathbf{E}[D_{w^*}(X)'\Sigma(X)^{-1} D_{w^*}(X)],$$
$$V_2 = \mathbf{E}[D_{w^*}(X)'\Sigma(X)^{-1}\Sigma_0(X)\Sigma(X)^{-1} D_{w^*}(X)].$$

Similar asymptotic normality results can be derived for other smooth functionals of the parameter vector. In applications, it is often of interest to study the functional $f(\theta) = \mathbf{E}[\partial G_l(\phi_l(X,\beta))/\partial X_j]$ which is the average partial effect of a change in a one-dimensional covariate $X_j$ given a choice model with parameters $\theta$. The estimated average partial effects are important objects in economic analyses and will also be of interest in the next section on Monte Carlo simulations illustrating the finite sample performance of the constrained sieve GLS estimator. The following results point out that, just like $\beta_0$, the average partial effects of discrete choice models under study here, are also $\sqrt{n}$ estimable and asymptotically normal. Under the following assumptions, we can derive an asymptotic normality result for the average partial effect of a discrete choice model.

**Assumption 6.** *For each choice $l$, index $i$ and covariate $X_j$, it holds that*
*(i)* $\mathbf{E}[|\partial\phi_{l,i}(X,\beta_0)/\partial X_j|] < \infty$,
*(ii)* $\|\nabla\phi_l(X,\beta_0)\|_{L_2(X)} < \infty$,
*(iii)* $\|\partial^2 G_{0,l}(\phi_l(X,\beta_0))/\partial\phi_{l,i}^2 \nabla\phi_{l,i}(X,\beta_0)_{L_2(X)}\| < \infty$,
*(iv)* $|\partial\phi_{l,i}(X,\beta)/\partial X_j - \partial\phi_{l,i}(X,\beta_0)/\partial X_j| \leq C_1\|\beta - \beta_0\|_E$ *and (v) for the directional derivative* $d(\partial\phi_{l,i}(X,\beta_0)/\partial X_j)/d\beta[\beta - \beta_0] \leq C_2\|\beta - \beta_0\|_E$ *for constants* $C_1, C_2 > 0$.

The assumptions are additional smoothness and regularity conditions, extending to the second derivatives of $G_{0,l}$ and $\phi_l$. Conditions on $\phi_l$ are directly satisfied in the case of linear indices. Then, the following asymptotic normality statement can be formulated.

**Theorem 3.4.** *Under Assumptions 1-6 it holds for the average partial effect functional $f(\theta)$ that*

$$\sqrt{n}(f(\widehat{\theta}) - f(\theta_0)) \to N(0, \sigma^2_{v_f^*})$$

*with*

$$\sigma^2_{v_f^*} = Var\left(\frac{dl(Z, \theta_0)}{d\theta}[v_f^*]\right)$$

*and where $v_f^*$ is the Riesz representer of the operator $\frac{df(\theta_0)}{d\theta}[\theta - \theta_0]$ satisfying*

$$\frac{df(\theta_0)}{d\theta}[\theta - \theta_0] = \langle \theta - \theta_0, v_f^* \rangle,$$

*with $\langle \cdot, \cdot \rangle$ the inner product induced by the Fisher-like norm $\|\cdot\|$ on $\overline{\boldsymbol{V}}$.*

The above Theorem follows from applying general results of Shen (1997), Chen and Shen (1998) and Chen (2007) and moves along the same lines as Theorem 3.3, as long as the functional $f$ obeys certain regularity conditions. For practical applications, a closed form of the Riesz representer $v_f^*$, as is available for the functional $f(\theta) = \beta$, would be helpful, though, does not appear to have been derived in the context of sieve minimum distance estimation of discrete choice models and is left to future research.

## 3.5 Monte Carlo Study

In this section, I assess the finite sample performance of the constrained sieve GLS estimator. The performance is contrasted to the one of a sieve GLS estimator that does not make use of shape constraints. This allows to single out the effect of shape constraints on the finite sample performance. Second, I also compare the estimator to popular parametric and semiparametric benchmark estimators from the literature.

First, we consider the data generating process,

$$d = \mathbb{1}\{X_1 + X_2\beta_1 + X_3\beta_2 + \epsilon \geq 0\},$$

with $X_1 \sim N(0,1)$, $X_2 \sim U[-1,1]$ and $X_3 \sim Poi(2)$, which is censored at the value 5, making it a censored poisson distribution. The error term $\epsilon$ follows a generelized extreme value (GEV) type I- distribution. Thus, the DGP specifies the popular binary Logit model. The parameter values are $\beta_1 = 1.25$ and $\beta_2 = -1.8$.

First, we compare the performance of the shape constrained sieve GLS estimator outlined in section 3.3.4 to an analogous sieve GLS procedure that does not make use of shape constraints. In particular, the constrained sieve GLS makes use of a constrained I-Spline sieve, whereas the latter uses a standard B-Spline sieve without any coefficient constraints.

For the experiments, I consider 1000 Monte Carlo replications and vary sample size and sieve dimension. The I-Spline basis functions are of degree 2 and by the analogy outlined in section 3.3.3, the B-Spline functions are of degree 3. Throughout this section only the identity weighted case is considered.

Table 3.1 shows the results of both estimators for the finite-dimensional parameter vector $\beta$ and the average partial effect, which are both estimable at $\sqrt{n}$-rate. The average partial effect is simple to compute, as by the relation of I- and M-Spline basis functions outlined in section 3.3.2, estimates of the sieve coefficients are sufficient to provide an estimate of the first derivative of the function. For $N = 500$, there are sizeable gains from imposing shape constraints. The MSE of the constrained estimates are more then half the size of the unconstrained estimates. The average bias is an order of magnitude lower. If we increase sample size to 1000, the results of both approaches become more similar. Yet, both in terms of bias and MSE, the shape constrained procedure still outperforms the unconstrained one. We can expect that for smaller sample sizes the gains from leveraging shape constraints are even more sizeable, whereas for larger sample sizes the differences vanish.

In Table 3.2, we assess the performance of structural quantities that hinge on the estimation of the functional component, i.e. the cdf of the logistic distribution. For this we take two coordinates namely $(0, 1, -2)$ and $(-1, 3, -3)$ and estimate the two predictions and partial effects with respect to $X_2$ at both points. The correct values are given in the table.

Since the true values are quite small, both bias and MSE are small in absolute magnitude. For $N = 500$ the shape constrained estimator outperforms the unconstrained one in terms of MSE. This is not the case for the bias in rows 1, 3 and 4, which is supposedly due to the fact that B-Spline estimates can be negative. This may lead to lower average biases for B-Spline estimates but in this context the average bias is not a suitable measure of comparison in rows 1,3 and 4. When focusing on

| N=500 | | I-Spline | | B-Spline | |
|---|---|---|---|---|---|
| | K | Bias | MSE | Bias | MSE |
| $\widehat{\beta}_1$ | 4 | **0.004** | **0.170** | 0.106 | 0.317 |
| | 8 | 0.028 | 0.204 | 0.128 | 0.359 |
| | 12 | 0.052 | 0.228 | 0.146 | 0.426 |
| $\widehat{\beta}_2$ | 4 | **-0.020** | **0.164** | -0.144 | 0.391 |
| | 8 | -0.049 | 0.212 | -0.159 | 0.422 |
| | 12 | -0.091 | 0.262 | -0.227 | 0.610 |
| APE | 4 | **0.000** | **0.001** | -0.077 | 0.006 |
| | 8 | 0.002 | 0.001 | -0.066 | 0.005 |
| | 12 | 0.002 | 0.001 | -0.061 | 0.004 |

| N=1000 | | I-Spline | | B-Spline | |
|---|---|---|---|---|---|
| | K | Bias | MSE | Bias | MSE |
| $\widehat{\beta}_1$ | 4 | -0.005 | **0.079** | 0.030 | 0.101 |
| | 8 | **-0.001** | 0.085 | 0.027 | 0.100 |
| | 12 | 0.019 | 0.093 | 0.035 | 0.116 |
| $\widehat{\beta}_2$ | 4 | 0.011 | **0.074** | -0.030 | 0.103 |
| | 8 | **-0.002** | 0.077 | -0.040 | 0.095 |
| | 12 | -0.025 | 0.088 | -0.044 | 0.113 |
| APE | 4 | **-0.000** | **0.000** | -0.078 | 0.006 |
| | 8 | 0.001 | 0.000 | -0.066 | 0.004 |
| | 12 | 0.001 | 0.000 | -0.061 | 0.004 |

Table 3.1: Shape constrained vs. unconstrained estimation of $\sqrt{n}$-consistent functionals of $\theta$

the MSE, we again see sizeable gains of imposing shape constraints. These gains are most pronounced for quantities close to the boundary such as the second prediction and partial effect. Increasing the sample size to 1000, takes away the differences between the first prediction and the first partial effect. For the second prediction and partial effect, there are still sizeable improvements by imposing shape constraints. So especially for high and low probabilities and partial effects, shape constrained estimation has its merits. Finally, it remains to compare the estimates to a parametric and semiparametric benchmark estimator. The binary Logit is in this case the correct parametric model to estimate $\beta$. An often applied semiparametric estimator for the binary model is the kernel quasi-likelihood estimator of Klein and Spady (1993). Klein and Spady's method uses a first stage kernel regression of the outcome dummy on the linear index. Then $\beta$ is estimated via maximum likelihood with plug-in nonparametric estimates for the conditional probabilities. A bandwidth parameter needs to be specified. Here, we choose a time-intensive cross-validation procedure to specify the bandwidth. The results in Table 3.3 show that the unconstrained sieve GLS estimator performs similar to Klein and Spady's estimator. The shape constrained sieve GLS estimator produces results that are somewhere in between the correctly specified parametric estimator and unconstrained semiparametric estimators. For the parametric components, shape constrained estimation can be viewed as middleground between the optimal parametric maximum likelihood estimator (which may be hard to attain in practice) and typical unconstrained semiparametric estimators.

This insight continues to hold for the case $N = 1000$ presented in Table 3.4. Here, we can also see that the unconstrained sieve GLS also outperforms Klein-Spady, which shows that sieve GLS may generally have favorable finite sample properties in comparison to computationally more involved kernel approaches.

## 3.6 Conclusion

This paper presents a novel estimator for a broad class of semi(non)parametric discrete choice models, where choice probability functions follow a certain multiple-index form and the model results in a set of moment conditions. The estimator imposes shape constraints on infinite-dimensional parameters that arise in discrete choice models, such as boundedness and monotonicity of the choice probability function. Therefore, it incorporates constrained I-Spline and B-Spline basis function into

| N=500 | | I-Spline | | B-Spline | |
|---|---|---|---|---|---|
| | K | Bias | MSE | Bias | MSE |
| Prediction: 8.7% | 4 | 0.017 | **0.001** | 0.024 | 0.002 |
| | 8 | 0.005 | 0.002 | 0.003 | 0.002 |
| | 12 | 0.002 | 0.002 | 0.007 | 0.003 |
| Prediction: 99.97% | 4 | **-0.059** | **0.004** | -0.134 | 0.032 |
| | 8 | -0.063 | 0.005 | -0.112 | 0.028 |
| | 12 | -0.065 | 0.005 | -0.113 | 0.029 |
| Partial Eff.: 9.94% | 4 | -0.031 | **0.002** | -0.024 | 0.003 |
| | 8 | 0.008 | 0.005 | -0.001 | 0.008 |
| | 12 | 0.008 | 0.010 | 0.011 | 0.037 |
| Partial Eff.: 0.036% | 4 | 0.008 | **6.63e-04** | 0.005 | 5.99e-03 |
| | 8 | 0.009 | 6.74e-04 | 0.008 | 4.50e-03 |
| | 12 | 0.009 | 7.34e-04 | 0.007 | 4.00e-03 |

| N=1000 | | I-Spline | | B-Spline | |
|---|---|---|---|---|---|
| | K | Bias | MSE | Bias | MSE |
| Prediction: 8.7% | 4 | 0.022 | 0.001 | 0.025 | 0.001 |
| | 8 | 0.002 | 0.001 | -0.000 | 0.001 |
| | 12 | 0.001 | 0.001 | 0.001 | 0.001 |
| Prediction: 99.97% | 4 | **-0.059** | **0.004** | -0.129 | 0.024 |
| | 8 | -0.064 | 0.005 | -0.100 | 0.018 |
| | 12 | -0.066 | 0.005 | -0.099 | 0.019 |
| Partial Eff.: 9.94% | 4 | -0.031 | 0.002 | -0.027 | 0.002 |
| | 8 | 0.007 | 0.003 | -0.005 | 0.002 |
| | 12 | 0.002 | 0.005 | -0.003 | 0.011 |
| Partial Eff.: 0.036% | 4 | 0.008 | 3.16e-04 | 0.005 | 6.07e-03 |
| | 8 | 0.009 | **3.08e-04** | 0.008 | 4.37e-03 |
| | 12 | 0.009 | 3.26e-04 | 0.008 | 3.87e-03 |

Table 3.2: Shape constrained vs. unconstrained estimation of predictions and partial effects

| N=500, K=4 | I-Spline | | B-Spline | | Logit | | Klein-Spady | |
|---|---|---|---|---|---|---|---|---|
| | Bias | MSE | Bias | MSE | Bias | MSE | Bias | MSE |
| $\widehat{\beta}_1$ | 0.004 | 0.170 | 0.106 | 0.317 | 0.007 | 0.095 | 0.008 | 0.278 |
| $\widehat{\beta}_2$ | -0.020 | 0.164 | -0.144 | 0.391 | -0.025 | 0.025 | -0.089 | 0.981 |

Table 3.3: Sieve GLS vs. Logit and Klein-Spady's Estimator with cross-validated bandwidth

| N=1000, K=4 | I-Spline | | B-Spline | | Logit | | Klein-Spady | |
|---|---|---|---|---|---|---|---|---|
| | Bias | MSE | Bias | MSE | Bias | MSE | Bias | MSE |
| $\widehat{\beta}_1$ | -0.005 | 0.079 | 0.030 | 0.101 | 0.002 | 0.044 | -0.023 | 0.136 |
| $\widehat{\beta}_2$ | -0.011 | 0.074 | -0.030 | 0.103 | -0.014 | 0.011 | -0.011 | 0.481 |

Table 3.4: Shape constrained vs. benchmark estimation, N=1000

the well-established sieve estimation framework. Applying a sieve GLS procedure provides joint estimates of both infinite- and finite-dimensional parameters. Imposing both boundedness and monotonicity constraints results in a faster convergence rate in a weak Fisher-like norm as opposed to the general case without shape constraints. This simplifies existing results on the asymptotic normality of smooth functionals of the parameter set, as rate restrictions in the weak norm are more easily satisfiable under shape constraints. This is a novel insight in how shape constraints aid in the context of semiparametric estimation apart from an effect on finite sample properties of an estimator. Also the finite sample performance of the estimator is illustrated in a series of Monte Carlo experiments. Shape-constraints are shown to be most effective in small samples and for predictions and partial effects in the boundary of the domain of the infinite-dimensional parameter.

## 3.7 Appendix

PROOF OF LEMMA 3.1. We begin the proof with part (i) of the Lemma. Choose some $g \in \mathcal{G}_K^c$ with $g = \sum_{j=1}^K p_j(\phi)\pi_j$. Then $g \geq 0$ follows from $p_j \geq 0$ by the positivity of I-Splines and B-Splines and the constraint $\pi_j \geq 0$ for every $j$, which is imposed on the sieve space. Similarly by positivity of basis functions and coefficients it follows that

$$\sum_{j=1}^K p_j(\phi)\pi_j \leq \sum_{j=1}^K \pi_j$$

by the fact that $p_j \leq 1$ uniformly for I-Spline and B-Spline basis functions. Then, by the constraint $\sum_{j=1}^K \pi_j \leq 1$, it follows that $g \leq 1$. Next, taking the derivative with respect to a monotonic argument of $g$, w.l.o.G. $\phi_m$, we obtain $\sum_{j=1}^K \frac{\partial}{\partial \phi_m} p_j(\phi)\pi_j \geq 0$ by the differentiability and non-decreasing property of I-Spline functions as well as the positivity of the other I- and B-Spline basis functions in the tensor product and the constraints on sieve coefficients. From this, we can conclude that for any $g \in \mathcal{G}_K^c$, we have $g \in \mathcal{G}^c$, see also the discussion in section 3.3.2. The last property of (i) follows immediately from the fact that for any $g \in \mathcal{G}_K^c$ we have $g \in \mathcal{G}_{K+1}^c$ since setting the additional sieve coefficient to zero does not violate the constraints imposed on the sieve coefficients.

Next, we consider part (ii). Define $\Pi_K g$ analogous to the definition of $Ag$ in the proof of Lemma 0.2 in Wu and Zhang (2012), see p. 4 before equation (0.3) in their supplementary material. The proof of Lemma 0.2 can be readily extended from the bivariate to the general multivariate case and can be applied to our setting with

$$\Pi_K g = \sum_{j_1=1}^{K_1} \cdots \sum_{j_1=1}^{K_M} g(\epsilon_{j_1}^{(1)}, \ldots, \epsilon_{j_M}^{(M)}) B_{j_1}^{l+1} \ldots B_{j_M}^{l+1},$$

where $\epsilon_{j_1}^{(1)}, \ldots \epsilon_{j_M}^{(M)}$ are sequences as defined in (0.5) and (0.6) of the proof of Lemma 0.2 for each dimension. The definition needs to be extended to the general multivariate case, i.e. for the first sequence we have

$$\epsilon_{j_1}^{(1)} = \begin{cases} u_1 + \frac{(l-1)(u_{l+1}-u_l)}{l}, & \text{if } j_1 = 1, \ldots, l \\ u_i, & \text{if } j_1 = l+1, \ldots, K_1 \end{cases}$$

where $\{u_{j_1}\}_{j_1=1}^{K_1}$ is the knot sequence of the B-Spline approximating the first dimension of $\Pi_K g$. Thus, the function $g(\epsilon_{j_1}^{(1)}, \ldots \epsilon_{j_M}^{(M)})$ constitutes the sieve coefficients of $\Pi_K g$. As $g \in \mathcal{G}^c$, we have $0 \leq g \leq 1$ resulting in all sieve coefficents of $\Pi_K g$ being

positive and bounded by 1. Further, w.l.o.g let $g$ be monotonic in its first argument, then we have $g(\epsilon^{(1)}_{j_1+1}, \ldots, \epsilon^{(M)}_{j_M}) \geq g(\epsilon^{(1)}_{j_1}, \ldots, \epsilon^{(M)}_{j_M})$ for any $j_1$ by monotonicity and the fact that the sequence $\{\epsilon^{(1)}_{j_1}\}_{j_1=1}^{K_1}$ is increasing. Therefore, we can conclude that $\Pi_K g \in \widetilde{\mathcal{G}}^c_K$, which is the constrained B-Spline sieve defined on p. 10 in this paper, which is equivalent to the space $\mathcal{G}^c_K$.

Applying Lemma 0.2. then yields

$$\|g - \Pi_K g\|_\infty \leq C|T|^\eta \sup_{[\alpha]\leq\eta} \sup_{\phi\in\Phi} |D^\alpha G(\phi)|$$

for some constant $C > 0$ and $|T|$ as defined in Lemma 0.2. Assumption 2 (ii) and the fact that $|T| = O(K^{-1/d})$ is sufficient to establish part (ii).

Finally, it remains to consider part (iii), for which we follow the tensor product definition of Chen (2007) pp. 5573. The space $\mathcal{G}^c_K$ is a tensor product of I-Spline sieves with monotonicity and positivity and boundedness constraints and B-Splines with positivity and boundedness constraints. In particular $\mathcal{G}^c_K = \mathcal{I}^c_{K_m} \otimes \mathcal{B}^c_{K_{(\neg m)}}$ and thus,

$$N_{[\,]}(\epsilon, \mathcal{G}^c_K, L^r(\Phi)) \leq N_{[\,]}(\epsilon/2, \mathcal{I}^c_{K_m}, L^r(\Phi)) \cdot N_{[\,]}(\epsilon/2, \mathcal{B}^c_{K_{(\neg m)}}, L^r(\Phi))$$

following Lemma 9.25 (ii) in Kosorok (2008). This Lemma can be applied since any function in the respective tensor sieve spaces is bounded by 1. The constrained I-Spline sieve space is a subset of the monotone function space outlined in Theorem 2.7.5 of van der Vaart and Wellner (2000) and satisfies for some constant $C$

$$\log N_{[\,]}(\epsilon, \mathcal{I}^c_{K_m}, L^r(\Phi)) \leq C(1/\epsilon).$$

This yields

$$\begin{aligned}\log N_{[\,]}(\epsilon, \mathcal{G}^c_K, L^r(\Phi)) &\leq 2C/\epsilon + \log N_{[\,]}(\epsilon/2, \mathcal{B}^c_{K_{(\neg m)}}, L^r(\Phi)) \\ &\leq 2C/\epsilon + C_0 \cdot K_{(\neg m)} \log(2/\epsilon) = O(K_{(\neg m)})\end{aligned}$$

for some constants $C, C_0 > 0$ where the bound for the bracketing number of $\mathcal{B}^c_{K_{(\neg m)}}$ follows e.g. from Chen (2007) p.5595. □

Proof of Theorem 3.1. For consistency, it suffices to check the conditions of the general consistency result in Lemma A.2 in Chen and Pouzo (2012). In our setting, we have $\overline{Q}_n(\theta) = \mathbf{E}[\rho(Z,\theta)'\Sigma(X)^{-1}\rho(Z,\theta)]$. There is no dependency on $n$ in our case

as I do not incorporate additional penalties in the model setup, yet we stick to the notation of Chen and Pouzo (2012). It holds that

$$\begin{aligned}
&\overline{Q}_n(\theta) - \overline{Q}_n(\theta_0) \\
=& \mathbf{E}[\rho(Z,\theta)'\Sigma(X)^{-1}\rho(Z,\theta) - \rho(Z,\theta_0)'\Sigma(X)^{-1}\rho(Z,\theta_0)] \\
=& \mathbf{E}[(\rho(Z,\theta) - \rho(Z,\theta_0))'\Sigma(X)^{-1}\rho(Z,\theta) \\
&+ \rho(Z,\theta_0)'\Sigma(X)^{-1}(\rho(Z,\theta) - \rho(Z,\theta_0))] \\
=& \mathbf{E}[(\rho(Z,\theta) - \rho(Z,\theta_0))'\Sigma(X)^{-1}\rho(Z,\theta)] \\
=& \mathbf{E}[[G_0(\phi(X,\beta_0)) - G(\phi(X,\beta))]'\Sigma(X)^{-1}[G_0(\phi(X,\beta_0)) - G(\phi(X,\beta))]]
\end{aligned}$$

where $G_0$ and $\phi$ are quantities such that the l-th element of $G_0(\phi(X,\beta_0)) - G(\phi(X,\beta))$ corresponds to $G_{0,l}(\phi_l(X,\beta_0)) - G_l(\phi_l(X,\beta))$. The third equality follows from the identifying condition $\mathbf{E}[\rho(Z,\theta_0)|X] = 0$ and the fourth equality from the identity $\rho(Z,\theta) = \rho(Z,\theta_0) + G_0(\phi(X,\beta_0)) - G(\phi(X,\beta))$ along with the identifiying condition. As the lowest eigenvalue of $\Sigma(X)$ is bounded away from zero by Assumption 3, (ii) we have for some constant $c > 0$

$$\overline{Q}_n(\theta) - \overline{Q}_n(\theta_0) \geq c \cdot \|G_0(\phi(X,\beta_0)) - G(\phi(X,\beta))\|^2_{L^2(X)}.$$

Define

$$g(K,n,\epsilon) := \inf_{\theta \in \Theta_K^c : \|\theta - \theta_0\| \geq \epsilon} \|G_0(\phi(X,\beta_0)) - G(\phi(X,\beta))\|^2_{L^2(X)}.$$

Then, let $(G_K^\star, \beta_K^\star)$ be the argument such that

$$g(K,n,\epsilon) = \|G_0(\phi(X,\beta_0)) - G_K^\star(\phi(X,\beta_K^\star))\|^2_{L^2(X)}$$

which is guaranteed to exist by the compactness of the finite dimensional sieve space $\Theta_K^c$. Let $\Pi_K\theta_0 = (\Pi_K G, \Pi_K\beta_0)$ as in Lemma 3.1(ii). Hence, it holds that

$$\begin{aligned}
g(K,n,\epsilon) =& \|G_K^\star(\phi(X,\beta_K^\star)) - \Pi_K G(\phi(X,\Pi_K\beta_0))\|^2_{L^2(X)} \\
&+ \|\Pi_K G(\phi(X,\Pi_K\beta_0)) - G_0(\phi(X,\beta_0))\|^2_{L^2(X)} \\
&+ 2\|G_K^\star(\phi(X,\beta_K^\star)) - \Pi_K G(\phi(X,\Pi_K\beta_0))\|^2_{L^2(X)} \\
&\cdot \|\Pi_K G(\phi(X,\Pi_K\beta_0)) - G_0(\phi(X,\beta_0))\|^2_{L^2(X)}
\end{aligned}$$

By the approximation properties in Lemma 3.1 (ii) and the rate in 3 (iv), we have $\|\Pi_K G(\phi(X,\Pi_K\beta_0)) - G_0(\phi(X,\beta_0))\|_{L^2(X)} = o(1)$ and as $\Pi_K\theta, (G_K^*, \beta_K^*) \in \Theta_K^c$, it also

holds that $\|G_K^\star(\phi(X,\beta_K^\star)) - \Pi_K G(\phi(X,\Pi_K\beta_0))\|^2_{L^2(X)} \leq 1$. Further, there exists a constant $c(\epsilon)$ such that $\|G_K^\star(\phi(X,\beta_K^\star)) - \Pi_K G(\phi(X,\Pi_K\beta_0))\|_{L^2(X)} \geq c(\epsilon) > 0$. This follows from the same reasoning as the NPIV example following Theorem 3.1. in Chen and Pouzo (2012). Summarizing,

$$\liminf_{n\to\infty} g(K,n,\epsilon) = c(\epsilon)^2,$$

which satisfies Condition a of Lemma A.2. in Chen and Pouzo (2012).

Condition b is satisfied by the sieve approximation properties in Lemma 3.1 (ii) and continuity of $\rho$. Condition c is trivially satisfied in this model setup. For Condition d we require that

$$\widehat{c}_n \equiv \sup_{\theta\in\Theta_K^c} \frac{1}{n}\sum_{i=1}^n \underbrace{\rho(d,X_i,\theta)'\Sigma(X)\rho(d,X_i,\theta) - \mathbf{E}[\rho(d,X,\theta)'\Sigma(X)\rho(d,X,\theta)]}_{:=f_\rho(Z_i,\theta)} = o_p(1).$$

Define $\mathcal{F}_{\rho,K} := \{f_\rho(\cdot,\theta) : \theta \in \Theta_K^c\}$ with envelope function $\overline{F}_\rho$. A necessary and sufficient condition to obtain uniform convergence over sieves is Condition 3.5 MD (i), (ii) in Chen (2007) with (i) having been already discussed. By Assumption 3 (ii) and the fact that each element of $\rho(Z,\theta)$ is uniformly bounded by 1 for all $\theta \in \Theta_K^c$ it follows for any $\theta_1,\theta_2 \in \Theta_K^c$ that

$$\begin{aligned}
&|f_\rho(Z,\theta_2) - f_\rho(Z,\theta_1)| \\
\leq&|(\rho(Z,\theta_2) - \rho(Z,\theta_1))'\Sigma(X)\rho(Z,\theta_2) + \rho(Z,\theta_1)'\Sigma(X)(\rho(Z,\theta_2) - \rho(Z,\theta_1))| \\
&+ \mathbf{E}[|(\rho(Z,\theta_2) - \rho(Z,\theta_1))'\Sigma(X)\rho(Z,\theta_2) + \rho(Z,\theta_1)'\Sigma(X)(\rho(Z,\theta_2) - \rho(Z,\theta_1))|] \\
\leq& c_1(\|\rho(Z,\theta_2) - \rho(Z,\theta_1)\|_E + \mathbf{E}[\|\rho(Z,\theta_2) - \rho(Z,\theta_1)\|_E])
\end{aligned}$$

for some constant $c_1 > 0$. For any element in the above vector we have $\rho_l(Z,\theta_2) - \rho_l(Z,\theta_1) = G_{1,l}(\phi_l(X,\beta_1)) - G_{2,l}(\phi_l(X,\beta_2))$ and thus, by way of the multivariate Taylors Theorem for any $\delta > 0$ and $\|\theta_2 - \theta_1\|_c \leq \delta$,

$$\begin{aligned}
|\rho_l(Z,\theta_2) - \rho_l(Z,\theta_1)| \leq& \sup_{x\in\mathcal{X}}\sup_{\beta\in B}|G_1(\phi_l(X,\beta)) - G_2(\phi_l(X,\beta))| \\
&+ \sup_{x\in\mathcal{X}}\sup_{\beta\in B}\|\nabla G_{2,l}(\phi(X,\beta))\|_E \cdot \|(\phi_l(X,\beta_2) - \phi_l(X,\beta_1))\|_E
\end{aligned}$$

and by Assumption 3 (iii), it holds that

$$\sup_{\theta_1,\theta_2\in\Theta_K^c:\|\theta_1-\theta_2\|_c\leq\delta} |\rho_l(Z,\theta_2) - \rho_l(Z,\theta_1)| \leq C_1(X)\|\theta_2 - \theta_1\|_c$$

for some $C_1(X)$ with $\mathbf{E}[C_1(X)] < \infty$ and ultimately

$$\sup_{\theta_1,\theta_2 \in \Theta_K^c : \|\theta_1 - \theta_2\|_c \leq \delta} |f_\rho(Z,\theta_2) - f_\rho(Z,\theta_1)| \leq C_2(X)\|\theta_2 - \theta_1\|_c$$

for some $C_2(X)$ with $\mathbf{E}[C_1(X)] < \infty$. Therefore Condition 3.5MD (ii) is satisfied and it remains to check Condition 3.5MD (iii) $\log N(\delta^{1/s}, \theta_K^c, \|\cdot\|_c) = o(n)$ which holds for standard linear sieves such as B-Splines, whenever $K = o(n)$, see Chen (2007) p.5595. This corresponds to our Assumption 3 (iv).

Then following Lemma A.2. of Chen and Pouzo (2012) consistency follows whenever

$$\max\{\widehat{c}_n, \|\Pi_K \theta - \theta_0\|_\infty\} = o(1),$$

which is the case under the rates in Assumption 3 (iv) and the results from Lemma 3.1 (ii). □

Proof of Theorem 3.2. The proof follows from applying the general convergence rate result for sieve M-estimators in Theorem 3.2 of Chen (2007). Some assumptions of Theorem 3.2 are implicitly mentioned in the beginning of the exposition on p. 5594 and need to be verified first. It is required that the estimator is consistent in some strong metric, which is in our case $\|\widehat{\theta} - \theta_0\|_c = o(1)$ (following from our Theorem 3.1) and that the metric in which the convergence rate is to be derived is weaker compared to the consistency metric, i.e. $\|\theta - \theta_0\| \lesssim \|\widehat{\theta} - \theta_0\|_c$[1]. in our case. Further, the convergence rate metric needs to be locally equivalent to the criterion function in the sense that $\|\theta - \theta_0\| \asymp \mathbf{E}[\rho(Z,\theta)'\Sigma(X)\rho(Z,\theta) - \rho(Z,\theta_0)'\Sigma(X)\rho(Z,\theta_0)]^{1/2}$ for $\theta \in \Theta^c$, which satisfy $\|\theta - \theta_0\|_c = o(1)$.

We begin the proof by showing the latter property. First, note that

$$\begin{aligned}
\mathbf{E}[l(\theta) - l(\theta_0)] &= \mathbf{E}[\rho(Z,\theta)'\Sigma(X)\rho(Z,\theta) - \rho(Z,\theta_0)'\Sigma(X)\rho(Z,\theta_0)] \\
&= \mathbf{E}[(\rho(Z,\theta) - \rho(Z,\theta_0))'\Sigma(X)(\rho(Z,\theta) - \rho(Z,\theta_0))] \\
&\quad + 2\,\mathbf{E}[(\rho(Z,\theta) - \rho(Z,\theta_0))\Sigma(X)\rho(Z,\theta_0)]
\end{aligned}$$

by the law of iterated expectation and the identifying condition in Assumption 1. Further, by Assumption 3(ii) and Assumption 1 the second term in the summation

[1] If $\{a_n\}$ and $\{b_n\}$ are sequences of positive numbers, we use the notation $a_n \lesssim b_n$ if $\limsup_{n\to\infty} a_n/b_n < \infty$ and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$

is zero and thus

$$\mathbf{E}[l(\theta) - l(\theta_0)] = \mathbf{E}[(\rho(Z,\theta) - \rho(Z,\theta_0))'\Sigma(X)(\rho(Z,\theta) - \rho(Z,\theta_0))]$$

and again by Assumption 3(ii)

$$\mathbf{E}[l(\theta) - l(\theta_0)]^{1/2} \asymp \|m(X,\theta)\|_{L^2(X)}$$

Hence, it suffices to establish that $\|m(X,\theta)\|_{L^2(X)} \asymp \|\theta - \theta_0\|$ for $\|\theta - \theta_0\|_c = o(1)$, which we now consider.
Therefore, the directional derivatives defined in Chapter 3.4.2 can be detailed to

$$\frac{dm(X,\theta_0)}{d\theta}[\theta_1 - \theta_0] = \left(\frac{dm_1(X,\theta_0)}{d\theta}[\theta_1 - \theta_0], \ldots, \frac{dm_{L-1}(X,\theta_0)}{d\theta}[\theta_1 - \theta_0]\right),$$

where the $l$-th element of the vector is of the form

$$\begin{aligned}&\frac{dm_l(X,\theta_0)}{d\theta}[\theta_1 - \theta_0]\\ =& -\nabla G_{0,l}(\phi_l(X,\beta_0))' J_{\phi_l}(X,\beta_0)(\beta - \beta_0) + G_{0,l}(\phi(X,\beta_0) - G_{0,l}(\phi(X,\beta)).\end{aligned}$$

Further, for each $l$ we have for the approximation error

$$\begin{aligned}&\|m_l(X,\theta) - \frac{dm_l(X,\theta_0)}{d\theta}[\theta - \theta_0]\|_{L_2(X)}\\ =&\|G_{0,l}(\phi_l(X,\beta)) - G_l(\phi_l(X,\beta)) + \nabla G_{0l}(\phi_l(X,\beta_0))' J_{\phi_l}(X,\beta_0)(\beta - \beta_0)\|_{L_2(X)}\\ \leq&\|\nabla G_{0,l}(\phi_l(X,\beta_0))' J_{\phi_l}(X,\beta_0)(\beta - \beta_0)\|_{L_2(X)} + \|G_{0,l}(\phi_l(X,\beta_0)) - G_l(\phi_l(X,\beta_0))\|_{L_2(X)}\\ &+ \|G_{0,l}(\phi_l(X,\beta)) - G_{0,l}(\phi_l(X,\beta_0))\|_{L_2(X)} + \|G_l(\phi_l(X,\beta_0)) - G_l(\phi_l(X,\beta))\|_{L_2(X)}\\ \lesssim&\|\beta - \beta_0\|_E + \|G_l - G_{0,l}\|\end{aligned}$$

by the fact that in our model $\|G_l - G_{0,l}\| = \|G_{0,l}(\phi_l(X,\beta_0)) - G_l(\phi_l(X,\beta_0))\|_{L_2(X)}$ and the other bounds following from Assumptions 2 (ii), 3 (iii) and 4(i). Under Assumption 4 (ii) we can invoke Lemma B.1 in Ai and Chen (2003) and obtain that both $\|\beta - \beta_0\|_E = O(\|\theta - \theta_0\|)$ and $\|G_l - G_{0l}\| = O(\|\theta - \theta_0\|)$ and thus for all $\theta \in \Theta_K^c$

$$\|m_l(X,\theta)\|_{L^2(X)} \lesssim \|\theta - \theta_0\|.$$

Further, by the same trick as in the proof of Proposition 3.2 in Ai and Chen (2003)

it holds for some constants $C, C_1 > 0$ that

$$\begin{aligned}\|\theta - \theta_0\| \leq & \|m(X,\theta)\|_{L_2(X)} + C(\|\beta - \beta_0\|_E + \|G_l - G_{0l}\|) \\ \leq & \|m(X,\theta)\|_{L_2(X)} + C_1(\|m(X,(G_0,\beta))\|_{L_2(X)} + \|m(X,(G,\beta_0))\|_{L_2(X)}) \\ \lesssim & \|m(X,\theta)\|_{L_2(X)}\end{aligned}$$

for any $\theta \in \Theta_K^c$ with $\|\theta - \theta_0\|_c = o(1)$, where the second inequality follows from Assumption 4(iii). This yields the claimed local equivalence of $\|\theta - \theta_0\|$ and $\mathbf{E}[l(\theta) - l(\theta_0)]^{1/2}$ for $\theta \in \Theta_K^c$ with $\|\theta - \theta_0\|_c = o(1)$.

Now we can proceed by checking the remaining conditions for Theorem 3.2 of Chen (2007). Condition 3.6 is our Assumption 3 (i). For Condition 3.7 we need that for some constant $C > 0$

$$\sup_{\theta \in \Theta_K^c : \|\theta - \theta_0\| \leq \delta} \mathbf{E}[(\rho(Z,\theta)'\Sigma(X)\rho(Z,\theta) - \rho(Z,\theta_0)'\Sigma(X)\rho(Z,\theta_0))^2] \leq C\delta^2.$$

In our case, we find that

$$\begin{aligned}& \mathbf{E}[(\rho(Z,\theta)'\Sigma(X)\rho(Z,\theta) - \rho(Z,\theta_0)'\Sigma(X)\rho(Z,\theta_0))^2] \\ = & \mathbf{E}[m(X,\theta)\Sigma(X)^{-1}\rho(Z,\theta)^2 + m(X,\theta)\Sigma(X)^{-1}\rho(Z,\theta_0)^2 \\ & + m(X,\theta)'\Sigma(X)^{-1}\rho(Z,\theta)\rho(Z,\theta)'\Sigma(X)^{-1}m(X,\theta)] \\ \leq & C\,\mathbf{E}[\|m(X,\theta)\|_E^2] \\ = & C\|m(X,\theta)\|_{L^2(X)}^2 \lesssim \|\theta - \theta_0\|^2\end{aligned}$$

with the first inequality holding for some constant $C > 0$ by Assumption 3 (ii) and the fact that $\rho(Z,\theta_0)$ and $\rho(Z,\theta)$ are uniformly bounded in $Z$ as well as for $\theta \in \Theta_K^c$. The last inequality holds by the local equivalence of $\|\theta - \theta_0\|$ and $\|m(X,\theta)\|_{L^2(X)}$.

It remains to check Condition 3.8, which states that for each $\delta > 0$ there exists a random variable $U(X)$ and some $s \in (0,2)$ such that

$$\sup_{\theta \in \Theta_K^c : \|\theta - \theta_0\| \leq \delta} \|m(X,\theta)\|_E \leq \delta^s U(X)$$

with $\mathbf{E}[U(X)^\gamma] < \infty$ for some $\gamma \geq 2$.

By the same reasoning as earlier and from Assumption 4 (ii) there exists a $C(X)$

with $\mathbf{E}[C(X)] < \infty$ such that

$$\begin{aligned}
&|m_l(X,\theta)| \\
\leq&|\nabla G_{0,l}(\phi_l(X,\beta_0))' J_{\phi_l}(X,\beta_0)(\beta-\beta_0)| + |G_{0,l}(\phi_l(X,\beta_0)) - G_l(\phi_l(X,\beta_0))| \\
&+ |G_{0,l}(\phi_l(X,\beta)) - G_{0,l}(\phi_l(X,\beta_0))| + |G_l(\phi_l(X,\beta_0)) - G_l(\phi_l(X,\beta))| \\
\leq&C(X)\|\beta-\beta_0\|_E + \sup_{x\in\mathcal{X}} |G_{0,l}(X,\beta_0) - G_{0,l}(X,\beta_0)|.
\end{aligned}$$

By Lemma 2 in Chen and Shen (1998) and Assumption 2 (ii) we have that

$$\begin{aligned}
&\sup_{x\in\mathcal{X}} |G_{0,l}(X,\beta_0) - G_{0,l}(X,\beta_0)| \\
\lesssim&\|G_{0,l}(X,\beta_0) - G_{0,l}(X,\beta_0)\|_{L^2(X)}^{2p/2p+1} \\
=&\|G_{0,l} - G_l\|^{2p/2p+1} \lesssim \|\theta-\theta_0\|^{2p/2p+1},
\end{aligned}$$

with the last inequality again following from Lemma B.1 in Ai and Chen (2003). Ultimately,

$$\begin{aligned}
\sup_{\theta\in\Theta_K^c:\|\theta-\theta_0\|\leq\delta} \|m(X,\theta)\|_E &\leq \sup_{\theta\in\Theta_K^c:\|\theta-\theta_0\|\leq\delta} \sqrt{\sum_{l=1}^{L-1} m_l(X,\theta)^2} \\
&\leq \sup_{\theta\in\Theta_K^c:\|\theta-\theta_0\|\leq\delta} \sqrt{\sum_{l=1}^{L-1} C(X)\|\theta-\theta_0\|^2}
\end{aligned}$$

which establishes the claim of Condition 3.8. As all necessary conditions are satisfied, we can apply Theorem 3.2 of Chen (2007) and obtain

$$\|\widehat{\theta} - \theta_0\| = O_p(\max\{\delta_n, \|\theta_0 - \Pi_K\theta_0\|\}),$$

where $\delta_n$ is the bracketing integral for the function space $\mathcal{F}_n = \{l(Z,\theta) - l(Z,\theta_0) : \|\theta-\theta_0\| \leq \delta, \theta \in \Theta_K^c\}$ and defined as follows

$$\delta_n = \inf\{\delta \in (0,1) : 1/\sqrt{(}n)\delta^2 \int_{b\delta^2}^{\delta} \sqrt{\log N_{[]}(w, \mathcal{F}_n, \|\cdot\|_2)}dw \leq const.\}.$$

By Chen (2007) p. 5595, we have $\log N_{[]}(w, \mathcal{F}_n, \|\cdot\|_2) \leq \log N(w^{1/s}, \Theta_K^c, \|\cdot\|) \leq N_{[]}(w^{1/s}, \Theta_K^c, \|\cdot\|)$ with the last inequality from e.g. Kosorok (2008) Lemma 9.18. Then, by Lemma 3.1 (iii) we have

$$\delta_n \lesssim \sqrt{K_{(\neg m)}}/\sqrt{n},$$

which then establishes the result stated in our Theorem 3.2. □

Proof of Theorem 3.3. In order to proof Theorem 3.3 we need to check if Assumptions 4.1 and 4.2 in Chen (2007) are satisfied. Assumption 4.1(i) is our Assumption 5 (i) and 4.1(ii) is our 4(ii), 4.1(iii) is our 5(iii). Assumption 4.2 (i) holds by Assumption 3(ii) and 5 (iv). 4.2(ii) is our 5(v). Assumption 4 (iv) is satisfied by our Assumption of i.i.d data, the model setup and the identification condition in Assumption 1.

For checking Condition 4.2' in Chen (2007) a more detailed assessment is necessary. The condition requires that for $\delta_n = o(1)$

$$\begin{aligned}
&\sup_{\bar{\theta}\in\Theta_K^c:\|\bar{\theta}-\theta_0\|\leq\delta_n} \mu_n\Big(\rho(Z,\bar{\theta})\Sigma(X)\frac{d\rho(Z,\bar{\theta})}{d\theta}[\Pi_K v^*] \\
&\quad - \rho(Z,\theta_0)\Sigma(X)\frac{d\rho(Z,\theta_0)}{d\theta}[\Pi_K v^*]\Big) \\
=&o_p(n^{-1/2})
\end{aligned}$$

In our case we have

$$\begin{aligned}
&\sup_{\bar{\theta}\in\Theta_K^c:\|\bar{\theta}-\theta_0\|\leq\delta_n} \mu_n\Big(\rho(Z,\bar{\theta})\Sigma(X)\frac{d\rho(Z,\bar{\theta})}{d\theta}[\Pi_K v^*] \\
&\quad - \rho(Z,\theta_0)\Sigma(X)\frac{d\rho(Z,\theta_0)}{d\theta}[\Pi_K v^*]\Big) \\
=&\sup_{\bar{\theta}\in\Theta_K^c:\|\bar{\theta}-\theta_0\|\leq\delta_n} \mu_n\Big(\rho(Z,\bar{\theta})\Sigma(X)\Pi_K v^*(\phi(X,\bar{\beta})) \\
&\quad - \rho(Z,\theta_0)\Sigma(X)\Pi_K v^*(\phi(X,\beta_0))\Big) \\
=&\sup_{\bar{\theta}\in\Theta_K^c:\|\bar{\theta}-\theta_0\|\leq\delta_n} \mu_n\left(\rho(Z,\bar{\theta})\Sigma(X)(\Pi_K v^*(\phi(X,\bar{\beta})) - \Pi_K v^*(\phi(X,\beta_0)))\right) + \\
&\mu_n\left(m(X,\bar{\theta}))\Sigma(X)\Pi_K v^*(\phi(X,\beta_0))\right)
\end{aligned}$$

and we have already provided arguments that both functions indexing the empirical process are Lipschitz with respect to $\|\cdot\|$. This follows for the first summand by Assumption 3(iii) and the boundedness of $\rho$ and $\Sigma$ and for the second summand analogous to the verification of Condition 3.8 in the proof of Theorem 3.2, where we established $m(X,\bar{\theta}) \lesssim \|\theta - \theta_0\|$. Using this Lipschitz property, we can invoke e.g. the last display of Theorem 2.14.2 in van der Vaart and Wellner (2000) and obtain that both empirical process are of order $O_p(\delta_n/n^{-1/2})$, which establishes the claim of Condition 4.2' as $\delta_n = o(1)$. Finally, we need to check for Condition 4.3', which

reads as

$$\begin{aligned}&\mathbf{E}\left[\rho(Z,\widehat{\theta})\Sigma(X)\frac{d\rho(Z,\widehat{\theta})}{d\theta}[\Pi_K v^*]\right]\\=&\mathbf{E}\left[\frac{d\rho(Z,\theta_0)}{d\theta}[\Pi_K v^*]\Sigma(X)\frac{d\rho(Z,\theta_0)}{d\theta}[\widehat{\theta}-\theta]\right]+o(n^{-1/2}).\end{aligned}$$

For the difference, we have

$$\begin{aligned}&\mathbf{E}\left[\rho(Z,\widehat{\theta})\Sigma(X)\frac{d\rho(Z,\widehat{\theta})}{d\theta}[\Pi_K v^*]\right]-\mathbf{E}\left[\frac{d\rho(Z,\theta_0)}{d\theta}[\Pi_K v^*]\Sigma(X)\frac{d\rho(Z,\theta_0)}{d\theta}[\widehat{\theta}-\theta]\right]\\=&I+II,\end{aligned}$$

where,

$$\begin{aligned}I&=\mathbf{E}\left[(m(X,\widehat{\theta})-\frac{dm(X,\theta_0)}{d\theta}[\widehat{\theta}-\theta_0])'\Sigma(X)\frac{dm(X,\widehat{\theta})}{d\theta}[\Pi_K v^*]\right],\\II&=\mathbf{E}\left[\frac{dm(X,\theta_0)}{d\theta}[\widehat{\theta}-\theta_0]'\Sigma(X)(\frac{dm(X,\widehat{\theta})}{d\theta}[\Pi_K v^*]-\frac{dm(X,\theta_0)}{d\theta}[\Pi_K v^*])\right].\end{aligned}$$

First consider I. By Cauchy-Schwarz and the Assumptions on $\Sigma(X)$,

$$|I|\leq\|(m(X,\widehat{\theta})-\frac{dm(X,\theta_0)}{d\theta}[\widehat{\theta}-\theta_0]\|_E\cdot\|\frac{dm(X,\widehat{\theta})}{d\theta}[\Pi_K v^*]\|_E,$$

where the first factor has for each element $l$

$$\begin{aligned}&m_l(X,\widehat{\theta})-\frac{dm_l(X,\theta_0)}{d\theta}[\widehat{\theta}-\theta_0]\\=&\widehat{G}_l(\phi_l(X,\beta_0))-\widehat{G}_l(\phi_l(X,\widehat{\beta}))+\nabla G_{0,l}(\phi_l(X,\beta_0))J_{\phi_l}(X,\beta_0)(\widehat{\beta}-\beta_0)\\=&O(\|\widehat{\beta}-\beta_0\|_E)\end{aligned}$$

by multivariate Taylors Theorem and Assumptions 4 (iii), 5 (iii), (iv) and the already frequently invoked Lemma B.1 of Ai and Chen (2003). This yields

$$\|(m(X,\widehat{\theta})-\frac{dm(X,\theta_0)}{d\theta}[\widehat{\theta}-\theta_0]\|_E=O(\|\widehat{\theta}-\theta_0\|^2).$$

For the second factor, we have $\|\frac{dm(X,\widehat{\theta})}{d\theta}[\Pi_K v^*]\|_E=O_p(1)$ by the equality $\frac{dm(X,\widehat{\theta})}{d\theta}[\Pi_K v^*]=\Pi_K v^*(\phi_l(X,\widehat{\beta}))$. This yields $|I|=o(n^{1/2})$ by the rate restriction in Assumption 5

(vi). Lastly, consider II. Again,

$$|II| \lesssim \|\widehat{\theta} - \theta_0\| \cdot \| \frac{dm(X, \widehat{\theta})}{d\theta}[\Pi_K v^*] - \frac{dm(X, \theta_0)}{d\theta}[\Pi_K v^*] \|_E$$

and it remains to analyze the behavior of the second factor. Here, for the $l$-th element,

$$\begin{aligned} &\frac{dm_l(X, \widehat{\theta})}{d\theta}[\Pi_K v^*] - \frac{dm_l(X, \theta_0)}{d\theta}[\Pi_K v^*] \\ =&\Pi_K v^*(\phi_l(X, \widehat{\beta})) - \Pi_K v^*(\phi_l(X, \beta_0)). \end{aligned}$$

Thus, as $\Pi_K v^* \in \Theta_K^c$ and by Assumption 3 (iii), the term is bounded by $\|\beta - \beta_0\|$ and therefore

$$|II| \lesssim \|\widehat{\theta} - \theta_0\|^2,$$

which results in the claim of Condition 4.3' under the rate restriction in Assumption 5 (vi). □

Proof of Theorem 3.4. Again it suffices to check whether the conditions are satisfied to invoke a general result for asymptotic normality of smooth functionals of $\theta$ such as Theorem 4.3 in Chen (2007). It suffices to check if the functional, in this case $f(\theta) = \mathbf{E}[dG_l(\phi_l(X, \beta))/dX_j]$, satisfies Condition 4.1 (i) and (ii) in Chen (2007). The remaining conditions have already been checked in the Proof of Theorem 3.3 for the context of sieve GLS estimation, as the results do not hinge on the particular form of the Riesz representer $v^*$.

In the case of the average partial effect functional, we have

$$\begin{aligned} f(\theta_0) =& \mathbf{E}[\nabla G_{0,l}(\phi_l(X, \beta_0)) \partial \phi_l(X, \beta_0)/\partial X_j] \\ =& \mathbf{E}\left[\sum_{i=1}^{M} \frac{\partial G_{0,l}(\phi_l(X, \beta_0))}{\partial \phi_{l,i}} \frac{\partial \phi_{l,i}(X, \beta_0)}{\partial X_j}\right] \end{aligned}$$

and for the directional derivative

$$\begin{aligned}\frac{df(\theta_0)}{d\theta}[\theta-\theta_0] = &\mathbf{E}\left[\sum_{i=1}^{M}\left(\frac{\partial^2 G_{0,l}(\phi_l(X,\beta_0))}{\partial\phi_{l,i}^2}\nabla\phi_{l,i}(X,\beta_0)'(\beta-\beta_0)\right.\right.\\ &\left.\left.+\frac{\partial G_l(\phi_l(X,\beta_0))}{\partial\phi_{l,i}}-\frac{\partial G_{0,l}(\phi_l(X,\beta_0))}{\partial\phi_{l,i}}\right)\frac{\partial\phi_{l,i}(X,\beta_0)}{\partial X_j}\right]\\ &+\mathbf{E}\left[\sum_{i=1}^{M}\frac{\partial G_{0,l}(\phi_l(X,\beta_0))}{\partial\phi_{l,i}}d(\frac{\partial\phi_{l,i}(X,\beta_0)}{\partial X_j})/d\beta[\beta-\beta_0]\right].\end{aligned}$$

Condition 4.1 (i) requires that there is some $w>0$ such that for any $\theta\in\Theta^c$ with $\|\theta-\theta_0\|=o(1)$, we have

$$|f(\theta)-f(\theta_0)-\frac{df(\theta_0)}{d\theta}[\theta-\theta_0]| = O(\|\theta-\theta_0\|^w).$$

In our setting,

$$\begin{aligned}&|f(\theta)-f(\theta_0)-\frac{df(\theta_0)}{d\theta}[\theta-\theta_0]|\\ =&\left|\mathbf{E}\left[\sum_{i=1}^{M}(\frac{\partial G_l(\phi_l(X,\beta))}{\partial\phi_{l,i}}-\frac{\partial G_l(\phi_l(X,\beta_0))}{\partial\phi_{l,i}})\frac{\partial\phi_{l,i}(X,\beta_0)}{\partial X_j}\right.\right.\\ &+\frac{\partial G_l(\phi_l(X,\beta))}{\partial\phi_{l,i}}(\frac{\partial\phi_{l,i}(X,\beta)}{\partial X_j}-\frac{\partial\phi_{l,i}(X,\beta_0)}{\partial X_j})\Big]\\ &-\mathbf{E}\left[\sum_{i=1}^{M}\frac{\partial^2 G_{0,l}(\phi_l(X,\beta_0))}{\partial\phi_{l,i}^2}\nabla\phi_{l,i}(X,\beta_0)'(\beta-\beta_0)\frac{\partial\phi_{l,i}(X,\beta_0)}{\partial X_j}\right]\\ &\left.-\mathbf{E}\left[\sum_{i=1}^{M}\frac{\partial G_{0,l}(\phi_l(X,\beta_0))}{\partial\phi_{l,i}}d(\frac{\partial\phi_{l,i}(X,\beta_0)}{\partial X_j})/d\beta[\beta-\beta_0]\right]\right|\\ \lesssim&\|\beta-\beta_0\|_E\lesssim\|\theta-\theta_0\|\end{aligned}$$

by the additional conditions in Assumption 6.

Part (ii) of Condition 4.1 assumes that $\|\frac{df(\theta_0)}{d\theta}[\theta-\theta_0]\|<\infty$ for $\theta\in\Theta^c$ and in our case, by the form of the directional derivative of $f(\cdot)$ above, we obtain,

$$\left|\frac{df(\theta_0)}{d\theta}[\theta-\theta_0]\right|\lesssim\|\beta-\beta_0\|_E+\|\nabla G_l(\phi_l(X,\beta_0))-\nabla G_{0,l}(\phi_l(X,\beta_0))\|_{L_2(X)}.$$

By the fact that $G_l, G_{l,0}\in\Lambda_\infty^p(\Phi)$, first derivatives are Hölder continuous and by Assumption 4 (iii), we obtain $\sup_{\|\theta-\theta_0\|>0}|\frac{df(\theta_0)}{d\theta}[\theta-\theta_0]|/\|\theta-\theta_0\|<\infty$, which establishes the claim of Condition 4.1. □

# Bibliography

J. Abrevaya and J. A. Hausman. Semiparametric estimation with mismeasured dependent variables: an application to duration models for unemployment spells. *Annales d'Economie et de Statistique*, pages 243–275, 1999.

J. Abrevaya and J. A. Hausman. Response error in a transformation model with an application to earnings-equation estimation. *The Econometrics Journal*, 7(2): 366–388, 2004.

J. Abrevaya and Y. Shin. Rank estimation of partially linear index models. *The Econometrics Journal*, 14(3):409–437, 2011.

C. Ai and X. Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71:1795–1843, 2003.

S. Athey, J. Tibshirani, and S. Wager. Generalized Random Forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

M. Banerjee, D. Mukherjee, and S. Mishra. Semiparametric binary regression models under shape constraints with an application to indian schooling data. *Journal of Econometrics*, 149(2):101–117, 2009.

A. Belloni, V. Chernozhukov, D. Chetverikov, and K. Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345 – 366, 2015. High Dimensional Problems in Econometrics.

D. Ben-Moshe, X. D'Haultfœuille, and A. Lewbel. Identification of additive and polynomial models of mismeasured regressors without instruments. *Journal of Econometrics*, 200(2):207–222, 2017.

R. Beran and P. Hall. Estimating coefficient distributions in random coefficient regressions. *Ann. Statist.*, 20(4):1970–1984, 12 1992.

R. Beran and P. W. Millar. Minimum distance estimation in random coefficient regression models. *Ann. Statist.*, 22(4):1976–1992, 12 1994.

R. Beran, A. Feuerverger, and P. Hall. On nonparametric estimation of intercept and slope distributions in random coefficient regression. *Ann. Statist.*, 24(6): 2569–2592, 12 1996.

A. Beresteanu. Nonparametric estimation of regression functions under restrictions on partial derivatives. Working Papers 04-06, Duke University, Department of Economics, 2004.

J. Berkson. Are there two regressions? *Journal of the American Statistical Association*, 45(250):164–180, 1950.

G. Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13:1063–1095, 2012.

R. Blundell, J. Horowitz, and M. Parey. Nonparametric estimation of a nonseparable demand function under the slutsky inequality restriction. *Review of Economics and Statistics*, 99(2):291–304, 2017.

R. W. Blundell and J. L. Powell. Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71(3):655–679, 2004.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Y. Breitmoser. The axiomatic foundation of logit. Discussion Paper No. 78, Collaborative Research Center Transregio 190, 2018.

C. Breunig. Varying random coefficient models. *Journal of Econometrics*, 221(2): 381–408, 2021.

C. Breunig and X. Chen. Adaptive, rate-optimal hypothesis testing in nonparametric iv models. *arXiv preprint arXiv:2006.09587v2*, 2023.

C. Breunig and P. Haan. Nonparametric regression with selectively missing covariates. *arXiv preprint arXiv:1810.00411*, 2018.

C. Breunig and S. Hoderlein. Specification testing in random coefficient models. *Quantitative Economics*, 9(3):1371–1417, November 2018.

C. Breunig, E. Mammen, and A. Simoni. Nonparametric estimation in case of endogenous selection. *Journal of Econometrics*, 202(2):268 – 285, 2018.

C. Breunig, S. Huck, T. Schmidt, and G. Weizsäcker. The standard portfolio choice problem in germany. *CRC TRR 190 Discussion Paper*, (171), 2019.

C. Cavanagh and R. P. Sherman. Rank estimators for monotonic index models. *Journal of Econometrics*, 84(2):351–381, 1998.

X. Chen. Large sample sieve estimation of semi-nonparametric models. volume 6, Part B of *Handbook of Econometrics*, pages 5549 – 5632. Elsevier, 2007.

X. Chen and D. Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012.

X. Chen and X. Shen. Sieve extremum estimates for weakly dependent data. *Econometrica*, 66(2):289–314, 1998.

X. Chen, O. Linton, and I. Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71:1591–1608, 2003.

X. Chen, H. Hong, and E. Tamer. Measurement Error Models with Auxiliary Data. *The Review of Economic Studies*, 72(2):343–366, 04 2005.

X. Chen, H. Hong, and D. Nekipelov. Nonlinear models of measurement errors. *Journal of Economic Literature*, 49(4):901–37, December 2011.

V. Chernozhukov, I. Fernandez-Val, and A. Galichon. Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575, 2009.

V. Chernozhukov, C. Hansen, and M. Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7(1):649–688, 2015.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 5:261–265, 2017.

V. Chernozhukov, M. Demirer, E. Duflo, and I. Fernandez-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. *arXiv*, 2019.

V. Chernozhukov, W. Newey, and R. Singh. De-biased machine learning of global and local parameters using regularized riesz representers, 2020.

D. Chetverikov and D. Wilhelm. Nonparametric instrumental variable estimation under monotonicity. *Econometrica*, 85(4):1303–1320, 2017.

D. Chetverikov, A. Santos, and A. M. Shaikh. The econometrics of shape restrictions. *Annual Review of Economics*, 10(1):31–63, 2018.

P.-A. Chiappori, I. Komunjer, and D. Kristensen. Nonparametric identification and estimation of transformation models. *Journal of Econometrics*, 188(1):22 – 39, 2015.

S. Clemencon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

G. Compiani. Market counterfactuals and the specification of multi-product demand: A nonparametric approach. *Quantitative Economics*, forthcoming, 2021.

H. Curry and I. Schoenberg. On polya frequency functions iv: The fundamental spline functions and their limits. *Journal d'Analyse Mathematique*, 17:71–107, 1966.

M. de Nadai and A. Lewbel. Nonparametric errors in variables models with measurement errors on both sides of the equation. *Journal of Econometrics*, 191(1): 19 – 32, 2016. ISSN 0304-4076.

M. Delecroix and C. Thomas-Agnan. *Spline and Kernel Regression under Shape Restrictions*, chapter 5, pages 109–133. John Wiley & Sons, Ltd, 2000.

X. D'Haultfoeuille. A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, 154(1):1–15, 2010.

T. Drerup, B. Enke, and H.-M. von Gaudecker. The precision of subjective data and the explanatory power of economic models. *Journal of Econometrics*, 200(2): 378 – 389, 2017. Measurement Error Models.

F. Dunker, J.-P. Florens, T. Hohage, J. Johannes, and E. Mammen. Iterative estimation of solutions to noisy nonlinear operator equations in nonparametric instrumental regression. Technical report, University of Göttingen, 2011.

B. Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014.

J. C. Escanciano, D. Jacho-Chávez, and A. Lewbel. Identification and estimation of semiparametric two-step models. *Quantitative Economics*, 7(2):561–589, 2016.

Y. Fan, F. Han, W. Li, and X.-H. Zhou. On rank estimators in increasing dimensions. *Journal of Econometrics*, 214:379–412, 2020.

J. Fox. Semiparametric Estimation of Multinomial Discrete-Choice Models Using a Subset of Choices. *RAND Journal of Economics*, 38(4):1002–1019, 2007.

J. Freyberger and M. Masten. Compactness of infinite dimensional parameter spaces. cemmap working paper CWP01/16, London, 2015.

J. Freyberger and B. Reeves. Inference under shape restrictions. Working paper, University of Wisconsin, 2019.

A. K. Han. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, 35(2-3):303–316, 1987.

J. A. Hausman, W. K. Newey, H. Ichimura, and J. L. Powell. Identification and estimation of polynomial errors-in-variables models. *Journal of Econometrics*, 50 (3):273 – 295, 1991.

S. Hoderlein and J. Winter. Structural measurement errors in nonseparable models. *Journal of Econometrics*, 157(2):432 – 440, 2010.

S. Hoderlein, J. Klemelä, and E. Mammen. Analyzing the random coefficient model nonparametrically. *Econometric Theory*, 26(3):804–837, 2010.

S. Hoderlein, B. Siflinger, and J. Winter. Identification of structural models in the presence of measurement error due to rounding in survey responses. 2015.

J. L. Horowitz. A Smoothed Maximum Score Estimator for the Binary Response Model. *Econometrica*, 60:505–531, 1992.

J. L. Horowitz. *Semiparametric Methods in Econometrics*. Springer New York, NY, 1998.

J. L. Horowitz and S. Lee. Nonparametric estimation and inference under shape restrictions. *Journal of Econometrics*, 201(1):108 – 126, 2017.

Y. Hu and S. M. Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.

S. Huck, T. Schmidt, and G. Weizsacker. The standard portfolio choice problem in germany. *CESifo Working Paper Series No. 5441*, 2015.

H. Ichimura. Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models. *Journal of Econometrics*, 58:71–120, 1993.

H. Ichimura and L.-F. Lee. *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, chapter Semiparametric Estimation of Multiple Index Models: Single Equation Estimation. Cambridge University Press, 1991.

G. W. Imbens and W. K. Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009. ISSN 1468-0262. doi: 10.3982/ECTA7108.

D. Jacho-Chavez, A. Lewbel, and O. Linton. Identification and nonparametric estimation of a transformed additively separable model. *Journal of Econometrics*, 156(2):392 – 407, 2010.

J. Jureckova, H. L. Koul, R. Navratil, and J. Picek. Behavior of r-estimators under measurement errors. *Bernoulli*, 22(2):1093–1112, 2016.

S. Khan. Two-stage rank estimation of quantile index models. *Journal of Econometrics*, 100(2):319–355, 2001.

R. Klein and F. Vella. A semiparametric model for binary response and continuous outcomes under index heteroscedasticity. *Journal of Applied Econometrics*, 24(5): 735–762, 2009. ISSN 1099-1255.

R. Klein, C. Shen, and F. Vella. Estimation of marginal effects in semiparametric selection models with binary outcomes. *Journal of Econometrics*, 185(1):82 – 94, 2015. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2014.10.006.

R. W. Klein and R. H. Spady. An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421, 1993.

M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer-Verlag New York, 1 edition, 2008. ISBN 978-0-387-74977-8.

L.-F. Lee. Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models. *Journal of Econometrics*, 65:381–428, 1995.

A. Lewbel. Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics*, 97(1):145–177, 2000.

A. Lewbel. An overview of the special regressor method, 08 2014.

E. Mammen. Estimating a smooth monotone regression function. *The Annals of Statistics*, 19(2):724–740, 1991. ISSN 00905364.

C. Manski. Maximum Score Estimation of the Stochastic Utility Model of Choice. *Journal of Econometrics*, 3:205–228, 1975.

C. Manski. Semiparametric Analysis of Discrete Response. *Journal of Econometrics*, 27:313–333, 1985.

M. A. Masten. Random Coefficients on Endogenous Variables in Simultaneous Equations Models. *The Review of Economic Studies*, 85(2):1193–1250, 08 2017.

R. Matzkin. Semiparametric Estimation of Monotone and Concave Utility Functions for Polychotomous Choice Models. *Econometrica*, 59(5):1315–1327, 1991a.

R. L. Matzkin. A nonparametric maximum rank correlation estimator. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, volume 5, page 277. Cambridge University Press, 1991b.

R. L. Matzkin. Nonparametric identification and estimation of polychotomous choice models. *Journal of Econometrics*, 58(1-2):137–168, 1993.

R. L. Matzkin. Restrictions of economic theory in nonparametric methods. *Handbook of econometrics*, 4:2523–2558, 1994.

R. L. Matzkin. Nonparametric identification. *Handbook of Econometrics*, 6:5307–5368, 2007.

W. K. Newey, J. L. Powell, and F. Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603, 1999.

D. Nolan and D. Pollard. U-processes: Rates of convergence. *The Annals of Statistics*, 15(2):780–799, 1987.

A. Pakes and D. Pollard. Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society*, pages 1027–1057, 1989.

D. Pollard. *Convergence of Stochastic Processes*. Springer Series in Statistics, 1984.

J. O. Ramsay. Monotone regression splines in action. *Statistical Science*, 3(4): 425–441, 1988.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4): 931–954, 1988.

C. Rothe. Semiparametric estimation of binary response models with endogenous regressors. *Journal of Econometrics*, 153(1):51 – 64, 2009. ISSN 0304-4076.

S. M. Schennach. Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica*, 75(1):201–239, 2007.

S. M. Schennach. Measurement error in nonlinear models - a review. *Advances in Economics and Econometrics, Theory and Applications: Tenth World Congress of the Econometric Society*, Jul 2013.

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *Ann. Statist.*, 43(4):1716–1741, 08 2015.

X. Shen. On methods of sieves and penalization. *The Annals of Statistics*, 25(6): 2555–2591, 1997.

R. P. Sherman. The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society*, pages 123–137, 1993.

Y. Shin. Local rank estimation of transformation models with functional coefficients. *Econometric Theory*, 26(6):1807–1819, 2010.

K. E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2 edition, 2009.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics (Springer Series in Statistics)*. Springer, corrected edition, Nov. 2000. ISBN 0387946403.

A. van der Vaart and J. Wellner. A note on bounds for vc dimensions. *IMS Collections: High Dimensional Probability*, 5:103–107, 2009.

A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018.

S. Wager and G. Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv*, 2016.

H. White and K. Chalak. Testing a conditional form of exogeneity. *Economics Letters*, 109(2):88–90, 2010.

X. Wu and R. Sickles. Semiparametric estimation under shape constraints. *Econometrics and Statistics*, 6:74–89, 2018. STATISTICS OF EXTREMES AND APPLICATIONS.

Y. Wu and Y. Zhang. Partially monotone tensor spline estimation of the joint distribution function with bivariate current status data. *The Annals of Statistics*, 40(3):1609–1636, 2012.

J. Yan. A Smoothed Maximum Score Estimator for Multinomial Discrete Choice Models. Working Paper, The Chinese University of Hongkong, 2014.

## Erklärung zu Selbstständigkeit und Hilfsmitteln

Hiermit erkläre ich, dass ich die Dissertation selbständig und nur unter der Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt habe.
Ich bezeuge durch meine Unterschrift, dass meine Angaben über die bei der Abfassung meiner Dissertation benutzten Hilfsmittel, über die mir zuteil gewordene Hilfe sowie über frühere Begutachtungen meiner Dissertation in jeder Hinsicht der Wahrheit entsprechen.

Datum: Unterschrift:

## Kumulative Dissertation, Erklärung zu Ko-Autoren, Eigenanteil und Publikationsstatus

| Lfd. Nr | Titel der Einzelarbeit | Namen der Ko-Autoren | Erklärung zum Eigenanteil | Publikationsstatus |
|---|---|---|---|---|
| 1 | Nonclassical Measurement Error in the Outcome Variable | Prof. Dr. Christoph Breunig | Identifikationsansatz, Herleitung der theoretischen Resultate, Implementierung, Simulation und empirische Anwendung | Veröffentlichung auf arxiv.org unter Nummer 2009.12665 am 26.09.2020, Revision am 30.05.2021 |
| 2 | Estimation of Conditional Random Coefficient Models using Machine Learning Techniques | - | einziger Autor | Veröffentlichung auf arxiv.org unter Nummer 2201.08366 am 20.01.2022 |
| 3 | A Simple Shape-Constrained Estimator for Semi(non)parametric Discrete Choice Models | - | einziger Autor | - |