Hoderlein, Stefan; Nesheim, Lars; Simoni, Anna

Working Paper

Semiparametric estimation of random coefficients in structural economic models

cemmap working paper, No. CWP09/12

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Hoderlein, Stefan; Nesheim, Lars; Simoni, Anna (2012) : Semiparametric estimation of random coefficients in structural economic models, cemmap working paper, No. CWP09/12, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2012.0912

This Version is available at:
https://hdl.handle.net/10419/64796

cemmap
centre for microdata methods and practice

Semiparametric estimation of random coefficients in structural economic models

Stefan Hoderlein
Lars Nesheim
Anna Simoni

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP09/12

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Semiparametric Estimation of Random Coefficients in Structural Economic Models

Stefan Hoderlein* Lars Nesheim Anna Simoni

Boston College University College London Università Bocconi

March 26, 2012

Abstract

In structural economic models, individuals are usually characterized as solving a decision problem that is governed by a finite set of parameters. This paper discusses the nonparametric estimation of the probability density function of these parameters if they are allowed to vary continuously across the population. We establish that the problem of recovering the probability density function of random parameters falls into the class of non-linear inverse problem. This framework helps us to answer the question whether there exist densities that satisfy this relationship. It also allows us to characterize the identified set of such densities. We obtain novel conditions for point identification, and establish that point identification is generically weak. Given this insight, we provide a consistent nonparametric estimator that accounts for this fact, and derive its asymptotic distribution. Our general framework allows us to deal with unobservable nuisance variables, e.g., measurement error, but also covers the case when there are no such nuisance variables. Finally, Monte Carlo experiments for several structural models are provided which illustrate the performance of our estimation procedure.

Keywords: Structural Models, Heterogeneity, Nonparametric Identification, Random Coefficients, Inverse Problems.

*Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA, email: stefan_hoderlein@yahoo.com. We have benefited from comments and discussions with Victor Aguirregabiria, Orazio Attanasio, Richard Blundell, Chris Carroll, Jeremy Fox, Emmanuel Guerre, Dirk Krueger, Arthur Lewbel, Enno Mammen, Rosa Matzkin, Jean-Marc Robin, Sami Stouli, Yuanyuan Wan, Hal White, and seminar participants at the conferences for "Nonparametrics and Demand" at MIT, for "Revealed Preferences and Partial Identification" at Montreal University, the NBER 2011 Summer Meeting on Consumption, the CEAR 2011 conference in Denver, the Royal Economic Society Meeting 2011 in London, the ESEM 2011 in Oslo, Boston College, CalTech, Cergy-Pontoise University, Johns Hopkins University, Mannheim University, NYU, Princeton, Queen Mary University, UC San Diego, UCL, UCLA, Stanford, and Toronto. All remaining errors are ours. Lars Nesheim gratefully acknowledges financial support from the UK Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001.

1 Introduction

Motivation. Many structural microeconomic models postulate that individual decision makers solve complicated optimization problems which are governed by a small number of structural parameters θ. These parameters are fixed for every individual. However, economic theory does not postulate that they be the same for every individual. Frequently, one expects that they will differ substantially across individuals. Yet, in most empirical applications, the extent to which individual decision makers are allowed to vary is severely constrained to depend entirely on observable variables, to involve only few discrete types, or to be monotonic in a scalar unobservable. These constraints on heterogeneity are unappealing and are typically not based on economic theory. A natural way to relax the constraints and make the structural model assumptions more appealing is to assume that the unobservable parameters θ in the individuals' decision problems are random parameters drawn from a fully flexible nonparametric continuous distribution that may be correlated with some observable regressors. Proposing and analyzing such an approach is the main innovation in this paper.

To give an example, in the workhorse Euler equation models of the consumption literature, the consumption function is characterized by the first order condition

$$\partial_c u(C_t, \theta) = \mathbb{E}\left[\partial_c u(C_{t+1}, \theta) | \mathcal{I}_t\right] \tag{1.1}$$

where u denotes instantaneous utility, C_t consumption in period t, $\mathcal{I}_t$ is the information set of the economic agent in period t, and ∂_x denotes the partial derivative with respect to x.[1] The information set $\mathcal{I}_t$ consists of exogenous observable variables Z_t, endogenous observable variables W_t and may also depend on unobservable variables. Interest centers on the distribution of (random) parameters θ. To not unduly restrict the generality of the model, we may want to allow for the possibility that there also be unobservable nuisance variables ε_t in the model. These may enter $\mathcal{I}_t$, e.g., heterogeneous beliefs about the income process, or they may reflect measurement error (e.g., $C_t = C_t^* + \varepsilon_t$, where C_t^* is the true consumption and ε_t is a measurement error). We remark that solutions to problems of the type displayed in equation (1.1) are often only defined implicitly and need to be obtained numerically. Our approach allows us to deal with these cases as well. Indeed, a benefit of our approach is that numerical solution of (1.1) is separated from econometric estimation of the distribution of unobserved heterogeneity.

In this paper we propose a general framework to analyze a large class of such structural models. Specifically, we consider all structural economic models that can be characterized by the following condition

$$\Psi(C, W, Z, \theta, \varepsilon) = 0 \tag{1.2}$$

[1] To focus on essentials, we have set the interest rate equal to the discount rate.

where C is a scalar observable outcome, θ is a d-vector of unobserved random parameters of interest, ε is an unobservable random variable, W is a k-vector of observable random variables allowed to be correlated with θ while Z is a l-vector of random observable variables that are uncorrelated with θ. All variables are assumed to be continuously distributed.

Our analysis covers two cases, one where there is no unobservable nuisance variable ε and one which includes ε.[2] Since the latter case is more general, we focus on it. Nevertheless, we give results for the first case and discuss in particular how the problem changes in this case.

Our model is structural in that we assume that Ψ is a known function coming from economic theory. Our aim is to identify and non-parametrically estimate the distribution of θ conditional on W. We do not require any monotonicity of Ψ in ε or θ. While our results apply to general structural models of the form (1.2), we develop further results that apply to the specific example of the Euler equation in order to fix ideas and motivate the discussion.

The two key notions we pursue in this paper are heterogeneity and knowledge of the structural equation. When we lack information about the probability distribution of heterogeneity in the population $\left(\text{for example the density } f_{\theta|W}\right)$ but have knowledge about the structural function Ψ, we can use this knowledge to define a mapping from $f_{\theta|W}$ to the population probability density function (pdf) of observables $f_{C|WZ}$. In our setting, this mapping is provided by the integral equation

$$f_{C|WZ} = T f_{\theta|W}, \quad P_W - \text{a.s.}, \tag{1.3}$$

where the integral operator T maps the density of random parameters into the density of the observable variables $f_{C|WZ}$ and P_W denotes the probability distribution of W. When Ψ can be uniquely solved for C as a measurable function of the other variables, the operator T can be explicitly characterized in terms of the structural economic model in (1.2). We focus on the case in which Ψ is differentiable and has a unique global solution $C = \varphi(W, Z, \theta, \varepsilon)$. In the case where φ is invertible in ε, i.e. $\varepsilon = \varphi^{-1}(w, z, \theta, c)$, the operator takes the form

$$f_{C|WZ}(c; w, z) = \int_\Theta f_{\varepsilon|WZ\theta} \circ \varphi^{-1} \left| \frac{\partial_\varepsilon \Psi(c, w, z, \theta, \varphi^{-1}(w, z, \theta, c))}{\partial_c \Psi(c, w, z, \theta, \varphi^{-1}(w, z, \theta, c))} \right| 1_{\mathcal{C}_\theta}(c) f_{\theta|W}(\theta; w) d\theta, \quad P_W - \text{a.s.},$$

where $f_{\varepsilon|WZ\theta}$ is the pdf of ε conditional on (W, Z, θ) and $\mathcal{C}_\theta$ denotes the support of the conditional distribution of C given (W, Z, θ). This paper focuses on this integral equation to establish identification and obtain an estimator for $f_{\theta|W}$. More specifically, given this representation, we can discuss the issues of existence, uniqueness and stability of the inverse. Translated into econometric terms, existence will correspond to conditions for at least partial identification and will allow us to characterize the partially identified set, and uniqueness will lead us to a

[2] For simplicity, we assume throughout this paper that ε is a random unobserved scalar; this could be relaxed without great difficulty.

novel condition for point identification called $\mathcal{T}$-completeness which exploits the structure of the problem. Finally, stability will relate to the question whether we can construct a feasible, consistent estimator, given the complex and high dimensional nature of the problem.

Contributions relative to the Literature. As already mentioned, this line of work extends the parametric structural models literature to allow for endogenous random coefficients. This literature is vast; the consumption literature which originally motivated this research is, for instance, surveyed in Deaton (1993) and Attanasio and Weber (2010). When parameter heterogeneity is introduced, identification becomes a crucial concern. The question is whether we can non-parametrically identify the distribution of preference parameters, and if yes, whether and how we can build an estimator based on the identification principle.

To be able to answer this question, we propose a nonparametric framework. The nonparametric features are not economically marginal generalizations. First, we answer the nonparametric identification question, i.e., where does the identifying power of the model come from, if not from the functional form, and what observable variables are required to ensure identification. Second, we provide insights into when identification is only partial, and we provide novel conditions for point identification. All of these steps are related to contributions in the literature, as we now explain.

Our work is complementary to the nonparametric nonseparable approach (see Matzkin (2007a) and Matzkin (2007b) for surveys). For example, Appendix A in Matzkin (2003) discusses how to estimate a nonparametric model with high dimensional heterogeneity using separability and the restriction that the random parameters are mutually independent. Our approach relaxes the independence and separability conditions in Matzkin (2003) and imposes alternative functional form restrictions. Alternatively, if computationally complexity makes a fully nonparametric approach infeasible, our approach may remain feasible. In particular, our approach completely separates computational issues related to approximation of φ from identification and estimation issues.

Most closely related to our approach are nonparametric econometric models involving random parameters. In particular, there is a literature that considers linear/single index nonparametric random coefficients models, as in Beran et al. (1996), Ichimura and Thompson (1998), Hoderlein et al. (2010), and Gautier and Kitamura (2010). In these papers, the random coefficients are continuously distributed and fully independent of regressors. In addition, there is the structural treatment effects literature (see Abbring and Heckman (2007) for a survey). In this literature, the random coefficients are allowed to be correlated with the treatment variables. We complement these literatures by moving away from linear models and allowing for nonparametric endogenous random coefficients in nonlinear structural models arising from economic theory; models in which the function mapping regressors into outcomes is often only implicitly

defined. Our most general model allows for nuisance unobservables and is hence more closely related to mixture models discussed below.

In the case when there is nuisance heterogeneity ε, our approach resembles somewhat deconvolution approaches to modeling unobservable variables, a line of research in econometrics that started with the seminal work of Heckman and Singer (1984), henceforth HS. In our notation, it is centered around the equation

$$f_{C|WZ}(c;w,z) = \int_{\Theta} f_{C|WZ\theta}(c;w,z,\theta) f_{\theta|W}(\theta;w) d\theta. \tag{1.4}$$

In HS's work, which focuses on duration analysis, the density $f_{C|WZ\theta}(c;w,z,\theta)$ is central. It is assumed to depend on a finite parameter σ which is the main structural object of interest while $f_{\theta|W}$ is a nuisance parameter. Closely related to HS are: Henry et al. (2011), who focus on estimating $f_{C|WZ\theta}(c;w,z,\theta)$ nonparametrically while restricting θ to be discretely distributed, Kasahara and Shimotsu (2009), who also considers finitely many types, and Bonhomme (2011), who like HS aims at estimating a finite parameter of interest σ when the exogenous variation comes from a panel. This line of work is closely related to mixture models. In contrast to all of these references, in our model interest centers on $f_{\theta|W}$, and the kernel of the operator in (1.4) obtains structure from the economic primitives of the model.

There is a recent line of work that discusses identification of random coefficients in models that are motivated by empirical IO, see in particular Bajari et al. (2012), Fox and Gandhi (2010). These models are close in spirit to our approach in terms of the nonparametric objectives of the analysis. However, there are a number of pronounced differences: their analysis is mainly based on a discrete (resp. countable) number of types, the identification results are not constructive, and all of their results establish point identification. In contrast, we focus on the nonparametric case, discuss the issue of partial identification, and focus on the ill posed character of the estimation problem. Consequently, these two concomitant approaches can be seen as complements, much as nonparametric instrumental variables with discrete, resp. continuous endogenous regressors complement each other.

These differences in the object of interest and the focus on the integral equation (1.3) in our approach make our work related to the general inverse problem literature, see Carrasco et al. (2007) for an overview. In particular, recovering the probability density of θ nonparametrically from (1.3) is equivalent to solving a convexly constrained integral equation of the first kind. Unconstrained integral equations of the first kind have been studied extensively in the literature on nonparametric instrumental regression, see *e.g.* Florens (2003), Newey and Powell (2003), Darolles et al. (2011) and Hall and Horowitz (2005). While our object of interest is very different from the estimation of a nonparametric IV regression function, we have some

overlap with these references in terms of the tools we employ. In particular, we use Tikhonov regularization which was proposed by Tikhonov (1963), and introduced into econometrics by Carrasco and Florens (2000), Florens (2003) and Darolles et al. (2011), among others.

Our estimating equation is also related to the approach of Hu and Schennach (2008). However, our model differs in many core aspects from their model, not least the different object of interest (i.e., the distribution of random parameters), and the structural nonseparability of the model considered. Moreover, our exclusion restrictions are fundamentally different from theirs (e.g., we do not assume conditional independence of C and Z given θ) and motivated by the structural economic application. Indeed, even in the case where we include measurement error, we allow for correlation between the error and the unobservable of interest θ. Finally, we do not assume or require injectivity of the operator defining the estimating equation, and we are able to characterize the identified set and provide conditions for point identification.

Structure of the Paper. We develop our analysis of the above class of models in the following way. The next section describes our basic setup including key assumptions and discusses several important economic example. Section 3 provides the main identification theorem. Section 4 discusses estimation by sample counterparts. Finally, we illustrate our approach with a simulation exercise in Section 5. An application using panel data on consumption constructed from the PSID and the CEX is transferred to a companion paper Hoderlein et al. (2012). Finally, Section 6 concludes.

2 The general structural model

In this section we introduce the basic building blocks of our model. We provide formal notation, clarify and discuss the assumptions, and establish that several important economic models fall into our framework.

2.1 Basic definitions and assumptions

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space and $(C, W, Z, \theta, \varepsilon)$ be a real-valued random vector defined on it, and partitioned into $C \in \mathbb{R}$, $W \in \mathbb{R}^k$, $Z \in \mathbb{R}^l$, $\theta \in \mathbb{R}^d$ and $\varepsilon \in \mathbb{R}$, with k, l and d finite integers. We denote by $\mathcal{B}_\mathcal{C}$, $\mathcal{B}_\mathcal{W}$, $\mathcal{B}_\mathcal{Z}$, $\mathcal{B}_\Theta$ and $\mathcal{B}_\varepsilon$ the corresponding Borel σ-fields in $\mathbb{R}$, $\mathbb{R}^k$, $\mathbb{R}^l$, $\mathbb{R}^d$ and $\mathbb{R}$, respectively.

We use capital Latin letters for observable random variables and lowercase Latin letters for their realizations. The unobservable random variables and their realization will be denoted by lowercase Greek letters without distinction. The first assumption specifies the structural data generating process (DGP) that we are considering.

Assumption 1. *The random element $(C, W, Z, \theta, \varepsilon)$ satisfies a structural economic model*

$$\Psi(C, W, Z, \theta, \varepsilon) = 0 \quad a.s. \tag{2.1}$$

*where Ψ is a **known** Borel measurable real-valued function. Moreover, we assume that (2.1) has a unique global solution in terms of C:*

$$C = \varphi(W, Z, \theta, \varepsilon), \quad a.s.$$

where $\varphi : \mathbb{R}^{k+l+d+1} \to \mathbb{R}$ is a Borel-measurable function.

This assumption describes how our structural model links observable variables (C, W, Z) to unobservable ones (θ, ε). We distinguish between three different observable variables: C is the dependent variable, while W and Z denote variables that cause C. The distinction between W and Z is made because we allow the former to be correlated with θ while the latter is assumed to be conditionally independent.[3] This distinction is motivated by applications in which some important explanatory variables are endogenous. Needless to mention, we can handle the case when no such variables are present.

The distinction between the unobservable variables θ and ε is made to separate random parameters of interest θ from an error term ε. In our analysis, we allow the distribution of θ, which is the distribution of interest, to be completely nonparametric, and assume that the distribution of ε is parametric in the sense that we allow unknown random parameters of finite-dimension in the distribution of ε.

We do not require that the function φ be available in closed-form. Given its existence and uniqueness, it may as well be available in numerical form only. Moreover, we do not require that φ be globally monotone in ε; instead it may only be piecewise monotone in ε. This is an important weakening of assumptions, as any monotonicity condition at this stage is rather implausible. To account for piecewise monotonicity, let $\mathcal{E}_1, \ldots, \mathcal{E}_s$ be a partition of $\mathbb{R}$ such that $\varphi(w, z, \theta, \cdot) : \mathcal{E}_i \to \mathbb{R}$ is one-to-one for each $i = 1, \ldots, s$, for given (w, z, θ), but not necessarily globally. We denote by $\varepsilon^i = \varphi_i^{-1}(w, z, \theta, \cdot) : \mathbb{R} \to \mathcal{E}_i$ the corresponding inverse mapping for given (w, z, θ). Thus, s is a function of (w, z, θ). In principle, s could be either countable or uncountable but for simplicity we assume throughout that s is a finite number.

We write $\partial_c \varphi_i^{-1}(w, z, \theta, c)$ and $\partial_\varepsilon \varphi(w, z, \theta, \varepsilon)$ to denote the partial derivatives of φ_i^{-1}, $i = 1, \ldots, s$ and φ, respectively, with respect to C and ε for given (w, z, θ). This allows us to introduce a differentiability assumption on Ψ.

[3] This obviously nests the case where all causal variables are independent of θ.

Assumption 2. *The structural function* $\Psi : \mathbb{R}^{k+l+d+2} \to \mathbb{R}$ *is almost everywhere differentiable in* C *and in* ε *with* $\partial_c \Psi(c, w, z, \theta, \varepsilon) \neq 0$ *and* $\partial_\varepsilon \Psi(c, w, z, \theta, \varepsilon) \neq 0$ *except, possibly, on a set of* $(c, w, z, \theta, \varepsilon)$ *values whose Lebesgue measure is* 0.

The next three assumptions we introduce concern the stochastic properties of the random vector $(C, W, Z, \theta, \varepsilon)$. Under Assumption 1, the random variable C has a degenerate distribution conditional on $(W, Z, \theta, \varepsilon)$. Hence, the joint distribution of (C, W, Z, θ) is completely characterized by the joint probability distribution of $(W, Z, \theta, \varepsilon)$ and the function φ. For the conditional distribution of the nuisance variable ε, conditional on (W, Z, θ), we only require that it is known up to a finite-dimensional parameter. This is formally stated in the following assumption, where we denote by $P_{WZ\theta}$ the joint distribution on $\mathcal{B}_\mathcal{W} \otimes \mathcal{B}_\mathcal{Z} \otimes \mathcal{B}_\Theta$.

Assumption 3. *The conditional probability distribution* $P_{\varepsilon|WZ\theta}$ *on* $\mathcal{B}_\varepsilon$ *given* (W, Z, θ) *admits a Radon-Nikodym derivative* $f_{\varepsilon|WZ\theta}$ *with respect to the Lebesgue measure. This probability density function (pdf, hereafter)* $f_{\varepsilon|WZ\theta}$ *is known, up to a finite-dimensional parameter* $\theta_\varepsilon \subset \theta$. *Moreover, there exists a constant* $m_\varepsilon > 0$ *such that* $m_\varepsilon \leq f_{\varepsilon|WZ\theta} < \infty$, $P_{WZ\theta}$*-a.s. on the support of* $P_{\varepsilon|WZ\theta}$.

This assumption allows ε to depend on all variables in the model but is also satisfied when ε is independent of (W, Z, θ), which may be relevant in the measurement error setup. Note also that, unlike in deconvolution, ε does not need to be independent of θ since we are not de-convolving a probability distribution. In general, we can allow for ε and θ to be dependent. In other applications, it may be useful to split the random parameter θ into two subparameters (θ_1, θ_2) and confine dependence between ε and θ to one of the two components of θ. Mathematically, this is equivalent to allowing the distribution of ε to depend on some ε-specific heterogeneity parameters. For instance, we could model $P_{\varepsilon|WZ\theta}$ to be normal with mean μ and variance σ^2. Thus, μ and σ^2 may be functions of (W, Z, θ) or even elements of the vector θ itself.

This allows a great deal of flexibility in structural modeling. By allowing $f_{\varepsilon|WZ\theta}$ to be known up to a finite dimensional random parameter (a parameter included in the vector θ), we allow for cases where not everything is known about $f_{\varepsilon|WZ\theta}$. We may further weaken this assumption: $f_{\varepsilon|WZ\theta}$ could also be a finite mixture of normal *pdfs* where, besides the vector of means and variances, also the mixture weights could depend on θ or be part of θ. Therefore, the specification can be very close to a nonparametric specification for $f_{\varepsilon|WZ\theta}$, provided there is enough independent variation in the data, as defined below. Adding flexibility in this fashion does come at a cost, however, as it will reduce the rate of the estimation or even lead to a failure of point identification, and we encounter the typical semiparametric trade-off of flexibility vs feasibility.

The probabilistic model relevant for our paper is the joint conditional probability distribution $P_{CZ\theta|W}$ on $\mathcal{B}_{\mathcal{C}} \otimes \mathcal{B}_{\mathcal{Z}} \otimes \mathcal{B}_{\Theta}$ conditional on W. Our assumptions will imply that this distribution is absolutely continuous with respect to the Lebesgue measure with Radon-Nikodym derivative $f_{CZ\theta|W}$. We denote by $\mathcal{C} \subset \mathbb{R}$, $\mathcal{Z} \subset \mathbb{R}^l$ and $\Theta \subset \mathbb{R}^d$ the supports of $P_{C|WZ}$, $P_{Z|W}$ and $P_{\theta|W}$, respectively, where $P_{C|WZ}$ is the conditional distribution on $\mathcal{B}_{\mathcal{C}}$ given (W, Z), $P_{Z|W}$ (resp. $P_{\theta|W}$) is the conditional distribution on $\mathcal{B}_{\mathcal{Z}}$ (resp. on $\mathcal{B}_{\Theta}$) given W. The marginal distribution on $\mathcal{B}_{\mathcal{W}}$ is denoted by P_W and has support $\mathcal{W} \subset \mathbb{R}^k$. We consider C and Z to be continuous random vectors while W can be either continuous or discrete. In contrast, θ is assumed to be continuously distributed, as stated in the following assumption.

Assumption 4. *The conditional probability distribution $P_{\theta|W}$ on $\mathcal{B}_{\Theta}$ given W admits a Radon-Nikodym derivative $f_{\theta|W}$ with respect to the Lebesgue measure. This pdf is strictly positive and bounded on its support P_W-a.s., i.e. there exists a constant $m_\theta > 0$ such that $m_\theta \leq f_{\theta|W} < \infty$. Moreover, the support Θ of $f_{\theta|W}$ does not depend on W.*

Remark 1. For our analysis, we need a parametric form for $f_{C|WZ\theta}$, that is, the conditional *pdf* of C given (Z, W, θ). This *pdf* can be recovered by using Assumptions 2, 3 and the function φ whose existence is assumed in Assumption 1. However, economic theory sometimes provides suggestions on the functional form of $f_{C|WZ\theta}$. For instance, Heckman and Singer (1984) and references therein give some examples in duration models where economic theory provides a structure for $f_{C|WZ\theta}$.

We conjecture that a setup where some or all of the variables (Z, θ) are discrete could be tackled by similar arguments. However, we expect that discreteness of Z would come at the cost of point identification as in linear exogenous random coefficient models. This case is currently beyond the scope of this paper.

The last assumption we introduce is an independence condition and is important for point identification of the *pdf* $f_{\theta|W}$ of the structural parameters of interest.

Assumption 5. *The random element Z is conditionally independent of θ given W, i.e. $Z \perp \theta|W$.*

Why do we invoke this assumption, and not a marginal independence condition? Assumption 5 would be satisfied if (Z, W) were independent of θ. In this case, all variables would be part of Z and there would be no W. However, many structural models imply that some regressors are likely to be correlated with unobserved heterogeneity. We illustrate this point through two important economic examples. We also illustrate the impact of Assumptions 3 and 5.

2.2 Examples

Example 1 (Linear endogenous random coefficient model)**.** *Let X measure log-expenditure on a set of goods and let C^* measure the true log-expenditure share for one good. Assume that C the observed log-expenditure share is measured with error and that $C = C^* + \eta$. Assume that the true outcome C^* is generated by a linear random coefficients model with*

$$C^* = \theta_0 + \theta_1 Z_1 + \theta_2 X$$

where Z_1 is the log-price and $\theta = (\theta_0, \theta_1, \theta_2)$ is a vector of random parameters. Since X is also a choice variable chosen by the same consumers, it is likely to be endogenous. To deal with this complication, we introduce instruments in a control function fashion: Let Z_2 measure log-income and suppose that $X = g(Z_2, W)$, where g is a nonparametric function that is strictly monotonic in W and W is the percentile of X conditional on Z_2. Moreover, let $Z = (Z_1, Z_2)$ and assume that $Z \perp (\theta, \eta, W)$. Impose the normalization $W \sim \mathcal{U}[0,1]$, and assume (for the moment) that $\eta | \theta W Z \sim \mathcal{N}[0,1]$.

Substituting all elements into the outcome equation, we obtain

$$C = \theta_1 Z_1 + \theta_2 g(Z_2, W) + \varepsilon, \tag{2.2}$$

where $\varepsilon = \eta + \theta_0$ and $\varepsilon | \theta W Z \sim \mathcal{N}[\theta_0, 1]$. We are interested in recovering the density of θ conditional on W. The example fits precisely into our framework since g can be treated as known; one can plug in a non-parametric estimate of g obtained from a first-stage analysis.[4] *Moreover, $Z \perp (\theta, \eta, W)$ implies that $Z \perp \theta | W$. This illustrates that our framework allows for a control function approach to endogeneity.*

To understand the flexibility made possible by Assumption 3, note that the above arguments remain valid if the density of ε depends on the entire vector θ or on (W, Z). On top of this, one could allow the measurement error component to be a mixture of normals with mixing weights that vary across the population. Alternatively, if a validation sample is available, a nonparametric pilot estimate of $f_{\eta|ZW}$ could be plugged in. Finally, if there is no measurement error, a modified version of our approach is detailed in Section 3.3.

Example 2 (Intertemporal consumption model)**.** *Consider the constant absolute risk aversion (CARA) intertemporal utility maximization problem with finite horizon T, constant interest rate r and random parameters θ_1 and θ_2 capturing heterogeneity in utility and subjective beliefs respectively. Define $R = (1 + r)$. Let A_t be a consumer's beginning-of-period assets after having*

[4] When a plug-in estimate of g is used, standard errors for our estimator can be adjusted using standard methods for plug-in estimators.

received all interest payments and let Y_t be his/her income. Suppose income follows a random walk. Let $S_t = (A_t, Y_t)$ be the state vector and let $v_t(S_t, \theta)$ be the value function for a consumer of type $\theta = (\theta_1, \theta_2)$ at date t. Let the terminal value function be $v_{T+1}(S_{T+1}, \theta) = -\frac{e^{\gamma A_{t+1}}}{\gamma}$ and let $\theta_1 = (\gamma, \beta)$ where γ is the coefficient of risk aversion and β is the discount factor. At each date $t \leq T$, a consumer's value function is defined by

$$v_t(S_t, \theta) = \max_{\{C_t^*\}} \left\{ \begin{array}{c} -\frac{e^{-\gamma C_t^*}}{\gamma} + \beta \mathbb{E}\left[v_{t+1}(S_{t+1}, \theta) \,|\, I_t(\theta_2)\right] \\ \text{subject to} \\ A_{t+1} = R(A_t + Y_t - C_t^*) \\ Y_{t+1} = Y_t + \eta_{t+1} \end{array} \right\}$$

where C_t^* is consumption and $\eta_t \sim N(0, \sigma_\eta^2)$. Here, the parameters are $\theta_1 = (\gamma, \beta)$ and $\theta_2 = \sigma_\eta^2$. At time t, a consumer's information set $I_t(\theta_2)$ consists of $\{\eta_s\}$ for all $s \leq t$. Suppose observed consumption C_t equals actual consumption C_t^* plus measurement error so that $C_t = C_t^* + \varepsilon_t$. Let $W_t = (A_t, Y_{t-1})$ and $Z_t = Y_t - Y_{t-1}$. Then this example fits precisely into our framework. In terms of the variables (C_t, W_t, Z_t), the Euler equation is

$$e^{-\gamma(C_t - \varepsilon_t)} - \beta \mathbb{E}\left[\partial_A v_{t+1}\left(R\left(W_t^1 + W_t^2 + Z_t - C_t + \varepsilon_t, \theta\right), W_t^2 + Z_t\right) | I_t(\theta_2)\right] = 0$$

where $W_t^1 = A_t$ and $W_t^2 = Y_{t-1}$. In particular, under the assumptions stated, the consumption function (with measurement error) takes the form

$$C_t = \phi_{1t} W_t^1 + \phi_{2t}\left(W_t^2 + Z_t\right) + m_t(\gamma, \beta, \theta_2) + \varepsilon_t \tag{2.3}$$

with

$$m_t(\gamma, \beta, \theta_2) = \phi_{3t} + \phi_{4t}\gamma + \phi_{5t}(\theta_2)\frac{\ln \beta}{\gamma}.$$

The vector $\phi_t = (\phi_{1t}, \phi_{2t}, \phi_{3t}, \phi_{4t}, \phi_{5t})$ consists of parameters that depend only on R, t and θ_2 (see, e.g., Caballero (1990)). The vector $\theta = (\theta_1, \theta_2) = (\gamma, \beta, \sigma_\eta^2)$ is assumed to be a time-invariant random coefficient vector, heterogeneously distributed in the population. We assume that the income process $(Y_t)_{t=1,..,T} \perp \theta_1$ and that $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$.[5]

Because θ is time invariant and determines both past and current consumption and savings decisions, it will be correlated with W_t. Obviously, as in Example 1, we could allow the density of ε_t to depend on (θ, W, Z).

With respect to Assumption 5, note that Y_t cannot be used as an exogenous variable. While it is assumed to be marginally independent of θ, it cannot be independent of θ conditional

[5] More generally, the consumption function $C = \varphi(W, Z, \theta, \varepsilon)$ could be computed numerically. For such an application, see Hoderlein et al. (2012).

on A_t. Observe, however, that by assumption, $Y_t - Y_{t-1}$ is independent of the entire history. Consequently, with the choice $W_t = (A_t, Y_{t-1})$ and $Z_t = Y_t - Y_{t-1}$, we obtain $\theta \perp Z_t | W_t$. Note also that the parameters (γ, β) enter (2.3) only through the single index $\delta = m_t(\gamma, \beta, \theta_2)$. As a result, the joint density of (γ, β) is not identified. Rather, the density of δ is identified. We discuss this point in more detail in Section 3.2.

Example 3 (Auction Model)**.** *Consider the independent symmetric private value* first-price sealed-bid auction model *with risk averse bidders considered in Campo et al. (2011). Let $I \geq 2$ denote the number of potential bidders and $\{v_i\}_{i=1,..,I}$ be the bidders' private values which are drawn independently from an unknown cumulative distribution function F. This distribution may depend on observed characteristics Z of the auctioned objects, and hence we write $F(\cdot|Z, I)$ for the CDF. We assume that $F(\cdot|Z, I)$ is differentiable with density $f(\cdot|Z, I)$ on a compact support $[\underline{v}(I), \bar{v}(I)] \subset \mathbb{R}_+$. Let $U(\cdot)$ be a bidder's von Neuman Morgenstern utility function with $U(0) = 0$, $U'(\cdot) > 0$ and $U''(\cdot) \leq 0$ because of potential risk aversion. Denote by $s(\cdot) \equiv s(\cdot; U, F, I)$ the Bayesian Nash equilibrium bidding strategy. From equation (1) in Campo et al. (2011), a bidder i's optimal bid $b_i = s(v_i)$ solves the following differential equation:*

$$s'(v_i) = (I-1)\frac{f(v_i|Z, I)}{F(v_i|Z, I)}\lambda(v_i - b_i) \tag{2.4}$$

for all $v_i \in [\underline{v}(I), \bar{v}(I)]$, where $\lambda(\cdot) = U(\cdot)/U'(\cdot)$. The boundary condition is $s(\underline{v}(I)) = \underline{v}(I)$.

Assume that $U(\cdot)$ is of CRRA type with $U(x) = x^{1-\theta_1}$ for $0 \leq \theta_1 < 1$. Then, $\lambda(v - b) = (1 - \theta_1)^{-1}(v - b)$ and using (2.4) bidder i's optimal bid is

$$b_i = v_i - [F(v_i|Z, I)]^{-\frac{I-1}{1-\theta_1}} \int_{\underline{v}}^{v_i} [F(t|Z, I)]^{\frac{I-1}{1-\theta_1}} dt. \tag{2.5}$$

In this example, b and v play the roles of C and ε, respectively, and the function on the right hand side of (2.5) plays the role of the function φ in Assumption 1. By assuming a parametric form for the pdf of v given (Z, I), we can recover the pdf of b conditional on θ_1 and eventually on Z. We remark that in this example the parameter θ_1 is assumed to be heterogeneous across different auctions but it is the same for bidders taking part in a given auction (as it is required in order to have the equilibrium). Finally, note that we could let the pdf of v_i depend on additional parameters θ that could be heterogeneous across the population of auctions.

This list of examples serves to illustrate the generality of our framework. The list could be greatly extended. For instance, one could apply this approach to study the original Heckman and Singer (1984) work on duration, or one could apply our framework to structural labor models of the form studied in Keane and Wolpin (1997). Instead of elaborating on the details, we leave the application of this framework to future research, and complete the formal definition

of our model.

2.3 A Hilbert-space setting

The natural space for probability density functions is the L^1 space with respect to the Lebesgue measure endowed with either the L^1- or the Hellinger- metric. Despite this fact, to exploit desirable properties of Hilbert spaces, we develop our analysis in L^2 spaces with respect to some suitable measures.

For this purpose, we introduce two non-negative weighting functions on Θ and $\mathcal{C} \times \mathbb{R}^l$ that we denote by π_θ and π_{cz}, respectively. Define the space $L^2_{\pi_\theta}(\Theta)$ (resp. $L^2_{\pi_{cz}}(\mathcal{C} \times \mathbb{R}^l)$) of real-valued functions defined on Θ (resp. on $\mathcal{C} \times \mathbb{R}^l$), and indexed by the random variable W, which are P_W-a.s. square integrable with respect to π_θ (resp. π_{cz}), that is,

$$\begin{aligned} L^2_{\pi_\theta}(\Theta) &= \left\{ h(\cdot; W) : \Theta \to \mathbb{R} \;\middle|\; \int_\Theta h^2(\theta; W)\pi_\theta(\theta)d\theta < \infty, \;\; P_W - a.s. \right\}, \\ L^2_{\pi_{cz}}(\mathcal{C} \times \mathbb{R}^l) &= \left\{ \psi(\cdot, \cdot; W) : \mathcal{C} \times \mathbb{R}^l \to \mathbb{R} \;\middle|\; \int_{\mathcal{C}} \int_{\mathbb{R}^l} \psi^2(c, z; W)\pi_{cz}(c, z)dcdz < \infty, \;\; P_W - a.s. \right\}. \end{aligned}$$

For brevity, we denote $L^2_{\pi_\theta}(\Theta)$ by $L^2_{\pi_\theta}$ and $L^2_{\pi_{cz}}(\mathcal{C} \times \mathbb{R}^l)$ by $L^2_{\pi_{cz}}$. Further, we denote the scalar product by $< \cdot, \cdot >$ and the induced norm by $|| \cdot ||$ in both $L^2_{\pi_\theta}$ and $L^2_{\pi_{cz}}$ without distinction. That is $\forall h_1, h_2 \in L^2_{\pi_\theta}$, $< h_1, h_2 >= \int h_1(\theta; W)h_2(\theta; W)\pi_\theta(\theta)d\theta$ and $\forall \psi_1, \psi_2 \in L^2_{\pi_{cz}}$, $< \psi_1, \psi_2 >= \int \psi_1(c, z; W)\psi_2(c, z; W)\pi_{cz}(c, z)d\theta$.

Since our analysis is conditional on W, we allow the weighting functions π_θ and π_{cz} to be indexed by W too. The sets of conditional probability density functions relevant for our analysis are denoted and defined as follows

$$\begin{aligned} \mathcal{F}_{\theta|W} &:= \left\{ f \in L^2_{\pi_\theta} \;\middle|\; f \text{ is a conditional } \textit{pdf} \text{ on } (\mathbb{R}^d, \mathcal{B}_\Theta) \text{ given } W \right\} \\ \mathcal{F}_{C|WZ} &:= \left\{ f \in L^2_{\pi_{cz}} \;\middle|\; f \text{ is a conditional } \textit{pdf} \text{ on } (\mathbb{R}, \mathcal{B}_\mathcal{C}) \text{ given } (Z, W) \right\} \\ \mathcal{F}_{C|WZ\theta} &:= \left\{ f \;\middle|\; f \text{ is a conditional } \textit{pdf} \text{ on } (\mathbb{R}, \mathcal{B}_\mathcal{C}) \text{ given } (W, Z, \theta) \right\}. \end{aligned}$$

While $\mathcal{F}_{\theta|W} \subset L^2_{\pi_\theta}$ and $\mathcal{F}_{C|WZ} \subset L^2_{\pi_{cz}}$, we do not assume that $\mathcal{F}_{C|WZ\theta} \subset L^2_{\pi_{cz} \times \pi_\theta}$ where

$$L^2_{\pi_{cz} \times \pi_\theta} := \left\{ h(\cdot, \cdot, \cdot; W) : \mathcal{C} \times \mathbb{R}^l \times \Theta \to \mathbb{R} \;\middle|\; \int_{\mathcal{C}} \int_{\mathbb{R}^l} \int_\Theta h^2(c, z, \theta; W)\pi_\theta(\theta)\pi_{cz}(c, z)d\theta dcdz < \infty, \;\; P_W - a.s. \right\}.$$

When this last condition is satisfied, that is the *pdf* of C conditional on (W, Z, θ) is square integrable, we can provide a simple characterization of the identified set. However, this condition is only sufficient and not necessary as we explain in Section 3.1 below. In the following we assume that $f_{\theta|W} \in \mathcal{F}_{\theta|W}$.

3 Identification of the distribution of parameters

In this section we use the structural relationship given in Assumption 1 to characterize the direct mapping from $f_{\theta|W}$, the *pdf* of unobservables θ conditional on W, to $f_{C|WZ}$, the conditional *pdf* of C given (W, Z), i.e., the mapping from the distribution of unobservables to the distribution of observables. We also characterize the inverse mapping from the observables to the distribution of unobservables. The economic model defines both mappings. The econometric problem is to analyze the inverse mapping.

3.1 Linear integral equation and non-linear inverse problem

Given the direct mapping described above, the econometrician is interested in the *inverse problem* of recovering the conditional *pdf* $f_{\theta|W}$ of θ given W (*i.e.* the cause) from the observed phenomenon. The following theorem characterizes both the structural mapping as an operator equation and the inverse problem as a convexly constrained inverse of the same operator equation.

Theorem 1. *Let Assumptions 1-5 be satisfied. Then*

$$f_{C|WZ} = T f_{\theta|W}, \qquad P_W - a.s.$$

where $\forall h \in L^2_{\pi_\theta}$,

$$Th = \int_\Theta \sum_{i=1}^{s} (f_{\varepsilon|WZ\theta} \circ \varphi_i^{-1})(c, w, z, \theta) \cdot \left| \frac{\partial_c \Psi(c, w, z, \theta, \varphi_i^{-1}(w, z, \theta, c))}{\partial_\varepsilon \Psi(c, w, z, \theta, \varphi_i^{-1}(w, z, \theta, c))} \right| 1_{\mathcal{C}_i}(c) h(\theta; w) d\theta, \quad (3.1)$$

$c = \varphi(w, z, \theta, \varepsilon)$ *is the a.s.-explicit solution of* $\Psi(c, w, z, \theta, \varepsilon) = 0$ *and* $\mathcal{C}_i = \{c \in \text{Im}(\mathcal{E}_i) \cap \mathcal{C}_\theta\}$. *Here* $\text{Im}(\mathcal{E}_i)$ *is the image of* $\mathcal{E}_i$ *through* φ *and* $\mathcal{C}_\theta$ *is the support of* $f_{C|WZ\theta}$. *This implies that* $f_{\theta|W}$ *is a solution of*

$$f_{C|WZ} = T f_{\theta|W} \quad \textit{subject to} \quad f_{\theta|W} \in \mathcal{F}_{\theta|W}, \quad P_W - a.s. \quad (3.2)$$

Note that although the operator T depends on W, we use the short-hand notation T and leave implicit the dependence on W.

The operator T in equation (3.2) is a mixing operator and the theorem characterizes the object of interest $f_{\theta|W}$ as the solution of a *convexly constrained Fredholm integral equation of the first kind*. Equation (3.2) states that $f_{C|WZ}$ is a P_W-a.s. $f_{\theta|W}$-mixture of $f_{C|WZ\theta} \in \mathcal{F}_{C|WZ\theta}$

where

$$f_{C|WZ\theta} = \sum_{i=1}^{s} (f_{\varepsilon|WZ\theta} \circ \varphi_i^{-1})(c, w, z, \theta) \cdot \left| \frac{\partial_c \Psi(c, w, z, \theta, \varphi_i^{-1}(w, z, \theta, c))}{\partial_\varepsilon \Psi(c, w, z, \theta, \varphi_i^{-1}(w, z, \theta, c))} \right| 1_{\mathcal{C}_\theta}(c).$$

In other words, this means that $f_{C|WZ}$ belongs P_W-a.s. to the convex hull $\mathbb{M}_\Theta$ of $\mathcal{F}_{C|WZ\theta}$:

$$f_{C|WZ} \in \mathbb{M}_\Theta := \left\{ h \;\middle|\; h = \int_\Theta f_{C|WZ\theta}(\cdot; w, \cdot, \theta) f_{\theta|W} d\theta;\ f_{C|WZ\theta} \in \mathcal{F}_{C|WZ\theta},\ f_{\theta|W} \in \mathcal{F}_{\theta|W} \right\}$$

where $f_{C|WZ\theta}$ denotes the *pdf* associated with $P_{C|WZ\theta}$ and $P_{C|WZ\theta}$ is the conditional probability on $\mathcal{B}_\mathcal{C}$ conditional on (W, Z, θ).

Recovering $f_{\theta|W}$ from (3.2) is an ill-posed inverse problem. The main contribution of the theorem is the characterization (3.1) of the operator T in terms of the structural quantities of the economic problem.

Throughout this paper, we work with compact operators on Hilbert spaces, because they have many similarities with linear operators on finite dimensional spaces. On top of that, they have appealing spectral properties and can be approximated by operators with finite-dimensional range that are norm convergent which is useful for estimation. To this end, we assume that π_θ and π_{cz} are suitably chosen so that T is bounded and compact and $\mathcal{R}(T) \subset L^2_{\pi_{cz}}$, where $\mathcal{R}(\cdot)$ denotes the range of an operator. See the discussion in the simulations below. Our assumptions imply that $f_{C|WZ} \in \mathcal{F}_{C|WZ} \subset L^2_{\pi_{cz}}$. A sufficient condition for $T : L^2_{\pi_\theta} \to L^2_{\pi_{cz}}$ and T being bounded and compact is that the *kernel* of the operator T, $\frac{f_{C|WZ\theta}}{\pi_\theta}$, is square integrable with respect to $\pi_{cz} \times \pi_\theta$, that is $\frac{f_{C|WZ\theta}}{\pi_\theta} \in L^2_{\pi_{cz} \times \pi_\theta}$.[6]

In practice, the econometrician specifies $f_{\varepsilon|WZ\theta}$ and π_θ instead of $f_{C|WZ\theta}$. Therefore, it is useful to give sufficient conditions for compactness and boundedness of the operator T in terms of $f_{\varepsilon|WZ\theta}$ and π_θ. To that end we introduce Assumption 6.

Assumption 6. *Let* $s^{\frac{1}{2}} f_{\varepsilon|WZ\theta} |\partial_c \Psi / \partial_\varepsilon \Psi|^{1/2} \Big|_{c=\varphi(w,z,\theta,\varepsilon)}$ *be square integrable in* (ε, Z, θ) *with respect to* $\frac{\pi_{cz}}{\pi_\theta} \Big|_{c=\varphi(w,z,\theta,\varepsilon)}$, P_W*-a.s.*

With this assumption, we can prove the next proposition.

[6] Remark that $\mathcal{R}(T) \subset L^2_{\pi_{cz}}$ if and only if $\forall h \in L^2_{\pi_\theta}$, $||Th|| < \infty$. By using the Cauchy-Schwarz inequality we obtain

$$\begin{aligned} ||Th||^2 &= \int_\mathcal{C} \int_{\mathbb{R}^l} \left\langle \frac{f_{C|WZ\theta}}{\pi_\theta}, h \right\rangle^2 \pi_{cz}(c, z) dc dz \leq \int_\mathcal{C} \int_{\mathbb{R}^l} \left|\left| \frac{f_{C|WZ\theta}}{\pi_\theta} \right|\right|^2 ||h||^2 \pi_{cz}(c, z) dc dz \\ &= ||h||^2 \int_\mathcal{C} \int_{\mathbb{R}^l} \int_\Theta \frac{f^2_{C|WZ\theta}}{\pi_\theta} \pi_{cz} d\theta dc dz < \infty. \end{aligned} \tag{3.3}$$

Proposition 1. *Let T be the operator defined in (3.1) with domain $L^2_{\pi_\theta}$ and let Assumptions 1 - 6 be satisfied. Then T is a P_W-a.s. bounded and compact operator with range included in $L^2_{\pi_{cz}}$.*

This proposition shows that compactness of T depends both on the structural model (characterized by the structural function Ψ and the density $f_{\varepsilon|WZ\theta}$) and on the weights π_{cz} and π_θ. With this additional assumption, the adjoint operator is bounded and linear, as the following proposition shows:

Proposition 2 (Adjoint of T)**.** *Let $T : L^2_{\pi_\theta} \to L^2_{\pi_{cz}}$ be the operator defined in (3.1) and $f_{C|WZ\theta}$ be the pdf associated with $P_{C|WZ\theta}$. The operator T^* defined as: $\forall \psi \in L^2_{\pi_{cz}}$,*

$$T^*\psi = \int_{\mathcal{C}} \int_{\mathbb{R}^l} f_{C|WZ\theta}(c; w, z, \theta)\psi(c, z; w)\frac{\pi_{cz}(c, z)}{\pi_\theta(\theta)} dcdz,$$

is the adjoint of T. The operator $T^* : L^2_{\pi_{cz}} \to L^2_{\pi_\theta}$ is bounded and linear.

A compact operator admits a singular value decomposition, which is briefly reviewed for completeness:

Remark 2 (Singular value decomposition - SVD)**.** When the operator T is compact with infinite-dimensional range, that is, its kernel is not degenerate, then T^*T is characterized by a countable number of eigenvalues which accumulate only at zero. Moreover, it admits the following *singular value decomposition*:

$$T\varphi_j = \lambda_j \psi_j, \qquad T^*\psi_j = \lambda_j \varphi_j, \quad j \in \mathbb{N}$$

where $\{\lambda_j\}_{j\in\mathbb{N}}$ and $\{\varphi_j, \psi_j\}_{j\in\mathbb{N}}$ are the sequences of singular values and singular functions, respectively. The singular values are the nonnegative square roots of the eigenvalues of T^*T (and also of TT^*). The set of functions $\{\varphi_j\}_{j\in\mathbb{N}}$ (resp. $\{\psi_j\}_{j\in\mathbb{N}}$) is a complete orthonormal system of eigenfunctions of T^*T (resp. of TT^*) which spans $\overline{\mathcal{R}(T^*)} = \overline{\mathcal{R}(T^*T)}$ (resp. $\overline{\mathcal{R}(T)} = \overline{\mathcal{R}(TT^*)}$) where $\overline{\mathcal{R}(T^*)}$ is the closure of the range of the operator T^*.

From now on we denote by $T|_{\mathcal{F}_{\theta|W}}$ the operator T restricted to $\mathcal{F}_{\theta|W}$. Thus, if $\mathcal{D}(\cdot)$ denotes the domain of an operator, we have $\mathcal{D}(T|_{\mathcal{F}_{\theta|W}}) = \mathcal{D}(T) \cap \mathcal{F}_{\theta|W} = \mathcal{F}_{\theta|W}$ and $\mathcal{R}(T|_{\mathcal{F}_{\theta|W}}) \subset \mathcal{F}_{C|WZ} \subset \mathcal{R}(T) \subset L^2_{\pi_{cz}}$. Following Example 2.4 in Carrasco et al. (2007), the adjoint $T|^*_{\mathcal{F}_{\theta|W}}$ of $T|_{\mathcal{F}_{\theta|W}}$ is given by $T|^*_{\mathcal{F}_{\theta|W}} = \mathcal{P}_c T^*$ where $\mathcal{P}_c$ denotes the metric projector onto $\mathcal{F}_{\theta|W}$. Notice that $\mathcal{F}_{\theta|W}$ is a convex set and hence the operator $T|_{\mathcal{F}_{\theta|W}}$ is an *affine* operator, *i.e.* $T|_{\mathcal{F}_{\theta|W}} : \mathcal{F}_{\theta|W} \to L^2_{\pi_{cz}}$ satisfies: $T|_{\mathcal{F}_{\theta|W}}((1-\lambda)f_1 + \lambda f_2) = (1-\lambda)T|_{\mathcal{F}_{\theta|W}}(f_1) + \lambda T|_{\mathcal{F}_{\theta|W}}(f_2)$ whenever $f_1, f_2 \in \mathcal{F}_{\theta|W}$, $0 < \lambda < 1$.

We introduce the following set that will be useful for discussing identification

$$\mathfrak{D} = \{f_1 - f_2 \;\big|\; \forall f_1, f_2 \in \mathcal{F}_{\theta|W}\} = \mathcal{F}_{\theta|W} \oplus \check{\mathcal{F}}_{\theta|W}$$

where $\check{\mathcal{F}}_{\theta|W} = \{-f;\ f \in \mathcal{F}_{\theta|W}\}$ and $\oplus$ denotes the Minkowski sum. The set $\mathfrak{D}$ is a convex set and we denote by $T|_{\mathfrak{D}} : \mathfrak{D} \to L^2_{\pi_{cz}}$ the operator T restricted to $\mathfrak{D}$. This is an affine operator.

Remark 3 (Nonlinear inverse problem)**.** Despite the linearity of T, the inverse problem in (3.2) is potentially non-linear because of the constraint $f_{\theta|W} \in \mathcal{F}_{\theta|W}$. By nonlinear inverse problem we mean that the $\mathcal{F}_{\theta|W}$-constrained pseudoinverse of T is a nonlinear operator. As we explain in more detail in section 4.3, this operator defines the constrained-best-approximate solution of (3.2).

Remark 4 (Existence of a solution)**.** If the model specification is correct, then the existence of at least one solution to (3.2) is guaranteed since $f_{C|WZ} \in \mathcal{R}(T|_{\mathcal{F}_{\theta|W}})$. In the terminology of inverse problems, this means that $f_{C|WZ}$ is "attainable".

3.2 The identified set

In this section, we discuss point and set identification of $f_{\theta|W}$. The distribution $f_{\theta|W} \in \mathcal{F}_{\theta|W}$ will be called *identified* (with respect to the class $\mathcal{F}_{\theta|W}$) if

$$T|_{\mathcal{F}_{\theta|W}}(f_{\theta|W}) = T|_{\mathcal{F}_{\theta|W}}(\tilde{f}_{\theta|W}) \quad \Rightarrow \quad f_{\theta|W} = \tilde{f}_{\theta|W}, \tag{3.4}$$

for all $\tilde{f}_{\theta|W} \in \mathcal{F}_{\theta|W}$.

Therefore, point identification of $f_{\theta|W}$ is equivalent to requiring that the operator $T|_{\mathfrak{D}}$ is injective. In fact, (3.4) is equivalent to "$T|_{\mathfrak{D}}(f_{\theta|W} - \tilde{f}_{\theta|W}) = 0$ implies $f_{\theta|W} = \tilde{f}_{\theta|W}$". The injectivity of $T|_{\mathfrak{D}}$ depends on the injectivity of T but it is not equivalent. In fact, if T is injective, that is, $\mathcal{N}(T) = \{0\}$ where $\mathcal{N}(\cdot)$ denotes the null space of an operator, then $T|_{\mathfrak{D}}$ is injective as well. However, when T is non-injective the restricted operator $T|_{\mathfrak{D}}$ may be injective. This is possible when the domain of $T|_{\mathfrak{D}}$ is sufficiently restricted.

The following proposition characterizes the identified set for the operator T, denoted by Λ. Using the notation from Remark 2, we denote the eigenvalues and eigenfunctions of T^*T by $\{\lambda_j, \varphi_j\}_{j\in\mathbb{N}}$. We denote by $f^{\dagger}_{\theta|W}$ the solution of the linear inverse problem $f_{C|WZ} = Tf_{\theta|W}$ which has minimal norm, and $J_0 = \{j\ \lambda_j = 0\}$. The following proposition is then immediate from the previous discussion:

Proposition 3. *The identified set is the set of all the solutions of (3.2):*

$$\begin{aligned}\Lambda &= \left\{h \in \mathcal{F}_{\theta|W} \;\middle|\; f_{C|WZ} = Th,\ P_W - a.s.\right\} \\ &= \left\{f^{\dagger}_{\theta|W} \oplus \mathcal{N}(T)\right\} \cap \mathcal{F}_{\theta|W}.\end{aligned}$$

If T is compact then

$$\Lambda = \left\{ h \in \mathcal{F}_{\theta|W} \;\middle|\; \begin{array}{c} h = f^{\dagger}_{\theta|W} + \sum_{\{j\in J_0\}} \zeta_j \varphi_j \text{ for } \{\zeta_j\} \\ \text{satisfying} \\ \sum_{j\in J_0} \zeta_j \int_{\Theta} \varphi_j d\theta = 1 - \kappa_w \ \wedge \ \sum_{\{j\in J_0\}} \zeta_j \varphi_j \geq -f^{\dagger}_{\theta|W},\ P_W - a.s. \end{array} \right\}$$

where $\kappa_w = \int_{\Theta} f^{\dagger}_{\theta|W}(\theta; w) d\theta$.

This proposition characterizes the identified set in terms of quantities that depend on the SVD of T, which is known, and on $f_{C|WZ}$ which can be easily estimated.

The model is point-identified when Λ is a singleton. This occurs in two cases:

(i) the operator T is injective, *i.e.* $\mathcal{N}(T) = \{0\}$. Then, $f^{\dagger}_{\theta|W} \in \mathcal{F}_{\theta|W}$ and is the unique solution of (3.2);

(ii) the operator T is not injective, *i.e.* $\mathcal{N}(T) \neq \{0\}$, but $T|_{\mathfrak{D}}$ is injective, *i.e.* (3.4) holds. In this case we have $\Lambda = (f^{\dagger}_{\theta|W} + h_{\theta|W})$ where $h_{\theta|W} \in \mathcal{N}(T)$ is such that $\int_{\Theta}(f^{\dagger}_{\theta|W} + h_{\theta|W})(\theta; W) d\theta = 1$ and $(f^{\dagger}_{\theta|W} + h_{\theta|W})$ is non-negative *a.e.* on Θ, P_W-a.s. In this case we can also have $\Lambda = f^{\dagger}_{\theta|W}$ if $f^{\dagger}_{\theta|W}$ is a probability density function.

The injectivity condition of T characterizes the strength of statistical dependence between C and θ conditionally on (W, Z). However, identification can be obtained even without injectivity of T. This shows that identification in our framework is a weaker concept than in the nonparametric instrumental variable (IV) literature. In fact, identification in nonparametric IV models is guaranteed only when the operator characterizing the estimating equation is injective, which corresponds to our case (i). In contrast, we may also achieve identification in case (ii), and this is due to the presence of the constraint in (3.2). This constraint makes the estimation problem more difficult because the estimation problem is nonlinear when the constraint is binding. On the other hand, it can help to shrink the size of the identified set.

Remark 5. If C, $Z = (Z_1, \ldots, Z_l)$ and θ are discretely distributed with finite supports $\{c_1, \ldots, c_s\}$, $\{z_{1j}, \ldots, z_{t^j j}\}$, for $j = 1, \ldots, l$ and $\{\theta_1, \ldots, \theta_m\}$, respectively, then a necessary condition for identification is that $sK \geq m$, where $K = \sum_{j=1}^{l} t^j$. This means that the number of support points in each dimension can differ as long as $sK \geq m$. If m is very large and the

supports of C and Z contain few points then we need to increase the dimension of Z compared to the dimension of θ, see also Newey and Powell (2003). This logic extends to the countably infinite case, and we hence conjecture that having a vector (C, Z) with large dimension relative to the dimension of θ is in general necessary to achieve identification.

Remark 6. The injectivity of the operator T is equivalent to the requirement that the conditional *pdf* $f_{\theta|CWZ}$ of θ given (C, W, Z) is $L^2_{\pi_\theta}$-complete, provided that $f_{C|WZ}$ and $f_{\theta|W}$ are bounded away from zero and infinity. The following argument shows this fact. The operator T is injective if and only if the only solution of $0 = \int_\Theta f_{C|WZ\theta}(c; w, z, \theta)h(\theta; w)d\theta$ in $L^2_{\pi_\theta}$ is $h(\theta; w) = 0$, P_W-a.s. Note that we can rewrite $0 = \int_\Theta f_{C|WZ\theta}(c; w, z, \theta)h(\theta; w)d\theta = \int_\Theta f_{\theta|CWZ}(\theta; c, w, z)\frac{f_{C|WZ}(c;w,z)}{f_{\theta|W}(\theta;w)}h(\theta; w)d\theta$. Now, under the assumptions of Theorem 1, $0 < m_\varepsilon\kappa \leq f_{C|WZ} < \infty$, for some constant κ. This is because by Assumption 2: $\partial_c\Psi \neq 0$ and $\partial_\varepsilon\Psi \neq 0$, and by Assumption 3: $m_\varepsilon \leq f_{\varepsilon|WZ\theta} < \infty$. Furthermore, Assumption 4 guarantees that $0 < m_\theta \leq f_{\theta|W} < \infty$. If we then assume that $f_{\theta|CWZ}$ is $L^2_{\pi_\theta}$-complete, then by definition of completeness, this implies that

$$\frac{f_{C|WZ}(c; w, z)}{f_{\theta|W}(\theta; w)}h(\theta; w) = 0, \quad P_W - a.s.$$

which in turn implies that $h(\theta; w) = 0$, P_W-a.s. since $0 < m_\varepsilon k \leq f_{C|WZ}(c; w, z) < \infty$ and $0 < m_\theta \leq f_{\theta|W}(\theta; w) < \infty$. On the other side, assume that T is injective, then $0 = Th$ implies $h(\theta; w) = 0, P_W$-a.s. Since $f_{C|WZ}$ and $f_{\theta|W}$ are bounded away from zero and infinity, $h(\theta; w) = 0, P_W$-a.s. is equivalent to $\frac{f_{C|WZ}(c;w,z)}{f_{\theta|W}(\theta;w)}h(\theta; w) = 0$, P_W-a.s. Thus, $f_{\theta|CWZ}$ is $L^2_{\pi_\theta}$-complete, and conditions that are sufficient for completeness ensure identification. The equivalence between L^2-completeness and injectivity was already noted in Florens et al. (1990), Florens (2003), Newey and Powell (2003) and Hu and Schennach (2008) in different setups. We would like to point out that in mixture models, like the one that underlies our framework or Hu and Schennach (2008), unlike the nonparametric IV literature, $L^2_{\pi_\theta}$-completeness does not refer to the *pdf* $f_{C|WZ\theta}$ which characterizes the kernel of the integral operator T. In fact, these approaches differ, as in our setting T is not a conditional expectation operator.[7]

However, in our framework, $L^2_{\pi_\theta}$-completeness is too strong for identification, since the solution to the integral equation is also constrained to be a *pdf*. In our case, identification is thus equivalent to a weaker concept of completeness, that we call $\mathcal{T}$-completeness of $f_{\theta|CWZ}$.

[7] There is one noticeable exception to this rule: the case without Z. To see this, first note that T is injective if and only if T^* is injective. The operator T^* is then defined as follows: $\forall\phi \in L^2_{\pi_c} \mapsto T^*\phi = \int_{\mathcal{C}} f_{C|W\theta}(c; w, \theta)\phi(c; w)\pi_c(c)dc\frac{1}{\pi_\theta} = \mathbb{E}(\phi\pi_c|W, \theta)\frac{1}{\pi_\theta}$. If $0 < \pi_\theta < \infty$ and $0 < \pi_c < \infty$, P_W-a.s., then $T^*\phi = 0$ is equivalent to $\mathbb{E}(\phi|W, \theta) = 0$. This shows that, in the less interesting case without Z, $L^2_{\pi_c}$-completeness of $f_{C|W\theta}$ is equivalent to injectivity of T.

This condition will turn out to be a *necessary and sufficient* condition for identification in our framework, as the following proposition shows:

Proposition 4 ($\mathcal{T}$-completeness)**.** *Let $L : \mathfrak{D} \to L^2_{\pi_\theta}$ be the multiplication operator $(Lh)(\theta) = \frac{1}{f_{\theta|W}(\theta;w)} h(\theta;w)$ where $f_{\theta|W}$ denotes the true pdf and denote $\mathcal{T} = \mathcal{R}(L) \subset L^2_{\pi_\theta}$. Under the assumptions of theorem 1, (3.4) holds if and only if $f_{\theta|CWZ}$ is $\mathcal{T}$-*complete*, that is, $\forall h \in \mathcal{T}$, $\int_\Theta h f_{\theta|CWZ} d\theta = 0$, P_{CWZ}-a.s., implies $h = 0$, P_W-a.s.*

We refer to Mandelbaum and Rüschendorf (1987) for more background on *completeness* of a probability distribution with respect to a general family of functions $\mathcal{T}$; in this paper, we adapt this concept to our problem. In the following, we give some examples of classes of distribution functions $\mathcal{F}_{C|WZ\theta}$, for which the corresponding *pdf* of θ given (W, Z, θ) satisfy $L^2_{\pi_\theta}$-completeness, resp. $\mathcal{T}$-completeness, and hence allow for point identification. We start with the well known class of exponential distributions that is complete.

Exponential family of distributions.

Lemma 3.1. *Let us assume that $\forall i = 1, \ldots, s$, $\partial_c \varphi_i^{-1}(w, z, \cdot, c) 1_{\mathcal{C}_\theta}(c)$ is bounded away from zero and infinity for every $(c, w, z) \in \mathcal{C} \times \mathcal{W} \times \mathcal{Z}$, and $(f_{\varepsilon|\theta WZ} \circ \varphi_i^{-1})(c, w, z, \theta)$ is of the form*

$$\exp\{\tau_i(c, w, z)' m_i(\theta)\} h_i(\theta) k_i(c, w, z), \qquad i = 1, \ldots, s$$

where for every $i = 1, \ldots, s$, $h_i(\cdot)$ is a positive function depending only on θ, $m_i(\cdot)$ is a vector-valued function whose image has dimension equal to the dimension of θ and each component is an increasing function depending only on θ. The functions τ_i and k_i do not depend on θ and k_i is a positive and bounded function. Then, if the support of (C, Z, W) has a nonempty interior, $f_{\theta|W}$ is identified with respect to the class $\mathcal{F}_{\theta|W}$.

In addition to this large class of distributions, we now provide additional examples of families $\mathcal{F}_{C|WZ\theta}$ of conditional distributions for which the operator $T|_{\mathcal{F}_{\theta|W}}$ is injective, that is, the corresponding $f_{\theta|CWZ}$ is $\mathcal{T}$-complete[8]. It turns out that the corresponding densities $f_{\theta|CWZ}$ satisfy a notion of completeness stronger than $\mathcal{T}$-completeness: they are complete with respect to the class of functions $\{(h/f_{\theta|W})(\theta;w);\ h = h_1 - h_2, : \forall h_1, h_2 \in L^1(\Theta) \cap L^2_{\pi_\theta}(\Theta)\}$. This class of functions contains $\mathcal{T}$, but is in general smaller than $L^2_{\pi_\theta}$. As a consequence, the *pdf*s $f_{\theta|CWZ}$

[8]In fact, this is easy see: Let $\mathcal{F}_{C|WZ\theta}$ be any of the two Teicher's families, and let $f_{C|WZ\theta} \in \mathcal{F}_{C|WZ\theta}$. Then, $T|_{\mathfrak{D}} h = 0$ implies $h = 0$, where $h \in \mathfrak{D}$. Now, $T|_{\mathfrak{D}} h = 0$ is equivalent to $\int_\Theta (f_{\theta|WZC} \frac{f_{C|WZ}}{f_{\theta|W}})(\theta, c; w, z) h(\theta; w) d\theta = 0$ which in turn is equivalent to $\int_\Theta (f_{\theta|WZC} \frac{1}{f_{\theta|W}})(\theta; w, z, c) h(\theta; w) d\theta = 0$. Now, suppose that $f_{\theta|WZC}$ is not $\mathcal{T}$-complete. This implies that $\frac{1}{f_{\theta|W}(\theta;w)} h(\theta;w)$ may be different from 0, P_W-a.s. which in turns implies that $h(\theta;w)$ may be different from 0, P_W-a.s. (since $0 < m_\theta < f_{\theta|W} < \infty$), but this is a contradiction with the fact that $T|_{\mathfrak{D}}$ is injective.

implied by the families $\mathcal{F}_{C|WZ\theta}$ listed below are not in general $L^2_{\pi_\theta}$- complete, but $\mathcal{T}$-complete. The examples are confined to the single random coefficient case, but illustrative for multiple parameters.

ADDITIVELY-CLOSED ONE-PARAMETER FAMILY OF DISTRIBUTIONS. Let $\Theta = \mathbb{R}_+$ and $\mathcal{F}_{C|WZ\theta}$ be *additively closed*, *i.e.* $\forall f_{C|WZ\theta},\ h_{C|WZ\theta} \in \mathcal{F}_{C|WZ\theta}$ and $\forall \theta_1, \theta_2 \in \Theta$, $f_{C|WZ\theta}(c; w, z, \theta_1) * h_{C|WZ\theta}(c; w, z, \theta_2) = f_{C|WZ\theta}(c; w, z, \theta_1 + \theta_2)$, where $*$ denotes the convolution operation. Then, $f_{\theta|W}$ is identified. Additive-closedness of $\mathcal{F}_{C|WZ\theta}$ depends on the functional form of $f_{\varepsilon|WZ\theta}$ and of the structural function φ and can be easily checked. In particular, some distributions that belongs to the additively-closed one-parameter family, and that are relevant for our application, are the following, see Teicher (1961).

- Type III distributions: $f_{C|WZ\theta} = \frac{z^\theta}{\Gamma(\theta)} c^{\theta-1} e^{-zc}$, $c > 0$, $z > 0$, $\theta > 0$ or $f_{C|WZ\theta} = \frac{\theta^z}{\Gamma(z)} c^{z-1} e^{-\theta c}$, $c > 0$, $z > 0$, $\theta > 0$. The role of W and Z can be interchanged.
- Uniform distributions: $f_{C|WZ\theta} = \mathcal{U}[\theta - g(Z,W), \theta + g(Z,W)]$, where $g(\cdot,\cdot)$ is some function of (Z, W). Therefore, $f_{C|WZ\theta} = \frac{1}{2g(Z,W)} 1\{[\theta - g(Z,W)] < c < [\theta + g(Z,W)]\}$. However, for uniform distributions for which the support does not depend on θ we have no identification of $\mathcal{F}_{\theta|W}$.

LOCATION-SCALE ONE-PARAMETER FAMILY OF DISTRIBUTIONS. Let $\Theta = \mathbb{R}_+$ and $\mathcal{F}_{C|WZ\theta}$ be the one-parameter family induced by $f_{C|WZ}$ via location or scale changes, *i.e.* $\forall f_{C|WZ\theta} \in \mathcal{F}_{C|WZ\theta}$, $f_{C|WZ\theta}(c; w, z, \theta) = f_{C|WZ}(c - \theta; w, z)$ or $f_{C|WZ\theta}(c; w, z, \theta) = f_{C|WZ}(c\theta; w, z)$. For the location (resp. scale) family: if the conditional characteristic function of C (resp. $\log C$), given (W, Z), does not vanish P_{WZ}-a.s. in some non-degenerate real interval, then the $f_{\theta|W}$ is identified, see Teicher (1961).

These examples illustrate nicely the degree to which our prior knowledge about the structure of the problem and the space of possible solutions – in our case, random coefficient densities – helps us to understand identification. Note, however, that the distribution of $f_{C|WZ\theta}(c; w, z, \theta)$ in our setup is determined by the distribution of ε and the structural model φ. To illustrate the type of arguments that would have to be considered, we now analyze identification in the two previous examples introduced above:

Example 1 (Continued)**.** *Suppose that $\varepsilon \sim Exp(\lambda)$ with $\lambda > 0$ a known parameter. Therefore, the functional equation that identifies $f_{\theta|W}$ is*

$$f_{C|WZ}(c, w, z) = \int_\Theta \lambda \exp\{-\lambda(c - z_1'\theta_1 - \theta_2 g(z_2, w))\} \cdot 1\{c \geq z_1'\theta_1 + \theta_2 g(z_2, w)\} \cdot f_{\theta|W} d\theta$$

where $z := (z_1', z_2)'$. *The function* $(f_{\varepsilon|\theta WZ} \circ \varphi_i^{-1})(c, w, z, \theta)$ *which characterizes the kernel of the operator can be rewritten as:*

$$\begin{aligned}(f_{\varepsilon|\theta WZ} \circ \varphi_i^{-1})(c, w, z, \theta) &= \lambda \exp\{-\lambda(c - z_1'\theta_1 - \theta_2 g(z_2, w))\} \\ &= \lambda \exp\{-\lambda c\} \exp\{\lambda[z_1', g(z_2, w)]\theta]\}\end{aligned}$$

and satisfies the assumptions of Lemma 3.1 with $h(\theta) = \lambda$, $m(\theta) = \theta = (\theta_1', \theta_2)'$ *is the identity function,* $\tau(c, w, z) = (z_1', g(z_2, w))'$ *and* $k(c, w, z) = \exp\{-\lambda c\}$. *Then, if the support of* (C, W, Z) *is nonempty,* $f_{\theta|W}$ *is point-identified.*

Example 2 (Continued)**.** *For simplicity, assume* θ_2 *is not random, eliminate the index* t *and assume classical measurement error. We make use only of cross-section data for the estimation. In this example,*

$$f_{C|WZ}(c, w, z) = \int_\Theta \frac{\exp\left(-\frac{1}{2}\left(\frac{c-\phi_1 w_1 - \phi_2(w_2+z) - m(\gamma,\beta,\theta_2)}{\sigma_\varepsilon}\right)^2\right)}{\sqrt{2\pi\sigma_\varepsilon^2}} f_{\gamma\beta|W}(\gamma, \beta; w)\, d\gamma d\beta. \tag{3.5}$$

Define $\delta = m(\gamma, \beta, \theta_2)$. *Denote by* D *the support of* δ *and by* Γ *the support of* γ. *After a change of variable, this integral equation can be rewritten*

$$\begin{aligned}f_{C|WZ}(c, w, z) &= \int_D \int_\Gamma \frac{\exp\left(-\frac{1}{2}\left(\frac{c-\phi_1 w_1 - \phi_2(w_2+z) - \delta}{\sigma_\varepsilon}\right)^2\right)}{\sqrt{2\pi\sigma_\varepsilon^2}} \widetilde{f}_{\gamma\delta|W}(\gamma, \delta; w)\, d\gamma d\delta \qquad (3.6) \\ &= \int_D \frac{\exp\left(-\frac{1}{2}\left(\frac{c-\phi_1 w_1 - \phi_2(w_2+z) - \delta}{\sigma_\varepsilon}\right)^2\right)}{\sqrt{2\pi\sigma_\varepsilon^2}} \widetilde{f}_{\delta|W}(\delta; w) \left(\int_\Gamma \widetilde{f}_{\gamma|W\delta}(\gamma, \delta; w)\, d\gamma\right) d\delta \\ &= \int_D \frac{\exp\left(-\frac{1}{2}\left(\frac{c-\phi_1 w_1 - \phi_2(w_1+z) - \delta}{\sigma_\varepsilon}\right)^2\right)}{\sqrt{2\pi\sigma_\varepsilon^2}} \widetilde{f}_{\delta|W}(\delta; w)\, d\delta\end{aligned}$$

where

$$\widetilde{f}_{\gamma\delta|W}(\gamma, \delta; w) = f_{\gamma\beta|W}\left(\gamma, m^{-1}(\gamma, \delta, \theta_2)\right) \left|\frac{\partial m^{-1}(\gamma, \delta, \theta_2)}{\partial \delta}\right|.$$

The joint density of (γ, δ) *is not point-identified because any proper conditional density* $\widetilde{f}_{\gamma|W\delta}(\gamma; \delta, w)$ *is consistent with the data. In fact, the conditions of lemma 3.1 are not satisfied. The marginal density* $\widetilde{f}_{\delta|W}(\delta; w)$ *is point-identified. The identified set of densities* $f_{\gamma\beta|W}$ *is the set whose element is*

$$f_{\gamma\beta|W}(\gamma, \beta; w) = \widetilde{f}_{\delta|W}(m(\gamma, \beta, \theta_2); w) \cdot \widetilde{f}_{\gamma|W\delta}(\gamma, m(\gamma, \beta, \theta_2); w) \left|\frac{\partial m}{\partial \gamma}\right|$$

for some conditional density $\widetilde{f}_{\gamma|W\delta}$.

While the second result appears negative, it is important to note that our analysis only uses information on $f_{C|WZ}$. As long as we condition on age (i.e. as long as age is an element of W), and as long parameters governing aggregate uncertainty have been pre-estimated on aggregate time series data, it does not matter that the cross-sectional distribution of C does not contain any cross-sectional variation in aggregate variables, e.g., interest rates. In fact, even in the presence of aggregate shocks, the distribution of $f_{\delta|W}$ may still be point identified by considering idiosyncratic variation only (in the example, a two period panel with large cross section dimension is required). Note that potentially the entire distribution of $f_{\gamma\beta|W}$ may be identified if utility is CRRA rather than CARA or with enough time series variation (e.g., a large panel with sufficient interest rate variation). See Hoderlein et al. (2012) for work in this direction.

3.3 Identification without nuisance unobservables

In this section we briefly describe the case where we do not have ε so that $f_{C|WZ\theta}$ cannot be recovered as in theorem 1. This is relevant in models where all the unobservable variables are of interest so, ε is included in θ. In our setup, this implies that the general structural model (1.2) reduces to

$$\Psi(C, W, Z, \theta) = 0. \tag{3.7}$$

Indeed, even in this setup where φ is not strictly monotonic in θ and θ is multivariate, we can characterize the structural *pdf* $f_{\theta|W}$ as a solution to a slightly different constrained functional equation. Let $F_{C|WZ}$ be the cumulative distribution function of C conditioned on (W, Z) assumed to be in $L^2_{\pi_{cz}}$. Then,

$$F_{C|WZ}(c; w, z) = Sf_{\theta|W}(\theta; w) \quad \text{and} \quad f_{\theta|W} \in \mathcal{F}_{\theta|W}, \qquad P_W - a.s. \tag{3.8}$$

where $S : L^2_{\pi_\theta} \to L^2_{\pi_{cz}}$ is a bounded linear operator defined as

$$Sh = \int_{\{\theta;\, \varphi(w,z,\theta)\leq c\}} h(\theta; w)d\theta \equiv \int_\Theta 1\{\varphi(w, z, \theta) \leq c\}h(\theta; w)d\theta, \quad \forall h \in L^2_{\pi_\theta}.$$

The kernel of the new operator is $\frac{1\{\varphi(w,z,\theta)\leq c\}}{\pi_\theta(\theta;w)}$ and the adjoint S^* is:

$$S^*h = \int_{\mathcal{C}} \int_{\mathbb{R}^l} 1\{\varphi(w, z, \theta) \leq c\}\frac{\pi_{cz}(c, z)}{\pi_\theta(\theta)}h(c; w, z)dcdz, \quad \forall h \in L_{\pi_{cz}}(c, z).$$

For this new problem Assumptions 1 and 5 are still required while Assumptions 3 and 4 can be weakened to

Assumption 4'. *The random variable $(C, \theta, Z|W)$ has a joint continuous distribution characterized by its cumulative distribution function $F_{C\theta Z|W}$ that is absolutely continuous with respect to the Lebesgue measure with Radon-Nikodym density $f_{C\theta Z|W}$. Moreover, the support Θ of $f_{\theta|W}$ does not depend on W.*

A sufficient condition for compactness of the operator S is that π_θ and π_{cz} are chosen such that $\int_\Theta \frac{1}{\pi_\theta} d\theta < \infty$ and $\int_\mathcal{C} \int_{\mathbb{R}^l} \pi_{cz} dc dz < \infty$. When there are nuisance unobservables, the estimating equation (3.2) can be trivially recovered from (3.8) by differentiating with respect to c.

Due to the structure of the operator S, this problem is somewhat different, and left to future research. The estimation procedure for this case is the same as that one proposed in Section 4. Our estimator defined in (4.2) is valid with T replaced by S. We conjecture that the rate of the MISE will improve since $F_{C|WZ}$ can be estimated at a better rate than $f_{C|WZ}$. Moreover, the degree of ill-posedness will not be as severe as in the case where the kernel of T is exponential.

A complete analysis of this case is beyond the scope of this paper and we leave it for future research. This short discussion simply shows that our estimation approach is quite general and can be extended to cover the case where ε is not part of the structural model.

4 Estimation

In this section we develop the estimation theory for $f_{\theta|W}$ based on the resolution of the nonlinear problem (3.2). We call our estimation method *Indirect Estimation* and our estimator an *Indirect Regularized Density Estimator*.

Equation (3.2) cannot be solved directly for $f_{\theta|W}$ since the (generalized) inverse of T is unbounded and in addition $f_{C|WZ}$ is unknown. To solve the latter problem, we assume to be able to estimate $f_{C|WZ}$ in (3.2) by a (standard nonparametric) consistent estimator from a cross-section. The next assumption formalizes this:

Assumption 8. *Let (c_i, w_i, z_i), $i = 1, \ldots, n$ be an i.i.d. sample of (C, W, Z) that is used to construct an estimator $\hat{f}_{C|WZ}$ of $f_{C|WZ}$ such that $\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2$ converges to 0 as $n \uparrow \infty$.*

To address the problem of unboundedness of the (generalized) inverse of T, we propose the following regularization procedure. Our procedure is valid for any consistent nonparametric estimator of $f_{C|WZ}$ that satisfies assumption 8. We give a more detailed analysis of the rate and show asymptotic normality for the special case when $f_{C|WZ}$ is estimated by kernel smoothing.

4.1 Existence of an estimated solution

When one replaces $f_{C|WZ}$ in (3.2) with a consistent estimator $\hat{f}_{C|WZ}$, it is neither guaranteed that $\hat{f}_{C|WZ} \in \mathcal{R}(T)$ nor that $\hat{f}_{C|WZ} \in \mathcal{R}(T|_{\mathcal{F}_{\theta|W}})$ even though, under mild conditions, $\hat{f}_{C|WZ} \in L^2_{\pi_{cz}}$. If $\hat{f}_{C|WZ} \notin \mathcal{R}(T|_{\mathcal{F}_{\theta|W}})$ then, a solution to (3.2) does not exist. Nevertheless, we may define a generalized approximate solution that solves (3.2) approximately and that also accounts for two additional issues: the possibility that T is not injective and the fact that the solution must be constrained to belong to the convex and closed subset $\mathcal{F}_{\theta|W}$. The solution concept that we adopt is the *$\mathcal{C}$-best-approximate solution*, see Neubauer (1988), denoted by $\hat{f}^{\dagger c}_{\theta|W}$ and defined in the following.

Definition 1. *The* $\mathcal{C}$-best-approximate solution $\hat{f}^{\dagger c}_{\theta|W}$ *is an element of* $\mathcal{F}_{\theta|W} \subset L^2_{\pi_\theta}$ *such that*

$$||T\hat{f}^{\dagger c}_{\theta|W} - \hat{f}_{C|WZ}|| = \inf\left\{||Th - \hat{f}_{C|WZ}|| \quad \textit{subject to } h \in \mathcal{F}_{\theta|W}\right\} \tag{4.1}$$

and

$$||\hat{f}^{\dagger c}_{\theta|W}|| = \min\left\{||h|| \;\middle|\; ||Th - \hat{f}_{C|WZ}|| = ||T\hat{f}^{\dagger c}_{\theta|W} - \hat{f}_{C|WZ}|| \quad \textit{and} \quad \hat{f}^{\dagger c}_{\theta|W} \in \mathcal{F}_{\theta|W}\right\}.$$

Therefore, the *$\mathcal{C}$-best-approximate solution* is the least-squares solution on $\mathcal{F}_{\theta|W}$ that has minimal norm among all minimizers. It can be shown that $\hat{f}^{\dagger c}_{\theta|W} = T|^{\dagger}_{\mathcal{F}_{\theta|W}}\hat{f}_{C|WZ}$ where $T|^{\dagger}_{\mathcal{F}_{\theta|W}}$ denotes the $\mathcal{F}_{\theta|W}$-constrained Moore-Penrose generalized inverse. We denote $f^{\dagger c}_{\theta|W} = T|^{\dagger}_{\mathcal{F}_{\theta|W}} f_{C|WZ}$ and our estimator is an estimator of $f^{\dagger c}_{\theta|W}$. The *$\mathcal{C}$-best approximate solution* $\hat{f}^{\dagger c}_{\theta|W}$ exists and is unique if and only if $\mathcal{Q}_c\hat{f}_{C|WZ} \in \mathcal{R}(T|_{\mathcal{F}_{\theta|W}})$, where $\mathcal{Q}_c$ is the metric projector onto $\overline{\mathcal{R}(T|_{\mathcal{F}_{\theta|W}})}$, see proposition 5.14 in Engl *et al.* (2000). However, $\hat{f}^{\dagger c}_{\theta|W}$ does not depend continuously on $\hat{f}_{C|WZ}$ since, in general, $\mathcal{R}(T|_{\mathcal{F}_{\theta|W}})$ is non-closed (even if it is convex because $\mathcal{F}_{\theta|W}$ is closed and convex). As a result, the inverse problem (3.2) is ill-posed. From a practical point of view this means that the estimator $\hat{f}^{\dagger c}_{\theta|W}$ is an inconsistent estimator for $f^{\dagger c}_{\theta|W}$ despite the consistency of $\hat{f}_{C|WZ}$.[9]

Thus, a regularization procedure must be used to compute a consistent estimator of $f^{\dagger c}_{\theta|W}$. Because $f^{\dagger c}_{\theta|W} \in \mathcal{F}_{\theta|W}$, it is natural to require that the regularized estimator is in $\mathcal{F}_{\theta|W}$, too. To this end, we use a constrained Tikhonov-type estimator defined as the minimizer of

$$||Th - \hat{f}_{C|WZ}||^2 + \alpha||h||^2_s, \qquad h \in \mathcal{F}_{\theta|W}, \tag{4.2}$$

with respect to h. Here, $||\cdot||_s$ denotes a norm to be specified later, indexed by the parameter s, and $\alpha > 0$ is a parameter that decreases to 0 at a suitable rate. If $s = 0$, we have the classical

[9] One might think that the constraint makes the problem stable. This would be the case if $\mathcal{F}_{\theta|W}$ were a compact set, but this is unfortunately not the case here.

norm in $L^2_{\pi_\theta}$ and the estimator is the classical constrained Tikhonov regularized solution.

We call our estimator an *Indirect Regularized Density Estimator*. Problem (4.2) is nonlinear and, in general, does not allow a solution in closed-form except in one case. Thus, in the estimation procedure we treat two cases separately: *(i)* the case in which $f^{\dagger c}_{\theta|W} = f^{\dagger}_{\theta|W}$, for which a closed-form solution exists and *(ii)* the case in which $f^{\dagger c}_{\theta|W} \neq f^{\dagger}_{\theta|W}$ for which the estimate of the solution has to be computed numerically. Remember that $f^{\dagger}_{\theta|W}$ denotes the unconstrained best-approximate solution. We point out that the estimators proposed below are estimators of $f^{\dagger c}_{\theta|W}$. This means that in the non-identified case our procedure gives estimators for only one element of the identified set. In the point-identified case $f^{\dagger c}_{\theta|W} = f_{\theta|W}$. In the case in which T is compact the identified set can be estimated in principle by using the characterization in Proposition 3 and the estimator of $f^{\dagger}_{\theta|W}$ given in (4.6).

In the following, we use the notation $M_n \asymp J_n$, for positive quantities M_n and J_n depending on the index n, to mean that the ratio M_n/J_n is bounded away from zero and infinity.

4.2 Estimation of $f^{\dagger c}_{\theta|W}$: a two-step approach

In this section we consider the case $f^{\dagger c}_{\theta|W} = f^{\dagger}_{\theta|W}$, that is, the *best-approximate solution* belongs to $\mathcal{F}_{\theta|W}$. This is possible for instance when T is injective or when T is not injective but $T|_{\mathfrak{D}}$ is and $f^{\dagger}_{\theta|W} \in \mathcal{F}_{\theta|W}$ (but it may also be possible in the non-identified case). In this particular case we can use a two-steps approach to compute a regularized solution in $\mathcal{F}_{\theta|W}$ that in many cases can be faster than directly solving the nonlinear problem in (4.2).

First step: compute the regularized solution, denoted by $\hat{f}^{\alpha}_{\theta|W}$, of the unconstrained problem:

$$\min_{h \in L^2_{\pi_\theta}} \left\{ ||Th - \hat{f}_{C|WZ}||^2 + \alpha ||h||^2_s \right\}. \tag{4.3}$$

For $s = 0$, this is the classical Tikhonov regularized estimator, while for $s > 0$ we obtain the Tikhonov regularized estimator in the Hilbert scale, see *e.g.* Engl et al. (2000) or Florens et al. (2010). In this paper, we focus on the case $s = 0$.

Second step: compute the metric projection of $\hat{f}^{\alpha}_{\theta|W}$ onto the set $\mathcal{F}_{\theta|W}$. We denote by $\mathcal{P}_c$ this metric projector. Thus, the *indirect Tikhonov regularized estimator* of $f^{\dagger c}_{\theta|W}$ is

$$\mathcal{P}_c \hat{f}^{\alpha}_{\theta|W} := \max \left\{ 0, \hat{f}^{\alpha}_{\theta|W} - \frac{c}{\pi_\theta} \right\}. \tag{4.4}$$

where c is such that $\int_\Theta \mathcal{P}_c \hat{f}^{\alpha}_{\theta|W} d\theta = 1$. In practice, the constant c cannot be explicitly computed.

For that, one can use the following algorithm proposed by Gajek (1986):

$\mathcal{P}_c-$algorithm:

1. Set $\hat{f}^{\alpha(0)}_{\theta|W} = \hat{f}^{\alpha}_{\theta|W}$ and $k = 0$;
2. set $\hat{f}^{\alpha(k+1)}_{\theta|W} = \max\{0, \hat{f}^{\alpha(k)}_{\theta|W}\}$ and check $C_{k+1} = \int_\Theta \hat{f}^{\alpha(k+1)}_{\theta|W}(\theta; w)d\theta$. If $C_{k+1} = 1$ stop. Otherwise:
3. set $\hat{f}^{\alpha(k+2)}_{\theta|W} = \hat{f}^{\alpha(k+1)}_{\theta|W} - \frac{(C_{k+1}-1)}{\pi_\theta \int_\Theta \frac{1}{\pi_\theta} d\theta}$;
4. set $k = k + 2$ and repeat 2 - 4 until $|C_{k+1} - 1| < \varepsilon$, for $\varepsilon > 0$.

Gajek (1986) shows that $\hat{f}^{\alpha(k+1)}_{\theta|W}$ is the orthogonal projection of $\hat{f}^{\alpha(k)}_{\theta|W}$ onto $\mathcal{F}^+_{\theta|W}$ and $\hat{f}^{\alpha(k+2)}_{\theta|W}$ is the orthogonal projection of $\hat{f}^{\alpha(k+1)}_{\theta|W}$ onto $\mathcal{F}^1_{\theta|W}$, where $\mathcal{F}^+_{\theta|W}$ and $\mathcal{F}^1_{\theta|W}$ are the subsets of all functions in $L^2_{\pi_\theta}$ (measurable as functions of W) which are positive *a.e.* on Θ and which integrate to 1, respectively. Step 3 of the $\mathcal{P}_c-$algorithm is well defined for each iteration if $\int_\Theta \frac{1}{\pi_\theta} d\theta < \infty$ and $C_{k+1} < \infty$. In particular, if $\int_\Theta \frac{1}{\pi_\theta} d\theta < \infty$ holds, then there exists a unique real number c such that the $\mathcal{P}_c$-algorithm converges pointwise and in norm to $\mathcal{P}_c \hat{f}^{\alpha}_{\theta|W}$ defined in (4.4).

Condition $\int_\Theta \frac{1}{\pi_\theta} d\theta < \infty$ is trivially satisfied if Θ is compact and π_θ is continuous. For the case where Θ is not compact, the condition $\int_\Theta \frac{1}{\pi_\theta} d\theta < \infty$ means that π_θ must assign high weights to arguments in the tail of the distribution of θ.

4.2.1 Tikhonov regularization in the first stage

Let I denote the identity operator in $L^2_{\pi_\theta}$. The minimizer of (4.3) for $s = 0$ is the classical Tikhonov regularized estimator:

$$\hat{f}^{\alpha}_{\theta|W}(\theta; w) = (\alpha I + T^*T)^{-1} T^* \hat{f}_{C|WZ} \tag{4.5}$$

$$= \sum_{j=1}^{\infty} \frac{\lambda_j}{\alpha + \lambda_j^2} < \hat{f}_{C|WZ}, \psi_j > \tag{4.6}$$

where the first expression is valid for any T while the second one is valid when T is compact. We recall that $\{\lambda_j^2\}_{j\in\mathbb{N}}$ and $\{\psi_j\}_{j\in\mathbb{N}}$ denote the eigenvalues and eigenfunctions, respectively, of the operator TT^*.The parameter $\alpha > 0$ is a *regularization parameter* that converges to zero as the estimation error $(\hat{f}_{C|WZ} - f_{C|WZ})$ converges to zero. The Tikhonov regularization method is very well-known and developed in econometric theory so that we do not detail it here[10].

In the following, we derive the rate of the Mean Integrated Square Error (MISE) associated

[10]The interested reader is referred to Kress (1999, Chapter 15) or Carrasco, Florens and Renault (2007, Section 3).

with the projection estimator $\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W}$. We note that it is a weighted MISE since the norm is the norm in $L^2_{\pi_\theta}$. In the following, we introduce regularity conditions on $f^{\dagger c}_{\theta|W}$. The smoothness of $f^{\dagger c}_{\theta|W}$ is measured relative to the smoothing properties of the operator T in terms of *source conditions*. This is possible since T is a smoothing operator.

Assumption 9. *For some $\beta > 0$ and $0 < M < \infty$ there exists an element $\nu \in L^2_{\pi_\theta}$ such that*

$$f^{\dagger c}_{\theta|W} = (T^*T)^{\frac{\beta}{2}}\nu \qquad and \qquad ||\nu|| \leq M.$$

This assumption introduces an Hölder-type smoothness condition. The function ν is called the *source* so that Assumption 9 is known as *source condition* and is rather standard in inverse problems literature. Assumption 9 can be interpreted as a smoothness condition in a Sobolev space.

We state in the following theorem the rate of the MISE.

Theorem 2. *Let Assumptions 1-9 be satisfied. Then, the MISE associated with $\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W}$ is*

$$\mathbb{E}||\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}\left(\alpha^{\beta\wedge 2} + \frac{1}{\alpha}\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2\right).$$

Moreover, if $\alpha \asymp (\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2)^{\frac{1}{\beta\wedge 2+1}}$ then,

$$\mathbb{E}||\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}\left([\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2]^{\frac{\beta\wedge 2}{\beta\wedge 2+1}}\right).$$

Thus, this theorem tells us that the rate of the MISE is at best of order $[\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2]^{\frac{2}{3}}$. In section 4.2.3 we compare the rate of our indirect estimator with the rate of the unfeasible kernel smoothing estimator. In general the first rate is slower than the latter. Indeed, it is the fact that we use indirect observations of θ to estimate $f_{\theta|W}$ that slows down the rate of convergence; below we will relate the properties of the economic model to this rate. A drawback of the classical Tikhonov method is that a β greater than 2 (i.e., a higher degree of smoothness of the object of interest $f^{\dagger c}_{\theta|W}$) cannot be exploited in order to improve the rate of the MISE, but for transparency of discussion we desist from presenting more involved estimation methods that could exploit higher values of β at this stage[11].

The rate obtained in the previous theorem is in general not minimax. Tikhonov estimators achieve minimax rates only in restricted smoothness classes and for a particular class of opera-

[11] This is known as *saturation effect* and its a drawback of the classical Tikhonov regularization scheme. It is the price to pay for having a regularization method that is easy to implement and intuitive. When an analyst is willing to assume that $f^{\dagger c}_{\theta|W}$ has a higher degree of smoothness, other regularization methods can be used to exploit this smoothness. One method that exploit higher smoothness is Tikhonov regularization in Hilbert scale obtained from (4.3) with $s > 0$. This discussion is similar to higher order kernels in ordinary density estimation.

tors. This is highlighted by the fact that the *qualification number* for Tikhonov regularization is 2. Bissantz et al. (2007) provide the conditions under which Tikhonov estimators achieve the same rate of convergence of the MISE as spectral cut-off estimators, which are known to be minimax. In the next section we give the rate of the MISE of our Tikhonov estimator under the assumption that T is an operator satisfying the conditions in Bissantz et al. (2007).

4.2.2 Rate optimality

In this subsection we show that if the operator T satisfies an additional smoothness condition, then our method yields optimal rates of convergence, where optimal is understood in the minimax sense.

We introduce an operator $L : \mathcal{D}(L) \subset L^2_{\pi_\theta} \to L^2_{\pi_\theta}$ which is unbounded, positive, self-adjoint and defined on a dense domain $\mathcal{D}(L) \subset L^2_{\pi_\theta}$.[12] We assume that the inverse $L^{-1} : L^2_{\pi_\theta} \to L^2_{\pi_\theta}$ is bounded. For example, L could be a differential operator with boundary constraints. We denote by $\mathcal{X}_s$, $s \in \mathbb{R}$, the completion of $\bigcap_{k\in\mathbb{R}_+} \mathcal{D}(L^k)$ under the norm generated by the inner product $< x, y >_s := < L^s x, L^s y >$, $\forall x, y \in \bigcap_{k\in\mathbb{R}_+} \mathcal{D}(L^k)$. This space is generated by L and is a Hilbert space. The scale of the Hilbert space generated by L is denoted by $(\mathcal{X}_s)_{s\in\mathbb{R}}$. The norm $||\cdot||_s$ in (4.2) is the norm in $\mathcal{X}_s$. If $s \geq 0$ we have $\mathcal{X}_s = \mathcal{D}(L^s)$. Moreover, we have $\mathcal{X}_s \subset \mathcal{X}_{s'}$ for $s, s' \in \mathbb{R}$ with $s > s'$. When T is injective we can choose $L = (T^*T)^{-1}$ and the corresponding $\mathcal{X}_s$ is known as the *canonical Hilbert scale*.

For simplicity we assume that L has a countable number of eigenvalues $\rho_k \to \infty$ with corresponding eigenfunctions $\{u_k\}$ which form an orthonormal basis of $L^2_{\pi_\theta}$.[13] We formulate the smoothness properties of T in terms of the operator L.

Assumption 10. *There exists a function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ continuous and strictly increasing with $\phi(0+) = 0$ and finite $\underline{m}, \overline{m} > 0$ such that:*

$$\underline{m}||[\phi(L^{-2})]^{\frac{1}{2}} f|| \leq ||Tf|| \leq \overline{m}||[\phi(L^{-2})]^{\frac{1}{2}} f|| \quad \textit{for all } f \in L^2_{\pi_\theta}(\theta).$$

This assumption is equivalent to $\mathcal{R}((T^*T)^{\frac{1}{2}}) = \mathcal{R}([\phi(L^{-2})]^{1/2})$. We refer to Nair et al. (2005) for a discussion and examples of this condition. The next theorem provides rates for the case in which T and L^{-1} are finite smoothing operators as well as for the case in which they are infinitely smoothing.

[12] Despite the same notation, this operator is not the same as the operator L used in Proposition 4.

[13] For non-discrete spectrum our results will still hold but the presentation would become more technical since it would require the use of spectral measures and abstract functional calculus.

Theorem 3. *Let Assumptions 1-10 be satisfied. Then, the MISE associated with* $\mathcal{P}_c \hat{f}^{\alpha}_{\theta|W}$ *is*

$$\mathbb{E}||\mathcal{P}_c \hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 \asymp \alpha^{\beta \wedge 2} + \mathbb{E}||U||^2 \sum_{j=1}^{\bar{\nu}} [\phi(\rho_j^{-2})]^{-1}$$

where $\bar{\nu} = \bar{\nu}(n) \to \infty$ *and* $U = \hat{f}_{C|WZ} - f_{C|WZ}$.

- Mildly ill-posed case: *let* $\tilde{\beta} = (\beta \wedge 2)$, $\phi(t) = t^a$ *and* $\rho_j \asymp j^{\frac{1}{d}}$ *for some* $a > 0$. *If* $\bar{\nu} = \alpha^{-\frac{d}{2a}}$ *and* $\alpha \asymp (\mathbb{E}||U||^2)^{a/(a\tilde{\beta}+a+d/2)}$ *then*

$$\mathbb{E}||\mathcal{P}_c \hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 \asymp \left(\mathbb{E}||U||^2\right)^{\frac{a\tilde{\beta}}{a\tilde{\beta}+a+d/2}}.$$

- Severely ill-posed case: *let* $\tilde{\beta} = (\beta \wedge 2)$, $\phi(t) = \exp(-t^{-a/2})$ *and* $\rho_j \asymp j^{\frac{1}{d}}$ *for some* $a > 0$. *If* $\bar{\nu} = \alpha^{-d}$ *and* $\alpha \asymp c(-\log(\mathbb{E}||U||^2))^{-1/a}$ *with a sufficiently small* $c > 0$ *then*

$$\mathbb{E}||\mathcal{P}_c \hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 \asymp (-\log\left(\mathbb{E}||U||^2\right))^{-\frac{\tilde{\beta}}{a}}.$$

The rates given in the theorem are minimax, see *e.g.* Bissantz et al. (2007). We remark that for the severely ill-posed case, the minimax rate is independent of the dimension of θ, but very slow, which is a standard result in deconvolution type problems. In contrast, for the mildly ill-posed case the rate depends in a negative way on d, a clear indication of curse of dimensionality. Indeed, this curse would be double, as not just the rate of convergence of the $\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2$ part would depend on the dimensionality, but also the outer factor $a\tilde{\beta}/(a\tilde{\beta}+a+d/2)$. We analyze this effect further in the perhaps leading case of a kernel density estimator.

4.2.3 A special case: Kernel estimation of the distribution of the data

In this section we elaborate on the rates of theorems 2 and 3 in the case where $\hat{f}_{C|WZ}$ is a kernel estimator. Without loss of generality we assume in this section that $\mathcal{C} = [0,1]$, $\mathcal{W} = [0,1]^k$, $\mathcal{Z} = [0,1]^l$. Let $K(\cdot,\cdot)$ denote a generalized kernel function of order $r = 2$ which will be used in order to avoid boundary effects (we refer to Hall and Horowitz (2005), Darolles et al. (2011) and references therein for an explicit definition of $K(\cdot,\cdot)$). By abuse of notation, we use the same notation K for the kernels involving c, w and z, and for simplicity we use a second order kernel. We also use the same notation h_n (resp. h_d) for the different bandwidths for c, w and z used to compute the numerator (resp. the denominator) of (4.7)[14]. Define $K_h(\cdot,\cdot) = K(\frac{\cdot}{h},\frac{\cdot}{h})$.

[14] In principle, these bandwidths could be different. The choice of them is an issue extensively discussed in the literature, see Roussas (1967, 1969) and Rosenblatt (1969) among many others. To not overly complicate

Then, the kernel estimator of $f_{C|WZ}$ is

$$\hat{f}_{C|WZ}(c;w,z) = \frac{\frac{1}{nh_n^{1+k+l}}\sum_{i=1}^n K_h(c_i-c,c)K_h(w_i-w,w)K_h(z_i-z,z)}{\frac{1}{nh_d^{k+l}}\sum_{l=1}^n K_h(w_l-w,w)K_h(z_l-z,z)}. \tag{4.7}$$

Rates By standard Taylor series arguments, as in Rosenblatt (1969), it is easy to show that

$$\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2 = \mathcal{O}\left(\frac{1}{n\min\{h_n,h_d\}^{k+l+1}} + \max\{h_n^4, h_d^4\}\right) \tag{4.8}$$

and if $h_n = h_d = h$ is chosen such that $\frac{1}{nh^{k+l+1}} \asymp h^4$ then $\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2 = \mathcal{O}(n^{-4/(k+l+1+4)})$. By plugging this rate in the optimal rate of theorem 2, with $\alpha \asymp n^{-4/[(k+l+1+4)(\beta\wedge 2+1)]}$, we obtain

$$\mathbb{E}||\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}\left(n^{-\frac{4(\beta\wedge 2)}{(k+l+1+4)(\beta\wedge 2+1)}}\right). \tag{4.9}$$

We show now that this rate can be improved and made independent of the dimension of Z. This is possible since the application of the operator T^* to the error term $(\hat{f}_{C|WZ} - f_{C|WZ})$ has a smoothing effect and integrates out (C,Z), so that the dimension of (C,Z) does not play any role in the rate. The following corollary to theorem 2 gives the new rate.

Corollary 1. *Under Assumptions 1 - 9*

$$\mathbb{E}||\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}\left(\alpha^{\beta\wedge 2} + \frac{1}{\alpha^2}\left(\max\{h_n^4,h_d^4\} + \frac{1}{n(\min\{h_n,h_d\})^k}\right)\right).$$

Moreover, if $h_n = h_d \asymp n^{-1/(4+k)}$ *we have*

$$\inf_{\alpha}\left\{\mathbb{E}||\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2\right\} = \mathcal{O}\left(n^{-4\frac{\beta\wedge 2}{(4+k)((\beta\wedge 2)+2)}}\right). \tag{4.10}$$

The rate in the corollary is faster than the rate in (4.9) if $(l+1)(\beta\wedge 2+1) > 4+k$. It is clear that, under the conditions of the corollary, if we have no W and if $h_n = h_d \asymp n^{-1/4}$ then $\mathbb{E}||T^*(\hat{f}_{C|WZ} - f_{C|WZ})||^2 = \mathcal{O}(n^{-1})$. Our rate is increasing in β and decreasing in the dimension k of the endogenous variables W. We have a curse of dimensionality only in the dimension of the endogenous variables W and not in the dimension of the instruments Z. This is due to the action of the operator T^* that integrates out (C,Z).

Remark 7. The rates (4.9) and (4.10) are obtained without making use of Assumption 10. Therefore, following the discussion at the end of section 4.2.1 they are in general not minimax.

our discussion, we do not distinguish among the bandwidths, but note that our results can be trivially adapted to the case with different bandwidths.

On the contrary, under the assumptions of theorem 3 the rate in (4.9) would be replaced by

$$\mathbb{E}||\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}\left(n^{-\frac{4a(\beta\wedge 2)}{(k+l+1+4)(a(\beta\wedge 2)+a+d/2)}}\right)$$

in the mildly ill-posed case if we choose $\alpha \asymp n^{-4a/[(k+l+1+4)(a(\beta\wedge 2)+a+d/2)]}$. This rate makes explicit the dimension of θ. When $(4+k)a < (l+1)(a(\beta\wedge 2)+d/2+a)$ we can even improve this rate by following the lines of the proof of corollary 1, and get the rate $n^{-4a(\beta\wedge 2)/[(4+k)((\beta\wedge 2)a+2a+d/2)]}$. The optimal rate of the MISE associated with the unfeasible standard kernel smoothing estimator for $f_{\theta|W}$ is $n^{-\frac{4}{d+k+4}}$. A Comparison of this rate with the rate $n^{-4a(\beta\wedge 2)/[(4+k)((\beta\wedge 2)a+2a+d/2)]}$ of our indirect estimator shows that our indirect estimator is in general slower, because it requires the inversion of an integral operator. However, our estimator may be faster than the unfeasible kernel smoothing estimator in the particular case in which $ad(\beta \wedge 2) > (2a + d/2)(4 + k)$, i.e., if the dimension d and the degree of smoothness β are quite large.

Asymptotic Normality. We now study pointwise asymptotic normality of the Tikhonov regularized estimator $\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W}$ in the case where $\hat{f}_{C|WZ}$ is computed by using kernel smoothing as in (4.7). For that we introduce the following technical assumption:

Assumption 11. *Let $h = \min\{h_n, h_d\}$. We assume that*

$$\mathbb{E}\frac{1}{h^k}\left|\left(\frac{f_{C|WZ\theta}\pi_{cz}(c_1;w,z_1,\theta)h^k}{\mathbb{E}(\hat{f}_{WZ})(w,z_1)h^k} - \frac{\int_{\mathcal{C}}\mathbb{E}(\hat{f}_{CWZ})(c,w,z_1)f_{C|WZ\theta}\pi_{cz}(c;w,z_1,\theta)dch^k}{\mathbb{E}^2(\hat{f}_{WZ})(w,z_1)\pi_\theta h_d^k}\right)K_h(w_1-w,w)\right|^3 < \infty$$

In the following lemma we use the notation '$\Rightarrow$' to denote pointwise convergence in distribution.

Lemma 4.1. *Let Assumptions 1 - 9 and 11 hold and $\hat{f}^{\alpha}_{\theta|W}$ be the Tikhonov regularized estimator computed by using $\hat{f}_{C|WZ}(c;w,z)$ defined in (4.7). If $h_n = h_d \asymp n^{-\frac{1+\epsilon}{4+k}}$, for $0 < \epsilon < (4/k)$ and $\alpha \asymp n^{-\frac{(4-k\epsilon)\eta}{(4+k)(\beta\wedge 2+2)}}$ for $1 < \eta < \frac{(\beta\wedge 2)+2}{2}$, then,*

$$\frac{\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W}(\theta;w) - f^{\dagger c}_{\theta|W}(\theta;w)}{\sqrt{V_c(\theta,w)}} \Rightarrow \mathcal{N}(0,1)$$

where

$$V_c(\theta,w) = \frac{1}{n}Var\left(\mathcal{P}_c^{\dagger}\left[(\alpha I + T^*T)^{-1}T^*\frac{1}{f_{WZ}}\left(\frac{K_h(c_i-c,c)}{h_n^{k+l+1}} - \frac{\mathbb{E}(\hat{f}_{CWZ})}{\mathbb{E}(\hat{f}_{WZ})h_d^{k+l}}\right)K_h(w_i-w,w)K_h(z_i-z,z)\right](\theta,w)\right)$$

and $\mathcal{P}_c^{\dagger}$ denotes the projection on the tangent cone of $\mathcal{F}_{\theta|W}$ at $f^{\dagger c}_{\theta|W}$ defined as $\overline{\{\lambda(f - f^{\dagger c}_{\theta|W});\ \lambda \geq 0,\ f \in \mathcal{F}_{\theta|W}\}}$.

In order to obtain this asymptotic normality result, we require an α and a bandwidth h that converge to 0 at a faster rate than the asymptotically optimal one. This guarantees that the bias of $\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W}(\theta;w)$ is asymptotically negligible.

4.3 Estimation of $f^{\dagger c}_{\theta|W}$: constrained Tikhonov regularization

When $f^{\dagger c}_{\theta|W} \neq f^{\dagger}_{\theta|W}$ the two-steps procedure can no longer be applied. Instead we have to compute the constrained Tikhonov regularized solution by directly solving the minimization problem

$$\min_{h \in \mathcal{F}_{\theta|W}} \left\{ ||Th - \hat{f}_{C|WZ}||^2 + \alpha ||h||^2 \right\}. \tag{4.11}$$

The existence of a unique solution to problem (4.11) is proved in Neubauer (1988). A closed-form solution of this problem does not exist and numerical methods must be used to compute a solution. We denote by $\check{f}^{\alpha,c}_{\theta|W}$ the estimator obtained from solution of (4.11).

In the case $f^{\dagger c}_{\theta|W} = f^{\dagger}_{\theta|W}$, *i.e.* $f^{\dagger c}_{\theta|W}$ is in the interior of $\mathcal{F}_{\theta|W}$, this procedure can be seen as an alternative to the two-steps approach and gives an estimator for $f^{\dagger c}_{\theta|W}$ that has the same rate as computed in Section 4.2, see proposition 5.1 of Neubauer (1988).

When $f^{\dagger c}_{\theta|W} \neq f^{\dagger}_{\theta|W}$ then $f^{\dagger c}_{\theta|W} \in \partial\mathcal{F}_{\theta|W}$, where $\partial\mathcal{F}_{\theta|W}$ denotes the boundary of the set $\mathcal{F}_{\theta|W}$. The next theorem, which is valid in both the cases $f^{\dagger c}_{\theta|W} = f^{\dagger}_{\theta|W}$ and $f^{\dagger c}_{\theta|W} \neq f^{\dagger}_{\theta|W}$, states the convergence of the solution of (4.11) to $f^{\dagger c}_{\theta|W}$.

Theorem 4. *Let $f^{\alpha,c}_{\theta|W}$ be the solution of (4.11) when $\hat{f}_{C|WZ}$ is replaced by the true $f_{C|WZ}$ and Q be the orthogonal projector of $L^2_{\pi_{cz}}$ onto $\mathcal{R}(T)$. Then:*

(i) $\lim_{\alpha \to 0} f^{\alpha,c}_{\theta|W} = f^{\dagger c}_{\theta|W}$;

(ii) moreover, if $\hat{f}_{C|WZ} \in L^2_{\pi_{cz}}$ is such that $||Q(\hat{f}_{C|WZ} - f_{C|WZ})|| = \mathcal{O}_p(\delta)$, for some $\delta \to 0$, if $\alpha \to 0$ and $\delta^2 \alpha^{-1} \to 0$ as $\delta \to 0$ then:

$$\lim_{\delta \to 0} \check{f}^{\alpha,c}_{\theta|W} = f^{\dagger c}_{\theta|W}.$$

We do not develop here the theory concerning the rate of convergence of the MISE associated with $\check{f}^{\alpha,c}_{\theta|W}$ in order to not unnecessarily burden the paper. However, some comments are in order. We have to distinguish two cases: the case in which $f^{\dagger c}_{\theta|W}$ is in the interior of $\mathcal{F}_{\theta|W}$ and the case in which $f^{\dagger c}_{\theta|W}$ is on the boundary of $\mathcal{F}_{\theta|W}$. In the first case the smoothness condition is given by assumption 9 and the rate of the MISE is the same as in the unconstrained case, which is the rate given in section 4.2.1. In contrast, when $f^{\dagger c}_{\theta|W} \in \partial\mathcal{F}_{\theta|W}$, assumption 9 has to be slightly modified and further conditions on the set $\mathcal{F}_{\theta|W}$ need to be added in order to obtain the same rate as in the unconstrained case.

5 Monte Carlo simulation

In this section, to illustrate the performance of our estimator, we discuss results from simulations of the models in Examples 1 and 2 in Section 2.2. Simulation results from the first example show that our estimator performs well in a linear endogenous random coefficient model. We compare the estimator's performance with an unfeasible "oracle" estimator that uses simulated data on the unobserved variables θ and illustrate how the estimator's performance changes with sample size. The results show that our estimator is capable of uncovering the distribution of θ, and, in particular, its dependence on W.

Simulation results from the second example show that the estimator performs well when estimating a lifecycle consumption function with random parameters with moderate data size. The example illustrates how economic assumptions are mapped into the statistical framework of the estimator and also shows that data can provide meaningful restrictions on the joint distribution of random parameters even when the joint distributions is not point identified.

5.1 Simulation 1: Linear endogenous random coefficient model

Consider model (2.2) from Example 1. To focus on the properties of our estimator, assume that $g(Z_2, W) = Z_2 W$.[15] Substitute this function into equation (2.2) to obtain

$$C = \theta_1 Z_1 + \theta_2 Z_2 W + \varepsilon. \tag{5.1}$$

Assume that $\varepsilon \sim N(0, 0.1)$, $W \sim U[1, 2]$, and $Z \sim N(0, \Sigma_z)$ with Σ_z equal to the identity matrix. Finally, assume that $\theta | W \sim N(\mu_\theta, \Sigma_\theta)$ with $\mu_\theta = \beta_0 + \beta_1 W$, $\beta_0 = (1, 1)'$, $\beta_1 = (1, 1)'$ and Σ_θ equal to 0.1 times the identity matrix.

We simulate 1500 Monte Carlo datasets from this model; 500 for each sample size ($N = 500$, $N = 100$, and $N = 2500$). For each dataset, we first estimate $\widehat{f}_{C|WZ}$ using a Gaussian product kernel with bandwidth chosen as discussed below. Then we compute $\hat{f}^{\alpha}_{\theta|W}$ using (4.5). Finally, we compute $\mathcal{P}_c \widehat{f}^{\alpha}_{\theta|W}$ as in (4.4). In each case, we computed the estimator at the 30th, 50th, and 70th percentiles of the distribution of W.

To facilitate accurate numerical integration when computing the operators in (4.5), we first

[15] If $g(Z_2, W)$ were not known, a first-stage nonparametric estimate could be plugged into equation (2.2). The standard errors for our estimator could then be adjusted using standard methods for plug in estimators.

make a change of variable, mapping (C, Z_1, Z_2) into the region $[-1,1]^3$. Specifically, we define

$$\begin{aligned} U_c &= 2\Phi\left(\frac{C-\mu_c}{\sigma_c}\right)-1 \\ U_1 &= 2\Phi\left(\frac{Z_1-\mu_{z_1}}{\sigma_{z_1}}\right)-1 \\ U_2 &= 2\Phi\left(\frac{Z_2-\mu_{z_2}}{\sigma_{z_2}}\right)-1 \end{aligned}$$

where Φ is the standard normal CDF and (μ_c, σ_c), $\left(\mu_{z_i}, \sigma_{z_i}\right)$, $i = 1, 2$ are the empirical mean and standard deviation of C, Z_1 and Z_2.[16] Substituting these new variables into (5.1), and solving for ε, the structural function $\varepsilon = \varphi^{-1}(W, Z, \theta, \varepsilon)$ can be written as

$$\begin{aligned} \varepsilon &= \mu_c + \sigma_c \Phi^{-1}\left(\frac{U_c+1}{2}\right) \\ &\quad -\theta_1\left(\mu_{z_1} + \sigma_{z_1}\Phi^{-1}\left(\frac{U_{z_1}+1}{2}\right)\right) \\ &\quad -\theta_2\left(\mu_{z_2} + \sigma_{z_2}\Phi^{-1}\left(\frac{U_{z_2}+1}{2}\right)\right) W. \end{aligned}$$

We use this function to define the operator in Theorem 1.

Using the weight functions $\pi_{cz} = 1$ and $\pi_\theta = 1$, for each $w \in \{w_{30}, w_{50}, w_{70}\}$, we then compute $\hat{f}^{\alpha}_{\theta|W}$ to solve

$$\min_{\{h\}}\left\{\int\left(\hat{f}_{U_c|W,U_{z_1}U_{z_2}} - Th\right)^2 dc dz_1 dz_2 + \alpha\int h(\theta)^2 d\theta\right\} \tag{5.2}$$

where $\hat{f}_{U_c|W,U_{z_1}U_{z_2}}$ is computed by kernel smoothing. The solution is given in equation (4.5). To compute the operators we approximated the integral over $[-1,1]^3$ with the tensor product of three unidimensional Gauss-Legendre quadature rules with 20 quadrature nodes in each dimension. We approximated the integral over Θ with the tensor product of two unidimensional Gauss-Legendre quadrature rules with 20 quadrature nodes in each dimension.

Figures 1-3 display plots of the true density and of the estimated density for the three different quantiles of W obtained from one of our Monte Carlo datasets (with $N = 1000$). In each figure, the top panel shows the true density and the bottom panel shows the estimate. In all cases both the shape and location of the estimate track the true density quite closely. In particular, the unimodality of the density is well covered, and the location of the mode almost exactly coincides with the true mode. Moreover, the spread also very much coincides in every

[16] We also experimented with linear changes of variables and with no change of variables but with alternative weight functions for π_{cz}.

dimension with the true spread of the density of random coefficients.

Figures 4-6 show contour plots of the density. Results are obtained using bandwidths $h_n = h_d = 0.05$ and the Tikhonov regularization parameter $\alpha = 0.01$. Bandwidths are chosen to minimize the average of the square root of the density weighted mean squared error:

$$\begin{aligned} AMSE &= E\left[\frac{1}{3}\sum_q \left(\int \left[\mathcal{P}_c \widehat{f}^{\alpha}_{\theta|W}\left(\theta; w_q\right) - f_{\theta|W}\left(\theta; w_q\right)\right]^2 f_{\theta|W}\left(\theta; w_q\right) d\theta\right)^{0.5}\right] \\ &= E\left[MSE\right] \end{aligned} \tag{5.3}$$

$w_q \in \{w_{30}, w_{50}, w_{70}\}$ and where the average is calculated as the empirical average across 100 Monte Carlo replications and the pointwise average across three quantiles of the distribution of W.

For sample size of 1000, Figure 7 shows the densities of the square root of the weighted MSE for the Tikhonov estimator and the oracle estimator (*i.e.* the infeasible kernel density estimator). In each case, the distribution is the distribution across 500 Monte Carlo replication and across five different values of W. As was to be expected, the oracle estimator performs better, yet there is significant overlap in the distributions of results. Table 1 shows that the AMSE (calculated as the average across 500 Monte Carlo replications) of both estimators:

Table 1: AMSE as a function of sample size

	Sample size		
	500	1000	2500
Tikhonov estimator	0.423	0.350	0.280
Oracle estimator	0.219	0.172	0.140
Ratio	1.93	2.03	2.00

Several features of this result are noteworthy: First, observe that the ratio is approximately twofold, which is not very large if one considers the small sample size and the complexity of the procedure. Second, note that absolute value decreases, showing consistency. Third, note also that the ratio of the two averages increases slightly from 1.93 to 2.03. This is to be expected given the fact that the unfeasible oracle estimator converges faster. Nevertheless, we like to point out that the ratio is almost constant, indicating that the theoretical large sample differences may sightly overstate the small sample differences in behavior.

5.2 Simulation 2: Intertemporal consumption model

This section illustrates small sample performance of our estimator using simulated data from Example 2 detailed in Section 2.2. In the simulation $n = 1000$ agents start at age $t = 21$,

work for 45 periods and then retire obtaining a terminal retirement utility. Income grows until retirement. In addition, in each period each agent faces a permanent income shock η_t. These shocks are independent over time and across individuals and distributed as $\eta_t \sim \mathcal{N}(0, 0.01668)$. The initial value of income is set to 0.2 (scaled so that 0.2 equals \$20,000) and the initial permanent shock is set to zero. The interest rate is set to $R = 1 + r = 1.05$ and the random parameters γ and β have support on $(0.5, 4.0)$ and $(0.700, 0.999)$ respectively, which covers all values suggested in the literature. The joint distribution of (γ, β) is generated from a non-linear transformation of a normal random variable with mean vector $(1, 0)'$ and identity covariance matrix. That is, we generate the distribution as follows. We define $x \sim N(\mu_x, I)$ with $\mu_x = (1, 0)'$ and generate

$$\begin{aligned} \gamma &= 0.5 + 3.5 * \Phi(x_1) \\ \beta &= 0.7 + 0.299 * \Phi(x_2) \end{aligned}$$

where Φ is the standard normal CDF.[17] In addition, measurement error in consumption is $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ with σ_ε^2 set equal to 25% of the the cross-sectional variance of consumption.

We show in Figures 8-10 the quartiles of the simulated Y_t, C_t^*, C_t and A_t for each t. For the estimation, we select one cross section to be used for our estimator. The results presented in Figures 11-18 refer to the period $t = 31$. We obtained similar results for other values of t. The dependence between δ and W varies with t as does the distribution of the data. However, the quality of the estimation results does not.

Recall that in this example, the joint distribution of (γ, β) is not identified because the variables enter the kernel of the operator only through a single index. Instead we estimate the distribution of

$$\delta = 0.5\phi_5\gamma + \phi_3 \frac{\ln(R\beta)}{\gamma} \tag{5.4}$$

where ϕ_3 and ϕ_5 are parameters that depend only on the interest rate R and the time period t.

For the simulations, we use a Gaussian kernel with Tikhonov regularization parameter $\alpha = 0.01$ and bandwidths $h_n = h_d = 0.3$. For the infeasible kernel density estimator we set the bandwidth to $h_\theta = 0.3$. Tuning parameters may be chosen using least-squares cross-validation. However, for the purposes of the illustration we use 1000 Monte Carlo replications to compute the square root of the density weighted mean squared error and chose tuning parameters using a grid search to minimize the square root of density weighted mean squared error.

In this case, the true distribution of δ conditional on W is difficult to compute because it is endogenously determined from the structural model. Therefore we compute the following

[17] Thus, while x is normally distributed, γ has support on $(0.5, 4.0)$ and β is uniformly distributed on $(0.7, 0.999)$.

square root of the density weighted mean squared error averaged across quantiles of the W distribution:

$$AMSE = E\left[\frac{1}{4}\sum_{q}\left(\int\left(\left[\mathcal{P}_c\widehat{f}^{\alpha}_{\delta|W}\left(\delta;w_q\right)-\widehat{f}^{Ker}_{\delta|W}\left(\delta;w_q\right)\right]^2\right)\left(\widehat{f}^{Ker}_{\delta|W}\left(\delta;w_q\right)\right)d\delta\right)^{0.5}\right]. \quad (5.5)$$

To compute the AMSE, we replace the expected values in (5.5) with the average across the 1000 Monte Carlo replications and compute the integral across δ using Gauss-Legendre quadrature nodes with 301 points of support. The average across W is computed as the pointwise average across vectors w with each coordinate of w equal to either its 25th or 75th percentile.

In Figures 11-14, we show an (infeasible) kernel density estimator of the *pdf* of δ (in solid black line) together with our Tikhonov estimator (in dashed green line) and pointwise 95% confidence intervals obtained using the bootstrap. In each figure, the estimate is conditional on fixed levels of assets and income. "Low" levels of each variable correspond to the 25th percentile and "high" levels correspond to the 75th percentile. To estimate the confidence intervals, we created 1000 bootstrap samples from the data, each a sample of 1000 observations drawn with replacement. We then use the pointwise 0.025 and 0.975 percentiles of the bootstrap estimates as our confidence bands. As the results reveal, the unfeasible oracle estimator which we take in place of the true density is, for every value w of W we consider, within the confidence intervals. This suggests that our estimator is reasonable accurate, in spite of the only moderate sample size of $n = 1000$.

To provide an economic interpretation of these results, note that while they characterize the density of δ conditional on W, these results also place constraints on the joint distribution of (β, γ) given W. For each quantile of the distribution of δ, we can draw a curve representing the values of (β, γ) satisfying (5.4) . This is a quantile level set. Suppose we draw such a curve for $\delta = \delta_q$ the $q'th$ quantile of the δ distribution. Since (5.4) is monotonic in β, it must be the case that with probability q, (β, γ) lie below this level set and with probability $1 - q$ they lie above this level set.

Figures 15–18 show these level set curves conditional on various values of $W_t = (A_{t=31}, Y_{t=30})$. For example, the blue solid line in Figure 15 shows the 0.1 quantile level set. With probability 0.1, (β, γ) lie below this curve. In each case, the quantile-level-sets partition the (β, γ) space into conical regions. The conical region in Figure 15 bounded by the 0.1 and 0.9 quantile level sets shows that people with low assets and low income are likely to be very impatient ($\beta < 0.9$) if they are risk averse ($\gamma > 3.5$) but are likely to be patient if they have low risk aversion. The other figures show that this conical region shifts upward for people with higher assets or income. As theory predicts, individuals with higher asset holdings are on average more patient and risk averse, but there is some evidence of trade off between patience and risk aversion.

6 Conclusion

This paper develops results on semi-parametric identification and estimation of the *pdf* of the unobserved heterogeneity in structural models. We identify the *pdf* of interest as the solution of a non-linear inverse problem. The identified set is a convex subset of $L^2_{\pi_\theta}$. The estimation methods we propose are based on methods developed in the Inverse Problem literature. Finally, a simulated exercise for the Euler Equation in finite horizon intertemporal consumption models shows good finite-sample properties of our procedure.

A Appendix: Proofs

A.1 Proof of Theorem 1

By Assumption 1, there exists a unique $c = \varphi(w, z, \theta, \varepsilon)$ that satisfies (2.1). Thus, using the transformation $\varphi(w, z, \theta, \cdot)$ mapping c to ε, the density of ε, $f_{\varepsilon|WZ\theta}$ specified in Assumption 3 and $f_{\theta|W}$ specified in Assumption 4, we can characterize the *pdf* of $f_{C\theta|WZ}$. Let $\mathcal{E}_1, \ldots, \mathcal{E}_s$ be a partition of $\mathbb{R}$ such that $\varphi(w, z, \theta, \cdot) : \mathcal{E}_i \to \mathbb{R}$ is one-to-one for each $i = 1, \ldots, s$, for given (w, z, θ). Let $\varepsilon^i = \varphi_i^{-1}(w, z, \theta, \cdot) : \mathbb{R} \to \mathcal{E}_i$ be the corresponding inverse mapping for given (w, z, θ). Then,

$$f_{C|WZ\theta}(c; w, z, \theta) = \sum_{i=1}^{s} f_{\varepsilon|WZ\theta}(\varphi_i^{-1}(w, z, \theta, c); w, z, \theta_2) \cdot \left|\partial_c \varphi_i^{-1}(w, z, \theta, c)\right| 1_{\mathcal{C}_i}(c). \tag{A.1}$$

Further using Assumption 5 we have $f_{C\theta|WZ} = f_{C|WZ\theta}\, f_{\theta|W}$. This implies that

$$f_{C|WZ}(c; w, z) = \int_\Theta f_{C|WZ\theta}(c; w, z, \theta) f_{\theta|W}(\theta; w) d\theta \tag{A.2}$$

Finally, since a unique solution in C to (2.1) exists, the chain rule implies that: $\partial_\varepsilon \Psi(c, w, z, \theta, \varepsilon) = \partial_c \Psi(c, w, z, \theta, \varepsilon)\partial_\varepsilon c + \partial_\varepsilon \Psi(c, w, z, \theta, \varepsilon) = 0$, by abuse of notation. Therefore, $\partial_\varepsilon c = \partial_\varepsilon \varphi(w, z, \theta, \varepsilon)$ and $\partial_\varepsilon \varphi(w, z, \theta, \varepsilon) = -\frac{\partial_\varepsilon \Psi(c,w,z,\theta,\varepsilon)}{\partial_c \Psi(c,w,z,\theta,\varepsilon)}$. We conclude that

$$\begin{aligned} \partial_c \varphi_i^{-1}(w, z, \theta, c) &= \frac{1}{\partial_\varepsilon \varphi(w, z, \theta, \varepsilon)|_{\varepsilon=\varphi_i^{-1}(w,z,\theta,c)}} = -\left[\frac{\partial_\varepsilon \Psi(c, w, z, \theta, \varepsilon)}{\partial_c \Psi(c, w, z, \theta, \varepsilon)}\right]^{-1} \Big|_{\varepsilon=\varphi_i^{-1}(w,z,\theta,c)} \\ &= -\left[\frac{\partial_c \Psi(c, w, z, \theta, \varphi_i^{-1}(w, z, \theta, c))}{\partial_\varepsilon \Psi(c, w, z, \theta, \varphi_i^{-1}(w, z, \theta, c))}\right]. \end{aligned} \tag{A.3}$$

By replacing (A.3) in (A.1) and (A.1) in (A.2) we get the result.

A.2 Proof of Proposition 1

First, we remark that an integral operator from $L^2_{\pi_\theta}$ to $L^2_{\pi_{cz}}$ is Hilbert-Schmidt if its kernel is square integrable with respect to $\pi_\theta \times \pi_{cz}$. An Hilbert-Schmidt operator is bounded and compact.

Under the conditions of the proposition we compute $\int_{\mathcal{C}} \int_{\mathbb{R}^l} \int_\Theta \frac{f^2_{C|WZ\theta}}{\pi^2_\theta} \pi_\theta \pi_{cz}$:

$$
\begin{aligned}
& \int_{\mathcal{C}} \int_{\mathbb{R}^l} \int_\Theta \left[\sum_{i=1}^s f_{\varepsilon|WZ\theta}(\varphi_i^{-1}(w,z,\theta,c);w,z,\theta) \cdot \left|\partial_c \varphi_i^{-1}(w,z,\theta,c)\right| 1_{\mathcal{C}_i}(c) \right]^2 \frac{\pi_\theta(\theta)\pi_{cz}(c,z)}{\pi^2_\theta} d\theta dc dz \\
\leq & \int_\Theta \int_{\mathbb{R}^l} \int_{\mathcal{C}} \sum_{i=1}^s f^2_{\varepsilon|WZ\theta}(\varphi_i^{-1}(w,z,\theta,c);w,z,\theta) \cdot \left|\partial_c \varphi_i^{-1}(w,z,\theta,c)\right|^2 \sum_{i=1}^s 1_{\mathcal{C}_i}(c) \frac{\pi_{cz}(c,z)}{\pi_\theta} dc dz d\theta \\
= & \int_\Theta \int_{\mathbb{R}^l} s \sum_{i=1}^s \int_{C_i} f^2_{\varepsilon|WZ\theta}(\varepsilon_i;w,z,\theta) \cdot \left|\partial_{\varepsilon_i} \varphi(w,z,\theta,\varepsilon_i)\right|^{-1} \frac{\pi_{cz}(\varphi(w,z,\theta,\varepsilon_i),z)}{\pi_\theta(\theta)} d\varepsilon_i dz d\theta < \infty
\end{aligned}
$$

where the first inequality follows from the Fubini's theorem and the Cauchy-Schwartz's inequality and the second equality follows from the change of variable $\varphi_i^{-1}(w,z,\theta,c) = \varepsilon_i$. The final inequality follows from Assumption 6. This result shows that T is Hilbert-Schmidt and then bounded and compact. By using the inequality in (3.3), this result shows that $\mathcal{R}(T) \subset L^2_{\pi_{cz}}$.

A.3 Proof of Proposition 2

By definition, the adjoint operator T^* of the bounded linear operator T satisfies: $\forall h \in L^2_{\pi_\theta}$, $\forall \psi \in L^2_{\pi_{cz}}$, $< Th, \psi > = < h, T^*\psi >$. Thus,

$$
\begin{aligned}
< Th, \psi > &= \int_{\mathcal{C}} \int_{\mathbb{R}^l} (Th)(c;w,z) \psi(c,z;w) \pi_{cz}(c,z) dc dz \\
&= \int_{\mathcal{C}} \int_{\mathbb{R}^l} \int_\Theta f_{C|WZ\theta}(c;w,z,\theta) h(\theta;w) d\theta \psi(c,z;w) \pi_{cz}(c,z) dc dz \\
&= \int_\Theta h(\theta;w)\pi_\theta(\theta) \int_{\mathcal{C}} \int_{\mathbb{R}^l} f_{C|WZ\theta}(c;w,z,\theta) \psi(c,z;w) \frac{\pi_{cz}(c,z)}{\pi_\theta(\theta)} dc dz d\theta = < h, T^*\psi >
\end{aligned}
$$

where the third equality follows from the Fubini's theorem.

A.4 Proof of Proposition 4

Suppose that $f_{\theta|CWZ}$ is $\mathcal{T}$-complete and that for $f_{\theta|W}, \tilde{f}_{\theta|W} \in \mathcal{F}_{\theta|W}$, $T|_{\mathcal{F}_{\theta|W}}(f_{\theta|W}) = T|_{\mathcal{F}_{\theta|W}}(\tilde{f}_{\theta|W})$ holds. By using the decomposition $f_{C|WZ\theta} = f_{\theta|CWZ} f_{C|WZ} / f_{\theta|W}$ this equality can be rewritten as

$$
\begin{aligned}
0 &= T|_{\mathcal{F}_{\theta|W}}(f_{\theta|W}) - T|_{\mathcal{F}_{\theta|W}}(\tilde{f}_{\theta|W}) = \int_\Theta f_{C|WZ\theta}(c;w,z,\theta) \left[f_{\theta|W}(\theta;w) - \tilde{f}_{\theta|W}(\theta;w) \right] d\theta \\
&= \int_\Theta f_{\theta|CWZ}(\theta;c,w,z) \frac{f_{C|WZ}(c;w,z)}{f_{\theta|W}(\theta;w)} \left[f_{\theta|W}(\theta;w) - \tilde{f}_{\theta|W}(\theta;w) \right] d\theta \qquad \text{(A.4)}
\end{aligned}
$$

which is equivalent to

$$0 = \int_\Theta f_{\theta|CWZ}(\theta; c, w, z) \frac{1}{f_{\theta|W}(\theta; w)} \left[f_{\theta|W}(\theta; w) - \tilde{f}_{\theta|W}(\theta; w) \right] d\theta \tag{A.5}$$

because, by Assumptions 2 and 3, $0 \leq m_\varepsilon \kappa \leq f_{C|WZ} < \infty$. Moreover, $\frac{1}{f_{\theta|W}(\theta;w)} \left[f_{\theta|W}(\theta; w) - \tilde{f}_{\theta|W}(\theta; w) \right] \in \mathcal{T}$ so that (A.5) implies $\frac{1}{f_{\theta|W}(\theta;w)} \left[f_{\theta|W}(\theta; w) - \tilde{f}_{\theta|W}(\theta; w) \right] = 0$ which in turns implies $f_{\theta|W}(\theta; w) = \tilde{f}_{\theta|W}(\theta; w)$ under Assumption 4.

On the other side, if (3.4) holds then $0 = \int_\Theta f_{\theta|CWZ}(\theta; c, w, z) \frac{1}{f_{\theta|W}(\theta;w)} \left[f_{\theta|W}(\theta; w) - \tilde{f}_{\theta|W}(\theta; w) \right] d\theta$ implies that $\frac{1}{f_{\theta|W}(\theta;w)} \left[f_{\theta|W}(\theta; w) - \tilde{f}_{\theta|W}(\theta; w) \right] = 0$ because, by Assumption 2 and 3, $0 \leq m_\varepsilon \kappa \leq f_{C|WZ} < \infty$. This concludes the proof.

A.5 Proof of Lemma 3.1

For simplicity we consider the case where θ is one-dimensional (the multi-dimensional case can be recovered in a similar way). Let us suppose that $T\phi(\theta; w) = 0$, P_W-a.e. for some function $\phi \in \mathcal{F}_{\theta|W}$. Then,

$$T\phi = \int_\Theta \sum_{i=1}^{s} f_{\varepsilon|\theta WZ} \left(\varphi_i^{-1}(w, z, \theta, c); \theta, w, z \right) \cdot \left| \partial_c \varphi_i^{-1}(w, z, \theta, c) \right| \phi(\theta; w) d\theta = 0$$

implies

$$\int_\Theta f_{\varepsilon|\theta Wz} \left(\varphi_i^{-1}(w, z, \theta, c); \theta, w, z \right) \cdot \left| \partial_c \varphi_i^{-1}(w, z, \theta, c) \right| \phi(\theta; w) d\theta = 0 \quad \forall i = 1, \ldots, s.$$

Then, $\forall i = 1, \ldots, s$ we have:

$$\begin{aligned}
0 &= \int_\Theta \exp\{\tau_i(c, w, z) m_i(\theta)\} h_i(\theta) k_i(c, w, z) \phi(\theta; w) \left| \partial_c \varphi^{-1}(w, z, \theta, c) \right| d\theta \\
&= \int_\Theta \exp\{\tau_i(c, w, z) m_i(\theta)\} h_i(\theta) k_i(c, w, z) \tilde{\phi}(\theta; w, z, c) | d\theta \\
&= \int_\Theta \exp\{\tau_i(c, w, z) \mu_i\} h_i\left(m_i^{-1}(\mu_i)\right) k_i(c, w, z) \tilde{\phi}_i\left(m_i^{-1}(\mu_i); w, z, c\right) dm_i^{-1}(\mu_i)
\end{aligned}$$

where we have used the notation $\tilde{\phi}_i(\theta; w, z, c) := \phi(\theta; w) \left| \partial_c \varphi_i^{-1}(w, z, \theta, c) \right|$ and the change of variable $m_i(\theta) = \mu_i$. Moreover, since $dm_i^{-1}(\mu_i)$ and h_i are positive functions, we can define a measure $\nu_i(d\mu_i) =$

$h_i\left(m_i^{-1}(\mu_i)\right) dm_i^{-1}(\mu_i)d\mu_i$.Thus,

$$\begin{aligned}
0 = T\phi &= k_i(c,w,z)\int_\Theta \exp\{\tau_i(c,w,z)\mu_i\}\tilde{\phi}_i\left(m_i^{-1}(\mu_i);w,z,c\right)\nu_i(d\mu_i) \\
&= k_i(c,w,z)\int_\Theta \exp\{\tau_i(c,w,z)\mu_i\}\zeta_i(\mu_i;w,z,c)\nu_i(d\mu_i) \\
&= k_i(c,w,z)\begin{pmatrix} \int_\Theta \exp\{\tau_i(c,w,z)\mu_i\}\zeta_i^+(\mu_i;w,z,c)\nu_i(d\mu_i) \\ -\int_\Theta \exp\{\tau_i(c,w,z)\mu_i\}\zeta_i^-(\mu_i;w,z,c)\nu_i(d\mu_i) \end{pmatrix} \\
&= k_i(c,w,z)\left(\int_\Theta \exp\{\tau_i(c,w,z)\mu_i\}F_i(d\mu_i;w,z,c) - \int_\Theta \exp\{\tau_i(c,w,z)\mu_i\}G_i(d\mu_i;w,z,c)\right)
\end{aligned}$$

where $\zeta_i(\mu_i;w,z,c) = \tilde{\phi}_i \circ m_i^{-1}$, $F_i(d\mu_i;w,z,c) = \zeta_i^+(\mu_i;w,z,c)\nu_i(d\mu_i)$, $G_i(d\mu_i;w,z,c) = \zeta_i^-(\mu_i;w,z,c)\nu_i(d\mu_i)$ and, for a function h, h^+ and h^- denote the positive and negative part of it, respectively. It follows that

$$\int_\Theta \exp\{\tau_i(c,w,z)\mu_i\}F_i(d\mu_i;w,z,c) = \int_\Theta \exp\{\tau_i(c,w,z)\mu_i\}G_i(d\mu_i;w,z,c),$$

that is, F_i and G_i are two measures with the same Laplace transform. Then, they are equal. This implies that $\zeta_i(\mu_i;w,z,c) = 0$ and then $\phi_i(\theta;w) = 0$, P_W-a.s. since $\partial_c \varphi_i^{-1}(w,z,\cdot,c)$ is bounded away from 0 and ∞, $\forall (c,w,z) \in \mathcal{C}\times\mathcal{Z}\times\mathcal{W}$.

A.6 Proof of Theorem 2

First, since $||\mathcal{P}_c|| \leq 1$ we have:

$$\mathbb{E}||\mathcal{P}_c\hat{f}^\alpha_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathbb{E}||\mathcal{P}_c(\hat{f}^\alpha_{\theta|W} - f^\dagger_{\theta|W})||^2 \leq ||\mathcal{P}_c||^2\mathbb{E}||\hat{f}^\alpha_{\theta|W} - f^\dagger_{\theta|W}||^2 \leq \mathbb{E}||\hat{f}^\alpha_{\theta|W} - f^\dagger_{\theta|W}||^2.$$

We develop here only the proof for the case where T is compact and refer to Engl *et al.* (2000 Section 5.1) for a proof in the general non-compact case. Let $f^\alpha_{\theta|W} := (\alpha I + T^*T)^{-1}T^* f_{C|WZ}$, then

$$\mathbb{E}||\hat{f}^\alpha_{\theta|W} - f^\dagger_{\theta|W}||^2 \leq 2\mathbb{E}||\hat{f}^\alpha_{\theta|W} - f^\alpha_{\theta|W}||^2 + 2\mathbb{E}||f^\alpha_{\theta|W} - f^\dagger_{\theta|W}||^2 := 2(\mathcal{A}_1 + \mathcal{A}_2). \tag{A.6}$$

Term $\mathcal{A}_1$ is

$$\begin{aligned}
\mathcal{A}_1 &= \mathbb{E}||(\alpha I + T^*T)^{-1}T^*(\hat{f}_{C|WZ} - f_{C|WZ})||^2 \leq ||(\alpha I + T^*T)^{-1}T^*||^2\mathbb{E}||(\hat{f}_{C|WZ} - f_{C|WZ})||^2 \\
&= \sup_{||u||\leq 1}\sum_j \frac{\lambda_j^2}{(\alpha+\lambda_j^2)^2} < u,\psi_j >^2 \mathbb{E}||(\hat{f}_{C|WZ} - f_{C|WZ})||^2 \\
&\leq \left(\sup_j \frac{\lambda_j}{(\alpha+\lambda_j^2)}\right)^2 \mathbb{E}||(\hat{f}_{C|WZ} - f_{C|WZ})||^2 = \frac{1}{2\alpha}\mathbb{E}||(\hat{f}_{C|WZ} - f_{C|WZ})||^2.
\end{aligned}$$

Next, we develop term $\mathcal{A}_2$:

$$\begin{aligned}
\mathcal{A}_2 &= \mathbb{E}||(\alpha I + T^*T)^{-1}T^* f_{C|WZ} - f^\dagger_{\theta|W}||^2 = |||[I - (\alpha I + T^*T)^{-1}T^*T]f^\dagger_{\theta|W}||^2 \\
&= ||\alpha(\alpha I + T^*T)^{-1} f^\dagger_{\theta|W}||^2 = \alpha^2 \sum_j \frac{\lambda_j^{2\beta}}{(\alpha+\lambda_j^2)^2} \frac{< f^\dagger_{\theta|W}, \varphi_j >^2}{\lambda_j^{2\beta}} \\
&\leq \alpha^2 \left(\sup_j \frac{\lambda_j^\beta}{(\alpha+\lambda_j^2)} \right)^2 \sum_j \frac{< f^\dagger_{\theta|W}, \varphi_j >^2}{\lambda_j^{2\beta}} \leq \alpha^\beta \frac{(2-\beta)^{2-\beta}}{4} \beta^\beta M \quad \text{if } \beta \leq 2
\end{aligned}$$

since $< f^\dagger_{\theta|W}, \varphi_j >^2 = < f^{\dagger c}_{\theta|W}, \varphi_j >^2 = \lambda_j^{2\beta} < \nu, \varphi_j >^2$ under Assumption 9. This shows that $\mathbb{E}||\hat{f}^\alpha_{\theta|W} - f^\dagger_{\theta|W}||^2 = \mathcal{O}\left(\alpha^{\beta\wedge 2} + \frac{1}{\alpha}\mathbb{E}||(\hat{f}_{C|WZ} - f_{C|WZ})||^2\right)$. If we choose α such that $\alpha^{\beta\wedge 2} \asymp \frac{1}{\alpha}\mathbb{E}||(\hat{f}_{C|WZ} - f_{C|WZ})||^2$ then we get the second result of the theorem.

A.7 Proof of Theorem 3

Let us consider the development made in the proof of Theorem 2:

$$\mathbb{E}||\hat{f}^\alpha_{\theta|W} - f^\dagger_{\theta|W}||^2 \leq 2\mathbb{E}||\hat{f}^\alpha_{\theta|W} - f^\alpha_{\theta|W}||^2 + 2\mathbb{E}||f^\alpha_{\theta|W} - f^\dagger_{\theta|W}||^2 := 2(\mathcal{A}_1 + \mathcal{A}_2).$$

We only develop term $\mathcal{A}_1$ as term $\mathcal{A}_2$ does not change.

We first show that the eigenvalue u_j of L belongs to $\mathcal{R}(T^*)$, $\forall j = 1, 2, \ldots$ To see this consider the operator $\tilde{T}^* := \left[\phi(L^{-2})\right]^{-1/2} T^*T \left[\phi(L^{-2})\right]^{-1/2}$ for which, under assumption 10, $< \tilde{T}^*\psi, \psi >= ||T\left[\phi(L^{-2})\right]^{-1/2}\psi||^2 \geq \underline{m}^2||\psi||^2$, $\forall \psi \in \mathcal{R}(\left[\phi(L^{-2})\right]^{1/2})$. This implies that $\tilde{T}^*$ is a strictly positive self-adjoint operator which is invertible with bounded inverse. Therefore, $\exists \tilde{\psi}_j \in \mathcal{D}(\tilde{T}^*)$ such that $\tilde{T}^*\tilde{\psi}_j = u_j$, $\forall j = 1, 2, \ldots$, and it follows that $\left[\phi(L^{-2})\right]^{1/2}\tilde{T}^*\tilde{\psi}_j = \left[\phi(L^{-2})\right]^{1/2} u_j = \left[\phi(\rho_j^{-2})\right]^{1/2} u_j$ which means that $u_j \in \mathcal{R}(\left[\phi(L^{-2})\right]^{1/2}\tilde{T}^*)$ and then $u_j \in \mathcal{R}(T^*)$.

By the Cauchy-Schwartz inequality and since $t/(\alpha+t^2) \leq t^{-1}$ term $\mathcal{A}_1$ can be bounded from above as

$$\begin{aligned}
\mathcal{A}_1 &= \mathbb{E}||(\alpha I + T^*T)^{-1}T^*U||^2 \leq \mathbb{E}||U||^2||(\alpha I + T^*T)^{-1}T^*||^2 \\
&\leq \mathbb{E}||U||^2||(\alpha I + T^*T)^{-1}T^*||^2_{HS} = \mathbb{E}||U||^2 \sum_{j=1}^\infty ||(\alpha I + T^*T)^{-1}T^* u_j||^2 \\
&\leq \mathbb{E}||U||^2 \sum_{j=1}^\infty ||(T^*)^{-1}u_j||^2 \leq \mathbb{E}||U||^2 \underline{m}^{-2} \sum_{j=1}^\infty ||\left[\phi(L^{-2})\right]^{-1/2} u_j||^2 \\
&\leq \mathbb{E}||U||^2 \underline{m}^{-2} C_{\bar{\nu}} \sum_{j=1}^{\bar{\nu}} \left[\phi(\rho_j^{-2})\right]^{-1} ||u_j||^2 \qquad \text{(A.7)}
\end{aligned}$$

where $||\cdot||_{HS}$ denotes the Hilbert-Schmidt norm of an operator and the last inequality follows by duality from Assumption 10 and $C_{\bar{\nu}}$ is a function of $\bar{\nu}$ which converges to 1 as $\bar{\nu} \to \infty$.

- *Mildly ill-posed case:* we replace $\phi(t) = t^a$, $a > 0$ and $\rho_j = j^{1/d}$ in (A.7) and we get

$$\mathcal{A}_1 \asymp \mathbb{E}||U||^2 \sum_{j=1}^{\bar{\nu}} j^{2a/d} \asymp \mathbb{E}||U||^2 \left(\bar{\nu}^{2a/d+1}\right).$$

If we choose $\bar{\nu} = \alpha^{-d/(2a)}$ then: $\mathcal{A}_1 \asymp \mathbb{E}||U||^2(\alpha^{-1-d/(2a)})$. Moreover, $\mathcal{A}_1 \asymp \mathcal{A}_2$ if and only if $\alpha \asymp \left(\mathbb{E}||U||^2\right)^{a/(a\tilde{\beta}+a+d/2)}$ which gives the minimax rate $\left(\mathbb{E}||U||^2\right)^{a\tilde{\beta}/(a\tilde{\beta}+a+d/2)}$ for the MISE.

- *Severely ill-posed case:* we replace $\phi(t) = \exp(-t^{-a/2})$, $a > 0$ and $\rho_j = j^{1/d}$ in (A.7) and we get

$$\mathcal{A}_1 \asymp \mathbb{E}||U||^2 \sum_{j=1}^{\bar{\nu}} \left[e^{-\rho_j^a}\right]^{-1} \asymp \mathbb{E}||U||^2 \exp\left(\bar{\nu}^{a/d}\right).$$

If we choose $\bar{\nu} = \alpha^{-d}$ then: $\mathcal{A}_1 \asymp \mathbb{E}||U||^2 \exp(\alpha^{-a})$. Moreover, $\mathcal{A}_1 \asymp \mathcal{A}_2$ if and only if $\alpha \asymp c\left(-\log(\mathbb{E}||U||^2)\right)^{-1/a}$ with a sufficiently small $c > 0$. This gives the minimax rate $\left(-\log(\mathbb{E}||U||^2)\right)^{-\tilde{\beta}/a}$ for the MISE.

A.8 Proof of Corollary 1

Following the decomposition (A.6) in the proof of Theorem 2 the upper bound for $\mathcal{A}_2$ remains unchanged while term $\mathcal{A}_1$ is now bounded above by $\mathcal{A}_1 \leq ||(\alpha I + T^*T)^{-1}||^2 \mathbb{E}||T^*(\hat{f}_{C|WZ} - f_{C|WZ})||^2 = \mathcal{O}\left(\alpha^{-2}\mathbb{E}||T^*(\hat{f}_{C|WZ} - f_{C|WZ})||^2\right)$. We have to compute the rate of $\mathbb{E}||T^*(\hat{f}_{C|WZ} - f_{C|WZ})||^2$. First, remark that $\mathbb{E}||T^*(\hat{f}_{C|WZ} - f_{C|WZ})||^2 = \int_\Theta \left(Var(T^*\hat{f}_{C|WZ}) + (\mathbb{E}(T^*\hat{f}_{C|WZ}) - T^*f_{C|WZ})^2\right)\pi_\theta(\theta)d\theta$. Moreover, $\hat{f}_{C|WZ} - f_{C|WZ} = \frac{1}{\hat{f}_{WZ}}(\hat{f}_{CWZ} - f_{C|WZ}\hat{f}_{WZ})[1 - (\hat{f}_{WZ} - f_{WZ})/\hat{f}_{WZ}]$ and since $(\hat{f}_{WZ} - f_{WZ})/\hat{f}_{WZ} = o_p(1)$ we can use the approximation $\hat{f}_{C|WZ} - f_{C|WZ} \simeq \frac{1}{f_{WZ}}(\hat{f}_{CWZ} - f_{C|WZ}\hat{f}_{WZ})$.

Let t be a k-dimensional vector and v a l-dimensional vector; we use the notation $\overrightarrow{vt} := (v', t')$ and $\overrightarrow{uvt} = (u, v', t')$. Moreover, we denote $p = k+l$, $D^2(h)$ is the Hessian matrix of a function h and we use a single integral $\int \cdot$ to denote the multiple integral either in $dvdt$ or in $dudvdt$. We start by computing the bias term $b(w, \theta) := \mathbb{E}(T^*\hat{f}_{C|WZ} - T^*f_{C|WZ})$. By standard Taylor series approximations we get:

$$b(w,\theta) \simeq T^*\frac{1}{f_{WZ}}(\mathbb{E}(\hat{f}_{CWZ}) - f_{C|WZ}\mathbb{E}(\hat{f}_{WZ})) = T^*\frac{1}{f_{WZ}}\left((\mathbb{E}(\hat{f}_{CWZ}) - f_{CWZ}) + f_{C|WZ}(f_{WZ} - \mathbb{E}(\hat{f}_{WZ}))\right);$$

$$\mathbb{E}(\hat{f}_{CWZ}) - f_{CWZ} = \frac{h_n^2}{2}tr\left(D^2(f_{CWZ})\int \overrightarrow{uvt}'\overrightarrow{uvt}K(u,c)K(v,z)K(t,w)dudvdt\right) + o(h_n^2);$$

$$\mathbb{E}(\hat{f}_{WZ}) - f_{WZ} = \frac{h_d^2}{2}tr\left(D^2(f_{WZ})\int \overrightarrow{vt}'\overrightarrow{vt}K(v,z)K(t,w)dvdt\right) + o(h_d^2);$$

$$\begin{aligned} b(w,\theta) \simeq & \int_{\mathcal{C}}\int_{\mathbb{R}^l} \frac{f_{C|WZ\theta}}{f_{WZ}}\Big[h_n^2 tr\left(D^2(f_{CWZ})(c,w,z)\int \overrightarrow{uvt}'\overrightarrow{uvt}K(u,c)K(v,z)K(t,w)dudvdt\right)dcdz \\ & - h_d^2 tr\left(D^2(f_{WZ})(w,z)\int \overrightarrow{vt}'\overrightarrow{vt}K(v,z)K(t,w)dvdt\right)\Big]\frac{\pi_{cz}(c,z)}{\pi_\theta(\theta)}dcdz + o\left(\max\{h_n^2, h_d^2\}\right) \\ := & \, h_n^2 b_1(w,\theta) - h_d^2 b_2(w,\theta) + o\left(\max\{h_n^2, h_d^2\}\right) \end{aligned}$$

Therefore, $b^2(w,\theta) = \mathcal{O}\left(\max\{h_n^4, h_d^4\}\right)$. Then we consider the variance term.

$$Var(T^*\hat{f}_{C|WZ}) = Var(T^*(\hat{f}_{C|WZ} - f_{C|WZ})) \simeq Var\left(T^*\frac{1}{f_{WZ}}(\hat{f}_{CWZ} - f_{C|WZ}\hat{f}_{WZ})\right) \tag{A.8}$$

$$= Var\left(T^*\frac{\hat{f}_{CWZ}}{f_{WZ}}\right) + Var\left(T^*\frac{f_{C|WZ}\hat{f}_{WZ}}{f_{WZ}}\right) - 2Cov\left(T^*\frac{\hat{f}_{CWZ}}{f_{WZ}}, T^*\frac{f_{C|WZ}\hat{f}_{WZ}}{f_{WZ}}\right);$$

$$Var\left(T^*\frac{\hat{f}_{CWZ}}{f_{WZ}}\right) = Var\Big(\int_{\mathbb{R}^l}\int_{\mathcal{C}}\frac{f_{C|WZ\theta}(c;w,z,\theta)}{f_{WZ}(w,z)nh_n^p}\sum_{i=1}^{n}\frac{K_h(c_i-c,c)}{h_n}K_h(z_i-z,z)K_h(w_i-w,w)\frac{\pi_{cz}(c,z)}{\pi_\theta}dcdz\Big)$$

$$= Var\Big(\frac{1}{nh_n^k}\sum_{i=1}^{n}f_{C|WZ\theta}(c_i;w,z_i,\theta)\frac{\pi_{cz}(c_i,z_i)}{f_{WZ}(w,z_i)}\frac{K_h(w_i-w,w)}{\pi_\theta}\Big) + o\left((nh_n^k)^{-1}\right)$$

$$= \frac{1}{nh_n^{2k}}\int f_{C|WZ\theta}^2(c_i;w,z_i,\theta)\frac{\pi_{cz}^2(c_i,z_i)}{f_{WZ}^2(w,z_i)}\frac{K_h^2(w_i-w,w)}{\pi_\theta^2}f_{CWZ}(c_i,w_i,z_i)dc_idw_idz_i \tag{A.9}$$

$$- \frac{1}{nh_n^{2k}}\left(\int f_{C|WZ\theta}(c_i;w,z_i,\theta)\frac{\pi_{cz}(c_i,z_i)}{f_{WZ}^2(w,z_i)}\frac{K_h(w_i-w,w)}{\pi_\theta}f_{CWZ}(c_i,w_i,z_i)dc_idw_idz_i\right)^2 + o\left((nh_n^k)^{-1}\right)$$

$$= \frac{1}{nh_n^k}\int f_{C|WZ\theta}^2(c_i;w,z_i,\theta)\frac{\pi_{cz}^2(c_i,z_i)}{f_{WZ}(w,z_i)}\frac{\int K^2(t,w)dt}{\pi_\theta^2}f_{C|WZ}(c_i;w,z_i)dc_idz_i + o\left((nh_n^k)^{-1}\right); \tag{A.10}$$

$$Var\Big(T^*\frac{f_{C|WZ}\hat{f}_{WZ}}{f_{WZ}}\Big) = Var\left(\int_{\mathbb{R}^l}\int_{\mathcal{C}}\frac{f_{C|WZ\theta}(c;w,z,\theta)}{f_{WZ}(w,z)nh_d^p}\sum_{i=1}^{n}f_{C|WZ}(c;w,z)K_h(z_i-z,z)K_h(w_i-w,w)\frac{\pi_{cz}(c,z)}{\pi_\theta}dcdz\right)$$

$$= Var\left(\frac{1}{nh_d^k}\sum_{i=1}^{n}\int_{\mathcal{C}}\frac{f_{C|WZ\theta}(c;w,z_i,\theta)}{f_{WZ}(w,z_i)}f_{C|WZ}(c;w,z_i)K_h(w_i-w,w)\frac{\pi_{cz}(c,z_i)}{\pi_\theta}dc\right) + o\left((nh_d^k)^{-1}\right)$$

$$= \frac{1}{nh_d^k}\int_{\mathcal{Z}}\left(\int_{\mathcal{C}}\frac{f_{C|WZ\theta}(c;w,z_i,\theta)}{f_{WZ}(w,z_i)}f_{C|WZ}(c;w,z_i)\pi_{cz}(c,z_i)dc\right)^2 \times$$

$$\frac{\int K^2(t,w)dt}{\pi_\theta^2}f_{WZ}(w,z_i)dz_i + o\left((nh_d^k)^{-1}\right), \tag{A.11}$$

where the results are obtained by standard Taylor series approximations. Finally, we have to compute the covariance term:

$$Cov\Big(T^*\frac{\hat{f}_{CWZ}}{f_{WZ}}, T^*\frac{f_{C|WZ}\hat{f}_{WZ}}{f_{WZ}}\Big) = \frac{1}{n^2h_n^kh_d^k}\sum_{i=1}^{n}Cov\Big(\frac{f_{C|WZ\theta}(c_i;w,z_i,\theta)}{f_{WZ}(w,z_i)}K_h(w_i-w,w)\frac{\pi_{cz}(c_i,z_i)}{\pi_\theta},$$

$$\int_{\mathcal{C}}\frac{f_{C|WZ\theta}(c;w,z_i,\theta)}{f_{WZ}(w,z_i)}K_h(w_i-w,w)\frac{\pi_{cz}(c,z_i)}{\pi_\theta}f_{C|WZ}(c;w,z_i)dc\Big) + o\left((n(\min\{h_n,h_d\})^k)^{-1}\right)$$

$$= \frac{1}{nh_d^k}\int\left(\int_{\mathcal{C}}f_{C|WZ\theta}(c;w,z_i,\theta)f_{C|WZ}(c;w,z_i)\frac{\pi_{cz}(c,z_i)}{\pi_\theta}dc\right)^2\frac{1}{f_{WZ}(w,z_i)}dz_i$$

$$\times\int K(t,w)K\left(\frac{th_n}{h_d},w\right)dt + o\left((n(\min\{h_n,h_d\})^k)^{-1}\right). \tag{A.12}$$

By putting (A.10), (A.11) and (A.12) together we obtain

$$
\begin{aligned}
Var(T^*\hat{f}_{C|WZ}) \simeq & \frac{1}{nh_n^k}\int_{\mathcal{Z}} \mathbb{E}\left(f^2_{C|WZ\theta}\pi^2_{cz}\,|w,z_i\right)\frac{\int K^2(t,w)dt}{f_{WZ}(w,z_i)\pi^2_\theta}dz_i + \frac{1}{nh_d^k}\int_{\mathcal{Z}}\left(\mathbb{E}\left(f_{C|WZ\theta}\pi_{cz}\,|w,z_i\right)\right)^2\frac{\int K^2(t,w)dt}{f_{WZ}(w,z_i)\pi^2_\theta}dz_i \\
& - 2\frac{1}{nh_d^k}\int\left(\mathbb{E}(f_{C|WZ\theta}\pi_{cz}\,|w,z_i)\right)^2\frac{1}{f_{WZ}(w,z_i)}dz_i\frac{\int K(t,w)K\left(\frac{th_n}{h_d}\right)dt}{\pi^2_\theta} + o\left(\frac{1}{n(\min\{h_n,h_d\})^k}\right) \\
:= & \frac{1}{nh_n^k}V_1(w,\theta) + \frac{1}{nh_d^k}V_2(w,\theta) - 2\frac{1}{nh_d^k}V_3(w,\theta) + o\left(\frac{1}{n(\min\{h_n,h_d\})^k}\right). \qquad \text{(A.13)}
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\mathbb{E}||T^*(\hat{f}_{C|WZ} - f_{C|WZ})||^2 = & \int_\Theta\Big(\frac{1}{nh_n^k}V_1(w,\theta) + \frac{1}{nh_d^k}V_2(w,\theta) - 2\frac{1}{nh_d^k}V_3(w,\theta) + h_n^4b_1^2(w,\theta) + h_d^4b_2^2(w,\theta) \\
& -2h_n^2h_d^2b_1(w,\theta)b_2(w,\theta)\Big)\pi_\theta d\theta + o\left(\frac{1}{n(\min\{h_n,h_d\})^k}\right) + o\left(\max\{h_n^4,h_d^4\}\right)
\end{aligned}
$$

and the rate of the MISE is:

$$
\mathbb{E}||\mathcal{P}_c\hat{f}^\alpha_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}\left(\alpha^{\beta\wedge 2} + \frac{1}{\alpha^2}\left(\max\{h_n^4,h_d^4\} + \frac{1}{n(\min\{h_n,h_d\})^k}\right)\right).
$$

A.9 Proof of Lemma 4.1

Let us consider the decomposition $(\hat{f}^\alpha_{\theta|W} - f^{\dagger c}_{\theta|W})(\theta;w) = [\hat{f}^\alpha_{\theta|W} - (\alpha I + T^*T)^{-1}T^*\mathbb{E}(f_{C|WZ})](\theta;w) + [(\alpha I + T^*T)^{-1}T^*\mathbb{E}(f_{C|WZ}) - f^\dagger_{\theta|W}](\theta;w) =: A + B$. The result of Lemma 4.1 follows from proving that $\frac{\mathcal{P}_cA}{\sqrt{V_c(\theta;w)}} \to^d \mathcal{N}(0,1)$ and $\frac{\mathcal{P}_cB}{\sqrt{V_c(\theta;w)}} = o_p(1)$. We start by proving that $\frac{A}{\sqrt{V(\theta;w)}} \to^d \mathcal{N}(0,1)$ where $V(\theta;w) = Var(A)$. Let $\{\lambda_j,\varphi_j,\psi_j\}_{j\in\mathbb{N}}$ denote the singular value decomposition of T, then

$$
\begin{aligned}
[\hat{f}^\alpha_{\theta|W} - (\alpha I + T^*T)^{-1}T^*\mathbb{E}(\hat{f}_{C|WZ})](\theta;w) &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^\infty\frac{1}{\alpha+\lambda_j^2} < T^*(\hat{f}_{C|WZ} - \mathbb{E}(\hat{f}_{C|WZ})),\varphi_j > \varphi_j(\theta;w) \\
&\simeq \sum_{j=1}^\infty\frac{1}{\alpha+\lambda_j^2} < T^*\frac{1}{\mathbb{E}(\hat{f}_{WZ})}\left(\hat{f}_{CWZ} - \frac{\mathbb{E}(\hat{f}_{CWZ})}{\mathbb{E}(\hat{f}_{WZ})}\hat{f}_{WZ}\right),\varphi_j > \varphi_j(\theta;w) \\
&= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^\infty\frac{1}{\alpha+\lambda_j^2} < T^*\frac{1}{\mathbb{E}(\hat{f}_{WZ})}\left(\frac{K_h(c_i-c,c)}{h_n^{k+l+1}} - \frac{\mathbb{E}(\hat{f}_{CWZ})}{\mathbb{E}(\hat{f}_{WZ})h_d^{k+l}}\right)K_h(z_i-z,z)K_h(w_i-w,w),\varphi_j > \varphi_j(\theta;w) \\
&=: \frac{1}{n}\sum_{i=1}^n Z_{ni}.
\end{aligned}
$$

By a triangular array version of the Liapounov's central limit theorem it follows that

$$
\frac{1}{n}\sum_{i=1}^n Z_{ni}/\sqrt{n^{-1}Var(Z_{ni})} \to^d \mathcal{N}(0,1)
$$

if $\sum_{i=1}^{n}\mathbb{E}\left|Z_{ni}/\sqrt{nVar(Z_{ni})}\right|^3 \to 0$ as $n\to\infty$. Therefore, we have to prove this latter convergence. We use the notation $h=\min\{h_n,h_d\}$. Remark that

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}\left|Z_{ni}/\sqrt{nVar(Z_{ni})}\right|^3 &= n(nVar(Z_{n1}))^{-\frac{3}{2}}\mathbb{E}|Z_{n1}|^3\\ &= n^{-\frac{1}{2}}(Var(Z_{n1}))^{-\frac{3}{2}}\mathbb{E}|Z_{n1}|^3 \end{aligned} \tag{A.14}$$

and $\mathbb{E}|Z_{n1}|^3$ is bounded above by

$$\begin{aligned}\mathbb{E}|Z_{n1}|^3 &= \mathbb{E}\left|\sum_{j=1}^{\infty}\frac{1}{\alpha+\lambda_j^2}<T^*\frac{1}{\mathbb{E}(\hat{f}_{WZ})}\left(\frac{K_h(c_1-c,c)}{h_n^{k+l+1}}-\frac{\mathbb{E}(\hat{f}_{CWZ})}{\mathbb{E}(\hat{f}_{WZ})h_d^{k+l}}\right)K_h(z_1-z,z)K_h(w_1-w,w),\varphi_j>\varphi_j(\theta;w)\right|^3\\ &= \mathbb{E}\Big|\sum_{j=1}^{\infty}\frac{1}{\alpha+\lambda_j^2}<\left(\frac{f_{C|WZ\theta}\pi_{cz}(c_1;w,z_i,\theta)}{\mathbb{E}(\hat{f}_{WZ})(w,z_1)h_n^k}-\frac{\int_{\mathcal{C}}\mathbb{E}(\hat{f}_{CWZ})(c,w,z_1)f_{C|WZ\theta}\pi_{cz}(c;w,z_1,\theta)dc}{\mathbb{E}^2(\hat{f}_{WZ})(w,z_1)\pi_\theta h_d^k}\right)\times \end{aligned} \tag{A.15}$$

$$\begin{aligned}&\left[K_h(w_1-w,w)+\mathcal{O}(h^2)\right],\varphi_j>\varphi_j(\theta;w)\Big|^3\\ &\le \frac{1}{\alpha^3h^{2k}}\mathbb{E}\frac{1}{h^k}\Big|\left(\frac{f_{C|WZ\theta}\pi_{cz}(c_1;w,z_1,\theta)h^k}{\mathbb{E}(\hat{f}_{WZ})(w,z_1)h_n^k}-\frac{\int_{\mathcal{C}}\mathbb{E}(\hat{f}_{CWZ})(c,w,z_1)f_{C|WZ\theta}\pi_{cz}(c;w,z_1,\theta)dch^k}{\mathbb{E}^2(\hat{f}_{WZ})(w,z_1)\pi_\theta h_d^k}\right)\times \end{aligned} \tag{A.16}$$

$$\begin{aligned}&K_h(w_i-w,w)\Big|^3\\ &:= \frac{1}{\alpha^3h^{2k}}A_2\end{aligned}$$

where A_2 is finite under Assumption 11. Therefore (A.14) becomes

$$\sum_{i=1}^{n}\mathbb{E}\left|Z_{ni}/\sqrt{nVar(Z_{ni})}\right|^3 \le \left(nh^k\right)^{-1/2}\left(h^k\alpha^2Var(Z_{n1})\right)^{-\frac{3}{2}}A_2.$$

Now we have to analyze $h^k\alpha^2Var(Z_{n1})$.

$$\begin{aligned}Var(Z_{n1}) &= \sum_{j=1}^{\infty}\frac{1}{(\alpha+\lambda_j^2)^2}Var\left(<T^*\frac{1}{\mathbb{E}(\hat{f}_{WZ})}\left(\frac{K_h(c_1-c,c)}{h_n^{k+l+1}}-\frac{\mathbb{E}(\hat{f}_{CWZ})}{\mathbb{E}(\hat{f}_{WZ})h_d^{k+l}}\right)K_h(z_1-z,z)K_h(w_1-w,w),\varphi_j>\right)\varphi_j^2\\ &\quad+2\sum_{j>m}^{\infty}\frac{1}{(\alpha+\lambda_j^2)(\alpha+\lambda_m^2)}Cov\Big(<T^*\frac{1}{\mathbb{E}(\hat{f}_{WZ})}\left(\frac{K_h(c_1-c,c)}{h_n^{k+l+1}}-\frac{\mathbb{E}(\hat{f}_{CWZ})}{\mathbb{E}(\hat{f}_{WZ})h_d^{k+l}}\right)K_h(z_1-z,z)\times\\ &\quad K_h(w_1-w,w),\varphi_j>,\\ &\quad<T^*\frac{1}{\mathbb{E}(\hat{f}_{WZ})}\left(\frac{K_h(c_1-c,c)}{h_n^{k+l+1}}-\frac{\mathbb{E}(\hat{f}_{CWZ})}{\mathbb{E}(\hat{f}_{WZ})h_d^{k+l}}\right)K_h(z_1-z,z)K_h(w_1-w,w),\varphi_m>\Big)\varphi_j\varphi_m\\ &= \sum_{j=1}^{\infty}\frac{1}{(\alpha+\lambda_j^2)^2}<\int_\Theta\left[\frac{V_1(w,\theta)\pi_\theta(\theta)}{h_n^k\pi_\theta(\tilde{\theta})}+\frac{V_4(w,\theta)}{h_d^k}-\frac{2V_5(w,\theta)}{h_d^k}+o\left(\frac{1}{h^k}\right)\right]\varphi_j(\theta),\varphi_j(\tilde{\theta})>\varphi_j^2(\theta;w)\\ &\quad+2\sum_{j>m}^{\infty}\frac{1}{(\alpha+\lambda_j^2)(\alpha+\lambda_m^2)}<\int_\Theta\left[\frac{V_1(w,\theta)\pi_\theta(\theta)}{h_n^k\pi_\theta(\tilde{\theta})}+\frac{V_4(w,\theta)}{h_d^k}-\frac{2V_5(w,\theta)}{h_d^k}+o\left(\frac{1}{h^k}\right)\right]\varphi_j(\theta),\varphi_m(\tilde{\theta})>\\ &\quad\times\varphi_j(\theta;w)\varphi_m(\theta;w).\end{aligned}$$

where $V_1(\theta, w)$ was defined in (A.13),

$$V_4(\theta, w) = \int \left(\int_{\mathcal{C}} f_{C|WZ\theta} \pi_{cz}(c, w, z_1, \theta) \frac{\mathbb{E}(\hat{f}_{CWZ})(c, w, z_1)}{\mathbb{E}^2(\hat{f}_{WZ})(w, z_1)} dc \right)^2 \int K^2(t) dt \frac{f_{WZ}(w, z_1) dz_1}{\pi_\theta(\theta)}$$

and $V_5(\theta, w) = \int \left(\int_{\mathcal{C}} f_{C|WZ\theta} \pi_{cz}(c, w, z_1, \theta) \frac{\mathbb{E}(\hat{f}_{CWZ})(c,w,z_1)}{\mathbb{E}^2(\hat{f}_{WZ})(w,z_1)} dc \right)^2 \int K(t, w) K(\frac{th_n}{h_d}, w) dt \frac{f_{WZ}(w,z_1)dz_1}{\pi_\theta(\theta)}$. Since, $\forall \alpha \neq 0$ and $\forall j \in \mathbb{N}$ we have that $(1 - \lambda_j^2/(\alpha + \lambda_j^2))^2 > 0$ and since the other terms in $Var(Z_{n1})$ not involving α are strictly positive, it is clear that $h^k \alpha^2 Var(Z_{n1}) > 0$. This shows that, under the condition that $nh^k \to \infty$, $\sum_{i=1}^n \mathbb{E} \left| Z_{n1} / \sqrt{nVar(Z_{n1})} \right|^3 \to 0$. This proves the first result that $\frac{A}{\sqrt{V(\theta;w)}} \to^d \mathcal{N}(0, 1)$. To prove $\frac{\mathcal{P}_c A}{\sqrt{V_c(\theta;w)}} \to^d \mathcal{N}(0, 1)$ we use the functional delta method (see van der Vaart (1998) Theorem 20.8). This requires that the projection operator $\mathcal{P}_c$ is Hadamard differentiable. The (one-sided) Hadamard derivative of $\mathcal{P}_c$ in $f^{\dagger c}_{\theta|W}$ is a projection as well, denoted by $\mathcal{P}_c^\dagger$, that projects on the tangent cone of $\mathcal{F}_{\theta|W}$ at $f^{\dagger c}_{\theta|W}$ defined as in Lemma 4.1.
To prove the second result, let us decompose B as

$$B = (\alpha I + T^*T)^{-1} T^* \left(\mathbb{E}(\hat{f}_{C|WZ}) - f_{C|WZ} \right)(\theta; w) - \left((\alpha I + T^*T)^{-1} T^* f_{C|WZ} - f_{\theta|W} \right)(\theta; w) := B_1 + B_2.$$

Therefore,

$$\frac{B_1}{\sqrt{V(\theta;w)}} = \frac{[(\alpha I + T^*T)^{-1} b(w, \theta)]}{\sqrt{V(\theta;w)}} = \mathcal{O}\left(\max\{h_n^2, h_d^2\} (nh^k)^{\frac{1}{2}} \right)$$
$$\frac{B_2}{\sqrt{V(\theta;w)}} = \frac{[\alpha(\alpha I + T^*T)^{-1} f_{\theta|W}](\theta; w)}{\sqrt{V(\theta;w)}} = \mathcal{O}\left(\alpha^{(\frac{\beta}{2} \wedge 1) - 1} (nh^k)^{\frac{1}{2}} \right)$$

which converge to 0 if α, h_n and h_d converges to 0 faster then the optimal ones (set as in Lemma 4.1). Since $\mathcal{P}_c$ is a nonexpansive map, these rates are not affected by replacing B_1 and B_2 by $\mathcal{P}_c B_1$ and $\mathcal{P}_c B_2$, respectively. Moreover, $V(\theta; w)$ and $V_c(\theta; w)$ have the same rate.

A.10 Proof of Theorem 4

(i) See proof of Theorem 2.4 in Neubauer (1988).

(ii) See proof of Theorem 2.7 in Neubauer (1988).

References

Attanasio, O.P. and G., Weber (2010). "Consumption and saving: models of intertemporal allocation and their implications for public policy", NBER Working Paper.

Abbring, J., and J., Heckman (2007), "Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation", *Handbook of Econometrics*, **6**, 5145-5303.

Bajari, P., Fox, J.T., Kim, K. and S., Ryan (2012). "The random coefficients logit model is identified", *Journal of Econometrics*, **166**, 204-212.

Beran, R., Feuerverger, A., and Hall, P. (1996). "On nonparametric estimation of intercept and slope distributions in random coefficient regression", *The Annals of Statistics*, **24**, 2569-2592.

Bissantz, N., Hohage, T., Munk, A. and F., Ruymgaart (2007). "Convergence rates of general regularization methods for statistical inverse problems and applications", *SIAM J. Numerical Analysis*, **45**, 2610-2636.

Blundell, R., Chen, X. and Kristensen, D. (2007). "Semi-nonparamtric IV estimation of shape-invariant engel curves", *Econometrica*, **75**, 1613-1669.

Bonhomme, S. (2011). "Functional differencing", *Econometrica*, forthcoming.

Campo, S., Guerre, E., Perrigne, I. and Q., Vuong (2011). "Semiparametric estimation of first-price auctions with risk averse bidders", *Review of Economic Studies*, **78**, 112-147.

Carrasco, M. and J.P., Florens (2000), "Generalization of GMM to a continuum of moment conditions", *Econometric Theory*, **16**, 797-834.

Carrasco, M. and J.P., Florens (2010). "Spectral Method for Deconvolving a Density", *Econometric Theory*, **27**, 546-581.

Carrasco, M., Florens, J.P., and E., Renault (2007). "Linear inverse problems in structural econometrics: estimation based on spectral decomposition and regularization", in: J., Heckman and E., Leamer, (Eds.), *Handbook of Econometrics*, Vol.6B, 5633-5751. Elsevier, North Holland.

Darolles, S., Fan, Y., Florens, J.P., and E., Renault (2011). "Nonparametric instrumental regression", *Econometrica*, **79**, 1541-1565.

Deaton, A., (1993). "Understanding Consumption".

Engl, H.W., Hanke, M. and A., Neubauer (2000). *Regularization of inverse problems*. Kluwer Academic, Dordrecht.

Escanciano, J.C., and S., Hoderlein (2010), "Nonparametric Identification of Euler Equations", unpublished manuscript.

Florens, J.P., (2003). "Inverse problems and structural econometrics: the example of instrumental variables". Invited Lectures to the World Congress of the Econometric Society, Seattle 2000. In: M., Dewatripont, L.-P., Hansen, and S.J., Turnovsky, (Eds.), *Advances in Economics end econometrics: theory and applications*, Vol. II, pp. 284-311. Cambridge University Press.

Florens, J.P., Johannes, J. and S., Van Bellegem (2010). "Identification and estimation by penalization in nonparametric instrumental regression", *Econometric Theory*, **27**, 472-496.

Florens, J.P., Mouchart, M. and J.M., Rolin (1990). *Elements of Bayesian statistics*. Dekker, New York.

Fox, J. and A. Gandhi (2011). "Nonparametric identification and estimation of random coefficients in multinomial choice models", *mimeo*.

Gajek, L. (1986). "On improving density estimators which are not bona fide functions", *Annals of Statistics*, **14**, 1612-1618.

Gautier, E. and Y. Kitamura (2010). "Nonparametric estimation in random coefficients binary choice models". Preprint.

Hall, R.E. (1978). "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence", *Journal of Political Economy*, **86**, 971-87.

Hall, P. and J., Horowitz (2005). "Nonparametric methods for inference in the presence of instrumental variables", *Annals of Statistics* **33**, 2904-2929.

Heckman, J.J., and Singer, B. (1984). "A method for minimizing the impact of distributional assumptions in econometric models for duration data", *Econometrica*, **52**. 271-320.

Henry, M., Kitamura, Y. and B. Salanié (2011). "Identifying finite mixtures in econometric models", Preprint.

Hoderlein, S., Klemelae, J. and Mammen, E. (2010). "Analyzing the Random Coefficient Model Nonparametrically", *Econometric Theory*, forthcoming.

Hoderlein, S., Nesheim, L. and Simoni, A. (2012). "Heterogeneous Euler Equations: a Semiparametric structural approach", mimeo.

Hong, H., and Shum, M. (2009) "Pairwise-Difference Estimation of a Dynamic Optimization Model", *Review of Economic Studies*, **77**, 273-304.

Hu, Y. and S., Schennach (2008). "Instrumental variable treatment of nonclassical measurement error models", *Econometrica*, **76**, 195-216.

Ichimura, H. and Thompson, T.S. (1998), "Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution," *Journal of Econometrics*, 86(2): 269-295.

Kasahara, H. and K. Shimotsu (2009). "Nonparametric identification of finite mixture models of dynamic discrete choices", *Econometrica*, **77**, 135-175.

Keane, M.P. and K.I., Wolpin (1997). "The Career Decisions of Young Men", *Journal of Political Economy*, **105**, 473-522.

Kress, R. (1999). *Linear integral equation*, Springer.

Mair, B.A. (1994). "Tikhonov regularization for finitely and infinitely smoothing operators", *SIAM J. Math.Anal.*, **25**, 135-47.

Mandelbaum, A. and Rüschendorf, L. (1987). "Complete and symmetrically complete families of distributions", *The Annals of Statistics*, **14**, 1229-1244.

Matzkin, R. (2003). "Nonparametric Estimation of Nonadditive Random Functions", Econometrica, **71**, 1339-1375.

Matzkin, R. (2007a), "Nonparametric Identification", in: J., Heckman and E., Leamer, (Eds.), *Handbook of Econometrics*, Vol.6B, 5307-5368. Elsevier, North Holland.

Matzkin, R. (2007b), "Heterogeneous Choice", *Econometric Society Monographs*, 43: 1-75.

Nair, M.T., Pereverzev, S.V. and Tautenhahn, U. (2005). "Regularization in Hilbert scales under general smoothing conditions", *Inverse Problems*, **21**, 1851-1869.

Neubauer, A. (1988). "Tikhonov regularization of ill-posed linear operator equations on closed convex sets", *J. Approx. Theory*, **53**, 304-320.

Newey, W.K. and J.L., Powell, (2003). "Instrumental variable estimation of nonparametric models", *Econometrica*, **71**, 1565-1578.

Rosenblatt, M. (1969). "Conditional probability density and regression estimators". In *Multivariate Analysis II* (P.R. Shnaiah, ed.) Academic Press, New York, 25-31.

Roussas, G. (1967). "Nonparametric estimation in Markov processes", *Ann. Inst. Statist. Math.*, Vol.21, 73-87.

Roussas, G. (1969). "Nonparametric estimation of the transition distribution function of a Markov process", *Ann. Math. Statist.*, **40**, 1386-1400.

Teicher, H. (1961). "Identifiability of Mixtures", *Ann. Math. Stat.*, **32**, 244-248.

Tikhonov, A.N. (1963). *Regularization of incorrectly posed problems*, Soviet Math. Dokl. 4, 1624-1627.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

B Figures

Figure 1: Example 1. $\mathcal{P}_c\widehat{f}_{\theta|W}$ (lower panel) vs. true $f_{\theta|W}$ (upper panel). $w = 1.2839$

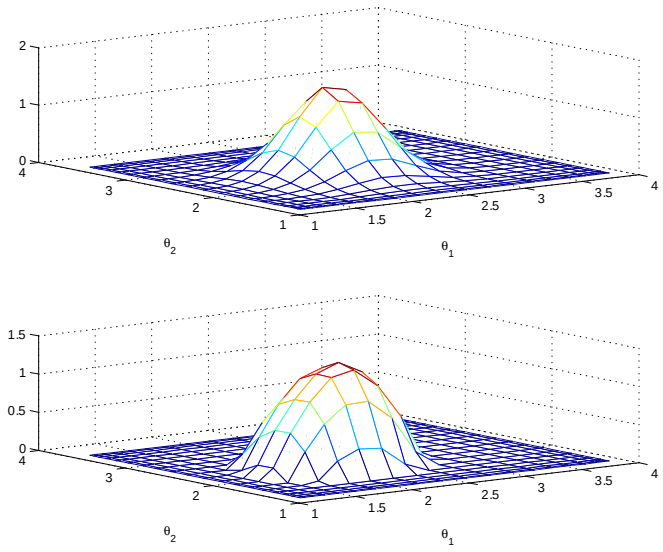

Figure 2: Example 1. $\mathcal{P}_c\widehat{f}_{\theta|W}$ (lower panel) $vs.$ true $f_{\theta|W}$ (upper panel). $w = 1.4849$

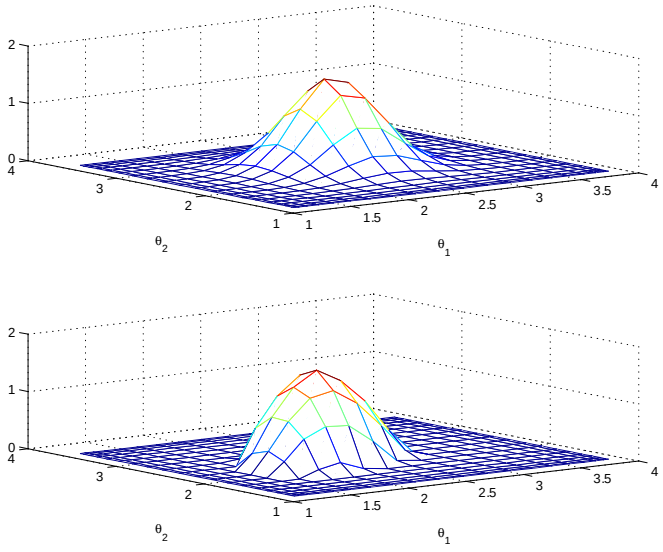

Figure 3: Example 1. $\mathcal{P}_c\widehat{f}_{\theta|W}$ (lower panel) vs. true $f_{\theta|W}$ (upper panel). $w = 1.673$

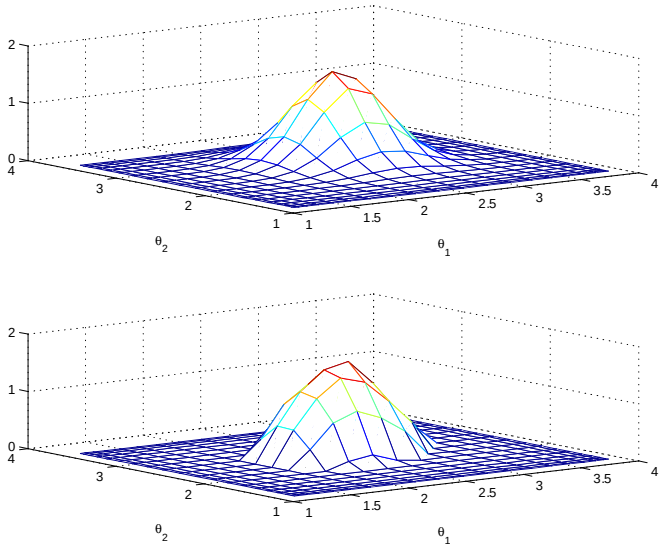

Figure 4: Example 1. $\mathcal{P}_c\widehat{f}_{\theta|W}$ (lower panel) vs. true $f_{\theta|W}$ (upper panel). $w = 1.2839$

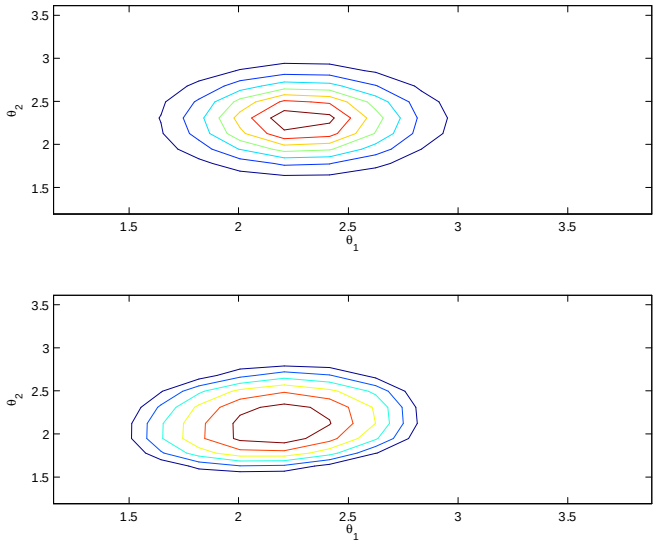

Figure 5: Example 1. $\mathcal{P}_c\widehat{f}_{\theta|W}$ (lower panel) vs. true $f_{\theta|W}$ (upper panel). $w = 1.4849$

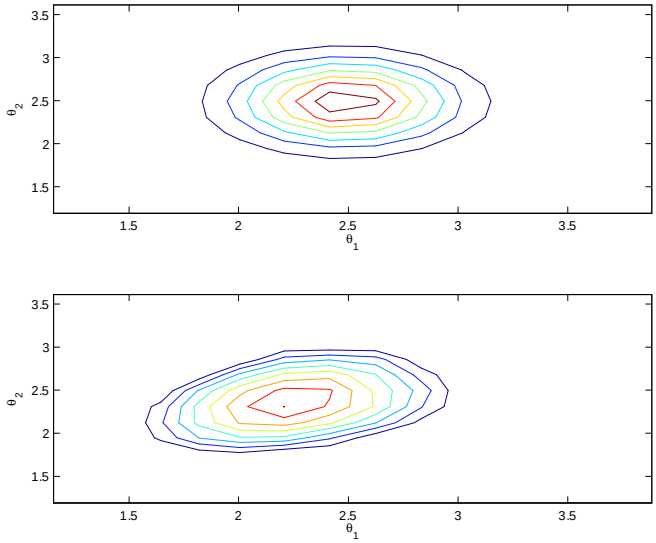

Figure 6: Example 1. $\mathcal{P}_c\widehat{f}_{\theta|W}$ (lower panel) vs. true $f_{\theta|W}$ (upper panel). $w = 1.673$

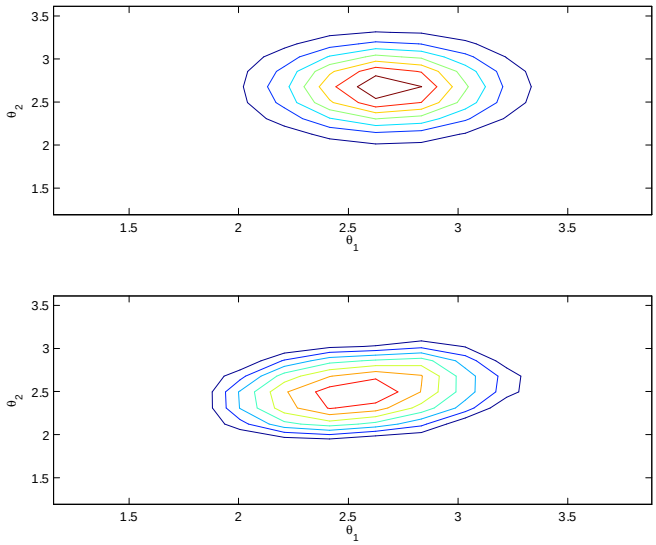

Figure 7: Example 1. Densities of WMSE for the Tikhonov estimator and the oracle estimator

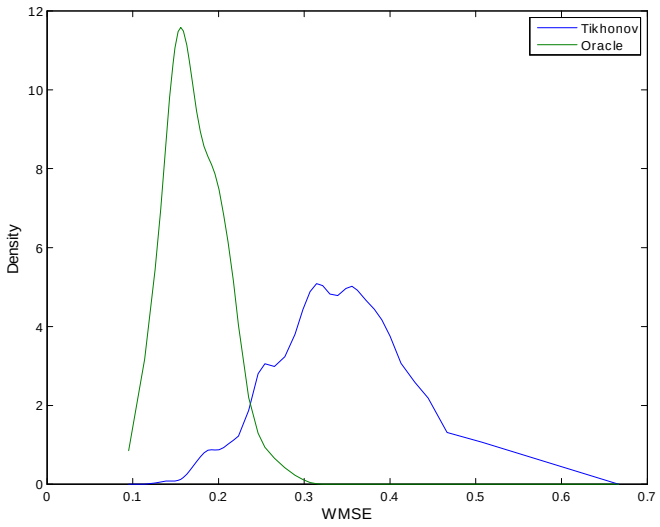

Figure 8: Example 2. Quartiles of income by age

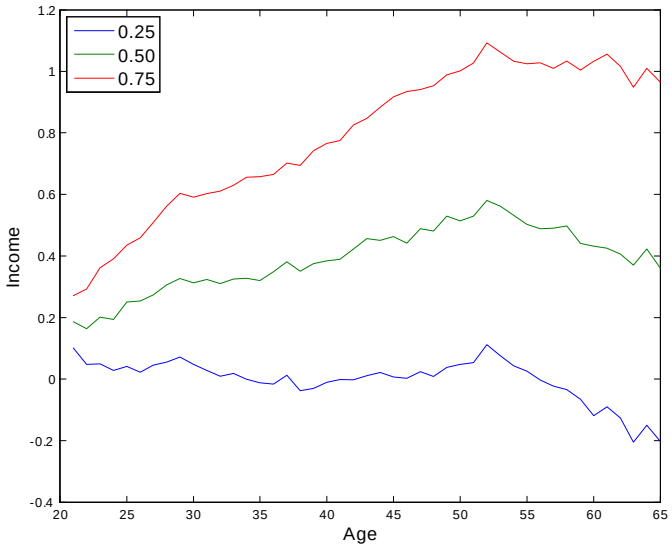

Figure 9: Example 2: Quartiles of consumption by age

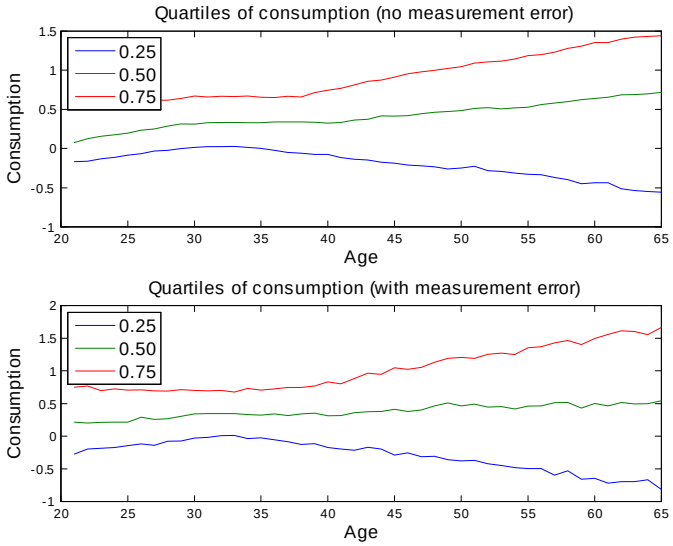

Figure 10: Example 2. Quartiles of assets by age

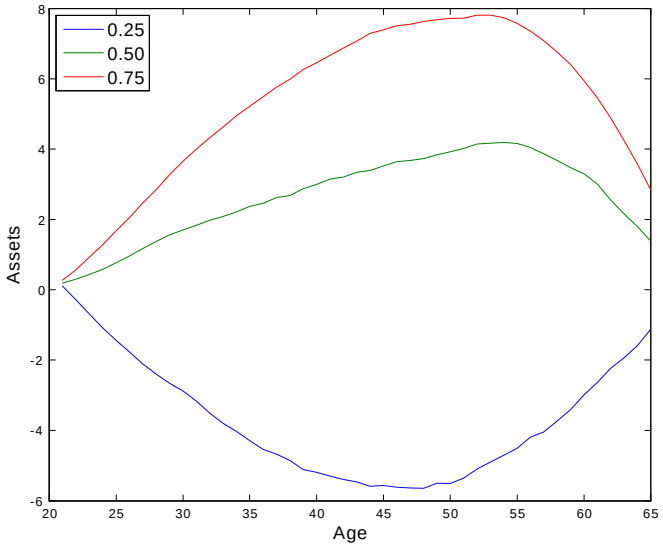

Figure 11: Example 2. Density of δ : (low assets, low income)

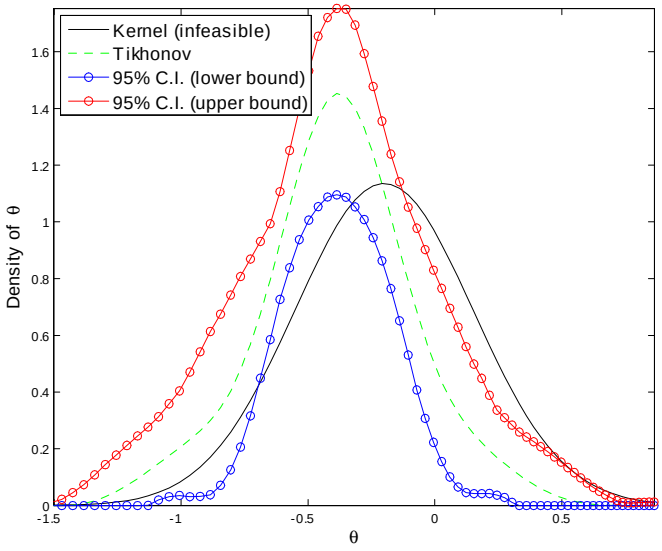

Figure 12: Example 2. Density of δ: (low assets, high income)

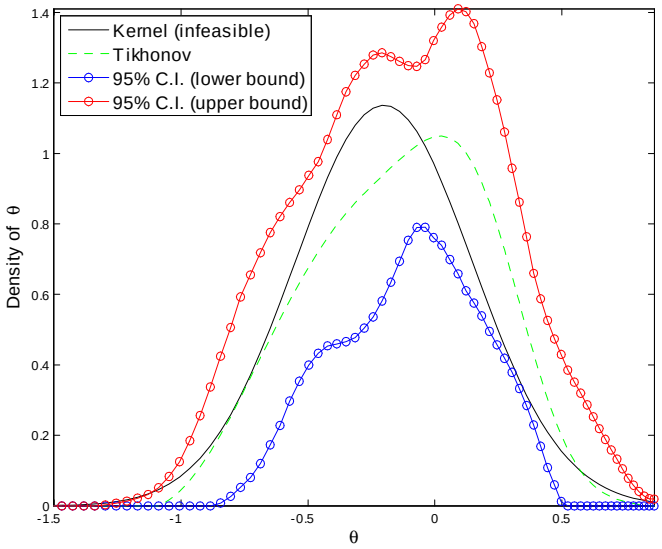

Figure 13: Example 2. Density of δ: (high assets, low income)

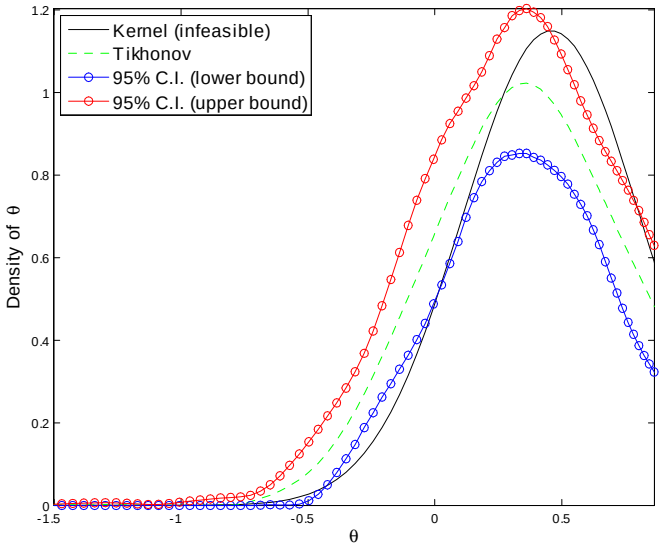

Figure 14: Example 2. Density of δ: (high assets, high income)

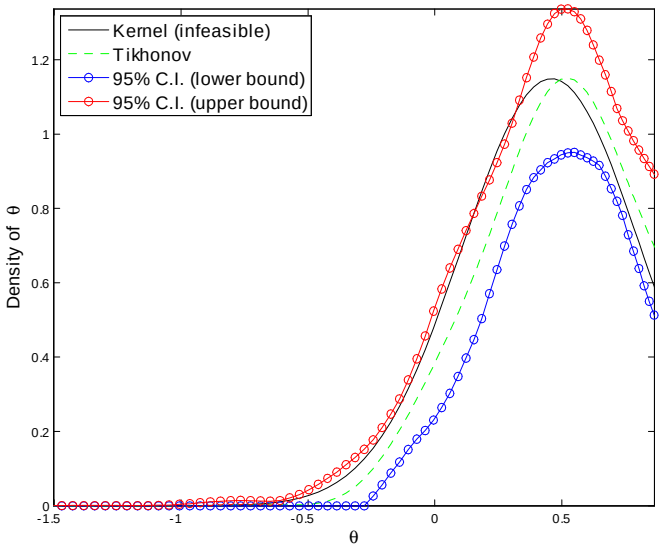

Figure 15: Example 2. Quantile level sets of δ : (low assets, low income)

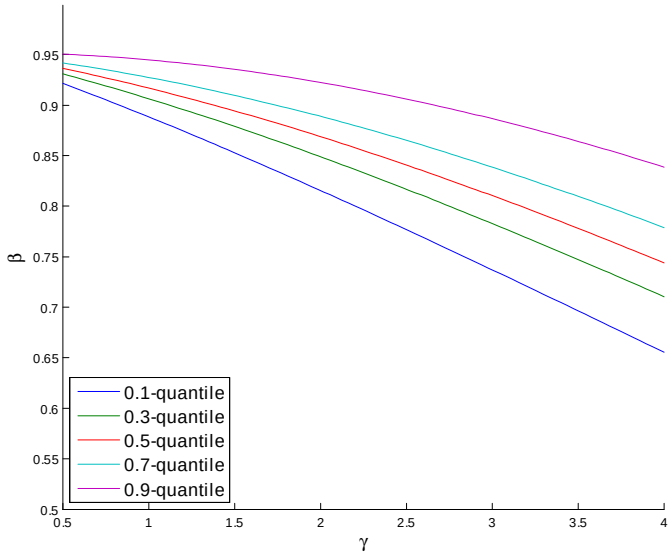

Figure 16: Example 2. Quantile level sets of δ : (low assets, high income)

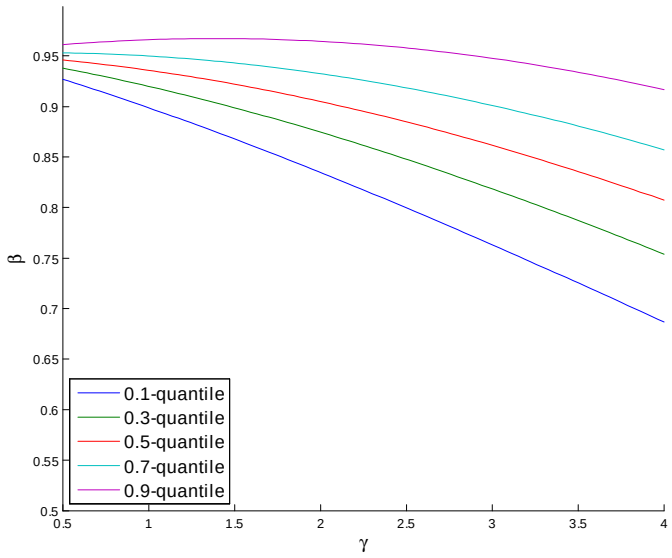

Figure 17: Example 2. Quantile level sets of δ : (high assets, low income)

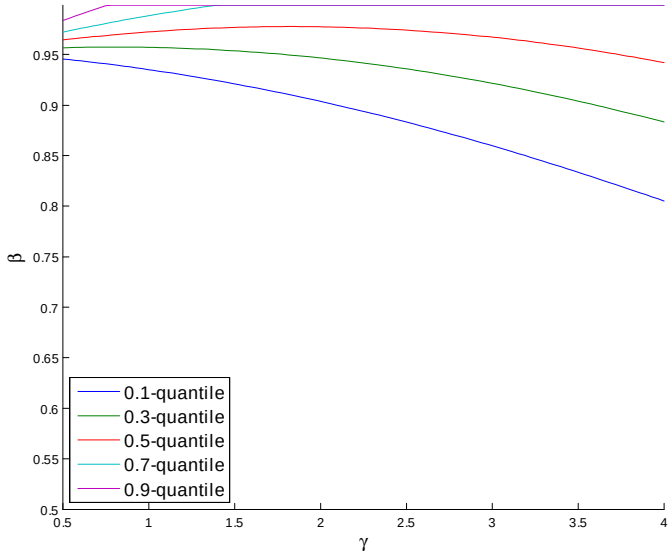

Figure 18: Example 2. Quantile level sets of δ : (high assets, high income)