Semiparametric Estimation of Random Coefficients in Structural Economic Models[1]

Stefan Hoderlein[2]
Boston College

Lars Nesheim [3]
University College London

Anna Simoni[4]
CNRS - CREST

April 14, 2015

[1] We thank the co-editor Yoon-Jae Whang and three anonymous referees whose comments greatly improved the paper. 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 Johns Hopkins University, Mannheim University, NYU, Princeton, UC San Diego, UCLA, and Stanford. 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. Anna Simoni gratefully acknowledges financial support from the University of Mannheim through SFB 884 and from ANR-13-BSH1-0004, and ANR-11-LABEX-0047.

[2] Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA, email: stefan_hoderlein@yahoo.com.

[3] Department of Economics, University College London, Gower Street, London, WC1E 6BT, UK, email: l.nesheim@ucl.ac.uk

[4] CREST, 15 Boulevard Gabriel Péri, 92240 Malakoff (France). email: simoni.anna@gmail.com

Abstract

This paper discusses nonparametric estimation of the distribution of random coefficients in a structural model that is nonlinear in the random coefficients. We establish that the problem of recovering the probability density function (*pdf*) of random parameters falls into the class of convexly-constrained inverse problems. The framework offers an estimation method that separates computational solution of the structural model from estimation. We first discuss nonparametric identification. Then, we propose two alternative estimation procedures to estimate the density and derive their asymptotic properties. 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: Nonlinear random coefficients, mixture models, structural models, heterogeneity, inverse problems.

1 Introduction

Many structural microeconomic models postulate that individual decision makers solve complicated optimization problems governed by a small number of structural parameters θ. While these parameters are fixed for every individual, economic theory does not postulate that they be the same for every individual. Yet, in most empirical applications, the extent to which individual decision makers are allowed to vary is severely constrained, but these constraints on heterogeneity are typically not based on economic theory.

A natural way to relax the constraints and make structural model assumptions more appealing is to assume that the unobservable parameters θ in individuals' decision problems are random parameters drawn from a fully flexible nonparametric continuous distribution that may be correlated with some observable explanatory variables. In this paper, we propose and analyze a method to estimate a nonparametric distribution of random coefficients θ in general structural economic models in which the mapping from random coefficients to outcomes is nonlinear and may only be implicitly defined. We allow the random coefficients θ to be correlated with some of the explanatory variables, analyze identification and propose an estimation method that completely separates computational solution of the economic model from estimation.

To give a stylized example,consider the workhorse Euler equation model of the consumption literature, where for simplicity we have set the discount rate to the interest rate:

$$\partial_c u(C_t, \theta, \varepsilon_t) = \mathbb{E}\left[\partial_c u(C_{t+1}, \theta, \varepsilon_{t+1}) | W_t, Z_t, \theta, \varepsilon_t\right]. \tag{1.1}$$

Here, $\partial_c u$ denotes the derivative of instantaneous utility with respect to consumption, C_t is consumption in period t, and the random parameters θ may include preference parameters such as the coefficient of risk aversion or parameters defining beliefs about future states. Moreover, $(W_t, Z_t, \varepsilon_t)$ are endogeneous observable, exogenous observable, and unobservable state variables, respectively, who affect the decision problem in period t. Equation (1.1) implicitly defines the consumption function $C_t = \varphi_t\left(W_t, Z_t, \theta, \varepsilon_t\right)$. When we lack information about the probability distribution of heterogeneity in the population (for example the *pdf* $f_{\theta|W}$) but have knowledge about the structural equation (so that we can explicitly compute φ), we can use this knowledge to define a mapping from $f_{\theta|W}$ to the population *pdf* of observables $f_{C|WZ}$, which is an integral equation of the form

$$f_{C|WZ} = T f_{\theta|W}, \tag{1.2}$$

where T is a known integral operator derived from the economic model.

This paper makes several contributions. First, it shows how nonlinear random coefficients in a structural economic model can be analyzed using the tools of linear inverse problems theory. Second, it shows how to derive the estimating equation (1.2) from the structural economic model without requiring the structural function φ to be monotonic in ε. Third, based on (1.2), the paper proposes two estimators of $f_{\theta|W}$, derives rates of convergence and shows asymptotic normality. The estimators are based on a simple Tikhonov regularization method modified to impose the constraint that the estimate must be a density function. Our main contributions in this respect are: *(i)* to extend source conditions and provide the rate of convergence for the convexly-constrained Tikhonov estimator in a stochastic setting and *(ii)* to provide and study properties of a step-by-step procedure to compute the orthogonal projection of the unconstrained Tikhonov-regularized estimate onto the set of densities. Fourth, the paper studies nonparametric identification of $f_{\theta|W}$ making use of a notion of completeness (that we call $\mathcal{T}$-completeness) which is weaker than L^2- or bounded-completeness, and characterizing the identified set when the model is not point identified.

This research therefore extends the parametric structural models literature to allow for nonlinear and endogenous random coefficients. This literature is vast. For a recent survey of the consumption literature which originally motivated this research, see **?**. Our analysis provides insights into when identification is only partial and provides novel conditions for point identification. In addition, our analysis makes clear that estimation in these contexts is fundamentally an ill-posed inverse problem.

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 **?**, **?**, **?**, and **?**. In these papers, the random coefficients are continuously distributed and fully independent of explanatory variables. Also related is **?** who considers a linear simultaneous equations model and focuses on estimation of the density of one random slope coefficient. We extend these literatures by allowing for endogenous random coefficients that enter in a nonlinear way and by allowing for models in which the function mapping explanatory variables into outcomes is often only implicitly defined.

Our work is also linked to the mixture models literature following **?** (HS). In a duration model setting, HS analyze the equation

$$f_{C|WZ}(c;w,z) = \int_{\Theta} f_{C|WZ\theta}(c;w,z,\theta) f_{\theta|W}(\theta;w) d\theta. \tag{1.3}$$

They focus on estimating a finite dimensional parameter that impacts the *pdf* $f_{C|WZ\theta}(c;w,z,\theta)$ while treating $f_{\theta|W}$ as a nuisance parameter. Closely related to HS are **?** and **?**. In contrast to these references, our analysis centers its interest on $f_{\theta|W}$, and the kernel of the operator in (1.3) is derived from the economic model.

Our work is also related to the stochastic inverse problem literature. See **?** for an overview. In particular, recovering the probability density of θ from (1.2) is equivalent to solving a convexly constrained integral equation of the first kind. Integral equations of the first kind have been studied extensively in different areas of econometrics (see e.g. **?**, for an overview). These areas include, among other: nonparametric instrumental regression estimation - see e.g. **?**, **?**, **?**, **?**, **?**, **?** and **?** - and moment estimation and deconvolution - see e.g. **?**, **?**, **?** and **?**. There are two important differences between our model and the models studied in these papers. First, the kernel of our integral operator is not estimated but is derived from a structural economic model. Second, we seek to estimate the density of random coefficients not a function of observables.

Our estimating equation is also related to **?**. However, our model differs in many core aspects from their model, not least for the different object of interest (i.e., the distribution of random parameters), and for the structural nonseparability of the model considered. Moreover, our exclusion restrictions are different from theirs (e.g., we do not assume conditional independence of C and Z given θ) and are motivated by the structural economic application.

2 A Roadmap

To illustrate the structure of the model and the main results in this paper, consider a common specification in consumer demand. Let X measure the true log-expenditure on all nondurable goods, and let C^* measure the true log expenditure for one good. Assume that 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. If θ is independent from (Z_1, X) and if (Z_1, X) have support equal to $\mathbb{R}^2$, then the joint density of random coefficients is nonparametrically identified and can be estimated using **?**. However, in consumer demand two important problems arise. First, since X is a choice variable, it is likely to be endogenous, because the same deep preferences parameters that determine C^* also determine X. Second, actual demand C is frequently measured with error, i.e., $C \neq C^*$.

To handle endogeneity, we follow the demand literature and use instruments in a control function fashion. Let Z_2 measure log-income and suppose there is a relation $X = g\left(Z_2, W\right)$, where g is a (identified) function that is strictly monotonic in an uniform unobservable W. Here, W can be obtained as the percentile of log-expenditure conditional on Z_2, and we let $Z = \left(Z_1, Z_2\right)'$, see **?** for such a structure in a demand application. To handle measurement error, we assume that observed C is generated as $C = C^* + \eta$ where $\eta | \theta, W, Z \sim \exp(\lambda)$, with $\lambda > 0$ and assume that $Z \perp (\theta, \eta, W)$. Substituting all elements into the outcome equation for observed C, we obtain

$$C = \theta_1 Z_1 + \theta_2 g(Z_2, W) + \varepsilon, \tag{2.1}$$

where $\varepsilon = \eta + \theta_0$ and $f_{\varepsilon|WZ\theta}(\varepsilon, \theta_0) = \lambda e^{\{-(\varepsilon - \theta_0)\lambda\}}$. This model is a special case of the general class of nonlinear models developed below, and it is useful to illustrate why we consider the specific structure put forward in this paper, and how our assumptions cause identification. First, observe the dual sources of unobserved variation in the model, ε and θ, where the latter is the preference heterogeneity of interest, and the former contains a measurement error. Since it is a nuisance part of our model, we follow the measurement error deconvolution literature and assume a (partially) parametric model for ε. Second, note that X is endogenous due to correlation between W and θ, which prohibits to use standard approaches like **?**. The main idea that we adapt from the control function literature is that conditional on W there is no endogeneity. Since the independence of the instruments from all unobservable implies that $Z \perp \theta|W$, conditional on $W = w$ equation (2.1) becomes exogenous and, provided there is enough variation in $g(Z_2, w)$, the pdf $f_{\theta|W}(\cdot; w)$ is nonparametrically identified, using standard arguments from the random coefficients literature.

Thus, our model introduces two general elements that are novel to this literature and stem from more complex structural models. First, θ may depend on some variables W, while the instruments Z provide exogenous variation in the sense that $Z \perp\!\!\!\perp \theta|\, W$. As we shall see below in the Euler equation and duration examples, allowing for some variables W to be correlated with the heterogeneity of interest while having others, the Z, be (conditionally) independent is a feature that arises more generally. Our strategy is thus to first perform all steps conditional on W, thus recovering $f_{\theta|W}$, and then to obtain f_θ by integrating out W. The case without endogenous variables is obviously a special case, where $Z \perp\!\!\!\perp \theta$, and we can directly obtain f_θ. The instruments Z, however, are generally necessary to identify f_θ, as it is their variation that is mapped into variation of θ.

The second general feature is a composite error ε, comprised of preference heterogeneity θ and a nuisance part η. In the "demand with measurement error" application, as well as in the Euler equation application, this is a necessary feature of the structural model. Our strategy in this part is motivated by the deconvolution literature, and requires a parametric assumption on the nuisance error. While we believe this to be an important feature of many applied models, our approach does not rely on the existence of ε and, as we demonstrate in a Supplementary Appendix, all arguments go through with minor modifications if there is no nuisance unobservable ε.

The rest of the paper formalises these ideas for a general nonlinear model in which $C = \varphi\,(W, Z, \theta, \varepsilon)$. In Theorem 1, we show that there is an integral operator T which maps $f_{\theta|W}$ into $f_{C|WZ}$. After characterising some properties of this operator, in Proposition 2 we characterise the set of solutions of the inverse of T using the singular value decomposition of T. Since φ and f_ε are known, this set can be computed. Next, for the general model, point identification requires a completeness condition on the probability distribution characterizing the operator. The completeness condition we require is weaker than that required in the nonparametric IV literature since our object of interest is a probability density and not an unrestricted function. We call this condition $\mathcal{T}$-completeness and must be established in every application, but we provide a sufficient condition that is easier to check. For instance, in the demand example (2.1), $f_{\varepsilon|WZ\theta}(\varepsilon, \theta_0) = \lambda e^{\{-(\varepsilon-\theta_0)\lambda\}}\varepsilon \sim Exp(\lambda)$, the equation that identifies $f_{\theta|W}$ is

$$f_{C|WZ}(c; w, z) = \int_\Theta \lambda e^{-\lambda(c-\theta_0-\theta_1 z_1-\theta_2 g(z_2,w))} 1\left\{c \geq \theta_0 + \theta_1 z_1 + \theta_2 g(z_2, w)\right\} f_{\theta|W} d\theta,$$

and 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:

$$\begin{aligned}
(f_{\varepsilon|\theta WZ} \circ \varphi_i^{-1})(c, w, z, \theta) &= \lambda \exp\{-\lambda(c - \theta_0 - z_1'\theta_1 - \theta_2 g(z_2, w))\} \\
&= \lambda \exp\{-\lambda c\} \exp\{\lambda[1, z_1', g(z_2, w)]\theta\}.
\end{aligned}$$

These expressions satisfy the sufficient conditions of Lemma 1, with $h(\theta) = \lambda$, $m(\theta) = \theta = (\theta_0, \theta_1', \theta_2)'$ the identity function, $\tau(c, w, z) = \lambda(1, z_1', g(z_2, w))'$ and $k(c, w, z) = \exp\{-\lambda c\}$, implying that $f_{\theta|W}$ is point-identified.

The plan of the rest of the paper is as follows. In Section 3, we present the model and assumptions. Then, we analyse identification in Section 4. Section 5 presents our two estimators and Section 6 concludes with results from two Monte Carlo simulations. Proofs of the main results are in Appendix A while minor and technical results are in the Supplementary Appendix.

3 The general structural model

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 $C \in \mathbb{R}$, $W \in \mathcal{W} \subset \mathbb{R}^k$, $Z \in \mathcal{Z} \subset \mathbb{R}^l$, $\theta \in \Theta \subset \mathbb{R}^d$ 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, and use capital and lowercase Latin letters for observable random variables and their realizations, but lowercase Greek letters for unobservable random variables, as well as their realizations. For two random vectors A and B we write: $P_{A|B}$ for the conditional distribution of A given B and $f_{A|B}$ for the density function (*pdf*, hereafter) of $P_{A|B}$ with respect to Lebesgue measure. We use the convention that $f_{A|B}(a; b) = 0$ if a is not in the support of $P_{A|B=b}$. We denote by $\mathcal{C} \subset \mathbb{R}$ (resp. $\mathcal{Z} \times \mathcal{W}$) the support of the marginal distribution of C (resp. (Z, W)).

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, π_θ and π_{cz}, with support on Θ and $\mathcal{C} \times \mathcal{Z}$ respectively.[1] Define the space $L^2_{\pi_\theta}$ (resp. $L^2_{\pi_{cz}}$) of real-valued functions defined on Θ (resp. $\mathcal{C} \times \mathcal{Z}$) that are square integrable with respect to π_θ (resp. π_{cz}). That is,

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

We denote the scalar product by $\langle \cdot, \cdot \rangle$ and the induced norm by $\| \cdot \|$ in both spaces without distinction, e.g., $\forall h_1, h_2 \in L^2_{\pi_\theta}$, $\langle h_1, h_2 \rangle = \int h_1(\theta)h_2(\theta)\pi_\theta(\theta)d\theta$. The sets of conditional *pdf*s relevant for our analysis are defined as follows

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

[1] The weighting functions should be chosen to ensure that the operator T defined below is compact and bounded as discussed after Proposition 1 and to reflect the researcher's statistical loss function as discussed after equation (5.2) in Section 5.

and analogously for $\mathcal{F}_{C|WZ\theta}$. The next assumption specifies the structural data generating process.

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

$$\Psi(C, W, Z, \theta, \varepsilon) = 0 \quad a.s. \tag{3.1}$$

*where Ψ is a **known** Borel measurable real-valued function.*[2] *We assume that (3.1) has a unique global solution in terms of C:*

$$C = \varphi(W, Z, \theta, \varepsilon), \quad a.s. \tag{3.2}$$

where $\varphi : \mathbb{R}^{k+l+d+1} \to \mathbb{R}$ is a Borel-measurable function. In addition, we assume (3.2) has a finite number s of solutions in terms of ε almost surely.[3]

This assumption describes how our structural model links observables (C, W, Z) to unobservables (θ, ε). We distinguish between three different observables. C is the dependent variable, while W and Z denote variables that cause C. W is allowed to be correlated with θ while Z is assumed to be conditionally independent of θ given W. As was discussed in section 2, this distinction is motivated by applications in which some important explanatory variables $W$are endogenous. The distinction between the unobservable variables θ and ε is made to separate objects of interest from an error term ε. Consequently, the distribution of θ is allowed to be completely nonparametric while the distribution of ε is flexible, but parametric.

Our approach does not require that the function φ be available in closed-form nor that it be globally monotone in ε. All that is required is that we can solve equation (3.1) numerically, and that the function φ be piecewise monotonic. Hence, its inverse can be written as a finite collection of one-to-one functions each defined for a subset of the domain of c. More precisely, for some set A, let $Im\left(A|\, w, z, \theta\right)$ be the image of A through φ conditional on (w, z, θ). We can then define a finite partition of $\mathbb{R}$, $(\mathcal{E}_1, ..., \mathcal{E}_s)$, such that $\varphi_i^{-1}(w, z, \theta, \cdot) : Im\left(\mathcal{E}_i|\, w, z, \theta\right) \to \mathcal{E}_i$ is one-to-one for each $i \in \{1, ..., s\}$. The elements of the partition and the inverse can be computed since they are implicitly defined by (3.1). In the following we denote, $\forall i \in \{1, \ldots, s\}$, by $\varphi_i^{-1}(w, z, \theta, \cdot)$ the function $\varphi^{-1}(w, z, \theta, \cdot)$ with domain $\varphi(w, z, \theta, \mathcal{E}_i)$ and image $\mathcal{E}_i$.

Allowing for this general form of the structural model is an important weakening of assumptions, as closed form expressions are frequently not available and monotonicity conditions are difficult to justify. In our Euler equation example, the consumption function is only implicitly defined and there is little reason to believe that there is a monotonic relationship between unobserved states and levels of consumption.

The only other assumption on Ψ is differentiability. Let $\partial_c\Psi(c, w, z, \theta, \varepsilon)$ and $\partial_\varepsilon\Psi(c, w, z, \theta, \varepsilon)$ denote the partial derivatives of Ψ with respect to C and ε.

[2] In the assumption, we state that Ψ is known. In fact, if Ψ is estimated in a first-stage, it affects neither our procedure nor the rate of convergence as long as the first-stage estimator converges faster than our estimator described below. In this case, the asymptotic normality result that we provide below still holds under further assumptions similar to assumption 6 in **?**.

[3] For simplicity, we assume that ε is a scalar. The analysis can be extended to the multivariate case without great difficulty.

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 $\partial_\varepsilon \Psi(c, w, z, \theta, \varepsilon) \neq 0$ for every $(c, w, z, \theta, \varepsilon)$ in the support of $(C, W, Z, \theta, \varepsilon)$ except, possibly, on a set of $(c, w, z, \theta, \varepsilon)$ values whose Lebesgue measure is 0.*

Finally, the next assumptions characterize the joint conditional distribution of (ε, Z, θ) given W.

Assumption 3. *The conditional probability distribution $P_{\varepsilon|WZ\theta}$ on $\mathcal{B}_\varepsilon$ given (W, Z, θ) admits a pdf $f_{\varepsilon|WZ\theta}$ with respect to the Lebesgue measure. This pdf $f_{\varepsilon|WZ\theta}$ is known up to a finite-dimensional parameter θ_ε, a subvector of the vector θ. Moreover, $f_{\varepsilon|WZ\theta}$ is strictly positive and bounded away from infinity a.s. on the support of $P_{\varepsilon|WZ\theta}$.*

Assumption 4. *The conditional probability distribution $P_{Z\theta|W}$ on $\mathcal{B}_Z \otimes \mathcal{B}_\Theta$ given W admits a pdf $f_{Z\theta|W}$ with respect to the Lebesgue measure. The pdf $f_{\theta|W}$ is strictly positive and bounded away from infinity a.s. on its support.*

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

Assumption 3 allows the conditional distribution of ε to depend on all variables in the model. Unlike in deconvolution, we can allow for ε and θ to be dependent. By allowing $f_{\varepsilon|WZ\theta}$ to be known up to a finite dimensional random parameter[4], we allow for cases where not everything is known about $f_{\varepsilon|WZ\theta}$. In theory, the specification for $f_{\varepsilon|WZ\theta}$ can be very close to a nonparametric specification. As in all semi-parametric models there is a trade-off between flexibility and feasibility. Adding flexibility in this fashion will generally increase the dimension of the estimation problem, reduce the convergence rate of the estimator and may even lead to a failure of point identification if there is not enough independent variation in the data.

Note that Z and θ are continuous random vectors while W may be discrete. If some elements of Z are discrete, then the analysis is unchanged as long as the *pdf* of Z is replaced with the probability mass function and integrals with respect to Z are replaced by sums. Note however that discrete Z are likely to have little identifying power. If some elements of θ are discrete and random with known support, then the analysis also is unchanged. In this case, all of the statements with respect to $f_{\theta|W}$ have their finite dimensional counterparts. If some elements of θ are deterministic (or equivalently discrete random variables with unknown, finite support), then the analysis is slightly different. We discuss this case in Section 5.3 and explain how to estimate the model when some components of θ are deterministic.

Finally, Assumption 5 is the key independence condition that is often required for point identification of the *pdf* $f_{\theta|W}$. Strictly speaking Z is not required for point identification. It is possible to specify a model in which $f_{\theta|W}$ is point identified solely by nonlinearities in $f_{C|W\theta}$. When θ is a scalar, such a specification may be reasonable. However, especially when θ is a large dimensional vector, it is easy to specify models in which $f_{\theta|W}$ is not point identified without exogenous variation in Z. The linear random coefficients model in Section 2 is a leading case. We now provide a second example that illustrates these points as well as how our model can be applied in a richer setting.

[4] This finite dimensional parameter may be either included in the vector θ and treated as a random parameter, or estimated as a fixed parameter.

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

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

where C_t^* is consumption, $\eta_t \sim N(0, \sigma_\eta^2)$, and $\mathbb{E}_t$ is the time t conditional expectation operator. We assume that σ_η^2 has been estimated in advance. 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 $Z_t = Y_t - Y_{t-1}$. Under the assumptions stated, the consumption function (with measurement error) takes the form

$$C_t = \phi_{1t} W_t^1 + \phi_{2t}(W_t^2 + Z_t) + m_t(\gamma, \beta) + \varepsilon_t \tag{3.3}$$

with

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

The vector $\phi_t = (\phi_{1t}, \phi_{2t}, \phi_{3t}, \phi_{4t}, \phi_{5t})$ consists of parameters that depend only on R, t and σ_η^2. The vector $\theta = (\gamma, \beta)$ is assumed to be a time-invariant random coefficient vector, heterogeneously distributed in the population. We assume that the income process $(Y_t)_{t=1,..,T} \perp \theta$ 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, it is correlated with W_t. However, by assumption, $Z_t = Y_t - Y_{t-1}$ is independent of θ. Consequently, with the choice $W_t = (A_t, Y_{t-1})$ and $Z_t = Y_t - Y_{t-1}$, we obtain $\theta \perp Z_t | W_t$.

4 Identification of the distribution of parameters

In the following, we use $(f_{\varepsilon|WZ\theta} \circ \varphi_i^{-1})(c, w, z, \theta)$ to denote $f_{\varepsilon|WZ\theta}(\varphi_i^{-1}(w, z, \theta, c); w, z, \theta)$, we suppress the dependence of the operator T on W (where T is defined below), and, for a subset $\mathcal{A} \subset L^2_{\pi_\theta}$, use the notation $T|_{\mathcal{A}}$ to denote the operator T restricted to $\mathcal{A}$, that is, $T|_{\mathcal{A}} : \mathcal{A} \to L^2_{\pi_{cz}}$. We also use $\mathcal{R}(T)$ to denote the range of the operator T.

[5] In the CARA example, the consumption function is available in closed form. **?** develop an application in which the consumption function must be computed numerically.

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

$$f_{C|WZ} = Tf_{\theta|W} \quad a.s. \tag{4.1}$$

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

$$Th = \int\limits_{\Theta} \sum_{i=1}^{s} (f_{\varepsilon|WZ\theta} \circ \varphi_i^{-1})(c,w,z,\theta) \cdot \left| \frac{\partial_c \Psi(c,w,z,\theta,\varphi_i^{-1}(w,z,\theta,c))}{\partial_\varepsilon \Psi(c,w,z,\theta,\varphi_i^{-1}(w,z,\theta,c))} \right| 1_{\mathcal{C}_i}(c) h(\theta;w) d\theta, \tag{4.2}$$

$\mathcal{C}_i = \{c \in Im\left(\mathcal{E}_i \,|w,z,\theta\right)\}$ *and* $Im\left(\mathcal{E}_i \,|w,z,\theta\right)$ *is the image of* $\mathcal{E}_i$ *through* φ *conditional on* (w,z,θ). *This implies that* $f_{\theta|W}$ *is a solution of*

$$f_{C|WZ} = Tf_{\theta|W} \quad \textit{subject to} \quad f_{\theta|W} \in \mathcal{F}_{\theta|W}, \quad a.s. \tag{4.3}$$

The operator T is a mixing operator and $f_{C|WZ}$ is an a.s. $f_{\theta|W}$-mixture of $f_{C|WZ\theta}$. Equation (4.2) provides an expression for the operator T that depends only on the elements of the structural model $\left(\Psi, \varphi, f_{\varepsilon|WZ\theta}\right)$. These elements are known by assumption and can be directly computed.

Equation (4.3) is the basis for our estimation strategy. This equation characterizes $f_{\theta|W}$ as the solution of a *convexly-constrained Fredholm integral equation of the first kind.* Under Assumptions 1-5 the existence of at least one solution to (4.3) is guaranteed since $f_{C|WZ} \in \mathcal{R}(T|_{\mathcal{F}_{\theta|W}})$. We note that recovering $f_{\theta|W}$ from (4.3) is an ill-posed inverse problem.

The properties of the solution (or solutions) to (4.3) depend on the properties of T and its adjoint. The next proposition characterizes the adjoint operator of T.

Proposition 1 (Adjoint of T). *Let* $T : L^2_{\pi_\theta} \to L^2_{\pi_{cz}}$ *be the operator defined in (4.2). Assume that* T *is bounded. Then, the operator* T^* *defined as:* $\forall \psi \in L^2_{\pi_{cz}}$, $T^*\psi = \int_{\mathcal{C}} \int_{\mathcal{Z}} f_{C|WZ\theta}(c;w,z,\theta)\psi(c,z)\frac{\pi_{cz}(c,z)}{\pi_\theta(\theta)}dcdz$, *with*

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

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

Because the kernel of T is known, we can choose weight functions π_θ and π_{cz} so that T is bounded and compact with $\mathcal{R}(T) \subset L^2_{\pi_{cz}}$. Assumption 6 gives sufficient conditions for compactness and boundedness of T in terms of Ψ, $f_{\varepsilon|WZ\theta}$, π_{cz} and π_θ.

Assumption 6. *The function* $s^{\frac{1}{2}} f_{\varepsilon|WZ\theta} |\partial_c \Psi / \partial_\varepsilon \Psi|^{1/2}\Big|_{c=\varphi(w,z,\theta,\varepsilon)}$ *is a.s. square integrable in* (ε, Z, θ) *with respect to* $\frac{\pi_{cz}}{\pi_\theta}\Big|_{c=\varphi(w,z,\theta,\varepsilon)}$, *where* $s < \infty$ *is the number of piecewise monotonic components of the inverse of* φ *as defined in Assumption 1.*

In the Supplementary Appendix, we show that this assumption ensures that T is compact and bounded (see Proposition **??**). This proposition is not necessary to define our estimator nor to derive its asymptotic properties. However, when it is true, one of our proposed estimators can be written simply in terms of the singular value decomposition of T. In practice, compactness can be checked by scrutinizing Assumption 6. To do this, simply compute the integral of the square of $s^{\frac{1}{2}} f_{\varepsilon|WZ\theta}|\partial_c \Psi / \partial_\varepsilon \Psi|^{1/2}\Big|_{c=\varphi(w,z,\theta,\varepsilon)}$ with respect to the weight functions. Under Assumptions 1 - 6, T^*T is characterized by a countable number of eigenvalues which accumulate only at zero and admits the following *singular value decomposition* (SVD):

$$T\varphi_j = \lambda_j \psi_j, \qquad T^*\psi_j = \lambda_j \varphi_j, \quad j \in \mathbb{N} \tag{4.4}$$

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 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^* in $L^2_{\pi_\theta}$. When $\mathcal{N}(T)$ is not a singleton, where $\mathcal{N}(\cdot)$ denotes the null space of an operator, we can complete this orthonormal system in order to form an orthonormal basis (o.n.b.) of $L^2_{\pi_\theta}$ denoted by $\{\{\varphi_j\}_{j\in\mathbb{N}}, \{\tilde{\varphi}_l\}_{l\in J_0}\}$ where $J_0 \subset \mathbb{N}$ and $\tilde{\varphi}_l$ are such that $\mathcal{N}(T) = span\{\tilde{\varphi}_l\}_{l\in J_0}$. In words, the null space is spanned by the elements in J_0.

In the following, we use the SVD to characterize the set of possible solutions of (4.3), i.e. the identified set. To gain intuition, consider the analogous case in which the support of θ was discrete so that θ took on only k distinct values. Suppose the support were known. In that case, if we evaluate T at a finite number of points (C, Z, W), the operator T is a finite dimensional matrix. The probability mass function of θ conditional on W is identified if the researcher can identify, for each value of W, k distinct values (C, Z) such that T is invertible. If the matrix is not invertible (because some of its eigenvalues equal zero), then the identified set can be computed using the singular value decomposition of T. In the limit, as k grows to infinity, the discrete case approaches the continuous case. Assumptions 4 and 6 are required to ensure that this discrete intuition holds true in the continuous case.

The *pdf* $f_{\theta|W} \in \mathcal{F}_{\theta|W}$ will be called *identified* (with respect to the class $\mathcal{F}_{\theta|W}$) if

$$T|_{\mathcal{F}_{\theta|W}}(f_{\theta|W}) = T|_{\mathcal{F}_{\theta|W}}(\tilde{f}_{\theta|W}) \quad \Rightarrow \quad f_{\theta|W} = \tilde{f}_{\theta|W}, \quad a.s. \tag{4.5}$$

for all $f_{\theta|W}, \tilde{f}_{\theta|W} \in \mathcal{F}_{\theta|W}$. In words, $f_{\theta|W}$ is point identified if the operator $T|_{\mathcal{F}_{\theta|W}}$ is injective. The injectivity of $T|_{\mathcal{F}_{\theta|W}}$ depends on the injectivity of T but it is not equivalent. If T is injective, that is, $\mathcal{N}(T) = \{0\}$, then $T|_{\mathcal{F}_{\theta|W}}$ is injective as well. However, when T is non-injective the restricted operator $T|_{\mathcal{F}_{\theta|W}}$ may be injective. This is possible when the domain of $T|_{\mathcal{F}_{\theta|W}}$ is sufficiently restricted.

The following proposition characterizes the set of possible solutions of (4.3). We denote by $f^\dagger_{\theta|W}$ the minimum-norm solution of the *unconstrained* linear inverse problem $f_{C|WZ} = Tf_{\theta|W}$, that is, $f^\dagger_{\theta|W} = \arg\min\{\|h\|; h \in L^2_{\pi_\theta} \text{ and } f_{C|WZ} = Th\}$.

Proposition 2. *Under Assumptions 1-5, the set of all the solutions of (4.3) is:*

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

If in addition, Assumption 6 holds, then T is compact and there exist $\zeta_l \in \mathbb{R}$ for $l \in J_0 \subset \mathbb{N}$ such that

$$\Lambda = \left\{ h(\theta; w) = f^{\dagger}_{\theta|W}(\theta; w) + \sum_{\{l \in J_0\}} \zeta_l \tilde{\varphi}_l(\theta; w); \ \sum_{\{l \in J_0\}} \zeta_l^2 < \infty \ \text{and} \ \sup_{\theta \in \Theta} h^-(\theta; w) = 0 \ a.s. \right\}.$$

where $h^-(\theta; w) := -\min(h(\theta; w), 0)$ denotes the negative part of h and span$\{\tilde{\varphi}_l\}_{l \in J_0} = \mathcal{N}(T)$.

The second part of this proposition characterizes the set Λ in terms of the SVD of T which is known and the density $f_{C|WZ}$ which can be easily estimated. When the null space of $T|_{\mathcal{F}_{\theta|W}}$ is a singleton, Λ is a singleton as well and the model is point-identified. 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 (4.3);

(ii) the operator T is not injective, i.e. $\mathcal{N}(T) \neq \{0\}$, but $T|_{\mathcal{F}_{\theta|W}}$ is injective, i.e. (4.5) 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, a.s. In this case we can also have $\Lambda = f^{\dagger}_{\theta|W}$ if $f^{\dagger}_{\theta|W}$ is a probability density function.

In our context, injectivity of T is determined by the structural economic model and depends on how C, Z and θ interact. When $T|_{\mathcal{F}_{\theta|W}}$ is not injective, computation of Λ requires computation of the complete singular value expansion of the kernel of the operator T. In theory, because T is known and is not estimated, a researcher can compute the SVD of T, calculate the elements $\{\tilde{\varphi}_j\}_{j \in J_0}$ by a simple procedure of basis completion, like the Gram-Schmidt orthonormalization, and then characterize the null space of the operator, see **?**. In practice, a researcher must truncate the expansion at some point and impose that all singular values not computed equal zero. The error of this approximation can be bounded using methods in **?**.

It is well known that shape restrictions may provide identifying power. For example, see **?** or **?**. Nonetheless, the econometric literature on inverse problems for the most part has not exploited the fact that point identification can be obtained even without injectivity of T because $T|_{\mathcal{F}_{\theta|W}}$ may be injective[6]. Restricting the domain of interest or imposing shape potentially has identifying power. We discuss this formally in the next section where we provide a necessary and sufficient condition for point identification that we call $\mathcal{T}$-completeness. This condition is weaker than the conditions of completeness or bounded completeness that have been used in the previous econometric literature on inverse problems.

[6] An exception is **?**.

4.1 Identification and completeness

Define $\mathcal{F}_{\theta|CWZ} := \{f \mid f \text{ is a conditional } pdf \text{ on } (\mathbb{R}^d, \mathcal{B}_\Theta) \text{ given } (C, W, Z)\}$ as the set of *pdf* of θ conditional on (c, w, z). Provided that $f_{C|WZ}$ and $f_{\theta|W}$ are bounded away from zero and infinity, injectivity of the operator T is equivalent to the requirement that $\mathcal{F}_{\theta|CWZ}$ is $L^2_{\pi_\theta}$-complete (or bounded-complete) as noted in **?**, **?**, **?**, **?** and **?** in different setups.

However, in our framework, neither $L^2_{\pi_\theta}$-completeness nor bounded completeness are equivalent to identification of $f_{\theta|W}$. In fact, because the solutions of (4.3) are constrained to be *pdf*'s, then identification of $f_{\theta|W}$ is equivalent to completeness of $\mathcal{F}_{\theta|CWZ}$ with respect to a class of functions smaller than $L^2_{\pi_\theta}$ and the class of bounded functions. This class, that we denote by $\mathcal{T}$, is the class of functions that equal the difference between two densities scaled by the true density of θ. That is, $\mathcal{T} = \left\{h \in L^2_{\pi_\theta} : h = \frac{f_1 - f_2}{f_{\theta|W}} \quad \text{for some} f_1, f_2 \in \mathcal{F}_{\theta|W}\right\} \subset L^2_{\pi_\theta}$ where $f_{\theta|W}$ is the true conditional *pdf* of θ given W. This is summarized in the following proposition.

Proposition 3 ($\mathcal{T}$-completeness)**.** *Under the assumptions of Theorem 1, (4.5) holds if and only if $\mathcal{F}_{\theta|CWZ}$ is complete with respect to $\mathcal{T} \subset L^2_{\pi_\theta}$.*

Since the set $\mathcal{T}$ is strictly smaller than $L^2_{\pi_\theta}$, identification can be achieved even when $L^2_{\pi_\theta}$-completeness fails. **?** provide more background on *completeness* of a probability distribution with respect to a general family of functions $\mathcal{T}$.

It is well-known that, if the elements of $\mathcal{F}_{\theta|CWZ}$ belong to the exponential family, then $\mathcal{F}_{\theta|CWZ}$ is $L^2_{\pi_\theta}$-complete. However, since $\mathcal{T} \subset L^2_{\pi_\theta}$, the elements of $\mathcal{F}_{\theta|CWZ}$ do not need to be in the exponential family in order for our model to be point identified. In general, checking that Proposition 3 is satisfied is a computational issue that must be checked on a case by case basis. The next lemma, while stronger than required, provides a sufficient condition for identification that can be more easily checked in practice and that provides some intuition as to the type of mathematical structure that is required to provide identification.

Lemma 1. *Let $\dim(\theta)$ denote the dimension of θ and Assumptions 1-5 hold. Assume that $\forall i = 1, \ldots, s$, $(f_{\varepsilon|\theta WZ} \circ \varphi_i^{-1})(c, w, z, \theta)$ 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 invertible function whose image has dimension equal to $\dim(\theta)$. The functions τ_i and k_i are real-valued, do not depend on θ and k_i is a positive and bounded function. Further, the rank of $\mathbb{E}\left(\tau_i' \tau_i\right)$ is equal to $\dim(\theta)$ and the vector τ_i varies over the entire real line. Then, $f_{\theta|W}$ is identified with respect to the class $\mathcal{F}_{\theta|W}$.

The conditions of this lemma are satisfied in the linear random coefficient model outlined in the roadmap section as long as $g(Z_2, W)$ has support on the entire real line. Moreover, remark that if Θ is bounded, then a more limited variation of τ_i is sufficient to get the result of the lemma. The conditions of the lemma are also satisfied in the classical examples of the additively-closed and the location-scale one-parameter family of distributions. We detail these classes in the Supplementary Appendix. In contrast, the conditions are not satisfied in Example 1.

Example 1 (Continued)**.** *Suppressing the time subscript,* (3.3) *can be written as*

$$\varphi(W,Z,\theta,\varepsilon)=\phi_1 W_1+\phi_2\left(W_2+Z\right)+m(\gamma,\beta)+\varepsilon.$$

This implies that the density of measured consumption is

$$f_{C|WZ}\left(c;w,z\right)=\int_{\Theta}\frac{\exp\left(-\frac{1}{2}\left(\frac{c-\phi_1 w_1-\phi_2(w_2+z)-m(\gamma,\beta)}{\sigma_\varepsilon}\right)^2\right)}{\sqrt{2\pi\sigma_\varepsilon^2}}f_{\gamma\beta|W}\left(\gamma,\beta;w\right)d\gamma d\beta. \tag{4.6}$$

Define $\delta=m\left(\gamma,\beta\right)$. 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

$$\begin{aligned}f_{C|WZ}\left(c;w,z\right) &= \int_D\int_\Gamma\frac{\exp\left(-\frac{1}{2}\left(\frac{c-\phi_1 w_1-\phi_2(w_2+z)-\delta}{\sigma_\varepsilon}\right)^2\right)}{\sqrt{2\pi\sigma_\varepsilon^2}}\widetilde{f}_{\gamma\delta|W}\left(\gamma,\delta;w\right)d\gamma d\delta \qquad (4.7)\\ &= \int_D\frac{\exp\left(-\frac{1}{2}\left(\frac{c-\phi_1 w_1-\phi_2(w_2+z)-\delta}{\sigma_\varepsilon}\right)^2\right)}{\sqrt{2\pi\sigma_\varepsilon^2}}\widetilde{f}_{\delta|W}\left(\delta;w\right)\left(\int_\Gamma\widetilde{f}_{\gamma|W\delta}\left(\gamma,\delta;w\right)d\gamma\right)d\delta\\ &= \int_D\frac{\exp\left(-\frac{1}{2}\left(\frac{c-\phi_1 w_1-\phi_2(w_1+z)-\delta}{\sigma_\varepsilon}\right)^2\right)}{\sqrt{2\pi\sigma_\varepsilon^2}}\widetilde{f}_{\delta|W}\left(\delta;w\right)d\delta\end{aligned}$$

where $\widetilde{f}_{\gamma\delta|W}\left(\gamma,\delta;w\right)=f_{\gamma\beta|W}\left(\gamma,m^{-1}\left(\gamma,\delta\right)\right)\left|\frac{\partial m^{-1}(\gamma,\delta)}{\partial\delta}\right|$. The joint density of (γ,δ) is not point-identified because any proper conditional density $\widetilde{f}_{\gamma|W\delta}\left(\gamma;\delta,w\right)$ is consistent with the data. In fact, the conditions of Lemma 1 are not satisfied. The marginal density $\widetilde{f}_{\delta|W}\left(\delta;w\right)$ is point-identified. The identified set Λ is the set containing all elements of the form

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

for some conditional density $\widetilde{f}_{\gamma|W\delta}$. In Section 6, we show in simulations that despite this failure of point identification of $f_{\gamma\beta|W}$, the model has identifying power because our estimate of $\widetilde{f}_{\delta|W}\left(\delta;w\right)$ places meaningful bounds on the identified set. For example, Figures 6-7 show that the probability that γ is between 2 and 2.5 and β is between 0.95 and 0.96 is identified. Joint densities of (β,γ) that are inconsistent with the estimated probability of this event are ruled out.

5 Estimation

Our estimation strategy is based on equation (4.3). While the solution of (4.3) need not be unique, there is a unique solution of minimal norm which we denote by $f^{\dagger c}_{\theta|W}$. This solution takes the form

$$f^{\dagger c}_{\theta|W} = T^{\dagger}_{\mathcal{F}_{\theta|W}} f_{C|WZ} \tag{5.1}$$

where $T^{\dagger}_{\mathcal{F}_{\theta|W}}$ denotes the constrained generalized inverse of the restricted operator $T|_{\mathcal{F}_{\theta|W}}$ (see e.g., ?, Definition 2.1). The definition of $f^{\dagger c}_{\theta|W}$ differs from the definition of $f^{\dagger}_{\theta|W}$ since the latter is not constrained to belong to $\mathcal{F}_{\theta|W}$. However, in some cases (for instance in the point identified case): $f^{\dagger c}_{\theta|W} = f^{\dagger}_{\theta|W}$. It is important to note that the operator $T^{\dagger}_{\mathcal{F}_{\theta|W}}$ is nonlinear and noncontinuous since, in general, $\mathcal{R}(T|_{\mathcal{F}_{\theta|W}})$ is non closed. As a result, the inverse problem of recovering $f_{\theta|W}$ from (4.3) suffers from ill-posedness. This means that the naive estimator obtained by replacing $f_{C|WZ}$ with a consistent estimator in (5.1) would be inconsistent and a regularization procedure must be used.[7]

To implement our estimation procedure we assume that a nonparametric consistent estimator of $f_{C|WZ}$ is available.

Assumption 7. *Let (c_i, w_i, z_i), $i = 1, \dots, n$ 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 $\hat{f}_{C|WZ} \in L^2_{\pi_{cz}}$ a.s. and $\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2 \to 0$ as $n \uparrow \infty$.*

Once an estimator $\hat{f}_{C|WZ}$ has been computed we use a constrained Tikhonov-type estimator for $f_{\theta|W}$. This is the infinite dimensional counterpart of Ridge regression. The constrained Tikhonov-type estimator is defined as the minimizer, with respect to h, of

$$||Th - \hat{f}_{C|WZ}||^2 + \alpha||h||^2, \qquad h \in \mathcal{F}_{\theta|W}, \tag{5.2}$$

where the *regularization* parameter $\alpha > 0$ decreases to 0 at a suitable rate.

We develop the classical case where the penalty term $||h||^2$ is simply based on the $L^2_{\pi_\theta}$ norm. This penalty has the benefit of being easy to compute and well understood in the literature. From an economic point of view, since the minimum norm element is closest to the origin, heuristically, it may have the smallest impact on counterfactual predictions and lead to the smallest variation in counterfactual predictions across a wide. Alternatively, if a researcher has a prior belief on $f_{\theta|W}$ based on previous research, then the penalty can be replaced by $\|h - f^o_{\theta|W}\|^2$ or by the entropy $\int \log(h/f^o_{\theta|W})h$ where $f^o_{\theta|W}$ is the researcher's prior belief about the density.

Since the norm in (5.2) depends on π_θ and π_{cz}, choice of the weighting functions can be important. As noted after Proposition 1, the weights should be chosen so that the operator T is compact. In

[7] An alternative estimator could be based on seminonparametric sieve maximum penalised likelihood estimation of equation (4.1). Three main advantages of the Tikhonov-type estimator are that it is computationally simple, it is guaranteed to converge, and the eigenvalues and eigenfunctions are computed as part of the estimation procedure. In contrast, sieve penalised MLE may lack these features. One advantage of the sieve MLE approach is that it is relatively straightforward to impose the constraints of the model when estimating the density of C. When the model is correctly specified, this may result in efficiency gains.

addition, the weights π_{cz} and π_θ should be chosen to reflect the researcher's loss function. For example, a researcher may choose to place greater weight on some values of C or Z than others to reflect greater economic importance. Or, they may place greater weights on some values of θ to reflect prior beliefs about the distribution of θ. In our simulations in Section 6, we use constant weights that weight all values equally.

We propose two methods to compute the minimizer of (5.2). The first method is a two-step procedure that first computes the unconstrained Tikhonov regularized estimator and then projects it onto the closed and convex set $\mathcal{F}_{\theta|W}$. The second method uses numerical methods to directly solve the constrained minimization problem in (5.2).

The main advantage of the first estimator is that it is simple. The first step has a closed-form and the second step consists of a simple iterative procedure. As a result, in many cases it will be very fast to compute. On the other hand, the two-step estimator is only consistent if $f^{\dagger c}_{\theta|W} = f^{\dagger}_{\theta|W}$. The second estimator we propose overcomes this problem. It does not have a closed form but works regardless of whether $f^{\dagger c}_{\theta|W} = f^{\dagger}_{\theta|W}$ holds or not.

When point-identification fails, our second estimator converges to the minimum norm element in the identified set Λ. This particular element of the identified set is easy to compute. Once it is computed, it can be used to estimate the set Λ using the formula detailed in Proposition 2. The procedure is straightforward. Estimate $f^{\dagger}_{\theta|W}$, compute the eigenfunctions of T, and then construct the identified set as described in Proposition 2.

The first step of our two-step estimator has been used in nonparametric instrumental variable regression estimation and deconvolution problems for instance by **?** and **?**. In our mixture model setting the expression for our estimator is somewhat different from the one in **?** [8]. We provide asymptotic properties of the two-step estimator and extend previous results by considering the important case where the problem is severely ill-posed and the pdf $f_{\theta|W}$ is not analytic. Therefore, the rates given in Corollary 1 below, and the asymptotic normality results are new and not provided in the previous literature. These rates are given for the case where $\hat{f}_{C|WZ}$ is obtained by using kernel smoothing.

Another contribution of this section is to provide the rate for the constrained estimator and discuss how the regularity condition on $f_{\theta|W}$ has to be modified in order to obtain the rate in this case. To the best of our knowledge, these results are available only for deterministic inverse problems and not for stochastic inverse problems which are relevant in econometrics.

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

The two-step estimator is computed as follows.

[8] In our case, the generalized Fourier coefficient $\langle f_{C|WZ}, \psi_j \rangle$, cannot be simplified as in **?**. Therefore, $f_{C|WZ}$ must be estimated nonparametrically and plugged-in. This allows us to obtain a rate of convergence which is in general faster.

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

$$\min_{h \in L^2_{\pi_\theta}} \left\{ ||Th - \hat{f}_{C|WZ}||^2 + \alpha ||h||^2 \right\}. \tag{5.3}$$

The solution is the classical Tikhonov regularized estimator:

$$\hat{f}^{\alpha}_{\theta|W}(\theta; w) = (\alpha I + T^* T)^{-1} T^* \hat{f}_{C|WZ} \tag{5.4}$$

where I denote the identity operator in $L^2_{\pi_\theta}$. When T is compact, expression (5.4) simplifies to $\hat{f}^{\alpha}_{\theta|W}(\theta; w) = \sum_{j=1}^{\infty} \lambda_j (\alpha + \lambda_j^2)^{-1} \langle \hat{f}_{C|WZ}, \psi_j \rangle \varphi_j(\theta; w)$ where $\{\lambda_j, \psi_j, \varphi_j\}_{j \in \mathbb{N}}$ denotes the SVD of T.

Second step. Compute the orthogonal projection, denoted by $\mathcal{P}_c \hat{f}^{\alpha}_{\theta|W}$, of $\hat{f}^{\alpha}_{\theta|W}$ onto the set $\mathcal{F}_{\theta|W}$:

$$\mathcal{P}_c \hat{f}^{\alpha}_{\theta|W} := \max \left\{ 0, \hat{f}^{\alpha}_{\theta|W} - \frac{c}{\pi_\theta} \right\} \tag{5.5}$$

where c is such that $\int_\Theta \mathcal{P}_c \hat{f}^{\alpha}_{\theta|W} d\theta = 1$.

We call $\mathcal{P}_c \hat{f}^{\alpha}_{\theta|W}$ the *indirect Tikhonov regularized estimator* of $f^{\dagger c}_{\theta|W}$. **?** shows that the projection can be computed simply using the following iterative algorithm.

$\mathcal{P}_c$**−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 $|C_{k+1} - 1| < \epsilon$, for $\epsilon > 0$.

While other projection methods exist, **?** shows that this algorithm converges pointwise and in norm to $\mathcal{P}_c \hat{f}^{\alpha}_{\theta|W}$ and that $\mathcal{P}_c \hat{f}^{\alpha}_{\theta|W}$ minimizes the weighted MISE $\mathbb{E} \| \cdot \|^2$.

5.1.1 Rates of convergence

The two-step estimator is consistent when $f^{\dagger c}_{\theta|W} = f^{\dagger}_{\theta|W}$, that is, when $f^{\dagger}_{\theta|W} \in \mathcal{F}_{\theta|W}$. This is possible for instance when T is injective, or when T is not injective but $T|_{\mathcal{F}_{\theta|W}}$ is and $f^{\dagger}_{\theta|W} \in \mathcal{F}_{\theta|W}$. Theorem 2 below provides the rate of the (weighted) Mean Integrated Square Error (MISE) associated with the two-step estimator $\mathcal{P}_c \hat{f}^{\alpha}_{\theta|W}$. The rate depends on the smoothness of $f^{\dagger c}_{\theta|W}$ and on the smoothness of T. The next assumption (which is a type of *source condition*[9]) quantifies the smoothness of $f^{\dagger c}_{\theta|W}$ relative to the smoothness of T. It is only required to derive the rate of convergence of the estimator.

[9] We refer to **?** for a discussion on different types of source conditions in inverse problems.

Assumption 8. *Let $\phi : [0, \infty) \to [0, \infty)$ be a continuous, strictly increasing function with $\phi(0) = 0$. Let $T : L^2_{\pi_\theta} \to L^2_{\pi_{cz}}$ be as defined in (4.2) and bounded. There exists a source $\nu \in L^2_{\pi_\theta}$ such that for some $0 < M < \infty$,*

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

When the operator T is finitely smooth (mildly ill-posed case) and $f^{\dagger c}_{\theta|W}$ belongs to a Sobolev class of functions, then an appropriate choice of ϕ is $\phi(t) = t^{\beta/2}$ for some $\beta > 0$. For example, in our intertemporal consumption model, this choice of ϕ is appropriate if $f_{\theta|W}$ is infinitely differentiable. In contrast, when T is infinitely smooth (severely ill-posed case) and $f^{\dagger c}_{\theta|W}$ is not analytic, then an appropriate choice of ϕ is $\phi(t) = (-\log(t))^{-\beta/2}$ for some $\beta > 0$. In this latter case, the rate of convergence is much slower.

The following theorem states the rate of convergence:

Theorem 2. *Let Assumptions 1-5 and 7-8 be satisfied, and $f^{\dagger c}_{\theta|W} = f^{\dagger}_{\theta|W} \in \mathcal{F}_{\theta|W}$. Assume that there exists a constant γ_ϕ such that*

$$\sup_{t \in \sigma(T^*T)} \left|\phi(t)\alpha(\alpha + t)^{-1}\right| \leq \gamma_\phi \phi(\alpha), \qquad \alpha \to 0 \tag{5.6}$$

where $\sigma(T^*T)$ denotes the spectrum of T^*T. Then, the weighted MISE associated with $\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W}$ is $\mathbb{E}||\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}(\phi^2(\alpha) + \alpha^{-1}\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2)$. If $\phi(t) = t^{\beta/2}$ with $\beta > 0$ then, (5.6) is satisfied for $\beta \leq 2$ and

$$\inf_{\alpha>0} \mathbb{E}||\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}\left([\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2]^{\frac{\beta\wedge 2}{\beta\wedge 2+1}}\right).$$

If $\phi(t) = (-\log(t))^{-\beta/2}$ with $\beta > 0$ and (5.6) is satisfied then

$$\inf_{\alpha>0} \mathbb{E}||\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}\left(\left[-\log\left(\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2\right)\right]^{-\beta}\right).$$

In the case $\phi(t) = (-\log(t))^{-\beta/2}$, it has been shown in **?** that (5.6) holds automatically for $0 < \beta \leq 2$. The rate given in the theorem is at most of order $[\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2]^{\frac{2}{3}}$. This rate is slower than the minimax rate for estimation of a density function because we use indirect observations of θ to estimate $f_{\theta|W}$. Let $\hat{f}^{\alpha(k)}_{\theta|W}$ be the two-step estimator obtained by using the $\mathcal{P}_c$-algorithm. It is possible to show that $\mathbb{E}||\hat{f}^{\alpha(k)}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 \leq \mathbb{E}||\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2$. Therefore, this theorem also provides the rate of convergence for the approximation of $\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W}$ obtained from the $\mathcal{P}_c$-algorithm.

The rate of Theorem 2 can be made explicit by replacing the rate of convergence for $\hat{f}_{C|WZ}$. We consider here the case where $\hat{f}_{C|WZ}$ is a kernel estimator, i.e.,

$$\hat{f}_{C|WZ}(c; w, z) = \frac{\frac{1}{nh_n^{1+k+l}} \sum_{i=1}^n K_h(c_i - c, c)K_h(w_i - w, w)K_h(z_i - z, z)}{\frac{1}{nh_d^{k+l}} \sum_{l=1}^n K_h(w_l - w, w)K_h(z_l - z, z)}, \tag{5.7}$$

where $K(\cdot,\cdot)$ is a generalized kernel function[10] of order $r = 2$, and we assume without loss of generality that $\mathcal{C} = [0,1]$, $\mathcal{W} = [0,1]^k$, $\mathcal{Z} = [0,1]^l$. By standard Taylor series arguments, as in **?**, it is easy to show that $\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2 = \mathcal{O}\left(\frac{1}{n\min\{h_n,h_d\}^{k+l+1}} + \max\{h_n^4, h_d^4\}\right)$, 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, we obtain for $\phi(t) = t^{\beta/2}$

$$\inf_{\alpha>0} \mathbb{E}||\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}\left(n^{-\frac{4(\beta\wedge 2)}{(k+l+1+4)(\beta\wedge 2+1)}}\right) \tag{5.8}$$

and for $\phi(t) = (-\log(t))^{-\beta/2}$, $\inf_{\alpha>0}\mathbb{E}||\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}((-\log(1/n))^{-\beta})$. 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. *Let Assumptions 1-5, 7-8 and* (5.6) *be satisfied, and* $f^{\dagger c}_{\theta|W} = f^{\dagger}_{\theta|W} \in \mathcal{F}_{\theta|W}$. *Then,* $\mathbb{E}||\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}(\phi^2(\alpha) + \alpha^{-2}\left(\max\{h_n^4, h_d^4\} + n^{-1}(\min\{h_n,h_d\})^{-k}\right))$. *Moreover, if* $h_n = h_d \asymp n^{-1/(4+k)}$ *and* $\phi(t) = t^{(\beta\wedge 2)/2}$ *we have* $\inf_{\alpha}\mathbb{E}||\mathcal{P}_c\hat{f}^{\alpha}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}\left(n^{-\frac{4(\beta\wedge 2)}{(4+k)((\beta\wedge 2)+2)}}\right)$.

The rate in Corollary 1 is faster than the rate in (5.8) 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 W. So, 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).

5.1.2 Asymptotic Normality.

We now study pointwise asymptotic normality of the two-step 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 (5.7). For that we introduce the following technical assumption which uses the SVD of T.

Assumption 9. *Let* $\mathbb{E}_{WZ}$ *denote the conditional expectation given* (W,Z) *and* $\hat{f}_{WZ}$ *denote the kernel estimator of the joint pdf of* (W,Z). *We assume that for every* $\theta\in\Theta$ *and* $w\in\mathcal{W}$: *(i)*

$$\mathbb{E}\left|\sum_{j=1}^{\infty}\frac{\lambda_j}{\alpha+\lambda_j^2}\left\langle (K_h(c_i-c,c) - \mathbb{E}_{WZ}(K_h(c_i-c,c)))\frac{K_h(z_i-z,z)K_h(w_i-w,w)}{h_n^{k+l+1}\hat{f}_{WZ}}, \psi_j\right\rangle \varphi_j(\theta;w)\right|^3 = \mathcal{O}\left(\alpha^{-3/2}h_n^{-2k}\right)$$

[10] We refer to **?**, **?** and references therein for an explicit definition of $K(\cdot,\cdot)$). By abuse of notation, we use the same second order kernel K for all the variables and the same bandwidth h_n (resp. h_d) for the different bandwidths, though they could in principle be distinct.

and (ii) there exists a constant $\kappa > 0$ *such that*

$$Var\left(\sum_{j=1}^{\infty}\frac{\lambda_j}{\alpha+\lambda_j^2}\left\langle (K_h(c_i-c,c)-\mathbb{E}_{WZ}(K_h(c_i-c,c)))\frac{K_h(z_i-z,z)K_h(w_i-w,w)}{h_n^{k+l+1}\hat{f}_{WZ}},\psi_j\right\rangle\varphi_j(\theta;w)\right) > \kappa\alpha^{-2}h_n^{-k}.$$

In the following lemma we use the notation '⇒' to denote pointwise convergence in distribution.

Lemma 2. *Let Assumptions 1-9 and* (5.6) *hold,* $f_{\theta|W}^{\dagger c} = f_{\theta|W}^{\dagger} \in \mathcal{F}_{\theta|W}$ *and* $\mathcal{P}_c\hat{f}_{\theta|W}^{\alpha}$ *be the two-step estimator computed by using* $\hat{f}_{C|WZ}(c;w,z)$ *defined in (5.7). Let* $\mathbb{E}_{WZ}$ *denote the conditional expectation given* (W,Z) *and* $\hat{f}_{WZ}$ *denote the kernel estimator of the joint pdf of* (W,Z)*. If* $n\alpha h_n^{k+4} \to 0$*,* $\alpha^3/(h_n^k n) \to 0$ *and if* $n\alpha^2 h_n^k \phi^2(\alpha) \to 0$*, then for every* $\theta \in \Theta$ *and* $w \in \mathcal{W}$*:*

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

where

$$V_c(\theta,w) \;=\; \frac{1}{n}Var\left(\mathcal{P}_c^{\dagger}\sum_{j=1}^{\infty}\frac{\lambda_j}{(\alpha+\lambda_j^2)}\left\langle (K_h(c_i-c,c)-\mathbb{E}_{WZ}(K_h(c_i-c,c)))\frac{K_{h,i}(z,w)}{h^{k+l+1}\hat{f}_{WZ}},\psi_j\right\rangle\varphi_j(\theta;w)\right),$$

$K_{h,i}(z,w) = K_h(z_i - z, z)K_h(w_i - w, w)$ *and* $\mathcal{P}_c^{\dagger}$ *denotes the projection on the tangent cone of* $\mathcal{F}_{\theta|W}$ *at* $f_{\theta|W}^{\dagger c}$ *defined as* $\overline{\{\lambda(f - f_{\theta|W}^{\dagger c});\ \lambda \geq 0,\ f \in \mathcal{F}_{\theta|W}\}}$.

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

5.2 Estimation of $f_{\theta|W}^{\dagger c}$: constrained Tikhonov regularization

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

$$\min_{h\in\mathcal{F}_{\theta|W}}\left\{||Th - \hat{f}_{C|WZ}||^2 + \alpha||h||^2\right\}. \tag{5.9}$$

The existence of a unique solution to problem (5.9) is proved in **?**. A closed-form solution of this problem does not exist and numerical methods must be used to compute a solution. We denote by $\check{f}_{\theta|W}^{\alpha,c}$ the estimator obtained by solving (5.9) and by $P_{\mathcal{F}_{\theta|W}}$ the orthogonal projector of $L^2_{\pi_\theta}$ onto $\mathcal{F}_{\theta|W}$. The next theorem states consistency of the estimator $\check{f}_{\theta|W}^{\alpha,c}$.

Theorem 3. *Let Assumptions 1-5 and 7 hold,* T *be a bounded operator from* $L^2_{\pi_\theta}$ *to* $L^2_{\pi_{cz}}$ *defined in (4.2) and* $f_{\theta|W}^{\dagger c} \in \mathcal{R}(P_{\mathcal{F}_{\theta|W}}T^*)$*. Then, if* $\alpha \to 0$ *and* $\alpha^{-1}\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2 \to 0$ *then:* $\mathbb{E}||\check{f}_{\theta|W}^{\alpha,c} - f_{\theta|W}^{\dagger c}||^2 \to 0$.

Under a smoothness assumption about $f_{\theta|W}^{\dagger c}$, it is possible to extend the result of Theorem 3 and recover the convergence rate for the constrained estimator. To derive this rate, a regularity

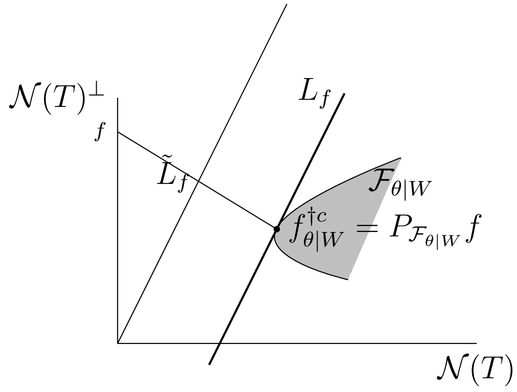


Figure 1: Representation of (part of) the set $\mathcal{F}_{\theta|W}$ (grey area), the supporting hyperplane $L_f := \{h \in L^2_{\pi_\theta}; \langle f^{\dagger c}_{\theta|W} - f, h - f^{\dagger c}_{\theta|W}\rangle = 0\}$ of $\mathcal{F}_{\theta|W}$ in $f^{\dagger c}_{\theta|W}$, the hyperplane $\tilde{L}_f$ and an element f of M.

condition for $f^{\dagger c}_{\theta|W}$ different from Assumption 8 is required. This regularity condition is stated in terms of a set of functions defined as follows.

Define: $M := \{\, f \in \mathcal{N}(T)^{\perp} \big|\, P_{\mathcal{F}_{\theta|W}} f = f^{\dagger c}_{\theta|W}\}$ and, $\forall f \in M$, $\tilde{L}_f := \{h \in L^2_{\pi_\theta}; \langle f^{\dagger c}_{\theta|W} - f, h\rangle = 0\}$. Define $\tilde{P}_f$ to be the orthogonal projector of $L^2_{\pi_\theta}$ onto $\tilde{L}_f$. Because we are in Hilbert spaces, $\tilde{P}_f$ is a linear operator. Finally, for $\beta > 0$ define $N_\beta := \{f \in M; \tilde{P}_f f^{\dagger c}_{\theta|W} \in \mathcal{R}\left((\tilde{P}_f T^* T \tilde{P}_f)^{\beta/2}\right)\}$. The regularity condition is stated in terms of the set N_β.

Theorem 4. *Let Assumptions 1-5 and 7 hold and let T be a bounded operator. Suppose $f^{\dagger c}_{\theta|W} \neq f^{\dagger}_{\theta|W}$ and $N_\beta \neq \varnothing$. Then,* $\mathbb{E}||\breve{f}^{\alpha,c}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}\left(\alpha^{\beta\wedge 2} + \alpha^{-1}\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2\right)$ *and if* $\alpha \asymp (\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2)^{\frac{1}{(\beta\wedge 2)+1}}$*:* $\mathbb{E}||\breve{f}^{\alpha,c}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}\left([\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2]^{\frac{\beta\wedge 2}{\beta\wedge 2+1}}\right)$.

In order to understand this result, consider Figure 1, adapted from **?**. From the figure, it is clear that N_β is the set of all functions $f \in \mathcal{N}(T)^{\perp}$ such that the orthogonal projection onto $\mathcal{F}_{\theta|W}$ equals $f^{\dagger c}_{\theta|W}$ and such that the orthogonal projection of $f^{\dagger c}_{\theta|W}$ onto the hyperplane $\tilde{L}_f$ is "smooth". The regularity condition on $f^{\dagger c}_{\theta|W}$ is imposed via the smoothness of its projection $\tilde{P}_f f^{\dagger c}_{\theta|W}$. Such smoothness is measured in terms of smoothness of the operator $(\tilde{P}_f T^* T \tilde{P}_f)^{\beta/2}$, which is a projection of an integral operator. The new regularity condition required for Theorem is that $N_\beta \neq \varnothing$. In words, there exists at least one $f \in M$ such that the projection of $f^{\dagger c}_{\theta|W}$ on the corresponding hyperplane $\tilde{L}_f$ has a degree of smoothness β.

5.3 Case with non-random parameters

Suppose that some components of θ are deterministic, that is, $\theta = (\theta_1', \theta_2')'$, where θ_1 is the vector of deterministic components of θ, assumed to belong to a compact subset $\Theta_1 \subset \mathbb{R}^{d_1}$, and θ_2 is the vector of random components of θ (with dimension d_2) distributed according to a probability distribution $P_{\theta_2|W}$ satisfying Assumptions 4 and 5. In this case, we can use either of the two estimation procedures we have proposed, after some minor modifications.

Here, we focus on the constrained Tikhonov regularization procedure.[11] The minimization problem (5.9) should be replaced by

$$\min_{\theta_1 \in \Theta_1,\, h \in \mathcal{F}_{\theta_2|W}} \left\{ ||T_{\theta_1} h - \hat{f}_{C|WZ}||^2 + \alpha ||h||^2 \right\} \tag{5.10}$$

where we write T_{θ_1} to make explicit the dependence of the operator on θ_1. The kernel of T_{θ_1} is (up to the factor $\frac{1}{\pi_\theta}$) equal to $f_{C|WZ\theta}(c; w, z, \theta_1, \theta_2)$ which has the same expression as in (4.2).

Let the parameter space be $\mathcal{G} = \Theta_1 \times \mathcal{F}_{\theta_2|W}$. Denote by $g = (\theta_1, h)$ a generic element of $\mathcal{G}$, with $\theta_1 \in \Theta_1$ and $h \in \mathcal{F}_{\theta_2|W}$. Let $g^0 = (\theta_1^0, f^0_{\theta_2|W})$ the true value of g. Moreover, denote $\widehat{Q}_n(\theta_1, h) = ||T_{\theta_1} h - \hat{f}_{C|WZ}||^2$ and $Q(\theta_1, h) = ||T_{\theta_1} h - f_{C|WZ}||^2$. The estimator computed by solving (5.10) will be denoted by $\hat{g} := (\hat{\theta}_1, \check{f}^{\alpha,c}_{\theta_2|W})$ and belongs to $\mathcal{G}$. Define $\|\hat{g} - g^0\| := \|\hat{\theta}_1 - \theta_1^0\|_E + \|\check{f}^{\alpha,c}_{\theta_2|W} - f^0_{\theta|W}\|$, where $\|\cdot\|_E$ denotes the Euclidean norm induced by the scalar product $\langle \cdot, \cdot \rangle_E$ in $\mathbb{R}^{d_1}$. Theorem 5 below states consistency of $\hat{g}$. We introduce the following assumption.

Assumption 9. *Let the following statements hold:*

1.

i. The subset $\Theta_1 \subset \mathbb{R}^{d_1}$ is compact.

ii. The family of functions $\{\widehat{Q}_n(\cdot, h) + \alpha\|h\|^2\}_{h \in \mathcal{F}_{\theta_2|W}}$ is equidifferentiable at every $\theta_1 \in \Theta_1$.

iii. Let $\widehat{Q}_{n,1}(\theta_1, h)$ denote the first derivative of $\widehat{Q}_n(\theta_1, h)$ with respect to θ_1 evaluated at θ_1. We assume that $\sup_{h \in \mathcal{F}_{\theta_2|W}} |\widehat{Q}_{n,1}(\theta_1, h)| < \infty$ for every $\theta_1 \in \Theta_1$.

iv. $Q(g^0) = 0$ and any $(\theta_1, h) \in \mathcal{G}$ that satisfies $Q(\theta_1, h) = 0$ also satisfies $\theta_1 = \theta_1^0$ and $h = f^0_{\theta_2|W}$ almost everywhere.

v. The function $f_{C|WZ\theta}(c; w, z, \theta_1, \theta_2)$ is continuous in θ_1.

vi. The criterion $\hat{Q}_n$ satisfies: $|\hat{Q}_n(g^0) - Q(g^0)| = O_p(\delta_n)$, where $\delta_n = o(1)$.

Theorem 5. *Let Assumptions 1-5, 7 and 9 hold. Then: (i) a solution to* (5.10) *exists and (ii), if $\delta_n = O(\alpha)$: $\|\hat{g} - g^0\| \to 0$ in probability.*

6 Monte Carlo simulation

6.1 Simulation 1: Linear endogenous random coefficient model

Consider model (2.1). Assume that $g(Z_2, W) = Z_2 W$. Then equation (2.1) becomes

$$C = \theta_1 Z_1 + \theta_2 Z_2 W + \varepsilon. \tag{6.1}$$

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 $\mu_\theta = \beta_0 + \beta_1 W$, $(\beta_0, \beta_1) = (1, 1)$ and Σ_θ equal to 0.1 times the identity matrix.

[11] The two-step procedure is described in Appendix **??**.

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 (5.4). Finally, we compute $\mathcal{P}_c\widehat{f^{\alpha}_{\theta|W}}$ as in (5.5) , at the 30th, 50th, and 70th percentiles of the distribution of W.

To facilitate accurate numerical integration, we first make a change of variable, mapping (C, Z_1, Z_2) into the region $[-1, 1]^3$. Specifically, we define $U_c = 2\Phi\left(\frac{C-\mu_c}{\sigma_c}\right) - 1$, $U_1 = 2\Phi\left(\frac{Z_1-\mu_{z_1}}{\sigma_{z_1}}\right) - 1, U_2 = 2\Phi\left(\frac{Z_2-\mu_{z_2}}{\sigma_{z_2}}\right) - 1$, where Φ is the standard normal CDF and (μ_c, σ_c), $(\mu_{z_i}, \sigma_{z_i})$, $i = 1, 2$ are the empirical mean and standard deviation of C, Z_1 and Z_2. Substituting these new variables into (6.1) , and solving for ε, the structural function $\varepsilon = \varphi^{-1}(W, Z, \theta, \varepsilon)$ can be written as

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

Next, let w_{30}, w_{50}, w_{70} denote the 30th, 50th and 70th percentile of W, resp.. 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

$$\min_{\{h\}}\left\{\int\left(\hat{f}_{U_c|WU_{z_1}U_{z_2}} - Th\right)^2 dcdz_1dz_2 + \alpha\int h(\theta)^2 d\theta\right\}. \tag{6.2}$$

The solution is given in equation (5.4) . We approximated the integral over $[-1, 1]^3$ with the tensor product of three unidimensional Gauss-Legendre quadrature rules with 20 quadrature nodes in each dimension, and analogously over Θ.

Figure 2 displays contour 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 panel of the 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 dimension with the true spread of the density of random coefficients.

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:

$$AMSE = E\left[\frac{1}{3}\sum_q\left(\int\left[\mathcal{P}_c\widehat{f^{\alpha}_{\theta|W}}(\theta; w_q) - f_{\theta|W}(\theta; w_q)\right]^2 f_{\theta|W}(\theta; w_q)\, d\theta\right)^{0.5}\right] = E[MSE] \tag{6.3}$$

$w_q \in \{w_{30}, w_{50}, w_{70}\}$ 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 3 shows the densities of the square root of the weighted MSE (WMSE) 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 replications 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 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 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 the 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, the ratio is almost constant, suggesting that the theoretical large sample differences may slightly overstate the small sample differences.

6.2 Simulation 2: Intertemporal consumption model

To analyse the CARA model, we simulated $n = 1000$ agents starting at age $t = 21$, working for 45 periods and then obtaining a terminal retirement utility. Income grows until retirement. In addition, in each period each agent faces a permanent i.i.d. income shock η_t 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, covering a range of values suggested in the literature. The joint distribution of (γ, β) is generated as follows. We define $x \sim N(\mu_x, I)$ with $\mu_x = (1, 0)'$ and generate $\gamma = 0.5 + 3.5\,\Phi(x_1)$, and $\beta = 0.7 + 0.299\,\Phi(x_2)$, where Φ is the standard normal CDF. 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.

While the data are simulated for 45 periods of life, we select one cross section at age 31 to be used for our estimator. We obtained similar results for other values of t. The dependence between δ and W (where δ is defined below) varies with t as does the distribution of the data. However, the quality of the estimation results does not.

In this CARA 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

$$\delta = 0.5\phi_5\gamma + \phi_3 \frac{\ln(R\beta)}{\gamma} \tag{6.4}$$

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

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

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

$$AMSE = E\left[\frac{1}{2}\sum_q \left(\int \left(\left[\mathcal{P}_c \widehat{f}^{\alpha}_{\delta|W}(\delta; w_q) - \widehat{f}^{Ker}_{\delta|W}(\delta; w_q)\right]^2\right)\left(\widehat{f}^{Ker}_{\delta|W}(\delta; w_q)\right) d\delta\right)^{0.5}\right]. \tag{6.5}$$

To compute the AMSE, we replace the expected values in (6.5) 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 4- 5 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 (6.4) . 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 (6.4) 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 6-7 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 6 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 convex regions. The convex region in Figure 6 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 convex 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.

A Proofs

A.1 Proof of Theorem 1

By Assumption 1, there exists a unique $c = \varphi(w, z, \theta, \varepsilon)$ that satisfies (3.1). Thus, using the transformation $\varphi(w, z, \theta, \cdot)$ mapping ε to c, 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 $\mathcal{E}_1, \dots, \mathcal{E}_s$ 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 $i = 1, \dots, s$, for given (w, z, θ) and $s \in \mathbb{N}_+$. Let $\varphi_i^{-1}(w, z, \theta, \cdot) : Im\left(\mathcal{E}_i \left| w, z, \theta\right.\right) \to \mathcal{E}_i$ be the corresponding inverse mapping for given (w, z, θ). Then,

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

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 (3.1) 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 2

The first characterization of Λ follows trivially from (4.3) since every element obtained as the sum of $f_{\theta|W}^{\dagger}$ and an element of $\mathcal{N}(T)$ is solution of the unconstrained inverse problem. Then, to obtain the set of solutions to the constrained problem, we only have to take the intersection with $\mathcal{F}_{\theta|W}$.

To obtain the second characterization of Λ remark that we can always write a generic element of $f_{\theta|W}^{\dagger} \oplus \mathcal{N}(T)$ in terms of the o.n.b. $\{\{\varphi_j\}_{j\in\mathbb{N}}, \{\tilde{\varphi}_l\}_{l\in J_0}\}$ which exists under Assumption 6. Then, we impose the following constraints to every $h \in \{f_{\theta|W}^{\dagger} \oplus \mathcal{N}(T)\}$: *(i)* $\int_{\Theta} h(\theta; w) d\theta = 1$, a.s., *(ii)* $\int_{\Theta} h^2(\theta; w) \pi_\theta(\theta) d\theta < \infty$, a.s. and *(iii)* $h(\theta; w) \geq 0$, a.s. The first constraint is automatically verified since for every (w, z, θ), $\int_{\mathcal{C}} f_{C|WZ\theta}(c; w, z, \theta) dc = 1$ and, by using Fubini's theorem: $\int_{\Theta} h(\theta; w) d\theta = \int_{\mathcal{C}} \int_{\Theta} f_{C|WZ\theta} d\theta h(\theta; w) dc = \int_{\mathcal{C}} f_{C|WZ} dc = 1$ (where we have used the fact that $f_{C|WZ\theta}$ integrates to 1 and $Th = f_{C|WZ}$). Constraint *(ii)* is equivalent to $||f^{\dagger}_{\theta|W}||^2 + \sum_{l\in J_0}\zeta_l^2 < \infty$ for some $\zeta_l \in \mathbb{R}$ and, by definition of $f^{\dagger}_{\theta|W}$, $||f^{\dagger}_{\theta|W}||^2 < \infty$. Finally, constraint *(iii)* is equivalent to require that the negative part of every function in Λ is equal to 0.

A.3 Proof of Proposition 3

Suppose that $\mathcal{F}_{\theta|CWZ}$ is $\mathcal{T}$-complete and that for $f^1_{\theta|W}, f^2_{\theta|W} \in \mathcal{F}_{\theta|W}$, $T|_{\mathcal{F}_{\theta|W}}(f^1_{\theta|W}) = T|_{\mathcal{F}_{\theta|W}}(f^2_{\theta|W})$ holds. By using the decomposition $f_{C|WZ\theta} = f_{\theta|CWZ}f_{C|WZ}/f_{\theta|W}$ this equality can be rewritten as

$$\begin{aligned} 0 &= T|_{\mathcal{F}_{\theta|W}}(f^1_{\theta|W}) - T|_{\mathcal{F}_{\theta|W}}(f^2_{\theta|W}) = \int_{\Theta} f_{C|WZ\theta}(c;w,z,\theta)\left[f^1_{\theta|W}(\theta;w) - f^2_{\theta|W}(\theta;w)\right]d\theta \\ &= \int_{\Theta} f_{\theta|CWZ}(\theta;c,w,z)\frac{f_{C|WZ}(c;w,z)}{f_{\theta|W}(\theta;w)}\left[f^1_{\theta|W}(\theta;w) - f^2_{\theta|W}(\theta;w)\right]d\theta \qquad \text{(A.4)} \end{aligned}$$

which is equivalent to

$$0 = \int_{\Theta} f_{\theta|CWZ}(\theta;c,w,z)\frac{1}{f_{\theta|W}(\theta;w)}\left[f^1_{\theta|W}(\theta;w) - f^2_{\theta|W}(\theta;w)\right]d\theta \qquad \text{(A.5)}$$

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

On the other hand, if (4.5) holds, then $0 = \int_{\Theta} f_{\theta|CWZ}(\theta;c,w,z)\frac{1}{f_{\theta|W}(\theta;w)}\left[f^1_{\theta|W}(\theta;w) - f^2_{\theta|W}(\theta;w)\right]d\theta$ implies that $\frac{1}{f_{\theta|W}(\theta;w)}\left[f^1_{\theta|W}(\theta;w) - f^2_{\theta|W}(\theta;w)\right] = 0$ because, by Assumptions 2, 3 and 4, $0 < f_{C|WZ} < \infty$ and $0 < f_{\theta|W} < \infty$. This concludes the proof.

A.4 Proof of Lemma 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 $T\phi(\theta;w) = 0$, a.s. for some function $\phi \in \mathfrak{D}$. Then, $\forall (c,z) \in \mathcal{C}\times\mathcal{Z}$

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

implies that $\forall (c,z) \in \mathcal{C}_i\times\mathcal{Z}$

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

Then, $\forall (c,z) \in \mathcal{C}_i\times\mathcal{Z}$ and $\forall i = 1,\dots,s$, we have:

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

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 $\nu_i(d\mu_i) = h_i\left(m_i^{-1}(\mu_i)\right) dm_i^{-1}(\mu_i)$. Thus, $\forall (c, z) \in \mathcal{C}_i \times \mathcal{Z}$ and $\forall i = 1, \ldots, s$,

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

a.s. 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 since $\tau_i(c, w, z)$ vary over $\mathbb{R}$. This implies that $\zeta_i(\mu_i; w, z, c) = 0$ and then $\phi_i(\theta; w) = 0$, a.s. since $\partial_c \varphi_i^{-1}(w, z, \theta, c)$ is bounded away from 0 and ∞ by Assumption 2, $\forall (c, w, \theta, z) \in \mathcal{C} \times \mathcal{W} \times \Theta \times \mathcal{Z}$.

A.5 Proof of Theorem 2

First, since $||\mathcal{P}_c|| \leq 1$ we have: $\mathbb{E}||\mathcal{P}_c \hat{f}^\alpha_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 = \mathbb{E}||\mathcal{P}_c(\hat{f}^\alpha_{\theta|W} - f^\dagger_{\theta|W})||^2 \leq ||\mathcal{P}_c||^2 \mathbb{E}||\hat{f}^\alpha_{\theta|W} - f^\dagger_{\theta|W}||^2 \leq \mathbb{E}||\hat{f}^\alpha_{\theta|W} - f^\dagger_{\theta|W}||^2$. Let $f^\alpha_{\theta|W} := (\alpha I + T^*T)^{-1} T^* f_{C|WZ}$, then

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

By the Halmos'spectral theorem (see, for instance, **?**) the operator T^*T admits a spectrum $\sigma(T^*T)$. Hence we can analyze the two terms $\mathcal{A}_1$ and $\mathcal{A}_2$ as follows. Term $\mathcal{A}_1$ is

$$\begin{aligned} \mathcal{A}_1 &= \mathbb{E}||(\alpha I + T^*T)^{-1} T^*(\hat{f}_{C|WZ} - f_{C|WZ})||^2 \leq ||(\alpha I + T^*T)^{-1} T^*||^2 \mathbb{E}||(\hat{f}_{C|WZ} - f_{C|WZ})||^2 \\ &\leq \sup_{t \in \sigma(T^*T)} |(\alpha + t)^{-1}\sqrt{t}|^2 \mathbb{E}||(\hat{f}_{C|WZ} - f_{C|WZ})||^2 = \mathcal{O}\left(\frac{1}{\alpha}\mathbb{E}||(\hat{f}_{C|WZ} - f_{C|WZ})||^2\right). \end{aligned} \tag{A.7}$$

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

$$\begin{aligned} \mathcal{A}_2 &= \mathbb{E}||(\alpha I + T^*T)^{-1} T^* f_{C|WZ} - f^\dagger_{\theta|W}||^2 = ||[I - (\alpha I + T^*T)^{-1} T^*T] f^\dagger_{\theta|W}||^2 \\ &= ||\alpha(\alpha I + T^*T)^{-1} f^\dagger_{\theta|W}||^2 = ||\alpha(\alpha I + T^*T)^{-1} \phi(T^*T)\nu||^2 \\ &\leq \sup_{t \in \sigma(T^*T)} |\phi(t)\alpha(\alpha + t)^{-1}|^2 \nu^2 = \mathcal{O}(\phi^2(\alpha)) \end{aligned} \tag{A.8}$$

where the last inequality follows from (5.6). This shows that $\mathbb{E}||\hat{f}^\alpha_{\theta|W} - f^\dagger_{\theta|W}||^2 = \mathcal{O}\left(\phi^2(\alpha) + \frac{1}{\alpha}\mathbb{E}||(\hat{f}_{C|WZ} - f_{C|WZ})||^2\right)$.

Next, consider the case $\phi(t) = t^{\beta/2}$. Then, $\sup_{t\in\sigma(T^*T)} |t^{\beta/2}\alpha(\alpha+t)^{-1}| = \frac{1}{2}\alpha^{\beta/2}$ if $\beta < 2$ and $\sup_{t\in\sigma(T^*T)} |\phi(t)\alpha(\alpha+t)^{-1}| = \alpha$ if $\beta = 2$. Hence, (5.6) is satisfied and we choose $\alpha \asymp (\mathbb{E}||(\hat{f}_{C|WZ} - f_{C|WZ})||^2)^{1/(\beta\wedge 2+1)}$ we get the result. Finally, consider the case $\phi(t) = (-\log(t))^{-\beta/2}$. If we choose $\alpha \asymp (\mathbb{E}||(\hat{f}_{C|WZ} - f_{C|WZ})||^2)^{\epsilon}$ for $0 < \epsilon < 1$ we get the final result of the theorem.

A.6 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 $\mathcal{A}_1 \leq ||(\alpha I + T^*T)^{-1}||^2\mathbb{E}||T^*(\hat{f}_{C|WZ} - f_{C|WZ})||^2$ so that $\mathcal{A}_1 = \mathcal{O}\left(\alpha^{-2}\mathbb{E}||T^*(\hat{f}_{C|WZ} - f_{C|WZ})||^2\right)$. We have to compute the rate of $\mathbb{E}||T^*(\hat{f}_{C|WZ} - f_{C|WZ})||^2$. Remark that $\mathbb{E}||T^*(\hat{f}_{C|WZ} - f_{C|WZ})||^2 = \int_\Theta \left(Var(T^*\hat{f}_{C|WZ}) + (\mathbb{E}(T^*\hat{f}_{C|WZ}) - T^*f_{C|WZ})^2\right)\pi_\theta(\theta)d\theta$. By using standard Taylor series approximations, it is easy to show (see Lemma **??** in the Supplementary Appendix) that the squared bias term is of order $\left(\mathbb{E}(T^*\hat{f}_{C|WZ} - T^*f_{C|WZ})\right)^2 = \mathcal{O}\left(\max\{h_n^4, h_d^4\}\right)$ and the variance term is $Var(T^*\hat{f}_{C|WZ}) = \mathcal{O}\left(n^{-1}(\min\{h_n, h_d\})^{-k}\right)$. Therefore, the rate of the MISE is:

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

A.7 Proof of Lemma 2

Denote by $\mathbb{E}_{WZ}$ the conditional expectation given (W, Z). Let us consider the decomposition $(\hat{f}^\alpha_{\theta|W} - f^{\dagger c}_{\theta|W})(\theta; w) = [\hat{f}^\alpha_{\theta|W} - (\alpha I + T^*T)^{-1}T^*\mathbb{E}_{WZ}(\hat{f}_{C|WZ})](\theta; w) + [(\alpha I + T^*T)^{-1}T^*\mathbb{E}_{WZ}(\hat{f}_{C|WZ}) - f^\dagger_{\theta|W}](\theta; w) =: A + B$. The result of Lemma 2 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 $\frac{A}{\sqrt{V(A)}} \to^d \mathcal{N}(0, 1)$ where $V(A) = Var(A)$. Let $\{\lambda_j, \varphi_j, \psi_j\}_{j\in\mathbb{N}}$ denote the SVD of T, $\hat{f}_{WZ}$ denote the kernel estimator of the joint pdf of (W, Z) and $K_{h,i}(z, w) = K_h(z_i - z, z)K_h(w_i - w, w)$, then

$$\begin{aligned} A &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^\infty \frac{1}{\alpha + \lambda_j^2}\left\langle T^*\left(K_h(c_i - c, c) - \mathbb{E}_{WZ}(K_h(c_i - c, c))\right)\frac{K_{h,i}(z, w)}{h_n^{k+l+1}\hat{f}_{WZ}}, \varphi_j\right\rangle\varphi_j(\theta; w) \\ &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^\infty \frac{\lambda_j}{\alpha + \lambda_j^2}\left\langle \left(K_h(c_i - c, c) - \mathbb{E}_{WZ}(K_h(c_i - c, c))\right)\frac{K_{h,i}(z, w)}{h_n^{k+l+1}\hat{f}_{WZ}}, \psi_j\right\rangle\varphi_j(\theta; w) =: \frac{1}{n}\sum_{i=1}^n Z_{ni}. \end{aligned}$$

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

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

if $\sum_{i=1}^n \mathbb{E}\left|Z_{ni}/\sqrt{nVar(Z_{ni})}\right|^3 \to 0$ as $n \to \infty$. Lemma **??** in the Supplementary Appendix shows that this latter convergence holds if $\alpha^3/(nh_n^k) \to 0$. To prove $\frac{\mathcal{P}_cA}{\sqrt{V_c(\theta;w)}} \to^d \mathcal{N}(0, 1)$ we use the functional delta method (see e.g. **?** 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 the statement of Lemma 2. Moreover, $V_c(\theta;w) = Var(\mathcal{P}^{\dagger}_c A)$ and $V(A)$ and $V_c(\theta;w)$ have the same rate.

To prove the second result we follow the strategy in the proof of Proposition 6 in **?** and prove that $\frac{B^2}{V(A)} \to 0$. Let us decompose B as

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

then $B^2 \leq 2\left|(\alpha I + T^*T)^{-1}T^*\left(\mathbb{E}_{WZ}(\hat{f}_{C|WZ}) - f_{C|WZ}\right)(\theta;w)\right|^2 + 2\left|\left((\alpha I + T^*T)^{-1}T^* f_{C|WZ} - f^{\dagger}_{\theta|W}\right)(\theta;w)\right|^2.$

Note that $\mathbb{E}_{WZ}(\hat{f}_{C|WZ}) = f_{C|WZ} + \mathcal{O}(h_n^2)$. Then, by using (A.7) and (A.8), under Assumption 8 we conclude that $\frac{B^2}{V(A)} = \mathcal{O}_p\left(n\alpha h_n^{k+4} + n\alpha^2\phi^2(\alpha)h_n^k\right)$ which converges to zero under the conditions of the theorem. Since $\mathcal{P}_c$ is a nonexpansive map, these rates are not affected by replacing B with $\mathcal{P}_c B$ and $V(A)$ with $V_c(\theta;w)$ since $V(A)$ and $V_c(\theta;w)$ have the same rate.

A.8 Proof of Theorem 3

The functional $J_\alpha(h) := ||Th - \hat{f}_{C|WZ}||^2 + \alpha||h||^2$ is a strictly convex and Fréchet differentiable functional with Fréchet derivative $2\left((T^*T + \alpha I)h - T^*\hat{f}_{C|WZ}\right)$. Hence, the convex problem (5.9) has a unique solution $\check{f}^{\alpha,c}_{\theta|W}$ that is characterized as the unique element in $\mathcal{F}_{C|WZ}$ such that the following variational inequality holds:

$$\langle(\alpha I + T^*T)\check{f}^{\alpha,c}_{\theta|W} - T^*\hat{f}_{C|WZ}, f - \check{f}^{\alpha,c}_{\theta|W}\rangle \geq 0, \quad \forall f \in \mathcal{F}_{\theta|W}. \tag{A.9}$$

For every $\alpha > 0$ define the inner product $\langle f_1, f_2\rangle_\alpha = \langle(\alpha I + T^*T)f_1, f_2\rangle$ on $L^2_{\pi_\theta}$. Then (A.9) is equivalent to

$$\langle\check{f}^{\alpha,c}_{\theta|W} - (\alpha I + T^*T)^{-1}T^*\hat{f}_{C|WZ}, f - \check{f}^{\alpha,c}_{\theta|W}\rangle_\alpha \geq 0, \quad \forall f \in \mathcal{F}_{\theta|W}. \tag{A.10}$$

Thus, $\check{f}^{\alpha,c}_{\theta|W} = \mathcal{P}^\alpha_c(\alpha I + T^*T)^{-1}T^*\hat{f}_{C|WZ}$ where $\mathcal{P}^\alpha_c$ denotes the projector onto $\mathcal{F}_{\theta|W}$ with respect to $\langle\cdot,\cdot\rangle_\alpha$. By denoting $f^{\alpha,c}_{\theta|W} = \mathcal{P}^\alpha_c(\alpha I + T^*T)^{-1}T^* f_{C|WZ}$ we can write

$$\begin{aligned}\mathbb{E}||\check{f}^{\alpha,c}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 &\leq 2\mathbb{E}\|\mathcal{P}^\alpha_c(\alpha I + T^*T)^{-1}T^*(\hat{f}_{C|WZ} - f_{C|WZ})\|^2 + 2||f^{\alpha,c}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 \\ &= \mathcal{O}\left(\alpha^{-1}\mathbb{E}||\hat{f}_{C|WZ} - f_{C|WZ}||^2\right) + 2||f^{\alpha,c}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2. \end{aligned} \tag{A.11}$$

It remains to show that $||f^{\alpha,c}_{\theta|W} - f^{\dagger c}_{\theta|W}||$ converges to 0. By definition of $f^{\alpha,c}_{\theta|W}$:

$$\langle(\alpha I + T^*T)f^{\alpha,c}_{\theta|W} - T^*f_{C|WZ}, f^{\dagger c}_{\theta|W} - f^{\alpha,c}_{\theta|W}\rangle \geq 0. \tag{A.12}$$

Define the closed and convex set $U := \{u \in \overline{\mathcal{R}(T)}; P_{\mathcal{F}_{\theta|W}}T^*u = f^{\dagger c}_{\theta|W}\}$ and let $\bar{u}$ be the element of U with minimal norm. It follows that $\langle f^{\dagger c}_{\theta|W} - T^*\bar{u}, f^{\alpha,c}_{\theta|W} - f^{\dagger c}_{\theta|W}\rangle \geq 0$ or, equivalently,

$$\langle T^*\bar{u} - f^{\dagger c}_{\theta|W}, f^{\dagger c}_{\theta|W} - f^{\alpha,c}_{\theta|W}\rangle \geq 0. \tag{A.13}$$

By summing (A.12), with $f_{C|WZ}$ replaced by $Tf_{\theta|W}^{\dagger c}$, and (A.13), multiplied by $\alpha > 0$, we obtain:

$$\langle (\alpha I + T^*T)(f_{\theta|W}^{\alpha,c} - f_{\theta|W}^{\dagger c}) + \alpha T^*\bar{u}, f_{\theta|W}^{\dagger c} - f_{\theta|W}^{\alpha,c}\rangle \geq 0$$

which is equivalent to $||T(f_{\theta|W}^{\alpha,c} - f_{\theta|W}^{\dagger c})||^2 + \alpha||f_{\theta|W}^{\alpha,c} - f_{\theta|W}^{\dagger c}||^2 \leq \alpha\langle T^*\bar{u}, f_{\theta|W}^{\dagger c} - f_{\theta|W}^{\alpha,c}\rangle$. Then, since $\alpha\langle T^*\bar{u}, f_{\theta|W}^{\dagger c} - f_{\theta|W}^{\alpha,c}\rangle \leq \alpha||\bar{u}||\ ||T(f_{\theta|W}^{\alpha,c} - f_{\theta|W}^{\dagger c})||$, it follows that $||T(f_{\theta|W}^{\alpha,c} - f_{\theta|W}^{\dagger c})|| \leq \alpha||\bar{u}||$ and hence, $||f_{\theta|W}^{\alpha,c} - f_{\theta|W}^{\dagger c}||^2 \leq \alpha^2||\bar{u}||^2$ which converges to 0. From (A.11) and this result, we conclude that: $\mathbb{E}||\check{f}_{\theta|W}^{\alpha,c} - f_{\theta|W}^{\dagger c}||^2 \to 0$.

A.9 Proof of Theorem 4

The first part of the proof is the same as the proof of Theorem 3. Thus, (A.11) still holds and we only have to determine the rate of $||f_{\theta|W}^{\alpha,c} - f_{\theta|W}^{\dagger c}||$. To do this we slightly modify the proof of Lemma 3.9 in **?**. For every $f \in \mathcal{N}(T)^{\perp}$, let $f_{\theta|W}^{\alpha,L_f}$ be the solution of (5.9) with $\hat{f}_{C|WZ}$ replaced by $f_{C|WZ}$ and $\mathcal{F}_{\theta|W}$ replaced by $L_f := \{h \in L^2_{\pi_\theta}; \langle f_{\theta|W}^{\dagger c} - f, h - f_{\theta|W}^{\dagger c}\rangle = 0\}$. Note that L_f is the supporting hyperplane of $\mathcal{F}_{\theta|W}$ in $f_{\theta|W}^{\dagger c}$. By the triangular inequality:

$$||f_{\theta|W}^{\alpha,c} - f_{\theta|W}^{\dagger c}|| \leq ||f_{\theta|W}^{\alpha,c} - f_{\theta|W}^{\alpha,L_f}|| + ||f_{\theta|W}^{\alpha,L_f} - f_{\theta|W}^{\dagger c}||. \tag{A.14}$$

We start by analyzing the second term in (A.14). Since $f_{\theta|W}^{\alpha,L_f}, f_{\theta|W}^{\dagger c} \in L_f$ then $\tilde{P}_f(f_{\theta|W}^{\dagger c} - f_{\theta|W}^{\alpha,L_f}) = f_{\theta|W}^{\dagger c} - f_{\theta|W}^{\alpha,L_f}$. Therefore, the rate of the second term can be easily determined if we show that $\tilde{P}_f f_{\theta|W}^{\alpha,L_f}$ is an unconstrained Tikhonov-regularized solution of the form $\tilde{P}_f f_{\theta|W}^{\alpha,L_f} = (\tilde{P}_f T^*T\tilde{P}_f + \alpha I)^{-1}\tilde{P}_f T^*T\tilde{P}_f f_{\theta|W}^{\dagger c}$. In order to show this, we start by showing that

$$\tilde{P}_f\left(T^*Tf_{\theta|W}^{\alpha,L_f} + \alpha f_{\theta|W}^{\alpha,L_f} - T^*f_{C|WZ}\right) = 0. \tag{A.15}$$

This is equivalent to show that $\langle T^*Tf_{\theta|W}^{\alpha,L_f} + \alpha f_{\theta|W}^{\alpha,L_f} - T^*f_{C|WZ}, h\rangle = 0$ for every $h \in \tilde{L}_f$. Let $h \in \tilde{L}_f$, then $h + f_{\theta|W}^{\dagger c} \in L_f$. By definition of $f_{\theta|W}^{\alpha,L_f}$, the variational inequality

$$\langle (\alpha I + T^*T)f_{\theta|W}^{\alpha,L_f} - T^*f_{C|WZ}, h' - f_{\theta|W}^{\alpha,L_f}\rangle \geq 0, \qquad \forall h' \in L_f \tag{A.16}$$

holds with equality (remark that L_f is a linear manifold). Therefore, for $h \in \tilde{L}_f$,

$$\langle (\alpha I + T^*T)f_{\theta|W}^{\alpha,L_f} - T^*f_{C|WZ}, h\rangle + \langle (\alpha I + T^*T)f_{\theta|W}^{\alpha,L_f} - T^*f_{C|WZ}, f_{\theta|W}^{\dagger c} - f_{\theta|W}^{\alpha,L_f}\rangle = 0 \tag{A.17}$$

where the second term is equal to 0 by applying (A.16) with equality and since $f_{\theta|W}^{\dagger c} \in L_f$. We conclude that $\langle (\alpha I + T^*T)f_{\theta|W}^{\alpha,L_f} - T^*f_{C|WZ}, h\rangle = 0$ for every $h \in \tilde{L}_f$. This proves (A.15).

By using the result of Lemma **??** in the Supplementary Appendix and by rearranging terms we get: $\tilde{P}_f f_{\theta|W}^{\alpha,L_f} = (\tilde{P}_f T^*T\tilde{P}_f + \alpha I)^{-1}\tilde{P}_f T^*T\tilde{P}_f f_{\theta|W}^{\dagger c}$. By Lemma 3.5 (a) in **?**: $f_{\theta|W}^{\dagger c} = f_{\theta|W}^{\dagger L_f}$, where $f_{\theta|W}^{\dagger L_f}$ satisfies:

$$\begin{aligned} ||Tf_{\theta|W}^{\dagger L_f} - f_{C|WZ}|| &= \inf\left\{||Th - f_{C|WZ}||; h \in L_f\right\} \\ ||f_{\theta|W}^{\dagger L_f}|| &= \min\left\{||h||; h \in L_f \text{ and } ||Th - f_{C|WZ}|| = ||Tf_{\theta|W}^{\dagger L_f} - f_{C|WZ}||\right\}. \end{aligned}$$

Finally, by using the regularity condition $N_\beta \neq \varnothing$ we conclude that

$$\begin{aligned} ||f^{\alpha,L_f}_{\theta|W} - f^{\dagger c}_{\theta|W}||^2 &= ||\tilde{P}_f(f^{\alpha,L_f}_{\theta|W} - f^{\dagger c}_{\theta|W})||^2 = ||(\tilde{P}_f T^* T \tilde{P}_f + \alpha I)^{-1} \tilde{P}_f T^* T \tilde{P}_f f^{\dagger c}_{\theta|W} - \tilde{P}_f f^{\dagger c}_{\theta|W}||^2 \\ &= ||\alpha(\tilde{P}_f T^* T \tilde{P}_f + \alpha I)^{-1} \tilde{P}_f f^{\dagger c}_{\theta|W}||^2 = \mathcal{O}(\alpha^{\beta \wedge 2}). \end{aligned} \quad (A.18)$$

We now consider the first term of (A.14). By Lemma 3.6 (b) in **?** the following inequality holds for $\alpha > 0$ sufficiently small and $f \in M$: $||T(f^{\alpha,c}_{\theta|W} - f^{\alpha,L_f}_{\theta|W})||^2 + \alpha||f^{\alpha,c}_{\theta|W} - f^{\alpha,L_f}_{\theta|W}||^2 \leq ||T(f^{\dagger c}_{\theta|W} - f^{\alpha,L_f}_{\theta|W})||^2 + \alpha||f^{\dagger c}_{\theta|W} - f^{\alpha,L_f}_{\theta|W}||^2$. This implies that $||f^{\alpha,c}_{\theta|W} - f^{\alpha,L_f}_{\theta|W}||^2 \leq \frac{1}{\alpha}||T(f^{\dagger c}_{\theta|W} - f^{\alpha,L_f}_{\theta|W})||^2 + ||f^{\dagger c}_{\theta|W} - f^{\alpha,L_f}_{\theta|W}||^2$. From (A.18) and the fact that $||T(f^{\dagger c}_{\theta|W} - f^{\dagger L_f}_{\theta|W})||^2 = \mathcal{O}(\alpha^{\beta \wedge 2+1})$ we conclude that $||f^{\alpha,c}_{\theta|W} - f^{\alpha,L_f}_{\theta|W}||^2 = \mathcal{O}(\alpha^{\beta \wedge 2})$. By putting all these results together we have proved the results of the theorem.

A.10 Proof of Theorem 5

Part *(i)* follows from Lemma **??** in the Supplementary Appendix. Hence, we prove *(ii)* Let $\mathcal{U}_w(g^0)$ denote an open neighborhood in $\mathcal{G}$ in the weak topology around g^0 and $\mathcal{U}_w(\theta_1^0)$ denote its projection onto Θ_1, that is, $\mathcal{U}_w(\theta_1^0) = \{\theta_1 \in \Theta_1; \exists h \in \mathcal{F}_{\theta_2|W} \text{ such that } (\theta_1, h) \in \mathcal{U}_w(g^0)\}$. Hence, because $\Theta_1 \subset \mathbb{R}^d$ and because the weak and norm topologies coincides on finite dimensional spaces, then $\|\hat{\theta}_1\|_E \to 0$ in probability if and only if $P(\hat{\theta}_1 \in \mathcal{U}_w(\theta_1^0)) \to 1$. This last result follows from Lemma **??** in the Supplementary Appendix and the inequality $P(\hat{\theta}_1 \in \mathcal{U}_w(\theta_1^0)) \geq P(\hat{g} \in \mathcal{U}_w(g^0))$.

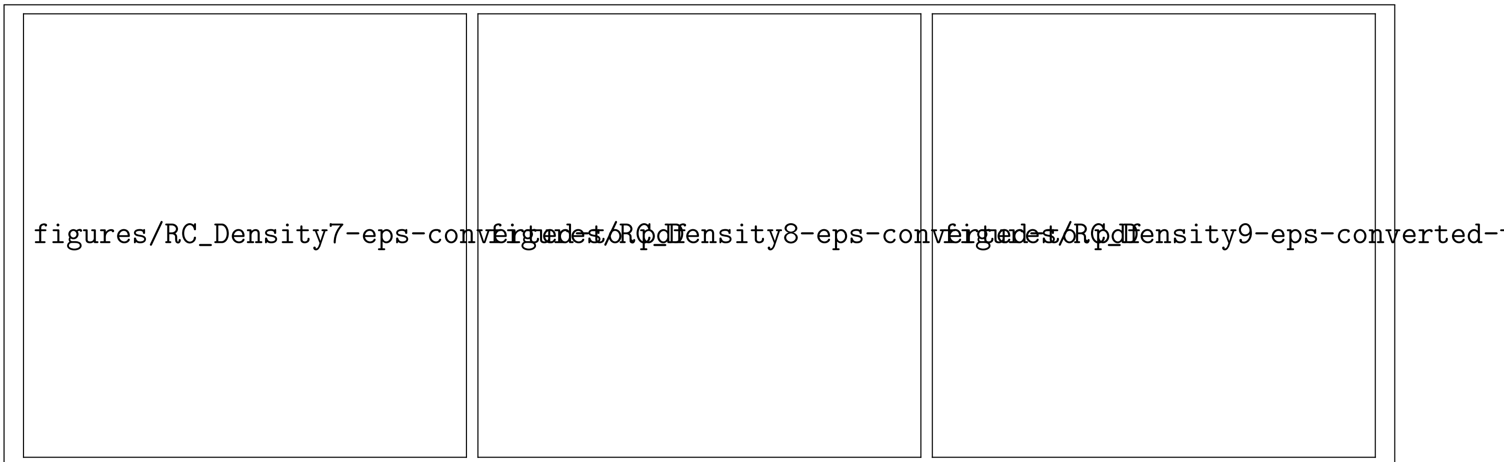
Next, we show $\|\hat{f}_{\theta_2|W} - f^0_{\theta_2|W}\| \to 0$ in probability. Lemma **??** shows consistency under the weak topology which implies $\langle \tilde{g}, \hat{f}_{\theta_2|W} - f^0_{\theta_2|W}\rangle$ for every $\tilde{g} \in \Theta_1 \times \mathcal{F}_{\theta|W}$. From Lemma **??**

$$\begin{aligned} \|\hat{f}_{\theta_2|W}\|^2 - \|f^0_{\theta_2|W}\|^2 &\geq \langle f^0_{\theta_2|W}, \hat{f}_{\theta_2|W} - f^0_{\theta_2|W}\rangle + c\|\hat{f}_{\theta_2|W} - f^0_{\theta_2|W}\|^2 \\ &= \langle g^0, \hat{g} - g^0\rangle - \langle \theta_1^0, \hat{\theta}_1 - \theta_1^0\rangle + c\|\hat{f}_{\theta_2|W} - f^0_{\theta_2|W}\|^2. \end{aligned}$$

The first two terms of the right hand side converge to zero in probability by Lemma **??** and the left hand side converges to zero in probability by Lemma **??** *(i)*. Hence, $\|\hat{f}_{\theta|W} - f^0_{\theta|W}\|^2 \to 0$ in probability.

B Figures

Figure 2:
Log linear demand example
$\mathcal{P}_c\widehat{f}_{\theta|W}$ vs. true $f_{\theta|W}$
(lower panel) (upper panel).


figures/RC_Density7-eps-converted-to.pdf
figures/RC_Density8-eps-converted-to.pdf
figures/RC_Density9-eps-converted-t

(a) $w = 1.2839$ (b) $w = 1.4849$ (c) $w = 1.673$

Figure 3: Log linear demand example
Oracle vs. Tikhonov estimator
(kernel density of WMSE)

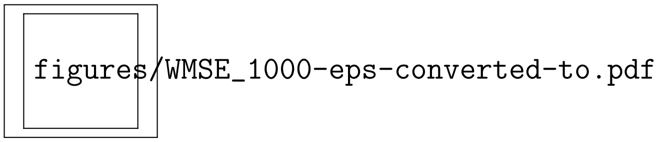


Figure 4:
CARA example
density of δ

figures/ThetaDensity1-eps-converted-to.pdf

(a) (low assets, low income)

figures/ThetaDensity2-eps-converted-to.pdf

(b) (low assets, high income)

Figure 5:
CARA example
density of δ

figures/ThetaDensity3-eps-converted-to.pdf

(a) (high assets, low income)

figures/ThetaDensity4-eps-converted-to.pdf

(b) (high assets, high income)

Figure 6:
CARA example 1
quantile level sets of δ

figures/levelsets5-eps-converted-to.pdf

(a) (low assets, low income)

figures/levelsets6-eps-converted-to.pdf

(b) (low assets, high income)

Figure 7:
CARA example
quantile level sets of δ

figures/levelsets7-eps-converted-to.pdf

(a) (high assets, low income)

figures/levelsets8-eps-converted-to.pdf

(b) (high assets, high income)