Erasmus Mundus Masters in Complex Systems

THE ROLE OF UNOBSERVED HETEROGENEITY IN THE RETURNS TO EDUCATION

July 11, 2012

H. Broome

Supervisor:
Dr. Eric Gautier, Centre de Recherche en Economie et Statistique, Paris

Mincer was one of the first people to study the return to education in the 1970s. He proposed a linear relationship between years of education, years of experience and wages that assumes the economic environment is static. The 1960 US Census data that was used by Mincer showed an approximately static economic environment insofar as the relationships between wages, education and experience that he predicted could be observed. However applying the Mincer model to data after 1980 no longer produces valid estimates [17].

The impact of education does not stay constant over time and it is necessary to build a model that is consistent with current empirical evidence. Recent studies highlight the large variation in the returns to college education experienced by individuals who are observationally identical in the data [10]. Consequently any model needs to account for heterogeneous returns to education that are influenced by unobservable factors. Furthermore an individual's expectations about what they stand to gain influence their decision about going to college [10]. Any decision model needs to include the influence of unobservables to account for the way individuals act as if they possess information that is not available in the data.

We work with a binary decision model that assumes individuals go to college if their expected net utility is positive. The net utility is calculated as a combination of observed data and unobserved factors that will include the individual's expected gains. In addition we use a model that relates an individual's observed wage to their college decision in a way that allows for individual specific effects to ensure the possibility of heterogeneous outcomes.

The main contribution of the project is to develop a strategy for estimating the unobserved returns to education conditional on the unobserved heterogeneity. We use nonparametric methods for estimation which requires us to choose a smoothing parameter. However typical methods for selecting the appropriate smoothness rely on having access to observed data. To be able to implement our estimator of unobservables we develop a criterion for choosing the level of smoothing and use a simulation study to test the performance.

Abstract

This paper studies the average returns to college education conditional on the unobservable heterogeneity that influences both education decisions and wage outcomes. The recent instrumental variables literature has tried to estimate this parameter for a scalar unobservable but it does so only under an unrealistic assumption that imposes monotonicity on the decision process. We show how to utilize the Radon transform with continuous instruments in a random coefficient model to estimate the average returns to college education conditional on a vector of unobservables that we call $CATE$. Our specification is flexible because it allows for complex unobserved heterogeneity of individuals and non-monotonic decisions. The main contribution of this project is to implement two estimation strategies for recovering $CATE$. In particular we construct a new regularized inverse for the Radon transform and a method for parameter selection based on observable data. A simulation study compares the performance of the estimation strategies and demonstrates that our parameter selection method is a reliable guide for tuning the estimators.

Contents

1 Introduction

The return to schooling has been the subject of continued interest for over fifty years and quantifying the returns remains an open question. Estimates for the causal impact of schooling on wages in the US are reported to be in the range of $4-7\%$ per year by some sources and $10-15\%$ by others [8]. The discrepancy is due to different statistical approaches and it remains even when the same data set is used.

The difficulty in untangling the influence of education on wage outcomes[1] is that an individual's decision about their education is potentially related to their unobserved abilities, tastes or social influences and, in turn, these unobserved qualities have an impact on the individual's wage. When a decision both depends on unobservables and is correlated with the gains that result from a decision then standard econometric tools produce biased results [10].

In this paper we focus on one major education decision; whether to go to college or not. We use a flexible model that allows for heterogeneity in returns to college and in decisions. Our aim is to investigate the relationship between unobserved heterogeneity and the returns to college education by estimating the average returns conditional on unobservables that we call $CATE$. By starting with a conditional average we can ask questions about the average returns for sub-populations and move beyond talking about 'a' return to schooling.

The structure of this report is as follows; section two introduces the model and the conditional average returns, $CATE$, which is the object we will estimate. Section three explains the estimation strategy which is based on the Radon transform while section four outlines the key parameters involved in the estimation and how they can be chosen. Finally section five presents a simulation study of the estimators that compares two approaches for estimation and looks at the implementation of parameter selection before an outlook concludes.

2 The Model

This section introduces the formal set up for the analysis of the returns to college education Δ which will be estimated via the conditional average returns $CATE$. We model the observed log wages Y in terms of the decision to go to college D with a separate model for the decision process itself.

$$Y = Y_0 + \Delta D, \quad \text{where } \Delta = Y_1 - Y_0 \tag{1}$$

$$D = \mathbf{1}\{\tilde{V} - \tilde{\Gamma}\tilde{Z} - \tilde{\Theta} > 0\} \tag{2}$$

The outcome equation (1) is in the form of a random coefficient model where we observe Y, D for each individual. The partial effect of D on Y is captured by the coefficient Δ which varies with the individual, finally Y_0 absorbs the constant and error term from the standard linear regression model. Y_0 is the wage in the base state corresponding to no college education and $\Delta = Y_1 - Y_0$ is the return to education for an individual under a hypothetical shift from state 0 to state 1.

[1] Recent work also consider influences on non-monetary outcomes such as health but we restrict out focus to wages in this project [11].

We explicitly model the state an individual is in through the decision model (2). Individuals choose college when their expected net utility is positive. The net utility depends on the cost factors or information contained in the observed instruments[2] $\tilde{V}, \tilde{Z}$ and on the unobserved scalar random coefficients $\tilde{\Gamma}, \tilde{\Theta}$.[3] The important aspect of our decision model is the inclusion of multiple sources of heterogeneity in the form of a vector of random coefficients. The random intercept $-\tilde{\Theta}$ contains unobserved contributions to the net utility such as the expected gains Δ while the random coefficient $\tilde{\Gamma}$ reflects the individual specific impact of the instruments on the net utility.

In this report we work with two dimensions[4] of heterogeneity $\Gamma = (\tilde{\Gamma}, \tilde{\Theta})^\top$ which is the vector of coefficients for $(\tilde{Z}, 1)^\top$. We rescale the decision model so that $(\tilde{Z}, 1)$ is a unit vector that we denote as $e_\Phi = (\cos(\Phi), \sin(\Phi))$ for some $\Phi \in [0, \pi]$ and let $V = \tilde{V}/|(\tilde{Z}, 1)|$ where $|\cdot|$ is the Euclidean norm. The rescaled model is

$$D = \mathbf{1}\{V > \Gamma^\top e_\Phi\}$$

Recent work on decision models that allow for dependence on unobservables and correlations with the returns Δ were based on an additive scalar unobservable [6][2]. A consequence of using an additive unobservable is that it places restrictions on the heterogeneity of decisions [4]. The key innovation in our approach is that we can recover a relationship between returns to education and unobservables from a decision model that allows for heterogeneity in decisions. Because of this generality Heckman and Vytlacil refer to (2) as a benchmark nonseparable, nonmonotonic model [2].

While we allow for a rich structure in terms of unobservables we require that the model is linear in the parameters which is potentially restrictive. Instead of trying to capture higher order terms in the observable variables we want to emphasize the dependence of the college decision on an unobserved structure. In addition we require the following assumptions for our estimation process.

Assumption 1 **(A-1)** $0 < \mathbb{P}(D = 1) < 1$*;*

(A-2) $V, \Phi \perp Y_0, \Gamma$ *and* $V, \Phi \perp Y_1, \Gamma$ *where* $\perp$ *denotes independence;*

(A-3) *The distribution of* (V, Φ, Γ) *is absolutely continuous with respect to the Lebesgue measure;*

(A-4) *The* $supp(f_\Phi) = [0, \pi]$ *and for every* $\phi \in [0, \pi]$

$$\left[\inf_{\gamma \in supp(f_\Gamma(\cdot))} \gamma^\top e_\phi, \quad \sup_{\gamma \in supp(f_\Gamma(\cdot))} \gamma^\top e_\phi \right] \subset supp(f_{V|\Phi}(\cdot|\phi).$$

Assumption **(A-1)** requires that this is a fraction of the population who goes to college and a fraction that does not to ensure there is some variation. **(A-2)** and **(A-3)** are standard assumptions in the instrumental variables literature and require

[2] An instrumental variable Z is defined to be correlated with D but independent of Y. In the case of education examples include distance to college or tuition fees. We use instruments to deal with the endogeneity problem due to the correlation between D and Δ.

[3] As the net utility can only be identified up to scale we set the coefficient of V to 1 under the assumption that the coefficient has a known sign and w.l.o.g. it is positive.

[4] This model can be generalized to higher dimensions when $\tilde{\Gamma}, \tilde{Z} \in \mathbb{R}^d$.

that the instruments V, Φ are continuous and independent of the random parameters. Lastly **(A-4)** is a large support assumption that is necessary for the Radon transform which is the core of our estimation strategy and is explained in the following sections.

The main aim of this project is to implement an estimator for the average returns to college education Δ conditioned on unobservable characteristics Γ,

$$CATE(\gamma) = \mathbb{E}[\Delta|\Gamma = \gamma].$$

$CATE$ is the average returns for the subpopulation with unobserved heterogeneity vector equal to γ. It is also the average returns for individuals who would be indifferent about attending college if they were exogenously assigned a value of $(\Phi, V) = (\phi, v)$ such that $\gamma^\top e_\phi = v$. Furthermore it is independent of the instruments (Φ, V) due to Assumption **(A-2)** which implies that it can be used for policy analysis. These are all essential properties of the marginal treatment effect (MTE) defined in [2]. We think that $CATE$ is a natural extension of MTE to a scenario with a vector of unobservables and the freedom for heterogeneous decisions. As in [2] a large variety of measures based on averages can be written as weighted averages of $CATE$.

If $CATE$ is not constant in the direction of $\tilde{\Gamma}$ it is an indication of heterogeneous cost factors associated with the instruments, namely a non-linearity in the influence of the instruments on the decision and hence on the returns. Similarly if $CATE$ is not constant in the direction of $\tilde{\Theta}$ it demonstrates the existence of heterogeneous unobservables that are included in the net utility of going to college.

3 Estimation with the Radon Transform

The difficulty with estimating $CATE$ is that it is an expectation of the *unobserved* gains Δ conditioned on *unobservable* characteristics Γ. To express $CATE$ in terms of observables we use the Radon transform R which is a bounded linear operator typically associated with tomography. Statistical inverse problems involving a Radon type operator have previously been used to estimate the distribution of random coefficients in a linear model [5].

Definition 1 *The **Radon transform** applied to a function $f \in L_1(\mathbb{R}^2)$ yields the integral of f over the line $L_{\phi,u} := \{\gamma : \gamma^\top e_\phi = u\}$ with respect to the Lebesgue measure on the line, $d_{L_{\phi,u}}(\gamma)$,*

$$R[f](\phi, u) = \int_{L_{\phi,u}} f(\gamma) d_{L_{\phi,u}}(\gamma).$$

Under Assumption 1 we can relate $CATE$ and the distribution of unobservables f_Γ to the observables Φ, V, Y, D by using the Radon transform.[5]

$$\begin{aligned}\mathbb{E}[Y|(\Phi, V) = (\phi, v)] &= \int_{\infty}^{v} R[\mathbb{E}[\Delta|\Gamma = \cdot] f_\Gamma(\cdot)](\phi, u) du \\ &= \int_{\infty}^{v} R[CATE(\cdot) f_\Gamma(\cdot)](\phi, u) du\end{aligned}$$

[5]Throughout this paper we use $f_X(x)$ to refer to the distribution of a random variable X at the point $X = x$.

The details of the derivation and further discussion of the Radon transform can be found in Appendix A.

3.1 Regularized Radon Inverse

The basic observation of our model is that a function of observables is equal to the integral of the Radon transform of an expression involving unobservables. To recover the unobservables we would like to invert the Radon operator. This is an ill-posed inverse problem as the inversion of the Radon transform is not continuous; introducing a small amount of error to the argument may result in large changes in the value. To manage this problem we use a regularized inverse A_T of the Radon transform. In our two dimensional setting it is an operator $A_T : \{h : [0, \pi] \times \mathbb{R} \to \mathbb{R}\} \to \{f : \mathbb{R}^2 \to \mathbb{R}\}$ defined as

$$A_T[h](\gamma) = \int_0^\pi \int_{-\infty}^\infty K_T(e_\phi^\top \gamma - u) h(\phi, u) du d\phi, \qquad \gamma \in \mathbb{R}^2 \tag{3}$$

with the property that $\lim_{T\to\infty} \|A_T R[h] - h\|_{L^2} = 0$. The function K_T which shows up in the integrand of A_T is similar to a smoothing kernel in nonparametric statistics and is classically defined with respect to a Fourier transform involving an indicator function [5]. Explicitly $K_T(x) = (2\pi^2)^{-1} \int_0^T \cos(tx) t dt$, where T is the regularization parameter that controls the degree of smoothing; a smaller T corresponds to greater smoothing and a decreased frequency of oscillations.

In the estimation of $CATE$ the argument of A_T is a function of the form $\partial_v E$ where $E : (\phi, v) \mapsto \mathbb{E}[g(Y, D)|(\Phi, V) = (\phi, v)]$ is a regression function. The argument has to be estimated prior to the regularized inversion which risks introducing substantial error early in the process. The estimation of the derivative of a regression function is more difficult than a standard regression and the choice of regularization parameter is a topic of ongoing research [7].

A theoretical contribution of this project is the construction of an alternative regularized inverse. By changing the form of K_T it is possible to use integration by parts to remove the difficult derivative estimation. As an alternative to the regularized inverse A_T we define an alternative regularized inverse B_T to be the same way as A_T (3) except in the integrand we replace the function K_T by J_T where

$$J_T(x) = \frac{1}{2\pi^2} \int_0^T \cos(tx) t \psi\left(\frac{t}{T}\right) dt$$

and $\psi(x) = c \exp\left(\frac{-1}{1-|x|^2}\right)$ is defined for $x \in B_2(0, 1)$ with a normalization constant c.[6] The details can be found in Appendix B. The important feature is that the function ψ is an element of the space of rapidly decreasing functions on $\mathbb{R}^2$ known as the Schwartz space $\mathcal{S}(\mathbb{R}^2)$. The properties of the Schwartz space[7] guarantee that we can do integration by parts on the regularized inverse $B_T[\partial_v E]$ and exploit the resulting iterated expectation to get

$$B_T[\partial_v E](\gamma) = \mathbb{E}\left[\frac{\tilde{J}_T(e_\Phi^\top \gamma - V) g(Y, D)}{f_{\Phi,V}(\Phi, V)}\right], \quad \text{where } \tilde{J}_T(x) = J_T'(x). \tag{4}$$

[6] We introduce the normalization constant c to guarantee that $\int_{\mathbb{R}} J_T(x) dx = 1$.
[7] For all $1 \le p \le \infty, \mathcal{S}(\mathbb{R}^2) \subset L^p(\mathbb{R}^2)$ and the Fourier transform as an isomorphism on $\mathcal{S}(\mathbb{R}^2)$.

The trimmed sample counterpart to the expectation (4) is,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\tilde{J}_T(e_{\phi_i}^\top\gamma - v_i)g(y_i,d_i)}{f_{\Phi,V}(\phi_i,v_i)}\mathbf{1}\{f_{\Phi,V}(\phi_i,v_i)>\tau\} \tag{5}$$

which requires an unknown, but estimatable, density $f_{\Phi,V}$ in the denominator. To evaluate (5) we replace $f_{\Phi,V}$ with an estimator which we call a plug-in. To avoid division by values that are too close to zero we including a trimming parameter τ. Throughout this report we trim 2% of the data and the results are robust for $1-5\%$ trimming. The density estimation of $f_{\Phi,V}$ can easily be done in an application and the sum is simple to evaluation which makes this method computationally appealing.

3.2 The Estimators

We can express $CATE(\gamma) = \mathbb{E}[\Delta|\Gamma = \gamma]$ as a ratio of functions of observables involving the Radon inverse

$$CATE(\gamma) = \frac{R^{-1}\left[\partial_v\mathbb{E}[Y|(\Phi,V)=\cdot]\right](\gamma)}{R^{-1}\left[\partial_v\mathbb{E}[D|(\Phi,V)=\cdot]\right](\gamma)} = \frac{R^{-1}\left[\partial_v\mathbb{E}[Y|(\Phi,V)=\cdot]\right](\gamma)}{f_\Gamma(\gamma)}.$$

The expression for the numerator of $CATE$ and the distribution of unobservables f_Γ in the denominator can be estimated using a regularized inverse with either a multidimensional integral to calculate A_T or a finite sum to evaluate B_T.

As an example we present two estimators for f_Γ based on A_T and B_T respectively and remark that the numerator for $CATE$ can be estimated in the same way with a Y replacing D in the expectation in (6) and y_i replacing d_i in the numerator of the sum in (7).

$$\hat{f}_\Gamma^A(\gamma) = \int_0^\pi\int_{-\infty}^{\infty} K_T(e_\phi^\top\gamma - u)\overline{\widehat{\partial_v\mathbb{E}[D|(\Phi,V)=(\phi,u)]}}dud\phi \tag{6}$$

$$\hat{f}_\Gamma^B(\gamma) = \frac{1}{N}\sum_{i=1}^{N}\frac{\tilde{J}_T(e_{\phi_i}^\top\gamma - v_i)d_i}{\hat{f}_{\Phi,V}(\phi_i,v_i)}\mathbf{1}\{\hat{f}_{\Phi,V}(\phi_i,v_i)>\tau\} \tag{7}$$

where $\bar{h}$ denotes the extension of the function h to zero outside of its domain of definition and we use $\hat{h}$ to show that we are using an estimator of h.

The estimation with A_T is a two stage process. First we must estimate the argument of A_T, $\partial_v\mathbb{E}[g(Y,D)|(\Phi,V)=(\phi,u)]$ where $g(Y,D)=Y$ for the numerator and $g(Y,D)=D$ for the denominator. The first stage estimation of the derivative requires its own regularization parameter. Secondly we need to choose the amount of smoothing for the regularized inverse A_T by appropriately picking a value of T as a function of sample size. With the two stage process both components of the integrand of (3) have a separate regularization parameter which provides more scope for fitting A_T.

By contrast the estimation with B_T only has one stage. By using integration by parts and rewriting the iterated expectation we removed the need to explicitly estimate $\partial_v\mathbb{E}[D|(\Phi,V)=(\phi,u)]$. The only task is to select the T which controls B_T and is now the only major parameter controlling the smoothing. We are effectively

smoothing both components of the integrand, J_T and $\partial_v \mathbb{E}[D|(\Phi, V) = (\phi, u)]$, at once. In addition this method requires a trimming parameter and an estimator for $f_{\Phi,V}$ as discussed above.

4 Parameter Selection

We begin with a description of standard criteria used for parameter selection and then develop a data-driven method for tuning the regularization parameter T that is feasible for an application with either A_T or B_T. In addition we discuss the parameter selection for the first stage plug-in estimator $\partial_v \mathbb{E}[D|(\Phi, V) = (\phi, u)]$.

4.1 Objective Functions for Parameter Selection

Given independent and identically distributed observations $(x_i)_{i=1}^N$ of a random variable X the density f_X can be estimated with non-parametric methods and the regularization parameter T can be chosen to minimize the mean squared error $\int (\hat{f}_X^T(x) - f_X(x))^2 dx$. As f_X is unknown the mean squared error cannot be calculated directly. If T minimizes the mean squared error then it minimizes the expression $\int (\hat{f}_X^T(x))^2 dx - 2 \int \hat{f}_X^T(x) f_X(x) dx$ for which there is an unbiased estimator

$$\int \left(\hat{f}_X^T(x)\right)^2 dx - \frac{2}{N} \sum_{i=1}^{N} \hat{f}_X^T(x_i). \tag{8}$$

However, we want to estimate the density of the *unobserved* random vector Γ where it is not possible to calculate the sum in (8) as we cannot evaluate $\hat{f}_\Gamma(\gamma_i)$ when γ_i are unobserved. We are particularly interested in the unobserved heterogeneity Γ and returns to education Δ because they provide an insight into how the benefit of schooling varies across a population. To recover these unobservables we have introduced an ill-posed inverse problem and now face a second challenge of how to adequately choose the amount of smoothing, as a function of sample size, to regularize the inverse.

In the simulation study we begin by choosing all the regularization parameters to minimize an objective function that relies on a comparison with the true function. We consider both the mean squared error and the sup-norm as objective functions that can provide a benchmark for the parameter selection. Subsequently we construct a third data-driven approach that is appropriate for an application as it minimizes an objective function that depends on observables.

A data-driven method for selecting T. To select T we minimize the difference between two estimators of the same object. We discuss the method in terms of f_Γ for which we compare two estimators, $\hat{P}, \bar{P}$ of the probability that an individual goes to college and has observable characteristics (Φ, V) within a certain range specified by the box B, namely $\mathbb{P}(D = 1 \cap \{\Phi, V\} \in B)$. The estimator $\hat{P}$ involves

a functional of $\hat{f}_\Gamma$ while $\bar{P}$ is calculated directly from the data as follows;

$$\hat{P}_B(T) = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbf{1}\{(\phi_i, v_i) \in B\} \int_{\mathbb{R}^2} \mathbf{1}\{v_i > \gamma^\top e_{\phi_i}\} \hat{f}_\Gamma^T(\gamma) d\gamma$$

$$\bar{P}_B = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} d_i \mathbf{1}\{(\phi_i, v_i) \in B\}$$

where we estimate $\hat{f}_\Gamma$ on 80% of the sample and calculate our criterion Err on the complementary sample of size N_{test}. A coarse grid of boxes is used to partition the data into equal sized groups and T is chosen to minimize

$$Err(T) = \sum_{j=1}^{M} |\hat{P}_{B_j}(T) - \bar{P}_{B_j}| \tag{9}$$

which is the sum of the difference in the estimators across all the boxes. Multiple boxes are used to ensure that the estimator performs reasonably in different areas.

Analogously we choose T for the estimator of the numerator of $CATE$ to minimize the difference between a direct, and indirect, estimate of $\mathbb{E}[Y\mathbf{1}\{(\Phi, V) \in B\}|(\Phi, V)]$.

4.2 Parameter Selection for the Argument of A_T

We use a non-parametric local polynomial estimators of $\partial_v \mathbb{E}[g(Y, D)|(\Phi, V) = (\phi, u)]$. The advantage of local polynomials is they have a lower bias at the boundaries compared to kernel or spline methods [9]. However the estimation of a derivative of a regression function is a difficult problem. The regularization parameter selection for estimators of derivatives is a field of active research [7]. We leave the implementation of a data-driven tuning of the derivative estimation for future research.

In the simulation study we choose the regularization parameter to minimize the mean squared error between the local polynomial estimator and the truth. In the estimation of f_Γ when $g(Y, D) = D$ we have $\partial_v \mathbb{E}[g(Y, D)|(\Phi, V) = (\phi, u)]$ is equal to $R[f_\Gamma(\cdot)](\phi, v)$ which is positive by definition of a slice integral of a density. We take the maximum of the local polynomial estimate and 0 which significantly reduces the error and contributes to the success of A_T for estimating f_Γ.

5 Simulation Study

The simulation study provides a guide for how to choose the regularization parameter T in an application and compares the two estimators of the regularized Radon transform (6) and (7) based on the regularized inverse operators A_T and B_T respectively. Three criteria for choosing T are compared in the estimation of f_Γ and it is shown that our data-driven method selects a reasonable T. After selecting suitable parameters for the estimation we compare the performance of the two estimation methods and show that they are able to recover the relationship between unobserved heterogeneity and the returns to education.

We use the statistical software R [12] with additional packages for local polynomial regression [13], adaptive quadrature [14][15] and parallel computing [16].

5.1 Simulation Set-up

Our simulation set up is based on independent normal distributions where Φ is a truncated normal with support on $[0, \pi]$. For the estimation of f_Γ we also experimented with a mixture of multivariate normals for which the estimators performed equally well but maintain a simpler set up here due to difficulties with $CATE$. We explicitly model the unobservables $\Delta, \Gamma = (\tilde{\Gamma}, \tilde{\Theta}), Y_0$ along with the observables Φ, V to define $D = \mathbf{1}\{V > \Gamma^\top e_\Phi\}$ and $Y = Y_0 + \Delta D$.

$$\begin{array}{lll} \tilde{\Gamma} \sim \mathcal{N}(1,1), & V \sim \mathcal{N}(-0.2, 2), & Y_0 = 2\tilde{\Gamma} + \tilde{\Theta} \\ \tilde{\Theta} \sim \mathcal{N}(-\frac{1}{2}, 1), & \Phi \sim \mathcal{N}_{[0,\pi]}(\frac{\pi}{2}, \frac{\pi}{4}), & \Delta = \frac{1}{2}\tilde{\Gamma} - 2\tilde{\Theta} \end{array}$$

The key intuition for the simulation set up comes from the role of the random intercept term in the decision model, $\tilde{\Theta}$, which contains the negative expected gains. We define $\tilde{\Theta}$ so that it is negatively correlated with the real gains Δ and is the dominant term in the expression for Δ. Subsequently the other parameters were chosen so that D takes the value one approximately half the time which is in line with empirical data [3].

5.2 Comparison of Objective Functions for Parameter Selection

In the simulation study we use two objective functions that are not available for an application, the mean squared error and the sup-norm. They provide a benchmark for the appropriate choice of T. In addition we use the data-driven objective function Err defined in section 4.1. We apply all three methods to choose T for the estimator $\hat{f}_\Gamma^{B,T}$ which is our preferred method for an application due to the difficulty in estimating the first stage plug-in for A_T. Values of $T \in [1.1, 1.9]$ produce an estimator $\hat{f}_\Gamma^{B,T}$ that has one central peak in a similar location to the true f_Γ. The three criteria focus on different features of the estimator but all choose a reasonable $T \in [1.1, 1.9]$.

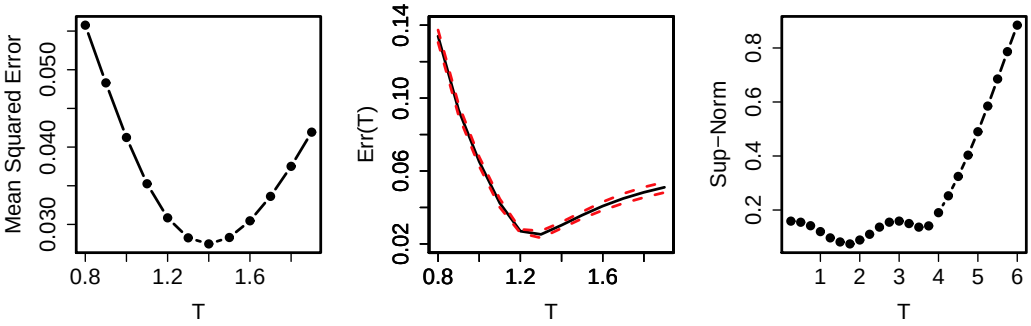

Figure 1: The choice of smoothing parameter T for the estimator $\hat{f}_\Gamma^{B,T}$ depends on which objective function is minimized. The results are averaged over 10 simulations with $N = 2000$ observations. The standard errors for the mean squared error and the sup-norm are too small to be seen. The corresponding plots of $\hat{f}_\Gamma^{B,T}$ for each T are in Appendix C.

5.3 Comparison of Estimator Performance

In this section we compare the performance of the two regularized inverse estimators (6) and (7) based on A_T and B_T. We look at the distribution of error when all the parameters are chosen to minimize a mean squared error criterion. The parameter values are recorded in Appendix C along with some indicative plots of the estimators. To measure the error in an estimator of h we use the average L^2 error (ALE)

$$ALE(h) = \left\{ \int [\hat{h}(\gamma) - h(\gamma)]^2 f_\Gamma(\gamma) d\gamma \right\}^{1/2}$$

which allows us to focus on how well the estimator performs at values of the unobservable Γ in the peak of the distribution [5]. As $CATE$ involves conditioning on f_Γ it is very hard to estimate this at points where the density of f_Γ is low.

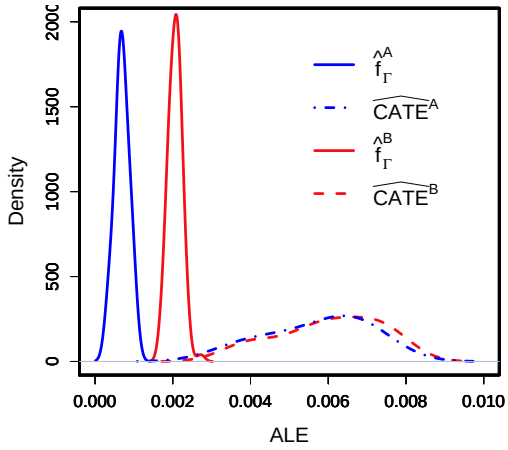

Figure 2: The distribution of the average L^2 error is calculated from 100 independent simulations where f_Γ is estimated with $N = 2000$ points and $CATE$ is estimated on $N = 10000$. The estimation of $CATE$ and corresponding calculation of ALE are done over the region where the mass of $\hat{f}_\Gamma$ is concentrated to avoid small values of $\hat{f}_\Gamma$ in the denominator of $\widehat{CATE}$.

In Figure 2 the distribution of error is shown for our two estimation strategies based on the regularized Radon inverses A_T and B_T respectively. The two estimators $\hat{f}_\Gamma^A$ and $\hat{f}_\Gamma^B$ show a clear difference in their performance however there is no difference in the estimation of $CATE$. The main reason A_T is comparatively more successful on f_Γ than $CATE$ is because of how well the first stage plug-in can be estimated.

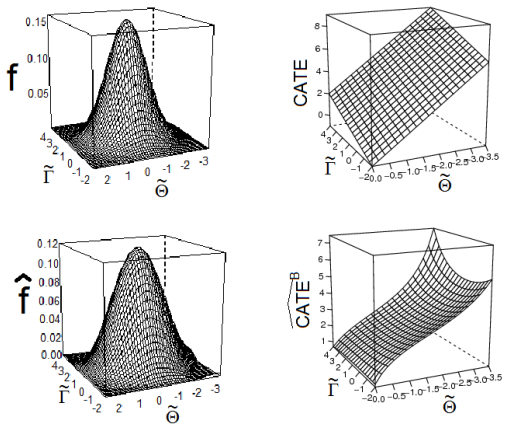

Figure 3: The true f_Γ and the estimator $\hat{f}_\Gamma^A$ are very similar. For $\hat{f}_\Gamma^A$ we used $h = 0.7$ for the first stage regularization and $T = 2$ for the regularized inverse A_T. The estimator $\widehat{CATE}^B$ has the appropriate smoothness and location in space but, unlike $\widehat{CATE}^A$, it struggles to pick up the variation in Γ. We use $T = 2.25$ for the numerator and $T = 1.4$ for the denominator. (See Appendix C)

Due to the computational time required to integrate estimators involving A_T and the difficulty in estimating the first-stage plug-in $\partial_v \mathbb{E}[g(Y, D)|(\Phi, V) = (\phi, u)]$ we recommend the use of the estimator (7) based on the regularized Radon inverse B_T.

However this estimator does not perform as well at estimating f_Γ and struggles with the dependence of the returns Δ on both dimensions of unobserved heterogeneity.

6 Outlook

This project estimates the average returns to college education conditional on unobservables. As the simulation study demonstrates, our estimators are able to recover information about the structure of unobservables that are intrinsic to the decision process and heterogeneous returns. To recover the unobservables we introduced an ill-posed inverse problem through the Radon transform. We constructed a new regularized inverse that allowed us to simplify the estimation strategy into a one-stage process involving a finite sum. To use the estimators in an application we presented a method for selecting the smoothing parameter T which worked well in the simulation study.

Future work will look at the application of these estimators to real data.

7 Bibliography

[1] E. Gautier and S. Hoderlein, *Estimating Treatment Effects with Random Coefficients in the Selection Equation*, Preprint http://arxiv.org/abs/1109.0362 Submitted Sep (2011)

[2] J. Heckman and E. Vytlacil, *Structural Equations, Treatment Effects and Econometric Policy Evaluation*, Econometrica, Vol. 73, (2005).

[3] P. Carneiro, J. Heckman and E. Vytlacil, *Estimating Marginal Returns to Education*, American Economic Review, Vol. 101, (2011).

[4] T. Klein, *Heterogenous Treatment Effects: Instrumental Variables without Monotonicity?*, Journal of Econometrics, Vol. 155, (2010).

[5] S. Hoderlein, J. Klemela, and E. Mammen, *Analyzing the Random Coefficient Model Nonparametrically*, Econometric Theory, Vol. 26, (2010).

[6] G. Imbens and J. Angrist, *Identification and Estimation of Local Average Treatment Effects*, Econometrica, Vol. 62, (1994).

[7] R. Charnigo, B. Hall and C. Scrinivasan, *A Generalized C_p Criterion for Derivative Estimation*, Technometrics, Vol. 53, (2011).

[8] C. Belzil, *The Return to Schooling in Structural Dynamic Models: a Survey*, European Economic Review, Vol. 51, (2007).

[9] C. Loader, Local Regression and Likelihood, Springer, (1999).

[10] J. Heckman, *Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy*, Journal of Economic Literature, Vol. 48, (2010).

[11] J. Heckman, J. Humphries, S. Urzua and G. Veramendi, *The Effects of Education Choices on Labor Market, Health and Social Outcomes*, Human Capital and Economic Opportunity Working Group, Working Paper No. 2011-002, (2011).

[12] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, (2011).

[13] C. Loader, *locfit: Local Regression, Likelihood and Density Estimation*, 2010.

[14] S. Johnson and B. Narasimhan, *cubature: Adaptive multivariate integration*, (2011).

[15] A. Bouvier, T. Hahn and K. Kieu, *R2Cuba: Multidimensional Numerical Integration*, (2010).

[16] J. Knaus, *snowfall: Easier cluster computing (based on snow)*, (2010).

[17] J. Heckman, L. Lochner and P. Todd, *Fifty Years of Mincer Earnings Regressions*, NBER Working Paper No. 9732, (2003).

A The Radon Transform

The application of the Radon transform to a function $f \in L_1(\mathbb{R}^d)$ yields the integral of f over the hyperplane $L_{s,u} := \{\gamma : \gamma^\top s = u\}$ with respect to $d_{L_{s,u}}(\gamma)$ which is the Lebesgue measure on the hyperplane,

$$R[f](s,u) = \int_{L_{s,u}} f(\gamma) d_{L_{s,u}}(\gamma).$$

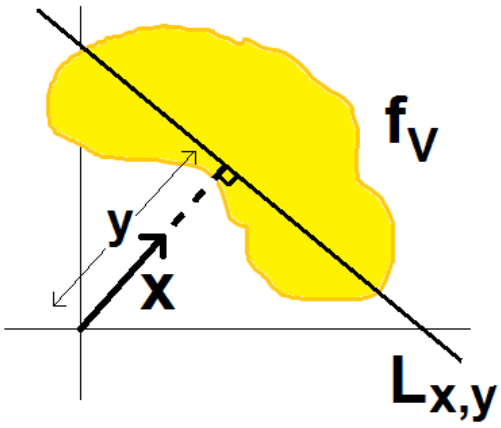

Figure 4: In $\mathbb{R}^2$ the Radon transform is the slice integral over the line $L_{x,y} = \{v : x^\top v = y\}$ indexed by a vector on the unit sphere $x = e_\phi$ and a scalar y.

We use a large support assumption **(A=4)** to ensure that the support of f_Γ is in a region that can be recovered by our instruments. Namely for any point γ in the support of f_Γ there exists some (ϕ, v) in the range of our instruments such that γ is on the line $L_{\phi,v}$ or equivalently $\gamma^\top e_\phi = v$.

A.1 Decision Model for Different Dimensions of Heterogeneity

The previous literature that accounted for the impact of unobservables on the treatment decision D relied on an additive scalar unobservable Θ [2][6]. In particular the decision model was in the form of a scalar threshold crossing model

$$D_1 = \mathbf{1}\{\mu(V,Z) > \Theta\} = \mathbf{1}\{F_\Theta(\mu(V,Z)) > \mathcal{U}\} = \mathbf{1}\{P(V,Z) < \mathcal{U}\} \tag{10}$$

for some function μ of the observables V, Z. Using the monotonic transformation of the cdf F_Θ yields the final expression in (10) which is in terms of a uniform random variable $\mathcal{U}$ on the interval$[0,1]$. A decision model in this form is equivalent to assuming that decisions are monotonic [10]. The advantage of this set up is the ease of estimation;

$$\mathbb{P}(D_1 = 1|(V,Z) = (v,z)) = \mathbb{P}(\mathcal{U} < P(V,Z)|(V,Z) = (v,z)) = P(v,z)$$

As a contract our decision model involves multiple sources of unobserved heterogeneity and requires the Radon transform to estimate $\mathbb{P}(D_d = 1|(V,Z) = (v,z))$; in d dimensions we have the model

$$D_d = \mathbf{1}\{\tilde{V} - \tilde{\Gamma}^\top \tilde{Z} - \tilde{\Theta} > 0\} \tag{11}$$

for $\tilde{\Gamma}, \tilde{Z} \in \mathbb{R}^d$. Rescaling by the norm of $(\tilde{Z}^\top, 1)$ gives $D_d = \mathbf{1}\{V > \Gamma^\top S\}$ where S is the unit vector on the upper hemisphere and $\Gamma \in \mathbb{R}^{d+1}$.

$$\begin{aligned}
\mathbb{P}(D_d = 1|(V,S) = (v,s)) &= \mathbb{E}[\mathbb{E}[D_d|V = v, S = s, \Gamma = \gamma]|V = v, S = s] \\
&= \mathbb{E}[\mathbb{E}[D_d|\Gamma = \gamma]|V = v, S = s], \qquad \text{using (A-2)} \\
&= \int \mathbf{1}\{v > \gamma^\top s\} f_\Gamma(\gamma) d\gamma \\
&= \int_{-\infty}^{v} \int_{\{\gamma:\gamma^\top s = u\}} f_\Gamma(\gamma) d_{L_{s,u}}(\gamma) du \\
&= \int_{-\infty}^{v} R[f_\Gamma(\cdot)](s,u) du
\end{aligned}$$

Using the Radon transform allows us to recover the distribution of unobservables f_Γ, namely

$$f_\Gamma(\gamma) = R^{-1}\left[\overline{\partial_v \mathbb{E}[D_d|(V,S) = \cdot]}\right](\gamma)$$

where $\bar{h}$ denotes the extension of the function h as 0 outside of its domain of definition. This is useful if the regression function $\partial_v \mathbb{E}[D_d|(V,S) = \cdot]$ has regressors with bounded support. It is an innocuous assumption as we assume the variables (V,S) have a large enough support to apprehend the whole distribution of unobservables f_Γ.

A.2 Central Result for Identification

Theorem 1 *Consider an arbitrary function ϕ such that $\mathbb{E}\left[|\phi(Y_0)| + |\phi(Y_1)|\right] < \infty$. Let $L \geq 2$, and assume that Assumption 1 holds. Then, for almost every x in $supp(X)$, the following statements are true:*

$$f_{\widetilde{\Gamma}|X}(\cdot|x) = R^{-1}\left[\overline{\partial_v \mathbb{E}\left[D \middle| \left(\widetilde{S}, \widetilde{V}\right) = \cdot, X = x\right]}\right] \tag{12}$$

$$\overline{\mathbb{E}\left[\phi(Y_1)\middle|\widetilde{\Gamma} = \cdot, X = x\right]} f_{\widetilde{\Gamma}|X}(\cdot|x) = R^{-1}\left[\overline{\partial_v \mathbb{E}\left[\phi(Y) D \middle| \left(\widetilde{S}, \widetilde{V}\right) = \cdot, X = x\right]}\right] \tag{13}$$

$$\overline{\mathbb{E}\left[\phi(Y_0)\middle|\widetilde{\Gamma} = \cdot, X = x\right]} f_{\widetilde{\Gamma}|X}(\cdot|x) = R^{-1}\left[\overline{\partial_v \mathbb{E}\left[\phi(Y)(D-1) \middle| \left(\widetilde{S}, \widetilde{V}\right) = \cdot, X = x\right]}\right]. \tag{14}$$

From the theorem it is simple to calculate the relationships for $CATE$ using the fact that $\Delta = Y_1 - Y_0$ and $Y = YD + Y(1-D)$. Namely

$$\begin{aligned}
\mathbb{E}[\Delta|\Gamma = \cdot] f_\Gamma(\cdot) &= \mathbb{E}[Y_1|\Gamma = \cdot] f_\Gamma(\cdot) - \mathbb{E}[Y_0|\Gamma = \cdot] f_\Gamma(\cdot) \\
&= R^{-1}\left[\overline{\partial_v \mathbb{E}\left[YD \middle| \left(\widetilde{S}, \widetilde{V}\right) = \cdot\right]}\right] - R^{-1}\left[\overline{\partial_v \mathbb{E}\left[Y(D-1) \middle| \left(\widetilde{S}, \widetilde{V}\right) = \cdot\right]}\right] \\
&= R^{-1}\left[\overline{\partial_v \mathbb{E}\left[Y \middle| \left(\widetilde{S}, \widetilde{V}\right) = \cdot\right]}\right]
\end{aligned}$$

A.3 Calculating $R[f_\Gamma(\cdot)](\phi, v)$ for the Simulation

For the two stage estimator (6) involving the regularized Radon inverse A_T it is difficult to tune the first stage plug-in estimator $\partial_v \mathbb{E}[g(Y,D)|(\Phi, V) = (\phi, u)]$. In

the simulation we use an oracle method of comparing the estimate against the truth where

$$\partial_v \mathbb{E}[D|(\Phi, V) = (\phi, u)] = R[f_\Gamma(\cdot)](\phi, u)$$
$$\partial_v \mathbb{E}[Y|(\Phi, V) = (\phi, u)] = R[\Delta f_\Gamma(\cdot)](\phi, u)$$

As the simulation set up was chosen to have the same variance it is straightforward to calculate these expressions using a parameterization of $L_{\phi,u}$ such that

$$\forall \gamma \in L_{\phi,u}, \exists t \in \mathbb{R}, \quad \gamma = u \begin{bmatrix} \cos(\phi) \\ \sin(\phi) \end{bmatrix} + t \begin{bmatrix} -\sin(\phi) \\ \cos(\phi) \end{bmatrix}$$

and then doing the Radon integral with respect to the parameter t.

As $f_\Gamma(\gamma, \theta) = \frac{1}{2\pi} e^{\frac{-1}{2}[(\gamma-1)^2+(\theta+0.5)^2]}$ we get

$$R[f_\Gamma(\cdot)](\phi, v) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}[v-(\cos(\phi)-0.5\sin(\phi))]^2}$$

Using these calculations we could tune the plug-in by comparing against the truth. This method is obviously not feasible for an application so we refer to it as an oracle.

B Alternative Regularized Radon Inverse

The inverse Radon transform is an ill-posed inverse problem; introducing a small amount of error to the argument may result in large changes in the value. To deal with this we need to construct a regularized inverse. To express our estimator in an alternative form that is more suitable for an application we want to be able to perform integration by parts on the regularized inverse. This is not possible with the classical regularized inverse involving K_T as defined in section 3.

B.1 Define the Regularized Inverse using the Schwartz Space

The classical regularized inverse is defined in terms of the inverse Fourier transform of an indicator function, we will use the inverse Fourier transform of an element of the Schwartz space and use the properties of the Schwartz space to guarantee that we can do integration by parts.

Definition 2 *The **Schwartz space** $\mathcal{S}(\mathbb{R}^d)$ is a space of functions where all their derivatives are rapidly decreasing*

$$\mathcal{S}(\mathbb{R}^d) := \{f \in \mathcal{C}^\infty(\mathbb{R}^d) : \forall \alpha, \beta \in \mathbb{N}^d \quad |x|^\alpha |\partial^\beta f(x)| \longrightarrow_{|x|\to\infty} 0\}$$

where α, β are multi-indices.

The Schwatz space has some important properties that are useful for our problem; first $\mathcal{S}(\mathbb{R}^d)$ is contained in $L^p(\mathbb{R}^d)$ for all $1 \leq p \leq \infty$ and secondly the Fourier transform $\mathcal{F}$ is an isomorphism on $\mathcal{S}(\mathbb{R}^d)$.

Example 1 $\psi(x) = \exp\left(\frac{-1}{1-|x|^2}\right)$ *is an element of $\mathcal{S}(\mathbb{R}^d)$ with support on $B_d(0,1)$*

For the regularized inverse we introduce the smoothing parameter T and observe that $\psi\left(\frac{x}{T}\right)$ is defined for $x \in B_d(0,T)$. By decreasing the value of T we truncate the Fourier transform which will increases the smoothing in the regularized inverse.

We will use the normalized Fourier inverse of ψ

$$\phi(x) = \frac{c}{(2\pi)^d} \int_{\xi \in B_d(0,1)} e^{-ix^\top \xi} \psi(\xi) d\xi$$

where c is chosen so that $\int_{\mathbb{R}^d} \phi(x)dx = 1$. By definition of the Schwatz space ϕ is integrable and so c exists. In particular we will define $\phi_T(x)$ for $x \in B_d(0,T)$ and show that $\phi_T(x) = T^d \phi(Tx)$.

$$\begin{aligned}
\phi_T(x) &= \frac{1}{(2\pi)^d} \int_{\xi \in B_d(0,T)} e^{-ix^\top \xi} \psi\left(\frac{\xi}{T}\right) d\xi \\
&= \frac{T^d}{(2\pi)^d} \int_{\eta \in B_d(0,1)} e^{-ix^\top \eta T} \psi(\eta)\, d\eta \\
&= T^d \phi(Tx)
\end{aligned}$$

Definition 3 *Let $A(\gamma) = R^{-1}[g](\gamma)$. We define the* ***regularized Radon inverse*** *A_T for a general function g through a convolution with ϕ_T*

$$A_T(\gamma) = (A * \phi_T)(\gamma)$$

where ϕ_T is defined above through the inverse Fourier transform of an element of the Schwartz space.

B.2 Bias in the Regularized Inverse

The bias from using the regularized inverse A_T as a replacement for the inverse Radon transform A is

$$\|A_T[g] - A[g]\|_\infty .$$

To calculate the bias we will work with $(A - A_T)(x) = \int_{\mathbb{R}^d}(A(x) - A(x-y))\phi_T(y)dy$ which comes from the fact c was chosen earlier to guarantee that $\int \phi(y)dy = 1 = \int \phi_T(y)dy$.

We need a smoothness assumption on A to be able to provide a bound on the bias. For the explanation we only assume the existence of one derivative but the result can be generalized for s derivatives.

Assumption 2 *The operator A has derivative $\mathcal{D}A \in L^\infty$ and $\|\mathcal{D}A\|_\infty \leq M$ for some $M < \infty$.*

Using assumption 2 we provide a bound for the absolute value of the difference between A and A_T which gives us the bound in sup-norm.

$$\begin{aligned}
|(A - A_T)(\gamma)| &\leq T^d M \int_{\mathbb{R}^d} |y||\phi(Ty)|dy \\
&\leq M \int_{\mathbb{R}^d} \left|\frac{\eta}{T}\right| |\phi(\eta)|d\eta \\
&= T^{-1} M \int_{\mathbb{R}^d} |\eta||\phi(\eta)|d\eta
\end{aligned}$$

As $\phi \in \mathcal{S}$ then $|\eta||\phi(\eta)| \to 0$ as $|\eta| \to \infty$ and the integral is finite. The fact that the bias depends on T^{-1} is due to the assumption that the first derivative is bounded. This can be generalized to having a bounded s^{th} derivative and a bound on the bias term that depends on T^{-s}.

B.3 Defining J_T and B_T

Analogous to the presentation of the classical regularized Radon inverse we rewrite A_T in terms of an integral where K_T is replaced with our new J_T. For our work in a two dimensional setting where $\partial_v E : [0, \pi] \times \mathbb{R} \to \mathbb{R}$ we have

$$A_T[\partial_v E](\gamma) = \int_{-\infty}^{\infty} \int_0^\pi J_T(\gamma^\top e_\phi - u)\partial_v E(\phi, u)dud\phi, \qquad \gamma \in \mathbb{R}^2$$

where

$$J_T(u) = \frac{2c}{(2\pi)^2} \int_0^T \cos(tu)t^{d-1}\psi\left(\frac{t}{T}\right)dt = \frac{c}{(2\pi)^2} \int_{-T}^T e^{itu}|t|^{d-1}\psi(\frac{|t|}{T})dt.$$

As $|t|\psi(|t|) \in \mathcal{S}(\mathbb{R})$ then the Fourier transform is also an element of the Schwatz class which implies that $J_T \in \mathcal{S}(\mathbb{R})$ and hence $J_T \in L^p(\mathbb{R})$ for all $1 \leq p \leq \infty$. This allows us to use integration by parts and write

$$A_T[\partial_v E](\gamma) = \int_{-\infty}^{\infty} \int_0^{\pi} \tilde{J}_T(\gamma^\top e_\phi - u) E(\phi, u) du d\phi, \qquad \gamma \in \mathbb{R}^2$$

where

$$\tilde{J}_T(u) = J_T'(u) = \frac{-c}{2\pi^2} \int_0^T \sin(tu) t^L \psi\left(\frac{t}{T}\right) dt.$$

C Plots of Estimates

C.1 Estimates of f_Γ

In this section we show a comparison between the true distribution of unobservables and various estimators. Recall that

$$\hat{f}_\Gamma^A(\gamma) = \int_0^\pi \int_{-\infty}^\infty K_T(e_\phi^\top \gamma - u)\overline{\partial_v \widehat{\mathbb{E}[D|(\Phi, V) = (\phi, u)]}} du d\phi$$

which requires a regularization parameter for the plug-in estimator $\overline{\partial_v \widehat{\mathbb{E}[D|(\Phi, V) = (\phi, u)]}}$ and a suitable choice for T. Due to the computation burden of running our data-driven method with this estimator we used the mean squared error criterion in the simulation to tune both parameters. When $N = 2000$ we use $h = 0.7$ as the regularization parameter associated with $\overline{\partial_v \widehat{\mathbb{E}[D|(\Phi, V) = (\phi, u)]}}$ and $T = 2$ for the regularized Radon inverse A_T.

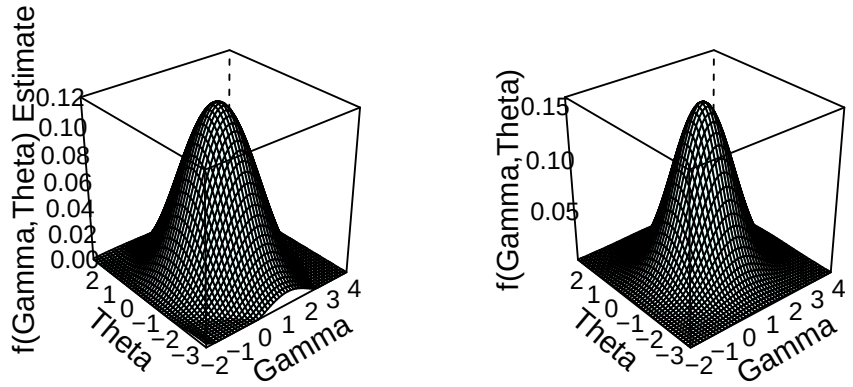

Figure 5: The estimator $\hat{f}_\Gamma^A$ (left) is a good match for the underlying distribution f_Γ (right). Estimation is done with a sample $N = 2000$ and the regularization parameters $(h, T) = (0.7, 2)$ based on a mean squared error criterion.

The alternative method of estimating f_Γ involves a sum which is computationally easier and allows us to use a data-driven approach for selecting T. In particular

$$\hat{f}_\Gamma^B(\gamma) = \frac{1}{N}\sum_{i=1}^N \frac{\tilde{J}_T(e_{\phi_i}^\top \gamma - v_i) d_i}{\hat{f}_{\Phi,V}(\phi_i, v_i)} \mathbf{1}\{\hat{f}_{\Phi,V}(\phi_i, v_i) > \tau\}$$

which is a one-stage estimator were the only smoothness parameter is T. In addition the plug-in $\hat{f}_{\Phi,V}$ is estimated with local polynomials and τ is chosen to trim 2% to avoid division by values too close to zero.

The choice of T depends on the criterion used; for the mean squared error $T = 1.4$, with the data-driven method $T = 1.2$ and for the sup-norm $T = 1.75$. The corresponding plots for each value of T are displayed in Figure 6.

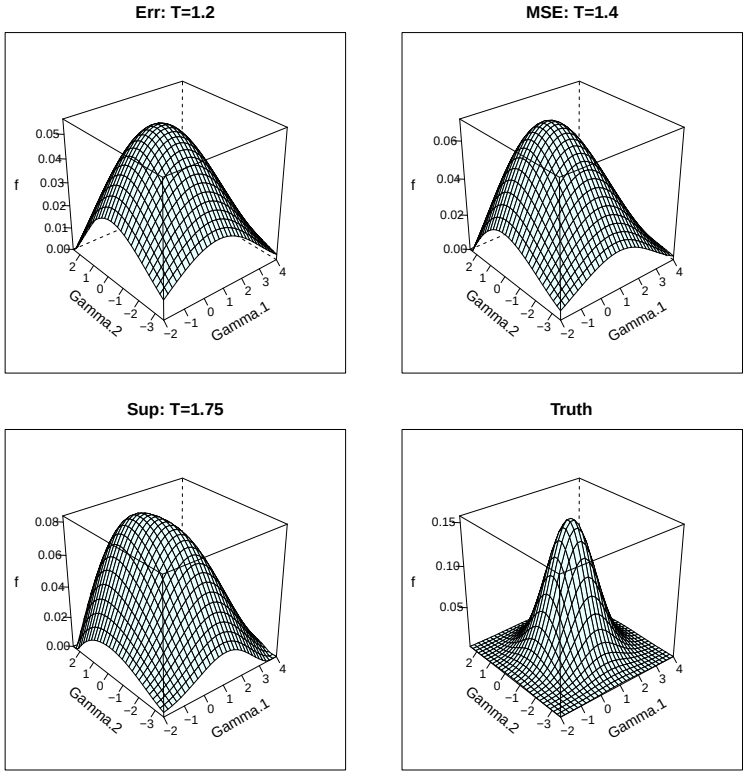

Figure 6: We consider three objective functions that can be minimized to select a suitable value of the parameter T. In the estimation with B_T the mean squared error (MSE), sup-norm (Sup) and our data-driven objective function Err all produce different values for T although they lie within a similar region and produce relatively similar estimates. For comparison the truth in the bottom right is substantially higher and more peaked, however the estimator does pick up on the correct location of the peak and is done with a comparatively small sample size of 2000.

C.2 Comparison with $\widehat{CATE}$

The estimates of $CATE$ are done with a sample size of 10000 in contrast to the 2000 that was used for the estimation of f_Γ in the report. For nonparametric estimation methods a sample size of 10000 is not considered large, however it can be difficult to find data sets of that size.

The plots are displayed over a central region where the majority of the mass of $\hat{f}_\Gamma$ lies as this guarantees a reasonable behaviour of the denominator in $CATE$. Even so on the boundaries of the central region the estimate starts to blow up.

Recall that in the simulation study we define

$$CATE(\Gamma) = 0.5\tilde{\Gamma} - 2\tilde{\Theta}.$$

The estimator $\widehat{CATE}^A$ is less reliable on the boundaries than $\widehat{CATE}^B$. The advantage of the $\widehat{CATE}^A$ is that it is not constant in either dimension of the unobservables indicating that there is a complex structure of unobservables involved in the outcome. By contrast $\widehat{CATE}^B$ is constant in the direction of $\tilde{\Gamma}$.

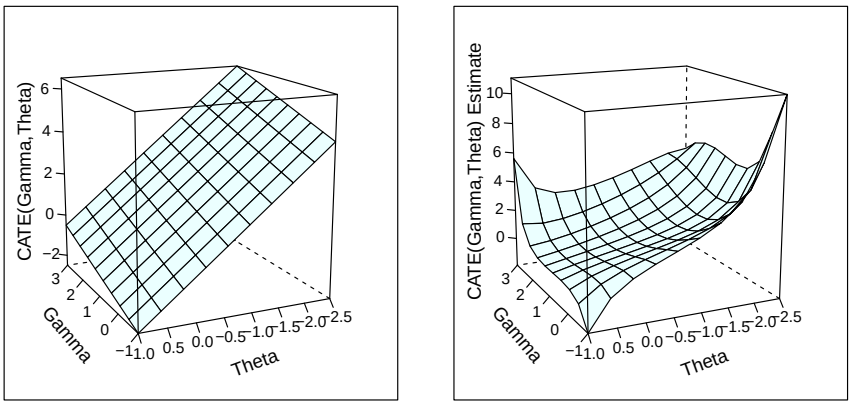

Figure 7: The true $CATE$ as a function of unobservables $(\tilde{\Gamma}, \tilde{\Theta})$ and an estimate $\widehat{CATE}^A$ where the estimation parameters for the numerator are $T = 2.25$ for the regularized Radon inversion and $h = 1.1$ for the local polynomial estimate of a derivative of a regression function and in the denominator $T = 2$ and $h = 0.7$. Decreasing T corresponds to more smoothing in the regularized inverse while increasing h introduces more smoothing in the local polynomial estimate. All parameters were chosen to minimize a MSE criterion.

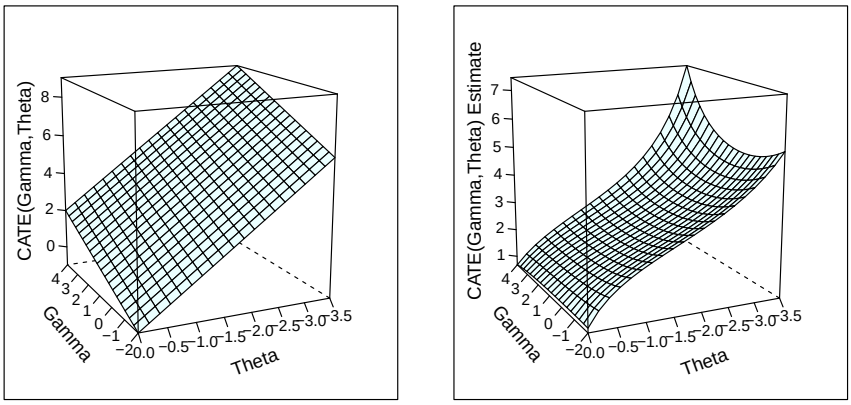

Figure 8: The true $CATE$ as a function of unobservables $(\tilde{\Gamma}, \tilde{\Theta})$ and an estimate using $\widehat{CATE}^B$ where the estimation parameters for the numerator are $T = 2.5$ with trimming set at 2% and in the denominator $T = 1.4$ again with 2% trimming. Decreasing T corresponds to more smoothing in the regularized inverse while increasing h introduces more smoothing in the local polynomial estimate. All parameters were chosen to minimize a MSE criterion.