Carnegie Mellon
Tepper
SCHOOL OF BUSINESS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
ECONOMICS

Titled
"MACHINE LEARNING TECHNIQUES IN APPLIED ECONOMETRICS"

Presented by
Nandana Sengupta

Accepted by

Fallaw Sowell **4/20/15**

Chair: Prof. Fallaw Sowell **Date**

Approved by The Dean

Robert M. Dammon **4/23/15**

Dean Robert M. Dammon **Date**

Machine Learning Techniques in Applied Econometrics

by

Nandana Sengupta

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

at

Carnegie Mellon University

David A. Tepper School of Business

Pittsburgh, Pennsylvania

Dissertation Committee:

Fallaw Sowell (Chair)

Dennis Epple

Bryan Routledge

Daniel Neill

Joachim Groeger

Spring 2015

Contents

Acknowledgments

The guiding forces behind my interest in mathematics and economics (and thus the guiding forces behind my PhD) are my parents, Arya and Suchitra Sengupta. My mother and father have always encouraged me to question things, they've encouraged me to try and understand society and they've always had a soft spot for mathematics and statistics. It really is no wonder that I've turned out to be an economist! A very special mention to my big sister Ananya Sengupta – who has been a combination of parent and friend to me and the person that invariably manages to put things in perspective for me.

I consider myself extremely lucky to have had amazing mentors throughout my academic journey. First and foremost I am most grateful to Professor Fallaw Sowell – I could not have asked for a better adviser. Professor Sowell has inspired me throughout my time at Tepper with his immense talent, academic integrity and ability to maintain balance between professional and personal lives. I am sure that his example will continue to inspire me in the future to be a good person as well as a sincere professional. I'm also thankful for the guidance of Professor Ravi whose dynamism is reflected in his ability to take on diverse projects with so much ease. I am very grateful to Professor Daniel Neill, Professor Bryan Routledge, Professor Joachim Groeger and Professor Dennis Epple for being available when I've needed guidance and for their honest and constructive feedback on my research.

No list of acknowledgements at Tepper would be complete without the mention of Lawrence Rapp who manages PhD student services so effortlessly. Lawrence has made our time at

Tepper one of great comfort – both by his efficiency and his ability to make each one of us find at Tepper a home away from home. I am also greatly indebted to my advisers in India – Professor Subrata Sarkar, Professor Jayati Sarkar, Professor Naveen Srinivasan and Professor Shubhro Sarkar – they encouraged me to follow my dreams when I was starting out and needed encouragement the most.

Thanks to my husband, Varun Gulshan for being my best friend and my support system. And thanks to Kamala Bopana – my first and original best friend! Both of you have taught me by example to be the best I can while keeping a smile on my face. Thanks also to my beer buddy, Shonali – it seems impossible that I would have survived the five long years without our regular catch-up sessions (and not all the credit goes to the beers!). Sudatta, my childhood buddy, thanks for being a delightful roommate, a shoulder to cry on and a vent for all my pent-up craziness! Camilo and Minyoung, my officemates from the 'Cool Office' – you guys are awesome! I'm sure you'll do great things and more importantly happy/fun things all your lives – thanks for your friendship. Finally a mention to some of my closest friends who have always been just a phone call away – Ruchika Mohanty, Pallavi Aron, Abhimanyu Sahai and Meenu Iyer. Having you guys in my life is an insurance against sadness and a reminder that a good laugh is never very far.

Preface

The main focus in econometrics is to provide an *explanation* of various observed outcomes. Structural econometricians obtain reliable estimates of parameters that describe an economic system to provide an understanding of the underlying processes that determine equilibrium outcomes. The estimation process is based on conditions implied by economic theory.

On the other hand, the main focus in machine learning is to provide accurate *predictions* of the variables of interest. While these techniques are extremely powerful for forecasting, it can be very hard to interpret the underlying structure implied by them.

As machine learning techniques become more popular and computers become capable of storing and processing large quantities of data, there have been some recent efforts to incorporate such techniques into structural econometric models. My research aims to extend this literature.

In my thesis I investigate whether it is possible to incorporate machine learning techniques in econometric models in a meaningful way. I explore two different approaches for doing so – first, I generalize the idea of regularization from machine learning to the Generalized Method of Moments framework; second, I apply pre-existing classification techniques from machine learning to the Propensity Score framework. Finally, I employ the empirical techniques developed in the second chapter to address a public policy question – that of the effectiveness of India's Safe Motherhood Scheme.

Chapter 1: Regularization Paths in Generalized Method of Moments

In the GMM framework, the objective function to be minimized is a weighted sum of squares of m moment conditions implied by economic theory. The derivative of the objective function with respect to the vector of parameters (θ) provides a system of k equations in k unknowns that is used to obtain parameter estimates. However if this matrix is nearly singular at the true parameter values, then the system of equations becomes highly unstable. This results in high standard errors of the parameter estimates. This is analogous to the problem of multicollinearity in linear regression. In the linear regression framework the problem is somewhat overcome by regularization. Ridge and spectral cut-off regularization are commonly used. However, due to the highly non-linear nature of the GMM objective function, ridge and spectral cut-off are not readily generalizable to the GMM framework.

In the first chapter (co-authored with Fallaw Sowell), we re-interpret regularization as a set of possible solutions that lie along a *path* between the unconstrained minimum of the objective function and the mean of a pre-defined prior. Using this interpretation, we propose algorithms for finding the 'regularized' parameter estimates. We use a holdout sample in GMM for parameter selection. We also show via simulations that our method performs very well when the system of equations is unstable. We discuss how to extend the techniques in higher dimensions and point to relevant algorithms in the Computer Science literature. As an empirical application we employ this method on the Consumption based Capital Asset Pricing Model.

Chapter 2: Propensity Score Model Selection using Machine Learning Classifiers

The basic issue in estimating the effect of a particular treatment using observational data is that the data suffers from *selection bias*. In other words those who receive treatment (the treatment group) are inherently different from those who don't (the control group).

Heckman (in his seminal 1978 paper) shows that a naive estimate of the regression parameter on a treatment dummy (say $W = 1$ if an individual is treated and $W = 0$ if the individual is a control) suffers from an *omitted variable bias*. The problem arises because we only observe outcomes under a single state (either treatment or control)– thus we have to control for factors which simultaneously affect both outcome and selection into the treatment group. Rubin and Rosenbaum (1983) pioneered the work on causal inference in the presence of selection bias. They suggest a two-step estimation procedure. In the first step the probability that an individual belongs to the treatment group is estimated. This is referred to as the individual's *Propensity Score*. The second step involves using the Propensity Score for pre-processing the data before estimating the Average Treatment Effect (ATE).

The use of Inverse Propensity Score Weighting (IPW) is now ubiquitous in the Causal Inference literature, however the estimation of propensity scores remains an open question. While many authors use logistic regression because of its interpretability, others argue in favor of non-parametric methods. We propose the use of machine learning classifiers (like Naive Bayes, Regression Trees and Support Vector Machines) for obtaining propensity scores. We also propose using a holdout sample to choose between different propensity score models. We show via theoretical arguments and simulation studies why its useful to consider a variety of propensity score models in the first step. We compare propensity scores estimates obtained from Linear Probit model as well as from semi-parametric classifiers like Naive Bayes, Random Forests and Support Vector Machines. We show via two sets of simulation studies why it's useful to choose from a variety of propensity score models. In particular we find that propensity score estimates with Minimum Covariate Imbalance perform very well in terms of Mean Squared Error of Average Treatment Effect estimates across all simulations.

Chapter 3: Evaluating India's Safe Motherhood Scheme using Inverse Propensity Score Weighting

Conditional Cash Transfers (CCT) programs are becoming an increasingly popular policy tool in developing countries to incentivize certain behavior such as school enrollment, vaccination and health check ups amongst a targetted section of the population. The beneficiaries of CCTs are typically from poorer communities and the final aim of such programs is to help such communities get out of poverty. India's Safe Motherhood scheme or Janani Suraksha Yojana, launched in 2005, incentivizes eligible women to give birth in health care facilities. With more than 9 million beneficiaries, it is the world's largest CCT program in terms of the number of beneficiaries. Eligibility criteria involves posession of Below-the-Poverty-Line (BPL) cards, belonging to a scheduled caste or tribe and order of birth. The financial assistance amounts range from Rs. 600 ($9.76) to Rs. 1400 ($22.78) depending on locality and focus. These incentives are communicated to the women locally by female health volunteers or Accredited Social Health Activists (ASHAs) who receive performance based compensation

We use estimation techniques developed in Chapter 2 (IPW using Minimum Covariate Imbalance criteria) to evaluate the effectiveness of the scheme. In particular we estimate the Average Treatment Effect (ATE) of receiving financial assistance via JSY on two health outcomes – number of stillbirths and infant mortality. We also estimate ATE on three behavioural outcomes – whether the mother had 3 or more ante-natal checkups, whether any post-natal check up was conducted within 2 weeks of delivery and the frequency of child check-ups within 10 days of delivery. We are not aware of any other paper that uses Propensity Score methods to evaluate JSY at the national level and therefore we consider this attempt as a substantial contribution towards the literature on the assessment of JSY. We are not aware of other studies which anayzed the effect of JSY on the frequency of child check-ups. Finally, our results indicate that in certain geographical regions propensity scores obtained via machine learning techniques were picked leading to results that are qualitatively different from those obtained by the standard linear probit.

Chapter 1

Regularization Paths in Generalized Method of Moments

1.1 Introduction

While most statistical and econometric estimation routines have traditionally focussed on obtaining *unbiased* parameter estimates with the lowest possible variance, over the last few decades the focus in statistics and machine learning has shifted to parameter estimates with the best *predictive* properties. In obtaining these estimates the estimation procedure is tweaked, often by augmenting the objective function with a penalty term resulting in *'regularized'* estimates. These estimates tend to have significantly lower variance which usually comes at the price of sacrificing the unbiasedness of estimates obtained by more traditional methods. This trade-off is popularly known as the *'bias-variance tradeoff'*.

Bickel and Li (2006) provide an excellent overview of the properties of various regularization procedures in statistics. They loosely define regularization as *"the class of methods needed to modify maximum likelihood to give reasonable answers in unstable situations"* where by unstable they refer to estimation problems containing an ill-posed inverse. These methods go beyond just maximum likelihood estimation. In linear regres-

sion especially methods like ridge, LASSO, elastic net and spectral cut-off regularization have become quite standard. Non-parametric density estimation methods are also based on the idea of regularization. Most recently the LASSO has been extended to the class of generalized linear models (GLM). For a review of methods currently in use refer to Hastie, Tibshirani and Friedman (2008).

Although regularization concepts have been extended to some non-linear parametric models the field of regularization in Generalized Method of Moments (GMM) in Econometrics is still largely open. Some of the notable exceptions being the work by Carrasco et all (2000, 2007, 2012), Caner (2009) and Liao (2010). The first set of papers (Carrasco et all) extend the m moment conditions to a continuum of moment conditions. The authors use ridge regularization in order to find the inverse of the optimal weighting *kernel* (instead of optimal weighting matrix in traditional GMM). Caner attaches a linear penalty term like in the LASSO framework to the usual GMM objective function and argues that this helps in variable selection by forcing variables whose co-efficients are not significant down to zero. Finally Liao augments the m moment conditions with another k moment conditions where the second set of augmented moment conditions is constructed from the subset of the original m moment conditions which may be misspecified. Using the new set of $m + k$ moment conditions and a LASSO-type penalty the author simultaneously performs estimation as well as moment selection.

This paper aims to take the current literature forward by using regularization to solve a GMM identification problem. In particular we deal with the case where the matrix of first derivatives comes close to losing rank – resulting in an *unstable system* and therefore an *identification* problem. The system is unstable in the sense that a minor change in the input variables can lead to disproportionately large changes in the parameter estimates. An example of such a situation is an exactly identified GMM estimation where some moments are *nearly* collinear i.e. some moments contain highly overlapping information. The analogous problem in linear regression is the problem of multicollinearity wherein

some covariates are nealy collinear resulting in parameter estimates which, though unbiased, are associated with very high standard errors.

We re-interpret regularization techniques in linear regression as a search for the 'best' parameter estimate over a unidimensional path in the parameter space between the global minimum and a prior. We define such a path as a *'Regularized Solution Path'*. The global minimum is associated with zero bias and high variance whereas the prior is associated with high bias and zero variance. The tradeoff between bias and variance is exploited to find the 'best' estimate where 'best' refers to the estimate which minimizes Mean Squared Error (MSE). Practically this is achieved by using a hold out sample, a procedure which is ubiquitous in machine learning algorithms. We randomly divide the data into testing and training sets, and then pick the point on the regularization path which minimizes GMM error on the testing set.

The rest of the paper is organized as follows: Section 2 sets up the theoretical GMM model and formally describes the identification problem we address. Section 3 revisits some commonly used regularization techniques in the linear regression framework and introduces the notion of *regularized solution paths*. Section 4 describes three procedures that generalize to GMM. Simulation results on a modified version of the Hall-Horowitz model as well as a linear GMM model are presented in Section 5. An application on the Consumption based Capital Asset Pricing Model in Section 6 is followed by a discussion on Possible Extensions in Section 7. Section 8 concludes.

1.2 The GMM Identification Problem

In this section we first set up the theoretical GMM Model and then describe the identification problem along with the conditions under which it occurs.

1.2.1 GMM Recap

Assumption 1: The data $x = \{x\}_{i=1}^{n}$ is assumed to be independently and identically distributed.

Parameters are represented by θ, which is a vector of dimension $(k \times 1)$. Economic theory implies m moment conditions $(m \geq k)$. For each of these m moment conditions at the true parameter value θ_0:

$$g_{i,j}(x, \theta) = g_{i,j}(\theta); \qquad E(g_{i,j}(\theta_0)) = 0; \qquad \forall j = 1, \cdots, m.$$

We represent the $(m \times 1)$ vector of moment conditions as $g(\theta)$ and its sample analogue (corresponding to the i^{th} data point) as $g_i(\theta)$.

The GMM objective function is:

$$\begin{aligned} Q_n(\theta) &= G_n(\theta)' W_n G_n(\theta) \\ \Rightarrow \hat{\theta} &= \underset{\theta}{\operatorname{argmin}} \left[G_n(\theta)' W_n G_n(\theta) \right] \end{aligned}$$

where W_n is a positive definite weighting matrix of dimension $m \times m$ and

$$G_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} g_i(\theta).$$

The **Central Limit Theorem** applicable here is $\sqrt{n} G_n(\theta_0) \sim^A N(0, \Sigma_g)$.

Assumption 2: The optimal choice of $W_n \to \Sigma_g^{-1}$, where;

$$\Sigma_g = E(g_i(\theta_0) g_i(\theta_0)'). \tag{1.1}$$

The first order condition obtained by taking the derivative of the objective $Q_n(\theta)$ with respect to the vector of parameters θ is

$$M_n(\hat{\theta})' W_n G_n(\hat{\theta}) = 0. \tag{1.2}$$

where $M_n(\theta) = \frac{\partial G_n(\theta)}{\partial \theta'}$ which is a matrix of dimension $(m \times k)$. In the remainder of the discussion we drop the n subscript from all the terms for notational convenience.

Define $\bar{M}(\hat{\theta}) = W^{1/2} M(\hat{\theta})$ and $\bar{G}(\hat{\theta}) = W^{1/2} G(\hat{\theta})$ respectively as standardized versions of $M(\hat{\theta})$ and $G(\hat{\theta})$. Then the first order condition can be rewritten as

$$\begin{aligned} M(\hat{\theta})' W G(\hat{\theta}) &= 0 \\ \Rightarrow M(\hat{\theta})' [W^{1/2}]' [W^{1/2}] G(\hat{\theta}) &= 0 \\ \Rightarrow [W^{1/2} M(\hat{\theta})]' [W^{1/2} G(\hat{\theta})] &= 0 \\ \Rightarrow \bar{M}(\hat{\theta})' \bar{G}(\hat{\theta}) &= 0 \end{aligned}$$

Written like this we note that $\bar{M}(\hat{\theta})$ controls the k linear combinations of the m standardized moment conditions $\bar{G}(\hat{\theta})$ that are being set to zero in order to estimate the k parameters.

1.2.2 The identification problem

This paper focuses on the identification problem caused by loss of rank in $\bar{M}(\hat{\theta})$. From the previous discussion this implies that if $\text{rank}(\bar{M}(\hat{\theta})) < k$ then we cannot identify the k parameters in θ exactly and independently.

Another way of looking at this is by considering the asymptotic distribution of the pa-

rameter estimates under all the assumptions of GMM:

$$\left(\hat{\theta} - \theta_0\right) \sim_a N\left(0, \left(M(\theta)'\Sigma_g^{-1}(\theta)M(\theta)\right)^{-1}\right)$$

Note that if $rank(\bar{M}(\hat{\theta})) < k$ then $rank\left(M(\theta)'\Sigma_g^{-1}(\theta)M(\theta)\right) < k$. This implies that the covariance matrix is singular and that variance is unbounded along some of the k dimensions. In other words all k parameters are not identified. This type of an identification problem can occur even if there are *finite* moment conditions (in contrast to the case considered by Carrasco et al who deal mainly with problems with infinitely many moments).

Consider the following stylized example with $m = 2$ moments and $k = 2$ parameters:

1. $$G(\theta_1, \theta_2) = \begin{pmatrix} g_1(\theta_1, \theta_2) \\ \\ g_2(\theta_1, \theta_2) \end{pmatrix}$$

   where $g_2(\theta_1, \theta_2) = \tau g_1(\theta_1, \theta_2) + \varepsilon$, $\varepsilon \perp (\theta_1, \theta_2)$ and τ is some constant.

2. $$M(\theta_1, \theta_2)' = \begin{pmatrix} \frac{\partial g_1(\theta_1,\theta_2)}{\partial \theta_1} & \frac{\partial g_2(\theta_1,\theta_2)}{\partial \theta_1} \\ \\ \frac{\partial g_1(\theta_1,\theta_2)}{\partial \theta_2} & \frac{\partial g_2(\theta_1,\theta_2)}{\partial \theta_2} \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1(\theta_1,\theta_2)}{\partial \theta_1} & \tau\frac{\partial g_1(\theta_1,\theta_2)}{\partial \theta_1} \\ \\ \frac{\partial g_1(\theta_1,\theta_2)}{\partial \theta_2} & \tau\frac{\partial g_1(\theta_1,\theta_2)}{\partial \theta_2} \end{pmatrix}$$

3. While $rank\left(G(\theta_1, \theta_2)\right) = 2 = m$, we note that since the columns of $M(\theta_1, \theta_2)'$ are linearly related therefore $rank\left(M(\theta_1, \theta_2)'\right) = 1 < k$, $\forall(\theta_1, \theta_2)$. This implies that the first order condition cannot be used to estimate both θ_1 and θ_2 independently.

1.2.3 The link with Multicolinearity in linear regression

Recall the problem of multicolinearity in linear regression where Y is the $n \times 1$ vector of the dependent variable and X is the $n \times k$ matrix of covariates. The model that is being

estimated is

$$Y = \beta X + \varepsilon, \quad \varepsilon \perp X, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

where $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$. The associated least squares objective function is

$$Q(\beta) = \frac{1}{n}(Y - X\beta)'(Y - X\beta)$$

The closed form solution for the parameter estimates (obtained by setting the derivative of $Q(\beta)$ wrt to β to zero) is:

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

The asymptotic distribution of the parameter estimates is

$$\left(\hat{\beta} - \beta\right) \sim_a N\left(0, \sigma_\varepsilon^2 (X'X)^{-1}\right)$$

Now, consider what happens if $rank(X) < k$ in small samples.

1 The matrix $X'X$ is nearly singular or $(X'X)^{-1}$ is an unstable inverse.

2 The closed form solution (which depends on $(X'X)^{-1}$) is also unstable.

3 The variance of the parameter estimates $\sigma_\varepsilon^2 (X'X)^{-1}$ blows up.

The method of moments formulation of linear regression has the following moment condition

$$E(X'\varepsilon) = 0 \quad \Rightarrow g(\beta) = X'\left(\varepsilon(\beta)\right) = X'(Y - X\beta)$$

where the X matrix controls the k linear combination of the n residual equations that are being set to zero in order to estimate the parameters. If $rank(X) < k$ then all k parameters in a linear regression cannot be identified simulatenously. Note that this is analogous to the GMM identification problem described in Section 2.2.

In such situations many statisticians and econometricians choose to lose some of the unbiasedness of the OLS estimates in exchange for gains in the stability of estimates. The common regularization procedures used to tackle the problem of multicolinearity are *Ridge Regularization*, *Spectral Cutoff Regularization* and *Iterative Landweber Regularization*. In the next section we discuss these three commonly used regularization procedures in detail in terms of the solution concepts and the geometry of the solutions. We then re-interpret regularization as a search for the 'best' parameter estimate over a unidimensional path in the parameter space between the global minimum and a prior, where 'best' refers to the estimate which minimizes Mean Squared Error (MSE).

1.3 Regularization Paths in Ordinary Least Squares

In this section we discuss some regularization techniques applied to linear regression in greater detail. We then discuss the interpretation of regularized estimates in terms of regularized solution paths which presents a natural generalization to GMM.

1.3.1 Geometry of Least Squares

Start with the usual least squares objective function

$$\begin{aligned} Q(\beta) &= \frac{1}{n}(Y - X\beta)'(Y - X\beta) \\ &= \frac{1}{n}\left[Y'Y - 2\beta'X'Y + \beta'X'X\beta\right] \\ \Rightarrow \frac{\partial Q(\beta)}{\partial \beta'} &= \frac{2}{n}\left[X'X\beta - X'Y\right] \\ \Rightarrow \frac{\partial^2 Q(\beta)}{\partial \beta \partial \beta'} &= \frac{2}{n}X'X. \end{aligned}$$

The curvature of the objective function is governed by the positive definite $(k \times k)$ covariance matrix of the regressors $\frac{X'X}{n}$ which is independent of parameter values and has the following eigen-decomposition

$$\frac{X'X}{n} = \begin{pmatrix} C_1 & C_2 & \cdots & C_k \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_k \end{pmatrix} \begin{pmatrix} C_1' \\ C_2' \\ \vdots \\ C_k' \end{pmatrix} = C\Lambda C'$$

where the eigenvectors $C_1, C_2 \cdots C_k$ form an orthonormal basis and correspond to the eigenvalues $\lambda_1, \lambda_2 \cdots \lambda_k$ arranged in descending order. λ_j represents how much variation in X is explained by the dimension spanned by eigenvector C_j. Also note that $C'C = CC' = I_k$.

The identification problem occurs when *some* of the eigenvectors (say $C_j, \forall j > r$) explain *very little* variation in X, as represented by the magnitude of the corresponding eigenvalues $\lambda_j < \epsilon, \forall j > r$ where $\epsilon > 0$ is a 'very small' positive real number. Geometrically, this implies that the objective function is very flat along these dimensions leading to multiple points along these dimensions at which the objective function value takes a value that is very close to its global minimum.

Mathematically,

$$\hat{\beta} = \left(\frac{X'X}{n}\right)^{-1} \frac{X'Y}{n}, \qquad Var(\hat{\beta}) \propto \left(\frac{X'X}{n}\right)^{-1}$$

where

$$\left(\frac{X'X}{n}\right)^{-1} = (C\Lambda C')^{-1} = C\Lambda^{-1}C'$$

which blows up if the eigenvalues $\lambda_j < \epsilon, \forall j > r$ where $\epsilon > 0$ is a 'very small' positive real number. This implies that $\hat{\beta}$ is unstable as its closed form solution and variance both depend on the unstable inverse $\left(\frac{X'X}{n}\right)^{-1}$.

Recall, however that the OLS solution is still *unbiased* and has the *minimum variance amongst all unbiased estimators*. This is a source of tension, especially if the main aim

of the estimation procedure is to obtain *stable* estimates. Ridge, spectral cutoff and iterative Landweber regularization techniques are often used in such contexts and work on the principal of introducing some bias in the estimation in order to reduce the disproportionately high associated variance.

1.3.2 Ridge Regularization

The Ridge regularization method augments the usual OLS objective with a well-defined parabola at a prior value (in most common cases the origin), weighted by a regularization parameter α as follows

$$\begin{aligned} Q(\beta) &= \frac{1}{n}(Y - X\beta)'(Y - X\beta) + \alpha\beta'\beta \\ &= \frac{1}{n}\left[Y'Y - 2\beta'X'Y + \beta'X'X\beta\right] + \alpha\beta'\beta \\ \Rightarrow \frac{\partial Q(\beta)}{\partial \beta'} &= \frac{2}{n}\left[X'X\beta - X'Y\right] + 2\alpha\beta \\ \Rightarrow \frac{\partial^2 Q(\beta)}{\partial\beta\partial\beta'} &= 2\left(\frac{X'X}{n} + \alpha I_k\right). \end{aligned}$$

Note that the curvature of this augmented objective function is defined by

$$\left(\frac{X'X}{n} + \alpha I_k\right) = C\Lambda C' + \alpha CC' = \begin{pmatrix} C_1 & C_2 & \cdots & C_k \end{pmatrix} \begin{pmatrix} \lambda_1 + \alpha & 0 & \cdots & 0 \\ 0 & \lambda_2 + \alpha & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_k + \alpha \end{pmatrix} \begin{pmatrix} C_1' \\ C_2' \\ \vdots \\ C_k' \end{pmatrix}$$

The addition of the scalar regularization parameter to each eigenvalue injects stability into the system since

$$\hat{\beta}_\alpha = \left(\frac{X'X}{n} + \alpha I_k\right)^{-1} \frac{X'Y}{n}, \qquad Var(\hat{\beta}) \propto \left(\frac{X'X}{n} + \alpha I_k\right)^{-1}$$

and

$$\left(\frac{X'X}{n} + \alpha I_k\right)^{-1} = \left(C[\Lambda + \alpha I_k]C'\right)^{-1} = C\left[\Lambda + \alpha I_k\right]^{-1} C'$$

which is stable for some value of the regularization parameter $\alpha \geq 0$.

Infact the ridge regression solution gives us a path between the *low bias high variance* OLS estimate (when $\alpha = 0$) to the *high bias low variance* prior (when $\alpha \approx \infty$). In other words *we can think of the solution to the ridge regularization regression as a path parametereized by* α.

Define the ridge solution given α as $\hat{\beta}_\alpha$:

$$\begin{aligned}
\hat{\beta}_\alpha &= \left(\frac{X'X}{n} + \alpha I_k\right)^{-1} \frac{X'Y}{n} \\
&= \left(C\Lambda C' + \alpha CC'\right)^{-1} \frac{X'Y}{n} \\
&= C\left(\Lambda + \alpha I\right)^{-1} C' \cdot \left[C\Lambda C' \cdot C\Lambda^{-1}C'\right] \frac{X'Y}{n} \\
&= C\left(\Lambda + \alpha I\right)^{-1} C' \cdot C\Lambda C' \left(\frac{X'X}{n}\right)^{-1} \frac{X'Y}{n} \\
&= C\left(\Lambda + \alpha I\right)^{-1} \Lambda C' \hat{\beta}_{ols}.
\end{aligned}$$

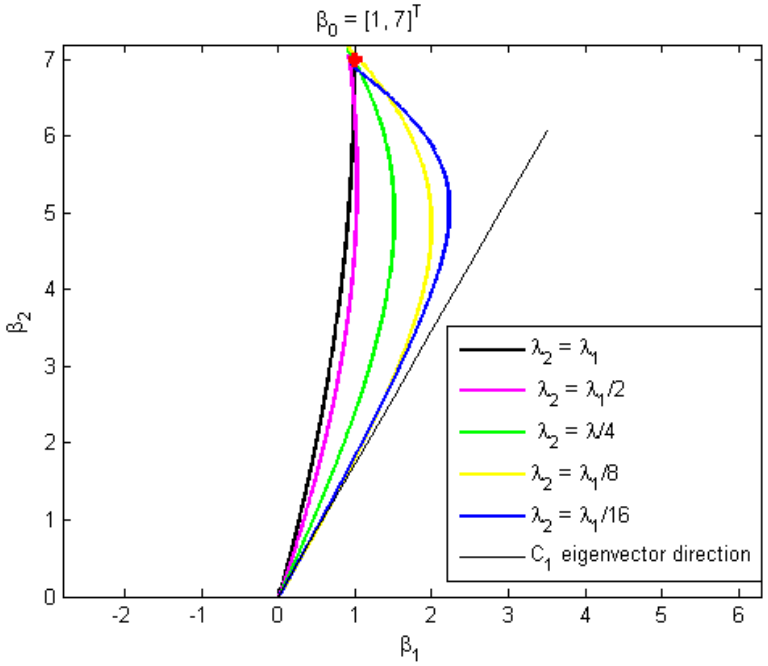
Consider the case where the number of regressors $k = 2$. We can rewrite the expression above as:

$$\hat{\beta}_\alpha = \left(C_1 C_1' \frac{\lambda_1}{\lambda_1 + \alpha} + C_2 C_2' \frac{\lambda_2}{\lambda_2 + \alpha}\right) \hat{\beta}_{ols}. \tag{1.3}$$

Note that C_1C_1' and C_2C_2' are projections onto dimensions spanned by the eigenvectors C_1 and C_2 respectively. Let $\lambda_2 < \lambda_1$. This implies that $\frac{\lambda_2}{\lambda_2+\alpha} < \frac{\lambda_1}{\lambda_1+\alpha}$. In words, ridge regression shrinks *both* dimensions towards the prior (the origin here), while penalizing the dimension associated with the smallest eigenvalue (C_2) more.

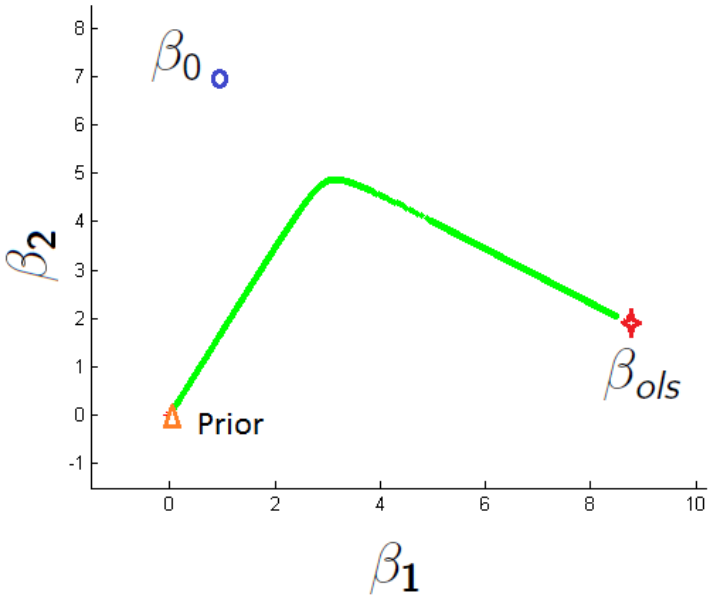
The following figure shows ridge paths for OLS problems corresponding to different λ_1/λ_2 ratios. Each path in the figure corresponds to a range of α values going from 0 to ∞

Figure 1.1: Ridge path in OLS for different values of λ_1/λ_2



We also present next the ridge path for a particular data sample which we will use throughout this section to describe the other two regularization paths too. The OLS estimate corresponds to $\alpha = 0$ and the origin corresponds to $\alpha \approx \infty$.

Figure 1.2: Ridge path in OLS characterized by tuning parameter α



1.3.3 Spectral Cutoff Regularization

The spectral cutoff regularization solution takes the OLS solution and removes all eigenvectors corresponding to eigenvalues less than some cut-off paramater γ (or the 'bad' dimensions). Mathematically,

$$\hat{\beta}_{ols} = \left(\frac{X'X}{n}\right)^{-1} \frac{X'Y}{n} = C\Lambda^{-1}C'\frac{X'Y}{n}.$$

If $\lambda_j < \gamma, \forall j > r$ then

$$\hat{\beta}_{spectral} = \left(C\Lambda_r C'\right)^{-1} \frac{X'Y}{n} = \left(C\Lambda_r^{+}C'\right) \frac{X'Y}{n}$$

where

$$\Lambda_r = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & 0 \\ 0 & 0 & \lambda_r & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

ignores all $(k-r)$ eigenvalues below the threshold γ and

$$\Lambda_r^{+} = \begin{pmatrix} 1/\lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & 0 \\ 0 & 0 & 1/\lambda_r & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

is the Moore-Penrose Inverse (MPI) of the singular matrix Λ_r.

Consider the case with $k = 2$ regressors where eigenvalue $\lambda_2 < \gamma$ i.e. the 'good' and

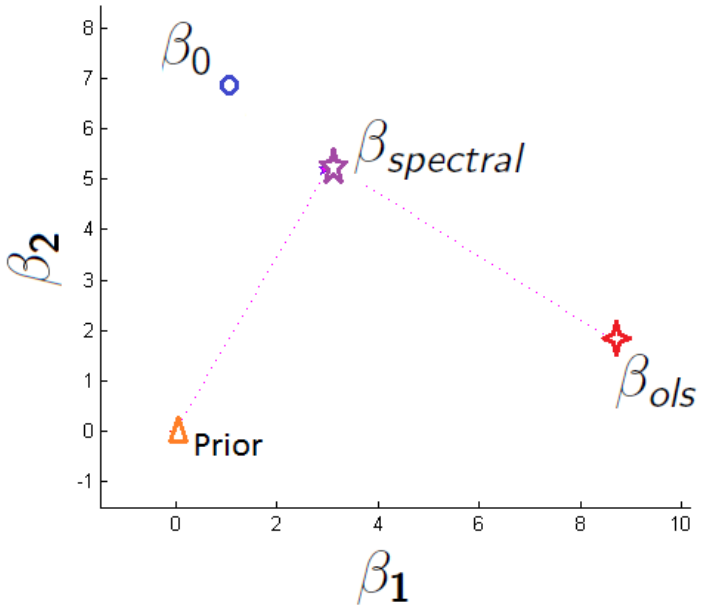
'bad' dimension corresponds to eigenvectors C_1 and C_2 respectively.

$$\begin{aligned}
\hat{\beta}_{spectral} &= \left[\begin{pmatrix} C_1 & C_2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} \right]^{-1} \frac{X'Y}{N} \\
&= \begin{pmatrix} C_1 & C_2 \end{pmatrix} \begin{pmatrix} 1/\lambda_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} \frac{X'Y}{N} \\
&= \frac{C_1 C_1'}{\lambda_1} \cdot \frac{X'Y}{N} \\
&= \frac{C_1 C_1'}{\lambda_1} \left[C \Lambda C' \left(C \Lambda C' \right)^{-1} \right] \frac{X'Y}{N} \\
&= \frac{C_1 C_1'}{\lambda_1} \left[\lambda_1 C_1 C_1' + \lambda_2 C_2 C_2' \right] \hat{\beta}_{ols} \\
&= C_1 C_1' \hat{\beta}_{ols}.
\end{aligned}$$

From this expression it can be seen that the spectral cutoff regularization solution is the projection of the OLS estimate onto the 'good' dimension. Note that unlike ridge regularization, where *all* dimensions are simultaneously penalized (and the magnitude of the penalty is inversely proportional to the magnitude of the corresponding eigenvalue), in spectral cut-off regularization *only* the 'bad' dimension is penalized by ignoring it completely.

Like ridge regularization, spectral cutoff regularization too is associated with a path between the prior and the OLS estimate. Here we go from the *high bias low variance* prior to the *low bias high variance* OLS estimate in a discrete 2-step path. The first step is the point on the 'bad dimension' closest to the prior and the second step is along the 'bad dimension' to the OLS estimate.

Figure 1.3: Spectral Path in GMM consisting of two discrete steps



1.3.4 Iterative Landweber Regularization

The Landweber Regularization algorithm belongs to the class of iterative algorithms for solving ill-posed inverse problems. Start with,

$$Y = X\beta \quad \Rightarrow \quad X'Y = X'X\beta$$

which leads to the fixed point equation

$$\beta = \beta - (X'X\beta - X'Y)$$

and the following iteration rule with a damping parameter τ

$$\beta^{(k+1)} = \beta^{(k)} - \tau \left(X'X\beta^{(k)} - X'Y\right).$$

The standard initial value is the origin, i.e. $\beta^{(0)} = 0$ and we can show that

$$\lim_{k\to\infty} \beta^{(k)} = \hat{\beta}_{ols} \quad \forall \quad 0 < \tau < 2/\sigma^2_{max}$$

where σ_{max} is the largest eigenvalue associated with the matrix $X'X$.

$$\begin{aligned}
\beta^{(k+1)} &= \beta^{(k)} - \tau\left(X'X\beta^{(k)} - X'Y\right) \\
&= \left(I - \tau\left(X'X\right)\right)\beta^{(k)} + \tau X'Y
\end{aligned}$$

(Recursively replace terms from previous iteration)

$$\begin{aligned}
&= \left(I - \tau\left(X'X\right)\right)\left[\left(1 - \tau\left(X'X\right)\right)\beta^{(k-1)} + \tau X'Y\right] + \tau X'Y \\
&= \left(I - \tau\left(X'X\right)\right)^2\beta^{(k-1)} + \tau\left[I + \left(I - \tau\left(X'X\right)\right)\right] \\
&\vdots \\
&= \left(I - \tau\left(X'X\right)\right)^k\beta^{(0)} + \tau\left[I + \left(I - \tau\left(X'X\right)\right) + \left(I - \tau\left(X'X\right)\right)^2 + \cdots\left(I - \tau\left(X'X\right)\right)^k\right]X'Y
\end{aligned}$$

(Since the prior $\beta^{(0)} = 0$,)

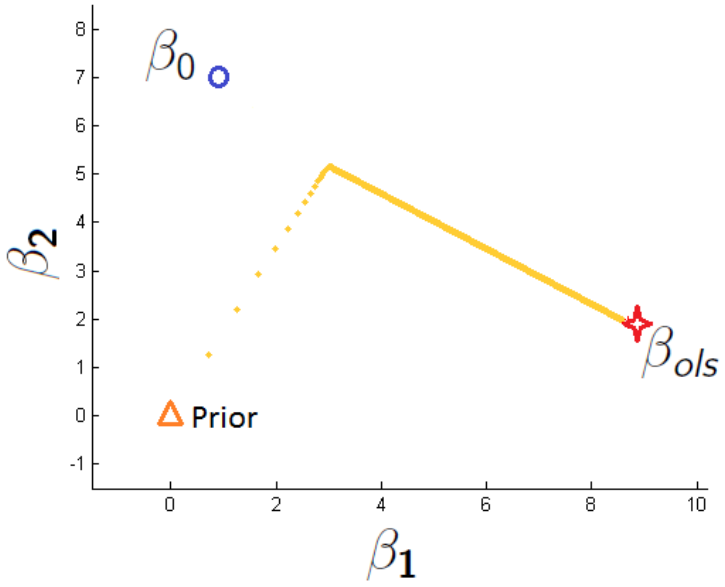
$$= \tau\left[I + \left(I - \tau\left(X'X\right)\right) + \left(I - \tau\left(X'X\right)\right)^2 + \cdots\left(I - \tau\left(X'X\right)\right)^k\right]X'Y$$

(If, $(0 < ||\tau|| < \sigma_{max}^2/2)$, then in the limit)

$$\begin{aligned}
\lim_{k\to\infty}\left[\beta^{(k+1)}\right] &= \tau\left[I - \left(I - \tau\left(X'X\right)\right)\right]^{-1}X'Y \\
&= \left(X'X\right)^{-1}X'Y \\
&= \hat{\beta}_{ols}.
\end{aligned}$$

Some of the popular stopping criteria are the Discrepancy Principle, the Monotone error Rule and Generalized crossvalidation. Note that here too the eigenvalues play an important role. Further, this regularization technique too can be viewed as a path between the prior at the origin and the OLS estimate.

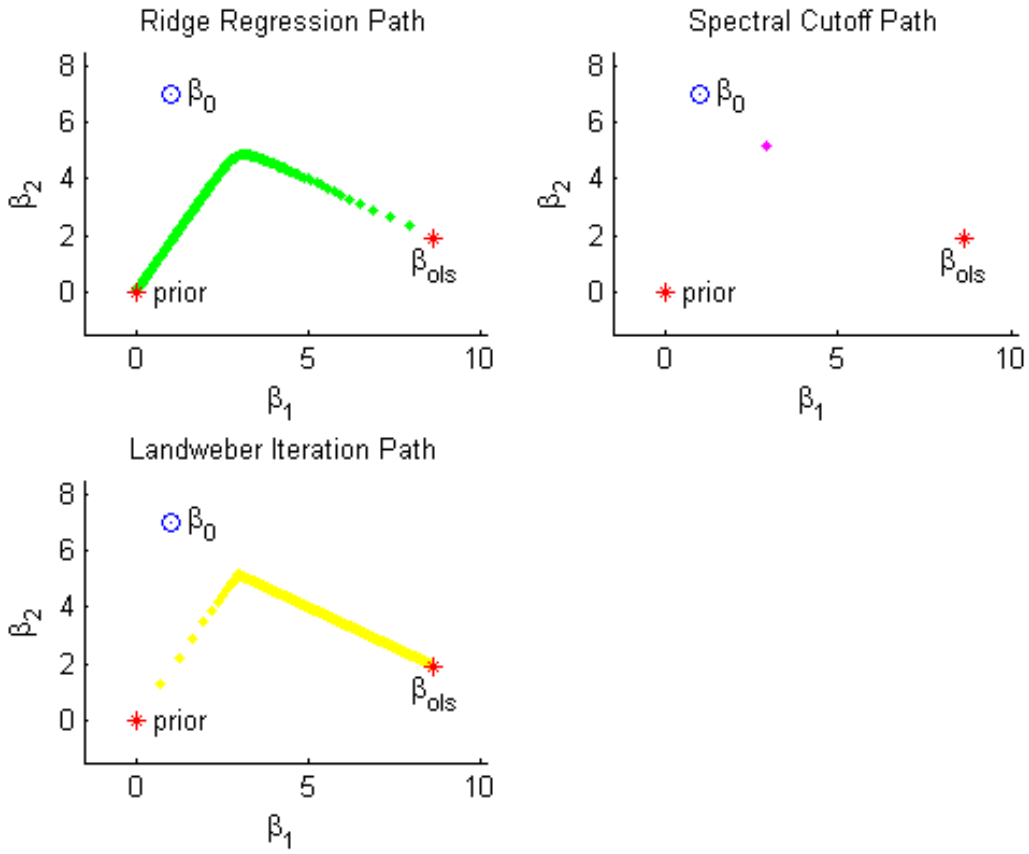
Figure 1.4: Landweber iteration path in GMM characterized by number of interations k



From the preceding discussion we note that all the above methods utilize the property that eigenvectors are fixed over the entire parameter space in linear optimization problems. However in non-linear optimization problems eigenvectors vary with the parameter space. Thus generalizations based on eigenvectors are non-trivial.

There exists another common property shared by the techniques discussed above – they are all characterized by a path between the *high bias-low variance* prior and the *low bias-high variance* OLS estimate.

Figure 1.5: Regularization paths in OLS



We define these paths as *'regularized solution paths'* Figure (1.3) presents the different regularized solution paths corresponding to Ridge Regularization, Spectral Cutoff Regularization, Landweber Iterative Regularization. We propose three regularized solution paths for GMM:

- Ridge-type solution path,
- Geodesic solution path,
- Local spectral cutoff path.

In the next section we describe these in greater detail.

1.4 Regularized Solution Paths in GMM

1.4.1 Detecting the Identification Problem

Before describing the regularization procedures, we provide guidance on how to detect the identication problem described in Section 1.2.2. Regularized GMM estimates should be calculated alongside traditional GMM estimates for samples that display these characteristcs.

- Very high condition number associated with $M(\hat{\theta})$.
- Unstable solution i.e. small changes in the sample (taking out some data points randomly) lead to disproportionately large changes in parameter estimate values.
- Very large asymptotic confidence intervals associated with the parameter estimates.

The next figure present what an identification problem in GMM looks like – note the curved ridge in the objective which contains on a number of local minima.

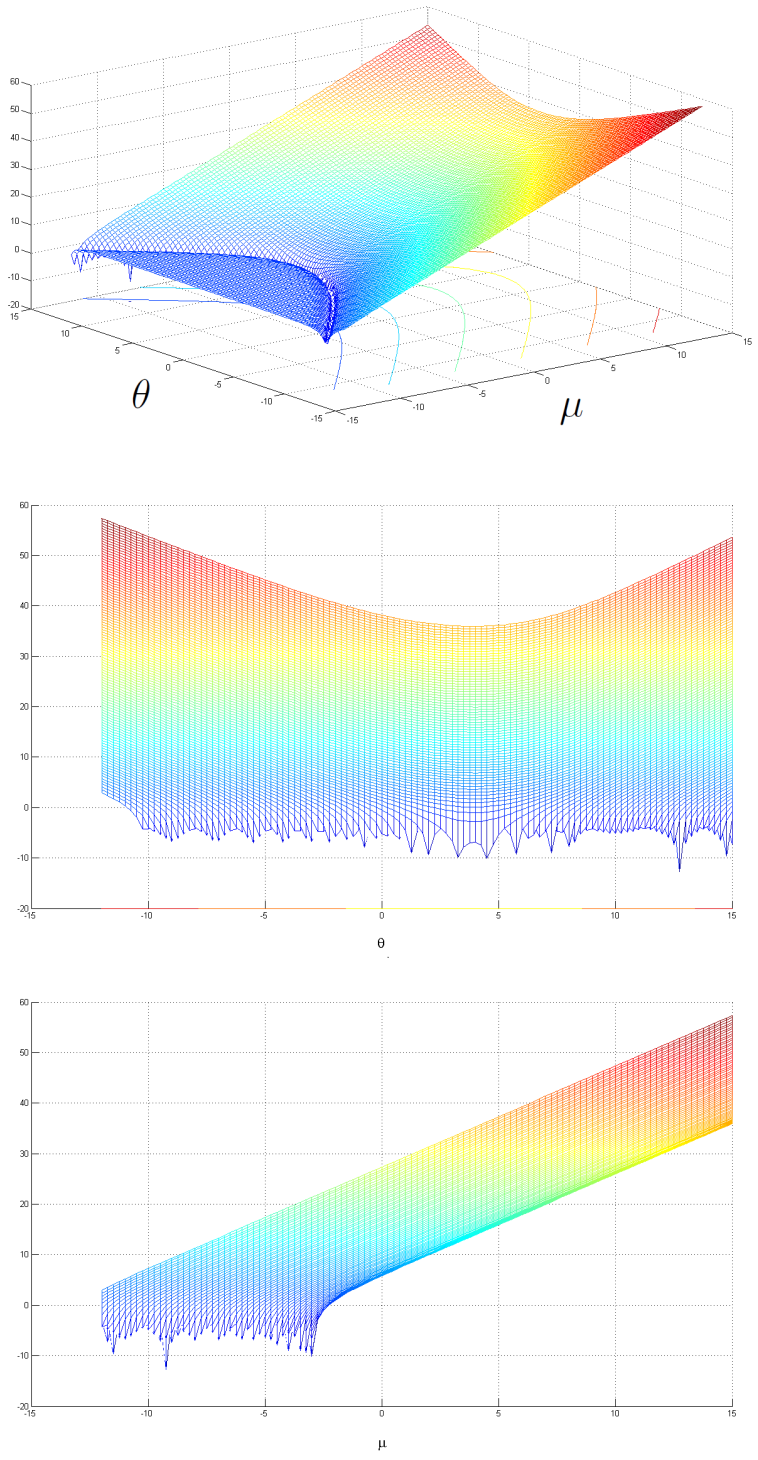
1.4.2 Hold out Sample in GMM

The three procedures described next all require using a hold out sample. The use of this method is ubiquitous in machine learning and involves randomly picking some of the data out of the full sample for testing. This is known as the *testing data*. The remaining data is known as the *training data*. A set of candidate parameter values is obtained from the *training data*. The parameter value from the candidate set that minimizes the objective function over the *testing data* is selected as the final regularized estimate.

Keeping with our notation in Section 1.2.1, we define the full data set as $\{x\}$ and the randomly selected training and testing data as $\{x_{trn}\}$ and $\{x_{tst}\}$ respectively. Candidate parameter values $\theta_{set}(x_{trn})$ are obtained using $\{x_{trn}\}$ and the loss function is:

$$Loss(\theta, x_{tst}) = G\left(x_{tst}, \theta\right)' W G\left(x_{tst}, \theta\right).$$

Figure 1.6: Nonlinear objective function with multiple local minima



The final regularized estimate is:

$$\widehat{\theta}_{reg} = \underset{\theta \in \theta_{set}(x_{trn})}{min} CV(\theta, x_{tst}).$$

It should be noted here that throughout this chapter we use the identity matrix for weighting i.e. $W = I_m$.

1.4.3 Ridge-type solution path

The first of the three procedures is an implementation of ridge regularization type algorithm. The implementation of the technique in the GMM setting is straight forward and involves appending the traditional GMM objective on the *training data* with a term that penalizes euclidian distance from the prior. Setting the prior θ_{prior}, the ridge type objective is:

$$Q(\theta, \alpha) = G(\theta)' \cdot W \cdot G(\theta) + \alpha \cdot (\theta - \theta_{prior})'(\theta - \theta_{prior}).$$

where α is known as a *tuning parameter* and θ_p is the prior which can be thought of as the starting point forthe regularization path. The solutions corresponding to different values of α are characterized by:

$$\widehat{\theta}_{\alpha} = \underset{\theta}{min}\, Q(\theta, \alpha, x_{trn})$$

When $\alpha = 0$ the objective function corresponds to traditional GMM and the solution is the traditional GMM estimate. When $\alpha \approx \infty$ the optimal estimate is the prior. The path between the GMM estimate and the prior is characterized by values of $0 < \alpha < \infty$.

To pick the optimal tuning parameter α^*, a hold out sample is employed. The value of the (traditional) objective is computed on the *testing set* (plugging in $\hat{\theta}_{\alpha}, \quad \alpha \in (0 \leq \alpha_1 < \alpha_2 < \cdots \alpha_l)$).

$$CV(\widehat{\theta}_{\alpha}, x_{tst}) = G\left(x_{tst}, \widehat{\theta}_{\alpha}\right)' \cdot W \cdot G\left(x_{tst}, \widehat{\theta}_{\alpha}\right).$$

The parameter estimate corresponding to the value of α that minimizes the loss function on the hold out sample is the regularized estimate.

$$\widehat{\theta}_{ridge} = \underset{\widehat{\theta}_\alpha(x_{trn})}{min}\ CV(\widehat{\theta}_\alpha, x_{tst}).$$

As discussed in Section 1.3.2, ridge regression utilizes the constant eigenvectors and eigenvalues associated with OLS. Therefore we expect this technique to perform especially well in GMM problems with moment conditions that are *linear* in parameters and thus associated with nonvarying eigenvalues and eigenvectors. However in general for non-linear GMM problems with varying eigendecompositions, its not necessary that this is the best regularization technique. We expect the next two techniques to perform better in such cases.

1.4.4 Geodesic solution path

Since most OLS regularization techniques can be characterized as unidimensional paths between a prior and the traditional OLS estimate, the natural generalization to non-linear GMM is the use of geodesics. The geodesic is defined as the shortest path between two points along a curved surface.

The second regularization procedure we propose is to compute the geodesic along the non-linear objective function surface between the prior and the GMM estimate on the *training data*. This forms the candidate set of parameter values which are then plugged into the loss function computed from the *testing data*. The point on the geodesic that minimizes the loss function is chosen as the regularized parameter estimate. Thus if $p = (0, \theta^1, \theta^2, \cdots \theta^l, \widehat{\theta}_{gmm})$ is a vector of $l + 2$ discrete points on the geodesic, then the regularized estimate is

$$\widehat{\theta}_g = \underset{\theta \in p}{min}\ CV(\theta, x_{tst}).$$

Implementation in cases where $k = 2$

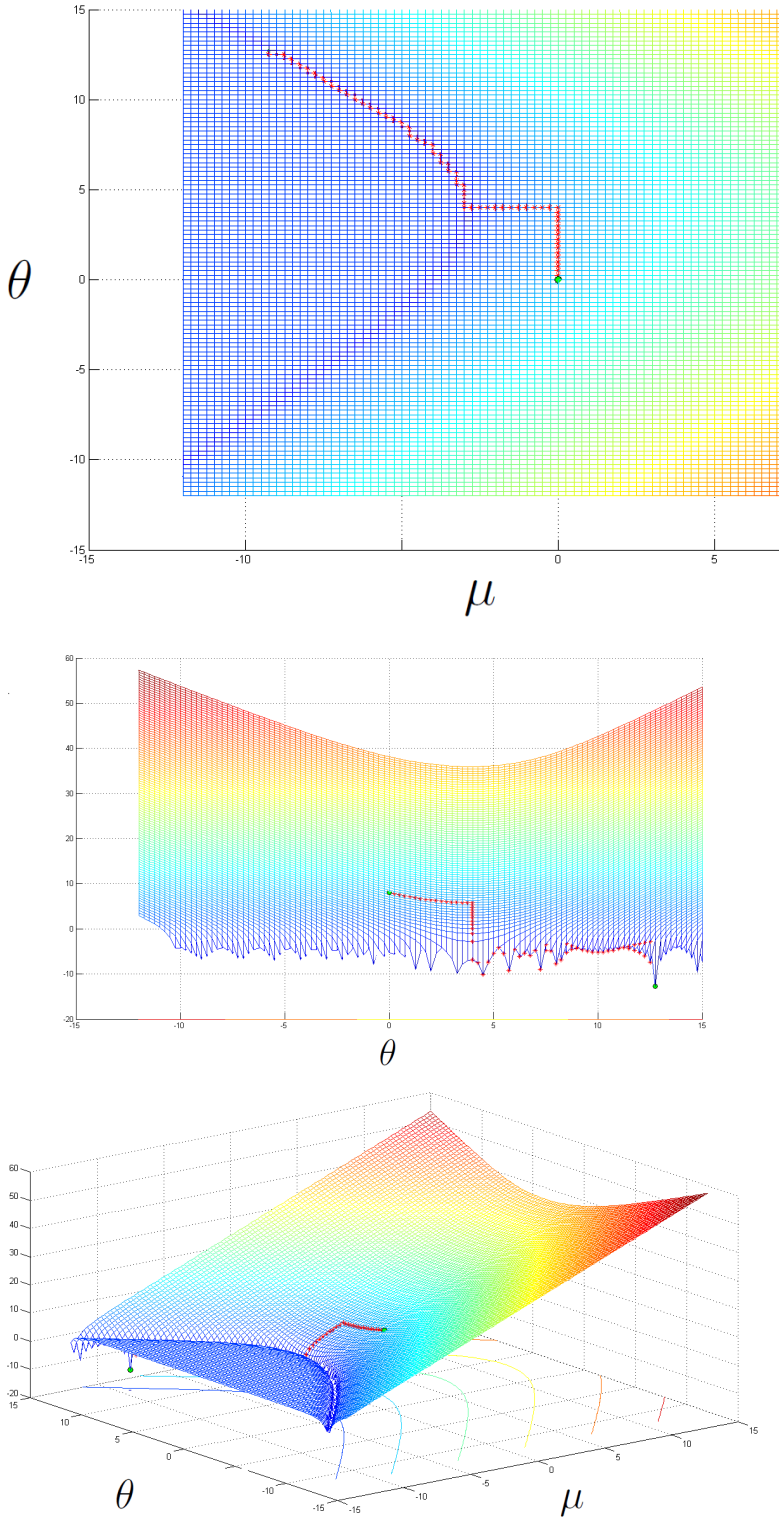
The most rigorous way to compute geodesics is via differential equations. However the differential equation approach may not be feasible for highly non-linear surfaces. Computer scientists calculate approximate geodesics on a surface using weighted graphs. *Dijkstra's Algorithm* is one of the most popular algorithm's to find geodesics between two nodes on a weighted graph with non-negative weights. We apply the Dijkstra's Algorithm to compute the approximate geodesic between the prior at the origin and the GMM estimate. This requires computation of the objective function over a grid of parameter values. Next we describe how the appropriate graph nodes, edges and weights are computed.

- **Nodes:** The nodes of the graph correspond to all the different pairs of parameter values over the grid. The finer the grid, the closer we will get to the true geodesic. Each node i is associated with parameter value $\theta(i)$ and value of the objective function $Q(i) = Q(\theta(i))$.
- **Edges:** Each node shares edges only with its closest neighbors.
- **Weights:** The weight on the edge between two connected nodes i and j is:

$$w_{ij} = \sqrt{(Q(i) - Q(j))^2 + s_g^2}$$

where s_g is the grid size used in computation. Once these parameters have been computed, most statistical and computational packages include in-built commands to obtain the geodesic. The next figure shows a sample geodesic for the nonlinear objective in Figure 1.7.

Figure 1.7: Numerical Geodesic on Nonlinear objective function



1.4.5 Local spectral cutoff path

The third regularization procedure we propose is based on finding a path that connects the prior and the GMM estimate via the *'ill-defined manifold'*. We refer to it as the *'ill-defined manifold'* because it is the dimension along which the solution varies the most – in other words this is the dimension associated with the smallest eigenvalue at each point in the parameter space. Note the similarity with spectral cutoff regularization where the solution is the projection of the OLS estimate onto the dimension(s) with the highest eigenvalue(s). The difference in non-linear GMM is that since the eigenvectors and eigenvalues vary with the parameters, therefore we need to search along the 'ill-defined manifold' too.

Implementation in cases where $k = 2$

In order to implement this regularization technique we follow these steps:

1. Using the *training data* find the global minimum as well as all the local minima. This gives us the 'ill-defined manifold'.

2. Now find θ^*, the point on the manifold that is closest to the prior. The points on the manifold between $\theta*$ and $\widehat{\theta}_{gmm}$ is the *local spectral cutoff path*, p.

3. For all parameter values on the local spectral cutoff path evaluate the loss function on the *testing data*.

4. The parameter value that minimizes the loss function is the regularized parameter estimate.

$$\widehat{\theta}_s = \min_{\theta \in p} CV(\theta, x_{tst}).$$

A point to note here – this path can also be thought of as one that minimizes total value of the objective function along the path. In order to find this path on the training set we suggest the following alternative algorithm:

Algorithm for finding 'ill-defined manifold':

Step 1: Find the unconstrained minimum of the objective function. Denote this as θ^{step}, where $step = 0$.

Step 2: Evaluate the value of the objective function in an ϵ-hemisphere of θ^{step} (to do this evaluate the objective using polar coordinates where $\rho = \epsilon$ and $\phi \in \{0, \pi\}$ or $\phi \in \{\pi, 2\pi\}$).

Step 3: Find the parameter values corresponding to the lowest value of the objective in the ϵ-hemisphere. Denote as $\theta^{step,\epsilon}$

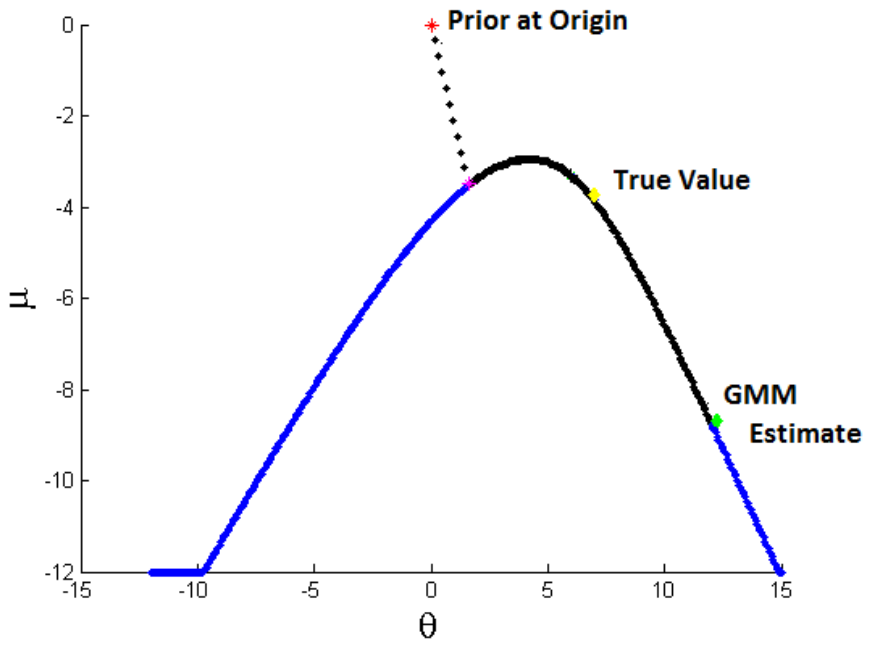
Step 4: $step = step + 1$ and $\theta^{step} \leftarrow \theta^{step-1,\epsilon}$.

Step 5: Repeat **Step 2** to **Step 4** for S number of steps for the hemispheres in two directions.

For the case where $k = 2$, a simple way around is to fix one parameter and find the value of the other parameter that minimizes the objective function.

The local spectral cutoff path for the non linear objective in Figure 1.7 is depicted here.

Figure 1.8: Local Spectral Cutoff path example with prior at the origin



1.5 Simulation Results

In order to carry out a simulation study that is appropriate for the non-linear GMM identification problem that we described in Section 1.2.2 any candidate model should satisy the following requirements:

- the model should be non-linear in at least $k = 2$ parameters;
- it should be possible to *tweak* the data in a way that gives rise to an identification problem via loss of rank of the first derivative of the vector of moments, $M(\theta)$;
- the true values of the parameter should lie away from the origin (since the origin is the conventional prior in regularization studies).

We chose to work with a modified version of the Hall and Horowitz (HH) model. The two modifications that we make are shifting the parameters away from the origin and allowing the ratio of standard deviations of the two data vectors to change.

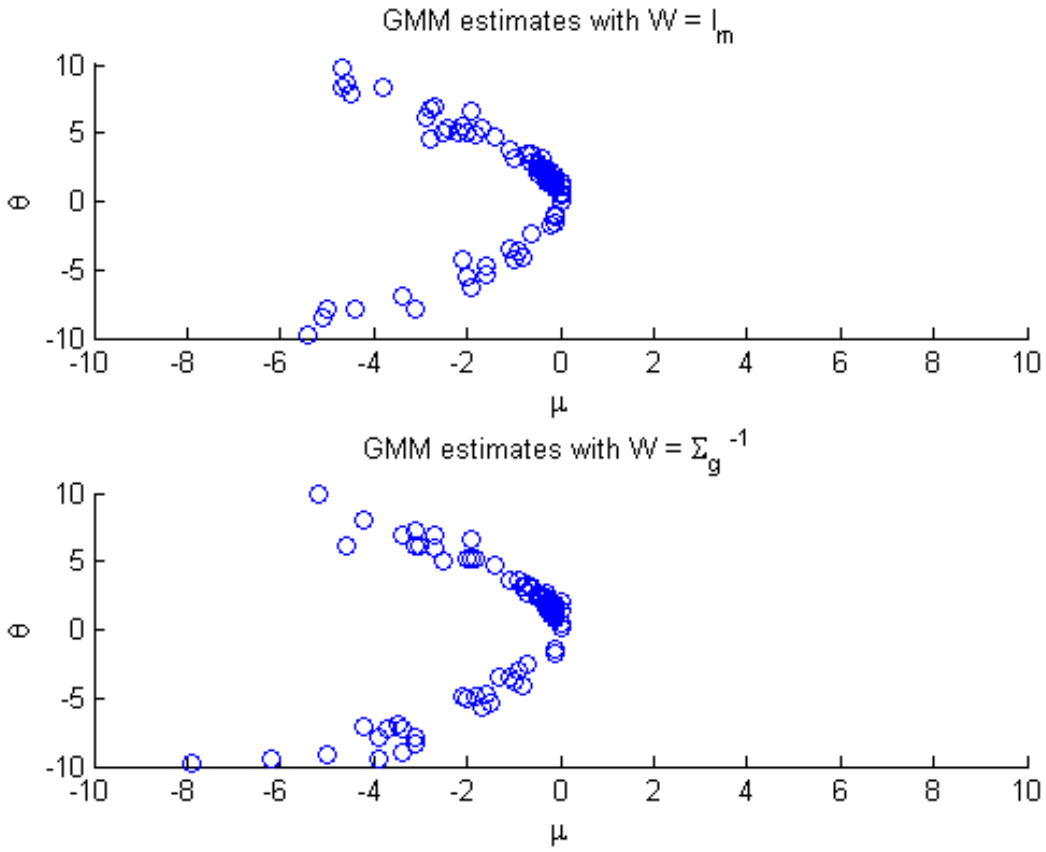
The base model that we use in our experiments is the HH model with the following moment conditions:

$$E(exp(\mu_0 + \theta_0(x + z) - 3z) - 1) = 0$$
$$E(z \cdot (exp(\mu_0 + \theta_0(x + z) - 3z) - 1)) = 0$$

where $\theta_0 = 3$, $\mu_0 = -0.72$ and $x \sim N(0, \sigma_x)$, $z \sim N(0, \sigma_z)$. Further $\sigma_x = \sigma_z = 0.4$.

Note that if $\sigma_z < \sigma_x$ the moment conditions above still hold but the one of them starts getting redundant as σ_z falls. Using this property we ran a number of simulations on models with small values of σ_z and noted that the estimated values were not clustered around the true parameter values but lay on a 'curved manifold'.

Figure 1.9: Curved 'Bad' Manifold obtained in Hall and Horowitz model with an identification problem



Moment Conditions with 'Shifted' Population Parameters

Since the true parameter values are close to the origin (which is our prior) we shifted them away from the origin. The moment conditions with the 'shifted' parameter values are:

$$E(exp(\tilde{\mu}_0 + 3 + (\tilde{\theta}_0 - 4)(x + z) - 3z) - 1) = 0$$
$$E(z \cdot \Big(exp(\tilde{\mu}_0 + 3 + (\tilde{\theta}_0 - 4)(x + z) - 3z) - 1\Big)) = 0$$

where $\tilde{\theta}_0 = 7$, $\tilde{\mu}_0 = -3.72$, $x \sim N(0, \sigma_x)$, $z \sim N(0, \sigma_z)$ and $\sigma_x = 0.4$. The value of σ_z is allowed to vary.

We present simulation results for a set of four experiments. The only paramater that

we change in the four sets of simulations is σ_z. In particular,

- Number of simulations in each set $N = 1,000$; number of datapoints in the full sample, in the training data and in the testing data is $n = 100$, $n_{train} = 70$ and $n_{test} = 30$ respectively.
- True values of the paramters $\theta_0 = 7$ and $\mu_0 = -3.72$.
- The X data is drawn from the distribution $N(0, \sigma_x^2)$ and the Z data is drawn from the distribution $N(0, \sigma_z^2)$. $\sigma_x = 0.4$ is fixed and $\sigma_z = \rho\sigma_x$ varies in the four sets ($\rho = 1, 0.1, 0.01, 0.001$).

For each of the four sets of simulations, we estimate the traditional GMM model as well as the three regularization procedures developed in this paper. The evaluation criteria for the results is the value of the Mean Squared Error.

$$
\begin{aligned}
MSE(\theta) &= \left[\overline{\theta} - \theta_0\right]^2, \\
MSE(\mu) &= \left[\overline{\mu} - \mu_0\right]^2, \\
MSE(\theta, \mu) &= MSE(\theta) + MSE(\mu).
\end{aligned}
$$

We find that except in the case where the assumption ($\sigma_z = \sigma_x = 0.4$) holds, all three regularization techniques perform better in terms of MSE than traditional GMM estimation.

The three regularization techniques also perform well in terms of MSE when the given GMM estimation problem is linear in parameters.

Case 1 : $\sigma_z = \sigma_x$

Figure 1.10: Scatterplots of GMM estimates and Regularized estimates

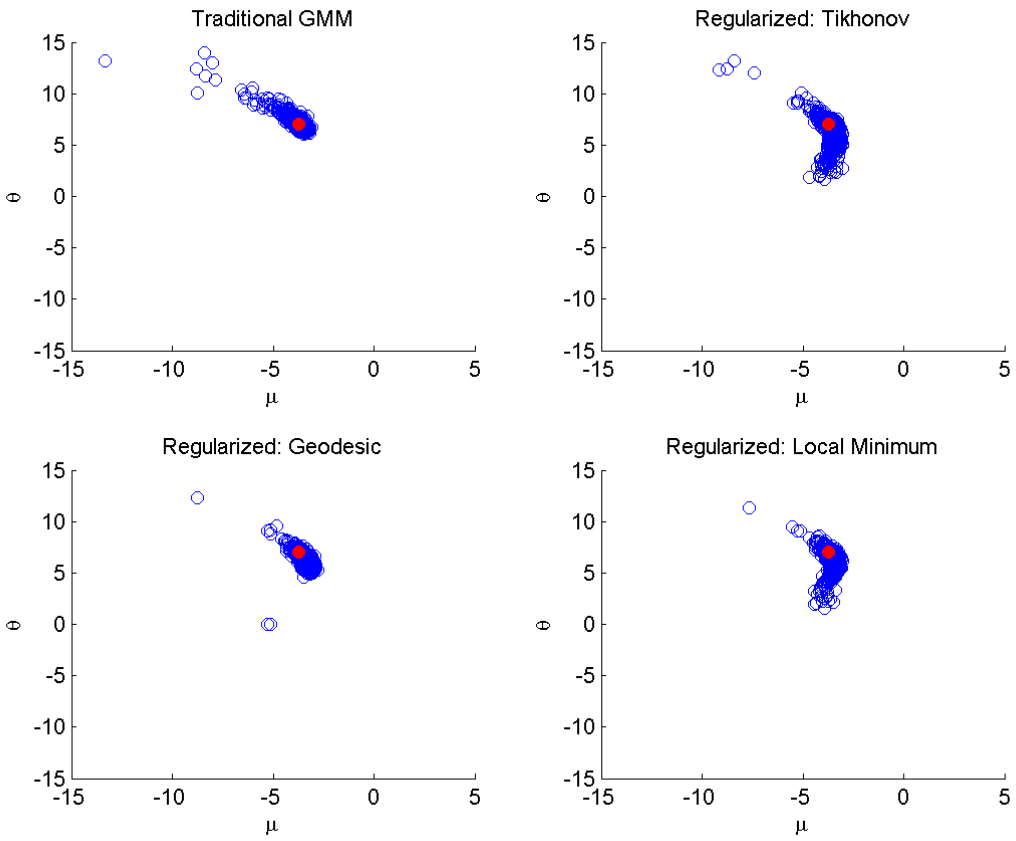


Table 1.1: Bias, Variance, MSE comparison of regularized estimates with GMM estimates

		Bias	Variance	MSE
GMM	$\hat{\mu}_{gmm}$	-0.084	0.380	0.386
	$\hat{\theta}_{gmm}$	0.146	0.511	0.532
	$(\hat{\mu}_{gmm}, \hat{\theta}_{gmm})$			0.918
Ridge-type	$\hat{\mu}_r$	0.135	0.177	0.195
	$\hat{\theta}_r$	-0.846	1.377	2.091
	$(\hat{\mu}_r, \hat{\theta}_r)$			2.286
Geodesic	$\hat{\mu}_g$	0.275	0.101	0.176
	$\hat{\theta}_g$	-0.697	0.442	0.927
	$(\hat{\mu}_g, \hat{\theta}_g)$			1.103
Local Spectral	$\hat{\mu}_s$	0.176	0.072	0.103
	$\hat{\theta}_s$	-0.789	0.783	1.405
	$(\hat{\mu}_s, \hat{\theta}_s)$			1.509

Case 2 : $\sigma_z = (0.1)\,\sigma_x$

Figure 1.11: Scatterplots of GMM estimates and Regularized estimates

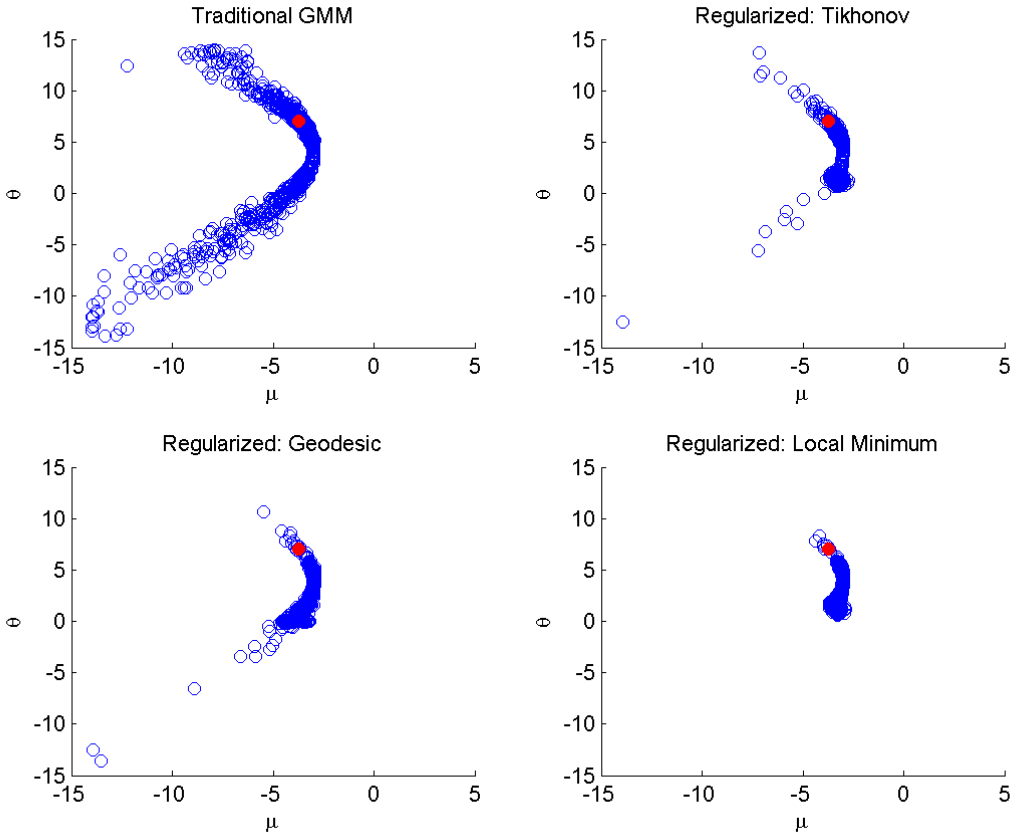


Table 1.2: Bias, Variance, MSE comparison of regularized estimates with GMM estimates

		Bias	Variance	MSE
GMM	$\hat{\mu}_{gmm}$	-0.743	4.484	5.032
	$\hat{\theta}_{gmm}$	-2.947	24.577	33.240
	$(\hat{\mu}_{gmm}, \hat{\theta}_{gmm})$			38.272
Ridge-type	$\hat{\mu}_r$	0.296	0.255	0.343
	$\hat{\theta}_r$	-4.671	4.211	26.025
	$(\hat{\mu}_r, \hat{\theta}_r)$			26.368
Geodesic	$\hat{\mu}_g$	0.420	0.463	0.639
	$\hat{\theta}_g$	-3.834	4.505	19.197
	$(\hat{\mu}_g, \hat{\theta}_g)$			19.836
Local Spectral	$\hat{\mu}_s$	0.550	0.047	0.349
	$\hat{\theta}_s$	-3.792	1.930	16.305
	$(\hat{\mu}_s, \hat{\theta}_s)$			16.654

Case 3 : $\sigma_z = (0.01)\,\sigma_x$

Figure 1.12: Scatterplots of GMM estimates and Regularized estimates

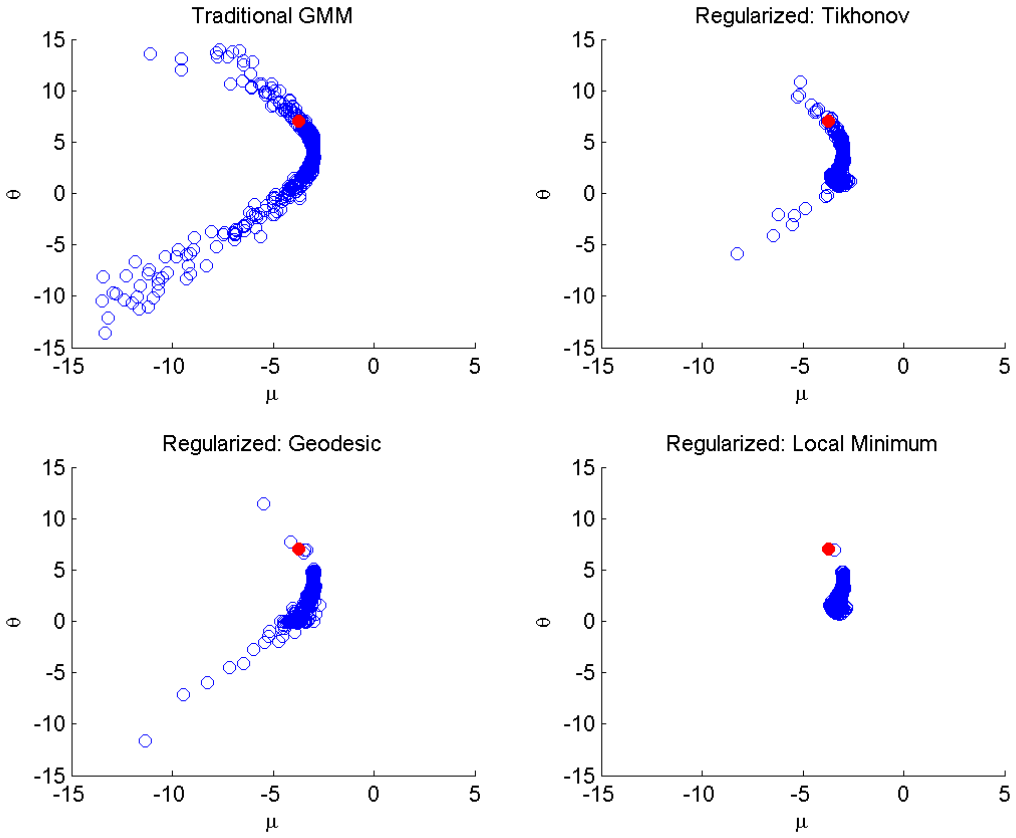


Table 1.3: Bias, Variance, MSE comparison of regularized estimates with GMM estimates

		Bias	Variance	MSE
GMM	$\hat{\mu}_{gmm}$	0.102	2.696	2.703
	$\hat{\theta}_{gmm}$	-3.420	10.333	22.018
	$(\hat{\mu}_{gmm}, \hat{\theta}_{gmm})$			24.722
Ridge-type	$\hat{\mu}_r$	0.450	0.117	0.320
	$\hat{\theta}_r$	-4.507	2.538	22.845
	$(\hat{\mu}_r, \hat{\theta}_r)$			23.164
Geodesic	$\hat{\mu}_g$	0.559	0.279	0.591
	$\hat{\theta}_g$	-3.739	2.520	16.501
	$(\hat{\mu}_g, \hat{\theta}_g)$			17.092
Local Spectral	$\hat{\mu}_s$	0.622	0.028	0.414
	$\hat{\theta}_s$	-3.788	1.111	15.459
	$(\hat{\mu}_s, \hat{\theta}_s)$			15.874

Case 4 : $\sigma_z = (0.001)\,\sigma_x$

Figure 1.13: Scatterplots of GMM estimates and Regularized estimates

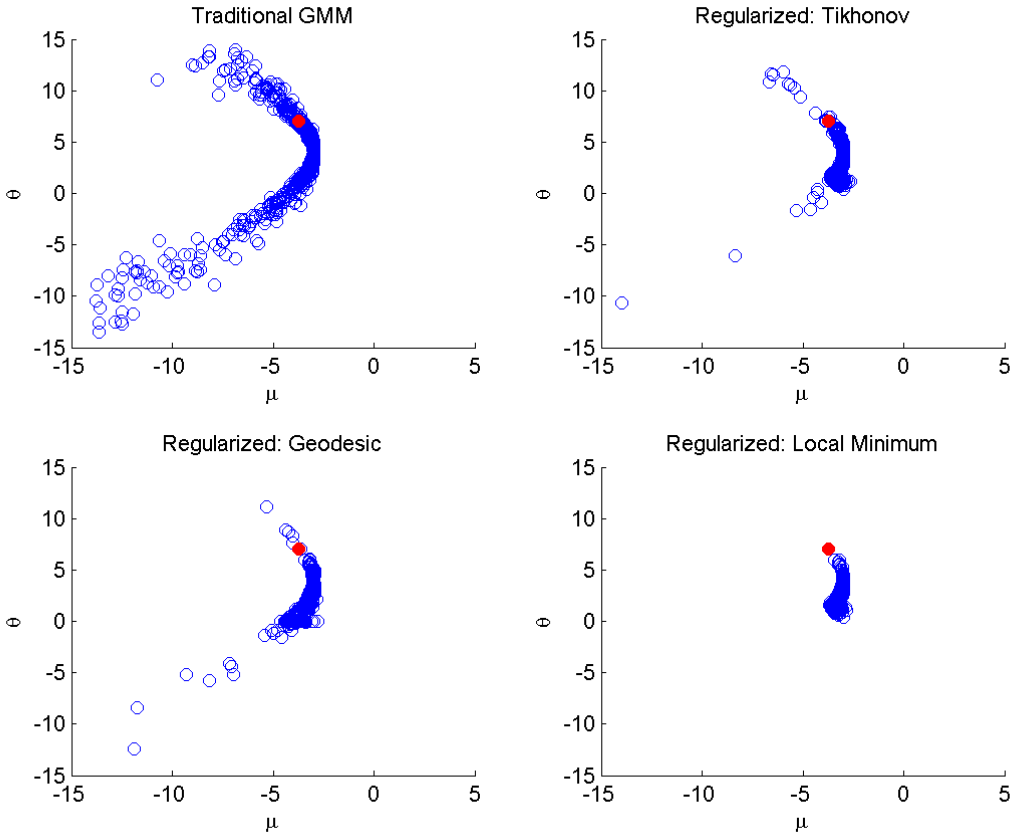


Table 1.4: Bias, Variance, MSE comparison of regularized estimates with GMM estimates

		Bias	Variance	MSE
GMM	$\hat{\mu}_{gmm}$	-0.201	3.818	3.855
	$\hat{\theta}_{gmm}$	-3.467	15.604	27.606
	$(\hat{\mu}_{gmm}, \hat{\theta}_{gmm})$			31.461
Ridge-type	$\hat{\mu}_r$	0.436	0.257	0.447
	$\hat{\theta}_r$	-4.384	3.011	22.229
	$(\hat{\mu}_r, \hat{\theta}_r)$			22.676
Geodesic	$\hat{\mu}_g$	0.487	0.402	0.639
	$\hat{\theta}_g$	-3.895	3.382	18.553
	$(\hat{\mu}_g, \hat{\theta}_g)$			19.192
Local Spectral	$\hat{\mu}_s$	0.598	0.037	0.394
	$\hat{\theta}_s$	-3.836	1.341	16.052
	$(\hat{\mu}_s, \hat{\theta}_s)$			16.447

1.6 Application

One of the leading application of the GMM technique is inferences on the consumption based capital asset pricing model. The model and the statistical properties of its GMM estimates have been widely studied in the literature including by Tauchen (1986), Kocherlakota (1990), Tauchen and Hussey (1991), Wright (2003) and most recently Inoue and Rossi (2010). In particular Kocherlakota (1990), Wright (2003) and Inoue and Rossi (2010) study the model in the context of identification issues.

In this section we apply the regularization techniques discussed in this paper on specifications of the consumption based capital asset pricing model that have been widely studied. We find that in cases where the model is not well-defined, regularization usually leads to estimates with lower MSE values. As in the first set of simulation results, when the model is poorly identified as well as when sample sizes are small, regularization leads to lower MSE values in general. However, the benefits decrease and eventually vanish as sample sizes become larger as well as when the prior moves very far from the true values.

Next we describe the consumption based asset pricing model in greater detail. We then discuss how the simulation data is generated using finite state markov chain approximations. This is followed by a discussion on the simulation setup and results.

1.6.1 Consumption Based Capital Asset Pricing Model

The consumption based capital asset pricing model using time separable preferences and constant relative risk coefficient (γ) is discussed in this section. The consumer's utility function is

$$u(c) = \frac{c^{1-\gamma}}{(1-\gamma)}.$$

The consumer is modeled as maximizing expected lifetime discounted utility

$$\max_{\{c_t\}_{t=0}^{\infty}} E\left[\sum_{t=0}^{\infty} \beta^t u(c_t)\right]$$

subject to the constraint

$$c_t + \Sigma_{i=1}^{M} q_{i,t} p_{i,t} = \sum_{i=1}^{M} q_{i,t-1}(p_{i,t} + d_{i,t})$$

where $d_{i,t}$ is dividends, $q_{i,t}$ is the holding of asset i at time period t and $p_{i,t}$ is asset i's price at time period t.

The first order conditions for this problem are

$$p_{i,t}u'(c_t) = E[\beta u'(c_{t+1})(p_{i,t+1} + d_{i,t+1})], \qquad i = 1, \cdots M.$$

Since the consumption and dividend series might exhibit nonstationarity, the transformed stationary series $v_{i,t} = \frac{p_{i,t}}{d_{i,t}}$ is used. The first order condition then becomes

$$\frac{p_{i,t}}{d_{i,t}} = E\left[\beta \frac{u'(c_{t+1})}{u'(c_t)} \left(\frac{p_{i,t+1}}{d_{i,t+1}} + \frac{d_{i,t+1}}{d_{i,t+1}}\right) \frac{d_{i,t+1}}{d_{i,t}}\right]$$

$$\Rightarrow v_{i,t} = E\left[\beta \frac{u'(c_{t+1})}{u'(c_t)}(v_{i,t+1} + 1)\frac{d_{i,t+1}}{d_{i,t}}\right], \qquad i = 1, \cdots M.$$

Define

$$x_{i,t} = \frac{d_{i,t}}{d_{i,t-1}}$$
$$w_t = \frac{c_t}{c_{t-1}}$$

and use the functional form of the utility function to get

$$\frac{u'(c_{t+1})}{u'(c_t)} = (w_{t+1})^{-\gamma}$$

to rewrite the first order condition as

$$v_{i,t} = E[\beta (w_{t+1})^{-\gamma}(v_{i,t+1} + 1)x_{i,t+1}].$$

Alternatively the first order condition can be written as

$$1 = E\left[\beta R_{i,t+1}(w_{t+1})^{-\gamma}\right]$$

where

$$R_{i,t+1} = \frac{(v_{i,t+1}+1)x_{i,t+1}}{v_{i,t}}$$

is the rate of return on asset i held from time period t to $t+1$. Consider a single dividend paying asset. Then the first order conditions imply the moment conditions:

$$E[g_t(\beta,\gamma)] = E[\beta R_{t+1} w_t^{-\gamma} - 1|I_t] = 0$$

where I_t is the information set available at time period t and includes lagged values of w_t and R_t. Following Wright (2003) we will use three instruments $[1 \quad w_t \quad R_t]$ with corresponding moment conditions

$$g_t(\beta,\gamma) = (\beta R_{t+1} w_t^{-\gamma} - 1) \begin{bmatrix} 1 \\ w_t \\ R_t \end{bmatrix}.$$

These are the moment conditions that we use for estimation in the simulation exercises.

1.6.2 Data Generation using Finite State Markov Chain Approximations

The data for carrying out the estimation exercise described above is obtained via simulations. The state variables for this system are w_t and x_t. Their stochastic behavior is captured with a VAR(1):

$$\begin{bmatrix} \ln(w_t) \\ ln(x_t) \end{bmatrix} = \mu + \Phi \begin{bmatrix} \ln(w_{t-1}) \\ ln(x_{t-1}) \end{bmatrix} + u_t \tag{1.4}$$

where $u_t \sim iidN(0,\Omega)$. In order to simulate the VAR(1) data we use the Finite State Markov Chain Approximation method outlined in Tauchen (1986). The VAR(1) in eq

(1.4) is of the form

$$Y_t = \mu + \Phi Y_{t-1} + u_t; \qquad u_t \sim N(0, \Omega).$$

Decompose $\Omega = B'\Sigma_\epsilon B$ where $BB' = B'B = I_m$ and Σ_ϵ is diagonal. Premultiply by B

$$\begin{aligned} BY_t &= B\mu + B\Phi B'BY_{t-1} + Bu_t \\ \Rightarrow W_t &= B\mu + DW_{t-1} + \varepsilon_t \end{aligned}$$

where $W_t = BY_t$, $D = B\Phi B'$ and $\varepsilon_t = Bu_t \sim (0, \Sigma_\epsilon)$. This is of the form of Tauchen (1986) and can be approximated by a discrete markov chain.

Unconditional Mean and Variance of W_t

The unconditional mean $\tilde{\mu}$ is calculated as

$$\begin{aligned} E(W_t) &= E(B\mu + DW_{t-1} + \varepsilon_t) \\ \Rightarrow \tilde{\mu} &= B\mu + D\tilde{\mu} \\ \Rightarrow \tilde{\mu} &= (I_m - D)^{-1}B\mu. \end{aligned}$$

For obtaining the unconditional variance Σ_W, use $W_t = \beta\mu + DW_{t-1} + \varepsilon_t$, $\tilde{\mu} = (I_m - D)^{-1}\beta\mu$ and $\beta\mu = (I_m - D)(I_m - D)^{-1}\beta\mu$ to obtain

$$\begin{aligned} (W_t - \tilde{\mu}) &= D(W_{t-1} - \tilde{\mu}) + \varepsilon_t \\ \Rightarrow \Sigma_W &= D\Sigma_W D' + \Sigma_\epsilon \end{aligned}$$

which is calculated using recursive substitution (assuming stationarity of the system).

Markov Chain States and Transition Matrix

In order to simulate the values using markov chains, we first discretize W_t^i (where i refers to the item corresponding to the i^{th} element of W) with a grid of values three standard

deviations σ_W^i on either side of zero. These N^i values $(s_1^i, s_2^i, \cdots s_{N^i}^i)$ are the discrete values for the state of the Markov process for W_t^i. This grid is calculated for each of the elements $i = 1, 2 \cdots m$. The state of the system will be denoted by S_t and will contain the values $S_{j,t} = (s_{j^1,t}^1, s_{j^2,t}^2, \cdots s_{j^m,t}^m)$ where j^i is one of the N^i states for W_t^i. Note that this implies $N^* = N^1 \cdot N^2 \cdot \cdots \cdot N^m$ total different states for the markov chain.

In our simulations $m = 2$, grid size $N^1 = N^2 = 9$ and total number of states $N^* = 9 \times 9 = 81$.

Because the ϵ_t are independent, the transition probability of each series can be calculated individually and multiplied to get the joint density.

$$
\begin{aligned}
\Pi(j,k) =& P[S_{k,t+1}|S_{j,t}] \\
=& P[(s_{k^1,t+1}^1, s_{k^2,t+1}^2, \cdots s_{k^m,t+1}^m)|(s_{j^1,t}^1, s_{j^2,t}^2, \cdots s_{j^m,t}^m)] \\
=& \prod_{i=1}^{m} P[s_{k^i,t+1}^i|(s_{j^1,t}^1, s_{j^2,t}^2, \cdots s_{j^m,t}^m)] \\
=& \prod_{i=1}^{m} P\left[W_{t+1} \in \left[s_{k^i}^i \pm \frac{6\sigma_{W^i}}{2(N^i-1)}\right] |(s_{j^1,t}^1, s_{j^2,t}^2, \cdots s_{j^m,t}^m)\right] \\
& \vdots \\
& \text{(algebra)} \\
& \vdots \\
=& \prod_{i=1}^{m} \left[F\left(\frac{s_{k^i}^i - (DS_{j,t})^i + \frac{6\sigma_{W^i}}{2(N^i-1)}}{\sigma_\epsilon^i}\right) - \left(\frac{s_{k^i}^i - (DS_{j,t})^i - \frac{6\sigma_{W^i}}{2(N^i-1)}}{\sigma_\epsilon^i}\right)\right].
\end{aligned}
$$

Given the transition density matrix the stationary transition density matrix can be found by repeated multiplication

$$\Pi^* = \lim_{k\to\infty} \Pi^k.$$

GMM Data

We have described above how to simulate values for x_t and w_t using discrete markov chains. In order to obtain values for v_t use the first order condition for the single dividend paying asset case

$$v_t = E[\beta(w_{t+1})^{-\gamma}(v_{t+1}+1)x_{t+1}].$$

Using the transition matrix and states from themarkov chain approximation, this can be rewritten as

$$v(s) = \beta\Pi(s,s')[w(s')^{-\gamma}(v(s')+1)x(s')].$$

Let Q be the $N^* \times 1$ vector with individual elements $w(s)^{-\gamma}x(s)$. The system of equations that defines the value of $v(s)$ for the different states can be written

$$\begin{aligned} v &= \beta\Pi[\text{diag}(Q)]v + \beta\Pi Q \\ &= (I_{N^*} - \beta\Pi[\text{diag}(Q)])^{-1}\beta\Pi Q \end{aligned}$$

Therefore given values of μ, Φ and Ω we simulate corresponding samples of

$$\begin{bmatrix} x_t \\ w_t \\ v_t \end{bmatrix} = \begin{bmatrix} \frac{d_t}{d_{t-1}} \\ \frac{c_t}{c_{t-1}} \\ \frac{p_t}{d_t} \end{bmatrix}.$$

We can now calculate the rate of return $R_t = \frac{1+v_t}{v_{t-1}}x_t$. This data is used to obtain the GMM moment conditions

$$g_t(\beta,\gamma) = (\beta R_{t+1} w_t^{-\gamma} - 1)\begin{bmatrix} 1 \\ w_t \\ R_t \end{bmatrix}.$$

to estimate the parameters of interest β and γ .

1.6.3 Simulation Setup and Results

We simulate three sets of data which correspond to three different GMM settings:

- **Full Rank (FR)** of $M(\beta, \gamma)$ matrix– in this case GMM estimation is well-defined and identification is not an issue. The parameter values are the same as those proposed in Tauchen (1986).
- **Near Rank Failure (NRF)** of $M(\beta, \gamma)$ matrix– in this case GMM estimation is poorly-defined and identification issues arise. The parameter values are the same as those proposed in Kocherlakota (1990).
- **Rank Failure (RF)** of $M(\beta, \gamma)$ matrix– in this case GMM estimation is unstable and identification of both parameters simultaneously is not possible. The parameter values are the same as those proposed in Wright (2003).

The three sets of parameters are presented in the Table below. Throughout the simulations the true values of the parameters of interest $\beta_0 = 0.97$ and $\gamma_0 = 2$. Also, throughout the simulations the first 75% of the data is used as training data and the remaining 25% is used as testing data.

Table 1.5: Parametrizations used in Consumption based CAPM simulations

Model	μ	Φ	Ω
FR	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -0.5 & 0 \\ 0 & -0.5 \end{pmatrix}$	$\begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}$
NRF	$\begin{pmatrix} 0.021 \\ 0.004 \end{pmatrix}$	$\begin{pmatrix} -0.161 & 0.017 \\ 0.414 & 0.117 \end{pmatrix}$	$\begin{pmatrix} 0.0012 & 0.00177 \\ 0.00177 & 0.014 \end{pmatrix}$
RF	$\begin{pmatrix} 0.018 \\ 0.013 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0.0012 & 0.0017 \\ 0.0017 & 0.0146 \end{pmatrix}$

We compare simulation results (on sets of $N = 1,000$ simulations) by MSE values. The effect of following initial conditions is studied:

- Time periods $T = \{60, \quad 100, \quad 200\}$ (sample size of $T = 100$ corresponds to the available US annual data sample size).

- Discount rate prior $\beta = \{0.95, \quad 0.99\}$ (an annual discount rate of $\beta = 0.97$ is standard in the literature).

- CRRA constant prior $\gamma = \{1, \quad 2, \quad 6\}$ (a standard value of γ remains an open question in the literature, with authors proposing values in the range of 1 to 10).

- Cases where the origin is used as the prior $\beta_{prior} = 0$, $\gamma_{prior} = 0$ are also studied.

Simulation Results – Broad Insights

1. GMM using an identity matrix performs better than GMM using optimal weighting matrix whenever the derivative of objective functions is not well-defined (i.e. in RF and NRF). This lends support to the practice by many authors who prefer using one-step GMM in asset pricing applications.

2. Both one-step and two-step GMM performance improves significantly when the sample size increases. In smaller samples both (particularly two-step GMM) perform poorly in the RF and NRF cases but their performance improves in larger samples. [1]

3. The prior used for regularization does matter – particularly for the local spectral cutoff path.

4. Of the three regularization techniques the Geodesic Solution performed the best. It had the lowest MSE values in most cases where regularization was useful as well as the lowest MSE values out of the regularization techniques in the cases where GMM estimates dominated the regularized solutions.

Selected graphs and tables are presented next.

[1] Apart from using the FMINUNC technique in MATLAB we implemented two alternative minimization techniques – Pattern Search and Genetic Algorithm. We found that Pattern Search led to dramatic improvements to the GMM minimization. However the broad insights of the simulations stayed the same.

Rank Failure: $T = 60$, $\beta_{prior} = 0.99$, $\gamma_{prior} = 2$

Figure 1.14: Scatterplots of GMM estimates and Regularized estimates

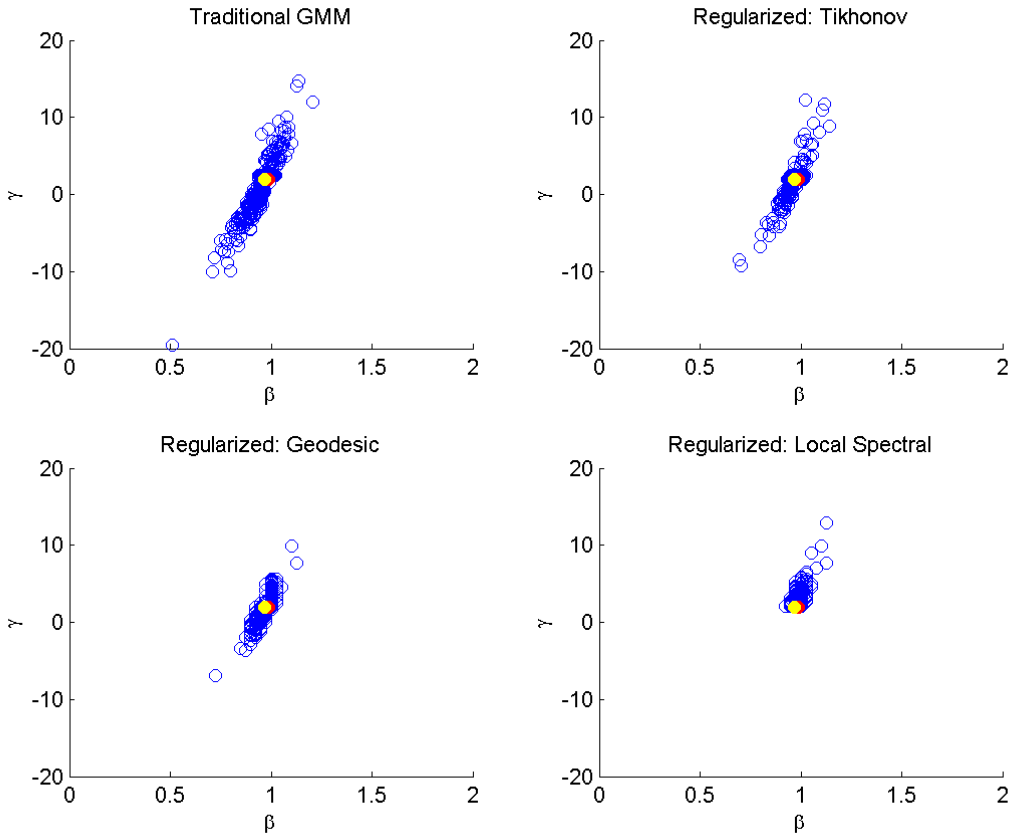


Table 1.6: Bias, Variance, MSE comparison of regularized estimates with GMM estimates

		Bias	Variance	MSE
GMM (Single Step)	$\hat{\beta}_{gmm,1}$	-0.0014	0.0021	0.0021
	$\hat{\gamma}_{gmm,1}$	-0.0230	5.1917	5.1870
	$(\hat{\beta}_{gmm,1}, \hat{\gamma}_{gmm,1})$	–	–	5.1891
GMM (Two Step)	$\hat{\beta}_{gmm,2}$	-0.0117	0.0052	0.0053
	$\hat{\gamma}_{gmm,2}$	-0.3366	18.1341	18.2293
	$(\hat{\beta}_{gmm,2}, \gamma_{gmm,2})$	–	–	18.2346
Ridge-type	$\hat{\beta}_r$	0.0012	0.0008	0.0008
	$\hat{\gamma}_r$	-0.0402	1.7072	1.7071
	$(\hat{\beta}_r, \gamma_r)$	–	–	1.7079
Geodesic	$\hat{\beta}_g$	0.0130	0.0006	0.0008
	$\hat{\gamma}_g$	0.1797	1.0186	1.0499
	$(\hat{\beta}_g, \gamma_g)$	–	–	1.0507
Local Spectral	$\hat{\beta}_s$	0.0184	0.0002	0.0005
	$\hat{\gamma}_s$	0.2911	0.6914	0.7754
	$(\hat{\beta}_s, \gamma_s)$	–	–	**0.7760**

Rank Failure: $T = 100$, $\beta_{prior} = 0.99$, $\gamma_{prior} = 2$

Figure 1.15: Scatterplots of GMM estimates and Regularized estimates

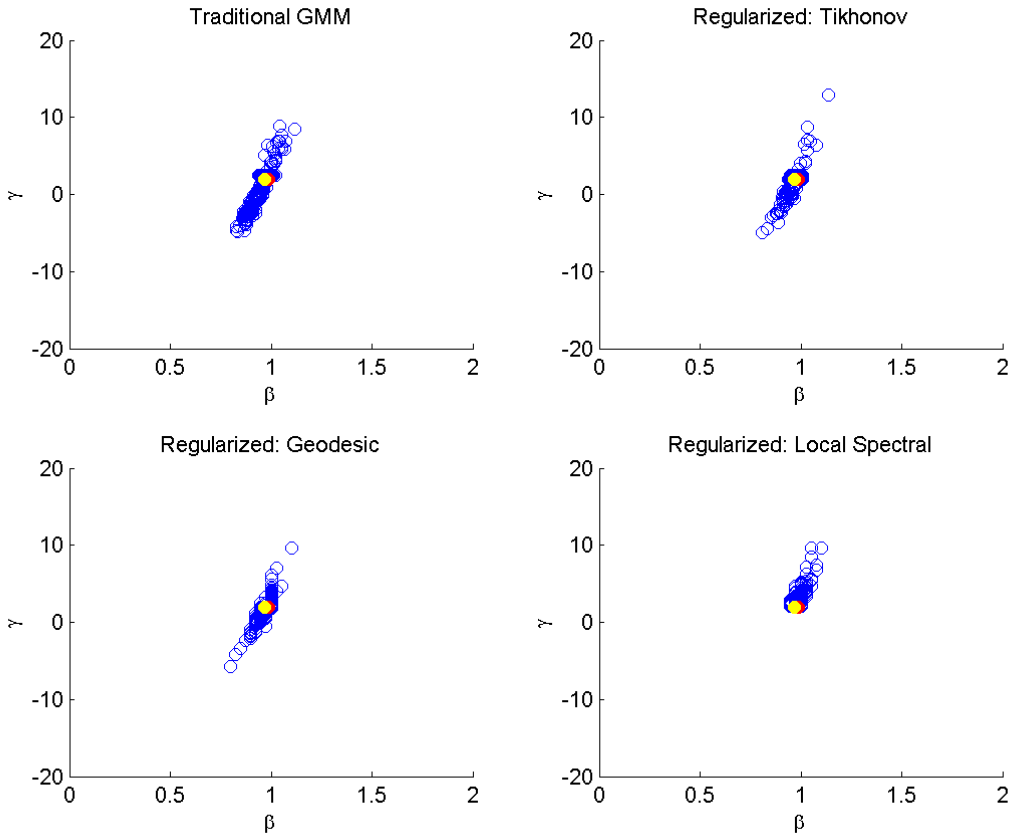


Table 1.7: Bias, Variance, MSE comparison of regularized estimates with GMM estimates

		Bias	Variance	MSE
GMM (Single Step)	$\hat{\beta}_{gmm,1}$	0.0019	0.0008	0.0008
	$\hat{\gamma}_{gmm,1}$	0.1575	1.9674	1.9903
	$(\hat{\beta}_{gmm,1}, \hat{\gamma}_{gmm,1})$	–	–	1.9910
GMM (Two Step)	$\hat{\beta}_{gmm,2}$	-0.0145	0.0041	0.0043
	$\hat{\gamma}_{gmm,2}$	-0.4776	15.7662	15.9786
	$(\hat{\beta}_{gmm,2}, \gamma_{gmm,2})$	–	–	15.9829
Ridge-type	$\hat{\beta}_r$	0.0020	0.0004	0.0004
	$\hat{\gamma}_r$	0.0057	0.7939	0.7931
	$(\hat{\beta}_r, \gamma_r)$	–	–	0.7935
Geodesic	$\hat{\beta}_g$	0.0125	0.0005	0.0007
	$\hat{\gamma}_g$	0.2370	0.7424	0.7978
	$(\hat{\beta}_g, \gamma_g)$	–	–	0.7985
Local Spectral	$\hat{\beta}_s$	0.0180	0.0002	0.0005
	$\hat{\gamma}_s$	0.2752	0.5478	0.6230
	$(\hat{\beta}_s, \gamma_s)$	–	–	**0.6235**

Rank Failure: $T = 200$, $\beta_{prior} = 0.99$, $\gamma_{prior} = 2$

Figure 1.16: Scatterplots of GMM estimates and Regularized estimates

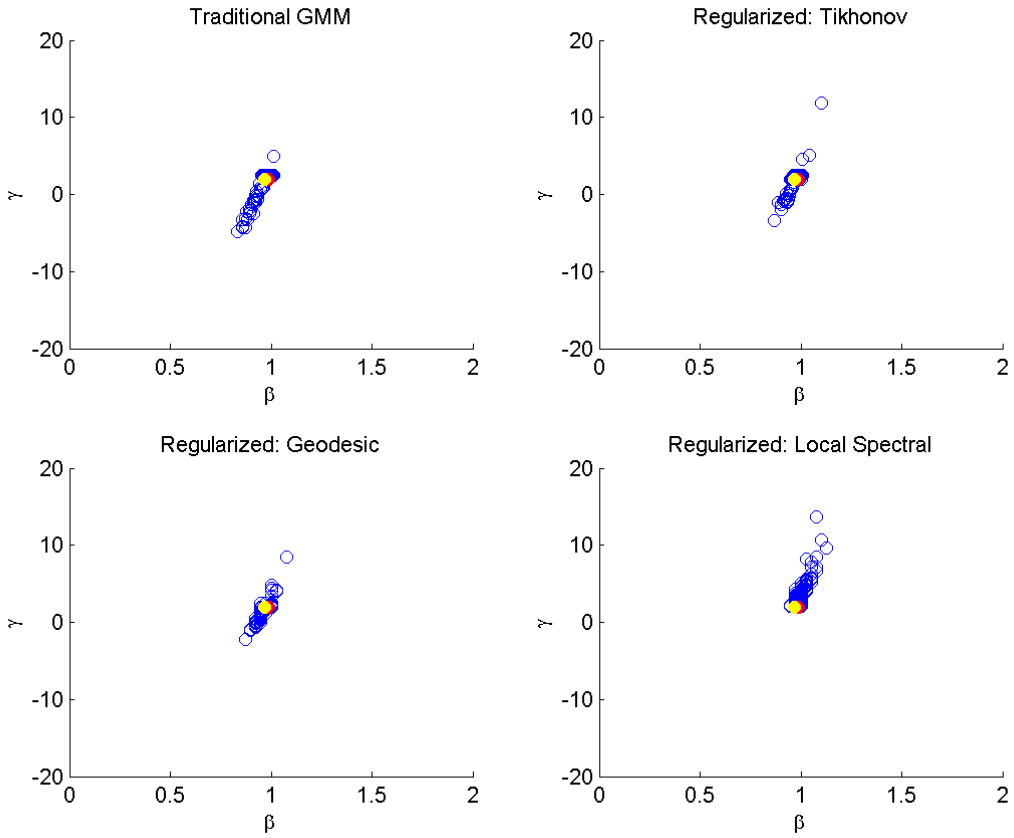


Table 1.8: Bias, Variance, MSE comparison of regularized estimates with GMM estimates

		Bias	Variance	MSE
GMM (Single Step)	$\hat{\beta}_{gmm,1}$	0.0062	0.0002	0.0003
	$\hat{\gamma}_{gmm,1}$	0.3797	0.5464	0.6900
	$(\hat{\beta}_{gmm,1}, \hat{\gamma}_{gmm,1})$	–	–	**0.6903**
GMM (Two Step)	$\hat{\beta}_{gmm,2}$	-0.0145	0.0039	0.0041
	$\hat{\gamma}_{gmm,2}$	-0.4984	13.8784	14.1130
	$(\hat{\beta}_{gmm,2}, \gamma_{gmm,2})$	–	–	14.1170
Ridge-type	$\hat{\beta}_r$	0.0033	0.0002	0.0002
	$\hat{\gamma}_r$	0.0660	0.3167	0.3208
	$(\hat{\beta}_r, \gamma_r)$	–	–	0.3209
Geodesic	$\hat{\beta}_g$	0.0132	0.0003	0.0005
	$\hat{\gamma}_g$	0.3744	0.2691	0.4090
	$(\hat{\beta}_g, \gamma_g)$	–	–	0.4095
Local Spectral	$\hat{\beta}_s$	0.0183	0.0002	0.0005
	$\hat{\gamma}_s$	0.3448	0.7839	0.9020
	$(\hat{\beta}_s, \gamma_s)$	–	–	0.9025

Near Rank Failure: $T = 60$, $\beta_{prior} = 0.99$, $\gamma_{prior} = 2$

Figure 1.17: Scatterplots of GMM estimates and Regularized estimates

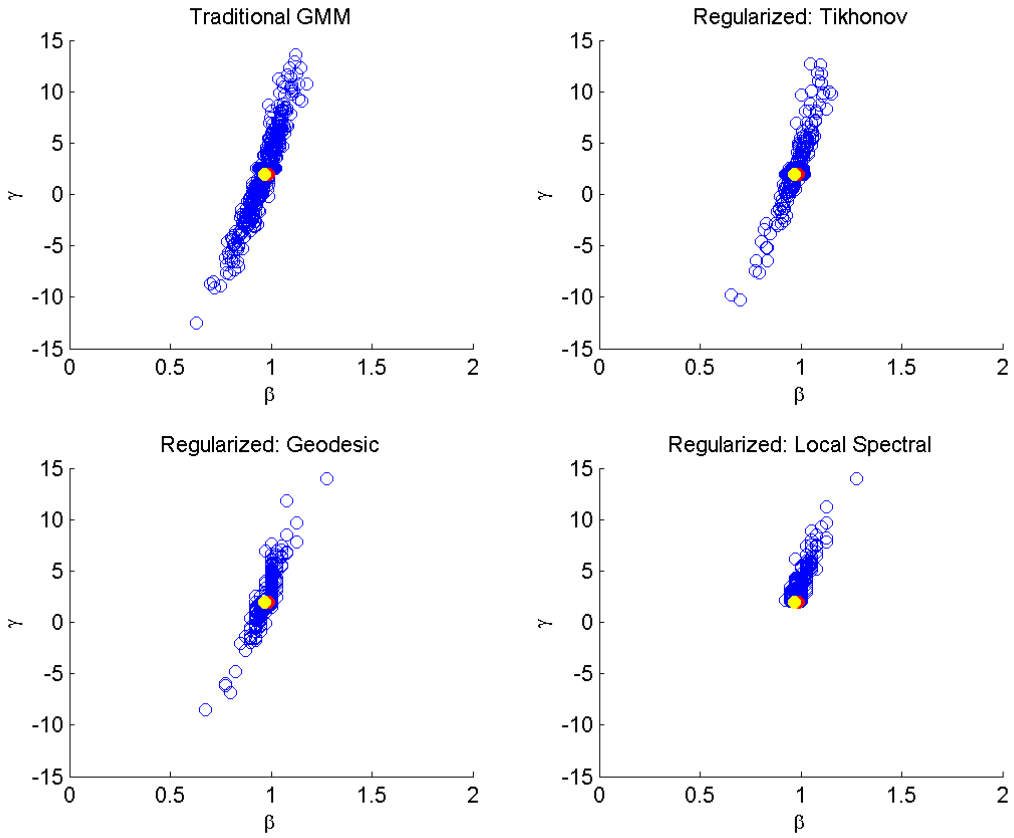


Table 1.9: Bias, Variance, MSE comparison of regularized estimates with GMM estimates

		Bias	Variance	MSE
GMM (Single Step)	$\hat{\beta}_{gmm,1}$	0.0015	0.0030	0.0030
	$\hat{\gamma}_{gmm,1}$	0.3003	8.5979	8.6795
	$(\hat{\beta}_{gmm,1}, \hat{\gamma}_{gmm,1})$	–	–	8.6825
GMM (Two Step)	$\hat{\beta}_{gmm,2}$	-0.0024	0.0041	0.0041
	$\hat{\gamma}_{gmm,2}$	0.1251	12.9954	12.9981
	$(\hat{\beta}_{gmm,2}, \gamma_{gmm,2})$	–	–	13.0022
Ridge-type	$\hat{\beta}_r$	0.0041	0.0013	0.0013
	$\hat{\gamma}_r$	0.2026	3.1318	3.1697
	$(\hat{\beta}_r, \gamma_r)$	–	–	3.1711
Geodesic	$\hat{\beta}_g$	0.0149	0.0010	0.0012
	$\hat{\gamma}_g$	0.3872	1.9674	2.1154
	$(\hat{\beta}_g, \gamma_g)$	–	–	2.1166
Local Spectral	$\hat{\beta}_s$	0.0204	0.0004	0.0008
	$\hat{\gamma}_s$	0.4695	1.2903	1.5095
	$(\hat{\beta}_s, \gamma_s)$	–	–	**1.5103**

Near Rank Failure: $T = 100$, $\beta_{prior} = 0.99$, $\gamma_{prior} = 2$

Figure 1.18: Scatterplots of GMM estimates and Regularized estimates

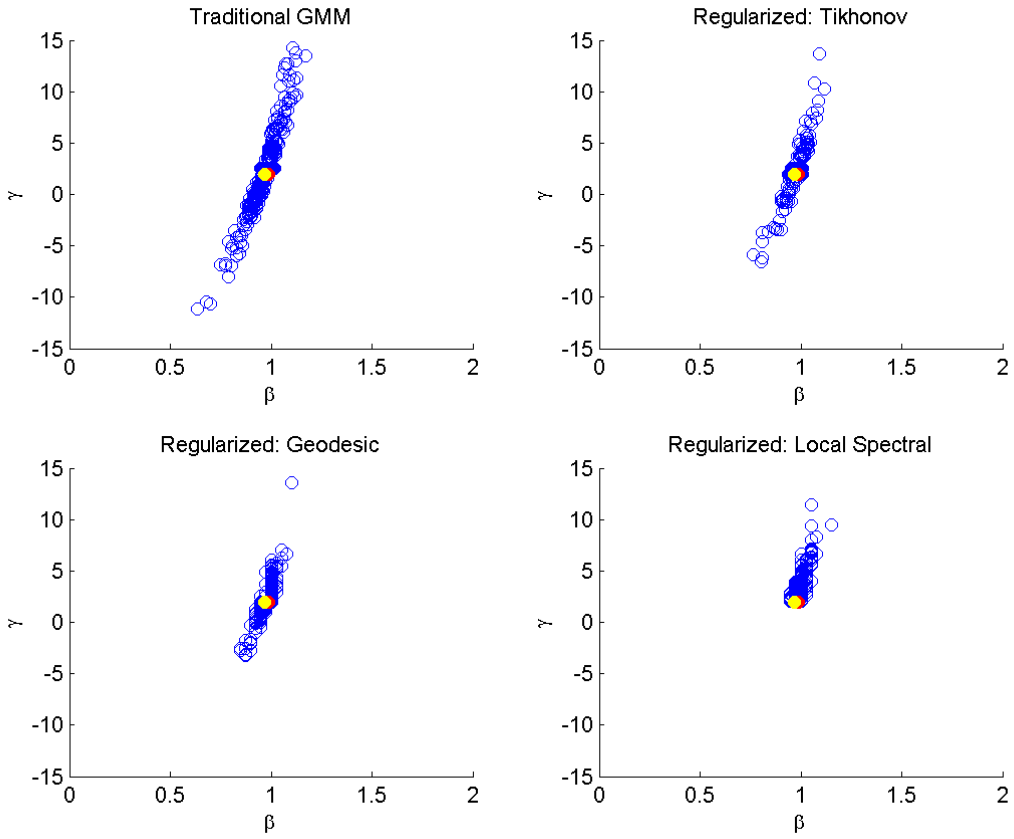


Table 1.10: Bias, Variance, MSE comparison of regularized estimates with GMM estimates

		Bias	Variance	MSE
GMM (Single Step)	$\hat{\beta}_{gmm,1}$	0.0044	0.0017	0.0017
	$\hat{\gamma}_{gmm,1}$	0.4142	4.9965	5.1631
	$(\hat{\beta}_{gmm,1}, \hat{\gamma}_{gmm,1})$	–	–	5.1648
GMM (Two Step)	$\hat{\beta}_{gmm,2}$	-0.0051	0.0036	0.0036
	$\hat{\gamma}_{gmm,2}$	-0.0590	10.0520	10.0454
	$(\hat{\beta}_{gmm,2}, \gamma_{gmm,2})$	–	–	10.0490
Ridge-type	$\hat{\beta}_r$	0.0037	0.0006	0.0006
	$\hat{\gamma}_r$	0.1444	1.3475	1.3670
	$(\hat{\beta}_r, \gamma_r)$	–	–	1.3676
Geodesic	$\hat{\beta}_g$	0.0142	0.0005	0.0007
	$\hat{\gamma}_g$	0.4028	0.9608	1.1221
	$(\hat{\beta}_g, \gamma_g)$	–	–	**1.1228**
Local Spectral	$\hat{\beta}_s$	0.0191	0.0002	0.0006
	$\hat{\gamma}_s$	0.4797	0.9565	1.1857
	$(\hat{\beta}_s, \gamma_s)$	–	–	1.1862

Near Rank Failure: $T = 200$, $\beta_{prior} = 0.99$, $\gamma_{prior} = 2$

Figure 1.19: Scatterplots of GMM estimates and Regularized estimates

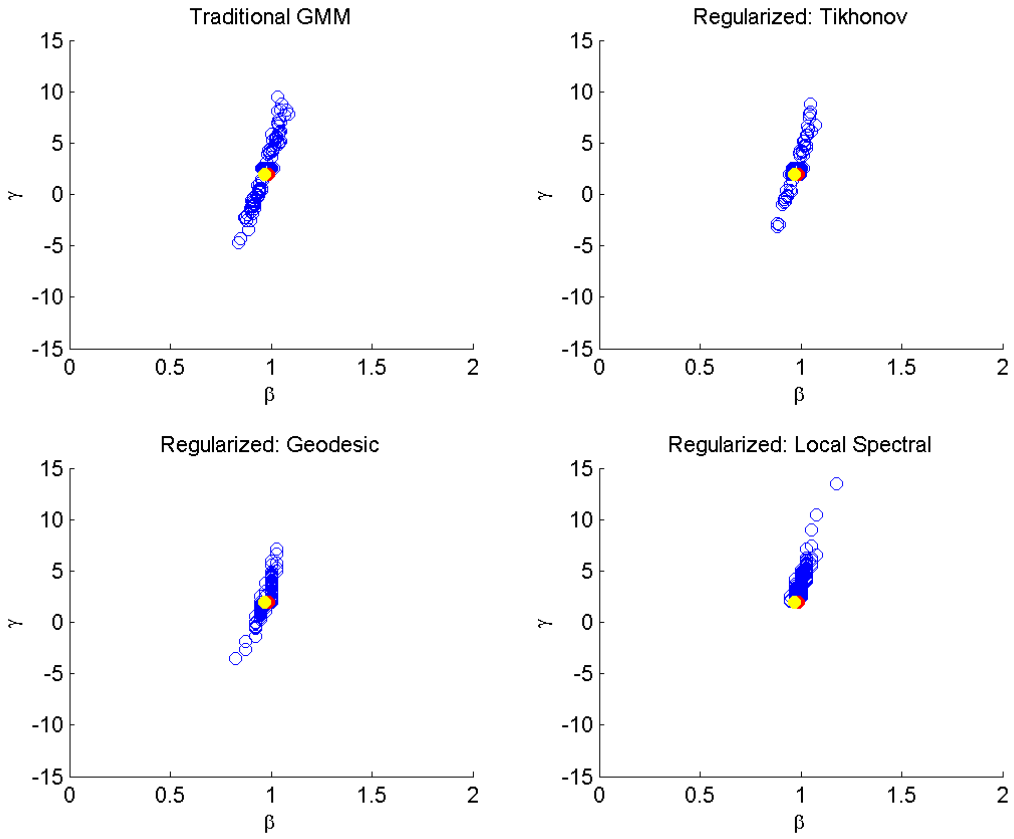


Table 1.11: Bias, Variance, MSE comparison of regularized estimates with GMM estimates

		Bias	Variance	MSE
GMM (Single Step)	$\hat{\beta}_{gmm,1}$	0.0080	0.0004	0.0004
	$\hat{\gamma}_{gmm,1}$	0.4817	1.0493	1.2803
	$(\hat{\beta}_{gmm,1}, \hat{\gamma}_{gmm,1})$	–	–	1.2807
GMM (Two Step)	$\hat{\beta}_{gmm,2}$	-0.0022	0.0017	0.0017
	$\hat{\gamma}_{gmm,2}$	-0.0313	4.6720	4.6683
	$(\hat{\beta}_{gmm,2}, \gamma_{gmm,2})$	–	–	4.6700
Ridge-type	$\hat{\beta}_r$	0.0052	0.0002	0.0002
	$\hat{\gamma}_r$	0.1776	0.4949	0.5259
	$(\hat{\beta}_r, \gamma_r)$	–	–	**0.5261**
Geodesic	$\hat{\beta}_g$	0.0139	0.0003	0.0005
	$\hat{\gamma}_g$	0.4693	0.4261	0.6459
	$(\hat{\beta}_g, \gamma_g)$	–	–	0.6464
Local Spectral	$\hat{\beta}_s$	0.0185	0.0002	0.0005
	$\hat{\gamma}_s$	0.4448	0.8706	1.0677
	$(\hat{\beta}_s, \gamma_s)$	–	–	1.0682

Full Rank: $T = 60$, $\beta_{prior} = 0.99$, $\gamma_{prior} = 2$

Figure 1.20: Scatterplots of GMM estimates and Regularized estimates

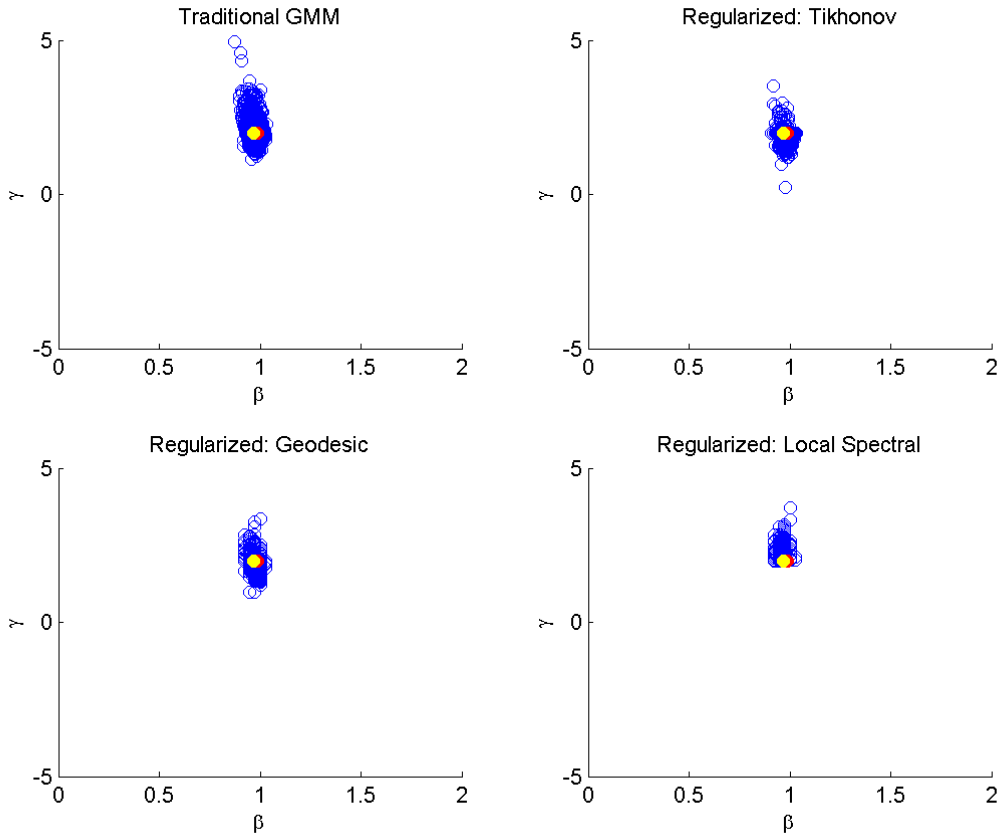


Table 1.12: Bias, Variance, MSE comparison of regularized estimates with GMM estimates

		Bias	Variance	MSE
GMM (Single Step)	$\hat{\beta}_{gmm,1}$	-0.0013	0.0004	0.0004
	$\hat{\gamma}_{gmm,1}$	0.0593	0.1635	0.1669
	$(\hat{\beta}_{gmm,1}, \hat{\gamma}_{gmm,1})$	–	–	0.1673
GMM (Two Step)	$\hat{\beta}_{gmm,2}$	-0.0002	0.0004	0.0004
	$\hat{\gamma}_{gmm,2}$	0.0123	0.1625	0.1625
	$(\hat{\beta}_{gmm,2}, \gamma_{gmm,2})$	–	–	0.1629
Ridge-type	$\hat{\beta}_r$	0.0036	0.0004	0.0004
	$\hat{\gamma}_r$	-0.0650	0.0399	0.0441
	$(\hat{\beta}_r, \gamma_r)$	–	–	0.0445
Geodesic	$\hat{\beta}_g$	0.0132	0.0004	0.0005
	$\hat{\gamma}_g$	-0.0926	0.0532	0.0617
	$(\hat{\beta}_g, \gamma_g)$	–	–	0.0622
Local Spectral	$\hat{\beta}_s$	0.0119	0.0002	0.0004
	$\hat{\gamma}_s$	0.0561	0.0277	0.0309
	$(\hat{\beta}_s, \gamma_s)$	–	–	**0.0313**

Full Rank: $T = 100$, $\beta_{prior} = 0.99$, $\gamma_{prior} = 2$

Figure 1.21: Scatterplots of GMM estimates and Regularized estimates

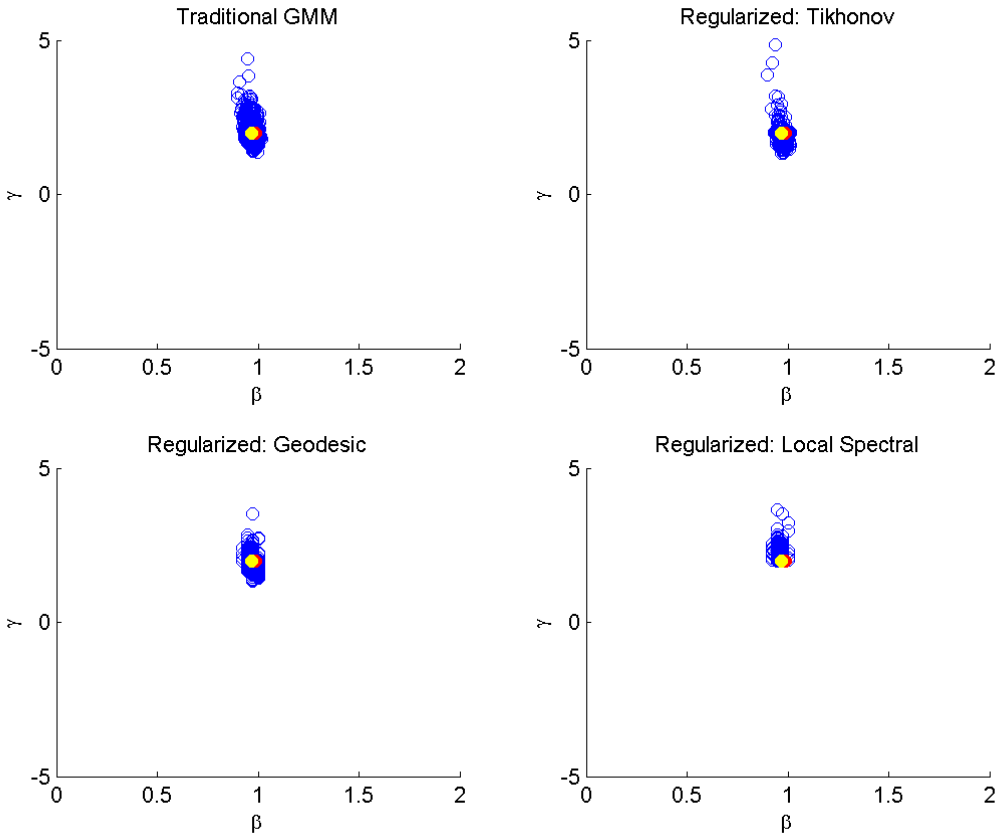


Table 1.13: Bias, Variance, MSE comparison of regularized estimates with GMM estimates

		Bias	Variance	MSE
GMM (Single Step)	$\hat{\beta}_{gmm,1}$	-0.0012	0.0003	0.0003
	$\hat{\gamma}_{gmm,1}$	0.0369	0.0987	0.1000
	$(\hat{\beta}_{gmm,1}, \hat{\gamma}_{gmm,1})$	–	–	0.1003
GMM (Two Step)	$\hat{\beta}_{gmm,2}$	-0.0004	0.0003	0.0003
	$\hat{\gamma}_{gmm,2}$	0.0041	0.0895	0.0895
	$(\hat{\beta}_{gmm,2}, \gamma_{gmm,2})$	–	–	0.0897
Ridge-type	$\hat{\beta}_r$	0.0026	0.0002	0.0002
	$\hat{\gamma}_r$	-0.0340	0.0417	0.0428
	$(\hat{\beta}_r, \gamma_r)$	–	–	0.0431
Geodesic	$\hat{\beta}_g$	0.0142	0.0005	0.0007
	$\hat{\gamma}_g$	0.4028	0.9608	1.1221
	$(\hat{\beta}_g, \gamma_g)$	–	–	0.0408
Local Spectral	$\hat{\beta}_s$	0.0111	0.0003	0.0004
	$\hat{\gamma}_s$	-0.0657	0.0361	0.0404
	$(\hat{\beta}_s, \gamma_s)$	–	–	**0.0280**

Full Rank: $T = 200$, $\beta_{prior} = 0.99$, $\gamma_{prior} = 2$

Figure 1.22: Scatterplots of GMM estimates and Regularized estimates

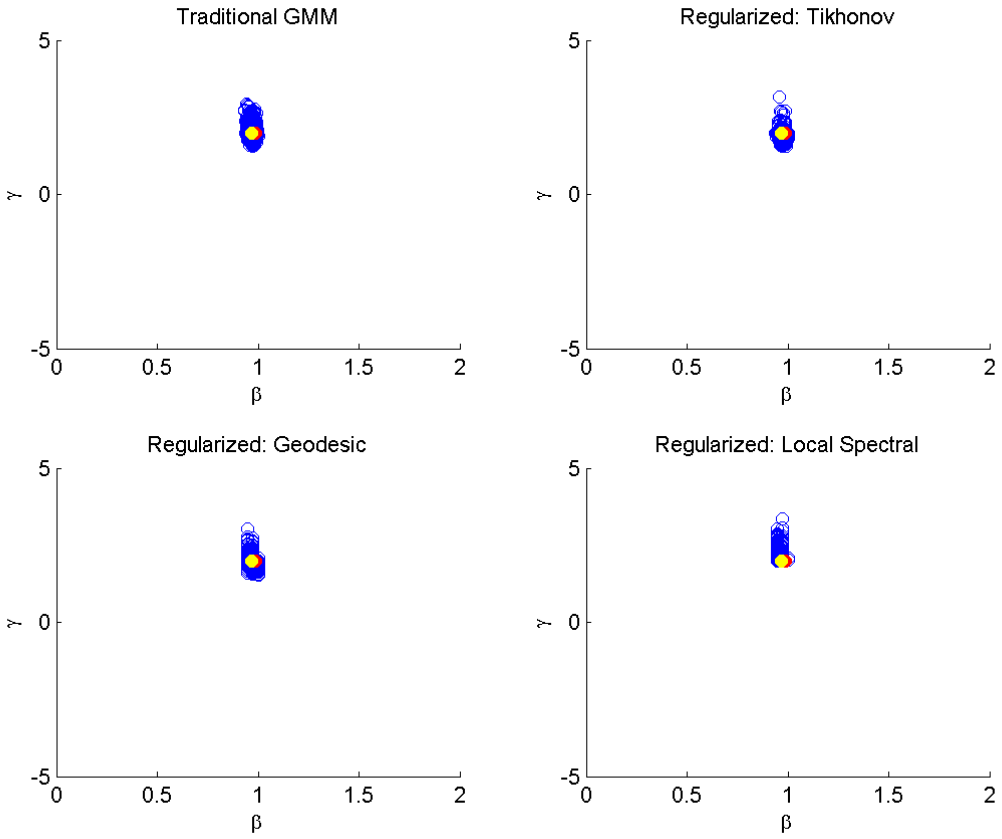


Table 1.14: Bias, Variance, MSE comparison of regularized estimates with GMM estimates

		Bias	Variance	MSE
GMM (Single Step)	$\hat{\beta}_{gmm,1}$	-0.0002	0.0001	0.0001
	$\hat{\gamma}_{gmm,1}$	0.0213	0.0453	0.0457
	$(\hat{\beta}_{gmm,1}, \hat{\gamma}_{gmm,1})$	–	–	0.0458
GMM (Two Step)	$\hat{\beta}_{gmm,2}$	0.0002	0.0001	0.0001
	$\hat{\gamma}_{gmm,2}$	0.0014	0.0405	0.0404
	$(\hat{\beta}_{gmm,2}, \gamma_{gmm,2})$	–	–	0.0405
Ridge-type	$\hat{\beta}_r$	0.0030	0.0001	0.0001
	$\hat{\gamma}_r$	-0.0248	0.0124	0.0130
	$(\hat{\beta}_r, \gamma_r)$	–	–	**0.0132**
Geodesic	$\hat{\beta}_g$	0.0095	0.0002	0.0003
	$\hat{\gamma}_g$	-0.0470	0.0235	0.0257
	$(\hat{\beta}_g, \gamma_g)$	–	–	0.0260
Local Spectral	$\hat{\beta}_s$	0.0099	0.0002	0.0003
	$\hat{\gamma}_s$	0.0684	0.0262	0.0309
	$(\hat{\beta}_s, \gamma_s)$	–	–	0.0312

Rank Failure Estimates for different priors and sample sizes

Table 1.15: MSE Comparison– Model: RF, Sample Size: $T = 60$, Prior: $\gamma = 1, \beta = 0.99$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	−0.001	0.002	0.002
	$\hat{\gamma}_1$	−0.023	5.192	5.187
	$(\hat{\beta}_1, \hat{\gamma}_1)$			5.189
Ridge	$\hat{\beta}_r$	−0.008	0.001	0.001
	$\hat{\gamma}_r$	−0.656	1.730	2.158
	$(\hat{\beta}_r, \gamma_r)$			2.159
Geodesic	$\hat{\beta}_g$	0.012	0.001	0.001
	$\hat{\gamma}_g$	−0.024	1.177	1.176
	$(\hat{\beta}_g, \gamma_g)$			**1.177**
Spectral	$\hat{\beta}_s$	0.011	0.000	0.001
	$\hat{\gamma}_s$	−0.519	1.077	1.345
	$(\hat{\beta}_s, \gamma_s)$			1.346

Table 1.16: MSE Comparison– Model: RF, Sample Size: $T = 100$, Prior: $\gamma = 1, \beta = 0.99$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.002	0.001	0.001
	$\hat{\gamma}_1$	0.158	1.967	1.990
	$(\hat{\beta}_1, \hat{\gamma}_1)$			1.991
Ridge	$\hat{\beta}_r$	−0.008	0.000	0.000
	$\hat{\gamma}_r$	−0.638	0.925	1.331
	$(\hat{\beta}_r, \gamma_r)$			1.332
Geodesic	$\hat{\beta}_g$	0.011	0.001	0.001
	$\hat{\gamma}_g$	0.101	0.964	0.973
	$(\hat{\beta}_g, \gamma_g)$			**0.974**
Spectral	$\hat{\beta}_s$	0.011	0.000	0.001
	$\hat{\gamma}_s$	−0.521	1.014	1.284
	$(\hat{\beta}_s, \gamma_s)$			1.285

Table 1.17: MSE Comparison– Model: RF, Sample Size: $T = 60$, Prior: $\gamma = 1, \beta = 0.95$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	−0.001	0.002	0.002
	$\hat{\gamma}_1$	−0.023	5.192	5.187
	$(\hat{\beta}_1, \hat{\gamma}_1)$			5.189
Ridge	$\hat{\beta}_r$	−0.011	0.001	0.001
	$\hat{\gamma}_r$	−0.590	2.145	2.490
	$(\hat{\beta}_r, \gamma_r)$			2.491
Geodesic	$\hat{\beta}_g$	−0.011	0.001	0.001
	$\hat{\gamma}_g$	−0.616	1.081	1.459
	$(\hat{\beta}_g, \gamma_g)$			1.460
Spectral	$\hat{\beta}_s$	−0.011	0.000	0.001
	$\hat{\gamma}_s$	−0.549	1.056	1.357
	$(\hat{\beta}_s, \gamma_s)$			**1.357**

Table 1.18: MSE Comparison– Model: RF, Sample Size: $T = 100$, Prior: $\gamma = 1, \beta = 0.95$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.002	0.001	0.001
	$\hat{\gamma}_1$	0.158	1.967	1.990
	$(\hat{\beta}_1, \hat{\gamma}_1)$			1.991
Ridge	$\hat{\beta}_r$	−0.010	0.001	0.001
	$\hat{\gamma}_r$	−0.539	1.118	1.407
	$(\hat{\beta}_r, \gamma_r)$			1.407
Geodesic	$\hat{\beta}_g$	−0.011	0.000	0.001
	$\hat{\gamma}_g$	−0.582	0.781	1.119
	$(\hat{\beta}_g, \gamma_g)$			**1.120**
Spectral	$\hat{\beta}_s$	−0.011	0.000	0.001
	$\hat{\gamma}_s$	−0.530	1.036	1.316
	$(\hat{\beta}_s, \gamma_s)$			1.316

Rank Failure Estimates for different priors and sample sizes

Table 1.19: MSE Comparison– Model: RF, Sample Size: $T = 200$, Prior: $\gamma = 1, \beta = 0.99$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.006	0.000	0.000
	$\hat{\gamma}_1$	0.380	0.546	0.690
	$(\hat{\beta}_1, \hat{\gamma}_1)$			0.690
Ridge	$\hat{\beta}_r$	−0.008	0.000	0.000
	$\hat{\gamma}_r$	−0.621	0.600	0.984
	$(\hat{\beta}_r, \gamma_r)$			0.985
Geodesic	$\hat{\beta}_g$	0.013	0.000	0.001
	$\hat{\gamma}_g$	0.322	0.389	0.492
	$(\hat{\beta}_g, \gamma_g)$			**0.493**
Spectral	$\hat{\beta}_s$	0.011	0.000	0.001
	$\hat{\gamma}_s$	−0.457	1.250	1.458
	$(\hat{\beta}_s, \gamma_s)$			1.458

Table 1.20: MSE Comparison– Model: RF, Sample Size: $T = 60$, Prior: $\gamma = 6, \beta = 0.99$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	−0.001	0.002	0.002
	$\hat{\gamma}_1$	−0.023	5.192	5.187
	$(\hat{\beta}_1, \hat{\gamma}_1)$			**5.189**
Ridge	$\hat{\beta}_r$	0.021	0.002	0.003
	$\hat{\gamma}_r$	1.580	5.976	8.466
	$(\hat{\beta}_r, \gamma_r)$			8.468
Geodesic	$\hat{\beta}_g$	0.017	0.001	0.001
	$\hat{\gamma}_g$	1.505	3.222	5.485
	$(\hat{\beta}_g, \gamma_g)$			5.486
Spectral	$\hat{\beta}_s$	0.024	0.000	0.001
	$\hat{\gamma}_s$	4.062	0.253	16.756
	$(\hat{\beta}_s, \gamma_s)$			16.757

Table 1.21: MSE Comparison– Model: RF, Sample Size: $T = 200$, Prior: $\gamma = 1, \beta = 0.95$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.006	0.000	0.000
	$\hat{\gamma}_1$	0.380	0.546	0.690
	$(\hat{\beta}_1, \hat{\gamma}_1)$			0.690
Ridge	$\hat{\beta}_r$	−0.009	0.000	0.000
	$\hat{\gamma}_r$	−0.483	0.731	0.964
	$(\hat{\beta}_r, \gamma_r)$			0.964
Geodesic	$\hat{\beta}_g$	−0.007	0.000	0.000
	$\hat{\gamma}_g$	−0.415	0.426	0.598
	$(\hat{\beta}_g, \gamma_g)$			**0.598**
Spectral	$\hat{\beta}_s$	−0.010	0.000	0.001
	$\hat{\gamma}_s$	−0.469	1.196	1.415
	$(\hat{\beta}_s, \gamma_s)$			1.416

Table 1.22: MSE Comparison– Model: RF, Sample Size: $T = 60$, Prior: $\gamma = 0, \beta = 0$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	−0.001	0.002	0.002
	$\hat{\gamma}_1$	−0.023	5.192	5.187
	$(\hat{\beta}_1, \hat{\gamma}_1)$			5.189
Ridge	$\hat{\beta}_r$	−0.038	0.001	0.003
	$\hat{\gamma}_r$	−1.434	2.075	4.128
	$(\hat{\beta}_r, \gamma_r)$			4.131
Geodesic	$\hat{\beta}_g$	−0.018	0.003	0.003
	$\hat{\gamma}_g$	0.009	7.564	7.556
	$(\hat{\beta}_g, \gamma_g)$			7.560
Spectral	$\hat{\beta}_s$	−0.353	0.211	0.336
	$\hat{\gamma}_s$	−1.152	1.819	3.145
	$(\hat{\beta}_s, \gamma_s)$			**3.480**

Rank Failure Estimates for different priors and sample sizes

Table 1.23: MSE Comparison– Model: RF, Sample Size: $T = 100$, Prior: $\gamma = 6, \beta = 0.99$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.002	0.001	0.001
	$\hat{\gamma}_1$	0.158	1.967	1.990
	$(\hat{\beta}_1, \hat{\gamma}_1)$			**1.991**
Ridge	$\hat{\beta}_r$	0.021	0.001	0.002
	$\hat{\gamma}_r$	1.556	4.565	6.981
	$(\hat{\beta}_r, \gamma_r)$			6.983
Geodesic	$\hat{\beta}_g$	0.017	0.001	0.001
	$\hat{\gamma}_g$	1.472	2.481	4.645
	$(\hat{\beta}_g, \gamma_g)$			4.646
Spectral	$\hat{\beta}_s$	0.024	0.000	0.001
	$\hat{\gamma}_s$	4.063	0.197	16.708
	$(\hat{\beta}_s, \gamma_s)$			16.709

Table 1.24: MSE Comparison– Model: RF, Sample Size: $T = 200$, Prior: $\gamma = 6, \beta = 0.99$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.006	0.000	0.000
	$\hat{\gamma}_1$	0.380	0.546	0.690
	$(\hat{\beta}_1, \hat{\gamma}_1)$			**0.690**
Ridge	$\hat{\beta}_g$	0.025	0.001	0.002
	$\hat{\gamma}_g$	1.809	3.620	6.890
	$(\hat{\beta}_g, \gamma_g)$			6.891
Geodesic	$\hat{\beta}_r$	0.021	0.000	0.001
	$\hat{\gamma}_r$	1.521	1.404	3.715
	$(\hat{\beta}_r, \gamma_r)$			3.715
Spectral	$\hat{\beta}_s$	0.026	0.000	0.001
	$\hat{\gamma}_s$	4.086	0.192	16.884
	$(\hat{\beta}_s, \gamma_s)$			16.885

Table 1.25: MSE Comparison– Model: RF, Sample Size: $T = 100$, Prior: $\gamma = 0, \beta = 0$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.002	0.001	0.001
	$\hat{\gamma}_1$	0.158	1.967	1.990
	$(\hat{\beta}_1, \hat{\gamma}_1)$			**1.991**
Ridge	$\hat{\beta}_r$	−0.034	0.001	0.002
	$\hat{\gamma}_r$	−1.387	1.690	3.613
	$(\hat{\beta}_r, \gamma_r)$			3.615
Geodesic	$\hat{\beta}_g$	−0.012	0.002	0.002
	$\hat{\gamma}_g$	0.035	3.562	3.560
	$(\hat{\beta}_g, \gamma_g)$			3.561
Spectral	$\hat{\beta}_s$	−0.370	0.216	0.352
	$\hat{\gamma}_s$	−1.188	1.788	3.198
	$(\hat{\beta}_s, \gamma_s)$			3.551

Table 1.26: MSE Comparison– Model: RF, Sample Size: $T = 200$, Prior: $\gamma = 0, \beta = 0$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.006	0.000	0.000
	$\hat{\gamma}_1$	0.380	0.546	0.690
	$(\hat{\beta}_1, \hat{\gamma}_1)$			**0.690**
Ridge	$\hat{\beta}_r$	−0.031	0.001	0.002
	$\hat{\gamma}_r$	−1.324	1.445	3.197
	$(\hat{\beta}_r, \gamma_r)$			3.199
Geodesic	$\hat{\beta}_g$	−0.003	0.001	0.001
	$\hat{\gamma}_g$	0.281	1.113	1.191
	$(\hat{\beta}_g, \gamma_g)$			1.192
Spectral	$\hat{\beta}_s$	−0.341	0.208	0.324
	$\hat{\gamma}_s$	−1.129	2.042	3.315
	$(\hat{\beta}_s, \gamma_s)$			3.639

Near Rank Failure Estimates for different priors and sample sizes

Table 1.27: MSE Comparison– Model: NRF, Sample Size: $T = 60$, Prior: $\gamma = 1, \beta = 0.99$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.001	0.003	0.003
	$\hat{\gamma}_1$	0.207	8.023	8.058
	$(\hat{\beta}_1, \hat{\gamma}_1)$			8.061
Ridge	$\hat{\beta}_r$	−0.006	0.001	0.001
	$\hat{\gamma}_r$	−0.457	2.307	2.514
	$(\hat{\beta}_r, \gamma_r)$			2.515
Geodesic	$\hat{\beta}_g$	0.013	0.001	0.001
	$\hat{\gamma}_g$	0.122	1.743	1.756
	$(\hat{\beta}_g, \gamma_g)$			**1.757**
Spectral	$\hat{\beta}_s$	0.011	0.001	0.001
	$\hat{\gamma}_s$	−0.330	1.702	1.809
	$(\hat{\beta}_s, \gamma_s)$			1.810

Table 1.28: MSE Comparison– Model: NRF, Sample Size: $T = 100$, Prior: $\gamma = 1, \beta = 0.99$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.005	0.001	0.001
	$\hat{\gamma}_1$	0.360	3.869	3.995
	$(\hat{\beta}_1, \hat{\gamma}_1)$			3.996
Ridge	$\hat{\beta}_r$	−0.007	0.001	0.001
	$\hat{\gamma}_r$	−0.509	1.637	1.895
	$(\hat{\beta}_r, \gamma_r)$			1.896
Geodesic	$\hat{\beta}_g$	0.014	0.001	0.001
	$\hat{\gamma}_g$	0.226	1.179	1.228
	$(\hat{\beta}_g, \gamma_g)$			**1.229**
Spectral	$\hat{\beta}_s$	0.011	0.000	0.001
	$\hat{\gamma}_s$	−0.310	1.439	1.534
	$(\hat{\beta}_s, \gamma_s)$			1.534

Table 1.29: MSE Comparison– Model: NRF, Sample Size: $T = 60$, Prior: $\gamma = 1, \beta = 0.95$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.001	0.003	0.003
	$\hat{\gamma}_1$	0.207	8.023	8.058
	$(\hat{\beta}_1, \hat{\gamma}_1)$			8.061
Ridge	$\hat{\beta}_r$	−0.009	0.001	0.001
	$\hat{\gamma}_r$	−0.401	2.681	2.840
	$(\hat{\beta}_r, \gamma_r)$			2.841
Geodesic	$\hat{\beta}_g$	−0.010	0.001	0.001
	$\hat{\gamma}_g$	−0.474	1.579	1.802
	$(\hat{\beta}_g, \gamma_g)$			**1.802**
Spectral	$\hat{\beta}_s$	−0.008	0.001	0.001
	$\hat{\gamma}_s$	−0.323	1.807	1.909
	$(\hat{\beta}_s, \gamma_s)$			1.910

Table 1.30: MSE Comparison– Model: NRF, Sample Size: $T = 100$, Prior: $\gamma = 1, \beta = 0.95$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.005	0.001	0.001
	$\hat{\gamma}_1$	0.360	3.869	3.995
	$(\hat{\beta}_1, \hat{\gamma}_1)$			3.996
Ridge	$\hat{\beta}_r$	−0.008	0.001	0.001
	$\hat{\gamma}_r$	−0.404	1.769	1.930
	$(\hat{\beta}_r, \gamma_r)$			1.931
Geodesic	$\hat{\beta}_g$	−0.008	0.001	0.001
	$\hat{\gamma}_g$	−0.410	1.147	1.315
	$(\hat{\beta}_g, \gamma_g)$			**1.315**
Spectral	$\hat{\beta}_s$	−0.007	0.001	0.001
	$\hat{\gamma}_s$	−0.334	1.490	1.600
	$(\hat{\beta}_s, \gamma_s)$			1.601

Near Rank Failure Estimates for different priors and sample sizes

Table 1.31: MSE Comparison– Model: NRF, Sample Size: $T = 200$, Prior: $\gamma = 1, \beta = 0.99$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.008	0.000	0.000
	$\hat{\gamma}_1$	0.446	0.935	1.134
	$(\hat{\beta}_1, \hat{\gamma}_1)$			1.134
Ridge	$\hat{\beta}_r$	−0.006	0.000	0.000
	$\hat{\gamma}_r$	−0.490	0.970	1.209
	$(\hat{\beta}_r, \gamma_r)$			1.209
Geodesic	$\hat{\beta}_g$	0.014	0.000	0.001
	$\hat{\gamma}_g$	0.368	0.476	0.611
	$(\hat{\beta}_g, \gamma_g)$			**0.612**
Spectral	$\hat{\beta}_s$	0.011	0.001	0.001
	$\hat{\gamma}_s$	−0.185	1.559	1.592
	$(\hat{\beta}_s, \gamma_s)$			1.592

Table 1.32: MSE Comparison– Model: NRF, Sample Size: $T = 200$, Prior: $\gamma = 1, \beta = 0.95$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	0.008	0.000	0.000
	$\hat{\gamma}_1$	0.446	0.935	1.134
	$(\hat{\beta}_1, \hat{\gamma}_1)$			1.134
Ridge	$\hat{\beta}_r$	−0.006	0.000	0.000
	$\hat{\gamma}_r$	−0.281	1.073	1.151
	$(\hat{\beta}_r, \gamma_r)$			1.151
Geodesic	$\hat{\beta}_g$	−0.006	0.000	0.000
	$\hat{\gamma}_g$	−0.335	0.536	0.648
	$(\hat{\beta}_g, \gamma_g)$			**0.648**
Spectral	$\hat{\beta}_s$	−0.004	0.001	0.001
	$\hat{\gamma}_s$	−0.199	1.480	1.518
	$(\hat{\beta}_s, \gamma_s)$			1.519

Full Rank Estimates for different priors and sample sizes

Table 1.33: MSE Comparison– Model: FR, Sample Size: $T = 60$, Prior: $\gamma = 1, \beta = 0.99$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	−0.001	0.001	0.001
	$\hat{\gamma}_1$	0.078	0.217	0.223
	$(\hat{\beta}_1, \hat{\gamma}_1)$			**0.223**
Ridge	$\hat{\beta}_r$	0.010	0.000	0.000
	$\hat{\gamma}_r$	−0.774	0.229	0.828
	$(\hat{\beta}_r, \gamma_r)$			0.828
Geodesic	$\hat{\beta}_g$	0.015	0.000	0.001
	$\hat{\gamma}_g$	−0.714	0.182	0.692
	$(\hat{\beta}_g, \gamma_g)$			0.692
Spectral	$\hat{\beta}_s$	0.011	0.000	0.000
	$\hat{\gamma}_s$	−0.734	0.200	0.738
	$(\hat{\beta}_s, \gamma_s)$			0.738

Table 1.34: MSE Comparison– Model: FR, Sample Size: $T = 60$, Prior: $\gamma = 1, \beta = 0.95$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	−0.001	0.001	0.001
	$\hat{\gamma}_1$	0.078	0.217	0.223
	$(\hat{\beta}_1, \hat{\gamma}_1)$			**0.223**
Ridge	$\hat{\beta}_r$	0.006	0.000	0.000
	$\hat{\gamma}_r$	−0.819	0.190	0.861
	$(\hat{\beta}_r, \gamma_r)$			0.861
Geodesic	$\hat{\beta}_g$	−0.009	0.000	0.000
	$\hat{\gamma}_g$	−0.386	0.317	0.466
	$(\hat{\beta}_g, \gamma_g)$			0.466
Spectral	$\hat{\beta}_s$	−0.001	0.000	0.000
	$\hat{\gamma}_s$	−0.646	0.247	0.665
	$(\hat{\beta}_s, \gamma_s)$			0.665

Full Rank Estimates for different priors and sample sizes

Table 1.35: MSE Comparison– Model: FR, Sample Size: $T = 100$, Prior: $\gamma = 1, \beta = 0.99$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	−0.001	0.000	0.000
	$\hat{\gamma}_1$	0.057	0.103	0.106
	$(\hat{\beta}_1, \hat{\gamma}_1)$			**0.107**
Ridge	$\hat{\beta}_r$	0.011	0.000	0.000
	$\hat{\gamma}_r$	−0.813	0.170	0.832
	$(\hat{\beta}_r, \gamma_r)$			0.832
Geodesic	$\hat{\beta}_g$	0.014	0.000	0.000
	$\hat{\gamma}_g$	−0.710	0.150	0.653
	$(\hat{\beta}_g, \gamma_g)$			0.654
Spectral	$\hat{\beta}_s$	0.010	0.000	0.000
	$\hat{\gamma}_s$	−0.726	0.175	0.702
	$(\hat{\beta}_s, \gamma_s)$			0.702

Table 1.36: MSE Comparison– Model: FR, Sample Size: $T = 200$, Prior: $\gamma = 1, \beta = 0.99$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	−0.001	0.000	0.000
	$\hat{\gamma}_1$	0.032	0.046	0.047
	$(\hat{\beta}_1, \hat{\gamma}_1)$			**0.047**
Ridge	$\hat{\beta}_r$	0.010	0.000	0.000
	$\hat{\gamma}_r$	−0.786	0.186	0.804
	$(\hat{\beta}_r, \gamma_r)$			0.804
Geodesic	$\hat{\beta}_g$	0.011	0.000	0.000
	$\hat{\gamma}_g$	−0.593	0.183	0.535
	$(\hat{\beta}_g, \gamma_g)$			0.536
Spectral	$\hat{\beta}_s$	0.009	0.000	0.000
	$\hat{\gamma}_s$	−0.636	0.210	0.614
	$(\hat{\beta}_s, \gamma_s)$			0.614

Table 1.37: MSE Comparison– Model: FR, Sample Size: $T = 100$, Prior: $\gamma = 1, \beta = 0.95$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	−0.001	0.000	0.000
	$\hat{\gamma}_1$	0.057	0.103	0.106
	$(\hat{\beta}_1, \hat{\gamma}_1)$			**0.107**
Ridge	$\hat{\beta}_r$	0.006	0.000	0.000
	$\hat{\gamma}_r$	−0.848	0.140	0.858
	$(\hat{\beta}_r, \gamma_r)$			0.858
Geodesic	$\hat{\beta}_g$	−0.009	0.000	0.000
	$\hat{\gamma}_g$	−0.249	0.228	0.290
	$(\hat{\beta}_g, \gamma_g)$			0.290
Spectral	$\hat{\beta}_s$	0.001	0.000	0.000
	$\hat{\gamma}_s$	−0.628	0.218	0.612
	$(\hat{\beta}_s, \gamma_s)$			0.612

Table 1.38: MSE Comparison– Model: FR, Sample Size: $T = 200$, Prior: $\gamma = 1, \beta = 0.95$

		Bias	*Var*	*MSE*
GMM 1	$\hat{\beta}_1$	−0.001	0.000	0.000
	$\hat{\gamma}_1$	0.032	0.046	0.047
	$(\hat{\beta}_1, \hat{\gamma}_1)$			**0.047**
Ridge	$\hat{\beta}_r$	0.006	0.000	0.000
	$\hat{\gamma}_r$	−0.815	0.167	0.831
	$(\hat{\beta}_r, \gamma_r)$			0.832
Geodesic	$\hat{\beta}_g$	−0.008	0.000	0.000
	$\hat{\gamma}_g$	−0.081	0.122	0.129
	$(\hat{\beta}_g, \gamma_g)$			0.129
Spectral	$\hat{\beta}_s$	0.002	0.000	0.000
	$\hat{\gamma}_s$	−0.472	0.219	0.442
	$(\hat{\beta}_s, \gamma_s)$			0.442

1.6.4 The Long Run Risk Model

The simulations in this section were based on the standard Consumption based Capital Asset Pricing model (CCAPM). While this is the standard starting point for asset prices, most empirical studies that assessed the empirical validity of CCAPM reject the model. In particular reasonable values for the parameters are unable to explain levels of risk premia, consumption growth rates and asset returns observed in practice.

One of the most influential recent papers in the literature is by Bansal and Yaron's (2004) Long Run Risk Model for Asset prices. In this model there exists a representative agent with Epstein and Zin (1989) type recursive preferences who maximizes her expeted lifetime utility,

$$V_t = \left[(1-\beta)C_t^{\frac{1-\gamma}{\theta}} + \beta\left(E_t[V_{t+1}^{1-\gamma}]\right)^{\frac{1}{\theta}}\right]^{\frac{\theta}{1-\gamma}}$$

subject to the budget constraint,

$$W_{t+1} = (W_t - C_t)R_{c,t+1}$$

where C_t is consumption at time t, W_t is the wealth of the agent and $R_{c,t}$ is the return on all invested wealth. β and γ are respectively the discount rate and CRRA as described before. $\theta = \frac{1-\gamma}{1-\frac{1}{\Psi}}$ and Ψ is the elasticity of intertemporal substitution.

The following joint dynamics describe consumption and dividends

$$\begin{aligned}
\Delta c_{t+1} &= \mu_c + x_t + \sigma_t \varepsilon_{c,t+1} \\
x_{t+1} &= \rho x_t + \sigma_{x,c}\sigma_t \varepsilon_{x,t+1} \\
\sigma^2_{t+1} &= \bar{\sigma}^2 + \mu(\sigma_t^2 - \bar{\sigma}^2) + \sigma_w w_{t+1} \\
\Delta d_{t+1} &= \mu_d + \phi_x x_t + \sigma_d \sigma_c \varepsilon_{d,t+1}
\end{aligned}$$

where Δc_{t+1} and Δd_{t+1} are the consumption and dividend growth rates respectively. Some of the important implications of these dynamica are:

1. the presence of a small but persistent component in the consumption growth equation, x_t which is referred to as the long run risk.

2. $\varepsilon_{c,t+1}$ represents the i.i.d. innovation which is referred to as short run risk

3. The conditional mean of dividend growth is proportional to the conditional mean of consumption growth.

The asset pricing Euler condition for asset i from the Epstein and Zin (1989) recursive preferences model is

$$E_t\left[\exp(m_{t+1} + r_{j,t+1})\right] = 1$$

where $r_{j,t+1}$ is the return on asset j and m_{t+1} is the log of the intertemporal marginal rate of substitution which is given by

$$m_{t+1} = \theta \log \beta - \frac{\theta}{\Psi}\Delta c_{t+1} + (\theta - 1) r_{c,t+1}$$

where $r_{c,t+1}$ is the continuous return on the consumption asset. Thus the relevant moment condition from this model is

$$E_t\left[\exp(\theta \log \beta - \frac{\theta}{\Psi}\Delta c_{t+1} + (\theta - 1) r_{c,t+1} + r_{j,t+1})\right] - 1 = 0.$$

Such Long Run Risk (LRR) models have important implications on asset prices and show promise in explaining the time series and cross-sectional properties of financial assets which traditional CCAPM have been unsuccessful in doing. Therefore any econometric exercise on asset prices should have an eye on ways to improve estimation of the LRR models. That being said, such an econometric exercise is out of the scope of this thesis for the following reasons:

1. Econometric estimation of LRR models is still in a somewhat nascent stage.

2. The empirical plausibilty for assessing these LRR models is hampered by the presence of many unobservable state variables.

3. The implied moment condition has $k = 3$ parameters to be estimated (θ, β, Ψ) and the practical scope of the current project is limited to models with $k = 2$ parameters.

We hope that we are able to extend our work to LRR models in the future as the econometrics for estimating LRR becomes more standard and as we improve techniques for GMM regularization in higher dimensions.

1.7 Notes on Extensions in Higher Dimensions

So far we have presented three possible algorithms for regularization in the GMM framework when the dimension of the parameter space $k = 2$. In this section we briefly discuss whether these techniques can be extending to higher dimensional parameter spaces. We investigate the possibility of extending each of the three techniques to a particular class of identification problems where $\text{rank}(M(\widehat{\theta})) \to k - 1$. In other words we restrict our attention to those cases where only one of the k dimensions is poorly identified.

Ridge-type solution path

Recall that the GMM objective function for this form of regularization requires augmenting the usual GMM objective with a term that penalizes the objective for moving away from the origin:

$$Q(\theta, \alpha) = G(\theta)' \cdot W \cdot G(\theta) + \alpha \cdot (\theta - \theta_{prior})'(\theta - \theta_{prior}).$$

where α is known as a *tuning parameter*. The solutions corresponding to different values of α are characterized by:

$$\widehat{\theta}_\alpha = \min_\theta Q(\theta, \alpha)$$

Extending this type of solution to $k > 2$ dimensions is straightforward and computationally feasible. The penalty term $\alpha \cdot (\theta - \theta_{prior})'(\theta - \theta_{prior})$ is simply the scalar product of the parameter vector θ multiplied with the tuning parameter α. In fact, ridge regularization

in linear regression is especially used in cases where the parameter vector is of very high dimension (including cases where $n > k$).

However, as we noted in the discussion in Section 1.3.2 and Section 1.4.3 as well as in the simulation results in Section 1.5 the ridge-type penalty may *not* be the best choice of regularization technique in the GMM framework because it does not consider the possible non-linearities of the GMM objective.

Next we consider extensions of the two techniques that take account of the non-linear geometry of the GMM objective.

Geodesic solution path

Recall that a geodesic is defined as the shortest path between two points along a curved surface. In order to find a regularized GMM parameter estimate, we suggested choosing from points on the geodesic between a prior (for instance the origin) and the unconstrained GMM estimate on the training set.

As discussed in Section 1.4.4, while the most rigorous way to obtain the geodesic between two points is via a solution to a differential equation problem, most practical applications (for e.g. computer science applications) obtain approximate geodesics using algorithms like the *'Djikstra's algorithm'*. However, these algorithms slow down considerably as the dimension of the parameter space increases.

One of the ways to extend this regularization technique is by using the differential equation approach to obtain geodesics on the GMM objective function surface, which is out of the scope of the current paper.

Another possible direction is to borrow the intuition behind ISOMAP – an algorithm developed to map points on a high-dimensional non-linear manifold to a lower dimen-

sional set of coordinates. The intuition is to randomly sample a large number of points (in our case compute the objective at random parameter values a large number of times), denote the K nearest neighbors of each point as connected by an edge and then apply 'Djikstra's Theorem' to compute the geodesic. The geodesic approximation improves as the number of sampled datapoints increases.

We expect that as advances are made in computer science which enable greater storage capabilities and faster computation, more algorithms for computing geodesics on higher dimensional surfaces will become available.

Local spectral cutoff path

Recall from Section 1.4.5 the steps to obtain the regularized estimate using the Local Spectral cutoff involved 1) finding a possibly nonlinear dimension in the parameter space where the GMM objective function is poorly defined (on the training set); 2) finding the point on this manifold which is closest to the origin, say θ^*; 3) the path between θ^* to the global minimum on the training set $\widehat{\theta}_{gmm}$ make up the spectral cutoff path; 4) picking the parameter value on the path, which minimizes the loss function on the testing set, as the regularized parameter estimate.

From the steps we note that main difficulty in extending this method to higher dimensions is in finding the 'ill-defined manifold' as the dimension of the parameter space increases.

A suggested algorithm proceeds as follows:

Step 1: Find the unconstrained global minimum denote it θ^{step} where $step = 0$

Step 2: At the parameter value θ^{step}, compute the Hessian of the objective function $H(\theta^{step})$. Find the eigen vector v^{step} corresponding to the smallest eigen value λ_1.

Step 3: The next parameter value on the path is an ϵ-step in the direction of v^{step} ie $\theta^{step+1} = \theta^{step} + \epsilon \cdot v^{step}$. Also let $step = step + 1$.

Step 4: Repeat **Steps 2 and 3** S number of times. Let $step = 0$

Step 5: At the parameter value θ^{step}, compute the Hessian of the objective function $H(\theta^{step})$. Find the eigen vector v^{step} corresponding to the smallest eigen value λ_1.

Step 6: The next parameter value on the path is an ϵ-step in the direction of $-v^{step}$ ie $\theta^{step-1} = \theta^{step} - \epsilon \cdot v^{step}$. Also let $step = step - 1$.

Step 7: Repeat **Steps 5 and 6** S number of times.

Step 8: The ill-defined manifold consists of the points $\theta^{-S}, \theta^{-S+1}, \cdots \theta^0, \theta^1, \cdots \theta^S$.

1.8 Summary and Possible Extensions

In this paper we have introduced the application of regularization techniques in GMM to solve a particular type of identification problem. We have shown via simulations that when the matrix of first order conditions is close to losing rank, traditional GMM estimates become unstable. We also show that the regularization techniques developed here perform well in terms of the Mean Squared Error of the estimates. This paper is the first step towards developing regularization techniques for GMM. There are a lot of possible extensions to this work.

First, since the regularization techniques developed here perform some form of grid search, the curse of dimensionality cannot be ignored. As the number of dimensions increase, the grid search method will take longer to complete and at some point will become infeasible. It is important to think of alternative characterizations of the regularized solution path which are less computationally intensive. One possible alternative may be to characterize the regularized solution paths in terms of differential equations.

Second, we have dealt only with exactly identified GMM systems. The effect of a poorly identified GMM system on overidentifying restrictions and the J-stat have not been addressed. Further the effect of regularized estimates on the J-stat also remains to be

explored.

Finally, there are a number of algorithms developed by Computer Scientists to find paths, modes, minima and maxima on irregular surfaces. GMM objective functions are typically associated with highly non-linear surfaces and algorithms and techniques from fields like Computer Science and Statistics should be utilized in order to improve the quality of those GMM problems that suffer from instability.

Appendix

A1: Results from a Linear System

As discussed in Section 1.5 of the paper we found that GMM estimation problems where the moment conditions are linear in paramaters, the three regularization techniques proposed in the paper perform well. In this section we present simulation results on GMM estimation of an OLS problem.

Suppose the $n \times 1$ vector Y is the outcome of interest, $X = [X_1 \quad X_2]$ is the $n \times 2$ matrix of covariates and the *true model* the data follows is:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \varepsilon \perp X.$$

$\frac{X'X}{n}$ has the eigenvalue decomposition:

$$\frac{X'X}{n} = \begin{pmatrix} C_1 & C_2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} C_1' \\ C_2' \end{pmatrix}$$

If $\lambda_2 \to 0$ then both parameters β_1 and β_2 cannot be estimated exactly. In our simulations we set $\beta_1 = \beta_2 = 2$, $\lambda_1 = 0.1$ and allow $\frac{\lambda_2}{\lambda_1} \to 0$. From the independence assumption on the error term the moment conditions for the exactly identified GMM problem are:

$$g(\beta_1, \beta_2) = X'(\varepsilon(\beta_1, \beta_2)) = X'(Y - \beta_1 X_1 - \beta_2 X_2).$$

The graphs and tables for different values of $\frac{\lambda_2}{\lambda_1}$ are presented next. Note that when $\frac{\lambda_2}{\lambda_1} \leq 0.01$ then all three regularization techniques lead to an improvement in MSE values of the parameter estimates.

Table 1.39: MSE Comparison $\lambda_2 = \lambda_1$

		Bias	*Var*	*MSE*
GMM	$\hat{\mu}$	−0.01	0.11	0.11
	$\hat{\theta}$	−0.01	0.10	0.10
	$(\hat{\mu}, \hat{\theta})$			0.21
Ridge-type	$\hat{\mu}_r$	−0.22	0.16	0.21
	$\hat{\theta}_r$	−0.24	0.16	0.22
	$(\hat{\mu}_r, \hat{\theta}_r)$			0.43
Geodesic	$\hat{\mu}_g$	−0.21	0.20	0.24
	$\hat{\theta}_g$	−0.22	0.18	0.23
	$(\hat{\mu}_g, \hat{\theta}_g)$			0.47
Spectral	$\hat{\mu}_s$	−0.35	0.24	0.36
	$\hat{\theta}_s$	−0.05	0.16	0.16
	$(\hat{\mu}_s, \hat{\theta}_s)$			0.53

Table 1.40: MSE Comparison $\lambda_2 = (0.1)\ \lambda_1$

		Bias	*Var*	*MSE*
GMM	$\hat{\mu}$	−0.01	0.82	0.82
	$\hat{\theta}$	0.00	0.34	0.33
	$(\hat{\mu}, \hat{\theta})$			1.15
Ridge-type	$\hat{\mu}_r$	−0.42	0.70	0.87
	$\hat{\theta}_r$	−0.10	0.28	0.29
	$(\hat{\mu}_r, \hat{\theta}_r)$			1.16
Geodesic	$\hat{\mu}_g$	0.01	1.07	1.06
	$\hat{\theta}_g$	−0.35	0.52	0.64
	$(\hat{\mu}_g, \hat{\theta}_g)$			1.70
Spectral	$\hat{\mu}_s$	−0.51	0.40	0.66
	$\hat{\theta}_s$	−0.35	0.20	0.20
	$(\hat{\mu}_s, \hat{\theta}_s)$			0.85

Table 1.41: MSE Comparison $\lambda_2 = (0.01)\ \lambda_1$

		Bias	*Var*	*MSE*
GMM	$\hat{\mu}$	−0.01	7.92	7.92
	$\hat{\theta}$	0.01	2.70	2.69
	$(\hat{\mu}, \hat{\theta})$			10.61
Ridge-type	$\hat{\mu}_r$	−0.33	5.81	5.91
	$\hat{\theta}_r$	−0.15	1.97	1.99
	$(\hat{\mu}_r, \hat{\theta}_r)$			7.90
Geodesic	$\hat{\mu}_g$	0.22	5.34	5.38
	$\hat{\theta}_g$	−0.46	1.91	2.12
	$(\hat{\mu}_g, \hat{\theta}_g)$			7.50
Spectral	$\hat{\mu}_s$	−0.25	2.05	2.12
	$\hat{\theta}_s$	−0.46	0.70	0.74
	$(\hat{\mu}_s, \hat{\theta}_s)$			2.85

Table 1.42: MSE Comparison $\lambda_2 = (0.001)\ \lambda_1$

		Bias	*Var*	*MSE*
GMM	$\hat{\mu}$	−0.20	63.03	63.00
	$\hat{\theta}$	0.11	21.07	21.06
	$(\hat{\mu}, \hat{\theta})$			84.06
Ridge-type	$\hat{\mu}_r$	−0.34	42.04	42.12
	$\hat{\theta}_r$	−0.14	14.12	14.12
	$(\hat{\mu}_r, \hat{\theta}_r)$			56.24
Geodesic	$\hat{\mu}_g$	0.30	21.64	21.70
	$\hat{\theta}_g$	−0.50	7.33	7.57
	$(\hat{\mu}_g, \hat{\theta}_g)$			29.27
Spectral	$\hat{\mu}_s$	0.40	9.76	9.91
	$\hat{\theta}_s$	−0.50	3.15	3.46
	$(\hat{\mu}_s, \hat{\theta}_s)$			13.37

Chapter 2

Propensity Score Model Selection using Machine Learning Classifiers

2.1 Introduction

The basic issue in estimating the average effect of a particular economic policy or program using observational data is that the data suffers from *selection bias*. In other words those who receive treatment (the treatment group) are inherently different from those who don't (the control group). Heckman (in his seminal 1978 paper) shows that a naive estimate of the regression parameter on a treatment dummy (say $W = 1$ if an individual is treated and $W = 0$ if the individual is a control) suffers from an *omitted variable bias*. Technically suppose we have a dataset with sample size n, the outcome of interest is Y and we are interested in the Average Treatment Effect

$$ATE = E[Y(W = 1) - Y(W = 0)]$$

which is the difference between the unconditional means of outcomes under treatment ($W = 1$) and under control ($W = 0$). If the data were truly random then a naive estimator which compares the mean outcomes of the two groups does indeed provide an unbiased estimate of the average treatment effect (ATE). The problem arises because we only observe outcomes under a single state (either treatment or control). In this situation

it becomes necessary to control for factors which *simultaneously affect both outcome and selection* into the treatment group.

In the Econometrics literature, Heckman was the first to explicitly model selection issues. Heckman's model includes a two-step estimation procedure, where the first step describes a selection equation and the second step involves the outcome equation. Under certain assumptions (notably bivariate normality of error terms) the paper provides estimation procedures that control for the *omitted variable bias*. However the model has come under scrutiny, mainly due to evidence of poorer performance in terms of Mean Squared Error (MSE) in cases where the rather strong distributional assumption on the error terms does not hold.

In the Statistics literature Rubin and Rosenbaum (1983) pioneered the work on causal inference in the presence of selection bias. The core idea here is also a two-step estimation procedure. In the first step the probability that an individual belongs to the treatment group is estimated. This is referred to the individual's *Propensity Score*. The second step involves using the *Propensity Score* for pre-processing of the data before running regression models to estimate the ATE. This is technically referred to as *balancing the data*.

A way of checking for selection bias *a priori* is by considering the distribution over covariates for the two groups. If the two groups have similar distribution across the different covariates then the dataset is said to have *covariate balance* and thus can be considered as an approximately random sample. By weighting the observations by Inverse Propensity Score in an intermediate step, we aims to *mimic* a random sample.

The use of Inverse Propensity Score Weighting (IPW) is now ubiquitous in the Causal Inference literature, however the procedure has come under recent scrutiny. Kang and Schafer (2007) argue that the method is sensitive to misspecification of the propensity score model. The method is also affected by extreme values of estimated propensity

scores which then leads some experimenters to *trim* the data points corresponding to the extreme values resulting in loss of data. Freedman and Berk (2008) show via simulation studies that weighting is likely to increase random error of the estimates *except* under three specific cases (1. participants are i.i.d., 2. selection is exogenous, 3. the selection equation is correctly specified).

Other authors have proposed the use of non-parametric and semi-parameteric methods to estimate propensity scores to minimize model misspecification errors. In particular, Hirano et al (2001) use a series logistic regression, Linton (2001) uses kernel density estimation and McCaffrey et al (2004) use Generalized Boosted Regresion. Recently Lee et al (2010) considered tree based classifiers like CART, Boosting and Random forests in the IPW framework. Diamond and Sekhon (2012) introduce the application of Genetic Matching algorithms for causal inference. Imbens (2004) provides an excellent overview of the Propensity Score Analysis framework, including methodological advances as well as future research directions for the applied econometrician.

We extend the literature by comparing estimates of propensity scores from logistic regression with estimates obtained from three popular classifiers from the statistics literature – Naive Bayes, Random Forests and Support Vector Machines (SVMs). We then evaluate four measures that can be used to choose between the different propensity score estimates– Covariate Imbalance, Calibration Error, Likelihood and Classification Error Rate. We are not aware of any other paper in the literature that considers a variety of propensity score models to pick the 'best' one which will be used for IPW.

Overall we present two sets of results – methodological and empirical. This chapter deals with the set of methodological results which are obtained via two simulation studies. The first is based on artificial data where we introduce non-linearities in the true Propensity Score equation. The second is based on a real-world dataset – the Dahejia and Wahba (1999) sample of the LaLonde (1986) data from a randomized job training

experiment. We find that propensity score estimates with Minimum Covariate Imbalance perform very well in terms of Mean Squared Error of Average Treatment Effect estimates across all our simulations. We also find that the best classifier (with the lowest Classification Error Rate) is not necessarily the best choice for propensity score estimation. The Minimum Covariate Imbalance measure picks probit estimates in a number of cases, implying that in many cases the probit does perform well even under model misspecification. The unweighted estimate is also picked by the measure in some cases, implying that sometimes the naive estimator has better balance than any of the weighted estimators, in which case IPW should not be used. The set of empirical results are from an application of the Minimum Covariate Imbalance estimator on a large public health dataset from India. These are presented in the last chapter of this document.

Our analysis deals only with the Inverse Propensity-Score Weighting (IPW) framework. There exist a number of alternative frameworks which utilize Propensity Scores such as Matching, Stratification and Difference in Difference estimators. While the current paper does not deal with these frameworks, we hope to extend our research to these frameworks in the future. Our method can be naturally extended to the Doubly Robust framework since it corresponds to reducing the effects of model mis-specification in the first step of the Doubly Robust framework.

The rest of the chapter is organized as follows. The Inverse Propensity Score Weighting framework is described in the next section. Section 3 presents a background on the classifiers we use in this paper. Section 4 outlines three data-driven measures to select between the candidate propensity score models. Section 5 follows with Simulation Results. Directions for Future Research are discussed in Section 7. Section 8 concludes. Some technical notes are presented in the Appendix.

2.2 Inverse Propensity Score Weighting Framework

2.2.1 Background

In this section we decribe the Inverse Propensity Score Weighting framework in greater detail. The discussion largely follows Guo and Fraser (2010).

Let the total number of individuals in the dataset be given by n, the number of individuals in the treatment and control groups are given by n_1 and n_0 respectively. For ease of exposition we consider a single treatment, a binary dummy variable W. Let $W_i = 1$ if individual i received treatment and $W_i = 0$ otherwise. For every individual in the sample, we observe W_i (the state the individual i is in), the outcome of interest Y_i as well as a vector of observed characteristics X_i, where X_i is a row vector of dimension $(K \times 1)$. The *observed* outcome variable Y_i can be expressed as

$$Y_i = Y_{i0}(1 - W_i) + Y_{i1}W_i$$

where Y_{i0} and Y_{i1} are the outcomes for individual i under control and treatment states respectively. Note that when $W_i = 0$ we only observe Y_{i0} and vice versa. Membership in the Treatment and Control groups is denoted by $i \in \mathcal{T}$ and $j \in \mathcal{C}$ respectively.

The selection bias issue arises when the process determining the value of the outcome variable, i.e. the outcome equation is described by

$$Y = f(W, X, Z_1, U)$$

and the process determining selection into the treatment group, i.e. the selection equation is described by

$$W = g(X, Z_2, V).$$

where covariates Z_1 and Z_2 affect outcome and selection independently, covariates X affect both outcome and selection simulataneously and U and V represent white noise error terms. This specification implies that a simple comparison of mean outcomes of the treatment and control group will be biased since outcome Y and selection W are not independent. To make meaningful comparisons we have to control for the confounding covariates X.

A possible method to correct for this bias is to balance on covariates X i.e. compare only those individuals who have similar covariate values. However, when the number of covariates in X becomes large, the *curse of dimensionality* kicks in and balancing over all K dimensions becomes computationally difficult. Propensity scores (also known as the *coarsest score*) address this problem by summarizing the information of vector X_i (also known as the *finest score*). "The most important property is that a coarsest score can sufficiently balance differences observed in the finest scores between treated and control participants" (Guo and Fraser (2010)).

Mathematically, define the *true* propensity score as:

$$e(X_i) = Pr(W_i = 1|X_i = X_i).$$

It can be shown[1] (under the Rubin, Rosenbaum (1983) framework) that given propensity scores, the treatment state and observed covariates are conditionally independent,

$$X_i \perp W_i|e(X_i).$$

If the true propensity score is known then an *unbiased* estimate of the ATE is provided by.

$$\widehat{\tau} = \frac{1}{n_1}\Sigma_{i=1}^{n_1}\left[\frac{W_i Y_i}{e(X_i)}\right] - \frac{1}{n_0}\Sigma_{i=1}^{n_0}\left[\frac{(1-W_i)Y_i}{1-e(X_i)}\right].$$

[1] under standard assumption of unconfoundedness/ignorability i.e. $Y_{io}, Y_{i1} \perp W_i|X_i$.

Alternatively, we can re-write this estimator as a weighted least squares regression function with weights λ_i

$$Y_i = \alpha + \tau W_i + \varepsilon_i, \qquad \lambda_i = \sqrt{\frac{W_i}{e(X_i)} + \frac{(1 - W_i)}{1 - e(X_i)}}. \tag{2.1}$$

Equation (2.1) forms the basis for the IPW method. The aim is to reweight the given sample of treated and control participants to create a sample that is representative of the entire population.

2.2.2 Understanding Inverse Propensity Score Weighting

Let $\widehat{e}(X_i)$ denote the estimated propensity score. Then the IPW weighting scheme for ATE is given by:

$$\omega_i(W_i, X_i) = \frac{W_i}{\widehat{e}(X_i)} + \frac{1 - W_i}{1 - \widehat{e}(X_i)}.$$

Consider the weighting scheme for observations in the treatment group,

$$\omega_i(W_i = 1, X_i) = [\widehat{e}(X_i)]^{-1}.$$

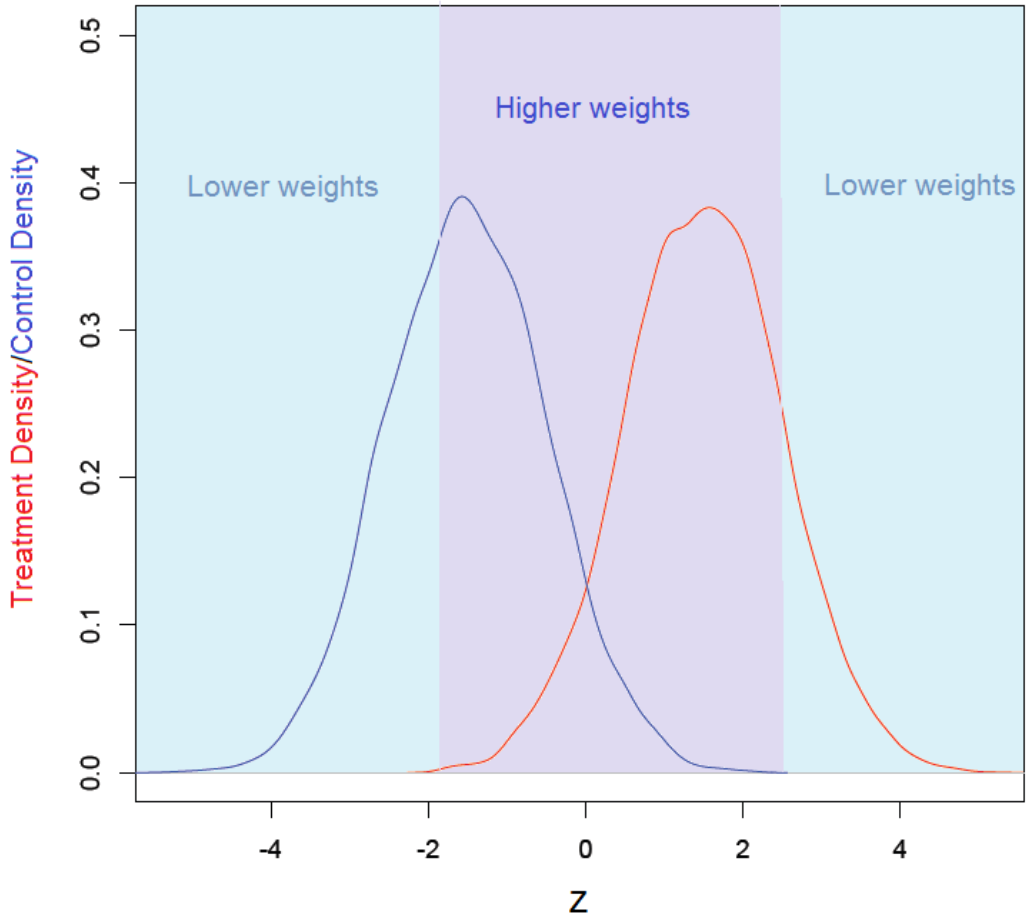
Since $1 \leq [\widehat{e}(X_i)]^{-1} \leq \infty$, therefore each observation gets a weight between 1 and $+\infty$, with those observations which are most likely to be treated getting the least weight and those most unlikely to be treated getting the maximum weight. Similarly, the weighting scheme for observations in the control group,

$$\omega_i(W_i = 0, X_i) = [1 - \widehat{e}(X_i)]^{-1}.$$

implies that each observation gets a weight between 1 and $+\infty$, with those observations which are most likely to be in the *control* group getting the least weight and those most unlikely to be in the control group getting the maximum weight.

Two properties of IPW deserve mention. First that the weighting scheme gives higher (lower) weights to those individuals who have low (high) probabilities of belonging to their actual group. Thus the weighting scheme aims to mimic a random distribution by giving greater weightage to those individuals in the treatment (control) group that were most likely to belong to the control (treatment) group.

Figure 2.1: Propensity Score Densities for Treatment and Control Groups – observations in the treatment group corresponding to higher values of propensity scores get lower weights and vice versa.



Second, extreme values of $\widehat{e}(X_i)$ can lead to disproportionately large weights ($\approx \infty$). This leads some researchers to only use those observations that satisfy $(1-\rho) < \widehat{e}(X_i) < \rho$, where ρ is a threshold probability value. In other words they *trim* the data, excluding those observations that are associated with extreme values of $\widehat{e}(X_i)$.

2.2.3 Estimation of Propensity Scores

Next we discuss the standard method used for the estimation of propensity scores. Recall that $W_i = 1$ if individual i received treatment and $W_i = 0$ otherwise. Therefore,

$$\begin{aligned} E(W_i) &= Pr(W_i = 1|X_i = x_i) \times 1 + Pr(W_i = 0|X_i = x_i) \times 0 \\ &= Pr(W_i = 1|X_i = x_i). \end{aligned}$$

Binary Logit and Probit regression are the standard estimation techniques used for this class of problems. $E(W_i)$ is modeled as:

$$E(W_i) = Pr(W_i = 1|X_i = x_i) = \begin{cases} \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} & \text{if using logit} \\ \\ \Phi(x_i'\beta) & \text{if using probit} \end{cases}.$$

The parameter vector β is estimated using Maximum Likelihood Estimation (MLE). Typically MLE problems are solved numerically. Once the $\widehat{\beta}$ estimates are obtained, these can be plugged into the Logit or Probit regression equation to obtain the *estimated propensity score* as:

$$\widehat{e}(x_i) = \widehat{Pr}(W_i = 1|X_i = x_i) = \begin{cases} \frac{e^{x_i'\widehat{\beta}}}{1+e^{x_i'\widehat{\beta}}} & \text{if using logit} \\ \\ \Phi(x_i'\widehat{\beta}) & \text{if using probit} \end{cases}.$$

However as discussed before the IPW method is sensitive to misspecification of the propensity score model. This has led some authors to propose non-parametric and semi-parametric techniques to estimate propensity scores.

The emphasis of this paper is on being *indifferent* between the different propensity score models. The aim is to pick the model which optimizes some desirable criteria. In Section 2.4 we discuss three criteria that may potentially be used to choose between the different candidate propensity score models. Next we introduce the candidate propensity score

models that we will be considering in this paper.

2.3 Classification Techniques and Class Probablities

The problem of predicting a discrete random variable Y from a set of random variable X is known as classification. Logistic regression itself is a traditional parametric method for classification. In the last few decades, classification techniques like 'Naive Bayes', 'Random Forests' and 'Support Vector Machines (SVM)' in particular have become very popular. These methods can also provide *classification probabilities* which are essentially *propensity scores*. We estimate propensity scores using these methods in the first stage and then select the best model by a data-driven procedure.

In the past other authors have proposed the use of non-parametric and semi-parameteric methods to estimate propensity score to minimize model mis-specification errors. In particular, Hirano, Imbens, Ridder (2003) propose the use a series logit estimator, Linton (2001) uses kernel density estimation and McCaffrey et al (2004) use Generalized Boosted Regresion. Recently Lee et al (2010) considered tree based classifiers like CART, Boosting and Random forests on the IPW framework. In this paper we work with Random Forests as well as two new possible candidates – Naive Bayes and SVMs. However, our main contribution is *not* to single out one particular estimation technique *a priori*. Instead we propose picking one of the many candidate models using a suitable selection criteria.

Machine learning alogorithms have proved to be very successful in providing good classifiers. The methods tend to be semi-parametric and data-driven. Next we briefly describe the different classifiers that we use in this paper. For a more detailed discussion refer to Hastie, Tibshirani and Friedman (2008).

Consider iid data $(X_1, W_1), \cdots, (X_n, W_n)$ where $W_i = \{0, 1\}$ and X_i that takes values $X_{i1}, \cdots X_{id} \in \Re^d$. A *classification rule* is a function $h : \mathbf{X} \to \{0, 1\}$. When we observe a

new X, we predict W to be $h(X)$.

2.3.1 Logit and Probit Models

As discussed previously, the traditional method for estimating propensity score is via logit and probit models which correspond to the *classification rule*:

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{r}(x) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

where

$$r(x) = P(W = 1|X = x) = \begin{cases} \frac{e^{X'\beta}}{1+e^{X'\beta}} & \text{if using logit} \\ \\ \Phi(X'\beta) & \text{if using probit} \end{cases}$$

Despite the strong distributional and linearity assumptions involved, Logit and Probit models are popular because of their ease of interpretability and implementation. Typically, both Logit and Probit models lead to very similar probability estimates. In the rest of the paper, we work with the Probit model.

2.3.2 Naive Bayes

The Naive Baye's classifier is derived from the application of Baye's theorem.

Let $\pi = Pr(W_i = 1)$, then by Baye's theorem:

$$P(W_i = 1|X = X_i) = \frac{f(X_i|W_i = 1)\pi}{f(X_i|W_i = 1)\pi + f(X_i|W_i = 0)(1-\pi)} = \frac{f_1(X_i)\pi}{f_1(X_i)\pi + f_0(X_i)(1-\pi)}$$

where $f(X_i|W_i = c) = f_c(X_i)$ is the probability density function of X conditional on $W_i = c$ evaluated at X_i. This leads to the classification rule:

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \frac{\widehat{f}_1(x)}{\widehat{f}_0(x)} > \left(\frac{1-\widehat{\pi}}{\widehat{\pi}}\right) \\ \\ 0 & \text{otherwise} \end{cases}$$

where $\widehat{\pi} = \frac{1}{N}\Sigma_i^N Y_i$. Under the assumption that $X_1, \cdots, X_d$ are independent;

$$\widehat{f}_0(x) = \widehat{f}_0(x_1, \cdots, x_d) = \prod_{j=1}^{d} \widehat{f}_{0j}(x_j)$$
$$\widehat{f}_1(x) = \widehat{f}_1(x_1, \cdots, x_d) = \prod_{j=1}^{d} \widehat{f}_{1j}(x_j)$$

and $\widehat{f}_{kj}, \quad k = \{0, 1\}, \quad j = \{1 \cdots d\}$ are estimated using either a functional form for the distribution (e.g. Gaussian, Multinomial) or by non-parametric one-dimensional density estimators.

This method has proved to be very successful and is relatively easy to estimate. However, the underlying independence assumption may not be applicable to many real world settings.

2.3.3 Trees and Random Forests

Trees are classification methods that partition the covariate space into disjoint partitions. Observations are then classified according to which partition element they fall in. Typically such algorithms minimize the probability of miscategorizing an item. The 'size' of tree is chosen via a process known as *pruning*.

Since trees are typically very noisy, most statisticians use methods like *bootstrapped bagging, boosting* and *random forests*. These are ensemble methods that average predictions from a set of trees. In this paper we work with Random Forests.

2.3.4 Support Vector Machines

For convenience relabel the treatment and control outcomes (corresponding to W_i) as 1 and -1 instead of 1 and 0 respectively. A *hard margin* linear support vector machine classifier can be written as

$$h(x) = sign(H(x))$$

where $x = (x_1, \cdots, x_d)$ and $H(x) = \alpha_0 + \Sigma_{j=1}^{d} \alpha_j x_j$. Note that if the classifier is correct then $W_i H(X_i) \geq 0$ and if the classifier is incorrect then $W_i H(X_i) \leq 0$.

Suppose that the data are linearly separable then there exists a a hyperplane that perfectly separates the two classes. Support Vector Machines choose the hyperplane that maximizes the *margin* or the distance of the separating hyperplane to the closest points. Points on the boundary of the margin are called *support vectors*.

In more realistic settings where the data is not linearly separable, non-negative slack variables ζ are introduced in $H(x)$. The resulting classifier is known as the *soft margin* linear support vector machine. In this paper we work with soft margin linear support vector machines.

2.4 Choosing between Propensity Score Estimators

Given the many potential propensity score models, the choice of final model becomes important. We examine four measures to choose between different propensity score estimates.

- Minimum Covariate Imbalance Propensity Scores.
- Minimum Error Rate Propensity Scores.
- Minimum Calibration Error Propensity Scores.
- Maximum Predicted Likelihood Propensity Scores.

2.4.1 Minimum Covariate Imbalance Propensity Scores

Recall that the central problem in the estimation of average treatment effect is that the distribution of covariates may differ significantly between the treatment and the control groups. As a result a naive comparison of average outcomes across the two groups captures the effect of confounding covariates along with the effect of the treatment. In their seminal paper Rubin and Rosenbaum (1983) show that the propensity score is a *balancing* score i.e. the covariates are independent of the treatment variable conditional on the propensity score. Mathematically,

$$X_i \perp W_i | e(X_i), \quad i \in (1, 2, \cdots n)$$

For checking covariate imbalance in practice, we use the Imbalance statistic presented in Imai and Ratkovic (2007), which is the multivariate version of standardized difference of means statistic previously proposed by Rubin and Rosenbaum (1985):

$$d(h(X)) = (\Gamma' X) \left(cov(X)\right)^{-1} (X'\Gamma),$$

where $X = X_{n \times k}$ is the matrix of covariates, $h(.)$ refers to the classifier and $\Gamma = [\gamma_1, \gamma_2, \cdots \gamma_n]$ is a vector of observation weights,

$$\gamma_i = \frac{W_i}{n[\widehat{e}(h(x_i))]} - \frac{1 - W_i}{n[1 - \widehat{e}(h(x_i)]}.$$

$\Gamma' X$ is the mean weighted difference between covariates in the treatment and control groups. The weights correspond to the propensity score estimates corresponding to classifier h.

For the unweighted estimator,

$$\gamma_i = \frac{W_i}{n} - \frac{1 - W_i}{n}$$

in which case $\Gamma' X$ corresponds to the mean difference between covariates of the two groups. The Covariate Imbalance score in this case is denoted by $d_0(X)$

The procedure for picking the Minimum Covariate Imbalance Propensity Score is described next.

1. Define outcome of interest Y, treatment W and covariates X.

2. Using W and X estimate propensity scores $\widehat{e}(h_1(x_i)), \widehat{e}(h_2(x_i)) \cdots \widehat{e}(h_m(x_i))$ for each of the m classifiers $h_1(.), h_2(.), \cdots h_m(.)$.

3. Calculate the covariate balance statistic $d_j(h(.)), \quad j \in [1, 2, \cdots m]$ for each of the m classifiers $h_1(.), h_2(.), \cdots h_m(.)$ and $d_0(.)$ for the naive estimator.

4. Pick the classifier that minimizes $d_j(h(.)), \quad j \in [0, 1, \cdots m]$.

    i. If $d_0(.)$ is selected then propensity score weighting should not be used.

    ii. If not then denote the selected classifier as $h^*(.)$ and the asociated estimated propensity scores as $\widehat{e}_{h^*}(.)$. The ATE estimate is given by $\widehat{\tau}_{imb}$ from the following weighted least squares regression with weights λ_i:

$$Y_i = \alpha + \tau_{imb} W_i + \varepsilon_i, \qquad \lambda_i = \sqrt{\frac{W_i}{\widehat{e}_{h^*}(X_i)} + \frac{(1 - W_i)}{1 - \widehat{e}_{h^*}(X_i)}}. \tag{2.2}$$

2.4.2 Minimum Classification Error Propensity Scores

This measure mirrors the Minimum Classification Error statistic used in many Machine Learning applications. In order to get an unbiased estimate of the prediction error, the data is randomly divided into a training and a testing set. The classifier is run on the training set. Class predictions are made on the testing set and denoted $\widehat{w}(h, tst)$. The error rate corresponding to classifier h, is given by:

$$error_h = \frac{\Sigma_i^{n_{tst}} I\{\widehat{w}_{i,tst}(h) \neq w_{i,tst}\}}{n_{tst}}$$

which simply calculates the number of observations that were mis-classified by classifier h divided by the total number of datapoints in the testing set (n_{tst}). This criteria is only applicable when we choose between different propensity score estimates (i.e. the unweighted estimator is not a candidate). Suppose classifier h^* minimizes the error rate, then the ATE estimate is given by $\widehat{\tau}_{err}$ from the following weighted least squares regression with weights λ_i:

$$Y_i = \alpha + \tau_{err} W_i + \varepsilon_i, \qquad \lambda_i = \sqrt{\frac{W_i}{\widehat{e}_{h^*}(X_i)} + \frac{(1 - W_i)}{1 - \widehat{e}_{h^*}(X_i)}}. \tag{2.3}$$

where the propensity score estimates come from running the best classifier on the full dataset. In this paper, we randomly pick a third of the total sample to make up the testing set.

2.4.3 Minimum Calibration Error Propensity Scores

Calibration error is used to measure how well the estimated probabilities match the data. The method is used to create 'reliability diagrams' which are often used to provide a visual aid to check whether estimated proabilities are well calibrated.

For a given classifier h,

1. Sort the sample in ascending order of estimated propensity scores $\widehat{e}_h(X_i)$
2. Pick a bin size, s. This corresponds to a total number of bins $B = n/s$. [2]
3. Within each bin b, compute

$$p_b = \overline{\widehat{e}_h(X_i)}, \quad f_b = \frac{\Sigma_i w_i}{s}; \quad i \in b$$

.

[2]For ease of exposition we assume that sample size n is a multiple of bin size.

4. Calibration Error corresponding to classifier h is given by

$$CAL_h = \Sigma_b^B (p_b - f_b)^2.$$

.

Suppose classifier h^* minimizes the Calibration Error, then the ATE estimate is given by $\widehat{\tau}_{cal}$ from the following weighted least squares regression with weights λ_i:

$$Y_i = \alpha + \tau_{cal} W_i + \varepsilon_i, \qquad \lambda_i = \sqrt{\frac{W_i}{\widehat{e}_{h^*}(X_i)} + \frac{(1 - W_i)}{1 - \widehat{e}_{h^*}(X_i)}}. \tag{2.4}$$

In case any of these measures leads to a *tie* between two candidate models, then the corresponding ATE estimate is obtained by taking the mean of the estimates coming from the tied models.

2.4.4 Maximum Likelihood Propensity Scores

The Maximum Likelihood Propensity Score measures the probability of observing the data in the testing set conditional on a given propensity score model. The likelihood function and the log likelihood function corresponding to classifier h are given by

$$\begin{aligned}
\mathcal{L}(W_{tst}|\widehat{e}_h) &= \prod_{i=1}^{n_{tst}} \widehat{e}_h(X_i)^{W_i} (1 - \widehat{e}_h(X_i))^{(1-W_i)} \\
l(W_{tst}|\widehat{e}_h) &= \log\left(L(W_{tst}|\widehat{e}_h)\right) \\
&= \sum_{i=1}^{n_{tst}} W_i \log\left(\widehat{e}_h(X_i)\right) + (1 - W_i) \log\left(1 - \widehat{e}_h(X_i)\right)
\end{aligned}$$

where observations i belong to the *testing data* and the propensity score model $\widehat{e}_h$ is obtained by fitting classifier h on the *training data*.

Suppose classifier h^* maximizes the predicted likelihood function, then the ATE estimate is given by $\widehat{\tau}_{lhd}$ from the following weighted least squares regression run on the

entire dataset with weights λ_i:

$$Y_i = \alpha + \tau_{lhd} W_i + \varepsilon_i, \qquad \lambda_i = \sqrt{\frac{W_i}{\widehat{e}_{h^*}(X_i)} + \frac{(1 - W_i)}{1 - \widehat{e}_{h^*}(X_i)}}. \tag{2.5}$$

The Maximum Predicted Likelihood criteria is related to both the Minimum Classification Error and Minimum Calibration Error criteria – first it uses the training set to model propensity scores and then predicts propensity scores on the testing set, second it measures the likelihood that the testing set data is generated by these predicted propensity scores.

Out of the four measures, we expect the first to perform the best since it is most closely associated with the aim of IPW i.e. to achieve maximum covariate balance. The second measure may not perform so well, since its possible that an estimator which pushes estimated probabilities towards extreme values does a good job of classifying observations into their correct groups but not of balancing the data. Further, the measure may be sensitive to the relative sizes of the testing and training sets. The third measure is promising since it is based on how well the models estimate probabilities. However, the measure may be sensitive to the choice of bin size and thus should be treated with caution. The fourth measure is also promisings ince it combines the objectives of the second and third criteria and does not have binning issues like the Minimum Calibration error criteria.

2.5 Simulation Experiments

In this section we present results from two sets of simulation experiments. The first simulation experiment is based on artificial data, whereas the second uses real world data. In both sets of experiments we introduce non-linearites in the outcome and selection equations. We then estimate propensity scores using the classification techniques described in Section 2.3 and select the propensity scores that minimize the criteria described in Section 2.4.

We use the R Statistical Package to run our simulations. A more detailed description of the implementation is presented in Appendix A1.

2.5.1 Evaluation Strategy

In the simulation setting we know the true value of ATE, therefore we are able to compare the ATE estimates based on the value of their Mean Squared Errors (MSE).

$$MSE(\hat{\tau}) = E\left[(\hat{\tau} - \tau)^2\right] = b(\hat{\tau})^2 + Var(\hat{\tau})$$

where $b(\hat{\tau})$ and $Var(\hat{\tau})$ refer to the bias and variance of the estimator $\hat{\tau}$ respectively. The MSE incorporates a trade-off between the bias and variance of an estimator and is a measure of the predictive ability of an estimator.

A major critique of the IPW framework is that the standard errors associated with the ATE estimates can be so large that they offset the gains from bias reduction. In our simulaton experiments we compare MSE values from the ATE estimates obtained from 1) the naive unweighted estimator, 2) IPW estimator using a linear probit model, 3) IPW with estimator selection via Minimum Covariate Imbalance criteria and 4) IPW with estimator selection via Minimum Classification Error criteria 5) IPW with estimator selection via Minimum Calibration Error criteria and 6) IPW with estimator selection via Maximum Predicted Likelihood criteria.

2.5.2 Simulation Experiment I – Artificial Data

In the first set of experiments, we simulate scenarios where a linear probit propensity score model may not be appropriate. We then choose from the four parametric and semi-parametric propensity score models described in Section 2.3. We investigate if selecting one of these model can lead to better performance in terms of MSE values. In particular, we consider the following five selection equations:

- Selection Equation 1: $w = 1 * \{-1 + x_1 + x_2 + v > 0\}$

- Selection Equation 2: $w = 1 * \{-1 + x_1 + x_2^2 + v > 0\}$
- Selection Equation 3: $w = 1 * \{-1 + x_1 + x_3^3 + v > 0\}$
- Selection Equation 4: $w = 1 * \{-1 + x_1 + x_2^2 + x_3^3 + v > 0\}$
- Selection Equation 5: $w = 1 * \{-1 + x_1 + \exp(x_3) + v > 0\}$

For each of the selection equations, we consider two outcome equations:

- Linear Outcome Equation: $y = 1 + 10w + 2x_1 + 2x_2 + 2x_3 + u$
- (Mildly) Nonlinear Outcome Equation: $y = 1 + 10w + 2x_1 + 2x_2^2 + 2x_3 + u$

The simulations are designed so that,

- all covariates are standard normal: $x_1, x_2, x_3 \sim N(0, 1)$,
- error terms are standard normal: $u, v \sim N(0, 1)$,
- covariates and error terms are mutually independent: $x_1 \perp x_2 \perp x_3 \perp u \perp v$.

We run $N = 1000$ simulations for each of these models for sample sizes $n = \{100, 250, 500, 1000\}$.

We find that for the smallest sample size of $n = 100$, IPW using the linear probit model estimates performed better than any of the other methods in most cases (7 out of 10). However for almost all cases (28 out of 30) with $n \geq 250$ and all cases (20 out of 20) with $n \geq 500$, the IPW estimator based on the Minimum Covariate Imbalance criteria performed better in terms of MSE than IPW using the linear probit model. Further the IPW estimator based on the Minimum Covariate Imbalance criteria had the lowest MSE values in most cases (27 out of the 30) with $n \geq 250$. The naive unweighted estimator consistently had the poorest MSE values of all methods considered. The performance of IPW estimator based on the other three criteria was mixed – in a large proportion of cases inferior to IPW using Linear Probit. A detailed discussion of the insights from the simulations is presented at the end of this Section.

Simulation I: Selection Equation 1

Selection Equation $w = 1 * \{-1 + x_1 + x_2 + v > 0\}$

– **Linear Outcome Equation** $y = 1 + 10w + 2x_1 + 2x_2 + 2x_3 + u$

Table 2.1: MSE Comparison $n = 100$

	Bias	*SD*	*MSE*
Unweighted	3.83	0.69	3.89
IPW: Probit	1.20	1.09	**1**.62
IPW: Min Imbalance	1.59	1.03	1.90
IPW: Min Error	1.93	1.14	2.24
IPW: Min Cal	1.71	1.15	2.06
IPW: Max Likelhd	1.73	1.14	2.07

Table 2.2: MSE Comparison $n = 250$

	Bias	*SD*	*MSE*
Unweighted	3.87	0.42	3.89
IPW: Probit	0.81	1.05	**1.32**
IPW: Min Imbalance	1.13	0.71	1.34
IPW: Min Error	1.61	1.14	1.97
IPW: Min Cal	1.21	0.92	1.52
IPW: Max Likelhd	1.23	1.03	1.61

Table 2.3: MSE Comparison $n = 500$

	Bias	*SD*	*MSE*
Unweighted	3.86	0.31	3.88
IPW: Probit	0.67	0.88	1.10
IPW: Min Imbalance	0.91	0.57	**1.07**
IPW: Min Error	1.38	0.96	1.68
IPW: Min Cal	0.91	0.84	1.24
IPW: Max Likelhd	0.97	0.83	1.27

Table 2.4: MSE Comparison $n = 1000$

	Bias	*SD*	*MSE*
Unweighted	3.86	0.21	3.87
IPW: Probit	0.52	0.84	0.99
IPW: Min Imbalance	0.74	0.45	**0.86**
IPW: Min Error	1.20	0.96	1.54
IPW: Min Cal	0.73	0.68	1.00
IPW: Max Likelhd	0.74	0.67	1.00

– **Nonlinear Outcome Equation** $y = 1 + 10w + 2x_1 + 2x_2^2 + 2x_3 + u$

Table 2.5: MSE Comparison $n = 100$

	Bias	*SD*	*MSE*
Unweighted	2.55	0.93	2.71
IPW: Probit	0.41	1.31	1.38
IPW: Min Imbalance	0.70	1.15	**1.35**
IPW: Min Error	1.00	1.15	1.52
IPW: Min Cal	0.88	1.27	1.54
IPW: Max Likelhd	0.83	1.24	1.49

Table 2.6: MSE Comparison $n = 250$

	Bias	*SD*	*MSE*
Unweighted	2.56	0.56	2.62
IPW: Probit	0.22	1.16	1.19
IPW: Min Imbalance	0.39	0.84	**0.92**
IPW: Min Error	0.79	1.04	1.31
IPW: Min Cal	0.50	1.08	1.19
IPW: Max Likelhd	0.54	1.07	1.20

Table 2.7: MSE Comparison $n = 500$

	Bias	*SD*	*MSE*
Unweighted	2.57	0.39	2.60
IPW: Probit	0.12	1.03	1.03
IPW: Min Imbalance	0.26	0.77	**0.81**
IPW: Min Error	0.70	0.93	1.16
IPW: Min Cal	0.28	0.93	0.97
IPW: Max Likelhd	0.32	0.96	1.01

Table 2.8: MSE Comparison $n = 1000$

	Bias	*SD*	*MSE*
Unweighted	2.57	0.29	2.58
IPW: Probit	0.15	0.98	0.99
IPW: Min Imbalance	0.23	0.61	**0.65**
IPW: Min Error	0.62	0.92	1.11
IPW: Min Cal	0.25	0.82	0.86
IPW: Max Likelhd	0.26	0.84	0.88

Simulation I: Selection Equation 2

Selection Equation $w = 1 * \{-1 + x_1 + x_2^2 + v > 0\}$

– **Linear Outcome Equation** $y = 1 + 10w + 2x_1 + 2x_2 + 2x_3 + u$

Table 2.9: MSE Comparison $n = 100$

	Bias	*SD*	*MSE*
Unweighted	1.80	0.73	1.95
IPW: Probit	0.10	0.58	0.59
IPW: Min Imbalance	0.24	0.51	**0.57**
IPW: Min Error	0.99	0.71	1.22
IPW: Min Cal	0.48	0.66	0.82
IPW: Max Likelhd	0.97	0.67	1.17

Table 2.10: MSE Comparison $n = 250$

	Bias	*SD*	*MSE*
Unweighted	1.78	0.45	1.84
IPW: Probit	−0.02	0.49	0.49
IPW: Min Imbalance	0.11	0.34	**0.36**
IPW: Min Error	0.99	0.48	1.10
IPW: Min Cal	0.18	0.52	0.55
IPW: Max Likelhd	0.88	0.43	0.98

Table 2.11: MSE Comparison $n = 500$

	Bias	*SD*	*MSE*
Unweighted	1.81	0.33	1.84
IPW: Probit	−0.05	0.34	0.34
IPW: Min Imbalance	0.04	0.26	**0.26**
IPW: Min Error	0.98	0.42	1.07
IPW: Min Cal	0.05	0.37	0.37
IPW: Max Likelhd	0.89	0.36	0.96

Table 2.12: MSE Comparison $n = 1000$

	Bias	*SD*	*MSE*
Unweighted	1.81	0.23	1.83
IPW: Probit	−0.10	0.31	0.33
IPW: Min Imbalance	−0.01	0.22	**0.22**
IPW: Min Error	0.94	0.37	1.01
IPW: Min Cal	−0.03	0.28	0.28
IPW: Max Likelhd	0.89	0.35	0.96

– **Nonlinear Outcome Equation** $y = 1 + 10w + 2x_1 + 2x_2^2 + 2x_3 + u$

Table 2.13: MSE Comparison $n = 100$

	Bias	*SD*	*MSE*
Unweighted	4.25	0.73	4.31
IPW: Probit	3.35	0.92	3.47
IPW: Min Imbalance	3.30	0.87	3.41
IPW: Min Error	2.99	0.90	3.12
IPW: Min Cal	3.19	0.95	3.33
IPW: Max Likelhd	2.76	0.89	**2.90**

Table 2.14: MSE Comparison $n = 250$

	Bias	*SD*	*MSE*
Unweighted	4.27	0.47	4.29
IPW: Probit	3.33	0.56	3.38
IPW: Min Imbalance	3.31	0.54	3.36
IPW: Min Error	2.81	0.75	2.91
IPW: Min Cal	3.28	0.63	3.34
IPW: Max Likelhd	2.61	0.66	**2.69**

Table 2.15: MSE Comparison $n = 500$

	Bias	*SD*	*MSE*
Unweighted	4.27	0.34	4.28
IPW: Probit	3.36	0.47	3.39
IPW: Min Imbalance	3.34	0.43	3.36
IPW: Min Error	2.69	0.66	2.77
IPW: Min Cal	3.38	0.45	3.41
IPW: Max Likelhd	2.58	0.59	**2.64**

Table 2.16: MSE Comparison $n = 1000$

	Bias	*SD*	*MSE*
Unweighted	4.26	0.22	4.27
IPW: Probit	3.34	0.30	3.35
IPW: Min Imbalance	3.34	0.30	3.35
IPW: Min Error	2.64	0.57	2.70
IPW: Min Cal	3.36	0.31	3.38
IPW: Max Likelhd	2.55	0.51	**2.60**

Simulation I: Selection Equation 3

Selection Equation $w = 1 * \{-1 + x_1 + x_3^3 + v > 0\}$

– **Linear Outcome Equation** $y = 1 + 10w + 2x_1 + 2x_2 + 2x_3 + u$

Table 2.17: MSE Comparison $n = 100$

	Bias	*SD*	*MSE*
Unweighted	3.84	0.68	3.90
IPW: Probit	1.26	1.03	**1.62**
IPW: Min Imbalance	1.49	1.01	1.79
IPW: Min Error	2.18	1.18	2.47
IPW: Min Cal	1.87	1.11	2.17
IPW: Max Likelhd	1.92	1.18	2.25

Table 2.18: MSE Comparison $n = 250$

	Bias	*SD*	*MSE*
Unweighted	3.82	0.43	3.84
IPW: Probit	0.89	0.92	**1.28**
IPW: Min Imbalance	1.08	0.69	**1.28**
IPW: Min Error	2.03	1.17	2.35
IPW: Min Cal	1.22	0.88	1.50
IPW: Max Likelhd	1.61	1.16	1.99

Table 2.19: MSE Comparison $n = 500$

	Bias	*SD*	*MSE*
Unweighted	3.84	0.29	3.85
IPW: Probit	0.74	0.85	1.12
IPW: Min Imbalance	0.90	0.55	**1.05**
IPW: Min Error	2.16	1.14	2.45
IPW: Min Cal	0.99	0.77	1.25
IPW: Max Likelhd	1.62	1.33	2.10

Table 2.20: MSE Comparison $n = 1000$

	Bias	*SD*	*MSE*
Unweighted	3.85	0.22	3.85
IPW: Probit	0.68	0.78	1.03
IPW: Min Imbalance	0.82	0.45	**0.93**
IPW: Min Error	2.25	1.19	2.54
IPW: Min Cal	0.89	0.62	1.08
IPW: Max Likelhd	1.69	1.41	2.20

– **Nonlinear Outcome Equation** $y = 1 + 10w + 2x_1 + 2x_2^2 + 2x_3 + u$

Table 2.21: MSE Comparison $n = 100$

	Bias	*SD*	*MSE*
Unweighted	3.84	0.81	3.93
IPW: Probit	1.13	1.40	**1.80**
IPW: Min Imbalance	1.39	1.33	1.92
IPW: Min Error	2.11	1.38	2.52
IPW: Min Cal	1.77	1.34	2.22
IPW: Max Likelhd	1.82	1.43	2.31

Table 2.22: MSE Comparison $n = 250$

	Bias	*SD*	*MSE*
Unweighted	3.85	0.50	3.88
IPW: Probit	0.92	1.09	1.42
IPW: Min Imbalance	1.08	0.91	**1.41**
IPW: Min Error	2.16	1.17	2.45
IPW: Min Cal	1.28	1.04	1.65
IPW: Max Likelhd	1.71	1.28	2.13

Table 2.23: MSE Comparison $n = 500$

	Bias	*SD*	*MSE*
Unweighted	3.86	0.34	3.88
IPW: Probit	0.74	0.96	1.21
IPW: Min Imbalance	0.89	0.72	**1.14**
IPW: Min Error	2.15	1.19	2.46
IPW: Min Cal	0.99	0.86	1.31
IPW: Max Likelhd	1.57	1.28	2.02

Table 2.24: MSE Comparison $n = 1000$

	Bias	*SD*	*MSE*
Unweighted	3.84	0.25	3.85
IPW:Probit	0.73	0.87	1.14
IPW: Min Imbalance	0.83	0.60	**1.02**
IPW: Min Error	2.28	1.11	2.54
IPW: Min Cal	0.92	0.70	1.15
IPW: Max Pred Lhd	1.69	1.36	2.17

Simulation I: Selection Equation 4

Selection Equation $w = 1 * \{-1 + x_1 + x_2^2 + x_3^3 + v > 0\}$

– **Linear Outcome Equation** $y = 1 + 10w + 2x_1 + 2x_2 + 2x_3 + u$

Table 2.25: MSE Comparison $n = 100$

	Bias	*SD*	*MSE*
Unweighted	3.21	0.65	3.27
IPW: Probit	0.51	0.92	**1.05**
IPW: Min Imbalance	0.79	0.73	1.07
IPW: Min Error	2.05	0.93	2.25
IPW: Min Cal	1.09	1.11	1.56
IPW: Max Likelhd	2.01	0.92	2.21

Table 2.26: MSE Comparison $n = 250$

	Bias	*SD*	*MSE*
Unweighted	3.25	0.40	3.28
IPW: Probit	0.29	0.85	0.90
IPW: Min Imbalance	0.55	0.60	**0.82**
IPW: Min Error	2.32	0.70	2.42
IPW: Min Cal	0.52	0.89	1.03
IPW: Max Likelhd	2.37	0.63	2.45

Table 2.27: MSE Comparison $n = 500$

	Bias	*SD*	*MSE*
Unweighted	3.24	0.29	3.25
IPW: Probit	0.18	0.73	0.75
IPW: Min Imbalance	0.38	0.50	**0.63**
IPW: Min Error	2.42	0.60	2.49
IPW: Min Cal	0.33	0.65	0.73
IPW: Max Likelhd	2.56	0.58	2.62

Table 2.28: MSE Comparison $n = 1000$

	Bias	*SD*	*MSE*
Unweighted	3.25	0.20	3.25
IPW: Probit	0.11	0.68	0.69
IPW: Min Imbalance	0.28	0.40	**0.49**
IPW: Min Error	2.55	0.53	2.61
IPW: Min Cal	0.24	0.61	0.66
IPW: Max Likelhd	2.81	0.42	2.84

– **Nonlinear Outcome Equation** $y = 1 + 10w + 2x_1 + 2x_2^2 + 2x_3 + u$

Table 2.29: MSE Comparison $n = 100$

	Bias	*SD*	*MSE*
Unweighted	5.25	0.65	5.29
IPW: Probit	3.71	1.07	**3.86**
IPW: Min Imbalance	3.73	0.95	3.85
IPW: Min Error	3.94	0.94	4.05
IPW: Min Cal	3.84	1.03	3.97
IPW: Max Likelhd	3.82	0.96	3.94

Table 2.30: MSE Comparison $n = 250$

	Bias	*SD*	*MSE*
Unweighted	5.25	0.42	5.27
IPW: Probit	3.78	0.77	3.86
IPW: Min Imbalance	3.78	0.67	**3.83**
IPW: Min Error	4.07	0.78	4.14
IPW: Min Cal	3.83	0.72	3.90
IPW: Max Likelhd	4.03	0.83	4.12

Table 2.31: MSE Comparison $n = 500$

	Bias	*SD*	*MSE*
Unweighted	5.26	0.28	5.27
IPW: Probit	3.80	0.60	3.85
IPW: Min Imbalance	3.79	0.54	**3.82**
IPW: Min Error	4.14	0.75	4.21
IPW: Min Cal	3.82	0.55	3.86
IPW: Max Likelhd	4.30	0.76	4.37

Table 2.32: MSE Comparison $n = 1000$

	Bias	*SD*	*MSE*
Unweighted	5.24	0.20	5.25
IPW: Probit	3.84	0.69	3.90
IPW: Min Imbalance	3.79	0.40	**3.81**
IPW: Min Error	4.21	1.09	4.35
IPW: Min Cal	3.82	0.50	3.86
IPW: Max Likelhd	4.63	0.53	4.66

Simulation I: Selection Equation 5

Selection Equation $w = 1 * \{-1 + x_1 + \exp(x_3) + v > 0\}$

– **Linear Outcome Equation** $y = 1 + 10w + 2x_1 + 2x_2 + 2x_3 + u$

Table 2.33: MSE Comparison $n = 100$

	Bias	*SD*	*MSE*
Unweighted	3.59	0.64	3.64
IPW: Probit	0.76	1.03	**1.28**
IPW: Min Imbalance	1.04	0.76	1.29
IPW: Min Error	1.72	1.13	2.06
IPW: Min Cal	1.35	1.10	1.74
IPW: Max Likelhd	1.56	1.16	1.94

Table 2.34: MSE Comparison $n = 250$

	Bias	*SD*	*MSE*
Unweighted	3.63	0.41	3.65
IPW: Probit	0.30	1.16	1.19
IPW: Min Imbalance	0.74	0.59	**0.94**
IPW: Min Error	1.53	1.14	1.91
IPW: Min Cal	0.71	1.05	1.27
IPW: Max Likelhd	1.21	1.20	1.71

Table 2.35: MSE Comparison $n = 500$

	Bias	*SD*	*MSE*
Unweighted	3.63	0.29	3.64
IPW: Probit	0.07	1.19	1.19
IPW: Min Imbalance	0.58	0.56	**0.80**
IPW: Min Error	1.48	1.22	1.92
IPW: Min Cal	0.39	1.05	1.12
IPW: Max Likelhd	1.13	1.22	1.67

Table 2.36: MSE Comparison $n = 1000$

	Bias	*SD*	*MSE*
Unweighted	3.62	0.20	3.63
IPW: Probit	−0.20	1.26	1.28
IPW: Min Imbalance	0.42	0.50	**0.65**
IPW: Min Error	1.41	1.23	1.87
IPW: Min Cal	0.10	0.97	0.98
IPW: Max Likelhd	1.10	1.28	1.69

– **Nonlinear Outcome Equation** $y = 1 + 10w + 2x_1 + 2x_2^2 + 2x_3 + u$

Table 2.37: MSE Comparison $n = 100$

	Bias	*SD*	*MSE*
Unweighted	3.62	0.77	3.70
IPW: Probit	0.71	1.30	**1.48**
IPW: Min Imbalance	1.04	1.07	1.49
IPW: Min Error	1.74	1.28	2.16
IPW: Min Cal	1.32	1.28	1.84
IPW: Max Likelhd	1.56	1.37	2.07

Table 2.38: MSE Comparison $n = 250$

	Bias	*SD*	*MSE*
Unweighted	3.62	0.48	3.65
IPW: Probit	0.28	1.29	1.32
IPW: Min Imbalance	0.74	0.81	**1.10**
IPW: Min Error	1.52	1.29	2.00
IPW: Min Cal	0.71	1.16	1.36
IPW: Max Likelhd	1.22	1.27	1.76

Table 2.39: MSE Comparison $n = 500$

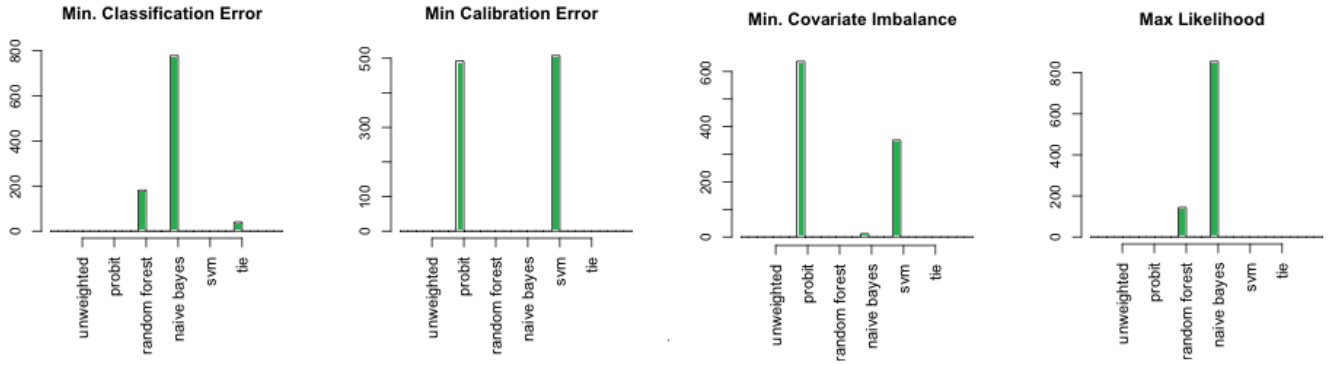
	Bias	*SD*	*MSE*
Unweighted	3.61	0.35	3.63
IPW: Probit	0.03	1.26	1.26
IPW: Min Imbalance	0.56	0.67	**0.87**
IPW: Min Error	1.40	1.24	1.88
IPW: Min Cal	0.36	1.04	1.10
IPW: Max Likelhd	1.08	1.30	1.69

Table 2.40: MSE Comparison $n = 1000$

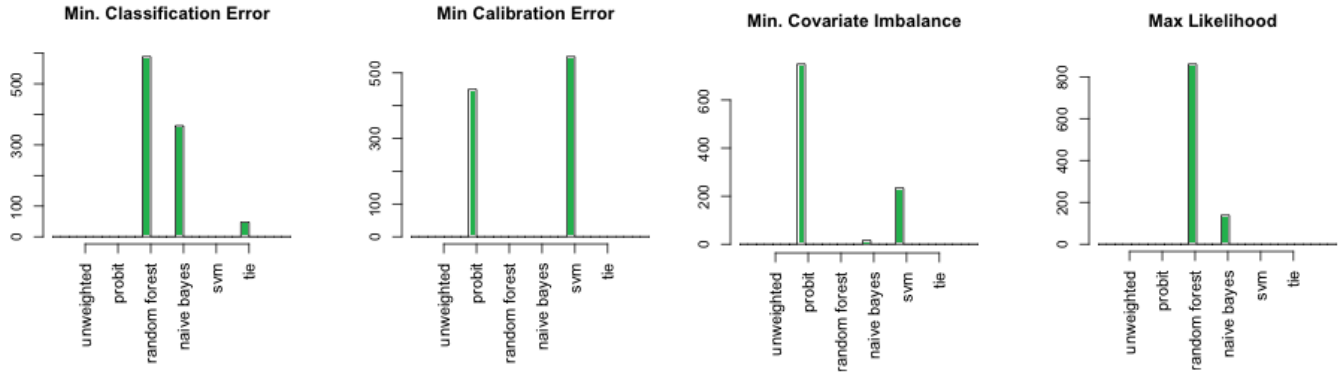
	Bias	*SD*	*MSE*
Unweighted	3.62	0.23	3.63
IPW: Probit	−0.17	1.41	1.42
IPW: Min Imbalance	0.46	0.58	**0.74**
IPW: Min Error	1.46	1.23	1.91
IPW: Min Cal	0.12	1.11	1.12
IPW: Max Likelhd	1.08	1.28	1.68

Figure 2.2: Propensity Score Models selected via Minimum Covariate Imbalance criteria, Minimum Classification Error criteria, Minimum Calibration Error criteria and Maximum Predicted Likelihood criteria corresponding to linear outcome equation and sample size $n = 1000$. Frequencies for the other cases are similar.

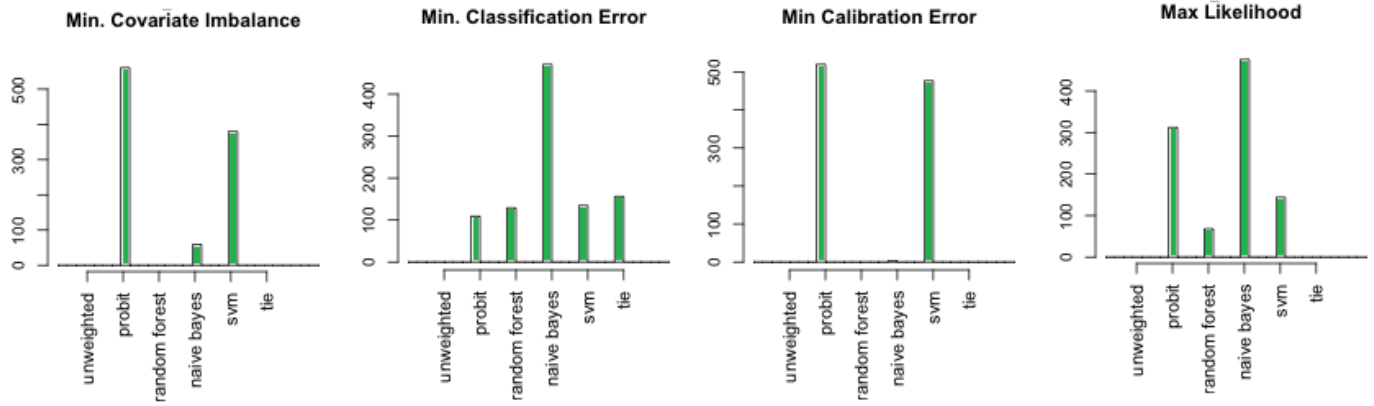
Selection Equation 1



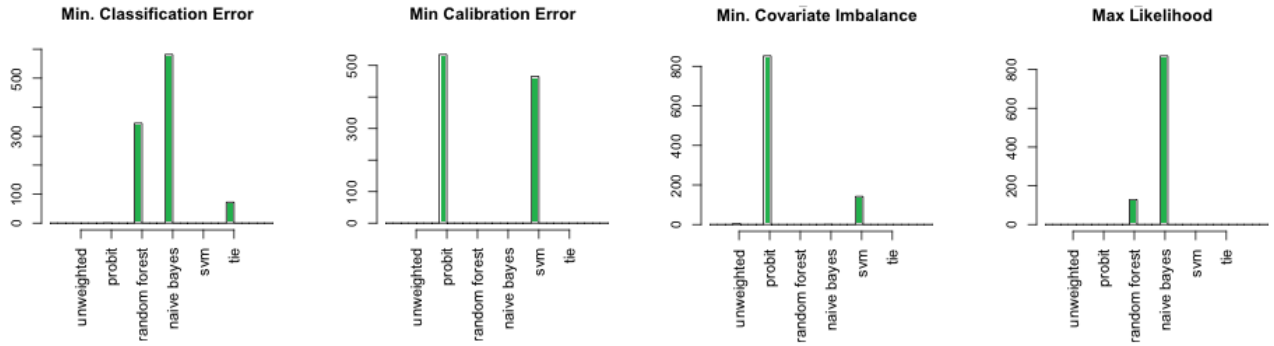
Selection Equation 2



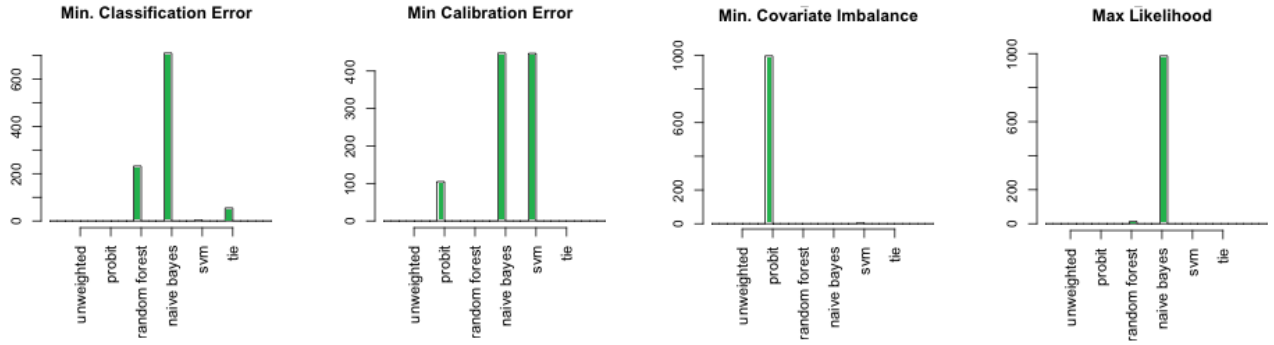
Selection Equation 3



Selection Equation 4



Selection Equation 5



2.5.3 Simulation Experiment II – Real-World Data

The data from the second simulation experiment is the Dahejia and Wahba (1999) sample of LaLonde's (1986) data from a randomized job training experiment. This is a more challenging scenario because unlike the previous experiment, the distribution of most covariates here is non-standard and many of the covariates are discrete. The maximum sample size is also limited to the 445 (260 control and 185 treatment) observations in the sample.The sample contains data on the following participant characteristics: age, education, whether black, whether hispanic, whether married, possession of a degree, real earnings in 1974, real earnings in 1975. The treatment variable is whether the individual was enrolled in the job training program. The outcome of interest is real earnings in 1978.

Our simulation design largely follows a simulation study outlined in Diamond and Sekhon (2012). The steps of the simulation are described next.

Step 1: We pick a random subsample of the data of size n after dropping the variables corresponding to treatment and outcome.

Step 2: For each individual in the random subsample we assign a probability of selection into the job training program (true propensity score) according to

$$\pi_i = \text{logit}^{-1}\left[1 + 0.5\widehat{\mu} + 0.01\ \text{age}^2 - 0.0675\ \text{edu}^2 - 0.01\log(\ \text{re74} + 0.01)^2 + 0.01\log(\ \text{re75} + 0.01)^2\right] \tag{2.6}$$

where $\hat{\mu}$ corresponds to the estimated propensity score model in the Dahejia and Wahba sample and is characterized by

$$\begin{aligned}\widehat{\mu} = 1 &+ 1.428 \times 10^{-4}\ \text{age}^2 - 2.918 \times 10^{-3}\ \text{edu}^2 - 0.2275\ \text{black} - 0.8276\ \text{hisp} + 0.2071\ \text{married} \\ &- 0.8232\ \text{nodegree} - 1.236 \times 10^{-9}\ \text{re74}^2 + 5.865 \times 10^{-10}\ \text{re75}^2 - 0.04328\ \text{u74} - 0.3804\ \text{u75}.\end{aligned}$$

where $u74$ and $u75$ are indicator variables for no real earnings in 1974 and 1975 respectively. The four extra terms in equation (2.6) added to $\widehat{\mu}$ ensure that the linear probit is badly misspecified. In our setup the only change we make to the experiment in Diamond and Sekhon is to change the coefficient of the education variable in the extra terms in equation (2.6). We make this change to calibrate the proportion of treated observations in the sample to approximately 40%, as is the case in the true sample.

Step 3: For each observation a Bernoulli trial with parameter π_i determines entry into the treatment group corresponding to $W_i = 1$.

Step 4: We fix the true value of ATE at $1,000$. For each observation the outcome is determined by

$$Y = 1000W + 0.1\exp\left[0.7\log(\text{re74} + 0.01) + 0.7\log(\text{re75} + 0.01)\right] + \varepsilon.$$

where $\varepsilon \sim N(0, 10)$ is a white noise error term.

Step 5: ATE estimates are obtained for this dataset using the different measures discussed in Section 2.3 and estimates corresponding to propensity score models selected via the three criteria discussed in Section 2.4. This is repeated for $N = 1,000$ random subsamples of size n, for $n = \{100, 150, 200, 250, 300, 350\}$

Step 6: We compare MSE values from the ATE estimates obtained from 1) the naive unweighted estimator, 2) IPW estimator using a linear probit model, 3) IPW with estimator selected using Minimum Covariate Imbalance criteria and 4) IPW with estimator selected using Minimum Classification Error criteria 5) IPW with estimator selected using Minimum Calibration Error criteria and 6) IPW with estimator selection via Maximum Predicted Likelihood criteria.

We find that the estimator based on IPW using the linear probit model had the lowest MSE values only for the smallest sample size of $n = 100$. In all other cases ($n \geq 150$) the IPW estimator with Minimum Covariate Imbalance had the lowest value of MSE. Interestingly the MSE values from the unweighted naive estimator are lower than those from IPW with the linear probit model for sample sizes of $n \geq 300$. However, this does not imply that IPW should not have been carried out – IPW using Minimum Covariate Imbalance criteria led to lower MSE values than the naive unweighted estimator. The results are discussed in greater detail in the next subsection.

Simulation II Results: Real-World Data

Table 2.41: MSE Comparison of different Selection Criteria. Sample size $n = 100$

	Bias	Std Dev	RMSE
Unweighted	1323.074	3060.429	3334.179
IPW:Probit	21.758	2703.193	2703.281
IPW: Min Imbalance	594.483	2531.254	2600.126
IPW: Min Error	272.230	3073.874	3085.905
IPW: Min Cal	300.052	2559.729	2577.255
IPW: Max Pred Lhd	382.014	2541.475	**2570.025**

Figure 2.4: Models selected via different Selection Criteria. Sample size $n = 100$

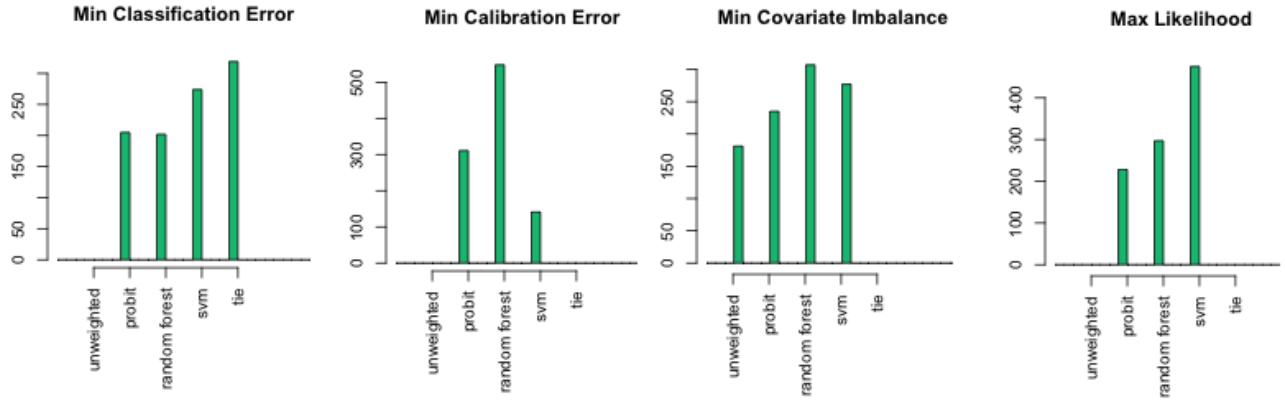


Table 2.42: MSE Comparison of different Selection Criteria. Sample size $n = 150$

	Bias	Std Dev	RMSE
Unweighted	1018.047	2452.824	2655.704
IPW:Probit	-84.357	2518.630	2520.042
IPW: Min Imbalance	239.168	2127.358	**2140.760**
IPW: Min Error	17.700	2573.285	2573.346
IPW: Min Cal	10.020	2594.955	2594.974
IPW: Max Pred Lhd	75.740	2564.426	2565.545

Figure 2.5: Models selected via different Selection Criteria. Sample size $n = 150$

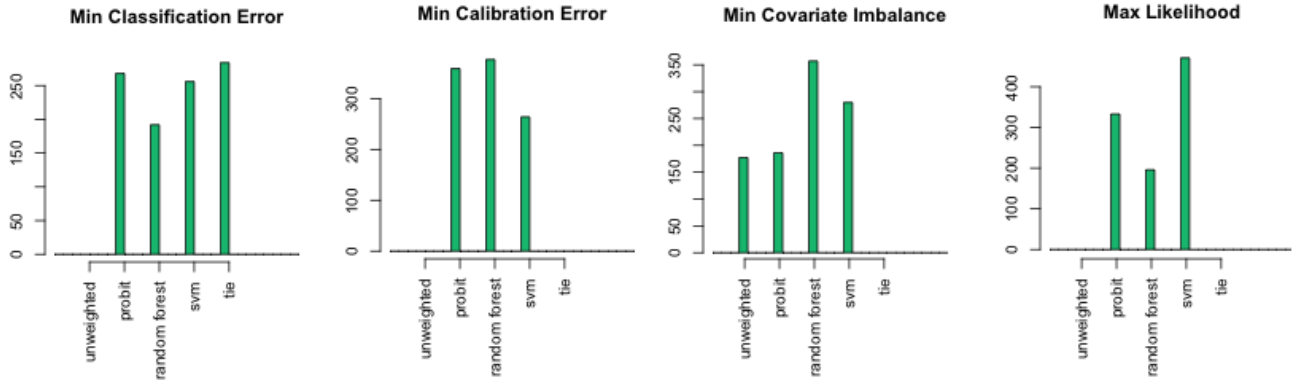


Table 2.43: MSE Comparison of different Selection Criteria. Sample size $n = 200$

	Bias	Std Dev	RMSE
Unweighted	1188.798	2042.307	2363.103
IPW:Probit	-157.001	2126.479	2132.267
IPW: Min Imbalance	214.026	1832.599	**1845.054**
IPW: Min Error	-146.113	2056.627	2061.811
IPW: Min Cal	-224.094	2443.016	2453.273
IPW: Max Pred Lhd	-120.755	2154.139	2157.521

Figure 2.6: Models selected via different Selection Criteria. Sample size $n = 200$

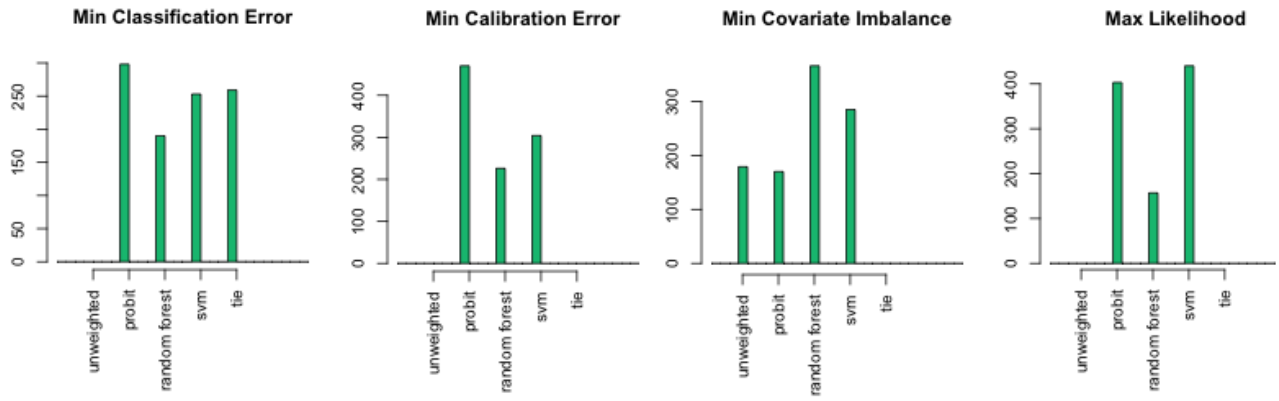


Table 2.44: MSE Comparison of different Selection Criteria. Sample size $n = 250$

	Bias	Std Dev	RMSE
Unweighted	1154.970	1695.977	2051.900
IPW:Probit	-136.612	2203.581	2207.811
IPW: Min Imbalance	192.308	1801.890	**1812.124**
IPW: Min Error	-194.114	1993.985	2003.411
IPW: Min Cal	-218.828	2321.352	2331.644
IPW: Max Pred Lhd	-215.283	2190.569	2201.122

Figure 2.7: Models selected via different Selection Criteria. Sample size $n = 250$

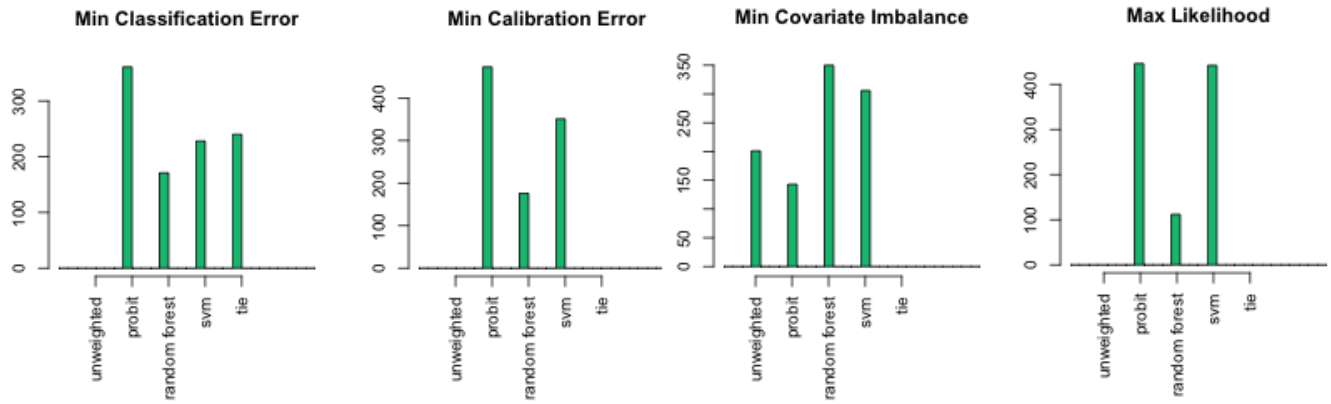


Table 2.45: MSE Comparison of different Selection Criteria. Sample size $n = 300$

	Bias	Std Dev	RMSE
Unweighted	1188.754	1526.210	1934.542
IPW:Probit	-192.041	2015.745	2024.872
IPW: Min Imbalance	223.211	1615.824	**1631.168**
IPW: Min Error	-274.553	1958.884	1978.031
IPW: Min Cal	-253.114	2004.573	2020.490
IPW: Max Pred Lhd	-244.354	2101.978	2116.133

Figure 2.8: Models selected via different Selection Criteria. Sample size $n = 300$

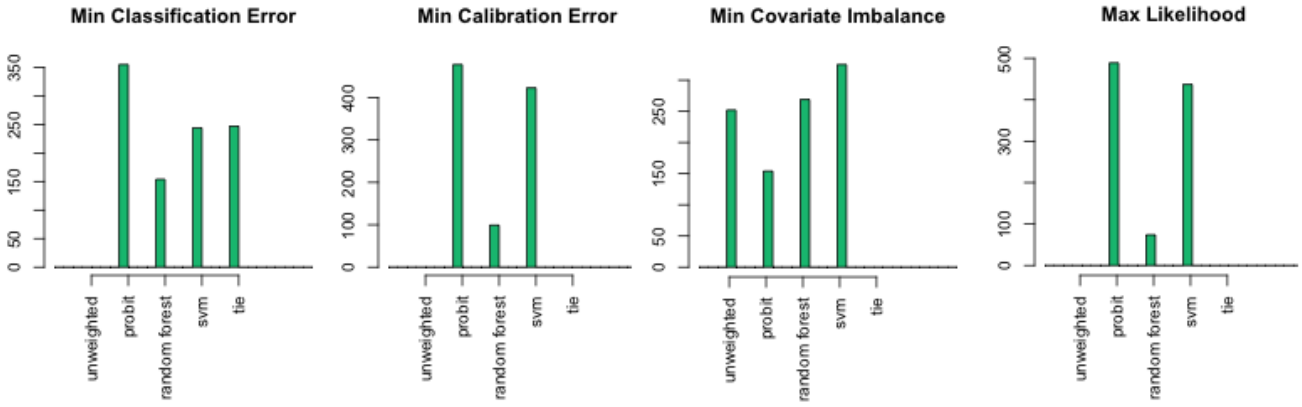
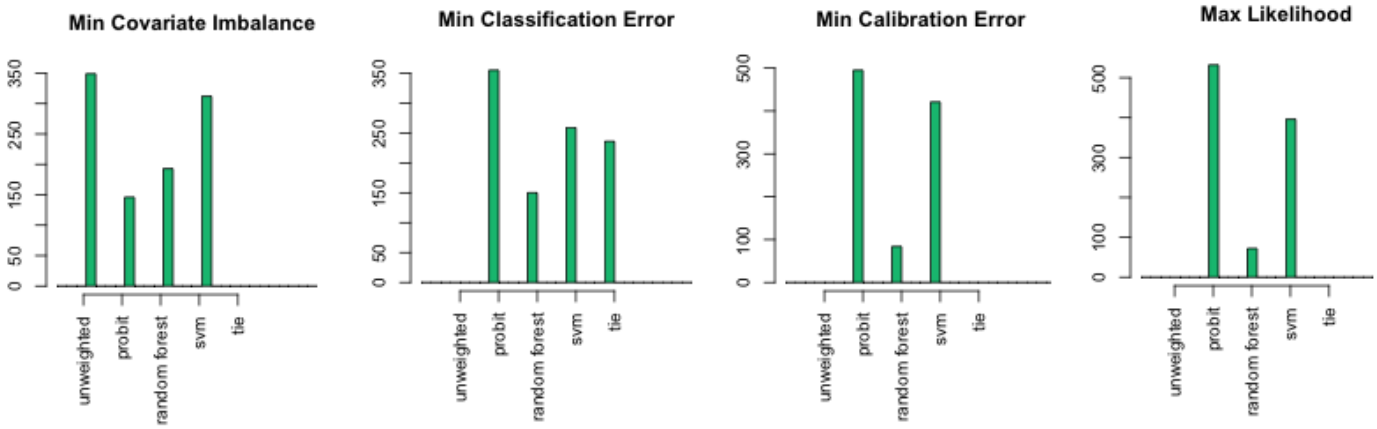


Table 2.46: MSE Comparison of different Selection Criteria. Sample size $n = 350$

	Bias	Std Dev	RMSE
Unweighted	1177.179	1364.867	1802.391
IPW:Probit	-228.027	2004.155	2017.086
IPW: Min Imbalance	338.219	1660.867	**1694.955**
IPW: Min Error	-413.370	2082.864	2123.487
IPW: Min Cal	-322.249	2123.254	2147.569
IPW: Max Pred Lhd	-281.865	1954.040	1974.264

Figure 2.9: Models selected via different Selection Criteria. Sample size $n = 350$



2.5.4 Insights from Simulation Experiments

Next, we discuss in greater detail the broad insights from the two simulation experiments described in Section 2.5.2 and Section 2.5.3. As in the case of all simulation studies, the results in this section come with the caveat of not being general. Further work is necessary to consolidate and advance this research direction.

Minimum Covariate Imbalance Selection Creiteria

We find throughout our simulations that for sample sizes of $n \geq 250$ observations, the IPW estimator using propensity score models selected via the Minimum Covariate Balance Criteria consistently dominates the IPW estimator using the linear probit model in terms of MSE. The Minimum Covariate Imbalance criteria picks a combination of the different estimation techniques. We note that in some cases, the balance of the data is *made worse* by weighting. In such cases the Minimum Covariate Imbalance measure picks the naive unweighted estimate of ATE. We also note that when there are non-linearities in the true selection equations IPW using linear probit can have very high associated standard errors. Techniques like Random Forest, Naive Bayes and Support Vector Machines may be very useful in these cases since they are associated with much lower standard errors.

This implies that for the applied econometrician it makes sense to consider a variety of parametric and non-parametric propensity score estimation models and pick the models that best serve the objective of IPW – i.e. to obtain better covariate balance between the treatment and control group. We recommend using the Minimum Covariate Imbalance criteria to select propensity score models for IPW.

Minimum Classification Error Selection Creiteria

The results on the IPW estimator using propensity score models selected via the Minimum Classification Error criteria is poor. In both sets of simulations the criteria has higher MSE values than IPW using linear probit in most cases. The poor performance of the selection criteria implies that the 'best classifier' may not necessarily be the best

propensity score model for IPW.

Minimum Calibration Error Selection Creiteria

From the simulation results we note that IPW models selected via the Minimum Calibration Error criteria perform poorly in both sets of experiments. The MSE values are higher than those from IPW with linear probit in a majority of the cases in both experiments. Moreover this selection criteria is sensitive to bin sizes. Thus there may be room for manipulation of results. Based on the simulation results and potential issues with bin sizes we recommend not using this selection criteria for applied research.

Maximum Predicted Likelihood Selection Criteria

From the simulation results we note that IPW models selected via the Maximum Predicted Likelihood criteria is poor. Whereas in a handful of cases the criteria has the lowest MSE values, in majority of the cases it performs worse than IPW using linear probit.

Linear Probit Model

There has been a lot of criticism of the linear probit model in the propensity score literature. In the second set of simulations we find evidence that the criticism is valid. In cases where the sample size $n \geq 300$, the naive unweighted estimator has lower MSE values than IPW with linear probit, despite the significantly lower values of bias associated with IPW with linear probit.

However, the MSE values of IPW using Minimum Covariate Imbalance criteria are lower than the unweighted naive estimator across all the simulations that we studied. And the linear probit propensity score model is picked very often by this criteria in all simulations. Moreover a study of the models picked by Minimum Calibration Error indicates that in many cases the linear probit does a good job of approximating the true probabilites, despite the non-linearities introduced into the model. Therefore we recommend keeping

the linear probit as one of the candidate IPW models when selecting a propensity score model, rather than avoiding it completely.

Naive Unweighted Estimator

In both sets of experiments, the unweighted estimator is picked by the Minimum Covariate Imbalance criteria in some of the simulations. This is consistent with the recommendation of many authors, who caution against IPW when covariate balance is not improved by weighting. By incorporating a check for covariate balance improvement into the selection procedure, the Minimum Covariate Imbalance criteria guards against careless use of IPW.

Bias Variance Trade-off

IPW using linear probit led to consistently lower bias than the unweighted naive estimator in all our simulations – therefore we find that *IPW with linear probit leads to a more unbiased estimate of ATE*, even with a poorly specified propensity score model. On the other hand, the *unweighted naive estimator consistently had much lower standard errors*, leading to lower MSE in some cases. Typically, estimates from IPW using the Minimum Covariate Imbalance criteria had higher bias than IPW with linear probit, but the lower standard error values lead to lower MSE values. This indicates that there may be a Bias Variance Trade-off embedded in the selection criteria.

2.6 Summary and Possible Extensions

The use of Propensity Score techniques for causal inference in social science has gained a lot of popularity in the last decade. However the most common models for the estimation of Propensity Scores i.e. Logit and Probit have come under a lot of scrutiny. The main criticism being that when these models are poorly specified, then the low bias property of propensity score techniques comes at the cost of extremely high values standard errors. Given this issue, the availability of large datasets, greater computing power as well as the many advances made in statistics and machine learning, it is natural that some authors

have recently introduced non-paramateric and machine learning techniques to replace Logit/Probit for the estimation of Propensity Scores. There is no reason why a single type of Propensity Score model should be the best choice in all scenarios, but there is a paucity of papers on choosing between different Propensity Score models. This paper aims to fill this gap in the literature.

We considered four criteria for selecting the 'best' propensity score model for estimating ATE in the IPW framework – Minimum Covariate Imbalance criteria, Minimum Classification Error criteria, Minimum Calibration Error criteria and Maximum Predicted Likelihood criteria. We found via two sets of simulations (using artificial data and real-world data) that ATE estimates from IPW using Minimum Covariate Imbalance propensity scores had lower MSE values than Linear Probit in all cases with sample size $n \geq 250$. We also found that the best classifier (with Minimum Classisfication Error) did not necessarily lead to lower MSE values than the Linear Probit, while the Minimum Calibration Error and Maximum Predicted Likelihood criteria performed poorly in general. Based on our results, we recommend estimating a number of propensity score models and picking the one with Minimum Covariate Imbalance before conducting IPW. We also found that while the Linear Probit estimates of ATE were in fact associated with large standard errors, in a large proportion of our simulations the model was selected via the Minimum Covariate Imbalance criteria. Similarly, the model was also selected frequently by the other three criteria. Therefore, in many cases the Linear Probit performs quite well in comparison to the other semiparametric models, despite being misspecified.

Future research directions include introducing model selection techniques and new propensity score estimation algorithms within the structural estimation framework.

Appendix

A1: Implementation in R

We used the R statistical package to run our simulations experiments and empirical estimation. R is an open source programming language which is becoming extremely popular among computer scientists and statisticians. There are a number of competing packages that can be used to perform Machine Learning estimation, including some we used in our analysis as well.

Estimation of Propensity Scores

- **Logit/ Probit:** These are in-built into R. They can be run using the *'glm'* command with an option of *family = "binomial" / family = "normal"*. Probability values are automatically returned with the *predict* command.

- **Naive Bayes:** We used the command *'naiveBayes'* from the package *'e1071'*. Probability values are automatically returned with the *predict* command.

- **Random Forest:** We used the command *'randomForest'* from the package *'randomForest'*. Class probabilities can be retrieved using the option *prob = T* in the prediction step.

- **Support Vector Machines:** We used the command *'LiblineaR'* with the option *type = 1*, from the package *'LiblineaR'* to classify the observations. To compute probabilities we refered to the source code of package *'ksvm'* and fit a sigmoid function to the predicted classes.

Miscellaneous

- **Trimming:** It is common practice to 'trim' datasets by dropping observations associated with extreme probability values when using propensity score techniques. Critics of trimming point out that this may lead to loss of important information.

In our code, we do allow for trimming by dropping observations with: $\widehat{e}_h \geq trim$ and $W = 0$ or $\widehat{e}_h \leq (1 - trim)$ and $W = 1$. However we only used trim values of 1, so that only those observations which corresponding to infinite weights were dropped.

- **Naive Bayes for Continuous Covariates:** We used the Naive Bayes classifier only when all covariates are continuous (Simulation Set I).

- **Random Forest with Large Dataset:**The only place where we did not use default package settings was in the Random Forest component of the empirical investigation due to the large sizes of the datasets ($> 100,000$). We set the option *nodesize =5* instead of the default *nodesize =1*.

The release of a package in R that runs the propensity score models as well as selection techniques described in this is being planned. Codes are available on request.

A2: Technical Notes

This section provides detailed analytical results that show that in the presence of selection bias, IPW leads to unbiased estimates of the treatment effect (in the linear reggression framework). These results are not new, however since we were unable to locate a detailed proof we attempted to reproduce the result independently.

2.7 Framework and Assumptions

- **Selection Equation:** An underlying process which determines which units get treatment. In our analysis this is:

$$x_i = z_i + v_i, \qquad i \in (i, \cdots, N).$$

where z_i is a scalar covariate (for example some individual characteristic) v_i is an error term. The rule determining whether a unit gets *selected* into the treatment group is:

$$w_i = \begin{cases} 1 & \text{if } x_i \geq 0 \qquad i \in (i, \cdots, N) \\ 0 & \text{if } x_i < 0 \end{cases}$$

- **Outcome Equation:** The true process that links the outcome of interest with the treatment. In our analysis this is:

$$y_i = \alpha_0 + b_0 w_i + \gamma_0 z_i + u_i, \qquad i \in (i, \cdots, N).$$

where u_i is an error term. Note that the coefficient of w_i in this equation, b_0 is the *true average treatment effect*.

- **Source of bias:** If the correctly specified outcome equation is estimated, then the estimate of the treatment effect is unbiased. However if some variables that affect both the selection and outcome equation are unobserved (or ommited), then the estimate of the treatment effect is biased upwards, i.e. the effect of the treatment

is *overestimated*. In our analysis, the source of bias comes from a **mis-specified outcome equation**:

$$y_i = \tau + \rho w_i + \epsilon_i, \qquad \epsilon_i = \gamma_0 z_i + u_i, \qquad i \in (i, \cdots, N).$$

This equation suffers from an **omitted variable bias** because the error term ϵ_i is correlated with regressor w_i via the unobserved varible z_i.

- **Assumptions:** For the above to hold true, we need some independence assumptions.

$$v_i \perp z_i, \quad u_i \perp z_i, \quad v_i \perp u_i.$$

In addition we make some distributional assumptions which we use in the proof (however we relax these assumptions towards the end).

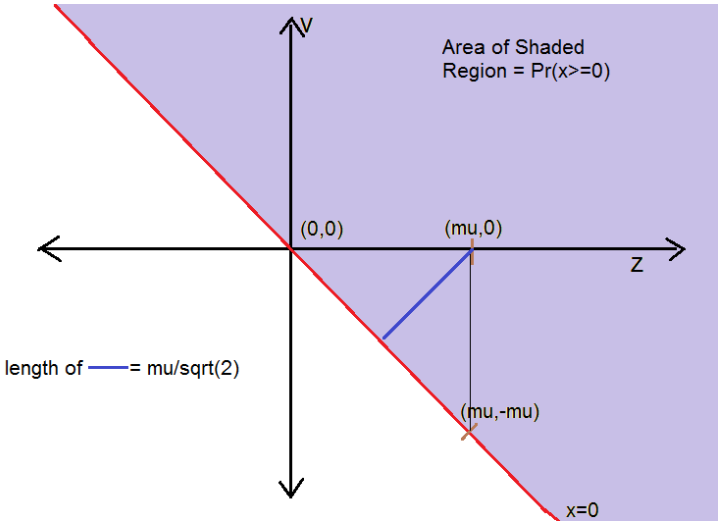
$$v_i \sim N(0,1), \quad u_i \sim N(0,1), \quad z_i \sim N(\mu, 1)$$

- **Bias Correction by Propensity Score Weighting:** In order to correct the bias in the estimate of the treatment effect, literature suggests running *Weighted Least Squares* **(WLS)** by weighting treated units with the *inverse of probability of being treated* and weighting control units with the *inverse of probability of being control*.

$$s_i = \begin{cases} (Pr(x_i \geq 0|z_i))^{-1} & \text{if } w_i = 1 \qquad i \in (i, \cdots, N) \\ \\ (1 - Pr(x_i \geq 0|z_i))^{-1} & \text{if } w_i = 0 \end{cases}$$

2.8 Some Useful Calculations

Here we calculate the values of some useful terms that will show up in subsequent derivations. We start with a figure that represents $Pr(x_i \geq 0)$.



In this figure the two axes map the values of the v_i and z_i. Recall that $v_i \perp z_i$ and $x_i = f(v_i, z_i) = v_i + z_i$. Thus the relevant probability distribution for x_i is $f_x(x) = f_{vz}(v, z) = f_v(v) \cdot f_z(z)$. Further recall that $v \sim N(0, 1)$, thus the pdf and cdf of v respectively are

$$f_v(v) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-v^2}{2}\right) = \phi_v(v),$$

$$\int_{-\infty}^{a} f_v(v)dv = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-v^2}{2}\right) dv = \Phi_v(a), \qquad -\infty < a < \infty$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ represent the pdf and cdf corresponding to the standard normal distribution respectively. Similarly since $z \sim N(\mu, 1)$, the pdf and cdf of z respectively are

$$f_z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right),$$

$$\int_{-\infty}^{a} f_z(z)dz = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) dz,, \qquad -\infty < a < \infty.$$

Thus the pdf of x_i is:

$$f_x(x) = f_v(v) \cdot f_z(z) = \phi_v(v) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right). \tag{2.7}$$

The shaded area represents the region where $x_i \geq 0$. Note that since $x_i = z_i + v_i$, therefore

$$x_i \geq 0 \Rightarrow z_i + v_i \geq 0 \Rightarrow v_i \geq -z_i.$$

Thus the shaded region also represents the region where $v_i \geq -z_i$.

- $Pr(x_i \geq 0)$: From (2.7) the relevant pdf is:

$$f_x(x) = f_v(v) \cdot f_z(z) = \phi_v(v) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right).$$

In order to find the area of the shaded region ($v_i \geq -z_i$), integrate over the entire range of z_i and over the range of v_i greater than $-z_i$. We drop the i subscript for now.

$$Pr(x \geq 0) = \int_{-\infty}^{\infty} \int_{-z}^{\infty} \phi_v(v) dv \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) dz$$

This double integral is easier to solve if we make the following transformation:

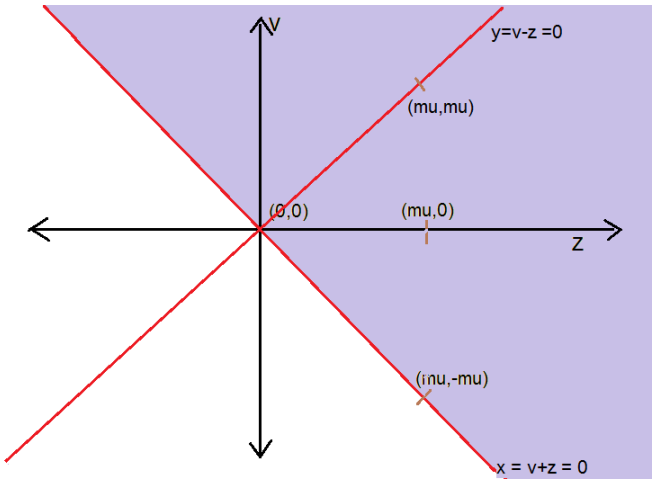
. $x = v + z$,

. $y = v - z$.

Notice that the ranges of integration in the new system are:

. x from 0 to ∞,

. y from $-\infty$ to ∞.



Using the Jacobian transformation we have:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} v \\ z \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} v \\ z \end{bmatrix} = \frac{1}{(-2)} \begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} v \\ z \end{bmatrix} = \begin{bmatrix} (x+y)/2 \\ (x-y)/2 \end{bmatrix}$$

Finally,

$$dxdy = \frac{1}{2} dvdz$$

where $dvdz = |J| \cdot dxdy$ and $J = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$.

Use the above to evaluate:

$$\begin{aligned} Pr(x \geq 0) &= \int_{-\infty}^{\infty} \int_{-z}^{\infty} \phi_v(v) dv \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) dz \\ &= \int_{-\infty}^{\infty} \int_{-z}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right) dv \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) dz \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{((x+y)/2)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{((x-y)/2-\mu)^2}{2}\right) dxdy \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \int_{0}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{((x+y)/2)^2 + ((x-y)/2-\mu)^2}{2}\right) dxdy \\ &\vdots \end{aligned}$$

(algebra – complete the squares in the exponential term)

$$\vdots$$

$$= \frac{1}{2}\int_{-\infty}^{\infty}\int_{0}^{\infty}\frac{1}{2\pi}\exp\left(-\frac{(x-\mu)^2}{2\times 2}-\frac{(y+\mu)^2}{2\times 2}\right)dxdy$$

$$= \frac{1}{\sqrt{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi}}\exp\left(-\frac{(y+\mu)^2}{2\times 2}\right)dy\int_{0}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\times 2}\right)dx$$

(the first integral is the normal pdf $N(-\mu, 2)$ evaluated over $\mathcal{R}$)

$$= \frac{1}{\sqrt{2}}\cdot 1\cdot\int_{0}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\times 2}\right)dx$$

$$\vdots$$

(change of variable) $t = \dfrac{x-\mu}{\sqrt{2}}$

$$\vdots$$

$$= \int_{-\mu/\sqrt{2}}^{\infty}\phi_t(t)dt$$

$$= \Phi(\mu/\sqrt{2})$$

Therefore,

$$Pr(x \geq 0) = \Phi(\mu/\sqrt{2}).$$

- $Pr(x_i \geq 0|z_i)$: Using the rules of probability and the distribution of $v \sim N(0,1)$,

$$Pr(x \geq 0|z) = Pr(v \geq -z) = \Phi(-z)$$

- $E(z_i)$: This is simply the unconditional mean of z which is μ. (Recall $z \sim N(\mu, 1)$.)

  Thus,

$$E(z) = \mu.$$

- $E(z_i|x_i \geq 0)$:

$$E(z|x \geq 0) = \int_{-\infty}^{\infty} z \int_{-z}^{\infty} \phi_v(v)dv \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) dz$$

(change of variables $t = z - \mu$)

$$= \int_{-\infty}^{\infty} (t+\mu) \int_{-(t+\mu)}^{\infty} \phi_v(v)dv \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

$$= \int_{-\infty}^{\infty} t \int_{-(t+\mu)}^{\infty} \phi_v(v)dv \cdot \phi_t(t)dt + \int_{-\infty}^{\infty} \mu \int_{-(t+\mu)}^{\infty} \phi_v(v)dv \cdot \phi_t(t)dt$$

$$= \text{Integral 1} + \text{Integral 2.}$$

Solve the two integrals separately,

$$\text{Integral 1} = \int_{-\infty}^{\infty} t \int_{-(t+\mu)}^{\infty} \phi_v(v)dv \cdot \phi_t(t)dt$$

$$= \int_{-\infty}^{\infty} t \cdot \Phi_v(t+\mu)\phi_t(t)dt$$

$(\int udv = uv| - \int vdu$ where $u = \Phi(t+\mu), dv = t\phi(t)$

$$= [\Phi(t+\mu)(-\phi(t))]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \phi(t)\phi(t+\mu)dt$$

$$(\because \Phi(-\infty) = 0, \Phi(\infty) = 0 \text{ and } \phi(\infty) = 0)$$

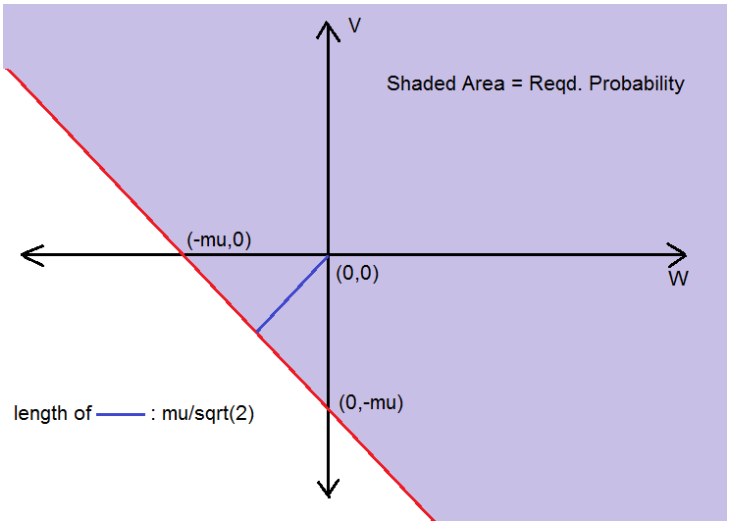
$$= 0 + \int_{-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{w^2 + w^2 + \mu^2 + 2w\mu}{2}\right)$$

(complete the square in the exponent to get)

$$= \frac{1}{2\sqrt{\pi}} \exp\left\{-\mu^2/4\right\}$$

$$= \frac{1}{\sqrt{2}} \cdot \phi(\mu/\sqrt{2}).$$

In order to solve Integral 2, consider the following graphical representation of the problem.



Note that this is (almost) identical to solving for $Pr(x_i \geq 0)$ (refer to (2.8)). Using similar calculations,

$$\text{Integral } 2 = \mu \cdot \Phi(\mu/\sqrt{2}).$$

Add the two integrals,

$$\therefore E(z|x \geq 0) = \frac{1}{\sqrt{2}} \cdot \phi(\mu/\sqrt{2}) + \mu \cdot \Phi(\mu/\sqrt{2}). \tag{2.8}$$

- $E(s_i)$: Recall that,

$$s_i = \begin{cases} (Pr(x_i \geq 0|z_i))^{-1} & \text{if } w_i = 1 \qquad i \in (i, \cdots, N) \\ \\ (1 - Pr(x_i \geq 0|z_i))^{-1} & \text{if } w_i = 0 \end{cases}$$

$$E(s) = \int_{-\infty}^{\infty} \frac{1}{1 - Pr(x_i \geq 0|z)} \int_{-\infty}^{-z} \phi(v)dv \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) dv$$

$$+ \int_{-\infty}^{\infty} \frac{1}{Pr(x_i \geq 0|z)} \int_{-z}^{\infty} \phi(v)dv \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) dv$$

$$= \int_{-\infty}^{\infty} \frac{1}{1 - Pr(v_i \geq -z)} \Phi(-z) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) dv$$

$$+ \int_{-\infty}^{\infty} \frac{1}{Pr(v_i \geq -z)} (1 - \Phi(-z)) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) dv$$

$(\because Pr(v_i \geq -z) = 1 - \Phi(-z)$ and the integration of a pdf over its support $= 1)$

$$= 1 \cdot 1 + 1 \cdot 1$$

$$= 2.$$

$$\therefore E(s) = 2.$$

- $E(s_i|x_i \geq 0)$: We repeat the same calculations as above with the modification that the support of v is only from $-z$ to ∞ to get:

$$E(s|x \geq 0) = 1. \tag{2.9}$$

- $E(s_i z_i)$: Using calculations similar to (2.8)

$$E(sz) = \int_{-\infty}^{\infty} \frac{z}{1 - Pr(x_i \geq 0|z)} \int_{-\infty}^{-z} \phi(v)dv \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) dv$$

$$+ \int_{-\infty}^{\infty} \frac{z}{Pr(x_i \geq 0|z)} \int_{-z}^{\infty} \phi(v)dv \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) dv$$

$$= \int_{-\infty}^{\infty} \frac{z}{1 - Pr(v_i \geq -z)} \Phi(-z) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) dv$$

$$+ \int_{-\infty}^{\infty} \frac{z}{Pr(v_i \geq -z)} (1 - \Phi(-z)) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) dv$$

$$(\because Pr(v_i \geq -z) = 1 - \Phi(-z) \text{ and } E(z) = \mu)$$

$$= 1 \cdot \mu + 1 \cdot \mu$$

$$= 2\mu.$$

$$\therefore E(sz) = 2\mu. \tag{2.10}$$

- $E(s_i z_i | x_i \geq 0)$: We repeat the same calculations as above with the modification that the support of v is only from $-z$ to ∞ to get:

$$E(sz|x \geq 0) = \mu. \tag{2.11}$$

Bias Calculation for Fully Specified Model

- **True Outcome Equation:** $y_i = \alpha_0 + b_0 w_i + \gamma_0 z_i + u_i$,
- **Estimated Outcome Equation:** $y_i = \tau + \rho w_i + \theta z_i + u_i$.

OLS Estimates

$$\begin{bmatrix} \widehat{\tau} \\ \widehat{\rho} \\ \widehat{\theta} \end{bmatrix} = \begin{bmatrix} \Sigma_i^N 1 & \Sigma_i^N w_i & \Sigma_i^N z_i \\ \Sigma_i^N w_i & \Sigma_i^N (w_i \cdot w_i) & \Sigma_i^N (z_i \cdot w_i) \\ \Sigma_i^N z_i & \Sigma_i^N (w_i \cdot z_i) & \Sigma_i^N (z_i \cdot z_i) \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_i^N y_i \\ \Sigma_i^N (w_i \cdot y_i) \\ \Sigma_i^N (z_i \cdot yi) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\Sigma_i^N 1}{N} & \frac{\Sigma_i^N w_i}{N} & \frac{\Sigma_i^N z_i}{N} \\ \frac{\Sigma_i^N w_i}{N} & \frac{\Sigma_i^N (w_i)}{N} & \frac{\Sigma_i^N (z_i \cdot w_i)}{N} \\ \frac{\Sigma_i^N z_i}{N} & \frac{\Sigma_i^N (w_i \cdot z_i)}{N} & \frac{\Sigma_i^N (z_i \cdot z_i)}{N} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\Sigma_i^N y_i}{N} \\ \frac{\Sigma_i^N (w_i \cdot y_i)}{N} \\ \frac{\Sigma_i^N (z_i \cdot yi)}{N} \end{bmatrix}$$

(note that $w_i^2 = w_i$ since w_i is a vector of zeros (control units) and ones (treated units))

(replacing with true model $y_i = \alpha_0 + b_0 w_i + \gamma_0 z_i + u_i$)

$$= \begin{bmatrix} \frac{\Sigma_i^N 1}{N} & \frac{\Sigma_i^N w_i}{N} & \frac{\Sigma_i^N z_i}{N} \\ \frac{\Sigma_i^N w_i}{N} & \frac{\Sigma_i^N (w_i)}{N} & \frac{\Sigma_i^N (z_i \cdot w_i)}{N} \\ \frac{\Sigma_i^N z_i}{N} & \frac{\Sigma_i^N (w_i \cdot z_i)}{N} & \frac{\Sigma_i^N (z_i \cdot z_i)}{N} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\Sigma_i^N \cdot (\alpha_0 + b_0 w_i + \gamma_0 z_i + u_i)}{N} \\ \frac{\Sigma_i^N (w_i \cdot (\alpha_0 + b_0 w_i + \gamma_0 z_i + u_i))}{N} \\ \frac{\Sigma_i^N (z_i \cdot (\alpha_0 + b_0 w_i + \gamma_0 z_i + u_i))}{N} \end{bmatrix}$$

(since w_i is a vector of ones, the dot product picks up values corresponding only to the treated units (units with $w_i = 1$ ie $x_i \geq 0$))

$$\underset{p}{\rightarrow} \begin{bmatrix} 1 & Pr(x_i \geq 0) & E(z_i) \\ Pr(x_i \geq 0) & Pr(x_i \geq 0) & E(z_i|x_i \geq 0) \\ E(z_i) & E(z_i|x_i \geq 0) & E(z_i^2) \end{bmatrix}^{-1} \times$$

$$\begin{bmatrix} \alpha_0 + b_0 Pr(x_i \geq 0) + \gamma_0 E(z_i) + E(u_i) \\ \alpha_0 Pr(x_i \geq 0) + b_0 Pr(x_i \geq 0) + \gamma_0 E(z_i|x_i \geq 0) + E(u_i|x_i \geq 0) \\ \alpha_0 E(z_i) + b_0 E(z_i|x_i \geq 0) + \gamma_0 E(z_i^2) + E(u_i|z_i) \end{bmatrix}$$

$$= \begin{bmatrix} 1 & Pr(x_i \geq 0) & E(z_i) \\ Pr(x_i \geq 0) & Pr(x_i \geq 0) & E(z_i|x_i \geq 0) \\ E(z_i) & E(z_i|x_i \geq 0) & E(z_i^2) \end{bmatrix}^{-1} \times$$

$$\begin{bmatrix} 1 & Pr(x_i \geq 0) & E(z_i) \\ Pr(x_i \geq 0) & Pr(x_i \geq 0) & E(z_i|x_i \geq 0) \\ E(z_i) & E(z_i|x_i \geq 0) & E(z_i^2) \end{bmatrix} \begin{bmatrix} \alpha_0 \\ b_0 \\ \gamma_0 \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ b_0 \\ \gamma_0 \end{bmatrix}$$

Bias Calculation for Mis-specified Model

- **True Outcome Equation:** $y_i = \alpha_0 + b_0 w_i + \gamma_0 z_i + u_i$,
- **Estimated Outcome Equation:** $y_i = \tau + \rho w_i + \varepsilon_i, \qquad (\because \varepsilon_i = \gamma_0 z_i + u_i.)$

OLS estimates

$$\begin{bmatrix} \widehat{\tau} \\ \\ \widehat{\rho} \end{bmatrix} = \begin{bmatrix} \Sigma_i^N 1 & \Sigma_i^N w_i \\ \\ \Sigma_i^N w_i & \Sigma_i^N (w_i \cdot w_i) \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_i^N y_i \\ \\ \Sigma_i^N (w_i \cdot y_i) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\Sigma_i^N 1}{N} & \frac{\Sigma_i^N w_i}{N} \\ \\ \frac{\Sigma_i^N w_i}{N} & \frac{\Sigma_i^N (w_i)}{N} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\Sigma_i^N y_i}{N} \\ \\ \frac{\Sigma_i^N (w_i \cdot y_i)}{N} \end{bmatrix}$$

(replacing with true model $y_i = \alpha_0 + b_0 w_i + \gamma_0 z_i + u_i$)

$$= \begin{bmatrix} \frac{\Sigma_i^N 1}{N} & \frac{\Sigma_i^N w_i}{N} \\ \\ \frac{\Sigma_i^N w_i}{N} & \frac{\Sigma_i^N (w_i)}{N} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\Sigma_i^N (\alpha_0 + b_0 w_i + \gamma_0 z_i + u_i)}{N} \\ \\ \frac{\Sigma_i^N (w_i \cdot (\alpha_0 + b_0 w_i + \gamma_0 z_i + u_i))}{N} \end{bmatrix}$$

$$\underset{p}{\to} \begin{bmatrix} 1 & Pr(x_i \geq 0) \\ \\ Pr(x_i \geq 0) & Pr(x_i \geq 0) \end{bmatrix}^{-1} \times$$

$$\begin{bmatrix} \alpha_0 + b_0 Pr(x_i \geq 0) + \gamma_0 E(z_i) + E(u_i) \\ \\ \alpha_0 Pr(x_i \geq 0) + b_0 Pr(x_i \geq 0) + \gamma_0 E(z_i | x_i \geq 0) + E(u_i | x_i \geq 0) \end{bmatrix}$$

$$= \begin{bmatrix} 1 & Pr(x_i \geq 0) \\ Pr(x_i \geq 0) & Pr(x_i \geq 0) \end{bmatrix}^{-1} \begin{bmatrix} 1 & Pr(x_i \geq 0) \\ Pr(x_i \geq 0) & Pr(x_i \geq 0) \end{bmatrix} \begin{bmatrix} \alpha_0 \\ b_0 \end{bmatrix}$$
$$+ \begin{bmatrix} 1 & Pr(x_i \geq 0) \\ Pr(x_i \geq 0) & Pr(x_i \geq 0) \end{bmatrix}^{-1} \begin{bmatrix} \gamma_0 E(z_i) + E(u_i) \\ \gamma_0 E(z_i|x_i \geq 0) + E(u_i|x_i \geq 0) \end{bmatrix}$$

$$= \begin{bmatrix} \alpha_0 \\ b_0 \end{bmatrix} + \gamma_0 \begin{bmatrix} 1 & Pr(x_i \geq 0) \\ Pr(x_i \geq 0) & Pr(x_i \geq 0) \end{bmatrix}^{-1} \begin{bmatrix} E(z_i) \\ E(z_i|x_i \geq 0) \end{bmatrix}$$

$$\therefore \textbf{ bias} \underset{p}{\rightarrow} \gamma_0 \begin{bmatrix} 1 & Pr(x_i \geq 0) \\ Pr(x_i \geq 0) & Pr(x_i \geq 0) \end{bmatrix}^{-1} \begin{bmatrix} E(z_i) \\ E(z_i|x_i \geq 0) \end{bmatrix}$$

$$= \frac{\gamma_0}{Pr(x_i \geq 0)\left(1 - Pr(x_i \geq 0)\right)} \begin{bmatrix} Pr(x_i \geq 0) & -Pr(x_i \geq 0) \\ -Pr(x_i \geq 0) & 1 \end{bmatrix} \begin{bmatrix} E(z_i) \\ E(z_i|x_i \geq 0) \end{bmatrix}$$

$$= \frac{\gamma_0}{Pr(x_i \geq 0)\left(1 - Pr(x_i \geq 0)\right)} \begin{bmatrix} Pr(x_i \geq 0)E(z_i) - Pr(x_i \geq 0)E(z_i|x_i \geq 0) \\ -Pr(x_i \geq 0)E(z_i) + E(z_i|x_i \geq 0) \end{bmatrix}$$

$$= \gamma_0 \begin{bmatrix} \frac{E(z_i)}{(1-p_i)} - \frac{E(z_i|x_i \geq 0)}{(1-p_i)} \\ \frac{E(z_i|x_i \geq 0)}{p_i(1-p_i)} - \frac{E(z_i)}{(1-p_i)} \end{bmatrix}$$

(Based on our assumptions of normality and using (2.8), (2.8) and (2.8)) and where $p_i = Pr(x_i \geq 0)$)

$$= \gamma_0 \begin{bmatrix} \frac{\mu}{1-\Phi(\mu/\sqrt{2})} - \frac{\frac{1}{\sqrt{2}}\cdot\phi(\mu/\sqrt{2})+\mu\cdot\Phi(\mu/\sqrt{2})}{1-\Phi(\mu/\sqrt{2})} \\ \\ \frac{\frac{1}{\sqrt{2}}\cdot\phi(\mu/\sqrt{2})+\mu\cdot\Phi(\mu/\sqrt{2})}{\Phi(\mu/\sqrt{2})(1-\Phi(\mu/\sqrt{2}))} - \frac{\mu}{1-\Phi(\mu/\sqrt{2})} \end{bmatrix}$$

$$= \gamma_0 \begin{bmatrix} \mu - \frac{1}{\sqrt{2}}\frac{\phi(\mu/\sqrt{2})}{\Phi(\mu/\sqrt{2})(1-\Phi(\mu/\sqrt{2}))} \\ \\ \frac{1}{\sqrt{2}}\frac{\phi(\mu/\sqrt{2})}{\Phi(\mu/\sqrt{2})(1-\Phi(\mu/\sqrt{2}))} \end{bmatrix}$$

Note that the coefficient measuring the average treatment effect, $\widehat{\rho}$ is **biased upwards** $\forall \mu \in \mathbf{R}$.

2.9 Bias Calculation for Mis-specied Model with Weighting

- **True Outcome Equation:** $y_i = \alpha_0 + b_0 w_i + \gamma_0 z_i + u_i$,
- **Estimated Outcome Equation:** $y_i = \tau + \rho w_i + \epsilon_i, \quad (\therefore \epsilon_i = \gamma_0 z_i + u_i)$ weighted by s_i:

$$s_i = \begin{cases} (Pr(x_i \geq 0|z_i))^{-1} & \text{if } w_i = 1 \qquad i \in (i, \cdots, N) \\ \\ (1 - Pr(x_i \geq 0|z_i))^{-1} & \text{if } w_i = 0 \end{cases}$$

WLS estimates

$$\begin{bmatrix} \widehat{\tau} \\ \\ \widehat{\rho} \end{bmatrix} = \begin{bmatrix} \Sigma_i^N 1 \cdot s_i & \Sigma_i^N (w_i \cdot s_i) \\ \\ \Sigma_i^N (w_i \cdot s_i) & \Sigma_i^N (w_i \cdot s_i \cdot w_i) \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_i^N (s_i \cdot y_i) \\ \\ \Sigma_i^N (w_i \cdot s_i \cdot y_i) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\Sigma_i^N s_i}{N} & \frac{\Sigma_i^N (w_i \cdot s_i)}{N} \\ \\ \frac{\Sigma_i^N (w_i \cdot s_i)}{N} & \frac{\Sigma_i^N (w_i \cdot s_i)}{N} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\Sigma_i^N (s_i \cdot y_i)}{N} \\ \\ \frac{\Sigma_i^N (w_i \cdot s_i \cdot y_i)}{N} \end{bmatrix}$$

(replacing with true model $y_i = \alpha_0 + b_0 w_i + \gamma_0 z_i + u_i$)

$$= \begin{bmatrix} \frac{\Sigma_i^N s_i}{N} & \frac{\Sigma_i^N (w_i \cdot s_i)}{N} \\ \\ \frac{\Sigma_i^N (w_i \cdot s_i)}{N} & \frac{\Sigma_i^N (w_i \cdot s_i)}{N} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\Sigma_i^N s_i (\alpha_0 + b_0 w_i + \gamma_0 z_i + u_i)}{N} \\ \\ \frac{\Sigma_i^N (w_i \cdot s_i (\alpha_0 + b_0 w_i + \gamma_0 z_i + u_i))}{N} \end{bmatrix}$$

$$\underset{p}{\to} \begin{bmatrix} E(s_i) & E(s_i | x_i \geq 0) \\ \\ E(s_i | x_i \geq 0) & E(s_i | x_i \geq 0) \end{bmatrix}^{-1} \times$$

$$\begin{bmatrix} \alpha_0 E(s_i) + b_0 E(s_i | x_i \geq 0) + \gamma_0 E(s_i z_i) + E(u_i s_i) \\ \\ \alpha_0 E(s_i | x_i \geq 0) + b_0 E(s_i | x_i \geq 0) + \gamma_0 E(s_i z_i | x_i \geq 0) + E(u_i s_i | x_i \geq 0) \end{bmatrix}$$

$$= \begin{bmatrix} E(s_i) & E(s_i | x_i \geq 0) \\ \\ E(s_i | x_i \geq 0) & E(s_i | x_i \geq 0) \end{bmatrix}^{-1} \begin{bmatrix} E(s_i) & E(s_i | x_i \geq 0) \\ \\ E(s_i | x_i \geq 0) & E(s_i | x_i \geq 0) \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \\ b_0 \end{bmatrix}$$

$$+\begin{bmatrix} E(s_i) & E(s_i|x_i \geq 0) \\ E(s_i|x_i \geq 0) & E(s_i|x_i \geq 0) \end{bmatrix}^{-1} \begin{bmatrix} \gamma_0 E(s_i \cdot z_i) + E(u_i \cdot s_i) \\ \gamma_0 E(s_i \cdot z_i|x_i \geq 0) + E(u_i \cdot s_i|x_i \geq 0) \end{bmatrix}$$

$$= \begin{bmatrix} \alpha_0 \\ b_0 \end{bmatrix} + \gamma_0 \begin{bmatrix} E(s_i) & E(s_i|x_i \geq 0) \\ E(s_i|x_i \geq 0) & E(s_i|x_i \geq 0) \end{bmatrix}^{-1} \begin{bmatrix} E(s_i \cdot z_i) \\ E(s_i \cdot z_i|x_i \geq 0) \end{bmatrix}$$

$$\therefore \textbf{ asymp. bias} = \gamma_0 \begin{bmatrix} E(s_i) & E(s_i|x_i \geq 0) \\ E(s_i|x_i \geq 0) & E(s_i|x_i \geq 0) \end{bmatrix}^{-1} \begin{bmatrix} E(s_i \cdot z_i) \\ E(s_i \cdot z_i|x_i \geq 0) \end{bmatrix} \tag{2.12}$$

Based on our normality assumptions and using (2.8), (2.9), (2.11)

$$\textbf{asymp. bias} = \gamma_0 \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2\mu \\ \mu \end{bmatrix}$$

$$= \frac{\gamma_0}{2-1} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 2\mu \\ \mu \end{bmatrix}$$

$$= \gamma_0 \begin{bmatrix} 2\mu - \mu \\ -2\mu + 2\mu \end{bmatrix} = \begin{bmatrix} \gamma_0 \cdot \mu \\ 0 \end{bmatrix}$$

Notice that the coefficient estimate associated with the treatment effect is now **unbiased**.

Coming up with the weights

Here we try to independently develop the appropriate weighting scheme to get an unbiased estimate of the treatment effect. We drop all distributional assumptions on v and z. Start with (2.12) and simplify to obtain the **asymptotic bias:**

$$\gamma_0 \begin{bmatrix} E(s_i) & E(s_i|x_i \geq 0) \\ E(s_i|x_i \geq 0) & E(s_i|x_i \geq 0) \end{bmatrix}^{-1} \begin{bmatrix} E(s_i \cdot z_i) \\ E(s_i \cdot z_i|x_i \geq 0) \end{bmatrix}$$

$$= \frac{\gamma_0}{E(s_i|x_i \geq 0)\left(E(s_i) - E(s_i|x_i \geq 0)\right)} \begin{bmatrix} E(s_i|x_i \geq 0) & -E(s_i|x_i \geq 0) \\ -E(s_i|x_i \geq 0) & E(s_i) \end{bmatrix} \begin{bmatrix} E(s_i z_i) \\ E(s_i z_i|x_i \geq 0) \end{bmatrix}$$

$$= \frac{\gamma_0}{E(s_i|x_i \geq 0)\left(E(s_i) - E(s_i|x_i \geq 0)\right)} \begin{bmatrix} E(s_i|x_i \geq 0)\left(E(s_i z_i) - E(s_i z_i|x_i \geq 0)\right) \\ -E(s_i|x_i \geq 0)E(s_i z_i) + E(s_i)E(s_i z_i|x_i \geq 0). \end{bmatrix}.$$

In order to get unbiased treatment effect estimate the numerator associated with $\widehat{\rho}$ should have no contribution. Therefore for an unbiased estimator,

$$E(s_i)E(s_i z_i|x_i \geq 0) - E(s_i|x_i \geq 0)E(s_i z_i) = 0 \tag{2.13}$$

Let's define:

$$s_i = \begin{cases} h(z_i) & \text{if } x_i \geq 0 \text{ i.e. } v_i \geq -z_i \\ r(z_i) & \text{if } x_i < 0 \text{ i.e. } v_i < -z_i \end{cases}$$

$$E(s_i) = \int_{-\infty}^{\infty} s_i \int_{-\infty}^{\infty} f_v(v)dv f_z(z)dz$$

$$= \int_{-\infty}^{\infty} r(z_i) \int_{-\infty}^{-z_i} f_v(v)dv f_z(z)dz + \int_{-\infty}^{\infty} h(z_i) \int_{-z_i}^{\infty} f_v(v)dv f_z(z)dz$$

$$= E(s_i|x_i < 0) + E(s_i|x_i > 0).$$

$$E(s_i z_i) = \int_{-\infty}^{\infty} z_i s_i \int_{-\infty}^{\infty} f_v(v)dv f_z(z)dz$$

$$= \int_{-\infty}^{\infty} z_i r(z_i) \int_{-\infty}^{-z_i} f_v(v)dv f_z(z)dz + \int_{-\infty}^{\infty} z_i h(z_i) \int_{-z_i}^{\infty} f_v(v)dv f_z(z)dz$$

$$= E(s_i z_i|x_i < 0) + E(s_i z_i|x_i > 0).$$

$$E(s_i)E(s_i z_i|x_i \geq 0) - E(s_i|x_i \geq 0)E(s_i z_i)$$

$$= \left(E(s_i|x_i < 0) + E(s_i|x_i > 0)\right) E(s_i z_i|x_i \geq 0)$$
$$- E(s_i|x_i \geq 0) \left(E(s_i z_i|x_i < 0) + E(s_i z_i|x_i > 0)\right)$$

$$= E(s_i|x_i < 0)E(s_i z_i|x_i \geq 0) - E(s_i|x_i \geq 0)E(s_i z_i|x_i < 0)$$
$$= \left[\int_{-\infty}^{\infty} r(z_i) \int_{-\infty}^{-z_i} f_v(v)dv f_z(z)dz\right] \cdot \left[\int_{-\infty}^{\infty} z_i h(z_i) \int_{-z_i}^{\infty} f_v(v)dv f_z(z)dz\right]$$
$$- \left[\int_{-\infty}^{\infty} h(z_i) \int_{-z_i}^{\infty} f_v(v)dv f_z(z)dz\right] \cdot \left[\int_{-\infty}^{\infty} z_i r(z_i) \int_{-\infty}^{-z_i} f_v(v)dv f_z(z)dz\right]$$

$$= \left[\int_{-\infty}^{\infty} \mathbf{r(z_i)F_v}(-\mathbf{z_i}) f_z(z)dz\right] \cdot \left[\int_{-\infty}^{\infty} \mathbf{h(z_i)(1 - F_v}(-\mathbf{z_i}))z_i f_z(z)dz\right]$$
$$- \left[\int_{-\infty}^{\infty} \mathbf{h(z_i)(1 - F_v}(-\mathbf{z_i}))f_z(z)dz\right] \cdot \left[\int_{-\infty}^{\infty} \mathbf{r(z_i)F_v}(-\mathbf{z_i})z_i f_z(z)dz\right]$$

The last term suggests the weighting scheme:

$$h(z_i) = \frac{1}{1 - F_v(-z_i)} = \frac{1}{Pr(x_i \geq 0|z_i)}, \qquad r(z_i) = \frac{1}{F_v(-z_i)} = \frac{1}{Pr(x_i < 0|z_i)}$$

Using these the expression in (2.13) to evaluate $E(s_i)E(s_i z_i|x_i \geq 0) - E(s_i|x_i \geq 0)E(s_i z_i)$

$$= \left[\int_{-\infty}^{\infty} \mathbf{r(z_i)F_v}(-\mathbf{z_i}) f_z(z)dz\right] \cdot \left[\int_{-\infty}^{\infty} \mathbf{h(z_i)(1 - F_v}(-\mathbf{z_i}))z_i f_z(z)dz\right]$$
$$- \left[\int_{-\infty}^{\infty} \mathbf{h(z_i)(1 - F_v}(-\mathbf{z_i}))f_z(z)dz\right] \cdot \left[\int_{-\infty}^{\infty} \mathbf{r(z_i)F_v}(-\mathbf{z_i})z_i f_z(z)dz\right]$$

$$= \left[\int_{-\infty}^{\infty} \frac{\mathbf{F_v}(-\mathbf{z_i})}{\mathbf{F_v}(-\mathbf{z_i})} f_z(z)dz\right] \cdot \left[\int_{-\infty}^{\infty} \frac{\mathbf{1 - F_v}(-\mathbf{z_i})}{\mathbf{1 - F_v}(-\mathbf{z_i})} z_i f_z(z)dz\right]$$
$$- \left[\int_{-\infty}^{\infty} \frac{\mathbf{1 - F_v}(-\mathbf{z_i})}{\mathbf{1 - F_v}(-\mathbf{z_i})} f_z(z)dz\right] \cdot \left[\int_{-\infty}^{\infty} \frac{\mathbf{F_v}(-\mathbf{z_i})}{\mathbf{F_v}(-\mathbf{z_i})} z_i f_z(z)dz\right]$$

$$= 1 \cdot \mu - 1 \cdot \mu = 0$$

Thus the weighting scheme,

$$s_i = \begin{cases} (Pr(x_i \geq 0|z_i))^{-1} & \text{if } w_i = 1 \qquad i \in (i, \cdots, N) \\ \\ (1 - Pr(x_i \geq 0|z_i))^{-1} & \text{if } w_i = 0 \end{cases}$$

leads to an unbiased estimate of the average treatment effect and this holds true generally in a linear setup with no distributional assumptions on v and z.

Chapter 3

Evaluating India's Safe Motherhood Scheme using Inverse Propensity Score Weighting

3.1 Introduction

In this chapter we apply Inverse Propensity Weighting using the Minimum Covariate Imbalance criteria to evaluate a large public health dataset from India. The aim of the study is to analyze the effectiveness of India's Safe Motherhood Scheme or Janani Suraksha Yojna (JSY) wherein eligible women receive monetary compensation from the government when they give birth in a health care facility. As per 2010 statistics, the perinatal mortality rate (number of stillbirths and deaths within first week of birth) and infant mortality rate (number of infants dying before reaching one year of age) in India were 49[1] and 46[2] per 1000 livebirths respectively. Maternal mortality rates were also very high at 220 [3] per $100,000$ livebirths. The longterm aim of JSY is to bring about a reduction in the incidence of neonatal deaths as well as maternal mortality rate in the country.

Previous analysis of the scheme includes mainly qualitative studies or quantitative evaluation on a geographical subset of the data. Notable exceptions include Lim et al (2010),

[1] National Health Survey, India.
[2] World Development Indicators, World Bank.
[3] World Development Indicators, World Bank.

Dongre and Kapur (2012) and Joshi and Sivaram (2014). However to our knowledge there has been no application of propensity score estimation techniques at the national level, a gap which this paper aims to fill.

The Safe Motherhood Scheme is not randomized – individuals self select into the program. This implies that there maybe selection bias in the data, making it suitable as an application of Inverse Propensity Score Weighting. Further the number of relevant observations exceeds $200,000$ – the large dataset is ideal for applying the Machine Learning techniques described in this paper. Previous evaluations of JSY have involved methods like Exact Matching, Difference in Difference and Regression Analysis. As per our knowledge, there has not been any work using Propensity Scores to evaluate JSY at the national level. In our empirical investigation we find that the Safe Motherhood Scheme has had a largely positive impact on health indicators like infant mortality, number of still births as well as behavioral indicators like the frequency of check ups for both mother and child. We also notice regional variations in the effects. In the datasets corresponding to the northeastern states our criteria picks propensity score estimates from the SVM model (whose results are substantially different from the linear probit).

The chapter is organized as follows. Section 2 describes the scheme and the data in greater detail. This is followed by a brief literature review in Section 3, the Empirical Design in Section 4 and a discussion of the results in Section 5. Section 6 concludes.

3.2 About the Scheme and the data

JSY which was launched by the Central Goverment of India in 2005 is the world's largest conditional cash transfer scheme in terms of the number of beneficiaries (over 9 million). As per the scheme, eligible women receive financial assistance from the government when they give birth in a health care facility . Eligibility criteria involves posession of Below-the-Poverty-Line (BPL) cards, belonging to a scheduled caste or tribe and order of birth.

The financial assistance amounts are Rs. 600 ($9.76) in urban areas and Rs.700 ($11.39) in rural areas. Further in 10 government identified high focus states where in-facility birth coverage is low – in these states all women are eligible for the cash benefit. The financial assistance amounts in these states are Rs. 1000 ($16.27) in urban areas and Rs. 1400 ($22.78) in rural areas. Also in these states, women aged 19 years and above with BPL cards also receive Rs 500 ($8.14) for births at home for the first two births.

These incentives are communicated to the women locally by female health volunteers or Accredited Social Health Activists (ASHA – which literally translates to HOPE in Hindi). ASHA volunteers receive performance based compensation for the promotion of universal immunization and raising awareness of health care delivery programmes and family planning initiatives. They earn monetary compensation for every beneficiary they escort to and accompany until discharge at a pre-determined healthcare facility according to a two-part disbursement scheme – first after the delivery, second after a post-natal check up and child vaccination visits. The following table summarizes the broad financial incentives of the scheme.

Table 3.1: Details of financial incentives (Rs.)

Rural Areas

	Mother	ASHA
High Focus State	1400	600
Non High Focus State	700	–

Urban Areas

	Mother	ASHA
High Focus State	1000	200
Non High Focus State	600	–

In this paper we estimate the Average Treatment Effect (ATE) of receiving financial assistance via JSY on two health outcomes – number of stillbirths and infant mortality. We also estimate ATE on three behavioural outcomes – whether the mother had 3 or more ante-natal checkups, whether any post-natal check up was conducted within 2 weeks of delivery and the frequency of child check-ups within 10 days of delivery. We are not aware of any other paper that uses Propensity Score methods to evaluate JSY and therefore we consider this attempt as a substantial contribution towards the literature on the assessment of JSY. Further, we are not aware of any other authors who have studied

the effect of JSY on infant mortality and frequency of child check-ups within 10 days of delivery.

Monitoring of the ongoing health and family welfare programs is carried out via the District Level Household and Facility Survey (DLHS). The International Institute for Population Sciences (IIPS) is the nodal agency for carrying out the survey. Three rounds of the survey data have been collected. The first round was collected between 1998 and 1999, the second between 2002 and 2004 and the third round between 2007 and 2008. We use the third round of the survey, DLHS-3 which was conducted between 2008 and 2009. The survey is one of the largest ever demographic and health surveys carried out in India. All districts of the country have been covered in the survey. The number of households covered by the survey is $720,320$. There is also individual level data on ever-married women aged 15-49 years – the sample size of the individual level data is $643,944$, The survey contains information on household socio-economic characteristics as well as individual level characteristics, fertility data and maternal health indicators.

Besides this, DLHS-3 also contains data on awareness and utilization of immunization services as well as family planning initiatives. An individual level dataset of unmarried women aged 15-24 years is also included. Finally we note that receipt of assistance via JSY was not implemented as a randomized experiment. This implies that any evaluation of the impact of JSY has to take into account selection issues.

3.3 Existing Literature

The first national-level analysis of the impact of JSY was carried out by Lim et al. (2010) where the authors used three strategies: Exact Matching, With Versus Without analysis and district-level Difference in Differences estimation. They found a largely positive impact of the scheme, particularly in the reduction of neonatal deaths and increase in in-facility births. Dongre and Kapur (2012) use a Difference in Differences specification

to study the impact of JSY on in-facility births with special focus on the gap between the high focus states with the rest of the states. Finally, Joshi and Sivaram (2014) also use a Difference in Differences specification to study the impact of JSY on ante-natal care, post-natal care and in-facility births. They control for Eligibility Criteria in their evaluation design. They find overall significant and positive impact of the scheme on all three outcomes but modest positive effects on those eligible to receive compensation under its guidelines.

Since the DLHS-3 differed in many significant ways with DLHS-2 (for example different sets of questions were asked in round 2 and round 3 to assess utilization of post natal care; the second round only includes married women uptil the age of 45 whereas the second includes both married and unmarried women up till age 50), we treat the pooling of the datasets required for Difference in Differences estimation with caution. In this paper we utilize only the most recent round of the longitudnal data, namely the DLHS-3 dataset. The only other work on estimating treatment effects using the single dataset was carried out by Lim et al (2010). In order to control for selection issues the authors implemented Exact Matching. The authors worked with a small subset of characteristics for matching, due to difficulty in finding exact matches with a large number of covariates.

By using IPW, we overcome this problem by working with a larger number of covariates which may contain important information about the selection process. We are not aware of any previous attempts to use Propensity Score related methods in the evaluation of JSY at the national level. Further, we use the insights from Chapter 2 to control for selection bias by choosing from a set of different propensity score estimation techniques. The empirical design and estimation strategy employed for the analysis are described next.

3.4 Empirical Design

3.4.1 Variables of Interest

We estimate the Average Treatment Effect (ATE) of receiving financial assistance via JSY on number of stillbirths, infant mortality, frequency of ante-natal care (ANC) visits, whether any post-natal check (PNC) was conducted and frequency of child check ups after delivery. Since the JSY scheme was introduced in 2005, we focus on the outcome of the last pregnancy which occured after 2004. This reduces the sample size to $227,039$ observations.

The treatment variable is

$$W_i = \begin{cases} 1 & \text{individual } i \text{ reported receipt of JSY assistance} \\ 0 & \text{otherwise.} \end{cases}$$

The health outcomes of interest are

$$Y_{i1} = \begin{cases} 1 & \text{individual } i\text{'s last pregnancy resulted in stillbirth} \\ 0 & \text{otherwise,} \end{cases}$$

$$Y_{i2} = \begin{cases} 1 & \text{individual } i\text{'s child died within 1 year} \\ 0 & \text{otherwise.} \end{cases}$$

The behavioral outcomes of interest are

$$Y_{i3} = \begin{cases} 1 & \text{individual } i \text{ reported at least 3 ANC visits} \\ 0 & \text{otherwise,} \end{cases}$$

$$Y_{i4} = \begin{cases} 1 & \text{individual } i \text{ reported receipt of PNC within 2 weeks} \\ 0 & \text{otherwise,} \end{cases}$$

$$Y_{i5} = \{\text{number of child checkups within 10 days: individual } i.\}$$

3.4.2 Estimation Techniques

In the first stage of the analysis we estimate propensity scores using Linear Probit, Random Forest and Support Vector Machines. The Covariate Imbalance scores from these techniques are then compared. If the unweighted estimator is picked then the the ATE is given by the naive difference of mean outcomes for the two groups. If one of the propensity score techniques is picked, then the Minimum Covariate Imbalance propensity score estimates $\widehat{e}_{h^*}$, corresponding to classifier h^* imply the ATE estimate is given by $\widehat{\tau}_{imb}$ from the following weighting least squares regression with weights λ_i:

$$Y_{ij} = \alpha + \tau_{imb} W_i + \varepsilon_i, \qquad \lambda_i = \sqrt{\frac{W_i}{\widehat{e}_{h^*}(X_i)} + \frac{(1 - W_i)}{1 - \widehat{e}_{h^*}(X_i)}}, \qquad j \in \{1, 2, 3, 4, 5\}$$

where X_i is the vector of covariates corresponding to individual i.

We also extend our method to the Doubly Robust framework by using IPW weights from the model picked by Minimum Covariate Imbalance criteria. In the Doubly Robust framework, weighted least squares (using IPW weights) is run on the outcome equation,

which includes control covariates along with the treatment variable.

$$Y_{ij} = \alpha + \tau_{dr} W_i + \beta_z Z + \varepsilon_i, \qquad \lambda_i = \sqrt{\frac{W_i}{\widehat{e}(X_i)} + \frac{(1 - W_i)}{1 - \widehat{e}(X_i)}}, \qquad j \in \{1, 2, 3, 4, 5\}.$$

where λ_i is the wieghting scheme, X_i and Z_i are the covariates in the selection and outcome equations respectively and $\widehat{e}(X_i)$ corresponds to the estimated propensity scores. The Doubly Robust ATE estimate is given by $\widehat{\tau}_{dr}$. The covariates included in X and Z are described in the following table.

When presenting our results, we report the change in numbers per 1000 births for the first two outcomes. For the last three we report the percentage change from the baseline $\left(\frac{\widehat{\tau}}{\widehat{\alpha}} \times 100\%\right)$.

Table 3.2: States listed by subgroups

High Focus States	Northeast States	Other States
Assam	Arunachal Pradesh	Andhra Pradesh
Bihar	Manipur	Goa
Chattisgarh	Meghalaya	Gujarat
Jammu and Kashmir	Mizoram	Haryana
Jharkhand	Sikkim	Himachal Pradesh
Madhya Pradesh	Tripura	Karnataka
Orissa		Kerala
Rajasthan		Maharashtra
Uttar Pradesh		Punjab
Uttarakhand		Tamil Nadu
		West Bengal

The 'Other States ' list does not list Union Territories which were included in the sample.

We create five separate datasets where missing/faulty observations corresponding to different covariates and outcome variables are dropped. We also slice the dataset into 12 smaller groups based on focus levels, geographical remoteness and type of locality. We conduct separate estimations at the National level, for High Focus States, Non-High Focus States and Northeast States. Data for each of these groups is further divided into Rural and Urban localities.

Table 3.3: List of Covariates Used in Empirical Analysis

Propensity Score Model (X: Selection Equation)	**Doubly Robust Estimation** (Z: Outcome Equation)
High Focus State, Northeast State, Urban, Rural: with healthcare facility, Whether Hindu, Whether Muslim, Whether Christian, Whether Sikh Whether Schedule Caste, Whether Schedule Tribe, Whether Other Backward Caste Ratio of Men to Women in Household, Wealth Quintie Possession of Below Poverty Line Card Whether any health insurance Type of House, Lack of toilet, Water treatment Years of Education: Respondent (0), Years of Education: Respondent $(1-5)$, Years of Education: Respondent $(6-12)$, Years of Education: Spouse (0), Years of Education: Spouse $(1-5)$, Years of Education: Spouse $(6-12)$, Number of livebirths, Age of living with husband, Age at delivery, Whether Pregnancy was known within 3 months.	Whether child is a girl, Number of sons at home, Number of children at home, Number of previous still-births, Number of previous spontaneous abortions, High Focus State, Northeast State, Urban, Rural: with healthcare facility, Whether Hindu, Whether Muslim, Whether Christian, Whether Schedule Caste, Whether Schedule Tribe, Whether Other Backward Caste Ratio of Men to Women in Household, Wealth Quintie Possession of Below Poverty Line Card Years of Education: Respondent (0), Years of Education: Respondent $(1-5)$, Years of Education: Respondent $(6-12)$, Years of Education: Spouse (0), Years of Education: Spouse $(1-5)$, Years of Education: Spouse $(6-12)$, Age at delivery,

Table 3.4: Number of observations in datasets corresponding to different outcomes and regions

	Y_1	Y_2	Y_3	Y_4	Y_5
National	205,390	205,394	202,845	195,399	198,205
High Focus	128,051	126,681	127,074	128,034	123,786
Northeast	14,667	14,581	13,962	14,662	13,738
Non-High Focus	77,339	76,713	75,771	67,365	74,419

3.5 Results

3.5.1 Sample Characteristics

A perusal of the household characteristics of the JSY beneficiaries in our sample indicates that a greater proportion of the sample comes from high focus states and rural areas. We also note that women in rural areas with no healthcare facility are under-represented. Similarly women from minority religions are under-represented. All wealth quintiles are well-represented. The individual characteristics of the beneficiaries indicates that in a large proportion of the sample both the woman and her spouse had less than 12 years of schooling (and over 30% women had no schooling). Most women in the sample are between the ages of 20 and 30 years and most reported 2 or fewer number of livebirths.

Figure 3.1: JSY Beneficieries grouped by Household Characteristics

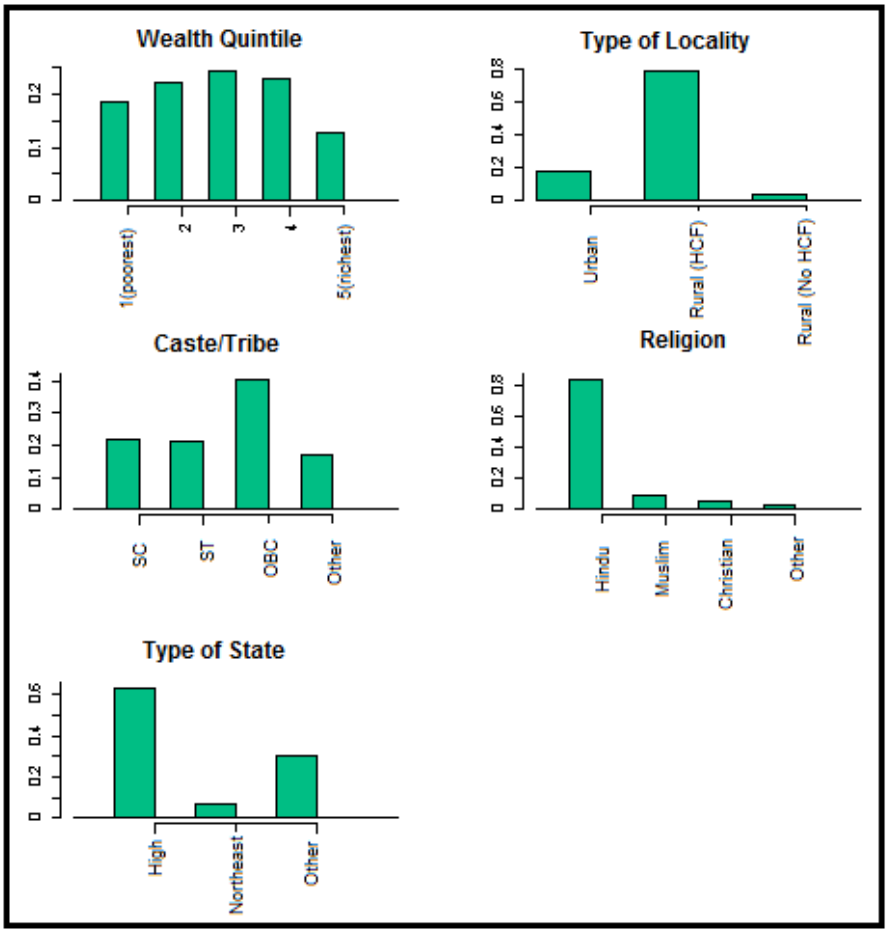
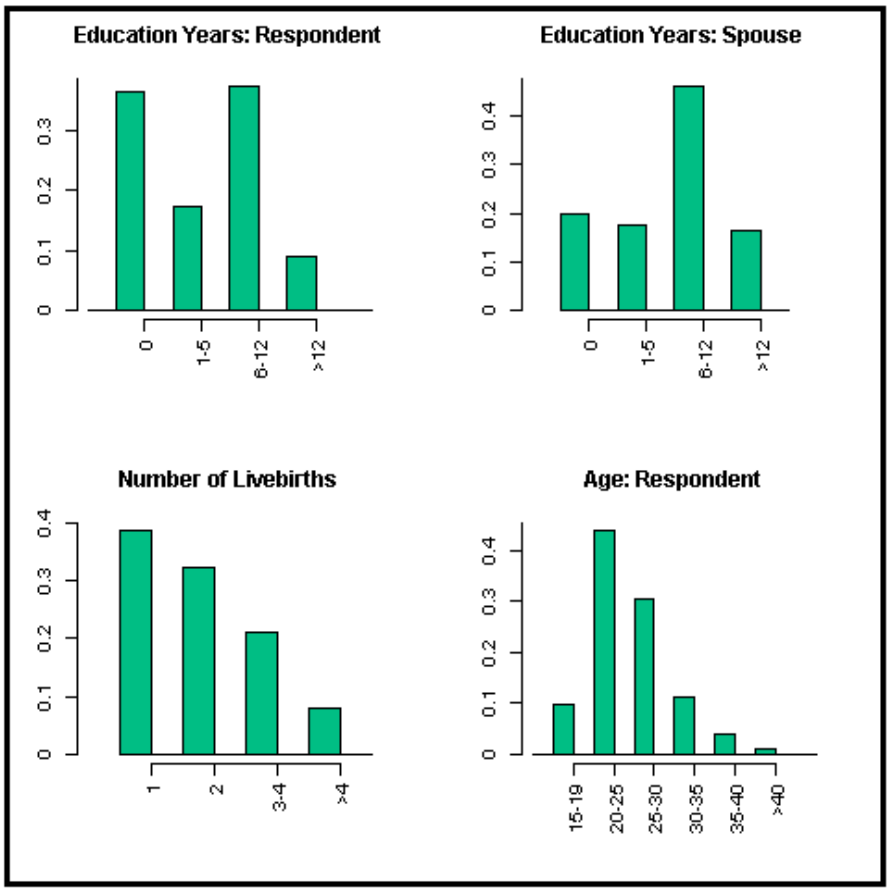


Figure 3.2: JSY Beneficieries grouped by Individual Characteristics



Household Characteristics used for estimation of Propensity Scores – High Focus State, Northeast State, Urban, Rural: with healthcare facility, Whether Hindu, Whether Muslim, Whether Christian, Whether Schedule Caste, Whether Schedule Tribe, ratio of Men to Women in Household, Type of House, Lack of toilet, Water treatment.

Individual Characteristics used for estimation of Propensity Scores– Years of Education: Respondent $(0, 1-5, 6-12, \geq 12)$ Years of Education: Spouse $(0, 1-5, 6-12, \geq 12)$, Number of livebirths, Age of living with husband, Age at delivery, Whether Pregnancy was known within 3 months.

Overall, we find encouraging results on the effect of JSY on both health and behavioural outcomes. However the effects are not uniform across different states, we note a large amount of

regional disparities.

The Linear Probit Model is picked by the Minimum Covariate Imbalance criteria in most of the regionally sliced datasets (49 out of 60). SVM is picked in the remaining 11 datasets. In the cases where SVM is picked, the corresponding results with Linear Probit are substantially different. For instance, weighting by Linear Probit imply larger health outcome effects and smaller behavioral outcome effects. In fact IPW using SVM implies no significant ATE of the scheme on infant mortality in the Northeast states, whereas IPW using Linear Probit implies large and significant reductions in infant mortality caused by the program in the same states. Therefore in this situation, relying on Linear Probit would lead to very different policy prescription inputs.

3.5.2 ATE estimates: Health Outcomes

The ATE estimate of JSY on the number of stillbirths per 1000 deliveries is -4.46 at the national level using IPW. Northeast States follow the national level effect. However we find that High focus states lag behind the Non-High Focus states. We also note that ATE estimates in urban areas is generally much higher than in the rural areas. The ATE estimate of JSY on infant mortalilty is -3.25 at the national level. Northeast States do not report any significant effect of JSY on infant mortality. Rural areas in High focus states lag behind rural areas in Non-High Focus states. ATE of JSY on infant mortality in urban areas is not significant for any of the study regions. The results from the Doubly Robust framework show the same general trends, albeit with slightly smaller effects.

There can be a number of explanations for the regional variations in the ATE estimates in terma of these health outcomes. In the High Focus states there were no eligibility requirements for being a beneficiary of JSY. On the other hand in the Non-High Focus states only those women who were at higher risk were targetted. Thus health outcome effects in High focus states is expected to be more modest in comparison to the Non-High focus states, since women who are less vulnerable are also present in the High focus states sample. Further the lack of eligibilty requirements in High focus states may have led to much greater participation in these regions leading to greater pressure on healthcare services and poorer quality of service. Finally better healthcare facilities in urban areas may help explain the larger negative effect of JSY on

number of stillbirths and better standards of living may help explain the smaller negative effect of JSY on infant mortality.

3.5.3 ATE Estimates: Behavioral Outcomes

In terms of behavioral outcomes we find increases in whether the mother had at least 3 ante-natal check ups across the datasets. Much larger increases are estimated for whether the woman had any post-natal check up (60.9% at the national level) and number of child check-ups within 10 days of delivery (41.8% at the national level). The percentage increase is larger in High Focus states and Northeast states than in the Non-High Focus States. The increase is also generally found to be higher in rural areas (more than 100%) than urban areas. The results from the Doubly Robust framework show similar trends with one clear exception. The Doubly Robust estimates show smaller positive ATE of the scheme on whether the mother went for at least 3 ante-natal checks. However, unlike the IPW estimates there is very little regional variation between the High Focus and Non-High Focus States.

The regional differences in financial incentives of the ASHA volunteers may help explain differences in estimated ATE of the scheme on behavioural outcomes. ASHA volunteers receive payment in two stages, the first is escorting women to healthcare facilities for childbirth and the second is escorting them for post-natal check up and child vaccination visits. This factor coupled with low bases corresponding to PNC and child check-ups can explain why rural areas, particularly in High focus states experienced large positive effects of JSY on behavioural outcomes.

Table 3.5: Estimates of ATE of JSY on Number of Stillbirths: IPW [1]

		All	Rural	Urban
National	ATE 0.95 C.I. Imbalance IPW Model per 1000	-0.00456 (-0.005 -0.004) 0.117 Probit -4.56	-0.00388 (-0.005 -0.003) 0.179 Probit -3.88	-0.00748 (-0.009 -0.006) 0.006 Probit -7.48
High Focus	ATE 0.95 C.I. Imbalance IPW Model per 1000	-0.00391 (-0.005 -0.003) 0.075 Probit -3.91	-0.00313 (-0.004 -0.002) 0.156 Probit -3.13	-0.00787 (-0.01 -0.006) 0.006 Probit -7.87
Northeast	ATE 0.95 C.I. Imbalance IPW Model per 1000	-0.00527 (-0.007 -0.003) 1.169 SVM -5.27 (-5.9)†	-0.00517 (-0.008 -0.003) 1.318 SVM -5.17 (-6.1)†	-0.00481 (-0.009 -0.001) 0.213 Probit -4.81
Non-High Focus	ATE 0.95 C.I. Imbalance IPW Model per 1000	-0.00585 (-0.007 -0.005) 0.267 Probit -5.85	-0.00548 (-0.007 -0.004) 0.662 Probit -5.48	-0.00761 (-0.009 -0.006) 0.131 Probit -7.61

Table 3.6: Estimates of ATE of JSY on Number of Stillbirths: Doubly Robust [2]

		All	Rural	Urban
National	ATE 0.95 C.I. IPW Model per 1000	-0.00425 (-0.005 -0.004) Probit -4.25	-0.00363 (-0.005 -0.003) Probit - 3.63	-0.00692 (-0.008 -0.006) Probit -6.92
High Focus	ATE 0.95 C.I. IPW Model per 1000	-0.00352 (-0.005 -0.003) Probit -3.52	-0.00277 (-0.004 -0.002) Probit -2.77	-0.00718 (-0.009 -0.005) Probit -7.12
Northeast	ATE 0.95 C.I. IPW Model per 1000	-0.00495 (-0.007 -0.003) SVM -4.95 (-5.5)†	-0.00536 (-0.008 -0.003) SVM -5.36 (-5.9)†	-0.00355 (-0.007 -3e^{-5}) Probit -3.55
Non-High Focus	ATE 0.95 C.I. IPW Model per 1000	-0.00511 (-0.006 -0.004) Probit -5.11	-0.00482 (-0.006 -0.004) Probit -4.82	-0.00698 (-0.009 -0.005) Probit -6.98

[1] Confidence Intervals reported are from Weighted Least Squares (WLS) Regression of outcome of interest Y on treatment W([2] and control variables Z). † IPW with probit results in parantheses.

Table 3.7: Estimates of ATE of JSY on Survival of Child for a Year: IPW [1]

		All	Rural	Urban
National	ATE 0.95 C.I. Imbalance IPW Model per 1000	-0.00288 (-0.004 -0.002) 0.135 Probit -2.88	-0.00319 (-0.005 -0.002) 0.204 Probit -3.19	0.00135 (-0.002 0.004) 0.007 Probit –
High Focus	ATE 0.95 C.I. Imbalance IPW Model per 1000	-0.00258 (-0.005 -0.001) 0.084 Probit -2.58	-0.00219 (-0.004 -0.0001) 0.17 Probit -2.19	-0.00304 (-0.008 0.002) 0.007 Probit –
Northeast	ATE 0.95 C.I. Imbalance IPW Model per 1000	-0.00276 (-0.007 0.001) 1.145 SVM – (-5.29)†	-0.00256 (-0.007 0.002) 1.271 SVM – (-5.2)†	-0.00699 (-0.015 0.001) 0.236 Probit –
Non-High Focus	ATE 0.95 C.I. Imbalance IPW Model per 1000	-0.00434 (-0.006 -0.002) 0.351 Probit -4.34	-0.00709 (-0.009 -0.005) 0.902 Probit -7.09	0.00253 (-0.001 0.006) 0.148 Probit –

Table 3.8: Estimates of ATE of JSY on Survival of Child for a Year: Doubly Robust[2]

		All	Rural	Urban
National	ATE 0.95 C.I. IPW Model per 1000	-0.00261 (-0.004 -0.001) Probit -2.61	-0.00287 (-0.005 -0.001) Probit -2.87	0.00152 (-0.001 0.004) Probit –
High Focus	ATE 0.95 C.I. IPW Model per 1000	-0.0016 (-0.004 0.000) Probit –	-0.0013 (-0.003 0.001) Probit –	-0.00155 (-0.006 0.003) Probit –
Northeast	ATE 0.95 C.I. IPW Model per 1000	-0.00648 (-0.011 -0.002) SVM -6.48 (-3.0)†	-0.00623 (-0.011 -0.001) SVM -6.23 (-2.8)†	-0.00517 (-0.013 0.003) Probit –
Non-High Focus	ATE 0.95 C.I. IPW Model per 1000	-0.00359 (-0.006 -0.002) Probit -3.59	-0.00571 (-0.008 -0.003) Probit -5.71	0.003 (-0.001 0.007) Probit –

[1] Confidence Intervals reported are from Weighted Least Squares (WLS) Regression of outcome of interest Y on treatment W([2] and control variables Z). † IPW with probit results in parantheses.

Table 3.9: Estimates of ATE of JSY on Whether at least 3 Ante-Natal Checks: IPW[1]

		All	Rural	Urban
National	ATE	0.113	0.096	0.067
	0.95 C.I.	(0.109 0.117)	(0.0911 0.101)	(0.058 0.077)
	Imbalance	0.047	0.183	0.006
	IPW Model	SVM	Probit	Probit
	% change	24.9 % (20 %)†	23.7 %	9.9 %
High Focus	ATE	0.081	0.080	0.065
	0.95 C.I.	(0.076 0.086)	(0.074 0.0853)	(0.050 0.080)
	Imbalance	0.075	0.155	0.006
	IPW Model	Probit	Probit	Probit
	% change	25.6 %	27.6 %	13 %
Northeast	ATE	0.275	0.293	0.091
	0.95 C.I.	(0.260 0.290)	(0.276 0.310)	(0.062 0.121)
	Imbalance	1.33	0.815	0.374
	IPW Model	SVM	SVM	Probit
	% change	52.7 % (34.9 %)†	62.4 % (45.6 %)†	11.8 %
Non-High Focus	ATE	0.097	0.111	0.074
	0.95 C.I.	(0.09120.104)	(0.103 0.118)	(0.065 0.083)
	Imbalance	0.331	0.791	0.146
	IPW Model	Probit	Probit	Probit
	% change	14.1%	17.4%	9%

Table 3.10: Estimates of ATE of JSY on Whether at least 3 Ante-Natal Checks: Doubly Robust [2]

		All	Rural	Urban
National	ATE	0.109	0.105	0.076
	0.95 C.I.	(0.105 0.113)	(0.101 0.109)	(0.068 0.084)
	IPW Model	SVM	Probit	Probit
	% change	18.4 % (17.7 %)†	17.7 %	12 %
High Focus	ATE	0.083	0.084	0.065
	0.95 C.I.	(0.078 0.088)	(0.079 0.09)	(0.051 0.079)
	IPW Model	Probit	Probit	Probit
	% change	20.9 %	20.5 %	18.8 %
Northeast	ATE	0.175	0.201	0.06
	0.95 C.I.	(0.16 0.19)	(0.184 0.219)	(0.031 0.09)
	IPW Model	SVM	SVM	Probit
	% change	34.1 % (60.2 %)†	35.5 % (60.5 %)†	11 %
Non-High Focus	ATE	0.112	0.127	0.079
	0.95 C.I.	(0.106 0.118)	(0.12 0.134)	(0.07 0.088)
	IPW Model	Probit	Probit	Probit
	% change	21.3 %	25.6 %	11.6 %

[1] Confidence Intervals reported are from Weighted Least Squares (WLS) Regression of outcome of interest Y on treatment W([2] and control variables Z). † IPW with probit results in parantheses.

Table 3.11: IPW: ATE of JSY on Any Post-Natal Check within 2 weeks of delivery [1]

		All	Rural	Urban
National	ATE	0.251	0.281	0.121
	0.95 C.I.	(0.247 0.256)	(0.277 0.286)	(0.112 0.130)
	Imbalance	0.123	0.183	0.006
	IPW Model	Probit	Probit	Probit
	% change	60.6%	78.3%	18.3%
High Focus	ATE	0.302	0.320	0.168
	0.95 C.I.	(0.297 0.307)	(0.315 0.326)	(0.154 0.182)
	Imbalance	0.075	0.155	0.006
	IPW Model	Probit	Probit	Probit
	% change	98.2%	117.4 %	31.9%
Northeast	ATE	0.348	0.373	0.212
	0.95 C.I.	(0.333 0.363)	(0.357 0.390)	(0.179 0.244)
	Imbalance	1.101	1.003	0.213
	IPW Model	SVM	SVM	Probit
	% change	95.9 % (56.8 %)†	120 % (75.1 %)†	34.3 %
Non-High Focus	ATE	0.142	0.154	0.080
	0.95 C.I.	(0.135 0.149)	(0.145 0.162)	(0.069 0.091)
	Imbalance	0.424	1.226	0.116
	IPW Model	Probit	Probit	Probit
	% change	22.9 %	27.6 %	10.1 %

Table 3.12: Doubly Robust: ATE of JSY on Any Post-Natal Check within 2 weeks of delivery [2]

		All	Rural	Urban
National	ATE	0.256	0.285	0.127
	0.95 C.I.	(0.252 0.26)	(0.28 0.289)	(0.119 0.136)
	IPW Model	Probit	Probit	Probit
	% change	54.3 %	62.1%	22.2%
High Focus	ATE	0.301	0.321	0.168
	0.95 C.I.	(0.296 0.306)	(0.316 0.327)	(0.154 0.182)
	IPW Model	Probit	Probit	Probit
	% change	92.2%	114.8%	29.3%
Northeast	ATE	0.272	0.306	0.178
	0.95 C.I.	(0.257 0.288)	(0.289 0.324)	(0.146 0.211)
	IPW Model	SVM	SVM	Probit
	% change	74.3% (65.6 %)†	69.6% (61.9 %)†	82.5%
Non-High Focus	ATE	0.156	0.18	0.083
	0.95 C.I.	(0.15 0.163)	(0.172 0.188)	(0.073 0.094)
	IPW Model	Probit	Probit	Probit
	% change	36.5%	38.1%	17.2%

[1] Confidence Intervals reported are from Weighted Least Squares (WLS) Regression of outcome of interest Y on treatment W([2] and control variables Z). † IPW with probit results in parantheses.

Table 3.13: IPW: ATE of JSY on Number of Child Checkups within 10 days of delivery[1]

		All	Rural	Urban
National	ATE	0.429	0.493	0.224
	0.95 C.I.	(0.416 0.443)	(0.479 0.507)	(0.188 0.259)
	Imbalance	0.139	0.209	0.007
	IPW Model	Probit	Probit	Probit
	% change	41.08 %	57.7 %	11.97 %
High Focus	ATE	0.478	0.514	0.216
	0.95 C.I.	(0.465 0.491)	(0.501 0.527)	(0.174 0.258)
	Imbalance	0.081	0.166	0.007
	IPW Model	Probit	Probit	Probit
	% change	82.73%	104.68 %	19.13 %
Northeast	ATE	0.641	0.825	0.390
	0.95 C,I,	(0.597 0.686)	(0.778 0.872)	(0.268 0.513)
	Imbalance	0.241	0.903	0.162
	IPW Model	SVM	SVM	Probit
	% change	88.1% (50.2 %)†	141 % (73.6 %)†	28.2 %
Non-High Focus	ATE	0.315	0.328	0.277
	0.95 C.I.	(0.289 0.341)	(0.298 0.357)	(0.228 0.327)
	Imbalance	0.438	1.181	0.21
	IPW Model	Probit	Probit	Probit
	% change	17.28%	20.72%	11.22 %

Table 3.14: Doubly Robust:ATE of JSY on Child Checkups within 10 days of delivery[2]

		All	Rural	Urban
National	ATE	0.458	0.514	0.261
	0.95 C.I.	(0.446 0.47)	(0.502 0.527)	(0.229 0.292)
	IPW Model	Probit	Probit	Probit
	% change	35.6%	40.7%	18.1%
High Focus	ATE	0.479	0.521	0.21
	0.95 C.I.	(0.467 0.491)	(0.508 0.534)	(0.170 0.250)
	IPW Model	Probit	Probit	Probit
	% change	81.9 %	98.2%	24.3%
Northeast	ATE	0.541	0.641	0.276
	0.95 C.I.	(0.498 0.583)	(0.592 0.691)	(0.158 0.393)
	IPW Model	SVM	SVM	Probit
	% change	275 % (209 %)†	134 % (154%)†	-49.5%
Non-High Focus	ATE	0.379	0.421	0.301
	0.95 C.I.	(0.356 0.402)	(0.394 0.448)	(0.255 0.347)
	IPW Model	Probit	Probit	Probit
	% change	63%	57.6 %	42.4%

[1] Confidence Intervals reported are from Weighted Least Squares (WLS) Regression of outcome of interest Y on treatment W([2] and control variables Z). † IPW with probit results in parantheses.

3.5.4 Discussion

At this point, we note that many of the covariates that we used in the outcome regression in the Doubly Robust framework had significant impacts on the outcomes of interest. Since the main concern of this paper is the estimation of ATE, the full regression results for the Doubly Robust estimates are not presented here.

Our results are on health outcomes are broadly consistent with the findings of Lim et al (2010)'s exact matching analysis. They found that JSY led to a reduction in perinatal and neonatal deaths and that the reduction was of a lower magnitude in High Focus states. We studied one related health outcome (number of still births) and found similar effects. We also studied a different health outcome – infant mortality and found that the general trends (encouraging results, lower effects in High Focus areas) are similar for this outcome too.

On the other hand our results differ with Lim et al's exact matching analysis on behavioral outcomes. They found an increase in the probability of at least 3 ANC visits caused by JSY and that the increase was of a greater magnitude in High focus states. While we found that while there is in fact a positive effect of the scheme on ANC visits, the effects are of similar magnitude across the High focus and the Non-High focus states[4]. We also studied two new behavioral outcomes – post-natal care and frequency of child check-ups.

Although the findings of our analysis are encouraging, this empirical investigation comes with a number of caveats. First, the validity of our results depends on the quality of the dataset. Since we did not collect the data firsthand, there may be data quality issues that we cannot account for. Second, the survey only reports information on the receipt of financial assistance via JSY. It is possible that due to corruption or inefficiencies, some women who took part in JSY did not receive compensation. Such a situation would lead us to underestimate the effect of the scheme. Third, despite the large list of covariates that we considered to control for selection effects, its possible that we missed some unobserved effects and that we couldn't control for selection all the way. There are several other possible Propensity Score related methods that we can use to evaluate the data such as Propensity Score Matching and Difference in Difference estimation using Propensity Scores that are out of the scope of this paper but warrant investigation. Finally, all the work on the effect of JSY has been based on the reduced form approach – a structural estimation model would be a very valuable contribution to the literature.

[4]Doubly Robust Estimates.

3.6 Summary and Directions for Future Research

In this chapter we evaluated the effect of one of the world's largest conditional cash transfer schemes, India's Safe Motherhood Scheme on certain health and behavioral outcomes using the Minimum Covariate Imbalance criteria in the IPW framework. As per our knowledge this is the first attempt to use Propensity Score techniques to evaluate the effectiveness of JSY using national level data. We found that in general JSY had positive effects on both health and behavioral outcomes. The effects varied a lot across different regions. Health outcome effects were typically smaller in High focus states while behavioral outcome effects were typically larger in High focus states. We attempted to explain these regional differences based on the incentive structure of the JSY.

We would like to extend the work in this paper to other techniques that rely on Propensity Scores for example Matching and Difference in Difference techniques. The work in this paper deals with a reduced form approach for the estimation of ATE. In the future, we would be very interested in introducing Machine Learning techniques in structural models of causal inference (for instance the Covariate Balancing propensity Score framework (Imai and Ratkovic (2013)). We would also like to build results on the consistency of the estimators introduced in this paper.

On the empirical side, we only covered a small subset of variables from the DLHS dataset. The survey also recorded a large amount of data on health policy, vaccinations, female reproductive health and awareness – all of which are critical from a Public Health point of view. We would like to continue working with the dataset to evaluate the status of health reforms in India. Finally, a structural model to evaluate the JSY can potentially be a very insightful exercise and we plan to work on developing such a model in the future.

Bibliography

[1] Hervé Abdi and Lynne J Williams, *Principal component analysis*, Wiley Interdisciplinary Reviews: Computational Statistics **2** (2010), no. 4, 433–459.

[2] Donald WK Andrews and Biao Lu, *Consistent model and moment selection procedures for gmm estimation with application to dynamic panel data models*, Journal of Econometrics **101** (2001), no. 1, 123–164.

[3] Ravi Bansal, Dana Kiku, and Amir Yaron, *An empirical evaluation of the long-run risks model for asset prices*, Tech. report, National Bureau of Economic Research, 2009.

[4] Mira Bernstein, Vin De Silva, John C Langford, and Joshua B Tenenbaum, *Graph approximations to geodesics on embedded manifolds*, Tech. report, Technical report, Department of Psychology, Stanford University, 2000.

[5] Mehmet Caner, *Lasso-type gmm estimator*, Econometric Theory **25** (2009), no. 01, 270–290.

[6] Mehmet Caner and Q Fan, *The adaptive lasso method for instrumental variable selection*, Tech. report, Working Paper, North Carolina State University, 2010.

[7] Marine Carrasco and Jean-Pierre Florens, *Generalization of gmm to a continuum of moment conditions*, Econometric Theory **16** (2000), no. 06, 797–834.

[8] Marine Carrasco, Jean-Pierre Florens, and Eric Renault, *Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization*, Handbook of econometrics **6** (2007), 5633–5751.

[9] Rich Caruana and Alexandru Niculescu-Mizil, *An empirical comparison of supervised learning algorithms*, Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 161–168.

[10] Nitesh V Chawla and David A Cieslak, *Evaluating probability estimates from decision trees*, American Association for Artificial Intelligence, 2006.

[11] Yizong Cheng, *Mean shift, mode seeking, and clustering*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **17** (1995), no. 8, 790–799.

[12] George M Constantinides and Anisha Ghosh, *Asset pricing tests with long-run risks in consumption growth*, Review of Asset Pricing Studies **1** (2011), no. 1, 96–136.

[13] Mitali Das, Whitney K Newey, and Francis Vella, *Nonparametric estimation of sample selection models*, The Review of Economic Studies **70** (2003), no. 1, 33–58.

[14] Morris H DeGroot and Stephen E Fienberg, *The comparison and evaluation of forecasters*, The statistician (1983), 12–22.

[15] Rajeev H Dehejia and Sadek Wahba, *Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs*, Journal of the American statistical Association **94** (1999), no. 448, 1053–1062.

[16] Alexis Diamond and Jasjeet S Sekhon, *Genetic matching for estimating causal effects*, (2012).

[17] Priyanka Dixit, Laxmi Kant Dwivedi, and Faujdar Ram, *Estimating the impact of antenatal care visits on institutional delivery in india: A propensity score matching analysis*, Health **5** (2013), 862.

[18] Dobrislav Dobreva and Ernst Schaumburgb, *Robust forecasting by regularization*, (2013).

[19] Ambrish Dongre, *Effect of monetary incentives on institutional deliveries: Evidence from india*, (2010).

[20] Ambrish A Dongre and Avani Kapur, *How is janani suraksha yojana performing in backward districts of india?*, Economic & Political Weekly **48** (2013), no. 42.

[21] Wei Fan, *On the optimality of probability estimation by random decision trees*, AAAI, vol. 2004, 2004, pp. 336–341.

[22] Frédérique Fève and Jean-Pierre Florens, *The practice of non-parametric estimation by solving inverse problems: the example of transformation models*, The Econometrics Journal **13** (2010), no. 3, S1–S27.

[23] David A Freedman and Richard A Berk, *Weighting regressions by propensity scores*, Evaluation Review **32** (2008), no. 4, 392–409.

[24] Isabelle Guyon, *A practical guide to model selection*, (2009).

[25] Isabelle Guyon, Amir Saffari, Gideon Dror, and Gavin Cawley, *Model selection: Beyond the bayesian/frequentist divide*, The Journal of Machine Learning Research **11** (2010), 61–87.

[26] Erinn M Hade and Bo Lu, *Bias associated with using the estimated propensity score as a regression covariate*, Statistics in medicine **33** (2014), no. 1, 74–87.

[27] Peter Hall, Joel L Horowitz, et al., *Nonparametric methods for inference in the presence of instrumental variables*, The Annals of Statistics **33** (2005), no. 6, 2904–2929.

[28] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani, *The elements of statistical learning*, vol. 2, Springer, 2009.

[29] James J Heckman and Richard Robb Jr, *Alternative methods for evaluating the impact of interventions: An overview*, journal of Econometrics **30** (1985), no. 1, 239–267.

[30] Keisuke Hirano and Guido W Imbens, *Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization*, Health Services and Outcomes research methodology **2** (2001), no. 3-4, 259–278.

[31] Keisuke Hirano, Guido W Imbens, and Geert Ridder, *Efficient estimation of average treatment effects using the estimated propensity score*, Econometrica **71** (2003), no. 4, 1161–1189.

[32] Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart, *Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference*, Political analysis **15** (2007), no. 3, 199–236.

[33] District Level Household, *Facility survey (dlhs-3), 2007–08*, IIPS/MoHFW **4** (2010).

[34] Hidehiko Ichimura and Oliver B Linton, *Asymptotic expansions for some semiparametric program evaluation estimators*, LSE STICERD Research Paper No. EM451 (2003).

[35] Kosuke Imai and Marc Ratkovic, *Covariate balancing propensity score*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **76** (2014), no. 1, 243–263.

[36] Guido W Imbens, *Nonparametric estimation of average treatment effects under exogeneity: A review*, Review of Economics and statistics **86** (2004), no. 1, 4–29.

[37] Atsushi Inoue and Barbara Rossi, *Testing for weak identification in possibly nonlinear models*, Journal of Econometrics **161** (2011), no. 2, 246–261.

[38] Serge Darolles Jean and Pierre Florens3 Eric Renault, *Nonparametric instrumental regression1*, (2007).

[39] Shareen Joshi and Anusuya Sivaram, *Does it pay to deliver? an evaluation of india's safe motherhood program*, World Development **64** (2014), 434–447.

[40] Joseph DY Kang and Joseph L Schafer, *Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data*, Statistical science (2007), 523–539.

[41] Narayana R Kocherlakota, *On tests of representative consumer asset pricing models*, Journal of Monetary Economics **26** (1990), no. 2, 285–304.

[42] John Langford and Bianca Zadrozny, *Estimating class membership probabilities using classifier learners*, Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, 2005, pp. 198–205.

[43] Brian K Lee, Justin Lessler, and Elizabeth A Stuart, *Improving propensity score weighting using machine learning*, Statistics in medicine **29** (2010), no. 3, 337–346.

[44] Zhipeng Liao, *Adaptive gmm shrinkage estimation with consistent moment selection*, Econometric Theory **29** (2013), no. 05, 857–904.

[45] Stephen S Lim, Lalit Dandona, Joseph A Hoisington, Spencer L James, Margaret C Hogan, and Emmanuela Gakidou, *India's janani suraksha yojana, a conditional cash transfer programme to increase births in health facilities: an impact evaluation*, The Lancet **375** (2010), no. 9730, 2009–2023.

[46] Sydney C Ludvigson, *Advances in consumption-based asset pricing: Empirical tests*, Tech. report, National Bureau of Economic Research, 2011.

[47] Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral, *Propensity score estimation with boosted regression for evaluating causal effects in observational studies.*, Psychological methods **9** (2004), no. 4, 403.

[48] David Mease and Abraham Wyner, *Evidence contrary to the statistical view of boosting*, The Journal of Machine Learning Research **9** (2008), 131–156.

[49] Alexandru Niculescu-Mizil and Rich Caruana, *Predicting good probabilities with supervised learning*, Proceedings of the 22nd international conference on Machine learning, ACM, 2005, pp. 625–632.

[50] John C Platt, *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods*, Advances in large margin classifiers, Citeseer, 1999.

[51] James Robins, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky, *Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable*, Statistical Science (2007), 544–559.

[52] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao, *Analysis of semiparametric regression models for repeated outcomes in the presence of missing data*, Journal of the American Statistical Association **90** (1995), no. 429, 106–121.

[53] Lorenzo Rosasco, Andrea Caponnetto, Ernesto D Vito, Francesca Odone, and Umberto D Giovannini, *Learning, regularization and ill-posed inverse problems*, Advances in Neural Information Processing Systems, 2004, pp. 1145–1152.

[54] Paul R Rosenbaum and Donald B Rubin, *The central role of the propensity score in observational studies for causal effects*, Biometrika **70** (1983), no. 1, 41–55.

[55] ———, *Constructing a control group using multivariate matched sampling methods that incorporate the propensity score*, The American Statistician **39** (1985), no. 1, 33–38.

[56] Gary Smith and Frank Campbell, *A critique of some ridge regression methods*, Journal of the American Statistical Association **75** (1980), no. 369, 74–81.

[57] Peter M Steiner, *S. guo & mw fraser (2010). propensity score analysis: Statistical methods and applications.*, Psychometrika **75** (2010), no. 4, 775–777.

[58] George Tauchen, *Statistical properties of generalized method-of-moments estimators of structural parameters obtained from financial market data*, Journal of Business & Economic Statistics **4** (1986), no. 4, 397–416.

[59] George Tauchen and Robert Hussey, *Quadrature-based methods for obtaining approximate solutions to nonlinear asset pricing models*, Econometrica: Journal of the Econometric Society (1991), 371–396.

[60] Joshua B Tenenbaum, Vin De Silva, and John C Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science **290** (2000), no. 5500, 2319–2323.

[61] Francis Vella, *Estimating models with sample selection bias: a survey*, Journal of Human Resources (1998), 127–169.

[62] Daniel Westreich, Justin Lessler, and Michele Jonsson Funk, *Propensity score estimation: machine learning and classification methods as alternatives to logistic regression*, Journal of clinical epidemiology **63** (2010), no. 8, 826.

[63] Jonathan H Wright, *Detecting lack of identification in gmm*, Econometric Theory **19** (2003), no. 02, 322–330.

[64] Mayuko Yatsu, *The impact of anganwadi centers' services on infant survival in india.*

[65] Janani Suraksha Yojana, *Effects of the janani suraksha yojana on maternal and newborn care practices: Women's experiences in rajasthan*, Population Council, 2011.

[66] Bianca Zadrozny and Charles Elkan, *Transforming classifier scores into accurate multiclass probability estimates*, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2002, pp. 694–699.