Title

Topics in Microeconometrics: Estimation of a Dynamic Model of Occupational Transitions, Wage and Non-Wage Benefits Cross Validation Bandwidth Selection for Derivatives of Various Dimensional Densities Testing the Additive Separability of the Teacher...

Author

Baird, Matthew David

Publication Date

2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Topics in Microeconometrics:
Estimation of a Dynamic Model of Occupational Transitions, Wage and Non-Wage Benefits
Cross Validation Bandwidth Selection for Derivatives of Various Dimensional Densities
Testing the Additive Separability of the Teacher Value Added Effect Semiparametrically

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Economics

by

Matthew David Baird

2012

ABSTRACT OF THE DISSERTATION

Topics in Microeconometrics:
Estimation of a Dynamic Model of Occupational Transitions, Wage and Non-Wage Benefits
Cross Validation Bandwidth Selection for Derivatives of Various Dimensional Densities
Testing the Additive Separability of the Teacher Value Added Effect Semiparametrically

by

Matthew David Baird
Doctor of Philosophy in Economics
University of California, Los Angeles, 2012
Professor Moshe Buchinsky, Chair

I study three separate questions in this dissertation. In Chapter 1, I develop and estimate a structural dynamic model of occupation and job choice to test hypotheses of the importance of wages and non-wages and learning in occupational transitions, and find that wages are approximately 3 times as important as non-wage benefits in decisions and that workers will pay 70 cents of their hourly wage to avoid the uncertainty surrounding occupational choice. Chapter 2 develops and tests criteria for cross-validation bandwidth selection for derivatives of multidimensional densities; for conditional density, joint estimation of the numerator and denominator bandwidths performs best. Chapter 3 tests the additive separability of the teacher effect assumption common in the teacher value added models using data from the Los Angeles Unified School District, and finds that interacting the teacher indicator variables with a function of the students' lagged test scores captures most of the nonlinearities, preserves the heterogeneity of teacher effects, and provides more accurate estimates.

The dissertation of Matthew David Baird is approved.

Jennie Brand

Maria Cassanova

Rosa Matzkin

Moshe Buchinsky, Committee Chair

University of California, Los Angeles

2012

I dedicate this dissertation to my patient and loving wife, Michelle.

Contents

List of Figures

List of Tables

Preface

This dissertation explores three separate topics in microeconometrics. The first chapter presents a stochastic dynamic discrete choice model of occupation and job choice. I model wage and non-wage job offers to understand why agents change occupations, and what the effects are. I estimate the model using data from the National Longitudinal Study of Youth 1979. Wages are the most important factor in decisions to change jobs and occupations, but non-wage benefits are also important. Agents are better off most of the time, but not always. They are willing to pay on average 70 cents of their hourly wage to avoid the uncertainty surrounding occupational choice. Accurate beliefs and tenure improve efficiency. I would like to thank Moshe Buchinsky and Maria Casanova for their invaluable and continual help in this project, as well as Rosa Matzkin, Arturo Harker, Dan Ben-Moshe, and Peter Bergman, and many other seminar participants.

In Chapter 2, I develop the criteria and consistency of cross validation methods for bandwidth estimators for high dimensional derivatives of densities. I perform a Monte Carlo simulation to test the various criteria. The joint estimation of bandwidths for derivatives of conditional densities outperforms separate estimation. I would like to thank Rosa Matzkin for financial support and research help, as well as seminar participants for helpful advice, including Jinyong Hahn and Conan Snider.

In Chapter 3, I present joint research with co-author Peter Bergman analyzing data from the Los Angeles Unified School District. We experiment with semiparametric estimation strategies to test the additive separability of the teacher effect assumption common in teacher value added models. Interacting lagged test score with the teacher indicator variables captures most of the nonlinearities captured in the more flexible Ichimura non-separable indexed model. We would like to thank Rosa Matzkin, Jinyong Hahn and Moshe Buchinsky for research advice, as well as seminar participants for helpful advice.

I would also like to thank my wonderful and patient wife Michelle for all of her help and support, and to my family for all of their support throughout my education.

Curriculum Vitae

Education

- PhD Candidate Economics, University of California Los Angeles 2009
- Masters of Science Economics, University of California Los Angeles 2009
- B.S. Economics, B.A. History, Minor Mathematics, graduated with University Honors Brigham Young University 2007

Work

- Research Assistant University of California Los Angeles Economics Department, 2009-2011
- Visiting Instructor and Researcher Brigham Young University Economics Department, 2010
- Teaching Assistant University of California Economics Department, 2010-2011
- Research Assistant California Center for Population Research, 2007-2008
- Research Assistant Brigham Young University Economics Department, 2006-2007
- Teaching Assistant Brigham Young University Economics Department, 2005-2007
- Research Assistant Boston University School of Management, 2001,2004,2005

Awards, Fellowships, and Other Activities

- Dissertation Year Fellowship, University of California Los Angeles, 2011-2012
- Teaching Assistant Fellowship University of California Los Angeles Economics Department 2010-2011

- Research Fellowship, California Center for Population Research 2009-2010
- Graduate Summer Research Mentorship, University of California Los Angeles 2009
- NICHD Traineeship, California Center for Population Research 2008-2009
- PhD Progress Award 2008
- Gordon B. Hinckley Scholar, Brigham Young University's most prestigious undergraduate academic scholarship 2000-2006
- Referee: Economics of Education Review
- National Merit Scholarship 2000

Current Research

- Estimation of a Dynamic Model of Occupational Transitions, Wage and Non-Wage Benefits
- Cross Validation Bandwidth Selection for Derivatives of Various Dimensional Densities
- Testing the Additive Separability of the Teacher Value Added Effect Semiparametrically (with Peter Bergman)
- Estimation of Nonparametric Distributions and Coefficients in an Earnings Dynamics Model (with Dan Ben-Moshe)

Chapter 1

Estimation of a Dynamic Model of Occupational Transitions, Wage and Non-Wage Benefits

1.1 Introduction

Occupational transitions are increasing in frequency.[1] Transitions are not costless; occupational-specific human capital is lost when switching away from an occupation and there are direct switching costs, such as moving. However, if workers are in sub-optimal occupations, the costs are mitigated by the benefits of moving into an occupation where they are more productive and receive better benefits.

The proportion of workers switching occupations is high enough to warrant attention, especially at younger ages. Kambourov and Manovskii (2008) use the Panel Study of Income Dynamics (PSID) and estimate the average level of occupational mobility is around 13% at

[1]This result is documented by Kambourov and Manovskii (2008) using the PSID, Moscarini and Vella (2003) using NLSY, and Perrado et al. (2007) using the PSID. For example, Kambourov and Manovskii (2008) find, from 1968 to 1997, occupational mobility has increased from 10% to 15% at the one-digit level, from 12% to 17% at the two-digit level, and from 16% to 20% at the three-digit level. Digit levels are the level of specificity that the census employs; thus, the one-digit level yields 9 different occupations.

the one-digit level, 15% at the two-digit level, and 18% at the three-digit level. Markey and Parks (1989) find similar rates in the January Current Population Survey 1987, and Parrado et al. (2007) also find a 7% to 11% change at one digit with the PSID. Using the Duncan Index they also find increasing transition rates over time (through the 1970s and 1980s), increasing from 20.1% for the 1970s to 26.3% for the 1980s.[2] Moscarini and Vella (2003) find higher transition rates at the three-digit level using the National Longitudinal Study of Youth (NLSY) at transition rates of 57% to 70% when measuring occupation at a three digit level. In the data sample used in this chapter from the NLSY, the transition rates, at 2 occupations, are higher than other one-digit levels because the data here is restricted to young men, below age 32. Young men transition more often than older men, a trend which this chapter will examine.

Occupational transitions are correlated with various factors, such as increased wages. The effects of the transitions can, as a result, have a strong impact on the welfare of the workers. Markey and Parks (1989), using data from the January Current Population Survey 1987, show that around 90% of workers who change occupations report receiving higher wages, while many cite improved working situations. Longhi and Brynin (2010) use British and German data sets, and in both samples find that both job and occupational transitions yield positive changes in both wage and job satisfaction, with job changes that are also occupation changes being the best off. The NLSY sample used in this chapter has a lower value, at around 60%, partially as a results of using a younger sample of workers. Delfgaauw (2007) uses 2003 survey data from the Netherlands and finds that lower job satisfaction leads to higher job search, and that the type of dissatisfaction leads to where they look for a new job (within the same organization or to a different organization). While Delfgaauw (2007) does not investigate the effect of actually changing jobs or occupation on satisfaction, his research illustrates the reasons for an increase in satisfaction when the change is voluntary.

Although occupation transitions on average are correlated with improved wage and non-

[2]The Duncan index is the sum across occupations of the absolute values of the change in the percentage of employment

wage benefits, there are also costs, including opportunity costs, associated with a transition. When a worker switches occupations, they forfeit the accrued benefits of the returns to occupation-specific knowledge. Kambourov and Manovskii (2008) find, ceteris paribus, that five years of occupational experience is correlated with at least a 12% increase in wages, while Buchinsky et al. (2010) use data from the PSID and find a return of 3.77-6.33% for each additional year of occupation experience. The research in this chapter finds values in between these, at 5.68-11.3%.

Using an estimated structural dynamic model, this chapter demonstrates that the uncertainty surrounding occupational choice is more important to the youngest workers than the uncertainty surrounding which job to choose. The chapter also shows that the common use of lagged wage or non-wage offers as proxy for unaccepted offers can be misleading. A one percent change in the wage offer changes the probability agents switch occupations by 3.1-28 percentage points on average. Wage is 1.8-8.9 times as important to workers as non-wage benefits in the decision to change occupations, and agents, while better off on average for an occupation switch, have a higher probability of making an ex-post efficient change with higher occupational experience, and different direction effects for job tenure, skill level, and accuracy of beliefs depending on which occupations they are switching from.

1.1.1 Models of Occupation Transitions

The model in this chapter is an expansion of previous economic models of occupational choice. The Roy model (1951) allows workers to self-select into occupations based on wage-specific skills and the returns to those skills. These elements are incorporated into the model used in this chapter. One common model used to explain occupation and job transition is a matching model. Matching models differentiate workers into types (e.g., high skill and low skill) and also differentiate the labor market into different skill segments.[3] Léné's model (2011) allows experience and education to be imperfect substitutes for each other, a feature

[3]Examples of matching models include Pissarides (1990), Van Ours and Ridder (1995), Gautier (2002), Léné (2011), Albrecht and Vroman (2002), and Dolado et al. (2009)

of my model. Léné's model and the results suggest that there is an entry cost to different labor segments; my model also includes these costs.

Blumen et al. (1955) develop a mover-stayer model, where high-productive workers are better able to retain jobs than low productive workers and therefore higher educated workers have lower transition rates. The search good model (Burdett 1978, Jovanovic 1979) has transitions that are chosen by workers to improve situations through higher utility. The model in this chapter is a form of a search good model. Search good models predict lower transition rates with increased ages. Older workers have had more time to search and find better jobs and occupations (in terms of overall satisfaction, which will include wage and non-wage factors) and stay there.

The occupation decision process is well explained and matched in the data with a non-linear, theory-backed dynamic model. Similar to this chapter, Keane and Wolpin (1997) estimate a dynamic model of occupational choice using data from NLSY on young men. Others have followed the use of their discrete choice dynamic model application to occupational choice.[4] Postel-Vinay and Robin (2002) create and estimate a model that examines firm and employee heterogeneity using data from the 1996-1998 in Déclarations Annuelles des Données Sociales, a data set of employers and employees from France, and look at equilibrium effects on wages. Meinicke's (2010) model is a dynamic occupational choice model with Bayesian updating for an individual-specific effect on wages. Agents have imperfect knowledge about their ability within different occupations, but each period they work they are able to observe a noisy measurement of their productivity and update their beliefs about the unobserved portion of their ability. This chapter also has agents update their beliefs about their productivity, but allows firms to update beliefs about the worker's productivity as well, and change wage offers according to their beliefs, similar to Felli and Harris (1996) and Gibbons and Waldman (1999) in a theoretical model. The model in this chapter also includes non-wage incentives as well as job choice and firing to expand upon the reasons

[4]For example, Lee 2005, Lee and Wolpin 2006, and Dix-Carneiro (2010)–see also Aguirregabiria and Mira (2010) and Keane and Wolpin (2008) for a review on these types of models

for the agent's occupational choice. Kunze and Troske (2011) show that workers are by age differentially affected by layoffs and the log wage changes before and after the layoff.

Motivated by the past research on these changes, I look to investigate the decision as a dynamic process that is a function of job offers that combine both wages and non-wage benefits, a unique contribution for these dynamic models. I allow for occupation and job changes, and firing from the firms. Both agents and firms update their beliefs about the productivity of the workers. The addition of non-wage benefits helps the data match the model better by more accurately reflecting the reasons for occupational transition. The model also allows for the comparison of the relative importance of wages and job satisfaction, a unique contribution. Including wage and non-wage offers in the model gives insight into how effective of a proxy the lagged wage and non-wage values are for the unaccepted offers. Also, modeling the selection process controls for selection bias.

1.1.2 Preview of Results

My estimated dynamic model adds to the body of research by establishing five main results: 1) Occupational uncertainty is important to agents, as measured by how much agents will pay to know which occupation is best for them. 2) The use of the lagged wage or non-wage offer which is often used as a proxy for the unobserved wage/non-wage offer to measure improvements can be inaccurate and potentially misleading when used to measure if agents are better off and the reasons they change occupations. 3) The discounted sum of expected future utility is the most important factor in an agent's decision to change occupations, and the change in the wage offer is 1.8-8.9 times as important as the change in the non-wage offer. 4) Workers on average have higher lifetime welfare from an occupational change, and the probability of a good transition increases with higher occupation experience; other factors, such as job experience, skill level, and accuracy of the agent's beliefs about their own productivity, have different signed effects depending on the occupation from which they are switching. 5) The information that firms gain on worker productivity has a greater influence

on whether a worker will change occupations than the information that the worker gains on his own productivity.

Willingness to Pay for Occupational Knowledge

One important contribution a structural model can make is measuring counterfactuals and the willingness to pay for information. This chapter adds to the literature on occupational change by providing the first estimation of how much agents would pay to know which occupation in any given period is optimal for them in terms of life-time discounted utility.

The importance of occupation changes, in particular for workers just entering the labor market, is emphasized by the results of these estimates. 19 year olds are willing, on average, to pay over 10 percent of their hourly wage to know the correct occupation to be in: 0.967 dollars of their hourly wage, which is on average around 8 dollars. On the other hand, they will pay 0.929 dollars of their wage to know which job offer to accept; early on, making the correct occupational decision has greater long-term effects than job choice. However, as workers get older job choice becomes more important than occupation choice; for the entire sample, workers will pay on average 0.73 dollars to know which occupation to be in, but 0.93 dollars to know which job to choose. These results strengthen the argument for researching occupational change and in particular for studying the youngest workers who are making their initial occupation decisions.

Use of Lagged Wage/Non-Wages as Proxy for Offers

Reduced form research that examines the reasons for or the effects of occupation changes often use the lagged wage or non-wage benefits as a proxy for the current, unobserved, and unaccepted wage/non-wage offers. This is common throughout the literature, such as in Kambourov and Manovskii (2008), Markey and Parks (1989), Parrado et al. (2007), and Longhi and Brynin (2010). However, workers typically make decisions based on the wage/non-wage offer in their current job or occupation from the current period, and not

from the previous period.

I test how suitable of a proxy the lagged wage/non-wage benefit is for the unobserved current period offer during a switch. The correlation coefficients from the simulations between these two measures are .453-.819 for occupation changes and .379-.585 for job changes, both low. Using the lagged values as proxy tends to underestimate the benefits from transitioning. In this chapter, lagged proxies are avoided by the use of a structural model that includes job offers. By doing without the lagged proxies and its inherent complications of mis-measurement, the effects of a transition and the reasons why an agent makes a transition can be more accurately ascertained.

Explaining Transitions

Previous research has demonstrated connections between occupational transitions and many variables. Kambourov and Manovskii (2008) with the PSID and Markey and Parks (1989) with CPS data find mobility rates decline with worker's age and education. Parrado et al. (2007) come to the same result with the PSID, but argue that both effects have decreased over time. With data from the British Household Panel Survey across 1984-2006, Longhi and Brynin (2010) show that overqualified workers (those that have higher education than the average in their occupation) are less likely to change jobs within occupation. They also show that those with high wage residuals from their model, a proxy for unobserved factors affecting wage, are more likely to change jobs but remain within the occupation, and are less likely to change their occupation. They argue that these workers have high individual-specific effects, and remain within the occupation to gain the returns from their fixed effects. The model in this chapter includes a fixed unobserved portion of the wage equation unique to each individual.

There are potential biases from not properly accounting for the selection process. By directly modeling how occupations are chosen and switched between, I am able to have better estimates of the reasons why agents change. Table A.2 demonstrates that workers

making the transition from white collar to blue collar jobs have significantly lower wages and non-wage benefits in all categories than their respective peers, at statistically significant levels. White collar workers might switch into blue collar jobs because they are unable to garner the wages and non-wage benefits of their peers. They might be able to find higher wages and non-wage benefits in blue collar jobs. Alternatively, some of them lose their white collar jobs and use this as an opportunity to switch occupations. The model in this chapter allows for all of these decisions, and finds that wage is consistently a stronger factor than non-wage benefits in transition probabilities. For white collar workers, a one percent change in the wage offer (either in their current occupation or the alternative) has a 3.4-6.0 percentage point marginal effect on the probability they switch to a blue collar job. The effect is even stronger for blue collar workers changing to white collar (10.9-28 percentage point marginal effects).

Non-wage benefits also play a role, albeit a smaller one, in the decision to switch occupations and jobs. The average hourly wage in simulations is 12.43 dollars with a standard deviation of 8.56 dollars; for non-wage benefits, the estimated mean is 4.22 with a standard deviation of 0.43. The lower standard deviation as well as the lower overall value shows how non-wage benefits contribute less to the decision process, but does affect it. At mean values of wage and non-wage benefits in the utility function, the difference between the average non-wage benefits in white collar and blue collar is equivalent to giving 1.09 dollars higher hourly wage to blue collar workers, suggesting a difference between the two occupations that is unmeasured just by wage and the observed non-wage variables. The provision of health insurance and retirement benefits are equivalent to 0.19 dollars and 0.195 dollars hourly wage increases respectively. A pleasant work environment, going from a rating of 1 (worst) to 4 (best), is equivalent to a 0.38 dollar hourly wage increase, and an increase in job security from 1 to 4 is valued at a 0.388 dollar hourly wage increase. Changes in the wage offers are 1.843-4.218 and 4.855-8.923 times as important for blue and white collar workers respectively as changes in the non-wage offer in their impact on the probability a worker changes

occupations.

Welfare Effects

Research has established positive wage outcomes for occupation changes.[5] Linear regressions on my model show that leaving voluntarily leads to higher wages (Table A.1), although if the worker stays within the same occupation while they switch jobs, they have even higher wages (with two occupations; with ten occupations, log wages are higher if they switch occupations as well). For the non-wage benefits of health insurance, retirement, pleasant environment, and job security, I observe similar trends (Table A.3).

Reduced form research tends to look at the change in wages or non-wage benefits in the next period (such as in Light and McGarry 1998, Longhi and Brynin 2010, Perrado et al. 2007 and Wilson and Green 1990), but fails to capture changes in welfare in the long run. Many occupation changes are done for the benefits that will accrue over a span of years. The use of a structural model allows for a natural summary statistic of welfare from an occupational choice, as the discounted sum of utility combines wage and non-wage benefits across all future years.

To explore the welfare effects of an occupation change, I estimate welfare through the ex-post discounted sum of utility. Whenever a job change or an occupation change is made, I simulate the model forward twice to the end of the model, once where the worker makes the transition that was ex-ante optimal and once where they do not make the transition. I then compare the proportion of workers that are ex-post better off, and see what characteristics improve the likelihood of the transition being better. Workers are on average better off

[5]As previously mentioned, Perrado et al. (2007) use the PSID and find a negative but insignificant relationship between occupational change and wages. On the other hand, Longhi and Brynin (2010) with the British and German data and Wilson and Green (1990) find a positive relationship using the PSID. There are similar measurements for job mobility; see for example Light and McGarry (1998), who use data from the NLSY IV/GLS in the regression with deviations from in-job means for time varying variables as their primary instruments. They find a negative correlation between job mobility and wages, which is near zero at the beginning of an individual's life cycle but becomes more negative as they get older. Also, Brand (2006), using the Wisconsin Longitudinal Study, finds worse outcomes for displaced workers in both wage and non-wage benefits.

for both types of changes: 75.9 and 57.4 percent are better off for white and blue collar occupation changes, and 51.7 and and 74.0 percent for white and blue collar job changes. Each additional year of occupation experience makes a worker -0.6 to 6.2 percentage points more likely to make an ex-post efficient occupation change, and each log wage dollar closer in their beliefs about their true productivity makes them as much as 16.2 percentage points more likely to make ex-post efficient occupation transition (for workers switching into white collar jobs). Job changers are also more likely to make an ex-post efficient job change for each additional year of job tenure (0.94-1.28 percentage points more likely).

Bayesian Learning

One of the possible factors for the change in occupation and job change over the worker's life cycle is learning on the part of the agents and the firms. For the agent, learning how productive they are in different occupations increases the probability they stay in that occupation. It also encourages longer job tenures for high-productivity workers, as they expect firms to discover how productive they are and reward them accordingly, while switching jobs would lead to firms undervaluing their unique contributions. A dynamic structural model allows for direct estimation of the learning and the effects of that learning.

Other research has incorporated Bayesian learning into a dynamic model. Felli and Harris (1996) build a dynamic model of workers and firms where both are Bayesian learners about the productivity of the workers, and they show what equilibrium in their model looks like. In their model, agents are either low or high productivity workers, and they update their beliefs on the probability they are one or the other. In my model, workers and firms update their beliefs on a continuous parameter of productivity. Gibbons and Waldman (1999) develop a theoretical model where firms are Bayesian learners of workers productivity and adjust wages as beliefs change. Meinecke (2010) includes Bayesian learning and finds that there is significant learning that happens using NLSY data. However, he does not model the firms learning, and does not investigate the impact on transition rates. Like Meinecke (2010), I use

the updating methodology for partially observed data demonstrated by Ansley and Kohn (1983). Arozamena and Centeno (2006) have learning in their theoretical model, but not in the estimated, reduced form model. Ackerberg (2003) builds a model with dynamic learning in an advertising as signaling model. He builds on Eckstein, Horsky and Raban (1988), as do Clay, Goettler and Wolff (2004). Understanding the role of learning and limited information opens up the possibility of testing policies that increase initial information accuracy or the speed of learning.

In order to examine the impact of learning, I conduct counterfactuals; the results show that the firms' beliefs about productivity affect choices in the model more than workers' beliefs. When firm's have perfect knowledge about the workers' productivity, the proportion of white collar workers and those in education increase, and has different effects on the probability of changing occupations depending on which occupation they are in (white collar decreases and blue collar increases). The firms' uncertainty in the model results in higher occupation changes, as firms' beliefs set the wages. If the firms know the type of each worker perfectly, blue collar workers change occupations about the same–on average, 1.09 times as often–while there is a larger reduction in the frequency of transitions for white collar workers, at 0.5898 times as often.

1.1.3 Outline

Section 1.2 presents the model and Section 1.3 discusses the data, assumptions and restrictions made, and stylized trends and statistics in the data. Section 1.4 describes the empirical strategy. Section 1.5 explores the results and the implications, including counterfactual studies. Section 1.6 then concludes.

1.2 Theoretical Model

In this model, occupation is restricted to two choices: white collar and blue collar. The division into two occupations retains many of the trends of interest of a more finely divided categorization. For example, Tables A.1 and A.3 show the coefficients from linear regressions of next periods wage and non-wage benefits on current wage and non-wage benefits and various controls for changing job and occupation, and whether they were fired. The analysis is performed both at the 2 occupation level (white collar vs. blue collar) and 10 occupation level. The coefficients are very similar, suggesting that these trends are retained with only two occupations; they also motivate understanding the evolution of wages and non-wage benefits through job and occupation transitions.

Agents choose every period whether to work or attend school. At the beginning of each period, they receive an offer from one firm in each occupation. Each offer is a bundle that includes both a wage and non-wage benefits. If the worker has a previous employer, they also receive a continuation offer from their previous employer in each occupation. Agents maximize present value lifetime welfare by choosing every period to be in school or work in a white collar or blue collar job, as well as whether to stay at their firm when possible or accept a new job offer. These choices can be summarized in a sequence of dummy variables $\{d_{kjt}\}_{k,j,t}$. k is the choice of schooling ($k = 0$), white collar ($k = 1$), or blue collar ($k = 2$). j is the choice of a new firm ($j = 0$) or to stay at the old firm where possible ($j = 1$). t is the period.[6] Agents maximize lifetime discounted utility

$$E\left[\sum_{t=1}^{T}\delta^{t-1}\sum_{k,j}d_{kjt}U_{kjt}|S_t\right]$$

U_{kjt} is the per-period utility function. The utility function is constant elasticity of substi-

[6]For example, if the worker in period t has a job they can return to but accepts instead a job offer in blue collar, then $d_{00t} = 0$, $d_{10t} = 0$, $d_{11t} = 0$, $d_{20t} = 1$, and $d_{21t} = 0$

tution (CES) as a function of wages and non-wage benefits with an additive random utility shock. The values of these factors differ depending on their choice of d_{kjt}. S_t is a vector of state variables, some of which are individual specific (education, experience in an occupation, etc.) and some of which are population wide (occupation characteristics that affect job satisfaction). As the model possibilities differ depending on whether the agent has a job from the previous period that they can return to, each option is separately explained.

1.2.1 Utility Function: Agent Has No Current Employer

First, consider the model when the agents have no previous job to return to, j=0. This happens either from being the first period of the model, from choosing schooling the previous period, or from being laid off from their previous job. They chose between 3 options: education, blue collar work, or white collar work.

Consider the utility function from choosing schooling ($k = 0$):

$$U_{00t} = (w_{00t}^{\rho} + b_{00t}^{\rho})^{1/\rho} + \xi_{00t}$$

w_{00t} is the wage from schooling; it is set equal to a minimum consumption level chosen exogenously from the model to be the poverty line. b_{00t} is the non-wage benefit from schooling, and is equal to

$$b_{00t} = \theta_0 + \beta^{ED}\text{AFQT}$$

θ_0 is a parameter that measures the relative attractiveness of education vis-à-vis working. AFQT is the Armed Forces Qualification Test score, a common measure from the NLSY for mental aptitude. The extent to which agents receive non-wage benefits from schooling differs by mental aptitude because schooling can be easier depending on their ability. ξ_{00t}

is a random utility shock, with a mean zero normal distribution and variance σ^{ξ}, and is uncorrelated across occupation or schooling.

The utility for choosing to work in occupation $k = 1, 2$ at a new job is given by

$$U_{k0t} = (w_{k0t}^{\rho} + b_{k0t}^{\rho})^{1/\rho} + M_k + \xi_{k0t}$$

b_{k0t} is the non-wage benefits, given by

$$b_{k0t} = \theta_k + \beta^{NW} X_t^{NW}$$

θ_k is the parameter that measures the relative attractiveness of occupation k against the other choices. X_t^{NW} is a vector of non-wage benefits, such as whether they offer health insurance, retirement benefits, provide a pleasant working environment, or have good job security. M_k is the entry cost associated with starting a new job. w_{k0t} is the wage offer, modeled by

$$w_{k0t} = \exp\{\theta_k^W + \lambda_k + \gamma_k \text{AFQT} + \beta_{1k}^{EXP} \text{educ}_t + \sum_{\ell=2,3} \beta_{\ell k}^{EXP} \text{exper}_{\ell t} + \varepsilon_{k0}\}$$

θ_k^W is the occupation log wage intercept; γ_k is the return to ability (AFQT). λ_k is a firm-employee match parameter, unique to a firm and employee. Job offers from new firms come with a new, λ_k that is constant as long as the agent is working at that firm. λ_k is distributed normally with variance σ^{λ}. Agents are more productive in some firms than others and so are better compensated when working for those firms. If a worker stays with the same firm, then future periods' wage offers are a function of the same match parameter. educ_t is the number of years they have chosen to do schooling, so that β_{1k}^{EXP} is the return to education. White collar and blue collar occupations will reward education differently. exper_k is the amount of

experience the agent has in occupation k, given by the number of years they have chosen to work in that occupation. $\beta_{\ell k}^{EXP}$ is the return to experience in occupation k for an additional year worked in occupation ℓ. ε_{k0} is a random shock, correlated across occupations but not across time or with other variables.

1.2.2 Utility Function: Agent Has Current Employer

If the agent has a job that they can return to, then they have 5 choices: education, 2 new job offers, one for each occupation, and 2 continuation offers to stay with their current firm, one in each occupation. This can be viewed as a promotion or change of duties between white and blue collar.

3 of these utility functions are described above, namely U_{k0t} for $k = 0, .., 2$. Next, I show the continuation offers, U_{k1t} for $k = 1, 2$.

$$U_{k1t} = \left(w_{k1t}^{\rho} + b_{k1t}^{\rho}\right)^{1/\rho} + \xi_{k1t}$$

The utility function is similar to that of returning to a job, except that the wage offer differs and there is no job entry cost to be borne.

The wage is given now by

$$w_{k1t} = \exp\{\theta_k^W + \lambda_k + \gamma_k \text{afqt} + \beta_1^{EXP}\text{educ}_t + \sum_{\ell=2,3} \beta_\ell^{EXP}\text{exper}_{\ell t} + \beta^{TEN}\text{ten}_t + \widehat{\eta}_{kt}^F + \varepsilon_{k1}\}$$

The continuation wage function is similar to the initial wage offer, with two differences. The first difference is a return to tenure in a firm, or the number of years the agent has worked at the firm, $\beta^{TEN}\text{ten}_t$. I include returns to job tenure in the model because of evidence of its importance in the wage equation and how it affects job transitions.[7] The

[7]See Topel (91) for empirical evidence and Felli and Harris (96) for theoretical support

second difference is that firms learn about their workers productivity after each year worked, and adjust their wage offer according to their new beliefs.[8] $\widehat{\eta}_{kt}^{F}$ is the firm's estimate. Each worker has some true occupation-specific, time-invariant fixed productivity heterogeneity, η_k. This fixed heterogeneity is unobserved by the agents and the firms. At the end of each period, both the agent and the firm they worked for that period observe a noisy measurement of this parameter, and update their beliefs according to Bayes' Law. Firms adjust their wage offers depending on their beliefs concerning this parameter.

At the end of each period, there is a certain probability that a worker is laid off. These probabilities differ depending on various variables, such as their firm tenure, their education, which occupation they are in, and their age. These parameters come from exogenous estimations using the NLSY data, so they can be of use to obtain the firing probabilities necessary to simulate whether or not a worker is fired. The consequences of being fired are that they have no job to return to, and so have only the choices between new job offers to choose between, and the entry cost must be borne.

Additionally, in the data I observe global job satisfaction on a 1-4 scale, with 1 being the highest report. I use this data to help fit the model, and assume that reports come from the period utility function, given by whether the report, r, falls into certain parameter thresholds. Specifically,

$$r_{kjt} = \begin{cases} 1 & \text{if } U_{kjt} \leq q_1 \\ 2 & \text{if } q_1 < U_{kjt} \leq q_2 \\ 3 & \text{if } q_2 < U_{kjt} \leq q_3 \\ 4 & \text{if } U_{kjt} > q_3 \end{cases}$$

1.2.3 Value Functions

Given this setup, the value functions for the Bellman equation are as follows:

[8] See Gibbons and Waldman (1999) for theoretical support for the inclusion of firms updating beliefs

$$V_t(S_t) = \max_{k,j} \{V_{kjt}(S_t)\}$$

$V_{00t}(S_t)$ is the value function for choosing education, and is given by

$$V_{00t}(S_t) = U_{00t} + \delta E \left[\max_{j} \{V_{j0t+1}(S_{t+1}|d_{00t} = 1)\} \right]$$

$V_{k0t}(S_t)$ for $k = 1, ..., K$

$$V_{k0t}(S_t) = U_{k0t} + \delta E \left[\max_{j,d} \{\pi_{k0} V_{j0t+1}(S_{t+1}|d_{k0t} = 1) + (1 - \pi_{k0}) V_{jdt+1}(S_{t+1}|d_{k0t} = 1)\} \right]$$

π_{k0} is the probability that the worker is fired from their job. The firing probability is a function of AFQT, firm tenure, age, and education level.

The other value functions are very similar:

$$V_{k1t}(S_t) = U_{k1t} + \delta E \left[\max_{j,d} \{\pi_{k1} V_{j0t+1}(S_{t+1}|d_{k1t} = 1) + (1 - \pi_{k1}) V_{jdt+1}(S_{t+1}|d_{k1t} = 1)\} \right]$$

The terminal value function is given as a scalar multiple of the last periods' utility:

$$V_{kjT}(S_T|d_{kjT} = 1) = \delta_T U_{kjT}$$

1.2.4 Bayesian Updating

I assume that both the agents and the firms that they work for learn about a fixed ability parameter heterogeneous to each worker and occupation, given by η_k^*. All learners start with a zero mean prior. At the end of every working period, workers and firms observe a noisy measurement of η_k^* for the occupation k worked in, and update beliefs according to Bayes' Law. This is comparable to each period, workers and firms observing the productivity of each worker and differencing out all contributing factors to the productivity, such as education and experience. All that is left after the differencing is η_k^*, or fixed ability, and other unobservables. From this, workers and firms are able to get a better sense of the productivity of the worker. Firms adjust wage offers accordingly. While firms' lowering wages might seem odd, there is evidence of real wage decreases in firms.[9] The beliefs of the workers about their own productivity matter insofar as workers form expectations about their future wage streams based on what they expect the firms to learn about their own ability. A worker with a low productivity parameter in occupation k might choose to avoid job offers from firms in occupation k, expecting the firms to learn about their poor ability and lower wage offers in the future and lower their wage offer accordingly.

The updating is based on the work of Ansley and Kohn (1983), as also explained by Meinecke (2010). I use similar notation to Meinecke (2010). Every period, firms and workers receive a noisy measurement of η_k^* given by $\eta_{kt} = \eta_k^* + v_t$

Assume that the noise shocks are i.i.d. across individuals and time, but possibly correlated across occupations, and given by

$$v_t \sim N(0, \Omega)$$

[9] see McLaughlin (1994) and Card and Hyslop (1997) for empirical evidence, and Gibbon and Waldman (1997) for theoretical support

Let $\eta^* = (\eta_1^*, ..., \eta_K^*)'$. Then, at the beginning, each agent receives a draw from

$$\eta^* \sim N(0, \Sigma)$$

The agent uses Bayesian updating according to the following rules. Let the agent have an initial prior on η^* be given by

$$\widehat{\eta}_0 \sim N(0, \Sigma)$$

Then, Meinecke (2010) demonstrates how by the independence of η^* and v_t, given an observation of η_{kt}, and recognizing that only one η_{kt} is observed each period, for one occupation,

$$\begin{aligned}
\widehat{\eta}_{t+1} &= \widehat{\eta}_t + G_t D_t'(\eta_{kt} - \widehat{\eta}_{kt}) \\
G_t &= \widehat{\Sigma}_t D_t'(D_t(\widehat{\Sigma}_t + \Omega)D_t')^{-1} \\
\widehat{\Sigma}_{t+1} &= Q_t \widehat{\Sigma}_t Q_t' + G_t \Omega G_t' \\
Q_t &= I_K - G_t
\end{aligned}$$

The estimation of the wage fixed effect can be estimated given a wage history and occupational choice by these rules.

1.3 Data and Summary Statistics

The data is from the National Longitudinal Study of Youth 1997 (NLSY) from the years that have the pertinent data on non-wage benefits and job satisfaction, from 1979-1994. I restrict the data to agents observed from at least age 18, and limit the upper age to 31 years old to ensure sufficient observations for the later periods on which to estimate the

model. This captures the years that I am most interested in modeling, when occupations are decided, early learning happens, and transitions occur. The sample is restricted to males to abstract from fertility and other cultural incentives of non-work for females. The data is also restricted to agents that do not report being self-employed to improve the reliability of wage measurements. Military interviewees are also dropped from the sample. This results in 3,484 agents in the estimated model, with varying amounts of years observed for each agent.

I use the self-reported occupational status to classify their occupation.[10] Agents are classified as in school if they report their primary activity for a year as schooling. A further description of the methods used, including for selection of job assignment and transitions, is in the Appendix in Section A.1.

ρ (the CES substitution parameter) and the minimum consumption parameter are chosen exogenously. ρ is not identified in this model because there are no savings, and so no intertemporal substitution. I choose a value equal to $\rho = 0.75$, consistent with the estimates in Mankiw, Rotemberg, and Summers (1985). This is equivalent to a substitution between consumption and non-wage benefits of $1/(1-\rho) = 4$. As for the minimum consumption parameter, I set it equal to the poverty line in 1989, inflated using the same GNP deflator into 2005 dollars to match the real wage data. The poverty line was 6,310 dollars (US Bureau of the Census 1993), which inflated into real 2005 dollars is 9,079 dollars. For a forty hour a week job, worked for 50 weeks in a year, this is equivalent to a 4.54 dollar hourly wage, which is the minimum consumption parameter value used in this model.

Table 1.1 presents some of the summary statistics associated with this model. The average log wage is higher in white collar, but so is the variance. This, along with job satisfaction being higher (lower numbers are better, on a scale of 1-4), shows that on average, white collar jobs are better. That more people are in blue collar jobs for many periods suggests individual heterogeneity allows some workers to gain the better white collar offers. That the average AFQT scores are so different, with white collar workers 14 points higher on average,

[10] I code their occupation as white collar if the 1990 census occupation recoding is under 400 and blue collar if the occupation code is over 400. see Table A.5 for a description of the classifications

or approximately half a standard deviation, reinforces the differences between the employees in the two occupations. The likelihood of a job change is about twice that of an occupational change, a trend that my model investigates.

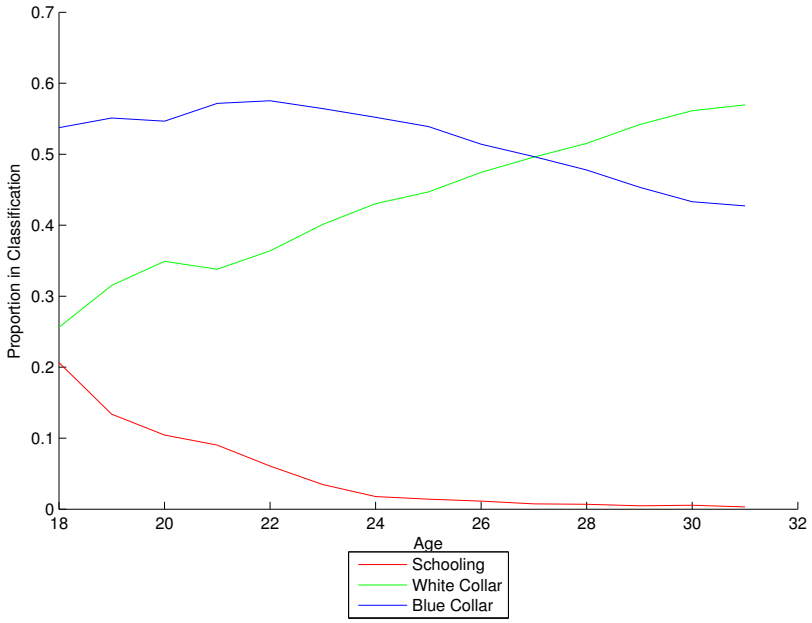
Table 1.1: Summary Statistics

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
ln(wage)	6.92207	0.565264	0.267677	15.1725	32745
ln(wage), White Collar	7.09365	0.603223	0.267677	13.2723	9864
ln(wage), Blue Collar	6.8481	0.531275	0.325748	15.1725	22881
AFQT	40.5254	29.3399	0	100	3484
AFQT, White Collar	47.2523	30.2847	0	100	19439
AFQT, Blue Collar	33.6001	26.4759	0	100	23730
Change Occs	0.206136	0.404535	0	1	38232
Change Job	0.397138	0.489312	0	1	35494
Change Job Voluntarily	0.618338	0.485813	0	1	12815
Job Satisfaction	1.79517	0.735337	1	4	22492
Job Sat., White Collar	1.6541	0.697299	1	4	10026
Job Sat., Blue Collar	1.79517	0.735337	1	4	22492

Figure 1.1 shows the proportion of the sample in each occupation by their age. The proportion of people in school steadily decreases over the entire period. School is an investment in the future, which a dynamic model is able to capture because of the future pay-offs. The highest proportion of people is in blue collar early on, and there is a sharp increase until age 21, after which there is a slow decline. On the other hand, there is only an increase in the proportion in white collar occupations, partially driven as high-ability agents in school enter the workforce into white collar jobs.

Figure 1.2 presents the proportion changing occupations, separated into the effect by origin occupation. Both trends are overall decreasing, but the proportion of workers leaving blue collar jobs does not decrease much over time. There are different types of white collar jobs that agents switch out of early or stay in later, captured in the model with higher wages and non-wage benefits. The model in this chapter attributes the overall decrease in switching

Figure 1.1: NLSY Sample Proportion in Each Choice by Age



to learning over time, to occupation-specific human capital that would be lost, to less time for benefits to offset new occupation entry costs, and to different types of jobs being offered.

Figure 1.2: NLSY Sample Proportion Changing Occupations, by Origin Occupation

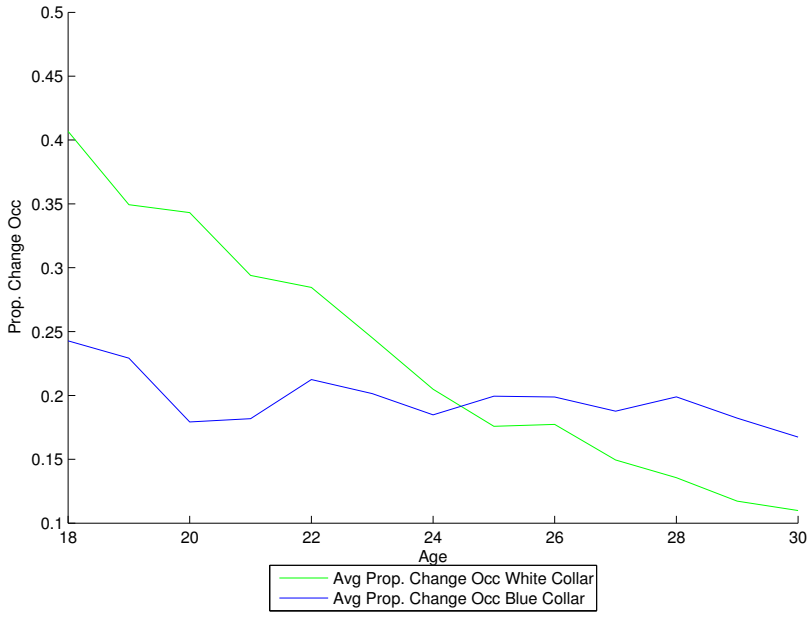


Figure A.1a shows average natural log of real wages by age. I deflate the wages using a

Gross National Product (GNP) deflater to put wages in terms of 2005 dollars. White collar workers have similar wages as blue collar workers on average in younger years, but follow a different trajectory after only a few years. Some of these white collar workers are finishing up school and joining the work force with higher wages, pulling up the average. White color jobs in general have a higher wage growth profile. There is also sorting into white collar by the highly productive, pulling up the average wage at a higher rate. Average job satisfaction (Figure A.1b) steadily improves (lower numbers) for both white and blue collar workers, and at about the same rate. However, white collar workers consistently report, on average, higher job satisfaction. The overall improvement in job satisfaction might be reflective of job and occupation sorting into both occupations, as workers are finding situations in which they are happy and comfortable more often as more time passes.

Figure A.1c shows the proportion of individuals changing jobs from one year to the next, by origin occupation. There is a decrease in this proportion as they get older, and blue collar workers regularly are more likely to change their jobs.

Figure A.2 presents the trends for the four non-wage benefits used in estimation of this model. The data is not available for all years for the final three variables; however, enough years are present to allow for the regressions used in the model. There seems to be an overall increase in each variable for both white and blue collar, except for perhaps in job security. Workers are getting into better matches as more time passes.

Figure A.3a shows two conditional probabilities. The first is the proportion of job changes that are within occupation (the complement being job changes that change occupations). This trend decreases over time, as more and more job changes are part of a change in occupation. Using a dynamic model that includes job offers will help capture the occupation changes happening. As the proportion of actual job changes is decreasing over this time as well, the decrease in overall job changes is outpaced by the decrease in those changing jobs within occupation, and when a job change does happen, workers are more likely to change occupations as well. The other plot in Figure A.3a shows the proportion of occupation

changes that are within the same firm (the complement being occupation changes that change jobs). This decreases slightly over time as well, but not as dramatically. Overall, about 20% of workers that are changing occupations do so within the same job. This could be a promotion or just a change in duties and responsibilities.

The model uses three groups of auxiliary regressions estimated from outside the model: 1) logit regressions to estimate probabilities of being fired, used to estimate firing probabilities (Table A.7) 2) logit and OLS regressions to estimate probabilities of receiving non-wage benefits or the level of the non-wage benefits, used to model non-wage benefit offers (Table A.6), and 3) a regression of the probability that workers change jobs regressed on their log wage, fringe benefits and other controls; the regression coefficients are part of the minimization criterion, used to help the fit of the model, and in particular to help with identification of the separate non-wage benefits (Table A.8).

The sample used to estimate this model is representative of young men in the United States in general, a group of particular interest with regards to occupation changes for many studies. More occupation switches happen at younger ages, and these decisions can have strong long-run effects, as highlighted by the higher amount of money workers would pay at younger ages to know which occupation is ideal for them (Table 1.3). Focusing on workers between the ages of 18 and 32 represents the high switching group. The other restrictions made (male, non-military, non-self-employed) keep the focus on the groups for which researchers would be primarily interested. The NLSY is a nationally representative sample of youth during this time period. While the data in use is from 1978-1994, and the economy and parameters of the model are possibly different 20 years later, the research in this chapter can be applied to understand historical movements and decisions and can generally be applied to understand the dynamics of changes today.

1.4 Estimation

The model is estimated using 14 periods. The value functions are solved recursively for all periods. The expectations of the maximum value functions in any given period are estimated using simulation, i.e.

$$E\left[\max_{k,j}\left\{V_{kjt+1}(S_{t+1}|d_{kjt}=1)\right\}|d_{kjt}=1\right]=\frac{1}{R}\sum_{r=1}^{R}\left[\max_{k,j}\left\{V^{r}_{kjt+1}(S_{t+1}|d_{kjt}=1)\right\}|d_{kjt}=1\right]$$

where $V^{r}_{kjt+1}(S_{t+1}|d_{kjt}=1)$ is the value function given specific draws of the random shocks ξ (random utility shocks), ε (wage shocks), v (measurement error on η^{*}), as well as shocks that determine whether they are fired or not and whether they receive different non-wage benefits.

Further, given the large state space, I use an interpolation technique, as suggested by Rust (1997), by taking random draws from the state space at every time period and estimating the value functions at these points in the state space. I estimate and store the coefficients from a flexible linear regression with quadratics and certain interaction terms of the state space on the value.[11]

I use two occupations: white collar and blue collar. I include four non-wage benefits of a job: whether it includes health insurance, whether it includes retirement benefits, overall pleasantness of the job environment, and the perceived job security. All of these variables are available in the data. For the first two, as discrete variables, I estimate a logit model outside of this system on the NLSY data. I estimate the probabilities that their job include health insurance, for example, as a function of their age, age squared, education, AFQT score, age interacted with education, and age interacted with AFQT score. Then, in the simulations, I take a random uniform draw, and if the random draw exceeds the probability that, given their state, they would receive health insurance, then they are modeled as getting a job offer

[11] Similar in spirit also to Keane and Wolpin (1994); see Aguirregabiria and Mira (2010) for a review of this methodology

that includes health insurance. Each occupation has its own set of coefficients.

The latter two non-wage benefits are not binary. In the data, they rank the questions (such as pleasantness of the job) on a scale of 1 to 4, four being they most strongly agree. For these regressors, I use the same state variables, but use OLS to estimate parameters and offer an average score for them, given their state. This, plus a normal random shock (with variance also determined from the data), yields the continuous variable included in their job offer for these non-wage benefits.

The firing probabilities are estimated from the data, exogenous to the parameters of my model. The probability of being fired differs by occupation, estimated using a logit model including AFQT score, education, job tenure, age, age squared, and age interacted with education, job tenure, and AFQT score as regressors. Similar to the case of the binary non-wage benefits, random uniform draws are taken that, if they exceed the probability that that worker would be fired, conditional on their state, then the agents are fired in the simulations and have no job to return to the next period.

I estimate the parameters of the model using simulated annealing on a minimization criterion determined by the method of indirect inference.[12] The minimization criterion is the squared distance between the moments in the data and those predicted by the model in the simulations. Specifically, the moments (and weight put on those moments) are the proportion of agents in schooling, white collar, and blue collar at each age (weight of 20); mean log wage by occupation and age (weight of 2.5); standard deviation of log wages by occupation but not by age (weight of 25); proportion changing occupations by origin occupation and age (weight of 15); proportion changing jobs by occupation and age (weight of 4); proportion changing job voluntarily by occupation and age (weight of 1); average job satisfaction report by occupation and age (weight of 2); and an auxiliary regression of whether they changed occupation regressed on the log wage, the non-wage benefits (health insurance, retirement, job pleasantness, and job security), AFQT score, education, and age

[12]See Gouriéroux and Montford (1996) for a review on the method of indirect inference.

(weight of 2).

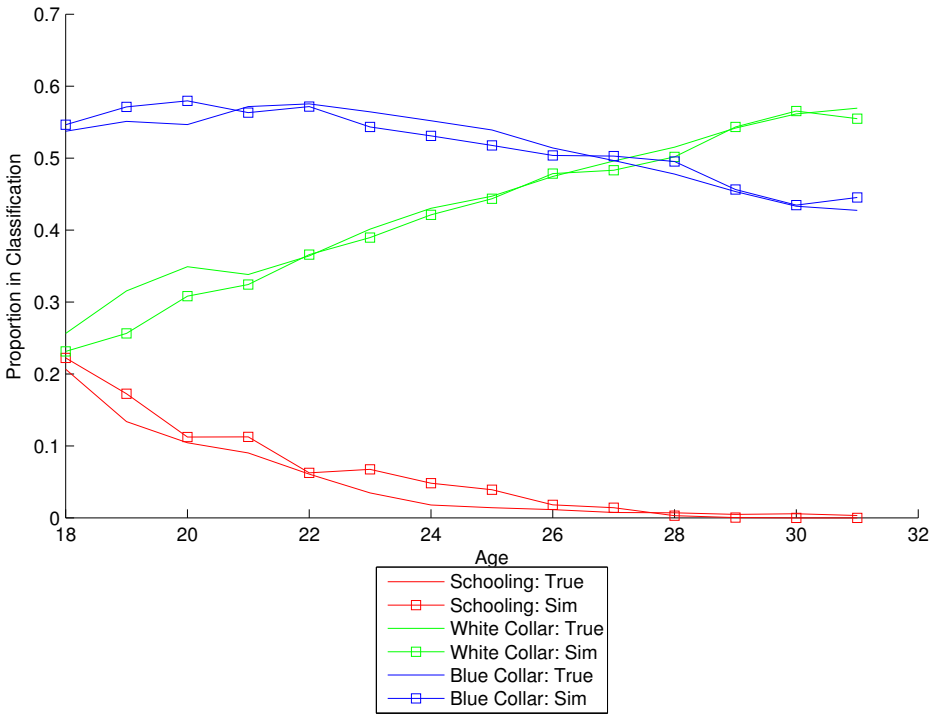
The average differences in occupation selection after controlling for the wage and observed non-wage benefits contribute to the identification of the non-wage threshold parameters $\theta_1 - \theta_3$. The parameters are identified mechanically because there is no constant in the utility function and no weights on wage or non-wage benefits. The margin across which agents with different observed AFQT scores but the same wages choose different occupations helps identify β^{ED}, the non-wage benefit return to AFQT. How non-wage benefits affect the probability agents choose different occupations and make occupation and job changes in response to non-wage offers contribute to the identification of β_j^{FV}. Observation of the log wage and its residual for different agents after controlling for agent fixed effects help identify the variance-covariance matrix of idiosyncratic log wage shocks. Observing the wage over time for individuals and separating out the fixed effects, and estimating the variance across individuals and occupations contributes to the identification of the variance-covariance matrix of individual fixed heterogeneity. The margin of observed log wages for workers with different education and occupation experience levels helps identify the returns to occupation experience and schooling. The different behavior of those already in a job and those not in a job (through firing or schooling) and the margin across which agents will switch jobs help towards the identification of M_k, the job entry costs. Direct observation of the log wage contributes to identification of the log wage intercept θ_k^w. Seeing the different behavior of agents in the final period of the model and previous periods, and how much importance they place in the second to last period on the value of the last period (which is a direction function of the scaling parameter) and the utility in the second to last period help identify the terminal value function scaling parameter δ_T. Agents with different AFQT scores and all else equal receiving different wages help identify the log wage returns to ability γ_k. The variance of utility shocks σ^ξ is identified through the model; it is not a parameter of interest, and so non-parametric identification is not important. The different wages of agents with different job tenures contribute to the identification of the return to job tenure β^{TEN}. The frequency of

job changes contributes to the identification of the firm-employee match parameter variance σ^{λ}, increasing with a higher variance in the parameter. The job satisfaction report thresholds q_j are identified through the different job satisfaction reports and the model utility; these thresholds are model-specific and not nonparametrically identified or of interest outside of this model.

1.5 Results

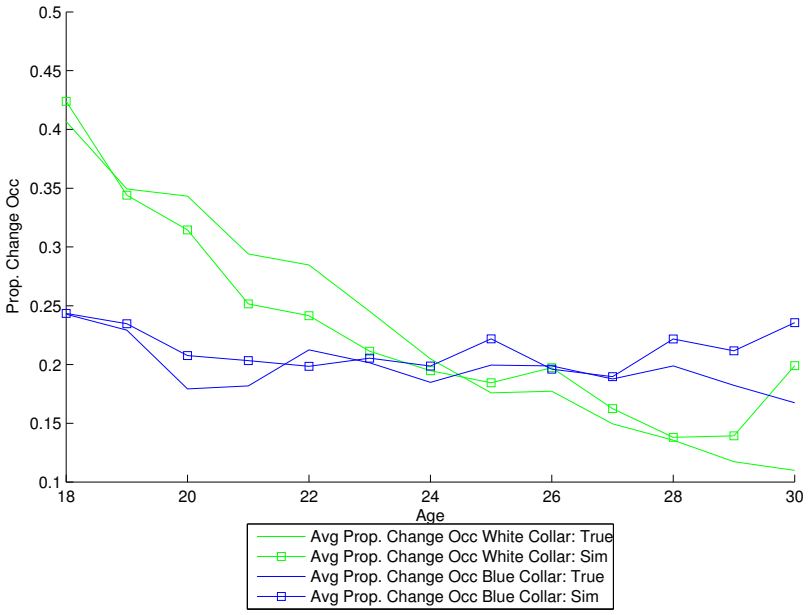
Figure 1.3 presents the true and simulated proportion in each occupation at different ages. The model does very well in matching these trends, providing partial evidence in favor of the model explaining the data. Figure 1.4 shows the proportion of workers switching occupations, by origin occupation.[13] Again, the model fits the data well.

Figure 1.3: Proportion that Choose Each Occupation: Simulated vs. True



[13]Origin occupation implies that the graph labeled Blue Collar represents the fraction of workers that were in a blue collar job and switched to a white collar job, out of the working population of blue collar workers at a given age. The graph labeled white collar documents transitions in the other direction.

Figure 1.4: Proportion Changing Occupations: Simulated vs. True



Figures A.4a-A.4c show the estimated vs. true trends for the proportion changes, average wages, and average job satisfaction report. The overall trends are captured and the model is performing well.

Given utility is a function of wages and non-wage benefits, the means and variances of the wage and the weighted sum of non-wage benefits give information about the decision process of the agents. Table 1.2 presents the average (over the simulations) for the sample means and standard deviations. The wage is the hourly wage in cents. Given the utility function, the units of the non-wage benefits, as a substitute for wage, can roughly be interpreted in terms of cents as well. The average wage is larger than the average non-wage benefits, which in sum is equal to 4.22 dollars of the hourly wage in addition to the 12.43 dollars they are paid. The standard deviation as a fraction of the mean for the wage is much larger than that of non-wage benefits, suggesting that the higher variation in wages is important in occupation choice and job sorting.

The estimated parameters for the structural model are given in Table A.10. The estimated parameters highlight the difference between white collar jobs and blue collar jobs.

Table 1.2: Average Means and Standard Deviations of Wages and Non-Wage Benefits

	Mean	Standard Deviation
Wage	1243.19	856.038
Non-Wage Ben	422.089	43.936

The log wage returns to education are higher for white collar than blue collar, although the difference is not as high as the difference in the returns to ability (AFQT, seen in γ_j). On the other hand, there is a higher non-wage benefit level for blue collar workers, all else equal, as seen by the differences in θ. White collar jobs tend to pay more and tend to have better observed benefits, but between white collar and blue collar jobs if all other factors were equal, workers on average prefer blue collar jobs.

The correlation between the different occupations' wage shocks are negative, while the correlation between unobserved heterogeneities in occupations are positive. The negative correlation helps spur occupation changes, while the positive heterogeneity correlation implies a large fraction of workers that are better at white collar jobs will also be better at blue collar jobs than the average. Relative differences will drive the sorting.

The return to job tenure is lower than the return to occupation experience, at about one half to one quarter the size; for example, an additional blue collar year in a blue collar job has an estimated return of 5.7%, while an additional year at the same employer increases log wages only 3.12%. The results are in line with previous research. Kambourov and Manovskii (2002) show five years of occupational experience have at least a 12% increase in wages, and Buchinsky et al. (2010) with PSID data estimate a return of 3.77-6.33% for each additional year of occupation experience. The results in this chapter estimate returns that range from 5.68-11.3, within the range of the other papers.

One way to investigate the dynamics of the model and the various contributing factors to occupation and job changes is to look at how a change in a parameter affects the proportions of workers changing occupation and job. I estimate the analytic gradient, the changes in the

proportion of workers making these changes as each parameter is changed on the margin.[14] The results are given in Tables A.11-A.13.

Changes in parameters that make the two occupations (or jobs) closer substitutes increase probabilities of changes. For example, white collar jobs tend to have higher wages, so any change in parameters that increase the wages of blue collar jobs (such as returns to blue collar experience) or decrease the wages of white collar lead to an increase of the probability that individuals change occupations. On the other hand, making the non-wage benefits less important also makes the jobs in the different occupations less different, and so decreases the proportion changing. An increase in the covariance between wage shocks increases the likelihood of switching occupations, as the wage offers are more likely to be similar. This reinforces the differences between white collar and blue collar jobs and workers. Increasing the appeal of education also makes workers less likely to change jobs voluntarily.

1.5.1 Willingness to Pay for Occupational Knowledge

Using a structural model enables calculation of how much agents will pay to avoid the uncertainty of not knowing the optimal occupation to be in for a given period. The calculation estimates 1) a worker's lifetime utility if they knew whether switching or not was optimal and 2) the worker's lifetime utility if they didn't know. From these estimates, I determine how much money agents would pay in the current period to know for sure which occupation to be in, i.e. the monetary transfer that would equate the lifetime utility from knowing and from not knowing. Table 1.3 shows the results for occupation changes, and Table 1.4 for job changes. Occupational uncertainty is almost as important to workers as job placement uncertainty, and even more important when workers are young. The average is approximately the same for white collar and blue collar workers and for whether they have a job to return

[14] Let Θ be the vector of all of the parameters. e_j is a vector of zeros with a value of one for the j^{th} element. The gradient of changes in the proportion changing occupations with respect to the various parameters is given by $\partial Pr(changeocc)/\partial\Theta$ by $\frac{\partial Pr(\widehat{changeocc})}{\partial\Theta_j} = \frac{Prop(changeocc|\Theta+e_j s)-Prop(changeocc|\Theta)}{s}$. s is set to be one percent of each parameter value.

to for occupation changes. There is a larger difference depending on the age of the worker. The youngest workers that are faced with the decision of switching or not (19 year olds) are willing to pay over a dollar an hour for white collar workers and 0.93 dollars for blue collar workers, on average. The average hourly wages for 19 year old workers are 8.09 and 8.24 dollars for white and blue collar workers, respectively, so that workers at age 19 are willing to give 12.9 and 11.3 percent of their hourly wage to be certain of which occupation they should be in that early on. The uncertainty associated with which occupation to be in, in particular early on, are quite large, and show understanding the dynamics surrounding occupational transitions are important to understand for young workers.

Table 1.3: Portion of Their Hourly Wage Agents Are Willing to Pay to Remove Uncertainty About Which Occupation

		White Collar	Blue Collar	Both Occupations
All Ages	Has No Current Job	0.675	0.745	0.718
		(0.00190)	(0.00178)	(0.00183)
	Has Current Job	0.750	0.724	0.737
		(0.00218)	(0.00201)	(0.00209)
	Both Job Situations	0.726	0.732	0.730
		(0.00210)	(0.00192)	(0.00200)
19 Year-Olds	Has No Current Job	0.953	0.915	0.927
		(0.00756)	(0.00667)	(0.00696)
	Has Current Job	1.081	0.938	0.988
		(0.00877)	(0.00787)	(0.00819)
	Both Job Situations	1.041	0.930	0.967
		(0.00841)	(0.00747)	(0.00780)

Standard Error of the Mean in Parentheses

For the uncertainty surrounding which job to choose, there is not a large difference between 19 year olds and the whole sample. The differing decisions of which occupation to be in have longer lasting effects through the returns to occupation experience and the agents' learning than which job to take, and so have a larger change in how much agents will pay to be certain as they grow older. The effects of choosing the right occupation early on have long-lasting effects. Choosing the wrong occupation can mean the worker is stuck there for years as the opportunity cost of switching increases with higher occupational experience.

Table 1.4: Hourly Wage Agents Are Willing to Pay to Remove Uncertainty About Which Job

	White Collar	Blue Collar	Both Occupations
All Ages	1.3178	0.5955	0.9312
	(0.0040)	(0.0023)	(0.0032)
19 Year-Olds	1.2829	0.7561	0.9286
	(0.0117)	(0.0084)	(0.0097)

Standard Error of the Mean in Parentheses

1.5.2 Use of Lagged Wage/Non-Wages as Proxy for Offers

The research on the welfare effects of occupational change use either changes in wages or non-wage benefits to measure welfare. For example, Longhi and Brynin (2010) and Wilson and Green (1990) conclude that transitions are efficient based on reduced form estimates of increased wages. However, typically unable to see wage and non-wage offers, these papers use the previous period's wage and non-wage benefits as proxies for what wage and non-wage offers the agents received before they switched.[15] Having a model that explicitly accounts for the counterfactual offers helps explain how often this is a valid proxy.

The question is what value to use for the wage in their old (origin) occupation: the wage offer they were given in the choice they did not make in the current switching period (typically unobserved) or the lagged observed wage from the previous period (the value often used). Figure A.5a presents kernel density estimates of the difference between the wage in a worker's new occupation and either the unaccepted wage offer in their old occupation or the lagged wage proxy. For both blue collar and white collar workers, using the lagged wage proxy underestimates the wage growth from switching occupations. Those switching from blue collar to white collar experience on average a higher increase. However, not all workers experience a wage increase from switching occupations among those fired. Table A.14 presents some of these trends. 95 percent have a wage increase for switching from blue collar after being fired from a white collar job (but only 83 percent if we use the lagged

[15] As done, for example, in Light and McGarry (1998), Longhi and Brynin (2010), Perrado et al. (2007) and Wilson and Green (1990).

test score). For those switching from white collar jobs, 71 percent have a higher wage after the change. Workers switching from white collar to blue collar are more likely to have better non-wage benefits, a strong contrast to the trends for blue collar. The correlations between using the true wage offers and the lagged values reveal that the common practice of using the previous periods' values can be misleading and understated. Even for the case of changing jobs, when both show very high rates of higher wages and non-wage benefits for changing jobs, the correlation between the differences is surprisingly low, and they tend to underpredict the true difference (except for the wage white collar case).

As Tables A.14-A.16 show, agents tend to have higher next period wages and non-wage benefits for a transition. There is a fraction of agents that accept job offers with lower wages or lower non-wage benefits, emphasizing the various motivations for changing occupations or jobs.

1.5.3 Explaining Transitions

Workers choose occupation (and thus whether or not to change) based on which choice maximizes

$$V_{kdt}(S_t) = (w_{kdt}^{\rho} + b_{kdt}^{\rho})^{1/\rho} + \xi_{kdt} + \delta E\left[V_{t+1}(S_{t+1}|d_{kdt} = 1)\right]$$

The choice is a function of wage, non-wage benefits, expectations for the future stream of utility, and a random shock. The change in the values of these factors and the occupation changing decisions of agents are informative for how agents respond to changes in the actors. If workers choose a wage/non-wage offer that has a lower wage, for example, we know that at least for some workers, they are not just deciding on wages. Using the structural model allows for estimation of the effect of wage and non-wage offers on transition probabilities, and not lagged proxy values, which would bias the results.

For every period for each agent in the model, I store the percentage changes in each of the four factors (wage, non-wage, expectations for the future utility, and the random shock), as well as the lag baseline levels for each. Using a probit model, I estimate predictions for how each factor affects the probability that they switch occupations. I estimate the regressions separately for which occupation the agents previously were in. The regressors are a quadratic in age, the percentage increase in the four factors, for both the current occupation and the alternative occupation, and the previous period's values for each of these variables. Table 1.5 shows the estimated marginal effects from the probit model by origin occupation for those without a previous firm to return to (because they were laid off, or had been in school). Tables A.17 and A.18 show the results for the probability of switching occupations for those who can return to a job and the results for the probability of changing jobs, respectively.

Table 1.5: Probit Estimated Marginal Effects on Changing Occupations: Has No Current Employer

Origin Occupation	White Collar	Blue Collar
Current Occupation: Wage	-0.0651***	-0.284***
	(0.00472)	(0.00589)
Current Occupation: Non-Wage Benefits	-0.0117***	-0.0734***
	(0.00329)	(0.00560)
Current Occupation: ξ	8.22e-06	-1.99e-05
	(9.80e-06)	(1.67e-05)
Current Occupation: EV	-2.849***	-10.25***
	(0.220)	(0.301)
Other Occupation: Wage	0.0636***	0.179***
	(0.00464)	(0.00487)
Other Occupation: Non-Wage Benefits	0.0131***	0.0971***
	(0.00259)	(0.00672)
Other Occupation: ξ	8.95e-06	-2.40e-06
	(1.09e-05)	(2.07e-06)
Other Occupation: EV	2.824***	10.18***
	(0.219)	(0.300)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Wage is an important factor in the decision process for every subset of the sample. For occupation changes, workers in blue collar jobs are much more responsive to wage changes

(and to every other change), while for job changes, white collar workers and blue collar workers respond to wage changes approximately equally. For workers who have jobs to return to, increases in the wage offer for the firm they are already working for have a larger effect than offers from other firms; workers who are offered a higher wage to change occupations within the same firm are more likely to do so. The marginal effects are relatively large, especially for blue collar. For white collar workers, the marginal effects of a one percent change in a wage offer range from 3.1 to 6 percentage point change in the probability of occupation changes, while for blue collars, the range is 10.9 to 28 percentage points. Even at the lower end of the range, a five percent increase in the white collar wage offer makes a blue collar worker over 50 percentage points more likely to switch occupations, all else equal. These results are smaller than those found by Parrado el al. (2007), who determine a one percentage point increase in the wage offer increases the probability of switching occupations by 27.8 percentage points. However, Parrado el al. analyze using 8 occupations instead of 2, so there are more frequent occupation changes. I find job changers are much more responsive to changes in the wage at the firm they are currently at, with marginal effects of 24.6 to 36.5 percent increased probability of changing jobs for a one percent change in the wage offer.

Expectations for the future stream of utility are the most important factor in decisions to change occupations or jobs. Expectations for the future value capture all future wages and non-wage benefits from switching a job. That the future value has approximately 40 times as large of a marginal effect as that of wage is reasonable. Non-wage benefits play an important role in both the decision to change occupations and to change jobs. Non-wage benefits are relatively and absolutely more important for workers in blue collar switching to white collar than workers switching in the other direction. Percentage changes in the wage offer are only 1.843-4.248 times as important as percentage changes in the non-wage offer (for white collar workers, the range is 4.855-8.923 times as important). Workers change jobs because of non-wage offer changes as well, with wage being only 1.55 to 4.626 as important in the new job offer as the non-wage factor. While wage is always more important than

non-wage benefits for changing occupations or jobs, the non-wage benefits are an important aspect of why agents change occupations and change jobs, especially for workers changing from blue collar to white collar jobs.

I also look at the estimated marginal effects when the percentage increases are interacted with a quadratic in age in the probit model to allow the effect to change over time. Figures A.6a-A.6c show the estimated marginal effect by age for the three main factors. Increases in the wage offer become slightly more important as agents get older, while non-wage benefits become less important and expected future value's marginal effects decrease at an even sharper rate. The marginal effects for expectations for the future utility are approximately 3 times as large for 19 year olds as 31 year olds. After interacting the effects with age, the marginal effect from a wage change is as large as 40 percent.

1.5.4 Ex-Post Efficiency of Transitions

Whenever an agent decides to change occupations or jobs, I simulate the model forward along parallel paths, one where they do make this change, and one where they do not make the transition. This allows calculations of the discounted sum of lifetime utility under the two options and direct estimates of the difference in ex-post welfare. Figures A.7a and A.7b are kernel estimations of the densities of the differences in welfare for making the change and not making the change. There are many positive and many negative differences. Table 1.6 shows the fraction of the density above zero (those that are better off); most are better off. The proportions better off in lifetime welfare are lower though than the proportions with higher wages; the common practice of using wage increase to measure welfare benefits of occupation or job changes overestimate improvements. Those in white collar occupations are more likely to make good occupation changes to blue collar, while those in blue collar are more likely to make good job changes.

Figures A.8a and A.8b show the proportion with higher welfare for changing by age of switch. The longer a worker in a white collar occupation has been in the labor market,

Table 1.6: Proportions Ex-Post Better Off in Simulations for Transitioning

	White Collar	Blue Collar
Prop. Better, Changed Occupation	0.759	0.574
	(0.428)	(0.494)
Prop. Better, Changed Job	0.517	0.740
	(0.500)	(0.439)

Standard Deviation in Parentheses

the more likely they are to make an ex-post better occupation transition. This reflects the learning of the agents and the divergence of job offers in different occupations as agents specialize. For both occupations, the older agents are, the more likely they are to make an ex-post efficient job transition. A more thorough investigation through estimating the contributing factors to ex-post efficient transitions using estimation of probit models helps to explain what characteristics of an agent and their situation improve the likelihood of an ex-post improving transition. I estimate the probit models for the probability the worker makes a welfare improving transition whenever they change. I also estimate probit models that include the difference in wage offers to see what portion of the probability that they do better is just from the higher wage offer in the changing period. The estimated marginal effects for occupation changers are given in Table 1.7. In the model, agents must either work or go to school each period; therefore, age is perfectly collinear with education and occupation experience, and is omitted as a control. Table A.19 presents the results for job changers.

Experience in the labor market matters, and in particular, the strongest positive effect for occupation changers is how long the worker has been in the occupation into which they are switching. Workers are as much as 6.2 percentage points more likely to be better off for a single year of occupation experience in the destination occupation, a hardly negligible factor. For job switchers, on the other hand, the effect is actually negative for white collar experience. The result is robust across the specifications, and is as large as -4.85 percentage point decrease. However, this is likely an indirect effect, in that workers would have been

Table 1.7: Probit Estimated Marginal Effects: Ex-Post Better Off for Occupation Change

	White Collar	Blue Collar	White Collar	Blue Collar
Years Education	0.00600	0.0153***	0.0306***	0.00266
	(0.00470)	(0.00512)	(0.00467)	(0.00520)
Years Experience WC	0.00150	0.0620***	0.00954***	0.0526***
	(0.00178)	(0.00208)	(0.00175)	(0.00215)
Years Experience BC	0.00646***	0.00516***	0.0190***	-0.00595***
	(0.00126)	(0.00136)	(0.00128)	(0.00142)
Years Job Tenure	0.0113***	-0.0191***	0.0185***	-0.00733***
	(0.00223)	(0.00220)	(0.00220)	(0.00227)
AFQT	-0.00341***	0.00781***	-0.00237***	0.00677***
	(0.000145)	(0.000169)	(0.000144)	(0.000173)
$\lvert\eta^*_{WC} - \widehat{\eta}_{WC}\rvert$	0.158***	-0.156***	0.186***	-0.162***
	(0.0416)	(0.0405)	(0.0409)	(0.0410)
$\lvert\eta^*_{BC} - \widehat{\eta}_{BC}\rvert$	0.0353	0.0433	0.0268	0.0383
	(0.0229)	(0.0286)	(0.0224)	(0.0290)
$\lvert\eta^*_{WC} - \widehat{\eta}^F_{WC}\rvert$	0.0937***	-0.0203	0.0974***	-0.0302
	(0.0306)	(0.0335)	(0.0300)	(0.0340)
$\lvert\eta^*_{BC} - \widehat{\eta}^F_{BC}\rvert$	-0.00111	0.0517***	-0.00214	0.0491***
	(0.0164)	(0.0174)	(0.0161)	(0.0176)
Difference in Wage Offers			0.335***	0.299***
			(0.00856)	(0.00936)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

better off with blue collar experience or education, and not that white collar experience actually makes them worse off.

Job tenure matters even more than occupation experience, but the effect is not always positive. For job switchers, more job tenure implies a better transition probability, by about 1 percentage point. The longer workers are with a firm, the more they know about the firm and are more likely to make informed switching decisions. However, for occupation changes, only workers switching from white collar to blue collar tend to be better off for higher job tenures. Workers switching the other direction sometimes leave a good job situation in a blue collar job for a seemingly better situation only to have the firms discover their type and lower their wage offers. On the other hand, AFQT score has the opposite effect. Only workers switching from blue collar to white collar jobs benefit from higher AFQT score. This

supports the idea that workers with lower AFQT scores can be worse off for switching into blue collar, where they might have hoped for a better long-term solution but not achieved it and missed out on blue collar years of experience. AFQT is correlated with all workers being worse off for a job change.

The accuracy of beliefs generally have no effect for job changers. The beliefs are concerning occupation-specific ability and not job-specific ability, so this comes as no surprise. However, accurate beliefs help the most for occupation switchers for their white collar productivity. For each log-wage dollar closer to their true productivity in white collar, workers are 15.6 percentage points more likely to be better off for switching into white collar. Firm beliefs' inaccuracy in the origin occupation is also correlated with improvements.

Workers are typically better off for their changes, but certain factors make them more likely to improve their situation from a change. Primarily, longer experience and tenure are the strongest factors, and the two different occupations are very similar for improving job change coefficients, but different for improving occupation changes, switching signs on many of the covariates.

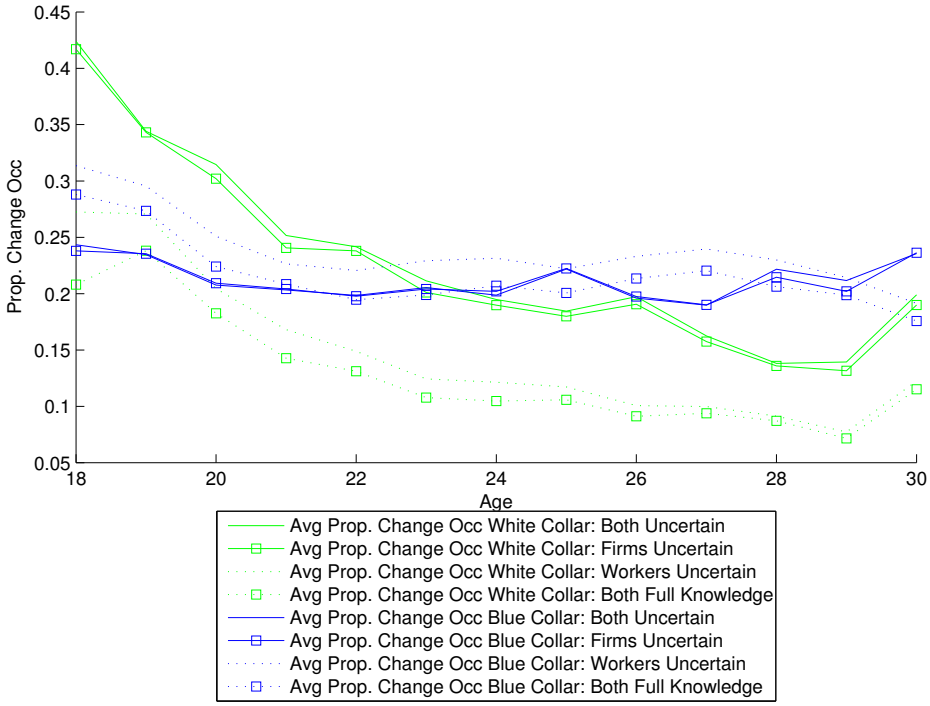
1.5.5 Counterfactual Test: Information on Type: Type is Known

Figure A.9 shows the average absolute deviations in beliefs about ability and actual ability. There is learning for both white collar and blue collar individual effects. Workers tend to have less uncertainty about their white collar capabilities. 15 years of learning reduces the deviations by about half.

This section tests what would happen if workers did not have to learn their occupation productivity, but already knew, or if the firm knew the workers' productivities, and workers did not, or if both knew. The results are given in a series of figures that compare the simulations for if they update and if they do not need to. Figure 1.5 shows the proportion switching occupations. Whether or not agents update does not have much of an effect, surprisingly. On the other hand, if the firms know the workers' type (and offer them wages

accordingly), this has a large effect on the probability that the workers switch occupations, overall decreasing the amount of switching, but all through reduction in those switching away from white collar jobs. These workers switch less than half as often as if the firms are uncertain about their productivity.

Figure 1.5: Proportion Switching Occupations: Simulations Testing Information



The remaining figures show more of what happens with less uncertainty. Again, the case of the employers having full knowledge about the worker's ability is the test that changes the most factors. The wage is determined by the firm's beliefs, and not the workers' beliefs. When the firms know the type of the workers, there are substantially more white collar workers and less blue collar workers, and more people go to school for longer. Firms knowing the type offer high wages to all productive white collar workers, and the workers do not need to learn their type or deal with the uncertainty associated, and choose to work in white collar jobs more.

1.6 Conclusion

This chapter presents a dynamic stochastic discrete choice model of occupational choice in order to examine the role of non-wage benefits and learning in occupational choice and transitions. The model is estimated using data from NLSY. The chapter reveals some interesting results that coincide with results of previous non-dynamic estimates, but offers further insight. Certain tests are only possible by using a structural model, such as the willingness of agents to pay to avoid the uncertainty with occupational choice, lifetime welfare, Bayesian learning, and the use of wage and non-wage offers instead of lagged proxies both for measuring improvements in these factors and how they induce occupation and job changes. Workers make switches mostly due to changes in their expectations concerning what wage they would earn in each occupation, but non-wage benefits are also important, at about one third the impact on transition probabilities. The structural model is able to show the effect of different factors in the decision to switch in ways that previous literature has not been able to. I find that workers have a change of 3.4-6.0 percentage points in their likelihood of changing occupations from a one percentage point change in the wage offer for white collar, and even higher for blue collar workers. The importance of expectations for the future become less important as agents get older, as the agents have less years to enjoy those benefits. The importance of non-wage benefits has ambiguous effects over time.

Agents are on average better off for both occupation and job transitions. However, the probability of being better off from an occupation change increases with accuracy of beliefs about their ability and with occupation and educational tenure, conditional on age. These results, along with higher utility for when workers know their type, suggest that workers are better off for any additional information they have about their productivity. Given workers will pay almost a dollar on average of their hourly wage to know which occupation they should switch into, understanding what affects the probability of an improvement transition is helpful.

Information is important, non-wage benefits are part of the decision process but wages

are more important, and policy makers would need to be careful with any policy changes, as there could be numerous unintended consequences.

A Appendix

A.1 Data Selection Procedure

Various methods have been used to code when an occupation has been changed and what their occupation is. The most common and natural definition is when there is a change in the reported occupation (for example, as used in Kambourov and Manovskii (2008) and Parrado et al. 2007). Mellow and Snider (1983) argue that there is a great deal of misclassification in occupation. They match self-reported occupation in the CPS with employer records and find only 58 percent match rates at the three-digit level and 81 percent at the one-digit level. Mathiowetz (1992) does a similar matching and finds a higher match rate at 87% at the three-digit level. Longhi and Brynin (2010) argue that this is unreliable, overestimating transitions when interviewees change their report when they actually haven't changed their occupation. They only record an occupation change when both a new occupation code is recorded and a change of job is also recorded. They find that this significantly decreases the measured amount of occupational transitions, and argue that this is the measure that should be used. However, this underreports changes, because some occupation changes are clearly within jobs. I allow for occupation changes within firms, such as through promotion. Sullivan (2009) uses the job that is recorded in the most number of weeks as their occupation. I use the definition of what the agent reports as their primary occupation each year, where possible, to link the job to the wage and non-wage benefits reported annually. The misclassification issues in this chapter are less severe because occupation is restricted to blue collar and white collar jobs.

The first restriction imposed is that the interviewee is male and observed from at least age 18 in the data. I drop those that report they are self-employed. The most consequential decision I had to impose on the data interpretation process is the selection of which of their various reported jobs each period was their primary job, and how long they had been there. Some years and for some variables they reported their current job, and this was used as the

most reliable classification. However, some years and some variables did not report current occupation; instead, the agents reported on up to five jobs they had held this year. The following rules, in order of the priority of the rule (rules with better, i.e. lower, priority numbers overruled potentially conflicting assignments from worse priority numbers) were used to select which was their primary current job:

1. The first in the priority list ($1 > 2 > 3 > 4 > 5$) that they say they are still in or do not record a stop date
2. If there is none from step 1, set equal to occupation with most recent stop date
3. If there is no stop date for any, choose highest priority occupation listed
4. If no occupations are listed that year, but difference in tenure across the two periods is greater than 52 (stayed at same job) and occupation is the same tenure before and after is the same, set occupation in the intermediate (missing year) equal to that value

Using this assignment procedure, I then assign which wage, tenure, and non-wage benefits belong to their current job.

Next, I need to determine when the interviewee changed jobs. The following rules were used for assignment (again, in order of priority), conditional on their working that period

1. The job chosen has a recorded stop date in that year
2. If the job chosen next period is not the earliest start date next period...
3. The next year has an occupation with the same occupation number and larger tenure but is not job X
4. The next year's tenure is less than tenure in the current year + 25
5. They record they are no longer there
6. Change to not changed job if past year and future year suggest the same job throughout

Finally, I need to assign whether they left their job voluntarily. Some of the time, the agents reported on the reason they left their job. Conditional on them changing jobs, the assignment rules are

1. If still working at the end of the year, but stop next year, among the jobs with tenures greater than this year's tenure and not still or unreported reason, choose the one with the earliest quit date and use that reason

2. Use the job chosen that has a recorded stop date in that year, and that period's quit reason

With all of these methods, I performed numerous inspections of the raw data to see if the assignments from the rules reflected what I would intuitively assumed happened on a case by case level, and found it to be reliable.

A.2 Tables

Table A.1: NLSY Sample Regressions on Next Period Log Wage to Compare 2 Occupations vs. 10

	2 occupations	10 occupations
ln(wage)	0.532***	0.531***
	(0.0159)	(0.0159)
Change Occupation	0.0223	0.0331***
	(0.0162)	(0.0123)
Change Job	-0.0670***	-0.0752***
	(0.0128)	(0.0132)
Period	0.0372***	0.0391***
	(0.00171)	(0.00182)
Change Job × Period	-0.0120***	-0.0107***
	(0.00167)	(0.00171)
Change Occupation × Period	-0.00230	-0.00575***
	(0.00214)	(0.00157)
Left Voluntarily	0.102***	0.101***
	(0.0103)	(0.0131)
Left Voluntarily × Change Occupation	-0.0338**	-0.00980
	(0.0170)	(0.0138)

Period is the Number of Years Older than 17
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.2: Average Job Characteristics, White Collar

Variable	ln(wage)	Health Ins.	Retire.	Pleasant Env.	Job Secure
Stayed in White Collar	7.13926	0.84011	0.70363	3.36347	3.24771
	(.00732)	(.00474)	(.00744)	(.02252)	(0.02508)
Switched to Blue Collar	6.84943	0.70666	0.54150	3.18322	3.05100
	(0.01252)	(0.01141)	(0.01839)	(0.03521)	(0.04483)

Standard Errors of the Mean in Parentheses

Table A.3: Regressions on Next Period Non-Wage Benefits

	Health Ins. 2 Occs	Health Ins. 10 Occs	Retire. 2 Occs	Retire. 10 Occs	Pleasant Env. 2 Occs	Pleasant Env. 10 Occs	Job Secure 2 Occs	Job Secure 10 Occs
Change Occupation	0.00147	0.0483***	0.0617	0.0692	-0.0545	0.106	0.0449	0.202
	(0.0234)	(0.0168)	(0.0684)	(0.0437)	(0.135)	(0.110)	(0.163)	(0.127)
Change Job	-0.108***	-0.121***	-0.0389	-0.0533	0.0121	-0.0354	0.0450	-0.0277
	(0.0208)	(0.0211)	(0.0520)	(0.0529)	(0.118)	(0.119)	(0.131)	(0.135)
Period	0.00595***	0.00774***	0.00717***	0.00883***	0.0135	0.0348	0.0263	0.0412
	(0.000992)	(0.00118)	(0.00219)	(0.00253)	(0.0356)	(0.0408)	(0.0397)	(0.0453)
Change Job × Period	-0.00755***	-0.00623***	-0.00948*	-0.00814	0.0414	0.0638	-0.135**	-0.116*
	(0.00216)	(0.00219)	(0.00493)	(0.00502)	(0.0521)	(0.0530)	(0.0618)	(0.0636)
Change Occupation × Period	0.00237	-0.00396**	-0.00438	-0.00625	-0.000677	-0.0554	0.0285	-0.0357
	(0.00262)	(0.00184)	(0.00658)	(0.00415)	(0.0653)	(0.0526)	(0.0767)	(0.0620)
Left Voluntarily	0.0888***	0.0933***	0.0881***	0.0932***	0.0273	0.181*	0.257***	0.420***
	(0.0137)	(0.0161)	(0.0190)	(0.0227)	(0.0742)	(0.0928)	(0.0871)	(0.113)
Left Voluntarily × Change Occupation	-0.0172	-0.0120	-0.0182	-0.0137	0.0236	-0.209**	-0.0830	-0.256**
	(0.0204)	(0.0158)	(0.0325)	(0.0241)	(0.120)	(0.0959)	(0.130)	(0.112)
Health Ins$_t$	0.453***	0.453***						
	(0.00786)	(0.00786)						
Retirement$_t$			0.612***	0.612***				
			(0.00825)	(0.00825)				
Pleasant Env.$_t$					0.350***	0.352***		
					(0.0285)	(0.0280)		
Job Secure$_t$							0.283***	0.282***
							(0.0307)	(0.0306)

Period is the Number of Years Older than 17
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.4: NLSY Subsample Transition Matrix Between Schooling and 2 Occupations

	1	2	3
1	.36828645	.33802217	.29369139
2	.02376949	.7660734	.21015711
3	.01821937	.19531895	.78646168

Table A.5: NLSY Subsample Transition Matrix Between Schooling and 10 Occupations

	1	2	3	4	5	6	7	8	9	10	11
1	.36828645	.13768116	.03878943	.03537937	.06564365	.06052856	.12958227	.04347826	.00682012	.03154305	.08226769
2	.02047361	.72890972	.01171682	.00653675	.01874692	.01825358	.05870745	.0426739	.0077701	.02294031	.06327084
3	.00832342	.04280618	.51605232	.03844629	.13238208	.07808165	.0554895	.05112961	.02259215	.02378121	.03091558
4	.03846154	.05769231	.11538462	.52125506	.07186235	.05465587	.0708502	.01821862	.00506073	.01923077	.02732794
5	.03652058	.05278884	.14143426	.02589641	.44156707	.07569721	.06640106	.04880478	.0126162	.03021248	.06806109
6	.02864782	.0565317	.09587471	.02559206	.10809778	.3961039	.06417112	.04163484	.01527884	.04430863	.12375859
7	.03138239	.09132277	.03122548	.01600502	.03765887	.03483446	.49850934	.0707673	.01474973	.0484858	.12505884
8	.00966996	.07714946	.03804919	.00336346	.02837923	.01997057	.0750473	.50704225	.03153248	.07567795	.13411814
9	.00365297	.05936073	.05114155	.00273973	.02648402	.03287671	.07488584	.13972603	.29771689	.18812785	.12328767
10	.00750976	.06878943	.02102734	.00600781	.03364374	.03724842	.07659958	.09912887	.06728747	.43256233	.15019525
11	.01956382	.09797947	.01796023	.00561257	.03848621	.0601347	.11289288	.11770366	.02116742	.0837075	.42479153

1: In School 2: Managerial and Specialty Occupations 3: Specialty Occupations 4: Technical Support and Sales Occupations 5: Administrative Support Occupations 6: Service Occupations and Farming 7: Production, Craft and Repair Occupations 8: Extraction, Precision Production, and Plant and System Operators 9: Operators, Laborers and Fabricators 10: Transportation and Material Moving, Handlers, Equipment Cleaners and Helpers 11: Military

Table A.6: Fringe Regression Results (Logit: 1-4; OLS: 5-8)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AFQT	0.000590	-0.000455	-0.00219	-0.000463	-0.00326**	-0.00245***	0.00475***	-0.000789
	(0.00286)	(0.00160)	(0.00654)	(0.00503)	(0.00157)	(0.000928)	(0.00182)	(0.00103)
Years of Education	-0.213***	-0.309***	0.173**	0.0465	0.126**	0.145**	0.00814	-0.0153
	(0.0584)	(0.0588)	(0.0874)	(0.110)	(0.0571)	(0.0660)	(0.0661)	(0.0730)
Period	0.462***	0.260***	0.252*	0.294***	-0.0934**	-0.00445	0.0208	-0.0153
	(0.0456)	(0.0234)	(0.137)	(0.105)	(0.0406)	(0.0247)	(0.0470)	(0.0273)
Period^2	-0.0247***	-0.0135***	-0.00954	-0.0110**	0.00804**	-0.000661	0.000889	-0.00128
	(0.00267)	(0.00141)	(0.00636)	(0.00495)	(0.00352)	(0.00220)	(0.00407)	(0.00244)
Period × AFQT	0.000941***	0.000733***	0.000634	0.000563	0.000415	0.000126	-0.000346	0.000497**
	(0.000337)	(0.000197)	(0.000605)	(0.000465)	(0.000253)	(0.000178)	(0.000292)	(0.000197)
Period × Years of Education	0.0226***	0.0357***	-0.00837	0.00661	-0.0145**	-0.0172*	-0.000925	0.00629
	(0.00604)	(0.00651)	(0.00794)	(0.0102)	(0.00720)	(0.00875)	(0.00832)	(0.00969)

1: Health Insurance, White Collar; 2: Health Insurance, Blue Collar
3: Retirement, White Collar; 4: Retirement, Blue Collar
5: Pleasant Environment, White Collar; 6: Pleasant Environment, Blue Collar
7: Job Security, White Collar; 8: Job Security, Blue Collar
Period is the Number of Years Older than 17

Table A.7: Probability Fired Coefficient Results

	White Collar	Blue Collar
AFQT	-0.00169	-0.0107***
	(0.00342)	(0.00183)
Years of Education	-0.0132	0.00414
	(0.0833)	(0.0781)
Years of Job Seniority	-0.336***	-0.135***
	(0.0846)	(0.0492)
Period	-0.114*	-0.0718**
	(0.0620)	(0.0306)
Period × AFQT	-0.000514	0.000541*
	(0.000466)	(0.000278)
Period × Years of Education	-5.53e-05	-0.0152
	(0.00925)	(0.00980)
Period × Years of Job Seniority	0.0242***	0.00862
	(0.00848)	(0.00526)
Period^2	0.00858**	0.00479**
	(0.00415)	(0.00209)

Period is the Number of Years Older than 17

Table A.8: Auxiliary Regression Results: OLS Regression of Probability Changed Job

ln(wage)	-0.0657***
	(0.0199)
Years of Educ	0.00344
	(0.00693)
Health Ins.	-0.135***
	(0.0283)
Retirement	-0.0693***
	(0.0246)
Pleasant Env.	0.0104
	(0.0130)
Job Secure	-0.0821***
	(0.0126)
AFQT	-0.000342
	(0.000424)
Period	-0.0274***
	(0.00764)

Period is the Number of Years Older than 17

Table A.9: Model Parameters

θ_k: non-wage benefits intercept for schooling, white collar, and blue collar
β^{ED}: non-wage marginal benefit for higher ability (AFQT score)
β_j^{FV}: non-wage return for fringe benefit health insurance ($j = 1$), retirement ($j = 2$), pleasant environment ($j = 3$), and job security ($j = 4$)
Σ^{ε}: variance-covariance matrix for the shock to wages for white collar and blue collar
Σ^{η}: variance-covariance matrix for the unobserved ability that agents and firms update beliefs on (this also effects the speed of learning)
β_{jk}^{EXP}: log wage returns to experience for schooling ($j = 1$), white collar experience ($j = 2$), and blue collar work ($j = 3$), having different returns in white collar jobs ($k = 1$) and blue collar jobs ($k = 2$)
M_k: job entry cost for occupation k
θ_k^W: log wage intercept for white collar ($k = 1$) and blue collar ($k = 2$)
δ_T: terminal value function scaling parameter
γ_k: log wage return to ability (AFQT) score for white collar ($k = 1$) and blue collar ($k = 2$)
σ^{ξ}: variance of random utility shocks
β^{TEN}: log wage return to job tenure
σ^{λ}: firm-employee match parameter variance
q_j: job satisfaction report thresholds
ρ: CES utility parameter

Table A.10: Parameter Values

θ_1	θ_2	θ_3	β^{ED}	β_1^{FV}	β_2^{FV}	β_3^{FV}
87.62175	270.95601	357.36965	-3.10820	14.86137	15.01811	9.89115
β_4^{FV}	$\Sigma_{11}^{\varepsilon}$	$\Sigma_{22}^{\varepsilon}$	$\Sigma_{12}^{\varepsilon}$	Σ_{11}^{η}	Σ_{22}^{η}	Σ_{12}^{η}
10.00293	0.04559	0.07226	-0.01410	0.03130	0.11755	0.00589
β_{11}^{EXP}	β_{21}^{EXP}	β_{31}^{EXP}	β_{12}^{EXP}	β_{22}^{EXP}	β_{32}^{EXP}	M_1
0.11732	0.11286	0.08143	0.07022	0.00025	0.05689	15.37853
M_2	θ_1^W	θ_2^W	δ_T	γ_1	γ_2	σ^{ξ}
-42.32591	5.07054	5.92431	8.10474	0.01825	0.00992	16.80672
β^{TEN}	σ^{λ}	q_1	q_2	q_3		
0.03125	0.04789	438.15201	740.43073	2243.70070		

Table A.11: $\partial Pr(ChangeOcc)/\partial\Theta$

θ_1	θ_2	θ_3	β^{ED}	β_1^{FV}	β_2^{FV}	β_3^{FV}
0.00002	-0.00189	0.00164	0.00065	0.00264	-0.00058	-0.00243
β_4^{FV}	$\Sigma_{11}^{\varepsilon}$	$\Sigma_{22}^{\varepsilon}$	$\Sigma_{12}^{\varepsilon}$	Σ_{11}^{η}	Σ_{22}^{η}	Σ_{12}^{η}
-0.00144	0.19766	0.23079	1.00408	-1.96154	-0.10814	1.36659
β_{11}^{EXP}	β_{21}^{EXP}	β_{31}^{EXP}	β_{12}^{EXP}	β_{22}^{EXP}	β_{32}^{EXP}	M_1
-0.21345	-13.09097	0.24027	0.76136	-3.35018	14.90318	-0.00037
M_2	θ_1^W	θ_2^W	δ_T	γ_1	γ_2	σ^{ξ}
0.00021	-0.59318	0.81308	-0.00619	-53.91620	52.37938	-0.00021
β^{TEN}	σ^{λ}	q_1	q_2	q_3		
0.05607	-1.14305	0.00000	0.00000	0.00000		

Table A.12: $\partial Pr(ChangeJob)/\partial\Theta$

θ_1	θ_2	θ_3	β^{ED}	β_1^{FV}	β_2^{FV}	β_3^{FV}
-0.00001	0.00059	-0.00052	0.00070	-0.00117	-0.00018	-0.00022
β_4^{FV}	$\Sigma_{11}^{\varepsilon}$	$\Sigma_{22}^{\varepsilon}$	$\Sigma_{12}^{\varepsilon}$	Σ_{11}^{η}	Σ_{22}^{η}	Σ_{12}^{η}
-0.00002	-0.13299	-0.28996	-0.19434	0.16505	0.07467	-0.79205
β_{11}^{EXP}	β_{21}^{EXP}	β_{31}^{EXP}	β_{12}^{EXP}	β_{22}^{EXP}	β_{32}^{EXP}	M_1
0.11554	6.74482	-1.65788	0.11452	0.53241	-4.08157	-0.00015
M_2	θ_1^W	θ_2^W	δ_T	γ_1	γ_2	σ^{ξ}
-0.00037	0.40231	-0.36844	0.00819	20.55622	-21.46550	-0.00022
β^{TEN}	σ^{λ}	q_1	q_2	q_3		
2.94679	0.31494	0.00000	0.00000	0.00000		

Table A.13: $\partial Pr(LeaveVolunt.)/\partial\Theta$

θ_1	θ_2	θ_3	β^{ED}	β_1^{FV}	β_2^{FV}	β_3^{FV}
-0.00020	0.00007	0.00015	-0.00273	0.00217	-0.00061	-0.00205
β_4^{FV}	$\Sigma_{11}^{\varepsilon}$	$\Sigma_{22}^{\varepsilon}$	$\Sigma_{12}^{\varepsilon}$	Σ_{11}^{η}	Σ_{22}^{η}	Σ_{12}^{η}
-0.00198	-0.27198	-0.26966	0.17893	0.45996	0.16133	-0.26433
β_{11}^{EXP}	β_{21}^{EXP}	β_{31}^{EXP}	β_{12}^{EXP}	β_{22}^{EXP}	β_{32}^{EXP}	M_1
-0.59384	-0.31435	0.79196	-1.15436	0.94397	1.59774	-0.00023
M_2	θ_1^W	θ_2^W	δ_T	γ_1	γ_2	σ^{ξ}
-0.00008	-0.06945	-0.07737	-0.00680	-0.01539	1.82162	-0.00003
β^{TEN}	σ^{λ}	q_1	q_2	q_3		
3.10272	-0.04343	0.00000	0.00000	0.00000		

Table A.14: Simulated Data Trends Comparing Accepted Offers in the Other Occupation and Current Occupation: Has No Current Employer

	(1)	(2)	(3)	(4)	(5)
Wage: White Collar	0.81224	0.94524	0.84609	509.50277	417.99300
Non-Wage Benefits: White Collar	0.42181	0.73195	0.36767	199.57978	-132.22954
Wage: Blue Collar	0.55496	0.37321	0.34211	-7.16490	-9.14055
Non-Wage Benefits: Blue Collar	0.45047	0.74012	0.70635	14.06388	11.69322

Table A.15: Simulated Data Trends Comparing Accepted Offers in the Other Occupation and Current Occupation: Has Current Employer

	(1)	(2)	(3)	(4)	(5)
Wage: White Collar	0.40767	0.95473	0.76575	535.25240	243.69172
Non-Wage Benefits: White Collar	0.33008	0.65808	0.25638	152.39689	-308.17574
Wage: Blue Collar	0.47969	0.83441	0.69150	435.49440	162.69235
Non-Wage Benefits: Blue Collar	0.35856	0.68641	0.39978	188.15580	-178.80537

Table A.16: Simulated Data Trends Comparing Accepted Offers in the Other Job and Current Job

	(1)	(2)	(3)	(4)	(5)
Wage: White Collar	0.35702	1.00000	1.00000	6917.89655	6980.04158
Non-Wage Benefits: White Collar	0.32531	1.00000	0.99885	3567.26469	3094.60481
Wage: Blue Collar	0.56997	1.00000	1.00000	69.23417	58.98680
Non-Wage Benefits: Blue Collar	0.57478	0.99994	0.99994	59.59530	49.62775

Tables A.14-A.16

1. Correlation Coefficient
2. Proportion Greater than Zero, Difference Offer in Other Occupation Minus Offer in Current Occupation
3. Proportion Greater Than Zero, Difference Offer in Other Occupation Minus Previous Period Wage/Non-wage
4. Mean Difference, Accepted Offer in Other Occupation Minus Offer in Current Occupation
5. Mean Difference, Accepted Offer in Other Occupation Minus Previous Period's Wage/Non-Wage

Table A.17: Probit Estimated Marginal Effects on Changing Occupations: Has Current Employer

Origin Occupation	White Collar	Blue Collar
Current Occupation, Current Job: Wage	-0.0430***	-0.213***
	(0.00177)	(0.00313)
Current Occupation, Current Job: ξ	-1.95e-06	-1.11e-05
	(2.86e-06)	(9.88e-06)
Current Occupation, Current Job: EV	-1.356***	-4.348***
	(0.0743)	(0.162)
Current Occupation, New Job: Wage	-0.0346***	-0.116***
	(0.00161)	(0.00341)
Current Occupation, New Job: Non-Wage Benefits	-0.00649***	-0.0275***
	(0.00240)	(0.00512)
Current Occupation, New Job: ξ	2.02e-07	3.00e-06
	(1.31e-06)	(4.60e-06)
Current Occupation, New Job: EV	-1.783***	-4.831***
	(0.0962)	(0.213)
Other Occupation, Current Job: Wage	0.0550***	0.120***
	(0.00188)	(0.00269)
Other Occupation, Current Job: ξ	-1.44e-06	5.80e-06
	(4.29e-06)	(1.02e-05)
Other Occupation, Current Job: EV	1.348***	4.241***
	(0.0772)	(0.151)
Other Occupation, New Job: Wage	0.0315***	0.109***
	(0.00121)	(0.00298)
Other Occupation, New Job: Non-Wage Benefits	0.00353**	0.0308***
	(0.00180)	(0.00616)
Other Occupation, New Job: ξ	-5.10e-07	-2.93e-06
	(1.27e-06)	(3.22e-06)
Other Occupation, New Job: EV	1.755***	4.931***
	(0.0963)	(0.209)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.18: Probit Estimated Marginal Effects Changing Job

Origin Occupation	White Collar	Blue Collar
Current Occupation, Current Job: Wage	-0.246***	-0.365***
	(0.00454)	(0.00415)
Current Occupation, Current Job: ξ	4.45e-05	-1.78e-07
	(3.19e-05)	(9.60e-06)
Current Occupation, Current Job: EV	-4.923***	4.341***
	(0.214)	(0.272)
Current Occupation, New Job: Wage	0.248***	0.282***
	(0.00441)	(0.00440)
Current Occupation, New Job: Non-Wage Benefits	0.0653***	0.0869***
	(0.00888)	(0.00756)
Current Occupation, New Job: ξ	2.25e-06	1.35e-06
	(3.77e-06)	(5.50e-06)
Current Occupation, New Job: EV	3.345***	-2.377***
	(0.243)	(0.292)
Other Occupation, Current Job: Wage	-0.0817***	-0.0832***
	(0.00279)	(0.00419)
Other Occupation, Current Job: ξ	1.34e-05	-5.22e-06
	(1.46e-05)	(1.48e-05)
Other Occupation, Current Job: EV	3.719***	-5.288***
	(0.223)	(0.258)
Other Occupation, New Job: Wage	0.0879***	0.0924***
	(0.00248)	(0.00450)
Other Occupation, New Job: Non-Wage Benefits	0.0190***	0.0594***
	(0.00657)	(0.00935)
Other Occupation, New Job: ξ	3.24e-06	6.92e-06
	(4.90e-06)	(1.77e-05)
Other Occupation, New Job: EV	-2.243***	3.260***
	(0.248)	(0.287)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.19: Probit Estimated Marginal Effects: Ex-Post Better Off for Job Change

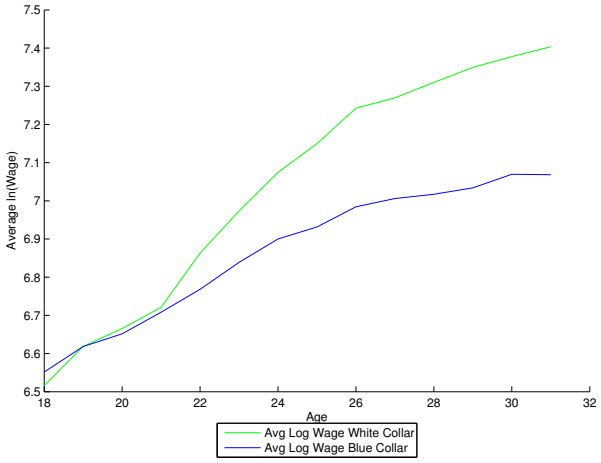
	White Collar	Blue Collar	White Collar	Blue Collar
Years Education	0.0428***	0.0353***	0.0262***	-0.000559
	(0.00734)	(0.00394)	(0.00750)	(0.00411)
Years Experience WC	-0.00957***	-0.0485***	-0.00544***	-0.0485***
	(0.00130)	(0.00246)	(0.00132)	(0.00244)
Years Experience BC	-0.00111	0.00840***	0.000838	0.00454***
	(0.00201)	(0.00112)	(0.00204)	(0.00113)
Years Job Tenure	0.0112***	0.0108***	0.0128***	0.00941***
	(0.00356)	(0.00328)	(0.00359)	(0.00329)
AFQT	-0.00109***	-0.00392***	-0.000914***	-0.00504***
	(0.000186)	(0.000142)	(0.000189)	(0.000147)
$\lvert\eta^*_{WC} - \widehat{\eta}_{WC}\rvert$	-0.0381	0.0553	-0.0350	0.0681**
	(0.0512)	(0.0343)	(0.0518)	(0.0344)
$\lvert\eta^*_{BC} - \widehat{\eta}_{BC}\rvert$	0.0308	-0.0220	0.0179	-0.0138
	(0.0283)	(0.0238)	(0.0286)	(0.0237)
$\lvert\eta^*_{WC} - \widehat{\eta}^F_{WC}\rvert$	-0.0519	0.0272	-0.0389	0.0266
	(0.0391)	(0.0273)	(0.0397)	(0.0273)
$\lvert\eta^*_{BC} - \widehat{\eta}^F_{BC}\rvert$	-0.0286	-0.0479***	-0.0311	-0.0454***
	(0.0210)	(0.0142)	(0.0213)	(0.0142)
Difference in Wage Offers			0.165***	0.0479***
			(0.00638)	(0.00147)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

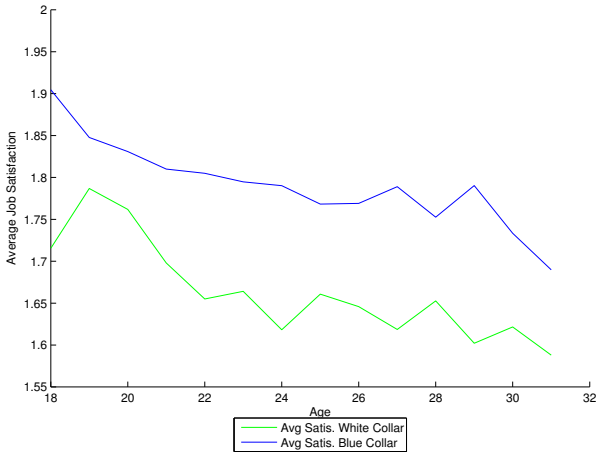
A.3 Figures

Figure A.1: NLSY Data Trends

(a) Average Log Wage by Period and Occupation



(b) Average Job Satisfaction by Age and Occupation



(c) Proportion Changing Job, by Age and Origin Occupation

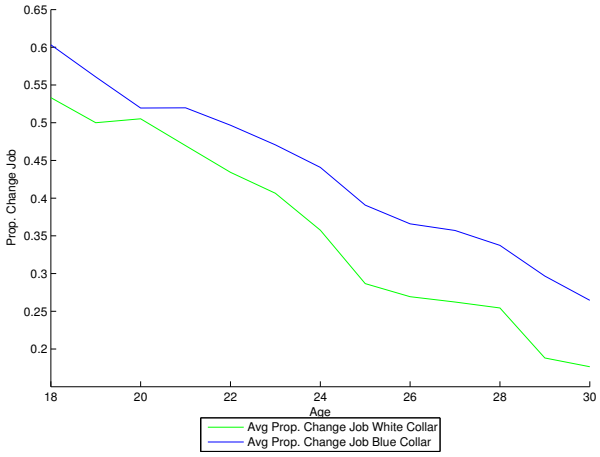


Figure A.2: NLSY Sample Non-Wage Benefit Averages

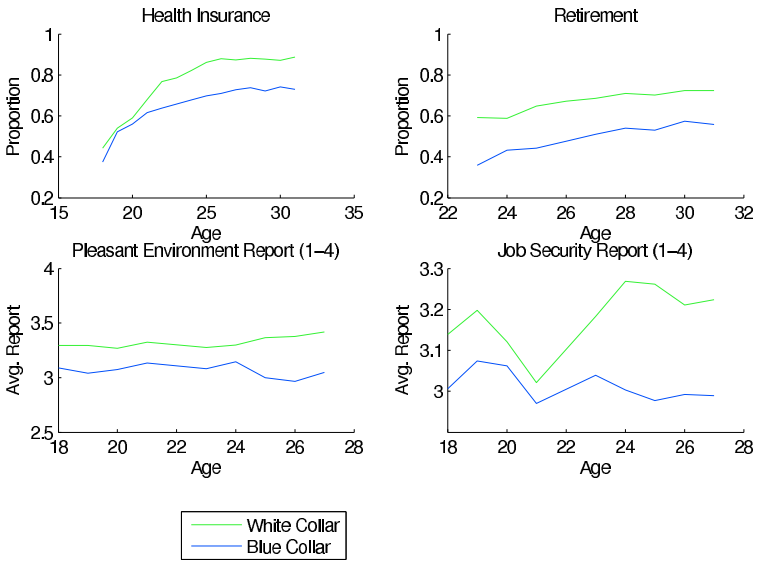
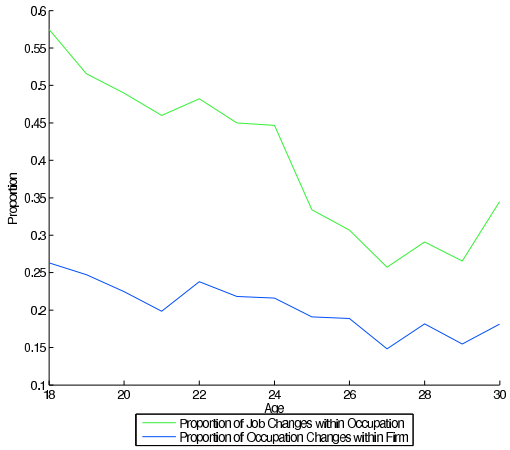


Figure A.3: NLSY Sample Transition Densities

(a) Conditional Proportions of Job and Occupation Transitions



(b) Density Estimates of Total Number of Job Changes and Occupational Changes in 14 Years

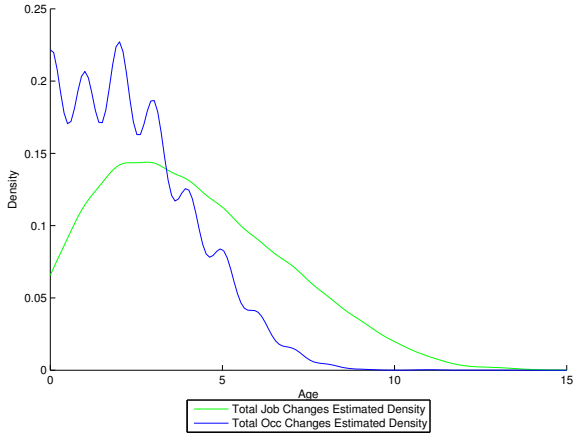
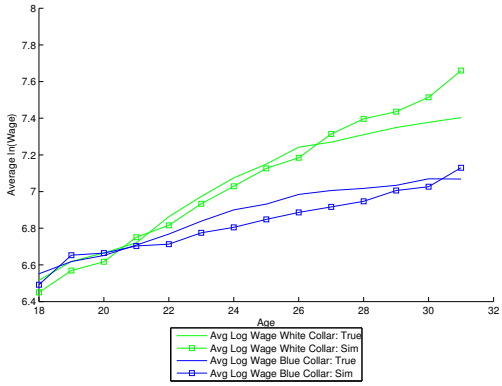
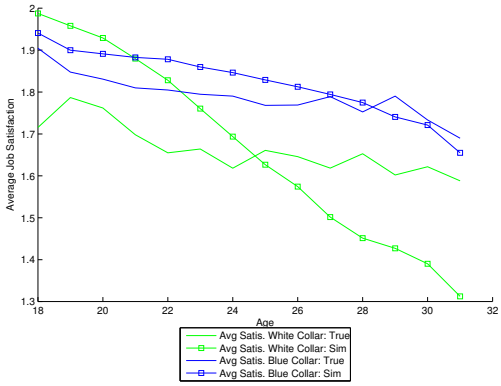


Figure A.4: Simulated vs. True Trends

(a) Average Wages by Occupation



(b) Average Job Satisfaction by Occupation



(c) Proportion Changing Job, by Age and Occupation

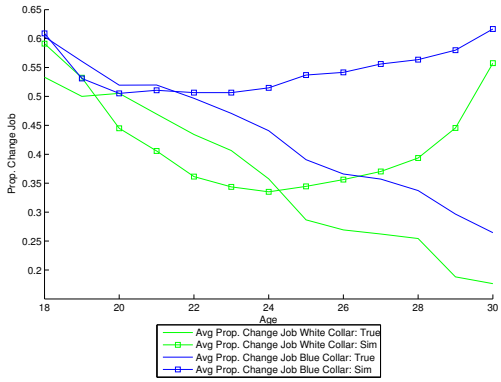
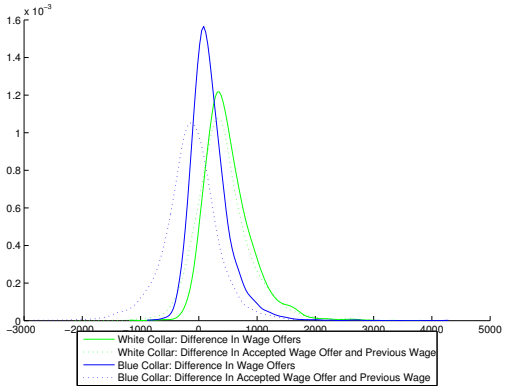


Figure A.5: Comparing Densities of Difference Between Accepted Wage Offer in the Other Occupation and Wage in Current Job/Occupation

(a) Occupation Change



(b) Job Change

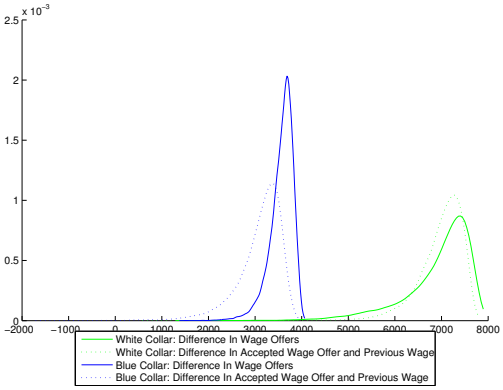
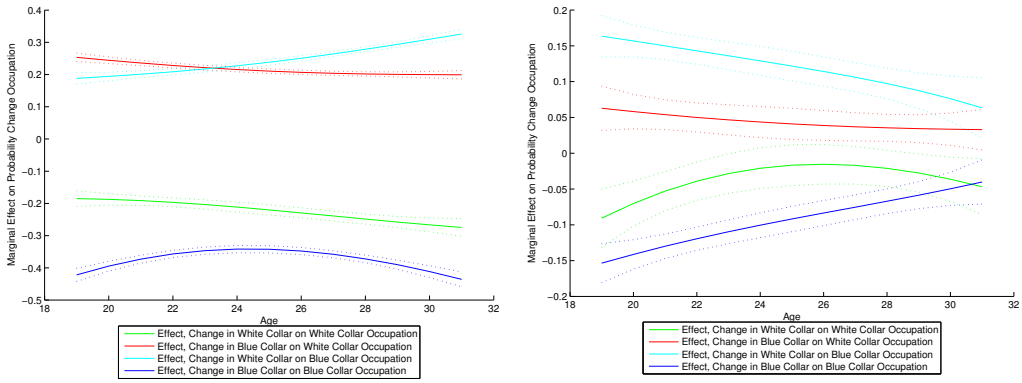


Figure A.6: Estimated Marginal Effects of Percentage Change in Offers on Δ Pr(Change Occ) Across Different Ages

(a) Wage

(b) Non-Wage Benefits



(c) Expected Future Utility

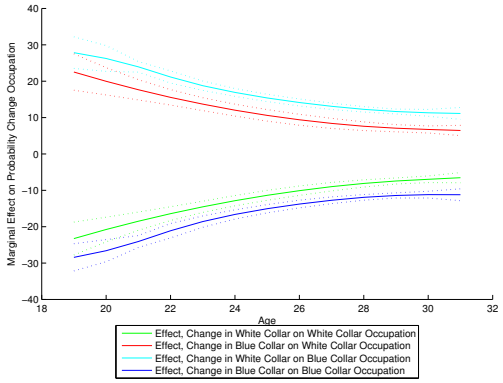


Figure A.7: Kernel Density Estimate of Switchers Lifetime Utility Minus Counterfactual Non-Switch Utility

(a) Occupation Changers

(b) Job Changers

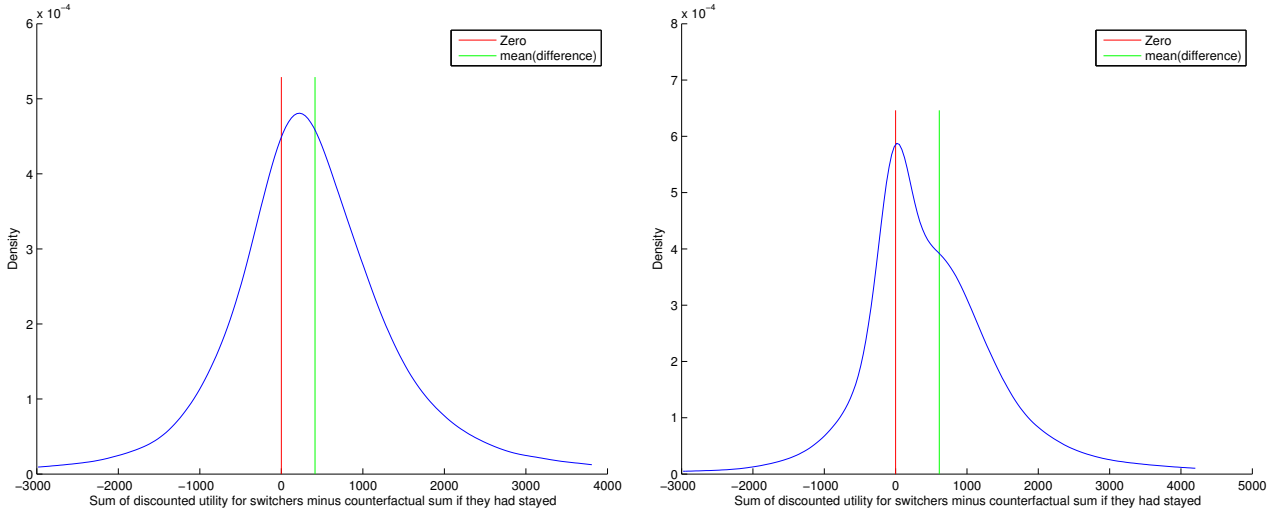
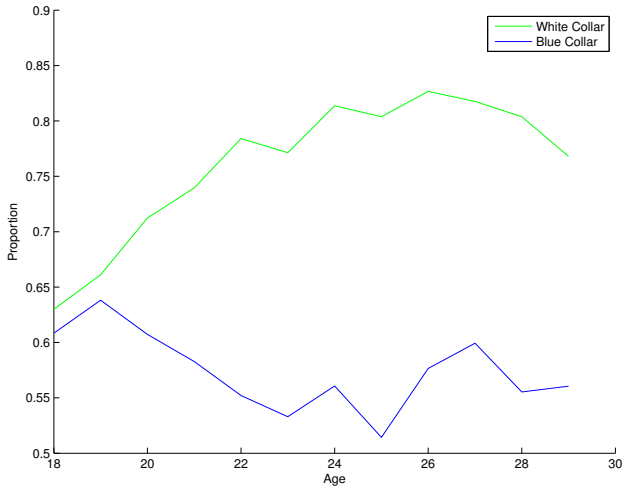


Figure A.8: Simulated Proportions Better Off by Age

(a) Occupation Changers



(b) Job Changers

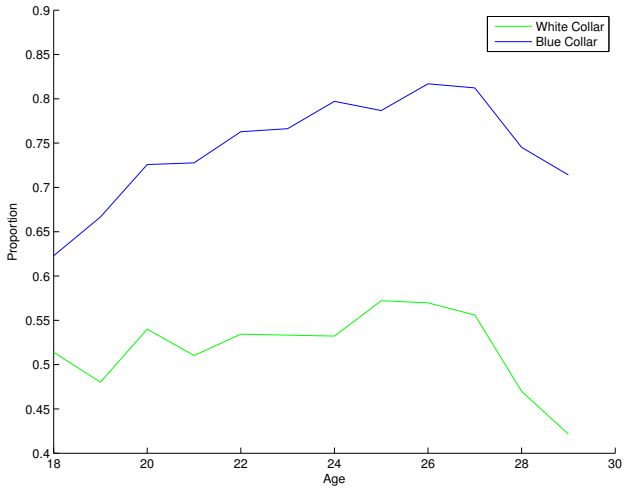


Figure A.9: Learning Curves by Occupation Experience: $|\widehat{\eta} - \eta|$

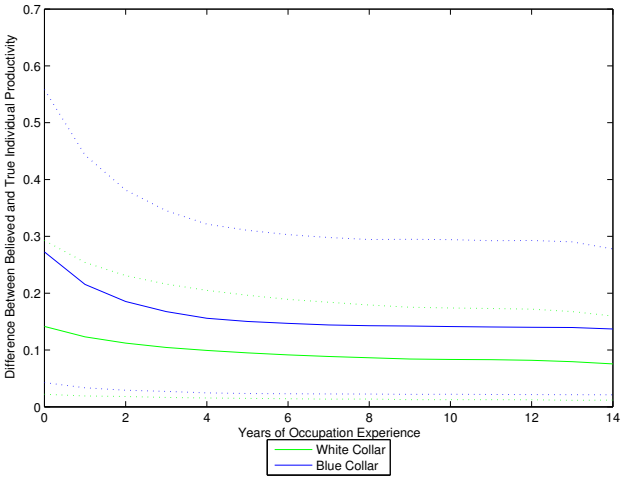
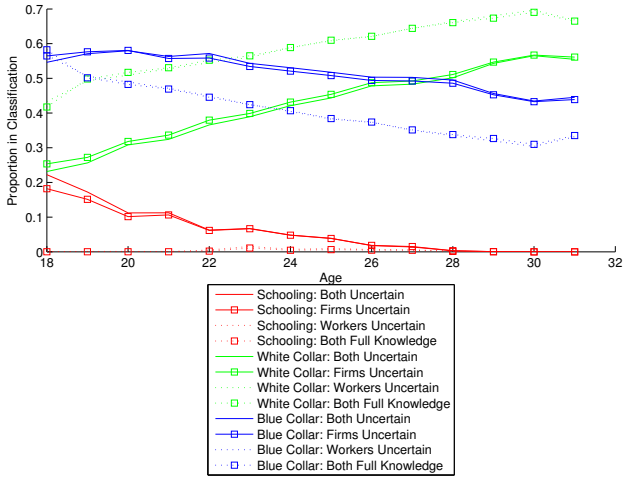
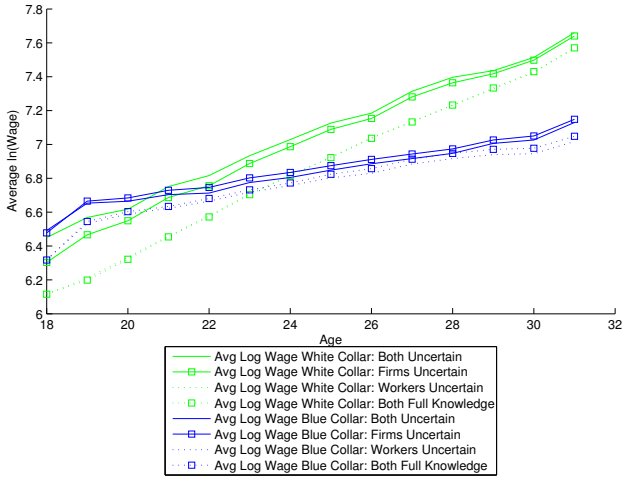


Figure A.10: Trends for Simulations Testing Information

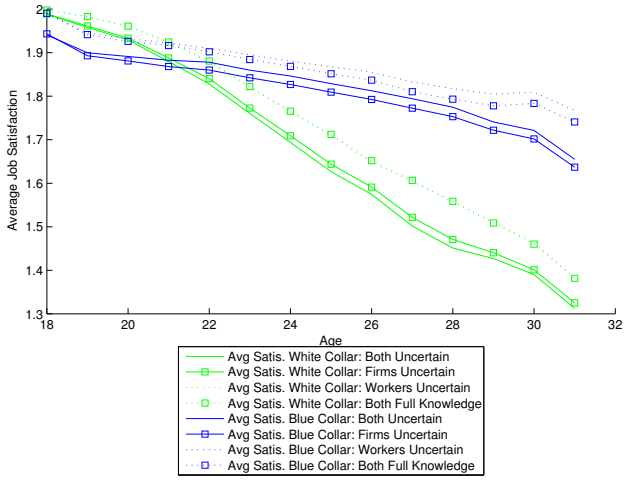
(a) Average Proportions in Occupations by Period



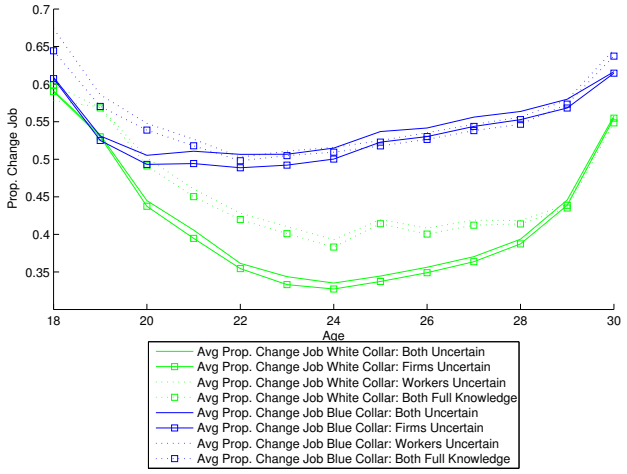
(b) Average Wage by Period and Changing Occupation



(c) Average Job Satisfaction by Age and Occupational



(d) Proportion Changing Job, by Age and Occupation



A.4 References

Ackerberg, D.A. (2003). "Advertising, Learning, and Consumer Choice in Experience Good Markets: A Structural Empirical Examination." *International Economic Review*, 44 1007-1040.

Aguirregabiria, V. and Mira, P. (2010) "Dynamic discrete choice structural models: A survey." *Journal of Econometrics*, 156, 38-67.

Akerlof, G., Rose, A., and Yelloe, J. (1988) "Job Switching and Job Satisfaction in the U.S. Labor Market," *Brookings Papers on Economic Activity*, 1988:2.

Albrecht, J., W. and Vroman, S. B., (2002). "A Matching Model with Endogenous Skill Requirements." *International Economic Review* 43, 283-305.

Amuedo-Dorantes, Catalina and Ricardo Serrano-Padial (2010). "Labor market flexibility and poverty dynamics." *Labour Economics*,17,632-642.

Ansley, C.F. and R. Kohn (1983). "Exact likelihood of vector autoregressive-moving average process with missing or aggregated data." *Biometrika*,70(1),275-278.

Blumen, Isadore, Marvin Kogan, and Philip J. McCarthy (1955). *The Industrial Mobility of Labor as a Probability Process.* Ithaca, NY: Cornell University Press.

Brand, J. (2006). "The effects of job displacement on job quality: Findings from the Wisconsin Longitudinal Study." *Research in Social Stratification and Mobility*, 24.

Buchinsky, M., Fougère, D., Kramarz, F. and Tchernis, R. (2010) "Interfirm Mobility, Wages and the Returns to Seniority and Experience in the United States." *Review of Economic Studies*, 77, 972-1001.

Burdett, Kenneth (1978), "A Theory of Employee Job Search and Quit Rates," *American Economic Review* 68, 212220.

Card, D., and D. Hyslop (1997). "Does Inflation 'Grease the Wheels of the Labor Market'?" in *Reducing Inflation: Motivation and Strategy*, C. Romer and D. Romer, eds. Chicago, IL: University of Chicago Press), pp. 71?114.

Clay, K., R. Goettler and E. Wolff (2004). "Consumer Learning about Experience Goods: Evidence from and Online Grocer." Working Paper, Carnegie Mellon University.

Del Bono, E. and D. Vuri (2011). "Job Mobility and the gender wage gap in Italy." *Labour Economics*,18,130-142.

Delfgaauw, J. (2007). "The effect of job satisfaction on job search: Not just whether, but also where." *Labour Economics* 14:299-317.

Dix-Carneiro, Rafael (2010) "Trade Liberalization and Labor Market Dynamics." CEPS Working Paper No. 212.

Dolado, J.J., Jansen, M., Jimeno J. F., (2009). "On-the-Job Search in a Matching Model with Heterogeneous Jobs and Workers." *The Economic Journal*, 119:534, 200-228.

Eckstein, Z., D. Horsky, and Y. Raban (1988). "An Empirical Dynamic Model of Optimal Brand Choice." Working Paper, Tel-Aviv University.

Farber (1999). "Mobility and stability: the Dynamics of job change in labor markets." Handbook of Labor Economics chapter...

Felli, L. and C. Harris (1996). "Learning, Wage Dynamics, and Firm-Specific Human Capital." *Journal of Political Economy*,104(4),838-868.

Freeman, Richard B. (1978), "Job satisfaction as an economic variable," *American Economic Review*, 68, 135-141.

Gautier, P., (2002). "Unemployment and search externalities in a model with heterogeneous jobs and workers." *Economica* 69, 21-40.

Gibbons, R. and M. Waldman (1999). "A Theory of Wage and Promotion Dynamics Inside Firms." *Quarterly Journal of Economics*,114, 1321-58.

Gouriéroux, C. and A. Monfort (1996). *Simulation-Based Econometric Methods*. Oxford University Press, New York.

Jovanovic, Boyan (1979), "Firm-Specific Capital and Turnover," *Journal of Political Economy,* 87, 12461259.

Kambourov, G., and Manovskii, I., (2008). "Rising occupational and industry mobility in the United States: 196897." *International Economic Review* 49 (1), 4179.

Keane, M. P. and Wolpin, K. I. (1994). "The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpoliation: Monte Carlo Evidence." *The Review of Economics and Statistics*, 76(4),648-672.

Keane, M. P. and Wolpin, K. I. (1997). "Career decisions of young men," *Journal of Political Economy* vol. 105, (June), pp. 473522.

Keane, Michael and Kenneth Wolpin (2009). "Empirical applications of discrete choice dynamic programming models," *Review of Economic Dynamics* 12(2009) 1-22.

Kunze, A. and K. Troske (2011). "Life Cycle Patters in Male/Female Differences in Job Search." *Labour Economics*,doi:10.1016/j.labeco.2011.09.009.

Lee, D. (2005). "An estimable dynamic general equilibrium model of work, schooling, and occupational choice." *International Economic Review* 46 (1), 134.

Lee, D., Wolpin, K., (2006). "Intersectoral labor mobility and the growth of the service sector." *Econometrica* 74, 1-46.

Léné, Alexandre (2011). "Occupational downgrading and bumping down: The combined effects of education and experience." *Labour Economics* 18(2), 257-269.

Light, A., and McGarry, K., (1998). "Job change patterns and the wages of young men." *Review of Economics and Statistics* 80, 276286.

Longhi, S. and Brynin, M. "Occupational change in Britain and Germany." *Labour Economics* 17 (2010) 655666.

Mankiw, G., J. Rotemberg, and L. Summers (1985). "Intertemporal Substitution in Macroeconomics." *Quarterly Journal of Economics*,100(1),225-251.

Markey, J. and Parks, W. (1989). "Occupational change: pursuing a different kind of work." *Monthly Labor Review*, 112(7), 312.

Mathiowetz, Nancy. (1992). "Errors in Reports of Occupation." *The Public Opinion Quarterly* 56(3):352-55.

McLaughlin, K. (1994). "Rigid Wages." *Journal of Monetary Economics*,34,383?414.

Meinecke, J. (2010). "Learning about job matches in a structural dynamic model."

Mellow, Wesley, and Hal Sider. (1983). "Accuracy of Response in Labor Market Surveys: Evidence and Implications." *Journal of Labor Economics* 1(4):331-44.

Moscarini, G., and F. Vella (2003): "Aggregate Worker Reallocation and Occupational Mobility in the United States: 1976-2000," mimeo, Yale University.

Moscarini, G., and F. Vella (2003b): "Occupational Mobility and Employment Reallocation: Evidence from the NLSY," mimeo, Yale University.

Neal, D. (1995). "Industry-Specific Human Capital: Evidence from Displaced Workers." *Journal of Labor Economics*,13(4),653-677.

Parrado, E., Caner, A., Wolff, E.N., (2007). "Occupational and industrial mobility in the United States." *Labour Economics* 14, 435455.

Pissarides, C. (1990). *Equilibrium Unemployment Theory.* Oxford, Basil Blackwell.

Postel-Vinay, F. and J. Robin (2002). "Equilibrium wage dispersion with worker and employer heterogeneity." *Econometrica*,70(6),2295-2350.

Roy, A. D. (1951). "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers*,3,135-146.

Rust, John (1997). "Using Randomization to Break the Curse of Dimensionality." *Econometrica*,65(3),487 516.

Sullivan, P. (2009). "Estimation of an Occupational Choice Model when Occupations are Misclassified." *Journal of Human Resources*, 44:2, 495-535.

Sullivan, P. (2010). "Empirical evidence on occupation and industry specific human capital." *Labour Economics*,17,567-580.

Topel,R. (1991). "Specific Capital, Mobility, and Wages: Wages Rise with Job Seniority." *Journal of Political Economy*,99(1),145-176.

U.S. Bureau of the Census (1993). *1990 Census of Population and Housing: Public Use Microdata Samples: Technical Documentation*. Washington, DC: U.S. Government Printing Office, pp. B-28-30.

van Ours, J. C., Ridder, G., (1995). "Job matching and job competition: Are lower educated workers at the back of job queues?" *European Economic Review* 39, 1717-1731.

Wilson, R.M., Green, C., (1990). "Occupation, occupational change and movement within the income distribution." *Eastern Economic Journal* 16, 209 220.

Zangelidis, A. (2008) "Occupational and Industry Specificity of Human Capital in the British Labour Market." *Scottish Journal of Political Economy* 55(4): 420-443.

Chapter 2

Cross Validation Bandwidth Selection for Derivatives of Various Dimensional Densities

2.1 Introduction

There are many cases when a researcher is interested in information about the shape of a distribution of variables. Methods are well developed for nonparametric estimation of the density created by the underlying data generating process, including methods for selecting the bandwidth parameters for kernel density estimators.[1] There is also substantial progress in developing techniques for estimating derivatives of distributions. More complicated methods have been developed for choosing bandwidths for multivariate densities (Zhang et al., 2005) and for derivatives of univariate densities (Wu 1997, Bearse and Rilstone 2008), but less work has been done on densities of multivariate densities. In this chapter, I develop cross validation criteria and investigate how well product kernels perform in estimating high dimension derivatives of densities, both for joint and conditional distributions.

[1] Li and Racine (2007) provide a good review.

I test various orders of Gaussian kernels with an extension of the univariate cross validation criteria as well as with two other cross validation criteria: a weighted integrated square error criterion and, for the derivative of a conditional density, a criterion that jointly estimates the bandwidths for the numerator and denominator densities. The direct cross validation is more robust against poor bandwidth selection than the weighted integrated square error criterion with regards both to a mean square error (MSE) criterion and a maximum deviation criterion. On the other hand, the criterion that jointly estimates the bandwidths for the derivative of a multidimensional conditional density does substantially better than the estimators that separately estimate the bandwidths. Higher order kernels tend to outperform lower order kernels, and its improvements increases both with sample size and with density dimension.

Throughout the chapter, I give my attention to product kernels for their ease of use. Cacoullos (1966) first presented product kernels as an option in estimating multivariate densities. This disallows potentially more accurate kernels; however, the product kernel is the easiest to use and to derive cross validation criteria for, and is frequently used in practice. This chapter also focuses on kernels from the Gaussian family. Turlach (1993) and Hansen (2005) show that the choice of kernel families has a much smaller impact on mean integrated square error (MISE) than the kernel order. It is not difficult to take the generalized criteria presented here and adapt them for use on a different kernel family. The orders of the kernel (defined as the first non-zero moment of the kernel) used are 2,4,6,8,10, and the infinite order (the Dirichlet kernel).

This chapter investigates only cross validation methods instead of plug-in methods, as plug-in methods become increasingly difficult to formulate with higher dimension kernels, necessitating solving systems of equations. Also, Loader (1999) challenges the previous work that suggested the superiority of plug-in methods over cross validation, and demonstrates that this is not true in many cases (e.g., when there is misspecification of the pilot bandwidths).

Hansen (2005) and Turlach (1993) both find that kernel order and bandwidth choice (respectively) are more important than choice of kernel family. For that reason, similar to the work of Hansen and Wand and Schucany (1990), I restrict attention to different orders of the Gaussian kernel. Marron (1994) shows that higher order kernels perform well when the curvature of what is being estimated is roughly constant, and poorly when there are abrupt changes in curvature on neighborhoods about the size of the bandwidth. Wand and Schucany (1990) examine Gaussian kernels from orders 2 to 10; they compare the efficiencies of these kernels theoretically to the optimal kernels, and show that for low order kernels, they are very close, and for ν derivative low (i.e., the zeroth derivative). The worst case they present for the first derivative has a relative efficiency of 0.76. Marron and Wand (1992) also show that the bandwidth that minimizes MISE is close to that which minimizes AMISE (the plug-in estimator) only for sample sizes close to 1 million, discouraging use of plug-in methods.

Wand and Schucany (1990) show that the $2r^{th}$ degree Gaussian kernel can be represented by

$$G_{2r}(x) = \frac{(-1)^r \phi^{(2r-1)}(x)}{2^{r-1}(r-1)!x}$$

I use this to derive the cross validation criteria for the different order kernels. The infinite order (Dirichlet) kernel, as Hansen (2005) presents it, is $K(x) = \frac{\sin x}{\pi x}$).

The cross validation methods are generalizations of the cross validation criteria set forth by Hardle, Marron and Wand (1990) for the univariate density derivative. They demonstrate that, for the univariate case and the first derivative, there is not much loss in efficiency (in the sense that Silverman uses the term) from using the Gaussian kernel instead of the optimal kernel. Various research connect to derivations of consistency and optimal convergence. Marron and Hardle (1986) generalizes the procedures for a variety of nonparametric estimators,

including density estimators. Li and Racine (2007) present numerous derivations.

Chacon, Duoung, and Wand (2011) investigate derivatives of multidimensional densities, just as in this chapter, but focus their attention just on second order kernels. They derive the MISE and show convergence rates. They also allow for a bandwidth matrix, and show that this general bandwidth matrix generally achieves better simulation results than that of using a diagonal matrix (as is done in product kernels, and so in this chapter). Hall, Racine and Li (2004) consider an estimation of a conditional density using bandwidths derived from weighted integrated square error cross validation. This chapter develops a similar estimator for the derivative of a conditional density.

Section 2.2 presents the direct cross validation criteria as well as a weighted cross validation criteria for the derivative of a multidimensional density and presents simulation results. Section 2.3 examines the derivative of multivariate conditional densities, showing the methodology for estimating the bandwidths for the marginal and joint densities separately and for estimating them jointly through a single cross validation criterion. Simulation results follow. Section 2.4 concludes. Proof and criteria derivations, along with results tables, are in the Appendix in Section B.

2.2 Bandwidth Selection for the Derivative of a Joint Density

This section develops criteria for cross validation estimators for bandwidths of high dimensional derivatives of densities, and investigates how effective these methods are in simulations. The direct criteria for the various orders as well as a weighted integrated square error for a second order Gaussian kernel are examined. The methods are compared across various sample sizes.

Assume that the density is of dimension q. Given the estimator

$$\widehat{f}(x) = \frac{1}{n\prod_{x\in G_x} h_s} \sum_{i=1}^{n} \prod_{s\in G_x} K\left(\frac{x_s - x_{is}}{h_s}\right)$$

Then, the estimator for the derivative of the density is

$$\frac{\partial \widehat{f}(x)}{\partial x_k} = \frac{1}{nh_k\prod_{x\in G_x} h_s} \sum_{i=1}^{n} K'\left(\frac{x_k - x_{ik}}{h_k}\right) \prod_{s\in G_x\setminus k} K\left(\frac{x_s - x_{is}}{h_s}\right)$$

I first demonstrate conditions for consistent estimation of the derivative of the density, and then derive two cross validation criteria connected to the integrated square error, and show that these criteria provide bandwidth estimators that satisfy the consistency criteria. I then perform a Monte Carlo study to compare the performance of different criteria and kernel orders across different data sizes and dimensions.

2.2.1 Consistency

Theorem 1. *Assume that the r^{th} order kernel has the following characteristics:*

1. $\int\cdots\int \prod_{s\in G_x} K(x_s)dx_1\cdots dx_q = 1$
2. $K(x_s) = K(-x_s)$
3. $\int\cdots\int x_k^{r-1} \prod_{s\in G_x} K(x_s)dx_1\cdots dx_q = 0$
4. $\int\cdots\int x_k^{r} \prod_{s\in G_x} K(x_s)dx_1\cdots dx_q > 0$
5. $\int\cdots\int x_k^{r} \prod_{s\in G_x} K(x_s)dx_1\cdots dx_q < \infty$
6. $f(x)$ *is* $(r+1)$ *times differentiable*

where $d_x = \prod_{s\in G_x} dx_s$. *Then,*

$$MSE\left(\frac{\partial \widehat{f}(x)}{\partial x_k}\right) = \left(\frac{\int K(x)\, x^r dx}{r!} \sum_{t=1}^{q} h_t^r \frac{\partial^{r+1} f(x)}{\partial x_k \partial^r x_t} + O\left(\sum_{t=1}^{q} h_t^{r+1}\right)\right)^2$$
$$+ \frac{f(x)}{n h_k^2 \prod_{s \in G_x} h_s} \int K'(x)^2 dx \left(\int K(x)^2 dx\right)^{q-1} + O\left(\frac{1}{n h_k \prod_{s \in G_x} h_s}\right)$$
$$= O\left(\left(\sum_{t=1}^{q} h_t^{r+1}\right)^2 + \frac{1}{n h_k \prod_{s \in G_x} h_s}\right)$$

and, if as $n \to \infty$, $\max_j\{h_j\} \to 0$ *and* $n h_k \prod_{s \in G_x} h_s \to \infty$, *then the last statement directly implies that* $\frac{\partial \widehat{f}(x)}{\partial x_k} \to \frac{\partial f(x)}{\partial x_k}$ *in mean square error (MSE), implying also convergence in probability, and consistency.*

The proof of this theorem is contained in Section B.1.

2.2.2 Cross Validation Criteria: Integrated Square Error

For the direct cross validation method, the criterion minimizes the integrated difference between the true and the estimated densities. This criterion is the same for any dimension of the density. Let $f(x)$ be the true density, and $\widehat{f}(x)$ be the product kernel estimator of $f(x)$. Then the cross validation criterion (Integrated Square Error, or ISE) is given by

$$ISE(h) = \int \cdots \int \left(\frac{\partial \widehat{f}(x)}{\partial x_k} - \frac{\partial f(x)}{\partial x_k}\right)^2 dx$$

Section B.1 provides the derivation for the cross validation: for given dimension, I express a criterion which yields equivalent minimizing arguments and is not a function of any unknowns. This is given by

$$ISE^{*}(h) = \frac{1}{n^2 h_k^2 \prod_{s\in G_x} h_s} \sum_{i=1}^{n}\sum_{j=1}^{n} \left[\int K'\left(\tilde{x}_{ijs} + \overline{x}_{is}\right) K'\left(\overline{x}_{is}\right) d\overline{x}_{is} \right] \times \cdots$$
$$\prod_{s\in G_x \setminus k} \left(\int K\left(\tilde{x}_{ijs} + \overline{x}_{is}\right) K\left(\overline{x}_{is}\right) d\overline{x}_{is} \right) + \frac{2\sum_{i=1}^{n}\sum_{j\neq i} K''\left(\tilde{x}_{ijk}\right) \prod_{s\in G_x\setminus k} K\left(\tilde{x}_{ijs}\right)}{n(n-1)\left(\prod_{s\in G_x} h_s\right)^2 h_k^2}$$

where $x = (x_1, ..., x_K)' \in G_x$ is the point of evaluation, and $x_i = (x_{i1}, ..., x_{iK})'$ is one observation in the data. h_s is the bandwidth for the s^{th} random variable, and $K(\cdot)$ is the kernel chosen for the estimation procedure.

2.2.3 Cross Validation Criteria: Weighted Integrate Square Error

I propose an alternative criterion which weights the difference between the true and the estimated derivatives of the densities. The weight is provided by an estimate of the density. The criterion is

$$WISE(h) = \int \cdots \int \left(\frac{\partial \widehat{f}(x;h)}{\partial x_k} - \frac{\partial f(x)}{\partial x_k} \right)^2 \widehat{f}(x;b) dx$$

where b (and equivalently $\widehat{f}(x;b)$) is estimated prior to searching for the optimal bandwidth h. The reason for this, and why they are not jointly searched for, is that a joint search would require including $\int \cdots \int \widehat{f}(x;b)^2 f(x)^2 dx$. As $f(x)$ is unknown, and the sample analogue cannot simply be used as is done in the other cases, this evaluation becomes more difficult to evaluate. For this reason, b is estimated before using cross validation methods.

Similar to the direct integrated square error cross validation derivation, I derive an expression that has the same minimizing argument and is a function only of observed data.

This is done in Section B.1, and comes out to be

$$WISE^*(h) = \frac{1}{h_k^2 n^3 \prod_{s\in G_x} h_s^2 b_s} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{m=1}^{n} \int K'\left(\frac{x_k - x_{ik}}{h_k}\right) K'\left(\frac{x_k - x_{jk}}{h_k}\right) K\left(\frac{x_k - x_{mk}}{b_k}\right) dx_k \cdots$$
$$\times \prod_{s\in G_x \backslash k} \int K\left(\frac{x_s - x_{is}}{h_s}\right) K\left(\frac{x_s - x_{js}}{h_s}\right) K\left(\frac{x_s - x_{ms}}{b_s}\right) dx_s$$
$$+ 2\frac{1}{n(n-1)^2 h_k^2 \prod_{s\in G_x} h_s^2} \sum_{m=1}^{n} \left[\sum_{i\neq m} K''\left(\frac{x_{mk} - x_{ik}}{h_k}\right) \prod_{s\in G_x} K\left(\frac{x_{ms} - x_{is}}{h_s}\right) \sum_{j\neq m} \prod_{s\in G_x} K\left(\frac{x_{ms} - x_{js}}{b_s}\right) \right.$$
$$\left. + \sum_{i\neq m} K'\left(\frac{x_{mk} - x_{ik}}{h_k}\right) \prod_{s\in G_x} K\left(\frac{x_{ms} - x_{is}}{h_s}\right) \sum_{j\neq m} K'\left(\frac{x_{mk} - x_{jk}}{b_k}\right) \prod_{s\in G_x} K\left(\frac{x_{ms} - x_{js}}{b_s}\right) \right]$$

As employed by Hardle, Marron and Wand (1990) for the derivative of a univariate density, the theorems of Marron and Hardle (1986) can be applied here for the derivative of a multivariate density to show that the criteria of ISE and WISE both converge to the optimal criteria of MISE, and that the bandwidths resulting are not only valid for consistent estimation, but converge to the optimal MISE bandwidths.

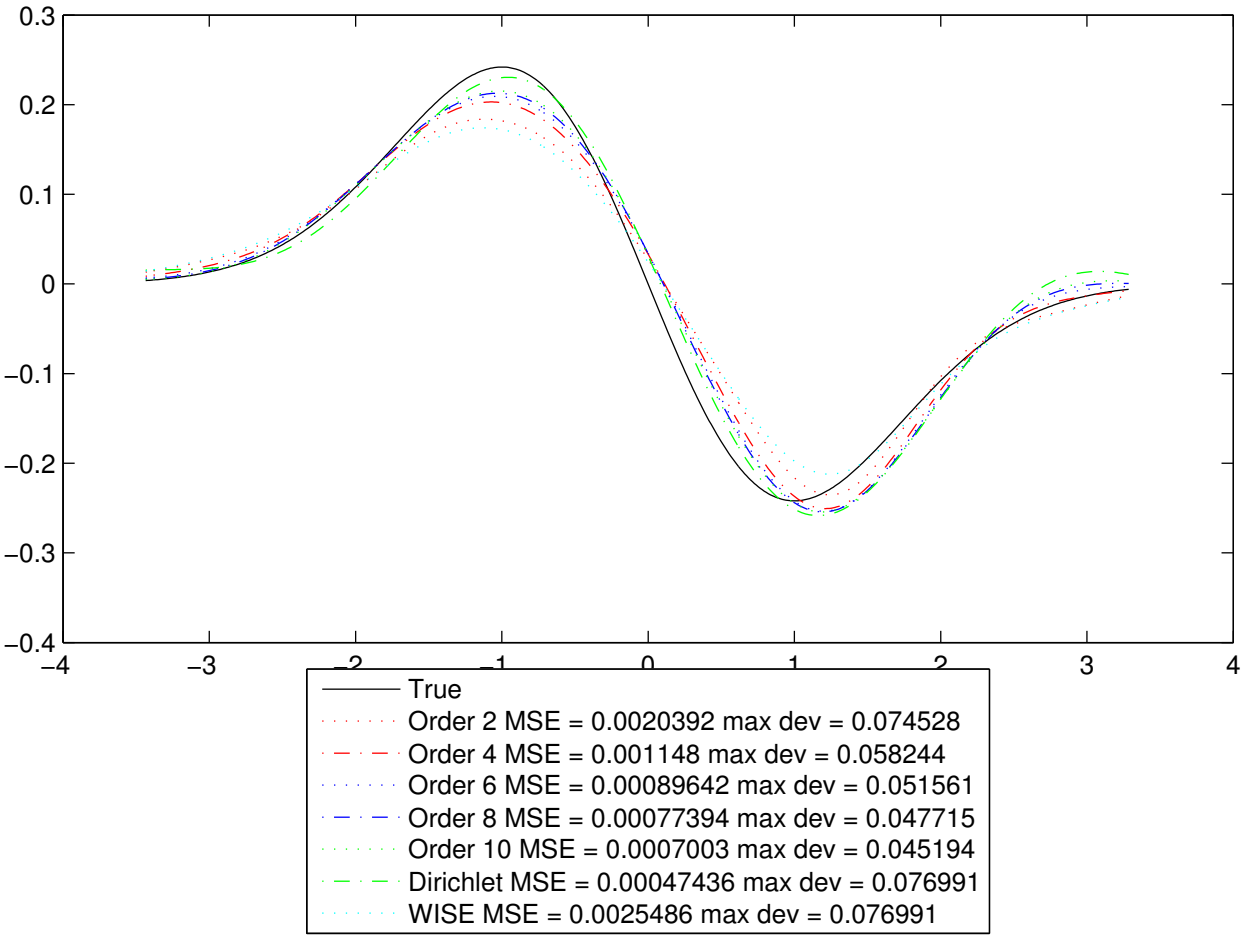
2.2.4 Simulation Results

I run a Monte Carlo with 100 simulations on the data, with various sample sizes (100,500,1000,5000) and various density dimensions (1,2,4,8). Data is simulated from a multivariate normal distribution with zero mean and variance-covariance matrix with variances given by $\sigma_i^2 = 1$ and covariance elements given by $\sigma_{ij} = 0.4$. The simulations use Gaussian kernels of order 2, 4, 6, 8, and 10, as well as the infinite dimension Dirichlet kernel. The weighted integrated square error is also tested, which uses a second order kernel. Two statistics of the simulations are examined: the mean square error ($\sum_i (\partial \widehat{f}(x_i)/\partial x_k - \partial f(x_i)/\partial x_k)^2$) and the maximum absolute deviation ($\max_i \left|\partial \widehat{f}(x_i)/\partial x_k - \partial f(x_i)/\partial x_k\right|$).

Figure 2.1, an estimated univariate derivative of a density for 500 observations, demonstrates different MSEs means. For this simulation, the Dirichlet kernel performed the best, both in terms of MSE and an eyeball test, while the weighted integrated square error cross

validation performs the worst.

Figure 2.1: Derivative of Univariate Normal Density and Estimations, $N = 500$



Tables 2.1 and 2.2 show the Monte Carlo results for $k = 2$, a bivariate distribution. Section B.2 contains tables of the results for the other values of k.

Higher order kernels outperform lower order kernels virtually universally. The Dirichlet kernel is overall the best performing, although often the higher order kernels are almost as good. Most of the gains from using higher order kernels are exhausted after a few increases in order–often, just increasing the kernel order to 4 improves the mean square error on average and the maximum deviations, but increasing the order beyond that does little more.

Table 2.1: Monte Carlo Simulation Results: Mean Square Error Comparisons for Derivative of 2-Dimensional Density

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100	N=5000 SIMS=58
2nd Order	0.00849	0.0064	0.00633	0.000602
	(0.0282)	(0.0276)	(0.0498)	(0.00148)
4th Order	0.00493	0.00154	0.000543	0.000298
	(0.0161)	(0.00329)	(0.000917)	(0.000955)
6th Order	0.00302	0.00152	0.000498	0.000149
	(0.00668)	(0.00377)	(0.000975)	(0.000187)
8th Order	0.00306	0.00159	0.000499	0.000147
	(0.0074)	(0.00424)	(0.00113)	(0.000199)
10th Order	0.00259	0.00067	0.000387	0.000104
	(0.0061)	(0.001)	(0.000745)	(8.47e-05)
Dirichlet	0.000835	0.000447	0.00149	6.95e-05
	(0.00173)	(0.00105)	(0.0113)	(0.000101)
WISE	1.12	0.014	0.00865	0.0014
	(11.1)	(0.0471)	(0.058)	(0.00423)

Standard Deviations of MSE Simulation Estimates in Parentheses

Table 2.2: Monte Carlo Simulation Results: Mean Maximum Difference Comparisons for Derivative of 2-Dimensional Density

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100	N=5000 SIMS=58
2nd Order	0.164	0.144	0.112	0.0658
	(0.238)	(0.234)	(0.297)	(0.0541)
4th Order	0.121	0.0849	0.0535	0.0406
	(0.161)	(0.0805)	(0.0399)	(0.0374)
6th Order	0.101	0.0804	0.0487	0.0314
	(0.0959)	(0.0816)	(0.0375)	(0.021)
8th Order	0.0972	0.0805	0.0471	0.03
	(0.0978)	(0.0856)	(0.0396)	(0.021)
10th Order	0.0901	0.0573	0.0408	0.0239
	(0.0916)	(0.0438)	(0.028)	(0.00973)
Dirichlet	0.0557	0.0413	0.0426	0.0177
	(0.0357)	(0.0281)	(0.0785)	(0.0108)
WISE	0.786	0.216	0.15	0.093
	(6.14)	(0.38)	(0.35)	(0.112)

Standard Deviations of Maximum Deviation Simulation Estimates in Parentheses

The second order term in the Taylor expansion, the bias of the 2nd order kernels, is more important to control for than higher order terms. Higher order kernels' improvement is most dramatic for low sample sizes and small dimensions, but the effect persists both with higher sample sizes and dimensions. Larger sample sizes yield more accurate results, and the weighted square error does poorly for small sample sizes, and only slightly worse than its 2nd order counterpart for large sample sizes. With the univariate case, it seems to do slightly better than its 2nd order counterpart, but higher dimension densities are poorly estimated using the WISE minimization criterion.

To offer comparison across dimension size, I look at how large the absolute maximum heights of the derivative of the densities are. While the higher dimensional densities have lower MSEs and maximum deviations, they are also coming from densities with much lower maximum and average absolute heights. The maximum and average heights of the densities are shown in Table B.13. The ratios of the best maximum deviation for the various sample sizes and densities (i.e., the maximum that is the smallest for all kernel orders) to both the average and the maximum absolute derivative of the density height are calculated (Tables B.14 and B.15). For example, for a univariate density derivative and $N = 100$, the best average maximum deviation comes from using the Dirichlet kernel, which has an average maximum deviation of 0.117 (Table B.4). Table B.13 shows that, for a univariate density of the type used in the Monte Carlo simulations, the maximum absolute height is 0.242, and the average absolute height is 0.1628.

This helps to frame how good of an estimate the .117 is–it is smaller than the average absolute height, but not by much. In fact, the ratio of the average maximum deviation in the simulations to the maximum absolute height is 0.117/0.2420=0.4835, meaning that the density estimator is off on average by almost half of the highest height of the true derivative of the density, or 0.718 of the average height. Clearly, this is not a satisfactory outcome.

Higher values of N quickly yield much better results. This accentuates how much worse some of the higher densities estimates are, even with the best estimators used here. Looking

at the ratio of the average maximum deviation to the average absolute height in Table B.15, for $N = 1000$, the univariate density yields the acceptable ratio of .3274 (compare this to the plot in Figure 2.1, where the best estimator, which looks very close, has a ratio of 0.473). For an 8-dimensional density derivative, this fraction jumps all the way up to 7.09, much too high for meaningful estimation of the underlying true derivative of the density. Theoretically, for some N high enough, the ratio would fall back into an acceptable range; however, the sample sizes that would be required for accurate estimation would be too large to be tractable.[2]

Overall, these results suggest that when trying to estimate a derivative of an unconditional multivariate density, higher order kernels and the straightforward cross validation criterion yield the best outcomes.

2.3 Bandwidth Selection for the Derivative of a Conditional Density

There are times when the researcher is interested in estimating the derivative of a conditional density. This chapter gives attention to the case when the derivative is with respect to the random variable, as opposed to the conditioning variable. It extends the work of Hall, Racine and Li (2004), who examine bandwidth choice for a conditional density, by looking instead at the derivative of a density, and by explicitly allowing for higher order kernels. They allow for a weighting function, which here is assumed to be one. Random vector y is conditioned on vector x; the conditional distribution is given by $f_{Y|X}(y|x)$, the numerator joint distribution is $f(y, x)$, and the multivariate marginal denominator distribution is $m(x)$. Then

$$\frac{\partial f_{Y|X}(y|x)}{\partial y_k} = \frac{\partial f(y,x)}{\partial y_k}(m(x))^{-1}$$

[2]The estimation of the bandwidth parameter using the methods in this chapter would take far too long, even with parallel processing, as was used in this project.

One natural estimator for this which has been suggested and used is

$$\frac{\partial \widehat{f}_{Y|X}(y|x)}{\partial y_k} = \frac{\partial \widehat{f}(y,x)}{\partial y_k}(\widehat{m}(x))^{-1}$$

The estimator requires bandwidths both for the marginal denominator density and for the derivative of the joint numerator density. Hall, Racine and Li (2004) restrict the bandwidths for variables x to be the same for the joint and for the marginal, an assumption not made in this chapter. I also test the initial candidate for bandwidth selection that use the bandwidths from the methods in the previous section, i.e. cross validation of ISE for $= \frac{\partial \widehat{f}(y,x)}{\partial y_k}$ and $\widehat{m}(x)$ separately, and investigate how well these perform. As before, I also establish conditions for consistent estimation of derivative of a conditional density.

2.3.1 Consistency

Theorem 2. *Assume that the r^{th} order kernel has the following characteristics:*

1. $\int \cdots \int \prod_{s \in G_x} K(x_s) dx_1 \cdots dx_q = 1$
2. $K(x_s) = K(-x_s)$
3. $\int \cdots \int x_k^{r-1} \prod_{s \in G_x} K(x_s) dx_1 \cdots dx_q = 0$
4. $\int \cdots \int x_k^{r} \prod_{s \in G_x} K(x_s) dx_1 \cdots dx_q > 0$
5. $\int \cdots \int x_k^{r} \prod_{s \in G_x} K(x_s) dx_1 \cdots dx_q < \infty$
6. $f(x)$ *is* $(r+1)$ *times differentiable*

where $d_x = \prod_{s \in G_x} dx_s$. *Then, if as* $n \to \infty$, $\max_j\{h_j\} \to 0$ *and* $nh_k \prod_{s \in G_x} h_s \to \infty$, *and* $n \prod_{s \in G_x} b_s \to \infty$, *then* $\frac{\partial \widehat{f}_{Y|X}(y|x)}{\partial y_k} \xrightarrow{p} \frac{\partial f_{Y|X}(y|x)}{\partial y_k}$.

The proof of this theorem is contained in Section B.1.

2.3.2 Joint Bandwidth Selection for the Derivative of Conditional Density

The criterion for the joint estimation of the bandwidths of a conditional density requires the joint integrated square error, given by

$$ISE(h,b,x) = \int \cdots \int \left(\frac{\partial \widehat{f}_{Y|X}(y|x)}{\partial y_k} - \frac{\partial f_{Y|X}(y|x)}{\partial y_k} \right)^2 dy$$

However, this chapter focuses on the case for bandwidths that would be good at any value of x, so I integrate over x as well, after multiplying through by $m(x)$:

$$ISE(h,b) = E_X(ISE(h,b,x)) = \int \cdots \int \left(\frac{\partial \widehat{f}_{Y|X}(y|x)}{\partial y_k} - \frac{\partial f_{Y|X}(y|x)}{\partial y_k} \right)^2 m(x) dy dx$$

Section B.1 shows that a cross validation criteria with the same minimizing argument can be given, which is equal to

$$\begin{aligned} &ISE(h,b) = \\ &\frac{\prod_{s\in G_X} b_s^2}{n h_k^2 \prod_{s\in G_y} h_s \prod_{s\in G_x} h_s^2} \sum_{i=1}^{n} \frac{\sum_{j\neq i} \sum_{\ell \neq i} \prod_{s\in G_x} R^h_{x,ij\ell s} R_{y,ij\ell k} \prod_{s\in G_y \setminus k} R_{y,ij\ell s}}{\left(\sum_{j=1,\neq i}^{n} \prod_{s\in G_x} K\left(\tilde{x}^b_{ijs}\right) \right)^2} \\ &+ 2 \frac{\prod_{s\in G_x} b_s}{n h_k^2 \prod_{s\in G_x \cup G_y} h_s} \sum_{i=1}^{n} \left[\frac{\sum_{j\neq i} K''\left(\tilde{y}_{ijk}\right) \left[\prod_{s\in G_y \setminus k} K(\tilde{y}_{ijk}) \right] \prod_{s\in G_x} K(\tilde{x}^h_{ijs})}{\sum_{j\neq i} \prod_{s\in G_x} K\left(\tilde{x}^b_{ijs}\right)} \right] \end{aligned}$$

where

$$\overline{y}_{is} = \frac{y_s - y_{is}}{h_s}$$
$$\tilde{y}_{ijs} = \frac{y_{is} - y_{js}}{h_s}$$
$$\overline{x}^h_{is} = \frac{x^*_s - x_{is}}{h_s}$$
$$\overline{x}^b_{is} = \frac{x^*_s - x_{is}}{b_s}$$

and

$$R^h_{x,ij\ell s} = K\left(\tilde{x}_{ijs}\right) K\left(\tilde{x}^h_{i\ell s}\right)$$
$$R_{y,ij\ell s} = \int K\left(\tilde{y}_{j\ell s} + \overline{y}_{js}\right) K\left(\overline{y}_{js}\right) d\overline{y}_{js}$$
$$R_{y,ij\ell k} = \int K\left(\tilde{y}_{j\ell k} + \overline{y}_{jk}\right) K'\left(\overline{y}_{jk}\right) d\overline{y}_{jk}$$

This is estimable from the data.

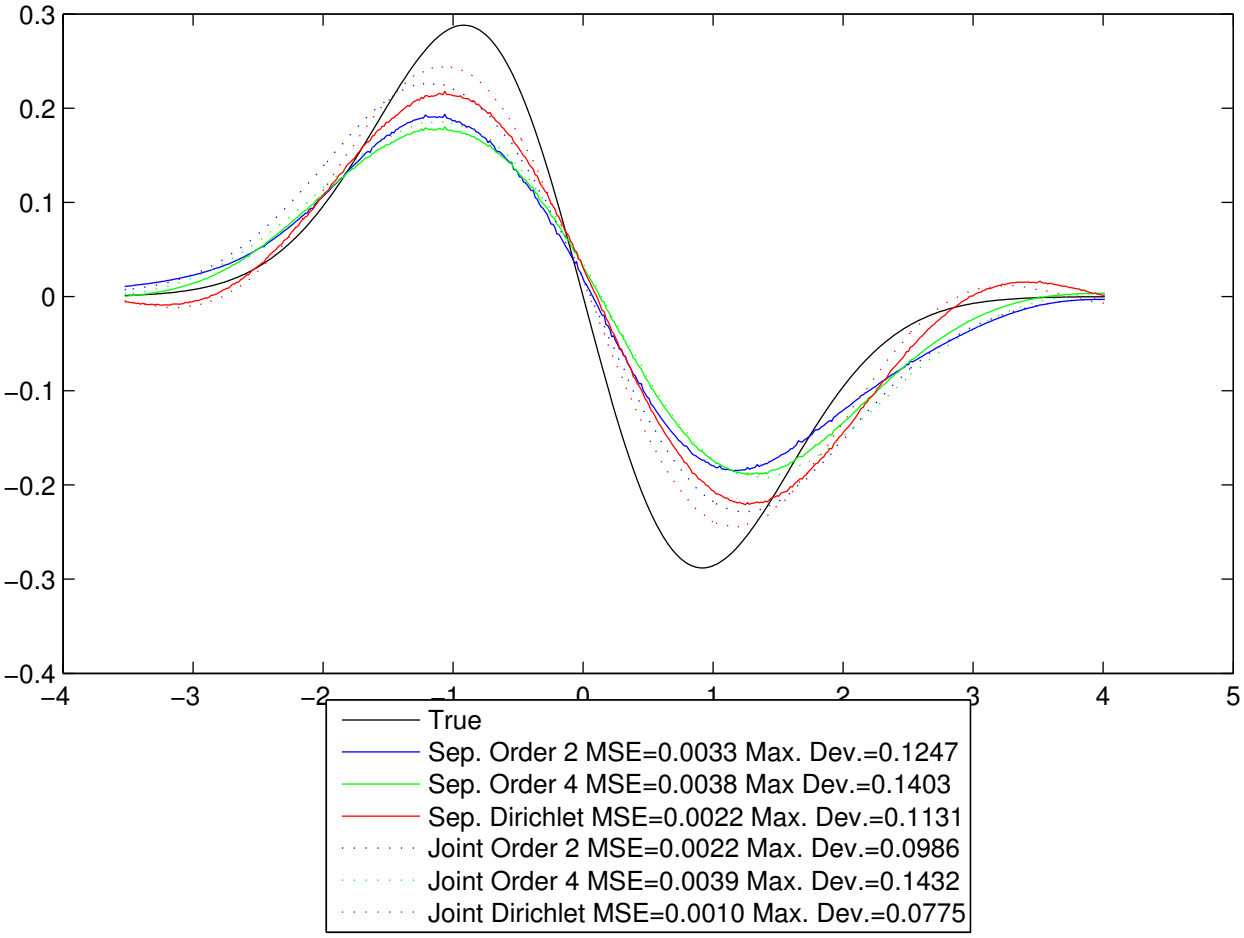
As for convergence of the parameters to the optimal parameters, it seems that an extension of the proof contained in Hall, Racine and Li (2004) for the joint estimation of a multivariate conditional density would apply here for the derivative of a multivariate conditional density.

2.3.3 Simulation

I run a Monte Carlo simulation. Data is simulated from a multivariate normal distribution with zero mean and variance-covariance matrix with variances given by $\sigma_i^2 = 1$ and covariance elements given by $\sigma_{ij} = 0.4$. The simulation tests use the criterion that estimates the bandwidths separately. A separate, smaller Monte Carlo study compares the results for using the criterion that estimates the bandwidths separately and the one that jointly estimates the bandwidths. Figure 2.2 presents one result for $N = 500$ and $k = 2$, conditioning on one

variable.

Figure 2.2: Derivative of Conditional Normal Density and Estimations, $N = 500$,$k = 1$



The following table are the results of $k = 2$; further results are in Section B.2.

Tables 2.5 and 2.6 show the results for the Monte Carlo study comparing joint and separate estimation for a two dimensional density derivative conditioned on one variable. The results for 4 dimensional densities conditioned on 1,2 or 3 variables are in the Appendix in Tables B.16-B.21.

When using the separate bandwidth criteria, the results are very mixed, but overall, 2nd or 4th order kernels actually seem to perform the best. The Dirichlet kernel often runs into

Table 2.3: Monte Carlo Simulation Results: Mean Square Error Comparisons for Derivative of 2-Dimensional Density Conditioned on 1 of the Variables

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100
2nd Order	0.112	0.0398	0.0774
	(0.35)	(0.0706)	(0.272)
4th Order	2.66	0.0379	0.241
	(26)	(0.053)	(1.48)
6th Order	0.161	0.621	1.71
	(0.648)	(5.25)	(14.5)
8th Order	0.466	0.132	2.48
	(3.52)	(0.448)	(22.3)
10th Order	0.696	2.06	0.641
	(6.04)	(19.5)	(4.57)
Dirichlet	6.04	132	48.6
	(49.9)	(1.31e+03)	(481)

Standard Deviations of MSE Simulation Estimates in Parentheses

Table 2.4: Monte Carlo Simulation Results: Mean Maximum Difference Comparisons for Derivative of 2-Dimensional Density Conditioned on 1 of the Variables

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100
2nd Order	0.73	0.804	1.37
	(1.22)	(0.76)	(2.92)
4th Order	2.27	1.47	3.77
	(16)	(2.04)	(12.5)
6th Order	1.28	3.94	8.93
	(3.17)	(16.5)	(40.2)
8th Order	1.61	3.03	9.42
	(6.37)	(5.83)	(48.8)
10th Order	1.89	6.02	7.27
	(7.94)	(31.5)	(21.7)
Dirichlet	5.82	31.7	27.8
	(23.8)	(256)	(220)

Standard Deviations of Maximum Deviation Simulation Estimates in Parentheses

problems when estimating numerator and denominator bandwidths separately. Looking at the ratio of deviations to function height in Tables B.14-B.15 shows that larger densities and more parameters to estimate decrease the accuracy. As more variables are conditioned on, even the best estimators get more unstable and the results are worse. Perhaps this

Table 2.5: Monte Carlo Simulation Results: Mean Square Error Comparisons for Derivative of 2-Dimensional Density Conditioned on 1 of the Variables

	Sep, N=100 SIMS=50	Joint, N=100 SIMS=50	Sep, N=500 SIMS=50	Joint, N=500 SIMS=50
2nd Order	17246	0.44272	9.0201e+10	0.10818
	(1.0341e+05)	(2.5078)	(5.8601e+11)	(0.42864)
4th Order	1.2819	0.078906	2.7963	0.028441
	(8.1813)	(0.14188)	(12.621)	(0.059)
Dirichlet	0.10719	0.034195	0.32468	0.015824
	(0.19442)	(0.1056)	(1.3278)	(0.0059094)

Sep. separately estimates the bandwidths for the numerator density and denominator density; Joint uses the joint criteria to estimate both bandwidths jointly
Standard Deviations of MSE Simulation Estimates in Parentheses

Table 2.6: Monte Carlo Simulation Results: Mean Maximum Difference Comparisons for Derivative of 2-Dimensional Density Conditioned on 1 of the Variables

	Sep, N=100 SIMS=50	Joint, N=100 SIMS=50	Sep, N=500 SIMS=50	Joint, N=500 SIMS=50
2nd Order	248.51	1.0512	1.1953e+06	1.5753
	(1302.6)	(2.0256)	(6.6756e+06)	(4.7528)
4th Order	2.6252	0.82402	11.411	0.87319
	(8.8802)	(0.61743)	(34.012)	(0.83264)
Dirichlet	1.7764	0.43967	4.6449	0.56448
	(2.0053)	(0.24731)	(9.2914)	(0.31097)

Sep. separately estimates the bandwidths for the numerator density and denominator density; Joint uses the joint criteria to estimate both bandwidths jointly
Standard Deviations of MSE Simulation Estimates in Parentheses

instability also leads to worse ratio results for higher values of N, a counter-intuitive result. Higher values of n increase the likelihood of drawing outliers, which, when estimated in the denominator, blows up the point estimate beyond what it should be.

However, when comparing the use of the separate bandwidth estimators vs. the criterion that jointly estimates the bandwidths, the new criterion that jointly estimates the bandwidths consistently outperforms the separate estimation, and the Dirichlet kernel also performs the best. The results are encouraging for estimation of a derivative of a conditional density, as the joint bandwidth estimator using a Dirichlet kernel consistently does the best, with results that are sufficiently accurate for analysis.

2.4 Conclusion

Researchers have developed and employed various methods to evaluate derivatives of univariate densities nonparametrically. However, little attention has been given to multivariate cases. In this chapter, I examine cross validation methods for higher dimension densities, comparing different kernel orders and various criteria. I develop various minimizing criteria for both the estimation of joint and conditional densities and establish consistency of the estimators.

A Monte Carlo simulation to test out the criteria suggests the complications inherent with estimating high dimensional density derivatives, as the accuracy of the estimates sharply decreases with increased dimension. The computational requirements are prohibitively large for increasing the sample size to compensate for the larger dimension when the dimension is overly large. However, when high dimensional derivative densities need to be estimated, the simulations suggest that higher order kernels are more effective, and in particular, use of the Dirichlet infinite order kernel performs the best in the class of Gaussian kernels. Weighted integrated square error methods are both slower and less effective generally, so I recommend for derivatives of multidimensional densities using the direct cross validation methods.

For the estimation of the derivative of a conditional density, the criterion that jointly estimates the bandwidths using a Dirichlet kernel yields the best results in the simulations, suggesting the best results would come from using this cross validation criterion. A possible extension is to test using a trimming function in the criterion and the estimator, which should lead to more accurate estimators for these cases.

B Appendix

B.1 Proofs and Derivations

Proofs

Theorem 1. *Assume that the r^{th} order kernel has the following characteristics:*

1. $\int \cdots \int \prod_{s \in G_x} K(x_s) dx_1 \cdots dx_q = 1$
2. $K(x_s) = K(-x_s)$
3. $\int \cdots \int x_k^{r-1} \prod_{s \in G_x} K(x_s) dx_1 \cdots dx_q = 0$
4. $\int \cdots \int x_k^{r} \prod_{s \in G_x} K(x_s) dx_1 \cdots dx_q > 0$
5. $\int \cdots \int x_k^{r} \prod_{s \in G_x} K(x_s) dx_1 \cdots dx_q < \infty$
6. $f(x)$ *is* $(r+1)$ *times differentiable*

where $d_x = \prod_{s \in G_x} dx_s$. *Then,*

$$
\begin{aligned}
MSE\left(\frac{\partial \widehat{f}(x)}{\partial x_k}\right) &= \left(\frac{\int K(x)\, x^r dx}{r!} \sum_{t=1}^{q} h_t^r \frac{\partial^{r+1} f(x)}{\partial x_k \partial^r x_t} + O\left(\sum_{t=1}^{q} h_t^{r+1}\right)\right)^2 \\
&\quad + \frac{f(x)}{n h_k^2 \prod_{s \in G_x} h_s} \int K'(x)^2 dx \left(\int K(x)^2 dx\right)^{q-1} + O\left(\frac{1}{n h_k \prod_{s \in G_x} h_s}\right) \\
&= O\left(\left(\sum_{t=1}^{q} h_t^{r+1}\right)^2 + \frac{1}{n h_k \prod_{s \in G_x} h_s}\right)
\end{aligned}
$$

and, if as $n \to \infty$, $\max_j\{h_j\} \to 0$ *and* $n h_k \prod_{s \in G_x} h_s \to \infty$, *then the last statement directly implies that* $\frac{\partial \widehat{f}(x)}{\partial x_k} \to \frac{\partial f(x)}{\partial x_k}$ *in MSE, implying also convergence in probability, and consistency.*

Proof.

$$MSE\left(\frac{\partial \widehat{f}(x)}{\partial x_k}\right) = var\left(\frac{\partial \widehat{f}(x)}{\partial x_k}\right) + \left[bias\left(\frac{\partial \widehat{f}(x)}{\partial x_k}\right)\right]^2$$

First, examine the bias:

$$\begin{aligned}
&bias\left(\frac{\partial \widehat{f}(x)}{\partial x_k}\right) = E\left(\frac{\partial \widehat{f}(x)}{\partial x_k}\right) - \frac{\partial f(x)}{\partial x_k} \\
&= E\left[\frac{1}{nh_k \prod_{s\in G_x} h_s} \sum_{i=1}^{n} K'\left(\frac{x_k - x_{ik}}{h_k}\right) \prod_{s\in G_x \backslash k} K\left(\frac{x_s - x_{is}}{h_s}\right)\right] - \frac{\partial f(x)}{\partial x_k} \\
&= \frac{1}{nh_k \prod_{s\in G_x} h_s} \sum_{i=1}^{n} E\left[K'\left(\frac{x_k - x_{ik}}{h_k}\right) \prod_{s\in G_x \backslash k} K\left(\frac{x_s - x_{is}}{h_s}\right)\right] - \frac{\partial f(x)}{\partial x_k} \\
&= \frac{1}{h_k \prod_{s\in G_x} h_s} E\left[K'\left(\frac{x_k - x_{ik}}{h_k}\right) \prod_{s\in G_x \backslash k} K\left(\frac{x_s - x_{is}}{h_s}\right)\right] - \frac{\partial f(x)}{\partial x_k} \\
&= \frac{1}{h_k \prod_{s\in G_x} h_s} \int \cdots \int K'\left(\frac{x_k - x_{ik}}{h_k}\right) \prod_{s\in G_x \backslash k} K\left(\frac{x_s - x_{is}}{h_s}\right) f(x_i) dx_i - \frac{\partial f(x)}{\partial x_k}
\end{aligned}$$

Substitute $z_{is} = \frac{x_{is} - x_s}{h_s}$, Note that this implies $x_{is} = x_s + h_s z_s$ and $dx_{is} = h_s dz_{is}$. This yields

$$\begin{aligned}
&\frac{1}{h_k \prod_{s\in G_x} h_s} \int \cdots \int K'\left(\frac{x_k - x_{ik}}{h_k}\right) \prod_{s\in G_x \backslash k} K\left(\frac{x_s - x_{is}}{h_s}\right) f(x_i) dx_i - \frac{\partial f(x)}{\partial x_k} \\
&= \frac{1}{h_k} \int \cdots \int K'\left(-z_{ik}\right) \prod_{s\in G_x \backslash k} K\left(-z_{is}\right) f(x + hz_i) dz_i - \frac{\partial f(x)}{\partial x_k}
\end{aligned}$$

Next, assume that $K\left(-z_{is}\right) = K\left(z_{is}\right)$. This also implies that $K'\left(-z_{is}\right) = -K'\left(z_{is}\right)$. Then

$$\begin{aligned}
&\frac{1}{h_k} \int \cdots \int K'\left(-z_{ik}\right) \prod_{s\in G_x \backslash k} K\left(-z_{is}\right) f(x + hz_i) dz_i - \frac{\partial f(x)}{\partial x_k} \\
&= \frac{-1}{h_k} \int \cdots \int K'\left(z_{ik}\right) \prod_{s\in G_x \backslash k} K\left(z_{is}\right) f(x + hz_i) dz_i - \frac{\partial f(x)}{\partial x_k}
\end{aligned}$$

Using integration by parts, and noting that $f(x_i + hz_i)K(z_k)|_{-\infty}^{\infty} = 0$,

$$\begin{aligned}
&\frac{-1}{h_k}\int\cdots\int K'(z_{ik})\prod_{s\in G_x\setminus k}K(z_{is})\,f(x+hz_i)dz_i - \frac{\partial f(x)}{\partial x_k} \\
&= \int\cdots\int \prod_{s\in G_x}K(z_{is})\,f_k(x+hz_i)dz_i - \frac{\partial f(x)}{\partial x_k}
\end{aligned}$$

Next, take a $(r+1)$ degree Taylor expansion of $f_k(x+hz_i)$ around x. Separating the expansion into relevant groupings:

$$\begin{aligned}
f_k(x+hz_i) = &\sum_{j_1,\ldots j_q|j_i\geq 0,\sum_i j_i\in[0,r-1]}\frac{\partial^{\sum_{i=1}^q j_i+1}f(x)}{\partial x_k\partial^{j_1}x_1\partial^{j_2}x_2\cdots\partial^{j_q}x_q}\prod_{\ell=1}^{q}\frac{(h_\ell z_{i\ell})^{j_\ell}}{j_\ell!} \\
&+\sum_{j_1,\ldots j_q|j_i\geq 0,\sum_i j_i=r}\frac{\partial^{\sum_{i=1}^q j_i+1}f(x)}{\partial x_k\partial^{j_1}x_1\partial^{j_2}x_2\cdots\partial^{j_q}x_q}\prod_{\ell=1}^{q}\frac{(h_\ell z_{i\ell})^{j_\ell}}{j_\ell!} \\
&+\sum_{j_1,\ldots j_q|j_i\geq 0,\sum_i j_i=r+1}\cdots\sum_{j_q=(r+1)\times 1(t=q)}^{r+1}\frac{\partial^{\sum_{i=1}^q j_i+1}f(\xi)}{\partial x_k\partial^{j_1}x_1\partial^{j_2}x_2\cdots\partial^{j_q}x_q}\prod_{\ell=1}^{q}\frac{(h_\ell z_{i\ell})^{j_\ell}}{j_\ell!}
\end{aligned}$$

where $\xi\in[x, x+hz_i]$. Substituting this in,

$$\begin{aligned}
bias\left(\frac{\partial\widehat{f}(x)}{\partial x_k}\right) = &\int\cdots\int\prod_{s\in G_x}K(z_{is})\sum_{j_1,\ldots j_q|j_i\geq 0,\sum_i j_i\in[0,r-1]}\frac{\partial^{\sum_{i=1}^q j_i+1}f(x)}{\partial x_k\partial^{j_1}x_1\partial^{j_2}x_2\cdots\partial^{j_q}x_q}\prod_{\ell=1}^{q}\frac{(h_\ell z_{i\ell})^{j_\ell}}{j_\ell!}dz_i \\
&+\int\cdots\int\prod_{s\in G_x}K(z_{is})\sum_{j_1,\ldots j_q|j_i\geq 0,\sum_i j_i=r}\frac{\partial^{\sum_{i=1}^q j_i+1}f(x)}{\partial x_k\partial^{j_1}x_1\partial^{j_2}x_2\cdots\partial^{j_q}x_q}\prod_{\ell=1}^{q}\frac{(h_\ell z_{i\ell})^{j_\ell}}{j_\ell!}dz_i \\
&+\int\cdots\int\prod_{s\in G_x}K(z_{is})\sum_{j_1,\ldots j_q|j_i\geq 0,\sum_i j_i=r+1}\frac{\partial^{\sum_{i=1}^q j_i+1}f(\xi)}{\partial x_k\partial^{j_1}x_1\partial^{j_2}x_2\cdots\partial^{j_q}x_q}\prod_{\ell=1}^{q}\frac{(h_\ell z_{i\ell})^{j_\ell}}{j_\ell!}
\end{aligned}$$

Consider the first term; this yields

$$\int \cdots \int \prod_{s \in G_x} K\left(z_{is}\right) \sum_{j_1, \ldots j_q | j_i \geq 0, \sum_i j_i \in [0, r-1]} \frac{\partial^{\sum_{i=1}^q j_i + 1} f(x)}{\partial x_k \partial^{j_1} x_1 \partial^{j_2} x_2 \cdots \partial^{j_q} x_q} \prod_{\ell=1}^{q} \frac{\left(h_\ell z_{i\ell}\right)^{j_\ell}}{j_\ell!} dz_i$$
$$= \sum_{j_1, \ldots j_q | j_i \geq 0, \sum_i j_i \in [0, r-1]} \frac{\partial^{\sum_{i=1}^q j_i + 1} f(x)}{\partial x_k \partial^{j_1} x_1 \partial^{j_2} x_2 \cdots \partial^{j_q} x_q} \int \cdots \int \prod_{s \in G_x} K\left(z_{is}\right) \prod_{\ell=1}^{q} \frac{\left(h_\ell z_{i\ell}\right)^{j_\ell}}{j_\ell!} dz_i$$

Given Assumption 3, all of the integrals evaluate to zero here except for the one where all j are equal to zero. When all s are equal to zero, Assumption 1 says that $\int \cdots \int \prod_{s \in G_x} K\left(z_{is}\right) dz_i = 1$. Therefore,

$$\sum_{j_1, \ldots j_q | j_i \geq 0, \sum_i j_i \in [0, r-1]} \frac{\partial^{\sum_{i=1}^q j_i + 1} f(x)}{\partial x_k \partial^{j_1} x_1 \partial^{j_2} x_2 \cdots \partial^{j_q} x_q} \int \cdots \int \prod_{s \in G_x} K\left(z_{is}\right) \prod_{\ell=1}^{q} \frac{\left(h_\ell z_{i\ell}\right)^{j_\ell}}{j_\ell!} dz_i = \frac{\partial f(x)}{\partial x_k}$$

Next, consider the second term. Now, by Assumption 3, all of the summands that have any $j_i \in (0, r)$ is equal to zero. Therefore, the only remaining elements are those where one is equal to r and the rest are equal to zero (these, by Assumption 2, integrate to one). Therefore, the second term is

$$\int \cdots \int \prod_{s \in G_x} K\left(z_{is}\right) \sum_{j_1, \ldots j_q | j_i \geq 0, \sum_i j_i = r} \frac{\partial^{\sum_{i=1}^q j_i + 1} f(x)}{\partial x_k \partial^{j_1} x_1 \partial^{j_2} x_2 \cdots \partial^{j_q} x_q} \prod_{\ell=1}^{q} \frac{\left(h_\ell z_{i\ell}\right)^{j_\ell}}{j_\ell!} dz_i$$
$$= \frac{1}{r!} \sum_{t=1}^{q} h_t^r \frac{\partial^{r+1} f(x)}{\partial x_k \partial^r x_t} \int \cdots \int \prod_{s \in G_x} K\left(z_{is}\right) z_{it}^r dz_i$$
$$= \frac{1}{r!} \sum_{t=1}^{q} h_t^r \frac{\partial^{r+1} f(x)}{\partial x_k \partial^r x_t} \int K\left(z_{it}\right) z_{it}^r dz_{it}$$
$$= \frac{\int K\left(x\right) x^r dx}{r!} \sum_{t=1}^{q} h_t^r \frac{\partial^{r+1} f(x)}{\partial x_k \partial^r x_t}$$

Next, consider the final term. Again, any of the summands that include $z_{i\ell}$ for $\ell \in [0, r]$

is equal to zero.

$$\int \cdots \int \prod_{s \in G_x} K(z_{is}) \sum_{j_1,...j_q | j_i \geq 0, \sum_i j_i = r+1} \frac{\partial^{\sum_{i=1}^q j_i + 1} f(\xi)}{\partial x_k \partial^{j_1} x_1 \partial^{j_2} x_2 \cdots \partial^{j_q} x_q} \prod_{\ell=1}^{q} \frac{(h_\ell z_{i\ell})^{j_\ell}}{j_\ell!} dz_i$$
$$= \frac{1}{(1+r)!} \sum_{t=1}^{q} h_t^{r+1} \frac{\partial^{r+2} f(\xi)}{\partial x_k \partial^{r+1} x_t} \int \cdots \int \prod_{s \in G_x} K(z_{is}) z_{it}^{r+1} dz_i$$
$$= \frac{1}{(1+r)!} \sum_{t=1}^{q} h_t^{r+1} \frac{\partial^{r+2} f(\xi)}{\partial x_k \partial^{r+1} x_t} \int K(z_{it}) z_{it}^{r+1} dz_{it}$$

Using Assumption 5, for some C

$$\frac{1}{(1+r)!} \sum_{t=1}^{q} h_t^{r+1} \frac{\partial^{r+2} f(\xi)}{\partial x_k \partial^{r+1} x_t} \int K(z_{it}) z_{it}^{r+1} dz_{it} < C \sum_{t=1}^{q} h_t^{r+1} = O\left(\sum_{t=1}^{q} h_t^{r+1} \right)$$

Combining these results,

$$bias\left(\frac{\partial \widehat{f}(x)}{\partial x_k} \right) = \frac{\int K(x) x^r dx}{r!} \sum_{t=1}^{q} h_t^r \frac{\partial^{r+1} f(x)}{\partial x_k \partial^r x_t} + O\left(\sum_{t=1}^{q} h_t^{r+1} \right)$$

Next, examine the variance.

$$var\left(\frac{\partial \widehat{f}(x)}{\partial x_k} \right)$$
$$= var\left(\frac{1}{n h_k \prod_{s \in G_x} h_s} \sum_{i=1}^{n} K'\left(\frac{x_k - x_{ik}}{h_k} \right) \prod_{s \in G_x \backslash k} K\left(\frac{x_s - x_{is}}{h_s} \right) \right)$$
$$= \frac{1}{n h_k^2 \prod_{s \in G_x} h_s^2} var\left(K'\left(\frac{x_k - x_{ik}}{h_k} \right) \prod_{s \in G_x \backslash k} K\left(\frac{x_s - x_{is}}{h_s} \right) \right)$$
$$= \frac{1}{n h_k^2 \prod_{s \in G_x} h_s^2} \left[E\left(K'\left(\frac{x_k - x_{ik}}{h_k} \right)^2 \prod_{s \in G_x \backslash k} K\left(\frac{x_s - x_{is}}{h_s} \right)^2 \right) \right.$$
$$\left. - E\left(K'\left(\frac{x_k - x_{ik}}{h_k} \right) \prod_{s \in G_x \backslash k} K\left(\frac{x_s - x_{is}}{h_s} \right) \right)^2 \right]$$

Consider the terms separately. The first term is

$$\begin{aligned}
&\frac{1}{nh_k^2\prod_{s\in G_x}h_s^2}\left[E\left(K'\left(\frac{x_k-x_{ik}}{h_k}\right)^2\prod_{s\in G_x\backslash k}K\left(\frac{x_s-x_{is}}{h_s}\right)^2\right)\right]\\
&=\frac{1}{nh_k^2\prod_{s\in G_x}h_s^2}\int\cdots\int K'\left(\frac{x_k-x_{ik}}{h_k}\right)^2\prod_{s\in G_x\backslash k}K\left(\frac{x_s-x_{is}}{h_s}\right)^2 f(x_i)dx_i\\
&=\frac{1}{nh_k\prod_{s\in G_x}h_s}\int\cdots\int K'\left(-z_{ik}\right)^2\prod_{s\in G_x\backslash k}K\left(-z_{is}\right)^2 f(x+z_ih)dz_i\\
&=\frac{1}{nh_k\prod_{s\in G_x}h_s}\int\cdots\int K'\left(z_{ik}\right)^2\prod_{s\in G_x\backslash k}K\left(z_{is}\right)^2\left[f(x)+\sum_{t=1}^{q}\frac{\partial f(\xi)}{\partial x_k}h_tz_{it}\right]dz_i\\
&=\frac{f(x)}{nh_k\prod_{s\in G_x}h_s}\int K'\left(x\right)^2dx\left(\int K\left(x\right)^2dx\right)^{q-1}+O\left(\frac{1}{nh_k\prod_{s\in G_x}h_s}\right)
\end{aligned}$$

Next consider the second term

$$\begin{aligned}
&\frac{1}{nh_k^2\prod_{s\in G_x}h_s^2}E\left(K'\left(\frac{x_k-x_{ik}}{h_k}\right)\prod_{s\in G_x\backslash k}K\left(\frac{x_s-x_{is}}{h_s}\right)\right)^2\\
&=\frac{1}{nh_k^2}\left(\int\cdots\int K'\left(z_{ik}\right)\prod_{s\in G_x\backslash k}K\left(z_{is}\right)dz_i\right)^2\\
&=O\left(\frac{1}{nh_k^2}\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&var\left(\frac{\partial\widehat{f}(x)}{\partial x_k}\right)=\frac{f(x)}{nh_k\prod_{s\in G_x}h_s}\int K'\left(x\right)^2dx\left(\int K\left(x\right)^2dx\right)^{q-1}\\
&+O\left(\frac{1}{nh_k\prod_{s\in G_x}h_s}\right)-O\left(\frac{1}{nh_k^2}\right)\\
&=\frac{f(x)}{nh_k\prod_{s\in G_x}h_s}\int K'\left(x\right)^2dx\left(\int K\left(x\right)^2dx\right)^{q-1}+O\left(\frac{1}{nh_k\prod_{s\in G_x}h_s}\right)
\end{aligned}$$

Combining these results

$$
\begin{aligned}
MSE\left(\frac{\partial \widehat{f}(x)}{\partial x_k}\right) &= \left(\frac{\int K(x)\, x^r dx}{r!} \sum_{t=1}^{q} h_t^r \frac{\partial^{r+1} f(x)}{\partial x_k \partial^r x_t} + O\left(\sum_{t=1}^{q} h_t^{r+1}\right)\right)^2 \\
&\quad + \frac{f(x)}{n h_k^2 \prod_{s \in G_x} h_s} \int K'(x)^2 dx \left(\int K(x)^2 dx\right)^{q-1} + O\left(\frac{1}{n h_k \prod_{s \in G_x} h_s}\right) \\
&= O\left(\left(\sum_{t=1}^{q} h_t^{r+1}\right)^2 + \frac{1}{n h_k \prod_{s \in G_x} h_s}\right)
\end{aligned}
$$

Therefore, if as $n \to \infty$, $\max_j\{h_j\} \to 0$ and $n h_k \prod_{s \in G_x} h_s \to \infty$, then the last statement directly implies that $\frac{\partial \widehat{f}(x)}{\partial x_k} \to \frac{\partial f(x)}{\partial x_k}$ in MSE, implying also convergence in probability, and consistency. □

Theorem 2. *Assume that the r^{th} order kernel has the following characteristics:*

1. $\int \cdots \int \prod_{s \in G_x} K(x_s) dx_1 \cdots dx_q = 1$
2. $K(x_s) = K(-x_s)$
3. $\int \cdots \int x_k^{r-1} \prod_{s \in G_x} K(x_s) dx_1 \cdots dx_q = 0$
4. $\int \cdots \int x_k^{r} \prod_{s \in G_x} K(x_s) dx_1 \cdots dx_q > 0$
5. $\int \cdots \int x_k^{r} \prod_{s \in G_x} K(x_s) dx_1 \cdots dx_q < \infty$
6. *$f(x)$ is $(r+1)$ times differentiable*

where $d_x = \prod_{s \in G_x} dx_s$. Then, if as $n \to \infty$, $\max_j\{h_j\} \to 0$ and $n h_k \prod_{s \in G_x} h_s \to \infty$, then $\frac{\partial \widehat{f}_{Y|X}(y|x)}{\partial y_k} \xrightarrow{p} \frac{\partial f_{Y|X}(y|x)}{\partial y_k}$.

Proof.

$$
\frac{\partial \widehat{f}_{Y|X}(y|x)}{\partial y_k} = \frac{\partial \widehat{f}_{Y,X}(y,x)}{\partial y_k} \widehat{f}_X(x)^{-1}
$$

By Theorem 1,

$$\frac{\partial \widehat{f}_{Y,X}(y,x)}{\partial y_k} \xrightarrow{p} \frac{\partial f_{Y,X}(y,x)}{\partial y_k}$$

By a similar theorem (such as is contained in Li and Racine 2007),

$$\widehat{f}_X(x) \xrightarrow{p} f_X(x)$$

Then, by Slutsky's Theorem,

$$\frac{\partial \widehat{f}_{Y,X}(y,x)}{\partial y_k} \widehat{f}_X(x)^{-1} \xrightarrow{p} \frac{\partial f_{Y,X}(y,x)}{\partial y_k} f_X(x)^{-1} = \frac{\partial f_{Y|X}(y|x)}{\partial y_k}$$

□

Derivation of CV bandwidth for Derivative of Joint

I consider the derivative of a joint density with respect to variable $x_k \in x$.

$$\begin{aligned} ISE(h) &= \int \cdots \int \left(\frac{\partial \widehat{f}(x)}{\partial x_k} - \frac{\partial f(x)}{\partial x_k} \right)^2 dx \\ &= \int \cdots \int \frac{\partial \widehat{f}(x)}{\partial x_k}^2 dx - 2 \int \cdots \int \frac{\partial \widehat{f}(x)}{\partial x_m} \frac{\partial f(x)}{\partial x_k} dx + \int \cdots \int \frac{\partial f(x)}{\partial x_k}^2 dx \\ &= ISE_1(h) - 2 * ISE_2(h) + ISE_3 \end{aligned}$$

Again, ISE_3 is not a function of the bandwidth selection, so minimizing ISE is identical to minimizing

$$ISE^*(h) = ISE_1(h) - 2 * ISE_2(h)$$

Examine each term separately

$$ISE_1(h) = \frac{1}{n^2\left(\prod_{s\in G_x} h_s\right)^2 h_k^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int\cdots\int K'\left(\frac{x_k - x_{ik}}{h_k}\right)K'\left(\frac{x_k - x_{jk}}{h_k}\right)\cdots$$
$$\times\prod_{s\in G_x\backslash k} K\left(\frac{x_s - x_{is}}{h_s}\right)K\left(\frac{x_s - x_{js}}{h_s}\right)dx$$

Let

$$\overline{x}_{is} = \frac{x_s - x_{is}}{h_s}$$
$$\tilde{x}_{ijs} = \frac{x_{is} - x_{js}}{h_s}$$

Then

$$ISE_1(h) = \frac{1}{n^2\left(\prod_{s\in G_x} h_s\right)^2 h_k^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int\cdots\int\left[\int h_k K'\left(\tilde{x}_{ijk} + \overline{x}_{ik}\right)K'\left(\overline{x}_{ik}\right)d\overline{x}_{ik}\right]\cdots$$
$$\times\prod_{s\in G_x\backslash k}\left(h_s K\left(\tilde{x}_{ijs} + \overline{x}_{is}\right)K\left(\overline{x}_{is}\right)d\overline{x}_{is}\right)$$
$$= \frac{1}{n^2 h_k^2\prod_{s\in G_x} h_s}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\int K'\left(\tilde{x}_{ijs} + \overline{x}_{is}\right)K'\left(\overline{x}_{is}\right)d\overline{x}_{is}\right]\prod_{s\in G_x\backslash k}\left(\int K\left(\tilde{x}_{ijs} + \overline{x}_{is}\right)K\left(\overline{x}_{is}\right)d\overline{x}_{is}\right) \tag{B.1}$$

As in the joint density case, the evaluation of these integrals depend on the kernel chosen. Section 1 provides an example of this, the analogy of which carries directly into this example, except now derivatives of the kernel must also be taken prior to integration of that section.

Next, examine $ISE_2(h)$:

$$ISE_2(h) = \int\cdots\int\frac{\partial\widehat{f}(x)}{\partial x_k}\frac{\partial f(x)}{\partial x_k}dx$$

Integrating by parts,

$$\begin{aligned}\int\cdots\int \frac{\partial \widehat{f}(x)}{\partial x_k}\frac{\partial f(x)}{\partial x_k}dx &= -\int\cdots\int \frac{\partial^2 \widehat{f}(x)}{\partial x_k^2}f(x)dx \\ &= -E\left(\frac{\partial^2 \widehat{f}(x)}{\partial x_k^2}\right) \\ &= \frac{-1}{n(n-1)\left(\prod_{s\in G_x} h_s\right)^2 h_k^2}\sum_{i=1}^{n}\sum_{j\neq i} K''\left(\tilde{x}_{ijk}\right)\prod_{s\in G_x\setminus k} K\left(\tilde{x}_{ijs}\right)\end{aligned}$$

Putting these together

$$\begin{aligned}&ISE^*(h) = \\ &\frac{1}{n^2h_k^2\prod_{s\in G_x} h_s}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\int K'\left(\tilde{x}_{ijs}+\overline{x}_{is}\right)K'\left(\overline{x}_{is}\right)d\overline{x}_{is}\right]\prod_{s\in G_x\setminus k}\left(\int K\left(\tilde{x}_{ijs}+\overline{x}_{is}\right)K\left(\overline{x}_{is}\right)d\overline{x}_{is}\right) \\ &+\frac{2}{n(n-1)\left(\prod_{s\in G_x} h_s\right)^2 h_k^2}\sum_{i=1}^{n}\sum_{j\neq i} K''\left(\tilde{x}_{ijk}\right)\prod_{s\in G_x\setminus k} K\left(\tilde{x}_{ijs}\right)\end{aligned}$$

In the example kernel, this would mean

$$\begin{aligned}ISE^*(h) = &\frac{1}{n^2h_k^2\prod_{s\in G_x} h_s}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(40\tilde{x}_{ijk}-84\tilde{x}_{ijk}^2\right)\prod_{s\in G_x\setminus k}\left(\frac{8}{3}-10\tilde{x}_{ijk}^2+7\tilde{x}_{ijk}^4\right) \\ &+\frac{2}{n(n-1)\left(\prod_{s\in G_x} h_s\right)^2 h_k^2}\sum_{i=1}^{n}\sum_{j\neq i}\left(84\tilde{x}_{ijk}^2-20\right)\prod_{s\in G_x\setminus k}\left(7\tilde{x}_{ijk}^4-10\tilde{x}_{ijk}^2+3\right)\end{aligned}$$

Derivation of Weighted CV bandwidth for Derivative of Joint

I consider the derivative of a joint density with respect to variable $x_k \in x$. Let h be the bandwidth that the researcher is trying to minimize with respect to, and $\widehat{f}(x;b)$ be the weighting function. b is estimated prior to this, so that the weighting is independent of the

bandwidth selection h. Then, the weighted integrated square error is given by

$$\begin{aligned}
WISE(h) &= \int \cdots \int \left(\frac{\partial \widehat{f}(x;h)}{\partial x_k} - \frac{\partial f(x)}{\partial x_k} \right)^2 \widehat{f}(x;b)dx \\
&= \int \cdots \int \frac{\partial \widehat{f}(x;h)}{\partial x_k}^2 \widehat{f}(x;b)dx - 2\int \cdots \int \frac{\partial \widehat{f}(x;h)}{\partial x_m} \frac{\partial f(x)}{\partial x_k} \widehat{f}(x;b)dx \\
&\quad + \int \cdots \int \frac{\partial f(x)}{\partial x_k}^2 \widehat{f}(x;b)dx \\
&= WISE_1(h) - 2 * WISE_2(h) + WISE_3
\end{aligned}$$

$WISE_3$ is not a function of the bandwidth (h) selection, so minimizing $WISE$ is identical to minimizing

$$WISE^*(h) = WISE_1(h) - 2 * WISE_2(h)$$

I examine each term

$$\begin{aligned}
WISE_1(h) &= \int \cdots \int \frac{\partial \widehat{f}(x;h)}{\partial x_k}^2 \widehat{f}(x;b)dx \\
&= \int \cdots \int \frac{1}{h_k^2 n^3 \prod_{s \in G_x} h_s^2 b_s} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{m=1}^{n} K'\left(\frac{x_k - x_{ik}}{h_k}\right) K'\left(\frac{x_k - x_{jk}}{h_k}\right) K\left(\frac{x_k - x_{mk}}{b_k}\right) \cdots \\
&\quad \times \prod_{s \in G_x \setminus k} K\left(\frac{x_s - x_{is}}{h_s}\right) K\left(\frac{x_s - x_{js}}{h_s}\right) K\left(\frac{x_s - x_{ms}}{b_s}\right) dx \\
&= \frac{1}{h_k^2 n^3 \prod_{s \in G_x} h_s^2 b_s} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{m=1}^{n} \int K'\left(\frac{x_k - x_{ik}}{h_k}\right) K'\left(\frac{x_k - x_{jk}}{h_k}\right) K\left(\frac{x_k - x_{mk}}{b_k}\right) dx_k \cdots \\
&\quad \times \prod_{s \in G_x \setminus k} \int K\left(\frac{x_s - x_{is}}{h_s}\right) K\left(\frac{x_s - x_{js}}{h_s}\right) K\left(\frac{x_s - x_{ms}}{b_s}\right) dx_s
\end{aligned}$$

This depends on the choice of the kernel. Evaluate the integrals and $WISE_1$ is estimable.

Next, I examine $WISE_2(h)$:

$$WISE_2(h) = \int \cdots \int \frac{\partial \widehat{f}(x;h)}{\partial x_k} \frac{\partial f(x)}{\partial x_k} \widehat{f}(x;b) dx$$

Integrating by parts, and assuming that the value of the derivative is bounded, this becomes

$$\begin{aligned}
WISE_2(h) &= -E\left[\frac{\partial^2 \widehat{f}(x;h)}{\partial x_k^2} \widehat{f}(x;b) + \frac{\partial \widehat{f}(x;h)}{\partial x_k} \frac{\partial \widehat{f}(x;b)}{\partial x_k}\right] \\
&= C_1 \sum_{m=1}^{n} \left[\sum_{i \neq m} K''\left(\frac{x_{mk} - x_{ik}}{h_k}\right) \prod_{s \in G_x} K\left(\frac{x_{ms} - x_{is}}{h_s}\right) \sum_{j \neq m} \prod_{s \in G_x} K\left(\frac{x_{ms} - x_{js}}{b_s}\right)\right] \\
&+ C_2 \sum_{m=1}^{n} \left[\sum_{i \neq m} K'\left(\frac{x_{mk} - x_{ik}}{h_k}\right) \prod_{s \in G_x} K\left(\frac{x_{ms} - x_{is}}{h_s}\right) \sum_{j \neq m} K'\left(\frac{x_{mk} - x_{jk}}{b_k}\right) \prod_{s \in G_x} K\left(\frac{x_{ms} - x_{js}}{b_s}\right)\right]
\end{aligned}$$

where

$$\begin{aligned}
C_1 &= \frac{1}{n(n-1)^2 h_k^2 \prod_{s \in G_x} h_s b_s} \\
C_2 &= \frac{1}{n(n-1)^2 h_k b_k \prod_{s \in G_x} h_s b_s}
\end{aligned}$$

All of the elements are now evaluated in order to estimate the $WISE$ and minimize with respect to h. However, note that it does become very complicated when trying to evaluate $WISE_1$ for a given kernel, as the integral across the derivatives of the kernels is in general not a simple task. Because of this, the researcher will most likely need to limit attention to simpler (generally lower order) kernels. Here, I give the example of the 2nd order Gaussian Kernel for exposition. In that case,

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\{-.5x^2\} \equiv \phi(x)$$
$$K'(x) = -x\phi(x)$$
$$K''(x) = (x^2 - 1)\phi(x)$$

For $\int K\left(\frac{x-r}{h}\right) K\left(\frac{x-s}{h}\right) K\left(\frac{x-t}{b}\right) dx$

$$= \int \frac{1}{(2\pi)^{3/2}} \exp\left\{-.5\left(\left(\frac{x-r}{h}\right)^2 + \left(\frac{x-s}{h}\right)^2 + \left(\frac{x-t}{b}\right)^2\right)\right\} dx$$
$$= \int \frac{1}{(2\pi)^{3/2}} \exp\left\{-\frac{1}{2\Omega}\left((x-\mu)^2 + D\right)\right\} dx$$
$$= \frac{\sqrt{\Omega}\exp\{\frac{1}{2}\Omega^{-1}D\}}{2\pi} \int \frac{1}{\sqrt{2\pi\Omega}} \exp\left\{-\frac{1}{2\Omega}(x-\mu)^2\right\} dx$$
$$= \frac{\sqrt{\Omega}\exp\{\frac{1}{2}\Omega^{-1}D\}}{2\pi}$$

where

$$\Omega = \frac{b^2h^2}{2b^2 + h^2}$$
$$\mu = \frac{b^2r + b^2s + h^2t}{2b^2 + h^2}$$
$$D = \frac{r^2b^2 + s^2b^2 + t^2h^2}{2b^2 + h^2} - \mu^2$$

Similarly, for $\int K'\left(\frac{x-r}{h}\right) K'\left(\frac{x-s}{h}\right) K\left(\frac{x-t}{b}\right) dx$

$$
\begin{aligned}
&= \int \frac{(x-r)(x-s)}{h^2(2\pi)^{3/2}} \exp\left\{-.5\left(\left(\frac{x-r}{h}\right)^2 + \left(\frac{x-s}{h}\right)^2 + \left(\frac{x-t}{b}\right)^2\right)\right\} dx \\
&= \int \frac{x^2 - (r+s)x + rs}{h^2(2\pi)^{3/2}} \exp\left\{-\frac{1}{2\Omega}\left((x-\mu)^2 + D\right)\right\} dx \\
&= \frac{\sqrt{\Omega}\exp\{\frac{1}{2}\Omega^{-1}D\}}{2h^2\pi} \int \frac{1}{\sqrt{2\pi\Omega}}\left(x^2 - (r+s)x + rs\right) \exp\left\{-\frac{1}{2\Omega}(x-\mu)^2\right\} dx \\
&= \frac{\sqrt{\Omega}\exp\{\frac{1}{2}\Omega^{-1}D\}}{2h^2\pi}\left(\mathrm{Var}(x) + E[x]^2 - (r+s)E[x] + rs\right) \\
&= \frac{\sqrt{\Omega}\exp\{\frac{1}{2}\Omega^{-1}D\}}{2h^2\pi}\left(\Omega + \mu^2 - (r+s)\mu + rs\right)
\end{aligned}
$$

where Ω, μ, and D are similarly defined.

Derivation of CV bandwidth for Derivative of Conditional

Let G_y be the set of bandwidths associated with variables y, and G_x the set of bandwidths associated with variables x. Then $x \in \mathbb{R}^{|G_x|}$ and $y \in \mathbb{R}^{|G_y|}$.

This procedure chooses bandwidths both for the full joint (numerator) and the marginal (conditioning, denominator) densities, given by h and b. The cross validation is given by

$$
\begin{aligned}
ISE(h,b) &= \int \cdots \int \left(\frac{\partial \widehat{f}_{Y|X}(y|x)}{\partial y_k} - \frac{\partial f_{Y|X}(y|x)}{\partial y_k}\right)^2 m(x) dy dx \\
&= \int \cdots \int \left(\frac{\partial}{\partial y_k}\left(\frac{\widehat{f}(y,x)}{\widehat{m}(x)}\right) - \frac{\partial}{\partial y_k}\left(\frac{f(y,x)}{m(x)}\right)\right)^2 m(x) dy dx \\
&= \int \cdots \int \left(\frac{\partial \widehat{f}(y,x)}{\partial y_k}\widehat{m}(x)^{-1} - \frac{\partial f(y,x)}{\partial y_k} m(x)^{-1}\right)^2 m(x) dy dx \\
&= \int \cdots \int \left(\frac{\partial \widehat{f}(y,x)}{\partial y_k}\right)^2 \widehat{m}(x)^{-2} m(x) dy dx - 2\int \cdots \int \frac{\partial \widehat{f}(y,x)}{\partial y_k}\frac{\partial f(y,x)}{\partial y_k}\widehat{m}(x)^{-1} dy dx \cdots \\
&\quad + \int \cdots \int \left(\frac{\partial f(y,x)}{\partial y_k}\right)^2 m(x)^{-1} dy \\
&= ISE_1(h,b) - 2 * ISE_2(h,b) + ISE_3
\end{aligned}
$$

As usual, ISE_3 is not a function of the bandwidth, so that the minimization of $ISE(h,b,)$ is identical to that of

$$ISE^*(h,b) = ISE_1(h,b) - 2 * ISE_2(h,b)$$

I examine each term

$$\begin{aligned} ISE_1(h,b) &= \int \cdots \int \left(\frac{\partial \widehat{f}(y,x)}{\partial y_k} \right)^2 \widehat{m}(x)^{-2} m(x) dy dx \\ &= E\left[\int \cdots \int \frac{\partial \widehat{f}(y,x)}{\partial y_k}^2 \widehat{m}(x)^{-2} dy \right] \end{aligned}$$

This I approximate with the sample analogue

$$\begin{aligned} \widehat{ISE}_1(h,b) &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{(n-1)^2 \prod_{s \in G_X} b_s^2} \left(\sum_{j=1, \neq i}^{n} \prod_{s \in G_x} K\left(\frac{x_{is} - x_{js}}{h_s} \right) \right)^2 \right)^{-1} \cdots \\ &\times \sum_{j \neq i} \sum_{\ell \neq i} \prod_{s \in G_x} K\left(\frac{x_{is} - x_{js}}{h_s} \right) K\left(\frac{x_{is} - x_{\ell s}}{h_s} \right) \left[\int K'\left(\frac{y_k - y_{jk}}{h_k} \right) K'\left(\frac{y_k - y_{\ell k}}{h_k} \right) dy_k \right] \cdots \\ &\times \frac{\left[\prod_{s \in G_y \setminus k} \int K\left(\frac{y_s - y_{js}}{h_s} \right) K\left(\frac{y_s - y_{\ell s}}{h_s} \right) dy_s \right]}{(n-1)^2 h_k^2 \prod_{s \in G_x \cup G_y} h_s^2} \end{aligned}$$

As before, let

$$\begin{aligned} \overline{y}_{is} &= \frac{y_s - y_{is}}{h_s} \\ \tilde{y}_{ijs} &= \frac{y_{is} - y_{js}}{h_s} \\ \overline{x}_{is}^h &= \frac{x_s^* - x_{is}}{h_s} \\ \overline{x}_{is}^b &= \frac{x_s^* - x_{is}}{b_s} \end{aligned}$$

Then

$$\widehat{ISE}_1(h,b) = \frac{\prod_{s\in G_X} b_s^2}{nh_k^2 \prod_{s\in G_y} h_s \prod_{s\in G_x} h_s^2} \sum_{i=1}^{n} \left(\sum_{j=1,\neq i}^{n} \prod_{s\in G_x} K\left(\tilde{x}_{ijs}^b\right) \right)^{-2} \cdots$$
$$\times \sum_{j\neq i} \sum_{\ell\neq i} \prod_{s\in G_x} K\left(\tilde{x}_{ijs}^h\right) K\left(\tilde{x}_{i\ell s}^h\right) \left[\int K'\left(\tilde{y}_{j\ell k} + \overline{y}_{js}\right) K'\left(\overline{y}_{js}\right) d\overline{y}_{js} \right] \cdots$$
$$\times \left[\prod_{s\in G_y \setminus k} \int K\left(\tilde{y}_{j\ell k} + \overline{y}_{js}\right) K\left(\overline{y}_{js}\right) d\overline{y}_{js} \right]$$

As for $ISE_2(h,b)$

$$\begin{aligned}
ISE_2(h,b) &= \int \cdots \int \frac{\partial \widehat{f}(y,x)}{\partial y_k} \frac{\partial f(y,x)}{\partial y_k} \widehat{m}(x)^{-1} dy dx \\
&= -\int \cdots \int \frac{\partial^2 \widehat{f}(y,x)}{\partial y_k^2} f(y,x) \widehat{m}(x)^{-1} dy dx \\
&= -E\left[\frac{\partial^2 \widehat{f}(y,x)}{\partial y_k^2} \widehat{m}(x)^{-1} \right] \\
&= -\frac{1}{n} \sum_{i=1}^{n} \left[\frac{\frac{1}{(n-1)h_k^2 \prod_{s\in G_x \cup G_y} h_s} \sum_{j\neq i} K''\left(\tilde{y}_{ijk}\right) \left[\prod_{s\in G_y\setminus k} K(\tilde{y}_{ijk}) \right] \prod_{s\in G_x} K(\tilde{x}_{ijs}^h)}{\frac{1}{(n-1)\prod_{s\in G_x} b_s} \sum_{j\neq i} \prod_{s\in G_x} K\left(\tilde{x}_{ijs}^b\right)} \right] \\
&= -\frac{\prod_{s\in G_x} b_s}{nh_k^2 \prod_{s\in G_x \cup G_y} h_s} \sum_{i=1}^{n} \left[\frac{\sum_{j\neq i} K''\left(\tilde{y}_{ijk}\right) \left[\prod_{s\in G_y\setminus k} K(\tilde{y}_{ijk}) \right] \prod_{s\in G_x} K(\tilde{x}_{ijs}^h)}{\sum_{j\neq i} \prod_{s\in G_x} K\left(\tilde{x}_{ijs}^b\right)} \right]
\end{aligned}$$

where I used integration by parts to go from line 1 to line 2. This is estimable as well, so putting together $ISE_1(h,b,x^*)$ and $ISE_2(h,b,x^*)$, the criterion becomes

$$\begin{aligned}
&ISE(h,b) = \\
&\frac{\prod_{s\in G_X} b_s^2}{nh_k^2 \prod_{s\in G_y} h_s \prod_{s\in G_x} h_s^2} \sum_{i=1}^{n} \frac{\sum_{j\neq i} \sum_{\ell\neq i} \prod_{s\in G_x} R_{x,ij\ell s}^h R_{y,ij\ell k} \prod_{s\in G_y\setminus k} R_{y,ij\ell s}}{\left(\sum_{j=1,\neq i}^{n} \prod_{s\in G_x} K\left(\tilde{x}_{ijs}^b\right) \right)^2} \\
&+ 2\frac{\prod_{s\in G_x} b_s}{nh_k^2 \prod_{s\in G_x \cup G_y} h_s} \sum_{i=1}^{n} \left[\frac{\sum_{j\neq i} K''\left(\tilde{y}_{ijk}\right) \left[\prod_{s\in G_y\setminus k} K(\tilde{y}_{ijk}) \right] \prod_{s\in G_x} K(\tilde{x}_{ijs}^h)}{\sum_{j\neq i} \prod_{s\in G_x} K\left(\tilde{x}_{ijs}^b\right)} \right]
\end{aligned}$$

where

$$
\begin{aligned}
R^{h}_{x,ij\ell s} &= K\left(\tilde{x}_{ijs}\right) K\left(\tilde{x}^{h}_{i\ell s}\right) \\
R_{y,ij\ell s} &= \int K\left(\tilde{y}_{j\ell s} + \overline{y}_{js}\right) K\left(\overline{y}_{js}\right) d\overline{y}_{js} \\
R_{y,ij\ell k} &= \int K\left(\tilde{y}_{j\ell k} + \overline{y}_{jk}\right) K'\left(\overline{y}_{jk}\right) d\overline{y}_{jk}
\end{aligned}
$$

B.2 Tables

Table B.1: Monte Carlo Simulation Results: Mean Square Error Comparisons for Derivative of 1-Dimensional Density

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100	N=5000 SIMS=100
2nd Order	0.914	0.0253	0.0186	0.0155
	(9.05)	(0.149)	(0.116)	(0.132)
4th Order	0.0102	0.0108	0.00253	0.000962
	(0.0201)	(0.0685)	(0.00696)	(0.00197)
6th Order	0.0114	0.00428	0.00186	0.000919
	(0.024)	(0.0097)	(0.00407)	(0.00234)
8th Order	0.0125	0.00454	0.0018	0.000923
	(0.0268)	(0.0104)	(0.00439)	(0.00267)
10th Order	0.015	0.00482	0.0019	0.000942
	(0.0301)	(0.0101)	(0.0048)	(0.00286)
Dirichlet	0.011	0.00448	0.00221	0.00114
	(0.0347)	(0.0106)	(0.0115)	(0.00427)
WISE	0.0525	0.092	0.0189	0.00444
	(0.233)	(0.473)	(0.0765)	(0.0204)

Standard Deviations of MSE Simulation Estimates in Parentheses

Table B.2: Monte Carlo Simulation Results: Mean Square Error Comparisons for Derivative of 4-Dimensional Density

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100
2nd Order	0.00257	0.000207	0.000113
	(0.012)	(0.000958)	(0.000652)
4th Order	8.43e-05	2.64e-05	1.7e-05
	(0.000147)	(4.01e-05)	(2.13e-05)
6th Order	6.05e-05	2.15e-05	1.37e-05
	(9.25e-05)	(2.8e-05)	(1.9e-05)
8th Order	5.58e-05	1.82e-05	1.56e-05
	(8.59e-05)	(2.84e-05)	(6.45e-05)
10th Order	3.23e-05	2.25e-05	2.04e-05
	(8.62e-06)	(4.41e-06)	(3.13e-06)
Dirichlet	2.21e-05	8.12e-06	5.51e-06
	(2.06e-05)	(4.92e-06)	(3.58e-06)
WISE	0.452	0.00493	0.0153
	(3.48)	(0.0233)	(0.146)

Standard Deviations of MSE Simulation Estimates in Parentheses

Table B.3: Monte Carlo Simulation Results: Mean Square Error Comparisons for Derivative of 8-Dimensional Density

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=12
2nd Order	2.42e-05	3.38e-08	1.64e-08
	(0.00013)	(5.54e-08)	(3.55e-09)
4th Order	3.56e-08	1.3e-08	1.04e-08
	(6.13e-08)	(3.37e-09)	(1.78e-09)
6th Order	2.04e-08	1.14e-08	8.84e-09
	(1.31e-08)	(2.77e-09)	(1.58e-09)
8th Order	1.68e-08	1.07e-08	8.22e-09
	(6.8e-09)	(2.26e-09)	(1.53e-09)
10th Order	2.25e-08	2.34e-08	2.29e-08
	(7.16e-09)	(3.07e-09)	(2.86e-09)
Dirichlet	1.71e-08	1.17e-08	8.65e-09
	(5.64e-09)	(1.87e-09)	(1.42e-09)
WISE	122	39.1	0.397
	(1.21e+03)	(314)	(1.36)

Standard Deviations of MSE Simulation Estimates in Parentheses

Table B.4: Monte Carlo Simulation Results: Mean Maximum Difference Comparisons for Derivative of 1-Dimensional Density

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100	N=5000 SIMS=100
2nd Order	0.439	0.184	0.146	0.113
	(2.92)	(0.385)	(0.366)	(0.357)
4th Order	0.139	0.114	0.0734	0.0479
	(0.131)	(0.248)	(0.0895)	(0.0498)
6th Order	0.141	0.0899	0.0631	0.0431
	(0.142)	(0.0907)	(0.0684)	(0.053)
8th Order	0.145	0.0915	0.0603	0.0408
	(0.15)	(0.0954)	(0.0686)	(0.0556)
10th Order	0.158	0.0951	0.061	0.0405
	(0.164)	(0.0963)	(0.0718)	(0.0564)
Dirichlet	0.117	0.0824	0.0533	0.0367
	(0.145)	(0.0873)	(0.0874)	(0.0611)
WISE	0.241	0.335	0.167	0.0985
	(0.49)	(0.89)	(0.358)	(0.171)

Standard Deviations of Maximum Deviation Simulation Estimates in Parentheses

Table B.5: Monte Carlo Simulation Results: Mean Maximum Difference Comparisons for Derivative of 4-Dimensional Density

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100
2nd Order	0.0805	0.0332	0.0268
	(0.221)	(0.0524)	(0.0457)
4th Order	0.0264	0.0179	0.015
	(0.0196)	(0.0099)	(0.00805)
6th Order	0.0225	0.0161	0.0132
	(0.0144)	(0.00893)	(0.00674)
8th Order	0.0213	0.0147	0.0122
	(0.0132)	(0.0086)	(0.00866)
10th Order	0.017	0.0161	0.0157
	(0.00269)	(0.00174)	(0.00118)
Dirichlet	0.0139	0.0101	0.00869
	(0.00416)	(0.00256)	(0.0024)
WISE	1.01	0.164	0.139
	(4.57)	(0.424)	(0.799)

Standard Deviations of Maximum Deviation Simulation Estimates in Parentheses

Table B.6: Monte Carlo Simulation Results: Mean Maximum Difference Comparisons for Derivative of 8-Dimensional Density

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=12
2nd Order	0.00859	0.00102	0.000851
	(0.0297)	(0.000702)	(0.000111)
4th Order	0.000715	0.000705	0.000729
	(0.000418)	(0.000117)	(0.000111)
6th Order	0.00061	0.000679	0.000686
	(0.000186)	(0.000114)	(9.81e-05)
8th Order	0.000584	0.00067	0.00067
	(0.000159)	(0.000111)	(9.53e-05)
10th Order	0.000659	0.000894	0.000979
	(0.000179)	(0.000145)	(9.82e-05)
Dirichlet	0.000603	0.000713	0.000705
	(0.000169)	(0.000125)	(9.01e-05)
WISE	8.94	17.8	4.67
	(78)	(97.7)	(13.9)

Standard Deviations of Maximum Deviation Simulation Estimates in Parentheses

Table B.7: Monte Carlo Simulation Results: Mean Square Error Comparisons for Derivative of 4-Dimensional Density Conditioned on 1 of the Variables

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100
2nd Order	0.0387	0.000761	0.000747
	(0.196)	(0.00147)	(0.00258)
4th Order	0.00472	0.000778	0.00332
	(0.0219)	(0.00244)	(0.0195)
6th Order	0.00178	0.00116	0.0194
	(0.00753)	(0.00587)	(0.141)
8th Order	0.00127	0.000582	0.00367
	(0.0027)	(0.000911)	(0.0219)
10th Order	0.00213	0.0187	0.00209
	(0.0146)	(0.146)	(0.00892)
Dirichlet	0.00264	0.00666	0.000906
	(0.00957)	(0.037)	(0.0032)

Standard Deviations of MSE Simulation Estimates in Parentheses

Table B.8: Monte Carlo Simulation Results: Mean Square Error Comparisons for Derivative of 4-Dimensional Density Conditioned on 2 of the Variables

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100
2nd Order	0.489	0.00947	0.0051
	(2.55)	(0.0208)	(0.00576)
4th Order	0.029	0.228	0.00652
	(0.128)	(2.07)	(0.0141)
6th Order	0.0113	0.0159	11.1
	(0.0325)	(0.0465)	(111)
8th Order	0.0217	0.657	0.0835
	(0.114)	(4.95)	(0.619)
10th Order	0.292	0.327	0.0762
	(2.88)	(3.07)	(0.688)
Dirichlet	0.0833	1.95	0.332
	(0.281)	(10.6)	(2.34)

Standard Deviations of MSE Simulation Estimates in Parentheses

Table B.9: Monte Carlo Simulation Results: Mean Square Error Comparisons for Derivative of 4-Dimensional Density Conditioned on 3 of the Variables

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100
2nd Order	4.41 (24.4)	0.138 (0.447)	0.13 (0.463)
4th Order	0.278 (1.27)	62.8 (627)	0.0397 (0.038)
6th Order	0.226 (1.35)	0.188 (0.68)	0.534 (3.14)
8th Order	0.219 (0.946)	7.04 (39.3)	4.01 (31.6)
10th Order	0.0321 (0.0168)	0.11 (0.782)	7.67 (51.8)
Dirichlet	4.77e+03 (4.75e+04)	6.03 (30.6)	22.5 (141)

Standard Deviations of MSE Simulation Estimates in Parentheses

Table B.10: Monte Carlo Simulation Results: Mean Maximum Difference Comparisons for Derivative of 4-Dimensional Density Conditioned on 1 of the Variables

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100
2nd Order	0.274 (0.785)	0.123 (0.146)	0.163 (0.291)
4th Order	0.182 (0.515)	0.202 (0.384)	0.495 (1.6)
6th Order	0.134 (0.282)	0.248 (0.585)	0.976 (4.27)
8th Order	0.148 (0.222)	0.205 (0.299)	0.586 (1.7)
10th Order	0.13 (0.397)	0.601 (2.98)	0.507 (1.21)
Dirichlet	0.226 (0.418)	0.568 (1.7)	0.334 (0.622)

Standard Deviations of Maximum Deviation Simulation Estimates in Parentheses

Table B.11: Monte Carlo Simulation Results: Mean Maximum Difference Comparisons for Derivative of 4-Dimensional Density Conditioned on 2 of the Variables

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100
2nd Order	1.02	0.381	0.353
	(3)	(0.444)	(0.327)
4th Order	0.41	2.15	0.812
	(0.768)	(10.4)	(1.21)
6th Order	0.361	1.18	12.8
	(0.448)	(1.92)	(105)
8th Order	0.507	3.88	3.06
	(1.13)	(17.7)	(8.35)
10th Order	0.759	2.14	1.95
	(5.35)	(11.6)	(8.11)
Dirichlet	1.42	10	7.05
	(2.39)	(29.6)	(16.3)

Standard Deviations of Maximum Deviation Simulation Estimates in Parentheses

Table B.12: Monte Carlo Simulation Results: Mean Maximum Difference Comparisons for Derivative of 4-Dimensional Density Conditioned on 3 of the Variables

	N=100 SIMS=100	N=500 SIMS=100	N=1000 SIMS=100
2nd Order	2.82	1.27	1.38
	(8.69)	(1.62)	(2.42)
4th Order	1.1	19.1	1.38
	(2.26)	(177)	(1.55)
6th Order	1.2	3.62	9.22
	(2.86)	(7.61)	(19.8)
8th Order	1.48	15.9	17.6
	(3.68)	(57.2)	(59.7)
10th Order	0.519	1.69	18.3
	(0.431)	(6.4)	(85.9)
Dirichlet	81.4	25.3	53.8
	(689)	(47.3)	(139)

Standard Deviations of Maximum Deviation Simulation Estimates in Parentheses

Table B.13: Derivative Density Size Comparisons

	Dimension	$\max_{x_i} \lvert\partial f(x_i)/\partial x_k\rvert$	$\frac{1}{n}\sum_i \frac{\partial f(x_i)}{\partial x_k}$
Joint	1	0.2420	0.1628
	2	0.1148	0.0520
	4	0.0253	0.0062
	8	0.0011	0.00009
Conditional	2|1	0.2881	0.1898
	4|1	0.0627	0.0174
	4|2	0.1497	0.0582
	4|3	0.3300	0.2073

Table B.14: Ratio of Best Average Maximum Deviation to Maximum True Derivative Density Height

	Dimension	$N = 100$	$N = 500$	$N = 1000$	$N = 5000$
Joint	1	0.4835	0.3405	0.2202	0.1517
	2	0.4852	0.3598	0.3554	0.1542
	4	0.5494	0.3992	0.2727	-
	8	0.5309	0.6091	0.6091	-
Conditional	2|1	2.5338	2.7907	4.7553	-
	4|1	2.0734	1.9617	2.5997	-
	4|2	2.4516	2.5451	2.3580	-
	4|3	3.3333	3.8485	4.1818	-

Table B.15: Ratio of Best Average Maximum Deviation to Average True Derivative Density Height

	Dimension	$N = 100$	$N = 500$	$N = 1000$	$N = 5000$
Joint	1	0.7187	0.5061	0.3274	0.2254
	2	1.0712	0.7942	0.7846	0.3404
	4	2.2419	1.6290	1.1129	-
	8	6.1855	7.0963	7.0963	-
Conditional	2|1	3.8462	4.2360	7.2181	-
	4|1	7.4713	7.0690	9.3678	-
	4|2	6.3058	6.5464	6.0653	-
	4|3	5.3063	6.1264	6.6570	-

Table B.16: Monte Carlo Simulation Results: Mean Square Error Comparisons for Derivative of 4-Dimensional Density Conditioned on 1 of the Variables

	Sep, N=100 SIMS=50	Joint, N=100 SIMS=50	Sep, N=500 SIMS=30	Joint, N=500 SIMS=30
2nd Order	1.8895e+18	0.059669	2.4899e+10	0.00056034
	(1.3361e+19)	(0.38274)	(1.3638e+11)	(0.0009355)
4th Order	0.001849	0.0010108	0.00050669	0.0003913
	(0.0049039)	(0.002624)	(0.00030772)	(0.0003876)
Dirichlet	0.00099285	0.00036763	0.69302	0.00022667
	(0.0021582)	(0.00032739)	(3.7877)	(2.9678e-05)

Sep. separately estimates the bandwidths for the numerator density and denominator density; Joint uses the joint criteria to estimate both bandwidths jointly
Standard Deviations of MSE Simulation Estimates in Parentheses

Table B.17: Monte Carlo Simulation Results: Mean Square Error Comparisons for Derivative of 4-Dimensional Density Conditioned on 2 of the Variables

	Sep, N=100 SIMS=50	Joint, N=100 SIMS=50	Sep, N=500 SIMS=22	Joint, N=500 SIMS=22
2nd Order	5.903e+44	4.1496e+06	6.5948e+05	0.036363
	(4.1741e+45)	(2.9339e+07)	(3.0922e+06)	(0.12374)
4th Order	71285	2.9459e+07	2.2161	0.0034022
	(5.0405e+05)	(2.0831e+08)	(10.209)	(0.00043934)
Dirichlet	1.2867	0.0041199	0.34795	0.0022534
	(6.4603)	(0.0058914)	(0.83613)	(0.00031385)

Sep. separately estimates the bandwidths for the numerator density and denominator density; Joint uses the joint criteria to estimate both bandwidths jointly
Standard Deviations of MSE Simulation Estimates in Parentheses

Table B.18: Monte Carlo Simulation Results: Mean Square Error Comparisons for Derivative of 4-Dimensional Density Conditioned on 3 of the Variables

	Sep, N=100 SIMS=50	Joint, N=100 SIMS=50	Sep, N=500 SIMS=11	Joint, N=500 SIMS=11
2nd Order	4.8371e+26	2.0834e+05	34549	1521.3
	(3.4203e+27)	(1.4214e+06)	(1.1439e+05)	(5035.3)
4th Order	118.65	0.62519	0.50565	0.036632
	(712.52)	(4.0691)	(1.1355)	(0.0053253)
Dirichlet	3.0883	0.055043	2.6181	0.020093
	(7.9263)	(0.088448)	(4.0135)	(0.001585)

Sep. separately estimates the bandwidths for the numerator density and denominator density; Joint uses the joint criteria to estimate both bandwidths jointly
Standard Deviations of MSE Simulation Estimates in Parentheses

Table B.19: Monte Carlo Simulation Results: Mean Maximum Difference Comparisons for Derivative of 4-Dimensional Density Conditioned on 1 of the Variables

	Sep, N=100 SIMS=50	Joint, N=100 SIMS=50	Sep, N=500 SIMS=30	Joint, N=500 SIMS=30
2nd Order	1.944e+09	0.46551	6.4419e+05	0.11837
	(1.3746e+10)	(1.7377)	(3.5284e+06)	(0.18745)
4th Order	0.18139	0.10103	0.23989	0.10341
	(0.29698)	(0.1259)	(0.18958)	(0.13978)
Dirichlet	0.14478	0.057056	3.7853	0.059145
	(0.20658)	(0.019158)	(18.534)	(0.012627)

Sep. separately estimates the bandwidths for the numerator density and denominator density; Joint uses the joint criteria to estimate both bandwidths jointly
Standard Deviations of MSE Simulation Estimates in Parentheses

Table B.20: Monte Carlo Simulation Results: Mean Maximum Difference Comparisons for Derivative of 4-Dimensional Density Conditioned on 2 of the Variables

	Sep, N=100 SIMS=50	Joint, N=100 SIMS=50	Sep, N=500 SIMS=22	Joint, N=500 SIMS=22
2nd Order	3.436e+22	2163.7	3945.2	0.59563
	(2.4296e+23)	(14912)	(18142)	(0.99435)
4th Order	380.84	5980.5	9.3335	0.22981
	(2669.4)	(42275)	(32.576)	(0.064033)
Dirichlet	3.9604	0.17571	6.8833	0.17421
	(10.697)	(0.10851)	(11.278)	(0.025509)

Sep. separately estimates the bandwidths for the numerator density and denominator density; Joint uses the joint criteria to estimate both bandwidths jointly
Standard Deviations of MSE Simulation Estimates in Parentheses

Table B.21: Monte Carlo Simulation Results: Mean Maximum Difference Comparisons for Derivative of 4-Dimensional Density Conditioned on 3 of the Variables

	Sep, N=100 SIMS=50	Joint, N=100 SIMS=50	Sep, N=500 SIMS=11	Joint, N=500 SIMS=11
2nd Order	3.1103e+13	734.06	1299.9	170.49
	(2.1993e+14)	(3829.1)	(4088.6)	(534.32)
4th Order	29.242	1.7745	7.1625	0.90023
	(104.67)	(7.4295)	(11.018)	(0.40122)
Dirichlet	8.9915	0.6608	25.668	0.58
	(14.757)	(0.39911)	(24.811)	(0.1121)

Sep. separately estimates the bandwidths for the numerator density and denominator density; Joint uses the joint criteria to estimate both bandwidths jointly
Standard Deviations of MSE Simulation Estimates in Parentheses

B.3 References

Bearse, Peter and Paul Rilstone, P. (2008). "Higher Order Bias Reduction of Kernel Density and Density Derivative Estimators at Boundary Points," 7th Annual Advances in Econometrics Conference, Baton Rouge, Louisiana.

Cacoullos, Theophilos (1966). "Estimation of a Multivariate Density," *Annals of the Institute of Statistical Mathematics" (Tokyo)*, 18:2, 179-189.

Chacon, J.E., Duong, T., and Wand, M.P. (2010). "Asymptotics for General Multivariate Kernel Density Derivative Estimators." *Statistica Sinica*, 21, 807-840.

Hall, Peter, Racine, Jeff, and Li, Qi (2004). "Cross-Validation and the Estimation of Conditional Probability Densities." *Journal of the American Statistical Association*, 99(468), 1015-1026.

Hansen, Bruce (2005). "Exact Mean Integrated Squared Error of Higher Order Kernel Estimators," *Econometric Theory*, 21: 10311057

Hardle, Wolfgang, J.S. Marron and M.P. Wand (1990), "Bandwidth Choice for Density Derivatives," *Journal of the Royal Statistical Society, Series B (Methodological)* 52:1 223-232.

Li, Qi, Racine, Jeffrey (2007). *Nonparametric Econometrics*. Princeton University Press, Princeton, NJ.

Loader, Clive R. (1999). "Bandwidth Selection: Classical or Plug-In?" *The Annals of Statistics* 27:2 415-438.

Marron, J.S. (1994). "Visual Understanding of Higher-Order Kernels," *Journal of Computational and Graphical Statistics*, 3:4 447-458.

Marron, J.S. and Hardle, W. (1986). "Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation," *Journal of Multivariate Analysis* 20, 91-113.

Marron, J.S. and M.P. Wand (1992). "Exact mean integrated squared error," *Annals of Statistics*, 20, 712-736.

Turlach, Berwin A. (1993). "Bandwidth Selection in Kernel Density Estimation: A Review," Discussion Paper 9307, Institut für Statistik und Ökonometrie, Humbolt-Universität zu Berlin, Berlin, Germany.

Wand and Schucany (1990). "Gaussian-Based Kernels," *The Canadian Journal of Statistics* 18:3 197-204.

Wu, Tiee-Jian (1997). "Root n Bandwidth Selectors for Kernel Estimation of Density Derivatives," *Journal of the American Statistical Association* 92:438 536:547.

Zhang, Xibin, Maxwell L. King, and Rob J. Hyndman (2006). "A Bayesian approach to bandwidth selection for multivariate kernel density estimation," *Computational Statistics & Data Analysis* 50, 3009-3031.

Chapter 3

Testing the Additive Separability of the Teacher Value Added Effect Semiparametrically

3.1 Introduction

Many school districts rely on subjective teacher assessments to evaluate teacher performance. However, there are reasons these measures may be inaccurate: assessments based on classroom observations can occur less than once a year and are scheduled in advance, allowing teachers to exert more effort on the announced day of the evaluation (Taylor and Tyler, 2012), and the highest rating can be given to nearly all teachers. For instance, in Los Angeles Unified School District, less than two percent of all teachers are rated as unsatisfactory and over 90% of teachers receive no negative ratings on any of the 25 ratings categories in the evaluation form (Buddin, 2011).

As an alternative, some school districts are also including in their teacher assessments objective evaluations of the teachers based on an output of the education production function: standardized test scores. Florida, Indiana, Rhode Island, Tennessee, and Colorado, as well

as school districts in Houston, Denver, Dallas, Minneapolis and Washington, D.C., already make use of test scores to estimate teachers' value added (Corcoran, 2010). The New York City Department of Education has estimated teacher effects on test scores for more than 10,000 teachers, and Los Angeles Unified School District has developed its own estimates as well.

Implicitly, these statistical models are based on estimations of an education production function. They are usually estimated by a linear regression of student test scores on previous scores, covariates from administrative data, with an additively separable teacher fixed effects variable. Additive separability of the teacher effect implies that the measured teacher value added does not change by different student characteristics. Todd and Wolpin (2003) show how the common, additively separable linear specification implies unobservable inputs and endowments must decay at a common geometric rate, one example of the restrictions assumed when using these models. Estimating the production function and the teacher value added with a linear model and additively separable teacher effect has at least two direct potential problems.

First, misspecification of the estimation model leads to biased estimation of the production function, the marginal effects of teachers, and thus the teacher rankings (the typical object of interest). The most common value-added specifications do not include flexible interactions between student and teacher characteristics. This assumption of additive separability of the teacher effect means the marginal effects of a teacher's value added is the same for all types of students. However, there may actually be a relationship between a teacher's value added and the ability of his or her students, so that one teacher works better with high performing students and another teacher performs better with low performing students. Incorrect specification of the model leads to biased estimates of the average marginal effect, and even more biased estimates of marginal effects or teacher rankings for low or high performing students.

Second, additive separability causes the teacher effect to be reduced to a single constant,

which misses the rich heterogeneity of the teacher effect, providing incomplete information for specific policy questions that are interested in the performance of teachers among low or high achieving students. Not only is there higher within-teacher variation of student's characteristics than between teacher variation, we estimate significantly higher within-teacher variation of their own value added than between-teacher variation of the mean value added. Teachers have more personal variation in their ability to help students than between other teachers' average ability. The standard deviations in value added *across* teachers is measured at .34 for English and .36 for math, while the standard deviations *within* teacher is .87 for English and .83 for math.[1] While a teacher's average value added across their students might be high (low), it might be very low (high) for a subset of their students. From a theoretical perspective, this implies there are potential complementarities between teachers and student characteristics. From a policy standpoint, initiatives that seek to move high-value added teachers to low-performing schools, such as the Talent Transfer Initiative, may want to take into account how a teacher's value added varies by student performance. Also, alternative rankings of teachers can be generated from choosing different measurement criteria. For instance, evaluators might be interested in the median value added effect or including a measure of the variation in the teacher's value added into the ranking. Any additively separable teacher effect model loses information at these different points in the support of student ability.

In this chapter, we estimate several semiparametric models, among them a baseline model that uses linear regression on an additively separable teacher effect model included to represent the common estimation practices currently being used. Although we allow for slightly more flexibility in how lagged test score enters the production function than is typically employed (using a cubic expression), the baseline model is representative of the class of estimation models researchers are using. We also estimate various semiparametric additively separable econometric specifications and additively non-separable specifications. Among the

[1]The measurement unit for the dependent variable and teacher value added is standard deviations from the sample mean of the student's test score

non-separable specifications estimated is one estimated by linear regression that interacts the teacher indicator variables with the cubic in lagged test scores. This method is fast and easy to implement, but allows for student-teacher variation; the results from this chapter suggest this specification should be used in practice.

We estimate all of the models on a subsample of high-tenure elementary school teachers, and estimate the linear regression models–additively separable and additively non-separable– on the full sample of elementary school teachers with sufficient student-year observations for valid estimation of the teacher effect (over forty student-year observations). We find three major results, all of which support the use of the simple to implement and estimate additively non-separable linear regression model.

First, we find that the baseline regression results for the covariates are in line with other value added models being estimated. In particular, the marginal effects are very similar to those of Buddin (2011), who also uses data from Los Angeles Unified School District elementary schools. The methods and data sample used in this chapter are representative of the work currently being done, and the comparisons between the baseline model and the alternative models are representative of the choices facing researchers using other data.

The other two main results show that there are larger differences across the margin of the additivity assumption of the teacher effect than across the generalization of the estimation of the production function through more flexible semiparametric models.

We find that the additively separable models (including the commonly used baseline model) yield substantially different results than the additively non-separable models, evidence against using the baseline model in practice. For even the average marginal effect, where the models match the most, the most flexible additively non-separable model finds in the subsample that 18% of teachers would be reclassified out of the lowest or highest quintiles for Math test scores (18% and 27% for English) when using the baseline model, with greater movement for middle-ranked teachers. For the full sample of teachers, our most flexible model reclassifies 18% (27%) of the lowest-quintile of teachers and 18% (23%) of

the highest quintile teachers for Math (English). There is even greater movement for value added at 10th and 90th percentiles of student lagged performance.

Last, the additively non-separable linear regression model matches well with the more flexible Ichimura (1993) single index model. It also still provides for heterogeneous teacher effects, overcoming the two major drawbacks of the linear additively separable model. The teacher value added methods' relative matchings are evaluated by estimating the correlation between the teacher value added between the various models at differing student lagged test score values, and by comparing which quantile teachers are ranked in using various models at differing student lagged test score values.

The additively non-separable linear regression model nests the baseline model. The only difference is that the cubic in lagged test score is interacted with the teacher indicator variables, so that teacher effects can differ with different lagged student test score. An F-test on whether these interaction terms are jointly zero is strongly rejected in all our samples.

There are several caveats to our analysis that also pertain to value-added estimations more generally. These estimates can be imprecise. Specific to our analysis, semiparametric estimations can also exacerbate imprecision. We try to mitigate this problem by using a large data set, from the second largest school district in the United States. There are also potential biases from a number of other sources, including student-to-teacher assignments based on unobservables (Rothstein 2010). In practice, value-added estimations assume these biases are small. Similarly, we make the assumption that assignment is random conditional on some function of observables in order to focus on the potential bias from model misspecification.

The additively non-separable linear regression model is easy to implement and estimate, even using a large data set of teachers such as the Los Angeles Unified School District. The small change of interacting the teacher effect with the lagged student test scores captures most of the effects given in the most flexible model we estimate which doesn't restrict the interactions to be only between teacher effect and lagged student test score. The additively non-separable linear regression model also yields more convincing estimates of the average

marginal teacher effect by more flexibly estimating the underlying production function, and allows for capturing heterogeneous teacher effects by lagged student test score.

This study focuses on elementary school teachers for two main reasons: first, teacher assignment for different subjects' tests is more straightforward, and second, as Heckman and Masterov (2007) argue, the most important learning and separation of students happens early, and early interventions are the most effective.

The rest of the chapter proceeds as follows. Section 3.2 presents the empirical strategy and econometric models we will use. Section 3.3 explains the data from Los Angeles schools. Section 3.4 shows the results and discusses the implications, and Section 3.5 concludes.

3.2 Empirical Strategies

Estimation of teacher value added requires assumptions on the education production function. Similar to the model proposed by Todd and Wolpin (2003), we consider the following production function for student achievement

$$T_{it} = m_j \left[Z_{it}, \mu_{i0}, \eta_{ijt} \right]$$

T_{it} is student i's test score in academic year t. m_j is an unknown function of family and school inputs, and j is the teacher assignment. The inputs can broadly be separated into Z_{it}, the history of family and school inputs, μ_{i0}, a student's initial human capital endowment, and η_{ijt}, idiosyncratic shocks. Given family and school inputs, student ability, and random shocks, the unknown production function m_j for each teacher which translates all of these inputs to the output, the test score.

A common assumption in the literature and current practice, that Todd and Wolpin (2003) classify as the value added specification, is to use the lagged test score as a sufficient statistic for unobserved family and school inputs and mental endowment. We also make this

assumption.

For each of the models used in this chapter, we assume that the idiosyncratic shocks, η_{ijt}, are additively separable and orthogonal to all other covariates, as is also commonly done in the literature. Let X_{it} be the observable family and school characteristics, and $W_{it} = [T_{it-1}, X_{it}]$. We separate the econometric models we use into whether the teacher value added effect is additively separable or not. The assumption of an additively separable teacher effect simplifies the estimation process, but has dramatic effects on the results. This chapter establishes that there is too much information lost in the assumption of additive separability, and the results are biased. However, there is a simple additively non-separable model that can be estimated, the linear regression model.

With these assumptions, the additively non-separable models of the production function, which we call the AN models, are

$$T_{it} = m_j\left(W_{ijt}\right) + \eta_{it} \tag{AN}$$

The more restrictive additively separable models, or AS, are

$$T_{it} = m(W_{it}) + \sum_{j=1}^{J} d_{ijt}\psi_j + \eta_{it} \tag{AS}$$

$\eta_{it} = \sum_{j=1}^{J} d_{ijt}\eta_{ijt}$ is the idiosyncratic shocks, ψ_j is the additively separable contribution of teacher j, and d_{ijt} is an indicator variable for whether student i was taught by teacher j in year t. The assumption of additively separable teacher effects implies that the teacher's contribution towards a student's test outcomes does not vary by student characteristics. Teachers are restricted to have the same effect on high performing and low performing students, on male and female, on students enrolled in the free lunch program and those not, and any other factor. The advantage of additive separability is its easy estimation.

We estimate three additively separable models: linear regression, Single-Index Ichimura Model, and Artificial Neural Networks (ANN). We label these models respectively AS1, AS2, and AS3. AS1, the linear regression additively separable teacher effect model, is the baseline model and is representative of the models currently in use. AS2 and AS3 allow for more flexible estimation of the production function $m(\cdot)$ while retaining the simplicity of the additivity assumption. AS2 and AS3 are included as comparisons, to see if the problems inherent in the baseline model are because of poor approximation of the $m(\cdot)$ production function (given the additivity assumption) or result from the additivity of the teacher effect assumption.

We also test two additively non-separable teacher effect models: a linear regression model where the teacher effect is interacted with the student lagged test score variables, and an Ichimura single-index model. We label these models AN1 and AN2, respectively. AN2, the Ichimura model, is the most flexible, allowing for heterogeneous teacher effects to differ by all of the inputs, including lagged student test score. However, we show that AN1 and AN2 have close results, suggesting that the simpler and quicker linear regression model that interacts student lagged test score with the teacher indicator variables is a suitable econometric model choice in application.

We explain each estimation method in detail in the Appendix in Section C.1, including the specification of the estimator of the teacher value added effect for each method. An overall review is presented here.

3.2.1 AS Models: Additively Separable Teacher Effects

The additively separable teacher effect models assume that the production function is the same for every teacher (with only inputs differing), and the teacher effect is the same for any student the teacher instructs. The model is written as

$$T_{it} = m(W_{it}) + \sum_{j=1}^{J} d_{ijt}\psi_j + \eta_{it}$$

Where $\eta_{it} = \sum_{j=1}^{J} d_{ijt}\eta_{ijt}$ is the idiosyncratic shock, ψ_j is the teacher value added for teacher j, and d_{ijt} is an indicator variable for whether student i was taught by teacher j in year t. The three different models we estimate, AS1-AS3, change how $m(\cdot)$ is estimated econometrically.

AS1: Additively Separable Linear Regression

AS1, the linear regression additively separable teacher effects model is the econometric estimation method commonly in use by researchers. For this reason, AS1 will serve as the baseline model. To allow for non-linearity of student ability and heterogeneity captured in the lagged test score, we use a cubic in lagged student achievement. The other controls are assumed to enter linearly.

The intuition behind this model is that the controls on average have linear effects at the margin, and that the teacher effect can be reduced to a summary statistic (the average marginal effect), which will be captured by restricting teachers to have the same effect (ψ_j) on all students that they teach. The absolute teacher effect is not identified, because no students are observed without any teacher. Instead, we identify teacher effects relative to other teachers (the normalization used for all methods in this chapter). The comparison group is the average of all the teachers in the sample by subject, so that teachers in different schools can be compared. The interpretation of the teacher effect is how many standard deviations on average a given teacher contributes. Any normalization (to the mean, to a given baseline teacher, or any other) will create different absolute values, but the same rankings of teachers.

AS2: Additively Separable Single Index Ichimura Model

The linear regression model requires some degree of assumptions about the functional form of $m(\cdot)$. Other semiparametric additively separable teacher effect models are estimated to distinguish if the linear model fails specifically because of the additively separable assumption or because of an insufficiently flexible specification of interactions and higher order terms in the production function. We estimate an additively-separable teacher effects model using a single index Ichimura model, AS2. The model is based on the work of Ichimura (1993). AS2 allows for a much more flexible estimation of the production function by using kernel density estimation of the conditional expectation of test score on the weighted sum of the controls (the index).

AS3: Additively Separable Artificial Neural Networks

We use a model of Artificial Neural Networks (ANN), using the Ridgelet sieve, as presented in Chen, Racine, and Swanson (2001). Chen (2007) presents a review of these estimators and demonstrates that ANN performs particularly well among the class of semiparametric index model estimators as the number of indexing variables increases. The model is more flexible than the Ichimura model by allowing the weights to differ by layer, which, in essence, allows for a very flexible estimation of the production function $m(\cdot)$ where any controls can have arbitrary dependencies with other control variables for marginal effects.

3.2.2 AN Models: Additively Non-Separable Teacher Effects

The next models relax the additivity of the teacher value added effect assumption. Teacher effects are allowed to vary by different student characteristics, a more reasonable approximation of the production function that should give more consistent estimates, and retain the teacher effect heterogeneity. The econometric model is given by

$$T_{it} = m_j(W_{it}) + \eta_{ij}$$

We test two different additively non-separable models, linear regression and Ichimura index model (AN1 and AN2, respectively).

AN1: Additively Non-Separable Linear Regression

AN1, the additively non-separable teacher effect linear regression model includes interactions of the three coefficients on lagged test score (up to the cubic effect) with the teacher indicator variables, d_{ijt}.

This allows a teacher's effect to differ depending on the lagged test score, a summary statistic for the ability of the student, while retaining the fast estimation of OLS. There might be some teachers that are effective in teaching high performing students, while others might contribute more to struggling students. The goal of this model is to determine whether, if the additively separable models seem to not capture the effect well, this version of a linear regression model (AN1) is close to the Ichimura model (AN2) which allows the teacher effect to differ by other controls as well. If so, it offers a suitable, tractable model for large-sample estimation and routine use.

AN2: Additively Non-Separable Single Index Ichimura Model

The intuitive difference between AS2 and AN2 is the $m(\cdot)$ function is allowed to differ by teacher. The conditional expectation of the test score on the weighted sum of the controls is done for each teacher separately (although we require the control weights to be the same across teachers). This allows for a much more flexible estimation of the production function by teacher, but requires sufficient data and takes a lot longer to run. For this reason, we only estimate AN2 for the subsamples, when we have at least 200 student-year observations

for each teacher on which to base the estimate.

3.3 LAUSD Data

The data comes from Los Angeles Unified School District (LAUSD), spanning from academic years 2002-2003 to 2009-2010. The data set includes students and teachers from all primary and secondary schools in the district. We limit our attention to teachers in grades three to five, where the teacher assignment is simpler and student learning has greater long-run effects (see Heckman and Masterov 2007). Earlier grades are not included because exams first occur in the second grade so there are no lagged test scores for use as the summary statistic before third grade.

The analysis is performed on two separate groups. First, extensive testing uses all of the estimation methods on a subsample of teachers with a substantial amount of student-year observations (over 200 student-year observations across the sample), for Math and English standardized tests. These sample restrictions help the analysis in at least three ways. First, it yields a higher number of student-year observations per teacher on which the estimates of value added are based, which is especially important for the precision of semiparametric estimation techniques. Second, it requires less parameters to be estimated for each model, as there are less teachers whose effects need to be estimated.[2] Third, the smaller sample, and in particular the smaller amount of teachers, means a much faster estimation time. Given some of the methods are slow to estimate, even programmed using multiple parallel-processor methods, we leave the full battery of tests for this subsample of high tenure teachers.

The second group we perform our analysis on is for all 3-5 grade teachers with at least 40 student-year observations, what we call the full sample. The full sample expands the number of teachers we are analyzing from just under 60 teachers (slightly different depending on whether it is Math or English scores) to over 7000 teachers. We limit the number of

[2]The Ichimura models and ANN all take their parameter values as the result of a simplex search; a smaller parameter space to search on decreases the likelihood of getting caught at local minima.

student-year observations to be 40 to insure a minimum threshold of accuracy estimating the teacher effect for each teacher. Given the much larger size of these data and computation constraints, we only perform two estimation techniques for the full sample, the two linear regression models AS1 and AN1. Even with these faster methods, the large sample size requires making further divisions in the sample: the analysis is performed by grade and subject, centering each grade-by-subject's teacher value added effects.

The data contain many of the standard variables in district-level teacher value added analysis. The most important variables are the student standardized test scores, in the current period and the lagged value. We measure this in terms of standard deviations from the mean (z scores), by year and grade. To match the teacher data to the student data, we look at which teacher gave the students their Math (Reading) marks in a given year, and assign them to be the teacher responsible for their Math (Reading) standardized test scores.

We generate three other continuous variables: the fraction of students in their class that are receiving free/reduced price lunch, the number of students in their class, and the standard deviation of the lagged test z scores in the classroom. The last provides a measure of how different abilities are in the classroom.

We also include a set of indicator variables of student characteristics: a set of race indicator variables (Black, Hispanic, Asian, and other, with White as the baseline group), an indicator variable for male, for being in the gifted program, for being in the free lunch program, and for whether one of their parents has 12 or more years of education. We also include a control indicator variable for students who either declined to report their parents' education or for whom it was missing.

Tables C.1-C.4 contain the summary statistics for the two test subjects and the subsample and full sample.

Teachers in the subsample work with students that are substantially different than the students in the full sample. Teachers are not randomly kept in the system. On average, teachers in the subsample teach higher achieving students, with lower proportions in free

lunch program and a much higher proportion in the gifted program. They teach a smaller fraction of black students, teach larger classes, and work with students almost three times as likely to have a parent that finished high school. These differences likely will bias an estimation of the population parameters, as seen in comparing the coefficients from the full sample and subsamples. However, the subsample does not bias results for the population of similar teachers in similar classrooms, and our goal is to investigate how well different estimation techniques perform within a given sample. To the extent that our subsample does not have systematic differences across the econometric models in the different samples, estimation on the entire population or on the subsample results are informative for the effects of using the different econometric specifications.

The within teacher estimates in Tables C.1-C.4 are the average standard deviations of each teachers' students characteristics, while the between teacher estimates are the standard deviation of the means of the students characteristics by teacher. In almost every case, the between teacher variance is larger than the within teacher variance.[3] This is helpful for our analysis, implying comparability of teachers on shared supports. Also, estimators that allow for different teacher effects for different students (non-separable models) could have higher within teacher variance than between if teacher effects vary by student characteristics. This emphasizes the potential importance of using non-separable methods to evaluate teacher value added more robustly.

[3]The one exception to higher within teacher variation is the fraction of the class on the free lunch program. However, the within teacher estimate will be the same for each given class, implying zero variance for any teacher teaching one class only.

3.4 Results

3.4.1 Baseline Regression Results and Covariate Marginal Effects

Tables C.5 and C.6 present the results of the baseline linear regression estimations from model AS1 for Math and English. The first column shows the results for the high-tenure sample[4], and the remaining columns have the results for the full samples in grades three to five. The last row shows the Average Marginal Effect (AME) for lagged test scores.[5] The coefficients are similar across samples with the exception of the fraction of the class receiving free lunch and the individual-level free-lunch recipient indicator variable. The latter is positive and significant for the high tenure sample, as opposed to near zero in the full samples, while the coefficient on the free lunch indicator is more negative for the high-tenured sample compared to the full sample coefficients.

We can compare several coefficients to the model estimated by Buddin (2011). Looking at the Math results, the coefficients on class size are similar: Buddin's range from -.005 to -.002 compared to -.005 to -.003 in our results. Our results on gender switch signs for the third grade versus grades four and five, but the coefficients are both small. The coefficients on parents' education in Buddin's paper are similar to our results.

Broadly speaking, the baseline linear regression results suggest that the relative differences across the full sample and high-tenure samples are small, and the model and data reasonably approximates Buddin's results.

For the other econometric models, we generate the distributions of marginal effects by taking each student in the sample and estimating the marginal effect for the variable of interest with numerical derivatives for continuous variables and the discrete difference in predicted values for binary variables. The distributions of marginal effects are presented in Figures C.4-C.16. The distributions are consistent with estimates of the same variables in

[4]The sample that only includes teachers who have at least 200 student-year observations

[5]The AME differs from the marginal effect because of the included higher-order lagged score terms. The AME for a continuous variable is calculated as $\frac{1}{\sum_{i,t} 1} \sum_{i,t} \frac{\partial E[T_{it}|X_{ijt}]}{\partial T_{it-1}}$.

the previous literature, supporting the external validity of our sample and overall methodology and demonstrating the different levels of flexibility don't generally affect the estimated marginal effects much for these control variables.

3.4.2 Correlations Across Models

We estimate four values of the teacher value added effect for each teacher: the average marginal effect and the value added at the 10th, 50th, and 90th percentiles of the lagged student test score. For the percentile estimated teacher effects, the other control variables are evaluated at the modes of the binary variables and means of the other variables. Figure C.1 shows the distributions across teachers for these four measures by estimation method. The teacher effects are normalized at each different percentile, which is why they are centered at zero. Generally speaking, the distributions of effects are similar across models.

However, the rankings of teachers within the distribution vary by model. Tables C.7 and C.8 report the correlation between the estimated teacher effects for the various econometric models. The average marginal effect comparisons give insight into how well the econometric models match for the typical measurement of teacher value added. The correlations of the model at different percentiles of the lagged test score emphasize the shortcomings of using an additively separable teacher value added effect econometric model. For the AS models, the teacher effects will not vary by lagged student test score, and will be the same for the AME and the marginal effect at the 10th, 50th, and 90th percentile of lagged student test score.

Table C.7 shows very high correlations across all percentiles for the AS models, usually around .99. For comparison, Johnson, Lipscomb and Gill (2012) find correlations from .92 to .99 (an average of .973) across models that vary the number of covariates. Under the assumption of additively separable teacher effects, various estimation methods of the education production function do not affect the value added estimates very much. Researchers should not be concerned over using more flexible semiparametric models; if the additively

separable assumption is applied, the results will not vary by much.

However, the results may not be accurate, as suggested by how much they vary from the additively non-separable models, between which the correlations are much lower. For the median lagged math test score, correlations range from .92 to .97. However at the 10th and 90th percentiles, correlations are particularly low: from .52 to .77 and .72 to .75 for the 10th and 90th percentiles respectively. Correlations of effects across models for English scores range higher at the 10th and 90th percentiles. Overall, they range from .76 to .89. The correlations in AME across models are similar to the results at the median.

The correlation coefficients are higher within the additive separability assumptions than across it. For example, AN1 and AN2 have higher correlation coefficients with each other than with AS models. This is the case across all percentiles and subjects, with the exception of the 10th percentile for math scores, which has a correlation of .39 (interestingly, the correlation at the 10th percentile of English scores is .93). It is unclear to us why this correlation is so low in the high tenure sample where there should be sufficient density around the 10th percentile of scores to rely on higher-order lagged score terms. However, the overall results for Math and English suggest that using AN1, which is fast and easily implemented, will be more accurate than the additively separable models and preserves the teacher effect heterogeneity.

The Ichimura and Neural Network models show how more flexible models affect teacher effect evaluation. However, these models are often intractable and too slow for larger samples, and certainly would be difficult to apply to a data set of teachers as large as LAUSD. AN1, on the other hand, is estimable even on a larger data set of over 2500 teachers for each estimation. Given that AN1 approximates AN2 relatively well, we test the correlation of the teacher effects for the two linear regression models, AS1 and AN1, for the full sample. Doing so gives additional insight into whether the subsample results are informative for the full sample by contrasting the relative differences between AS1 and AN1 across both samples. With many more teachers involved, it also gives a wider view at the benefits of

allowing for different teacher effects by lagged test scores in using the non-separable model AN1. Table C.9 shows that the correlations between the non-separable OLS model and the baseline model are overall low, suggesting the importance of interacting teacher effect with the lagged test score. At the 10th percentile of Math and English scores, the correlations are .48 and .52, respectively. The correlation at the median fares better, .94 for math and .90 for English, but suffers again at 90th percentile (.57 for math and .31 for English). The AME also suggests that the separable OLS model misses important within-teacher heterogeneity in value added, with correlations of .81 and .65 for Math and English.

The correlations are so low between the separable and non-separable models away from the median of lagged student test score because of their failure to allow for effects to differ by student ability. We estimate the within teacher distribution of teacher effects in the subsample by evaluating the teacher effect for each teacher and each given student in the sample. Figure C.2 presents four examples from the English value added estimated teacher distributions for a single teacher. These plots are typical of the remaining distribution estimates. Mechanically for the separable models, the distribution collapses to a vertical line because there can be no within-teacher heterogeneity by construction. The non-separable models document significant variation of the teacher effect. These plots demonstrate three important lessons. First, teachers with above (below) average value added according to the baseline OLS model can be below (above) average for a significant fraction of students, and the reduction to the single point loses quite a bit of information. Second, the distributions of the two non-separable methods AN1 and AN2 tend to be remarkably close to each other, reinforcing the justification for using the easy to implement AN1. The four teachers presented are typical of the results for all teachers. Third, sometimes even the average marginal effect estimates from the additively separable model can be substantially off base.

Overall, the semiparametric non-separable models show that teacher effects can differ substantially by student characteristics. As a summary measure, Table C.15 documents that the within teacher variation in value added is significantly greater than the across-

teacher variation in value added. This heterogeneity is not permitted in typical value-added models. The correlation results show the differences are larger between the groups AS and AN than within them.

3.4.3 Teacher Reclassification by Model

One of the primary goals of evaluating teacher value added effects is to provide a ranking of the teachers. However, the biases that come from using an additively separable model–even at the AME, and even more so for groups of students with outlying lagged test scores–cause the teachers to be incorrectly ranked. This section demonstrates the extent to which the teacher sorting can be wrong.

We look at the distribution in changes in the percentile rankings of teachers between the baseline AS1 model and the more flexible AN1 model. Figure C.3 shows these distributions according to various percentiles of lagged test scores as well as for the AME. Overall ranking changes appear normally distributed around zero. Changes are greatest at the 10th and 90th percentiles for Math, reaching changes of around 40 percentile points in the tails. Changes at the median and AME tail off at around 20 percentile points. For English scores, there is significantly more variation in the size of the ranking changes, tailing off at around 60 percentile points at the 10th and 90th percentile of lagged scores and around 40 percentile points for the median and AME. The baseline AS1 model can widely misclassify teachers relative to the more flexible AN1 model that nests the baseline model within it.

The correlation tables previously examined show that all of the additively separable methods yield very similar results in the teacher effects, so we limit the attention of the additively separable models to the baseline linear regression model. We compare how well AS1 and AN1 match with the most flexible AN2 for the subsample. Similar to Johnson, Limpscomb and Gill (2012), we examine the policy relevance of our results by separating the distribution of teacher effects into quintiles, and comparing how closely two estimation methods' quintiles match. The results for the subsample are in Tables C.10 and C.11. The

elements of the tables are proportions in each row conditioned on the column; for example, in the Math subsample in Table C.10, 45.5 percent of those ranked in the 2nd quantile of teacher effects on students in the 10th percentile lagged test score by AN2 are also ranked in the second quantile by AN1.

Generally, a greater fraction of teachers are ranked differently from the preferred method by the baseline AS1 than by AN1. For English scores, across the AME, 10th, and 90th percentiles of lagged scores, AN1 generally places teachers in the same quintiles as the Ichimura model. For Math scores, at various points in the lagged-score distribution, AN1 performs better, for instance at the 90th percentile; however this is not the case at the 10th percentile. Also, there is a much larger drop-off in the matching for the 10th and 90th percentiles as opposed to the AME for the baseline model than for the additively non-separable linear regression model, showing the heightened limitations of additively separable models away from average lagged test score.

In Tables C.12 and C.13, we use the full sample to compare the agreement of AS1 and AN1. Similar to the subsamples, agreement between the models for Math in the highest and lowest quintiles is 82% for the AME, with greater movement in the middle quintiles. However in the full sample there is greater reclassification to quintiles different than the immediately adjacent quintile. 2% (4%) of teachers placed in the lowest (highest) quintiles by the baseline OLS models are placed in non-adjacent quintiles by the the non-separable model, as opposed to zero percent in the subsample. For English scores, the agreement is less and the magnitude of the reclassification is greater: 73% to 77% of teachers at the upper and lower quintiles agree for the AME, while lower at extreme quantiles. At the 10th and 90th percentile of lagged test score range from 32%-63%, which is very low.

The separable model AS1 is a special case of the non-separable model AN1 in the OLS case where the interaction terms between the teacher effect and the lagged test score variables have coefficients are equal to zero. We conduct two sets of hypothesis tests for the full samples with null hypotheses of no difference between the additively separable model and the additively

non-separable model for each grade and subject. The results are reported in Table C.14. The first performs separate hypothesis tests by teacher on the teacher's interacted terms. The fraction of teachers for whom the null hypothesis is rejected is reported. Between 2 to 64 percent of the teachers have different effects by different student ability than the average, according to this hypothesis test. The proportions are indicative of the proportion of teachers for which there is statistical evidence that their production functions (but not necessarily their value added) are differently shaped than the average production function. The second tests all of the interacted terms together. The p-values from these tests show the null hypotheses are strongly rejected in each instance (p-value$<$0.0000). This is strong evidence that the higher-order interaction terms are an important inclusion to the model.

3.5 Conclusion

Value added models are common in empirical investigations into teacher quality. Almost universally, the literature uses a linear OLS model with additively separable teacher value added effects to estimate the teacher value added. In this chapter, we test the additive separability assumption by using various semiparametric methods. Our results show through correlations of the estimated teacher effects at different percentiles and AME, through rankings at different quantiles, and through hypothesis testing between the two OLS models that there is a high degree of within-teacher heterogeneity in the teacher effect, and that not accounting for this through an additively non-separable model, such as one interacting the lagged student test score with teacher assignment, will bias the results.

Our research suggests that estimators that don't allow for the within-teacher value added heterogeneity will provide a very limited view into the contributions the various teachers make, and incorrectly rank the teachers even using the metric of average marginal effect. AN1, the OLS non-separable model, provides a model that is easily estimable, but still captures most of the non-linearities in the Ichimura non-separable model, and retains the teacher

effect heterogeneity that is strongly displayed in the estimation and needed to evaluate for policy involved with teacher placement or rankings for low or high performing students. Within-teacher variation is much higher than between teacher variation, so a metric that reduces the distribution to one point, such as AME, even if it could be done accurately, will not provide a full view on the teachers' rankings. Misrankings of the teachers have effects in many school districts using value added methods to assess teacher performance. In Tennessee, measures of the teacher value added account for 35% of the teachers' evaluations, for example (Butrymowicz and Garland 2012). New policy in New York will allow school districts to base up to 40 percent of their teacher evaluation on standardized test performance, of which half must be based off a very simple value added model (Santos and Hu 2012).

The full sample estimations demonstrate that our conclusions are not unique to high observation teachers. In the full sample, we perform the analysis for over 7000 teachers, much larger than most school districts in the United States. Researchers and administrators can easily apply our model to other school districts. We also suggest looking at a combination of measures of the teacher's value-added distribution in their overall ranking. Doing so will yield better evaluations of teachers' varied contributions to student achievement. It will also provide for better estimations of the teacher's contribution for low or high performing students, as is of interest in many practical applications.

C Appendix

C.1 Econometric Specifications

AS1: Additively Separable Linear Regression

The specification is given by

$$T_{it} = \beta_0 + \sum_{\ell=1}^{3} \lambda_\ell T_{it-1}^{\ell} + X_{it}\beta + \sum_{j=1}^{J} d_{ijt}\psi_j + \eta_{it}$$

The normalized teacher effect is then given by

$$\widehat{VAM}_{ijt}^{AS1} = \widehat{\psi}_j - \frac{\sum_{q,r,s} \widehat{\psi}_r}{\sum_{q,r,s} d_{qrs}}$$

AS2: Additively Separable Single Index Ichimura Model

The estimation of Ichimura's (1993) model from the assumption $E[\widetilde{\eta}_{it}|T_{it-1}, X_{it}] = 0$, the same assumption made for all of the models. This implies that

$$E\left[T_{it}|W_{it}\right] = E\left[m\left(W_{it}\beta\right)|W_{it}\right] + \sum_{j=1}^{J} d_{ijt}\psi_j$$

The left hand side is observed, and estimated using kernel density estimation:

$$E\left[T_{it}|W_{it}\right] = \frac{\sum_{q\neq i \cap s\neq t} T_{qs} K\left(\frac{W_{qs}\beta - W_{it}\beta}{h}\right)}{\sum_{q\neq i \cap s\neq t} K\left(\frac{W_{qs}\beta - W_{it}\beta)}{h}\right)}$$

As suggested by Li and Racine (2007), we jointly estimate the bandwidths and the index

coefficients by minimizing the sum of squared residuals (SSR). The SSR come from the demeaned data, which serves to eliminate all ψ_j from the estimation equation.

$$\widehat{\eta}_{it} = T_{it} - \frac{\sum_{q \neq i \cap s \neq t} T_{qs} K\left(\frac{W_{qs}\beta - W_{it}\beta}{h}\right)}{\sum_{q \neq i \cap s \neq t} K\left(\frac{W_{qs}\beta - W_{it}\beta)}{h}\right)} - \sum_{j=1}^{J} d_{ijt}\psi_j$$

$$\widetilde{\eta}_{ijt} = \widehat{\eta}_{it} - \frac{\sum_{i,t} d_{ijt}\widehat{\eta}_{it}}{\sum_{i,t} d_{ijt}}$$

$$SSR = \sum_{i,j,t} \widetilde{\eta}_{ijt}^2$$

We estimate the teacher effects after we have gotten estimates for β and h by backing out the averaged difference between the student's exam outcomes and the predictions from $\widehat{m}(\cdot)$:

$$\widehat{VAM}_{ijt}^{AS2} = \frac{\sum_{i,t} d_{ijt}(T_{it} - \widehat{m}(W_{it}\widehat{\beta}))}{\sum_{i,t} d_{ijt}} - \frac{\sum_{q,r,s} \frac{\sum_{i,t} d_{irt}(T_{it} - \widehat{m}(W_{it}\widehat{\beta}))}{\sum_{i,t} d_{irt}}}{\sum_{q,r,s} d_{qrs}}$$

AS3: Additively Separable Artificial Neural Networks

The model depends on how many hidden neuron layers, or number of sieve terms, are included. Let r_N be the number of hidden neuron layers. The basic form is given by

$$\widehat{m}(W_{it}) = \alpha_0 + \sum_{r=1}^{r_N} \frac{\alpha_r}{\sqrt{a_r}} \phi\left(\frac{W_{it}\beta_r - b_r}{a_r}\right)$$

where the ridge function $\phi(\cdot)$ is given by

$$\phi(\mu) = -0.8311297508e^{-2}(-105 + 105\mu^2 - 21\mu^4 + \mu^6)e^{-.5\mu^2}$$

Conditional on the number of hidden layers, the parameters are estimated using nonlinear

least squares on the demeaned data:

$$\widetilde{T}_{ijt} = T_{it} - \frac{\sum_{i,t} d_{ijt} T_{it}}{\sum_{i,t} d_{ijt}}$$

and similarly for $m(\cdot)$ and η. The number of hidden neuron layers is chosen by which number of hidden layer gives nonlinear least squares estimators with the smallest Bayesian Information Criterion, given by

$$BIC(r) = \ln(SSR_r) + (r * (k+3)) \ln(n)/n$$

where $r * (k+3)$ is the number of parameters estimated, and n is the sample size.

Although very technical, this method is in the end just a highly flexible estimator evaluated using nonlinear least squares. We use it because of its good characteristics for multi-dimensional covariate spaces (Chen 2007). Once the model is estimated, estimations of the teacher value added parameters can be backed out through estimating

$$\widehat{VAM}_{ijt}^{AS3} = \frac{\sum_{i,t} d_{ijt}(T_{it} - \widehat{m}(W_{it}))}{\sum_{i,t} d_{ijt}} - \frac{\sum_{q,r,s} \left[\frac{\sum_{i,t} d_{irt}(T_{it} - \widehat{m}(W_{it}))}{\sum_{i,t} d_{irt}} \right]}{\sum_{q,r,s} d_{qrs}}$$

AN1: Additively Non-Separable Linear Regression

AN1, the additively non-separable teacher effect linear regression model includes interactions of the three coefficients on lagged test score (up to the cubic effect) with the teacher effects d_{ijt}:

$$T_{it} = \beta_0 + \sum_{\ell=1}^{3} \lambda_\ell T_{it-1}^{\ell} + \sum_{j=1}^{J} \sum_{\ell=0}^{3} \gamma_{\ell j} T_{it-1}^{\ell} d_{ijt} + X_{it}\beta + \eta_{it}$$

The value added is given by

$$\widehat{VAM}_{ijt}^{AN1} = \sum_{\ell=0}^{3} \gamma_{\ell j} T_{it-1}^{\ell} - \sum_{q,r,s} \frac{\sum_{\ell=0}^{3} \gamma_{\ell r} T_{qs-1}^{\ell}}{\sum_{q,r,s} d_{qrs}}$$

AN2: Additively Non-Separable Single Index Ichimura Model

Ichimura's (1993) index model comes from the assumption $E[\eta_{it}|W_{it}] = 0$. This implies that

$$E[T_{it}|W_{it}) = E[m_j(W_{it}\beta)|W_{it}]$$

We solve jointly for β and the bandwidths (now one for each teacher) by minimizing the SSR, given by

$$\widehat{\eta}_{ijt} = T_{it} - \frac{\sum_{\ell \neq i \cap s \neq t} d_{\ell j s} T_{\ell s} K\left(\frac{W_{\ell s}\beta - W_{it}\beta}{h_j}\right)}{\sum_{\ell \neq i \cap s \neq t} d_{\ell j s} K\left(\frac{W_{\ell s}\beta - W_{it}\beta}{h_j}\right)}$$

$$SSR = \sum_{i,j,t} \widehat{\eta}_{ijt}^2$$

Again, the average teacher value added effects are normalized to average to zero. The teacher value added effect is given by

$$\widehat{VAM}_{ijt}^{AN2} = \widehat{m}_j(W_{it}\widehat{\beta}) - \frac{\sum_{q,r,s} \widehat{m}_r(W_{qs}\widehat{\beta})}{\sum_{q,r,s} d_{qrs}}$$

C.2 Tables

Table C.1: Summary Statistics, Math High Tenure Subsample; Number of Students=11,484, Number of Teachers=56

	Mean	Std. Dev.	Between Std.	Within Std.	Min	Max
T_t	0.969	1.007	0.504	0.872	-2.049	3.514
T_{t-1}	0.976	0.982	0.464	0.867	-2.182	3.607
Frac Free Lunch	0.450	0.333	0.320	0.088	0	1
Class Size	31.121	3.229	1.414	2.234	16	60
Std(T_{t-1})	63.516	12.112	6.475	9.813	33.319	105.392
Hispanic	0.061	0.240	0.066	0.213	0	1
Asian	0.157	0.363	0.135	0.313	0	1
Black	0.381	0.486	0.282	0.386	0	1
Other Race	0.043	0.204	0.029	0.186	0	1
Gifted Prog.	0.609	0.488	0.217	0.433	0	1
Male	0.501	0.500	0.032	0.500	0	1
Free Lunch Prog.	0.450	0.497	0.321	0.366	0	1
Parents College	0.426	0.494	0.239	0.426	0	1
Missing Parents Ed.	0.150	0.357	0.135	0.312	0	1

Table C.2: Summary Statistics, Math Full Sample; Number of Students=657,406, Number of Teachers=7,072

	Mean	Std. Dev.	Between Std.	Within Std.	Min	Max
T_t	0.090	0.997	0.554	0.809	-4.527	3.514
T_{t-1}	0.096	0.989	0.491	0.843	-4.655	3.607
Frac Free Lunch	0.784	0.264	0.219	0.103	0	1
Class Size	24.167	5.150	4.484	2.251	1	61
Std(T_{t-1})	61.000	14.419	9.290	10.086	0.000	163.342
Hispanic	0.089	0.284	0.166	0.183	0	1
Asian	0.046	0.209	0.097	0.113	0	1
Black	0.748	0.434	0.272	0.290	0	1
Other Race	0.029	0.167	0.053	0.100	0	1
Gifted Prog.	0.158	0.365	0.173	0.255	0	1
Male	0.490	0.500	0.052	0.500	0	1
Free Lunch Prog.	0.785	0.411	0.220	0.311	0	1
Parents College	0.141	0.348	0.164	0.264	0	1
Missing Parents Ed.	0.281	0.449	0.218	0.382	0	1

Table C.3: Summary Statistics, English High Tenure Subsample; Number of Students=11,685, Number of Teachers=57

	Mean	Std. Dev.	Between Std.	Within Std.	Min	Max
T_t	1.136	0.921	0.488	0.785	-2.205	5.384
T_{t-1}	1.073	0.970	0.505	0.829	-2.753	5.408
Frac Free Lunch	0.459	0.337	0.325	0.087	0	1
Class Size	31.059	3.155	1.375	2.233	17	58
Std(T_{t-1})	42.862	9.050	4.968	7.264	24.639	77.633
Hispanic	0.061	0.239	0.066	0.211	0	1
Asian	0.154	0.361	0.135	0.310	0	1
Black	0.389	0.488	0.287	0.385	0	1
Other Race	0.044	0.204	0.029	0.187	0	1
Gifted Prog.	0.605	0.489	0.219	0.433	0	1
Male	0.502	0.500	0.031	0.500	0	1
Free Lunch Prog.	0.458	0.498	0.326	0.362	0	1
Parents College	0.420	0.494	0.241	0.423	0	1
Missing Parents Ed.	0.150	0.357	0.134	0.313	0	1

Table C.4: Summary Statistics, English Full Sample; Number of Students=658,561, Number of Teachers=7,081

	Mean	Std. Dev.	Between Std.	Within Std.	Min	Max
T_t	0.090	0.959	0.557	0.767	-6.261	5.384
T_{t-1}	0.103	0.988	0.556	0.799	-6.132	5.408
Frac Free Lunch	0.784	0.264	0.219	0.103	0	1
Class Size	24.122	5.147	4.475	2.258	1	62
Std(T_{t-1})	41.276	10.080	6.623	6.899	0.000	140.007
Hispanic	0.089	0.285	0.166	0.183	0	1
Asian	0.046	0.209	0.097	0.113	0	1
Black	0.747	0.434	0.272	0.290	0	1
Other Race	0.029	0.167	0.053	0.100	0	1
Gifted Prog.	0.158	0.365	0.173	0.254	0	1
Male	0.490	0.500	0.052	0.500	0	1
Free Lunch Prog.	0.785	0.411	0.220	0.310	0	1
Parents College	0.141	0.348	0.164	0.264	0	1
Missing Parents Ed.	0.281	0.449	0.218	0.382	0	1

Table C.5: Math OLS Regression Results, Additively Separable Teacher Effect Model (Baseline)

	High Tenure Subsample	Full, Grade 3	Full, Grade 4	Full, Grade 5
T_{t-1}	0.742***	0.718***	0.636***	0.838***
	(0.0133)	(0.00237)	(0.00195)	(0.00223)
T_{t-1}^2	-0.0380***	-0.0275***	-0.0249***	0.0113***
	(0.0110)	(0.00134)	(0.00116)	(0.00144)
T_{t-1}^3	-0.0162***	-0.0241***	-0.0165***	-0.0362***
	(0.00328)	(0.000623)	(0.000533)	(0.000641)
Frac Free Lunch	0.201***	-0.0323***	-0.00555	-0.0577***
	(0.0606)	(0.0101)	(0.00906)	(0.0101)
Class Size	-0.0105***	-0.00343***	-0.00445***	-0.00397***
	(0.00204)	(0.000707)	(0.000411)	(0.000419)
Std(T_{t-1})	-0.00298***	-0.00115***	0.000116	-0.00120***
	(0.000584)	(0.000110)	(0.000101)	(0.000115)
Black	-0.148***	-0.173***	-0.129***	-0.115***
	(0.0276)	(0.00806)	(0.00634)	(0.00661)
Asian	0.128***	0.172***	0.138***	0.176***
	(0.0194)	(0.00858)	(0.00689)	(0.00718)
Hispanic	-0.0336*	-0.0669***	-0.0527***	-0.0475***
	(0.0180)	(0.00645)	(0.00510)	(0.00533)
Other Race	-0.0186	0.0411***	0.0445***	0.0746***
	(0.0308)	(0.0101)	(0.00793)	(0.00837)
In Gifted Program	0.314***	0.411***	0.263***	0.381***
	(0.0157)	(0.00535)	(0.00389)	(0.00389)
Male	0.0471***	0.0462***	-0.0320***	-0.0101***
	(0.0119)	(0.00272)	(0.00218)	(0.00228)
Free Lunch Prog.	-0.0829***	-0.0442***	-0.0271***	-0.0258***
	(0.0168)	(0.00441)	(0.00349)	(0.00367)
Parents College	0.0641***	0.0641***	0.0429***	0.0534***
	(0.0151)	(0.00477)	(0.00377)	(0.00393)
Missing Parent's Educ.	0.0199	-0.00117	-0.00347	-0.000249
	(0.0190)	(0.00348)	(0.00281)	(0.00292)
AME(T_{t-1})	0.5741***	0.6353***	0.5760***	0.7470***
	(0.0080)	(0.0017)	(0.0014)	(0.0017)
Observations	11,484	199,557	221,118	236,731
R-squared	0.604	0.508	0.546	0.594
Number of Teachers	56	2,623	2,681	2,621

Table C.6: English OLS Regression Results, Additively Separable Teacher Effect Model (Baseline)

	High Tenure Subsample	Full, Grade 3	Full, Grade 4	Full, Grade 5
T_{t-1}	0.694***	0.705***	0.716***	0.773***
	(0.0111)	(0.00188)	(0.00189)	(0.00168)
T_{t-1}^2	-0.00113	0.0136***	-0.0152***	0.0246***
	(0.00719)	(0.00127)	(0.000956)	(0.00125)
T_{t-1}^3	-0.0130***	-0.0161***	-0.0158***	-0.0212***
	(0.00167)	(0.000393)	(0.000370)	(0.000457)
Frac Free Lunch	0.0648	0.0945***	-0.115***	0.0695***
	(0.0522)	(0.00918)	(0.00875)	(0.00829)
Class Size	0.00122	-0.00684***	-0.00289***	0.00170***
	(0.00178)	(0.000645)	(0.000397)	(0.000343)
Std(T_{t-1})	-0.00207***	7.60e-05	0.000448***	-0.00166***
	(0.000677)	(0.000137)	(0.000148)	(0.000139)
Black	-0.0989***	-0.117***	-0.119***	-0.0853***
	(0.0236)	(0.00732)	(0.00614)	(0.00542)
Asian	0.0352**	0.00361	0.0663***	0.0494***
	(0.0165)	(0.00781)	(0.00667)	(0.00589)
Hispanic	-0.0575***	-0.0862***	-0.0515***	-0.0516***
	(0.0154)	(0.00587)	(0.00494)	(0.00438)
Other Race	-0.0165	-0.0493***	0.0125	-0.00184
	(0.0262)	(0.00922)	(0.00768)	(0.00688)
In Gifted Program	0.210***	0.263***	0.279***	0.178***
	(0.0135)	(0.00499)	(0.00375)	(0.00323)
Male	-0.0340***	-0.0496***	-0.0578***	-0.0361***
	(0.0102)	(0.00247)	(0.00211)	(0.00188)
Free Lunch Prog.	-0.0578***	-0.0567***	-0.0469***	-0.0313***
	(0.0144)	(0.00402)	(0.00339)	(0.00302)
Parents College	0.0623***	0.0602***	0.0554***	0.0385***
	(0.0129)	(0.00435)	(0.00365)	(0.00324)
Missing Parent's Educ.	0.0227	0.0107***	-0.00182	-0.00270
	(0.0161)	(0.00317)	(0.00272)	(0.00240)
AME(T_{t-1})	0.6101***	0.6566***	0.6727***	0.7275***
	(0.0073)	(0.0016)	(0.0014)	(0.0014)
Observations	11,685	200,586	221,535	236,440
R-squared	0.653	0.544	0.597	0.627
Number of Teachers	57	2,635	2,683	2,614

Table C.7: Math Subsample: Correlation of Teacher Effects Between Models, at 10th, 50th, and 90th Percentiles of Lagged Test Score and Average Marginal Effect

		OLS Sep.	OLS Non. Sep.	Ich. Sep.	Ich. Non. Sep.	ANN Sep.
10th Perc.	OLS Sep.	1.0000	0.5203	0.9913	0.7660	0.9970
	OLS Non. Sep.	0.5203	1.0000	0.5529	0.3940	0.5222
	Ich. Sep.	0.9913	0.5529	1.0000	0.7769	0.9850
	Ich. Non. Sep.	0.7660	0.3940	0.7769	1.0000	0.7530
	ANN Sep.	0.9970	0.5222	0.9850	0.7530	1.0000
50th Perc.	OLS Sep.	1.0000	0.9684	0.9913	0.9163	0.9970
	OLS Non. Sep.	0.9684	1.0000	0.9568	0.9232	0.9587
	Ich. Sep.	0.9913	0.9568	1.0000	0.9150	0.9850
	Ich. Non. Sep.	0.9163	0.9232	0.9150	1.0000	0.9049
	ANN Sep.	0.9970	0.9587	0.9850	0.9049	1.0000
90th Perc.	OLS Sep.	1.0000	0.7536	0.9913	0.7171	0.9970
	OLS Non. Sep.	0.7536	1.0000	0.7359	0.8749	0.7511
	Ich. Sep.	0.9913	0.7359	1.0000	0.7118	0.9850
	Ich. Non. Sep.	0.7171	0.8749	0.7118	1.0000	0.7163
	ANN Sep.	0.9970	0.7511	0.9850	0.7163	1.0000
AME	OLS Sep.	1.0000	0.9350	0.9913	0.9619	0.9970
	OLS Non. Sep.	0.9350	1.0000	0.9380	0.8754	0.9305
	Ich. Sep.	0.9913	0.9380	1.0000	0.9596	0.9850
	Ich. Non. Sep.	0.9619	0.8754	0.9596	1.0000	0.9550
	ANN Sep.	0.9970	0.9305	0.9850	0.9550	1.0000

Table C.8: English Subsample: Correlation of Teacher Effects Between Models, at 10th, 50th, and 90th Percentiles of Lagged Test Score and Average Marginal Effect

		OLS Sep.	OLS Non. Sep.	Ich. Sep.	Ich. Non. Sep.	ANN Sep.
10th Perc.	OLS Sep.	1.0000	0.8719	0.5468	0.8936	0.9754
	OLS Non. Sep.	0.8719	1.0000	0.4823	0.9300	0.8620
	Ich. Sep.	0.5468	0.4823	1.0000	0.5864	0.4354
	Ich. Non. Sep.	0.8936	0.9300	0.5864	1.0000	0.8484
	ANN Sep.	0.9754	0.8620	0.4354	0.8484	1.0000
50th Perc.	OLS Sep.	1.0000	0.9814	0.5468	0.9452	0.9754
	OLS Non. Sep.	0.9814	1.0000	0.5109	0.9528	0.9662
	Ich. Sep.	0.5468	0.5109	1.0000	0.5794	0.4354
	Ich. Non. Sep.	0.9452	0.9528	0.5794	1.0000	0.9063
	ANN Sep.	0.9754	0.9662	0.4354	0.9063	1.0000
90th Perc.	OLS Sep.	1.0000	0.8651	0.5468	0.7568	0.9754
	OLS Non. Sep.	0.8651	1.0000	0.5640	0.8900	0.8185
	Ich. Sep.	0.5468	0.5640	1.0000	0.5432	0.4354
	Ich. Non. Sep.	0.7568	0.8900	0.5432	1.0000	0.6778
	ANN Sep.	0.9754	0.8185	0.4354	0.6778	1.0000
AME	OLS Sep.	1.0000	0.9760	0.5468	0.9568	0.9754
	OLS Non. Sep.	0.9760	1.0000	0.5532	0.9677	0.9526
	Ich. Sep.	0.5468	0.5532	1.0000	0.6675	0.4354
	Ich. Non. Sep.	0.9568	0.9677	0.6675	1.0000	0.8976
	ANN Sep.	0.9754	0.9526	0.4354	0.8976	1.0000

Table C.9: Full Sample: Correlation of Teacher Effects Between OLS Additively Separable and Non-Separable Models

	Math	English
10th Perc.	0.4815	0.5167
50th Perc.	0.9426	0.8967
90th Perc.	0.5718	0.3090
AME	0.8112	0.6476

Table C.10: Math Subsample: Proportion of Teacher Effects Ranked in Quantiles by Different Percentiles of Lagged Test Score, OLS Separable and Ichimura Non-Separable vs. OLS Non-Separable

			Ich Non-Sep.				
			1st	2nd	3rd	4th	5th
AME	OLS Sep.	1st	0.818	0.091	0.083	0.000	0.000
		2nd	0.182	0.636	0.167	0.000	0.000
		3rd	0.000	0.273	0.750	0.000	0.000
		4th	0.000	0.000	0.000	0.636	0.364
		5th	0.000	0.000	0.000	0.364	0.636
	OLS Non Sep.	1st	0.909	0.091	0.000	0.000	0.000
		2nd	0.091	0.818	0.083	0.000	0.000
		3rd	0.000	0.091	0.833	0.091	0.000
		4th	0.000	0.000	0.083	0.727	0.182
		5th	0.000	0.000	0.000	0.182	0.818
10th Percentile	OLS Sep.	1st	0.727	0.273	0.000	0.000	0.000
		2nd	0.091	0.455	0.417	0.000	0.000
		3rd	0.182	0.091	0.333	0.364	0.091
		4th	0.000	0.000	0.250	0.364	0.364
		5th	0.000	0.182	0.000	0.273	0.545
	OLS Non Sep.	1st	0.545	0.273	0.000	0.091	0.091
		2nd	0.364	0.273	0.167	0.091	0.091
		3rd	0.091	0.273	0.500	0.091	0.091
		4th	0.000	0.182	0.333	0.364	0.091
		5th	0.000	0.000	0.000	0.364	0.636
90th Percentile	OLS Sep.	1st	0.455	0.273	0.250	0.000	0.000
		2nd	0.364	0.364	0.083	0.182	0.000
		3rd	0.182	0.273	0.250	0.273	0.091
		4th	0.000	0.091	0.333	0.273	0.273
		5th	0.000	0.000	0.083	0.273	0.636
	OLS Non Sep.	1st	0.727	0.273	0.000	0.000	0.000
		2nd	0.182	0.727	0.083	0.000	0.000
		3rd	0.000	0.000	0.833	0.182	0.000
		4th	0.091	0.000	0.083	0.636	0.182
		5th	0.000	0.000	0.000	0.182	0.818

Table C.11: English Subsample: Proportion of Teacher Effects Ranked in Quantiles by Different Percentiles of Lagged Test Score, OLS Separable and Ichimura Non-Separable vs. OLS Non-Separable

			Ich Non-Sep.				
			1st	2nd	3rd	4th	5th
AME	OLS Sep.	1st	0.818	0.167	0.000	0.000	0.000
		2nd	0.182	0.667	0.182	0.000	0.000
		3rd	0.000	0.167	0.727	0.083	0.000
		4th	0.000	0.000	0.091	0.667	0.273
		5th	0.000	0.000	0.000	0.250	0.727
	OLS Non Sep.	1st	0.727	0.250	0.000	0.000	0.000
		2nd	0.273	0.500	0.273	0.000	0.000
		3rd	0.000	0.250	0.545	0.167	0.000
		4th	0.000	0.000	0.182	0.667	0.182
		5th	0.000	0.000	0.000	0.167	0.818
10th Percentile	OLS Sep.	1st	0.727	0.167	0.091	0.000	0.000
		2nd	0.182	0.667	0.182	0.000	0.000
		3rd	0.091	0.167	0.455	0.250	0.000
		4th	0.000	0.000	0.273	0.417	0.364
		5th	0.000	0.000	0.000	0.333	0.636
	OLS Non Sep.	1st	0.909	0.000	0.000	0.083	0.000
		2nd	0.091	0.750	0.182	0.000	0.000
		3rd	0.000	0.250	0.727	0.000	0.000
		4th	0.000	0.000	0.091	0.750	0.182
		5th	0.000	0.000	0.000	0.167	0.818
90th Percentile	OLS Sep.	1st	0.636	0.167	0.091	0.000	0.091
		2nd	0.273	0.500	0.273	0.000	0.000
		3rd	0.091	0.333	0.364	0.167	0.000
		4th	0.000	0.000	0.091	0.667	0.273
		5th	0.000	0.000	0.182	0.167	0.636
	OLS Non Sep.	1st	0.818	0.167	0.000	0.000	0.000
		2nd	0.182	0.583	0.273	0.000	0.000
		3rd	0.000	0.250	0.545	0.083	0.091
		4th	0.000	0.000	0.182	0.667	0.182
		5th	0.000	0.000	0.000	0.250	0.727

Table C.12: Math Full Sample: Proportion of Teacher Effects Ranked in Quantiles by Different Percentiles of Lagged Test Score, OLS Separable vs. OLS Non-Separable

			OLS Non-Sep.				
			1st	2nd	3rd	4th	5th
AME	OLS Sep.	1st	0.820	0.130	0.024	0.015	0.011
		2nd	0.140	0.679	0.156	0.018	0.006
		3rd	0.017	0.151	0.652	0.162	0.017
		4th	0.008	0.026	0.138	0.679	0.149
		5th	0.015	0.013	0.030	0.126	0.817
10th Perc.	OLS Sep.	1st	0.654	0.200	0.066	0.042	0.038
		2nd	0.250	0.442	0.220	0.073	0.016
		3rd	0.064	0.247	0.404	0.232	0.054
		4th	0.021	0.086	0.241	0.435	0.216
		5th	0.010	0.026	0.069	0.219	0.676
90th Perc.	OLS Sep.	1st	0.648	0.210	0.088	0.036	0.018
		2nd	0.248	0.445	0.208	0.080	0.020
		3rd	0.048	0.216	0.396	0.275	0.066
		4th	0.020	0.076	0.215	0.420	0.269
		5th	0.036	0.054	0.095	0.189	0.627

Table C.13: English Full Sample: Proportion of Teacher Effects Ranked in Quantiles by Different Percentiles of Lagged Test Score, OLS Separable vs. OLS Non-Separable

			OLS Non-Sep.				
			1st	2nd	3rd	4th	5th
AME	OLS Sep.	1st	0.733	0.180	0.044	0.026	0.017
		2nd	0.190	0.567	0.194	0.038	0.011
		3rd	0.041	0.181	0.566	0.189	0.022
		4th	0.015	0.049	0.153	0.601	0.183
		5th	0.021	0.023	0.043	0.146	0.767
10th Perc.	OLS Sep.	1st	0.620	0.189	0.073	0.065	0.052
		2nd	0.257	0.416	0.218	0.077	0.033
		3rd	0.080	0.262	0.373	0.229	0.055
		4th	0.030	0.092	0.238	0.406	0.234
		5th	0.012	0.041	0.098	0.222	0.626
90th Perc.	OLS Sep.	1st	0.542	0.256	0.112	0.056	0.033
		2nd	0.253	0.353	0.262	0.101	0.031
		3rd	0.077	0.198	0.327	0.322	0.075
		4th	0.047	0.101	0.198	0.350	0.305
		5th	0.081	0.091	0.100	0.171	0.556

Table C.14: Hypothesis Testing Interaction Terms Between Lagged Student Test Score Cubic and Teacher Effect

	Full F-Test p-value			Proportion Rejecting Null for by-Teacher Joint Tests		
	Grade 3	Grade 4	Grade 5	Grade 3	Grade 4	Grade 5
Math	0.0000	0.0000	0.0000	0.0835	0.1146	0.6748
English	0.0000	0.0000	0.0000	0.2289	0.1249	0.0241

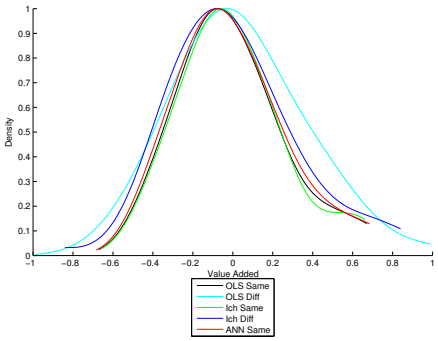
Table C.15: Teacher Value Added, Within vs. Between Standard Deviations

	Subsample		Full Sample	
	Between	Within	Between	Within
Math	0.2642	0.7618	0.3609	0.8257
English	0.2455	0.6758	0.3366	0.8725

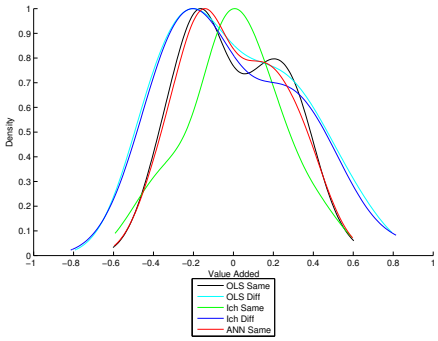
C.3 Figures

Figure C.1: Subsample: Kernel Estimates of Density of Teacher Effects by Different Lagged Student Score Percentiles

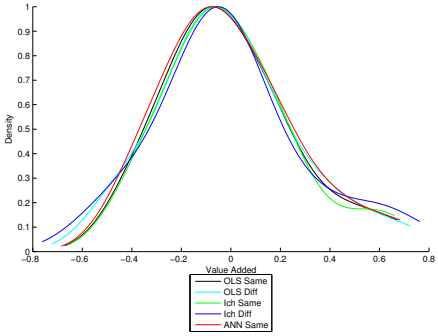
(a) Math: 10th Percentile



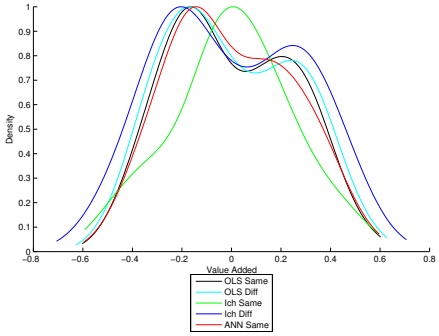
(b) English: 10th Percentile



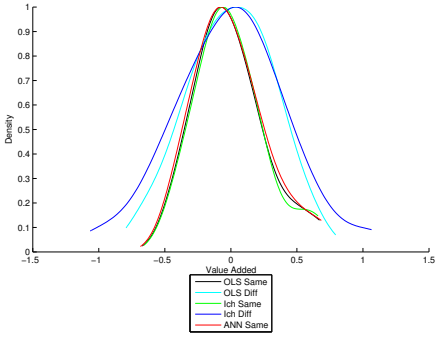
(c) Math: 50th Percentile



(d) English: 50th Percentile



(e) Math: 90th Percentile



(f) English: 90th Percentile

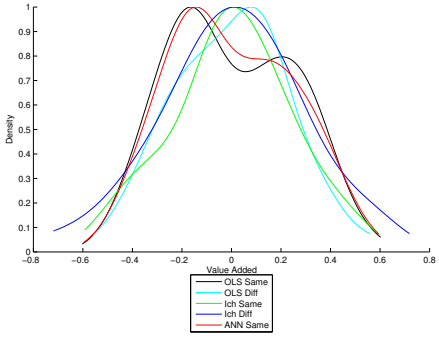


Figure C.2: Kernel Estimates of the Density of Within Teacher Effects for 4 Teachers, English Subsample

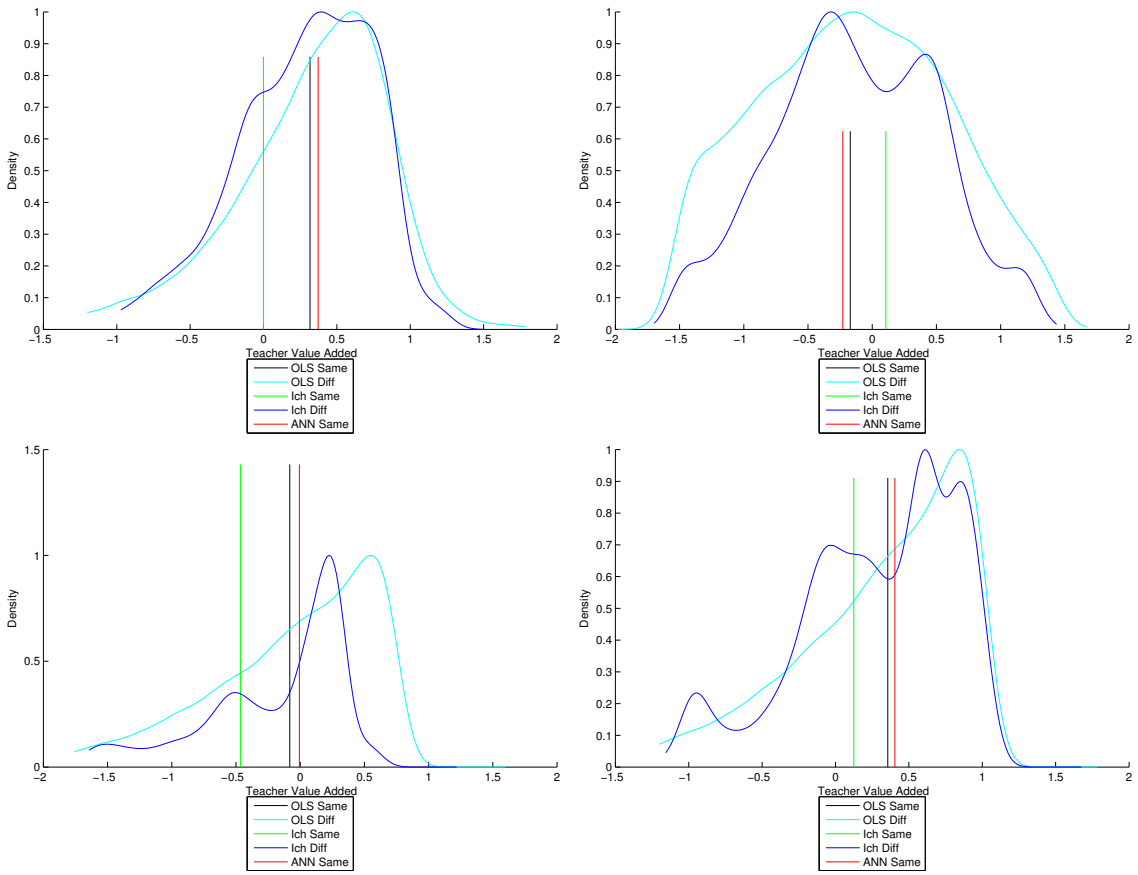
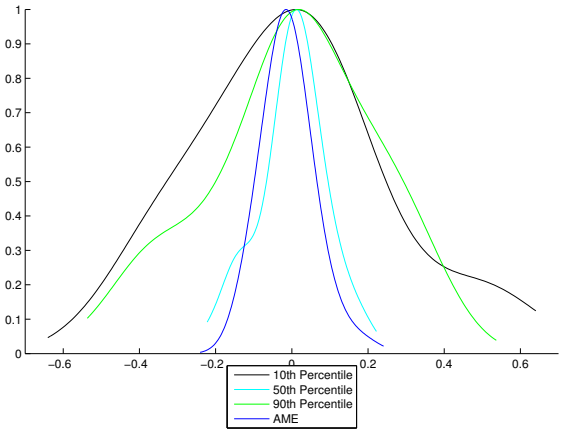
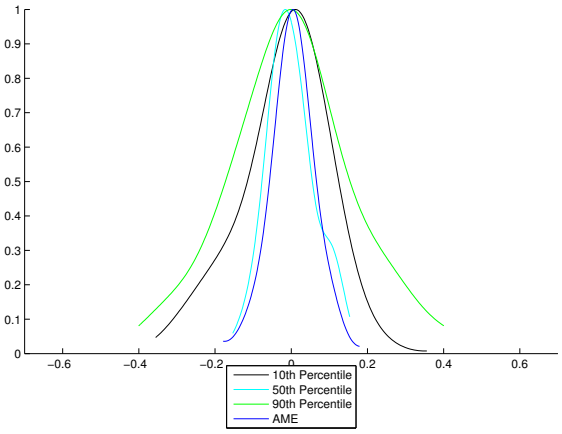


Figure C.3: Difference in Rankings by Econometric Models

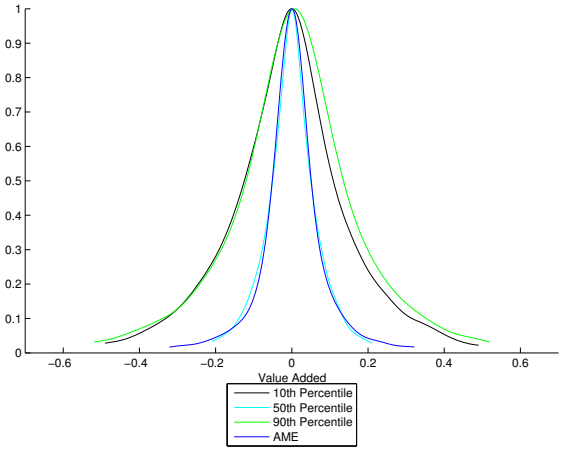
(a) Math Subsample



(b) English Subsample



(c) Math Full Sample



(d) English Full Sample

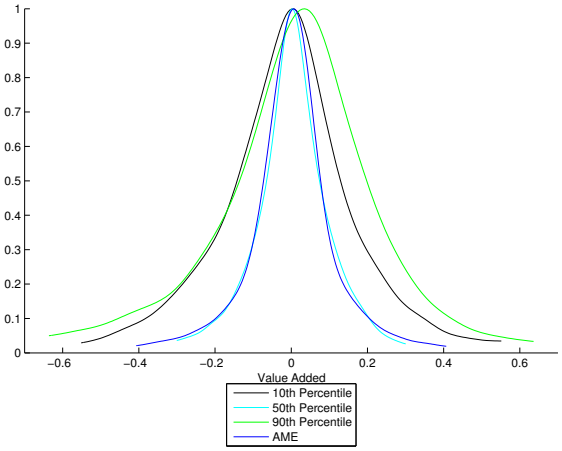
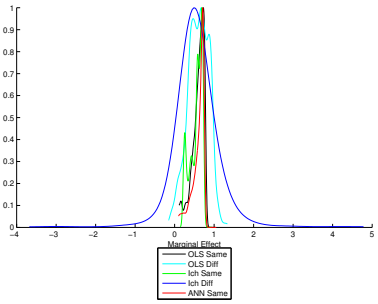


Figure C.4: Distribution of Estimated Marginal Effects: Lagged Test Score

(a) Math Subsample



(b) English Subsample

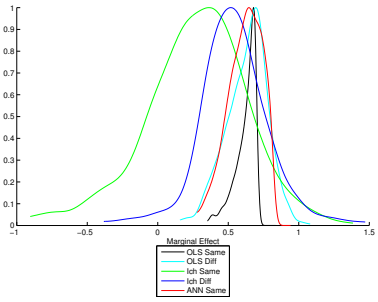


Figure C.5: Distribution of Estimated Marginal Effects: Fraction Free Lunch

(a) Math Subsample (b) English Subsample

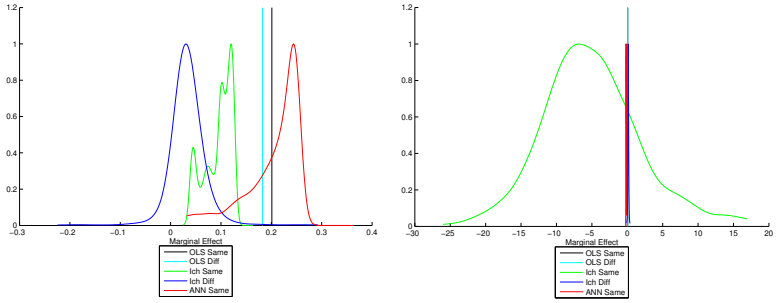


Figure C.6: Distribution of Estimated Marginal Effects: Class Size

(a) Math Subsample (b) English Subsample

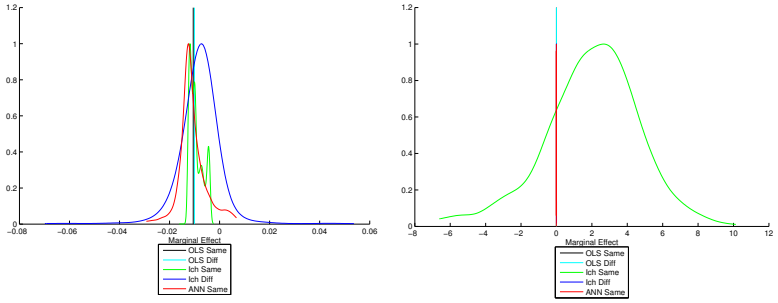


Figure C.7: Distribution of Estimated Marginal Effects: Standard Deviation of Class Lagged Test Score

(a) Math Subsample (b) English Subsample

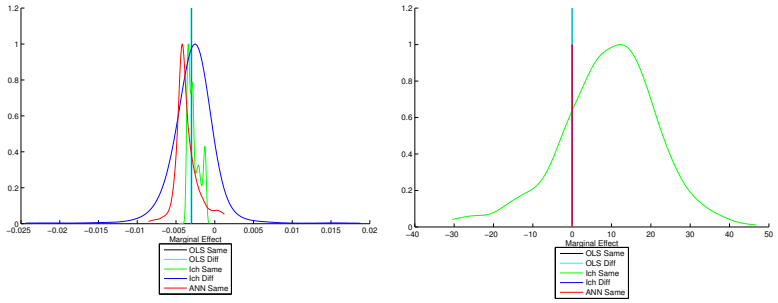


Figure C.8: Distribution of Estimated Marginal Effects: Hispanic

(a) Math Subsample (b) English Subsample

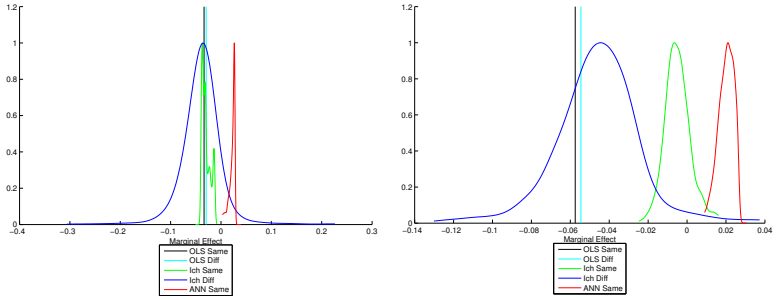


Figure C.9: Distribution of Estimated Marginal Effects: Asian

(a) Math Subsample (b) English Subsample

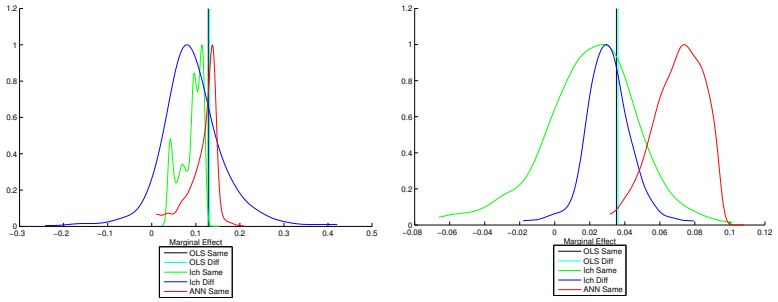


Figure C.10: Distribution of Estimated Marginal Effects: Black

(a) Math Subsample (b) English Subsample

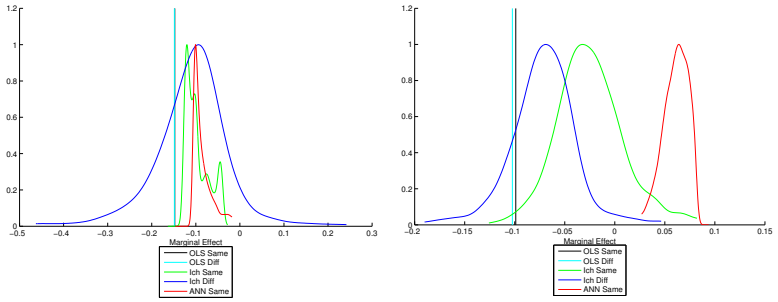


Figure C.11: Distribution of Estimated Marginal Effects: Other Race

(a) Math Subsample (b) English Subsample

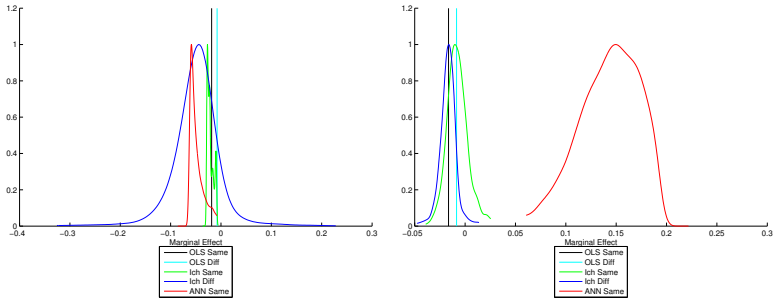


Figure C.12: Distribution of Estimated Marginal Effects: Male

(a) Math Subsample (b) English Subsample

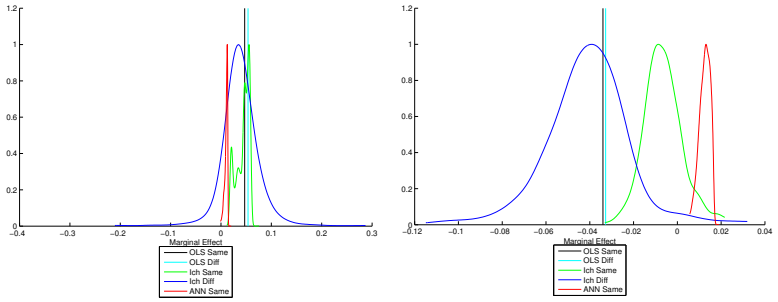


Figure C.13: Distribution of Estimated Marginal Effects: Participation in the Gifted Program

(a) Math Subsample (b) English Subsample

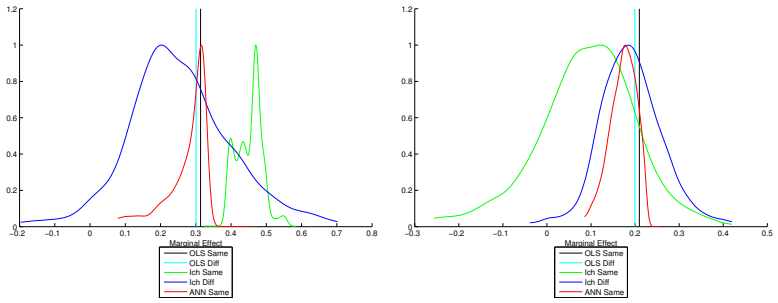


Figure C.14: Distribution of Estimated Marginal Effects: On Free Lunch Program

(a) Math Subsample (b) English Subsample

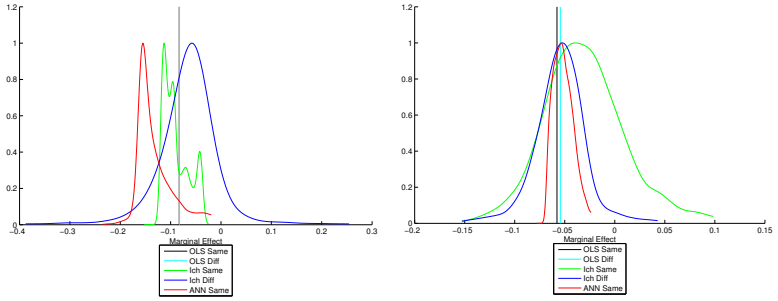


Figure C.15: Distribution of Estimated Marginal Effects: Parents Finished High School

(a) Math Subsample (b) English Subsample

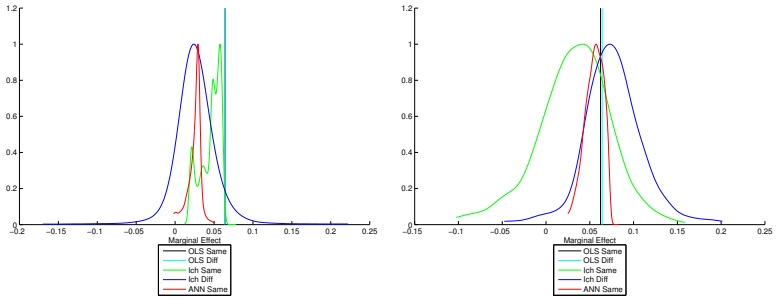
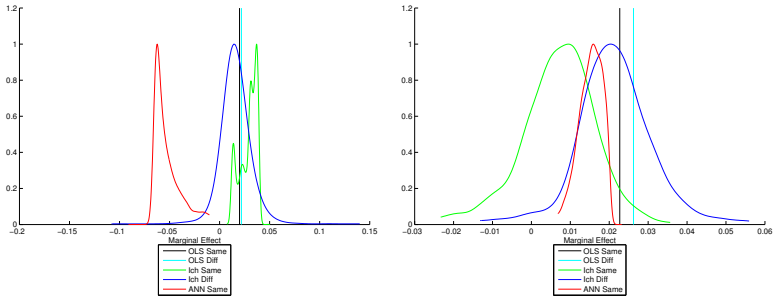


Figure C.16: Distribution of Estimated Marginal Effects: Missing Data on Parents' Education

(a) Math Subsample (b) English Subsample



C.4 References

Altonji, Joseph and Rosa L. Matzkin (2005) "Cross section and panel data estimators for nonseparable models with endogenous regressors," *Econometrica* 73, 10531102.

Ben-Porath, Yoram (1967) "The Production of Human Capital and the Lifecycle of Earnings," *Journal of Political Economy*, 75, 352-365.

Buddin, Richard. (2010) "How effective are Los Angeles elementary teachers and schools?" MPRA Paper No. 27366, posted 10. December 2010.

Buddin, Richard (2011) "Measuring teacher effectiveness at improving student achievement in Los Angeles elementary schools." unpublished paper, retrieved August 11, 1011 from `http://www.gse.uci.edu/person/buddin_r/measuring_teacher_effectiveness_march_2011.pdf`

Butrymowicz, S. and S. Garland (2012). "New York City Teacher Ratings: How Its Value-Added Model Compares To Other Districts." *The Hechinger Report,* posted 3/2/12. Retrieved 5/11/2012, from: `http://www.huffingtonpost.com/2012/03/02/new-york-city-teacher-rat_n_1316755.html`.

Chen, Xiaohong (2007). "Large Sample Sieve Estimatino of Semi-Nonparametric Models." chapter in J.J. Heckman & E.E. Leamer (ed.), 2007. *Handbook of Econometrics*, Elsevier, edition 1, volume 6, number 6b.

Chen,X., J. Racine, and N. Swanson (2001). "Semiparametric ARX Neural-Network Models with an Application to Forecasting Inflation." *IEEE Transactions on Neural Networks*,12(4),674-683.

Chen,X. and H. White (1999). "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators." *IEEE Transactions on Information Theory*,45(2),682-691.

Corcoran, Sean. P. (2010). "Can teachers be evaluated by their students test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice." Providence, Rhode Island: Annenberg Institute for School Reform at Brown University. Retrieved August 11, 2011, from: http://www.annenberginstitute.org/pdf/valueAddedReport.pdf.

Heckman, J. and Masterov, D. (2007). "The Productivity Argument for Investing in Small Children." *Applied Economic Perspectives and Policy*, 29(3).

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models", *Journal of Econometrics*, 58, 71-120.

Kane, Thomas J., and Douglas 0. Staiger, (2008) "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," National Bureau of Economic Research Working Paper No. 14607.

Li, Qi and Jeffrey Racine (2007). *Nonparametric Econometrics*. Princeton University Press.

Rockoff, Jonah E., and Cecilia Speroni. (2010). "Subjective and objective evaluations of teacher effectiveness." *American Economic Review* 100, 261-66.

Rothstein, Jesse (2010). "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics*, 125(1): 175-214.

Santos, F. and W. Hu (2012). "A Last-Minute Deal on Teacher Evaluations." 2/16/12. Retrieved 5/11/2012 from: http://www.nytimes.com/schoolbook/2012/02/16/as-deadline-nears-a-compromise-on-teacher-evaluations.

Todd, Petra, and Kenneth Wolpin (2003). "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 1 13(485): F3-F33.