UNIVERSITY OF CALIFORNIA,
IRVINE

Essays in Econometrics and Labor Economics

DISSERTATION

Submitted in Partial Satisfaction of the Requirements
for the degree of

DOCTOR OF PHILOSOPHY
in Economics

by

Kyle Colangelo

Dissertation Committee:
Assistant Professor Ying-Ying Lee (Chair)
Professor Matthew Harding
Associate Professor Yingying Dong

2021

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Professors Ying-Ying Lee, Yingying Dong and Matthew Harding for the extensive comments, feedback and assistance provided to me during the drafting of this dissertation.

Chapter 3 of the dissertation is derived from a joint work between myself and professor Lee. Professor Lee has provided me consent to include portions of our paper (including some figures and tables) from our unpublished paper. I would like to thank her for working with me on our paper "Double Debiased Machine Learning Nonparametric Inference with Continuous Treatment". Without her that chapter never would have become a reality. Ying-Ying Lee directed and supervised research which forms the basis for this thesis/dissertation.

I would also like to thank Gonzalo Dona for his hard work on our joint project on the effects of minimum wages on unemployment duration and re-employment outcomes. Chapter 2 of this dissertation is derived from "The Effects of Minimum Wages on Re-employment Outcomes," and the corresponding chapter in the dissertation"Empirical Studies on Policy Evaluation," with portions being a reprint of the material as it appears in these works. I have obtained the express consent of Gonzalo Dona to derive Chapter 2 of this dissertation from this material. Gonzalo Dona directed and supervised research which forms the basis for this thesis/dissertation.

I would also like to acknowledge the funding opportunities that I have received during my time here. During my time at UC Irvine, in addition to my baseline funding, I have received the Provost PhD Fellowship, the DTEI fellowship for summer of 2020, the Dissertation Completion Fellowship for Fall 2021, and a Research Assistantship under Professor Matthew Harding in summer of 2018.

Curriculum Vitae

Kyle Colangelo

2016 B.S. In Economics and B.S. in Mathematics, California State Polytechnic University, Pomona
Summa Cum Laude, Valedictorian

2017 M.A. In Economics, University of California, Irvine

2021 Ph.D. In Economics, University of California, Irvine

Fields of Study

Econometrics, Labor Economics

ABSTRACT OF THE DISSERTATION

Essays in Econometrics and Labor Economics

by

Kyle Colangelo

Doctor of Philosophy in Economics

University of California, Irvine, 2021

Professor Ying-Ying Lee, Chair

This dissertation is comprised of four chapters. Chapter 1 is a reprint of my job market paper "Estimation and Inference with Transfer Learning." In this chapter we propose and study machine learning algorithms which utilize "transfer learning," and their application to improving semi-parametric inference in cases of insufficient sample size. We consider multiple machine learning algorithms, including elastic net and deep feedforward neural networks with rectified linear unit (ReLU) activation functions, and allow for transfer through an ℓ_1 penalty term in the loss function which shrinks the estimates towards auxiliary estimates. Novel results on error bounds and convergence rates are established to justify the usage of these algorithms as nuisance function estimators for double machine learning. We evaluate the usage of these algorithms in conducting valid inference on treatment effects with Monte Carlo simulations and an empirical application on the Job Corps training program. Our numerical results show that transfer learning can substantially improve estimation in the presence of sizeable missing data.

Chapter two is comprised of an analysis of the effects of minimum wages on unemployment duration and re-employment outcomes. Using the Survey of Income and Program Participation, we build a sample of unemployment spells to study how the minimum wage affects several outcomes of the unemployed. We establish that the policy matters for the unemployed in two ways. A higher initial level of the minimum wage (at the start of a spell) leads workers to abandon their job search but has mostly null effects on other outcomes.

However, being unemployed at the time the minimum wage is raised is associated with longer spells, a higher rate of search quitting, and fewer working hours after re-employment.

Chapter three is derived from a joint work of mine with Ying-Ying Lee entitled "Double Machine Learning Nonparametric Inference with Continuous Treatment." In this chapter we numerically evaluate the effectiveness and characteristics of the double machine learning continuous treatment estimator discussed in Colangelo and Lee 2020. We conduct simulations using a variety of machine learning algorithms such as lasso, random forests and neural networks (and variations of these), and also provide an empriical application using data from the Job Corps program. The estimator is fairly robust to the choice of machine learning algorithm in the presence of cross-fitting, with all algorithms attaining near perfect coverage rates and unbiasedness. Some algorithms perform well even without cross-fitting, indicating that the procedure can be skipped in particular cases. The empirical results are similar regardless of which algorithm is used, and are consistent with previous research on the Job Corps program.

Chapter 4 discusses a new estimator for estimation and inference in long term panels with interactive effects. We modify the Common Correlated Effects (CCE) approach of Pesaran (2006) to produce a simpler and more computationally expedient estimator than the original CCE estimator. While CCE substitutes the factors with cross sectional means, the new method which we call Two Way CCE (TWCCE) substitutes out the individual specific factor loadings in addition to the factors themselves. This conveniently reduces the estimation problem to simple least squares without the need to estimate heterogeneous coefficients on each factor. We investigate the performance of TWCEE in comparison with the other most common factor model estimators such as the Interactive Effects Estimator (IFE) of Bai (2009) and the Augmented Mean Group Estimator (AMG) of Eberhardt and Teal (2010). We show that TWCEE has similar performance to the other methods, while also demonstrating the least computation time.

Chapter 1

Estimation and Inference with Transfer Learning

1.1 Introduction

Usage of machine learning algorithms for inferential purposes in economics has rapidly grown in recent years. Lasso for example is a common tool to perform variable selection in high dimensional settings. Neural networks can be used to conduct effectively non-parametric inference where traditional non-parametric methods are computationally infeasible. However, in some cases where one may wish to apply machine learning to conduct inference, the sample size may be insufficient for machine learning to be effective. The machine learning algorithms that are the primary focus of this paper are called "transfer learning" algorithms, which utilize information obtained from analyzing auxiliary outcome variables (usually an outcome variable which has more observations) to improve estimation in cases where the sample size would ordinarily be insufficient.

The development of Double Machine Learning (DML) in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins 2018 allows for the usage of any machine learning algorithm to conduct valid asymptotic inference, provided that the algorithms satisfy

certain conditions on their convergence rates. Thus, in order to justify the usage of a machine learning algorithm for the estimation of causal parameters in practice, their theoretical properties must first be established. There have been a number of recent studies in the econometrics literature which derive such results for a variety of machine learning algorithms. Farrell, Liang, and Misra 2021 and Athey, Tibshirani, Wager, et al. 2019 for example have derived results which justify the usage of deep neural networks and random forests for the estimation of causal parameters. However, while machine learning has been receiving much attention in the econometrics and causal inference literature, transfer learning algorithms specifically have received little to no attention in this regard.

The goal of the present work is to: (i) Establish key results for specific machine learning algorithms which utilize "transfer learning," in order to justify their usage in conducting valid asymptotic inference on parameters of interest (e.g. the Average Treatment Effect), and (ii) Investigate the effectiveness of these transfer learning algorithms through Monte Carlo simulations and an empirical application. We consider both linear models and deep feedforward neural networks with rectified linear unit (ReLU)[1] activation functions, allowing for transfer through an ℓ_1 penalty term in the loss function which shrinks the estimates towards auxiliary estimates.

In many empirical studies multiple outcomes of interest are analyzed. However, some outcomes may not be as well observed as others. For example, it may be common to fully observe an individual's employment status, but not their precise income. In these cases, the sample size of the poorly observed outcome may be insufficient to perform effective analysis using machine learning (neural networks for example can be very "data hungry", and lasso may perform poorly if the data is insufficiently sparse). To handle these types of scenarios for purposes of outcome prediction, "transfer learning" algorithms have been developed which incorporate auxiliary information from more well observed outcome variables to construct a

[1] A ReLU activation function takes the form $f(x) = max(0, x)$

model of the poorly observed outcome variable. One of the primary goals of transfer learning is to improve estimation when the samples size is insufficient.

Transfer learning, put simply, is the usage of knowledge gained from one task for the purpose of another (usually related) task. For example, suppose we have constructed an algorithm that can determine with a high degree of accuracy whether a photograph contains a cat. If we then wanted to construct an algorithm that determines if a photograph contains a dog, we may be able to better construct this new algorithm by leveraging knowledge gained from the previous algorithm[2]. In the context of causal inference, we might be able to better estimate the ATE for one outcome variable by utilizing information gained from studying some auxiliary outcome variable. In fact, this same concept is actually applied in the Seemingly Unrelated Regression model from Zellner 1962 for example, where efficiency can be gained by considering multiple outcomes jointly.

Transfer learning has enjoyed a great degree of success and growth in the world of prediction. It stands to reason that it may also be successful for the estimation of causal parameters. Andrew Ng, a leader in the development of machine learning and AI, said in his 2016 Neural Information Processing Systems (NIPS) talk that "transfer learning will be the next driver of machine learning success" (Ng 2016). Since then, the usage of transfer learning has grown, but not as substantially as predicted.

Transfer learning is a broad category which encompasses a wide range of different algorithms. In this paper we consider a specific class of transfer learning algorithms that utilize ℓ_1 penalization, to shrink parameter estimates towards auxiliary estimates. Suppose we have a nonparametric model for the relationship between a covariate set and outcome Y, such as

$$Y_i = f(\mathbf{x}_i, \mathbf{b}) + \epsilon_i$$

[2]In the computer science/machine learning literature, the task we wish to use auxiliary results to better accomplish is called the target task. The auxiliary task which we are using is called the source task.

where $\mathbf{x}_i$ is a set of covariates, $\mathbf{b}$ is the set of parameters (in lasso this would correspond to the regression coefficients, and for neural networks this would correspond to the weights in all layers), and ϵ_i is the error term. We would like to estimate the regression function f, using for example the deep ReLU neural networks in Farrell, Liang, and Misra 2021. The main idea of the algorithms considered in this paper is to shrink the estimates of the parameters $\mathbf{b}$ towards some auxiliary estimate $\tilde{\mathbf{b}}$. $\tilde{\mathbf{b}}$ can come for example by fitting the same model on some auxiliary outcome variable. If the auxiliary data or outcome variable has many observations that are not in our present sample, they can potentially provide much information to help us pin down f. All of the algorithms considered in this paper take variants of the following form:

$$\hat{\mathbf{b}} = \arg\min_{\mathbf{b}} \left\{ \ell(f, \mathbf{z}, \mathbf{b}) + \frac{\lambda}{n} \|\mathbf{b} - \tilde{\mathbf{b}}\|_1 \right\}$$

where ℓ is the associated loss function for the baseline machine learning algorithm, $\mathbf{z} = (\mathbf{x}, y)'$, λ is a tuning parameter and the second term shrinks the estimates towards the auxiliary estimates $\tilde{\mathbf{b}}$. Since for the purpose of these algorithms we only require knowledge of the auxiliary estimates ($\tilde{\mathbf{b}}$) without requiring direct access to auxiliary data, these algorithms belong to the class referred to as "Hypothesis Transfer Learning."

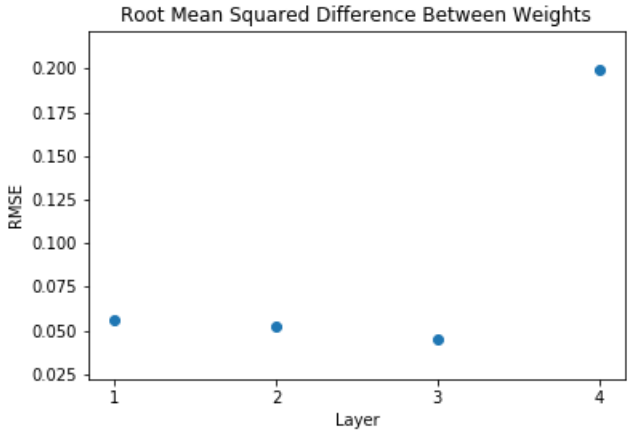
If the auxiliary estimates have sufficient similarity to the target parameters, then this additional shrinkage term will improve both finite and asymptotic performance. For example, a lasso estimator shrinks the parameters of a linear model arbitrarily towards zero, even though we can be fairly certain that the true coefficients are not all zero. If $\tilde{\mathbf{b}}$ is more representative of the true parameters than "all zeros" (e.g. it more correctly selects which coefficients are actually zero and which are not), then transfer learning in this setting should out-perform lasso, which we confirm both theoretically and through our numerical examples.

Another example in which transfer learning has a natural application is deep neural networks. It has been found that for deep neural networks the weights in the early layers tend to be very similar when fitting models with different outcomes on the same set of

covariates. This is hypothesized to be because the early layers are learning a basis from the covariates, and then the later layers taking this basis to predict the outcome. Thus when fitting the same neural network with different outcomes, there are substantial similarities in many of the weights, even when the outcomes have very little correlation.

To illustrate this, we conducted a brief simulation study (for complete details see Appendix B). In this simulation we fit the same neural network structure for 2 different outcomes of interest, where the outcomes of interest only had an approximate 40% correlation with each other. Figure 1.1 displays the root mean squared difference between the weights of the neural networks by layer. As can be seen, even though the auxiliary outcome has a low correlation with the outcome of interest, the weights are found to be fairly similar in the early layers, with most of the differences arising in the last layer. Since deep neural networks tend to have a large number of parameters and are data hungry, utilizing this relationship between different models can result in drastic improvement. Beyond the fact that transfer learning can improve performance for a fixed neural network architecture, an additional advantage is that it can allow us to select more complex neural networks than we otherwise could have.

Figure 1.1: RMSE of Neural Network Weight Differences



The rest of this paper is organized as follows: In Section 1.2 we produce a review of the relevant literature. In Section 1.3 we provide the general framework and main results. Section

1.4 discusses how the new algorithms can be used to estimate different parameters of interest. In Section 1.5 we perform simulations to test the effectiveness of the proposed algorithms in comparision with "non-transfer" machine learning algorithms. Section 1.6 provides an empirical illustration of the new methods, and Section 1.7 provides a summary/conclusion.

1.2 Literature Review

Recently there has been a large growth in the volume of research on machine learning algorithms in econometrics and causal inference. Most of this research has focused on variants of lasso, random forest and neural networks. This paper contributes to several areas in both the machine learning and econometrics literature. Our work builds primarily on the results of Farrell, Liang, and Misra 2021 (which will be referred to as FLM) and Takada and Fujisawa 2020 (which will be referred to as TF).

TF develop and evaluate the convergence properties of a lasso algorithm with transfer via an ℓ_1 penalty. However, looking exclusively at lasso is limiting since transfer is far more effective when used for neural networks. Besides TF there has been some other limited work on the properties of transfer learning, such as in Weiss, Khoshgoftaar, and Wang 2016 and Pan and Yang 2009. Kuzborskij and Orabona 2013, Du et al. 2016, Kuzborskij 2018, . Prior to TF there have been a number of studies evaluating the theoretical properties of regular lasso such as Belloni and Chernozhukov 2011, Belloni, Chernozhukov, and Hansen 2014, Bickel, Ritov, and Tsybakov 2009, Farrell 2015, and Chetverikov, Liao, and Chernozhukov 2021.

Unlike lasso, neural networks have been far less studied from a theoretical standpoint. FLM developed novel results on the error bounds and convergence properties of deep ReLU neural networks. As noted in FLM, previous research pinning down the properties of neural networks have generally focused on simple architectures, such as in Chen and White 1999, White 1989, and White et al. 1992, whereas FLM is one of the first papers to pin down the

theoretical properties of more complex deep neural networks. Some work has been done on establishing results for specific types of neural networks with transfer in Künzel et al. 2018, and Kuzborskij and Orabona 2017, but the neural networks we consider here take a different form. Furthermore, they do not consider the usage of their neural networks in combination with double machine learning.

There is a large volume of literature regarding transfer learning, of which we only mention a few. Most of these papers however do not develop clear results about the convergence properties of the algorithms. Pratt 1993 is arguably the first paper to describe transfer learning for neural networks. A number of review papers exist which give a comprehensive summary of transfer and multi-task learning algorithms, see Xu and Yang 2011, Zhuang et al. 2020, and Zhang and Yang 2017. These surveys also highlight the gains in the performance of algorithms which utilize transfer under the correct circumstances. The work here mainly expands upon a branch known as "hypothesis transfer learning", which is defined as only requiring access to the auxiliary estimates and not necessarily the auxiliary data itself.

Our work also builds upon the double machine learning (DML) estimator developed in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins 2018, which provided an estimator for causal parameters based on any machine learning algorithm which satisfied certain high level conditions. In our simulations we assess the usage of the transfer learning algorithms in the DML estimator for the ATE.

Although we do not build on it directly, it is also worth noting that there have been recent developments in assessing how "transferable" a neural network is to a new problem, such as in Rosenstein et al. 2005 and Yosinski et al. 2014, or more recently in Nguyen et al. 2020 which proposes a new measure of how transferable a neural network is to a new problem. One open problem which we are unable to address in the current work is how best to determine whether transfer will be effective in a particular empirical study.

To the best of my knowledge this is the first paper to develop error bounds and convergence results for elastic net and to evaluate the performance of deep ReLU neural networks with transfer via ℓ_1 penalization.

1.3 Theory

In this section we provide the general framework which we will be working with, as well as key results for a baseline linear model, elastic net, and deep neural network algorithms.

1.3.1 General Framework

Consider one outcome of interest which we will denote Y, and $p-$dimensional covariate set X, with n observations. Suppose we have a single auxiliary outcome W which we observe alongside X on an auxiliary dataset with n_1 observations[3]. We can model these outcomes non-parametrically:

$$Y_i = f(X_i) + \varepsilon_i \quad i = 1, ..., n$$
$$W_i = g(X_i) + \nu_i \quad i = 1, ..., n_1$$

where $f(X_i)$ and $g(X_i)$ can be estimated via any machine learning method, with f being the primary function of interest. The above nests the special case of the partial linear model in which Y (and W) can be expressed as $Y_i = \beta X_1 + f(X_i) + \varepsilon_i$, and also the fully linear model. If we wished, we could apply a machine learning algorithm (neural network for example) to each of these outcomes separately. In the case in which all outcomes are well observed for a large sample size, this may work well. However, in many situations we may be able to improve performance in the estimation of f by utilizing information gained by estimating g.

[3]The "auxiliary" data may be from the same dataset, for example we might observe W for a different (greater) subset of the data than Y. This is the situation we examine in the simulation and empirical sections. In general we only need access to auxiliary estimates to be able to implement the transfer learning algorithms considered in this paper, which could theoretically be made up out of thin air.

For example, in the case where an outcome has insufficient observations it may be difficult to justify the the usage of "traditional" machine learning methods. For example, elastic net (or specifically lasso or ridge regression as special cases) typically requires sparsity restrictions, that is restrictions on the number of non-zero β_j^*, in order to have desirable properties. In a high dimensional setting where $p \gg n$, these restrictions are less likely to be satisfied. With transfer learning however, as shown in Takada and Fujisawa 2020, these sparsity restrictions can be relaxed in the presence of auxiliary information. With respect to neural networks, they require the specification of an architecture which may be restricted depending on the sample size and number of covariates. This selection of architecture can be less restricted with the use of transfer learning. Furthermore, even if the above are not concerns, the estimation of f may not be of desired efficiency.

The algorithms considered in this paper generally work by first obtaining an auxiliary estimate $\tilde{f} = \hat{g}$, and shrinking $\hat{f}$ towards $\tilde{f}$. We proceed by deriving error bounds for linear regression, elastic net and deep neural networks with transfer.

Remark 1.1 In machine learning applications where transfer learning is used, it is typical that $n_1 \gg n$, where n_1 is the sample size of the more well observed outcome being utilized for transfer. For practical purposes it may be very important to consider the difference between n_1 and n. For example, the size of n_1 relative to n may certainly affect the practical choice of neural network architecture. However, for the theoretical purposes of this paper we make no formal assumptions on the relationship between n_1 and n.

Remark 1.2 It is typcal in practice to utilize a "pre-trained" neural network, which is fit on a completely separate dataset than the dataset at hand by someone other than the reseracher. That is, there exists already "pre-trained" neural networks for a variety of circumstances that can be used for transfer learning depending on the problem. In the above setup we speak as if g is estimated directly by the researcher, but this need not be the case. To implement the transfer algorithm we simply need access to some $\tilde{f} = \hat{g}$ which can come from any number of sources.

1.3.2 Elastic Net

Transfer can be used in conjuction with Lasso, Ridge Regression, Elastic Net and other similar estimators. Takada and Fujisawa 2020 derive results for lasso with transfer via and ℓ_1 penalty. In this section we extend the results of Takada and Fujisawa 2020 to the case of elastic net. The new estimator we term "Transfer Elastic Net (TEN)". Similar algorithms were also considered in the "prior lasso" of Jiang, He, and Zhang 2016 and the "Trans Lasso" in Li, Cai, and Li 2020.

We begin by assuming that the underlying data generating process is linear:

$$Y = X\boldsymbol{\beta}^* + \varepsilon \tag{1.1}$$

We wish to estimate the "true" parameter $\boldsymbol{\beta}^*$, which we assume has cardinality $s = \|\boldsymbol{\beta}^*\|_0$ non-zero elements, with corresponding support S. The TEN estimator can be defined by following minimizer:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} L(\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}) = \arg\min_{\beta} \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1(\alpha(\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2) + (1-\alpha)\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_1). \tag{1.2}$$

Where $0 \leq \alpha \leq 1$, λ_1 and λ_2 are tuning parameters, and $\tilde{\boldsymbol{\beta}}$ are the auxiliary estimates of $\boldsymbol{\beta}^*$, which can come for example by fitting an elastic net for a regression of a different outcome variable on the same set of covariates.[4] Note that the tuning parameters may be allowed to vary with n, and so it may be more appropriate to denote them $\alpha_n, \lambda_{1,n}$, and $\lambda_{2,n}$ but for notational simplicity we leave out the n subscript. α regulates how much weight is put on the baseline elastic net penalty terms, and how much is placed on the new transfer penalty term.

[4]The first step estimation need not be elastic net, but can be any algorithm which has a similiar structure (linear regression, lasso, ridge regression, etc).

We can note that equation 1.2 is similar to the normal elastic net, with the exception of the additional penalty term which shrinks estimates towards $\tilde{\boldsymbol{\beta}}$. This estimator reduces to the vanilla elastic net if we choose $\alpha = 1$. If we choose $\alpha = 0$ then we only penalize for distance from $\tilde{\boldsymbol{\beta}}$ with no other regularization. If we choose $\lambda_1 = 0$ then this reduces to ordinary least squares. We refer to the case where $\alpha = 0$ as "Transfer Ordinary Least Squares (TOLS)."

Note that if we select $\alpha = 0$, then the estimator looks nearly identical to the lasso estimator, but rather than shrinking the parameters towards 0, we shrink them towards $\tilde{\boldsymbol{\beta}}$. If we choose $\tilde{\boldsymbol{\beta}} = 0$, then it would reduce to lasso exactly. In this sense, $\alpha = 0$ can be thought of as a generalization of lasso in which we can choose what vector to shrink towards, rather than arbitrarily shrinking towards 0. If the auxiliary estimates are more representative of the truth than "all zeros," we would intuitively expect the transfer algorithm to outperform the baseline lasso/elastic net. In the case where $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$, it would be guaranteed to be optimal to choose $\alpha = 0$.

In order to derive the error bound of the TEN algorithm, we require the following main assumptions:

Assumption 1.1 *$\boldsymbol{\varepsilon}$ is a sub-gaussian random vector with variance $\sigma^2 I$*

Assumption 1.2 *For a set $\mathcal{B} := \{v \in \mathbb{R}^p \colon (\alpha - c)\|v_{S^c}\|_1 + (1-\alpha)\|v - \Delta\|_1 \leq (\alpha + c)\|v_S\|_1 + \lambda_2 \alpha \|v\|_2^2 + (1-\alpha)\|\Delta\|_1\}$, we have*

$$\phi = \phi(\mathcal{B}) = \inf_{v \in \mathcal{B}} \frac{v' \frac{1}{n} X' X v}{\|v\|_2^2} > 0$$

Assumption 1.3 (Boundedness)

(i) For some constant M, $\beta_j^* \leq M$ for all j.

(ii) $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^*$.

Assumption 1.1 requires that tails of the distribution of the errors are at least as thin as those of a normal distribution. This includes Gaussian errors but also allows for some other cases. A more restrictive but possibly not any less realistic assumption would be to simply assume that errors are Gaussian. Assumption 1.2 is analogous to the main assumption of Takada and Fujisawa 2020, and is key in constructing the error bound. Assumption 1.3, is an additional assumption not made in TF. This is required in order to bound the additional ℓ_2 penalty term, as the power on this term prevents an application of the triangle inequality. Although we have to make this additional assumption, we believe that it is reasonable to assume the coefficients are bounded and that estimates will converge as sample size goes to infinity. We do not delve into any results for the consistency of the new estimator and leave that for future work. We now present the main result for elastic net.

Theorem 1.1 (TEN Error Bound) *Given assumptions 1.1-1.3 hold, then with probability approaching* 1*, for some constants* $c > 0, M > 0$

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{[\lambda_1(\alpha + c)\sqrt{s} + 2\lambda_1\lambda_2\alpha\sqrt{s}M]}{(\phi - 2\lambda_1\lambda_2\alpha)} + \sqrt{\frac{2\lambda_1(1-\alpha)}{\phi - 2\lambda_1\lambda_2\alpha}|\Delta\|_1 + \frac{[\lambda_1(\alpha + c)\sqrt{s} + 2\lambda_1\lambda_2\alpha\sqrt{s}M]^2}{(\phi - 2\lambda_1\lambda_2\alpha)^2}}$$

Where $\Delta = \boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}$[5]*, and* s *is the cardinality of* $\boldsymbol{\beta}^*$*.*

In the case where $\lambda_2 = 0$, we recover the result of TF exactly. To gain some intuition on this result, it can help to consider the infeasible case in which $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$, which would be the best case scenario for the neural network estimator with transfer. In this case, $\Delta = 0$ and the error bound is strictly less than the case when $\|\Delta\|_1 \neq 0$. This makes intuitive sense because in this case we would be shrinking the estimates towards the true values. In principal similar rates may be able to be found for different penalties (such as some arbitrary ℓ_q penalty). As the penalties λ_1 and λ_2 approach 0 as sample size increases to infinity, the error bound

[5]Again, it is recommended that the data is normalized as is common with lasso so that the scale of $\boldsymbol{\beta}^*$ and $\tilde{\boldsymbol{\beta}}$ are comparable. Alternatively we could incorporate scaling directly into the estimation procedure.

approaches 0. However, since we are required to include the constant M into the result, we can not make definitive statements about the relationship of this bound and those for elastic net. However, if $\lambda_2 = 0$, the terms involving M reduce to 0, and we retain the analysis of TF, that if $\|\Delta\|_1 = 0$ this is strictly less than the error bound for lasso.

We can further note that the bound depends on p and s. The bound will be strictly less for smaller values of p and s, which makes sense intuitively. If the number of covariates is reduced, our estimator should be more readily able to pick up the true data generating process. If the true data generating process is simple (e.g. 1 variable instead of 1000), then we would expect our estimator to be able to estimate it more precisely. On the other hand, if the data is insufficiently sparse, the error bound will increase.

A natural extension of our results is to consider the case of cross-validation as in Chetverikov, Liao, and Chernozhukov 2021, which derives convergence properties for the LASSO under cross-validated choice of λ. It may also be of interest to implement this transfer algorithm for the purpose of variable selection. Similar to post-lasso, we can use the above procedure to select variables to include and then use OLS on these variables. Other work such as Belloni and Chernozhukov 2011 have derived convergence results for variable selection. We leave this to future work.

1.3.3 Deep Neural Networks

Now we turn to the application of transfer with ℓ_1 penalization for deep ReLU neural networks. In this section we expand upon the feedforward deep neural networks described in Farrell, Liang, and Misra 2021 (henceforth to be referred to as FLM). We construct the estimator in the same exact fashion, but we add an ℓ_1 penalty. As discussed in that paper, performing regularization in general for the estimation is not necessary. However, we find that applying transfer via regularization can be quite effective at improving finite sample estimates (see Sections 1.5 and 1.6).

We begin by first briefly describing the neural networks which we will be working with. We would like to estimate an optimal function f^* where,

$$f^* = \arg\min_f E[\ell(f, z)]$$

Where any loss function $\ell(f, z)$ which satisfies the following assumption will work as in FLM.

Assumption 1.4 (loss function)

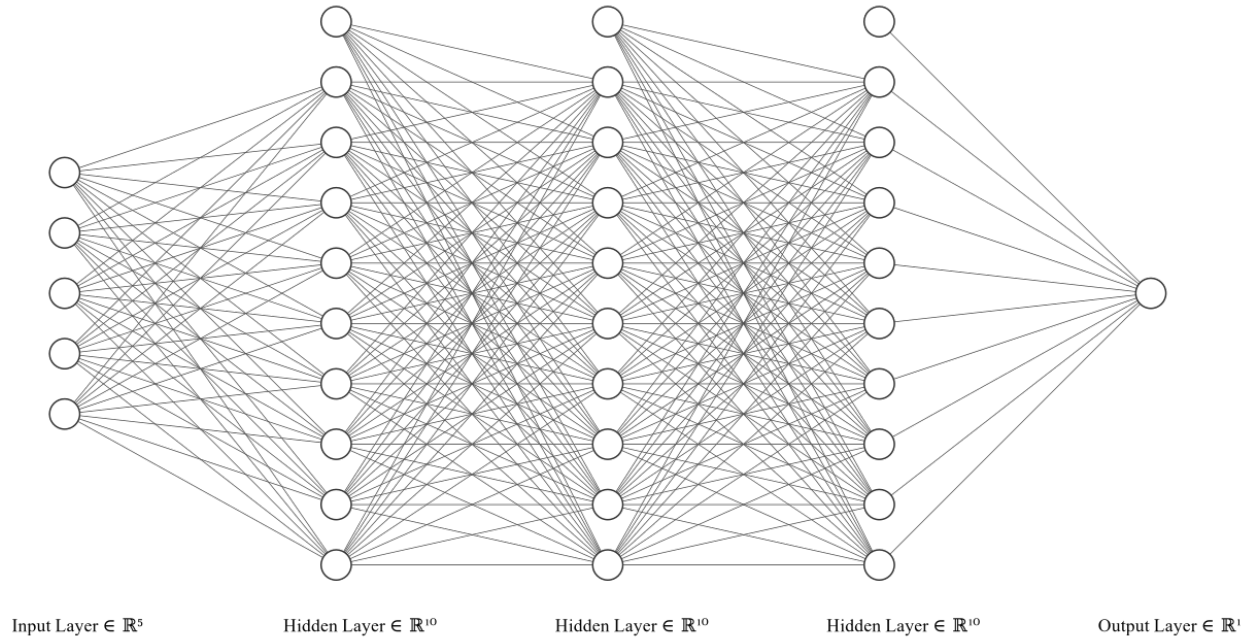
$$(i) \quad |\ell(f, z) - \ell(g, z)| \leq C_\ell |f(x) - g(x)|$$

$$(ii) \quad c_1 E[(f - f^*)^2] \leq E[\ell(f, z)] - E[\ell(f^*, z)] \leq c2E[(f - f^*)^2]$$

As shown in FLM, these conditions are satisfied for both mean squared error loss, and logistic loss. The choice is irrelevant for the derivation of convergence rate, but for the purpose of our numerical exercises we use a mean squared loss.

A user will specify a neural network architecture $\mathcal{F}_{DNN}$, which may or may not contain f^*. The choice of $\mathcal{F}_{DNN}$ will determine how well our network can approximate f^*. An example of a specific neural network architecture can be found in Figure 1.2, which shows a neural network with 5 covariates (represented by the 5 circles on the left-most column, called neurons), and 3 "hidden layers," in this case each consisting of 10 neurons. A neural network is in essence a composite function, consisting of nested regressions. The first layer (depicted on the leftmost side), consists of all covariates used for the model. From these inputs, a weighted average is constructed for each neuron in the following layer. That is, for each neuron in the first hidden layer, we construct $\sum_{i=1}^{p} w_i x_i$ (in essence, a linear regression of a latent variable on x) where the weights are allowed to vary between neurons. Each weighted average is then passed through an "activation funtion," which can be any number of functions. In our case we exclusively consider the commonly used rectified linear unit (ReLU) activation function defined as $h(x) = max(0, x)$. This constructs the output of the first hidden layer $\tilde{\mathbf{x}}_1$, which is then used as an input to the second hidden layer in the same

Figure 1.2: Neural Network Architecture



fashion (a weighted average is constructed and then passed through an activation function). This process is repeated for all layers until the output layer is reached.

The number of layers in the specified architecture we denote by L, and the number of neurons per layer we denote $H_l,\ l = 1, ..., L$. For simplicity define $z_i = (y_i, \mathbf{x}_i')'$. If we define the matrix of weights for layer l as $\mathbf{W}_l$, then we can express the output of hidden layer l as $h(\mathbf{W}_{l-1}\tilde{x}_{l-1} + b_{l-1})$. The final output is obtained by a linear regression of y on $\tilde{\mathbf{x}}_L$, $\hat{y} = \mathbf{W}_L\tilde{x}_L + b_L$. In this sense, the neural network is simply learning a set of basis functions which are then used as inputs for a linear regression of y on the set of basis functions. By nesting all layers, we can obtain the following formulation of the final output of a neural network:

$$\hat{y} = \mathbf{W_1}h(\mathbf{W}_{l-1}h(\mathbf{W}_{l-2}h(...h(\mathbf{W_0}\mathbf{x} + b_0) + ...) + b_{l-2}) + b_{l-1}) + b_L$$

Ordinarily an estimator is constructed by minimizing the mean squared prediction error (or another loss function) of the neural network to estimate the weights of the chosen network.

We define the newly proposed transfer estimator as

$$\hat{W}_{TDNN} = \underset{\substack{\mathbf{W} \\ f \in \mathcal{F}_{DNN} \\ \|f\|_\infty \leq 2M}}{\arg\min} \sum_{i=1}^{n} \ell(f(\mathbf{x_i}, \mathbf{W}), y_i) + \lambda_n \|\mathbf{W} - \tilde{\mathbf{W}}\|_1$$

Where $\tilde{\mathbf{W}}$ is an estimate of $\mathbf{W}$ obtained from auxiliary data, and $\ell(f(\mathbf{x_i}, \mathbf{W}), y_i)$ may for example be $\frac{1}{2}(y_i - f(\mathbf{x}_i, \mathbf{W}))^2$. With the estimator of f^* being

$$\hat{f}_{TDNN} = f(\mathbf{x}, \hat{\mathbf{W}})$$

In practice it may be beneficial to choose a different λ_n for each layer of the neural network, but for theoretical purposes we assume only one penalty parameter. Our main results additionally require the next three assumptions

Remark 1.3 A method which is quite popular in machine learning is the method of "weight freezing." In this method, specific weights in our neural network are chosen to be frozen at the values estimated on the auxiliary sample. The optimization of the neural network is then carried out with only the remaining non-frozen weights being optimized. If the frozen weights are assumed to be equal to the weights associated with f^*, and we freeze the entirety of the first H layers, then the results of FLM still apply to the reduced neural network.

Remark 1.4 Another method which is commonly employed in transfer learning for neural networks is weight initialization. When fitting the neural network, for optimization purposes the initial values of the weights can be selected according to the weights associated with $\tilde{f}$. In some cases, the simple act of weight initialization without any further transfer can improve performance. This is presumably because neural networks have very complex loss functions with potentially many local minimums, and the initialization results in the optimization converging to a more optimal estimate. We found in our numerical experiments that weight initialization alone can result in improved performance.

Partial Transfer

In the previous section we develop results in which we penalize for differences from $\tilde{W}$. This causes all of the weights in the neural network to shrink towards the weights associated with $\tilde{W}$. It is of interest to adjust the estimator of the previous section to allow for the partial penalization of the neural network. That is, we would like to be able to construct a transfer algorithm such that we only penalize specific weights/layers for distance from the auxiliary esitmates. As shown in 1.1, it is known that weights are more similar for early layers compared to later layers. Therefore we ideally want to be able to apply more shrinkage to the early layers. As discussed in Section 1.1, the assumption that the auxiliary estimates are close to the truth is much more reasonable if this only needs to hold for specific (early) layers of the neural network. We can construct an estimator where we penalize only the first h layers,

$$\hat{W}_{TDNN} = \underset{\substack{\mathbf{W} \\ f \in \mathcal{F}_{DNN} \\ \|f\|_\infty \leq 2M}}{\arg\min} \sum_{i=1}^{n} \ell(f(\mathbf{x_i}, \mathbf{W}), y_i) + \lambda_n \sum_{j=1}^{h} \|\mathbf{W_j} - \tilde{\mathbf{W}}_\mathbf{j}\|_1$$

This is the method which we employ in our numerical examples.

1.4 Estimation of Parameters of Interest

In practice the ultimate goal of using these transfer algorithms as nuisance estimators is to conduct valid inference for the ultimate parameter of interest. In our numerical examples in Sections 1.5 and 1.6 we consider the estimation of causal parameters of interest, specifically the Average Treatment Effect (ATE) in a binary treatment setting. In order to estimate the ATE we use two forms of the Double Machine Learning (DML) procedure. As mentioned in Section 1.1, DML requires particular assumptions on the machine learning algorithms which are used with it.

We begin by describing the ATE setting. Suppose we have an outcome of interest Y, binary treatment variable $T \in \{0, 1\}$, and covariate set $\mathbf{X}$. The outcome variable Y is such that $Y = TY(1) + (1-T)Y(0)$, where $Y(1)$ is the potential outcome associated with $T = 1$, and $Y(0)$ is the potential outcome associated with $T = 0$. The ATE in this setting can be defined as $E[Y(1) - Y(0)]$, the expected difference in potential outcomes, i.e. the average effect of treatment on Y.

We use two DML models discussed in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins 2018, the Partial Linear Model (PLM) introduced in Robinson 1988 and a full non-parametric model which we term the "Interactive Regression Model"(IRM). The partial linear model allows for additive separability between the treatment and control variables, but allows for Y and T to be dependent on $\mathbf{X}$ through a fully non-parametric function. The general formulation of the partial linear model is the following:

$$Y_i = \beta T_i + f(\mathbf{x}_i) + \varepsilon_i \tag{1.3}$$

$$T_i = m(\mathbf{x_i}) + \nu_i \tag{1.4}$$

Double Machine Learning allows for the estimation of f and m using machine learning or any non-parametric estimator in order to construct an asymptotically valid estimator of the ATE. Like many other estimators, DML relies on the unconfoundedness assumption given the available covariate set in order to produce valid causal estimates.

Assumption 1.5 (Unconfoundedness) *T is independent of the error term ε conditional on covariates set $\mathbf{X}$, $\epsilon \perp T|X$*

Given that assumption 1.5 holds, and that the algorithms used to estimate f and m converge sufficiently fast then we can apply the DML estimator of the PLM to produce valid estimates of the ATE. The basic procedure of DML to estimate the model given by equations 1.3 and 1.4 is the following:

1. Estimate f via any desired method (e.g. a neural network). Obtain residuals $\hat{e}_i \; \forall \; i$.

2. Estimate m via any desired method. Obtain residuals $\hat{\nu}_i \ \forall \ i$.

3. Regress $\hat{e}$ on $\hat{\nu}$. The coefficient on $\hat{\nu}$ is the estimate of the ATE.

More formally, the above procedure reuslts in the following estimator of the ATE:

$$\widehat{ATE} = \big(\frac{1}{n}\sum_{i=1}^{n}\hat{\nu}_i\hat{\nu}_i\big)^{-1}\frac{1}{n}\sum_{i=1}^{n}\hat{\nu}_i(Y_i - \hat{f}(\mathbf{x}_i)) \tag{1.5}$$

If the true data generating process can be approximated by a PLM, then estimation via PLM is preferable as it is more efficient. If the data generating process is more complex however, then the PLM model will be misspecified while the IRM model may not be. Regardless of which model we use, we require the standard unconfoundedness assumption.

The IRM model on the otherhand, allows for cases of non-additive separability, and is in this sense a more non-parametric model than the PLM model:

$$Y_i = f(T_i, \mathbf{X}_i) + \epsilon_i \tag{1.6}$$

Just as in the partial linear case, we can estimate f and propensity score m using a machine learning algorithm or non-parametric estimator in order to construct an estimate of the ATE. Unlike the PLM, the IRM model requires the treatment to be binary. In order to use the DML estimator for the IRM model, we require the common support assumption on the propensity score.

Assumption 1.6 (Common Support) $0 < m(\mathbf{x}) < 1$

Given that assumptions 1.5 and 1.6 are satisfied, then the IRM model estimator is given by

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(1,\mathbf{x}_i) - \hat{f}(0,\mathbf{x}_i) + \frac{I\{T_i = 1\}(Y_i - \hat{f}(1,\mathbf{x}_i))}{\hat{m}(\mathbf{x}_i)} - \frac{I\{T_i = 0\}(Y_i - \hat{f}(0,\mathbf{x}_i))}{1 - \hat{m}(\mathbf{x}_i)}\right] \tag{1.7}$$

In practice both the PLM and IRM estimators are combined with cross-fitting via sample splitting. This involves splitting the sample into L sub-samples. For subset I_ℓ, fit the machine learning models on the complement set I_ℓ^C. Then extrapolate the models to I_ℓ to estimate any necessary quantities (e.g. $\hat{f}(\mathbf{x_i})$). Sample splitting is an important part of DML and has been shown to reduce bias in estimaton and improve coverage rates (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins 2018, Colangelo and Lee 2020, Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey 2017). In our numerical exercises we utilize sample splitting with $L = 2$.

1.5 Simulations

This section provides Monte Carlo simulations to explore the efficacy of the proposed algorithms. In Section 1.5.1 the algorithms are assessed through their out of sample prediction performance. In Section 1.5.2 we assess their performance with regard to the estimation of the average treatment effect through double machine learning.

1.5.1 Prediction Performance

In this section we assess the proposed algorithms by their out of sample predictive performance, measured by root mean squared error (RMSE). We use the following data generating process with $1,000$ replications:

$$
\begin{aligned}
&Y_1 = 1 + \mathbf{X}\theta_1 + (\mathbf{X}^2)\theta_1 + \varepsilon_1 \\
&Y_2 = 2 + \mathbf{X}\theta_2 + (\mathbf{X}^2)\theta_2 + \varepsilon_2 \\
&\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma}) \\
&\varepsilon_1, \varepsilon_2 \overset{i.i.d.}{\sim} N(0,1) \\
&\theta_1 = [1\ 1\ 1\ 1\ 0\ 0\ ...\ 0]'
\end{aligned}
$$

where $\mathbf{\Sigma}$ is a tri-diagonal matrix with the non-zero off-diagonal elements equal to 0.2, and the diagonal elements are 1. $\mathbf{X}$ is a 100 dimensional covariate set, and we use two different definitions of θ_2

$$
\begin{aligned}
&(1) \quad \theta_2 = [.9\ .9\ .9\ .9\ 0\ 0\ ...\ 0]' \\
&(2) \quad \theta_2 = [1\ .8\ .6\ .4\ 0\ 0\ ...\ 0]'
\end{aligned}
$$

We refer to the DGP under (1) as DGP 1 or the "high correlation" case. We refer to the DGP under (2) as DGP 2 or the "low correlation" case. By considering both high and low correlation scenarios we can explore how robust the new algorithms are.

The goal is to construct out-of-sample predictions for Y_1 with the highest level of accuracy. For each replication we generate a sample size of $n = 10,000$. To perform penalized transfer, we first estimate the desired algorithm on the full sample for Y_2. We then use the parameter estimates/weights from this preliminary step for usage in the transfer algorithm on a subset of the data for Y_1. Tuning parameters (e.g. the penalty for lasso) are chosen via 10-fold cross-validation.

Root mean squared error is computed for all algorithms on the remaining observations of the data not used for fitting. That is, for a sample size of $n = 1,000$, we would have $9,000$ remaining observations to compute the out-of-sample RMSE on. For comparison we plot relative root mean squared error (The ratio of RMSE with transfer to RMSE without) of the transfer algorithms compared with the analagous algorithm with no transfer. Values less than 1 indicate that the transfer learning algorithm performed better than its counterpart.

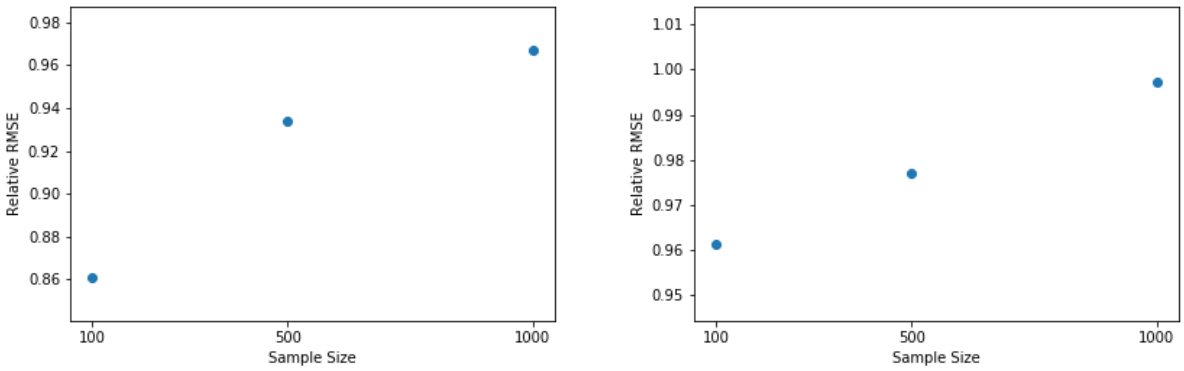
Linear Models

In this section we consider Transfer OLS (TOLS) and Transfer Elastic Net (TEN). In addition to the original covariate set, we construct basis functions for second and third powers of $\mathbf{X}$.

Figure 1.3a and 1.3b display the relative RMSE of TOLS relative to lasso for the high and low correlation cases respectively. We can see that for the high correlation case, the RMSE advantage of TOLS is larger than the low correlation case. As we would expect, the advantage of transfer learning decreases with sample size. As n increases, the RMSE of the transfer algorithm approaches the RMSE of the lasso.

While the relative RMSE of TOLS is not as small as for the high correlation case, we still observe a clear advantage for the sample sizes of $n = 100$ and $n = 500$. However

this difference vanishes by $n = 1,000$ whereas in the high correlation case there is still an advantage to TOLS even at $n = 1,000$.



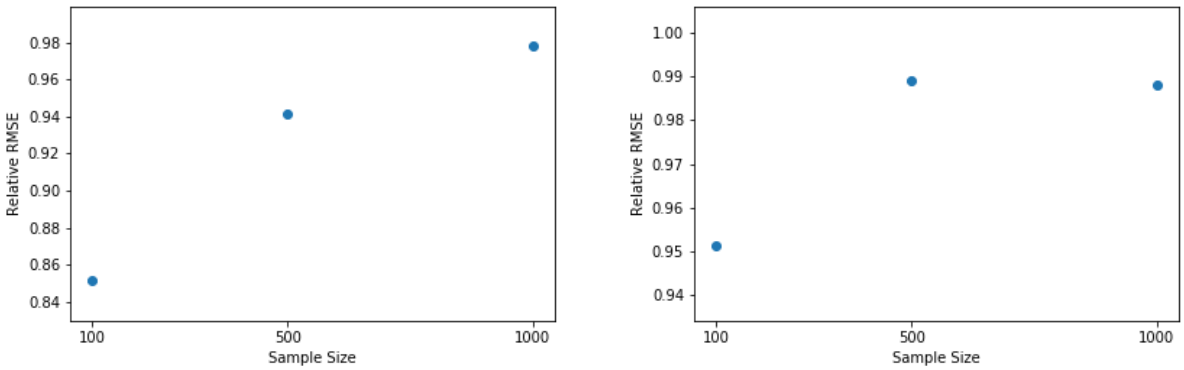
(a)
High Correlation

(b)
Low Correlation

Figure 1.3: Relative RMSE of TOLS

Figure 1.4a and 1.4b display the the relative RMSE of TEN relative to elastic net for the high and low correlation cases respectively. The resuls are consistent with the 1.3, showing similar patterns for both cases. Just like Figure 1.3, the RMSE converges as n increase. In the low correlation case it appears that the relative RMSE of TEN increases faster than for TOLS, with the advantage mostly disappearing by $n = 500$. Overall, both 1.3 and 1.4 are consistent with theoretical predictions.

Of particular note is that the transfer algorithms appear to not at any point make the performance worse than the analogous non-transfer algorithm. This could be due to the fact that we chose the tuning parameters via cross validation. In the case that transfer would be less helpful or detrimental, the selected tuning parameter should be smaller. Thus our results would appear to indicate that proper tuning of the models can prevent a worsening of performance.



(a)
High Correlation

(b)
Low Correlation

Figure 1.4: Relative RMSE of TEN

Neural Networks

We consider a fully connected feedforward neural network architecture with 3 hidden layers. All hidden layers have the same number of neurons, which is allowed to vary to compare the peformance with respect to different levels of neural network complexity. We refer to the number of hidden neurons in each hidden layer as the "complexity" of the network. Figure 1.2 displays what the architecture looks like with a complexity of 10. Activation functions for the 3 hidden layers are Rectified Linear Units (ReLU) as described in Section 1.3.

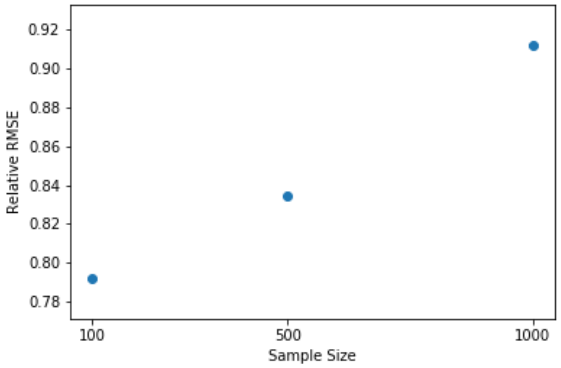
To implement the transfer neural network algorithm, we first fit a neural network (the "pre-trained" network) on the full sample ($n = 10,000$) using Y_2 as the outcome variable and the exact same architecture as our target network. We then fit the neural network with transfer using Y_1 as the outcome variable on a sub-set of the data. All variables are standardized before fitting. A single penalty is applied to the first 3 layers, but not the final output layer. As explained in Section 1.3, we penalize the weights for distance from the weights of the pre-trained network. The penalty is chosen via cross validation. For comparison, we fit the same neural network with no penalty (no transfer). Unlike the linear models, we do not add any additional basis functions.

Figure 1.5 displays the results for DGP 1 which has a high correlation between Y_1 and Y_2, and 1.6 displays analogous results for the low correlation case. Looking at Figures 1.5a and 1.6a we can see how the relative RMSE of the transfer neural network changes with sample size, fixing the complexity at 25. The advantage of the transfer algorithm for both the high and low correlation cases is clearly stronger for smaller sample sizes. The relative RMSE increases rapidly as the sample size is increased. This is especially the case for the low correlation case in which the difference almost completely vanishes by $n = 1,000$, whereas there is still a substantial difference in the high correlation case. We can also note that the transfer neural network is at least as good as the neural network without transfer for all cases considered.

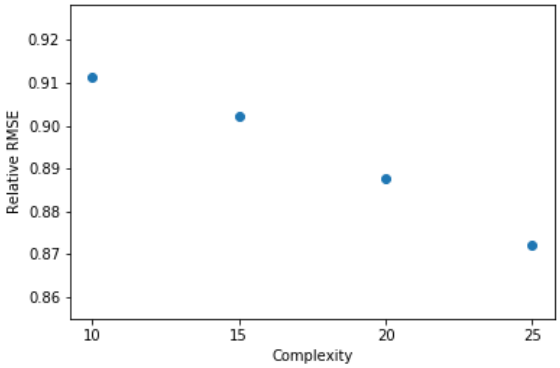
Figure 1.5b displays the relative RMSE for a fixed sample size ($n = 1,000$), but for varying complexity of the network. This is to illustrate that for small sample sizes, the chocie of neural network architecture is more limited when transfer is not used. This is because you generally do not want to have too many weights to estimate relative to the sample size. As complexity is increased, we can see that the transfer algorithm begins to perform substantially better when compared to the neural network without transfer. This difference however is toned down for the low-correlation case, albeit still present.

Overall, these results are consistent with what was found for the case of linear models. We do find however that the relative RMSE for TDNN is lower than for the linear models. This potentially tells us that transfer can be more advantageous in cases where one wishes to use neural networks.

Remark 1.5 The DGP used for the simulations is quite simplistic, and fairly sparse and linear. Because of this, we would not expect the non-transfer algorithms to perform poorly, even on a small dataset, as the true DGP has only 4 relevant covariates. Hence, while the evidence points to the transfer learning algorithms performing better, the difference is not drastic. The DGP was chosen to be simple for computational constraints, but the advantage of transfer learning would be much clearer in a more complex high-dimensional setting.

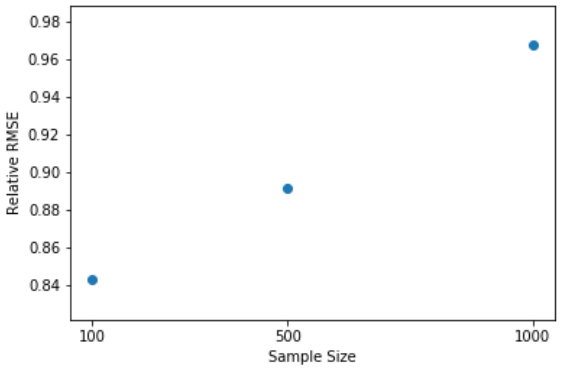


(a)
Relative RMSE for fixed complexity, varying sample size

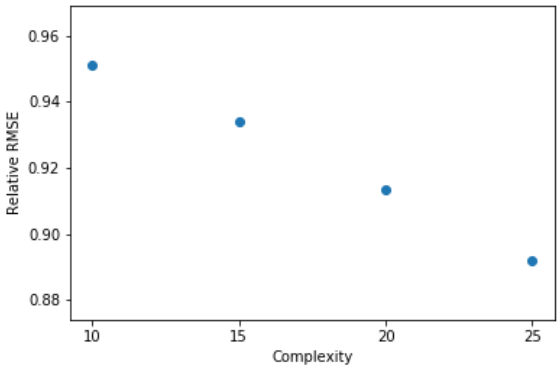


(b)
Relative RMSE for fixed sample size (N=1000), varying complexity

Figure 1.5: RMSE of Transfer Neural Network with Highly Correlated Outcomes



(a)
Relative RMSE for fixed complexity, varying sample size



(b)
Relative RMSE for fixed sample size (N=1000), varying complexity

Figure 1.6: RMSE of Transfer Neural Network with Weakly Correlated Outcomes

1.5.2 ATE Estimation

In this section we implement the proposed algorithms for the purposes of estimating the average treatment effect in the partial linear model and fully non-parametric models described

in section 1.4. We consider the following data generating process:

$$
\begin{aligned}
&Y_1 = 1 + \beta_1 T + \mathbf{X}\theta_1 + (\mathbf{X}^2)\theta_1 + \varepsilon_1 \\
&Y_2 = 1 + \beta_2 T + \mathbf{X}\theta_2 + (\mathbf{X}^2)\theta_2 + \varepsilon_2 \\
&T = I\{\mathbf{X}\theta_3 + (\mathbf{X}^2)\theta_3 > 0\} \\
&\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma}) \\
&\varepsilon_1, \varepsilon_2 \overset{i.i.d.}{\sim} N(0,1) \\
&\theta_{1j} = 1, \quad j = 1, 2, ..., 100 \\
&\theta_{21} = 0.5, \theta_{22} = 0.4, \theta_{2j} = 0.2, j = 3, ..., 100 \\
&\beta_1 = 1, \beta_2 = 0.5
\end{aligned}
$$

Where $\mathbf{\Sigma}$ is again tri-diagonal matrix where the non-zero off-diagonal elements are equal to 0.2, and the diagonal elements are 1.

The goal is to estimate the average treatment effect β_1 through Double Machine Learning. Please see Section 1.4 for a full description of the double machine learning procedure. We will be computing the average RMSE and bias, and coverage rates. The coverage rate being defined as the percent of the time that the true parameter of interest lies in the constructed 95% confidence interval (95% being optimal).

For the partial linear model we use the specified algorithm for the outcome model ($g(X)$), and a linear model for the treatment model ($m(X)$). For the fully heterogeneous DML model, we use the specified algorithm to estiamte $g(T, X)$, and a logistic regression to estimate the propensity score ($m(X)$). The tuning parameters were again tuned via 10-fold cross validation for all models. The architecture of our deep neural network is a 5 layer network with 3 hidden layers, each with 25 neurons and a ReLU activation function. For the linear algorithms we add basis functions for second and third degree powers of $\mathbf{X}$, and interactions with the treatment variable.

Table 1.1 displays the results on RMSE, bias, and coverage rates for lasso and TOLS. Table 1.2 displays the results for deep neural networks with and without transfer (TDNN and DNN respectively). Looking at 1.1 we can see that there is a clear improvement with transfer in both the PLM and IRM models for all three statistics at a sample size of $n = 500$. However, this advantage vanishes for $n = 1,000$ in both cases. This is contrary to what we found in Section 1.5.1, and shows that certainly in some cases transfer could make matters worse. However, the biases, RMSE and coverage rates are not bad. The coverage rate for TOLS is 92% in the PLM case and 91% in the IRM case which is relatively close to the desired 95%. We can also note that while TOLS does improve performance at a sample size of $n = 500$, it still does not obtain optimal coverage rates at the sample size. This tells us that transfer is not to be thought of as a panacea for insufficient samples size, and that while it can improve estimation it is not guaranteed to fully alleviate biases.

Turning to Table 1.2, we see that the DNN with transfer vastly outperforms the baseline DNN in bias, RMSE and coverage rate for both sample sizes. At a sample size of $n = 500$, both algorithms still result in substantial biases and poor coverage rates for both the PLM and IRM cases. But even though the performance of TDNN at $n = 500$ is still poor, it vastly outperforms DNN. At $n = 1,000$, TDNN obtains near 0 bias, and near 95% coverage rates, while DNN still has substantial bias and the coverage rate actually decreased even with the larger sample size. This could be an indication that as the sample size increased, the DNN without transfer is converging slower that DML requires, thus resulting in worsening coverage rates as sample size increases. In both the DNN and TDNN cases we do see that the RMSE and bias is reduced with a larger sample size, in spite of the potentially counter-intuitive results on coverage rate.

The results appear to indicate that transfer may be a more important tool when using neural networks than for other models. But when comparing the DNN case to the linear case, it should be emphasized that for the linear case the "correct" set of basis functions are given as inputs to the model. On the other hand, for DNN we allow the neural network to

estimate the model in a more non-parametric fashion, allowing the neural network to pick up the necessary basis functions for unbiased estimation. In this sense, the neural network estimation is more difficult and could explain why we see such large differences in Table 1.2 compared to Table 1.1. In practice we are unlikely to properly guess the correct basis functions and so Table 1.1 may underplay the advantage of transfer learning.

Table 1.1: Comparison of Performance of Lasso and TOLS

		Lasso			TOLS		
Model		RMSE	Bias	Coverage	RMSE	Bias	Coverage
PLM	500	0.10	0.07	0.65	0.06	0.03	0.76
	1000	0.02	0.00	0.95	0.03	0.00	0.92
IRM	500	0.15	0.08	0.84	0.09	0.05	0.87
	1000	0.04	0.00	0.95	0.03	0.02	0.91

[1] Results are from 1,000 Monte Carlo replications.

Table 1.2: Performance of Deep Neural Networks with and without Transfer

		DNN			TDNN		
Model		RMSE	Bias	Coverage	RMSE	Bias	Coverage
PLM	500	0.16	0.14	0.32	0.11	0.08	0.73
	1000	0.13	0.11	0.02	0.04	0.00	0.91
IRM	500	0.14	0.10	0.43	0.09	0.05	0.81
	1000	0.10	0.08	0.04	0.04	0.00	0.97

[1] Results are from 1,000 Monte Carlo replications.
[2] DNN (Deep Neural Network) refers to the neural network with no transfer. TDNN (Transfer Deep Neural Network) refers to the neural network with transfer.

To summarize, we find that the results for both outcome prediction and ATE estimation are consistent with theoretical predictions. Transfer learning performed at least as well as the non-transfer counterparts in all cases considered, succeeding in reducing bias and RMSE. Furthermore, coverage rates were drastically improved for the ATE estimation. The advantage of transfer learning for both prediction and ATE estimation appears to be larger for neural networks than the linear models.

1.6 Empirical Application

In this section we will demonstrate the effectiveness of the previously described algorithms on actual data. One could certainly make the case that the Monte Carlo data generating process may not be realistic, and so even though the algorithms seem to be effective for our chosen data generating process, it does not mean that they are necessarily effective in actual empirical applications. To assess performance on empirical data, we artificially create the problem of missing outcome Y by only using a fraction of the sample for estimation, and then compare results to the full-sample estimation. As we will show in this section, the algorithms also perform well on real data. We will assess the performance of transfer learning in DML by estimating on subsets of the empirical data, and comparing with the estimates on the full sample of data.

We analyze the effectiveness of our proposed methods by re-analyzing the Job Corps program from the mid 1990s. We use the same data from the Job Corps program used in Schochet, Burghardt, and McConnell 2008. The Job Corps program is the largest publically funded program to help train/educate under-privileged youth, ages 16-24, to improve their future outcomes. Schochet, Burghardt, and McConnell 2008 analyze the Job Corps program in the presence of a random experiment, in which a subset of eligible youth who applied to participate between 1994-96 were randomly allowed to enroll in the program. 9,409 youths were assigned to the treatment group, and 5,977 were assigned to the control group. The random assignment was imperfect however, in that some of the control group still enrolled in the program, and some of the assigned group did not enroll in the Job Corps program. Furthermore, there is sample attrition in the surveys and missing responses for specific survey questions. This is assessed in Schochet, Burghardt, and McConnell 2008 which find that in spite of this, there still exists pre-treatment sample balance between treatment and control. We furthermore have a rich set of covariates to attempt to control for any selection bias. For a more detailed and thorough discussion of the specifics of the Job Corps program, we refer the reader to the original paper.

We chose the Job Corps program specifically for this empirical demonstration because: (i) it has a rich set of covariates. (ii) It has a plethora of different outcome variables, includ-

ing: weekly earnings, welfare receipts, employment status, crime, drug use and educational attainment. (iii) The outcome variables are all well observed, which will allow us to assess how our transfer algorithms perform relative to a full sample estimation. For this demonstration we will limit our analysis to two outcome varibles: Earnings in the third year after the start of the program, and the percent of weeks employed in the third year after the start of the program. These two outcomes have a clear relationship in that a higher proportion of weeks employed should be directly correlated with earnings. We estimate the correlation between these two variables after standardization[6] on the full dataset to be 0.785. We also chose these two outcomes as it is a reasonable occurrence in practice that individuals will report their employment status but not their exact earnings. Fortunately we have nearly complete coverage for both of these outcome variables in our sample.

We drop observations which were missing data on key pre-treatment characteristics such as pre-treatment employment history (e.g. wages). We further dropped the remaining observations which were missing one of the two specified outcome variables. The resulting sample size is $11,250$.

For estimation we use both the partial linear and fully nonparametric DML estimators (PLM-DML and IRM-DML) as described in Section 1.4. We first estimate the effect on earnings using DML on the full dataset, using both lasso and deep neural networks for the outcome model. These estimates will serve as a baseline for comparison to assess the performance of ATE estimation with and without transfer. Once we have these estimates as a baseline, we split the sample into equally sized pieces and perform DML estimation with and without transfer on each subset of the data. We transfer based on the full-sample estimation of the corresponding machine learning model.

For deep neural networks we use partial transfer, and do not apply any penalty to the final layer of the network. The neural network considered had 5 hidden layers, each with 25 neurons and ReLU activation functions. Tuning parameters were selected via 10-fold cross-validation. We construct additional basis functions including second and third degree

[6]We recommend to always standardize data when using transfer learning so that the scale of coefficients/weights is comparable between the two models. It is common to standardize generally for machine learning algorithms, but it is of particular importance for transfer learning

polynomials, and interactions between the treatment and covariates for use as inputs in the linear model. No added basis functions are used for the neural network model.

Table 1.3 displays the coefficient estimates for both double Machine Learning models, using both deep neural networks and lasso to estimate the outcome model. Columns (1) and (2) display results for both outcome variables for deep neural networks in both the PLM and IRM Double Machine Learning models. Columns (3) and (4) display analogous results for lasso. There are some differences between the coefficients depending on which algorithm and DML model is used, however the conclusions are generally the same regardless of the chosen model. The effect on earnings is positive and significant, with coefficients ranging from 7.82 to 8.93 dollars per week increase on average for earnings. The coefficients for earnings are statisitcally significant at the 99% confidence level across the board.

Turning to the results for % of weeks employed, we also find positive coefficients across the board. However, the standard errors are larger relative to the coefficients in comparison to earnings, resulting in only the estimates for the IRM models being signfiicant at the 95% level. Focusing on the point estimates we see that they are fairly similar in all columns, ranging from 1.16 to 1.31, indicating agreement between the models that the effect is generally in the positive direction, similar in this regard to the estimates of the effect on earnings. This confirms what we would expect, that the effect of the program is in the same direction for both of these outcomes. This is exactly what we would hope before attempting to implement a transfer algorithm. Furthermore, these results are consistent with the conclusions of Schochet, Burghardt, and McConnell 2008, which also found significantly positive effects without the use of DML.

Now we turn to our evaluation of the described transfer learning algorithms. We first split the sample into 10 equal slices. We perform the estimation on all 10 sub-samples, and average the resulting coefficients and standard errors. That is, for the primary estimation of the ATE for earnings (Y) we allow 90% of the sample to be missing, while the auxiliary estimation for the secondary outcome utilizes the full sample. Ideally we would like (i) the average coefficient to be as similar as possible to the estimates in Table 1.3 for the corresponding algorithm, and (ii) the variance of the estimates to be lower with transfer

Table 1.3: Results with Full Sample

Outcome		(1) DNN-PLM	(2) DNN-IRM	(3) Lasso-PLM	(4) Lasso-IRM
Earnings	**coef.**	8.76***	7.82***	8.93***	8.18***
	std. error	(3.15)	(2.6)	(3.18)	(2.6)
% Employed	**coef.**	1.24	1.31**	1.16	1.2**
	std. error	(0.77)	(0.61)	(0.74)	(0.61)

[1] * Corresponds to significance at the 90% level, ** at the 95% level, and *** at the 99% level.

than the corresponding algorithms without transfer. We choose to focus on the average estimates across the sub-samples as this should be more informative about the overall bias of the estimator. Estimates for individual sub-samples have high variances due to the small sample size and so are not informative by themselves.

Since we don't know what the true treatment effect is, we use the full sample estimates with the corresponding algorithm as a baseline to compare to. Table 1.4 displays the coefficients and standard errors for the effect on earnings, averaged over the 10 sub-samples. Columns (1) and (2) display the results for deep neural networks for the PLM and IRM models, and columns (3) and (4) display the results for a linear outcome model (with added basis functions), for the PLM and IRM models. Results without transfer are displayed in the first row, and the results with transfer are displayed in the second row.

Table 1.4: Average results with and without Transfer (90% missing Y)

Method		(1) DNN-PLM	(2) DNN-IRM	(3) Linear-PLM	(4) Linear-IRM
No Transfer	**Avg. coef.**	6.87	5.36	7.34	9.38
	Avg. std. error	(14.12)	(15.47)	(14.77)	(15.21)
With Transfer	**Avg. coef.**	8.16	9.53	7.87	9.21
	Avg. std. error	(10.98)	(9.04)	(11.28)	(9.88)

[1] * Corresponds to significance at the 90% level, ** at the 95% level, and *** at the 99% level.
[2] Displayed results are for the *average* coefficients and *average* standard errors for the 10 sub-samples.

We can see that in all columns there are two main takeaways: (i) The coefficients when transfer is used are closer to the full sample estimates in 1.3. (ii) The standard errors are

universally smaller when transfer is used. Since the coefficients are closer to the full sample estimates, this likely indicates that the penalized transfer allows for a more accurate capture of the data generating process on a smaller dataset, and hence reduces bias. The difference with/without transfer is much more extreme for the deep neural network case, which is likely due to the fact that the neural networks considered are much more complex than linear models to account for the lack of added basis functions. That is, for neural networks we allow the neural network architecture itself to pick up the additional basis functions that are necessary to capture the data generating process.

While the transfer algorithms do seem to perform better, they are not perfect. The DNN-PLM model achieves the closest average to the baseline estimate (8.16 vs. 8.76), but the other columns have more substantial differences. This is important to note as it tells us that even with transfer learning we can still get substantial biases in our estimates when the sample size is too small. This is in contrast to our Monte Carlo simulations which found that there was little bias even in small samples, especially when compared with the case of no transfer. We can also notice that even though the standard errors are smaller than without transfer, all coefficients are statistically insignificant even at confidence levels much below 95%. The more important takeaway though is that the transfer learning algorithms performed their job in that they largely outperformed their non-transfer counterparts.

To further evaluate the transfer learning algorithms, we re-estimated the models when using only 2 equal sized subsets of the data. The results from this estimation are displayed in Table 1.5. The average coefficients with and without transfer are substantially improved from the 10 sub-sample case, which is what we would expect with the larger sample size. The difference between transfer/no transfer is lower than in 1.4, both with regard to the average coefficients and standard errors. Interestingly, the only significant results at the 90% level are for the deep neural network with transfer. This shows how in some circumstances, a result can become significant with the implementation of transfer learning.

Our results are consistent with expectations as we would expect the advantage of transfer learning to decrease as the sample size approaches the full sample. This is because there is less "information" to be learned from auxiliary estimates when the outcome is well observed.

Table 1.5: Average results with and without Transfer (50% Missing Y)

Method		(1) DNN-PLM	(2) DNN-IRM	(3) Linear-PLM	(4) Linear-IRM
No Transfer	**Avg. coef.**	8.32	7.65	9.12	8.41
	Avg. std. error	(6.33)	(6.94)	(6.57)	(6.68)
With Transfer	**Avg. coef.**	8.71*	7.97*	8.91	8.02
	Avg. std. error	(5.25)	(4.49)	(5.86)	(5.23)

[1] * Corresponds to significance at the 90% level, ** at the 95% level, and *** at the 99% level.
[2] Displayed results are for the *average* coefficients and standard errors for the 10 sub-samples.

To summarize, our results in this section corroborate the results from the Monte Carlo simulations in that transfer learning when used as part of double machine learning, can reduce bias and improve efficiency. By using a dataset which has near fully observed outcomes, we were able to compare the estimates on sub-samples to the full sample estimates, and demonstrate that in the presence of substantial missing data transfer learning can be used to improve DML estimation. As the difference between the sizes of the sub-samples and the full samples decreases, the advantage of the transfer learning approach decreases.

1.7 Summary and Conclusion

In recent years there has been an explosion in theoretical results for machine learning algorithms, paving the way for their usage in economics and other fields. However, while transfer learning has been growing in usage for predictive problems, the theoretical behavior of these algorithms has not been well explored or understood.

This paper develops theoretical results for elastic net with transfer via ℓ_1 penalization, and proposes transfer algorithms for deep neural networks. We then use these algorithms as nuisance function estimators for double machine learning to obtain estimates of the average treatment effect.

We find that in practice the proposed transfer algorithms, especially the neural network, can be used to great effectiveness. Simulation results show that there exists a range of scenarios in which all algorithms can substantially improve the estimation of the ATE in the presence of auxiliary data. More research is required to develop a theoretical framework for practically identifying when transfer is appropriate for improving estimation. Nevertheless, the numerical results in this paper demonstrate that these algorithms are effective not just in theory but also in practice

A number of natural extensions arise by applying the transfer via ℓ_1 penalty to other machine learning algorithms not considered here. These include variations of random forests, support vector machines, and different types of neural networks. It should additionally be noted that our results are not completely general in that we generate the sub-samples at random. In the very likely case that the missingness is non-random, additional adjustments may need to be made to adjust for this problem. We can however hypothesize/speculate that transfer learning in the case of non-random missingness may be even more advantageous, as the penalization could force the estimates to be closer to what they would be under random missingness. However, further investigation is necessary on this subject, which we leave for future work.

Chapter 2

The Effects of Minimum Wages on Unemployment Duration and Re-employment Outcomes

2.1 Introduction

The effect of the minimum wage policy is a very contentious issue that has amassed a very large body of literature. Although a great deal of attention has been given to the impact of the minimum wage on employment, no clear consensus exists yet. One strand of research, such as Powell 2017, Meer and West 2016, Thompson 2009, and Neumark, Schweitzer, and Wascher 2004, finds disemployment effects (see Neumark and Wascher 2008 for a comprehensive review of previous work). A parallel strand of research questions the methods that lead to negative results and finds instead no dis-employment effects (Dube, Lester, and Reich, 2010; Allegretto, Dube, and Reich, 2011; Giuliano, 2013; Card and Krueger, 1993; Cengiz et al., 2019; Dustmann et al., 2019), and even positive effects (Card, 1992). However, despite this great body of research and effort given to the topic of employment effects; virtually no effort has been put on evaluating how the policy affects unemployed workers specifically. In fact, to our knowledge the only published paper that asks this question is Pedace and Rohn 2011.

We assert that studying the unemployed directly can give us new insights into the effect of minimum wage policies. We consider how the minimum wage impacts unemployment duration, workers' search behavior, and their re-employment outcomes (the trajectories of their wages and hours after re-employment), in order to get a full picture. This way we can determine whether a potentially negative effect (i.e., a longer spell) is compensated in any way (i.e., higher wages at re-employment), and also whether the effect is short lived or long lasting.

Furthermore, we believe that studying the unemployed may be particularly insightful because unemployed workers are likely to experience the negative effects of a minimum wage increase more immediately than workers who remain employed. Recent research has suggested that the effects of the minimum wage are felt on employment growth (Meer and West, 2016) and flows (Dube, Lester, and Reich, 2016) rather than immediate disemployment. Under these circumstances, the unemployed are more likely to exhibit negative effects sooner and more intensely, if they exist. Moreover, their outcomes should be less affected by sticky prices than those of the employed population (Barattieri, Basu, and Gottschalk, 2014).

For our analysis we use a sample of unemployment spells built using four panels[1] of the Survey of Income and Program Participation (SIPP) to establish the relationship between minimum wages and several outcomes of unemployed individuals. More precisely, we focus on the following outcomes: unemployment duration, probability of abandoning the search for work, starting wages and starting hours after re-employment, and the trajectories of wages and hours after re-employment (for two to three years). Using the individual level panel data provided by the SIPP has the advantage of allowing us to follow individuals over long periods of time, enabling us to better determine whether and how minimum wages affect employment matching.

We take advantage of the weekly frequency of the SIPP data on employment status to pair it precisely with the state minimum wage data. This allows us to introduce another novelty to our analysis, the distinction between initial value of the minimum wage at the start of a spell (which we may refer to as "minimum wage level") from within-spell changes (which we may refer to as "minimum wage change"). We are able to determine who experienced a

[1] Panels 2001, 2004, 2008, and 2014

change of the minimum wage precisely because we have weekly frequency data on everyone for their employment status. Alternatively, we could simply include in our models the initial minimum wage level ignoring any within-spell changes, but there is reason to believe these changes are itself important information. Previous literature (presented below) has shown that these events (minimum wage increases) shock markets, leading to higher prices and exit of firms, which could have an important effect on workers looking for work. Dividing the minimum wage into these two components should identify each effect better.

Our findings suggest that indeed this distinction between minimum wage levels and changes is very important. Higher levels of the minimum wage are associated with null or very mild effects. At best, they can lead to a moderate reduction in unemployment duration for the least productive workers; at worst they can create important distortions by significantly increasing the probability a worker quits the search for work. On the other hand, minimum wage changes are much more important for the unemployed. A minimum wage change leads to longer spells, an effect that may more than double for the least skilled workers (proxied by educational attainment); it increases the probability of quitting the job search by double digits; and leads to long-term decline in working hours, although this latter effect is not present for the least skilled group.

We find that our results are robust to a number of common critiques in the minimum wage literature. They are robust to the inclusion of division-specific time effects as suggested by Allegretto, Dube, and Reich 2011, and do not seem to be driven by local policy preferences either (which we tested taking advantage of the federal minimum wage changes in 2007-9). Additionally, our effects do not seem to be driven by bias due to longer spells becoming more likely to experience minimum wage changes, which we tested by implementing inverse probability weighting. We also ran various tests to make sure our results are not sensitive to inclusion or exclusion of extreme cases, and that our results for quitting the job search are not overly sensitive to our definition.

This paper has important contributions to the literature. Pedace and Rohn 2011 is the only other paper we are aware of that measures directly the impact of higher minimum wages on unemployed workers. Using a hazard model under four different distributions of the survival function and a sample of unemployed individuals from the Displaced Worker

Survey, the authors study the impact of initial minimum wage levels on unemployment duration. They find an effect that varies along the lines of sex and educational attainment. Specifically, spells become longer for older women in low skill occupations and men that are high school dropouts; while more educated men reduced the length of their unemployment period. However, their analysis focuses exclusively on unemployment duration, and so they cannot conclusively say whether a higher minimum wage is beneficial or not for less educated women or more educated men, as longer spells are not necessarily a bad thing (they can be good if they lead to better wages for example). Additionally, our analysis reveals the existence of two different effects acting on the unemployed that are not distinguished by Pedace and Rohn.

Clemens and Wither 2019 is one of the most recent and relevant papers that evaluates the impact of the minimum wage using individual level panel data. Furthermore, they even use the same survey in their analysis, albeit only the 2008 panel, and show that employment suffers when minimum wages are higher. However, they do not concentrate on the impact of higher minimum wages on the unemployed but rather on workers in general. Furthermore, they focused on a very particular moment in history in which two rare events met: the Great Recession and a 41% increase of the federal minimum wage. Our analysis seeks to provide policy relevant information in a much wider set of scenarios.

Previous research that has recognized the need for individual level panel data has typically reached similar conclusions to those of Clemens and Wither using other sources. Neumark and Wascher 1995 and Abowd et al. 1997 find dis-employment effects using the Current Population Survey (CPS), and Currie and Fallick 1993 find dis-employment effects using the National Longitudinal Survey of Youth (NLSY). However, Zavodny 2000 uses the CPS survey and finds no dis-employment effects for low productivity teenagers compared to other low productivity teens.

We improve on these analyses with the use of the SIPP. The SIPP has two important advantages over the CPS data: It follows individuals for a longer time period, four years compared to only one; and provides repeated data on employment status, wages, hours, and search behavior throughout the survey. Compared to the NLSY the key advantage is that the survey follows a larger number of individuals, which span a larger number of birth dates.

Finally, the nature of the SIPP allowed us to distinguish initial minimum wage levels from its changes during a spell. We find that the effects from initial levels and changes are in fact very different. Previous literature has hinted at this result before, with findings that increases in prices due to a change in the minimum wage occurs shortly after the hike (Aaronson, 2001; Aaronson, French, and MacDonald, 2008; Basker and Khan, 2016), and that firm exit and entry is accelerated by minimum wage increases (Aaronson, French, Sorkin, et al., 2018).

The rest of the paper is organized as follows: in Section 2.2 we describe the data used; in Section 2.3 we summarize the econometric models used and address their advantages and limitations; in Section 2.4 we present our findings and our understanding of what they mean; in Section 2.5 we cover various robustness checks we implemented; finally, in Section 2.6 we conclude and summarize the most important findings and their implications for the minimum wage policy.

2.2 Data

For this analysis we consider four panels of the Survey of Income and Program Participation (SIPP), which are denoted by the first complete year which that panel observes (i.e. we consider the 2001, 2004, 2008, and 2014 panels). It should be emphasized that despite the names of the individual panels, they do include some observations preceding the panel year (e.g. the 2001 panel includes observations from the last quarter of 2000). Additionally, each panel spans multiple years, albeit they are all different lengths. The 2001 panel is the shortest and spans 3.25 years, it is followed by the 2008 panel that is just a quarter short of four years. The 2014 panel is the second longest, at exactly four years, and lastly the 2004 panel spans four years and one extra quarter.

The 2001, 2004, and 2008 panels interviewed participants 3 times a year on a rotating basis, with different groups of participants entering the survey at different times (only a small fraction enter the survey at the first observed month). The 2014 panel interviewed participants once per year, however we find no evidence that this resulted in any abnormalities or reporting issues (we did not find for example that very short spells are less likely in the 2014 panel). Any individual over the age of 15 in a participating household is inter-

viewed if possible. The survey asks participants a large volume of questions regarding their demographic characteristics and participation in the labor force. This includes (but is not limited to) the participant's age, sex, race, state of residence, number of children, education level, employment status, approximate dates of job loss (if they lost a job), approximate dates of hiring (if they were hired), the specific reason for any job loss, at what times was the individual looking for a job, wage, income, hours worked, and household assets.

Recall that our analysis focuses on unemployment, so first and foremost we must define our measure of unemployment spells. The data available does not include the exact dates of job loss/hiring and the exact dates that an individual was looking for work. Instead we have the employment status for every week of each panel for each individual, and an indicator for whether they were looking for a job in each week. From here we are able to create a variable identifying the "survival time" (in weeks) of each unemployment spell for our hazard model, which we will discuss in detail in Section 2.3. Using the survey, we define a spell as a period of unemployment that begins when a person first declares to be looking for work and ends when said individual finds a job. Therefore, individuals who cease to be employed are not defined as unemployed unless they declare to be looking for work. However, this also means that a spell does not end merely because someone declares not to be looking any longer, even if the search is never resumed within sample (these people may instead be defined as quitting the search).

The next outcome of interest we include in our analysis is whether an unemployment spell resulted in quitting the search. We define a spell as having ended in quitting the search if the individual declares to not be looking for work for the last 8 weeks or more of observed joblessness, even if the spell ends in employment.

Starting hourly wages and weekly hours at re-employment are directly reported by the survey, however some individuals do not report both and some imputation is necessary. The sample includes 43,201 spells in which re-employment is observed: for 14,966, wage upon re-employment is not reported, and for 11,647 weekly hours upon re-employment are not reported. When income and hours are reported we can impute hourly wage, we do this for 11,806 spells.

All outcomes are defined with respect to individual unemployment spells rather than individual people. For example, if an individual experienced multiple unemployment spells, then we will have two separate re-employment trajectories associated with each spell. However, in most cases an individual experiences only one spell in our sample. All individuals who never experience an unemployment spell are dropped from the sample. By combining the 2001 panel (19,093 spells), 2004 panel (25,175 spells), 2008 panel (26,978 spells), and 2014 panel (12,557 spells) we obtain a total of 83,803 spells for 56,335 individuals. From these, we exclude 6,592 spells that only last one week, and 11,767 for workers that moved to a different state at any point during the duration of the panel. Therefore, our analyses never consider more than 66,473 spells (1,029 are both one week long and move to another state) from 46,437 different individuals. Although most movers moved before or after their spell rather than while unemployed, working hours and wage trajectories could be affected by including movers in the analysis, and to make results more comparable across outcomes we thought it better to exclude them completely.

In addition to the SIPP data, we use data on state and federal minimum wages and on monthly state unemployment rates from the Bureau of Labor Statistics. The monthly data on state minimum wages comes from Vaghul and Zipperer 2016. During the time span covered by the SIPP data described above, the minimum wage takes values between \$5.15 and \$10.50 and was changed for 233 month-state pairs within sample. Of these, 89 changes come from increases of the federal wage floor in 2007, 2008, and 2009; and almost 90% occurred either in January (107) or July (96). The average nominal increase was \$0.54, but the largest changes surpassed \$1.

The different outcomes we study in this paper are better analyzed using different models and samples for each. The analysis of unemployment duration can use a larger sample because it can deal with censored data and use all the observations from a spell, so we do not need to exclude spells that are not observed until re-employment (censored spells). This means that for the duration analysis we consider all spells that last at least two weeks and where the workers remained in the same state throughout the survey. For our Accelerated Failure Time (AFT) model which we discuss in Section 2.3, we include repeated weekly observations for the same spell. This allows us to capture the effect of a change in the

minimum wage by identifying the exact week during a spell where the change took effect. If we only included the observations corresponding to the end of a spell, then this would not adequately identify the effect of the change since the model wouldn't take into account when the change took place.

For re-employment outcomes we chose to limit our analysis to spells that last at most a year, reducing the sample by 7,737 spells. Additionally, it is only possible to include in this analysis uncensored spells with information on hourly wage and weekly hours after re-employment, which leads to a loss of 15,712 spells for the hourly wages, and 4,649 spells for lack of sensible re-employment data[2], about 11% of the uncensored spells (starting hours is not available for 20% of uncensored spells). For the long-term trajectories, the sample size shrinks as we consider re-employment further after a spell ends. Our analysis for trajectories starts with almost 40,000 observations, reaches 24,275 after 52 weeks, and just 12,586 by the end of the second year.

This shrinking sample size we encounter for re-employment outcomes is concerning as it may introduce bias depending on the nature of this decrease. We find that the overwhelming majority of these individuals disappear from our sample simply because the panel in which they participated ended before we could fully observe their trajectory over the considered time frame. We do not find evidence that spells of a certain length cannot be properly compared to spells of the same length that ended a few months earlier (conditional on controls). Furthermore, longer spells are more likely to have ended near the end of a panel simply due to their length. If the re-employment wages/hours following a longer spell differ importantly from those of shorter spells, then we may get biased estimates in our regressions. If for example the wages and hours of those with longer spells have a tendency to be lower, then this could bias our findings upward. To deal with this we limit the observation of long spells. We also conducted a correlation analysis between starting wage/hours and unemployment duration. Our test revealed no correlation between spell length and wage, with or without controls. However, we did find a correlation between unemployment duration and starting hours after the spell, although the coefficient is very small in magnitude. An

[2]We exclude extreme cases from the analysis: wages under $4 and over $300 an hour.

increase in 1% on unemployment duration produces an increase of only 0.05% in starting hours on average.

2.3 Model

We identify both the short-term effect of increasing the minimum wage on the currently unemployed as well as longer-term effects on their labor market outcomes. Short term effects are impacts of the policy change on spell length, chance of abandoning the search, starting wages and starting hours; and the longer term outcomes that we analyze are the impacts on future wages and working hours for the following two years after the unemployment period. For the unemployment duration analysis we use an accelerated failure time (AFT) hazard model that takes better advantage of the data than a linear model. For workers getting discouraged from their search (a binary outcome), we use a binary logit model. All other outcomes are studied using a linear model.

2.3.1 Accelerated Failure Time Model

To determine the effect of the policy on spell length we use an accelerated failure time (AFT) model. We chose this model over a cox proportional hazard because the proportionality assumption, that requires the hazard ratio to be constant, is not met by the data. Instead we decided on an accelerated failure time (AFT) model under a lognormal distribution because the proportionality tests show that a proportional hazard would underestimate the hazard at short spells and overestimate it for longer spells; a lognormal distribution should fit better such data. With it we can establish how initial levels and increases of the minimum wage reduce or increase the length of unemployment. In our regressions we add several controls as well as the two variables of interest: log of initial real minimum wage, to capture the effect of the initial level; and the log difference between initial and actual minimum wage, to capture the effect of changes while unemployed. The hazard model takes the following form:

$$\ln(T_{iswt}) = \beta \times \ln(imw_{ist}) + \delta \times \Delta mw_{ist} + \gamma X_{iswt} + \lambda_s + \sigma_t + \epsilon_{iswt} \tag{2.1}$$

Where $ln(T_{iswt})$ is the natural *log* of the failure time (spell length) for individual spell i in state s and week w of month t (identified uniquely for each year), imw_i is the initial value of the minimum wage at the start of the spell i, and Δmw_{ist} is the difference between the log of the current minimum wage and the log of the initial minimum wage for spell i. This model includes state (λ_s) and month-year (σ_t) fixed effects, as well as individual covariates (X_{iswt}): age and its square, sex, race, number of children under 18 in the household, current unemployment rate (by month, by state), educational attainment, the reason they left their last job, unemployment insurance eligibility, and this program's generosity (as measured by maximum benefits at each point in time).

Traditionally for this literature we would have β be our parameter of interest, the impact of the minimum wage level on unemployment duration, and we would implicitly assume that $\delta = 0$. However, with minimum wages changing during some unemployment spells but not others, the interpretation of the parameter β may be difficult because of the differing effect of these increases, if they have any. Previous research (referred to in the introduction) hints that changes of the minimum wage may in fact have an important short term effect. Therefore we separate initial level of the minimum wage from changes in it during a spell of unemployment. As a result, we get two parameters of interest, β and δ. The former informs us of the impact of a higher initial minimum wage on spell duration; and the latter about how these outcomes are impacted when the minimum wage is increased during an unemployment spell.

2.3.2 Linear Model

Most other outcomes we study can be evaluated using a standard linear model. We use this linear model for log of starting wages, log of starting work hours, trajectories of wages, and trajectories of working hours (attributing zero hours to those not working). However, in the case of whether a worker abandoned the search for employment (defined as not looking for employment the last eight weeks observed) we instead use a logit model, given that the outcome is binary. Nevertheless, the covariates used are very similar to those used in the duration analysis, although in this case we cannot use real time values and instead choose to keep initial values of the covariates (at the start of a spell). The resulting equation is:

$$y_{ist} = \alpha + \beta \times \ln(imw_{ist}) + \delta \times \Delta mw_{st} + \gamma X_{ist} + \lambda_s + \sigma_t + \epsilon_{ist} \tag{2.2}$$

Where y_{ist} is the outcome of interest (wage, and hours). The logit model is specified analogously with the same controls. Another distinction is made for the analysis of trajectories, in which we include both the initial and current unemployment rate by month and state. In other respects the models mirror each other.

2.3.3 Sample Balance - Initial Minimum Wage Levels

A difference-in-difference strategy may be unable to identify the proper effect of initial minimum wage levels if the comparison group is inadequate. We are particularly concerned with the effect a higher minimum wage has on the composition of the unemployed. Research on minimum wages commonly finds some important effect that may lead to a composition change of the unemployed under different initial minimum wage levels. Such a change of composition of the unemployed could compromise our results, as they might instead arise from comparing two different groups and not be informative of the effects of the policy.

To investigate whether changes of composition occurred we performed a correlation analysis by simultaneously regressing the initial minimum wage levels on the following characteristics: The age of an individual ("age"), a dummy indicating whether an individual has completed more than grade 10 ("1st decile be Education"), a dummy indicating whether an individual is a teenager ("Teenager"), the unemployment rate at the start of a spell ("Unemployment"), the number of kids in a household ("Kids"), a dummy for whether an individual was fired ("Fired"), and a dummy for whether an individual quit the labor force ("Quit"). For these regressions we use only time and state dummies for controls in addition to these characteristics. The results are reported in Table 2.1, which we divide into 3 columns for the 3 different sub-samples used for different outcomes as explained in Section 2.2. We can see that for all three sub-samples the point estimate for all variables considered is near zero and non-significant, with the exception of the coefficient for "1st decile by Education" in column (2). However, the relationship seems particularly weak as the significance is not observed in either of the other two samples. We conclude that this correlation analysis does not provide evidence of there being important changes in composition with respect to initial minimum

wage levels. Furthermore, the non-significant estimate for unemployment rate tells us that higher initial minimum wage levels are not associated with local labor market downturns. We also address this issue in a robustness check (see Section 2.5) that exploits federal minimum wage changes in 2007, 2008, and 2009, which are not driven by local conditions.

Table 2.1: Correlation Analysis for Initial Minimum Wage Levels

			Starting:	
	Hazard	Quitting	Wage	Hours
	(1)	(2)	(3)	(4)
Age	-0.00005	0.0001	0.0001	0.0001
	(0.0002)	(0.0002)	(0.0003)	(0.0003)
1st decile	0.005	0.0128*	0.009	0.007
by Education	(0.005)	(0.005)	(0.008)	(0.009)
Teenager	0.0005	0.0007	0.0014	0.0010
	(0.001)	(0.001)	(0.001)	(0.001)
Unemployment	-0.743	-0.712	-0.717	-0.697
rate	(0.693)	(0.644)	(0.660)	(0.644)
Number	0.0002	0.0001	-0.0001	-0.0001
of Children	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Fired from	0.0038	0.0028	0.0062	0.0075
last Job	(0.0028)	(0.0039)	(0.0049)	(0.0057)
Quit	-0.0027	-0.009	0.006	0.008
	(0.010)	(0.013)	(0.013)	(0.013)
N	66,455	50,921	38,229	34,406

Significant at: *** 0.1% ** 1% * 5% + 10%. Displayed are point estimates and standard errors for regressing initial minimum wage levels on covariates at the start of the spell. These regressions include all the same standard covariates used to obtain our main results. Results use data from the 2001, 2004, 2008 and 2014 SIPP panels. Different columns refer to different subsamples used for different short term outcomes (unemployment spell duration, starting wage following re-employment, and starting weekly hours following re-employment). We exclude spells one week long, and spells for individuals who lived in more than one state. Standard errors are clustered at the state level.

Even with no evidence that the sample's composition is changing due to different initial minimum wage levels, there can be differences between high and low minimum wage places that entail a problem for a difference-in-difference analysis. Allegretto, Dube, and Reich 2011 and Dube, Lester, and Reich 2010 make the argument that for any state a proper control state should come from within the same geographical division, because of important differences in labor market dynamics across U.S. census divisions. Consequently, we include division specific time effects and find that our results hold (see Section 2.5).

2.3.4 Sample Balance - Minimum Wage Changes

For a subset of spells in our sample, a minimum wage change occurred while the worker was looking for employment. We observe 3,729 uncensored spells that experience this occurrence, representing 8.6% of uncensored spells (that are between two and fifty-two weeks, for non-movers).

We showed above that there is little evidence to suggest that composition changes are occurring, validating the use of a difference-in-difference model. Further, for minimum wage changes we can argue this is not a concern, because in this case we also compare workers to others that may have started their spell under the same initial minimum wage level.

However, we might be concerned that changing minimum wages leads to other composition changes in the sample of unemployed. Minimum wage changes are typically known in advance and can be anticipated. We may have a problem if individuals or firms adjust in response to a minimum wage change that is about to happen. Table 2.2 compares four different groups we might be concerned about: those with spell length of one and those that moved to a different state, who we exclude from the analysis; those that do not experience a variation of the minimum wage while unemployed, and those that do. Important to highlight, we only see small differences in average wages, and almost no difference in unemployment rates when spells start (particularly for the last two groups). The latter observation being of interest because it contradicts the notion that minimum wage changes may be correlated to macroeconomic conditions, to which we alluded above.

Migration could be a response to changes in the wage floor. When minimum wage is increased in a state, a worker may choose to look for employment in another state as a

Table 2.2: Summary Statistics

	Spell length=1		Movers		$\Delta mw = 0$		$\Delta mw \neq 0$	
	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
Age	32.1	12.2	33.0	12.4	32.8	12.7	34.1	13.7
Women	0.565	0.496	0.518	0.500	0.546	0.498	0.535	0.499
$\leq$ High School	0.185	0.388	0.159	0.367	0.210	0.407	0.214	0.410
Teenager	0.136	0.343	0.108	0.311	0.154	0.361	0.177	0.382
Race:								
White	0.781	0.414	0.734	0.442	0.749	0.434	0.717	0.450
Black	0.139	0.139	0.174	0.379	0.173	0.378	0.192	0.394
Asian	0.035	0.183	0.041	0.198	0.034	0.181	0.036	0.187
Other	0.045	0.208	0.052	0.221	0.045	0.207	0.055	0.227
Unemp. Rate	0.061	0.020	0.066	0.018	0.063	0.021	0.064	0.021
Fired	0.031	0.173	0.028	0.166	0.037	0.189	0.037	0.190
Unemp. Duration	1	0	18.2	21.5	18.6	22.2	51.2	41.7
Δmw	0	0	0.005	0.044	0	0	0.081	0.082
Spells	6,588		11,776		74,332		9,637	
Hourly wage	\$ 11.70	\$ 8.94	\$ 12.77	\$ 11.89	\$ 12.27	\$ 11.71	\$ 13.19	\$ 14.44
Spells	5,351		6,118		53,973		6,910	

Summary statistics for data from the 2001, 2004, 2008 and 2014 SIPP panels. Columns denoted "s.d." give the standard deviation of the variable, for each group. Summary statistics are shown for 4 different groups: for spells 1 week long, for workers who changed states, those who experience a minimum wage change during their unemployment spell, and those who do not experience a minimum wage change during their unemployment spell.

response. In our sample of uncensored spells 6,355 unemployed workers moved to a different state, with 615 of them moving while unemployed. If the worker left because of the increased minimum wage, the group migrating may be a less (more) productive one than the average. If this is the case, ignoring them will likely produce shorter (longer) spell lengths as a response to the minimum wage hike. However, even though the bias could theoretically go in either direction, we would expect that, since we are talking of a higher minimum wage, the migrating worker who moved *because* of this new wage floor is most likely the less productive one. Further, this worker would be looking specifically for lower minimum wages, in order to improve his chances of getting employed. To understand if this is the reason for these workers to be moving, we studied how the new minimum wage for these individuals related to the wage floor in their home state, and found no evidence supporting this theory. In 154 cases the receiving state had a lower minimum wage, and in 176 cases a higher minimum wage, leaving 285 migrations that resulted in no change in the minimum wage that the individual experienced. Overall, a very balanced situation with even a small bias toward movements to higher minimum wage states, which is not consistent with migration motivated by increases in the wage floor. Nevertheless, they are significantly more likely to experience a minimum wage change than the rest of the population, which contributed to our decision not to include movers in the analysis, even if the movement did not happen during the spell. Although they are likely not reacting to minimum wage changes, by moving they create a change in minimum wage that is not comparable in nature to the change experienced by other workers.

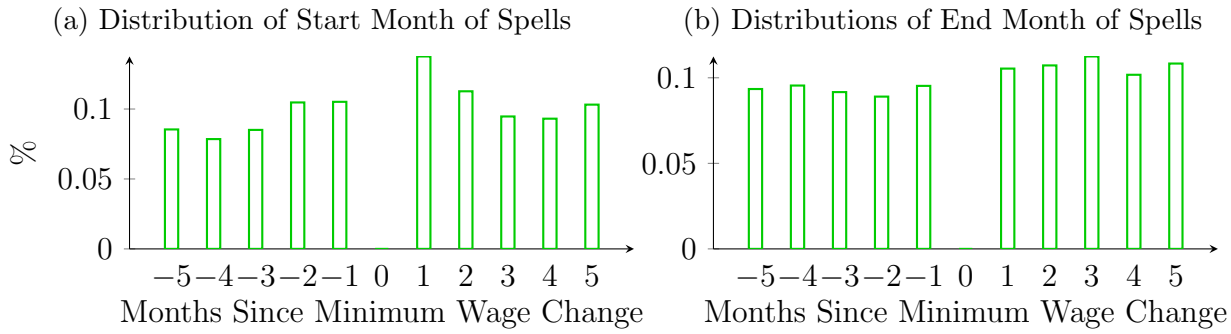
Another important potential source of selection into or out of the sample comes from personal choices, both by firms and workers. To evaluate whether firms or workers are induced by minimum wage changes to modify their behavior in any way that may impact our sample we investigated spells that started or ended around a minimum wage change. Changes in minimum wages are known in advance by both parties, which makes it possible for either of them to anticipate the increase and modify their behavior in potentially important ways. If firms were to lay-off low productivity workers right before the change in policy for example, our sample of workers affected by the policy change would be unbalanced, and our results likely biased. Workers could make decisions that may unbalance our sample too. In their case, it may lead workers to increase their efforts in order to start a job before the

minimum wage is increased, if they believe either that the market will be less dynamic after, or that they will benefit from the increase by finding employment before it happens.

In order to test whether firms or workers were partaking in this type of behaviors we graphed spells starting and ending within five months of a minimum wage increase (Figure 2.1). We only considered changes in January, or July for the years 2007, 2008, and 2009; but overall almost 90% of all changes in the wage floor. The figures show in green bars the proportion of spells started each month from five months before the wage floor increase to five months after.

Figure 2.1a show spells that start within an eleven month window. It seems most spells are starting after the policy is changed rather than before. We do not observe any sharp increase of spells in the months prior to the change either, which suggest to us that firms are not anticipating it by laying off workers. Next to it, Figure 2.1b, similarly graphs spells that ended in the same time frame. In this case there are even less noticeable differences in the bars prior to the minimum wage being changed, that means no evidence that workers are increasing their employment rates in advance of the minimum wage change.

Figure 2.1: Distribution of Start and End Months of Spells Relative to Minimum Wage Changes



Each bar represents a percentage of total spells started/ended within the observation window. For example, a −1 on the x-axis indicates that these spells started/ended within one month before the policy change. Values were computed using data from the 2001, 2004, 2008 and 2014 SIPP panels.

2.4 Results

We will now proceed to the key results of the paper, from which we observe two key take-aways. The first is that the policy's impact on the market is either small or, in cases where it is more important, short-lived. However, the second conclusion is that the policy can be very impactful on the labor supply of certain workers. The policy elicits some interesting immediate responses from the unemployed, but seems overall more impactful by virtue of its changes. We will start by discussing overall effects, continue the analysis with a focus on the least productive workers, and close the section with a look at differences in responses by each sex.

2.4.1 Overall Effects

Possibly the most novel part of our analysis is the separation of initial minimum wage levels at the start of a spell (policy level) from minimum wage changes during a spell (policy change), which means we decompose the effect of the minimum wage on the unemployed into two elements. However, this decomposition is in fact necessary to identify the impact of minimum wage levels if changes to the policy have a non zero impact of their own on unemployed workers' outcomes. Furthermore, it allows us to distinguish the short term effects of a minimum wage policy change from the long-term sustained effects of minimum wage differences between locations. Therefore, by separating the two components of the minimum wage (level and change) we can better understand how minimum wages affect unemployed workers.

Table 2.3 presents the overall results for our 'short-term' outcomes: unemployment spell duration, quitting the search, starting wage, and starting hours following re-employment. In this table $\ln(imw)$ denotes the log of initial minimum wage levels and Δmw denotes the percentage change (log-difference) in minimum wage during a spell; both are expressed in such a manner that their coefficients can be understood as elasticities. Column (1) displays the coefficients and standard errors for the accelerated failure time hazard model, with unemployment duration (Spell) as the outcome variable. Column (2) displays the results for

Table 2.3: Effect of Minimum Wage on Short-term Outcomes

	(1)	(2)	(3)	(4)
			Log Starting:	
	Spell	Quit	Wage	Hours
ln(imw)	-0.079	3.098***	-0.001	-0.115^{+}
	(0.118)	(0.573)	(0.051)	(0.063)
Δmw	1.412***	11.21***	-0.018	-0.462***
	(0.327)	(0.624)	(0.066)	(0.125)
Intercept	2.902***	-5.897***	1.001***	3.015***
	(0.222)	(0.939)	(0.123)	(0.107)
Spells	66,650	50,921	38,229	34,406

Significant at: *** 0.1% ** 1% * 5% $^{+}$ 10%. Actual N for AFT model is 1,345,077. All models control for state and time (month-year) fixed effects, and a complete set of covariates. All analyses exclude spells lasting only one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level.

our logit model for quitting the job search (Quit). Columns (3) and (4) display the linear regression results for log of starting hourly wage and log of starting hours (weekly) following re-employment, respectively.

Table 2.3 confirms our suspicions: changes to the minimum wage have an impact of their own on unemployed workers. With respect to unemployment duration (column 1), policy level has no statistically significant impact, whereas within-spell changes have a significant and relevant effect. More specifically, a relatively standard 10% change of the minimum wage during unemployment is associated with an approximately 14% increase in unemployment duration, on average. For the median unemployed person in our sample (spell length of 15 weeks for the sub-sample used for the AFT model) this would translate to a duration average increase of approximately 2.1 weeks. For such a worker, assuming an hourly wage of $10 working full time (40 hours a week), this translates to a cost of approximately $840 in earnings.

Although higher initial minimum wages do not seem to make unemployment spells longer, they do affect unemployed workers' quitting behavior in the same general direction as minimum wage increases (Table 2.3, column 2), albeit substantially lower magnitudes[3].The coef-

[3]However, the smaller coefficient can be misleading in certain cases. For example, during our observation period California's minimum wage was at a minimum 10% higher than Texas', it was on average 22% higher,

ficient on Δmw is almost four times larger than the coefficient for initial level. The former translates to an average increase in quitting probability of 4.5 percentage points for a 10% increase in minimum wage, with the effect for $\ln(imw)$ being less than a third of that. However, it is important to keep in mind that while changing the minimum wage affects only those currently unemployed, minimum wage levels impact every unemployed worker. In the extreme example of Texas and California, with a minimum wage difference averaging 22% during the observation window, the effect of this difference is to increase discouragement by 2.8 percentage points on average, about 20% of the sample baseline, and this difference would be relevant for every unemployed worker in California.

In the presence of no impact on their unemployment duration this is an interesting effect. A possible explanation may be that quitters are more likely to have children and to have a higher household income, in our sample. It is possible that these quitters are in fact more likely to be secondary earners within their household. If this is the case, then their incentive to find a job may be diminished when a minimum wage change comes into effect if they believe it will make the job search harder (in fact, their quitting may be the reason the search is not becoming harder after all).

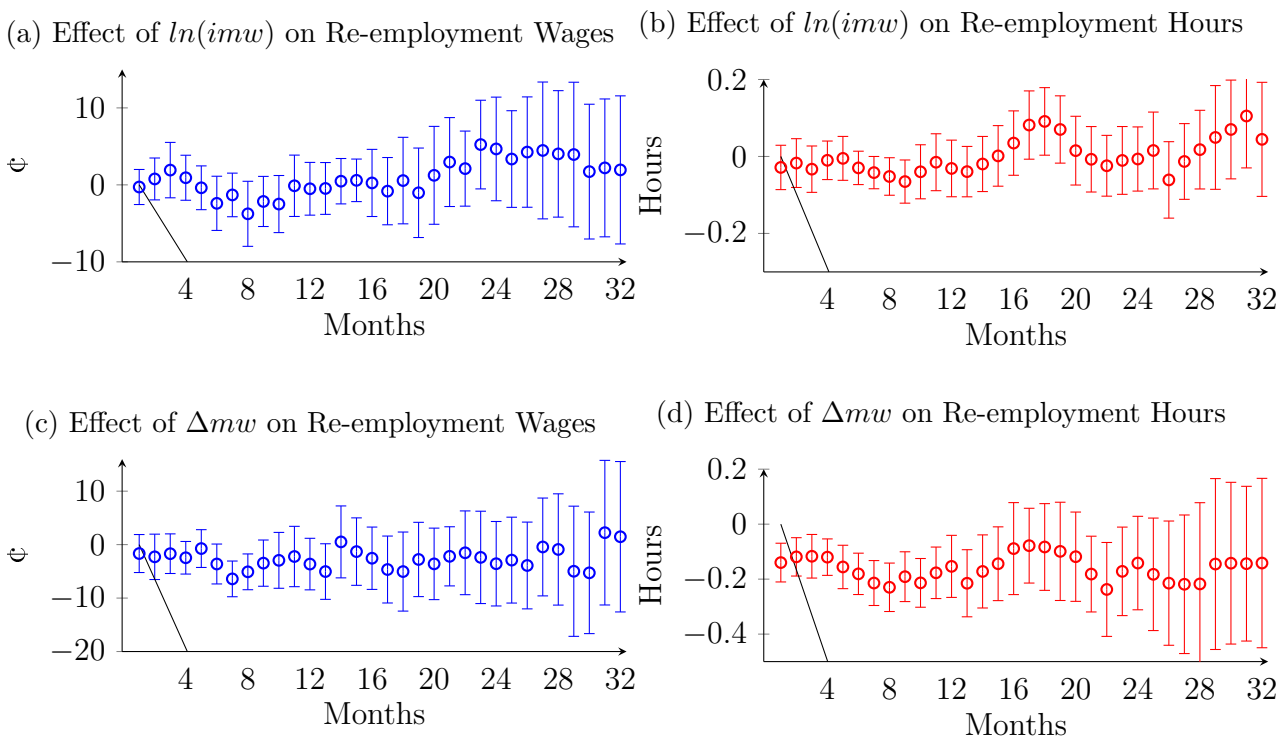
The last two columns of this table look at re-employment outcomes. Columns (3) and (4) of Table 2.3 show the impact on the log of starting wage and log of starting working hours, following re-employment. Starting wage is not significantly affected by the policy, through level or change. Starting hours on the other hand do respond to the policy. Both initial level and minimum wage changes have significant effects. As for previous results, the coefficient for within-spell changes is about four times greater than that for initial levels. A 10% increase in the minimum wage during one's spell would be associated with a reduction in weekly hours of 4.6%, while a similar difference in minimum wage level would only translate into a 1.4% difference. For our exemplary full time worker earning \$10 an hour these coefficients would create a loss over a month of \$18 due to the level difference (\$40 for the Californian worker when compared to the Texan) and an extra \$74 if the minimum wage is increased while he is unemployed.

and at its peak the difference reached 31%. If we take the average difference, the effect on quitting for levels in this case is about 61% of the effect attributable to a 10% minimum wage increase.

In order to get a complete picture on how individuals are affected, we need to follow them for the coming months after re-employment. Figure 2.2 does this using four graphs which describe the effect of minimum wages on hourly wages and weekly hours at different time periods following re-employment. Graphs 2.2a and 2.2b describe the effect of $\ln(imw)$ on wages and hours respectively and graphs 2.2c and 2.2d do the same for minimum wage changes. Each graph shows the coefficients and corresponding 95% confidence intervals for 4 week intervals, starting 4 weeks after re-employment and reaching up to 144 weeks after re-employment. We assign wages and hours of zero to those that lose their jobs again after re-employment (however, we include in Figure 2.3 conditional trajectories that only consider workers with positive wages and hours). We find in these graphs no evidence that long-term wages are affected on average by either feature of the minimum wage policy. There are some periods with significant coefficients, but we believe they are more likely to be the result of type one error than actual wage responses, due to their transitory nature. When we concentrate on weekly hours we find that again initial minimum wage level has no effect, but in the long-term there is a sustained negative impact as a result of minimum wage changes. These results for initial minimum wages confirm that the previously discussed significant effect on starting hours is not likely indicative of a non-negligible effect on the unemployed. The significant negative impact of minimum wage changes on hours is sustained for 15 months before becoming statistically insignificant for most of the remaining periods. It should be noted however that the coefficients becomes less precise due to decreasing sample size after this point, but the point estimates remain at similar levels in the vicinity of -0.2 even more than two years after re-employment. For the first year the coefficients average -0.17; a worker who would normally work 40 hours per week, that experienced a 10% increase of the minimum wage while unemployed will work over his first year of employment 88 less hours on average (1.7×52), with a total cost for our case study (a \$10 an hour full time worker) of \$880.

It is also of interest to analyze the effect of the policy on re-employment outcomes conditional on working. Figure 2.3 displays the same graphs excluding individuals for whom hours or wages are zero at any point after re-employment (i.e. lose their jobs). We can see that our conclusions change little, however the impact of minimum wage changes on weekly

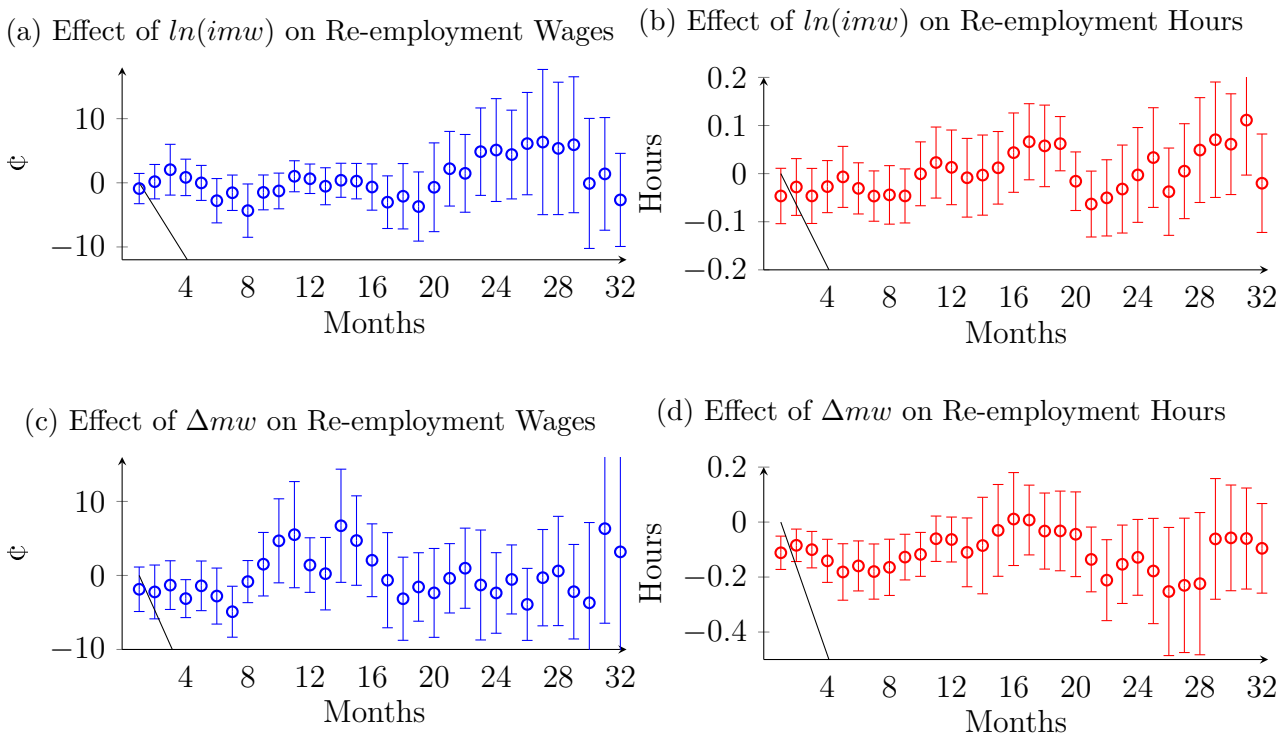
Figure 2.2: Overall Effects on Re-employment Wages and Hours



All graphs include 95% confidence intervals around the estimated coefficients. Separate linear models were fit with the outcome being the wage or hours at a specific time period after re-employment. All models control for the same set of covariates used throughout the paper, plus current unemployment rate. We exclude spells longer than 52 weeks, equal to one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level.

hours disappears a little sooner now, after 10 months (and additionally the point estimates after this point come closer to zero). This suggests that maybe the persistence of the effect beyond this period has more to do with workers being pushed out of the labor market after re-employment than in employed workers working for shorter hours. On average, point estimates are smaller now, -0.12 during the first year, and impact workers for a slightly shorter period of time. Overall the cost to our hypothetical worker would be 43 hours for the year, or \$430, half the unconditional effect for the first year.

Figure 2.3: Overall Effects on Re-employment Wages and Hours - Conditional on Working



All graphs include 95% confidence intervals around the estimated coefficients. Separate linear models were fit with the outcome being the wage or hours at a specific time period after re-employment. All models control for the same set of covariates used throughout the paper, plus current unemployment rate. We exclude spells longer than 52 weeks, equal to one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level.

In quick summation, we get to a total cost for our unemployed worker that experiences a level that is 10% higher than his benchmark, and a change of 10% while unemployed, of \$1720 over a year. Considering his total income of \$20,800, the policy creates a 8.2% reduction in income for him and his family.

Up to this point we have implicitly assumed that treatment deactivates immediately after the minimum wage change occurs (i.e. if a worker becomes unemployed just a week after the increase, Δmw is zero). However, it would be sensible to presume the effects of the increase take a little time to diffuse entirely. For this reason we ran a test in which if a worker became unemployed in the vicinity of a minimum wage increase, we replaced actual minimum wage levels and changes with the recently past level and change. The results are reported in Table 2.4, and the only difference to our standard regression is that we modified minimum wage level for selected unemployed workers, and created four new variables to assign the correct Δmw to each group of workers. Our first conclusion from this table is that the effect of minimum wage increases diffuses almost immediately. In fact, it seems that workers respond to the policy right after it comes into effect. Going back momentarily to Figure 2.1 we see that there is an important surge in spells starting right after the minimum wage is changed, which is not accompanied by a similarly important surge in spells ending (size is fairly similar between the groups) at the same time. Back to Table 2.4, the group becoming unemployed the same month the minimum wage was changed (but after the change) is unemployed for less time and significantly less likely to quit the search. For this reason, we believe that minimum wage increases motivate people to enter the labor force. These positive results suggest that this group of people is also more productive than the average unemployed worker, which would be in line with standard theory. This analysis generally suggest as well that the effect of minimum wage changes dissipates fast and has no effect on workers losing their positions after the change occurs.

In summary, our overall results indicate that minimum wage changes act as an important, though passing, shock to labor supply (and demand maybe). At the same time, we find no important monetary costs associated to higher levels of the minimum wage for the average unemployed worker. However, we do have evidence that higher minimum wage levels lead to some important behavioral responses, a subject that may be worth investigating further with data more appropriate to this end. Minimum wage increases are different, and create both monetary costs for workers and strong behavioral responses. We estimate the total monetary cost for the average worker to be quite high. Minimum wage changes also seem

Table 2.4: Dissipation Test for Δmw

	(1)	(2)	(3)	(4)
			Log Starting:	
	Spell	Quit	Wage	Hours
ln(imw)	0.013	0.112**	0.002	-0.007
	(0.008)	(0.037)	(0.004)	(0.004)
Δmw	1.623***	10.64***	-0.020	-0.430**
	(0.330)	(0.684)	(0.063)	(0.124)
Same month	-1.130***	-7.73***	0.070	0.183
	(0.338)	(1.549)	(0.235)	(0.175)
Next month	0.406	-0.744	-0.132	-0.072
	(0.369)	(1.595)	(0.183)	(0.222)
2nd month	0.499	-1.411	0.132	0.032
after change	(0.338)	(1.877)	(0.201)	(0.188)
3rd month	0.024	-2.372	-0.096	0.361^{+}
after change	(0.400)	(1.415)	(0.194)	(0.196)
Spells	66,650	50,921	38,229	34,406

Significant at: *** 0.1% ** 1% * 5% $^{+}$ 10%. Actual N for AFT model is 1,345,077. All models control for state and time (month-year) fixed effects, and a complete set of covariates. All analyses exclude spells lasting only one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level.

to push workers off the labor force indefinitely, a development that can result in even worse consequences for workers and their families.

Most of our results are a consequence of what we believe is a shock to the market, not necessarily a movement to a new equilibrium. For this reason it is not straightforward that the affected population be necessarily the target of the policy. Additional uncertainty in the labor market could easily affect most workers, even if the policy itself is not relevant for them directly. We will study this matter in the following section.

2.4.2 Heterogeneity

In this section we look at how the effects on the different outcomes differs by education and by sex. Education is the best way to proxy for productivity in our sample in our opinion. We created a binary variable to indicate whether an individual has completed more than grade 10 (which gives us a 10/90 division approximately, i.e. 10% of the sample has less than tenth grade). For the purpose of this section we refer to the people who have completed more than grade 10 as "high educated" and the people who have not as "low educated".

Although historical information on wages is available on the SIPP, and would be an ideal way to distinguish workers' productivity, there are issues with how this information is reported. A very important consideration is the fact that over half the sample does not include information on past wages. Furthermore, also important is the fear that the information may be given less accurately than other characteristics that are both simpler to report and perceived to be less sensitive. Nevertheless, we still made use of past wage distributions to help us determine which demographic characteristics mapped workers better to productivity. We found that education level was one of the strongest determinants in our sample with respect to past wages. Those that did not surpass grade 10 averaged \$11.1 per hour while workers that advanced further averaged \$15.1 per hour, a difference 60% larger than that between men and women or black and whites. The dispersion for the less educated workers is also relatively small, with approximately 75% earning an hourly wage under \$12. For this reason we will concentrate on discussing in this section the results by educational attainment.

Table 2.5: Effect of Minimum Wage by Education Attainment

	(1)	(2)	(3)	(4)
			Log Starting:	
	Spell	Quit	Wage	Hours
ln(imw)	-0.056	3.122***	-0.0005	-0.130*
	(0.119)	(0.568)	(0.048)	(0.063)
ln(imw)× low educ.	-0.372***	-0.280	-0.008	0.261***
	(0.100)	(0.289)	(0.062)	(0.056)
Δmw	1.265***	11.22***	-0.045	-0.483***
	(0.314)	(0.569)	(0.074)	(0.125)
Δmw× low educ.	2.076^{+}	0.027	0.413	0.268
	(1.147)	(0.907)	(0.439)	(0.487)
Intercept	3.524***	-5.435***	1.012***	2.578***
	(0.306)	(1.144)	(0.141)	(0.143)
Spells	66,650	50,921	38,229	34,406

Significant at: *** 0.1% ** 1% * 5% $^{+}$ 10%. Actual N for AFT model is 1,345,077. All models control for state and time (month-year) fixed effects, and a complete set of covariates. All analyses exclude spells lasting only one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level.

Table 2.5 shows the results for our short-term outcomes (spell length, quitting, log of starting wage, and log of starting hours) with education interactions. As we expected, the table suggest more heterogeneity by policy level than by policy changes, although there is weak evidence of some heterogeneity in the latter as well. A higher minimum wage level translates to some mild gains for the least productive worker. A difference of 10% would make a worker's unemployment period 4% shorter and would not reduce starting hours at re-employment. The spell for an average low productivity worker would become 0.6 weeks shorter (sample median of 15 weeks), giving our example worker a benefit of $240. However, for more productive workers there is still a small cost associated to a higher minimum wage level, as they would work less hours at re-employment on average.

If we extrapolate this finding throughout the wage distribution, it could mean that the demand for higher-skilled labor is more elastic in industries affected by the minimum wage. For example, a fast food restaurant may not be able to reduce hours for their cashiers in the short term if the same number of customers needs to be served; but they may be able

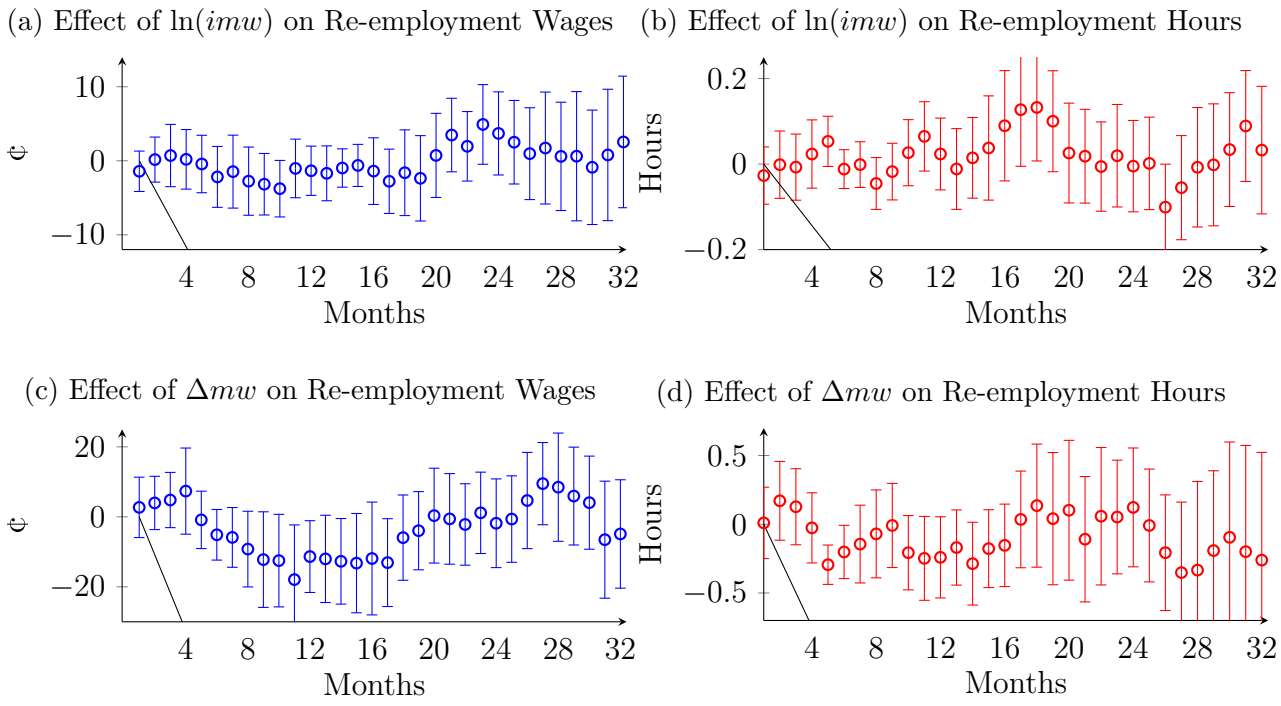
to reduce the hours of the managerial staff. This is very interesting as it tells us it is not necessarily just workers near minimum wage that are affected by the policy.

With respect to minimum wage changes, we find only very weak statistically significant heterogeneity in unemployment duration. However, the point estimate for the interaction between education and Δmw (column 1) is very large. If taken seriously this would imply that lower educated workers experience much longer spells as a result of minimum wage hikes (at least in the short term following a change) compared to their more educated counterparts. If this coefficient is taken at face value, it would mean that increasing the minimum wage by 10% would make the spell of the least productive workers by 33%, almost eight extra weeks of unemployment for the average unemployed worker and almost 15% of what they could earn working full time during a year.

Figure 2.4 shows the effects of the policy on long term hours and wages for the low educated, with the graphs having the same structure as Figure 2.2. Figure 2.5 shows the analogous results for the rest of the workers. In both figures, graphs a) and b) show the effect of minimum wage levels on wages and hours, respectively. We find no significant effect or heterogeneity, suggesting no heterogeneity by skill and confirming the null effects for both groups. Also in both figures, graphs c) and d) show the effect of minimum wage increases on wages and hours, respectively. Both graphs c) exhibit very similar patterns, suggesting again no heterogeneity in the effect the policy has (or does not have) on wage trajectories. However, graph d) shows that the policy may be bringing worse consequences in terms of hours worked for the more educated workers. In fact, the least educated workers seem unaffected by minimum wage changes, in terms of hours worked over the next two years. Nevertheless, though it may be judged as good news that workers that left their education before ending grade 10 are not being injured by the changes of the minimum wage, the other 90% is not necessarily rich. Indeed, 50% of them earn less than \$9 in our sample (and under 1.5 minimum wages). For one of these workers the policy would bring a reduction in weekly working hours of 1.7 hours for at least a year (after a year we cannot distinguish the coefficient from zero because of decreasing precission, but the coefficient remains at similar values), costing our exemplary worker[4] \$884.

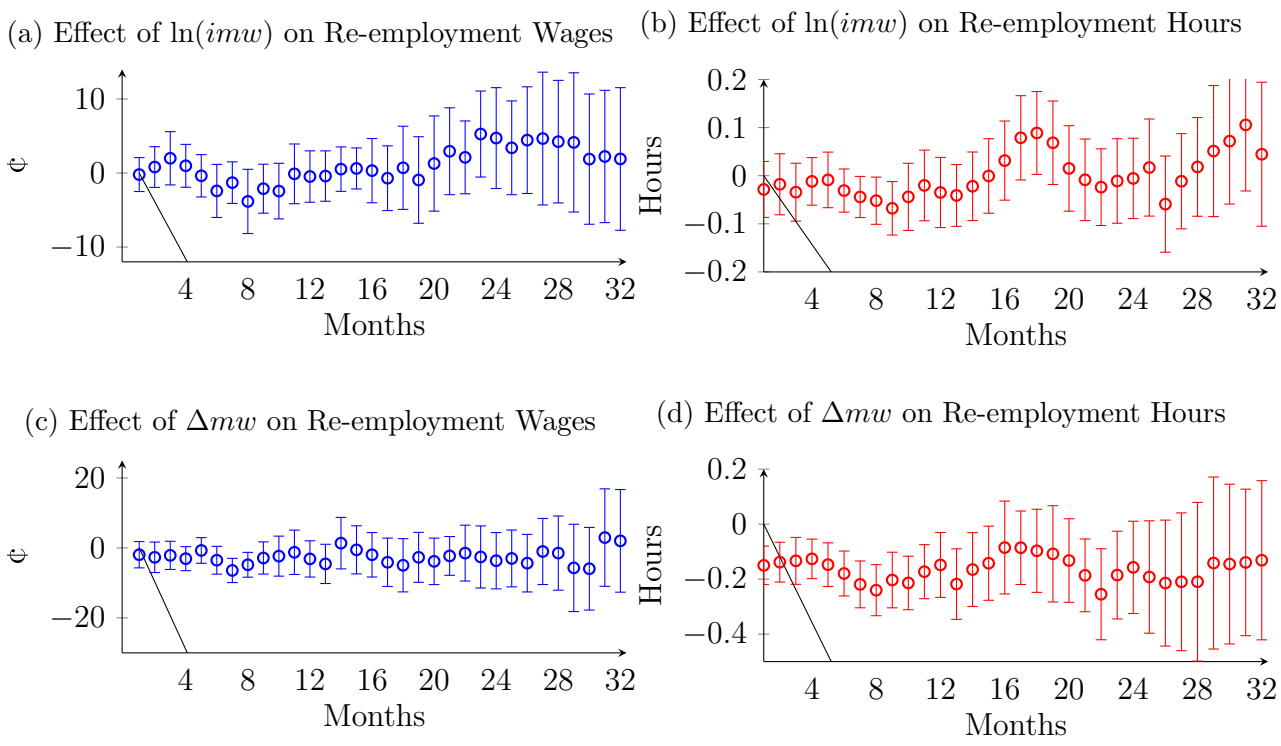
[4]Works full time for \$10 an hour.

Figure 2.4: Effect on Low Educated Unemployed Workers



All graphs include 95% confidence intervals around the estimated coefficients. Separate linear models were fit with the outcome being the wage or hours at a specific time period after re-employment. All models control for the same set of covariates used throughout the paper, plus current unemployment rate. We exclude spells longer than 52 weeks, equal to one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level.

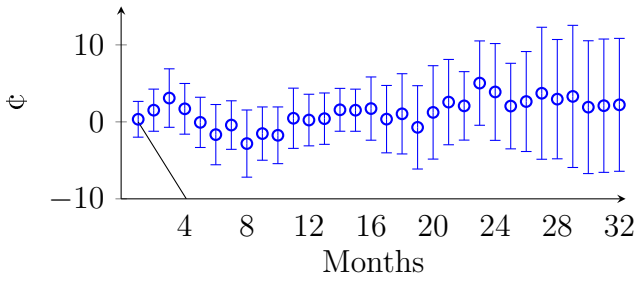
Figure 2.5: Effect on High Educated Unemployed Workers



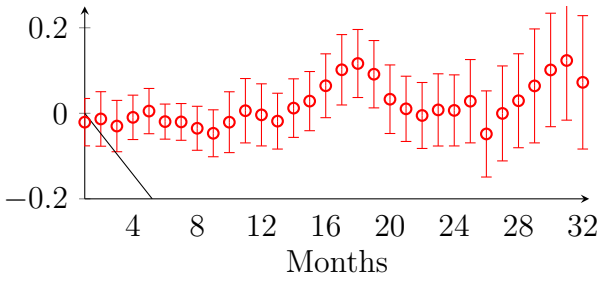
All graphs include 95% confidence intervals around the estimated coefficients. Separate linear models were fit with the outcome being the wage or hours at a specific time period after re-employment. All models control for the same set of covariates used throughout the paper, plus current unemployment rate. We exclude spells longer than 52 weeks, equal to one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level.

Figure 2.6: Effect on Female Unemployed Workers (by educational attainment)

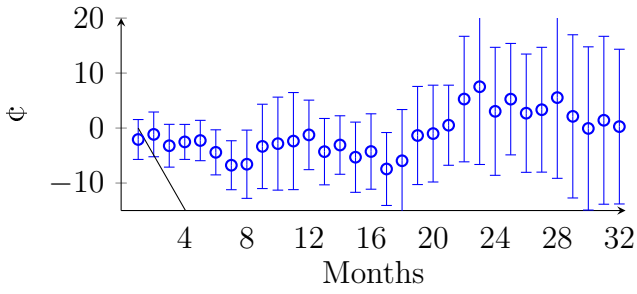
(a) Effect of ln(imw) on Re-employment Wages



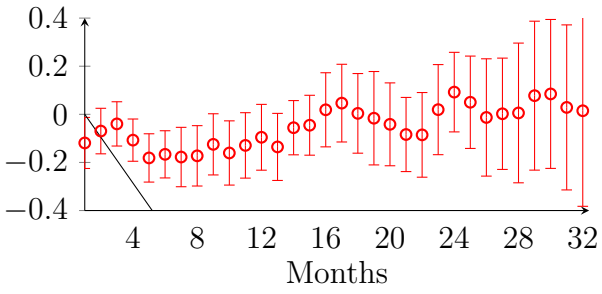
(b) Effect of ln(imw) on Re-employment Hours



(c) Effect of Δmw on Re-employment Wages

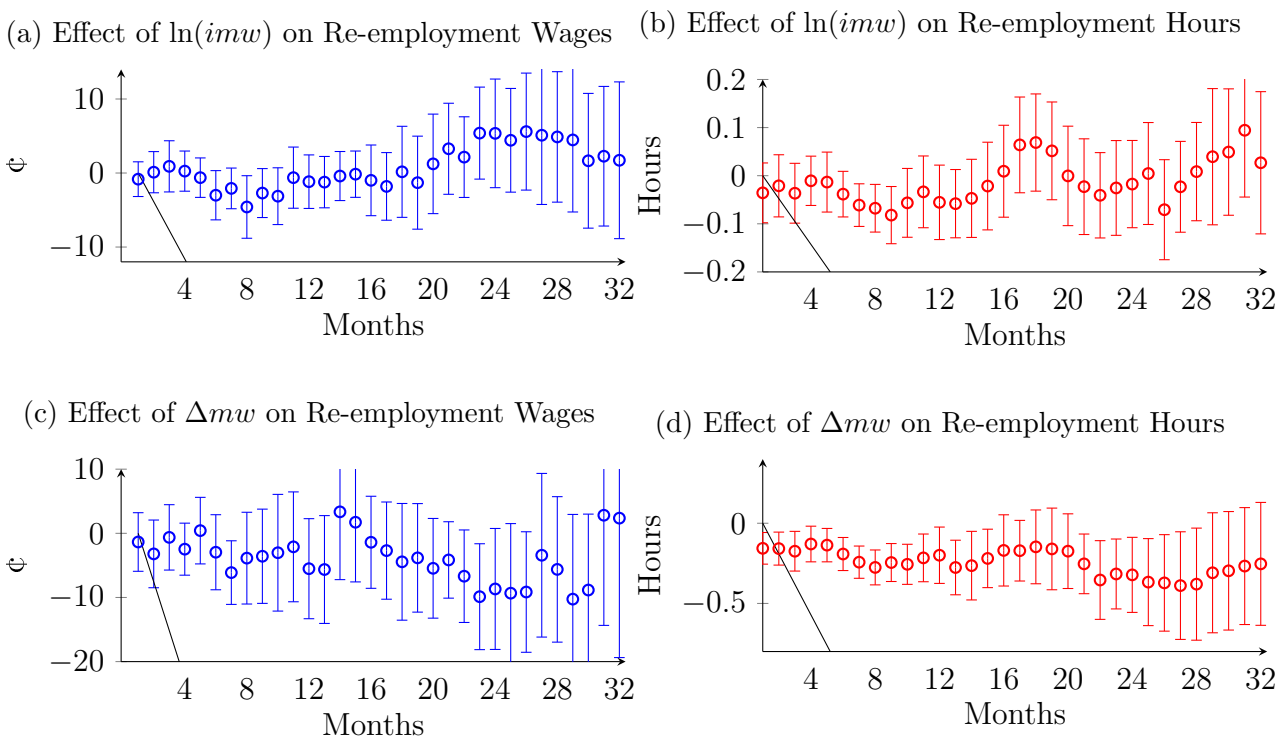


(d) Effect of Δmw on Re-employment Hours



All graphs include 95% confidence intervals around the estimated coefficients. Separate linear models were fit with the outcome being the wage or hours at a specific time period after re-employment. All models control for the same set of covariates used throughout the paper, plus current unemployment rate. We exclude spells longer than 52 weeks, equal to one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level.

Figure 2.7: Effect on Male Unemployed Workers



All graphs include 95% confidence intervals around the estimated coefficients. Separate linear models were fit with the outcome being the wage or hours at a specific time period after re-employment. All models control for the same set of covariates used throughout the paper, plus current unemployment rate. We exclude spells longer than 52 weeks, equal to one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level.

Summarizing, we find that the policy level has a positive impact for the least productive workers, who are unemployed for 0.6 weeks less and so gain $240. However, it is still producing a negative behavioral response with consequences that are much harder to measure. A minimum wage level 10% higher will also have a negative impact in starting hours for the other ninety percent of workers, with a total cost of 1.6% of their monthly income (if we suppose it lasts for a month), $25 for our chosen example. The heterogeneity analysis for minimum wage changes shows low levels of heterogeneity, but it does suggest a potentially damaging effect on unemployment duration. If we take the coefficient at face value a minimum wage increase of 10% would lead to an unemployment period that is 33% longer for workers in the first decile by education, with a cost to them of 15% of their yearly earnings, $3,090 for our worker receiving $10 an hour.

We also wanted to consider heterogeneity based on sex, because women may have a slightly different relationship to the market and we found that it is particularly large. Table 2.6 displays the results for which we include interactions with a binary indicator for women. The first thing to note on this table is how much results differ for women with respect to policy levels. Women's unemployment duration, starting wages and starting hours respond to the policy significantly. Although relatively small, the effect on unemployment duration would shorten the average spell by half a week when the minimum wage level is 10% higher, creating a benefit of $200 for the full time worker with an hourly wage of $10. Women's wages also increase with respect to those of men, although very modestly (over a month it would amount to a benefit close to $15 for our chosen parameters). Furthermore, her hours decrease with respect to those of men, although very modestly too.

Women's response to minimum wage changes are less distinguishable from that of men, at least in the short term. The only noticeable difference is that wage becomes lower for women when their unemployment coincides with a minimum wage change. However, the difference is not large and the coefficient barely significant.

Turning now to the long-term effects on wages and hours after re-employment, Figure 2.6 shows the effects of both the log initial minimum wage and Δmw on hours and wages for women. Figure 2.7 displays the corresponding results for men. Comparing the two figures we see that neither group appears to be consistently affected with respect to wages by either

Table 2.6: Effect of Minimum Wage by Sex

	(1)	(2)	(3)	(4)
			Log Starting:	
	Spell	Quit	Wage	Hours
$\ln(imw)$	0.008	3.194***	-0.045	-0.078
	(0.121)	(0.580)	(0.047)	(0.067)
$\ln(imw)\times$ Women	-0.187*	-0.199	0.094***	-0.077^{+}
	(0.074)	(0.207)	(0.026)	(0.044)
Δmw	1.440**	10.89***	0.118	-0.462***
	(0.503)	(0.667)	(0.093)	(0.113)
$\Delta mw\times$ Women	-0.053	0.704	-0.305^{+}	0.002
	(0.660)	(0.848)	(0.154)	(0.232)
Intercept	2.743***	-6.054***	1.076***	2.949***
	(0.240)	(0.930)	(0.122)	(0.112)
Spells	66,650	50,921	38,229	34,406

Significant at: *** 0.1% ** 1% * 5% $^{+}$ 10%. Actual N for AFT model is 1,345,077. All models control for state and time (month-year) fixed effects, and a complete set of covariates. All analyses exclude spells lasting only one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level.

aspect of the policy. with respect to hours, clearly men are getting more of a negative impact as a result of the policy. However, this is only for minimum wage increases, as minimum wage level has the same effects for both groups. Furthermore, men hours do not seem to recover whithin our observation window even though at times it becomes statistically indistinguishable from zero. During the 32 months it averages -2.5 weekly hours, 6.25% of a full time contract. Over one year our worker would lose $1,300, and this would double over the course of the second year. Women on the other hand seem to be affected for a much smaller period of time (6-7 months) by within-spell changes. Still, over six months the reduction averages -1.2 hours and would equate to a cost of $312.

In summary, we find results for women that seem better than those for men. However, even the one positive response observed to higher minimum wage levels among the least educated workers and female unemployed workers brings with it a negative behavioral response that increases substantially the probability a worker will stop looking for work, which can have much larger consequences over the long time.

Our simple accounting excercise done for our exemplary worker can allow us a crude comparison between benefits and costs of the policy. We found that the cost to the average worker of being unemployed when the minimum wage is raised 10% is $950, and the benefit to workers in the first decile by productivity of a minimum wage level higher by 10% is $240. Now, the former only applies to unemployment periods that coincide with minimum wage changes, that is 8.4% of spells that last at most a year (14% without the time restriction), which would bring the cost per unemployed worker to $80 ($133 considering longer spells, which may be a way to factor in quitting behavior). On the other hand, the higher level can potentially benefit all workers in the first decile by education, so that would mean a benefit by worker of $24. Of course, the distribution of this cost/benefit is anything but irrelevant, but the fact that the cost is almost three times the benefit is not irrelevant either. In fact, while we are talking of distribution, we should care particularly about quitting behavior. These workers are substantially less likely to find employment than non-quitters, and the distributional effects of this failure to find employment can be much larger than this monetary penalties. The effect of the policy on quitting behavior seems unequivocally negative and strong.

2.5 Robustness Checks

In this section we show that our results are robust to a number of potential criticisms. First, we show through inverse probability weighting that our results are not driven by the fact that individuals with longer spells are more likely to experience a change. Additionally, we show the results seem to be robust to the inclusion of division specific time effects as suggested in Allegretto, Dube, and Reich 2011. We also check that the results are not being driven by federal minimum wage changes which suggests that our results are not driven by local labor market conditions. Lastly, we allow the definition of quitting the search (originally defined as 8 weeks of not looking) to vary. This is to show that our results are not sensitive to the choice of definition.

The first potential issue we decided to address is the possibility of a bias in our results. In particular, we are concerned that the probability of experiencing within spell variation of

the minimum wage is increasing in unemployment duration. To address this we used inverse probability weighting conditional on spell length for the untreated group (those that don't experience a change). Since we can not know the counterfactual spell length for those that experienced a change, we can not assign them a propensity score. Because of this limitation we must use the inverse probability weighting estimator for the average treatment effect on the treated (ATT), which only requires that we weight the untreated observations. We calculated propensity scores by panel, estimating the probability of experiencing a minimum wage increase during unemployment given spell duration. A longer spell will have a higher propensity score, which will tend to increase its weight compared to other spells. Scores are assigned assuming that an individual is randomly assigned to a state/time pair within a panel. However, this procedure can only be used with uncensored spells, because we need to know the spell length to attribute propensity scores, which is particularly problematic for search behavior outcomes because we are forced to exclude any spell that does not end on re-employment. Furthermore, because this estimation can make use of about half of the sample, we consider these results solely for the purpose of testing the existence of bias in our original findings. Table 2.7 shows our results with this propensity score weighting methodology. The effect of within-spell variation on unemployment duration remains significant, although the effect suggests a bias does exist as the magnitude is reduced notably. The coefficient on Δmw in column (1) suggests the impact on spell is roughly half what it was in our original results. However, the negative response for hours remains close to what it was before. The effects on quitting is no longer significant. It should also be noted that this procedure approximates the average treatment to the treated (ATT). It should also be noted that since the original methodology finds the average treatement effect (ATE), we cannot be completely sure that the only reason for this difference is bias.

Methods are the object of important discussions within the minimum wage literature. We use a nation-wide approach that has been questioned by many important researchers as unable to account for local heterogeneity. Allegretto, Dube, and Reich 2011 show that disemployment findings were not robust to adding census division time fixed effects to the model most commonly used, and that we use in our analysis. Table 2.8 shows the effect of instead using census division time fixed effects. For both initial levels and within spell

Table 2.7: Regressions with Inverse Propensity Score Weighting

	(1)	(2)	(3)	(4)
			Log Starting:	
	Spell	Quit	Wage	Hours
$\ln(imw)$	0.012	1.000^{+}	-0.209*	-0.308**
	(0.071)	(0.598)	(0.088)	(0.109)
Δmw	0.648***	0.547	0.076	-0.350**
	(0.178)	(0.743)	(0.069)	(0.158)
Intercept	0.732***	-0.717	1.435***	3.142***
	(0.174)	(1.070)	(0.177)	(0.190)
Spells	43,190	43,190	38,229	34,406

Significant at: *** 0.1% ** 1% * 5% $^{+}$ 10%. All models control for state and time (month-year) fixed effects, and a complete set of covariates. All analyses exclude spells lasting only one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level

changes, we find that the effects are unchanged in direction, with some differences in magnitude. The effect of within-spell changes on hours is notably larger, and the effect of initial levels and changes on quitting is larger. The effect of within-spell changes on unemployment duration is also larger. Therefore our results are not driven by the choice of fixed effects in our models.

We also explored potential correlation between local labor markets and minimum wage levels or changes and the effect of differences in preferences for the policy that may lead to selection issues. Table 2.9 evaluates a potentially different effect from local state increases and general federal changes of the minimum wage. This table shows that federal minimum wage changes do not have a significantly different effect from other minimum wage changes, providing no evidence that this potential issue is causing any bias in our analysis. Therefore, local market conditions are likely not an important concern when interpreting our findings.

Finally, we address the possibility that our arbitrary choice of eight weeks not looking at the end of the observed spell may be driving the results on unemployed workers getting discouraged in their job search. Table 2.10 shows the results of choosing ten weeks instead of eight for initial minimum wage level and for minimum wage variation within the spell. We see in both cases only small changes in the coefficients, that however remain in other ways the same: same sign and overall meaning. We conclude from this that our choice of eight

Table 2.8: Regressions with Census Division Time Fixed Effects

	(1)	(2)	(3)	(4)
			Log Starting:	
	Spell	Quit	Wage	Hours
ln(imw)	-0.081	5.449***	-0.055	-0.101
	(0.170)	(0.825)	(0.077)	(0.077)
Δmw	1.728***	13.07***	-0.115	-0.601***
	(0.284)	(0.934)	(0.070)	(0.124)
Spells	66,650	50,938	38,229	34,406

Significant at: *** 0.1% ** 1% * 5% + 10%. Actual N for AFT model is 1,345,077. All models control for state and time (month-year) fixed effects, and a complete set of covariates. All analyses exclude spells lasting only one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level

Table 2.9: Regressions with Federal Minimum Wage Change Interaction

	(1)	(2)	(3)	(4)
	Spell	Quit	Log Starting Wage	Log Starting Hours
Δmw	1.302***	10.56***	0.027	-0.553***
	(0.411)	(0.623)	(0.078)	(0.157)
×Federal	0.368	4.077**	-0.192	0.372
	(0.667)	(1.396)	(0.171)	(0.254)
Spells	66,650	50,938	38,229	34,406

Significant at: *** 0.1% ** 1% * 5% + 10%. Actual N for AFT model is 1,345,077. All models control for state and time (month-year) fixed effects, and a complete set of covariates. All analyses exclude spells lasting only one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level

Table 2.10: Regressions for Quitting Search Defined Around 10 Weeks

	(1) 8 Weeks	(2) 10 Weeks	(3) 12 Weeks
ln(imw)	3.079***	3.067***	3.218***
	(0.291)	(0.281)	(0.330)
ln(imw)×:			
1st decile by Education		-0.327	
		(0.333)	
Women			-0.286
			(0.212)
Δmw	11.68***	11.42***	11.32***
	(0.546)	(0.517)	(0.644)
Δmw× :			
1st decile by Education		1.930	
		(1.319)	
Women			0.790
			(0.804)
Spells		50,938	

Significant at: *** 0.1% ** 1% * 5% $^{+}$ 10%. Actual N for AFT model is 1,345,077. All models control for state and time (month-year) fixed effects, and a complete set of covariates. All analyses exclude spells lasting only one week, and those of individuals who lived in more than one state. Standard errors are clustered at the state level

weeks is not indispensable to the finding and can be modified without need to modify our conclusions

We believe these robustness checks confirm that the effects we attribute to initial minimum wage and its within spell variation are causal. We find no evidence to suggest that these effects could be an artifact of the way we analyze the data, instead they seem to confirm that our findings are strong to alternative specifications.

2.6 Conclusions

We examine how the minimum wage policy impacts the unemployed, both through its initial level at the start of an unemployment spell and its increases while a worker is unemployed. Most notable, we find almost no positive impact of the policy on the unemployed worker.

As we expected, levels and increases of the policy have different impacts on unemployed workers and the market as a whole. Although our analysis is focused on individuals, it does suggest that the minimum wage policy has an important impact on the labor market through its increases. Specifically, increasing the market's minimum wage may tighten the market shortly, creating penalties for the unemployed workers at the time.

Although we do find some heterogeneity that suggests that the effects of the policy through levels is less negative on the less productive workers, we are still concerned with the very strong impact of the policy, even through levels, on the probability a worker will abandon the search for work. This could lead to distributional issues much more important than the potential half a week of extra work obtained in exchange. Futhermore, while this positive effect is remarkably localized, the negative impact on search behavior is suffered by all workers, even the least productive.

At the same time we find that minimum wage changes can be very onerous on the unemployed. Its long-term effects working hours can be particularly concerning. Our analysis shows that some of this effect is likely due to some workers quitting the labor force altogether, a less than good development. However, we still observe that hours will be reduced even conditional on working. Effect on men are particularly strong and concerning, extending over at least two years and costing them over 6% of their yearly income.

Overall, these results are very critical of the minimum wage policy. Irrespective of how the debate on dis-employment effects progresses, our findings here show that the policy losers do exist, and experience real disadvantages due to it. Furthermore, we find surprisingly little evidence that the minimum wage can be advantageous to workers in this position, another big concern from the perspective of the policy maker.

We believe more work like this needs to be done. For any public policy the search for those negatively affected by it should be a priority. The measuring of these negative effects can help us improve the way we use and modify the policy itself.

Chapter 3

A Numerical Evaluation of Double Machine Learning Non-parametric Inference with Continuous Treatment

3.1 Introduction

The goal of the present work is to evaluate the effectiveness of the double machine learning (DML) estimator for continuous treatment developed in Colangelo and Lee 2020 (Henceforth referred to as CL). CL extend the DML Average Treatment Effect (ATE) estimator of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins 2018 to the case where the treatment is continuous, and derive asymptotic results for conducting valid non-parametric inference. The DML continuous treatment (DMLCT) estimator allows for a fully non-parametric outcome model, and for a high-dimensional covariate set in which the number of covariates is allowed to be large relative to the sample size.

CL consider a fully non-parametric outcome equation model

$$Y = g(T, X, \varepsilon) \tag{3.1}$$

with no assumptions on the functional form (i.e. separability). The potential outcome for a treatment level t can be defined as $Y(t) = g(t, X, \varepsilon)$. The primary object of interest is

the average dose response function $\beta_t = \mathbb{E}[Y(t)] = \int\int g(t, X, \varepsilon) dF_{X\varepsilon}$, which is simply the expected value of Y for a given value of treatment. The well-studied average treatment effect of switching from treatment t to s is $\beta_s - \beta_t$. CL further define the partial (or marginal) effect of the first component of the continuous treatment T at $t = (t_1, ...t_{d_t})'$ to be the partial derivative $\theta_t = \partial\beta_t/\partial t_1$.

The *doubly robust* estimator is defined by CL (and also used in Su, Ura, and Zhang 2019) as

$$\hat{\beta}_t^{DR} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\gamma}(t, X_i) + \frac{K_h(T_i - t)}{\hat{f}_{T|X}(t|X_i)}\big(Y_i - \hat{\gamma}(t, X_i)\big)\right\}, \tag{3.2}$$

where $\hat{\gamma}(t, x)$ is an estimator of the conditional expectation function $\gamma(t, x) = \mathbb{E}[Y|T = t, X = x]$, $\hat{f}_{T|X}(t|x)$ is an estimator of the conditional density (or generalized propensity score) $f_{T|X}(t|x)$, and a kernel $K_h(T_i - t)$ weights observation i with treatment value around t in a distance of h. The number of such observations shrinks as the bandwidth h vanishes with the sample size n.

The first term of the estimator is the "naive" machine learning estimator, in which we simply use a machine learning algorithm to predict the outcome of each observation at a treatment level t, and average over all observations. In some situations this may produce an effective estimator, but as discussed in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey 2017,Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins 2018, and Colangelo and Lee 2020, the "naive" estimator often leads to substantial finite sample biases. As such, the "adjustment term" is added which gives the estimator the double robustness property, and allows for this estimator to be used for valid inference even when a nuisance estimator attains a slower than $\sqrt{n}$ convergence rate.

Typically, rather than a direct application of the doubly robust estimator, it is also combined with cross-fitting. L-fold cross-fitting splits the sample into L subsamples. The nuisance estimators $\hat{\gamma}(t, X_i)$ and $\hat{f}_{T|X}(t|X_i)$ use observations in the other $L - 1$ subsamples that do not contain the observation i. The DML estimator averages over the subsamples to obtain $\hat{\beta}_t$. We then estimate the partial effect θ_t by a numerical differentiation. The function of cross-fitting is to prevent overfitting. If the model we construct performs well on

the data it is fit on, but not out of sample, this may give us an unrealistic picture of how good our estimates are. Cross-fitting makes sure that the estimates we obtain are based on out-of-sample performance, and thus alleviate this problem.

CL show that the estimator is asymptotically normal and converges at nonparametric rates, and provides low level conditions on the nuisance function estimators, which can be traditional non-parametric estimators (e.g. kernel or series estimators) or machine learning algorithms (e.g. random forest, deep neural networks, lasso).

The estimation of γ is straightforward, it simply requires a regression of Y on T and X using any desired model/algorithm. Once the model is fitted, it can be evaluated for each observation at a specified treatment level t. The estimation of the generalized propensity score is less straightforward and much less developed in the literature. CL propose a novel estimator of the GPS, in which a regression is estimated to estimate $f_{T|X}(t|X) = E(g_{h_1}(T_i - t)|X = x)$. Where h_1 is the bandwidth, and g is a kernel function which can be a gaussian kernel. This can be estimated via any machine learning or non-parametric method by regressing $g_{h_1}(T_i - t)$ on X. The model can then be evaluated at X_i for each individual.

The full estimation procedure from CL to estimate the average dose-response and the partial effects with cross-fitting, can be summarized in the following:

Estimation procedure

Step 1. (Cross-fitting) For some $L \in \{2, ..., n\}$, partition the observation indices into L groups I_ℓ, $\ell = 1, ..., L$. For each $\ell = 1, ..., L$, the estimators $\hat{\gamma}_\ell(t, x)$ for $\gamma(t, x) = \mathbb{E}[Y|T = t, X = x]$ and $\hat{f}_\ell(t|x)$ for $f_{T|X}(t|x)$ use observations not in I_ℓ.

Step 2. (Doubly robust) The double debiased ML (DML) estimator is defined as

$$\hat{\beta}_t = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \left\{ \hat{\gamma}_\ell(t, X_i) + \frac{K_h(T_i - t)}{\hat{f}_\ell(t|X_i)} \left(Y_i - \hat{\gamma}_\ell(t, X_i)\right) \right\}. \tag{3.3}$$

Step 3. (Partial effect) Let $t^+ = (t_1 + \eta/2, t_2, ..., t_{d_t})'$ and $t^- = (t_1 - \eta/2, t_2, ..., t_{d_t})'$, where η is a positive sequence converging to zero as $n \to \infty$. We estimate the partial effect of the first component of the continuous treatment $\theta_t = \partial\beta_t/\partial t_1$ by $\hat{\theta}_t = (\hat{\beta}_{t^+} - \hat{\beta}_{t^-})/\eta$.

3.2 Numerical Exercises

This section provides numerical examples of Monte Carlo simulations and an empirical illustration. The estimation procedure of the proposed double debiased machine learning (DML) estimator is described in Section 3.1. For both the regression $\gamma(t,x) = \mathbb{E}[Y|T=t, X=x]$ and the generalized propensity score (GPS) $f_{T|X}$, We consider five algorithms in total: The ordinary LASSO derived in Tibshirani 1996, the feedworward fully connected ReLU[1] neural networks of Farrell, Liang, and Misra 2021 (which we will simply refer to as the "Neural Network (NN)") and the modified version derived in Colangelo and Lee 2020 (which we refer to as the K Neural Network (KNN)), the Random Forest (RF) developed in Breiman 2001, and finally the Generalized Random Forest (GRF) of Athey, Tibshirani, Wager, et al. 2019. Colangelo and Lee 2020 derive low level conditions for the usage of the KNN and GRF. We implement our DML estimator with these five algorithms in Python, using the packages scikit-learn, pytorch, numpy, pandas, rpy2 and scipy. We utilize the R package "grf" for the generalized random forest implementation, implementing it in Python via the rpy2 package. Software is available from the author. For full details on the specific algorithms, we refer the reader to the original papers.

3.2.1 Simulation Study

We begin by describing the nuisance estimators for the simulation in more detail. For each method considered, we use that method for the estimation of both $\gamma(t,X)$ and $f_{T|X}$. In practice the two nuisance functions may be estimated using different algorithms. For lasso, we augment the covariate set with basis functions, but for the other algorithms we do not. This is because since Lasso is linear, it can not pick up a non-linear data generating process without added basis functions, but random forests and neural networks are universal approximators.

Lasso: The penalization parameter is chosen via grid search utilizing tenfold cross validation in both estimators of γ and $f_{T|X}$ separately. The included basis functions contain third-order polynomials of X and T, and interactions among X and T.

[1] ReLU stands for Rectified Linear Unit, which is a function $f(x) = max(0,x)$

Random Forest: We use the random forest in Breiman 2001, with 2,000 trees, and with the minimum leaf size, maximum depth, and minimum number of observations for a split, tuned via cross-validation. No additional basis functions are used.

Generalized Random Forest: We use Generalized Random Regression forests as in Athey et al. (2020), with 2,000 trees and all other parameters chosen via cross validation on every Monte Carlo replication. The parameters tuned via cross validation are: The fraction of data used for each tree, the number of variables tried for each split, the minimum number of observations per leaf, whether or not to use "honesty splitting," whether or not to prune trees such that no leaves are empty, the maximum imbalance of a split, and the amount of penalty for an imbalanced split. Unlike lasso, we do not add any additional basis functions as inputs into the random forest.

Neural Network and K Neural Network: We use a fully connected feedforward neural network with 4 hidden layers for both types of neural networks. Each hidden layer has 10 neurons and uses rectified linear unit (ReLU) activation functions. The output layer uses no activation function. The weights are fit using stochastic gradient descent with a weight decay of 0.2 and a learning rate of 0.01.[2] For the selection of the neural network models, we performed a train-test split of the data and chose the models based on out-of-sample performance.

The only difference between the ordinary neural network and KNN, is that the KNN multiplies the loss function by a kernel function $k\left(\frac{T-t}{b}\right)$, where b is the bandwidth, and k is a kernel function satisfying the bounded differentiable assumption of CL, which may be for example a gaussian or epanechnikov kernel. This results in a more "localized" fitting of the model, in which only observations close to the desired treatment level t are used to construct the neural network. This has the disadvantage of using fewer observations to fit the neural network, but the included observations may be more relevant to the estimation of β_t and can result in better performance. The neural network of CL is referred to as the "K neural network" for its use of the additional kernel function in the loss function.

[2]Weight decay is a form of regularization to prevent overfitting. Weight decay is a penalty where after each iteration the weights in the network are multiplied by $(1 - \alpha\lambda)$ before adding the adjustment in the direction of the gradient, where α is the learning rate (step size) and λ is the weight decay.

We consider the data-generating process: $\nu \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{N}(0,1)$,

$$X = (X_1, ..., X_{100})' \sim \mathcal{N}(0, \Sigma), \quad T = \Phi(3X'\theta) + 0.75\nu, \quad Y = 1.2T + 1.2X'\theta + T^2 + TX_1 + \varepsilon,$$

where $\theta_j = 1/j^2$, $diag(\Sigma) = 1$, the (i,j)-entry $\Sigma_{ij} = 0.5$ for $|i-j| = 1$ and $\Sigma_{ij} = 0$ for $|i-j| > 1$ for $i, j = 1, ..., 100$, and Φ is the CDF of $\mathcal{N}(0,1)$. Thus the potential outcome $Y(t) = 1.2t + 1.2X'\theta + t^2 + tX_1 + \varepsilon$. Let the parameter of interest be the average dose response function at $t = 0$, i.e., $\beta_0 = \mathbb{E}[Y(0)] = 0$.

We compare estimations with fivefold cross-fitting and without cross-fitting, and with a range of bandwidths to demonstrate robustness to bandwidth choice. We consider sample size $n \in \{500, 1000\}$ and the number of subsamples used for cross-fitting $L \in \{1, 5\}$. We use the second-order Epanechnikov kernel with bandwidth h. For the GPS estimator described in Section 3.1, we choose bandwidth h_1 to also be equal to h. Let the bandwidth of the form $h = c\sigma_T n^{-0.2}$ for a constant $c \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$ and the standard deviation σ_T of T. We computed the AMSE-optimal bandwidth h_0^* given in CL that has the corresponding $c^* = 1.43$. Thus using some undersmoothing bandwidth with $c < c^*$, the 95% confidence interval $\left[\hat{\beta}_t \pm 1.96 s.e.\right]$ is asymptotically valid, where the standard error ($s.e.$) is computed using the sample analogue of the estimated influence function, as described in Section 3.1. All the results are based on 1,000 Monte Carlo simulations. Table 3.1 reports the results.

The estimators using these machine learning methods perform very well in the case of cross-fitting ($L = 5$), with coverage rates near the nominal 95% for low bandwidths for all five algorithms considered. Under no cross-fitting ($L = 1$), the confidence intervals generally have lower coverage rates than under cross-fitting. However, the coverage rate and bias is improved by far the most for random forests and generalized random forests, and results in only marginal improvements for our lasso and neural network. Cross-fitting should only improve our estimation in the case that our machine learning algorithm is over-fitting. Given that cross-fitting only results in small improvements for lasso and neural network, this tells us that those algorithms as we have implemented them do not have a severe over-fitting problem, but generalized random forest does. It may be possible to alleviate this over-fitting via more regularization.

Interestingly, cross-fitting improves the coverage rate for our K neural network estimator, but worsens the bias. This is contrary to the regular neural network in which bias i substantially improved. Since the K neural network uses an effectively smaller subset of the data, it may be more sensitive to cross-fitting as this could result in a less than adequate sample size. Nonetheless, coverage rates are still in the vicinity of 95% for $n = 1000$ in spite of this bias. Further investigation on larger sample sizes may be informative about whether this bias will vanish at a sufficiently fast rate to maintain the near 95% coverage rates. It is important to see that in some ways cross-fitting can potentially worsen performance. This could suggest that in some applications it may be more prudent to utilize different forms of regularization (e.g. weight decay) as opposed to cross-fitting.

All five methods seem somewhat robust to bandwidth choice under cross-fitting, but performance typically worsens as the bandwidth grows. This is expected as a higher bandwidth is associated with a higher bias in theory. At a sample size of $1,000$ under cross fitting, generalized random forests appear to be the most robust to bandwidth choice.

Overall these results demonstrate consistency with the theoretical results of this paper, confirming the importance of cross-fitting and bandwidth choice in reducing bias.

3.2.2 Empirical Illustration

We illustrate our method by reanalysing the Job Corps program in the United States, which was conducted in the mid-1990s. The Job Corps program is the largest publicly funded job training program, which targets disadvantaged youth. The participants are exposed to different numbers of actual hours of academic and vocational training. The participants' labor market outcomes may differ if they accumulate different amounts of human capital acquired through different lengths of exposure. We estimate the average dose response functions to investigate the relationship between employment and the length of exposure to academic and vocational training.

As our analysis builds on Flores et al. 2012, Hsu et al. 2018, and Lee 2018, we refer the readers to the reference therein for further details of Job Corps.

We use the same dataset in Hsu et al. 2018. We consider the outcome variable (Y) to be the proportion of weeks employed in the second year following the program assignment.

Table 3.1: Simulation Results for Lasso, GRF and KNN

			Lasso			GRF			KNN		
n	L	c	Bias	RMSE	Coverage	Bias	RMSE	Coverage	Bias	RMSE	Coverage
500	1	0.50	0.017	0.142	0.938	0.142	0.211	0.795	-0.029	0.259	0.948
		0.75	0.013	0.132	0.948	0.117	0.183	0.831	-0.038	0.229	0.926
		1.00	0.019	0.124	0.949	0.105	0.168	0.838	-0.032	0.204	0.932
		1.25	0.030	0.118	0.945	0.100	0.156	0.857	-0.029	0.188	0.922
		1.50	0.044	0.118	0.938	0.099	0.152	0.859	-0.027	0.187	0.913
	5	0.50	-0.041	2.468	0.960	-0.003	0.196	0.954	-0.087	0.277	0.972
		0.75	0.017	0.320	0.949	0.000	0.166	0.953	-0.088	0.239	0.951
		1.00	0.015	0.131	0.954	0.006	0.147	0.955	-0.081	0.210	0.950
		1.25	0.027	0.120	0.948	0.011	0.134	0.963	-0.082	0.194	0.946
		1.50	0.041	0.118	0.943	0.018	0.126	0.967	-0.082	0.184	0.944
1000	1	0.50	0.008	0.120	0.947	0.122	0.171	0.717	-0.060	0.207	0.955
		0.75	0.010	0.105	0.945	0.106	0.149	0.764	-0.062	0.181	0.939
		1.00	0.014	0.096	0.934	0.094	0.135	0.799	-0.063	0.170	0.924
		1.25	0.021	0.092	0.929	0.090	0.128	0.803	-0.061	0.158	0.920
		1.50	0.031	0.091	0.920	0.088	0.124	0.794	-0.058	0.148	0.909
	5	0.50	0.005	0.123	0.951	0.009	0.145	0.948	-0.073	0.221	0.954
		0.75	0.008	0.105	0.942	0.011	0.122	0.951	-0.067	0.184	0.948
		1.00	0.014	0.096	0.934	0.013	0.108	0.948	-0.070	0.163	0.938
		1.25	0.021	0.092	0.927	0.017	0.101	0.947	-0.069	0.151	0.940
		1.50	0.030	0.091	0.919	0.023	0.097	0.941	-0.067	0.145	0.922

Notes: $L = 1$: no cross-fitting. $L = 5$: fivefold cross-fitting. The bandwidth is $h = c\sigma_T n^{-0.2}$, and $c = 1.43$ for the AMSE-optimal bandwidth. The nominal coverage rate of the confidence interval is 0.95.

Table 3.2: Simulation Results for RF and NN

			RF			NN		
n	L	c	Bias	RMSE	Coverage	Bias	RMSE	Coverage
500	1	0.50	0.208	0.252	0.665	0.171	0.240	0.803
		0.75	0.160	0.205	0.756	0.164	0.220	0.791
		1.00	0.128	0.179	0.800	0.161	0.210	0.780
		1.25	0.108	0.162	0.857	0.159	0.205	0.778
		1.50	0.094	0.149	0.866	0.159	0.202	0.757
	5	0.50	-0.026	0.198	0.962	-0.082	0.413	0.960
		0.75	-0.010	0.159	0.974	-0.060	0.211	0.965
		1.00	-0.007	0.146	0.968	-0.055	0.189	0.956
		1.25	-0.006	0.134	0.961	-0.051	0.170	0.956
		1.50	-0.004	0.126	0.963	-0.050	0.161	0.951
1000	1	0.50	0.165	0.193	0.556	0.047	0.143	0.937
		0.75	0.133	0.162	0.636	0.044	0.124	0.940
		1.00	0.105	0.137	0.733	0.047	0.114	0.933
		1.25	0.095	0.127	0.764	0.050	0.109	0.917
		1.50	0.086	0.119	0.793	0.056	0.107	0.907
	5	0.50	-0.014	0.140	0.958	-0.035	0.163	0.962
		0.75	-0.002	0.115	0.955	-0.029	0.135	0.959
		1.00	-0.004	0.102	0.962	-0.024	0.119	0.960
		1.25	0.006	0.096	0.954	-0.018	0.107	0.963
		1.50	0.007	0.089	0.959	-0.012	0.100	0.963

Notes: $L = 1$: no cross-fitting. $L = 5$: fivefold cross-fitting. The bandwidth is $h = c\sigma_T n^{-0.2}$, and $c = 1.43$ for the AMSE-optimal bandwidth. The nominal coverage rate of the confidence interval is 0.95.

The continuous treatment variable (T) is the total hours spent in academic and vocational training in the first year. We follow the literature to assume the conditional independence assumption, required by the DML estimator, meaning that selection into different levels of the treatment is random, conditional on a rich set of observed covariates, denoted by X. The conditional independence assumption is indirectly assessed in Flores et al. 2012.

Our sample consists of 4,024 individuals who completed at least 40 hours (one week) of academic and vocational training. There are 40 covariates measured at the baseline survey. Figure 3.1 shows the distribution of T by a histogram, and Table 3.3 provides brief descriptive statistics.

Implementation Details We estimate the average dose response function $\beta_t = \mathbb{E}[Y(t)]$ and partial effect $\theta_t = \partial\mathbb{E}[Y(t)]/\partial t$ by the proposed DML estimator with fivefold cross-fitting. We implement the DML estimator five times with Lasso, Random Forest, Generalized random forest, Deep Neural Network and K Deep Neural Network for the nuisance parameters, using the same algorithm for both γ and $f_{T|X}$. The parameters for these five methods are selected as described in the simulation Section 3.2.1. For neural network and K neural network, the regression estimation of γ uses a neural network with two hidden layers and a weight decay of 0.1. The first and second hidden layers have twenty neurons. The hidden layers use Rectified Linear Unit (ReLU) activation functions. The output layer uses no activation function. The GPS estimation uses a network with 4 hidden layers and a weight decay of 0.1. Each hidden layer with 10 neurons and with ReLU activation functions. The output layer uses a linear activation function.

We use the second-order Epanechnikov kernel with bandwidth h. For the GPS estimator, we use the Gaussian kernel with bandwidth $h_1 = h$. We compute the optimal bandwidth h_w^* that minimizes an asymptotic integrated MSE derived in CL after an initial choice of bandwidth $3\hat{\sigma}_T n^{-0.2} = 563.339$. A practical implementation is to choose the weight function $w(t) = \mathbf{1}\{t \in [\underline{t}, \bar{t}]\}/(\bar{t}-\underline{t})$ to be the density of $Uniform[\underline{t}, \bar{t}]$ on the interior support $[\underline{t}, \bar{t}] \subset \mathcal{T}$ of the support of the continuous treatment. Set m equally spaced grid points over $[\underline{t}, \bar{t}]$: $\{\underline{t} = t_1, t_2, ..., t_m = \bar{t}\}$. A plug-in estimator can be defined as $\hat{h}_w^* = \left(\hat{\mathsf{V}}_w/\left(4\hat{\mathsf{B}}_w\right)\right)^{1/5} n^{-1/5}$, where $\hat{\mathsf{V}}_w = m^{-1}\sum_{j=1}^m \hat{\mathsf{V}}_{t_j}\mathbf{1}\{t_j \in [\underline{t}, \bar{t}]]\}/(\bar{t}-\underline{t})$ and $\hat{\mathsf{B}}_w = m^{-1}\sum_{j=1}^m \hat{\mathsf{B}}_{t_j}^2\mathbf{1}\{t_j \in [\underline{t}, \bar{t}]]\}/(\bar{t}-\underline{t})$.

We use $[\underline{t}, \bar{t}] = [160, 1840]$ and $t_2 - t_1 = 40$ in this empirical application. We then obtain bandwidths $0.8h_w^*$ for undersmoothing that are 418.87 for Lasso, 363.40 for random forest, and 322.25 for neural network.

Results Figure 3.2 presents the estimated average dose response function β_t along with 95% point-wise confidence intervals for the case where the chosen bandwidth is the estimated optimal bandwidth in CL multiplied by 0.8. Figure 3.4 displays the estimates of the average dose response function when the initial rule of thumb bandwidth is used.

The results for the five ML nuisance estimators have similar patterns. The estimates suggest for both the rule of thumb and optimal bandwidth show an inverted-U relationship between the employment and the length of participation. The optimal estimated bandwidth results in more erratic estimates, whereas the rule of thumb bandwidth is more smooth. This is somewhat expected, as the rule of thumb bandwidth is larger we would expect those estimates to be smoother. However, it is concerning that the estimates are "wobbly" under the estimated optimal bandwidth. This may indicate that in practice undersmoothing the optimal bandwidth may result in it being too small and thus resuling in this behavior. In any case, the results for both choices of bandwidth point to the same conclusion, that the effect of the program peaks in the vicinity of $1,000$ hours.

Figure 3.3 reports the partial effect estimates $\hat{\theta}_t$ with step size $\eta = 160$ (one month) using the undersmooth estimated optimal bandwidth. Figure 3.5 displays the analogous results for the initial rule of thumb choice. Across all procedures and both bandwidth choices, we see positive partial effects when hours of training are less than around 500 (three months) and negative partial effect around 1,500 hours (9 months). All five algorithms result in remarkably similarly shaped graphs, which is a testament to the robustness of the continuous treatment DML developed in CL, as a wide array of completely different algorithms all point to the same conclusion.

Taking the estimates by lasso for example, $\hat{\beta}_{400} = 47.03$ with standard error $s.e. = 1.31$ and $\hat{\theta}_{400} = 0.0221$ with $s.e. = 0.0062$. This estimate implies that increasing the training from two months to three months increases the average proportion of weeks employed in the second year by 3.37% (about one working week) with $s.e. = 1.00\%$.

We can also note that the different machine learning algorithms produce estimates of varying stability. The neural network for example appears to be the most erratic. This is likely related to the fact that the optimal bandwidth calculation results in a smaller bandwidth choice for neural network. Under a larger bandwidth choice (like the bandwidth for lasso), the neural network results in similarly smooth estimates.

We note that the empirical practice has focused on semiparametric estimation; see 2012, 2018, Lee 2018, for example. The semiparametric methods are subject to the risk of misspecification. Our DML estimator provides a solution to the challenge of implementing a fully nonparametric inference in practice.

Figure 3.1: Histogram of Hours of Training

Table 3.3: Descriptive statistics

Variable	Mean	Median	StdDev	Min	Max
share of weeks employed in 2nd year (Y)	44.00	40.38	37.88	0	100
total hours spent in 1st-year training (T)	1219.80	992.86	961.62	40	6188.57

Notes: Summary statistics for 4,024 individuals who completed at least 40 hours of academic and vocational training.

Figure 3.2: Estimated Average Dose Response Functions and 95% Confidence Intervals

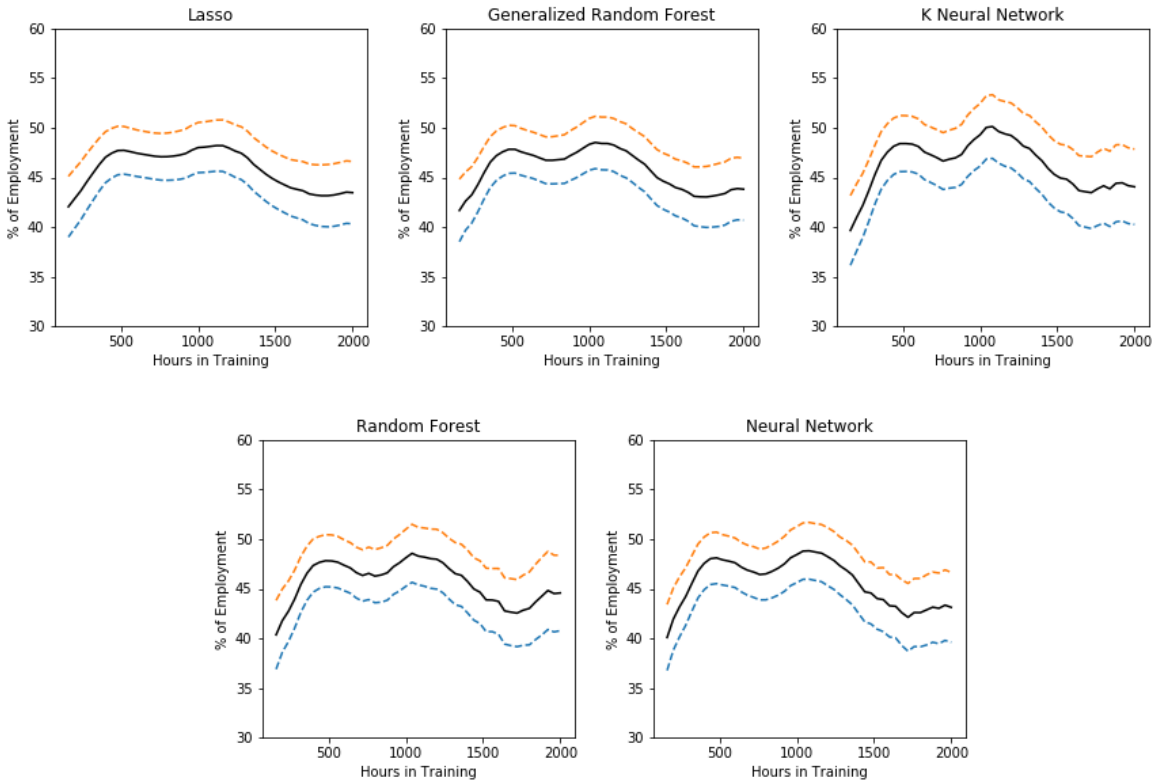


Figure 3.3: Estimated Partial Effects and 95% Confidence Interval

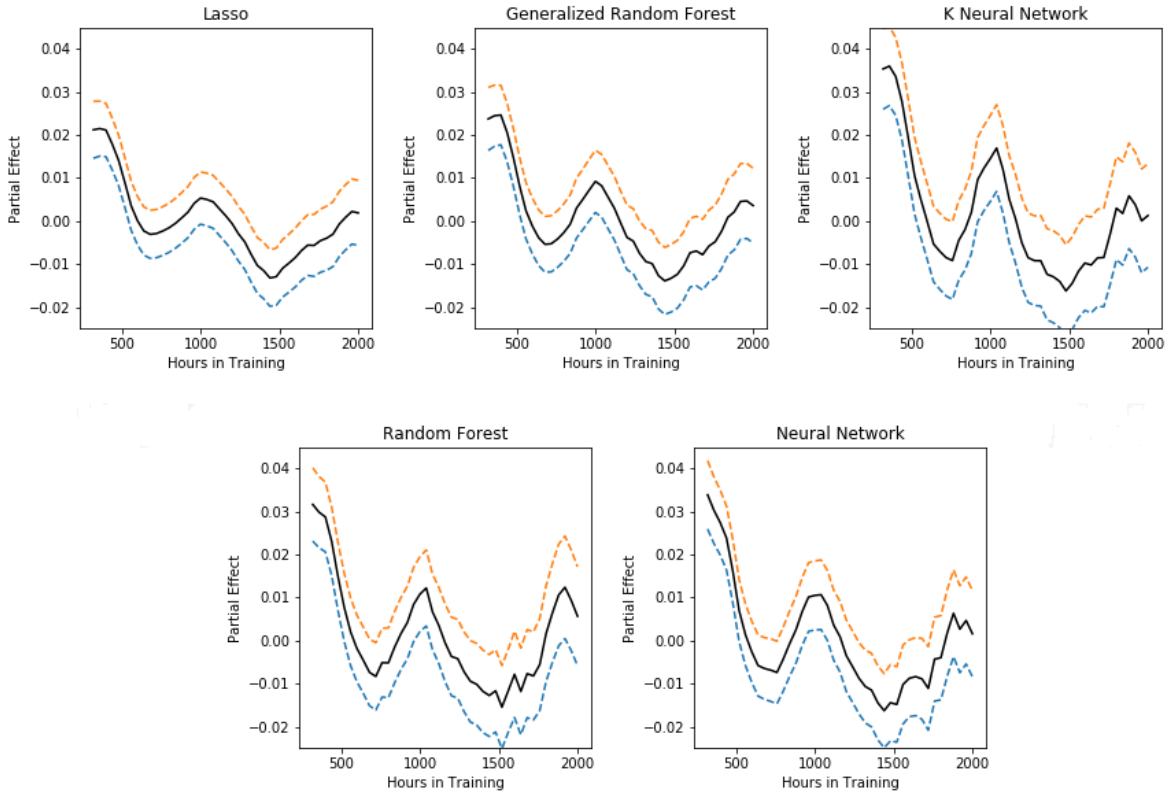


Figure 3.4: Estimated Average Dose Response Functions and 95% Confidence Intervals (Rule of Thumb bandwidth)

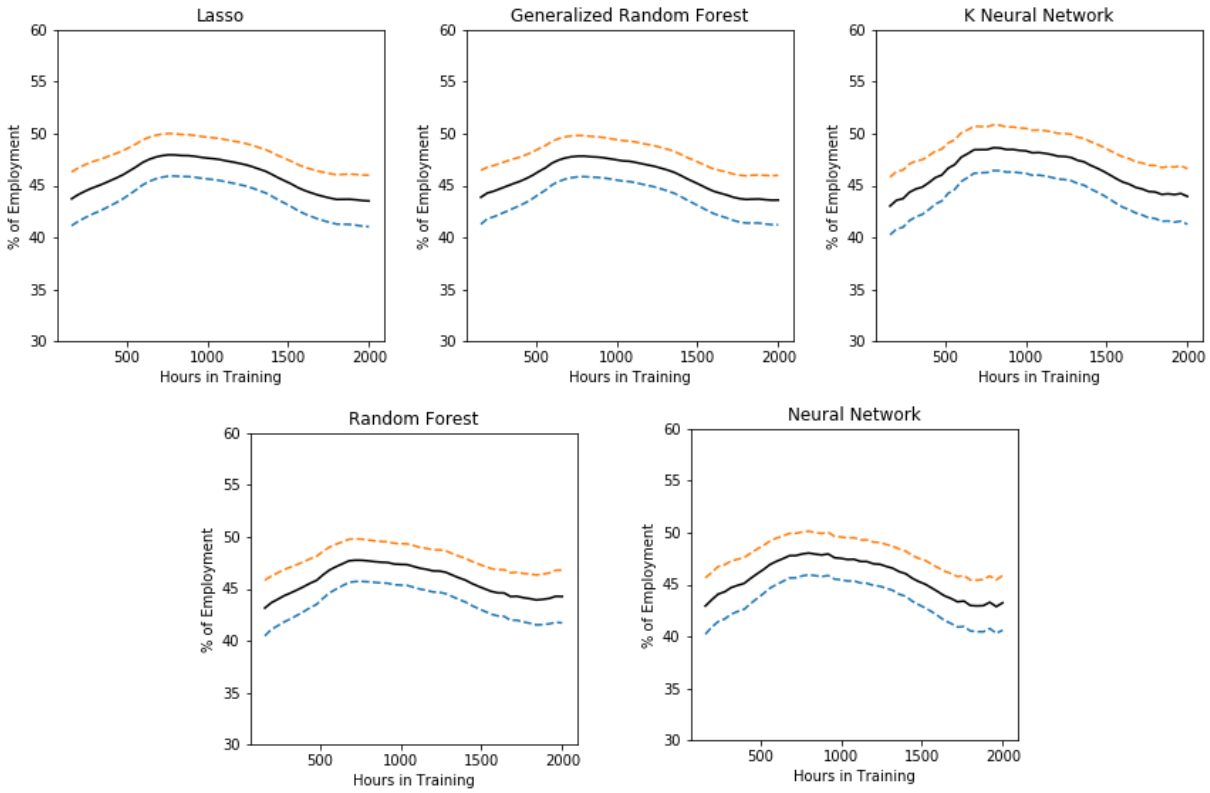
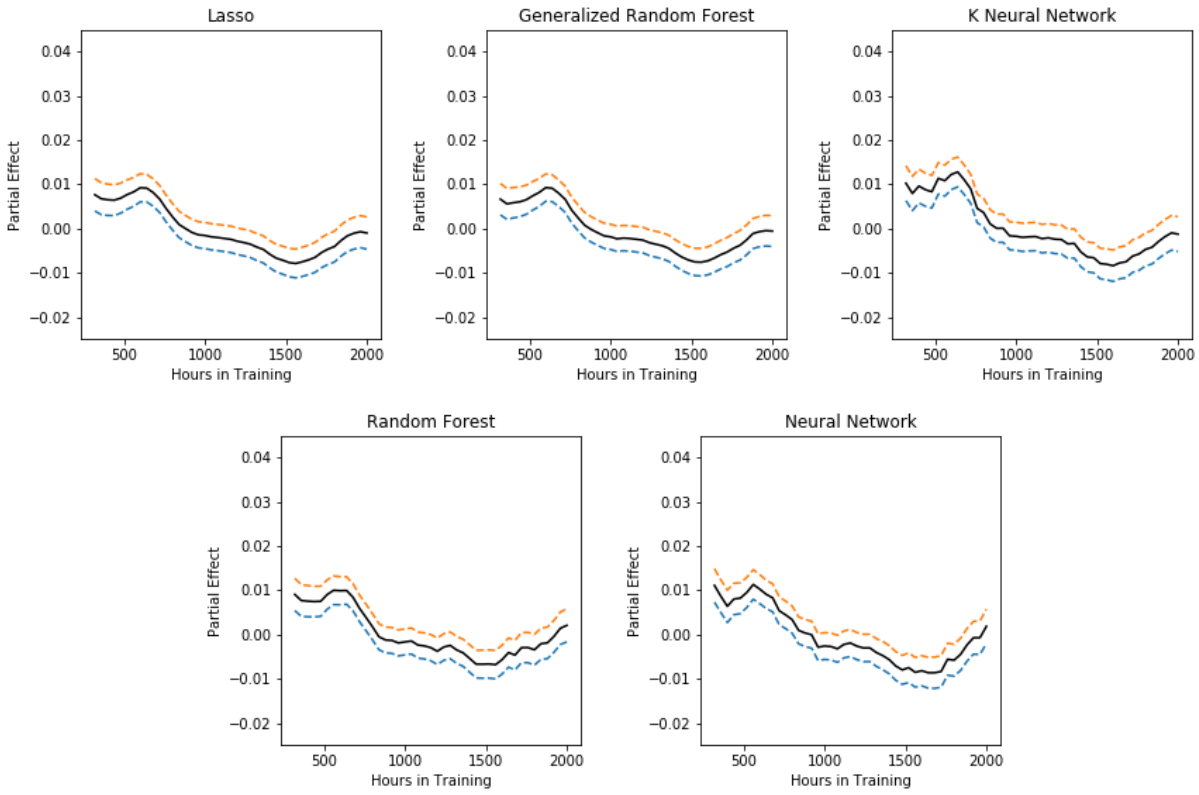


Figure 3.5: Estimated Partial Effects and 95% Confidence Interval (Rule of Thumb Bandwidth)



3.3 Conclusion and Outlook

This paper numerically analyzes the DML nonparametric inference method for continuous treatment effects developed in Colangelo and Lee 2020.

For our numerical exercises we consider five algorithms: Lasso, Deep Neural Network, K Deep Neural Network, Random Forest, and Generalized Random Forest. All algorithms perform well under cross-fitting, with only random forest and generalized random forest performing extremely poorly without cross-fitting. Empirical results for all five algorithms were nearly identical, attesting to the robustness of the continuous treatment DML estimator. The improvements shown for random forests under cross-fitting shows us how important it is to combine the doubly robust estimator with cross-fitting.

We further evaluated the effect of bandwidth choice on the performance of the estimators. We find that for the simulations, smaller (less than the optimal) bandwidth choices resulted in coverage rates closer to the desired 95% which is consistent with the theoretical predictions of CL. However, the undersmoothing bandwidth resulted in more erratic estimates for the empirical application. Overall the results were similar in both the simulation and empirical application regardless of the bandwidth chosen.

Our analysis in this work is very limited however in that we only consider one data generating process and one empirical dataset. Future work can further investigate the performance of this estimator for different data generating processes that have different forms on non-linearity, and different amounts of sparsity. It may also be of interest to investigate how performance differs on larger sample sizes. Due to computational constraints the samples sizes considered are rather small. Lastly, there are of course a plethora of machine learning algorithms which we do not assess in this paper, future work can investigate whether results will be consistent with other machine learning algorithms or non-parametric estimators.

Chapter 4

Estimation in Large Panels with Interactive Effects

4.1 Introduction

In the presence of panel data, fixed effects estimation is a very powerful way to control for many unobserved variables that could be causing endogeneity. Typically some variant of the following regression is estimated using OLS.

$$Y_{it} = \beta X_{it} + \gamma_i + \delta_t + \epsilon_{it}, \quad i = 1, ..., N; \; t = 1, ..., T$$

where Y_{it} is the outcome of interest, X_{it} denotes any individual and/or time varying covariates, γ_i is the individual specific effect and δ_t is the time effect. By controlling for individual and time effects we hopefully remove much of the unobserved heterogeneity causing X to be endogenous. The key assumption with fixed effects models is that the unobserved confounders either vary only with individual, or only with time. However, it may not be sufficient in practice to control only for individual specific and time specific effects as unobserved endogeneity may vary with both individuals and time.

In order to be more general, we can study individual specific effects that vary with time (we call these interactive effects). These interactive effects models, (sometimes called factor models) will hopefully control for more of the unobserved endogenous variation than tradi-

tional fixed effect estimation.

In this paper we are concerned with the estimation of the following panel data factor model where N is the number of cross-sectional units and T is the number of time periods.

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + e_{it}, \quad i = 1, ..., N \; t = 1, ..., T \tag{4.1}$$

$$e_{it} = \boldsymbol{\gamma}'_i \mathbf{f}_t + \epsilon_{it} \tag{4.2}$$

$$\mathbf{x}_{it} = \boldsymbol{\Gamma}'_i \mathbf{f}_t + \mathbf{v}_{it} \tag{4.3}$$

where $\mathbf{x}_{it}$ is a $k \times 1$ vector of covariates, $\boldsymbol{\beta}_i$ is a $k \times 1$ vector of parameters. $\mathbf{f}_t$ is a $m \times 1$ vector of the unobserved factors and $\boldsymbol{\gamma}_i$, $\boldsymbol{\Gamma}_i$ are $m \times 1$ vectors of the unobserved factor loadings. The factors thus generate both $\mathbf{x}_{it}$ and y_{it}, introducing endogeneity through $\mathbf{f}_t$. The factor loadings $\boldsymbol{\gamma}_i$ and $\boldsymbol{\Gamma}_i$ are assumed to be independent so the only source of endogeneity are the factors. The goal is to estimate $\boldsymbol{\beta}_i$ consistently by somehow controlling for the unobserved factors. Note that if we let $m = 2$, $\boldsymbol{\gamma}'_i = (1 \; \lambda_i)$ and $\mathbf{f}'_t = (g_t \; 1)$, then $\boldsymbol{\gamma}'_i \mathbf{f}_t = \lambda_i + g_t$ and thus we recover the canonical two-way fixed effects model. Hence any factor model estimator should also be able to estimate a fixed effect model provided that equation 4.3 holds (i.e. the fixed effects are relevant). In the general case however, since the time varying effects and individual varying effects enter the model interactively we can not apply normal fixed effect estimators.

Factor models have typically seen applications in macro and finance. In the past decade however, access to long term micro panels has made factor models far more applicable to various other fields. In the past basic fixed effect models have been sufficient as control in short term micro studies. However, in longer term studies it seems more plausible that there could be a number of different factors that effect individuals in a heterogeneous way. Thus while factor models have generally been limited to macro and finance, we also believe that they have applications to micro panels (see Section 4.5 for an application to the minimum wage).

By restricting our model to the above form in 4.1, 4.2, and 4.3, a number of approaches become available to us. In roughly the past decade a number of different approaches have

been developed for the estimation of panel data factor models. Three very common ones are the Common Correlated Effects (CCE) Estimator proposed by Pesaran 2006, the Interactive Fixed Effects (IFE) Estimator proposed by Bai 2009 and the Augmented Mean Group Estimator (AMG) proposed by Eberhardt and Teal 2010.

The CCE estimator allows for the above specification in 4.1-4.3 with large N and T. CCE augments the regression with the cross sectional means of the dependent variable and the covariates. By doing this the CCE estimator approximates controlling for the factors by controlling for cross sectional averages. That is, if we let $\bar{z}_t' = (\bar{y}_t \ \bar{x}_t')$, where $\bar{x}_t$ is a column vector containing the means of all k covariates, then $\delta_i \bar{z}_t \approx \gamma_i' f_t$. Hence we just need to add $\bar{z}_t$ as a control with heterogeneous coefficients. A consistent estimate of $\boldsymbol{\beta}$ can then be obtained via ordinary least squares by estimating the following model:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \delta_i \bar{z}_t + e_{it} \tag{4.4}$$

One problem with both CCE and IFE is that they do not provide any effective way to estimate the factors themselves. This can be problematic if the factors have a particular interpretation that we are interested in. For example, they can represent total factor productivity in a macro setting. To solve this issue Eberhardt and Teal 2010 proposed the Augmented Mean Group estimator which actually attempts to estimate the factors.

In this paper we propose a new approach to estimation of large panels with a multifactor error structure based on the approach of Pesaran 2006. In our new approach we augment the regression with cross sectional averages for each time period as well as for each individual, plus their interactions. By doing so we eliminate the need to compute the individual coefficients on the cross sectional means as is necessary in CCE. One issue with CCE is that for large samples it quickly becomes very computationally inefficient due to the large number of parameters requiring estimation via OLS. Two-way CCE provides a similar quality estimator that is also more feasible for larger samples. While this might not be a concern for macro panels with small N and large T, it can become an issue for panels with large N and large T.

This paper proceeds as follow: in Section 4.2 we summarize the CCE approach and also

explain the concept behind Two-Way CCE. This includes a basic derivation of the estimator similar to Pesaran 2006. In Section 4.3 we formally state and prove consistency and asymptotic normality for the Two-Way CCE estimator. In Section 4.4 we compare Two-way CCE with other factor model estimators (specifically CCE, IFE, and AMG) through Monte Carlo simulations. In Section 4.5 we apply our Two-Way CCE to estimate the effects of minimum wages. In Section 4.6 we provide a summary and conclusion, and discuss possible extensions of Two-Way CCE.

4.2 Estimation

Pesaran 2006 demonstrated that the unobserved factors $\mathbf{f}_t$ can be approximated by a linear combination of cross sectional means $\bar{\mathbf{y}}_{\cdot t} = \sum_{i=1}^{n} \frac{1}{n} y_{it}$ and $\bar{\mathbf{x}}_{\cdot t} = \sum_{i=1}^{n} \frac{1}{n} \mathbf{x}_{it}$. In other words, we can control for the unobserved factors by controlling for the cross sectional means, allowing for individual specific coefficients. A similar derivation to Pesaran (2006) can also be applied to the factor loadings $\boldsymbol{\gamma}_i$, such that $\boldsymbol{\gamma}_i$ can be approximated by a linear combination of the means $\bar{\mathbf{y}}_{i\cdot} = \sum_{t=1}^{T} \frac{1}{T} y_{it}$ and $\bar{\mathbf{x}}_{i\cdot} \sum_{t=1}^{T} \frac{1}{T} \mathbf{x}_{it}$. This follows easily because of the symmetry of the interactive effects model.

We will be making the following primary assumptions for the rest of this paper:

Assumption 4.1 (Stationary Factor) *The vector of factors f_t is stationary with finite mean, and covariance stationary with absolute summable auto-covariances.*

Assumption 4.2 (Independence of Errors) *The errors ϵ_{it} and $\mathbf{v}_{it}$ are independent*

Assumption 4.3 (Factor Distribution) *The factor loadings ($\boldsymbol{\gamma}_i$,$\boldsymbol{\Gamma}_i$) have finite mean and variance. Furthermore $\boldsymbol{\gamma}_i$ and $\boldsymbol{\Gamma}_i$ are independent*

Assumption 4.4 (Coefficient Normality) *The slope coefficients are normally distributed.*

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \eta_i, \quad \eta_i \sim N(0, \boldsymbol{\Omega}) \tag{4.5}$$

Assumption 4.5 (Zero-mean Errors) $E_i(\epsilon_{it}) = E_t(\epsilon_{it}) = 0$, $E_i(\mathbf{v}_{it}) = E_t(\mathbf{v}_{it}) = \mathbf{0}$

The following is a basic derivation of the CCE estimator. First, let us stack the equations for y_{it} and $\mathbf{x}_{it}$ to obtain:

$$\mathbf{z}_{it} = \begin{pmatrix} y_{it} \\ \mathbf{x}_{it} \end{pmatrix} = \mathbf{C}_i' \mathbf{f}_t + \mathbf{u}_{it} \tag{4.6}$$

$$\mathbf{u}_{it} = \begin{pmatrix} \epsilon_{it} + \beta_i' \mathbf{v}_{it} \\ \mathbf{v_{it}} \end{pmatrix} \tag{4.7}$$

$$\mathbf{C}_i = (\gamma_i \quad \mathbf{\Gamma}_i) \begin{pmatrix} 1 & \mathbf{0} \\ \beta_i & \mathbf{I}_k \end{pmatrix} \tag{4.8}$$

Following Pesaran 2006, we can take the cross section mean of the equation for $\mathbf{z}_{it}$. This gives us:

$$\bar{\mathbf{z}}_{\cdot t} = \bar{\mathbf{C}}_{\cdot}' \mathbf{f}_t + \bar{\mathbf{u}}_{\cdot t} \tag{4.9}$$

where $\bar{\mathbf{z}}_{\cdot t} = \sum_{i=1}^{N} \frac{1}{N} \mathbf{z}_{it}$, $\bar{\mathbf{C}}_{\cdot} = \sum_{i=1}^{N} \frac{1}{N} \mathbf{C}_i$, and $\bar{\mathbf{u}}_{\cdot t} = \sum_{i=1}^{N} \frac{1}{N} \mathbf{u}_{it}$. Assume that $\text{Rank}(\bar{\mathbf{C}}_{\cdot}) = m \leq k+1$) We can then solve for f_t

$$f_t = (\bar{\mathbf{C}}_{\cdot}' \bar{\mathbf{C}}_{\cdot})^{-1} \bar{\mathbf{C}}_{\cdot} (\bar{\mathbf{z}}_{\cdot \mathbf{t}} - \bar{\mathbf{u}}_{\cdot \mathbf{t}}) \tag{4.10}$$

Given that the error term has mean 0, as $N \to \infty$ $\bar{\mathbf{u}}_{\cdot t} \to_p 0$, and also $\bar{\mathbf{C}}_{\cdot} \to_p C$ where C is a constant matrix. Therefore we will claim that $(\bar{\mathbf{C}}_{\cdot}' \bar{\mathbf{C}}_{\cdot})^{-1} \bar{\mathbf{C}}_{\cdot} \to_p R_1$ where R_1 is an unknown constant matrix. Thus by Slutsky's Theorem, we are left with:

$$p \lim [R_1 \bar{\mathbf{z}}_{\cdot t} - f_t] = 0 \tag{4.11}$$

This justifies the substitution of the unknown factors with the cross sectional means which are known, giving us the CCE estimator.

For the new approach which we dub the Two-Way Common Correlated Effects Estimator (TWCCE), we analogously perform the same derivation for $\boldsymbol{\gamma}_i$ but we instead take the mean over all time periods. In order to derive the new estimator we need to place the assumption that $\boldsymbol{\Gamma}_{ji} f_t = \boldsymbol{\gamma}_i A_j \mathbf{f}_t$, where A is a $m \times m$ matrix of coefficients, $j = 1, .., k$ is an indicator for covariate j. In other words, we need that each entry of $\boldsymbol{\Gamma}_i$ can be written as a linear combination of the entries of $\boldsymbol{\gamma}_i$. The entry corresponding to the first factor and the first covariate must be the same linear combination of $\boldsymbol{\gamma}_i$ for all i. We need to start by rewriting the stacked model in 4.6 to the following form:

$$\mathbf{z}_{it} = \begin{pmatrix} y_{it} \\ \mathbf{x}_{it} \end{pmatrix} = \underset{(k+1)\times m}{\mathbf{D}_t'} \underset{m\times 1}{\boldsymbol{\gamma}_i} + \mathbf{u}_{it} \tag{4.12}$$

where

$$\mathbf{D}_t = \underset{(km+m)\times(k+1)}{(\mathbf{f}_t \otimes I_{k+1})} + \underset{(km+m)\times(k+1)}{\begin{pmatrix} 0 & 0 & ... & 0 \\ (\mathbf{f}_t \otimes I_k)\boldsymbol{\beta} & 0 & ... & 0 \end{pmatrix}} \tag{4.13}$$

$$\tilde{\boldsymbol{\Gamma}}_i = \begin{pmatrix} \boldsymbol{\gamma}_i \\ \boldsymbol{\Gamma}_i \end{pmatrix} \tag{4.14}$$

$\mathbf{D}_t$ is a $(k+1) \times (km+m)$ matrix. $\tilde{\boldsymbol{\Gamma}}_i$ is a $(km+m) \times 1$ vector.

$$\bar{\mathbf{z}}_{i\cdot} = \bar{\mathbf{D}}_{\cdot}' \tilde{\boldsymbol{\Gamma}}_i + \bar{\mathbf{u}}_{i\cdot} \tag{4.15}$$

We need that $\text{Rank}(\bar{\mathbf{D}}_{\cdot}') = k+1 \geq m$. Note that here we wish to estimate $\boldsymbol{\gamma}_i$, whereas $\tilde{\boldsymbol{\Gamma}}_i$ are nuisance parameters. $\mathbf{a}'\tilde{\boldsymbol{\Gamma}}_i$ is estimable if and only if $\mathbf{a} \in C(\bar{\mathbf{D}})$. Define $\mathbf{a}' = e_j$, $j = 1, ..., m$. Where e_j is the unit vector with a 1 for the j^{th} entry. For $\mathbf{a} \in C(\bar{\mathbf{D}})$ to hold for all j, we need that all m factors have different and non-zero means. Given this condition implies that $\boldsymbol{\gamma}_i$ is estimable. Given the expression in 4.15 It follows that we have:

$$\tilde{\boldsymbol{\Gamma}}_i = (\bar{\mathbf{D}}_{\cdot}\bar{\mathbf{D}}_{\cdot}')^{-}\bar{\mathbf{D}}_{\cdot}(\bar{\mathbf{z}}_{i\cdot} - \bar{\mathbf{u}}_{i\cdot}) \ \rightarrow \bar{\mathbf{z}}_{i\cdot} R_2 \tag{4.16}$$

where we have used the generalized inverse. Assuming that u_{it} has zero mean for each individual, we have that $\bar{\mathbf{u}}_{i\cdot} \to_p 0$. Also similarly $\bar{\mathbf{D}}_{\cdot} \to_p D$ where D is a constant matrix (Since by assumption the loadings have finite mean). Hence $(\bar{\mathbf{D}}_{\cdot}\bar{\mathbf{D}}_{\cdot}')^{-1}\bar{\mathbf{D}}_{\cdot} \to_p R_2$, where R_2 is an unknown constant matrix. This justifies that $(\tilde{\mathbf{\Gamma}} - R_2\bar{\mathbf{z}}_{i\cdot}) \to_p 0$. Substituting 4.11 and 4.16 into the original model we have the following approximation:

$$y_{it} \approx \boldsymbol{\beta}_i'\mathbf{x}_{it} + \bar{\mathbf{z}}_{i\cdot}'R_2'R_1\bar{\mathbf{z}}_{\cdot t} + \epsilon_{it} \tag{4.17}$$

Stacking the equations we have:

$$\mathbf{y} \approx \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta}\bar{\mathbf{H}} + \epsilon \tag{4.18}$$

Where $\boldsymbol{\beta} = (\beta_{11}, \beta_{1k}, ..., \beta_{N1}, ..., \beta_{Nk})$, and $\boldsymbol{\delta}$ is an unknown $(k+1)(k+3)$ vector of coefficients. Also

$$\mathbf{X} = \begin{pmatrix} \mathbf{X_1} & 0 & ... & 0 \\ 0 & \mathbf{X}_2 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & ... & 0 & \mathbf{X}_N \end{pmatrix} \tag{4.19}$$

With $\mathbf{X}_i$ being a $T \times k$ matrix of covariates for each individual

$$\bar{\mathbf{H}} = \begin{pmatrix} \bar{\mathbf{z}}_{1\cdot} & \bar{\mathbf{z}}_{\cdot 1} & h_{11}^{(11)} & ... & h_{1,k+1}^{(11)} & ... & h_{k+1,k+1}^{(11)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{\mathbf{z}}_{N\cdot} & \bar{\mathbf{z}}_{\cdot T} & h_{11}^{(NT)} & ... & h_{1,k+1}^{(NT)} & ... & h_{k+1,k+1}^{(NT)} \end{pmatrix} \tag{4.20}$$

With $h_{ij}^{(ql)} = \bar{\mathbf{z}}_{i\cdot}^{(q)} \cdot \bar{\mathbf{z}}_{\cdot j}^{(l)}$. Where $i, j = 1, ..., (k+1)$ indicate which variables are included in the interaction and $q = 1, ..., N$ and $l = 1, ..., T$ indicate the observation.

In simpler terms, what this means is that it is sufficient to include the means with respect to individuals and time ($\bar{\mathbf{z}}_{i\cdot}$ and $\bar{\mathbf{z}}_{\cdot t}$), as well as the interactions of all of the terms of $\bar{\mathbf{z}}_{i\cdot}$ and $\bar{\mathbf{z}}_{\cdot t}$ as proxies for the interactive effects. This method can potentially result in large computational efficiency gains over CCE which is a very computationally intense estimator.

There is also what I believe to be a simpler derivation of CCE and TWCCE using expectations which goes as follows, starting from equation 4.6 and taking expectations:

$$\mathbf{z}_{it} = \mathbf{C}_i'\mathbf{f}_t + \mathbf{u}_{it} \tag{4.21}$$

$$E_i(\mathbf{z}_{it}) = E_i(\mathbf{C}_i')\mathbf{f}_t + E_i(\mathbf{u}_{it}) \tag{4.22}$$

By the law of large numbers, the samples means converge in probability to the population means, so we can substitute to get the approximation. Assuming the errors have zero mean the error term drops out, and we have

$$\bar{\mathbf{z}}_{.t} \approx \bar{\mathbf{C}}.\mathbf{f}_t \tag{4.23}$$

We can then solve for $\mathbf{f}_t$ in the same manner as previously to get the same result. Analogously for the factor loadings

$$E_t(\mathbf{z}_{it}) = \tilde{\mathbf{\Gamma}}_i' E_t(\mathbf{D}_t) + E_t(\mathbf{u}_{it}) \tag{4.24}$$

Again, by the law of large numbers, the samples means converge in probability to the population means, so we can substitute to get the approximation. Assuming the errors have zero mean the error term drops out, and we have

$$\bar{\mathbf{z}}_{i.} \approx \tilde{\mathbf{\Gamma}}_i'\bar{\mathbf{D}}. \tag{4.25}$$

The rest follows identically. Define

$$\hat{\beta} = (X'MX)^{-1}(X'MY) \tag{4.26}$$

where $M = I_{NT} - \bar{H}(\bar{H}'\bar{H})^{-1}\bar{H}' = I_{NT} - P_H$ and $\bar{H}$ is defined in 4.20. Note that the original model 4.1 can be rewritten using M,

$$MY = MX\boldsymbol{\beta} + M\boldsymbol{e} \tag{4.27}$$

$$\tilde{Y} = \tilde{X}\boldsymbol{\beta} + \tilde{\boldsymbol{e}} \tag{4.28}$$

Where we refer to the constituent elements of $\tilde{X}$ and $\tilde{x}_i$.

4.3 Consistency and Asymptotic Normality

In this section we will state and prove the asymptotic properties of Two-way CCE more formally. Note first that we can stack the original model in equation 4.1 as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{F} \otimes I_N)\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{4.29}$$

With $\mathbf{X}$ defined above, $\boldsymbol{\gamma} = (\gamma_{11}, ...\gamma_{1m}, ...\gamma_{Nm})$ and

$$\mathbf{F} = \begin{pmatrix} f_{11} & f_{12} & ... & f_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ f_{T1} & f_{T2} & ... & f_{Tm} \end{pmatrix} \tag{4.30}$$

With f_{ij} indicating the j^{th} factor in the i^{th} time period.

Theorem 4.1 (Consistency) *Consider the panel data model in equations 4.1, 4.2, and 4.3. Given assumptions 4.1-4.5, further assume that the rank conditions Rank($\bar{\mathbf{C}}$.) $= m \leq k+1$, Rank($\bar{\mathbf{D}}$.) $= m \leq k+1$, are satisfied. Further assume that $E(\tilde{x}_i\tilde{x}') < \infty$, and that $\boldsymbol{\Gamma}_{ji}f_t = \boldsymbol{\gamma}_i A_j \mathbf{f}_t$. Then as $(N,T) \to \infty$, $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$.*

Theorem 4.2 (Asymptotic Normality) *Given the assumptions of Theorem 1,*

$$\sqrt{NT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d N(0, \boldsymbol{\Sigma}_\beta) \tag{4.31}$$

Where $\boldsymbol{\Sigma}_\beta = C\boldsymbol{\Omega}C'$, where $\Omega = E(\tilde{x}_i\tilde{x}_i'\varepsilon_i^2)$, and $C = E(\tilde{x}_i\tilde{x}_i')$

4.4 Simulations

In this section we conduct Monte Carlo simulations to investigate the finite sample properties of our estimator. We compare the bias, RMSE, power, and type 1 error rate of Two-Way CCE with the original CCE estimator in addition to the augmented mean group (AMG), and Interactive Effects (IFE) estimators. For reference we also compute statistics for OLS, where we simply regress y on x without accounting for the factors, and the infeasible estimator in which we assume the factors are known. In addition, we compare the computation times for the above methods.

Consider the following data generating process similar to Pesaran (2006).

$$y_{it} = \alpha_{i1} d_{1t} + \beta_{i1} x_{1it} + \beta_{i2} x_{2it} + \gamma_{i1} f_{1t} + \gamma_{i2} f_{2t} + \epsilon_{it} \tag{4.32}$$

and

$$x_{ijt} = a_{ij1} d_{1t} + a_{ij2} d_{2t} + \gamma_{ij1} f_{1t} + \gamma_{ij3} f_{3t} + \nu_{ijt}, \quad j = 1, 2 \tag{4.33}$$

for $i = 1, 2, ..., N$ and $t = 1, 2, ..., T$. The factors and errors are generated as $AR(1)$ processes.

$$d_{1t}, \quad d_{2t} = \rho_d d_{2,t-1} + \nu_{dt}, \quad t = -49, ..., 1, ..., T$$
$$\nu_{dt} \sim IIDN(0, 1 - \rho_d^2), \quad \rho_d = 0.5, \quad d_{2,-50} = 0$$

$$f_{jt} = \rho_{fj} f_{j,t-1} + \nu_{fjt} \quad \text{for} \quad j = 1, 2, 3, \quad t = -49, ..., 0, ..., T$$
$$v_{fjt} \sim IIDN(0, 1 - \rho_{fj}^2), \quad \rho_{fj} = 0.5, \quad f_{j,-50} = 0, \text{ for } j = 1, 2, 3,$$

$$\nu_{ijt} = \rho_{vij} \nu_{ij,t-1} + \eta_{ijt}$$
$$\eta_{ijt} \sim IIDN(0, 1 - \rho_{vij}^2), \quad v_{ji,-50} = 0$$

$$\rho_{vij} \sim IIDU[0.05, 0.95] \text{ for } j = 1, 2.$$

So we have 3 factors ($m = 3$) and 2 covariates ($k = 2$). The error of y_{it} are generated as an AR(1) process for half of the units and MA(1) for the other half as in Pesaran 2006. This will help us investigate the properties of TWCCE under serial correlation and heteroskedasticity (namely type 1 error and power with normal standard errors).

$$\epsilon_{it} = \rho_{i\epsilon} \epsilon_{i,t-1} + \sigma_i (1 - \rho_{i\epsilon}^2)^{1/2} \xi_{it}, \quad i = 1, 2, ..., \frac{N}{2}$$
$$\epsilon_{it} = \sigma_i (1 + \theta_{i\epsilon}^2)^{-1/2} (\xi_{it} + \theta_{i\epsilon} \xi_{i,t-1}), \quad i = \frac{N}{2} + 1, ..., N$$
$$\xi_{it} \sim IIDN(0, 1)$$
$$\sigma_i^2 \sim IIDU[0.5, 1.5]$$
$$\theta_{i\epsilon} \sim IIDU[0, 1]$$

In the context of the theory presented in Section 4.2, this is a restricted form of that more general framework with $\gamma_i' = (\gamma_{i1}, \gamma_{i2}, 0)$, $f_t' = (f_{1t}, f_{2t}, f_{3t})$. Pesaran 2006 highlights that

CCE is robust to serial correlation and heteroskedasticity. This DGP will also allow us to investigate how the robustness of TWCCE to serial correlation.

We will be comparing the case of both heterogeneous slopes and homogeneous slopes. While the Two-Way CCE estimator was not designed specifically with heterogeneous slopes in mind, it will be of interest to see how robust it is to that case. To summarize, we will conduct 2 different experiments:

- **Experiment 1**: Homogeneous slopes

$$\beta_{i1} = \beta_{i2} = 1 \tag{4.34}$$

- **Experiment 2**: Heterogeneous slopes

$$\beta_{i1} = 1 + N(0, 0.4) \tag{4.35}$$

$$\beta_{i2} = 1 + N(0, 0.4) \tag{4.36}$$

Each experiment was simulated 100 times for each (N,T) pair. Simulation results are summarized in Tables 4.1-4.10.

4.4.1 Bias and RMSE

Tables 4.1 and 4.2 show that bias and RMSE results for the first case of homogeneous slope coefficients. Tables 4.3 and 4.4 display the bias and RMSE for the case of heterogeneous slopes. Column (1)-(6) denote TWCCE, CCE, OLS, IFE, AMG and the infeasible estimator (INF) respectively. Unsurprisingly the OLS estimates are severely biased for all sizes of N and T, exhibiting bias roughly in the range of 8% to 16%, with no improvement for large N or T. In fact, OLS seems to become more biased with larger T. The RMSE of OLS is also quite high in all cases. In contrast, the proposed new TWCCE estimator displays negligible bias in the samples considered. The highest bias is exhibited with the smallest N and T ($(N, T) = (50, 5)$). This demonstrates that the new estimator doesn't appear to

require outrageously large sample sizes in order to attain negligible bias.

TWCCE also performs best in the case of large T and large N which is the opposite of OLS. Interestingly TWCCE has very low RMSE compared to other factor model estimators, being comparable to the infeasible estimator. While the infeasible estimator displays the lowest bias of any of the estimators, TWCCE must be competitive with it in terms of variability. CCE displays similar properties to TWCCE, in that it performs better as both N and T increase. However the bias for small T is noticeably larger. In the case of $N = 200$ and $T = 5$, CCE has bias in excess of OLS. This is one of the main drawbacks of ordinary CCE. If T is small and N is not sufficiently large, there can be some instability in the estimates.

Not surprisingly IFE and AMG also perform well in this experiment. All factor model estimators appear to display properties of consistency for increasing sample size, with near zero biases in all cases except the case of $N = 50, T = 5$. The RMSE indicates that in smaller samples, TWCCE has a lower variability than the other factor model estimators. In the case of $N = 50$,$T = 5$ the RMSE for TWCCE is 0.070 compared to 0.114 for CCE, 0.330 for AMG, and 0.082 for IFE. However this changes for larger values of N and T, with IFE displaying the lowest level of variability for all cases of $T = 120$

Table 4.1: Bias of Factor Model Estimators (Experiment 1)

		Estimators					
N	T	TWCEE	CCE	OLS	AMG	IFE	INF
50	5	-0.007	-0.018	0.079	0.025	0.007	-0.004
	12	0.002	0.005	0.132	0.002	-0.002	0.001
	120	0.002	0.001	0.159	0.003	0.001	-0.001
200	5	-0.004	-0.113	0.084	-0.001	0.001	0.001
	12	-0.002	-0.004	0.122	-0.003	-0.002	0.000
	120	-0.002	-0.002	0.162	-0.001	-0.001	0.000
1000	5	-0.001	0.002	0.093	0.003	0.001	0.001
	12	-0.001	0.000	0.114	0.000	0.000	-0.001
	120	0.000	0.000	0.171	0.000	0.000	0.000

[1] OLS = Ordinary Least Squares, CCE = Common Correlated Effects Estimator, IFE = Interactive Effects Estimator, AMG = Augmented Mean Group Estimator, INF = Infeasible Estimator.

Table 4.2: RMSE of Factor Model Estimators (Experiment 1)

		Estimators					
N	T	TWCEE	CCE	OLS	AMG	IFE	INF
50	5	0.070	0.114	0.181	0.330	0.082	0.071
	12	0.042	0.062	0.184	0.044	0.034	0.045
	120	0.024	0.019	0.173	0.018	0.013	0.013
200	5	0.030	0.304	0.162	0.074	0.036	0.039
	12	0.022	0.029	0.184	0.021	0.017	0.021
	120	0.013	0.015	0.170	0.009	0.006	0.007
1000	5	0.014	0.022	0.177	0.053	0.013	0.017
	12	0.010	0.013	0.166	0.010	0.008	0.010
	120	0.007	0.008	0.177	0.004	0.003	0.003

[1] OLS = Ordinary Least Squares, CCE = Common Correlated Effects Estimator, IFE = Interactive Effects Estimator, AMG = Augmented Mean Group Estimator, INF = Infeasible Estimator.

Table 4.3: Bias of Factor Model Estimators (Experiment 2)

		Estimators					
N	T	TWCEE	CCE	OLS	AMG	IFE	INF
50	5	-0.003	-0.025	0.091	0.026	-0.010	-0.002
	12	-0.012	-0.003	0.132	-0.005	-0.010	-0.003
	120	-0.005	-0.001	0.156	-0.006	-0.008	-0.010
200	5	-0.001	-0.003	0.115	-0.011	0.003	0.003
	12	-0.008	-0.010	0.133	-0.005	-0.003	-0.008
	120	0.003	0.005	0.158	0.003	0.003	0.001
1000	5	-0.004	0.000	0.101	-0.018	0.000	-0.003
	12	0.000	0.000	0.119	0.001	0.000	-0.001
	120	0.001	0.002	0.164	0.001	0.001	0.000

[1] OLS = Ordinary Least Squares, CCE = Common Correlated Effects Estimator, IFE = Interactive Effects Estimator, AMG = Augmented Mean Group Estimator, INF = Infeasible Estimator.

Table 4.4: RMSE of Factor Model Estimators (Experiment 2)

		Estimators					
N	T	TWCEE	CCE	OLS	AMG	IFE	INF
50	5	0.124	0.159	0.245	0.199	0.133	0.113
	12	0.091	0.101	0.191	0.080	0.093	0.085
	120	0.075	0.078	0.181	0.064	0.071	0.058
200	5	0.050	0.061	0.197	0.102	0.064	0.065
	12	0.044	0.045	0.188	0.040	0.041	0.043
	120	0.034	0.033	0.167	0.030	0.034	0.032
1000	5	0.026	0.029	0.173	0.055	0.027	0.030
	12	0.019	0.020	0.176	0.015	0.018	0.020
	120	0.017	0.016	0.171	0.014	0.015	0.015

[1] OLS = Ordinary Least Squares, CCE = Common Correlated Effects Estimator, IFE = Interactive Effects Estimator, AMG = Augmented Mean Group Estimator, INF = Infeasible Estimator.

4.4.2 Type 1 Error and Power

Tables 4.5 and 4.6 show the type 1 error and power results for the first case of homogeneous slope coefficients. Tables 4.7 and 4.8 display the size and power for the case of heterogeneous slopes. Type 1 error is computed averaging the results of testing the hypothesis:

$$H_0 : \ \beta_1 = 1 \tag{4.37}$$

$$H_1 : \ \beta_1 \neq 1 \tag{4.38}$$

Where the null hypothesis represents the true value of β. Optimally the type 1 error rate should be 0.05. Power is computed by computing the average rejection rate of the following hypothesis:

$$H_0 : \ \beta_1 \leq 0.95 \tag{4.39}$$

$$H_1 : \ \beta_1 > 0.95 \tag{4.40}$$

Where the null hypothesis in this case represents an incorrect value of β_1. Hypothesis testing was conducted using normal OLS standard errors. This was done to see how much effect

serial correlated errors have on TWCCE and other factor model estimators when using regular standard errors.

Looking at Table 4.5, TWCCE appears to have a problem as the type 1 error rate is always largely in excess of 0.05 and increases with T. This implies that ordinary standard errors are not sufficient when using TWCCE under serial correlation. CCE also performs poorly in most cases. AMG and IFE on the other hand have type 1 error rates near 0.05 when using regular standard errors.

Regarding power, CCE appears to do quite well, with power exceeding other factor model estimators. However, this is clearly at the cost of a higher type 1 error rate.

Table 4.5: Type 1 Error Rate of Factor Model Estimators (Experiment 1)

		Estimators					
N	T	TWCEE	CCE	OLS	AMG	IFE	INF
50	5	0.09	0.10	0.45	0.05	0.06	0.02
	12	0.10	0.08	0.68	0.10	0.02	0.08
	120	0.30	0.04	0.99	0.07	0.06	0.06
200	5	0.12	0.23	0.67	0.04	0.08	0.07
	12	0.12	0.13	0.84	0.05	0.06	0.05
	120	0.31	0.39	0.99	0.04	0.08	0.09
1000	5	0.12	0.06	0.84	0.06	0.01	0.03
	12	0.14	0.15	0.94	0.07	0.06	0.06
	120	0.37	0.45	1.00	0.09	0.06	0.09

[1] OLS = Ordinary Least Squares, CCE = Common Correlated Effects Estimator, IFE = Interactive Effects Estimator, AMG = Augmented Mean Group Estimator, INF = Infeasible Estimator.

4.4.3 Computation Time

In this section we compare the average computation time of each estimator over 100 simulations. Table 4.9 displays computation times in seconds for experiment 1. Table 4.10 displays computation times in seconds for experiment 2. We can see in Table 4.9 that the computation time of TWCCE is lower in all sample sizes when compared to any other factor model estimator. In fact, it has computation time even less than the infeasible estimator. The dif-

Table 4.6: Power of Factor Model Estimators (Experiment 1)

		Estimators					
N	T	TWCEE	CCE	OLS	AMG	IFE	INF
50	5	0.22	0.12	0.51	0.05	0.14	0.09
	12	0.33	0.18	0.76	0.26	0.33	0.19
	120	0.89	0.78	1.00	0.90	0.98	0.96
200	5	0.47	0.01	0.68	0.10	0.34	0.35
	12	0.73	0.53	0.81	0.68	0.81	0.69
	120	1.00	0.97	1.00	1.00	1.00	1.00
1000	5	0.96	0.70	0.74	0.27	0.94	0.83
	12	0.99	0.99	0.85	1.00	1.00	0.99
	120	1.00	1.00	1.00	1.00	1.00	1.00

[1] OLS = Ordinary Least Squares, CCE = Common Correlated Effects Estimator, IFE = Interactive Effects Estimator, AMG = Augmented Mean Group Estimator, INF = Infeasible Estimator.

Table 4.7: Type 1 Error Rate of Factor Model Estimators (Experiment 2)

		Estimators					
N	T	TWCEE	CCE	OLS	AMG	IFE	INF
50	5	0.25	0.16	0.44	0.05	0.12	0.04
	12	0.26	0.25	0.55	0.10	0.33	0.18
	120	0.72	0.70	0.90	0.10	0.71	0.61
200	5	0.20	0.08	0.54	0.06	0.17	0.09
	12	0.29	0.25	0.82	0.09	0.25	0.21
	120	0.72	0.75	1.00	0.02	0.72	0.63
1000	5	0.25	0.09	0.84	0.07	0.10	0.09
	12	0.31	0.28	0.86	0.03	0.30	0.24
	120	0.07	0.77	1.00	0.03	0.65	0.56

[1] OLS = Ordinary Least Squares, CCE = Common Correlated Effects Estimator, IFE = Interactive Effects Estimator, AMG = Augmented Mean Group Estimator, INF = Infeasible Estimator.

Table 4.8: Power of Factor Model Estimators (Experiment 2)

		Estimators					
N	T	TWCEE	CCE	OLS	AMG	IFE	INF
50	5	0.22	0.09	0.41	0.05	0.10	0.09
	12	0.23	0.23	0.67	0.10	0.29	0.16
	120	0.56	0.62	0.98	0.15	0.60	0.54
200	5	0.35	0.22	0.65	0.09	0.23	0.24
	12	0.47	0.41	0.83	0.25	0.50	0.37
	120	0.84	0.85	1.00	0.50	0.85	0.88
1000	5	0.66	0.50	0.79	0.15	0.53	0.54
	12	0.93	0.93	0.89	0.87	0.92	0.85
	120	1.00	1.00	1.00	0.97	1.00	1.00

[1] OLS = Ordinary Least Squares, CCE = Common Correlated Effects Estimator, IFE = Interactive Effects Estimator, AMG = Augmented Mean Group Estimator, INF = Infeasible Estimator.

ference is largest between TWCCE and CCE. This was one of the motivations for creating TWCCE, since CCE quickly becomes intractable for large sample sizes. We can see that for $N = 1000$ CCE takes nearly 2 minutes to compute for $T = 5$ and $T = 12$, while TWCCE takes 0.038 and 0.079 seconds respectively. IFE also performs quite well computationally, however TWCCE is still much faster to compute (TWCCE appears to be roughly 5 times as fast). AMG performs better than CCE but worse than IFE. we can see that computation time begins to rapidly increase for larger values of N. These results lead us to conclude that TWCCE is the most computationally efficient of the estimators considered.

4.5 Application to Minimum Wage Research

In this section we apply the new estimator TWCCE and also CCE to an application in minimum wage research. An in depth discussion of the validity or importance of factor models in micro panels is beyond the scope of this paper, this section simply illustrates how it can be applied in practice. In this section we would like to make two key points. First, TWCCE and CCE produce similar estimates in practice. A concern regarding the new estimator is that the new assumptions may be unreasonable, however this does not appear

Table 4.9: Computation Time of Factor Model Estimators (Experiment 1)

		Estimators					
N	T	TWCEE	CCE	OLS	AMG	IFE	INF
50	5	0.023	0.093	0.010	0.241	0.092	0.121
	12	0.029	0.240	0.010	0.267	0.105	0.058
	120	0.056	0.282	0.012	1.333	0.301	0.069
200	5	0.021	0.829	0.006	0.782	0.118	0.370
	12	0.035	0.937	0.009	0.950	0.155	0.398
	120	0.158	1.141	0.036	6.439	0.993	0.536
1000	5	0.038	116.080	0.010	4.986	0.319	30.615
	12	0.079	117.310	0.019	6.426	0.438	32.597
	120	0.672	122.662	0.161	68.643	3.741	33.505

[1] OLS = Ordinary Least Squares, CCE = Common Correlated Effects Estimator, IFE = Interactive Effects Estimator, AMG = Augmented Mean Group Estimator, INF = Infeasible Estimator.

Table 4.10: Computation Time of Factor Model Estimators (Experiment 2)

		Estimators					
N	T	TWCEE	CCE	OLS	AMG	IFE	INF
50	5	0.022	0.097	0.010	0.240	0.095	0.062
	12	0.023	0.096	0.009	0.256	0.097	0.062
	120	0.059	0.123	0.014	1.259	0.323	0.072
200	5	0.033	1.355	0.018	1.478	0.288	0.722
	12	0.058	1.455	0.019	1.636	0.328	0.752
	120	0.256	1.674	0.086	13.417	2.628	0.632
1000	5	0.038	119.967	0.013	5.792	0.378	32.314
	12	0.072	120.864	0.019	7.404	0.499	33.094
	120	0.695	125.908	0.166	73.350	3.966	34.868

[1] OLS = Ordinary Least Squares, CCE = Common Correlated Effects Estimator, IFE = Interactive Effects Estimator, AMG = Augmented Mean Group Estimator, INF = Infeasible Estimator.

to be a major issue when it comes to this particular application. Second, if we believe the assumptions of TWCCE and CCE, then they can help us determine the true effect of minimum wage. However the second point requires much further investigation beyond this paper.

In the past few years there has been significant debate about the model specifications for minimum wage research. This can mainly be boiled down to the question of what kinds of fixed effects should we include? The paper by Dube, Lester, and Reich 2010 (which will be called DLR) and Allegretto, Dube, and Reich 2011 (which we will call ADR) find minimal impact on employment when census division specific time effects are included in the model. These research designs have been criticized by Neumark, Salas, and Wascher 2014, and Neumark and Wascher 2015. Previous literature has generally found the minimum wage to have a significant impact on employment.

ADR fit a few different models that are variants of the following:

$$\text{(Two-Way FE)} \quad y_{ist} = \beta MW_{st} + X_{ist}\Gamma + \lambda \cdot unemp_{st} + \phi_s + \tau_t + \epsilon_{ist}$$

Where y_{ist} is the variable of interest (employment status, log wage, or log hours worked), MW_{st} is the minimum wage in state s and time t, $unemp_{st}$ is the quarterly unemployment, ϕ_s is the state fixed effect and τ_t are the time dummies. The above is the canonical model used for minimum wage research. ADR then consider three slightly different specifications:

$$\text{(ADR1)} \quad y_{ist} = \beta MW_{st} + X_{ist}\Gamma + \lambda \cdot unemp_{st} + \phi_s + \tau_{dt} + \epsilon_{ist}$$
$$\text{(ADR2)} \quad y_{ist} = \beta MW_{st} + X_{ist}\Gamma + \lambda \cdot unemp_{st} + \phi_s + \psi_s \cdot t + \tau_t + \epsilon_{ist}$$
$$\text{(ADR3)} \quad y_{ist} = \beta MW_{st} + X_{ist}\Gamma + \lambda \cdot unemp_{st} + \phi_s + \psi_s \cdot t + \tau_{dt} + \epsilon_{ist}$$

Where d refers to census division, τ_{dt} are the division specific time effects, and $\psi_s \cdot t$ is a state specific time trend. When ADR estimate models ADR1-3, they find that there is no significant impact of the minimum wage on employment. Additionally, they find that with models ADR1-3 the effect on wages is estimated to be larger than for Two-Way FE and the reduction in hours worked is found to be less than for Two-Way FE.

As discussed in Section 4.1, all of the above model specifications are actually special cases of an interactive effects (or common factor) model. Totty 2017 suggests that instead of just using fixed effect models, we should apply factor model estimators, as there could be additional time varying factors that are not being controlled for by the fixed effects. If there aren't additional factors, then in theory the factor model estimators should agree with the fixed effects estimators (although this may not be the case as the factor model estimators might pick up a large amount of exogenous variation). That is, all of the fixed effect models can be described in the general form:

$$y_{ist} = \beta MW_{st} + X_{ist}\Gamma + \lambda \cdot unemp_{st} + \lambda_s \cdot f_t + \epsilon_{ist} \tag{4.41}$$

Where each specification simply applies a restriction on the factor loadings (λ_s) and/or common factors (f_t). For example, we can recover the canonical two-way fixed effects model by letting $\lambda_s = (\phi_s \quad 1)$ and $f_t = (1 \quad \tau_t)'$. Hence the application of factor model estimators need not necessarily be restricted to cases where we are concerned with this more complicated factor error structure. This means that if we use an estimator that is robust to most forms of $\lambda_s \cdot f_t$, then it should agree with which of the above models is most correct (if any). Totty applied the Common Correlated Effects (CCE) estimator developed by Pesaran 2006, and the Interactive Effects Estimator (IFE) developed by Bai 2009 for this purpose and his results tended to agree with ADR.

The data is CPS data spanning from 1990-2009, containing data on demographic characteristic, wages, hours worked, employment status, and location. For comparability with ADR we will be using the same controls as them, however we will be substituting the fixed effects with interactive effects. Controls are for gender, race (four categories), age (four categories), education (twelve categories), and marital status (four categories), as well as controls for the non-seasonally adjusted unemployment rate, and the relevant population share for each demographic group. The following table displays the estimates for the original 4 specifications for β taken from ADR in columns 1-4, as well as for CCE in column 5, and Two-Way CCE in column 6.

Interestingly, we can see that for the effect of minimum wage on employment, the CCE

and TWCCE estimates are almost identical to the estimates from the two-way fixed effects specification. The positive effect on wages is estimated to be higher than any of the original 4 specifications, and the reduction in hours is found to be insignificant. However there is no significant difference between CCE and the fixed effects models. If our hypothesis is correct, then this implies that the canonical model is correct for measuring the effect of the minimum wage on employment. Interestingly it appears at a glance that none of the 4 models is correct for measuring the effect of the minimum wage on wages.

4.6 Summary, Conclusion and Extensions

This paper proposes a new approach to the estimation of factor models in the presence of panel data. We apply a simple modification to the Common Correlated Effects estimator resulting in the Two-Way CCE estimator. We derived the asymptotic theory regarding this new estimator as $N \to \infty$ and $T \to \infty$.

Simulations demonstrate that the new estimator appears to be comparable in Bias, RMSE, and power to other factor model estimators. Specifically we compared it to the CCE, IFE and AMG estimators. It exhibits similar properties and displays consistency even in finite samples. We suspected that the new estimator would require a larger sample for convergence due to the increased number of added covariates but this appears not to be the case.

One potential extension I believe regarding the Two-Way CCE estimator (and also CCE and other factor model estimators) is its potential ability to test fixed effect model specifications. I believe that it may be possible to argue from this an additional test that can help select fixed effect model specifications. As discussed in Section 4.5, there is significant debate in labor research regarding which fixed effects to include in minimum wage research.

Additionally, as was accomplished in Chudik and Pesaran 2015, we likely can expand two-way CCE to the case of dynamic panels. Dynamic CCE simply augments the model with additional lags of the cross sectional means. Since the dynamic component is only with respect to time I expect two-way CCE can be modified in exactly the same way, adding lags of $\bar{z}_t$ to the model as well as the interactions with said lags. However, this requires further investigation to prove.

Table 4.11: Estimates for the Effect of Minimum Wages on Employment Outcomes

		Estimators					
N	T	TWFE	ADR1	ADR2	ADR3	CCE	TWCCE
Wage							
All Teens	Coeff	0.123***	0.161***	0.165***	0.149***	0.155***	0.161***
	se	(0.026)	(0.030)	(0.025)	(0.024)	(0.0168)	(0.0275)
Teens 16-17	Coeff	0.197***	0.224***	0.221***	0.220***	0.197***	0.235***
	se	(0.032)	(0.036)	(0.030)	(0.033)	(0.0195)	(0.0377)
Teens 18-19	Coeff	0.074**	0.115***	0.120***	0.093***	0.125***	0.112***
	se	(0.030)	(0.037)	(0.038)	(0.033)	(0.0230)	(0.0402)
Employment							
All Teens	Coeff	-0.047**	-0.015	-0.014	0.019	-0.042**	0.051*
	se	(0.022)	(0.034)	(0.027)	(0.024)	(0.0169)	(0.0268)
Teens 16-17	Coeff	-0.069**	-0.023	-0.021	0.030	-0.069***	-0.078**
	se	(0.028)	(0.043)	(0.032)	(0.032)	(0.0177)	(0.034)
Teens 18-19	Coeff	-0.027	-0.005	-0.010	0.009	0.014	.0.063**
	se	(0.021)	(0.034)	(0.027)	(0.027)	(0.0255)	(0.025)
Hours							
All Teens	Coeff	-0.074**	-0.054	-0.001	-0.032	-0.022	0.051
	se	(0.035)	(0.048)	(0.040)	(0.042)	(0.0544)	(0.0418)
Teens 16-17	Coeff	-0.070	0.002	-0.011	0.038	-0.031	-0.070
	se	(0.042)	(0.074)	(0.044)	(0.073)	(0.055)	(0.0473)
Teens 18-19	Coeff	-0.090**	-0.082*	-0.011	-0.079	-0.024	-0.052
	se	(0.042)	(0.049)	(0.050)	(0.042)	(0.0544)	(0.361)

[1] CCE = Common Correlated Effects Estimator, TWFE = Two-way Fixed Effects, TWCCE = Two-Way CCE

[2] Significance levels ***1 percent, **5 percent, *10 percent.

[3] Each specification contains controls for gender, race (4), age(4), education (12), and marital status (4), and nonseasonally adjusted unemployment rate. Log hourly wages is the dependent variable. Wage regressions restricted to those making between $1 and 100 per hour. Standard errors are cluster robust.

Bibliography

Aaronson, Daniel (2001). "Price pass-through and the minimum wage". In: *Review of Economics and statistics* 83.1, pp. 158–169.

Aaronson, Daniel, Eric French, and James MacDonald (2008). "The minimum wage, restaurant prices, and labor market structure". In: *Journal of Human Resources* 43.3, pp. 688–720.

Aaronson, Daniel, Eric French, Isaac Sorkin, et al. (2018). "Industry dynamics and the minimum wage: a putty-clay approach". In: *International Economic Review* 59.1, pp. 51–84.

Abowd, John M et al. (1997). *Minimum wages and youth employment in France and the United States*. Tech. rep. National Bureau of Economic Research.

Allegretto, Sylvia A, Arindrajit Dube, and Michael Reich (2011). "Do minimum wages really reduce teen employment? Accounting for heterogeneity and selectivity in state panel data". In: *Industrial Relations: A Journal of Economy and Society* 50.2, pp. 205–240.

Athey, Susan, Julie Tibshirani, Stefan Wager, et al. (2019). "Generalized random forests". In: *The Annals of Statistics* 47.2, pp. 1148–1178.

Bai, Jushan (2009). "Panel data models with interactive fixed effects". In: *Econometrica* 77.4, pp. 1229–1279.

Barattieri, Alessandro, Susanto Basu, and Peter Gottschalk (2014). "Some evidence on the importance of sticky wages". In: *American Economic Journal: Macroeconomics* 6.1, pp. 70–101.

Basker, Emek and Muhammad Taimur Khan (2016). "Does the minimum wage bite into fast-food prices?" In: *Journal of Labor Research* 37.2, pp. 129–148.

Belloni, Alexandre and Victor Chernozhukov (2011). "High dimensional sparse econometric models: An introduction". In: *Inverse Problems and High-Dimensional Estimation*. Springer, pp. 121–156.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014). "Inference on treatment effects after selection among high-dimensional controls". In: *The Review of Economic Studies* 81.2, pp. 608–650.

Bickel, Peter J, Ya'acov Ritov, and Alexandre B Tsybakov (2009). "Simultaneous analysis of Lasso and Dantzig selector". In: *The Annals of statistics* 37.4, pp. 1705–1732.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Card, David (1992). "Do minimum wages reduce employment? A case study of California, 1987–89". In: *ILR Review* 46.1, pp. 38–54.

Card, David and Alan B Krueger (1993). *Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania*. Tech. rep. National Bureau of Economic Research.

Cengiz, Doruk et al. (2019). "The effect of minimum wages on low-wage jobs". In: *The Quarterly Journal of Economics* 134.3, pp. 1405–1454.

Chen, Xiaohong and Halbert White (1999). "Improved rates and asymptotic normality for nonparametric neural network estimators". In: *IEEE Transactions on Information Theory* 45.2, pp. 682–691.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey (2017). "Double/debiased/neyman machine learning of treatment effects". In: *American Economic Review* 107.5, pp. 261–65.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018). *Double/debiased machine learning for treatment and structural parameters*.

Chernozhukov, Victor, Whitney Newey, and Rahul Singh (2018). "De-biased machine learning of global and local parameters using regularized Riesz representers". In: *arXiv e-prints*, arXiv–1802.

Chetverikov, Denis, Zhipeng Liao, and Victor Chernozhukov (2021). "On cross-validated lasso in high dimensions". In: *The Annals of Statistics* 49.3, pp. 1300–1317.

Chudik, Alexander and M Hashem Pesaran (2015). "Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors". In: *Journal of Econometrics* 188.2, pp. 393–420.

Clemens, Jeffrey and Michael Wither (2019). "The minimum wage and the Great Recession: Evidence of effects on the employment and income trajectories of low-skilled workers". In: *Journal of Public Economics*.

Colangelo, Kyle and Ying-Ying Lee (2020). "Double debiased machine learning nonparametric inference with continuous treatments". In: *arXiv preprint arXiv:2004.03036*.

Currie, Janet and Bruce Fallick (1993). *The minimum wage and the employment of youth: Evidence from the NLSY*. Tech. rep. National Bureau of Economic Research.

Du, Simon Shaolei et al. (2016). "Hypothesis transfer learning via transformation functions". In: *arXiv preprint arXiv:1612.01020*.

Dube, Arindrajit, T William Lester, and Michael Reich (2010). "Minimum wage effects across state borders: Estimates using contiguous counties". In: *The review of economics and statistics* 92.4, pp. 945–964.

— (2016). "Minimum wage shocks, employment flows, and labor market frictions". In: *Journal of Labor Economics* 34.3, pp. 663–704.

Dustmann, Christian et al. (2019). *Reallocation effects of the minimum wage: Evidence from Germany*. Tech. rep. mimeo.

Eberhardt, Markus and Francis Teal (2010). "Productivity Analysis in Global Manufacturing Production." In.

Farrell, Max H (2015). "Robust inference on average treatment effects with possibly more covariates than observations". In: *Journal of Econometrics* 189.1, pp. 1–23.

Farrell, Max H, Tengyuan Liang, and Sanjog Misra (2021). "Deep neural networks for estimation and inference". In: *Econometrica* 89.1, pp. 181–213.

Flores, Carlos A. et al. (Feb. 2012). "Estimating the Effects of Length of Exposure to Instruction in a Training Program: The Case of Job Corps". In: *The Review of Economics and Statistics* 94.1, pp. 153–171. URL: http://ideas.repec.org/a/tpr/restat/v94y2012i1p153-171.html.

Giuliano, Laura (2013). "Minimum wage effects on employment, substitution, and the teenage labor supply: Evidence from personnel data". In: *Journal of Labor Economics* 31.1, pp. 155–194.

Hsu, Yu-Chin et al. (2018). *Direct and indirect effects of continuous treatments based on generalized propensity score weighting*. SES Working Paper 495, University of Fribourg.

Jiang, Yuan, Yunxiao He, and Heping Zhang (2016). "Variable selection with prior information for generalized linear models via the prior LASSO method". In: *Journal of the American Statistical Association* 111.513, pp. 355–376.

Künzel, Sören R et al. (2018). "Transfer learning for estimating causal effects using neural networks". In: *arXiv preprint arXiv:1808.07804*.

Kuzborskij, Ilja (2018). *Theory and algorithms for hypothesis transfer learning*. Tech. rep. EPFL.

Kuzborskij, Ilja and Francesco Orabona (2013). "Stability and hypothesis transfer learning". In: *International Conference on Machine Learning*. PMLR, pp. 942–950.

— (2017). "Fast rates by transferring from auxiliary hypotheses". In: *Machine Learning* 106.2, pp. 171–195.

Lee, Ying-Ying (2018). *Partial Mean Processes with Generated Regressors: Continuous Treatment Effects and Nonseparable Models*. arXiv:1811.00157.

Li, Sai, T Tony Cai, and Hongzhe Li (2020). "Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality". In: *arXiv preprint arXiv:2006.10593*.

Meer, Jonathan and Jeremy West (2016). "Effects of the minimum wage on employment dynamics". In: *Journal of Human Resources* 51.2, pp. 500–522.

Neumark, David, JM Ian Salas, and William Wascher (2014). "Revisiting the minimum wage—Employment debate: Throwing out the baby with the bathwater?" In: *Ilr Review* 67.3_suppl, pp. 608–648.

Neumark, David, Mark Schweitzer, and William Wascher (2004). "Minimum wage effects throughout the wage distribution". In: *Journal of Human Resources* 39.2, pp. 425–450.

Neumark, David and William Wascher (1995). *The effects of minimum wages on teenage employment and enrollment: Evidence from matched CPS surveys*. Tech. rep. National Bureau of Economic Research.

— (2015). "The effects of minimum wages on employment". In: *FRBSF Economic Letter* 37, p. 2015.

Neumark, David and William L Wascher (2008). *Minimum wages*. MIT press.

Ng, Andrew (2016). *The State of Artificial Intelligence*. Youtube. URL: https://www.youtube.com/watch?v=NKpuX_yzdYs.

Nguyen, Cuong et al. (2020). "Leep: A new measure to evaluate transferability of learned representations". In: *International Conference on Machine Learning*. PMLR, pp. 7294–7305.

Pan, Sinno Jialin and Qiang Yang (2009). "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10, pp. 1345–1359.

Pedace, Roberto and Stephanie Rohn (2011). "The impact of minimum wages on unemployment duration: Estimating the effects using the Displaced Worker Survey". In: *Industrial Relations: A Journal of Economy and Society* 50.1, pp. 57–75.

Pesaran, M Hashem (2006). "Estimation and inference in large heterogeneous panels with a multifactor error structure". In: *Econometrica* 74.4, pp. 967–1012.

Powell, David (2017). "Synthetic Control Estimation Beyond Case Studies: Does the Minimum Wage Reduce Employment?" In.

Pratt, Lorien Y (1993). "Discriminability-based transfer between neural networks". In: *Advances in neural information processing systems*, pp. 204–211.

Robinson, Peter M (1988). "Root-N-consistent semiparametric regression". In: *Econometrica: Journal of the Econometric Society*, pp. 931–954.

Rosenstein, Michael T et al. (2005). "To transfer or not to transfer". In: *NIPS 2005 workshop on transfer learning*. Vol. 898, pp. 1–4.

Schochet, Peter Z, John Burghardt, and Sheena McConnell (2008). "Does job corps work? Impact findings from the national job corps study". In: *American economic review* 98.5, pp. 1864–86.

Su, Liangjun, Takuya Ura, and Yichong Zhang (2019). "Non-separable models with high-dimensional data". In: *Journal of Econometrics* 212.2, pp. 646–677.

Takada, Masaaki and Hironori Fujisawa (2020). "Transfer Learning via ℓ_1 Regularization". In: *arXiv preprint arXiv:2006.14845*.

Thompson, Jeffrey P (2009). "Using local labor market data to re-examine the employment effects of the minimum wage". In: *ILR Review* 62.3, pp. 343–366.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Totty, Evan (2017). "The effect of minimum wages on employment: A factor model approach". In: *Economic Inquiry* 55.4, pp. 1712–1737.

Vaghul, Kavya and Ben Zipperer (2016). "Historical state and sub-state minimum wage data". In: *Washington Center for Equitable Growth*.

Weiss, Karl, Taghi M Khoshgoftaar, and DingDing Wang (2016). "A survey of transfer learning". In: *Journal of Big data* 3.1, p. 9.

White, Halbert (1989). "Learning in artificial neural networks: A statistical perspective". In: *Neural computation* 1.4, pp. 425–464.

White, Halbert et al. (1992). *Artificial neural networks*. Blackwell Cambridge, Mass.

Xu, Qian and Qiang Yang (2011). "A survey of transfer and multitask learning in bioinformatics". In: *Journal of Computing Science and Engineering* 5.3, pp. 257–268.

Yosinski, Jason et al. (2014). "How transferable are features in deep neural networks?" In: *arXiv preprint arXiv:1411.1792*.

Zavodny, Madeline (2000). "The effect of the minimum wage on employment and hours". In: *Labour Economics* 7.6, pp. 729–750.

Zellner, Arnold (1962). "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias". In: *Journal of the American statistical Association* 57.298, pp. 348–368.

Zhang, Yu and Qiang Yang (2017). "A survey on multi-task learning". In: *arXiv preprint arXiv:1707.08114*.

Zhuang, Fuzhen et al. (2020). “A comprehensive survey on transfer learning”. In: *Proceedings of the IEEE* 109.1, pp. 43–76.

Appendix A

Chapter 1

A.1 Proof of Theorem 1.1

First note that by construction we have that,

$$L(\hat{\beta}) \leq L(\beta^*)$$

Subsituting the definition of L,

$$\begin{aligned}
&\frac{1}{2n}\|Y - X\hat{\beta}\|_2^2 + \lambda_1(\alpha(\|\hat{\beta}\|_1 + \lambda_2\|\hat{\beta}\|_2^2) + (1-\alpha)\|\hat{\beta} - \tilde{\beta}\|_1) \\
&\leq \frac{1}{2n}\|Y - X\beta^*\|_2^2 + \lambda_1(\alpha(\|\beta^*\|_1 + \lambda_2\|\beta^*\|_2^2) + (1-\alpha)\|\beta^* - \tilde{\beta}\|_1)
\end{aligned}$$

For notational simplicity, we write λ_j instead of $\frac{\lambda_j}{n}$. Note that $\|Y - X\beta^*\|_2^2 = \|\varepsilon\|_2^2$, and that $\|Y - X\hat{\beta}\|_2^2 = \|X\beta^* + \varepsilon - X\hat{\beta}\|_2^2 = \|\varepsilon - X(\hat{\beta} - \beta^*)\|_2^2 = (\varepsilon - X(\hat{\beta} - \beta^*))'(\varepsilon - X(\hat{\beta} - \beta^*)) = \varepsilon'\varepsilon - (X(\hat{\beta} - \beta^*))'\varepsilon - \varepsilon' X(\hat{\beta} - \beta^*) + (X(\hat{\beta} - \beta^*))'(X(\hat{\beta} - \beta^*)) = \|\varepsilon\|_2^2 - 2\varepsilon' X(\hat{\beta} - \beta^*) + \|X(\hat{\beta} - \beta^*)\|_2^2$.

Subsituting this we have:

$$\frac{1}{2n}\big(\|\varepsilon\|_2^2 - 2\varepsilon' X(\hat{\beta}-\beta^*) + \|X(\hat{\beta}-\beta^*)\|_2^2\big) + \lambda_1(\alpha(\|\hat{\beta}\|_1 + \lambda_2\|\hat{\beta}\|_2^2) + (1-\alpha)\|\hat{\beta}-\tilde{\beta}\|_1)$$
$$\leq \frac{1}{2n}\|\varepsilon\|_2^2 + \lambda_1(\alpha(\|\beta^*\|_1 + \lambda_2\|\beta^*\|_2^2) + (1-\alpha)\|\beta^*-\tilde{\beta}\|_1)$$

Therefore,

$$\frac{1}{2n}\|X(\hat{\beta}-\beta^*)\|_2^2 + \lambda_1(\alpha(\|\hat{\beta}\|_1 + \lambda_2\|\hat{\beta}\|_2^2) + (1-\alpha)\|\hat{\beta}-\tilde{\beta}\|_1)$$
$$\leq \frac{1}{n}\boldsymbol{\varepsilon}' X(\hat{\beta}-\beta^*) + \lambda_1(\alpha(\|\beta^*\|_1 + \lambda_2\|\beta^*\|_2^2) + (1-\alpha)\|\beta^*-\tilde{\beta}\|_1)$$

Note that the term $\frac{1}{n}\boldsymbol{\varepsilon}' X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)$ is less than or equal to $\|\frac{1}{n}X'\boldsymbol{\varepsilon}\|_\infty\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1$. That is, $\frac{1}{n}\boldsymbol{\varepsilon}' X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) \leq |\frac{1}{n}\boldsymbol{\varepsilon}' X\|_\infty \iota_k'(\hat{\beta}-\beta^*) \leq |\frac{1}{n}\boldsymbol{\varepsilon}' X\|_\infty\|\hat{\beta}-\beta^*\|_1$, where ι_k is a $k\times 1$ vector.

$$\frac{1}{2n}\|X(\hat{\beta}-\beta^*)\|_2^2 + \lambda_1(\alpha(\|\hat{\beta}\|_1 + \lambda_2\|\hat{\beta}\|_2^2) + (1-\alpha)\|\hat{\beta}-\tilde{\beta}\|_1)$$
$$\leq \|\frac{1}{n}X'\boldsymbol{\varepsilon}\|_\infty\|\hat{\beta}-\beta^*\|_1 + \lambda_1(\alpha(\|\beta^*\|_1 + \lambda_2\|\beta^*\|_2^2) + (1-\alpha)\|\beta^*-\tilde{\beta}\|_1)$$

Since $\boldsymbol{\varepsilon}$ is assumed to be sub-gaussian in assumption 1.1, we have by the properties of sub-gaussian random variables,

$$P(\|\frac{1}{n}X'\boldsymbol{\varepsilon})\|_\infty \leq \gamma_n) \geq 1 - \exp\{-\frac{n\gamma_n^2}{2\sigma^2} + \log(2p)\}$$

Where $\gamma_n = c\lambda_1$. So with probability approaching 1 for large n.

$$\frac{1}{2n}\|X(\hat{\beta}-\beta^*)\|_2^2 + \lambda_1(\alpha(\|\hat{\beta}\|_1 + \lambda_2\|\hat{\beta}\|_2^2) + (1-\alpha)\|\hat{\beta}-\tilde{\beta}\|_1)$$
$$\leq c\lambda_1\|\hat{\beta}-\beta^*\|_1 + \lambda_1(\alpha(\|\beta^*\|_1 + \lambda_2\|\beta^*\|_2^2) + (1-\alpha)\|\beta^*-\tilde{\beta}\|_1)$$

For simplicity, we refer to the left hand side of the equation as A. Further note that $\|\boldsymbol{\beta}^*\|_1 \leq \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 + \|\hat{\boldsymbol{\beta}}\|_1$. For notational purposes, note that β^S denotes the elements

of β corresponding to the non-zero elements of β^*, and β^{S^c} denotes the elements of β corresponding tot he zero elements of β^*.

$$\begin{aligned} A &\leq c\lambda_1\|\hat{\beta} - \beta^*\|_1 + \lambda_1\alpha\|\beta^*\|_1 + \lambda_1\lambda_2\alpha\|\beta^*\|_2^2 + (1-\alpha)\lambda_1\|\Delta\|_1 \\ &\leq c\lambda_1\|\hat{\beta}^s - \beta^{*,s}\|_1 + c\lambda_1\|\hat{\beta}^{S^c}\|_1 + \lambda_1\alpha\|\beta^{*,s}\|_1 + \lambda_1\lambda_2\alpha\|\beta^*\|_2^2 + (1-\alpha)\lambda_1\|\Delta\|_1 \end{aligned}$$

Note that by the triangle inequality, $\|b\| \leq \|a-b\| + \|a\|$.

$$\begin{aligned} &\leq c\lambda_1\|\hat{\beta}^s - \beta^{*,s}\|_1 + c\lambda_1\|\hat{\beta}^{S^c}\|_1 + \lambda_1\alpha\|\hat{\beta}^s - \beta^{*,s}\|_1 + \lambda_1\alpha\|\hat{\beta}^s\|_1 \\ &+ \lambda_1\lambda_2\alpha\|\beta^*\|_2^2 + (1-\alpha)\lambda_1\|\Delta\|_1 \end{aligned}$$

Combining like terms,

$$\begin{aligned} &\frac{1}{2n}\|X(\hat{\beta} - \beta^*)\|_2^2 + \lambda_1(\alpha - c)\|\hat{\beta}^{S^c}\|_1 + \lambda_1(1-\alpha)\|\hat{\beta} - \tilde{\beta}\|_1 \\ &\leq \lambda_1(\alpha + c)\|\hat{\beta}^s - \beta^{*,s}\|_1 + \lambda_1\lambda_2\alpha\|\beta^*\|_2^2 - \lambda_1\lambda_2\alpha\|\hat{\beta}\|_2^2 + \lambda_1(1-\alpha)\|\Delta\|_1 \\ &\frac{1}{2n}\|X(\hat{\beta} - \beta^*)\|_2^2 + \lambda_1(\alpha - c)\|\hat{\beta}^{S^c}\|_1 + \lambda_1(1-\alpha)\|\hat{\beta} - \tilde{\beta}\|_1 \\ &\leq \lambda_1(\alpha + c)\|\hat{\beta}^s - \beta^{*,s}\|_1 + \lambda_1\lambda_2\alpha\|\hat{\beta} - \beta^*\|_2^2 + 2\lambda_1\lambda_2\alpha\|\hat{\beta}\|_2\|\hat{\beta} - \beta^*\|_2 + \lambda_1(1-\alpha)\|\Delta\|_1 \end{aligned}$$

By assumption 1.3, for large n, $\|\hat{\beta}\|_2 \leq \sqrt{s}M$ with probability approaching 1. This is because we know that the subset of coefficients in S^c are 0, and thus the upper bound on these coefficients is 0. The upper bound on all other coefficient is M, so with high probabiltiy $\|\hat{\beta}\|_2 = \sqrt{\hat{\beta}_1^2 + ... + \hat{\beta}_p^{\,2}} \leq \sqrt{\sum_S M^2 + \sum_{S^c} 0} = \sqrt{sM^2} = \sqrt{s}M$. Since the first term on the left hand side is positive, the second term on the right hand side is positive, and λ_1 scales all other terms equally, this tells us that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \{v \colon (\alpha - c)\|v^{S^c}\|_1 + (1-\alpha)\|v - \Delta\|_1 \leq (\alpha + c)\|v^S\|_1 + \lambda_2\alpha\|v\|_2^2 + (1-\alpha)\|\Delta\|_1\}$. Furthermore, by assumption 1.2

$$\frac{1}{2n}\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 \geq \frac{\phi}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$$

$$\frac{\phi}{2}\|\hat{\beta}-\beta^*\|_2^2 \leq \lambda_1(\alpha+c)\|\hat{\beta}^s-\beta^{*,s}\|_1 + \lambda_1\lambda_2\alpha\|\hat{\beta}-\beta^*\|_2^2 + 2\lambda_1\lambda_2\alpha\sqrt{s}M\|\hat{\beta}-\beta^*\|_2 + \lambda_1(1-\alpha)\|\Delta\|_1$$

$$\begin{aligned}\big(\frac{\phi}{2}-\lambda_1\lambda_2\alpha\big)\|\hat{\beta}-\beta^*\|_2^2 &\leq \lambda_1(\alpha+c)\|\hat{\beta}-\beta^{*,s}\|_1 + 2\lambda_1\lambda_2\alpha\sqrt{s}M\|\hat{\beta}-\beta^*\|_2 + \lambda_1(1-\alpha)\|\Delta\|_1 \\ &\leq [\lambda_1(\alpha+c)\sqrt{s}]\|\hat{\beta}-\beta^*\|_2 + 2\lambda_1\lambda_2\alpha\sqrt{s}M\|\hat{\beta}-\beta^*\|_2 + \lambda_1(1-\alpha)\|\Delta\|_1 \\ &\leq [\lambda_1(\alpha+c)\sqrt{s}+2\lambda_1\lambda_2\alpha\sqrt{s}M]\|\hat{\beta}-\beta^*\|_2 + \lambda_1(1-\alpha)\|\Delta\|_1\end{aligned}$$

$$\begin{aligned}&\big(\frac{\phi}{2}-\lambda_1\lambda_2\alpha\big)\|\hat{\beta}-\beta^*\|_2^2 - [\lambda_1(\alpha+c)\sqrt{s}+2\lambda_1\lambda_2\alpha\sqrt{s}M]\|\hat{\beta}-\beta^*\|_2 \leq \lambda_1(1-\alpha)\|\Delta\|_1 \\ &\|\hat{\beta}-\beta^*\|_2^2 - \frac{2[\lambda_1(\alpha+c)\sqrt{s}+2\lambda_1\lambda_2\alpha\sqrt{s}M]}{\phi-2\lambda_1\lambda_2\alpha}\|\hat{\beta}-\beta^*\|_2 \leq \frac{2\lambda_1(1-\alpha)}{\phi-2\lambda_1\lambda_2\alpha}|\Delta\|_1\end{aligned}$$

At this point we can complete the square on the left hand side by adding half of the second term to both sides.

$$\begin{aligned}&\|\hat{\beta}-\beta^*\|_2^2 - \frac{2[\lambda_1(\alpha+c)\sqrt{s}+2\lambda_1\lambda_2\alpha\sqrt{s}M]}{\phi-2\lambda_1\lambda_2\alpha}\|\hat{\beta}-\beta^*\|_2 + \frac{[\lambda_1(\alpha+c)\sqrt{s}+2\lambda_1\lambda_2\alpha\sqrt{s}M]^2}{(\phi-2\lambda_1\lambda_2\alpha)^2} \\ &\leq \frac{2\lambda_1(1-\alpha)}{\phi-2\lambda_1\lambda_2\alpha}|\Delta\|_1 + \frac{[\lambda_1(\alpha+c)\sqrt{s}+2\lambda_1\lambda_2\alpha\sqrt{s}M]^2}{(\phi-2\lambda_1\lambda_2\alpha)^2}\end{aligned}$$

The left side now can be factored into a square, and we can solve for $\hat{\beta}-\beta^*\|_2$.

$$\begin{aligned}&\left(\|\hat{\beta}-\beta^*\|_2 - \frac{[\lambda_1(\alpha+c)\sqrt{s}+2\lambda_1\lambda_2\alpha\sqrt{s}M]}{\phi-2\lambda_1\lambda_2\alpha}\right)^2 \\ &\leq \frac{2\lambda_1(1-\alpha)}{\phi-2\lambda_1\lambda_2\alpha}|\Delta\|_1 + \frac{[\lambda_1(\alpha+c)\sqrt{s}+2\lambda_1\lambda_2\alpha\sqrt{s}M]}{(\phi-2\lambda_1\lambda_2\alpha)^2}\end{aligned}$$

$$\|\hat{\beta} - \beta^*\|_2 - \frac{[\lambda_1(\alpha + c)\sqrt{s} + 2\lambda_1\lambda_2\alpha\sqrt{s}M]}{(\phi - 2\lambda_1\lambda_2\alpha)}$$
$$\leq \sqrt{\frac{2\lambda_1(1-\alpha)}{\phi - 2\lambda_1\lambda_2\alpha}|\Delta\|_1 + \frac{[\lambda_1(\alpha + c)\sqrt{s} + 2\lambda_1\lambda_2\alpha\sqrt{s}M]^2}{(\phi - 2\lambda_1\lambda_2\alpha)^2}}$$

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{[\lambda_1(\alpha + c)\sqrt{s} + 2\lambda_1\lambda_2\alpha\sqrt{s}M]}{(\phi - 2\lambda_1\lambda_2\alpha)}$$
$$+ \sqrt{\frac{2\lambda_1(1-\alpha)}{\phi - 2\lambda_1\lambda_2\alpha}|\Delta\|_1 + \frac{[\lambda_1(\alpha + c)\sqrt{s} + 2\lambda_1\lambda_2\alpha\sqrt{s}M]^2}{(\phi - 2\lambda_1\lambda_2\alpha)^2}}$$

Appendix B

Chapter 2

Summary Statistics for Trajectories Subsamples

	Time after re-employment:					
	4 weeks		52 weeks		104 weeks	
	Mean	s.d.	Mean	s.d.	Mean	s.d.
Age	32.0	12.4	32.6	12.6	32.8	12.6
Women	0.558	0.497	0.557	0.497	0.559	0.497
At Most High School	0.223	0.416	0.227	0.419	0.230	0.421
Teenager	0.176	0.381	0.177	0.381	0.167	0.373
Race:						
White	0.771	0.420	0.776	0.417	0.767	0.423
Black	0.157	0.364	0.154	0.361	0.160	0.367
Asian	0.028	0.164	0.028	0.165	0.029	0.167
Other	0.044	0.206	0.042	0.201	0.044	0.205
Unemployment Rate	0.062	0.022	0.062	0.021	0.058	0.018
Fired From Last Job	0.037	0.189	0.036	0.187	0.030	0.171
Unemp. Duration	15.1	12.1	15.1	12.0	13.9	11.1
Δmw	0.008	0.030	0.007	0.026	0.008	0.028
Spells	40,031		24,358		12,630	
Last wage	$ 11.65	$ 9.62	$ 12.86	$ 10.30	$ 13.78	$ 12.14
Observations	35,889		17,883		9,020	

Summary statistics for data from the 2001, 2004, 2008 and 2014 SIPP panels. Columns denoted "s.d." give the standard deviation of the variable, for each group. Summary statistics are shown for 3 different groups: spells whith re-employment data four weeks after re-employment, those with data 52 weeks after re-employment, and those with data 104 weeks after re-employment.

Appendix C

Chapter 4

C.1 Proof of Theorem 4.1

Begin with the definition of the TWCCE estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{M}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}\mathbf{Y} \tag{C.1}$$

We can now substitute for $\mathbf{Y}$ and expand

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{M}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}\big(\mathbf{X}\boldsymbol{\beta} + (\mathbf{F}\otimes I_N)\boldsymbol{\gamma} + \boldsymbol{\epsilon}\big) \tag{C.2}$$

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{M}\mathbf{X}}{NT}\right)^{-1}\frac{\mathbf{X}'\mathbf{M}(\mathbf{F}\otimes I_N)\boldsymbol{\gamma}}{NT} + \left(\frac{\mathbf{X}'\mathbf{M}\mathbf{X}}{NT}\right)^{-1}\frac{\mathbf{X}'\mathbf{M}\boldsymbol{\epsilon}}{NT} \tag{C.3}$$

Now taking the probability limits

$$p\lim\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + p\lim\left(\frac{\mathbf{X}'\mathbf{M}\mathbf{X}}{NT}\right)^{-1}\left[\frac{\mathbf{X}'\mathbf{M}(\mathbf{F}\otimes I_N)\boldsymbol{\gamma}}{NT}\right] \tag{C.4}$$

$$+ p\lim\left(\frac{\mathbf{X}'\mathbf{M}\mathbf{X}}{NT}\right)^{-1}\left[\frac{\mathbf{X}'\mathbf{M}\boldsymbol{\epsilon}}{NT}\right] \tag{C.5}$$

By Slutsky's theorem, since by assumption and the law of large numbers, $p\lim(X'MX)^{-1}/(NT)$ is a finite matrix (denote it C), we have

$$p\lim\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + C \cdot p\lim\left[\frac{\mathbf{X}'\mathbf{M}(\mathbf{F}\otimes I_N)\boldsymbol{\gamma}}{NT}\right] + C \cdot p\lim\left[\frac{\mathbf{X}'\mathbf{M}\boldsymbol{\epsilon}}{NT}\right] \tag{C.6}$$

Note that

$$\frac{\mathbf{X}'\mathbf{M}\boldsymbol{\epsilon}}{NT} = \mathbf{X}\boldsymbol{\epsilon} + \mathbf{X}'\bar{\mathbf{H}}(\bar{\mathbf{H}}'\bar{\mathbf{H}})^{-1}\bar{\mathbf{H}}'\boldsymbol{\epsilon} \tag{C.7}$$

By the Law of Large Numbers

$$\frac{\mathbf{X}'\mathbf{M}\boldsymbol{\epsilon}}{NT} \to_p E(\mathbf{x}_{it}\epsilon_{it}) + E(\mathbf{x}_{it}\bar{\mathbf{h}}_{it}(\bar{\mathbf{h}}'_{it}\bar{\mathbf{h}}_{it})^{-1}\bar{\mathbf{h}}'_{it}\epsilon_{it}) \tag{C.8}$$

Where $\bar{\mathbf{h}}_{it}$ is defined as the $(i \cdot t)$ row of $\bar{\mathbf{H}}$. By the assumption of exogeneity $E(\mathbf{x}'_{it}\epsilon_{it}) = \mathbf{x}'_{it}E(\epsilon_{it}) = \mathbf{x}'_{it} \cdot 0 = 0$. Assuming that $\mathbf{x}_{it}$ maintains this property when projected onto the column space of $\bar{\mathbf{H}}$ then the second term also equals 0. Hence the term in (34) containing the error reduces to 0 in the limit.

Now we must show that the column spaces of $(\mathbf{F}\otimes I_N))\boldsymbol{\gamma}$ and $\bar{\mathbf{H}}$ converge. That is, we wish to show that in the limit $\mathbf{M}((\mathbf{F}\otimes I_N)\boldsymbol{\gamma}) \to_p (I_{NT} - P_{F\gamma})((\mathbf{F}\otimes I_N)\boldsymbol{\gamma}) = \mathbf{M}_{F\gamma}((\mathbf{F}\otimes I_N)\boldsymbol{\gamma}) = 0$, where $P_{F\gamma}$ is the projection matrix associated with the interactive effects.

$$P_{F\gamma} = ((\mathbf{F}\otimes I_N)\boldsymbol{\gamma})\left[((\mathbf{F}\otimes I_N)\boldsymbol{\gamma})'((\mathbf{F}\otimes I_N)\boldsymbol{\gamma})\right]^{-1}((\mathbf{F}\otimes I_N)\boldsymbol{\gamma})' \tag{C.9}$$

This will result in the second term converging to zero. By 4.9 we have that $A\bar{\mathbf{H}} \to (\mathbf{F}\otimes I_N)\boldsymbol{\gamma}$, where $A\bar{\mathbf{H}}$ is a linear combination of the columns of $\bar{\mathbf{H}}$. It follows that in the limit $(\mathbf{F}\otimes I_N)\boldsymbol{\gamma}) \in C(\bar{\mathbf{H}})$. This implies that in the limit, $\mathbf{M}$ will "absorb" the interactive effects and reduce that term to 0. This is because the column space of the cross sectional means

and the column space of the interactive effects converge in the limit. Therefore

$$p\lim\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + C \cdot 0 + C \cdot 0 \tag{C.10}$$

$$p\lim\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} \tag{C.11}$$

■

C.2 Proof of Theorem 4.2

$$\sqrt{NT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{NT}\left[\left(\frac{\mathbf{X}'\mathbf{M}\mathbf{X}}{NT}\right)^{-1}\frac{\mathbf{X}'\mathbf{M}(\mathbf{F}\otimes I_N)\boldsymbol{\gamma}}{NT} + \left(\frac{\mathbf{X}'\mathbf{M}\mathbf{X}}{NT}\right)^{-1}\frac{\mathbf{X}'\mathbf{M}\boldsymbol{\epsilon}}{NT}\right] \tag{C.12}$$

$$= \left[\left(\frac{\mathbf{X}'\mathbf{M}\mathbf{X}}{NT}\right)^{-1}\frac{\mathbf{X}'\mathbf{M}(\mathbf{F}\otimes I_N)\boldsymbol{\gamma}}{\sqrt{NT}} + \sqrt{NT}\left(\frac{\mathbf{X}'\mathbf{M}\mathbf{X}}{NT}\right)^{-1}\frac{\mathbf{X}'\mathbf{M}\boldsymbol{\epsilon}}{NT}\right] \tag{C.13}$$

As shown in proof of theorem 4.1, $\mathbf{X}'\mathbf{M}(\mathbf{F}\otimes I_N)\boldsymbol{\gamma} \xrightarrow{p} 0$, and $\left(\frac{\mathbf{X}'\mathbf{M}\mathbf{X}}{NT}\right)^{-1} \xrightarrow{p} C$, therefore we have that the entire first term converges to 0. Thus we have,

$$\begin{aligned}
\sqrt{NT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \sqrt{NT}\left(\frac{\mathbf{X}'\mathbf{M}\mathbf{X}}{NT}\right)^{-1}\frac{\mathbf{X}'\mathbf{M}\boldsymbol{\epsilon}}{NT} \\
&= \left(\frac{\mathbf{X}'\mathbf{M}\mathbf{X}}{NT}\right)^{-1}\sqrt{NT}\frac{\mathbf{X}'\mathbf{M}\boldsymbol{\epsilon}}{NT} \\
&\xrightarrow{d} CN(0, \boldsymbol{\Omega}) \\
&\xrightarrow{d} N(0, C\boldsymbol{\Omega}C') \\
&\xrightarrow{d} N(0, \boldsymbol{\Sigma}_\beta)
\end{aligned}$$

■