Estimation of Conditional Average Treatment Effects

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Craig Anthony Rolling

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Yuhong Yang, Advisor

July, 2014

Acknowledgements

I would first like to thank my wife Margie, whose love and patience made this thesis possible. This thesis is dedicated to her and to our children, Luke and Leo.

Joan Rolling, Gary Rolling, and Ginger Rolling always have supported me and encouraged me to aim high. This thesis is a result of their encouragement. Tom and Joan DeWard also have generously supported us throughout my time in graduate school.

Thanks to Birgit Grund, Sandy Weisberg, and Mike Steinbach for serving on my committee. The helpful comments they made during my oral prelim helped steer this project in the right direction. I would like to especially acknowledge Birgit Grund for providing the FIRST clinical trial data discussed in Chapter 6.

I have learned a tremendous amount from the faculty of the School of Statistics. In particular, I want to acknowledge Lan Wang, for teaching statistical theory so clearly; and Dennis Cook, for his course on regression graphics that led to the ideas discussed in Section 4.3.4. The University of Minnesota Supercomputing Institute provided resources for some of the simulations in Section 4.4.

Thanks to professors Stephen Stigler and Ronald Thisted at Chicago for inspiring my love of statistics and for writing me letters of recommendation long after I had graduated. My former coworkers Bob Meyerhoff and Will McCleskey also wrote me letters of recommendation for graduate school and are great people.

Finally, I would like to express deep gratitude to my advisor, Yuhong Yang, for his wisdom, patience, and good humor as we explored this new topic (the “Wild West”) together. Thank you.

Dedication

To Margie, Luke, and Leo

Abstract

Researchers often believe that a treatment's effect on a response may be heterogeneous with respect to certain baseline covariates. This is an important premise of personalized medicine and direct marketing. Within a given set of regression models or machine learning algorithms, those that best estimate the regression function may not be best for estimating the effect of a treatment; therefore, there is a need for methods of model selection targeted to treatment effect estimation. In this thesis, we demonstrate an application of the focused information criterion (FIC) for model selection in this setting and develop a treatment effect cross-validation (TECV) aimed at minimizing treatment effect estimation errors. Theoretically, TECV possesses a model selection consistency property when the data splitting ratio is properly chosen. Practically, TECV has the flexibility to compare different types of models and estimation procedures.

In the usual regression settings, it is well established that model averaging (or more generally, model combining) frequently produces substantial performance gains over selecting a single model, and the same is true for the goal of treatment effect estimation. We develop a model combination method (TEEM) that properly weights each model based on its (estimated) accuracy for estimating treatment effects. When the baseline covariate is one-dimensional, the TEEM algorithm automatically produces a treatment effect estimate that converges at almost the same rate as the best model in a candidate set.

We illustrate the methods of FIC, TECV, and TEEM with simulation studies, data from a clinical trial comparing treatments of patients with HIV, and a benchmark public policy dataset from a work skills training program. The examples show that the methods developed in this thesis often exhibit good performance for the important goal of estimating treatment effects conditional on covariates.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Statisticians have been concerned with problems of causal inference for many years. In the early 20th century, seminal works of Neyman (1935) and Fisher (1935) discussed the use of randomized experiments to attribute differences in outcomes to the causal effect of a treatment. In the century's second half, Rubin (1974) and others developed formal insights into inferring about causal effects from randomized and nonrandomized studies. These 20th-century works on causal inference were primarily concerned with the average effect of a treatment across a population of interest. However, a population's average treatment effect (ATE) gives no insight into whether (or how) individuals within the population may be diversely affected by a treatment.

In many applications for which a treatment's effect on a response is of interest, it is believed that the treatment effect may be heterogeneous within the population. To use a common example from mental health, a number of different medications and types of psychotherapy are available to treat people with clinical depression, and different patients may respond differently to these different treatments. Often, treatment effect heterogeneity can be at least partially identified using one or more baseline covariates that are measured before the application of the treatment. Returning to the example, a patient's initial depression level, age, family history, intelligence, and various indicators of physical health may help inform whether she will respond well to a particular treatment for depression.

The current century's Information Age has seen an explosion in the amount of data available to researchers. In this era of Big Data, there appears to be increased potential for identifying and estimating heterogeneous treatment effects by conditioning on baseline covariates. In the medical field, the mapping of the human genome completed in the century's first decade has given rise to "personalized medicine", the idea that medical decisions and treatments are to be individually tailored, rather than broadly applied. In business, personalized marketing "treatments" also are becoming more common. These include the printing of different images and language on promotional mail items depending on the addressee's demographics and mining an individual's internet browsing history to select the advertisement to play before an online video.

Despite the apparent usefulness of estimating a treatment effect conditional on covariates, this issue has not been comprehensively studied in the statistics literature. Including interaction terms in linear models is perhaps the most common way of estimating conditional (mean) treatment effects, but this approach has serious limitations. Whenever the underlying linear model is misspecified, any inference about conditional treatment effects based on interaction terms may be compromised. Even if the linear model is correct, limiting the number of interaction terms in a model often is desirable or necessary when the number of baseline covariates is large. In this case, practitioners have had to rely on *ad hoc* methods, or on model selection statistics with questionable relevance to treatment effect estimation, to decide which interactions to include.

In the last few years, authors in several scientific disciplines have started to develop alternative methods to estimate and infer about conditional treatment effects. These papers typically propose alternative methods of conditional treatment effect estimation (and sometimes inference) and apply them to a data example within their discipline. These recent works include Cai et al. (2011) in biostatistics, Crump et al. (2008) in economics, Imai and Ratkovic (2013) in political science, and Radcliffe and Surry (2011) in marketing.

The fact that this topic is being studied in several application areas indicates its current importance. However, some of these recent works fail to acknowledge the work in other disciplines, lack theoretical justification for their proposed methods, or fail to compare their method to others. The lack of a common framework and language for this problem may contribute to the disconnect; phrases used for conditional treatment effect estimation include heterogeneous treatment effect estimation, subgroup analysis, incremental response modeling, uplift modeling, and true lift modeling.

These somewhat disjointed research threads indicate the need for a clear framing of the problem in a statistical setting. For a researcher trying to estimate conditional treatment effects using a particular dataset, these new methods also beg the question of which method is best for the data at hand, or if more accurate estimates could be obtained by combining the estimates from the different procedures. This thesis aims to develop a general framework for conditional treatment effect estimation and to shed some light on the issues of model selection and combination in this setting.

1.1 Dissertation Objectives and Structure

We have three main objectives for this dissertation:

- Clearly present the issue of conditional treatment effect estimation in a general statistical framework.
- Develop a method of model selection targeted toward estimation of the conditional treatment effect.
- Develop a similarly targeted method of model combination.

There are a number of interesting topics for further research related to conditional treatment effect estimation, and we will discuss some of these in the dissertation's final chapter.

The remainder of the dissertation is organized as follows. The final section in Chapter 1 formally defines the conditional treatment differences and effects that are the focal objects of this thesis. Chapter 2 summarizes the previous literature on conditional treatment effect estimation and motivates our work on model selection in this setting. An application of the focused information criterion (FIC) (Claeskens and Hjort, 2003) for the purpose of treatment effect estimation is presented in Chapter 3. Chapter 4 introduces the TECV method of model selection for the conditional treatment effect, and the TEEM method of model combination is introduced in chapter 5. Chapter 6 applies the new methods on two real data examples. Chapter 7 discusses further issues of model selection in the setting of treatment effect estimation and presents some ideas for future research directions. Detailed proofs of our theoretical results are presented in Appendix A and Appendix B.

1.2 Δ and the Conditional Average Treatment Effect

The definitions and notations in this section are partially adopted from Rosenbaum and Rubin (1983), Holland (1986), Imbens and Wooldridge (2009), and Cai et al. (2011).

We consider a general regression framework in which the distribution of the response Y may depend on the treatment assignment T and one or more baseline covariates $\mathbf{U}$. In order to isolate the treatment difference, which is of primary interest, we represent the observed data in the following way:

$$Y_i = \{f_t(\mathbf{U}_i) + \xi_i\} I(T_i = t) + \{f_c(\mathbf{U}_i) + \nu_i\} I(T_i = c), \qquad 1 \leq i \leq n. \tag{1.1}$$

The data consist of $(Y_i, T_i, \mathbf{U}_i)_{i=1}^n$, where Y_i is the response, $T_i \in \{t, c\}$ is a binary treatment assignment (this work considers only binary treatments), and $\mathbf{U}_i$ represents a collection of p baseline covariates observed before the treatment is applied. We assume the covariates $\mathbf{U}_i$ to be i.i.d. from an unknown probability density $P_{\mathbf{U}}$ with support $\mathcal{U} \subset \mathbb{R}^p$. (This assumption will hold if the n observed units represent a simple random sample from the population.) The random errors under treatment are denoted by ξ_i, while ν_i are the errors under control. Each collection of random errors is assumed to be i.i.d. with zero mean, but the treatment and control error distributions are allowed to differ. The primary object of interest in our work is

$$\Delta(\mathbf{u}) := f_t(\mathbf{u}) - f_c(\mathbf{u}), \tag{1.2}$$

the difference between the regression functions for the treatment and control groups.

This thesis defines causal effects using the framework of potential outcomes known as the Rubin Causal Model (Holland, 1986). In this framework, the causal effect of the treatment T on the outcome Y for a given unit i is understood as the difference between $Y_{i,(t)}$, the Y_i that would have been observed had $T_i = t$, and $Y_{i,(c)}$, the Y_i that would have been observed had $T_i = c$. Of course, the random variable $Y_{i,(t)} - Y_{i,(c)}$ representing the causal effect is not observed for any i; Holland calls this the “fundamental problem of causal inference".

Holland calls the averaging of treatment effects over groups the statistical solution to this fundamental problem. For example, we may consider the average of $Y_{i,(t)} - Y_{i,(c)}$ over all the units i in a population of interest. Inference can be done on this average treatment effect (sometimes called the ATE) using standard methods (e.g., a two-sample t-test) under some conditions.

A key concept for our work is that information about a treatment's average effect for subgroups *within* a population can be obtained by observing a set of baseline covariates before the treatment application. If the treatment effect is heterogeneous with respect to these covariates, this heterogeneity can be captured described by the conditional distribution of the random variable $Y_{(t)} - Y_{(c)}$ on the covariate vector $\mathbf{U}$. We will focus on the *expectation* of this conditional distribution. This conditional expectation is sometimes called the conditional average treatment effect, or CATE:

$$\text{CATE}(\mathbf{u}) := E[\{Y_{i,(t)} - Y_{i,(c)}\}|\mathbf{U}_i = \mathbf{u}] = E\{Y_{i,(t)}|\mathbf{U}_i = \mathbf{u}\} - E\{Y_{i,(c)}|\mathbf{U}_i = \mathbf{u}\}. \quad (1.3)$$

Next we show conditions under which $\Delta(\mathbf{u}) = \text{CATE}(\mathbf{u})$. Note that the two are not equal in general. Expression (1.2) simply represents the difference between two regression functions (we sometimes call Δ the conditional average treatment *difference*), while (1.3) refers specifically to a causal effect.

1.2.1 Interpreting Δ as a Causal Effect

In this section we present two conditions under which the conditional average treatment effect is identifiable and equal to the conditional average treatment difference, Δ.

Unconfoundedness: This condition requires that all potential confounding information for the relationship between the treatment and the potential outcomes is observed in the covariates. Mathematically, we express this as

$$\{Y_{(t)}, Y_{(c)}\} \perp\!\!\!\perp T|\mathbf{U}, \quad (1.4)$$

where $\perp\!\!\!\perp$ denotes independence. The unconfounded assignment assumption always holds in randomized experiments. In observational studies, it is typically unknown whether this condition holds; increasing the number of observed baseline covariates may increase the chance that all confounding information has been captured in $\mathbf{U}$.

Overlap: The overlap condition is necessary for the identifiability of the CATE on the support $\mathcal{U}$. It requires any observation, regardless of its covariate values, to have a chance to be assigned to either the treatment or control group.

$$P(T_i = t | \mathbf{U}_i = \mathbf{u}) \in (0, 1), \qquad \text{for all } \mathbf{u} \in \mathcal{U}. \tag{1.5}$$

The unconfoundedness and overlap conditions together are called the assumption of strong ignorability in Rosenbaum and Rubin (1983).

The following argument (from Imbens and Wooldridge, 2009, p. 26-27) shows that under these two conditions, Δ and the CATE defined in (1.2) and (1.3), respectively, are equal and identifiable.

$$\begin{aligned} \text{CATE}(\mathbf{u}) &= E\{Y_{i,(t)} | \mathbf{U}_i = \mathbf{u}\} - E\{Y_{i,(c)} | \mathbf{U}_i = \mathbf{u}\} \\ &= E\{Y_i | \mathbf{U}_i = \mathbf{u}, T_i = t\} - E\{Y_i | \mathbf{U}_i = \mathbf{u}, T_i = c\} \qquad \text{(by (1.4))} \\ &= f_t(\mathbf{u}) - f_c(\mathbf{u}) \\ &= \Delta(\mathbf{u}). \end{aligned}$$

By (1.5), f_t and f_c (and thus Δ) are identifiable for every $\mathbf{u} \in \mathcal{U}$. The identifiability of Δ uses (1.5) only; the less-realistic assumption of unconfoundedness is not required. For this reason, the rest of the thesis will target estimation of Δ. Estimation of Δ is more broadly attainable and may be of interest even when Δ is not the CATE. For simplicity, we sometimes refer to Δ as the conditional treatment effect (or simply the treatment effect) during the remainder of this thesis. We ask the reader to keep in mind that "treatment effect" is used as a shorthand for the treatment's effect on the conditional mean and that condition (1.4) is needed to formally bestow a causal interpretation on Δ.

Chapter 2

Literature Review and Motivation

In this chapter, we attempt to summarize the most relevant literature on conditional treatment effect estimation from statistics and related disciplines. Through this summary, we motivate our thesis work on model selection and combination in this context.

2.1 Causal Inference in Statistics

Inferring about the causal effect of a treatment on a response often is a primary goal of a statistical analysis. Holland (1986) provides an overview of the relationship between causation and statistics. His article includes a historical review of the ways in which the early works of Neyman, Fisher, and D.R. Cox addressed issues of causation. He also discusses various philosophers' ideas about causation and how these relate to statistical models of causality.

Holland formulates the potential outcomes model for causal inference, which he attributes to Rubin and which we introduced in Section 1.2. We use slightly different notation than Holland; he denotes potential outcomes under treatment and control, respectively, by $Y_t(u)$ and $Y_c(u)$, for a unit u in a population U. Holland does not directly define the *conditional* average treatment effect, but on p. 949 he points out the limitations of the average treatment effect (which we call the ATE and he calls T):

> The average causal effect T is an average and as such enjoys all of the advantages and disadvantages of averages. For example, if the variability in the causal effects $Y_t(u) - Y_c(u)$ is large over U, then T may not represent the causal effect of a specific unit, u_o, very well. If u_o is the unit of interest, then T may be irrelevant, no matter how carefully we estimate it!

Holland defines the assumption of constant effect, or additivity, in a population U as

$$T = Y_t(u) - Y_c(u), \quad \text{for all } u \text{ in } U. \tag{2.1}$$

If (2.1) holds, the average treatment effect T is relevant to every unit. However, if (2.1) does not hold, then the treatment effect is heterogeneous in U and T may not be relevant to some (or perhaps any) units.

2.2 Testing Existence of Heterogeneity

Before estimating Δ or the CATE for a particular population, it is sensible to ask whether the data provide any evidence for treatment effect heterogeneity in the population. In other words, we may want to measure the evidence against the additivity assumption in (2.1). Under the normal linear model and randomized treatment assignments, a hypothesis test for treatment effect heterogeneity can be done by simultaneously testing all treatment-covariate interaction terms in the regression model via an F test. However, the validity of such a test depends on the model assumptions being correct.

Crump et al. (2008) propose a nonparametric test for treatment effect heterogeneity that does not require the specification of a functional form for the treatment effect. More specifically, their test is for heterogeneity of the Δ function:

$$H_0 : \exists\ \Delta \text{ s. t. } \forall\ \mathbf{u} \in \mathcal{U}, \Delta(\mathbf{u}) = \Delta$$
$$H_A : \forall\ \Delta, \exists\ \mathbf{u} \in \mathcal{U} \text{ s. t. } \Delta(\mathbf{u}) \neq \Delta$$

The authors construct a test statistic W that is essentially the distance between two estimated nonparametric regression functions $\hat{\xi}_t$ and $\hat{\xi}_c$ built on the treatment and control groups separately and estimated using series estimators. This distance is scaled by the inverse of the estimated covariance matrix of the estimators, and a correction is made for the number of terms in the model. Under some regularity conditions, the authors prove that under $H_0, W \xrightarrow{d} N(0,1)$ and that the test is consistent for a sequence of local alternatives.

2.3 Interactions

Under the normal linear model

$$Y_i = \alpha + \beta^T \mathbf{U}_i + (\tau + \eta^T \mathbf{U}_i) I(T_i = t) + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \tag{2.2}$$

$\Delta(\mathbf{u}) = \tau + \eta^T \mathbf{u}$ and ordinary least squares regression can be used to do estimation and inference about Δ. However, if one or more of the covariates in $\mathbf{U}$ are continuous, it often seems unreasonable to believe that treatment-covariate interaction terms are linear as described by (2.2). Consider a marketing treatment T representing a direct mail promotion from a bank for a home equity line of credit (HELOC), and a single covariate U representing the age of the head of household. One would suspect that the HELOC promotion would be more attractive for middle-aged households rather than for young adults (who are more likely to rent or have little equity in their homes) or for those in retirement (who may have reduced income and be less likely to qualify for the loan). If this is true, the nonlinearity in $\Delta(u)$ cannot be captured by the model (2.2).

Feller and Holmes (2009) propose extending interaction terms to additive models in order to estimate potentially nonlinear Δ functions. In an additive model, each term is a smooth, possibly nonlinear function of a single covariate. Allowing these functions to be differently specified for the treatment and control groups enables the estimation of a nonlinear Δ. For example, representing a p-dimensional $\mathbf{U}_i$ as $(U_{i,1}, \ldots, U_{i,p})$, the model is written as

$$Y_i = \alpha + \left(\tau + \sum_{j=1}^{p} m_{t,j}(U_{i,j})\right) I(T_i = t) + \left(\sum_{j=1}^{p} m_{c,j}(U_{i,j})\right) I(T_i = c) + \varepsilon_i,$$

for smooth functions $m_{t,j}$ and $m_{c,j}$, $1 \leq j \leq p$, approximated by splines. For this model, we have

$$\Delta(\mathbf{u}) = \tau + \sum_{j=1}^{p} \Big\{ m_{t,j}(u_j) - m_{c,j}(u_j) \Big\}.$$

These types of additive models are used as candidate models in the numerical examples described later in this thesis.

2.4 Algorithmic Approaches

2.4.1 Trees

One of the earliest methods to target estimation of conditional average treatment effects is proposed by Hansotia and Rukstales (2002) in a direct marketing analytics journal. They use a type of regression tree to estimate what they call *incremental response* or *incremental value*. Their method constructs a tree via a sequence of binary splits on single covariates, where each split based on a covariate U_j creates two nodes that are as different as possible in terms of their observed "incremental" response, $(\overline{Y}_i|T_i = t) - (\overline{Y}_i|T_i = c)$. In other words, a cutoff point c_j for a covariate U_j is chosen so that the absolute value of

$$\begin{aligned} &\big\{(\overline{Y}_i|T_i = t, U_{i,j} \geq c_j) - (\overline{Y}_i|T_i = c, U_{i,j} \geq c_j)\big\} \\ &\quad - \big\{(\overline{Y}_i|T_i = t, U_{i,j} < c_j) - (\overline{Y}_i|T_i = c, U_{i,j} < c_j)\big\} \end{aligned} \tag{2.3}$$

is maximized, subject to a minimum number of observations used to compute each $\overline{Y}_i$. The idea is that a tree constructed from splits based on (2.3) will identify subgroups ("leaves") for which the treatment effect is very large and other subgroups for which the treatment is not effective or perhaps has a negative impact. Radcliffe and Surry (2011) propose a modification to this method that they call Significance-Based Uplift Trees. Essentially, they use the statistical significance of the difference in (2.3), instead of its magnitude, as a splitting criterion. Radcliffe and Surry (2011) also propose bootstrap-based methods to prune and average trees, improving their stability.

Political scientists Green and Kern (2010) use a different tree-based approach to estimate the CATE. They point out that ordinary regression trees, where the treatment variable is simply considered as an additional covariate in $\mathbf{U}$ taking values in $(0, 1)$ (call the $(p+1)$-dimensional covariate vector with the treatment variable $\mathbf{V}$) and splits are done to maximize

$$(\overline{Y}_i | V_{i,j} \geq c_j) - (\overline{Y}_i | V_{i,j} < c_j), \tag{2.4}$$

will indirectly estimate treatment-covariate interactions by the nature of successive splitting. For example, if the first split in an ordinary regression tree is on the treatment variable, the differences between the subsequent splits in the treatment branch and those in the control branch represent estimates of treatment effect heterogeneity. They use the Bayesian Additive Regression Trees (BART) method (Chipman et al., 2010) to build many regression trees based on the type of splits in (2.4), some of which will be based on the treatment variable. Averaging the predicted responses from these trees at any given $\mathbf{u}$, while assuming first treatment and then control status, will produce estimates of $f_t(\mathbf{u})$ and $f_c(\mathbf{u})$. Subtracting these estimates then gives an estimate of $\Delta(\mathbf{u})$.

2.4.2 Support Vector Machines

Imai and Ratkovic (2013) use support vector machine classifiers with LASSO constraints to estimate the CATE for a binary response. They propose separating the predictors of the response Y into groups $\mathbf{Z}$ and $\mathbf{V}$. The $\mathbf{Z}$ vector (with dimension L_Z) represents the interactions between the treatment indicator and all baseline covariates, while $\mathbf{V}$ (with dimension L_V) represents the baseline covariates only.

The "hinge-loss" function is defined as $|x|_+ = \max(x, 0)$. Then the SVM minimizes the objective function

$$\sum_{i=1}^{n} |1 - Y_i(\mathbf{Z}_i^T \beta + \mathbf{V}_i^T \gamma)|_+ + \lambda_z \sum_{j=1}^{L_z} |\beta_j| + \lambda_v \sum_{j=1}^{L_v} |\gamma_j|.$$

The idea is to put separate LASSO constraints on the groups of model coefficients β and γ that estimate heterogeneous treatment effects and baseline covariate effects, respectively. The paper describes an algorithm to estimate the regression coefficients and the tuning parameters.

2.4.3 A Two-Stage Approach

Cai et al. (2011) propose a two-stage approach to estimate the conditional average treatment effect. The first stage is to formulate generalized linear "working" models for the treatment and control groups separately:

$$E(Y_i|\mathbf{U}_i, T_i = k) = g_k(\beta_k^T \mathbf{U}_i), \quad \text{for } k = (t, c), \tag{2.5}$$

where the g_k are known link functions and the β_k are unknown vectors of regression coefficients. Then the temporary stage 1 estimator for the CATE is

$$\hat{s}(\mathbf{u}) = g_t(\hat{\beta}_t^T \mathbf{u}) - g_c(\hat{\beta}_c^T \mathbf{u})$$

Note that when the generalized linear models in (2.5) are correctly specified, $\hat{\beta}_t$ and $\hat{\beta}_c$ converge to β_t and β_c, respectively, and $\hat{s}(\mathbf{u})$ is a consistent estimator of $\Delta(\mathbf{u})$. The authors argue that even when the models in (2.5) are not correctly specified, $\hat{s}(\mathbf{u})$ may be used as an index to group subjects with potentially similar values of the CATE. They consider dividing the population into many strata based on the index such that patients in the same subset $\Omega_v = \{\mathbf{u} : \hat{s}(\mathbf{u}) = v\}$ have the same parametric score value v.

In stage 2, the authors write the mean functions for the treatment and control groups as conditional on the parametric score variable V:

$$\mu_k(v) = E(Y_i|V_i = v, T_i = k) \;\; \text{for } k = (t, c)$$

One-dimensional kernel regression methods are used to estimate $\mu_t(v)$ and $\mu_c(v)$, and the final estimator of the CATE is

$$\widehat{\Delta}(v) = \hat{\mu}_t(v) - \hat{\mu}_c(v).$$

Theoretical results include a proof that for each subgroup Ω_v, $\widehat{\Delta}(v)$ is a consistent estimator of the average treatment effect for that subgroup.

A Criticism

While the consistency of $\widehat{\Delta}(v)$ for each Ω_v is mathematically correct, it may not be practically relevant. Adding the conditional independence condition

$$Y \perp\!\!\!\perp \mathbf{U} | (T, V) \tag{2.6}$$

would guarantee that $\widehat{\Delta}(v)$ is consistent for $\Delta(\mathbf{u})$, but the authors do not discuss the sufficiency of the dimension reduction, and there is little reason to expect (2.6) to hold if the models in (2.5) are not correct.

The authors are using parametric regression methods to do dimension reduction, forcing the dimension down to one by classifying observations with a univariate V. Then they carry out non-parametric regression in the one-dimensional world, where it is easier. However, without ensuring that the dimension reduction is sufficient, forcing the dimension down to one may cause loss of information about the true object of interest $\Delta(\mathbf{u})$. Some or all of the treatment effect heterogeneity captured in the covariates $\mathbf{U}$ may be lost in the dimension reduction from $\mathbf{U}$ to V if the models in (2.5) are not accurate.

2.5 Model Evaluation

In the typical regression setting where prediction of a response or estimation of its conditional mean function is of interest, a common way to evaluate a model or algorithm is to summarize its prediction errors on out-of-sample data. That is, we summarize the differences between observed responses and predicted responses for individual observations that were not used to fit the model. However, when estimation of the conditional treatment effect is of interest, individual treatment effects cannot be directly observed; as a result, individual differences between observed and predicted treatment effects are not available for model evaluation. Therefore, new methods are needed in order to evaluate candidate models and algorithms in this setting. Two recent papers in the statistics literature address the problem of evaluating conditional treatment effect estimators.

Qian and Murphy (2011) consider the use of a high-dimensional linear regression model to construct an individualized treatment rule that maximizes the value, or mean response, resulting from the treatment rule in the population. They illustrate that within a set of models, the model that minimizes the prediction error may not be the model that yields the optimal treatment rule. They use a modified version of the LASSO (Tibshirani, 1996) that uses a cross-validated estimate of the value of the treatment rule suggested by the model, rather than a cross-validated estimate of the model's prediction error, to select the tuning parameter. The authors provide a finite sample upper bound of the difference between the value of the optimal treatment rule among the candidate models and the value of the rule chosen by their procedure.

Zhao et al. (2013) propose a more general cumulative average treatment difference curve for model evaluation and comparison. To evaluate a model's performance on a set of data, the range of quantiles q from 0 to 1 is plotted on the horizontal axis, and on the vertical axis, the average difference between the responses in the treatment and control groups for those data points with $H\left\{\widehat{\Delta}(\mathbf{U}_i)\right\} \geq q$ is drawn, where H is the empirical cumulative distribution function for the data. The more effectively the model ranks patients by their $\Delta(\mathbf{U}_i)$, the greater the expected value of the area under this curve. This method is not model-dependent and provides a way to analyze how well an estimator $\widehat{\Delta}$ holds up on out-of-sample data. One limitation of this method is that the area under the curve depends only on the ordering of the individuals in a particular sample, not on the accuracy of the treatment effect estimates.

2.5.1 A Motivating Example

For a subset of model selection problems, including some situations for which the true model is in the set of candidate models, the best model in the set for estimating the full regression function may also be the best model for estimating the conditional treatment effect. However, the two goals generally do not agree; indeed, they frequently conflict when all the models in a candidate set are misspecified.

Consider a situation in which two covariates $\mathbf{U} = (U_1, U_2)$ and a binary treatment variable T independent of $\mathbf{U}$ are available to predict a response Y. For simplicity of calculation, suppose U_1 and U_2 are independent standard normal and that $P(T_i = t) = 0.5$ for all i. The true model is

$$Y_i = \alpha_{\text{true}} + \tau_{\text{true}} I(T_i = t) + \beta_{\text{true}} U_{i,1} + \gamma_{\text{true}} I(T_i = t) U_{i,2} + \varepsilon_i,$$

with $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ and $\beta_{\text{true}}^2 > \gamma_{\text{true}}^2/2$. Suppose that due to the cost of observing U_1 and U_2, we must select a model which uses at most one of these. In particular, we consider candidate models

$$M_1 : Y_i = \alpha_1 + \tau_1 I(T_i = t) + \beta_1 U_{i,1} + \gamma_1 I(T_i = t) U_{i,1} + \varepsilon_i$$

and

$$M_2 : Y_i = \alpha_2 + \tau_2 I(T_i = t) + \beta_2 U_{i,2} + \gamma_2 I(T_i = t) U_{i,2} + \varepsilon_i,$$

and our object of interest is the mean treatment effect conditional on covariates. While this is an artificial example meant to illustrate the problem, it is not entirely unrealistic; data can be expensive, and often measurements come with specific costs or from different sources or providers.

The treatment effect is $\Delta(\mathbf{u}) = \tau_{\text{true}} + \gamma_{\text{true}} u_2$. Straightforward calculations show that the risk for estimating Δ under the L_2 loss, as $n \to \infty$, converges to γ_{true}^2 for M_1 while converging to 0 for M_2. Clearly, M_2 is preferred for accurate estimation of the treatment effect.

However, as $n \to \infty$, the average squared residuals of models M_1 and M_2 will converge to $\gamma_{\text{true}}^2/2$ and β_{true}^2, respectively. Since M_1 and M_2 estimate the same number of parameters, $\beta_{\text{true}}^2 > \gamma_{\text{true}}^2/2$ implies that AIC, BIC, and traditional cross-validation will select M_1 with probability tending to 1 as $n \to \infty$.

Therefore, in this situation the tools traditionally used for model selection will likely select the model that is worse for treatment effect estimation. P-values may be used to judge the statistical significance of the interaction term in this case; however, in general p-values are not geared toward estimation or prediction, and their interpretation is unclear whenever the model is misspecified.

2.6 Discussion

Researchers in several disciplines recently have developed methods to estimate the effect of a treatment conditional on covariates. Some of these methods are based on an explicit statistical model, while others are algorithmic in nature. In the presence of so many competing methods, the ability to evaluate and compare them on a particular dataset is important. Given a set of candidate models, the ones that best estimate the full regression function may not be the ones that best estimate Δ. This further motivates study of model selection and model combination in the context of treatment effect estimation.

Recent methods of model evaluation have been developed to judge an estimator $\widehat{\Delta}$ by its ability to effectively determine an individualized treatment rule or to effectively rank individuals in a sample with respect to their actual $\Delta(\mathbf{U}_i)$. In this thesis, we take a different approach to model evaluation. We develop methods for model selection and combination that evaluate a particular $\widehat{\Delta}$ by its (estimated) average closeness to the true Δ. The methods of TECV and TEEM developed in this thesis have the flexibility to compare different types of regression models and estimation algorithms, including all of those described in this chapter. Additionally, the TECV and TEEM methods evaluate the candidate models for Δ without assuming any of them are correct.

Chapter 3

FIC for Treatment Effects

3.1 The Focused Information Criterion

Claeskens and Hjort (2003) develop the focused information criterion (FIC) based on the idea that the determination of which model is best may depend on the purpose of the model. Indeed, as illustrated in Section 2.5.1, different models within a consideration set may possess minimal estimation risks for different quantities of interest. FIC aims to identify the model that minimizes the risk for the estimation of a particular "focus parameter", which is generally some function of the model parameters. This is done using a local misspecification asymptotic framework to estimate each model's estimation risk for the focus parameter of interest. See Claeskens and Hjort (2003) and Claeskens and Hjort (2008b) for more background on FIC.

Vansteelandt et al. (2012) derive the FIC for estimation of the marginal treatment effect on a binary response in an observational study. Specifically, their focus parameter is the marginal log odds ratio, and their concern is the selection of potentially confounding covariates that affect the estimate of the overall treatment effect. In this chapter, we illustrate the use of FIC to select a model for estimating *conditional* treatment effects.

Specifically, we will derive the formula for the FIC in the setting of linear regression with Gaussian errors where the focus parameter is $\Delta(\mathbf{u})$ at a particular $\mathbf{u}$. In this form, the FIC of each candidate model would depend on $\mathbf{u}$, the value of the covariate vector. In practice, this use of FIC would select different models for different subjects. Such localized model selection may sometimes be advantageous for estimating individual treatment effects; see also Yang (2008) for a discussion of applying cross-validation locally. However, when we discuss model selection in this work, our goal is to select a single model or procedure that accurately estimates $\Delta(\mathbf{u})$ for any $\mathbf{u}$. To use FIC for this type of global model selection, the weighted FIC (wFIC) described in Claeskens and Hjort (2008a) may be used.

3.2 Derivation of the FIC for Δ

The derivation of FIC is based on an underlying parametric model for the data. The FIC assumes that the true model is among the candidates in a consideration set that contains a smallest (narrow) model and a biggest (wide) model. The narrow model, the smallest model that might be used for the data, involves estimation of a parameter vector θ of length r. In the wide model, there are an additional q parameters $\gamma = (\gamma_1, \ldots, \gamma_q)$. The narrow model is a special case of the wide model, in that there is a value γ_0 such that with $\gamma = \gamma_0$ in the wide model, the narrow model is obtained. The models in the consideration set are submodels of the wide model that correspond to including some of the γ_j parameters while excluding others. These submodels are indexed by subsets S of $\{1, \ldots, q\}$; $S = \emptyset$, for example, identifies the narrow model with $\gamma = \gamma_0$. Up to 2^q candidate models may therefore be considered.

The FIC aims to select a model that provides the lowest risk under squared error loss for estimating some focus parameter $\mu = \mu(\theta, \gamma)$ of particular interest. This is done by estimating the risk for μ of each model under a local misspecification setting. Within this local misspecification framework, the authors derive the asymptotic normal distribution of $\sqrt{n}(\widehat{\mu}_S - \mu_{\text{true}})$, where $\widehat{\mu}_S$ is the maximum likelihood estimate of μ under submodel S. All quantities involved in the limiting risk of this distribution are estimated by (asymptotically unbiased) plug-in estimators, terms common to all submodels are removed, and the result is the FIC for submodel S.

Linear regression with Gaussian errors can be described within the above framework as

$$Y_i = \beta^T \mathbf{X}_i + \gamma^T \mathbf{Z}_i + \sigma \varepsilon_i,$$

where the error terms ε_i are i.i.d. $N(0, 1)$. Let $\mathbf{X}$ be the $n \times (r - 1)$ design matrix with ith row equal to $\mathbf{X}_i$; the columns of $\mathbf{X}$ represent variables that are included in all models in the consideration set. $\mathbf{Z}$ is the corresponding $n \times q$ matrix with rows $\mathbf{Z}_i$. The model selection problem in this setting is the decision of which variables (columns of $\mathbf{Z}$) to include in the model. In the language of FIC, $\theta = (\beta, \sigma)^T$, $\gamma = (\gamma_1, \ldots, \gamma_q)^T$, and γ_0 is the zero vector of length q. The general formula for FIC in the normal linear model is derived in Claeskens and Hjort (2008b). We will not give the entire formula here; we will only mention that in this setting, the formula for FIC depends on the focus parameter μ only through the quantity $\omega(\mathbf{u})$, defined as

$$\omega(\mathbf{u}) = \mathbf{Z}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\frac{\partial\mu}{\partial\beta} - \frac{\partial\mu}{\partial\gamma}. \tag{3.1}$$

We now derive the formula for $\omega(\mathbf{u})$ for the focus parameter Δ in a linear model with Gaussian errors when the coefficient associated with the treatment main effect is considered a "protected" parameter and there are p additional baseline covariates available. This would be an appropriate use of FIC if we believe the treatment has at least an additive effect on the regression function but we wish to determine which of the p (potentially confounding) covariate main effects or p treatment-covariate interactions to include. Similar formulas can be derived for situations in which it is desired to protect other terms or to leave the treatment variable unprotected.

The normal linear model with a treatment main effect and treatment-covariate interactions can be represented as

$$Y_i = \alpha + \tau I(T_i = t) + \eta^T\mathbf{U}_i + \xi^T\mathbf{U}_i I(T_i = t) + \sigma\varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0,1). \tag{3.2}$$

In the FIC framework, $\beta = (\alpha, \tau)^T$, the protected regression coefficients; $\theta = (\alpha, \tau, \sigma)^T$ represents all protected parameters, so $r = \dim(\theta) = 3$; and $\gamma = (\eta, \xi)^T$ constitutes the $q = 2p$ unprotected parameters.

Our focus parameter here is $\Delta(\mathbf{u})$, the effect of the treatment variable on the mean of Y. That is, $\mu(\mathbf{u}) = \Delta(\mathbf{u}) = E(Y_i|T_i = t, \mathbf{U}_i = \mathbf{u}) - E(Y_i|T_i = c, \mathbf{U}_i = \mathbf{u})$. Using (3.2), we see that $\mu(\mathbf{u}) = \tau + \xi^T\mathbf{u}$. Therefore, $\frac{\partial\mu}{\partial\beta} = (0,1)^T$ and $\frac{\partial\mu}{\partial\gamma} = (0, \ldots, 0, \mathbf{u})^T$, a vector with p zeros followed by the p-vector $\mathbf{u}$. Plugging these into (3.1), we obtain

$$\omega(\mathbf{u}) = \mathbf{Z}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(0,1)^T - (0, \ldots, 0, \mathbf{u})^T,$$

where $\mathbf{Z} = (\mathbf{U}, T^T\mathbf{U})$ and $\mathbf{X} = (1, T)$ represent the unprotected and protected portions, respectively, of the design matrix. Thus, $\omega(\mathbf{u})$ can be computed and used in the general formula, given in Claeskens and Hjort (2008b), for FIC in the normal linear model.

Since $\omega(\mathbf{u})$, and thus the FIC, depends on $\mathbf{u}$, in this form the FIC may recommend different models for different individuals. As mentioned earlier, to perform global model selection for Δ we may use the weighted FIC (wFIC). Generally, the wFIC aims to select the model with the minimal weighted average risk for estimating a focus parameter $\mu(\mathbf{u})$ for a given weight distribution $W(\mathbf{u})$. For the simulations and data analysis in this thesis, we simply take $W(\mathbf{u})$ to be the empirical distribution of the covariates in our use of wFIC. We follow the authors' suggestion in Claeskens and Hjort (2008a) to use a truncated version of the wFIC in case the estimated squared bias is negative. To calculate the wFIC in our simulations, we used the R code provided by Gerda Claeskens on her website for Claeskens and Hjort (2008b) and made the changes necessary to target our focus parameter Δ, the conditional average treatment difference function.

Chapter 4

Treatment Effect Cross-Validation

4.1 Introduction

As discussed in Section 2.3, a common way to estimate and infer about conditional treatment effects is via a regression of the response on a treatment indicator variable and one or more baseline covariates. Within this regression framework, we may consider performing variable selection or including transformations, higher-order interactions, or polynomial terms. In Chapter 2 we discussed additional methods of estimating conditional treatment effects outside of the traditional regression framework. For almost any application, there are a large number of potential models or estimation procedures we might consider to estimate the treatment effect.

At this point, as in most applications of regression, we are confronted with the question of which model or procedure to use. The model selection problem is central to much of statistical theory and methodology, and many works have been devoted to the topic. However, model assessment and selection methods have mainly focused on the models' ability to estimate the conditional mean of the response, or to predict or classify the response given a set of predictors. AIC and delete-one cross-validation are examples of such methods targeted toward estimation of the conditional mean function or prediction.

Our goal in this work is different. We wish to assess models based on their ability to estimate the treatment effect. This problem is worth exploring because given a set of candidate models, the model that is best for estimating the mean of the response may not be the best for estimating the treatment effect; this was illustrated by an example in Section 2.5.1. Therefore, there is a need for model selection tools targeted to treatment effect estimation.

We develop two such targeted model selection criteria in this thesis. In Chapter 3, we described the use of the focused information criterion (FIC) for estimation of treatment effects. In this chapter, we extend the powerful and flexible idea of cross-validation (CV) to the estimation of treatment effects. The traditional use of CV in the regression context (see, e.g., Stone (1974), Geisser (1975), and Opsomer et al. (2001)) assesses a model's performance in terms of its prediction accuracy on the response variable. We devise a new form of CV for our purpose and call it treatment effect cross-validation (TECV).

Several of the methods described in Chapter 2 are nonparametric or semi-parametric in nature. One limitation of the FIC is that it is a parametric method. FIC can only compare models within the same parametric family, so it is not suitable for choosing between, for example, a linear model with interactions, the two-stage method of Cai et al. (2011), and the different tree-based methods described in Section 2.4.1. Therefore, there is a need for a model selection method that is able to compare these newer methods against each other and against traditional parametric models so that the best procedure can be chosen for the data at hand. The TECV method described in this chapter considers a general model selection problem, within which different types of models and procedures may be compared.

A more theoretical difference between FIC and TECV regards the assumption of a true model. The FIC assumes that the true data generating process is contained in the largest parametric model considered. On the other hand, TECV can evaluate any prediction algorithm and aims to select the best procedure for estimating Δ regardless of whether any of the models being considered represent the truth. Indeed, we prove that under certain conditions, the TECV method will asymptotically select the model within a candidate set that is globally the most accurate for estimating the treatment effect. To demonstrate the method, we provide examples where the TECV method successfully chooses the best model from a candidate set, while traditional methods of model selection often fail by targeting the wrong quantity and the FIC is not always suited to make a proper comparison.

The organization of this chapter is as follows. In Section 4.2 we formalize a general model selection problem in the context of treatment effect estimation. Section 4.3 introduces the treatment effect cross-validation method, states our result on selection consistency, and discusses modifications to the method for applications and high-dimensional problems. Simulation results comparing model selection methods for estimating treatment effects are presented in Section 4.4. In Section 4.5, we mention some other potential uses for treatment effect cross-validation. A detailed proof of the selection consistency theorem stated in this chapter can be found in Appendix A.

4.2 Model Selection for Δ

In this chapter of the thesis, our goal is to choose a treatment effect estimation procedure from a finite collection of candidate procedures $\{\phi_1, \phi_2, \dots, \phi_K\}$. The candidate procedures may be from linear regression models (with or without treatment-covariate interactions), different tree-based prediction methods, the two-stage estimators of Cai et al. (2011), or a collection of several different types of procedures. We do not assume that any of the ϕ's represent the true model; nevertheless, we would like to find the best procedure ϕ from among the candidate set. By the best procedure, we mean the one with the lowest risk according to some loss function L.

To formalize these ideas, we adopt the following definitions of asymptotic procedure comparisons and selection consistency from Yang (2007). Let $L(\Delta, \widehat{\Delta})$ be a loss function for estimating Δ. For simplicity, consider only two estimation procedures ϕ_1 and ϕ_2, and let $\{\widehat{\Delta}_{n,1}\}_{n=1}^{\infty}$ and $\{\widehat{\Delta}_{n,2}\}_{n=1}^{\infty}$ be the resulting estimators when applying the two procedures at sample sizes $n = 1, 2, \dots,$, respectively.

Definition Procedure ϕ_1 is *asymptotically better* than ϕ_2 for estimating Δ under the loss function $L(\Delta, \widehat{\Delta})$ if for every $0 < \epsilon < 1$, there exists a constant $c_\epsilon > 0$ such that when n is large enough,

$$P\{L(\Delta, \widehat{\Delta}_{n,2}) \geq (1 + c_\epsilon) L(\Delta, \widehat{\Delta}_{n,1})\} \geq 1 - \epsilon.$$

When comparing a collection of candidate procedures for estimating Δ, if one procedure in the collection is asymptotically better than each of the other procedures, we can define the concept of selection consistency with regard to choosing a procedure from the collection of candidates.

Definition Assume that one of the candidate procedures ϕ^* is asymptotically better than each of the other candidate procedures for estimating Δ under the loss function L. A selection rule is said to be *consistent* if the probability of selecting ϕ^* from among the candidates approaches 1 as $n \to \infty$.

Although there are exceptions, in most comparisons one procedure will be asymptotically better than the other for estimating the treatment effect, even when neither procedure represents a correct model. Therefore, the concept of model selection consistency in this context will be well-defined and often practically important.

We now define what we mean by a procedure's exact rate of convergence for estimating Δ.

Definition Let $\{a_n\}$ be a sequence of positive numbers with $\lim_{n\to\infty} a_n = 0$. A procedure ϕ (or $\{\widehat{\Delta}_n\}_{n=1}^{\infty}$) for estimating Δ is said to converge at exactly rate $\{a_n\}$ in probability under the loss L if $L(\Delta, \widehat{\Delta}_n) = O_p(a_n)$, and if for every $0 < \epsilon < 1$, there exists $c_\epsilon > 0$ such that when n is large enough, $P\{L(\Delta, \widehat{\Delta}_n) \geq c_\epsilon a_n\} \geq 1 - \epsilon$.

We define the L_q norm with respect to the distribution of the covariates.

$$\|f\|_q = \begin{cases} \left\{\int |f(\mathbf{u})|^q P_{\mathbf{U}}(d\mathbf{u})\right\}^{1/q}, & \text{for } 1 \leq q < \infty \\ \text{ess sup}|f|, & \text{for } q = \infty, \end{cases}$$

where $P_{\mathbf{U}}$ denotes the probability distribution of $\mathbf{U}_i$ for $1 \leq i \leq n$.

Since $\Delta = f_t - f_c$, if a procedure ϕ provides estimates of f_t and f_c which converge at rates a_n and b_n, respectively, then ϕ converges to Δ at rate $\max(a_n, b_n)$. Thus, for example, if the true model for Y is linear with treatment-covariate interaction terms, then a procedure ϕ representing a linear model with all the proper terms will converge to Δ at rate $n^{-1/2}$ under the L_2 loss.

4.3 Treatment Effect Cross-Validation

Cross-validation is a commonly used model selection tool. To use cross-validation for model selection, one splits the data into training and evaluation parts, fits each candidate model (or procedure) to the training part of the data, and selects the model that performs the best on the evaluation part. Often multiple splittings of the data are done and the performance of each model is assessed over the multiple evaluation parts, thus putting the 'cross' in cross-validation.

For regression problems, model assessment in cross-validation is typically based on a summary of the individual prediction errors $Y_i - \widehat{Y}_i$ in the evaluation data. However, for the goal of estimating Δ, individual prediction errors are not available because each subject is in the treatment or control group and so Δ is not observed for any individual. Therefore, the typical usage of cross-validation must be modified in order to target estimation of Δ.

Our treatment effect cross-validation (TECV) method is based on pairing each individual in the treatment group with an individual in the control group that has similar covariate values. From each pair, we approximate the treatment effect by subtracting the response of the untreated individual from that of the treated individual. We then compare these approximated treatment effects with the estimated treatment effects from the candidate procedures to assess the accuracy of the various competitors. In this section, we describe the TECV method in detail and show that under some conditions, it is selection consistent for estimating Δ under the L_2 loss defined earlier.

4.3.1 TECV for Theoretical Development

The following steps outline the treatment effect cross-validation method for which we prove model selection consistency for Δ.

Step 0. Select a number of observations $n_1 < n$ that will be used to fit the models. The remaining $n_2 = n - n_1$ observations will be used to evaluate the models.

Step 1. Randomly permute the order of the n observations; call this permutation π. Split the data into two parts: the training part $Z^{(1)} = (Y_i, T_i, \mathbf{U}_i)_{i=1}^{n_1}$ and the evaluation part $Z^{(2)} = (Y_i, T_i, \mathbf{U}_i)_{i=n_1+1}^{n}$.

Step 2. Fit the K candidate models (or generally, the K candidate estimation procedures) $\phi_1, \phi_2, \ldots, \phi_K$ to the data $Z^{(1)}$ to obtain K estimates $\widehat{\Delta}_{n_1,1}, \widehat{\Delta}_{n_1,2}, \ldots, \widehat{\Delta}_{n_1,K}$ of the treatment effect function. Use these procedures to estimate the treatment effect for each observation in the evaluation part, and denote these estimates as $\widehat{\Delta}_{n_1,k}(\mathbf{U}_i)$ for $1 \leq k \leq K$ and $n_1 + 1 \leq i \leq n$.

Step 3. Partition $\mathcal{U}$ into cells of suitably-chosen size. (The size of the cells is discussed in remark (f) following the theorem and in Appendix A.) For each cell j, if the cell contains at least one treatment observation from $Z^{(2)}$ and at least one control observation from $Z^{(2)}$, randomly select a pair of observations (j_t, j_c) such that $T_{j_t} = t$ and $T_{j_c} = c$. Denote the resulting number of pairs by $\widetilde{n}_2$, and let W_j be the number of observations from $Z^{(2)}$ in cell j.

Step 4. For each of the $\widetilde{n}_2$ pairs (j_t, j_c), create approximate treatment effects $\widetilde{\delta}_j := Y_{j_t} - Y_{j_c}$. Then for each candidate model k, compute the TECV statistic

$$\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,k}) = \sum_{j=1}^{\widetilde{n}_2} W_j \{\widetilde{\delta}_j - \widehat{\Delta}_{n_1,k}(\mathbf{U}_{jt})\}^2,$$

where π denotes the permutation applied in Step 1. The purpose of the weights W_j in the formula for TECV_π is to relate the statistic to the global L_2 loss by giving higher weights to the model assessments in regions where the density of $\mathbf{U}$ is greater.

Step 5. Assign the vote for π to the procedure ϕ_k with the lowest value of $\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,k})$.

Step 6. Repeat steps 1-5 over a collection of permutations Π, and count the number of votes each procedure receives in step 5. The procedure ϕ_k with a plurality of the votes over the collection Π is chosen. If multiple procedures receive a plurality, any tiebreaking method may be used to choose one.

4.3.2 Selection Consistency of TECV

In this section we show that, under some regularity conditions and conditions on the data splitting ratio, the treatment effect cross-validation algorithm just described will, with probability tending to one as n grows, select the procedure that is asymptotically best among the candidates for estimating Δ with respect to the global L_2 loss. For simplicity, our result focuses on the situation where two procedures, ϕ_1 and ϕ_2, are being compared. However, the result can be generalized to the situation where any finite collection of procedures is being evaluated and one of them is asymptotically best.

Conditions

We now enumerate and discuss the conditions used to obtain the consistency of TECV under a proper data splitting ratio.

(a) Under the global L_2 loss for Δ, either ϕ_1 is asymptotically better than ϕ_2, or ϕ_2 is asymptotically better than ϕ_1.

(b) The covariate density $P_{\mathbf{U}}$ has compact support $\mathcal{U} \subset \mathbb{R}^p$ for some $p \geq 1$. Without further loss of generality, we take $\mathcal{U} = [0,1]^p$.

(c) The covariate densities for the treatment and control groups, $P_{\mathbf{U}_t}$ and $P_{\mathbf{U}_c}$, are each lower and upper bounded by constants $\underline{c} > 0$ and $\overline{c} > 0$ on $[0,1]^p$.

(d) The number of treatment and control observations are asymptotically of the same order. That is, for n large enough, there exist constants $(a,b) \in (0,1)$ such that $0 < a < n_t/n < b < 1$, where n_t is the number of the n observations for which $T_i = t$.

(e) The regression functions for the treatment and control groups, f_t and f_c, and both estimators $\widehat{\Delta}_{n,1}$ and $\widehat{\Delta}_{n,2}$, have all p partial derivatives, and each of these partial derivatives is upper bounded in absolute value by a positive constant L.

(f) The collections ξ_i and ν_i, $1 \leq i \leq n$, denoting the random errors under treatment and control, have finite variances σ_t^2 and σ_c^2, respectively. Neither these variances nor the shapes of the error distributions are assumed to be known or identical.

(g) There exists a sequence of positive numbers A_n such that for $k = 1,2, \|\Delta - \widehat{\Delta}_{n,k}\|_\infty = O_p(A_n)$, where $\|\cdot\|_\infty$ denotes the essential supremum (i.e., the sup-norm).

(h) There exists a sequence of positive numbers M_n such that for $k = 1,2$, $\|\Delta - \widehat{\Delta}_{n,k}\|_4 / \|\Delta - \widehat{\Delta}_{n,k}\|_2 = O_p(M_n)$.

Condition (a) is naturally necessary for the idea of selection consistency to be meaningful. In order to uniformly bound the bias terms induced by the treatment-control pairing, the compact support mentioned in (b) is used. Conditions (c) and (d) are needed to ensure that Δ is identifiable and that when the sample size is large enough, the pairing scheme will be able to find treatment-control pairs that are near each other. Condition (e) ensures that the values of the regression functions and their estimates for the treatment and control pairs are not too far apart. The final two conditions are analogous to conditions required for Theorem 1 in Yang (2007). That paper contains discussion of each condition and examples of situations where the conditions are known to hold.

We use p_n and q_n to denote the exact rates at which $\widehat{\Delta}_{n,1}$ and $\widehat{\Delta}_{n,2}$, respectively, converge to Δ. Let $I^* = 1$ if ϕ_1 is asymptotically better than ϕ_2 for estimating Δ under the L_2 loss, and let $I^* = 2$ if ϕ_2 is asymptotically better. Let $\widehat{I}_n = 1$ if the TECV method (applied to the n data points) chooses ϕ_1; otherwise, $\widehat{I}_n = 2$.

Theorem 1 *Under the conditions provided above, if the data splitting satisfies*

(1) $n_1 \to \infty$ *and* $n_2 \to \infty$,

(2) $n_2 M_{n_1}^{-4} \to \infty$, *and*

(3) $\{n_2 \max(p_{n_1}^{2p}, q_{n_1}^{2p})\}/\{(\log n_2)(1 + A_{n_1}^{2p})\} \to \infty$,

then the TECV algorithm is consistent; that is, $P(\widehat{I}_n \neq I^*) \to 0$ *as* $n \to \infty$.

Remarks

(a) The theorem is valid for any nonempty collection of random permutations Π, including just a single permutation. Although TECV is consistent for any nonempty Π, in practice multiple permutations will average out the variability in data splitting and tend to give better results.

(b) It may be that neither ϕ_1 nor ϕ_2 converge to Δ. If $L(\Delta, \widehat{\Delta}_{n,1})$ and $L(\Delta, \widehat{\Delta}_{n,2})$ converge to different limits, then the job of discerning between them is considerably easier and the third condition on the splitting is not needed (and indeed is meaningless).

(c) For $p = 1$, the conditions on the splitting ratio are similar to those in Theorem 1 of Yang (2007); our situation adds only a $\log n_2$ term in the denominator of condition (3). However, for $p > 1$, if both candidate procedures converge, the number of evaluation observations will typically need to be much larger than the size of the training set to achieve selection consistency for Δ.

(d) Our theoretical result assumes the underlying regression functions f_t and f_c and the estimates $\widehat{\Delta}_{n,1}$ and $\widehat{\Delta}_{n,2}$ are smooth; specifically, that they have all partial derivatives bounded. This assumption may not hold for certain nonparametric estimators such as those based on a regression tree. However, since the smoothness is only needed within the cells containing the TECV pairs, the method can be adapted to maintain selection consistency in such situations. For example, in the case of a regression tree, a constraint that all TECV pairs belong to the same node of the tree may be added to the algorithm.

(e) A detailed proof of Theorem 1 can be found in Appendix A of this document.

(f) A cell size within the partition that enables the achievement of selection consistency while ensuring that $\widetilde{n}_2 \xrightarrow{p} \infty$ is given in the proof. For the use of partitioning in finite samples, the cell size may be chosen heuristically to achieve a reasonable value for $\widetilde{n}_2$. The version of TECV in Section 4.3.3 finds pairs using nearest-neighbor searches instead of partitioning, and we recommend this version for applications in most cases.

4.3.3 Modifications to TECV for Applications

Theorem 1 shows that a voting-based treatment effect cross-validation method is consistent in selection, but for finite samples, averaging over the multiple splittings may produce better results than voting. Simulations in Yang (2007) indicate that cross-validation with averaging often outperforms CV with voting for the typical use of estimating a response with the L_2 loss, although sometimes the reverse is true. Intuitively, averaging seems to be a more effective use of the data because the voting method considers only the ranking of the procedures for each splitting and ignores the magnitude of the performance differences.

For the theoretical development of TECV, within each data splitting we have assumed each observation in the evaluation part is used in at most one treatment-control pairing. This is done in order to maintain independence of the treatment effect estimates resulting from the pairs. More treatment-control pairings with shorter within-pair distances can be created by allowing observations to belong to multiple treatment-control pairs. Although the resulting approximate treatment effects will not be independent in this scheme, more of the information in the data will be used. If we remove the requirements for independence and for a uniform bound on the distance between members of each pair, we can simply create n_2 approximated treatment effects from the evaluation set by pairing each observation with its nearest neighbor in the other group. The algorithm below uses nearest-neighbor pairing and averaging over permutations to select a model for Δ. We call this algorithm TECV(a) to distinguish it from the version of TECV used for theoretical development.

Steps 0, 1, and 2. These are the same as the corresponding steps in the TECV algorithm.

Step 3. Center and scale the covariates within $Z^{(2)}$ so that each covariate has common mean and variance. Denote these standardized covariates $\widetilde{U}$. For each observation in $Z^{(2)}$, find its nearest neighbor with respect to the standardized covariates from the other treatment group within $Z^{(2)}$. Specifically, for each i find

$$i^* = \underset{n_1+1 \le i' \le n}{\arg\min}\ d(\widetilde{U}_i, \widetilde{U}_{i'}) \text{ subject to } T_i \neq T_{i'},$$

where $d(\cdot)$ represents the Euclidean distance.

Step 4. For each pair (i, i^*), create approximate treatment effects $\widetilde{\delta}_i$. If $T_i = t$, $\widetilde{\delta}_i = Y_i - Y_{i^*}$; otherwise, $\widetilde{\delta}_i = Y_{i^*} - Y_i$. Then for each candidate model, compute the TECV statistic

$$\text{TECV}_\pi(\widehat{\Delta}_{n_1,k}) = \sum_{i=n_1+1}^{n} \{\widetilde{\delta}_i - \widehat{\Delta}_{n_1,k}(\mathbf{U}_i)\}^2.$$

Step 5. Repeat Steps 1-4 multiple times over a collection of permutations Π. Average the TECV_π statistics for each procedure ϕ_k, and select the procedure with the lowest average TECV_π over Π. That is, select the procedure ϕ_k with the lowest value of

$$\overline{\text{TECV}}(\widehat{\Delta}_{n_1,k}) = \frac{1}{|\Pi|}\sum_{\pi\in\Pi}\text{TECV}_{\pi}(\widehat{\Delta}_{n_1,k}).$$

TECV(a) has some practical advantages over the theoretical version of TECV, particularly when the sample size is small. We conjecture that TECV(a) shares the selection consistency property of TECV under similar conditions. The main ideas behind the two algorithms are similar, so we will not pursue theoretical results for TECV(a) here.

4.3.4 Sufficient Dimension Reduction for TECV

When p, the dimension of the covariate vector $\mathbf{U}$, is large or even moderate, it may be difficult for the TECV algorithm to identify nearby treatment-control pairs. For example, if $p = 10$ and each column of $\mathbf{U}$ is coarsely divided into only two bins, $2^{10} = 1024$ bins will result, and thus a sample size of at least 2048 would be needed to find a treatment-control pair in each bin. Our proposed solution to this "curse of dimensionality" is to pursue low-dimensional linear combinations of $\mathbf{U}$ that capture all of the information in the treatment and control regression functions $f_t(\mathbf{u})$ and $f_c(\mathbf{u})$. Pairs of neighboring observations based on these low-dimensional linear combinations of $\mathbf{U}$ are typically much easier to find, and if the dimension reduction has not resulted in any loss of information regarding the regression functions, these pairs will be similar with regard to their values of f_t and f_c (and therefore Δ). In the current section, we develop these ideas further using the framework of sufficient dimension reduction (Cook, 1998). In Sections 4.4.4 and 4.4.5, we provide simulation examples showing the effectiveness of TECV following the application of dimension reduction techniques.

The usual goal of sufficient dimension reduction is to find a low-dimensional representation of the predictors that contains all the information that the predictors have about the response variable. Because our object of interest Δ involves only $E(Y|T = t, \mathbf{U} = \mathbf{u})$ and $E(Y|T = c, \mathbf{U} = \mathbf{u})$, we restrict our attention to these mean functions. Cook and Li (2002) define a mean dimension reduction subspace and the central mean subspace for regressions where only the conditional mean is of interest. A mean dimension reduction subspace S for the regression of Y on $\mathbf{U}$ is a subspace such that $E(Y|\mathbf{U}) \perp\!\!\!\perp \mathbf{U}|\eta^T\mathbf{U}$, where $\perp\!\!\!\perp$ denotes independence and η is a matrix whose columns form a basis in S.

When the intersection of all such subspaces is itself a mean dimension reduction subspace, it is called the central mean subspace and denoted by $S_{E(Y|\mathbf{U})}$. Existence and uniqueness of $S_{E(Y|\mathbf{U})}$ are guaranteed under mild conditions (Cook and Li, 2002). In many cases, $d = \dim(S_{E(Y|\mathbf{U})})$ is much less than the original dimension p. For example, $d = 1$ in linear models, generalized linear models, and single-index models.

Within the framework developed in Chapter 1 of the thesis, $f_t(\mathbf{u}) = E(Y|T = t, \mathbf{U} = \mathbf{u})$, $f_c(\mathbf{u}) = E(Y|T = c, \mathbf{U} = \mathbf{u})$, and $\Delta(\mathbf{u}) = f_t(\mathbf{u}) - f_c(\mathbf{u})$. Let S_t denote the central mean subspace for f_t and S_c the central mean subspace for f_c, and let d_t and d_c be the respective dimensions of S_t and S_c. It is easy to show that $\Delta(\mathbf{u}) = \Delta(\eta_{tc}^T\mathbf{u})$, where η_{tc} is a matrix whose columns span both S_t and S_c. Let $S_{tc} = \{s_t + s_c | s_t \in S_t, s_c \in S_c\}$. S_{tc} has dimension at most $d_t + d_c$ and carries all the information that $\mathbf{U}$ has about $\Delta(\mathbf{u})$. Thus the central mean subspaces for the regressions under treatment and control can be combined to produce a typically low-dimensional representation of $\mathbf{U}$ that is sufficient for our object of interest.

If S_{tc} were known, its utilization would increase the statistical and computational efficiency of the TECV method. If $d_{tc} := \dim(S_{tc}) \ll p$, the modified task of finding a treatment-control pair (i, i^*) such that $d(\eta_{tc}^T\mathbf{U}_i, \eta_{tc}^T\mathbf{U}_{i^*})$ is small is much easier than finding a pair for which $d(\mathbf{U}_i, \mathbf{U}_{i^*})$ is small, because it is easier to find nearby neighbors in low-dimensional space. In addition, because $\Delta(\mathbf{u}) = \Delta(\eta_{tc}^T\mathbf{u})$, these low-dimensional representations would result in no loss of information about Δ. Therefore, Theorem 1 would still hold if $\eta_{tc}^T\mathbf{U}$ were substituted for $\mathbf{U}$ in the TECV pairing algorithm and the reduced dimension d_{tc} were substituted for the original dimension p. Indeed, the third condition on the data splitting could be relaxed if d_{tc} were less than p.

In practice, S_t and S_c are typically unknown and must be estimated. Fortunately, there are several methods available to estimate vectors in the central subspace (within which the central mean subspace is contained; see Cook and Li (2002)) without assuming a model for the data. These include sliced inverse regression (SIR) (Li, 1991) and sliced average variance estimation (SAVE) (Cook and Weisberg, 1991). Iterative Hessian transformation (IHT) (Cook and Li, 2002) specifically targets vectors in the central mean subspace. Cook et al. (2007) propose a method for estimating vectors in the central subspace without matrix inversion; the method is therefore applicable regardless of the (n, p) relationship. Li (2007) and Li and Yin (2008) apply regularization techniques to obtain sparse estimates of the central subspace when p is large. If a linear model is assumed for the data, estimation of the central mean subspace is equivalent to estimation of the regression coefficients. Simulation studies in Sections 4.4.4 and 4.4.5 illustrate the use of dimension reduction followed by treatment effect cross-validation.

4.4 Simulation Studies

In this section, we compare wFIC and TECV with the commonly used AIC (Akaike, 1974), BIC (Schwarz, 1978), and traditional cross-validation methods in a variety of simulation settings. Traditional CV is implemented by splitting the data 100 times with a 50-50 splitting ratio, computing the average out-of-sample squared prediction error of each model for each validation sample, and choosing the model with the lowest average squared error over the 100 splits. Likewise, for TECV, 100 different data splittings are done. TECV in these simulations denotes the TECV(a) algorithm for applications described in Section 4.3.3. In simulations where the set of candidate models is small, a voting-based version of the TECV(a) algorithm is also tried; it is denoted by TECV(v).

In all simulation examples, the performance of each model selection method is aggregated over 100 different sample realizations. In each of the examples, the treatment assignments are i.i.d. with $P(T_i = t) = 0.5$ and independent of the covariate vector $\mathbf{U}$, while the errors also are independent of $\mathbf{U}$ and have a Gaussian distribution with mean 0 and equal variance σ^2 for the treatment and control groups. All simulations were performed using the R software (R Core Team, 2014).

4.4.1 Nonlinear Regression Function with Constant Δ

Even when only one covariate is available, many models can be considered, and there may be a conflict between estimation of the regression function and estimation of the treatment effect. Suppose we have a situation like the one observed in Figure 4.1, in which the response is a nonlinear function of the covariate but the effect of the treatment is independent of the covariate. Specifically, Figure 4.1 is a realization of $n = 300$ from

$$Y_i = I(T_i = t) + 3U_i + 3\exp\{-100(U_i - 0.5)^2\} + \varepsilon_i,$$

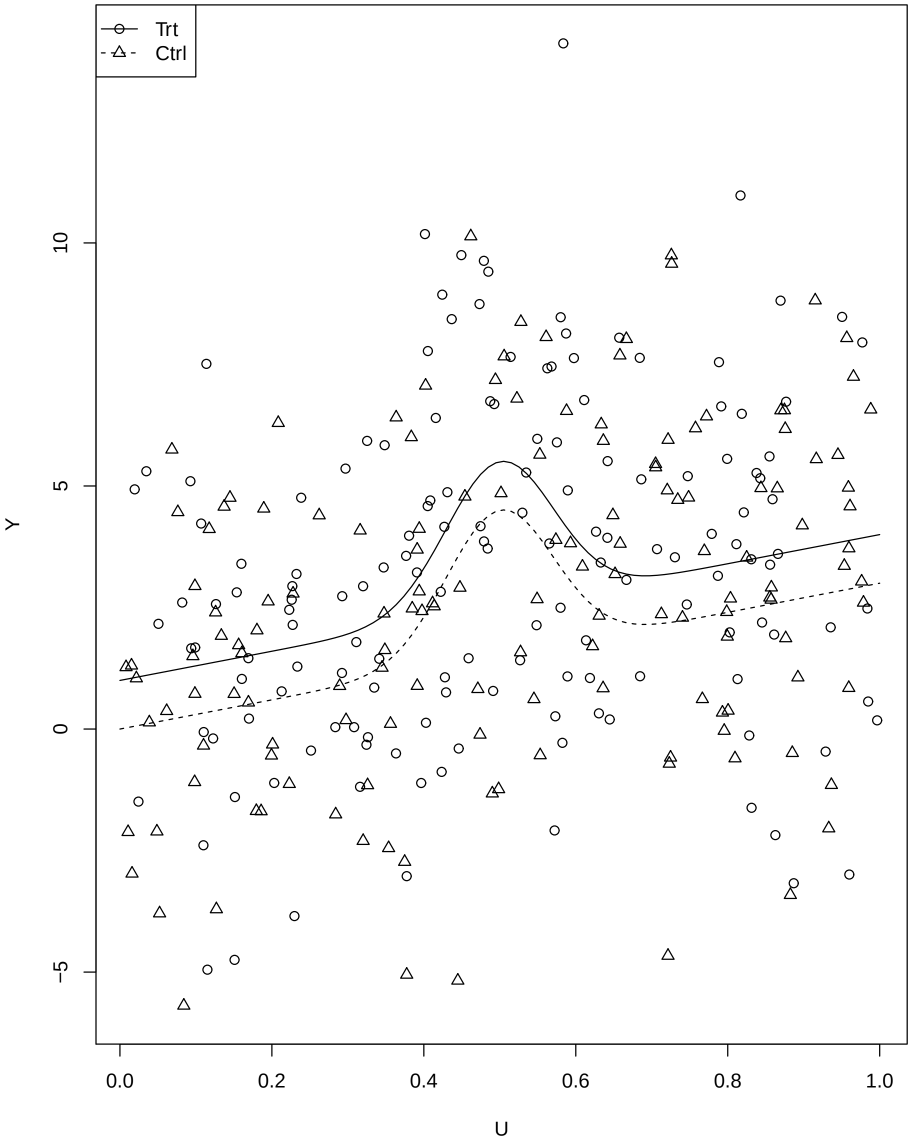


Figure 4.1: Realization of $n = 300$ from the example with a nonlinear regression function and constant Δ. The treatment and control regression functions are drawn on the plot.

where U is uniformly distributed on $[0, 1]$ and $\sigma = 3$. Three models are considered:

1. A linear model with main effects for T and U but no interaction term.

Table 4.1: Simulation results: nonlinear mean functions with constant Δ

	Linear[a]	Quadratic[b]	Sm. Spline[b]	
Avg MSE for $E(Y\|T,U)$	0.95	0.73	0.51	
Avg MSE for Δ	0.15	0.41	0.84	
Times Selected by				Avg MSE for Δ (SE)
AIC	1	0	99	0.84 (0.05)
BIC	79	19	2	0.23 (0.03)
CV	30	27	43	0.63 (0.05)
wFIC	79	21	–	0.26 (0.03)
TECV	90	8	2	0.25 (0.04)
TECV(v)	85	10	5	0.30 (0.04)

[a] No treatment-covariate interactions are included in the linear model.

[b] Treatment-covariate interactions are included in the quadratic and smoothing spline models.

2. A quadratic model with interaction terms between T and U, and between T and U^2.

3. A smoothing spline model in which $E(Y|U)$ is allowed to depend on T.

To fit the smoothing spline model, we use the gam function in the R mgcv package and use the default (generalized cross-validation) choice of smoothing parameter. This package is described in Wood (2001) and Wood (2006).

To evaluate the effectiveness of a model ϕ for estimating the regression function, within each sample realization the average of $\{E(Y|T,U) - \widehat{Y}_\phi\}^2$ is taken over an independent evaluation set of 100,000 randomly generated values of (T,U) from the design distribution, where $\widehat{Y}_\phi$ is the estimate of Y obtained by fitting model ϕ to the realized sample. These values are then averaged over the 100 realizations to estimate the average MSE of ϕ for $E(Y|T,U)$. The average MSE of each ϕ for Δ is estimated likewise. For each model selection method M, the average MSE for Δ is obtained by performing a similar double averaging of $(\Delta - \widehat{\Delta}_{\phi_M})$, where $\widehat{\Delta}_{\phi_M}$ is the estimate of Δ obtained by fitting the model ϕ_M chosen by M for that realization of data.

As seen in Table 4.1, in this example the simplest model is best for treatment effect estimation, while the more complicated models improve the estimation of the regression

Table 4.2: Simulation results: $E(Y|U)$ linear with nonlinear Δ

	Linear[a]	Quadratic[b]	Sm. Spline[b]	
Avg MSE for $E(Y\|T,U)$	0.95	0.73	0.50	
Avg MSE for Δ	3.53	2.53	1.18	
Times Selected by				Avg MSE for Δ (SE)
AIC	0	0	100	1.19 (0.05)
BIC	73	23	4	3.26 (0.06)
CV	18	43	39	2.12 (0.11)
wFIC	3	97	–	2.56 (0.04)
TECV	5	14	81	1.43 (0.08)
TECV(v)	2	14	84	1.36 (0.07)

[a] No treatment-covariate interactions are included in the linear model.

[b] Treatment-covariate interactions are included in the quadratic and smoothing spline models.

function. AIC and traditional CV pursue the estimation of the regression function and so tend to choose the quadratic or smoothing spline models in this case. Meanwhile, the methods targeted to estimation of the treatment effect, wFIC and TECV, both correctly prefer the linear model with no interactions a majority of the time, with TECV choosing it the most often (90 times out of 100).

4.4.2 $E(Y|U)$ Linear with Nonlinear Δ

We next examine a situation that is, in a sense, opposite of the previous example. The values of n and σ and the distribution of U are the same as the previous example, but the mean functions take the form

$$\begin{aligned} f_t(u) =& 1 + 3u + 3\exp\{-100(u-0.5)^2\} \\ f_c(u) =& 3u - 3\exp\{-100(u-0.5)^2\} \\ \Delta(u) =& 1 + 6\exp\{-100(u-0.5)^2\} \end{aligned}$$

A realization of $n = 300$ with the mean functions superimposed is shown in Figure 4.2.

The same three models are considered; results are found in Table 4.2. The smoothing

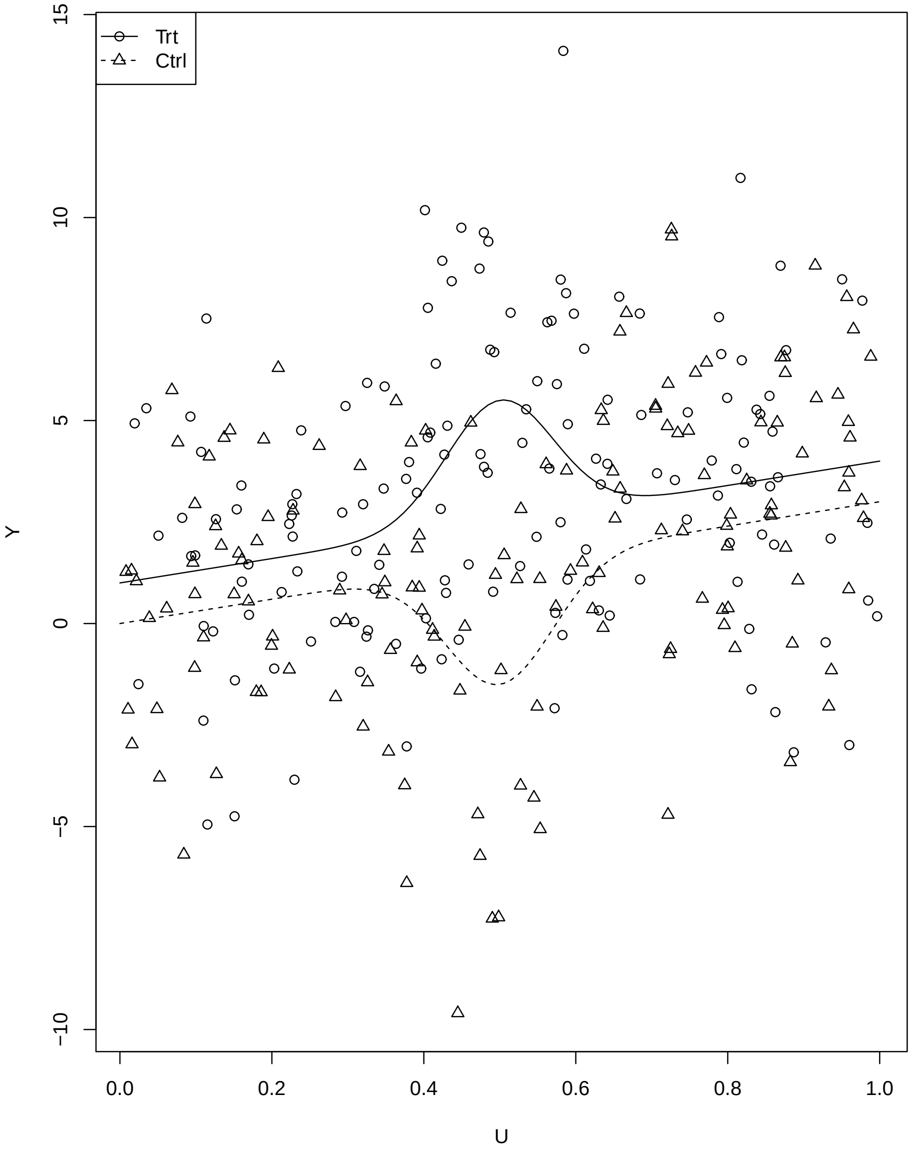


Figure 4.2: Realization of $n = 300$ from the example with $E(Y|U)$ linear but a nonlinear treatment effect. The treatment and control regression functions are drawn on the plot.

spline model performs the best both for estimating the regression function and for estimating Δ, but the performance difference between the spline model and the simpler

Table 4.3: Selection frequencies of M_2 in motivating example

Model Selection Method	$n = 100$	$n = 200$	$n = 400$	$n = 800$
AIC/BIC	27	19	12	2
CV	26	18	13	2
wFIC	84	93	100	100
TECV	75	89	99	100
TECV(v)	76	85	100	100

models is larger for Δ than for the regression function. Both versions of TECV choose the smoothing spline model a majority of the time. wFIC cannot compare the smoothing spline model against the others, so its potential is limited in this setting.

TECV is the only method to perform well in treatment effect estimation for both this example and the previous one. This shows the potential of TECV to select a good model for the treatment effect when the treatment effect is simple and when it is more complicated than is described by typical parametric models. This is in accord with our theoretical result, which shows selection consistency is achievable for TECV without assuming a parametric form for Δ.

4.4.3 Motivating Example of Section 2.5.1

In this section we compare the success of different model selection methods for the example described in Section 2.5.1 of this thesis. Using the notation described in that section, we set $\beta_{\text{true}} = 3$, $\gamma_{\text{true}} = 3$, $\alpha_{\text{true}} = 0$, $\tau_{\text{true}} = 1$ and $\sigma = 10$. Under this configuration, at any sample size model M_1 will have lower risk for estimating the conditional mean while M_2 will have lower risk for estimating Δ under the global L_2 loss.

Four different sample size levels (n = 100, 200, 400, and 800) are considered in the simulation. Table 4.3 shows, for each sample size, how often model M_2 was selected by AIC/BIC, traditional CV, wFIC, TECV, and TECV(v). Note that the same number of parameters are estimated in M_1 and M_2, so AIC and BIC are equivalent for this problem.

This simple simulation example shows the need for targeted model selection methods. As expected, the traditional model selection methods (AIC/BIC and traditional CV) all tend to target model M_1 as the sample size grows because M_1 has lower risk for estimating the full regression function. Meanwhile, the methods targeted toward treatment effect estimation all tend to choose M_2 for large n. The three targeted methods all perform similarly, with wFIC having a slight edge in this setting.

4.4.4 Dimension Reduction by Penalized Regression

In order for treatment effect cross-validation to be effective when p is large, the dimension of $\mathbf{U}$ typically will need to be reduced prior to identifying the treatment-control pairs central to the TECV algorithm. In this section we demonstrate the use of dimension reduction followed by TECV, following the ideas presented in Section 4.3.4.

When $f_t(\mathbf{u}) = E(Y_i|\mathbf{U}_i = \mathbf{u}, T_i = t)$ is linear in $\mathbf{u}$ (i.e., $f_t(\mathbf{u}) = \beta_t^T\mathbf{u}$), $\beta_t^T\mathbf{U}$ is a sufficient dimension reduction of $\mathbf{U}$ for f_t. Likewise, if $f_c(\mathbf{u}) = \beta_c^T\mathbf{u}$, $\beta_c^T\mathbf{U}$ is sufficient for f_c. Thus the matrix $(\beta_t^T\mathbf{U}, \beta_c^T\mathbf{U})$, with only two columns, retains all of the information for Δ originally contained in the p-dimensional $\mathbf{U}$. Good estimates of β_t and β_c may thereby substantially reduce the dimension of the problem without much loss of information about Δ.

Therefore, if a linear model is assumed, we may utilize the estimated coefficient vectors for the treatment and control groups prior to the pairing step of the TECV algorithm. Let $\widehat{\beta}_{tc}$ denote the $p \times 2$ matrix whose columns are $\widehat{\beta}_t$ and $\widehat{\beta}_c$. In the pairing step of the TECV algorithm, we measure the distance between subjects i and i' by $d(\widehat{\beta}_{tc}^T\mathbf{U}_i, \widehat{\beta}_{tc}^T\mathbf{U}_{i'})$. There are many ways to estimate β_t and β_c; penalized regression methods are commonly used when the number of covariates p exceeds the sample size n. Two popular penalized regression methods are the LASSO (Tibshirani, 1996) and the minimax concave penalty (MCP) (Zhang, 2010). In this section we illustrate the use of these penalized regression methods with our TECV algorithm.

The regression coefficients under treatment and control, β_t and β_c, are estimated by first performing variable selection via penalized regression to identify the nonzero elements, then estimating the nonzero coefficients by ordinary least squares (OLS) regression. This two-step approach allows the important variables to be identified in a screening step while avoiding the possible downward bias in the coefficients caused by the penalization.

To construct LASSO solution paths, the R `glmnet` package (Friedman et al., 2010) was used, while the `ncvreg` package (Breheny and Huang, 2011) was used to obtain MCP solutions. The LASSO tuning parameter λ was chosen by 10-fold CV (the default in `glmnet`). For MCP, the tuning parameter γ was set to 3 while λ was chosen by 10-fold CV (both defaults in `ncvreg`). Calls to `glmnet` and `ncvreg` by default fit models using 100 values of λ. When $p > n$, both packages by default set the minimum λ to be fairly large to avoid fitting models that are too rich. In our simulations, we observed that the minimum CV statistic often occurred at the minimum λ considered under the default setting; in such situations, we expanded the range of λ values under consideration to give larger models an opportunity to be selected. Specifically, when the minimum CV statistic occurred at the minimum λ considered, we divided the minimum λ by two, doubled the number of λ values considered, and ran the penalized method again. This process was iterated until the minimum CV statistic was achieved at a value of λ other than the minimum λ considered.

A separate penalized regression is applied to obtain the set of candidate models for Δ. We consider $2p + 1$ potential predictors: the treatment main effect, p covariate main effects, and p treatment-covariate interactions. For each set of active variables on the penalized regression solution path, a model is fit using OLS on the entire set of n observations. The range of λ values is expanded as before so that the minimum CV statistic does not occur at the minimum λ considered.

Example: $p = 200$, Quadratic Δ

In this example, the true regression function is nonlinear but the nonlinearity is not clearly visible at the given sample size. We set $p = 200$, $n = 250$, and $\sigma = 1.5$. The covariates $\mathbf{U}$ are normal with mean zero and covariance matrix $\Sigma_{ij} = 0.9^{|i-j|}$. For the regression functions, we set

$$f_t(\mathbf{u}) = 10 + 10.5u_1 + 5u_1^2 + 8.5u_2 + 2.5u_3 + 3u_3^2 + 2.5u_4 + 7.5\textstyle\sum_{k=5}^{8} u_k + 2.5\sum_{l=9}^{12} u_l$$

and

$$f_c(\mathbf{u}) = 0 + 2.5u_1 + 2.5u_2 + 4u_2^2 + 6.5u_3 + 4.5u_4 + 2u_4^2 + 7.5\textstyle\sum_{k=5}^{8} u_k + 2.5\sum_{l=9}^{12} u_l.$$

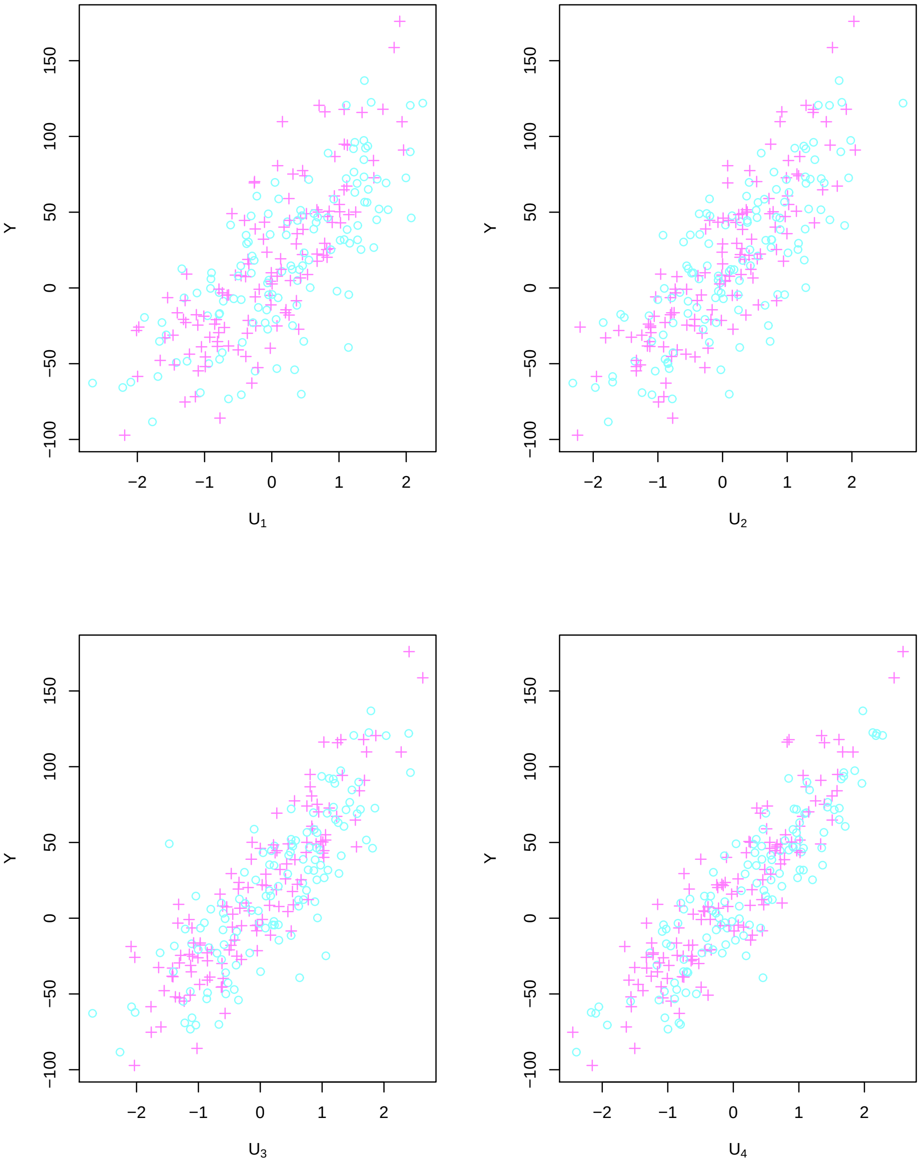


Figure 4.3: Plots of Y vs. $U_1 - U_4$: $n = 250$, $p = 200$, quadratic Δ. Points marked by + represent observations from the treatment group, while points marked by ○ represent control observations. The nonlinear relationships are not clearly visible.

Table 4.4: Simulation results: $p = 200$, quadratic Δ

Model Selection After MCP Screening

Selection Method	Avg MSE for Δ (SE)	TPR (%)	FPR (%)	Pairwise (%)
AIC	74.7 (3.2)	25.3	1.9	72.5
BIC	71.2 (3.0)	25.3	1.5	77.0
CV	73.9 (3.2)	25.3	1.8	73.3
wFIC	75.3 (3.3)	25.6	1.5	70.5
TECV	58.9 (2.7)	22.3	0.4	89.7
MCP Estimate	59.0 (3.2)	24.0	1.0	–

Model Selection After LASSO Screening

Selection Method	Avg MSE for Δ (SE)	TPR (%)	FPR (%)	Pairwise (%)
AIC	77.4 (2.6)	37.3	6.1	63.9
BIC	46.9 (1.4)	33.0	0.6	93.0
CV	55.9 (1.7)	35.0	1.9	85.6
wFIC	76.5 (3.1)	36.8	4.9	65.2
TECV	45.4 (1.1)	33.5	0.7	94.3
LASSO Estimate	47.3 (1.0)	38.0	3.4	–

Therefore,

$$\Delta(\mathbf{u}) = 10 + 8u_1 + 5u_1^2 + 6u_2 - 4u_2^2 - 4u_3 + 3u_3^2 - 2u_4 - 2u_4^2.$$

In Figure 4.3, it is not clear that the regression function for the treatment group is nonlinear in the left two panels (u_1 and u_3), while the regression function under control is nonlinear in the right two panels (u_2 and u_4). Since there is no obvious nonlinearity, a researcher may believe a linear model is appropriate and apply the LASSO or MCP to do variable selection.

Table 4.4 compares the five model selection methods, as well as the estimates from the penalized regressions, in this setting. The TPR column denotes the percentage of true linear interaction terms that are included, on average, in the model selected by each method; FPR is the false positive rate of linear interaction terms included. The

Table 4.5: Simulation results: TECV with and without dimension reduction

Selection Method	Avg MSE for Δ (SE)	TPR (%)	FPR (%)	Pairwise (%)
AIC	0.133 (0.007)	44.5	19.0	61.6
BIC	0.160 (0.006)	19.0	3.3	53.3
CV	0.139 (0.007)	34.8	10.8	58.6
wFIC	0.139 (0.007)	33.5	10.0	60.3
TECV (with DR)	0.135 (0.006)	24.5	4.0	59.1
TECV (without DR)	0.164 (0.006)	21.0	6.8	52.6

Pairwise column shows the percentage of pairwise comparisons for which the model with the lower average risk for Δ in the evaluation set was assigned the lower value of the model selection statistic. Because of the correlation among the covariates involved in the active linear interaction terms, it is difficult for either screening method to identify all four interactions. In this difficult situation in which the covariates are correlated and all candidate models are misspecified, TECV overall performs as well as or better than its competitors for providing accurate estimates of the treatment effect.

4.4.5 TECV With and Without Dimension Reduction

Here we consider an example with $p = 8$, as might be realistic in a clinical trial setting where eight covariates are known to influence the response. While this setting would not typically be considered high-dimensional, it still may be difficult to find nearby neighbors in eight dimensions and dimension reduction prior to implementation of TECV may therefore be beneficial. We evaluate the performance of TECV with and without dimension reduction against other model selection methods in this setting. The sample size is $n = 300$ with $\sigma = 1$. The eight columns of $\mathbf{U}$ are mean-zero normal with $\Sigma_{ij} = 0.5^{|i-j|}$. We have

$$\begin{aligned}
f_t(\mathbf{u}) =& 0.5 + 0.8u_1 - 0.5u_2 + 0.2u_4 - 0.1u_5 + 0.4u_7 - 0.2u_8, \\
f_c(\mathbf{u}) =& 0 + 0.4u_1 - 0.2u_2 + 0.4u_4 - 0.2u_5 + 0.4u_7 - 0.2u_8, \text{ and} \\
\Delta(\mathbf{u}) =& 0.5 + 0.4u_1 - 0.3u_2 - 0.2u_4 + 0.1u_5.
\end{aligned}$$

The candidate models considered in this setting are all linear; the treatment and covariate main effects are included in all candidate models, and all 256 combinations of the eight treatment-covariate interaction terms are considered. For the dimension reduction, all eight covariates are used to estimate $\widehat{\beta}_t$ and $\widehat{\beta}_c$. We observe in Table 4.5 that the performance of TECV is improved by first applying dimension reduction in this setting, then finding treatment-control pairs in two-dimensional space rather than in eight dimensions.

Overall, the simulation results show the need to consider model selection methods targeted toward treatment effect estimation, and they show the flexibility of TECV in this context. We have considered situations where p is as low as one, or as high as 200; situations in which Δ is constant, linear, or nonlinear in $\mathbf{u}$; and situations in which several candidate models are correct, or all are misspecified. For the goal of selecting a good model for the treatment effect Δ, TECV often performs as well or better than its competitors.

4.5 Other Uses for TECV

The proposed TECV method identifies treatment-control pairs with similar covariate values in the evaluation set. In the current work, the response differences between these pairs are utilized to select a good model for $\Delta(\mathbf{u})$. The TECV pairing strategy could also be used to evaluate estimates of other functions of treatment-control response pairs, such as the average of the two responses, their ratio, or their squared difference (or other moments of the difference distribution). For categorical responses, different estimates of the concordance rate (conditional on covariates) could be evaluated with TECV.

TECV could also be modified to compare more than two groups, which may be useful in evaluating a treatment with more than two levels. For instance, the FIRST clinical trial analyzed in Chapter 6 compares one treatment regimen of three drug classes (call this treatment t) and two different treatment regimens of two drug classes (call them c_1 and c_2). Different models may be proposed to estimate the difference between a patient's expected response to the three-drug treatment, $E(Y_i|T_i = t, \mathbf{U}_i = \mathbf{u})$, and the average of the patient's expected responses to the two-drug treatments, $(E(Y_i|T_i = c_1, \mathbf{U}_i = \mathbf{u}) + E(Y_i|T_i = c_2, \mathbf{U}_i = \mathbf{u}))/2$. Triplets containing one member of each treatment group could be formed by partitioning the covariate space, and the values of $Y_{(t)} - (Y_{(c_1)} + Y_{(c_2)})/2$ from these triplets could be used to compare the different models. In general, other functions of a subgroup's responses (e.g., the maximum of the responses, the variability between responses, or the level of agreement between categorical responses) could be used to compare different models that attempt to estimate the conditional target quantity of interest.

Chapter 5

Combining Estimates of Conditional Treatment Effects

5.1 Introduction

Estimating the causal effect of a treatment on a response is a primary goal of many statistical applications, particularly in fields such as business, medicine, and public policy. Most causal inference research has focused on estimation of the average treatment effect within a population. For example, Ho et al. (2007) and Abadie and Imbens (2011) provide two recent methods for estimating the average effect of a treatment from an observational study.

While knowledge of the average treatment effect in a population can be useful, treatment effects are often heterogeneous within the population. In the presence of treatment effect heterogeneity, when a treatment can be applied at the individual level (or at the level of subgroups), accurate estimation of the treatment's effect on each individual (or subgroup) can be used to increase the effectiveness of the treatment program in maximizing the outcome of interest. For example, a retailer with a limited marketing budget would be able to optimize a seasonal catalog mailing if it knew the effect of the catalog on the purchasing behavior of each household. In the public sector, an economic development agency often needs to decide which applicants will create the best utilization of grant dollars.

It also is often possible for the treatment to have a negative effect on the outcome for some individuals, even if the treatment is beneficial on average. For example, some consumers may find direct marketing efforts such as telemarketing invasive, even if the telemarketing campaigns are profitable overall. In these settings, it is important to identify such individuals to avoid the prescription of harmful treatments.

With the increasing volume of data becoming available to many organizations, estimating heterogeneous treatment effects has become more feasible. Treatment effect heterogeneity can be identified by conditioning on baseline covariates observed before the treatment is applied. There is a growing literature on the estimation of such conditional treatment effects. Cai et al. (2011) introduced a two-stage method that consistently estimates treatment effects for subgroups created by an initial parametric model, while Imai and Ratkovic (2013) developed a method to estimate treatment effect heterogeneity through L_1-penalized regression.

In most statistical applications, there are many plausible models for the data-generating process. Typically in practice, one of the plausible candidate models is selected by the researcher, and estimation, inference, and prediction are performed using the selected model. Rolling and Yang (in press) discussed the model selection process in the context of estimating the treatment effect conditional on covariates. They found that within a given candidate set of models, the best model for treatment effect estimation may be different than the best model for response estimation or prediction. This issue also was discussed in Qian and Murphy (2011) in the context of optimizing treatment decisions.

While such targeted model selection tools are a step in the right direction for accurate estimation of treatment effects, post-model selection estimators still may have large variability in finite samples because of model selection instability. In many situations, treatment effect estimates resulting from a combination of procedures may exhibit less variability and more accuracy than estimates chosen by model selection. It has been well-established (e.g., Yang, 2003) that model combination algorithms often lead to more accurate estimates and predictions than model selection procedures when model selection instability is high.

Different researchers have taken different approaches to model combination. Methods introduced from a machine learning perspective, such as stacking (Wolpert, 1992; Breiman, 1996), boosting (Freund and Shapire, 1996), and random forests (Breiman, 2001), were motivated initially by intuition and empirical performance, although some theoretical understanding of these methods was later developed (e.g., Breiman, 2004). The method of frequentist model averaging (Hjort and Claeskens, 2003) is motivated by asymptotic arguments and is justified within a parametric local misspecification setting. Bayesian model averaging (Hoeting et al., 1999) originates from a Bayesian perspective, with the model weights based on posterior probabilities of the models.

Yang (2001) viewed model combination from an adaptation point of view. His combination algorithm, called adaptive regression by mixing (ARM), possesses an oracle inequality that bounds the risk of the resulting estimator in terms of the minimum risk among the candidate procedures. This approach, which has connections with information theory, was shown to perform almost as well (up to a constant) as the best procedure among the candidates, without knowing in advance which procedure is best. An important practical advantage of ARM is its flexibility; it can combine different classes of regression models and machine learning algorithms.

The method we present in this chapter is similar in spirit to ARM but is targeted to estimate *the conditional effect of a treatment* rather than the full regression function. Since models that are good for response estimation or prediction may not be good for treatment effect estimation (and vice versa), an algorithm targeted to the specific goal of treatment effect estimation is needed to ensure that models doing a good job of estimating the treatment effect receive higher weights. This algorithm, which we call Treatment Effect Estimation by Mixing (abbreviated TEEM), is to the best of our knowledge the first model combination method specifically aimed at the important goal of accurately estimating the effect of a treatment conditional on covariates. Like the method of ARM, the TEEM method relies on data splitting to evaluate the candidate models and can combine multiple types of regression procedures and estimates; any procedure that, given data, produces an estimate of the treatment effect conditional on covariates can be used as a candidate in TEEM. Furthermore, the theoretical results we present for TEEM do not assume that any of the candidate models are correct. These features give the method tremendous flexibility to be used in a wide variety of settings.

5.2 Framework

We consider a general regression framework in which the distribution of the response Y may depend on a binary treatment variable $T \in \{t, c\}$ and one or more baseline covariates $\mathbf{U} \in \mathbb{R}^p$. In order to isolate the treatment difference of primary interest, we express the observations in the following way:

$$Y_i = \{f_t(\mathbf{U}_i) + \sigma_t \xi_i\} I(T_i = t) + \{f_c(\mathbf{U}_i) + \sigma_c \nu_i\} I(T_i = c), \qquad 1 \leq i \leq n. \tag{5.1}$$

The error terms under treatment and control are denoted by $\{\xi_i\}_{i=1}^n$ and $\{\nu_i\}_{i=1}^n$, respectively, and the error variance is σ_t^2 for those in the treatment group and σ_c^2 for control. The only difference between the expression in (5.1) and the previous representation in (1.1) is that in (5.1), we have pulled the error standard deviations out of the error terms to make them more noticeable. The estimation of these σ_t and σ_c will be part of the TEEM algorithm derived in this chapter.

As in previous chapters, the object of interest in our work is $\Delta(\mathbf{u}) := f_t(\mathbf{u}) - f_c(\mathbf{u})$, the difference between the regression functions under treatment and under control. We define causal effects using the potential outcomes framework of the Rubin Causal Model (Holland, 1986). That is, let $Y_{i,(t)}$ denote the response that would have been observed had $T_i = t$, and let $Y_{i,(c)}$ denote the corresponding potential outcome if $T_i = c$. Then the causal effect of the treatment T on unit i is the unobserved random variable $Y_{i,(t)} - Y_{i,(c)}$. Following Imbens and Wooldridge (2009), we define the Conditional Average Treatment Effect (CATE) as the expectation of this random variable conditional on the observed value of the covariate vector $\mathbf{U}_i$:

$$\text{CATE}(\mathbf{u}) := E\{(Y_{i,(t)} - Y_{i,(c)}) | \mathbf{U}_i = \mathbf{u}\}.$$

Note that $\Delta(\mathbf{u}) = E(Y_i | T_i = t, \mathbf{U}_i = \mathbf{u}) - E(Y_i | T_i = c, \mathbf{U}_i = \mathbf{u})$. We sometimes call Δ the Conditional Average Treatment Difference (CATD) to distinguish it from the CATE because Δ may be influenced by unobserved confounding variables and therefore not accurately represent the conditional effect of the treatment. However, for simplicity we often refer to Δ as the "treatment effect". Again, it is important to keep in mind that in order for $\Delta(\mathbf{u})$ to represent a causal effect of the treatment variable T, we need to assume that (1.4) holds.

The data consist of $(Y_i, T_i, \mathbf{U}_i)_{i=1}^n$, where Y_i is the response, $T_i \in \{t, c\}$ is a binary treatment assignment, and $\mathbf{U}_i$ represents a collection of p baseline covariates observed before the treatment is applied. We assume the covariates $\mathbf{U}_i$ to be i.i.d. from an unknown probability distribution $P_{\mathbf{U}}$ with compact support $\mathcal{U} \subset \mathbb{R}^p$. We further assume that within each treatment group, the covariate values are independent and identically distributed; that is, $\mathbf{U}_i|T_i = t$ are i.i.d. with distribution $P_{\mathbf{U}_t}$, and $\mathbf{U}_i|T_i = c$ are i.i.d. with distribution $P_{\mathbf{U}_c}$. In order for Δ to be identifiable on $\mathcal{U}$, it is necessary for the densities of $P_{\mathbf{U}_t}$ and $P_{\mathbf{U}_c}$ each to be nonzero on $\mathcal{U}$.

In this chapter, the errors $\{\xi_i\}_{i=1}^n$ and $\{\nu_i\}_{i=1}^n$ are assumed to follow standard normal distributions, with $\xi_i \perp\!\!\!\perp \xi_j$ and $\nu_i \perp\!\!\!\perp \nu_j$ for any $i \neq j$. We assume each set of errors is i.i.d. and independent of $\mathbf{U}$. The error variances σ_t^2 under treatment and σ_c^2 control are assumed to be homoscedastic with respect to the covariates $\mathbf{U}$. However, we do allow the response variances under treatment and control groups to differ. The normality and homoscedasticity assumptions are made to simplify the presentation and can be relaxed without affecting the risk bound.

We define the L_2 norm with respect to the probability distribution of the covariates:

$$\|f\|_2 := \left\{ \int |f(\mathbf{u})|^2 P_{\mathbf{U}}(d\mathbf{u}) \right\}^{1/2},$$

where $P_{\mathbf{U}}$ denotes the probability distribution of $\mathbf{U}_i$ for $1 \le i \le n$. This norm will be used to evaluate the average discrepancy between Δ and various estimates $\widehat{\Delta}$ over $\mathcal{U}$.

The TEEM method involves combining a finite collection of regression procedures proposed for estimating the treatment effect Δ. Here a regression procedure or strategy, say ψ, refers to a method of estimating Δ and σ based on $\mathbf{Z}^m = (Y_i, T_i, \mathbf{U}_i)_{i=1}^m$ at each sample size m. Here, ψ could be any sort of statistical regression model or machine learning algorithm that produces an estimate of $\Delta(\mathbf{u})$ for any $\mathbf{u} \in \mathcal{U}$. Let ψ_j, $1 \le j \le J$, denote the proposed treatment effect estimation procedures, and let $\widehat{\Delta}_{m,j}(\mathbf{u})$ and $\hat{\sigma}_{m,j}$ denote the estimators of Δ and σ, respectively, resulting from the application of procedure ψ_j to the data $\mathbf{Z}^m$.

5.3 The TEEM Algorithm

The TEEM algorithm for combining estimates of conditional treatment effects is based on data splitting (as in cross-validation). The candidate procedures are fit to a "training" subset of the data and evaluated on the remaining subset. Since each observational unit is in either the treatment or control group, individual treatment effects are not available to evaluate the procedures. Our solution to this problem is to approximate individual treatment effects in the evaluation data by using pairs of nearby observations, one from each treatment group.

Suppose we have a pair of observations (i, j) such that $T_i = t$ and $T_j = c$. If individuals i and j have the same baseline covariates ($\mathbf{U}_i = \mathbf{U}_j$), then within the framework of the previous section, $Y_i - Y_j$ is an observation from $N(\Delta(\mathbf{U}_i), \sigma_t^2 + \sigma_c^2)$ and this difference can be used to evaluate the accuracy of estimates $\widehat{\Delta}(\mathbf{U}_i)$. If the covariates of i and j do not match exactly but $d(\mathbf{U}_i, \mathbf{U}_j)$ is small with respect to some distance measure $d(\cdot)$, then $Y_i - Y_j \sim N(\Delta(\mathbf{U}_i) + (f_c(\mathbf{U}_i) - f_c(\mathbf{U}_j)), \sigma_t^2 + \sigma_c^2)$, and the bias for $Y_i - Y_j$ as an estimate of $\Delta(\mathbf{U}_i)$ is represented by $f_c(\mathbf{U}_i) - f_c(\mathbf{U}_j)$. If the distance between $\mathbf{U}_i$ and $\mathbf{U}_j$ is small and the control regression function f_c is smooth, this bias will be small and the paired difference $Y_i - Y_j$ will be a nearly unbiased estimate of $\Delta(\mathbf{U}_i)$. The TEEM algorithm uses these differences between treatment/control pairs to evaluate the candidate estimates of treatment effects and assign to them appropriate weights.

In this section we present two versions of the TEEM algorithm: one version in which each individual response Y_i is used in at most one treatment-control pair, and a second version in which each observation i is paired with its nearest neighbor i^* in the other treatment group and observations are allowed to belong to more than one pair. The two versions of the algorithm are similar; the main difference between them can be thought of as the difference between matching without replacement and matching with replacement. The first version is used for theoretical development because the method we use to bound the combined estimator's risk requires that the treatment-control differences used to evaluate the models be independent of each other. The second version may perform better in applications, because as argued in Abadie and Imbens (2006), matching with replacement will produce higher-quality (closer) matches and therefore introduce less bias.

5.3.1 TEEM with Independent Pairs

Here we describe in detail the version of the TEEM algorithm with independent pairs for which we derive the risk bound in Section 5.3.2.

Step 0. Select a fraction $\rho \in (0, 1)$ of the n observations that will be used to fit the models. Denote $\lfloor \rho n + 0.5 \rfloor$ by n_1; n_1 is the number of observations used to fit the models. Similarly denote the size of the evaluation set, $\lceil (1-\rho)n - 0.5 \rceil$, by n_2. Note that asymptotically, n_1 and n_2 are both of order n.

Step 1. Randomly permute the order of the n observations; call this permutation π. Split the resulting ordered data into two parts: the training part $\mathbf{Z}^{(1)} = (Y_i, T_i, \mathbf{U}_i)_{i=1}^{n_1}$ and the evaluation part $\mathbf{Z}^{(2)} = (Y_i, T_i, \mathbf{U}_i)_{i=n_1+1}^{n}$.

Step 2. Within the evaluation data $\mathbf{Z}^{(2)}$, let n_{t_2} denote the number of observations for which $T_i = t$ and n_{c_2} the number for which $T_i = c$. Let $n_2^* = \min(n_{t_2}, n_{c_2})$. Partition $\mathcal{U} = [0, 1]^p$ into hypercubes each with side length h such that

$$\frac{1}{h} = \left\lfloor \left(\frac{\underline{c} n_2^*}{2 \log n_2^*} \right)^{1/p} \right\rfloor .$$

Let $\widetilde{n}_2$ denote the number of these hypercubes containing at least one realized covariate value from each treatment group in $\mathbf{Z}^{(2)}$. Within each of these $\widetilde{n}_2$ cells (which we index by m), randomly select a pair of observations (i_m, i_m^*) such that $T_{i_m} = t$ and $T_{i_m^*} = c$.

Step 3. For each resulting matched pair (i_m, i_m^*), create approximate treatment effects $\widetilde{\delta}_m = Y_{i_m} - Y_{i_m^*}$. These approximate "individual" treatment effects will be used to evaluate the candidate procedures and assign them weights.

Step 4. Fit the J candidate models (or generally, the J candidate estimation procedures) $\psi_1, \ldots, \psi_J$ to the data $\mathbf{Z}^{(1)}$ to obtain J estimates $\widehat{\Delta}_{n_1,1}, \ldots, \widehat{\Delta}_{n_1,J}$ of the treatment effect function. Similarly, apply the J candidate estimation procedures to $\mathbf{Z}^{(1)}$ to estimate $\sigma := \sqrt{\sigma_t^2 + \sigma_c^2}$. Estimates of σ can be shared between procedures if desired. Denote these estimates as $\hat{\sigma}_{n_1,1}, \ldots, \hat{\sigma}_{n_1,J}$.

Step 5. For each procedure indexed by $j = 1, 2, \ldots, J$, assign initial weights (or prior probabilities) $W_{1,j} = \omega_j$, where the ω_j's are positive numbers that sum to 1. Then for $2 \leq m \leq \widetilde{n}_2$, let

$$W_{m,j} = \frac{\omega_j \prod_{l=1}^{m-1} \phi\left(\left(\widetilde{\delta}_{l+1} - \widehat{\Delta}_{n_1,j}(\mathbf{U}_{i_{l+1}})\right)/\hat{\sigma}_{n_1,j}\right)/\hat{\sigma}_{n_1,j}}{\sum_{k=1}^{K} \omega_k \prod_{l=1}^{m-1} \phi\left(\left(\widetilde{\delta}_{l+1} - \widehat{\Delta}_{n_1,k}(\mathbf{U}_{i_{l+1}})\right)/\hat{\sigma}_{n_1,k}\right)/\hat{\sigma}_{n_1,k}},$$

where ϕ is the standard normal density function. Note that $\sum_{j \geq 1} W_{m,j} = 1$ for each $m = 1, \ldots, \widetilde{n}_2$.

Step 6. For $m = 1, \ldots, \widetilde{n}_2$, let

$$\widetilde{\Delta}_m(\mathbf{u}) = \sum_{j=1}^{J} W_{m,j} \widehat{\Delta}_{n_1,j}(\mathbf{u}).$$

Step 7. For every cell m containing at least one treatment-control pair, let $\mathcal{U}_m$ denote the region of the covariate space representing the cell. Then let

$$\widetilde{\widetilde{\Delta}}_\pi(\mathbf{u}) = \begin{cases} \widetilde{\Delta}_m(\mathbf{U}_{i_m}) & \text{if } \mathbf{u} \in \mathcal{U}_m \\ 0 & \text{if the cell containing } \mathbf{u} \text{ has no treatment-control pair in } \mathbf{Z}^{(2)}. \end{cases}$$

The subscript π indicates the estimator's dependence on the permutation π applied in Step 1.

Step 8. Repeat Steps 1-7 a total of P times for some $P \geq 1$, and average the resulting $\widetilde{\widetilde{\Delta}}_\pi$ to obtain the TEEM estimator

$$\overline{\widehat{\Delta}}(\mathbf{u}) = \frac{1}{P} \sum_{p=1}^{P} \widetilde{\widetilde{\Delta}}_{\pi_p}(\mathbf{u}),$$

where for each iteration $1 \leq p \leq P$, π_p denotes the permutation applied in Step 1 of the iteration.

5.3.2 Risk Bound for the TEEM Estimator

In this section we bound the risk of the estimator produced by the TEEM algorithm. Our proof uses the following assumptions on the data-generating process:

Regularity Conditions

(a) Boundedness: The regression functions f_t and f_c are uniformly bounded in absolute value by $A < \infty$, and the standard deviations σ_t and σ_c each are bounded above and below by $\overline{\sigma} < \infty$ and $\underline{\sigma} > 0$, respectively. We assume correspondingly that the estimators $\widehat{\Delta}_{l,j}$ and $\hat{\sigma}_{l,j}$ satisfy $\|\widehat{\Delta}_{l,j}\|_\infty \leq 2A$ and $\hat{\sigma}_{l,j} \in [\sqrt{2}\underline{\sigma}, \sqrt{2}\overline{\sigma}]$, for each $l \geq 1$ and $j \geq 1$. In addition we assume the densities of the distributions $P_{\mathbf{U}_t}$ and $P_{\mathbf{U}_c}$ each are bounded above and below by $\bar{c} < \infty$ and $\underline{c} > 0$ on $\mathcal{U}$.

(b) Smoothness: The regression functions for the treatment and control groups, f_t and f_c, and the estimators $\widehat{\Delta}_{l,j}$ for $l \geq 1$ and $j \geq 1$ have all p first-order partial derivatives, and each of these first-order partial derivatives is upper bounded in absolute value by a constant L on $\mathcal{U}$. We also assume the densities of the distributions $P_{\mathbf{U}_t}$ and $P_{\mathbf{U}_c}$ are continuous on $\mathcal{U}$.

(c) Asymptotic Order of Treatment and Control Groups: For n large enough, there exist constants (a, b) not depending on n such that $0 < a < n_t/n < b < 1$, where n_t is the number of the n observations for which $T_i = t$.

The theorem below bounds the risk of the TEEM estimator in terms of the risks of the individual procedures, the size of the evaluation set, and the dimension of the covariate vector. A detailed proof of this theorem can be found Appendix B.

Theorem 2 *Under the above regularity conditions, the risk of $\overline{\widehat{\Delta}}$ from the TEEM algorithm with independent pairs under the L_2 loss has the following bound:*

$$
\begin{aligned}
&E\|\Delta - \overline{\widehat{\Delta}}\|_2^2 \\
&\quad \leq C\left(\left(\frac{\log n_2}{n_2}\right)^{1/p} + \inf_j \left[\left(\frac{\log n_2}{n_2}\right) \log \frac{1}{\omega_j} + E(\sigma - \hat{\sigma}_{n_1,j})^2 + E\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2 \right] \right),
\end{aligned}
$$

where the constant C depends on a, b, $\underline{c}$, $\bar{c}$, $\underline{\sigma}$, $\overline{\sigma}$, A, p, and L (but not on n).

Remarks

(a) Since estimation of σ can usually be done at the parametric rate (see Remark (d)), the above oracle inequality says that the combined estimator of Δ converges at the best rate offered by the candidate procedures, up to $(\log n_2/n_2)^{1/p}$, which is the "curse of dimensionality" in pairing the individuals.

(b) By choosing a fixed fraction of n to fit the estimators and using the remainder to construct the combining weights, n_1 and n_2 both are $O(n)$ asymptotically. Therefore, if one of the candidate models (say j^*) is a correctly specified parametric representation of the data-generating process, then $E(\sigma - \hat{\sigma}_{n_1,j^*})^2$ and $E\|\Delta - \widehat{\Delta}_{n_1,j^*}\|_2^2$ each will converge to zero at a rate of n^{-1}. In this case, if $p = 1$, the risk of the combined estimator will converge to zero at rate $(\log n)n^{-1}$, almost as fast as an oracle that knows the true model in advance.

(c) The constant p representing the number of continuous covariates does affect the convergence rate of the risk of the combined estimator, as expected. It is helpful (in both theory and practice) to reduce the dimension of the covariate vector before applying the TEEM algorithm. Ideally, the dimension reduction would not result in any loss of information about Δ. Such dimension reduction often can be done using variable selection techniques or by finding a few linear combinations of the covariates that are sufficient for the regression of Δ on $\mathbf{U}$. (See Cook (1998) for an overview of sufficient dimension reduction.) Practically, we have found that TEEM features good performance for accurate estimation of the treatment effect in a variety of simulation settings and for different values of p.

(d) In the setting we have assumed, with homoscedastic errors within the treatment and control groups and smoothness conditions on f_t and f_c, the variance terms σ_t^2 and σ_c^2 (and therefore σ^2) can be estimated at rate n_1^{-1} independently of the candidate models (see, e.g., Rice, 1984). Thus, by modifying the algorithm the term $E(\sigma - \hat{\sigma}_{n_1,j})^2$ could be removed from the risk bound. However, we believe that in practice, model-based estimators of σ often are helpful in assigning proper weights to each of the candidate procedures.

(e) Theorem 2 can be generalized to handle heteroscedastic errors and/or non-Gaussian errors. Our proof assumes homoscedastic Gaussian errors only for simplicity of presentation.

5.3.3 Modifications to TEEM for Applications

In this section, we describe a modified version of the TEEM algorithm for use in applications. This algorithm is computationally simpler than the algorithm in 5.3.1, and we believe it makes more efficient use of the available data. Because of the techniques used to establish the risk bound in Section 5.3.2, the algorithm in 5.3.1 is presented for theoretical development. The two algorithms are fundamentally similar: each is based on data splitting, pairing of nearby treatment and control observations, and the use of a likelihood based on these pairings to compute the combination weights. The differences between the two algorithms involve how the n data points in the sample are utilized in the different steps.

In the TEEM algorithm for applications, we create the treatment-control pairs from the evaluation data by simply searching for the nearest neighbor of each observation in the other treatment group, rather than by creating a partition of the covariate space. This modification allows each observation in the evaluation data to possibly belong to more than one treatment-control pair. Because of this, the treatment-control pairs used are no longer independent, but the pairs are closer with respect to their covariate values. The combination weights in this section are not constructed sequentially, as in Step 5 of the algorithm in 5.3.1; instead, they are based on the entire evaluation data, independent of the ordering of observations. Finally, the TEEM estimator is based on a weighted average of $\widehat{\Delta}_{n,j}$, rather than the weighted average of $\widehat{\Delta}_{n_1,j}$ in Step 6 of the algorithm in 5.3.1. In the algorithm described in this section, only the weights are averaged over the different data splittings. These averaged weights are then used to combine the candidate estimates $\widehat{\Delta}_{n,j}$ built on the full data.

Steps 0A and 1A. These steps are the same as Steps 0 and 1 of the algorithm in 5.3.1.

Step 2A. For each unit i in $\mathbf{Z}^{(2)}$ (regardless of its treatment status), let i^* denote its nearest neighbor from the other control group. Specifically, i^* represents the observation in $\mathbf{Z}^{(2)}$ such that $T_i \neq T_{i^*}$ and

$$\sum_{k=n_1+1}^{n} I(T_i \neq T_k) I(d(\mathbf{U}_i, \mathbf{U}_k) \leq d(\mathbf{U}_i, \mathbf{U}_{i^*})) = 1.$$

For each unit i, we will be able to find its nearest neighbor in the other control group. (If in practice the covariates $\mathbf{U}$ are discrete, there may be multiple i^* equidistant from i. Any tiebreaking method may be used to choose one i^* in this case.) Therefore, the number of treatment-control pairs (i, i^*) will be n_2, but the pairs may not be independent. In some situations, it may be useful to apply a caliper to bound the matching discrepancy so that we do not end up matching pairs that are too far apart. Althauser and Rubin (1970) and others have argued that caliper matching can remove a large percentage of the total bias induced by matching while only removing a small percentage of the matched pairs. For TEEM, if a caliper is applied, observations without a "close enough" match in the other treatment group are not used in the evaluation of the candidate estimation procedures.

Step 3A. For each of the n_2 treatment-control pairs produced by Step 2A (or the $\widetilde{n}_2$ pairs for some $\widetilde{n}_2 \leq n_2$ if a caliper is applied), create approximate treatment effects

$$\widetilde{\delta}_i = (2I(T_i = t) - 1)(Y_i - Y_{i^*}).$$

In other words, for each pair (i, i^*), $\widetilde{\delta}_i$ is the response of the treated unit minus the response of the control unit.

Step 4A. Same as Step 4 of the algorithm in 5.3.1.

Step 5A. Use the performance of the estimates $\widehat{\Delta}_{n_1,j}$ on the evaluation data to create the weights $W_{\pi,j}$, where the subscript π represents the permutation applied in Step 1A.

$$W_{\pi,j} = \frac{\prod_{i=n_1+1}^{n} \phi\left(\left(\widetilde{\delta}_i - \widehat{\Delta}_{n_1,j}(\mathbf{U}_i)\right)/\hat{\sigma}_{n_1,j}\right)/\hat{\sigma}_{n_1,j}}{\sum_{k=1}^{K}\prod_{i=n_1+1}^{n} \phi\left(\left(\widetilde{\delta}_i - \widehat{\Delta}_{n_1,k}(\mathbf{U}_i)\right)/\hat{\sigma}_{n_1,k}\right)/\hat{\sigma}_{n_1,k}},$$

where ϕ is the standard normal density function.

Step 6A. Repeat Steps 1A-5A a total of P times for some $P \geq 1$, and for each j average the weights $W_{\pi,j}$ from each permuation π. Call these averaged weights $\overline{W}_j$.

$$\overline{W}_j = \frac{1}{P}\sum_{p=1}^{P} W_{\pi_p,j},$$

where for each iteration $1 \leq p \leq P$, π_p denotes the permutation applied in Step 1A of the iteration.

Step 7A. Fit each candidate procedure j to the entire sample of n observations to obtain estimates $\widehat{\Delta}_{n,j}(\mathbf{u})$ for any $\mathbf{u} \in \mathcal{U}$.

Step 8A. Create the final TEEM estimator

$$\overline{\widehat{\Delta}}_A(\mathbf{u}) = \sum_{j=1}^{J} \overline{W}_j \widehat{\Delta}_{n,j}(\mathbf{u}).$$

Observe that $\overline{\widehat{\Delta}}_A$ is a weighted average of the estimates $\widehat{\Delta}_{n,j}$ created in Step 7A, where the weights are the $\overline{W}_j$ created in Step 6A.

We believe the TEEM estimator $\overline{\widehat{\Delta}}_A$ will typically exhibit better performance than the estimator $\overline{\widehat{\Delta}}$ in Section 5.3.1 because it makes more efficient use of the data, but for technical reasons we derive the risk bound of TEEM using the algorithm in Section 5.3.1. The numerical analyses in Sections 5.4 and 6.2 of this thesis use the version of TEEM described in the current section, and generally we recommend this version of TEEM for applications.

5.4 Simulation Study

5.4.1 Data-Generating Process and Candidate Models

In this section, we compare a variety of model selection and combination methods for estimation of the treatment effect in a situation where all the candidate models are misspecified. Additional numerical results comparing TEEM with other selection and combination methods are presented in Sections 6.2.3 and 6.2.4. In each of the guided simulation settings in Section 6.2.4, one of the candidate models *is* correctly specified. In settings with and without model misspecification, we see that TEEM performs well relative to other methods.

We generate data from the following process:

$$Y_i = 0.5U_{i,1}^2 + 0.5U_{i_2} + I(T_i = t) * (0.5U_{i,1} + 0.5U_{i,2}^2) + \varepsilon_i, \tag{5.2}$$

where $(U_{i,1}, U_{i,2}, \varepsilon_i)$ are i.i.d. $N(\mathbf{0}, \mathbf{I}_3)$ and the T_i are i.i.d. with $P(T_i = t) = 0.5$. The nine candidate models are enumerated in Table 5.1. These models are hierarchical in the sense that if a treatment-covariate interaction is included in the model, the main effect of that covariate also is included.

Table 5.1: Candidate models for simulation study in Section 5.4

Model Number	Model Terms
1	$T, U_1, U_2, T : U_1, T : U_2$
2	$T, U_1, U_2, T : U_1$
3	$T, U_1, U_2, T : U_2$
4	T, U_1, U_2
5	$T, U_1, T : U_1$
6	T, U_1
7	$T, U_2, T : U_2$
8	T, U_2
9	T

We create 100 realizations of (5.2) at each of two sample sizes: $n = 100$ and $n = 300$. For each realization, we use various model selection and combination methods to choose a model/combination and use the chosen model/combination to estimate Δ. The squared L_2 risks (for Δ) for each candidate model and for each selection/combination method at each sample size are estimated by averaging the risks over the 100 realizations, where each realization-risk is estimated from the sample mean of $(\Delta(\mathbf{U}_i) - \widehat{\Delta}(\mathbf{U}_i))^2$ based on an independent evaluation data set of 1 million independent draws from the distribution of $(U_{i,1}, U_{i,2})$.

5.4.2 Model Selection and Combination Methods Considered

We compare the risks of the nine candidate models in Table 5.1 with the risks of the five model selection methods described in Section 4.4 and five model combination methods. The first three combination methods form convex combinations of the candidate procedures based on the value of some information criterion. That is, each produces a $\widehat{\Delta}$ function by

$$\widehat{\Delta}(\mathbf{u}) = \sum_{j=1}^{J} w_j \widehat{\Delta}_{n,j}(\mathbf{u}),$$

where $J = 9$ in this case, and $\widehat{\Delta}_{n,j}$ is the estimate of Δ produced by applying procedure j to the entire sample of size n.

Perhaps the most common method of model combination is Bayesian Model Averaging (BMA), in which each w_j represents the posterior probability of model j. Raftery (1995) derived the approximation

$$w_j = \frac{\exp(-\frac{1}{2}\mathrm{BIC}_j)}{\sum_{j=1}^{J} \exp(-\frac{1}{2}\mathrm{BIC}_j)}$$

to the posterior probability of model j when the models have equal prior probabilities. Buckland et al. (1997) suggested combining models based on AIC by replacing BIC_j with AIC_j in the above expression for w_j. We refer to these two weighting schemes as BMA and cAIC, respectively.

Claeskens and Hjort (2008a) suggest combining estimators of a focus parameter using a similar weighting scheme,

$$w_j = \frac{\exp(-\frac{1}{2}\lambda \mathrm{wFIC}_j)}{\sum_{j=1}^{J} \exp(-\frac{1}{2}\lambda \mathrm{wFIC}_j)},$$

where λ is a tuning parameter. It is unclear how λ is to be selected; in this simulation we try values of $\lambda = 1$ and $\lambda = 0.1$. We denote this method as cwFIC (combination based on wFIC).

Adaptive Regression by Mixing (Yang, 2001), or ARM, is (like TEEM) a method of model combination based on data splitting, but it targets estimation of the response. Essentially, each model's weight is based on its ability to predict the response on outside data. We construct the ARM weights w_j by assuming normal errors, using each procedure separately to estimate the error variance, and averaging the weights over 100 different 50/50 data splittings. We implement TEEM using the same 100 50/50 data splittings. The version of TEEM described in Section 5.3.3 (sampling with replacement) is used. No caliper is applied, so all observations in the evaluation set are simply matched to their nearest neighbor (with respect to Euclidean distance among the covariates) in the other treatment group.

Table 5.2: Simulation results: 9 misspecified candidates, $n = 100$

	Model/Method	Estimated Risk of $\widehat{\Delta}$ (SE)
Candidate Models	Model 2	0.7121 (0.0173)
	Model 5	0.7323 (0.0192)
	Model 4	0.8240 (0.0086)
	Model 6	0.8268 (0.0105)
	Model 8	0.8311 (0.0086)
	Model 9	0.8327 (0.0096)
	Model 1	0.8347 (0.0251)
	Model 3	0.9613 (0.0236)
	Model 7	0.9705 (0.0227)
Model Selection Methods	TECV	0.8335 (0.0174)
	AIC	0.8475 (0.0248)
	BIC	0.8560 (0.0209)
	CV	0.8577 (0.0205)
	wFIC	0.8597 (0.0232)
Model Combination Methods	TEEM	0.7142 (0.0150)
	ARM	0.7325 (0.0156)
	cwFIC ($\lambda = 1/10$)	0.7790 (0.0182)
	BMA	0.7897 (0.0178)
	cAIC	0.8045 (0.0212)
	cwFIC ($\lambda = 1$)	0.8443 (0.0232)

5.4.3 Results

In this setting, there is a conflict between the goals of estimating the full regression function and estimating the treatment effect. For example, Model 5 is a relatively effective model for treatment effect estimation, because it contains the linear interaction term between T and U_1, but it is not effective for estimating the full regression function because it omits the main effect for U_2. Because of this conflict, we should expect that the selection and combination methods targeted toward Δ will perform better than the methods targeted toward the full regression function. Because all of the candidate models are of the same parametric family, we can use wFIC to compare them. However,

Table 5.3: Simulation results: 9 misspecified candidates, $n = 300$

	Model/Method	Estimated Risk of $\widehat{\Delta}$ (SE)
Candidate Models	Model 2	0.5866 (0.0087)
	Model 5	0.5877 (0.0087)
	Model 1	0.6314 (0.0104)
	Model 4	0.7791 (0.0035)
	Model 6	0.7802 (0.0035)
	Model 8	0.7811 (0.0041)
	Model 9	0.7813 (0.0039)
	Model 3	0.8246 (0.0072)
	Model 7	0.8285 (0.0075)
Model Selection Methods	TECV	0.6220 (0.0117)
	AIC	0.6292 (0.0119)
	wFIC	0.6328 (0.0121)
	BIC	0.6561 (0.0134)
	CV	0.6818 (0.0138)
Model Combination Methods	TEEM	0.6074 (0.0089)
	cAIC	0.6272 (0.0109)
	cwFIC ($\lambda = 1/10$)	0.6293 (0.0105)
	cwFIC ($\lambda = 1$)	0.6316 (0.0118)
	ARM	0.6355 (0.0094)
	BMA	0.6444 (0.0112)

the true model is not contained in the candidate set; this violates one of the assumptions of wFIC and may lead to its poor performance in this case.

Tables 5.2 and 5.3 show the risks of the model selection and combination methods, as well as the risks of the individual models, at $n = 100$ and $n = 300$, respectively. Among the model selection methods, TECV performs the best at both sample size levels. Its performance is much better than that of traditional CV in both cases.

The method of TEEM proposed in this chapter features the lowest risk among all 11 methods of selection and combination methods at both sample size levels. At $n = 100$, TEEM results in much better performance than any of the model selection methods due to the high model selection instability at this sample size. At $n = 300$, TEEM is approximately two standard errors better than the second-place combination method, cAIC.

5.5 Conclusion

When model selection uncertainty is high for estimating the conditional effect of a treatment given covariates, combining estimates from the different candidate procedures often results in an estimator for the treatment effect that is more accurate than one from a single selected candidate. While some interpretability is lost when combining many models rather than selecting one, the increase in accuracy is often substantial enough to justify the trade-off. In this chapter, we propose the TEEM model combination algorithm that relies on approximating treatment effects by pairing each observation with a nearby neighbor in the other treatment group. Our oracle inequality for the TEEM estimator under squared error loss implies that the combined estimator will converge to the true Δ as $n \to \infty$ as long as at least one of the candidate estimators converges to Δ. The convergence will be slower when there are many covariates due to the "curse of dimensionality" in pairing the individuals.

A simulation study in Section 5.4 and an analysis of the benchmark LaLonde data in Section 6.2 suggest that TEEM compares favorably with other selection and combination methods in providing an accurate estimate of the treatment effect conditional on covariates. There is much theoretical and empirical evidence in the literature to support model combination when the goal is accurate estimation or prediction of a response. The goal of accurate treatment effect estimation is fundamentally different in some ways and may result in different candidate models needing to receive higher weights; models that are good for prediction may not be good for treatment effect estimation. A properly targeted model combination method may enjoy important advantages over model selection in this setting.

Chapter 6

Real Data Applications

6.1 Application: FIRST Clinical Trial

6.1.1 Background

In this section we apply the five model selection methods considered in the simulations in Section 4.4 to a dataset from the Community Programs for Clinical Research on AIDS (CPCRA). (Model combination methods are not used in this analysis, but they are used in Section 6.2.) The data contain results from a clinical trial known as the FIRST (Flexible Initial Retrovirus Suppressive Therapies) trial. The purpose of the FIRST trial was to evaluate different treatment strategies for HIV-positive patients. Patients were assigned to either a treatment strategy combining three classes of drugs or to one of two different two-class strategies. The primary medical publication to analyze these data was MacArthur et al. (2006). Their main analysis compared the average change in CD4 cell count under the three-class treatment strategy to the average change under either of the two-class strategies. CD4 cell counts are often used to assess the strength of an HIV-positive patient's immune system.

Using p-values from linear regression, MacArthur et al. (2006) analyzed data from 1,196 patients and concluded that there was no significant difference between the three-class and two-class strategies with respect to the average change in CD4 cell count. The authors also checked for interactions between the treatment and pre-specified subgroups, including subgroups based on age ($<$ 40 vs. 40+), baseline CD4 cell count ($\leq$ 200 vs.

> 200), and baseline HIV RNA concentration ($< 100,000$ copies per mL vs. $100,000+$). No statistically significant interaction terms appeared among any of the subgroups, so the authors concluded that there was no evidence that the treatment strategy made a difference in CD4 count for any of the pre-specified subgroups of the HIV-positive population. A separate analysis found that toxic effects led to treatment discontinuation more often for the three-class strategy than for the two-class strategies. Because the more potent treatment strategy was associated with more frequent toxic effects and did not appear to provide a greater increase in CD4 count than the less potent strategies, the authors recommended that either of the two-class treatment strategies be used.

The data used in our analysis differs in some minor details from the data used in MacArthur et al. (2006). When analyzing CD4 cell counts, the square root transformation is commonly used to stabilize variance and improve the approximation of normality (see, e.g., Buclin et al., 2011). In our analysis, we apply the transformation, although this was not done in MacArthur et al. (2006). Specifically, the response variable Y in our analysis is the difference between the square root of the patient's average CD4 cell count over all measurements taken at or after 32 months from enrollment and the square root of the patient's CD4 cell count at the baseline enrollment date. For our baseline covariate vector $\mathbf{U}$, we consider three variables: the square root of the baseline CD4 cell count ($\mathrm{CD4}_0$), the log of the baseline HIV RNA concentration (RNA_0), and the age of the patient (Age). The three-class vs. two-class comparison in MacArthur et al. (2006) included 1,196 patients; we removed five observations with missing values of RNA_0 to arrive at $n =$ 1,191. The treatment variable was set to $T = t$ for the more potent three-class strategy (assigned to 392 patients) and $T = c$ for either of the two-class strategies (799 patients).

It is clear that the initial CD4 count has a strong relationship with the change in CD4 count, so $\mathrm{CD4}_0$ is included in all candidate models; this variable is "protected" in the language of FIC. In addition to $\mathrm{CD4}_0$, we consider models with different combinations of T, RNA_0, Age, and two-way treatment-covariate interactions. We allow interaction terms to be considered only in models for which both main effects are present. These guidelines lead to 22 distinct variable subsets. For each subset, a linear model and an additive model are considered (see Section 2.3 for a discussion of interactions in additive models), bringing the total number of candidate models to 44.

The method of wFIC is not suitable to compare linear models to additive models or additive models with different nonlinear terms to each other; therefore, wFIC selects among only the 22 linear models in our analysis. The likelihood-based criteria of AIC and BIC can be used to compare additive models to linear models because the additive model can be represented as a penalized likelihood, where the penalty is a measure of the wiggliness (roughness) of the function. TECV and traditional cross-validation also can compare different model types because of their inherent flexibility. Eight of the 44 candidate models do not include T or any treatment-covariate interactions, so these models are equivalent with respect to estimation of the treatment effect, implying that $\Delta(\mathbf{u}) = 0$ for all $\mathbf{u}$. These models all will have equal wFIC and TECV statistics; however, because the models differ in their main effects, they will have different values of AIC, BIC, and traditional CV.

Table 6.1: Models chosen for FIRST trial data analysis

			$\widehat{\Delta}(\mathbf{U}_i)$ Values[b]	
Method	Model Type	Active Variables[a]	Mean	SD
AIC	Additive	T, CD4_0, *Age*, RNA_0, T:CD4_0, T:Age	0.20	0.90
BIC	Linear	CD4_0, Age	0	0
CV	Additive	CD4_0, Age, RNA_0	0	0
wFIC	Linear	T, CD4_0, Age, T:CD4_0, T:Age	0.25	0.85
TECV	Linear	No treatment effects[c]	0	0

[a] The presence of interaction terms implies the presence of both main effects.

[b] The mean and standard deviation of $\widehat{\Delta}(\mathbf{U}_i)$ over the $n = 722$ values of $\mathbf{U}_i$.

[c] The TECV statistic for the eight models suggesting no treatment effect was lower than the TECV of each of the other candidate models.

6.1.2 Results

Table 6.5 shows the model that was selected by each model selection method. AIC chooses a model in which Δ is a nonlinear function of CD4_0 and Age. wFIC selects a linear model with the same treatment-covariate interaction terms selected by AIC. The models selected by AIC and wFIC suggest that the average effect of the three-drug treatment relative to the two-drug treatment is positive but small relative to the

Table 6.2: Linear model selected by wFIC for FIRST trial data

Term	Estimate	P-value
Intercept	10.30	< 0.01
T	3.99	0.02
$\mathrm{CD4}_0$	-0.46	< 0.01
Age	0.07	< 0.01
$\mathrm{T{:}CD4}_0$	-0.09	0.06
T:Age	-0.07	0.08

variability of the treatment effect among patients. According to these models, the three-drug treatment (relative to the two-drug treatment) would be expected to increase the CD4 cell count of some patients and decrease the count of others. Meanwhile, BIC and both forms of CV select models with no treatment effects. The TECV statistic is identical for all models with no treatment effects, and this statistic is lower than the TECV statistic for any of the 36 models with one or more treatment effects. Thus, the BIC, CV, and TECV models thus agree with the conclusions of MacArthur et al. (2006) that the treatment strategies are equal in their effect on CD4 cell counts of patients with HIV.

A summary of the model selected by wFIC is shown in Table 6.2. This model indicates the effect of the three-drug treatment on the response decreases as baseline CD4 and age increase. It makes some sense that the more potent treatment would be more beneficial to those with weaker initial immune systems and those who are younger. Notice that the three terms involving the treatment variable T have associated p-values between 0.02 and 0.08. With p-values in this range, it is not surprising that different model selection methods disagree about the effect of the treatment variable on the regression function.

The following section describes a guided simulation that gives further insight into how each method performs under three scenarios consistent with the data observed in the FIRST clinical trial. The method of wFIC seems to adapt well to each scenario, correctly identifying the presence (or absence) of a treatment effect over 75 percent of the time under each scenario. However, given the increased risk of toxic effects for those taking the three-drug treatment, a more conservative model such as the one suggested by BIC, CV, TECV, and the hypothesis tests of the original *Lancet* paper would perhaps be more appropriate for guiding treatment in this setting.

6.1.3 Cross-examination of FIRST Trial Data

The previous section illustrated the use of five model selection methods to estimate the treatment effect function in the FIRST clinical trial. The five model selection methods provide three distinct estimates of the treatment effect function Δ. BIC, traditional CV, and TECV all suggest that the value of the treatment variable T indicating a two-drug or three-drug combination does not affect the mean of the response regardless of the values of the baseline covariates. In the language of this thesis, BIC, traditional CV, and TECV all are choosing a model for which $\Delta(\mathbf{u}) = 0$ for all $\mathbf{u}$. However, the global focused information criterion (wFIC) for Δ selects a model with a treatment main effect and two linear interaction terms, while AIC chooses an additive model with a treatment main effect and two nonlinear interaction terms. The methods are giving quite different answers, and which answer is the best for guiding treatment is unknown.

To gain further insight, in this section we analyze a guided simulation experiment that tests how each method performs under scenarios in which the data are simulated from models selected by other methods. Li et al. (2000) used this scheme in a different context and called it "cross-examination". For each model selection method M, we generate a response vector Y^*_M by adding noise to the estimated regression functions (from the model selected by M) at the original values of $(T_i, \mathbf{U}_i)$. The noise is generated by a mean-zero Gaussian distribution with variance equal to the variance estimate from the model selected by M. For each M, all five model selection methods are then applied to $(Y^*_{i,M}, T_i, \mathbf{U}_i)$ to select a model for Δ. Since the true Δ_M is known for each M, we can examine the features of each model selection method under different versions of the truth compatible with the data. To average out the variability in the random errors, the results are aggregated over 100 different realizations of the error vector.

Table 6.3: Results of simulation guided by FIRST trial data

Scenario		AIC	BIC	wFIC	CV	TECV
AIC	Trt Effects in Model	89	11	78	57	55
	Avg MSE for Δ (SE)	0.81 (0.05)	0.89 (0.01)	0.68 (0.03)	0.78 (0.03)	0.77 (0.03)
	Pairwise Success (%)	72.8	62.9	72.7	66.7	64.2
	Falsely Treated (%)	27.0	4.9	26.9	21.2	20.6
	Undertreated (%)	29.7	90.5	33.4	51.7	53.6
wFIC	Trt Effects in Model	85	10	77	57	51
	Avg MSE for Δ (SE)	0.74 (0.04)	0.83 (0.01)	0.65 (0.03)	0.70 (0.03)	0.78 (0.03)
	Pairwise Success (%)	74.0	68.9	70.6	67.1	62.7
	Falsely Treated (%)	24.9	4.4	23.7	17.3	19.8
	Undertreated (%)	31.3	91.5	35.7	52.2	58.1
BIC	Trt Effects in Model	38	1	18	11	17
	Avg MSE for Δ (SE)	0.35 (0.05)	0.01 (0.01)	0.14 (0.03)	0.07 (0.02)	0.11 (0.03)
	Pairwise Success (%)	67.2	85.5	86.9	72.6	87.4
	Falsely Treated (%)	19.0	1.0	9.0	6.6	10.2
	Undertreated (%)	–	–	–	–	–

We present the results of the cross-examination under three scenarios: a model with two nonlinear treatment-covariate interactions (corresponding to AIC's preferred model), a model with two linear treatment-covariate interaction terms (the model selected by wFIC), and BIC's model, which indicated no treatment effects at all. Results were similar for different models indicating no treatment effects, so we present only the BIC scenario here and omit the CV and TECV scenarios. Table 6.3 contains the results of the guided simulation. For each scenario, we first count how many times of the 100 realizations the selected model contained a treatment main effect or any treatment-covariate interactions. The falsely treated percentage is the average percentage of patients with a nonpositive Δ that are assigned a positive $\widehat{\Delta}$ by the estimate chosen by the model selection method. We also compute the undertreated percentage; that is, the average percentage of patients with a positive Δ that have a nonpositive $\widehat{\Delta}$.

Neither the use of AIC nor the use of BIC seem to be satisfactory under all three scenarios because of their respective tendencies to overfit and underfit. In the BIC scenario, under which no treatment effect exists, AIC falsely selects a model with the treatment variable in 38 of 100 realizations. As a result, it has by far the highest MSE for Δ as well as a high percentage of patients for which it recommends unnecessary treatment. On the other hand, when a treatment effect is present under the AIC and wFIC scenarios, fewer than 10 percent of patients for whom the treatment is beneficial receive a treatment recommendation from BIC.

The targeted model selection methods perform better for our purpose, with wFIC featuring the best performance overall. The restriction of wFIC to consider only the set of linear models appears to work in favor of wFIC in this case. For example, under the AIC scenario the true regression function is nonlinear, but still the additive models possess greater variability and often suffer worse performance than the corresponding linear models. Notice that even under the AIC scenario, wFIC features significantly lower MSE for Δ than AIC. In all three scenarios, wFIC correctly identified whether or not a treatment effect was present over 75 percent of the time. Overall, this cross-examination suggests that model selection methods targeted toward estimation of the treatment effect can be useful when analyzing clinical trial data.

6.2 Application: Labor Training Program

6.2.1 The LaLonde Data

In this section we apply wFIC, TECV, TEEM, and various other model selection and combination methods to the well-known LaLonde (1986) National Supported Work (NSW) Demonstration data set. The NSW Demonstration was a federally and privately funded program in the 1970s that provided work experience to individuals who were struggling financially. Eligible participants were randomly assigned to the treatment or control group, and follow-up interviews were conducted with both groups to obtain information about post-intervention earnings. LaLonde (1986) analyzed the male and female participants separately, and we will focus on the study's male participants. The male participants from this experiment were previously analyzed by Dehejia and Wahba (1999) in a study of propensity scores and by Imai and Ratkovic (2013), who used a penalized regression method to estimate heterogeneity of the treatment effect.

Table 6.4: Variables used in LaLonde NSW data analysis

Name	Type	Description	Mean
Y	Outcome	$\sqrt{\text{1978 income}} - \sqrt{\text{1975 income}}$	20.4
T	Treatment	T=1 if enrolled in training; otherwise, T=0	0.411
Inc75	Covariate	$\sqrt{\text{1975 income}}$	37.5
Educ	Covariate	Years of education	10.3
Age	Covariate	Age in years	24.5
Married	Covariate	Married=1 if married; otherwise; Married=0	0.162

There were $n = 722$ male participants in the experiment; 297 were treated and 425 were in the control group. The outcome variable Y in our analysis is the change in the square root of income from 1975 (pre-treatment) to 1978 (post-treatment). Square root transformations on income were done to reduce skewness. The treatment variable T equals 1 if the person was treated in the NSW demonstration; otherwise, T=0. Four baseline covariates ($\sqrt{\text{1975 income}}$, age, education, and marital status), measured before the treatment was applied, are used to identify heterogeneity of the treatment effect in some of the candidate models. Racial indicators for black and Hispanic individuals also are available in the LaLonde dataset, but these variables were not used because a preliminary analysis provided no evidence that race moderated the treatment effect. The six two-way interactions of the four baseline covariates were not effective moderators in a preliminary linear model, so these were not considered further. Descriptions and mean values for each of the variables used in our analysis are given in Table 6.4.

6.2.2 Candidate Models and Methods

For this analysis, as for many other examples of treatment effect heterogeneity, it seems plausible that the effect of the job training program on the outcome may be nonlinear with respect to some of the covariates. For example, the program may be most beneficial for those in the middle of the income, education, or age distributions. Therefore, our set of candidate models includes linear models and additive models, which are possibly nonlinear.

Preliminary analysis shows that baseline income (Inc75) is clearly related to the response, so this covariate is included in every candidate model. The linear candidate models are those containing different subsets of the variables {T, Educ, Age, Married, T:Inc75, T:Educ, T:Age, T:Married}. Model hierarchy is enforced, meaning if a treatment-covariate interaction is included in the model, both corresponding main effects must be included as well. This constraint applied to this set of variables allows for 62 possible linear models.

The additive candidate models are estimated with the `gam` function from the R `mgcv` package (Wood, 2006). Each term in the additive model is a smooth, possibly nonlinear function of a single covariate. Treatment-covariate interactions are estimated by allowing these functions to vary for each covariate depending on the value of the treatment variable. The default choice of smoothing parameter (based on generalized cross-validation) is used to fit each model. Since Married is a categorical variable, terms involving Married and T:Married are simply linear (no smoothing). Model hierarchy is similarly enforced as in the linear model consideration set, generating 62 possible additive models and 124 candidate models in all.

Sixteen of the models have no treatment effects (i.e., $\widehat{\Delta} = 0$). Since these models differ with respect to their estimation of the response, they are considered as separate models for the TEEM model combination algorithm so that all model selection and combination methods were presented with the same set of candidates. For the TEEM model combination algorithm, this effectively gives the null estimate of $\widehat{\Delta} = 0$ a prior weight of 16 times the other candidate estimates of $\widehat{\Delta}$ and serves to shrink the estimate of $\widehat{\Delta}$ resulting from the algorithm.

We apply several common methods of model selection and model combination to this set of candidate models for the LaLonde data. The five model selection methods used in the analysis of Section 6.1 are used to select a model, while the five model combination methods used in Section 5.4 combine the candidate models. The data-splitting based statistics of CV, TECV, ARM, and TEEM are averaged over 100 different 50/50 splits, as before. The methods of wFIC and cwFIC are applied only to the 62 linear models and do not consider the additive models. Previously we have discussed the issue of choosing the tuning parameter λ for the combination method of cwFIC. The default value of $\lambda = 1$ clearly is too large for this data analysis, as it results in one model (the one with minimum wFIC) receiving almost all of the weight. We apply the method at three (somewhat arbitrary) values of λ $(10^{-2}, 10^{-3}$, and $10^{-4})$ in our analysis of the LaLonde data.

6.2.3 Results

Table 6.5 summarizes the results of the model selection and combination methods when applied to the LaLonde data. For each model selection method, the table lists the model that was selected. For each selection and combination method, the mean and standard deviation (SD) of the resulting estimator $\widehat{\Delta}(\mathbf{U}_i)$ over the 722 data values of $\mathbf{U}_i$ are reported. The standard deviation of $\widehat{\Delta}(\mathbf{U}_i)$ indicates the level of treatment effect heterogeneity indicated by each estimator.

Each of the five model selection methods identifies a different model, with each implying something different about the treatment's effect on the outcome. For example, the additive model selected by AIC implies the treatment effect varies nonlinearly with pre-treatment income and age, as well as being different for married vs. single people. The model selected by BIC implies the NSW treatment has no effect at all on the outcome, while the model chosen by traditional CV implies a homogeneous positive treatment effect. The two model selection methods targeted to estimation of Δ, wFIC and TECV, each select a linear model implying a heterogeneous treatment effect.

Table 6.6 shows the model that was selected by the wFIC criterion. The interactions

Table 6.5: $\Delta(\mathbf{u})$ estimates chosen by model selection and combination methods for NSW data analysis

			$\widehat{\Delta}(\mathbf{U}_i)$ Values[b]	
Method	Model Type	Active Variables[a]	Mean	SD
AIC	Additive	T*s(re75), T*s(Age), T*Married, Educ	5.7	12.5
BIC	Linear	re75	0	0
wFIC	Linear	T*re75, T*Married	6.6	8.6
CV	Linear	T, re75	6.6	0
TECV	Linear	re75, T*Married	6.7	6.4
cAIC			5.6	5.2
BMA			1.4	0.1
cwFIC ($\lambda = 10^{-2}$)			6.6	8.7
cwFIC ($\lambda = 10^{-3}$)			6.6	8.6
cwFIC ($\lambda = 10^{-4}$)			6.1	4.2
ARM			5.3	2.9
TEEM			5.6	3.8

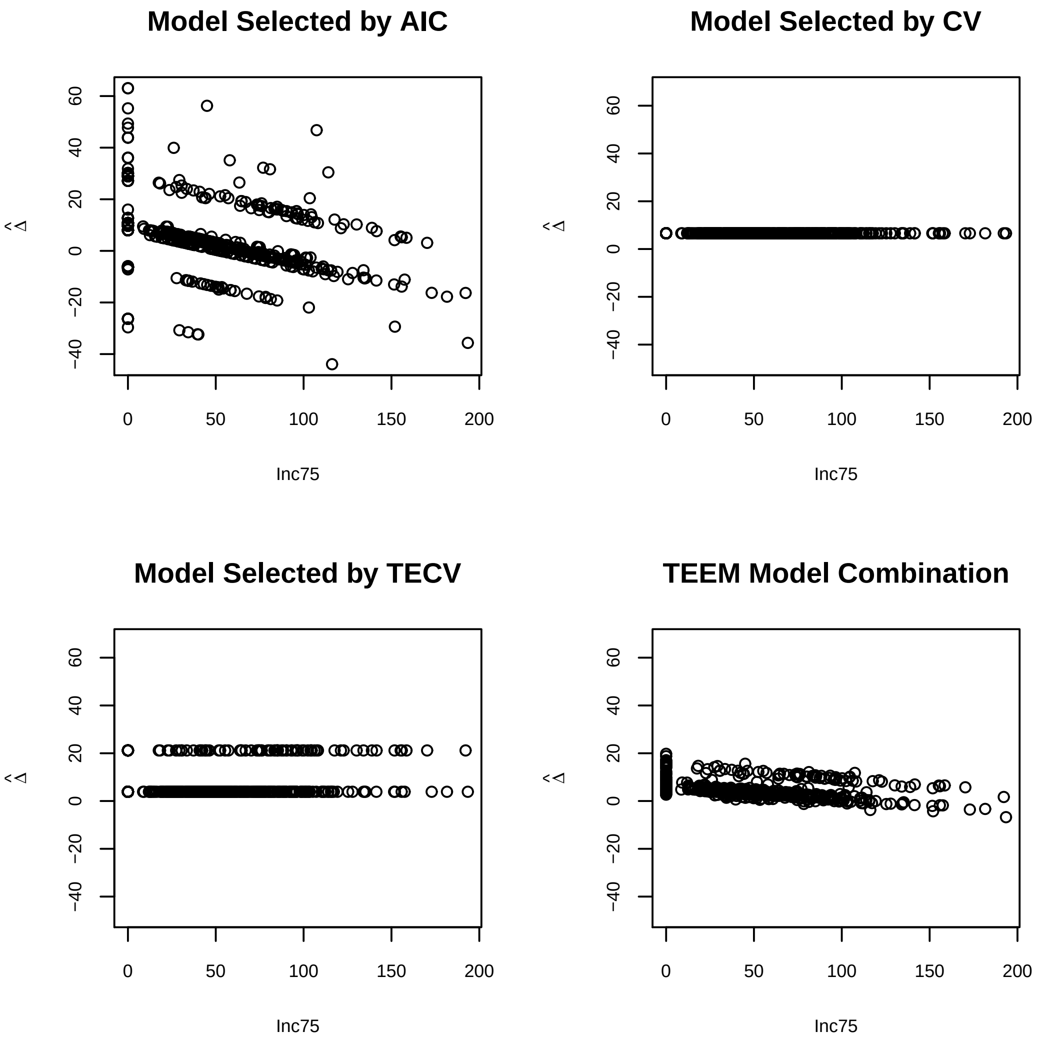
[a] The presence of interaction terms implies the presence of both main effects.
[b] The mean and standard deviation of $\widehat{\Delta}(\mathbf{U}_i)$ over the $n = 722$ values of $\mathbf{U}_i$.

Table 6.6: Linear model selected by wFIC for LaLonde NSW data

Term	Estimate	P-value
Intercept	47.7	< 0.001
T	9.2	0.055
Inc75	-0.77	< 0.001
Married	-6.7	0.276
T:Inc75	-0.15	0.074
T:Married	19.5	0.038

imply that the treatment is more effective for people with a lower pre-treatment income and for people who are married. All three terms involving the treatment indicator have p-values between 0.038 and 0.074. With p-values in this range, it is perhaps not surprising that each of the model selection methods are giving different estimates of the

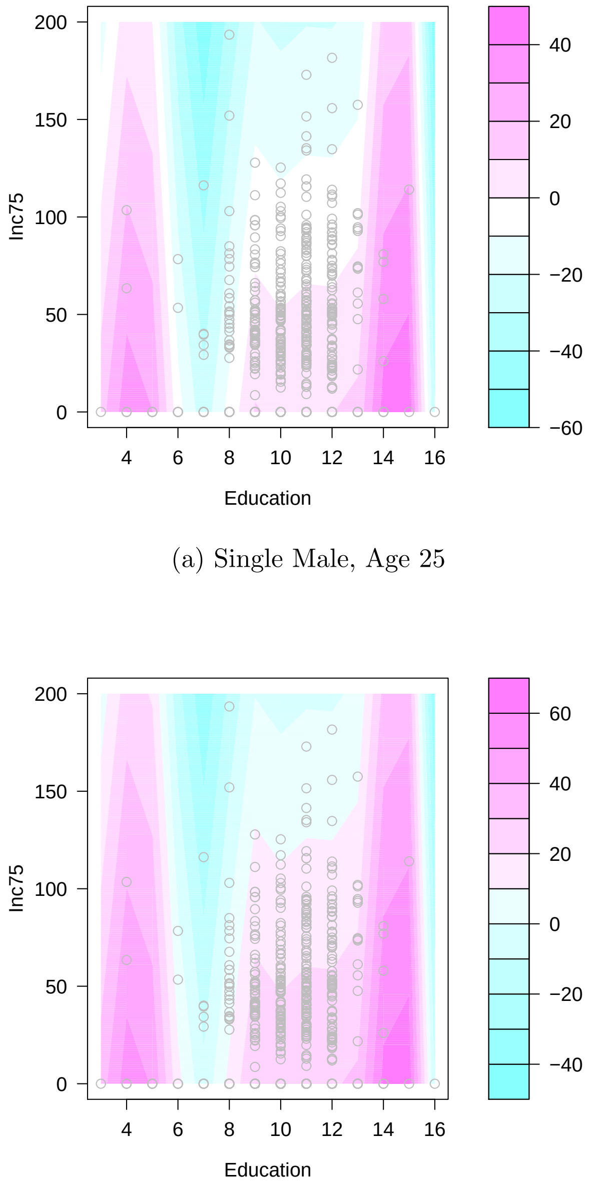
Figure 6.1: $\widehat{\Delta}$ from the models selected by AIC, CV, and TECV, and from the combination produced by TEEM are plotted against the Inc75 variable.



Δ function.

Table 6.5 and Table 6.6 suggest there is substantial model selection uncertainty in this analysis. In such situations, model combination often provides a good compromise between similar-performing models that give quite different estimates. In Figure 6.1, the $\widehat{\Delta}$ values from the LaLonde data resulting from AIC, CV, TECV, and TEEM are

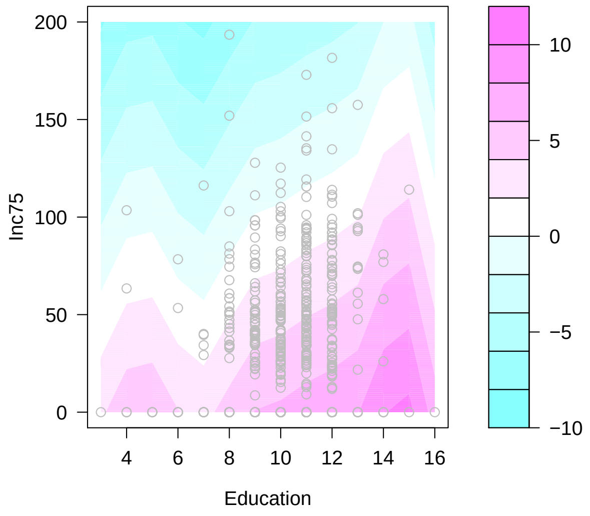
Figure 6.2: Contour plots for $\widehat{\Delta}$ from the model selected by AIC for the LaLonde NSW data. The circles indicate the locations of one or more original data points.



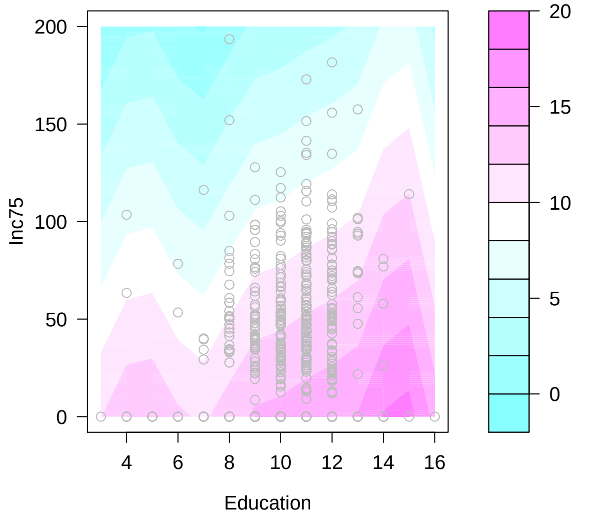
(a) Single Male, Age 25

(b) Married Male, Age 25

Figure 6.3: Contour plots for $\widehat{\Delta}$ from the TEEM algorithm applied to the LaLonde NSW data. The circles indicate the locations of one or more original data points.



(a) Single Male, Age 25



(b) Married Male, Age 25

plotted against the square root of pre-treatment income. For the model selected by AIC, there is a lot of variability in $\widehat{\Delta}$, with an overall negative association between $\widehat{\Delta}$ and pre-treatment income. For the models selected by CV and TECV, there is no interaction between the treatment and baseline income. The TEEM model combination lies somewhere in between these extremes, exhibiting the negative association between $\widehat{\Delta}$ and pre-treatment income but with much less variability in $\widehat{\Delta}$. We don't know how the treatment effect truly varies with pre-treatment income, but it is certainly plausible that those who entered the program with a higher income benefited less from the program overall. At the same time, it seems unlikely that the treatment effect is as heterogeneous as the additive model selected by AIC suggests.

The decreased variability produced by TEEM can also be seen in Figures 6.2 and 6.3. These are contour plots of $\widehat{\Delta}$ from AIC and TEEM, respectively, over the range of Inc75 and Education in the LaLonde data. Age is set to 25 (the sample mean), and subfigures (a) and (b) show results for single and married males, respectively. The points on the contour plots represent individuals in the LaLonde sample.

The AIC plots in Figure 6.2 show the large heterogeneity in $\widehat{\Delta}$ with respect to both variables. In particular, the heterogeneity and nonlinearity of the estimated treatment effect with respect to education in Figure 6.2 seems implausible. The plots in Figure 6.3 show a much more reasonable degree of heterogeneity (note the differing scales in Figures 6.2 and 6.3). The contour plots of $\widehat{\Delta}$ from TEEM show a positive estimated treatment effect for most of the individuals, including all of those with no pre-treatment income. TEEM does suggest some treatment effect heterogeneity, with the program estimated to be more beneficial for those with little or no income, those a higher degree of education, and married participants.

6.2.4 Cross-examination of LaLonde Data

To gain further insight into the performance of these model selection and combination methods on the LaLonde data, we perform a guided cross-examination simulation experiment like the one in Section 6.1.3. In this type of simulation experiment, each model selection method gets a chance to compete against the other model selection (and combination) methods on its "home field". The five methods of model selection all provide different answers about the treatment effect in the LaLonde data, so each of

Table 6.7: Estimated risks[a] (SEs) of $\widehat{\Delta}$ under 5 scenarios

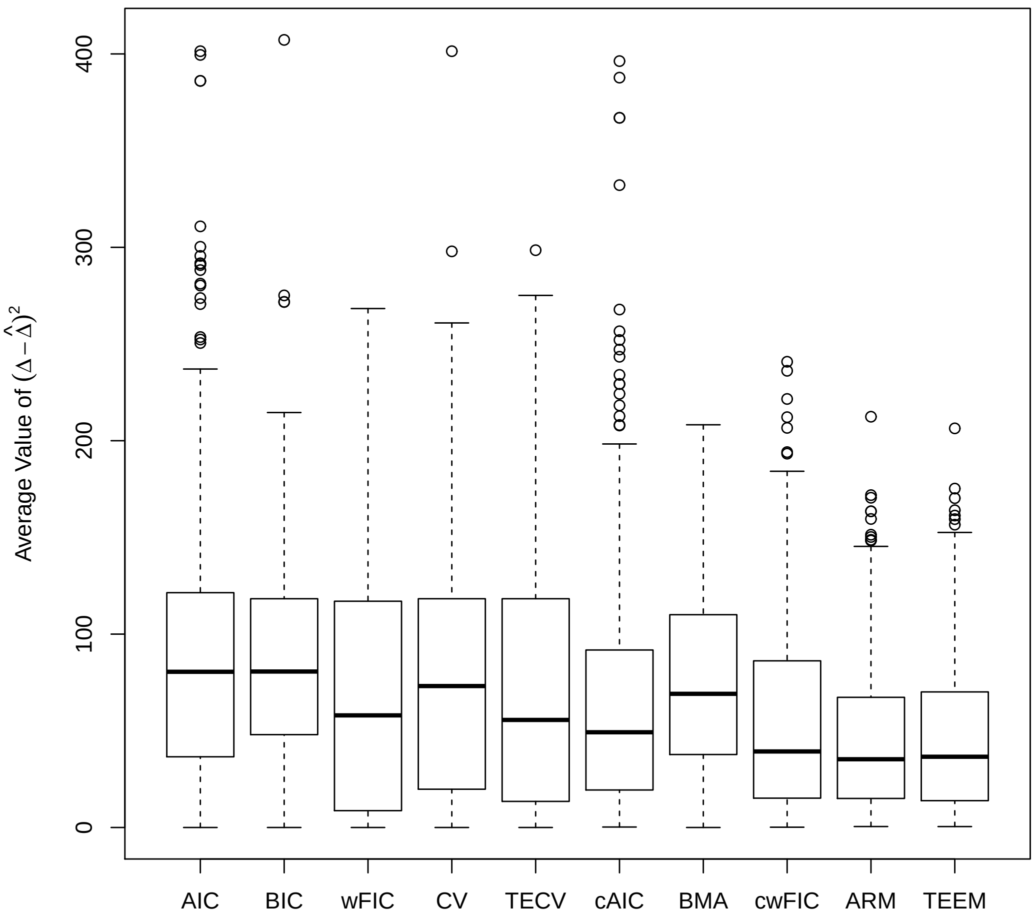
	Model Selection Method Determining $E(Y\|T,\mathbf{U})$				
Method	AIC	BIC	wFIC	CV	TECV
AIC	158.6 (6.1)	41.3 (6.2)	86.8 (6.3)	54.6 (5.5)	74.2 (6.0)
BIC	191.6 (2.8)	**1.4** (1.4)	109.8 (1.6)	37.0 (2.1)	72.3 (1.5)
wFIC	149.3 (4.0)	21.0 (4.5)	69.9 (4.9)	39.7 (3.9)	53.2 (4.2)
CV	169.3 (4.4)	12.6 (3.7)	81.0 (4.3)	31.1 (3.4)	59.6 (3.7)
TECV	160.9 (4.0)	21.4 (4.2)	75.8 (4.2)	36.4 (3.6)	48.7 (4.0)
cAIC	134.6 (5.4)	28.0 (4.9)	65.5 (5.1)	35.8 (4.8)	51.4 (5.0)
BMA	166.4 (2.1)	**1.3** (1.1)	88.4 (2.2)	**24.1** (1.7)	57.4 (1.8)
cwFIC ($\lambda = 10^{-2}$)	149.2 (4.0)	20.6 (4.4)	69.0 (4.8)	39.1 (3.9)	52.9 (4.2)
cwFIC ($\lambda = 10^{-3}$)	146.6 (4.0)	19.6 (4.4)	66.5 (4.6)	37.1 (3.8)	50.9 (4.2)
cwFIC ($\lambda = 10^{-4}$)	127.8 (3.4)	15.6 (2.8)	**51.8** (3.4)	**22.1** (2.7)	**38.3** (3.0)
ARM	**112.1** (2.7)	11.6 (1.7)	**47.1** (2.5)	**16.6** (1.8)	**32.5** (2.1)
TEEM	**108.6** (2.9)	15.1 (1.9)	**46.4** (2.7)	**20.4** (2.0)	**32.3** (2.3)

[a] Numbers in bold represent the methods with the lowest estimated risks for each scenario and those not statistically indistinguishable (using Tukey's HSD method of multiple comparisons) from the lowest-risk method.

these answers is given a chance to serve as the true data-generating process. The active variables in each of the five scenarios can be found by looking at the model selected by each method in Table 6.5. The AIC scenario features a nonlinear Δ, while the BIC scenario has $\Delta = 0$. The scenario under CV has a constant Δ of 6.6, while the wFIC and TECV scenarios feature linearly heterogeneous Δ.

Table 6.7 shows the estimated risk (average mean squared error) of $\widehat{\Delta}$ as an estimator of Δ for each of the 12 model selection and combination methods under each of the 5 scenarios. For the scenario with a nonlinear Δ (the model selected by AIC), the model combination methods of ARM and TEEM significantly outperform all other methods, including AIC. AIC does not perform so well because even if the true (additive) model is selected, there is substantial variability involved in estimating the true mean function and the true Δ. When $\Delta = 0$ (the BIC scenario), BIC usually chooses a model with no

Figure 6.4: Results of the cross-examination over the five scenarios combined. Each data point in the boxplot is an average of $(\Delta - \widehat{\Delta})^2$ over the evaluation set, where $\widehat{\Delta}$ is an estimate resulting from the application of the method to one realization of sample data. Since 100 realizations are generated for each method, each boxplot summarizes 500 data points.



treatment effect and thus performs the best.

In each of the other three scenarios (based on the models chosen by wFIC, CV, and TECV), the model combination methods of ARM, TEEM, and cwFIC (with $\lambda = 10^{-4}$)

perform the best. Although cwFIC with a small value of λ performed well, in practice a good setting for λ would not be known in advance and would need to be selected heuristically or by some sort of cross-validation. Also, cwFIC is not equipped to combine linear and non-parametric models; this appears to hurt its performance in the AIC scenario, when the nonlinear model is correct. It is notable that in each of the three scenarios with a heterogeneous treatment effect, TEEM featured the lowest estimated risk.

A summary of each method's estimated risks over all five scenarios combined can be found in Figure 6.4. Each data point in the boxplot represents an average of $(\Delta - \widehat{\Delta})^2$ for one realization (from one of the five scenarios) over the $n = 722$ $(T, \mathbf{U})$ values. There are therefore 500 data points represented in each boxplot. Among the model selection methods, wFIC and TECV (the only methods targeted to selecting a model for the treatment effect) had the lowest median risk. Model combination methods possessed lower risk than model selection methods, with the data-splitting based methods of ARM and TEEM performing the best overall. Between these two, ARM performed slightly better in the two scenarios where the treatment effect was constant, while TEEM was slightly better in the three scenarios where Δ was heterogeneous.

In this section and the previous one, we compared the methods of TECV and TEEM developed in this thesis to several other methods of model selection and combination on a famous dataset from public policy known as the LaLonde labor training data. For the LaLonde data, TEEM seems to provide a sensible data-driven weighting of the treatment effect estimates recommended by various model selection methods. A guided simulation demonstrates that in various settings consistent with the LaLonde data, TECV and TEEM compare favorably with other selection and combination methods for providing an accurate estimate of the conditional treatment effect.

Chapter 7

Discussion

7.1 Future Research Directions

The topic of conditional treatment effect estimation presents a number of interesting theoretical issues and has great potential for practical use. Several methods to estimate Δ have recently been proposed, and in this thesis we discuss selecting and combining different estimators of Δ to achieve small risk under a global L_2 loss. In this section we mention some future research directions that may lead to further understanding of this important topic.

Two-stage designs. The methods of TECV and TEEM both involve the pairing of nearby individuals in the treatment and control groups. This pairing suffers from the "curse of dimensionality" in that it can be difficult to find nearby pairs when p, the dimension of the covariate vector, is large. This property shows up in our theoretical results; for example, the term $(\log n_2/n_2)^{1/p}$ appears in the risk bound of TEEM in Theorem 2. We wish to propose a two-stage experimental design that could be applied in clinical trials (and other studies) to produce an estimator from model combination with an improved convergence rate in high-dimensional settings. The first stage of the design would test for the overall effectiveness of the treatment in the population and allow for dimension reduction of the covariate vector (say, to q dimensions) through a screening step. In the second stage, a large control pool could be used to find nearby treatment-control pairs (with respect to the q covariates) so that the bias induced by the treatment-control pairings would be at most $O_p\left(n^{-1/2}\right)$. Under this design, one could guarantee that (under conditions similar to those in Theorem 2) the combined estimator would converge at the same rate as the best procedure among the candidates.

Penalized regression. Penalized regression methods have received much attention in recent years. The use of methods such as the LASSO (Tibshirani, 1996) for variable selection is common when the number of covariates p exceeds the sample size n. While most penalized regression methods target estimation of the full regression function, we wish to create a penalized regression for treatment effect estimation. The method of Imai and Ratkovic (2013) discussed in Section 2.4.2 uses L_1 LASSO penalties to identify treatment effect heterogeneity, but their method focuses on binary responses and does not directly target estimation of Δ in selecting the tuning parameters. A properly targeted method of penalized regression may be useful to many researchers.

Time-to-event outcomes. In many medical and business applications where a treatment is analyzed, the outcome of interest is the time until some event. For example, in medicine the outcome could be the time until death or until the onset of a disease, while in a marketing study the outcome may be the time until the customer's next purchase. Many studies express the estimated treatment effect on a time-to-event outcome as a hazard ratio, the ratio of the instantaneous risk of the event under treatment to the same risk under control. For a given application, several different models might be considered to estimate hazard ratios conditional on covariates. Therefore, model selection and combination methods targeted to the estimation of conditional hazard ratios could have many applications.

Other loss functions. The methods developed in this thesis target estimation of Δ under squared error loss. Although this loss function is the most commonly used in statistics (for continuous outcomes), it may not be the most practically relevant. When making practical treatment decisions, one may be more interested in the value function of the treatment rule resulting from the model (as in Qian and Murphy, 2011) or in the reliability of the model for correctly estimating the sign of the treatment effect. Asymmetric loss functions, for which either over-estimation or under-estimation of the treatment effect is considered more harmful, may also be of interest. Because of the flexibility of cross-validation as a model evaluation tool, our methods could be extended to different types of loss functions.

Kernel-based estimates of treatment differences. The methods proposed in this thesis estimate treatment-control differences at different values of the covariate vector using independent pairs arising from a partition of the covariate space. Treatment-control differences alternatively could be estimated using kernel-based weights instead of non-overlapping partitioning. The use of kernel weights could increase the bias of $\widetilde{\delta}_i$ as an estimate of $\Delta(\mathbf{U}_i)$ by allowing more observations, including those further from $\mathbf{U}_i$, to influence $\widetilde{\delta}_i$. On the other hand, kernel-based weights could reduce variability of the $\widetilde{\delta}_i$ by incorporating more observations. Selection of a proper bandwidth for the kernel function would permit a theoretical solution to the bias-variance trade-off.

7.2 Conclusion

There are a number of subtle issues involved in problems of model selection and combination, and in this section we discuss some of these in the context of treatment effect estimation.

Although it is not the focus of our current work, quantifying the uncertainty of treatment effect estimates is an important goal. For example, confidence intervals of individual-level treatment benefits can be helpful for medical practitioners, particularly when the treatment may carry serious risks. Given the proper model assumptions, asymptotically valid confidence intervals for Δ can be computed from most regression procedures. If the conditions of Theorem 1 hold, then TECV asymptotically chooses the best model (under L_2 loss) for Δ with probability tending to 1. Consequently, if the consideration set contains one or more models that capture a true representation of Δ, we can expect the model selected by TECV to reflect the true Δ with high probability asymptotically. Thus for any fixed Δ, a confidence interval from the regression procedure selected by TECV will attain an asymptotically correct coverage rate.

Leeb and Pötscher (2005) argue against such post model selection inference, citing the non-uniform convergence of the regression parameters' finite-sample distributions. Their view is that even when a consistent model selection procedure is used, inference done after model selection cannot accurately account for the uncertainty involved in the selection step. While this view is an important message to keep in mind to avoid overly optimistic analysis that ignores model selection uncertainty, taking this view too rigidly would lead to rejection of any model-based inference because for whatever given model, there are always larger models that can be considered. In situations where there is one model or a set of similar models that stands out amongst others, model selection uncertainty is less of a concern and inference based on the selected model is largely trustworthy.

Model selection uncertainty can be quantified statistically to indicate the reliability of post model selection inference for the problem at hand. In the setting of linear candidate models with Gaussian errors, Liu and Yang (2011) propose a parametricness index (PI) that can be used to determine if one model stands out as a stable parametric description of the data. If the PI is high for the model selection problem at hand, then post-model selection inference will be reasonably accurate; however, if the PI does not support the selected model as the right parametric model for the data, the usual post model selection inference should be viewed with skepticism. In our setting of treatment effect estimation, another way to estimate model selection uncertainty is to count the votes received by each model in a voting-based version of TECV. If one model or a set of similar models wins the voting by a large margin, we can be reasonably confident in the inference based on the chosen model. However, if very different models each receive many votes, one should not claim the resulting confidence intervals from the selected model are valid.

A somewhat related issue is the conflict between the goals of pursuing consistency on the one hand and estimating the regression function on the other. It is well-known that in many cases (e.g., a sequence of increasing models), BIC is consistent if the true parametric model is among the candidates, while AIC is asymptotically efficient and minimax-rate optimal for estimating the regression function. Yang (2005) showed that in a parametric setting, any consistent model selection procedure cannot be minimax-rate optimal; i.e., the strengths of AIC and BIC cannot be shared. The same conflict holds in general for estimating the treatment effect; a model selection procedure that consistently identifies the true model for Δ in a parametric situation cannot be minimax-rate optimal for estimating Δ.

Our definition of selection consistency is more general than the usual definition that a model selection method is consistent if it selects the true model (when existing and being considered) with probability tending to one. Our definition allows for comparisons between parametric and nonparametric models and does not assume the existence of a true model. Pursuit of selection consistency as we have defined it does not necessarily conflict with the goal of estimating Δ. When both parametric and nonparametric models are being compared, it generally seems possible for a model selection method to achieve both selection consistency and minimax-rate optimality for estimating Δ.

When combining a given candidate set of statistical models, there can be two different goals. One might desire a combination method that will automatically perform as well as the best model in the candidate set, without requiring the knowledge of which model is best. A more ambitious goal would be to combine the models so that the combined estimator improves on even the best-performing candidate. Yang (2004) calls the first goal *combining for adaptation* and the second goal *combining for improvement*. Our method of TEEM is an adaptive method. When $p = 1$, Theorem 2 says the TEEM estimator automatically performs almost as well (up to a constant) as the best model in the candidate set. There is no theoretical guarantee that TEEM will be able to improve on the best model, although our numerical work suggests that improvement can sometimes be achieved by the TEEM combination.

TECV and TEEM allow for flexibility in the data splitting ratio. For TECV to achieve selection consistency in a parametric framework, Theorem 1 requires the size of the evaluation set to dominate that of the estimation set when $p > 1$. If the goal is to identify which estimation procedure is best, a higher proportion of evaluation observations is needed to enable TECV to differentiate between two procedures with similar performance. On the other hand, if the goal is to choose a model with low risk for estimating Δ, it will often be better to include more observations in the estimation part. In our view, a 50/50 splitting of the data into estimation and evaluation provides a nice balance to achieve the goals of estimating Δ and evaluating competing procedures for estimating Δ. Therefore, in our numerical examples we have used a 50/50 splitting for TECV and TEEM when comparing them with other methods. For a given application, another data splitting ratio may be preferred depending on the purpose of the investigation.

This dissertation studies the estimation of treatment effects at the individual level or at the level of subgroups. Conditional treatment effect estimation is an important goal of many investigations, and we anticipate that methods to estimate such local treatment effects will increase in popularity as practices such as personalized medicine and targeted online advertising become more common. While several methods of estimating heterogeneous treatment effects have recently been proposed, this thesis addresses the problem of selecting or combining estimation procedures in this setting. By applying reliable methods of model evaluation tailored to estimation of the treatment effect, practitioners can select a model that is best for the data they have and the purpose they need.

References

Abadie, A. and Imbens, G. W. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267.

— (2011), "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business and Economic Statistics*, 29, 1–11.

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *"IEEE Transactions on Automatic Control"*, 19, 716–723.

Althauser, R. P. and Rubin, D. (1970), "The Computerized Construction of a Matched Sample," *American Journal of Sociology*, 76, 325–346.

Barron, A. R. (1987), "Are Bayes Rules Consistent in Information?" in *Open Problems in Communication and Computation*, eds. Cover, T. M. and Gopinath, B., Springer-Verlag, pp. 85–91.

Breheny, P. and Huang, J. (2011), "Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection," *The Annals of Applied Statistics*, 5, 232–253.

Breiman, L. (1996), "Stacked Regressions," *Machine Learning*, 24, 49–64.

— (2001), "Random Forests," *Machine Learning*, 45, 5–32.

— (2004), "Population Theory for Boosting Ensembles," *The Annals of Statistics*, 32, 1–11.

Buckland, S., Burnham, K., and Augustin, N. (1997), "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618.

Buclin, T., Telenti, A., Perera, R., Csajka, C., Furrer, H., Aronson, J., and Glasziou, P. (2011), "Development and Validation of Decision Rules to Guide Frequency of Monitoring CD4 Cell Count in HIV-1 Infection Before Starting Antiretroviral Therapy," *PloS one*, 6, e18578.

Cai, T., Tian, L., Wong, P. H., and Wei, L. (2011), "Analysis of Randomized Comparative Clinical Trial Data for Personalized Treatment Selections," *Biostatistics*, 12, 270–282.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, 4, 266–298.

Chvátal, V. (1979), "The Tail of the Hypergeometric Distribution," *Discrete Mathematics*, 25, 285–287.

Claeskens, G. and Hjort, N. L. (2003), "The Focused Information Criterion," *Journal of the American Statistical Association*, 98, 900–916.

— (2008a), "Minimizing Average Risk in Regression Models," *Econometric Theory*, 24, 493–527.

— (2008b), *Model Selection and Model Averaging*, Cambridge: Cambridge University Press.

Cook, R. D. (1998), *Regression Graphics*, New York: Wiley.

Cook, R. D. and Li, B. (2002), "Dimension Reduction for Conditional Mean in Regression," *The Annals of Statistics*, 30, 455–474.

Cook, R. D., Li, B., and Chiaromonte, F. (2007), "Dimension Reduction in Regression Without Matrix Inversion," *Biometrika*, 94, 569–584.

Cook, R. D. and Weisberg, S. (1991), "Discussion of Li (1991)," *Journal of the American Statistical Association*, 86, 328–332.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008), "Nonparametric Tests for Treatment Effect Heterogeneity," *The Review of Economics and Statistics*, 90, 389–405.

Dehejia, R. H. and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.

Feller, A. and Holmes, C. C. (2009), "Beyond Toplines: Heterogeneous Treatment Effects in Randomized Experiments," (Available from http://www.stat.columbia.edu/~gelman/stuff_for_blog/feller.pdf).

Fisher, R. A. (1935), *The Design of Experiments*, London: Oliver and Boyd.

Freund, Y. and Shapire, R. E. (1996), "Experiments with a New Boosting Algorithm," in *Machine Learning: Proceedings of the Thirteenth International Conference*, pp. 148–156.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models Via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22.

Geisser, S. (1975), "The Predictive Sample Reuse Method With Applications," *Journal of the American Statistical Association*, 70, 320–328.

Green, D. P. and Kern, H. L. (2010), "Detecting Heterogeneous Treatment Effects in Large-Scale Experiments using Bayesian Additive Regression Trees," in *The Annual Summer Meeting of the Society of Political Methodology, University of Iowa.*

Hansotia, B. and Rukstales, B. (2002), "Incremental Value Modeling," *Journal of Interactive Marketing*, 16, 35–46.

Hjort, N. L. and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007), "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis*, 15, 199–236.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417.

Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.

Imai, K. and Ratkovic, M. (2013), "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation," *The Annals of Applied Statistics*, 7, 443–470.

Imbens, G. and Wooldridge, J. M. (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

LaLonde, R. J. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review*, 76, 604–620.

Leeb, H. and Pötscher, B. M. (2005), "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

Li, K.-C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–327.

Li, K.-C., Lue, H.-H., and Chen, C.-H. (2000), "Interactive Tree-Structured Regression Via Principal Hessian Directions," *Journal of the American Statistical Association*, 95, 547–560.

Li, L. (2007), "Sparse Sufficient Dimension Reduction," *Biometrika*, 94, 603–613.

Li, L. and Yin, X. (2008), "Sliced Inverse Regression with Regularizations," *Biometrics*, 64, 124–131.

Liu, W. and Yang, Y. (2011), "Parametric or Nonparametric? A Parametricness Index for Model Selection," *The Annals of Statistics*, 39, 2074–2102.

MacArthur, R. D., Novak, R. M., Peng, G., Chen, L., Xiang, Y., Hullsiek, K. H., Kozal, M. J., van den Berg-Wolf, M., Henely, C., Schmetter, B., and Dehlinger, M. (2006), "A Comparison of Three Highly Active Antiretroviral Treatment Strategies Consisting of Non-Nucleoside Reverse Transcriptase Inhibitors, Protease Inhibitors, or Both in the Presence of Nucleoside Reverse Transcriptase Inhibitors as Initial Therapy (CPCRA 058 FIRST Study): A Long-Term Randomised Trial," *The Lancet*, 368, 2125 – 2135.

Neyman, J. (1935), “Statistical Problems in Agricultural Experiments,” *Journal of the Royal Statistical Society II*, 2, 107–180.

Opsomer, J., Wang, Y., and Yang, Y. (2001), “Nonparametric Regression with Correlated Errors,” *Statistical Science*, 16, 134–153.

Pollard, D. (1984), *Convergence of Stochastic Processes*, New York: Springer.

Qian, M. and Murphy, S. A. (2011), “Performance Guarantees for Individualized Treatment Rules,” *The Annals of Statistics*, 39, 1180–1210.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Radcliffe, N. J. and Surry, P. D. (2011), “Real-World Uplift Modelling with Significance-Based Uplift Trees,” Tech. rep., Stochastic Solutions, Edinburgh, UK.

Raftery, A. E. (1995), “Bayesian Model Selection in Social Research,” *Sociological Methodology*, 25, 111–163.

Rice, J. (1984), “Bandwidth Choice for Nonparametric Regression,” *The Annals of Statistics*, 12, 1215–1230.

Rolling, C. A. and Yang, Y. (in press), “Model Selection for Estimating Treatment Effects,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, doi = 10.1111/rssb.12043.

Rosenbaum, P. R. and Rubin, D. B. (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–55.

Rubin, D. B. (1974), “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology*, 66, 688–701.

Schwarz, G. (1978), “Estimating the Dimension of a Model,” *The Annals of Statistics*, 6, 461–464.

Stone, M. (1974), “Cross-Validatory Choice and Assessment of Statistical Predictions,” *Journal of the Royal Statistical Society. Series B*, 36, 111–147.

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society. Series B*, 58, 267–288.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012), "On Model Selection and Model Misspecification in Causal Inference," *Statistical Methods in Medical Research*, 21, 7–30.

Wolpert, D. (1992), "Stacked Generalization," *Neural Networks*, 5, 241–259.

Wood, S. N. (2001), "`mgcv`: GAMs and Generalized Ridge Regression for R," *R News*, 1, 20–25.

— (2006), *Generalized Additive Models: An Introduction with R*, Boca Raton, Florida: Chapman and Hall/CRC.

Yang, Y. (2001), "Adaptive Regression by Mixing," *Journal of the American Statistical Association*, 96, 574–588.

— (2003), "Regression with Multiple Candidate Models: Selecting or Mixing?" *Statistica Sinica*, 13, 783–809.

— (2004), "Combining Forecasting Procedures: Some Theoretical Results," *Econometric Theory*, 20, 176–222.

— (2005), "Can the Strengths of AIC and BIC Be Shared? A Conflict Between Model Identification and Regression Estimation," *Biometrika*, 92, 937–950.

— (2007), "Consistency of Cross Validation for Comparing Regression Procedures," *The Annals of Statistics*, 35, 2450–2473.

— (2008), "Localized Model Selection for Regression," *Econometric Theory*, 24, 472–492.

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942.

Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L. J. (2013), "Effectively Selecting a Target Population for a Future Comparative Study," *Journal of the American Statistical Association*, 108, 527–539.

Appendix A

Proof of TECV Selection Consistency

This appendix presents a detailed proof of Theorem 1 in Section 4.3.2. Conditions (a) through (h) referenced in this section are enumerated in Section 4.3.2. We first discuss two lemmas used in the proof of this theorem.

A.1 Lemma A.1: Bias Term in TECV

Our TECV algorithm's solution to the lack of observed individual treatment effects is to estimate individual treatment effects via a matching approach. Specifically, for a given observation j_t within the treatment group we find an observation j_c within the control group with similar covariate values and create approximate individual treatment effects $\widetilde{\delta}_j := Y_{j_t} - Y_{j_c}$.

In general, we can expect this matching to cause bias for $\widetilde{\delta}_j$ as an estimator of $\Delta(\mathbf{U}_{j_t})$. Each approximate treatment effect $\widetilde{\delta}_j$ has expectation $f_t(\mathbf{U}_{j_t}) - f_c(\mathbf{U}_{j_c})$ or, expressed another way, $\Delta(\mathbf{U}_{j_t}) + f_c(\mathbf{U}_{j_t}) - f_c(\mathbf{U}_{j_c})$. Thus the bias of $\widetilde{\delta}_j$ as an estimator of $\Delta(\mathbf{U}_{j_t})$ is $f_c(\mathbf{U}_{j_t}) - f_c(\mathbf{U}_{j_c})$, the difference between the regression function under control at $\mathbf{U}_{j_t}$ and at $\mathbf{U}_{j_c}$.

In order for TECV to be effective, the bias terms resulting from the matching need to be suitably controlled. The following lemma describes how the support of the covariate space can be partitioned so that as the size of the evaluation set goes to infinity, the number of pairs for model evaluation goes to infinity (in probability) while the bias terms are uniformly bounded (in probability) by a sequence that tends to zero.

Lemma A.1 *A partition of* $[0,1]^p$ *can be constructed such that Step 3 of the TECV algorithm will produce a set of* $\widetilde{n}_2$ *pairs* $(j_t, j_c), 1 \leq j \leq \widetilde{n}_2$ *such that as* $n_2 \to \infty$, $\widetilde{n}_2 \xrightarrow{p} \infty$ *and*

$$\sup_{1 \leq j \leq \widetilde{n}_2} |f_c(\mathbf{U}_{j_t}) - f_c(\mathbf{U}_{j_c})| = O_p \left\{ \left(\frac{\log n_2}{n_2} \right)^{1/p} \right\}.$$

Proof Let h denote the side length of each cell (hypercube) resulting from the partition of $[0,1]^p$. The basic idea is to construct a sequence h_{n_2} that converges to 0 at an appropriate rate. As the size of each cell shrinks, the number of cells will grow. We will show that if h_{n_2} is suitably chosen, a treatment-control pair will be found in all of the cells with high probability, yielding the first result. Meanwhile, the shrinking of each cell, combined with the smoothness condition on the regression functions, provides the uniform bound in probability on the magnitude of the bias terms.

Let n_{t_2} be the number of observations in the evaluation set for which $T = t$ and n_{c_2} be the corresponding number of control observations. By condition (c), the probability that any observation from the treatment group falls into a given cell is at least $\underline{c}h^p$. Since the covariate values of the n_{t_2} treatment observations are i.i.d., the probability that at least one of the treatment observations falls into a given cell is at least

$$1 - (1 - \underline{c}h^p)^{n_{t_2}} = 1 - e^{n_{t_2} \log(1 - \underline{c}h^p)} \geq 1 - e^{-\underline{c}n_{t_2}h^p},$$

where the last inequality results from the fact that $\log x \leq x - 1$. Denote by $B_{n_2,t}$ the event that all cells in our partition contain at least one observation from the treatment group. Since there are $(1/h)^p$ such cells, we have

$$P(B_{n_2,t}) \geq 1 - (1/h)^p e^{-\underline{c}n_{t_2}h^p} = 1 - \exp[-\{\underline{c}n_{t_2}h^p - p\log(1/h)\}].$$

In order for $B_{n_2,t}$ to happen with probability tending to 1, we need

$$\underline{c}n_{t_2}h^p - p\log(1/h) \to \infty. \tag{A.1}$$

Let $n_2^* = \min(n_{t_2}, n_{c_2})$. Note that the expression in (A.1) is an increasing function of h

and that $(1/h)$ must be an integer. So consider

$$\frac{1}{h} = \left\lfloor \left(\frac{\underline{c} n_2^*}{\log n_2^*} \right)^{1/p} \right\rfloor .$$

Then we have $h \geq \{\log(n_2^*)/\underline{c} n_2^*\}^{1/p}$. With this choice of h,

$$\begin{aligned} \underline{c} n_{t_2} h^p - p \log \left(\frac{1}{h} \right) &\geq \frac{n_{t_2} \log n_2^*}{n_2^*} - \log(\underline{c} n_2^*) + \log\log n_2^* \\ &= \left(\frac{n_{t_2}}{n_2^*} - 1 \right) \log n_2^* - \log \underline{c} + \log\log n_2^*. \end{aligned}$$

As $n_2 \to \infty$, $n_2^* \to \infty$; therefore, by the above expression (A.1) holds and thus $P(B_{n_2,t}) \to 1$. A similar calculation can be done for the control group to show that $P(B_{n_2,c}) \to 1$, where $B_{n_2,c}$ is the event that all cells contain at least one observation from the control group. Thus $P(B_{n_2,t} \cap B_{n_2,c}) \to 1$.

Conditional on $B_{n_2,t} \cap B_{n_2,c}$, the number of pairs $\widetilde{n}_2$ generated by this pairing algorithm is

$$\widetilde{n}_2 = \left(\frac{1}{h} \right)^p = \left\{ \left\lfloor \left(\frac{\underline{c} n_2^*}{\log n_2^*} \right)^{1/p} \right\rfloor \right\}^p .$$

Thus we can conclude $\widetilde{n}_2 \xrightarrow{p} \infty$ as $n_2 \to \infty$.

Step 3 of the TECV algorithm involves randomly selecting one treatment observation and one control observation from within each cell. Since all pairs (j_t, t_c) used for model evaluation reside in the same cell, $\mathbf{U}_{j_t}$ will be approximately equal to $\mathbf{U}_{j_c}$. Formally, letting $d(\cdot)$ represent the Euclidean distance, we have

$$\sup_{1 \leq j \leq \widetilde{n}_2} d(\mathbf{U}_{j_t}, \mathbf{U}_{j_c}) \leq \sqrt{p} h = \sqrt{p} \left\{ \left\lfloor \left(\frac{\underline{c} n_2^*}{\log n_2^*} \right)^{1/p} \right\rfloor \right\}^{-1} .$$

From condition (e), there exists a constant L such that all partial derivatives of f_c on $[0,1]^p$ are bounded by L. The Mean Value Theorem and the Cauchy-Schwarz Inequality can be used to show that f_c satisfies a Lipschitz condition with Lipschitz constant $\sqrt{p}L$.

That is, $d(\mathbf{U}_{j_t}, \mathbf{U}_{j_c}) \le x$ implies $|f_c(\mathbf{U}_{j_t}) - f_c(\mathbf{U}_{j_c})| \le \sqrt{p}Lx$. Thus we have

$$\sup_{1 \le j \le \tilde{n}_2} |f_c(\mathbf{U}_{j_t}) - f_c(\mathbf{U}_{j_c})| \le pL \left\{ \left\lfloor \left(\frac{\underline{c} n_2^*}{\log n_2^*} \right)^{1/p} \right\rfloor \right\}^{-1}.$$

If we can show that n_2^* and n_2 are of the same order in probability, then

$$pL \left\{ \left\lfloor \left(\frac{\underline{c} n_2^*}{\log n_2^*} \right)^{1/p} \right\rfloor \right\}^{-1} = O_p \left\{ \left(\frac{\log n_2}{n_2} \right)^{1/p} \right\}$$

and the second result of Lemma A.1 is obtained. Since n_2^*/n_2 is upper bounded by 0.5, it suffices to show there exists $\rho \in (0,1)$ such that $P(n_2^* \le \rho n_2) \to 0$ as $n_2 \to \infty$.

The random variable n_{t_2}, representing the number of observations in the evaluation set for which $T_i = t$, follows a hypergeometric distribution with population size n, sample size n_2, and number of treatment observations n_t. Let f_t denote the fraction of all observations for which $T_i = t$; that is, $f_t = n_t/n$. Applying the bound provided by Chvátal (1979) for the upper tail probability of the hypergeometric distribution, we have

$$P\{n_{t_2} \ge n_2(f_t + s)\} \le e^{-2s^2 n_2} \qquad \text{for all } s \ge 0.$$

By condition (d), there exist constants a and b such that $0 < a < f_t < b < 1$ for n large enough. Let $s = (1-b)/2$. Then $n_2(f_t + s) < n_2\{(b+1)/2\}$, so

$$P[n_{t_2} \ge n_2\{(b+1)/2\}] \le e^{-0.5(1-b)^2 n_2} \to 0 \text{ as } n_2 \to \infty.$$

Next we apply the corresponding bound on the lower tail of n_{t_2}, which can be obtained by bounding the upper tail of n_{c_2}. For $\tilde{s} \ge 0$, we obtain

$$P\{n_{t_2} \le n_2(f_t - \tilde{s})\} \le e^{-2\tilde{s}^2 n_2}.$$

Let $\tilde{s} = a/2$. Then $n_2(f_t - \tilde{s}) > n_2(a - \tilde{s}) = n_2(a/2)$, so

$$P\{n_{t_2} \le n_2(a/2)\} \le e^{-0.5a^2 n_2}.$$

Since both tail probabilities go to 0, we have $P\{a/2 < n_{t_2}/n_2 < (b+1)/2\} \to 1$.

Equivalently, $P\{(1-b)/2 < n_{c_2}/n_2 < 1-a/2\} \to 1$ since $n_{c_2} = n_2 - n_{t_2}$. Now let $\rho = \min\{a/2, (1-b)/2\}$. Then $\rho \in (0,1)$ and

$$P(n_2^* \leq \rho n_2) \leq e^{-0.5(1-b)^2 n_2} + e^{-0.5a^2 n_2} \to 0 \text{ as } n_2 \to \infty.$$

This completes the proof of Lemma A.1.

A.2 Lemma A.2: Size of the W_j

The partition of Lemma A.1 ensures that as n_2 grows, the volume of each cell shrinks and the number of cells increases. This behavior, combined with the upper bounds assumed for the densities $P_{\mathbf{U}_t}$ and $P_{\mathbf{U}_c}$, ensures that having a very large number of observations in any one cell is unlikely. Let $W_j, j = 1, \ldots, \widetilde{n}_2$ denote the number of evaluation observations in each cell; these are the cell weights in the TECV statistic. Using the partition described in Lemma A.1, each cell will be expected to contain on the order of $\log n_2$ observations. Moreover, the following lemma shows that with high probability, the supremum of the W_j values is of the order $\log n_2$ in probability.

Lemma A.2 *Applying the partition described in Lemma A.1, with*

$$h = \left\{ \left\lfloor \left(\frac{\underline{c} n_2^*}{\log n_2^*} \right)^{1/p} \right\rfloor \right\}^{-1},$$

we have

$$\sup_{1 \leq j \leq \widetilde{n}_2} W_j = O_p(\log n_2).$$

Proof By condition (c), $\underline{c}$ and $\overline{c}$ are lower and upper bounds, respectively, on the density of $\mathbf{U}$. Recall that each cell in the partition of Lemma A.1 is a hypercube with side length h. The W_j represent the number of observations in cell j from the evaluation set. Each W_j is binomial with n_2 trials and success $p_j \in [\underline{c}h^p, \overline{c}h^p]$. Let $\overline{p}$ denote the supremum of the success probabilities over the $\widetilde{n}_2$ cells for which at least one treatment and one control observation is found. Using the formula for h given in Lemma A.1,

$$\overline{p} \leq \overline{c} \left\{ \left\lfloor \left(\frac{\underline{c} n_2^*}{\log n_2^*} \right)^{1/p} \right\rfloor \right\}^{-p} \leq C \frac{\log n_2^*}{n_2^*},$$

for some positive constant C depending on $\underline{c}$ and $\bar{c}$. Thus for arbitrary $\beta > 0$,

$$P\left\{\sup_{1\le j\le \widetilde{n}_2} W_j \ge C(1+\beta)\log n_2^*\right\} \le P\left\{\max_{1\le j\le \widetilde{n}_2} W_j \ge n_2^*\bar{p}(1+\beta)\right\}.$$

Since $n_2^* \to \infty$ as $n_2 \to \infty$ (by the hypergeometric argument used in the proof of Lemma A.1), it suffices to show there exists $\beta > 0$ for which the upper bound of the above expression goes to 0 as $n_2^* \to \infty$.

Let j represent the index for an arbitrary cell. By Bernstein's inequality (see, e.g., Pollard (1984), p. 193), treating W_j as the sum of independent Bernoulli(p_j), we have, for arbitrary $\beta > 0$,

$$P\{W_j \ge n_2^* p_j(1+\beta)\} \le \exp\left\{\frac{-(n_2^* p_j \beta)^2}{2(n_2^* p_j + \beta n_2^* p_j/3)}\right\} = \exp\left\{\frac{-n_2^* p_j \beta^2}{2(1+\beta/3)}\right\}. \tag{A.2}$$

We know

$$p_j \ge \underline{c}\left\{\left\lfloor\left(\frac{\underline{c} n_2^*}{\log n_2^*}\right)^{1/p}\right\rfloor\right\}^{-p} \ge \frac{\log n_2^*}{n_2^*}.$$

Therefore, using (A.2),

$$P\{W_j \ge n_2^* p_j(1+\beta)\} \le \exp\left\{\log(n_2^*)\frac{-\beta^2}{2(1+\beta/3)}\right\} = n_2^{*\,-\beta^2/\{2(1+\beta/3)\}}. \tag{A.3}$$

Since $p_j \le \bar{p}$, the upper bound in (A.3) also holds for $P\{W_j \ge n_2^*\bar{p}(1+\beta)\}$. Since j is arbitrary,

$$P\left\{\sup_{1\le j\le \widetilde{n}_2} W_j \ge n_2^*\bar{p}(1+\beta)\right\} \le \widetilde{n}_2 n_2^{*\,-\beta^2/\{2(1+\beta/3)\}} \le n_2^{*\,1-\beta^2/\{2(1+\beta/3)\}}. \tag{A.4}$$

Choose any $\beta > (1+\sqrt{19})/3$. Then the upper bound in (A.4) goes to 0 as $n_2^* \to \infty$. This completes the proof of Lemma A.2.

A.3 Proof of Theorem 1

Theorem 1 is formulated in Section 4.3.2 of this document.

Proof We first show that selection consistency holds for a single permutation; that is,

$|\Pi| = 1$. At the end of the proof, we will show that the consistency result extends to any nonempty collection of permutations.

Without loss of generality, assume that ϕ_1 is the asymptotically better procedure by condition (a). Let π denote any random permutation. Apply the permutation π, and let $(j_t, j_c)_{j=1}^{\widetilde{n}_2}$ index the pairs of observations from the evaluation set that are selected by the pairing algorithm. Let $b(\mathbf{U}_j)$ denote $f_c(\mathbf{U}_{j_t}) - f_c(\mathbf{U}_{j_c})$, the bias of $\widetilde{\delta}_j$ as an estimator of $\Delta(\mathbf{U}_{j_t})$. Let W_j denote the total number of observations (including treatment and control) in bin j. Then we have

$$\begin{aligned}
\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,k}) &= \sum_{j=1}^{\widetilde{n}_2} W_j\{\widetilde{\delta}_j - \widehat{\Delta}_{n_1,k}(\mathbf{U}_{j_t})\}^2 \\
&= \sum_{j=1}^{\widetilde{n}_2} W_j[\{f_t(\mathbf{U}_{j_t}) + \xi_{j_t}\} - \{f_c(\mathbf{U}_{j_c}) + \nu_{j_c}\} - \widehat{\Delta}_{n_1,k}(\mathbf{U}_{j_t})]^2 \\
&= \sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) + \xi_{j_t} - \nu_{j_c} - \widehat{\Delta}_{n_1,k}(\mathbf{U}_{j_t})\}^2 \\
&= \Bigg[\sum_{j=1}^{\widetilde{n}_2} W_j(\xi_{j_t} - \nu_{j_c})^2 + \sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,k}(\mathbf{U}_{j_t})\}^2 \\
&\qquad + 2\sum_{j=1}^{\widetilde{n}_2} W_j(\xi_{j_t} - \nu_{j_c})\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,k}(\mathbf{U}_{j_t})\}\Bigg].
\end{aligned}$$

$\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,1}) \geq \mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,2})$ is thus equivalent to

$$\begin{aligned}
&2\sum_{j=1}^{\widetilde{n}_2} W_j\{\xi_{j_t} - \nu_{j_c})(\widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_{j_t})\} \\
&\geq \Bigg[\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2 \\
&\qquad - \sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_{j_t})\}^2\Bigg]. \qquad (A.5)
\end{aligned}$$

Note that the error terms ξ_{j_t} and ν_{j_c} are independent, so the variance of their difference is $\sigma_t^2 + \sigma_c^2$. Then conditional on the estimation data (which we denote as $Z^{(1)}$) and the covariate values of the evaluation data (denoted as $U^{(2)}$), and assuming the right-hand side of the inequality in (A.5) is positive, by Chebyshev's inequality we have

$$
\begin{aligned}
&P\{\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,1}) \geq \mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,2}) | Z^1, U^2\} \\
&\leq \min\Bigg\{1, 4(\sigma_t^2 + \sigma_c^2) \sum_{j=1}^{\widetilde{n}_2} W_j^2 \{\widehat{\Delta}_{n_1,2}(\mathbf{U}_{jt}) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_{jc})\}^2 \\
&\qquad \times \Bigg(\Bigg[\sum_{j=1}^{\widetilde{n}_2} W_j \{\Delta(\mathbf{U}_{jt}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{jt})\}^2 \\
&\qquad\qquad - \sum_{j=1}^{\widetilde{n}_2} W_j \{\Delta(\mathbf{U}_{jt}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_{jt})\}^2 \Bigg]^2 \Bigg)^{-1} \Bigg\}.
\end{aligned}
$$

Let Q_n denote the ratio in the upper bound in the preceding inequality, and let S_n be the event that the right-hand side of the inequality in (A.5) is positive. Then

$$
\begin{aligned}
&P\{\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,1}) \geq \mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,2})\} \\
&= \Big(P[\{\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,1}) \geq \mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,2})\} \cap S_n] \\
&\qquad + P[\{\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,1}) \geq \mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,2})\} \cap S_n^c] \Big) \\
&\leq E[P(\mathrm{TECV}_\pi\{\widehat{\Delta}_{n_1,1}) \geq \mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,2}) | Z^1, U^2\} I_{S_n}] + P(S_n^c) \\
&\leq E\{\min(1, Q_n)\} + P(S_n^c).
\end{aligned}
$$

If we can show that $P(S_n^c) \to 0$ and $Q_n \xrightarrow{p} 0$ as $n \to \infty$, then due to the boundedness of $\min(1, Q_n)$ (which implies that the random variables $\min(1, Q_n)$ are uniformly integrable), we have

$$
P\{\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,1}) \geq \mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,2})\} \to 0 \quad \text{as } n \to \infty,
$$

from which selection consistency follows for the single permutation pair π.

Let $\epsilon > 0$ be arbitrary. Suppose we can show that there exists $\alpha_\epsilon > 0$ such that when n is large enough,

$$P\left[\frac{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2}{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_{j_t})\}^2} \geq 1+\alpha_\epsilon\right] \geq 1-\epsilon. \tag{A.6}$$

The probability in (A.6) is at most $P(S_n)$, so that would imply $P(S_n^c) \to 0$ as $n \to \infty$.

The statement in (A.6) can also be used to show $Q_n \xrightarrow{p} 0$, as follows. Note that by the triangle inequality,

$$\begin{aligned}
&\sum_{j=1}^{\widetilde{n}_2} W_j^2\{\widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_{j_t})\}^2 \\
&\leq 2\sum_{j=1}^{\widetilde{n}_2} W_j^2\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_{j_t})\}^2 \\
&\quad + 2\sum_{j=1}^{\widetilde{n}_2} W_j^2\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2.
\end{aligned}$$

By Lemma A.2, the weights W_j are uniformly $O_p(\log n_2)$. Thus there exists a constant M_ϵ such that when n_2 is large enough, $P\{\sup_j W_j \leq M_\epsilon \log n_2\} \geq 1-\epsilon$. Using this inequality and supposing we can show (A.6), then we can conclude that with probability no less than $1-2\epsilon$, Q_n is upper bounded by

$$\begin{aligned}
&\frac{8(\sigma_t^2+\sigma_c^2)M_\epsilon \log n_2}{[\{1-1/(1+\alpha_\epsilon)\}\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2]^2} \\
&\times\left(\sum_{j=1}^{\widetilde{n}_2}\left[W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_{j_t})\}^2\right.\right. \\
&\qquad\left.\left. + W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2\right]\right) \\
&\leq \frac{16(\sigma_t^2+\sigma_c^2)M_\epsilon \log n_2\{1+1/(1+\alpha_\epsilon)\}}{\{1-1/(1+\alpha_\epsilon)\}^2\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2}.
\end{aligned}$$

So to show $P(S_n^c) \to 0$ and $Q_n \xrightarrow{p} 0$ as $n \to \infty$, it suffices to show (A.6) and

$$\frac{\log n_2}{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2} = o_p(1). \tag{A.7}$$

We will essentially show that for each estimator $k = 1, 2$,

$$\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,k}(\mathbf{U}_{j_t})\}^2 \approx \sum_{i=1}^{n_2}\{\Delta(\mathbf{U}_i) - \widehat{\Delta}_{n_1,k}(\mathbf{U}_i)\}^2. \tag{A.8}$$

That is, the bias terms are negligible, and for each procedure, the weighted sum of the errors of the single representatives from each bin is a good approximation of the errors of an i.i.d. sample.

Suppose we can show that with high probability, there exists $\widetilde{\alpha} > 0$ such that when n is large enough,

$$\frac{\sum_{i=1}^{n_2}\{\Delta(\mathbf{U}_i) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_i)\}^2}{\sum_{i=1}^{n_2}\{\Delta(\mathbf{U}_i) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_i)\}^2} \geq 1 + \widetilde{\alpha}. \tag{A.9}$$

Then (A.9), together with the approximations we will formalize in (A.8), will give us the result in (A.6).

To show (A.7), we will apply the approximation in (A.8) to the estimator $\widehat{\Delta}_{n_1,2}$ to show that the denominator of (A.7) grows faster than $\log n_2$.

Now we begin to provide specifics. First we show

$$\frac{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2}{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2} \xrightarrow{p} 1. \tag{A.10}$$

The ratio in (A.10) is

$$\begin{aligned}
&\Big(\textstyle\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2 + \sum_{j=1}^{\widetilde{n}_2} W_j b^2(\mathbf{U}_j)\\
&\qquad + 2\textstyle\sum_{j=1}^{\widetilde{n}_2} W_j b(\mathbf{U}_j)\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}\Big)\\
&\qquad \times \frac{1}{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2}\\
&= 1 + \frac{\sum_{j=1}^{\widetilde{n}_2} W_j b^2(\mathbf{U}_j)}{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2} + 2\frac{\sum_{j=1}^{\widetilde{n}_2} W_j b(\mathbf{U}_j)\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}}{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2}.
\end{aligned}$$

So to show (A.10) it suffices to show

$$\frac{\sum_{j=1}^{\widetilde{n}_2} W_j b^2(\mathbf{U}_j)}{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2} \xrightarrow{p} 0 \tag{A.11}$$

and

$$\frac{\sum_{j=1}^{\widetilde{n}_2} W_j b(\mathbf{U}_j)\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}}{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2} \xrightarrow{p} 0. \tag{A.12}$$

By the Cauchy-Schwarz Inequality,

$$\begin{aligned}&\left|\sum_{j=1}^{\widetilde{n}_2} W_j b(\mathbf{U}_j)\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}\right| \\ &\leq \sqrt{\left(\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2\right) \times \left(\sum_{j=1}^{\widetilde{n}_2} W_j b^2(\mathbf{U}_j)\right)},\end{aligned}$$

so the fraction in (A.12) is bounded in absolute value by

$$\sqrt{\frac{\sum_{j=1}^{\widetilde{n}_2} W_j b^2(\mathbf{U}_j)}{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2}},$$

which is simply the square root of the fraction in (A.11). Therefore, to show (A.10) it suffices to show

$$\frac{\sum_{j=1}^{\widetilde{n}_2} W_j b^2(\mathbf{U}_j)}{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2} \xrightarrow{p} 0. \tag{A.13}$$

Each weight W_j is a count of the observations in each bin, so the denominator in (A.13) can be written as the following double sum:

$$\sum_{j=1}^{\widetilde{n}_2} \sum_{i \in \text{bin } j} \{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2.$$

Denote this sum as $\widetilde{S}$, and let S denote the corresponding sum over all observations in the evaluation set:

$$S := \sum_{i=1}^{n_2} \{\Delta(\mathbf{U}_i) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_i)\}^2$$

It was shown in the proof of Lemma A.1 that with our choice of bin width, each cell will contain at least one treatment observation and at least one control observation with probability tending to 1. Here we denote the event that each cell contains at least one observation from each group as B_{n_2}. Conditional on B_{n_2},

$$S = \sum_{j=1}^{\widetilde{n}_2} \sum_{i \in \text{bin } j} \{\Delta(\mathbf{U}_i) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_i)\}^2.$$

We will show that conditional on B_{n_2}, $\widetilde{S}/S \xrightarrow{p} 1$, so that to show (A.10) it will suffice to show

$$\frac{\sum_{j=1}^{\widetilde{n}_2} W_j b^2(\mathbf{U}_j)}{\sum_{i=1}^{n_2}\{\Delta(\mathbf{U}_i) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_i)\}^2} \xrightarrow{p} 0.$$

Note that showing $\widetilde{S}/S \xrightarrow{p} 1$ is equivalent to showing $(S - \widetilde{S})/S \xrightarrow{p} 0$. Observing that, conditional on B_{n_2}, S and $\widetilde{S}$ are both the sum of n_2 terms, and using the basic algebraic fact that $a^2 - b^2 = (a-b)(a+b)$, conditional on B_{n_2} we can write $S - \widetilde{S}$ as

$$\begin{aligned}\sum_{j=1}^{\widetilde{n}_2} \sum_{i \in \text{bin } j} &\left[\left\{\Delta(\mathbf{U}_{j_t}) - \Delta(\mathbf{U}_i) + \widehat{\Delta}_{n_1,2}(\mathbf{U}_i) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\right\}\right. \\ &\quad \left. \times \left\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t}) + \Delta(\mathbf{U}_i) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_i)\right\}\right]. \end{aligned} \tag{A.14}$$

Due to the smoothness conditions on Δ and $\widehat{\Delta}_{n_1,2}$, the Mean Value Theorem and the Cauchy-Schwarz Inequality can be used to show that both are Lipschitz. That is, for any $\delta > 0$, there exist constants M_1 and M_2 such that $d(x, y) \leq \delta$ implies

$$\begin{aligned}&|\Delta(\mathbf{U}_x) - \Delta(\mathbf{U}_y)| \leq \delta M_1 \text{ and} \\ &|\widehat{\Delta}_{n_1,2}(\mathbf{U}_x) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_y)| \leq \delta M_2,\end{aligned}$$

where $d(\cdot)$ represents the Euclidean distance. From the partitioning and pairing scheme, for every i in bin j, we have

$$d(\mathbf{U}_{j_t}, \mathbf{U}_i) \leq \sqrt{p} L \left(\frac{\log n_2}{n_2}\right)^{1/p},$$

for some constant L. Letting $\widetilde{L} = \max(\sqrt{p} L M_1, \sqrt{p} L M_2)$, we can then use the Lipschitz properties of Δ and $\widehat{\Delta}_{n_1,2}$ to conclude that, for every j_t that represents the treatment representative from the cell in which observation i resides,

$$|\Delta(\mathbf{U}_{j_t}) - \Delta(\mathbf{U}_i)| \leq \widetilde{L}\left(\frac{\log n_2}{n_2}\right)^{1/p} \tag{A.15}$$

and

$$|\widehat{\Delta}_{n_1,2}(\mathbf{U}_i) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})| \leq \widetilde{L}\left(\frac{\log n_2}{n_2}\right)^{1/p}. \tag{A.16}$$

Now we derive a lower bound for S using Bernstein's Inequality. By condition (g), there exists a sequence $A_{n_1,\epsilon}$ such that for n_1 large enough, $P(\|\Delta-\widehat{\Delta}_{n_1,k}\|_\infty \geq A_{n_1,\epsilon}) \leq \epsilon$ for $k = 1, 2$. Let H_{n_1} be the event $\{\max(\|\Delta - \widehat{\Delta}_{n_1,1}\|_\infty, \|\Delta - \widehat{\Delta}_{n_1,2}\|_\infty) \leq A_{n_1,\epsilon}\}$. Then conditional on H_{n_1}, the other part of $S - \widetilde{S}$ can be bounded by $A_{n_1,\epsilon}$. That is,

$$\left|\{\Delta(\mathbf{U}_{j_t}) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\} + \{\Delta(\mathbf{U}_i) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_i)\}\right| \leq 2A_{n_1,\epsilon}. \tag{A.17}$$

Therefore, conditional on H_{n_1} and B_{n_2}, using (A.14), (A.15), (A.16), (A.17), and the triangle inequality, we have

$$|S - \widetilde{S}| \leq 4n_2\widetilde{L}A_{n_1,\epsilon}\left(\frac{\log n_2}{n_2}\right)^{1/p}. \tag{A.18}$$

Again conditional on H_{n_1}, we have $V_i := \{\Delta(\mathbf{U}_i) - \widehat{\Delta}_{n_1,k}(\mathbf{U}_i)\}^2 - \|\Delta - \widehat{\Delta}_{n_1,k}\|_2^2$ is bounded between $-(A_{n_1,\epsilon})^2$ and $(A_{n_1,\epsilon})^2$ for $k = 1, 2$. Conditional on the estimation data $Z^{(1)}$ and the event H_{n_1}, for the $i = 1, \ldots, n_2$ observations in the evaluation data, we have

$$\mathrm{Var}_{Z^{(1)}}(V_i) \leq E_{Z^1}\{\Delta(\mathbf{U}_i) - \widehat{\Delta}_{n_1,k}(\mathbf{U}_i)\}^4 = \|\Delta - \widehat{\Delta}_{n_1,k}\|_4^4,$$

where the subscript $Z^{(1)}$ denotes the conditional expectation given $Z^{(1)}$. Since S is the sum of n_2 i.i.d. terms, each with mean $\|\Delta - \widehat{\Delta}_{n_1,2}\|_2^2$, on H_{n_1} we can apply Bernstein's Inequality to obtain the following for all $0 < \beta < 1$:

$$\begin{aligned} &P_{Z^1}\{S \leq (1-\beta)n_2\|\Delta - \widehat{\Delta}_{n_1,2}\|_2^2\} \\ &\quad \leq \exp\left\{-\frac{1}{2}\frac{(\beta n_2\|\Delta - \widehat{\Delta}_{n_1,2}\|_2^2)^2}{n_2\|\Delta - \widehat{\Delta}_{n_1,2}\|_4^4 + (A_{n_1,\epsilon}^2/3)(\beta n_2\|\Delta - \widehat{\Delta}_{n_1,2}\|_2^2)}\right\}. \end{aligned}$$

If we have

$$\frac{n_2\|\Delta - \widehat{\Delta}_{n_1,2}\|_2^4}{\|\Delta - \widehat{\Delta}_{n_1,2}\|_4^4} \to \infty \quad \text{in probability} \tag{A.19}$$

and

$$\frac{n_2\|\Delta - \widehat{\Delta}_{n_1,2}\|_2^2}{(A_{n_1,\epsilon})^2} \to \infty \quad \text{in probability,} \tag{A.20}$$

then the upper bound in the last inequality above converges to zero in probability.

To show (A.19), note that by condition (h), for n_1 large enough, $\|\Delta - \widehat{\Delta}_{n_1,2}\|_2^4/\|\Delta - \widehat{\Delta}_{n_1,2}\|_4^4$ is lower bounded in probability by $M_{n_1}^{-4}$ times a constant. By the second condition on the data splitting, $n_2 M_{n_1}^{-4} \to \infty$, so (A.19) follows.

Because ϕ_2 converges at rate q_n in probability, for n_1 large enough the expression in (A.20) is lower bounded in probability by $n_2 c_\epsilon q_{n_1}^2 A_{n_1,\epsilon}^{-2}$. By condition (g), when n_1 is large enough we can take $A_{n_1,\epsilon} = O(A_{n_1})$ so that H_{n_1} occurs with probability at least $1-\epsilon$. Therefore, with probability at least $1-2\epsilon$ for n_1 large enough, the expression in (A.20) is lower bounded by $Cn_2 q_{n_1}^2/A_{n_1}^{-2}$ for some constant C. The third condition on the data splitting implies $(n_2/\log n_2)^{1/p} q_{n_1}^2/A_{n_1}^{-2} \to \infty$, so $n_2 q_{n_1}^2/A_{n_1}^{-2} \to \infty$ and (A.20) holds.

Thus for n large enough, conditional on H_{n_1} and B_{n_2},

$$P_{\{Z^{(1)},W^{(2)}\}}\left\{\frac{|S-\widetilde{S}|}{S} \le \frac{4\widetilde{L}A_{n_1,\epsilon}\left(\frac{\log n_2}{n_2}\right)^{1/p}}{\|\Delta - \widehat{\Delta}_{n_1,2}\|_2^2(1-\beta)}\right\} \ge 1-\epsilon,$$

where the subscript $\{Z^{(1)}, W^{(2)}\}$ denotes the conditional probability given $Z^{(1)}$ and $W^{(2)} := (\mathbf{U}_i, T_i)_{i=n_1+1}^n$. Denote the event $\|\Delta - \widehat{\Delta}_{n_1,2}\|_2^2 \ge c_\epsilon q_{n_1}$ as D_{n_1}. Putting the pieces together, since $n_1 \to \infty$ and $n_2 \to \infty$ as $n \to \infty$, for n large enough, we have

$$
\begin{aligned}
&P\left\{ \frac{|S-\widetilde{S}|}{S} \geq \frac{4\widetilde{L}A_{n_1,\epsilon}\left(\frac{\log n_2}{n_2}\right)^{1/p}}{(1-\beta)c_\epsilon^2 q_{n_1}^2} \right\} \\
&\leq P(H_{n_1}^c) + P(D_{n_1}^c) + P(B_{n_2}^c) \\
&\qquad + P\left[H_{n_1} \cap D_{n_1} \cap B_{n_2} \cap \left\{ \frac{|S-\widetilde{S}|}{S} \geq \frac{4\widetilde{L}A_{n_1,\epsilon}\left(\frac{\log n_2}{n_2}\right)^{1/p}}{(1-\beta)c_\epsilon^2 q_{n_1}^2} \right\} \right] \\
&\leq P(H_{n_1}^c) + P(D_{n_1}^c) + P(B_{n_2}^c) \\
&\qquad + P\left[H_{n_1} \cap D_{n_1} \cap B_{n_2} \cap \left\{ \frac{|S-\widetilde{S}|}{S} \geq \frac{4\widetilde{L}A_{n_1,\epsilon}\left(\frac{\log n_2}{n_2}\right)^{1/p}}{(1-\beta)\|\Delta - \widehat{\Delta}_{n_1,2}\|_2^2} \right\} \right] \\
&\leq 3\epsilon + EP\left[H_{n_1} \cap D_{n_1} \cap B_{n_2} \cap \left\{ \frac{|S-\widetilde{S}|}{S} \geq \frac{4\widetilde{L}A_{n_1,\epsilon}\left(\frac{\log n_2}{n_2}\right)^{1/p}}{(1-\beta)\|\Delta - \widehat{\Delta}_{n_1,2}\|_2^2} \right\} \middle| Z^{(1)}, W^{(2)} \right] \\
&\leq 3\epsilon + E\exp\left\{ -\frac{1}{2} \frac{(\beta n_2 \|\Delta - \widehat{\Delta}_{n_1,2}\|_2^2)^2}{n_2\|\Delta - \widehat{\Delta}_{n_1,2}\|_4^4 + (A_{n_1,\epsilon}^2/3)(\beta n_2\|\Delta - \widehat{\Delta}_{n_1,2}\|_2^2)} \right\}.
\end{aligned}
$$

The expectation in the upper bound of the last inequality above converges to zero due to the convergence in probability to zero of the random variables of the exponential expression (provided that (A.19) and (A.20) hold) and their uniform integrability (since they are bounded above by 1). By the third condition on the data splitting, we also have that

$$
\frac{4\widetilde{L}A_{n_1,\epsilon}\left(\frac{\log n_2}{n_2}\right)^{1/p}}{(1-\beta)c_\epsilon^2 q_{n_1}^2} \to 0 \text{ as } n \to \infty.
$$

Thus we conclude that $|S-\widetilde{S}|/S \xrightarrow{p} 0$, which implies $\widetilde{S}/S \xrightarrow{p} 1$.

To show

$$
\frac{\sum_{j=1}^{\widetilde{n}_2} W_j b^2(\mathbf{U}_j)}{\sum_{i=1}^{n_2}\{\Delta(\mathbf{U}_i) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_i)\}^2} \xrightarrow{p} 0, \tag{A.21}
$$

we first use Lemma A.1 to observe that for some constant C_ϵ,

$$P\left\{\sum_{j=1}^{\widetilde{n}_2} W_j b^2(\mathbf{U}_j) \leq C_\epsilon n_2 \left(\frac{\log n_2}{n_2}\right)^{2/p}\right\} < \epsilon.$$

By arguments made previously in showing $\widetilde{S}/S \xrightarrow{p} 1$, $P\{S \leq (1-\beta)n_2 c_\epsilon^2 q_{n_1}^2\} < \epsilon$ when n is large enough. The assertion of (A.21) then follows because, by the conditions on the data splitting,

$$\lim_{n\to\infty} \frac{C_\epsilon \left(\frac{\log n_2}{n_2}\right)^{2/p}}{(1-\beta)c_\epsilon^2 q_{n_1}^2} = \lim_{n\to\infty} \frac{\log n_2}{n_2 q_{n_1}^p} = 0.$$

Having shown $\widetilde{S}/S \xrightarrow{p} 1$ and (A.21), the conclusion of (A.10) follows.

Now we work on the other estimator, $\widehat{\Delta}_{n_1,1}$. Let T and $\widetilde{T}$ denote the analogs to S and $\widetilde{S}$, respectively, for the estimator $\widehat{\Delta}_{n_1,1}$. That is,

$$\begin{aligned}\widetilde{T} &:= \sum_{j=1}^{\widetilde{n}_2} \sum_{i\in \text{bin } j} \{\Delta(\mathbf{U}_{jt}) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_{jt})\}^2 \text{ and}\\ T &:= \sum_{i=1}^{n_2} \{\Delta(\mathbf{U}_i) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_i)\}^2 = \sum_{j=1}^{\widetilde{n}_2} \sum_{i\in \text{bin } j} \{\Delta(\mathbf{U}_i) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_i)\}^2.\end{aligned}$$

Starting with the ratio in (A.6), we can use results shown previously to conclude that with probability at least $1-\epsilon$ for n large enough, the following three inequalities hold for any $\delta_1 > 0$, $\delta_2 > 0$, and $\delta_3 > 0$:

$$\begin{aligned}&\frac{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{jt}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{jt})\}^2}{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{jt}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_{jt})\}^2}\\ &\geq \frac{(1-\delta_1)\widetilde{S}}{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{jt}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_{jt})\}^2}\\ &\geq \frac{(1-\delta_1)(1-\delta_2)S}{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{jt}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_{jt})\}^2}\\ &\geq \frac{(1-\delta_1)(1-\delta_2)}{\widetilde{T}/S + \delta_3 + 2\sqrt{\delta_3}\sqrt{\widetilde{T}/S}}.\end{aligned} \tag{A.22}$$

The first inequality is from (A.10), and the second is from the fact that $\widetilde{S}/S \xrightarrow{p} 1$. The third inequality results from applying the Cauchy-Schwarz Inequality to upper bound $2\sum_{j=1}^{\widetilde{n}_2} W_j b(\mathbf{U}_j)\{\Delta(\mathbf{U}_{jt}) - \widehat{\Delta}_{n_1,1}(\mathbf{U}_{jt})\}$ by $2\sqrt{\sum_{j=1}^{\widetilde{n}_2} W_j b^2(\mathbf{U}_j)}\sqrt{\widetilde{T}}$ and using (A.13) to bound $\sum_{j=1}^{\widetilde{n}_2} W_j b^2(\mathbf{U}_j)/A$ in probability by δ_3 for n large enough.

If we can also show that

$$\frac{\widetilde{T} - T}{S} \xrightarrow{p} 0, \tag{A.23}$$

and that there exists an $\widetilde{\alpha}_\epsilon > 0$ such that

$$P\left(T/S < \frac{1}{1+\widetilde{\alpha}_\epsilon}\right) \geq 1 - \epsilon, \tag{A.24}$$

then with probability at least $1 - 2\epsilon$ for n large enough, for any $\delta_4 > 0$, we have

$$\frac{\widetilde{T}}{S} = \frac{\widetilde{T} - T}{S} + \frac{T}{S} < \delta_4 + \frac{1}{1+\widetilde{\alpha}_\epsilon}. \tag{A.25}$$

Thus (A.22) combined with (A.25) shows that with probability at least $1 - 5\epsilon$ for n large enough, the ratio in (A.6) is lower bounded by

$$\frac{(1+\widetilde{\alpha}_\epsilon)(1-\delta_1)(1-\delta_2)}{1 + (1+\widetilde{\alpha}_\epsilon)(\delta_4+\delta_3) + 2\sqrt{1+\widetilde{\alpha}_\epsilon}\sqrt{\delta_3\delta_4(1+\widetilde{\alpha}_\epsilon)+\delta_3}}. \tag{A.26}$$

Given an $\widetilde{\alpha}_\epsilon > 0$ from (A.24), we can choose $\delta_1, \delta_2, \delta_3$, and δ_4 so that the expression in (A.26) is greater than 1. Then take α_ϵ to be the expression in (A.26) minus one, and the conclusion of (A.6) follows. Thus to show (A.6), it suffices to show (A.23) and (A.24).

To show (A.23), we observe that, conditional on H_{n_1} and B_{n_2}, the bound on $|S - \widetilde{S}|$ found in (A.18) also serves as a bound on $|\widetilde{T} - T|$ by the conditions on Δ and $\widehat{\Delta}_{n_1,1}$. Therefore, the same argument used to show $(S - \widetilde{S})/S \xrightarrow{p} 0$ can be used to show (A.23).

The proof of (A.24) was done in Yang (2007) for general regression functions f as part of the proof of his Theorem 1. Recall that $\Delta := f_t - f_c$, so like f, Δ also can be thought of as a function of $\mathbf{U}$. The conditions on f and $\widehat{f}$ required in Yang (2007) are a subset of our conditions on Δ and $\widehat{\Delta}$, respectively. Our conditions on the data splitting ratio also imply the conditions of Yang (2007). Therefore, the logic used there can be directly applied to our situation to show (A.24).

Showing (A.7) is equivalent to showing

$$\frac{\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2}{\log n_2} \to \infty \quad \text{in probability.} \tag{A.27}$$

Using steps done earlier, for n large enough and arbitrary $\delta_1 > 0$, $\delta_2 > 0$, and $0 < \beta < 1$ and some $c_\epsilon > 0$, each of the following inequalities hold with probability at least $1 - \epsilon$:

$$\begin{aligned}
&\sum_{j=1}^{\widetilde{n}_2} W_j\{\Delta(\mathbf{U}_{j_t}) + b(\mathbf{U}_j) - \widehat{\Delta}_{n_1,2}(\mathbf{U}_{j_t})\}^2 \geq (1-\delta_1)\widetilde{S} \\
&\geq (1-\delta_1)(1-\delta_2)S \\
&\geq (1-\delta_1)(1-\delta_2)(1-\beta)n_2\|\Delta - \widehat{\Delta}_{n_1,2}\|_2^2 \\
&\geq (1-\delta_1)(1-\delta_2)(1-\beta)n_2 c_\epsilon q_{n_1}^2.
\end{aligned}$$

Therefore, to show (A.27) it suffices to have $(n_2/\log n_2)q_{n_1}^2 \to \infty$, which holds by the third condition on the data splitting because $p \geq 1$. Thus we have shown

$$P\{\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,1}) \leq \mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,2})\} \to 1 \text{ as } n \to \infty.$$

Finally, we generalize the result to any nonempty collection of permutations Π. Let W denote the values of $(Y_i, T_i, \mathbf{U}_i)_{i=1}^n$, ignoring the orders. Let $\tau_\pi = I\{\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,1}) \leq \mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,2})\}$. Conditional on W, every ordering of these values has the same probability under the i.i.d. assumptions. Since π is arbitrary, we have

$$\begin{aligned}
&P\{\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,1}) \leq \mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,2})\} \\
&\qquad = EP\{\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,1}) \leq \mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,2})|W\} = E\left(\frac{\sum_{\pi\in\Pi}\tau_\pi}{|\Pi|}\right).
\end{aligned}$$

Our proof has shown that $P\{\mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,1}) \leq \mathrm{TECV}_\pi(\widehat{\Delta}_{n_1,2})\}$ as $n \to \infty$, so we must also have $E(\sum_{\pi\in\Pi}\tau_\pi/|\Pi|) \to 1$. Since $\sum_{\pi\in\Pi}\tau_\pi/|\Pi|$ is between 0 and 1,

$$E\left(\sum_{\pi\in\Pi}\tau_\pi/|\Pi|\right) \to 1 \text{ implies } \sum_{\pi\in\Pi}\tau_\pi/|\Pi| \xrightarrow{p} 1.$$

Thus the event $\sum_{\pi\in\Pi}\tau_\pi > \frac{|\Pi|}{2}$, under which ϕ_1 is selected, occurs with probability tending to 1. This completes the proof of Theorem 1.

Appendix B

Proof of TEEM Risk Bound

This chapter presents a detailed proof of Theorem 2, which is formulated in Section 5.3.2. The regularity conditions mentioned in this proof are enumerated before the theorem in Section 5.3.2.

Proof First let $P = 1$, where P is the number of permutations from Step 8 of the algorithm. Use the indices i of the treated units in $\mathbf{Z}^{(2)}$ to create the ordering $m = 1, \ldots, \widetilde{n}_2$, where each m represents the treatment-control pair (i, i^*) with the mth-smallest value of i among the pairs created in Step 2 of the algorithm. Using this assignment, we hereafter denote each (i, i^*) as (m_t, m_c) for simplicity of notation. For each pair m, denote the realized values of $(\mathbf{U}_{m_t}, \mathbf{U}_{m_c})$ as $(\mathbf{u}_{m_t}, \mathbf{u}_{m_c})$, and let $\widetilde{\delta}_m = Y_{m_t} - Y_{m_c}$. Conditional on $(\mathbf{U}_{m_t}, \mathbf{U}_{m_c}) = (\mathbf{u}_{m_t}, \mathbf{u}_{m_c})$, the density of $\widetilde{\delta}_m$ under Δ, f_c and σ can be expressed as

$$p_{\Delta, f_c, \sigma}(\widetilde{\delta}_m | \mathbf{u}_{m_t}, \mathbf{u}_{m_c}) = \frac{1}{\sigma}\phi\left(\frac{\widetilde{\delta}_m - \Delta(\mathbf{u}_{m_t}) - (f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c}))}{\sigma}\right).$$

The estimated density of $\widetilde{\delta}_m$ under $\widehat{\Delta}$ and $\hat{\sigma}$ and supposing $f_c(\mathbf{u}_{m_t}) = f_c(\mathbf{u}_{m_c})$ is

$$p_{\widehat{\Delta}, \hat{\sigma}}(\widetilde{\delta}_m | \mathbf{u}_{m_t}, \mathbf{u}_{m_c}) = \frac{1}{\hat{\sigma}}\phi\left(\frac{\widetilde{\delta}_m - \widehat{\Delta}(\mathbf{u}_{m_t})}{\hat{\sigma}}\right).$$

Define

$$q_1(\widetilde{\delta}_1 | \mathbf{u}_{1_t}, \mathbf{u}_{1_c}) = \sum_{j=1}^{J} \omega_j p_{\widehat{\Delta}_{n_1,j}, \hat{\sigma}_{n_1,j}}(\widetilde{\delta}_1 | \mathbf{u}_{1_t}, \mathbf{u}_{1_c}),$$

and for $2 \leq m \leq \widetilde{n}_2$, define

$$q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c}) = \frac{\sum_{j=1}^{J} \omega_j \left(\prod_{l=1}^{m-1} p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}_{n_1,j}}(\widetilde{\delta}_l|\mathbf{u}_{l_t}, \mathbf{u}_{l_c})\right) p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}_{n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})}{\sum_{j=1}^{J} \omega_j \prod_{l=1}^{m-1} p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}_{n_1,j}}(\widetilde{\delta}_l|\mathbf{u}_{l_t}, \mathbf{u}_{l_c})}.$$

The error density ϕ has mean 0; therefore, given π, γ, $\mathbf{Z}^{(1)}$, $(\mathbf{u}_{l_t}, \mathbf{u}_{l_c}, y_{l_t}, y_{l_c})_{l=1}^{m-1}$, and $(\mathbf{u}_{m_t}, \mathbf{u}_{m_c})$, $q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})$ has mean $\sum_j W_{m,j}\widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t}) = \widetilde{\Delta}_m(\mathbf{u}_{m_t})$, where $W_{m,j}$ represent the weights defined in Step 5 of the TEEM algorithm.

Let

$$g_j\left((\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right) = \prod_{m=1}^{\widetilde{n}_2} p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}_{n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c}),$$

and let

$$\widetilde{g}\left((\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right) = \sum_{j=1}^{J} \omega_j g_j\left((\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right).$$

Note that $\prod_{m=1}^{\widetilde{n}_2} q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c}) = \widetilde{g}\left((\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right)$. One can view $q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})$ as an estimator of the conditional density of $\widetilde{\delta}_m$ given $(\mathbf{u}_{m_t}, \mathbf{u}_{m_c})$. The cumulative risk, under the Kullback-Leibler divergence, of $q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})$ at the design points $(\mathbf{u}_{m_t}, \mathbf{u}_{m_c})_{m=1}^{\widetilde{n}_2}$ can be bounded in terms of the risks of the individual procedures using an idea from Barron (1987). Letting E_π denote the expectation conditional on the permutation π, we have

$$\begin{aligned}
&\sum_{m=1}^{\widetilde{n}_2} E_\pi D\{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})||q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\}\\
=&\sum_{m=1}^{\widetilde{n}_2} E_\pi \int p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \log \frac{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} d\widetilde{\delta}_m\\
=&\sum_{m=1}^{\widetilde{n}_2} E_\pi \int \left\{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right\} \log \frac{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} d\widetilde{\delta}_m\\
=&E_\pi \int \left\{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right\} \left\{\sum_{m=1}^{\widetilde{n}_2} \log \frac{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}\right\} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2}\\
=&E_\pi \int \left\{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right\} \log \frac{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{\prod_{m=1}^{\widetilde{n}_2} q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2}\\
=&E_\pi \int \left\{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right\} \log \frac{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{\widetilde{g}\left((\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right)} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2}.
\end{aligned}$$

Since ϕ is a positive-valued function and $\log(x)$ is an increasing function, we have that for any $j \geq 1$,

$$\begin{aligned}
&E_\pi \int \left\{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right\} \log \frac{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{\widetilde{g}\left((\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right)} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2}\\
\leq& E_\pi \int \left\{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right\} \log \frac{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{\omega_j g_j\left((\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right)} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2}\\
=& \log \frac{1}{\omega_j}\\
&\quad + E_\pi \int \left\{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right\} \log \frac{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{g_j\left((\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right)} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2}.
\end{aligned}$$

The last term in the preceding equation is the cumulative risk, under the Kullback-Leibler divergence, of $p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}_{n_1,j}}$ at the design points $(\mathbf{u}_{m_t}, \mathbf{u}_{m_c})_{m=1}^{\widetilde{n}_2}$, given the permutation π. This is because

$$
\begin{aligned}
&E_\pi \int \left\{ \prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c}) \right\} \log \frac{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})}{g_j\left((\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right)} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2} \\
=&E_\pi \int \left\{ \prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c}) \right\} \left\{ \sum_{m=1}^{\widetilde{n}_2} \log \frac{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})}{p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}_{n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})} \right\} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2} \\
=&\sum_{m=1}^{\widetilde{n}_2} E_\pi \int p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c}) \log \frac{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})}{p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}_{n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})} d\widetilde{\delta}_m \\
=&\sum_{m=1}^{\widetilde{n}_2} E_\pi D\{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})||p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}_{n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})\}.
\end{aligned}
$$

By definition,

$$
\begin{aligned}
&D\{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})||p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}_{n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})\} \\
&= \int \left\{ \frac{1}{\sigma}\phi\left[\frac{\widetilde{\delta}_m - \Delta(\mathbf{u}_{m_t}) - \{f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\}}{\sigma}\right] \right. \\
&\qquad \left. \times \log \frac{(1/\sigma)\phi\left[\left(\widetilde{\delta}_m - \Delta(\mathbf{u}_{m_t}) - \{f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\}\right)/\sigma\right]}{(1/\hat{\sigma}_{n_1,j})\phi\left[(\widetilde{\delta}_m - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t}))/\hat{\sigma}_{n_1,j}\right]} \right\} d\widetilde{\delta}_m.
\end{aligned}
$$

Letting

$$z = \frac{\widetilde{\delta}_m - \Delta(\mathbf{u}_{m_t}) - \{f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\}}{\sigma},$$

we perform an integral transformation to obtain

$$
\begin{aligned}
&D\{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})||p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}_{n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})\} \\
=&\int \phi(z) \log \frac{\phi(z)}{(\sigma/\hat{\sigma}_{n_1,j})\phi\left[\sigma z + \Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t}) + f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\}/\hat{\sigma}_{n_1,j}\right]} dz.
\end{aligned}
$$

The standard normal p.d.f. ϕ has the property that for each pair $0 < s_0 < 1$ and

$T > 0$, there exists a constant B_0 (depending on s_0 and T) such that

$$\int \phi(x) \log \frac{\phi(x)}{(1/s)\phi((x-t)/s)} dx \leq B_0((1-s)^2 + t^2)$$

for all $s_0 \leq s \leq 1/s_0$ and $-T < t < T$ (see Assumption A2 in Yang, 2001). Using this fact and taking

$$\begin{aligned} & s_0 = \underline{\sigma}/\overline{\sigma},\ s = \hat{\sigma}_{n_1,j}/\sigma,\ T = 4A/\underline{\sigma},\ \text{and} \\ & t = -\left[\frac{\left\{\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right\} + \left\{f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\right\}}{\sigma}\right], \end{aligned}$$

it follows that

$$\begin{aligned} & D\{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})||p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}_{n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\} \\ & \leq B_0\left(\left[1 - \frac{\hat{\sigma}_{n_1,j}}{\sigma}\right]^2 + \left[\frac{\left\{\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right\} + \left\{f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\right\}}{\sigma}\right]^2\right), \end{aligned}$$

for a constant B_0 depending on A, $\underline{\sigma}$, and $\overline{\sigma}$. Using $\sigma^2 \geq 2\underline{\sigma}^2$ and the parallelogram law, we obtain that for any $j \geq 1$,

$$\begin{aligned} & D\{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})||p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}_{n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\} \\ & \leq \frac{B_0}{\underline{\sigma}^2}\left(\frac{1}{2}\left\{\sigma - \hat{\sigma}_{n_1,j}\right\}^2 + \left\{\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right\}^2 + \left\{f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\right\}^2\right). \end{aligned}$$

Thus we have shown

$$\begin{aligned} & \frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2} E_\pi D\{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})||q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\} \\ & \quad \leq \frac{B_0}{\underline{\sigma}^2\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2} E_\pi\left\{f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\right\}^2 + \inf_j\left[\frac{1}{\widetilde{n}_2}\log\frac{1}{\omega_j}\right. \\ & \qquad \left. + \frac{B_0}{\underline{\sigma}^2}\left\{\frac{1}{2}E_\pi(\sigma - \hat{\sigma}_{n_1,j})^2 + \frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2} E_\pi\left(\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right)^2\right\}\right]. \quad \text{(B.1)} \end{aligned}$$

Let $d_H^2(f, g) = \int (\sqrt{f} - \sqrt{g})^2 d\nu$ denote the squared Hellinger distance between the densities f and g with respect to the measure ν. The squared Hellinger distance is upper bounded by the K-L divergence, so

$$\frac{1}{\widetilde{n}_2} \sum_{m=1}^{\widetilde{n}_2} E_\pi d_H^2 \{ p_{\Delta, f_c, \sigma}(\widetilde{\delta}_m | \mathbf{u}_{m_t}, \mathbf{u}_{m_c}), q_m(\widetilde{\delta}_m | \mathbf{u}_{m_t}, \mathbf{u}_{m_c}) \}$$

is bounded above by (B.1).

As mentioned earlier, for each m, given π, γ, $\mathbf{Z}^{(1)}$, $(\mathbf{u}_{l_t}, \mathbf{u}_{l_c}, y_{l_t}, y_{l_c})_{l=1}^{m-1}$, and $(\mathbf{u}_{m_t}, \mathbf{u}_{m_c})$, $q_m(\widetilde{\delta}_m | \mathbf{u}_{m_t}, \mathbf{u}_{m_c})$ has mean $\widetilde{\Delta}_m(\mathbf{u}_{m_t})$ with respect to $\widetilde{\delta}_m$. For this estimator, we have

$$
\begin{aligned}
&\left(\int \widetilde{\delta}_m p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m - \int \widetilde{\delta}_m q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m\right)^2\\
=&\left(\int \widetilde{\delta}_m \left\{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) - q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right\} d\widetilde{\delta}_m\right)^2\\
=&\Bigg\{\int \widetilde{\delta}_m \left(\sqrt{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} + \sqrt{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}\right)\\
&\qquad \times \left(\sqrt{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} - \sqrt{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}\right) d\widetilde{\delta}_m\Bigg\}^2\\
\leq&\int \widetilde{\delta}_m^2 \left(\sqrt{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} + \sqrt{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}\right)^2 d\widetilde{\delta}_m\\
&\qquad \times \int \left(\sqrt{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} - \sqrt{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}\right)^2 d\widetilde{\delta}_m\\
\leq&2\left(\int \widetilde{\delta}_m^2 p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) + \int \widetilde{\delta}_m^2 q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m\right)\\
&\qquad \times \int \left(\sqrt{p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} - \sqrt{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}\right)^2 d\widetilde{\delta}_m\\
=&2\left(E(\widetilde{\delta}_m^2|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) + \int \widetilde{\delta}_m^2 q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m\right)\\
&\qquad \times d_H^2\left(p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}), q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right)\\
=&2\left(\left\{E(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right\}^2 + \sigma^2 + \int \widetilde{\delta}_m^2 q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m\right)\\
&\qquad \times d_H^2\left(p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}), q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right)\\
=&2\left(\left\{\Delta(\mathbf{u}_{m_t}) + f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\right\}^2 + \sigma^2 + \int \widetilde{\delta}_m^2 q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m\right)\\
&\qquad \times d_H^2\left(p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}), q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right),
\end{aligned}
$$

where the first and second inequalities follow from the Cauchy-Schwarz inequality and the parallelogram law, respectively.

By the first regularity condition, $\{\Delta(\mathbf{u}_{m_t}) + f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\}^2 \leq (4A)^2$. Now $\int \widetilde{\delta}_m^2 q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m = E_{q_m}(\widetilde{\delta}_m^2|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \leq \{E_{q_m}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\}^2 + \overline{\sigma}^2$, and $q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})$ is a convex combination of J densities in the location-scale family

$\phi((x-b)/a)/a$, each with mean $\widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})$ with respect to $\widetilde{\delta}_m$. Therefore,

$$\int \widetilde{\delta}_m^2 q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m \leq (2A)^2+\overline{\sigma}^2.$$

It follows that

$$\begin{aligned}&\left(\int \widetilde{\delta}_m p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m - \int \widetilde{\delta}_m q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m\right)^2\\ &\qquad \leq (40A^2+4\overline{\sigma}^2)d_H^2\left(p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}), q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right).\end{aligned}$$

Together with

$$\int \widetilde{\delta}_m p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m = E(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) = \Delta(\mathbf{u}_{m_t})+f_c(\mathbf{u}_{m_t})-f_c(\mathbf{u}_{m_c})$$

and

$$\int \widetilde{\delta}_m q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m = \widetilde{\Delta}_m(\mathbf{u}_{m_t}),$$

we have, for each $1 \leq m \leq \widetilde{n}_2$,

$$\begin{aligned}&\left(\Delta(\mathbf{u}_{m_t})+f_c(\mathbf{u}_{m_t})-f_c(\mathbf{u}_{m_c})-\widetilde{\Delta}_m(\mathbf{u}_{m_t})\right)^2\\ &\qquad \leq (40A^2+4\overline{\sigma}^2)d_H^2\left(p_{\Delta,f_c,\sigma}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}), q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right).\end{aligned} \tag{B.2}$$

The expression (B.2) also is an upper bound for

$$\left(\Delta(\mathbf{u}_{m_t})-(f_c(\mathbf{u}_{m_t})-f_c(\mathbf{u}_{m_c}))-\widetilde{\Delta}_m(\mathbf{u}_{m_t})\right)^2.$$

So by the parallelogram law, (B.2) is an upper bound for $\left(\Delta(\mathbf{u}_{m_t})-\widetilde{\Delta}_m(\mathbf{u}_{m_t})\right)^2$. Then by using the earlier risk bound on the average squared Hellinger distance and combining constants, we obtain

$$\begin{aligned}
&\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2} E_\pi \left(\Delta(\mathbf{u}_{m_t}) - \widetilde{\Delta}_m(\mathbf{u}_{m_t})\right)^2 \\
&\qquad \le B_2 \left(\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2} E_\pi \Big\{ f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c}) \Big\}^2 + \inf_j \left[\frac{1}{\widetilde{n}_2} \log \frac{1}{\omega_j} \right.\right. \\
&\qquad\qquad \left.\left. + E_\pi(\sigma - \hat{\sigma}_{n_1,j})^2 + \frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2} E_\pi \left(\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right)^2 \right] \right), \qquad \text{(B.3)}
\end{aligned}$$

where B_2 depends on $\underline{\sigma}$, $\overline{\sigma}$, and A.

Now we connect the global risk of the estimator $\tilde{\tilde{\Delta}}_\pi$ to the average risk of the individual estimators $\widetilde{\Delta}_m$ at the design points. Let D_π denote the event that $\widetilde{n}_2 = (1/h)^p$; that is, the event that every cell in the partition of $\mathcal{U}$ contains at least one treatment-control pair from $\mathbf{Z}^{(2)}$ after the permutation π. Let $\mathcal{U}_m$ denote the cell in the partition containing the mth treatment-control pair. Conditional on D_π,

$$\begin{aligned}
&E_\pi \|\Delta - \tilde{\tilde{\Delta}}_\pi\|_2^2 \\
&= E_\pi \int_{\mathcal{U}} \left(\Delta(\mathbf{u}) - \tilde{\tilde{\Delta}}_\pi(\mathbf{u})\right)^2 dP_{\mathbf{U}} \\
&= E_\pi \sum_{m=1}^{\widetilde{n}_2} \int_{\mathcal{U}_m} \left(\Delta(\mathbf{u}) - \tilde{\tilde{\Delta}}_\pi(\mathbf{u})\right)^2 dP_{\mathbf{U}}.
\end{aligned}$$

By the definition of $\tilde{\tilde{\Delta}}_\pi$, for any $\mathbf{u} \in \mathcal{U}_m$, $\tilde{\tilde{\Delta}}_\pi(\mathbf{u}) = \widetilde{\Delta}_m(\mathbf{u}_{m_t})$. Therefore, for $\mathbf{u} \in \mathcal{U}_m$,

$$\begin{aligned}
&\left(\Delta(\mathbf{u}) - \tilde{\tilde{\Delta}}_\pi(\mathbf{u})\right)^2 \\
&= \left(\Big\{ \Delta(\mathbf{u}) - \Delta(\mathbf{u}_{m_t}) \Big\} + \Big\{ \Delta(\mathbf{u}_{m_t}) - \widetilde{\Delta}_m(\mathbf{u}_{m_t}) \Big\} \right)^2 \\
&\qquad \le 2\Big(\Delta(\mathbf{u}) - \Delta(\mathbf{u}_{m_t})\Big)^2 + 2\Big(\Delta(\mathbf{u}_{m_t}) - \widetilde{\Delta}_m(\mathbf{u}_{m_t})\Big)^2.
\end{aligned}$$

Combining the previous two displays, and using the fact that for any m, $\int_{\mathcal{U}_m} dP_{\mathbf{U}} \le$

$\overline{c}/\widetilde{n}_2$, we have

$$\begin{aligned}
&E_\pi\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \\
&\leq 2E_\pi\left(\sum_{m=1}^{\widetilde{n}_2}\int_{\mathcal{U}_m}\Big(\Delta(\mathbf{u}) - \Delta(\mathbf{u}_{m_t})\Big)^2 dP_{\mathbf{U}}\right. \\
&\qquad \left. + \frac{\overline{c}}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2}\Big(\Delta(\mathbf{u}_{m_t}) - \widetilde{\Delta}_m(\mathbf{u}_{m_t})\Big)^2\right). \qquad \text{(B.4)}
\end{aligned}$$

For the first summation on the right-hand side of (B.4), by the Mean Value Theorem for integrals and the fact that every cell $\mathcal{U}_m$ has volume $1/\widetilde{n}_2$, we have

$$\sum_{m=1}^{\widetilde{n}_2}\int_{\mathcal{U}_m}\Big(\Delta(\mathbf{u}) - \Delta(\mathbf{u}_{m_t})\Big)^2 dP_{\mathbf{U}} = \frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2} f(\mathbf{u}_m^*)\Big(\Delta(\mathbf{u}_m^*) - \Delta(\mathbf{u}_{m_t})\Big)^2,$$

where $\mathbf{u}_m^*$ is some point in the hypercube $\mathcal{U}_m$ and $f(\mathbf{u}_m^*)$ represents the design density at this point. The smoothness condition for Δ ensures that it satisfies a Lipschitz condition with Lipschitz constant $\sqrt{p}L$. Thus for any m, since the distance between $\mathbf{u}_m^*$ and $\mathbf{u}_{m_t}$ is at most $\sqrt{p}h$, $\Delta(\mathbf{u}_m^*) - \Delta(\mathbf{u}_{m_t}) \leq pLh$. Thus we have

$$\sum_{m=1}^{\widetilde{n}_2}\int_{\mathcal{U}_m}\Big(\Delta(\mathbf{u}) - \Delta(\mathbf{u}_{m_t})\Big)^2 dP_{\mathbf{U}} \leq \overline{c}(pLh)^2. \qquad \text{(B.5)}$$

Combining (B.3), (B.4), and (B.5), we have established

$$\begin{aligned}
&E_\pi\left\{\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2\Big|D_\pi\right\} \\
&\leq 2\overline{c}(pLh)^2 + 2\overline{c}B_2\left(\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2}E_\pi\Big\{f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\Big\}^2 + \inf_j\left[\frac{1}{\widetilde{n}_2}\log\frac{1}{\omega_j}\right.\right. \\
&\qquad \left.\left. + E_\pi(\sigma - \hat{\sigma}_{n_1,j})^2 + \frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2}E_\pi\left(\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right)^2\right]\right). \qquad \text{(B.6)}
\end{aligned}$$

Next we relate the global risk of each $\widehat{\Delta}_{n_1,j}$ to its average risk at the design points. Again using the Mean Value Theorem for integrals and conditioning on D_π, we have for any $j \geq 1$,

$$\begin{aligned}
&\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2} E_\pi \left(\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right)^2 - E_\pi \|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2 \\
&\qquad \leq \frac{c^*}{\widetilde{n}_2} E_\pi \sum_{m=1}^{\widetilde{n}_2} \left\{ \left(\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right)^2 - \left(\Delta(\mathbf{u}^*_m) - \widehat{\Delta}_{n_1,j}(\mathbf{u}^*_m)\right)^2 \right\},
\end{aligned}$$

where c^* is a constant bounded by $\max(1/\underline{c}, \overline{c})$ that exists by the boundedness of $P_{\mathbf{U}}$. The difference in the squared differences after the summation can be bounded for each m by the smoothness of Δ and $\widehat{\Delta}_{n_1,j}$.

Indeed, for each m we have

$$\begin{aligned}
&\left(\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right)^2 - \left(\Delta(\mathbf{u}^*_m) - \widehat{\Delta}_{n_1,j}(\mathbf{u}^*_m)\right)^2 \\
&= \left\{ \left(\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right) + \left(\Delta(\mathbf{u}^*_m) - \widehat{\Delta}_{n_1,j}(\mathbf{u}^*_m)\right) \right\} \\
&\qquad \times \left\{ \left(\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right) - \left(\Delta(\mathbf{u}^*_m) - \widehat{\Delta}_{n_1,j}(\mathbf{u}^*_m)\right) \right\}.
\end{aligned}$$

Since Δ and $\widehat{\Delta}_{n_1,j}$ both are bounded between $-2A$ and $2A$,

$$\left(\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right) + \left(\Delta(\mathbf{u}^*_m) - \widehat{\Delta}_{n_1,j}(\mathbf{u}^*_m)\right) \leq 4A.$$

Meanwhile, the smoothness of Δ and $\widehat{\Delta}_{n_1,j}$ ensure that both satisfy a Lipschitz condition with Lipschitz constant $\sqrt{p}L$. Thus for any m, since each $\mathcal{U}_m$ has diameter $\sqrt{p}h$,

$$\begin{aligned}
&\left(\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right) - \left(\Delta(\mathbf{u}^*_m) - \widehat{\Delta}_{n_1,j}(\mathbf{u}^*_m)\right) \\
&= \left(\Delta(\mathbf{u}_{m_t}) - \Delta(\mathbf{u}^*_m)\right) + \left(\widehat{\Delta}_{n_1,j}(\mathbf{u}^*_m) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right) \leq 2pLh.
\end{aligned}$$

Therefore, conditional on D_π,

$$\frac{1}{\widetilde{n}_2} E_\pi \sum_{m=1}^{\widetilde{n}_2} \left(\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right)^2 \leq E_\pi \|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2 + 8c^* ApLh. \tag{B.7}$$

Thus combining (B.7) with (B.6), we have established that

$$\begin{aligned}
&E_\pi \left\{ \|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \middle| D_\pi \right\} \\
&\leq 8c^* ApLh + \overline{c}(pLh)^2 + B_2 \Bigg(\frac{1}{\widetilde{n}_2} \sum_{m=1}^{\widetilde{n}_2} E_\pi \Big\{ f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c}) \Big\}^2 \\
&\qquad + \inf_j \left[\frac{1}{\widetilde{n}_2} \log \frac{1}{\omega_j} + E_\pi(\sigma - \hat{\sigma}_{n_1,j})^2 + E_\pi \|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2 \right] \Bigg).
\end{aligned}$$

Using the Lipschitz condition for f_c within each cell, in a similar fashion as before, we can show that

$$\frac{1}{\widetilde{n}_2} \sum_{m=1}^{\widetilde{n}_2} E_\pi \Big\{ f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c}) \Big\}^2 \leq (pLh)^2.$$

Thus we have

$$\begin{aligned}
&E_\pi \left\{ \|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \middle| D_\pi \right\} \\
&\leq 8c^* ApLh + B_3 \Bigg((pLh)^2 \\
&\qquad + \inf_j \left[\frac{1}{\widetilde{n}_2} \log \frac{1}{\omega_j} + E_\pi(\sigma - \hat{\sigma}_{n_1,j})^2 + E_\pi \|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2 \right] \Bigg), \qquad \text{(B.8)}
\end{aligned}$$

for a constant B_3 depending on $\underline{\sigma}$, $\overline{\sigma}$, A, and $\overline{c}$.

Now,

$$E_\pi \|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \leq E_\pi \left\{ \|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \middle| D_\pi \right\} + E_\pi \left\{ \|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \middle| D_\pi^c \right\} \times P(D_\pi^c). \qquad \text{(B.9)}$$

By the boundedness of Δ and $\widetilde{\widetilde{\Delta}}_\pi$ between $-2A$ and $2A$,

$$E_\pi \left\{ \|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \middle| D_\pi^c \right\} \leq 16A^2. \qquad \text{(B.10)}$$

To use (B.9), we need to bound $P(D_\pi^c)$. Denote the event that all cells in our partition contain at least observation from the treatment group by $D_{\pi,t}$, and let $D_{\pi,c}$ denote the corresponding event for the control group. Since $D_\pi = D_{\pi,t} \cap D_{\pi,c}$, $P(D_\pi^c) \leq$

$P(D_{\pi,t}^c) + P(D_{\pi,c}^c)$.

Let $\mathcal{U}_g$ denote an arbitrary cell in the partition. By the first regularity condition, the probability that any observation from the treatment group falls into $\mathcal{U}_g$ is at least $\underline{c}h^p$. Since the covariate values of the n_{t_2} treatment observations are i.i.d., the probability that $\mathcal{U}_g$ contains no treatment observations from $\mathbf{Z}^{(2)}$ is at most

$$(1 - \underline{c}h^p)^{n_{t_2}} = e^{n_{t_2} \log(1-\underline{c}h^p)} \leq e^{-n_{t_2}\underline{c}h^p},$$

where the last inequality results from the fact that $\log x \leq x - 1$.

Since $\mathcal{U}_g$ is arbitrary and there are $(1/h)^p$ such cells in the partition of $\mathcal{U}$, the probability that any of them contain no treatment observations is at most

$$(1/h)^p e^{-n_{t_2}\underline{c}h^p} = \exp\{-n_{t_2}\underline{c}h^p + p\log(1/h)\}.$$

By the choice of h in Step 2 of the TEEM algorithm, $h \geq \{2\log(n_2^*)/\underline{c}n_2^*\}^{1/p}$. Therefore,

$$\begin{aligned}
&- n_{t_2}\underline{c}h^p + p\log(1/h) \\
&\leq \frac{-2n_{t_2}\log(n_2^*)}{n_2^*} + \log\left(\frac{\underline{c}n_2^*}{2\log n_2^*}\right) \\
&\leq \log\left(\frac{\underline{c}}{2n_2^*\log n_2^*}\right) \\
&\leq \log\left(\frac{\underline{c}}{2\widetilde{n}_2\log \widetilde{n}_2}\right).
\end{aligned}$$

The second inequality in the above expression results from $n_{t_2} \geq n_2^*$. Thus

$$P(D_{\pi,t}^c) \leq \exp\left\{\log\left(\frac{\underline{c}}{2n_2^*\log n_2^*}\right)\right\} = \left(\frac{\underline{c}}{2n_2^*\log n_2^*}\right).$$

The same bound may be established for $P(D_{\pi,c}^c)$; therefore,

$$P(D_\pi^c) \leq \frac{\underline{c}}{n_2^*\log n_2^*}. \tag{B.11}$$

Using (B.9) together with (B.8), (B.10), and (B.11), and using the fact that $h =$

$B_4\{\log(n_2^*)/n_2^*\}^{1/p}$ for some B_4 depending on $\underline{c}$ and p, we have

$$\begin{aligned}
&E_\pi\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \\
&\leq 8c^* ApLB_4 \left(\frac{\log n_2^*}{n_2^*}\right)^{1/p} + B_3(B_4pL)^2\left(\frac{\log n_2^*}{n_2^*}\right)^{2/p} + 16A^2\underline{c}\left(\frac{1}{n_2^*\log n_2^*}\right) \\
&\quad + B_3 \inf_j \left[\frac{1}{\widetilde{n}_2}\log\frac{1}{\omega_j} + E_\pi(\sigma - \hat{\sigma}_{n_1,j})^2 + E_\pi\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2\right]. \qquad \text{(B.12)}
\end{aligned}$$

With the exception of small n_2^*,

$$\frac{1}{n_2^*\log n_2^*} \leq \left(\frac{\log n_2^*}{n_2^*}\right)^{2/p} \leq \left(\frac{\log n_2^*}{n_2^*}\right)^{1/p},$$

so we can rewrite expression (B.12) as

$$\begin{aligned}
&E_\pi\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \\
&\leq B_5\left(\left\{\frac{\log n_2^*}{n_2^*}\right\}^{1/p} + \inf_j\left[\frac{1}{\widetilde{n}_2}\log\frac{1}{\omega_j} + E_\pi(\sigma - \hat{\sigma}_{n_1,j})^2 + E_\pi\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2\right]\right),
\end{aligned}$$

for a constant B_5 depending on $\underline{c}$, $\overline{c}$, $\underline{\sigma}$, $\overline{\sigma}$, A, p, and L.

Now n_2^* and $\widetilde{n}_2$, which heretofore we have treated as fixed, are random variables determined by the values of $(\mathbf{U}_i, T_i)_{i=1}^n$ and the permutation π. By the iterated expectation law, unconditional on the permutation π,

$$\begin{aligned}
&E\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 = E\left\{E_\pi\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2\right\} \\
&\leq B_5\left(E\left\{\frac{\log n_2^*}{n_2^*}\right\}^{1/p} + \inf_j\left[E\frac{1}{\widetilde{n}_2}\log\frac{1}{\omega_j} + E(\sigma - \hat{\sigma}_{n_1,j})^2 + E\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2\right]\right).
\end{aligned} \qquad \text{(B.13)}$$

Let $\alpha \in (0,1)$ be a fixed constant and let $H_{\alpha,\pi}$ denote the event that $P(n_2^* \geq \alpha n_2)$.

Since $\{\log n_2^*/n_2^*\}^{1/p} \leq 1$, we have

$$\begin{aligned} E\left\{\frac{\log n_2^*}{n_2^*}\right\}^{1/p} &\leq E\left[\left\{\frac{\log n_2^*}{n_2^*}\right\}^{1/p} \middle| H_{\alpha,\pi}\right] + P(H_{\alpha,\pi}^c) \\ &\leq \alpha^{-1/p}\left(\frac{\log n_2}{n_2}\right)^{1/p} + P(H_{\alpha,\pi}^c). \end{aligned}$$

For $P(H_{\alpha,\pi}^c)$, the exponential bound on the upper tail probability of the hypergeometric distribution established by Chvátal (1979) can be used to show that we can find $\alpha \in (0,1)$ depending on a and b from Regularity Condition 3 such that

$$P(H_{\alpha,\pi}^c) \leq B_6 e^{-n_2},$$

for a constant B_6 depending on a and b. Thus

$$E\left\{\frac{\log n_2^*}{n_2^*}\right\}^{1/p} \leq B_7\left(\frac{\log n_2}{n_2}\right)^{1/p}, \tag{B.14}$$

for B_7 depending on a and b.

For $E(1/\widetilde{n}_2)$, conditional on D_π,

$$\frac{1}{\widetilde{n}_2} = h^p = \left\{\left\lfloor\left(\frac{\underline{c}n_2^*}{2\log n_2^*}\right)^{1/p}\right\rfloor\right\}^{-p} \leq B_8\left(\frac{\log n_2^*}{n_2^*}\right) \leq B_7 B_8\left(\frac{\log n_2}{n_2}\right), \tag{B.15}$$

for a constant B_8 depending on $\underline{c}$. As established earlier in this proof, $P(D_\pi^c)$ converges faster than $O(1/n_2^*) = O(1/n_2)$.

Using (B.14) and (B.15) to replace the random variables in (B.13) with fixed constants, we obtain a bound for the risk of $\widetilde{\widetilde{\Delta}}_\pi$:

$$\begin{aligned} &E\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \\ &\leq B_9\left(\left(\frac{\log n_2}{n_2}\right)^{1/p} + \inf_j\left[\left(\frac{\log n_2}{n_2}\right)\log\frac{1}{\omega_j} + E(\sigma - \hat{\sigma}_{n_1,j})^2 + E\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2\right]\right), \end{aligned} \tag{B.16}$$

for a constant B_9 depending on a, b, $\underline{c}$, $\overline{c}$, $\underline{\sigma}$, $\overline{\sigma}$, A, p, and L.

For $P > 1$, the estimator $\overline{\widehat{\Delta}}$ from Step 8 of the algorithm is the average (over the set of P permutations) of $\widetilde{\widetilde{\Delta}}_{\pi_p}$. Therefore, by the convexity of the L_2 loss, an application of Jensen's inequality gives us

$$E\|\Delta - \overline{\widehat{\Delta}}\|_2^2 \leq \frac{1}{P}\sum_{p=1}^{P} E\|\Delta - \widetilde{\widetilde{\Delta}}_{\pi_p}\|_2^2.$$

Since the permutation π used to establish the bound in (B.16) was arbitrary, the bound also holds for $E\|\Delta - \overline{\widehat{\Delta}}\|_2^2$. This completes the proof of Theorem 2.