Title
High-dimensional and causal inference

Author
Walter, Simon

Publication Date
2019

Peer reviewed|Thesis/dissertation

High-dimensional and causal inference

by

Simon James Sweeney Walter

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bin Yu, Co-chair
Professor Jasjeet Sekhon, Co-chair
Professor Peter Bickel
Assistant Professor Avi Feller

Fall 2019

Abstract

High-dimensional and causal inference

by

Simon James Sweeney Walter

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Bin Yu, Co-chair

Professor Jasjeet Sekhon, Co-chair

High-dimensional and causal inference are topics at the forefront of statistical research. This thesis is a unified treatment of three contributions to these literatures. The first two contributions are to the theoretical statistical literature; the third puts the techniques of causal inference into practice in policy evaluation.

In Chapter 2, we suggest a broadly applicable remedy for the failure of Efron's bootstrap in high dimensions is to modify the bootstrap so that data vectors are broken into blocks and the blocks are resampled independently of one another. Cross-validation can be used effectively to choose the optimal block length. We show both theoretically and in numerical studies that this method restores consistency and has superior predictive performance when used in combination with Breiman's bagging procedure. This chapter is joint work with Peter Hall and Hugh Miller.

In Chapter 3, we investigate regression adjustment for the modified outcome (RAMO). An equivalent procedure is given in Rubin and van der Laan [2007] and then in Luedtke and van der Laan [2016]; philosophically similar ideas appear to originate in Miller [1976]. We establish new guarantees when the procedure is applied in designed experiments (where the propensity score is known *a priori*) and confirm that the procedure is doubly robust. RAMO can be implemented in only a few lines of code and it can be immediately combined with existing regression models, including random forests and deep neural networks, used in classical prediction problems. This chapter is joint work with Bin Yu and Jasjeet Sekhon.

In Chapter 4, we investigate the specific deterrent effect of traffic citations. In Queensland, Australia many speeding and red-light running offenses are detected by traffic cameras and drivers are notified of the citation, not at the time they commit the offense, but when the citation notice is delivered by mail about two weeks later. We use a regression discontinuity design to assess whether the chance of crashing or recidivism changes at the moment of notification. We analyzed a population of nearly 3 million drivers who committed camera-detected offenses. We conclude that there is not a significant change in the incidence of crashes but there is a marked decrease in recidivism of about 25%. This chapter is joint work with David Studdert and Jeremy Goldhaber-Fiebert.

Contents

1 Introduction and overview **1**
1.1 A bootstrap for high dimensional classification problems 1
1.1.1 The block bootstrap 2
1.1.2 Moving block bootstrap 4
1.1.3 Circular block bootstrap 4
1.1.4 Comparison of block bootstrap variants 5
1.1.5 The asynchronous bootstrap 6
1.1.6 References 8
1.2 Causal inference 9

2 The asynchronous bootstrap **10**
2.1 Introduction 10
2.2 Methodology 12
2.2.1 Block resampling 12
2.2.2 Distribution estimation 13
2.2.3 Block-bagged classifiers and their error rates 14
2.2.4 Estimating the error rate of a block-bagged classifier 14
2.2.5 Block remnants 15
2.3 Numerical properties 16
2.3.1 Simulation settings 16
2.3.2 Empirical effectiveness 16
2.4 Theory 17
2.4.1 Summary and remarks 17
2.4.2 Distribution estimation 19
2.4.3 Classification 23
2.5 Outlines of technical arguments 27
2.5.1 Outline of the proof of Theorem 1 27
2.5.2 Outline of the proof of Theorem 2 29
2.5.3 Outline of the proof of Theorem 3 30

2.5.4 Outline of the proof of Theorem 4 31
2.5.5 Outline of the proof of Theorem 5 34

3 Adjustment for the modified outcome 36
3.1 Introduction 36
3.2 The modified outcome method 37
3.3 Regression adjustment for the modified outcome 39
3.4 Simulations 43
3.5 Conclusion 45

4 Once ticketed, twice shy 46
4.1 Introduction 46
4.2 Pathways and targets 48
4.3 What is known about the deterrent effects of traffic laws? 50
4.3.1 Drunk driving studies 50
4.3.2 General deterrence 51
4.3.3 Specific deterrence 53
4.4 The causal inference challenge 54
4.4.1 Known unknowns and unknown unknowns 54
4.4.2 Attempts at stronger causal inference 55
4.4.3 Overview of study approach 56
4.5 Study approach 57
4.5.1 Setting 57
4.5.2 Offenses and penalties 57
4.5.3 Data and variables 58
4.5.4 Study design 59
4.5.5 Study sample 60
4.5.6 Statistical analysis 61
4.5.7 Ethics 63
4.6 Results 63
4.6.1 Sample characteristics 63
4.6.2 Effects of notification on crashes 64
4.6.3 Effects of notification on recidivism 64
4.7 Discussion 65
4.7.1 The bifurcation of specific deterrence 66
4.7.2 Is specific deterrence working in Queensland? 67
4.7.3 Other trends in levels risk over time 68
4.7.4 Limitations 70
4.8 Conclusion 71

4.8.1 Acknowledgements 71
4.8.2 Funding 72

Bibliography **82**

A Technical results for the asynchronous bootstrap **88**
A.1 Detailed proofs 88
A.1.1 Proof of Theorem 1 88
A.1.2 Proof of Theorem 2. 92
A.1.3 Proof of Theorem 3 99
A.1.4 Proof of Theorem 4 103
A.1.5 Proof of Theorem 5. 110

Acknowledgments

I am deeply indebted to my advisors: Bin Yu and Jasjeet Sekhon, who formed a perfectly complementary team. I am very grateful to Bin for sharing her rigorous approach to academic research. It was extremely kind of her to include her group in Yu family gatherings and to introduce us to Ken, Maya and Matthew. I owe no less a debt to Jas for sharing his extraordinary ideas with me and for helping me to make my work practically useful; I am in awe of his facility with the social aspects of academic research. I am also grateful to Peter Bickel, Avi Feller and Peng Ding who served on my qualifying or dissertation committees. Peter Bickel generously shared his vast knowledge with me and I will always remember our discussions about the history of our discipline; Avi Feller was an endless source of enthusiasm and encouragement; and Peng Ding provided me with many useful references in the causal inference literature that I had not seen. David Studdert provided me with my first experience of academic research and he has been extraordinarily supportive of me; I regard the work I have done with him as the most consequential. I must thank Peter Hall for his unrivaled insight, generosity, and kindness in life. As H. W. Fowler wrote, "I think of [this work] as it should have been, with its prolixities docked, its dullness enlivened, its fads eliminated, its truths multiplied."

I must also record my gratitude to our department staff: to La Shana Porlaris for coping with my disorganization with grace; to Mary Melinn for not being too disapproving when I lost my office key three times.

Many of my fellow students were and are an indispensible source of friendship: Reza Abbasi-Asl, Geoff Bacon, Yasaman Bahri, Siva Balakrishnan, Rebecca Barter, Zsolt Bartha, Riddhipratim Basu, Sumanta Basu, Merle Behr, Eli Ben-Michael, Adam Bloniarz, Joe Borja, Hyesoo Choi, Yuansi Chen, Zihao Chen, Raaz Dwivedi, Andrew Do, Monica Farid, Ryan Giordano, David Graham-Squire, Wooseok Ha, Johnny Hong, Steve Howard, Miyabi Ishihara, Kejing Jiang, Christine Kuang, Karl Kumbier, Sören Künzel, Kyueun Lee, Lihua Lei, Lisha Li, Xiao Li, Ivana Malenica, Tyler Maltba, W. James Murdoch, Kellie Ottoboni, Biyue Pan, Fanny Perraudeau, Yannik Pitcan, Frank Qiu, Jeff Regier, Dominik Rothenhäusler, Sujayam Saha, Jake Soloff, Bradly Stadie, Sara Stoudt, Sze-chuan Suen, Yan Shuo Tan, Tiffany Tang, Hoang Duong Trong, Zoe Vernon, Christine Wang, Hue Wang, Yuting Wei, Aisha Wilson, Jason Wu, Siqi Wu, Fanny Yang, Chelsea Zhang, and Chi Zhang. Of them I ask only that you "think of me sometimes when oceans and continents divide us—but they never will, unless you should wish it."

I am also grateful to the many faculty who have generously answered my questions and deepened my understanding of statistics including Jay Battacharya, Peter Bickel, Joan Bruna, Aurore Delaigle, Persi Diaconis, Ping Ding, Steve Evans, Will Fithian,

Lisa Goldberg, Jeremy Goldhaber-Fiebert, Avi Feller, Haiyan Huang, Michael Jordan, Mark van der Laan, Jon McAuliffe, Luke Miratrix, Sam Pimentel, Elizabeth Purdom, Jim Pitman, Terry Speed and Willem R. van Zwet.

Finally I acknowledge the love and support of my family. They know that I would be lost without them.

Chapter 1

Introduction and overview

This chapter provides an accessible overview of this thesis. We will eschew technical details and emphasize ease of understanding and interpretability wherever possible.

1.1 A bootstrap for high dimensional classification problems

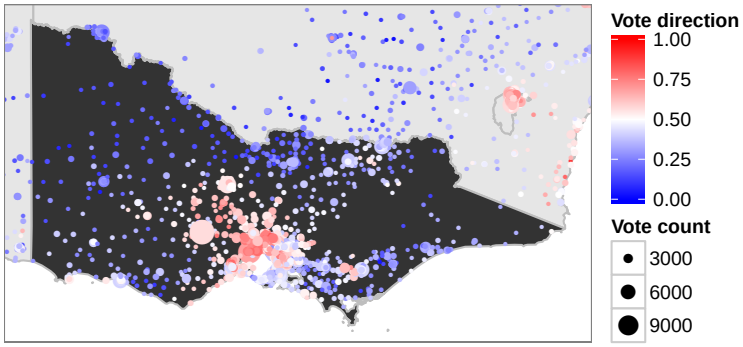


Figure 1.1: Proportion of voters by polling booth favouring the Labor party compared to the Coalition on a two party preferrred basis in Victoria, Australia.

The first of our theoretical contributions is an extension of a technique called the bootstrap. We will use an example to describe the bootstrap. Figure 1.1 records the voting results for a state government election in Victoria, Australia. Each point records the two party preferred vote share at a polling booth; the color of the point

records the direction of the vote: the deeper red a point is, the more strongly the corresponding polling booth favored the center-left Labor party, the deeper blue, the more strongly the booth favored the center-right Liberal-National Coalition.

We have a procedure for converting these points into a smoothed two party preferred voting surface that describes geographic variation in voting preferences and we are interested in assessing the uncertainty of the voting surface constructed. The simplest way of doing this is to draw samples from the population and see how the surface varies across samples. We can simulate this by dividing the data into four sets and constructing an estimate for each set.

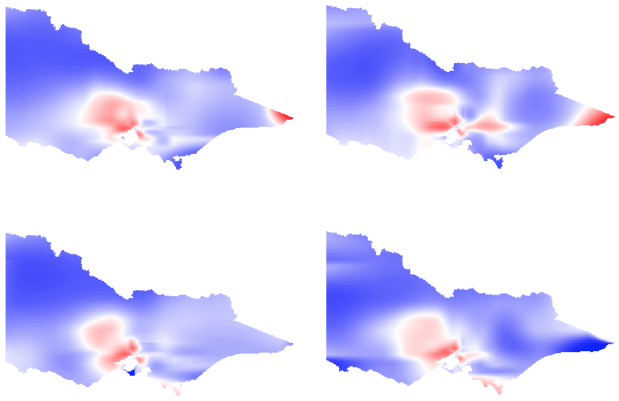


Figure 1.2: Independent replicates of the two party preferred surface in Victoria

Although this works, it is, in practice, an extremely conservative estimate of the uncertainty of the surface estimated with the full dataset because it relies on datasets one quarter the size. The bootstrap is a refinement of this intuition; rather than drawing from the population we (re)sample, with replacement, from the original sample and construct estimates for each resample. This is illustrated in Figure 1.3.

1.1.1 The block bootstrap

The bootstrap is versatile, but in some settings, it requires customization. For example, we do not get good results if we apply the bootstrap to a time series. If we apply the bootstrap assuming the observations of a time series are independent, we do not capture the time-dependent structure of the observations, nor should we treat the time series as a single observation because then the bootstrap replicates will not vary and the bootstrap distribution of a statistics will be a poor estimate of the true distribution.

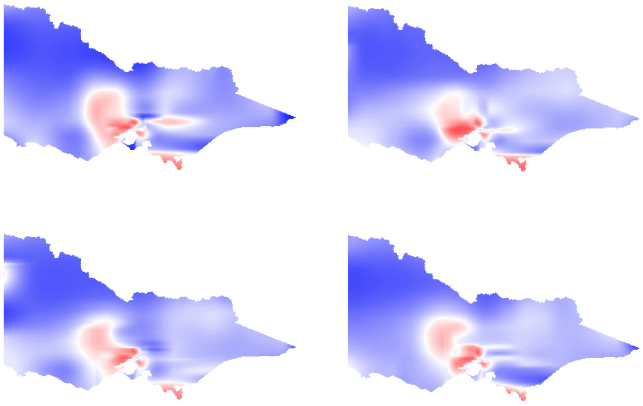


Figure 1.3: Bootstrap replicates of the 2PP surface in Victoria

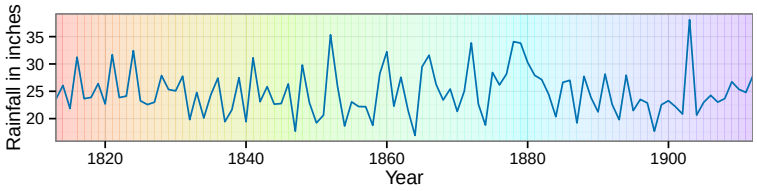
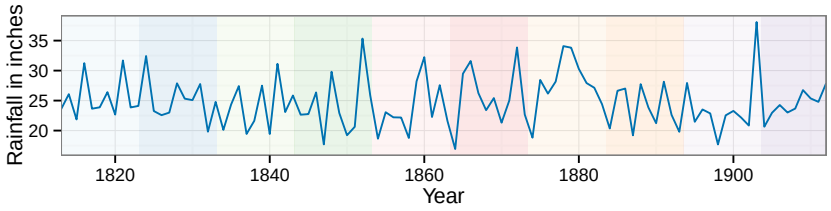


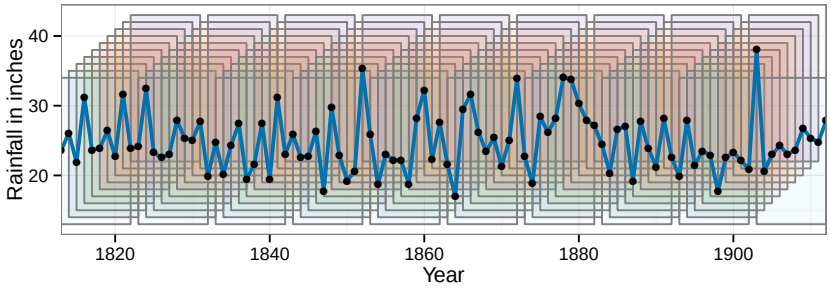
Figure 1.4: Annual rainfall in London 1813–1912

The solution is to divide the time series into contiguous blocks and then stitch together resampled versions of the blocks. This procedure only works if the time series is strongly stationary (meaning we can shift it back or forwards in time and it still has the same distribution). There are many procedures for constructing blocks. To illustrate a selection of procedure we use the example dataset in Figure 1.4, which shows the annual rainfall in London between 1813 and 1912. The fixed block bootstrap defines disjoint blocks so that each observation is part of one and only one block. Notice that this discards the information about the dependence structure at block transitions.



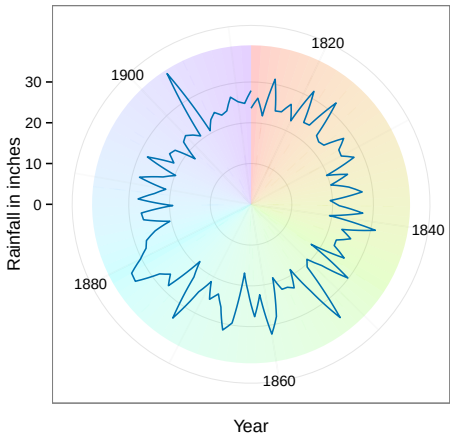
1.1.2 Moving block bootstrap

A more data conservative approach is available in the moving block bootstrap. Blocks are defined as *every* contiguous segment of the time series of an appropriate length. Notice that the first and last values of the time series are only part of one block and are therefore less likely to be includuded in a bootstrap replicate.

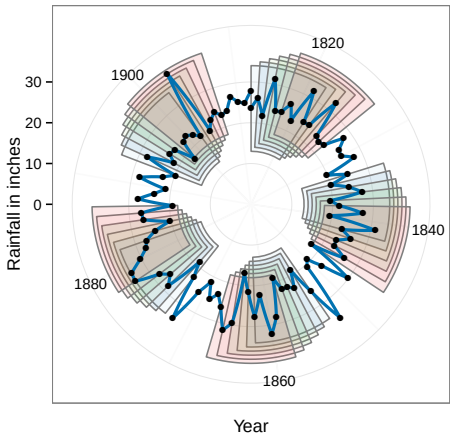


1.1.3 Circular block bootstrap

The circular block bootstrap addresses this asymmetry by wrapping the time series around a circle. This way every value is a member of the same number of blocks, even the first and last values.

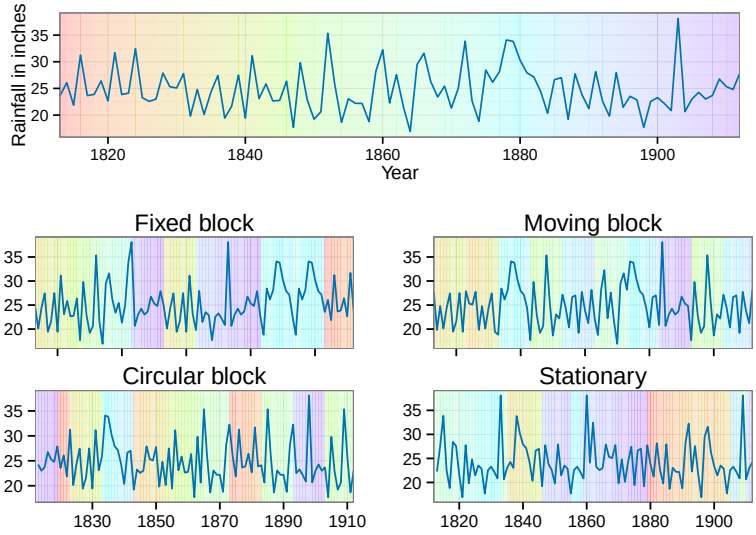


A selection of blocks identified by the circular are shown in the Figure below, in reality the blocks are defined uniformly over the circle.



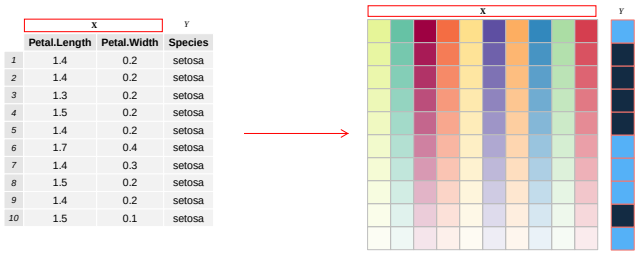
1.1.4 Comparison of block bootstrap variants

Examples of bootstrap replicates constructed using each of these procedures are shown here:



1.1.5 The asynchronous bootstrap

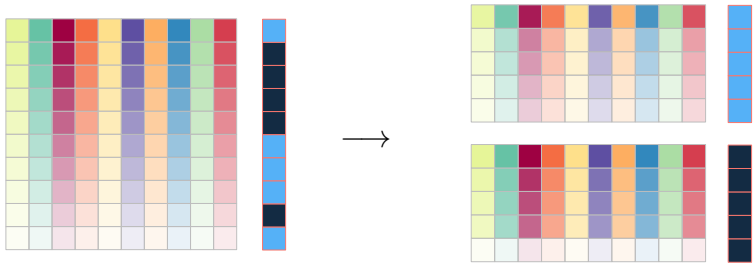
Our proposal for the asynchronous bootstrap asks: can the strategy of dividing a sequence of variables into blocks and resampling the blocks be used without assuming the variables have identical distributions and relationships to their neighbours? The answer is, yes, at least in classification problems. To describe our algorithm, we will represent the covariate matrix $\mathcal{X}$ and a categorical outcome Y as coloured arrays:



Our task is to estimate the distribution of $Y|\mathcal{X}$ and we adopt the convention that components of $\mathcal{X}$ that depend strongly on one another are similar colors.

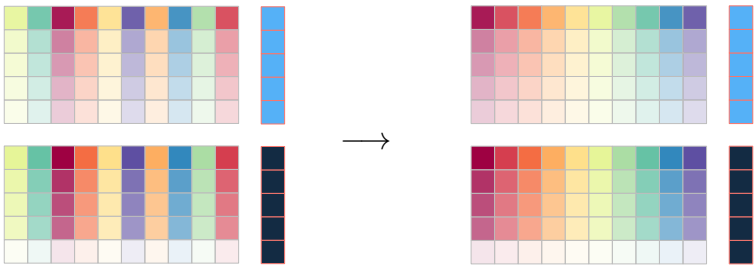
Step 1: Divide

First we divide $\mathcal{X}$ using the values of Y:

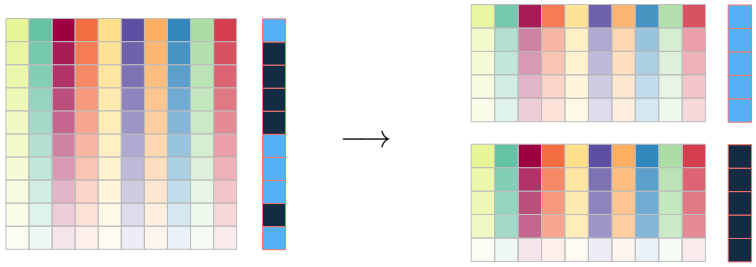


Step 2: Sort

Next we sort the components of $\mathcal{X}$ within each division so that highly dependent columns are near each other:

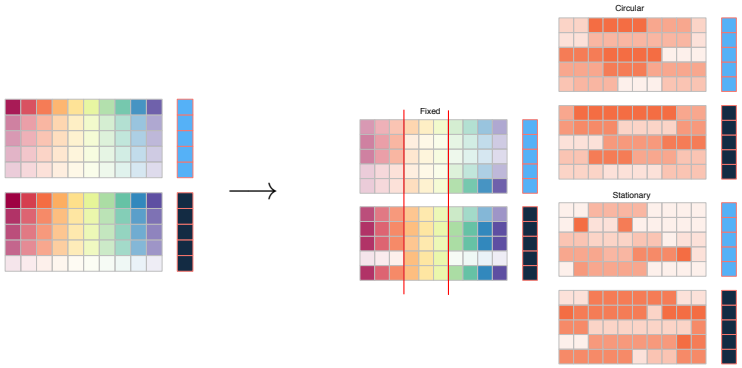


In this hypothetical data set the ordering of the components is the same for each value of Y, but this need not be the case in general.



Step 3: Resample

Now we produce replicates of $\mathcal{X}$ and Y by resampling blocks:



As for the block bootstrap, a variety of methods can be used to allocate blocks. Notice that here replicate components are always drawn from the corresponding component of $\mathcal{X}$. It is this alteration that permits us to relax the assumption of stationarity.

Step 4: Aggregate

Finally we compute the statistic for each replicate data set and aggregate the results.

$$\{G_k(\mathcal{X},Y)\}_{k\in 1,\ldots B} = \left\{ G\left(\right), G\left(\right), \ldots, G\left(\right) \right\}$$

The method of aggregation depends on the motivation a practitioner has in using the bootstrap.

1.1.6 References

The notion of explaining the standard bootstrap with a statistical map in a non-technical setting is borrowed from Diaconis and Efron [1983]. The fixed, moving and circular block bootstraps are due to Hall [1985]; similar ideas are expressed in Carlstein [1986], Künsch [1989], and Politis and Romano [1992]. A further variant of the block bootstrap, the stationary bootstrap, is due to Politis and Romano [1994].

1.2 Causal inference

The remaining two chapters are more accessible so we treat them only briefly here. Chapter 2 assesses whether we can improve on a procedure for estimating heterogeneous treatment effects, the modified outcome method. In that procedure we first construct a very coarse estimate for the individual treatment effect and then use that estimate as the outcome for a classical regression algorithm. We ask whether it is possible to improve the coarse estimate constructed in the first step. We describe a procedure for making such an improvement and present some guarantees. In the final chapter of the thesis we put the techniques of causal inference into practice, and demonstrate the power and limitations of these techniques when they are applied to answer problems of practical significance.

Chapter 2

A bootstrap for high-dimensional classification problems

2.1 Introduction

Because of its strong intuitive appeal and ostensibly conservative assumptions, the bootstrap is used throughout applied statistics. However, it does not always work, and, when it fails, there may be few warnings. Classical examples of failure are given by Bickel and Freedman [1981], Beran [1982] and Andrews [2000]. More recently a literature has developed on the performance of the bootstrap in moderate and high-dimensional problems. Karoui and Purdom [2016, 2018] demonstrate that some varieties of the bootstrap do not provide correct coverage for the parameters of a linear model or the eigenvalues of a covariance matrix.

The issues causing the failure of the bootstrap in high dimensions are subtle, so we shall address them by reference to the better understood problem of estimating a covariance matrix, Σ say, for high-dimensional data, for example when the number of dimensions, p, is much greater than the sample size, n. In this setting it is generally understood that conventional nonparametric estimators of Σ are of little value; there are far too many degrees of freedom, relative to sample size, and at least some parametric structure must be introduced to produce useful estimators. One approach is to assume that the covariance matrix is banded and the number of nonvanishing bands is small; see, for example, Bickel et al. [2004] and the references in the last paragraph of this section. If there are $2m - 1$ bands, one of which is the main diagonal, then an approach such as this is tantamount to assuming that the p-vectors that comprise the data each represent segments, of length p, of an m-dependent time series.

A similar approach can be used to overcome the problem of information scarcity in high-dimensions, by resampling the vectors not as whole entities (this is referred to below as synchronous bootstrap resampling) but as independent blocks (a form of block bootstrap resampling we call the asynchronous bootstrap). The case where the blocks are of length 1, so that the vectors are resampled as though they were all independent, was suggested by Hall et al. [2009] in the context of assessing the authority of rankings. More generally, if the blocks are of length m then the data vectors are being resampled under an assumption of m-dependence, as is imposed implicitly when constructing banded estimators of covariance matrices.

To clarify the discussion above, assume we have a sample $\mathcal{X} = \{X_1, ..., X_n\}$ of independent p-vectors $X_i = (X_{i1}, ..., X_{ip})$. Resampling using the standard, synchronous bootstrap amounts to producing the resample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ by sampling the whole vectors randomly, with replacement, from $\mathcal{X}$. If we were to use the completely independent, asynchronous bootstrap we would replace each X_i^* by the vector that was obtained by independently, for each j, choosing a value randomly from among $X_{1j}, X_{2j}, ..., X_{nj}$, and putting it in the jth position in the resampled vector; and then repeating this operation n independent times, i.e. for $i = 1, ..., n$. The approach suggested in the present paper amounts to a compromise between these two extremes.

In section 3 we shall show, in applications to simulated data, that this modification of the bootstrap can improve significantly the performance of classifiers. Section 4 will explain why, by showing that the asynchronous bootstrap, but not its standard synchronous counterpart, can estimate consistently a nonlinear function of high-dimensional data. The particular relevance of this property to classification, which is the application considered in this paper, will be identified.

Bagging methods for the bootstrap, introduced by Breiman [1996], are often an effective approach to reducing error in relatively complex statistical problems. Bagging works because it reduces variability, and so it has been pressed into use in a wide range of settings. Indeed, the literature on bagging is particularly extensive. To give a flavour of it we mention only a few relatively recent contributions, in particular the work of Bergmeir et al. [2016] and Collell et al. [2018] who studied the application of bagging to neurological data; Bergmeir et al. [2016], who used bagging to forecast economic time series; Biau et al. [2010], who explored properties of bagged nearest neighbour regression estimators; and Hillebrand and Medeiros [2010], who applied bagging to the modelling of volatility in stockmarkets. When used in classification problems, bagging amounts to applying the classifier repeatedly to resampled versions of the dataset, and allocating a new data value to the population to which it is most frequently assigned in these resampling steps.

Prediction problems comprise an important application of the asynchronous boot-

strap methods, where they can be used in combination with Breiman's bagging procedure. In both classification and prediction it is straightforward to use leave-one-out methods to estimate the appropriate block size, or equivalently the number of blocks, and in fact this is a major attractive feature of our methodology. In particular, even though block bagging depends on a tuning parameter, determining its value does not present any practical challenges. However, addressing prediction requires a very different theoretical treatment, as well as a different class of numerical examples; and moreover, the case of prediction is not encountered nearly as commonly, for relatively high-dimensional data, as classification. Therefore we do not treat it here.

We conclude by discussing more of the literature on covariance estimation in relatively high-dimensional problems, since the challenges (although not the methodologies) addressed there are related to those encountered when using the bootstrap with classifiers for high-dimensional data. Ledoit and Wolf [2004] addressed variance estimation by shrinking the covariance matrix towards the identity; Wu and Pourahmadi [2003] suggested, in effect, constructing the estimator as though the data vectors came from varying-coefficient, varying-order regression models; Huang et al. [2006] proposed penalisation methods related to the lasso; Furrer and Bengtsson [2007] used parametric modelling based on the ensemble Kalman filter or the square-root filter; and Fan et al. [2008] employed methods founded on dimension reduction. Related work includes that of Bickel and Freedman [1981], Meinshausen and Bühlmann [2006], Zou et al. [2006], Paul [2007], Johnstone and Lu [2009], Cai et al. [2010, 2012], Cai and Liu [2011], Chen et al. [2011], Fisher and Sun [2011], Negahban and Wainwright [2011] and Rohde and Tsybakov [2011].

2.2 Methodology

2.2.1 Block resampling

Suppose we have a training sample $\mathcal{X} = \{X_1, ..., X_n\}$ comprised of independent p-vectors X_i. Interpret $X_i = (X_{i1}, ..., X_{ip})$ as k blocks of length b:

$$X_i = (X_{i1}, ..., X_{ib}, X_{i,b+1}, ..., X_{i,2b}, ..., X_{ip}) = (B_{i1}, ..., B_{ik}) \tag{2.1}$$

where $B_{ij} = (X_{i,(j-1)b+1}, ..., X_{i,jb})$. Here, for the sake of simplicity, we assume that $p = kb$, but the case where k does not divide p evenly is readily handled; see section 2.2.5.

Compute the resampled blocks B^*_{ij}, for $1 \le i \le n$ and $1 \le j \le k$, by sampling randomly, with replacement, from $B_{1j}, ..., B_{nj}$. Do this independently for each j,

obtaining the following block-resampled analogue of X_i, at (2.1):

$$X_i^*(k) = (B_{i1}^*, \ldots, B_{ik}^*) = (X_{i1}^*, \ldots, X_{ib}^*, X_{i,b+1}^*, \ldots, X_{i,2b}^*, \ldots, X_{ip}^*) \qquad (2.2)$$

Here $X_i^*(k)$ is a block resampled p-vector, and the approach that produced it generalizes the completely independent asynchronous bootstrap suggested by Hall et al. [2009]. We recover the synchronous bootstrap by taking $b = p$, or equivalently, $k = 1$.

The block resampling method here should not be confused with its counterpart for spatial data and time series (e.g. Hall, 1985; Carlstein, 1986; Kunsch, 1989), which, in the context of a time series represented as a vector, takes the positions of the blocks to be random. For example, one of several variants of the conventional block bootstrap for time series would place the resampled block B_{ij}^* into any of the k positions in the representation $X_i^*(k) = (B_{i1}^*, \ldots, B_{ik}^*)$. Those positions would be chosen independently and uniformly from $1, \ldots, k$. This approach exploits the assumed stationarity of the time series, and is necessary because, in the time series case, there is usually only one realisation of the vector X_i. That is, the sample size is $n = 1$. In the context of our work it is generally inappropriate to assume stationarity of the time series $X_{i1}, \ldots, X_{ip}$, but we have the advantage of access to n realisations.

The algorithm discussed above tacitly assumes that there is some sense of order in the components of the dataset. This is not necessarily the case for all datasets, but for these data the block bootstrap can still be very useful; simply apply an algorithm, such as hierarchical clustering, to impose a natural ordering on the components.

2.2.2 Distribution estimation

Assume we wish to construct a classifier to discriminate, on the basis of the dataset $\mathcal{X}$, among L mutually exclusive populations $\Pi_1, ..., \Pi_L$. We take $\mathcal{X}$ to be a training sample for this problem, and in particular, for each $i = 1, ..., n$ we assume that we know the value, $J(X_i)$ say, of the index of the population from which X_i was drawn. Of course, $1 \le J(X_i) \le L$. A general classifier, which we denote by $\mathcal{C}(\cdot|n, \mathcal{X})$, can be viewed as a function from $\mathbb{R}^p$ to the set $\{1, ..., L\}$ and is interpreted as consigning a new data value X to Π_ℓ if $\mathcal{C}(X|n, \mathcal{X}) = \ell$.

Let $\mathcal{X}^* = \mathcal{X}^*(k) = \{X_1^*(k), \ldots X_n^*(k)\}$ denote the version of the training sample $\mathcal{X}$ when it is drawn using block bagging with k blocks. The p-vectors comprising $\mathcal{X}^*(k)$ are independent (conditional on $\mathcal{X}$) versions of $\mathcal{X}_i^*(k)$, at (2.2). Construct the classifier using $\mathcal{X}^*(k)$ rather than $\mathcal{X}$, and compute the conditional probability, $\hat{\pi}_\ell(X|k)$ say, that the classifier assigns X to Π_ℓ, where $1 \le \ell \le L$:

$$\hat{\pi}_\ell(X|k) = \mathbb{P}[\mathcal{C}\{X|n, \mathcal{X}^*(k)\} = \ell \mid \mathcal{X}, X]$$

We interpret $\hat{\pi}_\ell(x|k)$ as an estimator of $\pi_\ell(x) = \mathbb{P}\{\mathcal{C}(x|n,\mathcal{X}) = \ell\}$. Of course, a key assumption in our definition of $\hat{\pi}_\ell$ is that the asynchronous bootstrap estimates consistently the distribution of classification decisions. We shall show in section 4 that it does, but that the standard synchronous bootstrap generally does not.

In practice we compute $\hat{\pi}_\ell(X|k)$ by simulation. That is, we draw B independent versions $\mathcal{X}_1^*(k), \ldots, \mathcal{X}_B^*(k)$ of $\mathcal{X}^*(k)$ and we take our numerical approximation to $\hat{\pi}_\ell(X|k)$ to be:

$$\frac{1}{B}\sum_{r=1}^{B} \mathbb{I}\left[\mathcal{C}\{X|n,\mathcal{X}_r^*(k)\} = \ell\right].$$

2.2.3 Block-bagged classifiers and their error rates

The block-bagged classifier, $\mathcal{C}_{\rm bb}(\cdot|n,\mathcal{X})$, is defined to be the classifier that assigns X to the population Π_ℓ whose index ℓ maximises $\hat{\pi}_\ell(X|k)$:

$$\mathcal{C}_{\rm bb}(X|n,\mathcal{X}) = \arg\max_\ell \hat{\pi}_\ell(X|k). \tag{2.3}$$

Again, this methodology is appropriate only if our estimator $\hat{\pi}_\ell(\cdot|k)$ is consistent for π_ℓ, and that is again a motivation for the block-bagged bootstrap.

The error rate of the classifier $\mathcal{C}_{\rm bb}(\cdot|n,\mathcal{X})$ is

$$\mathrm{err}_{\rm bb}(k) = \mathbb{P}\{\mathcal{C}_{\rm bb}(X|k,n,\mathcal{X}) \neq J(X)\}, \tag{2.4}$$

where X denotes a random p-vector drawn randomly from the mixture of the populations $\Pi_1, ..., \Pi_L$, and is taken to be independent of the training data $\mathcal{X}$; and $J(X) = \ell$ if and only if X was drawn from Π_ℓ. Equivalently we can define $\mathrm{err}_{\rm bb}(k)$ in terms of the prior probability ρ_ℓ of Π_ℓ, for $1 \le \ell \le L$:

$$\mathrm{err}_{\rm bb}(k) = \sum_{\ell=1}^{L} \rho_\ell \mathbb{P}\left[\mathcal{C}_{\rm bb}(X(\ell)|k,n,\mathcal{X}) \neq \ell\right]$$

where the random p-vector $X(\ell)$ has the distribution of a p-vector drawn randomly from Π_ℓ, and $\rho_1 + ... + \rho_L = 1$.

2.2.4 Estimating the error rate of a block-bagged classifier

First we define a version of the classifier $\mathcal{C}_{\rm bb}(\cdot|k,n,X)$, for $(n-1)$- rather than n-samples. Given i_1 in the range $1 \le i_1 \le n$, let $\mathcal{X}(i_1) = \mathcal{X}\backslash\{X_{i1}\}$ denote the $(n-1)$-sample that remains if X_{i1} is deleted from $\mathcal{X}$; let $X_{i_2}^*(i_1,k)$, for $1 \le i_2 \le n$,

denote independent (conditional on $\mathcal{X}(i_1)$) versions of $X^*_{i_2}(k)$, at (2.2), when $\mathcal{X}$ is replaced by $\mathcal{X}(i_1)$; define $\mathcal{X}^*(i,k) = \{X^*_1(i,k) \ldots X^*_n(i,k)\}$; put

$$\hat{\pi}_\ell(X|i,k) = \mathbb{P}\left[\mathcal{C}\{X|n-1, \mathcal{X}^*(i,k)\} = \ell | \mathcal{X}, X\right];$$

analogously to the definition of $\mathcal{C}_{\rm bb}(X|k,n,\mathcal{X})$ at (2.3), define the leave-one-out classifier $\mathcal{C}_{\rm bb}(X|k,n-1,X(i))$ by

$$\mathcal{C}_{\rm bb}(X|k,n-1,X(i)) = \arg\max_\ell \hat{\pi}_\ell(X|i,k)$$

and put

$$\widehat{\rm err}(k) = \frac{1}{n}\sum_{i=1}^n \mathbb{P}\left[\mathcal{C}_{\rm bb}(X_i|k,n-1,\mathcal{X}(i)) \neq J(X_i) \,\middle|\, \mathcal{X}\right]$$

If the prior probabilities ρ_ℓ are asymptotically, and respectively, proportional to the numbers of data in X that come from Π_ℓ, then $\widehat{\rm err}(k)$ is a consistent estimator of ${\rm err}(k)$, at (2.4). In other cases, $\widehat{\rm err}(k)$ can be re-defined by incorporating the probabilities ρ_ℓ as weights, if they are known, or in terms of their estimators, if they are estimated from the data. The value of k can be chosen to minimise $\widehat{\rm err}(k)$.

2.2.5 Block remnants

For the sake of simplicity, in the discussion above we assumed that k divides p. In the majority of cases this will not be the case, and there are a variety of ways of dealing with the problems that this creates. The simplest approaches involve appreciating that the blocks B_{ij}, in (2.1), need not be of the same width, and that they can be increased to accommodate any block remnants that arise through the fact that p/k is not an integer. For example, if $p = kb + c$, where b, c and k are nonnegative integers and $0 \leq c \leq b-1$, then the length of the kth block can be increased from b to $b+c$. Alternatively, if $k \leq c$ then c of the blocks can have their lengths increased by 1 unit. Using either of these approaches, the total number of blocks equals k. On the other hand, we could simply regard the block remnant of length c as a block by itself, in which case there would be a total of $k+1$ blocks. There are many other options of this type.

Alternatively, the exact position of the breaks in the blocking procedure could be randomised in each resample, so that while all blocks that do not include either of the very ends of the p-vectors could be of size b, the length of the first (or last) block would be randomly chosen from the set $\{1, ..., b\}$. This is actually the method adopted in section 3.

2.3 Numerical properties

2.3.1 Simulation settings

To test the effectiveness of the asynchronous bootstrap, we simulate a classification problem: Y_i, indicates the binary class label of the i^{th} observation and the predictor matrix, $X_{ij} = \mu_{kj} + \epsilon_{ij}$ records the value of the j^{th} predictor for the i^{th} observation, here $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, p\}$. The class means, μ_{kj}, are defined such that k indicates the value of Y_i and j the component of X. The mean vector for the first class, μ_{0j}, is zero always and for the second class, μ_{1j}, is zero with probability $1/2$ and otherwise is uniform on $[0, 1]$ or $\left[0, \frac{1}{2}\right]$. The error structure ϵ_{ij} is a zero mean, unit variance Gaussian AR(1) process with autocorrelation function $R(\epsilon_{ij_1}, \epsilon_{ij_2}) = \rho^{|j_1 - j_2|}$. The classifier used is logistic regression with ℓ^1 regularisation.

2.3.2 Empirical effectiveness

A dataset with $p = 500$, $n = 500$ and error autocorrelation parameter $\rho = 0.6$ was analysed. Figure 2.1 records the misclassification error on a test set of $10,000$ observations of ten block bagged classifiers trained on this data set. For each classifier, 60 bootstrap replicates were used and varying numbers of blocks were considered. The improvement attributable to block bagging is dramatic with a relative reduction in mean misclassification rate of 58% over standard bagging (corresponding to block bagging with a single block) and 40% over bagging with a bootstrap that assumes independence amongst the components of X (corresponding to block bagging with 500 blocks).

When we vary n and ρ the advantage of block bagging endures. A variety of choices of n and ρ are shown in Figure 2.2 and summarised in Table 2.1. Block bagging always achieves a substantial and significant reduction in misclassification rate compared to bagging save one simulation when $\rho = 0.3$ and $n = 500$. This may well be a type II error, particularly as the method used to account for the multiple comparisons implicit in the selection of the optimal number of blocks, Bonferroni's correction, is highly conservative. When $\rho = 0.3$ block bagging seems not to confer an advantage over bagging assuming independent components but neither does it seem to degrade performance. Elsewhere, block bagging with an intermediate block size is clearly the best performer except when $\rho = 0.6$ and $n = 200$.

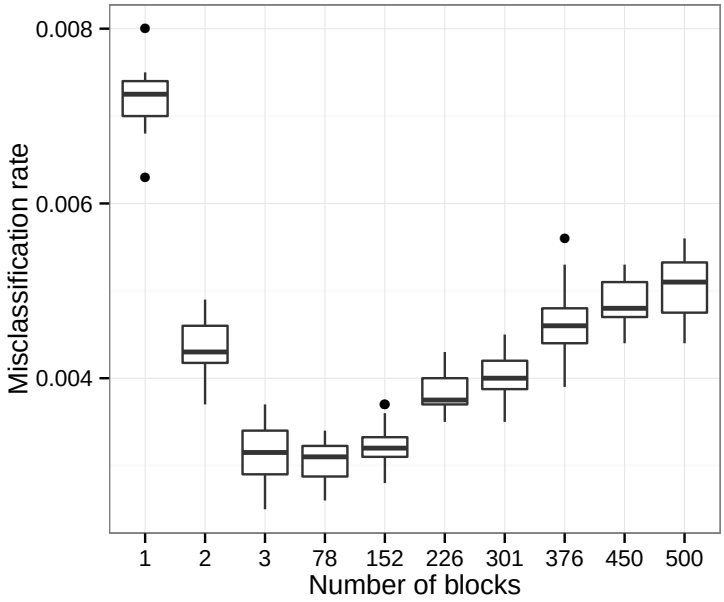


Figure 2.1: The performance of block bagging with regularised logistic regression when $p = 500$, $n = 500$ and $\rho = 0.6$. Note that the numbers of blocks used are not equally spaced and may not include the optimal number of blocks.

2.4 Theory

2.4.1 Summary and remarks

Section 2.4.2 will highlight the role that nonlinearity plays in failure of the standard synchronous bootstrap. In particular, both the synchronous bootstrap and its asynchronous counterpart can estimate consistently a linear function of highly multivariate data, but of these two approaches only the block-bagged, asynchronous bootstrap is effective in nonlinear settings. This is relevant from a practical viewpoint because classifiers are generally highly nonlinear functions of the training data. The simplest way to access this issue seems to be by treating block bagging as a technique for distribution estimation, and so we shall take that approach in section 2.4.2. Section 4.3 will point specifically to failure of the synchronous bootstrap in the context of classification, and demonstrate that the block-bagged, asynchronous bootstrap overcomes that shortcoming.

Throughout section 4 we take dimension, p, to be the basic parameter, and in-

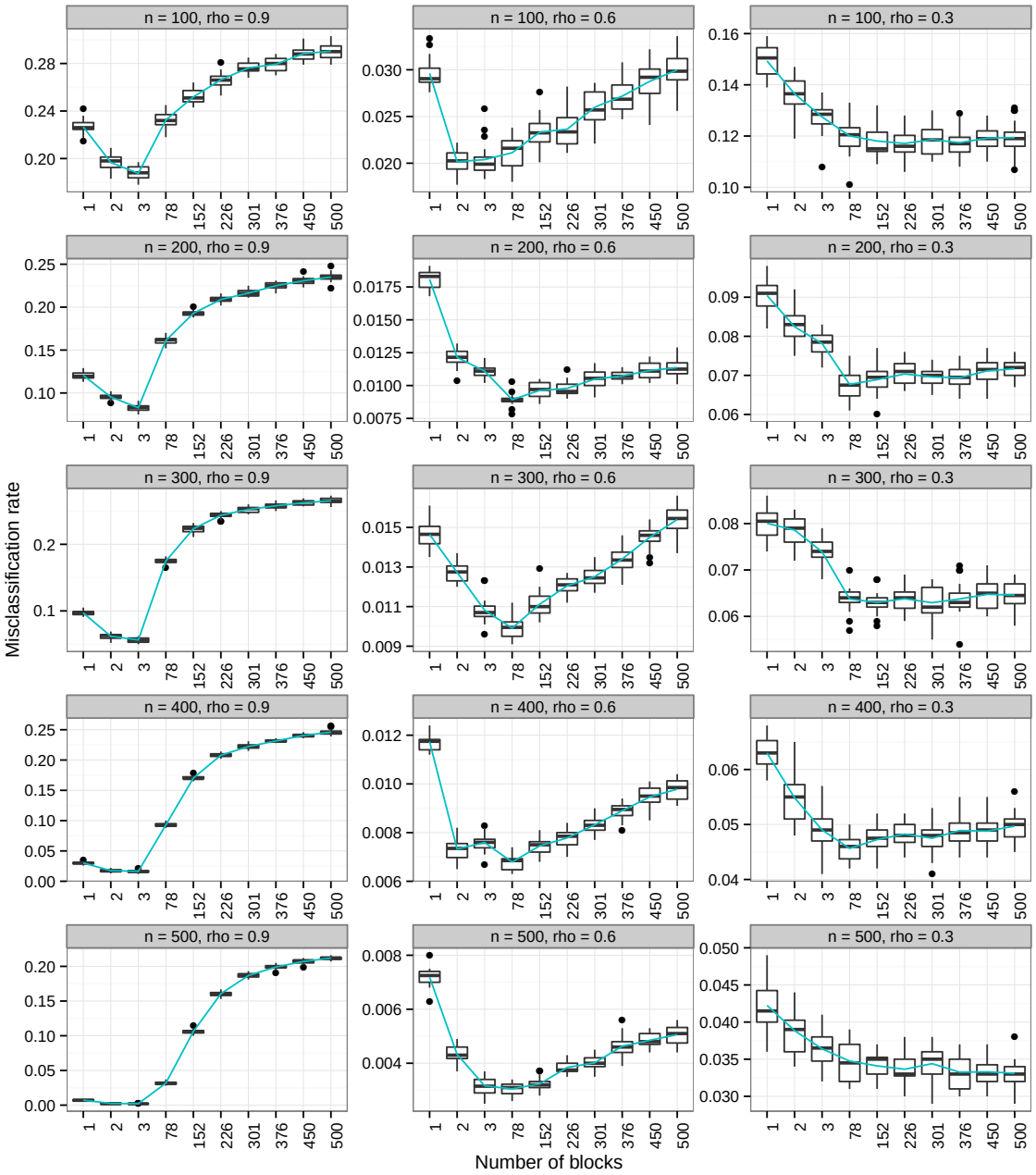


Figure 2.2: Performance of block bagging for a variety of sample sizes and error autocorrelations. In all cases $p = 500$. For $\rho = 0.3$, μ_{1j} is chosen uniformly from $\left[0, \frac{1}{2}\right]$ with probability $\frac{1}{2}$ and is zero otherwise; elsewhere it is chosen uniformly from $[0, 1]$ with probability $\frac{1}{2}$. The special treatment of $\rho = 0.3$ is necessary to make the classification task harder because estimating very low misclassification rates poses computational problems (although these are not insurmountable). For $\rho = 0.6$ a test set of 10000 observations was used, elsewhere a test set of 1000 observations was used.

Table 2.1: Relative reduction in misclassification rate for the optimal block choice (of those tried) compared to bagging ($\hat{\theta}_B$) and bagging assuming independent components ($\hat{\theta}^p_{BB}$). Confidence intervals are computed using Fieller's Theorem. Family-wise error attributable to selecting the optimal block size is controlled using Bonferroni's correction.

ρ	n	θ_B	Bonferroni 95%CI	$\hat{\theta}^p_{BB}$	Bonferroni 95% CI
0.9	500	70%	(35% to 87%)	99%	(98% to 100%)
0.9	400	46%	(3% to 75%)	93%	(90% to 97%)
0.9	300	42%	(26% to 55%)	79%	(74% to 83%)
0.9	200	31%	(16% to 45%)	65%	(57% to 71%)
0.9	100	17%	(8% to 25%)	35%	(29% to 41%)
0.6	500	58%	(45% to 69%)	40%	(10% to 60%)
0.6	400	42%	(32% to 51%)	31%	(15% to 43%)
0.6	300	32%	(17% to 45%)	36%	(19% to 49%)
0.6	200	51%	(39% to 61%)	22%	(-7% to 42%)
0.6	100	32%	(13% to 47%)	33%	(13% to 48%)
0.3	500	22%	(-11% to 42%)	0%	(0% to 0%)
0.3	400	28%	(10% to 42%)	8%	(-19% to 29%)
0.3	300	21%	(2% to 38%)	3%	(-16% to 19%)
0.3	200	25%	(8% to 40%)	6%	(-16% to 25%)
0.3	100	22%	(0% to 38%)	2%	(-27% to 24%)

terpret sample size, n, and (in the case of the asynchronous bootstrap) block length, b, to be functions of p that diverge with increasing p. When using the asynchronous bootstrap we tacitly assume, in statements of results below, that block remnants are dealt with in either of the two ways suggested in section 2.2.5. In technical arguments we suppose for simplicity that $p = bk$, since the two ways of addressing block remnants can be handled using identical arguments, with only notational changes.

2.4.2 Distribution estimation

Our main result in this section is that, while the standard synchronous bootstrap accurately estimates the distribution of linear functions of the data, it can fail spectacularly in nonlinear cases; whereas the asynchronous bootstrap is successful in both linear and nonlinear settings.

To model dependence we take the p-vectors $X_i = (X_{i1}, ..., X_{ip})$, for $1 \le i \le n$, to be independent and identically distributed as translated versions of the first p

components of a zero-mean time series $X_0 = (X_{01}, X_{02}, ...)$. Define

$$Q_j = n^{-1/2} \sum_{i=1}^{n} (X_{ij} - \mu_j),$$

and let g be a function. Given a random variable R, write $(1 - \mathbb{E})R$ to denote $R - \mathbb{E}(R)$, and put

$$S = p^{-1/2} \sum_{j=1}^{p} (1 - \mathbb{E}) g(Q_j). \tag{2.5}$$

A bootstrap estimator of the distribution function $F(x) = P(S \le x)$ is given by

$$\widehat{F}(x) = \mathbb{P}(S^* \le x | X),$$

where

$$S^* = p^{-1/2} \sum_{j=1}^{p} \Big[g(Q_j^*) - \mathbb{E}\{g(Q_j^*) | \mathcal{X}\} \Big], \qquad Q_j^* = n^{-1/2} \sum_{i=1}^{n} \{X_{ij}^*(k) - \bar{X}_j\}$$

$\bar{X}_j = n^{-1} \sum_i X_{ij}$ is the jth component-wise sample mean, X_{ij}^* is defined as at (2.2) in the block-bagged, asynchronous case, and more generally the bootstrap can have either the standard synchronous or the asynchronous form.

We assume that:

(2.6)

(a) for each p the vectors $X_i - \mathbb{E}(X_i) = (X_{i1} - \mathbb{E}\, X_{i1}, ..., X_{ip} - \mathbb{E}\, X_{ip})$, for $1 \le i \le n$, are independent and identically distributed as the vector of the first p components of the time-series X_0, and

(b) X_0 is a stationary, m-dependent process, with $\mathbb{E}\,|X_{01}|^3 < \infty$, zero mean and fixed $m \ge 0$,

Each of conditions (2.6)(a) and (2.6)(b) can be relaxed at the expense of longer theoretical arguments in the proof of Theorems 1 and 2, below, and in particular m can be permitted to increase with p. Third moments are needed principally for the Berry-Esseen bound in (2.9).

Case of linear g

Here, if $g(x) = a + bx$ for constants a and $b \ne 0$, both the conventional synchronous bootstrap and its asynchronous counterpart consistently estimate both the variance

and the distribution of S, as n and p diverge. In particular, if S^* is computed from data vectors X_i^* obtained using either the synchronous bootstrap, or the asynchronous bootstrap with $b \to \infty$ (in this case S^* is interpreted as in (2.5)), then, if (2.6) holds,

$$\frac{\mathbb{E}(S^{*2}|\mathcal{X})}{\mathbb{E}(S^2)} \to 1, \qquad \sup_{-\infty<x<\infty} |\mathbb{P}(S^* \le x|\mathcal{X}) - \mathbb{P}(S \le x)| \to 0, \tag{2.7}$$

where both convergences are in probability. For example, in the case of the synchronous bootstrap the first result in (2.7) follows from the following easily proved results:

$$\mathbb{E}(S^{*2}|\mathcal{X}) = \frac{1}{np}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p}(X_{ij}-\bar{X}_j)\right\}^2 = \frac{1}{np}\sum_{i=1}^{n}\mathbb{E}\left\{\sum_{j=1}^{p}(X_{ij}-\bar{X}_j)\right\} + o_p(1),$$

$$\frac{1}{np}\sum_{i=1}^{n}\mathbb{E}\left\{\sum_{j=1}^{p}(X_{ij}-\bar{X}_j)\right\} = (1-n^{-1})\,\mathbb{E}(S^2).$$

Case of nonlinear g

Here a general theoretical exposition is made more difficult by the wide variety of forms that g can take. We shall simplify matters by tailoring our account to the case where $g(x) = \mathbb{I}(x \le y)$, for a fixed number y. Then, if the vectors (W_{i1}, W_{i2}) are distributed with component-wise variances σ_{i1}^2 and σ_{i2}^2, respectively; and if (Z_1, Z_2) is taken to be normally distributed with the same mean and covariance matrix as the 2-vector $(n^{-1/2}\sum W_{i1}, n^{-1/2}\sum W_{i2})$; then, using the Berry-Esseen theorem for independent two-vectors (see e.g. Götze, 1991), we deduce that:

$$\begin{aligned}
&\left|\operatorname{cov}\left\{g\big(n^{-1/2}\sum_{i=1}^{n}W_{i1}\Big), g\Big(n^{-1/2}\sum_{i=1}^{n}W_{i2}\Big)\right\}\right| \\
&\quad\le \left|\mathbb{P}\Big(n^{-1/2}\sum_{i=1}^{n}W_{i1}\le y, n^{-1/2}\sum_{i=1}^{n}w_{i2}\le y\Big) - \mathbb{P}(Z_1\le y, Z_2\le y)\right| \\
&\qquad\quad + \sum_{\ell=1}^{2}\left|\mathbb{P}\Big(n^{-1/2}\sum_{i=1}^{N}W_{i\ell}\le y\Big) - \mathbb{P}(Z_\ell\le y)\right| \\
&\qquad\quad + \Big|\mathbb{P}(Z_1\le y, Z_2\le y) - \mathbb{P}(Z_1\le y)P(Z_2\le y)\Big| \\
&\quad\le C_1\left\{(1-\theta)^{-3/2}\frac{1}{n^{3/2}}\sum_{i=1}^{n}\big(\mathbb{E}\,|W_{i1}/\sigma_{i1}|^3 + \mathbb{E}\,|W_{i2}/\sigma_{i2}|^3\big) + \theta\right\}
\end{aligned}$$

where $C_1 > 0$ does not depend on n or on the distributions of the W_{ij}s, and

$$\theta = \left| \operatorname{corr} \left(\sum_{i=1}^{n} W_{i1}, \sum_{i=1}^{n} W_{i2} \right) \right|. \tag{2.8}$$

Therefore we shall assume that, with θ given by (2.8),

$$\begin{aligned} &\left| \operatorname{cov} \left\{ g\left(n^{-1/2} \sum_{i=1}^{n} W_{i1} \right), g\left(n^{-1/2} \sum_{i=1}^{n} W_{i2} \right) \right\} \right| \\ &\qquad \le C_1 \left\{ (1-\theta)^{-3/2} \frac{1}{n^{3/2}} \sum_{i=1}^{n} (\mathbb{E}\,|W_{i1}/\sigma_{i1}|^3 + \mathbb{E}\,|W_{i2}/\sigma_{i2}|^3) + \theta \right\}. \end{aligned} \tag{2.9}$$

Theorem 1. *Assume that* (2.6) *holds, and* $n = n(p) \to \infty$ *and*

$$n/p \to 0 \tag{2.10}$$

as $p \to \infty$. *When implementing the block-bagged, asynchronous bootstrap, take the block length,* $b = b(p)$, *to diverge with increasing dimension,* p, *such that* $b^{2+\delta}/n = \mathcal{O}(1)$ *for some* $\delta > 0$. *Then: (a) When using the standard synchronous bootstrap to estimate the distribution of* S^* *in the case* $g(x) \equiv \mathbb{I}(x \le y)$, *both results in* (2.7) *fail. (b) When using the block-bagged, asynchronous bootstrap for a* g *satisfying both* (2.9) *and* $\sup |g| \le C_2$ *(this includes the case* $g(x) \equiv \mathbb{I}(x \le y)$*), both results in* (2.7) *hold.*

Proofs are given in outline in section 5, and in detail in a longer version of the paper available from the authors.

To stress that the synchronous bootstrap fails even when the marginals are independent, and to demonstrate that these problems arise principally when p is of order n or larger, we show below that if the X_{ij}s are completely independent then the first part of (2.7) holds for the synchronous bootstrap if and only if (2.10) holds. A proof of part (a) of Theorem 1 is similar to but simpler than that of the following result.

Theorem 2. *If* (2.6) *holds with* $m = 0$, *and in particular if the variables* X_{ij} *are independent and identically distributed; if*

$$\begin{aligned} &p = O(n^{C_3}) \text{ for some } C_3 > 0, \text{ and } \mathbb{E}\,|X_{01}|^{C_4} < \infty, \text{ where } C_4 \text{ is} \\ &\text{chosen sufficiently large, depending on } C_3; \end{aligned} \tag{2.11}$$

and if we use the standard synchronous bootstrap to estimate the distribution of S; *then the condition* $p/n \to 0$ *is necessary and sufficient for* $\mathbb{E}(S^{*2}|X)/\,\mathbb{E}(S^2)$ *to converge in probability to* 1.

In nonlinear cases, if (2.10) holds, and in particular if the number of dimensions is an order of magnitude larger than the sample size, then the synchronous bootstrap variance estimator, $\mathbb{E}(S^{*2}|X)$, is greater than the true variance by a factor that is asymptotically proportional to p/n.

2.4.3 Classification

For simplicity we assume that there are just two populations, $\Pi^{(0)}$ and $\Pi^{(1)}$. That is, in the notation of section 2.2.2 we suppose that $L = 2$. Assume too that the sample $\mathcal{X} = \{X_1, ..., X_n\}$ is the union of $\mathcal{X}^{(0)}$ and $\mathcal{X}^{(1)}$, where the training sample of size $n^{(r)}$ is drawn by sampling randomly from $\Pi^{(r)}$, and $X^{(r)} = (X_{i1}^{(r)}, \ldots, X_{ip}^{(r)})$ for $r = 0, 1$. Put $\bar{X}^{(r)} = (n^{(r)})^{-1} \sum_{1 \le i \le n^{(r)}} X_{ij}^{(r)}$.

We shall study the centroid classifier, which assigns a new data vector $V = (V_1, ..., V_p)$, independent of the training samples, to $\Pi^{(1)}$ if the quantity

$$
\begin{aligned}
D(V) &\equiv \sum_{j-1}^{p} \left\{ \left(\bar{X}_j^{(0)} - V_j \right)^2 - \left(\bar{X}_j^{(1)} - V_j \right)^2 \right\} = \sum_{j=1}^{p} \left(\bar{X}_j^{(0)} - \bar{X}_j^{(1)} \right) \left(\bar{X}_j^{(0)} + \bar{X}_j^{(1)} - 2V_j \right) \\
&= \sum_{r=0}^{1} \sum_{j=1}^{p} (-1)^r \left\{ \left(\bar{X}_j^{(r)} \right)^2 - 2V_j \bar{X}_j^{(r)} \right\} \\
&= \sum_{r=0}^{1} \sum_{j=1}^{p} (-1)^r \Big\{ \left(\bar{X}_j^{(r)} - \mu_j^{(r)} \right)^2 + 2 \left(\mu_j^{(r)} - V_j \right) \left(\bar{X}_j^{(r)} - \mu_j^{(r)} \right) \\
&\qquad\qquad + (\mu_j^{(r)})^2 - 2V_j \mu_j^{(r)} \Big\} \qquad (2.12)
\end{aligned}
$$

satisfies $D(V) > 0$, and assigns V to $\Pi^{(0)}$ otherwise. (We use the notation V here and below, rather than X as in section 2.2.2, to avoid confusing the jth vector component V_j with the jth data value X_j.) The centroid classifier is popularly applied in contexts ranging from genomics to speech recognition; see, for example, Cootes et al. [1993], Franco-Lopez et al. [2001], Bilmes and Kirchhoff [2003], Dabney [2005], Dabney and Storey [2007], Schoonover et al. [2003], Tibshirani et al. [2002], McKinney et al. [2006], Wang and Zhu [2007] and Sharma and Paliwal [2010]. It also enjoys optimality properties [Hall et al., 2010].

The version of $D(V)$ in the bootstrap case is

$$
D^*(V) \equiv \sum_{j=1}^{p} \left\{ (\bar{X}_j^{*(0)} - V_j)^2 - (\bar{X}_j^{*(1)} - V_j)^2 \right\}
$$

where we define $\bar{X}_j^{*(r)} = (n^{(r)})^{-1} \sum_{j \le i \le n^{(r)}} X_{ij}^{*(r)}$. Here, in the context of the standard synchronous bootstrap, the random variables $X_j^{*(r)}$ are defined by noting that the vectors $X_i^{*(r)} = (X_{i1}^{*(r)}, \ldots, X_{ip}^{*(r)})$, comprising the resample $\mathcal{X}^{*(r)} = \{X_i^{*(r)}, \ldots, X_{n^{(r)}}^{*(r)}\}$, are drawn by sampling randomly, with replacement, from $\mathcal{X}^{(r)}$; and the resampling that produces $\mathcal{X}^{*(0)}$ is independent (conditional on $\mathcal{X}$) of that which gives $\mathcal{X}^*(1)$. The variant of $D^*(V)$ in the case of the blockbagged, asynchronous bootstrap is constructed identically, except that the vectors are now drawn from $\mathcal{X}^{(r)}$ using the block resampling algorithm introduced in section 2.1.

Our theoretical analysis will lead to the following two conclusions. (i) Despite $D(V)$ behaving very conventionally, when viewed as a random variable, the synchronous bootstrap distribution of $D^*(V)$, conditional on $\mathcal{X}$ and V, is not a good estimator of the unconditional distribution of $D(V)$ when dimension is high relative to sample size. We shall demonstrate this by showing first, in Theorem 3, that $D(V)$ behaves in a particularly regular fashion, and then, in Theorem 4, that despite this regularity, if p increases modestly faster than the training sample sizes then the ratio of the synchronous bootstrap estimator of the variance of $D(V)$, to the true variance, diverges to infinity. (ii) If the asynchronous bootstrap is used to estimate the distribution of $D(V)$ then both the variance and the distribution are estimated consistently; see Theorem 5.

These results have immediate impact on use of the bootstrap to estimate error rate in high-dimensional data analysis. In particular, they imply that standard synchronous bootstrap estimators of error rate are inconsistent, and that alternative methodology, such as the asynchronous bootstrap, is necessary. Thus, the theoretical results in this section motivate the techniques introduced in section 2, and explain the numerical properties summarized in section 3. In asymptotic terms, appropriate implementation of the block bootstrap requires the block size, b, or alternatively, the number of blocks, k, to diverge with increasing dimension, but not to increase too fast. Our regularity conditions prescribe rates of increase, but in practice, as illustrated in sections 2.3 and 3, leave-one-out methods can be used very effectively to estimate error rate and thereby to choose b or k.

In preparation for Theorems 3 and 4 we extend (2.6) by allowing the components of X_i to be both non-stationary and dependent; see (2.13) below. There $\mathbb{E}^{(r)}$ denotes the expectation operator applied to functions of data from $\Pi^{(r)}$:

(a) for each $p \geq 1$, and for data from $\Pi^{(r)}$ where $r = 0$ or 1, the p-vector $(X_{i1} - \mathbb{E}^{(r)}(X_{i1}), ..., X_{ip} - \mathbb{E}^{(r)}(X_{ip}))$ is distributed as a zero-mean, m-dependent but not necessarily stationary stochastic process $X_0 = (X_{01}, ..., X_{0p})$;

and X_0 enjoys the properties:

(b) $\sup_{j \geq 1} \mathbb{E}(|X_{0j}|^{4+2\varepsilon}) \leq C < \infty$, and

(c) $\text{var}\left(\sum_{j_1+1 \leq j \leq j_2} X_{0j}^k\right) \leq C(j_2 - j_1)$ whenever $0 \leq j_1 < j_2 \leq p$ and $k = 1$ or 2,

where $\varepsilon > 0$ and C do not depend on p, but m and the distribution of X_0 (and in particular, the marginal distributions) may so depend. (2.13)

Condition (2.13) still makes the simplifying assumption that the populations $\Pi^{(0)}$ and $\Pi^{(1)}$ differ only in terms of their means, although this constraint can be removed at the expense of a longer argument. If the distribution of X_0 were that of the first p components of a stationary process $(X_{01}, X_{02}, ...)$, and if m were bounded, then (2.13) would imply the following condition for $q = 1$ and for a constant τ^2:

(a) $\sum_j \sum_k \{\text{cov}(X_{0j}, X_{0k})\}^2 = p\tau^2 + o(p)$,

(b) $\text{var}(\sum_j X_{0j}^2) = \mathcal{O}(p)$ and

(c) for some $q \in [1, 2]$, $\sup_j \sum_k |\,\text{cov}(X_{0j}, X_{0k})|^q = \mathcal{O}(1)$ as $p \to \infty$. (2.14)

To further simplify discussion we ask that, in one population (say, $\Pi^{(0)}$), the marginal means are all zero; in the other population, just ν of the p marginal means equal η, say, and the other means all vanish; and the training sample sizes, $n^{(0)}$ and $n^{(1)}$, diverge and are of similar sizes:

(a) $\mathbb{E}(X_{1j}^{(0)}) = 0$ for each j,

(b) $\mathbb{E}(X_{1j}^{(1)}) = 0$ for all but ν values of j, for which $\mathbb{E}(X_{1j}^{(1)}) = \eta$, a function of p, (2.15)

(c) $\nu^{2-(1/q)}\eta^2 = o(p/n)$, where q is as in (2.14)(c), and

(d) $n^{(0} \asymp n^{(1)}$ and $n \equiv n^{(0)} + n^{(1)} \to \infty$ as $p \to \infty$,

where $a_n \asymp b_n$ means tha a_n/b_n is bounded away from zero and infinity as n increases.

Theorem 3, below, shows that, under these conditions, the discriminator $D(V)$ behaves very regularly. In particular it has variance of size p/n, where $n = n^{(0)} + n^{(1)}$, and is asymptotically normally distributed.

Theorem 3. *If* (2.13)–(2.15)*, hold with* $\tau > 0$ *in* (2.14)*, and if* $m^{2+(2/\varepsilon)/p} \to 0$ *where* $\varepsilon > 0$ *is as in* (2.13)*, then (a)*

$$\mathrm{var}\{D(V)\} \sim 4p\tau^2 \sum_{r=0}^{1} \left(n^{(r)}\right)^{-1} \tag{2.16}$$

and (b) the distribution of $D(V)$ *is asymptotically normally distributed with this variance, in the sense that*

$$\sup_{-\infty<x<\infty} \left| \mathbb{P}\left\{D(V) \le x\right\} - \mathbb{P}\left[\left\{\mathrm{var}\, D(V)\right\}^{1/2} \mathcal{N} + \mathbb{E}\{D(V)\} \le x\right] \right| \to 0 \tag{2.17}$$

where the random variable $\mathcal{N}$ *has the standard normal distribution.*

Next we show that the standard synchronous bootstrap typically overestimates the variance of $D(V)$ by an order of magnitude. Likewise, the synchronous bootstrap fails to capture the broader distribution of $D(V)$, despite the regularity of that distribution evinced by (2.17). We treat the case of relatively large p, considered as a function of n, but the synchronous bootstrap leads to inconsistency for smaller orders too. In lower dimensional settings, different terms come into play, and so result (2.18) below changes.

Theorem 4. *If* (2.13) *holds with* (2.13)(b) *strengthened to* $\sup_{j\ge1} \mathbb{E}(|X_{0j}|^8) \le C < \infty$*, if the marginal means for both populations are uniformly bounded, if* $D^*(V)$ *is constructed using the conventional synchronous bootstrap, and if* $mn^3/p \to 0$ *as* p

diverges, then the ratio $\mathrm{var}\{D^*(V)|X,V\}/\mathrm{var}\{D(V)\}$ *diverges. More particularly, no matter whether* V *is drawn from* $\Pi^{(0)}$ *or* $\Pi^{(1)}$*,*

$$\frac{\mathrm{var}\{D^*(V)|\mathcal{X},V\}}{\mathrm{var}\{D(V)\}} = \{1+o_p(1)\}\frac{\sigma^4 p}{2\tau^2}\frac{\sum_{r=0,1}(n^{(r)})^{-4}}{\sum_{r=0,1}(n^{(r)})^{-1}} \tag{2.18}$$

Finally we show that the asynchronous bootstrap overcomes these these difficulties.

Theorem 5. *Assume the conditions of Theorems 1 and 4. If* $D^*(V)$ *is constructed using block bagging, where the block size,* $b = b(p)$*, satisfies* $b \to \infty$ *and* $b/p \to 0$ *as* $p \to \infty$*, then, no matter whether* V *is drawn from* $\Pi^{(0)}$ *or* $\Pi^{(1)}$*, the ratio on the left-hand side of* (2.18) *converges in probability to 1; that is,*

$$\mathrm{var}\{D^*(V)|\mathcal{X},V\} = \{1+o_p(1)\{\mathrm{var}\{D(V)\} \tag{2.19}$$

where $\mathrm{var}\{D(V)\}$ *satisfies* (2.16)*; and*

$$\sup_{-\infty<x<\infty}\Bigg|\mathbb{P}\Bigg[D^*(V)-\mathbb{E}\{D^*(V)|\mathcal{X},V\}\le x\Bigg|\mathcal{X},V\Bigg] - \mathbb{P}\big[D(V)-\mathbb{E}\{D(V)\}\le x\big]\Bigg| \to 0 \tag{2.20}$$

in probability.

2.5 Outlines of technical arguments

2.5.1 Outline of the proof of Theorem 1

Preliminaries

We derive only part (b) of the theorem; establishing part (a) is similar to, but simpler than, the derivation of Theorem 2. Assume for notational simplicity that $p = bk$, where b, denoting block length, and k, the number of blocks, are both positive integers. Writing $\mathbb{E}_{\mathcal{X}}$ for expectation conditional on $\mathcal{X}$, we have:

$$S^* = \frac{1}{p^{1/2}}\sum_{j-1}^{p}(1-\mathbb{E}_{\mathcal{X}})g\left[\frac{1}{n^{1/2}}\sum_{i=1}^{n}\{X_{ij}^*(k)-\bar{X}_j\}\right] = \frac{1}{k^{1/2}}\sum_{j=1}^{k}T_j^*,$$

where

$$T_j^* = \frac{1}{b^{1/2}} \sum_{r=1}^{b} (1 - \mathbb{E}_{\mathcal{X}}) g(V_{jr}^*), \quad V_{jr}^* = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left\{ X_{i,(j-1)b+r}^*(k) - \bar{X}_{(j-1)b+r} \right\},$$

and, by the definition of the block-bagging algorithm, if block-bagging is used then the variables T_j^*, for $1 \le j \le k$, are independent conditional on $\mathcal{X}$. Let $\mathrm{var}_{\mathcal{X}}$ and $\mathrm{cov}_{\mathcal{X}}$ denote variance and covariance, respectively, conditional on $\mathcal{X}$, and note that

$$b \,\mathrm{var}_{\mathcal{X}}(T_j^*) = \sum_{r_1=1}^{b} \sum_{r_2=1}^{b} \mathrm{cov}_{\mathcal{X}} \left\{ g(V_{jr_1}^*), g(V_{jr_2}^*) \right\}. \tag{2.21}$$

Outline proof of first part of (2.7)

Careful calculations from (2.9) show that

$$\mathbb{E} \left| \mathrm{cov}_{\mathcal{X}} \{ g(V_{jr_1}^*), g(V_{jr_2}^*) \} \right| = \mathcal{O}(n^{-1/2})$$

uniformly in $1 \le i \le n$, $1 \le j \le p$, $1 \le r_1, r_2 \le b$ and $|r_1 - r_2| > m$. This result, the fact that $n^{-1/2} b \to 0$ (since we assumed that $b^{2+\delta}/n = \mathcal{O}(1)$), and the m-dependence property can be used to prove that

$$\sum_{r_1=1}^{b} \sum_{1 \le r_2 \le b, |r_1 - r_2| > m} \mathbb{E} \left[\left| \mathrm{cov}_{\mathcal{X}} \{ g(V_{jr_1}^*), g(V_{jr_2}^*) \} \right| \right] = \mathcal{O}(n^{-1/2} b^2) = o(b) \tag{2.22}$$

$$\mathrm{var} \left[\sum_{r_1=1}^{b} \sum_{1 \le r_2 \le b, |r_1 - r_2| > m} \mathrm{cov}_{\mathcal{X}} \{ g(V_{jr_1}^*), g(V_{jr_2}^*) \} \right] = o(b^2) \tag{2.23}$$

Combining (2.21), (2.22) and (2.23) we deduce that

$$b \,\mathrm{var}_{\mathcal{X}}(T_j^*) = \sum_{r_1=1}^{b} \sum_{r_2=\max(1, r_1 - m)}^{\min(b, r_1 + m)} \mathbb{E}[\mathrm{cov}_{\mathcal{X}} \{ g(V_{jr_1}^*), g(V_{jr_2}^*) \}] + o_p(b). \tag{2.24}$$

Standard arguments show that

$$\mathbb{E}[\mathrm{cov}_{\mathcal{X}} \{ g(V_{jr_1}^*), g(V_{jr_2}^*) \}] = \mathrm{cov} \{ g(Q_{r_1}), g(Q_{r_2}) \} + o(1) \tag{2.25}$$

and together (2.24) and (2.25) imply that

$$\mathrm{var}_{\mathcal{X}}(T_j^*) = \sum_{r=1}^{2m} \mathrm{cov} \{ g(Q_r), g(Q_m) \} + R_j \tag{2.26}$$

where the random variables R_j are identically distributed and satisfy $R_j = o_p(1)$. Combining (2.26) with analogous results for $\mathrm{cov}_{\mathcal{X}}(T_j^*, T_{j\pm1}^*)$ it can be proved that

$$\mathrm{var}(S^*|\mathcal{X}) = \sum_{r=1}^{2m} \mathrm{cov}\{g(Q_r), g(Q_m)\} + o(1) \tag{2.27}$$

Result (2.5) and the m-dependence property can be employed to show that

$$\mathrm{var}(S) = \sum_{r=1}^{2m} \mathrm{cov}\{g(Q_r), g(Q_m)\} + o(1) \tag{2.28}$$

Together (2.27) and (2.28) imply the following, result equivalent to the first part of (2.7):

$$\mathrm{var}(S^*|X) = \mathrm{var}(S) + o_p(1)$$

Outline proof of second part of (2.7)

It suffices to note that the variables

$$U_j^* = (1 - \mathbb{E}_{\mathcal{X}}) g\left[\frac{1}{n^{1/2}} \sum_{i=1}^{n} \left\{X_{ij}^*(k) - \bar{X}_j\right\}\right]$$

in the formula $S^* = p^{-1/2} \sum_{1\le j\le p} U_j^*$, are b-dependent conditional on $\mathcal{X}$, and to check that the conditions for Berk's (1973) theorem. Appropriate conditional versions of those conditions can be established using stochastic analysis.

2.5.2 Outline of the proof of Theorem 2

In the case of independent marginals, and when $g(x) = \mathbb{I}(x \le y)$, we have $\pi \equiv \mathbb{P}(Q_1 \le y) \to \pi_0$, say, and $v^2 = \pi_0(1 - \pi_0)$ is the limit of $\mathbb{E}(S^2)$. Define $\hat{\pi}_j = \mathbb{P}(Q_j^* \le y|\mathcal{X})$, an estimator of π, and put

$$\Delta_{jk} = \mathbb{P}(Q_j^* \le y, Q_k^* \le y|\mathcal{X}) - \mathbb{P}(Q_j^* \le y|\mathcal{X})\mathbb{P}(Q_k^* \le y|\mathcal{X}).$$

In this notation,

$$\mathrm{var}(S^*|\mathcal{X}) = \frac{1}{p} \sum_{j=1}^{p} \sum_{k=1}^{p} \Delta_{jk} = A_1 + A_2 \tag{2.29}$$

where

$$A_1 = \frac{1}{p} \sum_{j=1}^{p} \Delta_{jj} = \frac{1}{p} \sum_{j=1}^{p} \hat{\pi}_j(1 - \hat{\pi}_j), \qquad A_2 = \frac{1}{p} \sum_{j,k:j\ne k} \sum \Delta_{jk}$$

Now, $\mathbb{E}(\hat{\pi}_1 - \pi)^2 \to 0$ as $n \to \infty$, whence it follows, since the variables $\hat{\pi}_j$ are identically distributed, that $A_1 = \pi(1-\pi) + o_p(1)$. From this property, (2.29), a central limit theorem and the fact that $\pi(1-\pi) \to v^2$, it follows that either part of (2.7) holds, and in particular that $\mathbb{E}(S^{*2}|X) \to v^2$, if and only if

$$A_2 \to 0 \text{ in probability.} \tag{2.30}$$

as $p \to \infty$. The proof of the equivalence of (2.10) and either part of (2.7) is completed by showing that (2.10) is necessary and sufficient for (2.30). First we derive Edgeworth expansions for $\mathbb{P}(Q_j^* \le y\hat{\sigma}_j|\mathcal{X})$ and $\mathbb{P}(Q_j^* \le y\hat{\sigma}_j, Q_k^* \le y\hat{\sigma}_k|\mathcal{X})$, where $\hat{\sigma}_j = \mathrm{var}(Q_j^*|\mathcal{X})$, and use those to show that for a constant $a(y) \ne 0$,

$$\begin{aligned}\Delta \equiv \sum_{j,k:j\ne k}\sum \big\{&\mathbb{P}(Q_j^* \le y\hat{\sigma}_j, Q_k^* \le y\hat{\sigma}_k|\mathcal{X}) \\ &- \mathbb{P}(Q_j^* \le y\hat{\sigma}_j|\mathcal{X})\mathbb{P}(Q_k^* \le y\hat{\sigma}_k|\mathcal{X})\big\} \\ &= a(y)n^{-1}p^2 + o_p(p + n^{-1}p^2).\end{aligned} \tag{2.31}$$

Taylor expansion in the Edgeworth expansion can be used to show that if $\hat{\sigma}_j$ and $\hat{\sigma}_k$ in the definition of Δ are replaced by the true standard deviations, which without loss of generality both equal 1, then (2.31) still holds. If this replacement is made then Δ changes to pA_2, and so (2.31) becomes: $A_2 = a(y)n^{-1}p + o_p(1 + n^{-1}p)$. It is clear from this property that (2.10) is necessary and sufficient for (2.30).

2.5.3 Outline of the proof of Theorem 3

Let $W_j = \sum_{r=0,1}(-1)^r W_j^{(r)}$ where

$$W_j^{(r)} = (\bar{X}_j^{(r)} - \mu_j^{(r)})^2 + 2(\mu_j^{(r)} - V_j)(\bar{X}_j^{(r)} - \mu_j^{(r)}) + \left(\mu_j^{(r)}\right)^2 - 2V_j\mu_j^{(r)}$$

In this notation, $D(V) = \sum_j W_j$; see (2.12). Minor changes to the proof of Berk's (1973) central limit theorem enable it to be shown that, if there exist constants $C, \varepsilon > 0$, not depending on p and such that

$$\mathbb{E}\,|(1-\mathbb{E})n^{1/2}W_j|^{2+\varepsilon} \le C, \quad \mathbb{E}\left\{(1-\mathbb{E})n^{1/2}\sum_{j=j_1+1}^{j_2} W_j\right\}^2 \le C(j_2 - j_1), \tag{2.32}$$

for all $1 \le j \le p$ and $0 \le j_1 < j_2 \le p$; if, as assumed in Theorem 3, the quantity $m = m(p)$, introduced in (2.13), satisfies $m^{2+(2/\varepsilon)}/p \to 0$ as $p \to \infty$, where ε is as in

(2.32); and if, as $p \to \infty$,

$$v_n^{(r)} \equiv \frac{1}{p}\left\{(1-\mathbb{E})(n^{(r)})^{1/2}\sum_{j=1}^{p} W_j^{(r)}\right\}^2 \to 4\tau^2, \tag{2.33}$$

for $r = 0, 1$, where τ is as in (2.14); then, defining $v_n = (n^{(0)})^{-1}v_n^{(0)} + (n^{(1)})^{-1}v_n^{(1)}$, and writing $\mathcal{N}$ for a normally distributed $\mathcal{N}(0,1)$ variable, $D(V)$ satisfies:

$$\sup_{-\infty<x<\infty}\left|\mathbb{P}\{D(V) \le x\} - \mathbb{P}\left[(pv_n)^{1/2}\mathcal{N} + \mathbb{E}\{D(V)\} \le x\right]\right| \tag{2.34}$$

Result (2.34) is equivalent to (2.17) and so implies part (b) of the theorem. The proof of Theorem 3 is completed by establishing (2.32) and

$$\mathrm{var}\left(\sum_{j=1}^{p} W_j^{(r)}\right) \sim 4p\tau^2(n^{(r)})^{-1}, \text{ for } r = 0, 1. \tag{2.35}$$

Note that (A.39) implies both (2.16) and (2.38). The first part of (2.32) follows by direct calculation. A proof of the second part is more complex, and exploits (2.13)(c), (2.14)(c) and (2.15)(a)-(2.15)(c). To derive (A.39), first define $t^2 = \mathrm{var}(\sum_j W_j^{(r)}))$, $t_4^2 = \mathrm{var}\{\sum_j (V_j - \mathbb{E}\,V_j)\mu_j^{(r)}\}$,

$$t_1^2 = \mathrm{var}\left[\sum_{j=1}^{p}\left\{(\mu_j^{(r)} - V_j)(\bar{X}_j^{(r)} - \mu_j^{(r)}) - (V_j - \mathbb{E}\,V_j)\mu_j^{(r)}\right\}\right],$$

$$t_2^2 = \mathrm{var}\left\{\sum_{i=1}^{p}(\bar{X}_j^{(r)} - \mu_j^{(r)})^2\right\}, \qquad t_3^3 = \mathrm{var}\left\{\sum_{j=1}^{p}(\mu_j^{(r)} - V_j)(\bar{X}_j^{(r)} - \mu_j^{(r)})\right\}.$$

It can be proved by lengthy calculations that t_2^2 and t_4^2 both equal $o(p/n)$, and that $n^{(r)}t_3^2 = p\tau^2 + o(p)$. Result (2.35) follows from these properties and the fact that $|t - 2t_1| \le t_2$ and $|t_1 - t_3| \le t_4$, implying that $|t - 2t_3| \le t_2 + t_4$.

2.5.4 Outline of the proof of Theorem 4

Defining $D(V)$ as at (2.12), and

$$D_1^*(V) \equiv \sum_{r=0,1}(-1)^r \sum_{j=1}^{p}\left\{(\bar{U}_j^{*(r)})^2 + 2\bar{U}_j^{*(r)}(\bar{X}_j^{(r)} - V_j)\right\}$$

where $\bar{U}_j^{*(r)} = (n^{(r)})^{-1} \sum_{1 \le i \le n^{(r)}} (\bar{X}_j^{*(0)} - V_j)$, we have: $D^*(V) = D(V) + D_1^*(V)$. The resamples $\mathcal{X}^*(0)$ and $\mathcal{X}^*(1)$ were drawn independently, conditional on $\mathcal{X}$, hence:

$$\begin{aligned} \mathrm{var}\{D_1^*(V)|\mathcal{X}, V\} = \sum_{r=0}^{1} \sum_{j=1}^{p} \sum_{k=1}^{p} \Bigg[& \mathrm{cov}\left\{(\bar{U}_j^{*(r)})^2, (\bar{U}_k^{*(r)})^2 | \mathcal{X}\right\} \\ & + 4(\bar{X}_j^{(r)} - V_j)\, \mathrm{cov}\left\{\bar{U}_j^{*(r)}, (\bar{U}_k^{*(r)})^2\right\} \\ & + 4(\bar{X}_j^{(r)} - V_j)(\bar{X}_k^{(r)} - V_k)\, \mathrm{cov}(\bar{U}_j^{*(r)}, \bar{U}_k^{*(r)} | \mathcal{X}) \Bigg] \end{aligned} \tag{2.36}$$

From (A.50) it can be proved that

$$\begin{aligned} \mathrm{var}(D^*(V)|\mathcal{X}, V) = \sum_{r=1}^{1} \Bigg[& \frac{2}{(n^{(r)})^4} \sum \sum_{i_1 \neq i_2} (\hat{G}_{i_1 i_2}^{(r)})^2 \\ & + \frac{4}{(n^{(r)})^2} \sum_{i=1}^{n^{(r)}} \left\{ \frac{1}{n^{(r)}} \hat{H}_{i1}^{(r)} \hat{H}_{i2}^{(r)} + (\hat{H}_{i1}^{(r)})^2 \right\} \Bigg] \end{aligned} \tag{2.37}$$

where, for $s = 1, 2$,

$$\hat{G}_{i_1 i_2}^{(r)} = \sum_{j=1}^{p} (X_{i_1 j}^{(r)} - \bar{X}_j^{(r)})(X_{i_2 j}^{(r)} - \bar{X}_j^{(r)}), \quad \hat{H}_{is}^{(r)} = \sum_{j=1}^{p} (\bar{X}_j^{(r)} - V_j)^{2-s} (X_{ij}^{(r)} - \bar{X}_j^{(r)})^s.$$

Let $Z_{ij}^{(r)} = X_{ij}^{(r)} - \mathbb{E}(X_{ij}^{(r)})$, $\bar{Z}_j^{(r)} = (n^{(r)})^{-1} \sum_i Z_{ij}^{(r)}$ and $\sigma_j^2 = \mathrm{var}(X_{0j})$. For each j in the range $1 \le j \le p$,

$$\mathbb{E}(\hat{H}_{i1}^{(r)}) = \sum_{j=1}^{p} \mathbb{E}\{\bar{Z}_j^{(r)}(Z_{ij}^{(r)} - \bar{Z}_j^{(r)})\} = \sum_{j=1}^{p} \left\{\sigma_j^2 (n^{(r)})^{-1} - \sigma_j^2 (n^{(r)})^{-1}\right\} = 0$$

and so, using the m-dependence property, $\mathbb{E}\{(\hat{H}_{i1}^{(r)})^2\} = \mathrm{var}(\hat{H}_{i1}^{(r)}) = \mathcal{O}(mp)$. Similarly, since by assumption in Theorem 4 the components of X_0 have eight finite moments, $\mathbb{E}\{(\hat{H}_{i1}^{(r)})^4\} = \mathcal{O}((mp)^2)$. From these results it can be shown by lengthy argument, including bounds to the variance of $\sum_i (\hat{H}_{i1}^{(r)})^2$, that

$$\frac{1}{(n^{(r)})^2} \sum_{i=1}^{n^{(r)}} (\hat{H}_{i1}^{(r)})^2 = \mathcal{O}_p(mp/n) \tag{2.38}$$

Analogously it can be proved that $\mathbb{E}(\hat{H}_{i1}^{(r)}\hat{H}_{i2}^{(r)}) = \mathrm{cov}(\hat{H}_{i1}^{(r)}\hat{H}_{i2}^{(r)}) = \mathcal{O}(mp)$, and similarly, $\mathbb{E}\{(\hat{H}_{i1}^{(r)}\hat{H}_{i2}^{(r)})^2\} = \mathcal{O}(mp^3)$. Therefore,

$$\frac{1}{(n^{(r)})^3}\sum_{i=1}^{n^{(r)}}\hat{H}_{i1}^{(r)}\hat{H}_{i2}^{(r)} = \mathcal{O}_p\big\{(mp/n^2) + (mp^3/n^5)^{1/2}\big\} \tag{2.39}$$

Combining (2.38) and (2.39) we find that

$$\frac{1}{n^{(r)}}\sum_{i=1}^{n^{(r)}}\Big\{\hat{H}_{i1}^{(r)}\hat{H}_{i2}^{(r)} + (\hat{H}_{i1}^{(r)})^2\Big\} = \mathcal{O}_p\left\{(mp/n) + (mp^3/n^5)^{1/2}\right\} \tag{2.40}$$

Noting the definitions of $\hat{G}_{i_1i_2}^{(r)}$, $Z_{ij}^{(r)}$ and $\bar{Z}_j^{(r)}$ we see that if $i_1 \neq i_2$ then

$$\mathbb{E}(\hat{G}_{i_1i_2}^{(r)}) = -\sum_{j=1}^{p}\mathbb{E}\left\{Z_{i_1j}^{(r)}\bar{Z}_j^{(r)} + Z_{i_2j}^{(r)}\bar{Z}_j^{(r)} - (\bar{Z}_j^{(r)})^2\right\} = -\frac{1}{n^{(r)}}\sum_{j=1}^{p}\mathbb{E}(X_{0j}^2)$$

Since the process X_0 in (2.13) is m-dependent then $\mathrm{var}(\hat{G}_{i_1i_2}^{(r)}) = \mathcal{O}(mp)$. Therefore,

$$\begin{aligned}
\sum\sum_{i_1\neq i_2}\mathbb{E}(\hat{G}_{i_1i_2}^{(r)})^2 &= \sum\sum_{i_1\neq i_2}\left\{(\mathbb{E}\,\hat{G}_{i_1i_2}^{(r)})^2 + \mathrm{var}(\hat{G}_{i_1i_2}^{(r)})\right\} \\
&= \sum\sum_{i_1\neq i_2}\left[\frac{1}{(n^{(r)})^2}\left(\sum_{j=1}^{p}\mathbb{E}\,X_{0j}^2\right)^2 + \mathcal{O}(mp)\right] \\
&= \{1+o_p(1)\}\left(\sum_{j=1}^{p}\mathbb{E}\,X_{0j}^2\right)^2 + \mathcal{O}(mn^2p).
\end{aligned}$$

From this formula, the property $mn^2/p \to 0$ (a consequence of the assumption that $mn^3/p \to 0$), and lengthy arguments that include bounds to the variance of $\sum\sum_{i_1\neq i_2}(\hat{G}_{i_1i_2}^{(r)})^2$, it can be shown that

$$\sum\sum_{i_1\neq i_2}(\hat{G}_{i_1i_2}^{(r)})^2 = \{1+o_p(1)\}\left(\sum_{j-1}^{p}\mathbb{E}\,X_{0j}^2\right) \tag{2.41}$$

From (2.37), (2.40) and (2.41) we deduce that, since $mn^3/p \to 0$,

$$\mathrm{var}\{D^*(V)|\mathcal{X}, V\} = \{1+o_p(1)\}2\left(\sum_{j-1}^{p}\mathbb{E}\,X_{0j}^2\right)\sum_{r=0}^{1}(n^{(r)})^{-4} \tag{2.42}$$

Result (2.18) follows from (2.16) and (2.42).

2.5.5 Outline of the proof of Theorem 5

Outline of the proof of (2.19)

As in the proof of Theorem 1 we assume, for simplicity, that $p = bk$, where b and k are positive integers. A variant of property (2.37) can be shown to hold, written to express the block structure of $D_1^*(V)$:

$$\mathrm{var}\{D_1^*(V)|\mathcal{X}, V\} = \sum_{r=0}^{1}\sum_{j=1}^{k} \mathrm{var}\{A_j^{*(r)}(V)|\mathcal{X}, V\}, \tag{2.43}$$

where

$$A_j^{*(r)}(V) = (-1)^r \sum_{t=1}^{b}\left[\left(\bar{U}^{*(r)}_{(j-1)b+t}\right)^2 + 2\bar{U}^{*(r)}_{(j-1)b+t}\left(X^{(r)}_{(j-1)b+t} - V_{(j-1)b+t}\right)\right]$$

So the block variances are:

$$\mathrm{var}\{A_j^{*(r)}(V)|\mathcal{X}, V\} = \sum_{r=0}^{1}\left[\frac{2}{(n^{(r)})^4}\sum\sum_{i_1\neq i_2}(\hat{G}^{(r)}_{i_1 i_2})^2 + \frac{4}{(n^{(r)})^2}\sum_{i=1}^{n^{(r)}}\left\{\frac{1}{n^{(r)}}(\hat{H}^{(r)}_{ji_1}\hat{H}^{(r)}_{ji_2} + (\hat{H}^{(r)}_{ji_1})^2\right\}\right],$$

$$\hat{G}^{(r)}_{i_1 i_2} = \sum_{t=1}^{b}(X^{(r)}_{i_1,(j-1)b+1} - \bar{X}^{(r)}_{(j-1)b+1})(X^{(r)}_{i_2,(j-1)b+t} - \bar{X}^{(r)}_{(j-1)b+t})$$

$$\hat{H}^{(r)}_{ji_s} = \sum_{j=1}^{p}(\bar{X}^{(r)}_{(j-1)b+t} - V_{(j-1)b+t})^{2-s}(X^{(r)}_{i,(j-1)b+t} - \bar{X}^{(r)}_{(j-1)b+t})^s$$

Lengthy calculations can be used to show from (2.43) that

$$\mathrm{var}\{D_1^*(V)|\mathcal{X}, V\} = 4p\tau^2\sum_{r=1}^{1}(n^{(r)})^{-1} + o(p/n). \tag{2.44}$$

Properties (2.16) and (2.44) together imply (2.19).

Outline proof of (2.20)

It can be shown that:

$$(1 - \mathbb{E}_{\mathcal{X}})D^*(V) = \sum_{j=1}^{k}(1 - \mathbb{E}_{\mathcal{X}})A_j^*(V) = \sum_{r=0}^{1}(-1)^r \sum_{j=1}^{k} B_j^{*(r)}, \tag{2.45}$$

where the random variables

$$B_j^{*(r)} = \sum_{t=1}^{b} \Big\{ (1 - \mathbb{E}_{\mathcal{X}})(\bar{U}_{(j-1)b+t}^{*(r)})^2 + 2(\bar{X}_{(j-1)b+t}^{(r)} - V_{(j-1)b+t})(1 - \mathbb{E}_{\mathcal{X},V})\bar{U}_{(j-1)b+t}^{*(r)} \Big\}$$

for $1 \leq j \leq k$ and $r = 0, 1$, are independent and have zero mean, both statements holding conditional on $\mathcal{X}$, and $E_{\mathcal{X}}$ and $E_{\mathcal{X},V}$ denote expectation conditional on $\mathcal{X}$, and expectation conditional on both $\mathcal{X}$ and V, respectively. Therefore Lyapounov's central limit theorem for sums of independent random variables, using a bound on fourth rather than third moments, can be used to give (2.20).

Chapter 3

A simple and doubly robust estimate for heterogeneous treatment effects

3.1 Introduction

Statistical inquiries investigate phenomena that have different manifestations in different circumstances and these differences cannot be captured by statistics that estimate population average or sample average effects. Identifying this heterogeneity has assumed increasing importance in the last quarter century in several domains. For example, technology companies can now conduct experiments on tens or hundreds of millions of subjects, which might provide the power to detect fine-grained heterogeneity; this is desirable in practice because many interventions explored by technology companies have negligible effects on outcomes of interest for all but a small fraction of subjects, so it is important to identify the subjects most likely to have a positive response to avoid wasted expenditure. In addition, there is an increasing focus in medicine in providing treatments that are tailored to the peculiarities of individual patients; this has become viable because there are rich datasets on patients — particularly for those patients in intensive care or those for whom genomic datasets are available.

The notion that the expected treatment effect may differ with observed characteristics has a long history. In medical statistics, the canonical early citation is Bernard [1865]:

> ... in physiology, we must never make average descriptions of experiments, because the true relations of phenomena disappear in the average;

> when dealing with complex and variable experiments, we must study their various circumstances ...

Subgroup estimation is widely used to estimate heterogeneity: discrete (or ordered) covariates are used to define subgroups for which separate treatment effects can be estimated [Byar and Corle, 1977, Simon, 1982, Foster et al., 2011]. Much of this literature provides methods that ensure the validity of tests conducted on estimates for subgroups and discourages the use of data dredging to conduct tests for which Type I error is not controlled. There is also a literature on optimal designs for studies conducted online that acknowledges the potential for heterogeneity: the goal here is to arrive at a treatment policy that allocates each individual to the treatment that will yield the best expected outcome conditional on covariate information [Aitchison, 1970, Zhang et al., 2012, Luedtke and van der Laan, 2016]. More recently a literature has developed that focuses on estimation when there are a large number of covariates relative to the number of samples [Su et al., 2009, Tian et al., 2014, Athey and Imbens, 2016, Wager and Athey, 2017, Künzel et al., 2017]. In particular, our work is complementary to the procedures of Künzel et al. [2017], whose X-learner could be viewed as a kind of regression adjustment.

3.2 The modified outcome method

We will use the Neyman-Rubin model to describe the method under study: each individual i is characterized by the quadruple $(W_i, Y_i(0), Y_i(1), X_i)$, where $W_i \sim$ Bernoulli(p_i) indicates treatment assignment, $Y_i(0)$ is the potential outcome we observe if individual i is assigned to the control group, and $Y_i(1)$ is the outcome we observe if individual i is assigned to the treatment group. The individual treatment effect is then $Y_i(1) - Y_i(0)$; X_i is a d-vector of covariates thought to be associated in some way with the individual treatment effect. This notation implicitly assumes there is no interference: for each i and $W, W' \in \{0,1\}^n$,

$$Y_i(W_1, \dots W_i, \dots, W_n)) = Y_i(W'_1, \dots W_i, \dots, W'_n)) = Y_i(W_i)$$

and there is only one version of treatment: if $W_i = w$, then, with probability one, the observed outcome, Y_i, is equal to the corresponding potential outcome $Y_i(w)$.

Our data consists of the iid sample $(W_i, Y_i(W_i), X_i)_{i=1}^n$ and our goal is to estimate the superpopulation *conditional average treatment effect*:

$$\tau(x) = \mathbb{E}(Y_i(1) - Y_i(0) | X_i = x)$$

where $\mathbb{E}(\cdot)$ denotes expectation with respect to the infinite superpopulation from which the data are drawn. We assume throughout that strong ignorability holds:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i | X_i$$

Using $p(X_i) = \mathbb{E}(p_i|X_i)$ to denote the propensity score, the *modified outcome* is:

$$Y_i^* = \frac{W_i - p(X_i)}{p(X_i)(1-p(X_i))} Y_i = \frac{W_i Y_i(1)}{p(X_i)} - \frac{(1-W_i)Y_i(0)}{1-p(X_i)}$$

It is straightforward to see that under the strong ignorability assumption, $\mathbb{E}_{W_i}(Y_i^*|X_i) = \mathbb{E}(Y_i(1) - Y_i(0)|X_i)$ (where $\mathbb{E}_{W_i}$ means the expectation is taken over the distribution of W_i). So we conclude that any consistent estimator for $\mathbb{E}(Y_i^*|X = x)$ is also consistent for the treatment effect $\tau(x)$. The *modified outcome method* fits a regression model with response Y_i^* and this model is interpreted as an estimate of $\tau(x)$.

It will also be useful to work with a generalized version of this transformation:

Definition 1. *Generalized modified outcome transformation*

$$f^* = \frac{W_i - p(X_i)}{p(X_i)(1-p(X_i))} f$$

Here f can be any function of the data.

This procedure has been discovered several times in the literature: Miller [1976] explored an analogous approach in survival analysis, which can be mapped to our problem by letting p_i be the censoring probability, W_i indicate censoring and Y_i be the survival time. Miller [1976] then advocates using the outcome $W_i Y_i / p_i$ to estimate $\mathbb{E}(Y_i|X_i = x)$. Tian et al. [2014] suggest extensions of the modified outcome method and attribute its first use to Signorovitch [2007]. An approach based on a different transformation (which is in fact the refinement we describe in this paper) is given in Rubin and van der Laan [2007] and Luedtke and van der Laan [2016].

The modified outcome method is seldom used in practice, even when the propensity score is known, because the modified outcome has unnecessarily high variance. By way of example, suppose Fisher's sharp null holds: $\forall i \;\; Y_i(1) = Y_i(0)$; then Y_i^* has mass concentrated on two atoms $Y_i(1)/p(X_i)$ and $-Y_i(1)/(1-p(X_i))$. Now if $Y_i(1)$ is relatively large, then Y_i^* will have very high variance, even if the potential outcomes and the treatment effect are constant. Aside from inducing high variance, the bimodal structure of the transformation is particularly problematic for adaptive local methods. For example, when using unpruned decision trees there is a tendency for observations in leaves to belong entirely to one treatment status. This means for fixed covariate value x, the tree tends to converge to a mixture distribution composed of $Y_i(1)/p(X_i)|X_i = x$ and $-Y_i(0)/p(X_i)|X_i = x$. This phenomenon will typically make decision trees fit to the modified outcome inconsistent. Similarly if p is close to

zero or close to one then the variance of the modified outcome is exceptionally large because we must divide by a number very near zero. When the propensity score itself is unknown, the modified outcome will be worse still, as it relies on a consistent estimator of the propensity score. This is often unachievable, particularly in high dimensions. This situation is remedied by doubly robust estimation.

Definition 2. *Let $\hat{Y}_i(W_i, X_i)$ be an estimate of $\mathbb{E}(Y_i(W_i)|X_i)$ and $\hat{p}(X_i)$ estimate $p(X_i)$. An estimate $\hat{\tau}(\hat{Y}, \hat{p}_i)$ of a causal effect τ is doubly robust iff whenever $\hat{Y}_i(W_i, X_i) \to \mathbb{E}(Y_i(W_i)|X_i)$ or $\hat{p}(X_i) \to p(X_i)$ then $\hat{\tau} \to \tau$*

Double robustness is a desirable property of many estimators of causal effects and we will see in the sequel that the refinement we propose enjoys this property.

3.3 Regression adjustment for the modified outcome

We aim to repair the modified outcome method by using regression adjustment. To develop this we consider the modified outcome an estimate of the individual treatment effect, so with no regression adjustment the modified outcome has risk:

$$\mathcal{R}(Y_i^*) = \mathbb{E}_{Y_i(1),Y_i(0),W_i}\left[(Y_i^* - [Y_i(1) - Y_i(0)])^2\right] \tag{3.1}$$

We aim to find a better transformation by considering the risk of a regression adjusted modified outcome:

$$\mathcal{R}(Y_i^* - f_i(Y, X, W)) = \mathbb{E}_{Y_i(1),Y_i(0),W_i}\left[([Y_i^* - f_i(Y, X, W)] - [Y_i(1) - Y_i(0)])^2\right] \tag{3.2}$$

We call estimators of this form **RAMO** standing for *regression adjustment for the modified outcome*. Unfortunately, optimizing this criterion directly is cumbersome and depends on detailed properties of the family from which f_i is chosen and the functions $x \mapsto Y_i(1)|X_i = x$ and $x \mapsto Y_i(0)|X_i = x$, which are not known *a priori*. Before addressing this complexity, we can at least determine the correct estimand for regression adjustment. To do this we idealize our problem by allowing our regression adjustment to depend on all potential outcomes (even unobserved potential outcomes) and suppose that we seek to minimize the variance of the modified outcome after regression adjustment, subject to the condition that the adjustment does not introduce bias:

$$\begin{aligned} f_i = \arg\min_{f_i} \quad & \mathbb{V}_{W_i}\left[Y_i^* - f_i(Y(1), Y(0), p(X_i), W)|Y_i(1), Y_i(0)\right] \\ \text{subject to} \quad & \mathbb{E}_{W_i}[f_i(Y(1), Y(0), p(X_i), W)|Y_i(1), Y_i(0)] = 0 \end{aligned} \tag{3.3}$$

This optimization can be motivated by analogy to uniform minimum variance unbiased estimation. The advantage of the idealization is that it permits an easy solution:

Proposition 1. *The unique solution to (3.3) is:*

$$f_i = \frac{W_i - p(X_i)}{p(X_i)(1-p(X_i))}\left[Y_i(1)(1-p(X_i)) + Y_i(0)p(X_i)\right]$$

and this satisfies:

$$\mathbb{V}_{W_i}(Y_i^* - f_i) = 0$$

Proof. We observe

$$\begin{aligned} Y_i^* &= \frac{W_i Y_i(1)}{p(X_i)} - \frac{(1-W_i)Y_i(0)}{1-p(X_i)} \\ &= Y_i(1) - Y_i(0) + \left(\frac{W_i - p(X_i)}{p(X_i)}\right) Y_i(1) + \left(\frac{W_i - p(X_i)}{1-p(X_i)}\right) Y_i(0) \\ &= Y_i(1) - Y_i(0) + \frac{W_i - p(X_i)}{p(X_i)(1-p(X_i))}\left[Y_i(1)(1-p(X_i)) + Y_i(0)p(X_i)\right] \end{aligned}$$

We can verify that the expectation of the third summand is zero and the remainder after subtracting it from the modified outcome is equal to the treatment effect and so $\mathbb{V}_{W_i}(Y_i^* - f_i) = 0$. □

So, if $\hat{Y}_i(1)$ estimates $Y_i(1)$ and $\hat{Y}_i(0)$ estimates $Y_i(0)$, we expect that many desirable adjustments can be written in the form:

$$Y_i^* - \hat{f}_i = \left[Y_i - (\hat{Y}_i(1)(1-p(X_i)) + \hat{Y}_i(0)p(X_i))\right]^*$$

The effectiveness of regression adjustment is illustrated in a simple synthetic data example in Figure 3.1.

In practice this adjustment is not feasible because it depends on both potential outcomes and the best unbiased regression adjustments that is feasible is:

$$\begin{aligned} \hat{f}_i = \arg\min_{\hat{f}_i} \quad & \mathbb{V}\left[Y_i^* - f_i(X, Y, W, p(X_i))\right] \\ \text{subject to} \quad & \mathbb{E}[f_i(X, Y, W, p(X_i))] = 0 \end{aligned} \tag{3.4}$$

Note that in (3.4) we permit all observations to be used to compute the regression adjustment and only allow observed potential outcomes to be used when constructing the regression adjustment. This makes things more difficult and, in general,

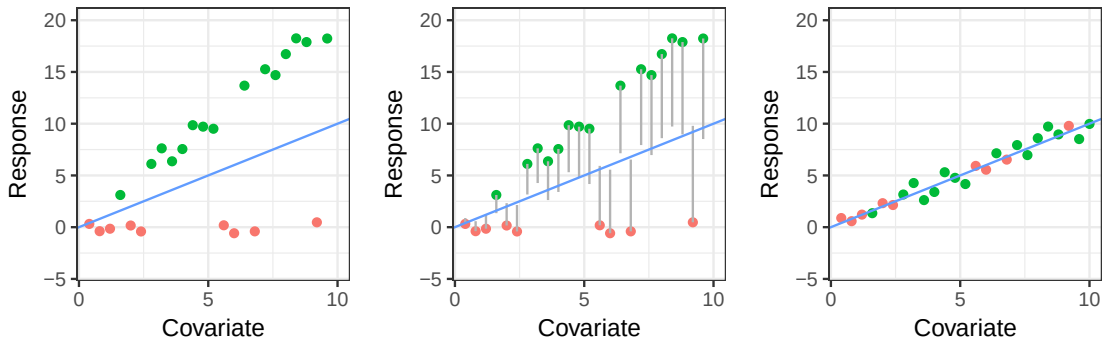


Figure 3.1: The first panel shows the modified outcome, with each point colored according to its treatment assignment; the bimodal struture of the transformation is evident; the second panel shows our estimates of the optimal regression adjustment estimand; and the final panel shows the modified outcome after regression adjustment. In this setting $Y_i(0) \sim \mathcal{N}(0,1)$, $Y_i(1) \sim \mathcal{N}(2X_i, 1)$ and $X_i \sim \mathcal{U}[0,10]$. The blue line is the CATE, the green points are the outcomes for individuals assigned to the treatment group and the red points are the outcomes for individuals assigned to the control group.

there is no uniformly optimal solution as the optimum depends on properties of the random functions $x \mapsto Y_i(1)|X_i = x$ and $x \mapsto Y_i(0)|X_i = x$. We will give strategies for finding optimal regression adjustments that hold for all UMVU, minimax and admissible estimators at the same time. We begin by formally defining the three types of optimality. Throughout we suppose the potential outcomes are generated from a family of distributions indexed by a (possibly infinite dimensional) parameter: $(Y_i(1), Y_i(0))|X_i \sim \mathbb{P}_\theta$ where $\mathbb{P}_\theta \in \mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$.

In this section we will work with a generic loss function $L_\theta(\cdot, \cdot)$ which could be, for example, the squared error loss: $L_\theta(\tau_i, \hat{\tau}_i) = (\tau_i - \hat{\tau}_i)^2$.

Definition 3 (Optimality)**.** *A regression adjustment $\hat{f}_i$ is **admissible** using the loss $L_\theta(\cdot, \cdot)$ with respect to a family of distributions on the potential outcomes $\mathcal{P}$ if there is no other regression adjustment $\tilde{f}$ which is uniformly better:*

$$\forall \theta \in \Theta, \quad \sum_{i=1}^{n} \mathbb{E}[L_\theta(Y_i^* - \hat{f}_i, \tau_i)|X_i] \leq \sum_{i=1}^{n} \mathbb{E}[L_\theta(Y_i - \tilde{f}_i, \tau_i)|X_i]$$

and

$$\exists \theta \in \Theta, \quad \sum_{i=1}^{n} \mathbb{E}[L_\theta(Y_i^* - \hat{f}_i, \tau_i)|X_i] < \sum_{i=1}^{n} \mathbb{E}[L_\theta(Y_i - \tilde{f}_i, \tau_i)|X_i]$$

A regression adjustment is ***minimax*** *if*

$$\sum_{i=1}^{n} \mathbb{E}[L_\theta(Y_i^* - \hat{f}_i, \tau_i)|X_i] = \inf_{\tilde{f}} \sup_{\theta} \sum_{i=1}^{n} \mathbb{E}[L_\theta(Y_i^* - \tilde{f}_i, \tau_i)|X_i]$$

A regression adjustment is ***uniform minimum variance unbiased*** *[UMVU] if* $\forall\, \mathbb{P}_\theta \in \mathcal{P}$, $\mathbb{E}(\hat{f}_i|X_i) = 0$ *and*

$$\sum_{i=1}^{n} \mathbb{V}(Y_i^* - \hat{f}_i|X_i) = \inf_{\tilde{f}_i : \mathbb{E}(\tilde{f}_i)=0} \sum_{i=1}^{n} \mathbb{V}(Y_i^* - \tilde{f}_i|X_i)$$

Next we consider a correspondence between optimal estimates of the treatment effect and optimal regression adjustments. These results collectively suggest that all optimal estimators can be cast as regression adjustment estimators.

Definition 4. *A modified outcome estimator with regression adjustment,* $\hat{\tau}$*, is doubly admissible, minimax, or UMVU if the adjustement is admissible minimax or UMVU according to Definition 3 and* $\hat{\tau}$ *is admissible, minimax, or UMVU estimate of the treatment effect with respect to the data* $(Y_i^* - \hat{f}_i, X_i)_{i=1}^n$.

Theorem 6. *If* $\hat{\tau}$ *is an admissible, minimax, or UMVU estimate of the treatment effect then it is equal to a doubly admissible, minimax, or UMVU regression adjustment estimator of the modified outcome with respect to the expected sample loss:*

$$\sum_i \mathbb{E}(L_\theta[\tau_i(X_i), \hat{\tau}(X_i)]|X_i)$$

Proof. We prove this holds for admissible estimators, the proofs for minimax and UMVU estimators are identical. Suppose that $\hat{\tau}$ is an admissible estimator and define the adjustment:

$$\hat{f}_i = \hat{\tau}_i - Y_i^*$$

We can rewrite the loss for the regression adjustment using $\hat{f}_i$ as

$$\sum_{i=1}^{n} L_\theta(Y_i^* - \hat{f}_i, \tau_i) = \sum_{i=1}^{n} L_\theta(\hat{\tau}_i, \tau_i)$$

So we conclude that this must be an admissible adjustment because $\hat{\tau}$ is admissible by assumption. Next consider the following estimator based on the modified outcome with regression adjustment:

$$\tilde{\tau}_i(X, Y^* - \hat{f}_i) = Y_i^* - \hat{f}_i = Y_i^* - (\hat{\tau}_i - Y_i^*) = \hat{\tau}_i$$

We observe that $\hat{\tau} = \tilde{\tau}$ and because the class of estimators over which $\hat{\tau}$ is admissible is strictly larger than the class of estimators over which $\tilde{\tau}$ is admissible the conclusion follows.

□

To show the modified outcome with regression adjustment is doubly robust it suffices to show that if $\hat{p}_i \to_p \mathbb{E}(W_i|X_i)$ or $\hat{Y}_i(W_i) \to_p \mathbb{E}(Y_i(W_i)|X = X_i)$ then

$$\mathbb{E}_{W_i}(Y_i^* - \hat{f}_i|X_i) \to_p Y_i(1) - Y_i(0)$$

Establishing this is a straightforward computation.

3.4 Simulations

A comparison of RAMO to the methods and a selection of the simulation settings explored in Künzel et al. [2017] are shown in figure 3.2. In all simulations $Y_i(W_i) = \mu_i(W_i) + \varepsilon_i$ where the definition of $\mu_i(W_i)$ varies across settings and $\varepsilon_i \sim \mathcal{N}(0, 1)$. The rows of the covariate matrix X are iid and multivariate normal with mean 0 and covariance matrix generated by the C-vine method described in Lewandowski et al. [2009]; this method requires the selection of a tuning parameter α, which we set to 0.3. We generate training and testing sets each with 100 observations and allow the dimension of X to vary so that its aspect ratio (p/n) ranges between 0.1 and 1. Four response functions are considered:

(1) $\mu_i(0) = X\beta_0$ where the entries of β_0 are iid $\mathcal{U}(1, p)$
$\mu_i(1) = X\beta_1$ where the entries of β_1 are iid $\mathcal{U}(1, p)$
$p_i = 0.5$

(2) $\mu_i(0) = \sin(X_{i1}) + \sin(X_{i2}) + \sin(X_{i3}) + \sin(X_{i4})$
$\mu_i(1) = \mu_i(0) + 0.3\mathbb{I}(X_{i2} > 0.1)$
$p_i = 0.5$

(3) $\mu_i(0) = 3(X_{i1}) + 5X_{i2}$
$\mu_i(1) = \mu_i(0) + 30X_{i3}$
$p_i = 0.1$

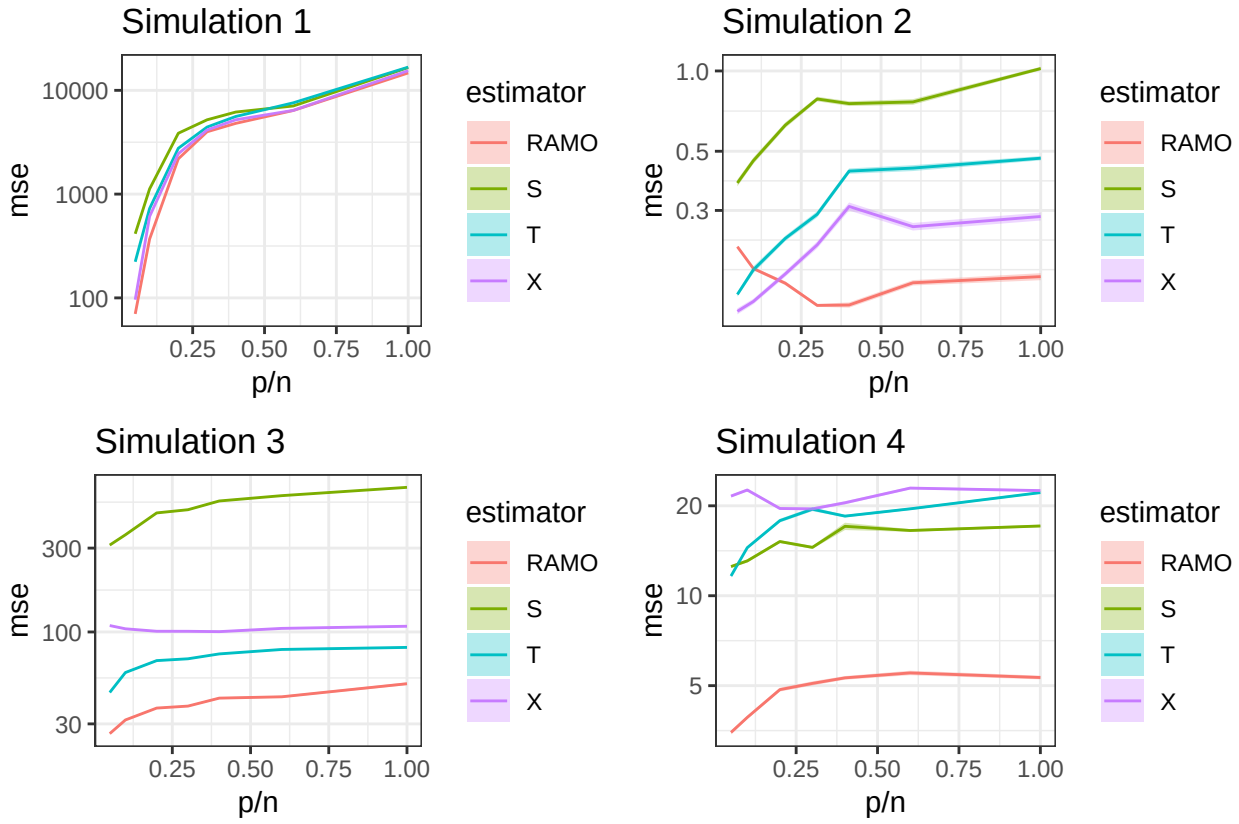


Figure 3.2: Performance of RAMO on synthetic datasets.

(4) $\mu_i(0) = 3(X_{i1}) + 5X_{i2}$
$\mu_i(1) = \mu_i(0) + 30X_{i3}$
$p_i = 0.5$

Inspection of the top left panel of Figure 3.2 suggests that no method appears to enjoy a substantial advantage or disadvantage for the first simulation setting; however in the remaining three settings RAMO is superior with its superiority increasing with the dimension of the problem.

3.5 Conclusion

This work demonstrates that, when regression adjustment is used, the modified outcome deserves a prominent place in the toolbox of applied statisticians working to estimate heterogeneous treatment effects in practice. It enjoys a variety of optimality properties and is simpler to use than several competing procedures for estimating heterogeneous treatment effects.

Chapter 4

Once ticketed, twice shy

4.1 Introduction

Motor vehicle accidents inflict a devastating toll on human life and well-being. In 2010, they killed 1.3 million people worldwide (3% of all deaths) and caused 78 million injuries serious enough to require medical care.[1] They rank 8th among the leading causes of premature mortality,[2] and are projected to rise to 4th by 2030.[3]

Much of this injury burden falls on developing countries. Developed countries, and some middle-income countries, have made huge gains in road safety over the last 50 years. The decline in road traffic injuries—due primarily to safer vehicle and roadway redesign, seatbelts, and reductions in speeding and drunk driving—stands as one of the great public health victories of the twentieth century.[4] However, motor vehicle accidents remain a major cause of mortality and morbidity in rich countries.

[1]Global Road Safety Facility. The World Bank & Institute for Health metrics and Evaluation, University of Washington. Transport for health: the global burden of disease from motorized road transport. Seattle, WA; Washington DC: The World Bank, 2014; Lozano R, et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. The Lancet, 380(9859), 2095-2128.

[2]Murray CJL, et al. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. The Lancet, 380: 2197-2223.

[3]Mathers CD, Loncar D (2006) Projections of Global Mortality and Burden of Disease from 2002 to 2030. PLoS Med 3(11): e442.

[4]Peden M, Scurfield R, Sleet D, et al (eds). World Report on Road Traffic Injury Prevention. Geneva: World Health Organization, 2004; Dellinger AM, Sleet DA, Jones BH. Drivers, Wheels, and Roads: Motor Vehicle Safety in the Twentieth Century, in Ward JW, Warren C. Silent Victories: The History and Practice of Public Health in Twentieth-Century America. New York, NY: Oxford University Press, 2006.

In the United States, for example, nearly 35,000 people die on the road each year and 2.3 million are injured,[5] at an estimated total cost of $ 100 billion.[6] Part of the immense social cost stems from the disproportionately high incidence of car crashes among the young: car crashes are the leading cause of death and injury among Americans 4-34 years of age.[7]

Laws governing the use of mechanically propelled vehicles appeared in the mid-nineteenth century, well before petrol-powered automobiles were commercially available.[8] The "red flag" traffic laws, enacted by the British Parliament in 1865, are recognized as among the first. They imposed a speed limit of four miles per hour and mandated a three-man crew for every vehicle, one of whom was to walk ahead to warn bystanders of the approaching car by means of "a red flag constantly displayed".[9]

Today, traffic laws are ubiquitous. Getting a ticket, or hearing of a friend or family member who got one, is not exactly a routine event, but it might be the occasion for no more surprise and anguish than would greet, say, losing one's credit card or having a particularly bad day at work. Governments promulgate elaborate lists of road rules, which are designed to mitigate hundreds of different behaviours—from speeding and stoplight breaches to carriage of excessive loads and driving under the influence of alcohol. Many—though clearly not all—of these sanctioned behaviours are empirically-established risk factors for accidents.[10]

Enforcement occurs on a vast scale. A 2010 analysis of US state courts counted 58 million traffic offenses under judicial management in that year; they accounted for 54% of the aggregate trial court caseload.[11] The vast majority of these charges

[5]National Highway Transportation Safety Agency. Traffic safety facts: 2012 Data. DOT HS 812 016 (May 2014).

[6]Naumann RB, Dellinger AM, Zaloshnja E, Lawrence BA, Miller TR. Incidence and total lifetime costs of motor vehicle-related fatal and nonfatal injury by road user type, United States, 2005. Traffic Inj Prev 2010;11:353-60.

[7]Centers for Disease Control. Nonfatal, motor vehicle-occupant injuries (2009) and seat belt use (2008) among adults—United States. MMWR 2011;1681-86; Subramanian R. Motor vehicle traffic crashes as a leading cause of death in the United States, 2006. DOT HS 811 226 (Oct 2009).

[8]Oliphant K. Tort law, risk, and technological innovation in England. 59 McGill L.J. 819 (2014).

[9]The Locomotive Act 1865, 28 & 29 Vict. C. 83, s.3.

[10]Some behaviours are sanctioned because of their role in reducing the severity of injuries when accidents occur, rather than the incidence of accidents. Rules regarding use of seatbelts and motorcycle helmets are two examples.

[11]LaFountain R, Schauffler R, Strickland S, Holt K. Examining the work of state courts: An analysis of 2010 state court caseloads. (National Center for State Courts 2012). (Note: The category used in this analysis is actually titled "traffic/violations" , and it includes some non-traffic related violations, such as breaches of ordinances. However, state-specific sub-analyses presented in the report suggest that traffic offenses account for about 90-95% of the total counts in this category.)

are resolved outside court.[12] Nonetheless, with a median incidence of 18 offenses per 100 persons per year,[13] traffic law violations must surely rank as the most common point of contact Americans have with any punitive side of the legal system.

Although the detail, scale, and reach of traffic laws expanded dramatically over the twentieth century, the core rationale of these regimes remains essentially unchanged from the red flag days: they exist to protect the public's health. The standard account of how this outcome is achieved turns on deterrence: sanctioning risky driving practices discourages them, thereby improving safety.[14]

To what extent do traffic laws actually achieve their foundational safety objectives? And in what ways does deterrence shape driver behaviour? This study aimed to add to the empirical evidence base available to answer those fundamental questions. Our focus was specific deterrence. We searched for evidence of its imprint on recidivism and crash rates, drawing on a large dataset of driver, offense, and crashes records from the Australian state of Queensland.

In the next section, we sketch a simple theoretical model of deterrence in the context of traffic penalties. Part III reviews the literature on traffic law deterrence. Part IV discusses some of the difficulties with causal inference in this area. Part V describes our study approach. Part VI reports results. Part VII discusses the study findings and considers their implications for law and road safety policy. Part VIII concludes.

4.2 Pathways and targets

Classic deterrence theory describes two distinct mechanisms of action.[15] "General deterrence" refers to the threat of punishment prevailing in society at large. In the road traffic context, this is the diffuse signal emanating from the very existence of

[12]The figures reported by the National Center for State Courts are based on tallies across state courts. Cases from single-tiered courts, courts of general jurisdiction, courts of limited jurisdiction are combined. In some states parking tickets fall under court jurisdiction, in which case they contributed to the caseload totals from these jurisdictions. However, many states assign parking ticket enforcement to a separate administrative agency, in which case they do not figure in caseload totals. But even in states where parking tickets are excluded, traffic offenses remain a very large proportion of total court caseloads. Take California, whose case total exclude parking tickets. The state had 6.4 million incoming traffic offenses in 2010, which represents 61% of all civil, criminal, and other cases in the state court system.

[13]LaFountain et al, supra note 11.

[14]Andenaes J. Punishment and deterrence. Ann Arbor: University of Michigan Press, 1974.

[15]Zimring FE, Hawkins GJ. Deterrence: the legal threat in crime control. Chicago: Univ Chicago Press, 1973; Shavell S. Foundations of economic analysis of law. Cambridge, MA: Harvard University Press, 2004; Kahan D. The secret ambition of deterrence. Harv Law Rev. 1999;113:413–500

a rule or regime; drivers seek to obey road rules because they realize they risk fines and penalties if they break them. "Specific deterrence" comes from direct personal experience. Drivers who infringe road rules get caught and are penalized, and then learn their lesson; they become less likely to reoffend, which indirectly leads them to drive more safely and have fewer accidents.

Figure 1 presents a simple illustration of the pathways through which general and specific deterrence are theorized to influence road safety. General deterrence is scattershot; it may simultaneously affect drivers' propensity to offend, the safety with which they drive, and their risk of crashing. Specific deterrent signals travel along a more structured pathway. The penalty experience reduces propensity to offend and to drive dangerously, in that order, or possibly simultaneously. The net effect of those behavioral changes is a reduction in crash risk.

The classic model of deterrence has long been an intellectual punching bag. There are many lines of attack. The clarity of the conceptual distinction between general and specific deterrence, for example, is hotly debated. Criminal justice scholars have also questioned the severability of deterrent effects from other determinants of behavior change, such as incapacitation and rehabilitation. And Stafford and Warr's widely-discussed reconceptualization of deterrence theory posits that far too little attention has been paid to the "anti-deterrent" effects that flow from personal experiences with committing offenses that go unsanctioned.[16]

We would readily agree that the orthodox accounts of deterrence are incomplete and, of particular relevance to our study, that the dividing lines between different forms of deterrence are not always bright. Nonetheless, in describing the behavioral effects under investigation among Queensland drivers we generally stick with the traditional constructs and nomenclature. We do so partly for reasons of expediency. But it is also worth pointing out that a couple of aspects of our study design help to deflect some of most trenchant theoretical attacks. Specifically, we search for evidence of specific deterrence within a time frame that is short—short enough to discount rival explanations for any behavioral changes observed.[17] In addition, we observe which Queensland drivers had their licenses lapse or become suspended or revoked, when, and for how long. "Censoring" those periods from the analysis helps to separate true deterrent effects from incapacitation effects (i.e. lack of exposure to the risks under study).

[16]Stafford M, Warr M. A reconceptualization of general and specific deterrence. Journal of Research in Crime and Delinquency. 1993;30:123-35; Piquero A, Paternoster R. An application of Stafford and Warr's reconceptualization of deterrence to drinking and driving. J Res Crime Delinq. 1998;35(1):3-39.

[17]The tradeoff is that we are not positioned to draw inferences about how enduring specific deterrent effects are.

4.3 What is known about the deterrent effects of traffic laws?

Over the last 40 years dozens of studies have sought to measure deterrent effects associated with traffic laws. The body of research converges tightly around two outcomes: recidivism and crashes. In other respects, however, it is quite heterogeneous.

4.3.1 Drunk driving studies

Perhaps the most striking feature of the traffic law deterrence literature is that the overwhelming majority of studies focus on drunk driving. An obvious explanation is the important causal role of alcohol in crashes.[18] Other factors are also at work. The sharp rise in the prevalence and severity of drunk driving laws since the 1970s, coupled with the availability of high-quality population-level data on offenders and road accidents, have created abundant opportunities for research. Further, the large number of studies examining recidivist drunk driving almost certainly reflects the leadership of criminologists in this area.

Drunk driving studies are included in the literature review of general and specific deterrence that follows—without them the review would be short and the evidence base remarkably thin. There are good reasons, however, to be cautious about extrapolating from findings in those studies to surmise the nature of deterrent effects of traffic laws more broadly. First, drunk drivers account for a small minority of penalized offenders. In Queensland, for example, over the 16-year period we examined, less than 2% of the 11.6 million recorded offenses were DUIs. Second, driving under the influence of alcohol or drugs sits at the egregious end of the traffic offense spectrum. It is commonly treated as a crime, whereas most other types of traffic offenses are civil or administrative in nature, and are sanctioned by fines and license demerit points.[19] Third, owing to their criminal nature, drunk driving offenses often trigger penalties such as license suspension, vehicle impoundment, and, for repeat offenders,

[18] In the US in 2012, for example, 10,322 deaths, or 31 percent of all road fatalities, occurred in alcohol-impaired-driving crashes. These figures have declined steeply over the last 30 years. In 1982, 57% of the 43,945 traffic fatalities were alcohol-related. See National Highway Transportation Safety Administration, Traffic Safety Facts: Alcohol-impaired driving DOT HS 81 1 870. December 2013.

[19] While it may be tempting to infer from that distinction that drunk driving laws should therefore set the high-water mark for deterrent effects, that conclusion ignores research suggesting that: (1) severity of punishment is a poor predictor of safer driving; and (2) drunk driving offenders, especially recidivist drunk drivers, tend to be an atypical kind of traffic offender. See Nochajski TH, Stasiewicz PR. Relapse to driving under the influence (DUI): a review. Clin Psychol Rev 2006;26(2):179-95.

incarceration.[20] Such incapacitating interventions have been associated with some of the most impressive deterrent effects detected in the literature.[21] However, as noted above, these studies generally do not disentangle safer driving responses from drivers' reduced or complete lack of exposure to driving during the penalty period, and the latter is not a species of deterrence. Finally, the prevalence of alcohol addiction among drunk drivers may mute their susceptibility to deterrent effects, although this theory is controversial.[22]

4.3.2 General deterrence

Nearly all studies of general deterrence from traffic laws employ the same design: they are before-and-after comparisons of accident rates and/or recidivism, centered on the introduction of new penalties or enhancements of existing ones. Two laws that have consistently demonstrated large safety effects are the lowering of permissible levels of blood-alcohol concentration (BAC) [23] and pre-conviction license suspensions for drunk drivers.[24] Isolating general deterrent effects in ecological pre/post studies is challenging. The difficulty is compounded in the case of drunk driving laws by the

[20]Under so-called "administrative per se" laws, around 30 states now suspend licenses immediately, at the point of test failure. See Wagenaar et al, infra note 24.

[21]See, for e.g., Zaldor PL, Lunk AK, Fields M, Weinberg K. Fatal crash involvement and laws against alcohol-impaired driving. Journal of Public Health Policy 1989;10:467-485; Voas RB Tippetts AS, Fell JC. The relationship of alcohol safety laws to drinking drivers in fatal crashes. Accident Analysis and Prevention 2000;32;483-92.

[22]Nochajski, Miller, Parks. Comparison of first time and repeat DWI offenders Alcoholism: Clinical and Experimental Research 1994;18:48; Wiczorek and Nochaski. Characteristics of persisent drinking drivers; Yu and Williford 1993, Problem drinking and hi-risk driving ; Yu J, Chin Evans P, Perfetti Clark L. Alcohol addiction and perceived sanction risks: Deterring drinking drivers. J Crim Justice. 2006;34:165-174.

[23]Wagenaar AC, Maldonado-Molina MM, Ma L, Tobler AL, Komro KA. Effects of legal BAC limits on fatal crash involvement: analyses of 28 States from 1976 through 2002. J Safety Res. 2007;38:493-499; Fell JC, Voas RB. The effectiveness of reducing illegal blood alcohol concentration (BAC) limits for driving: evidence for lowering the limit to .05 BAC. J Saf Res 2006;37:233-243; Whetten-Goldstein et al. Civil liability, criminal law, and other policies and alcohol-related motor vehicle fatalities in the United States: 1984-1995. Accid Anal Prev 2000;32:723-33; Williams AF, Zador PL, Harris SS, Karpf RS. The effect of raising the legal minimum drinking age on involvement in fatal crashes. J Legal Stud. 1983 Jan;12(1):169-179; Deshapriya EB, Iwase N. Impact of the 1970 Legal BAC 0.05 mg% limit legislation on drunk-driver-involved traffic fatalities, accidents, and DWI in Japan. Subst Use Misuse. 1998;33(14):2757-2788

[24]Wagenaar AC, Maldonado-Molina MM. Effects of drivers' license suspension policies on alcohol-related crash involvement: long-term follow-up in forty-six states. Alcohol Clin Exp Res. 2007 Aug.;31(8):1399-1406; Rogers PN. The general deterrent impact of California's 0.08% blood alcohol concentration limit and administrative per se license suspension laws. An evaluation of the effectiveness of California's 0.08% blood alcohol concentration and administrative per se license

incapacitating nature of the penalties involved which, as note above, can cloud the true size and nature of the deterrent effect.

BAC and pre-conviction license suspension laws aside, the evidence for general deterrence from drunk driving laws is variable. Evans et al[25] found no reduction in accident risk in the US from escalations in the drunk driving penalties, nor did Briscoe[26] in Australia. Wagenaar et al[27] identified a modest negative association between mandatory minimum fines for drunk driving and fatal crash rates, but the effects were not consistent across the 32 US states examined; this study also found no strong deterrent effects from mandatory minimum jail policies. On the whole, international reviews of general deterrence have concluded that traffic laws that promise increased certainty of punishment lead to temporary reductions in alcohol-related fatalities, whereas laws aimed at increased severity are ineffective.[28]

Outside the drunk driving context, there is limited evidence on the general deterrent effects of traffic laws. What published studies exist are mostly positive. For example, Bar-Ilan and Sacerdote[29] found that red-light running in San Francisco and Israel decreased in response to an increase in the applicable fine. In Portugal, Tavares et al found that fine increases and the introduction of an "on-the-spot" fine payment policy were associated with decreases in both accident and injury rates.[30]

---

suspension laws, Volume 1. Sacramento, California: California Department of Motor Vehicles, Research and Development Section. CAL-DMV-RSS-95-158; 1995; McArthur DL, Kraus JF. The specific deterrence of administrative per se laws in reducing drunk driving recidivism. Am J Prev Med. 1999;16(1):68-75; Klein TM. Changes in alcohol-involved fatal crashes associated with tougher state alcohol legislation. Washington, DC: US Department of Transportation, National Highway Safety Administration. DOT HS 807 511; 1989; Zador P, Lund AK, Fields M, Weinberg K. Fatal crash involvement and laws against alcohol-impaired driving. Washington, DC: Insurance Institute for Highway Safety; 1988.

[25]Evans WN, Neville D, Graham JD. General deterrence of drunk driving: evaluation of recent American policies. Risk Anal. 1991;11(2):279-289

[26]Briscoe S. Raising the bar: can increased statutory penalties deter drink-drivers? Accid Anal Prev. 2004;36(5)919-929.

[27]Wagenaar AC, Maldonado-Molina MM, Erickson DJ, Ma L, Tobler AL, Komro KA. General deterrence effects of U.S. statutory DUI fine and jail penalties: long-term follow-up in 32 states. Accid Anal Prev. 2007;39:982-994.

[28]Ross HL. Deterring the drinking driver: legal policies and social control. Lexington, MA: Lexington Books, 1980; Homel R. Policing and Punishing the Drinking Driver. A Study of General and Specific Deterrence. New York: Springer, 1988; Nagin DS, Pogarsky G. Beyond Stafford and Warr's reconceptualization of deterrence: personal and vicarious experiences, impulsivity, and offending behaviors. J Res Crime Delinq 2001;39:153-186.

[29]Bar-Ilan A, Sacerdote B. The response of criminals and noncriminals to fines. J Law Econ. 2004 April;47(1):1-17.

[30]Tavares AF, Mendes SM, Costa CS. The impact of deterrence policies on reckless driving: the case of Portugal. Eur J Crim Policy Res. 2008;14:417-429.

The introduction of a penalty points system in Italy in 2003 was associated with reductions in both crashes and fatalities there.[31] And Canadian laws aimed at stopping street racing and stunt driving have been linked to a small but significant reduction in speeding-related casualties among male drivers.[32]

4.3.3 Specific deterrence

Drunk driving studies also dominate the specific deterrence literature. The standard approach here is to compare the effects of different forms and levels of punishment on recidivism.[33] The evidence is somewhat mixed. A few studies have detected significant specific deterrent effects,[34] but most have found no effects, very small effects, or effects only in discrete subpopulations (e.g. first-time offenders).[35]

Two studies are noteworthy for extending specific deterrence investigations beyond the drunk driving context. Both reported evidence of specific deterrence. Li et al[36] examined a cohort of nearly 30,000 Maryland drivers who were ticketed for speeding. The researchers found lower risks of subsequent speeding citations but higher risks of crashes among drivers who elected to appear in traffic court, compared with drivers who chose to simply mail in payment of their fines. Among court-goers,

[31]De Paola M, Scoppa V, Falcone M. The deterrent effects of the penalty point system for driving offenses: a regression discontinuity approach. Empir Econ. 2013;45:965-985.

[32]Meirambayeva A, Vingilis E, McLeod AI, Elzohairy Y, Xiao J, Zou G, Lai Y. Road safety impact of Ontario street racing and stunt driving law. Accid Anal Prev. 2014;71:72-81.

[33]Salzberg PM, Paulsrude SP. An evaluation of Washington's driving while intoxicated law: Effect on drunk driving recidivism. J Safety Res. 1984;15(3):117–124; Yu J. Punishment celerity and severity: testing a specific deterrence model on drunk driving recidivism. J Crim Justice. 1994;22(4):355-366; Taxman FS, Piquero A. On preventing drunk driving recidivism: an examination of rehabilitation and punishment approaches. J Crim Justice. 1998;26(2):129-143; McArthur DL, Kraus JF. The specific deterrence of administrative per se laws in reducing drunk driving recidivism. Am J Prev Med. 1999;16(1):68-75; Briscoe S, New South Wales, Bureau of Crime Statistics and Research. The impact of increased drink-driving penalties on recidivism rates in NSW. Bureau of Crime Statistics and Research. 2004;5:11; Weatherburn D, Moffatt S. The specific deterrent effect of higher fines on drunk-driving offenders. Br J Criminol. 2011;51(5):789-803; Ahlin EM, Zador PL, Rauch WJ, Howard JM, Duncan GD. First-time DWI offenders are at risk of recidivating regardless of sanctions imposed. J Crim Justice. 2011;39(2):137-142.

[34]Yu et al 1994; McArthur et al 1999

[35]Taxman and Piquero 1998; Salzberg and Paulsrude 1984; Ahlin et al 2011; Weatherburn and Moffatt 2011; Briscoe 2004.

[36]Li J, Amr S, Braver ER, Langenberg P, Zhan M, Smith GS, Dischinger PC. Are current law enforcement strategies associated with a lower risk of repeat speeding citations and crash involvement? A longitudinal study of speeding Maryland drivers. Ann Epidemiol. 2011;21(9):641-647.

those whose case was not prosecuted or suspended had significantly lower rates of subsequent crashing and reoffending than drivers with other case outcomes.[37]

Redelmeier et al[38] studied a sample of drivers in Ontario, Canada, who were convicted of a wide range of traffic offenses. The drivers' risks of having fatal crashes in the month after a conviction were about 35% lower than in another comparable period; 2 months after the conviction this "benefit" had dwindled, and by 3-4 months it was no longer significantly different from the drivers' baseline risks. These results suggested a short-run specific deterrent effect.

4.4 The causal inference challenge

4.4.1 Known unknowns and unknown unknowns

The causal relationship between traffic laws and road safety is nuanced and challenging to isolate. Several inter-related factors conspire to complicate causal inference. One is that penalties do not occur in isolation; they are one of a host of variables that influence a driver's risk of crashing. Another complication is the well-established association (as distinct from causal relationship) between offenses and accidents: numerous studies have shown that drivers at high risk of incurring traffic citations are also at relatively high risk of crashing.[39]

A more general way of describing these causal inference problems is to say that differences *between* drivers that influence both their risks of offending and their risks of crashing—and thus modulate the effects of penalties on driving behavior—cannot be fully observed and adjusted for, at least not in large population-level studies. Such "confounders" undercut researchers' ability to make strong causal claims about the effect of penalties on rates of accidents and recidivism.

Two of the most important between-person differences that usually cannot be observed in population-based studies are driving "exposure" (how much and when a driver is on the road) and driving performance (how safely a driver drives relative to others). A simple example helps to illustrate the problem. Imagine a driver

[37]The investigators compared drivers in four outcome groups: (1) not guilty; (2) suspended/no prosecution; (3) probation before judgment and fines; and (4) fines and demerit points.

[38]Redelmeier DA, Tibshirani RJ, Evans L. Traffic-law enforcement and risk of death from motor-vehicle accidents: case-crossover study. Lancet 2003; 361: 2177–2182.

[39]See, for e.g., Gebers MA. An inventory of California driver accident risk factors. Technical report; California Department of Motor Vehicles: 2003; Blows S, Ameratunga S, Ivers RQ, Lo SK, Norton R. Risky driving habits and motor vehicle driver injury. Accident Analysis and Prevention 2005; 37: 619–624; DeYoung DJ, Gebers MA. An examination of the characteristics and traffic risks of drivers suspended/revoked for different reasons. J Safety Res. 2004;35(3):287-295.

who commutes to work and averages 100 kilometers of driving per day; she incurs two citations per year on average and crashes twice over a 10-year period. Another driver, who works at home and drives 10 kilometers per day, averages one citation per year and is involved in one crash over a decade. A comparison of the 10-year track records of these two drivers, based only on information about their offense and crash histories, would draw the specious conclusion that citations increase crash risk.[40]

A range of other unobserved differences besides driving exposure have the potential to bias estimates of the causal effects of traffic law penalties. In interpreting results from their study of specific deterrence among Maryland speeders, for example, Li et al concede that "[t]he increased risk of crashes associated with court appearances likely reflects the high-risk characteristics of drivers who chose this approach rather than being a true causal relationship." In sum, problems of unobserved heterogeneity bedevil large-scale studies that rely on unadjusted or under-adjusted comparisons of different classes of offenders[41]—which is to say all but a couple of the specific deterrence studies conducted to date.

4.4.2 Attempts at stronger causal inference

Two studies of the safety effects of traffic laws were carefully designed to try to combat potential confounding bias. The studies reached opposite conclusions about whether penalties specifically deterred, although they were focused on different types of offenses.

Weatherburn and Moffat's analysis of the specific deterrent effects of high fines on recidivist drunk driving exploited a quirk in the management of these cases in the New South Wales court system. The cases were assigned randomly within a panel of magistrates, yet court records showed wide variation in the severity of the penalties the magistrates imposed on offenders. The researchers took advantage of this variation to measure the effect of different penalty levels on recidivism. They detected no evidence of specific deterrence. Offenders who received more severe penalties did not have lower rates of recidivism.

In the Ontario study discussed above, Redelmeier, Tibshirani, and Evans made creative use of a case-crossover design to estimate the effect of convictions for a range of different traffic offenses on crash risk. In the case-crossover design, each case serves as its own control. All drivers in the study sample were involved in a fatal crash.

[40] A number of the studies of the deterrent effects of traffic laws make an inferential leap not unlike this.

[41] Nagin D, Cullen FT, Jonson CL. Imprisonment and reoffending. In Tonry M (ed.), Crime and Justice: A review of Research, 115-200. Chicago: Univ Chic Press 2009; Weatherburn and Moffat 2011.

The researchers compared each driver's probability of incurring a penalty in the month before the crash with the probability that driver incurred a penalty in the same month one year earlier (when they did not crash). A lower penalty risk in the pre-crash period was interpreted as evidence of specific deterrence. The researchers found such an effect, albeit a transient one.

Although these two studies had very different designs, focused on different offenses or measures of safety, and reached opposite conclusions regarding specific deterrence, their methodologies were shaped by a common goal: to counteract the threat of biases from unobserved between-driver differences. We shared this goal, and pursued it through a novel study design.

4.4.3 Overview of study approach

We exploited a "wrinkle in time" created by the way drivers are notified of certain offenses in Queensland. We followed a large cohort of drivers who were caught speeding or red-light running by traffic cameras. The drivers did not learn of their offense and the penalty to be imposed until 2-3 weeks after the offense occurred. We compared crash and offending rates in the periods immediately before and after the drivers received notification.

This approach addresses the potentially pernicious effects of unobserved heterogeneity in two main ways. First, the comparison is between crash risk for a cohort of drivers immediately before and immediately after notification. The pre and post groups are homogeneous because they contain the same drivers (barring small losses for drivers who are censored because they crash or reoffend). Second, because the date of notification is determined by the regulator's internal processes and the variability of the postal service, both of which are beyond the influence of drivers, it is difficult to imagine any connection, unrelated to specific deterrence, yielding an instantaneous change in drivers' risk profile precisely at notification.[42]

We turn to now to describe the methods in more detail.

[42] As well as tackling unobserved between-driver differences, the short time-frame of our regression discontinuity design also addresses within-driver differences that may emerge over time, provided any such differences do not arise instantaneously at notification.

4.5 Study approach

4.5.1 Setting

With 4.8 million residents, Queensland is the third most populous state in Australia.[43] Its regime for driver licensing and regulation is broadly similar to Australia's other states and territories, and to regimes in many other developed countries, including those currently in force in US states. [44] Queensland operates a graduated licensing scheme. Residents 17 years or older may apply for a learner license. Learners who log 100 hours of supervised driving, are at least 18 years of age, and pass both a written road rules test and a practical driving test are issued a provisional license. Provisional license holders become eligible for an "open" general license after one to three years.

4.5.2 Offenses and penalties

In Queensland's penalty scheme, each traffic offense triggers a fine and carries a specified number of demerit points. Fine amounts vary widely, from around one hundred dollars for minor infractions to several thousand for the most serious ones. In 2014, the lowest speeding fine for first time offenders was $ AU151 and the highest was $ AU1,062.[45] The demerit points assigned to each offense are codified by statute, and range from 1 point for minor transgressions (e.g. failure to dip high-beam headlights for oncoming traffic, a small defect that renders the vehicle unroadworthy but not necessarily unsafe) to 8 points for exceeding the speed limit by more than 40 kilometres per hour. One-point and three-point offenses are by far the most common: they account for approximately 30% and 45% , respectively, of all citations issued.

The Queensland Department of Transport and Main Roads (DTMR) keeps a running tally of cumulative demerit points against every licensed driver in the state. Demerit points are removed three years after the offense, and fully reset after a period of license suspension or good driving behaviour.[46] Drivers who accumulate 12 or more demerit points in a three-year period typically face a period of license

[43] Australian Bureau of Statistics. Australian Demographic Statistics. Canberra: 2013.

[44] Transport Operations (Road Use Management—Driver Licensing) Regulation 2010 (Queensland, Australia).

[45] The Australian dollar is roughly equivalent in value to the US dollar.

[46] When Queensland drivers reach 12 demerit points, they are given a choice of accepting a suspension for a relatively short period (usually 3 months) or continuing to drive under threat of heavier suspension (usually 6 months) if one further offense occurs during a defined period of good driving behaviour (usually 12 months). Our data allowed us to observe these choices and drivers were excluded from our analysis for any periods in which their license was suspended.

suspension of 3-6 months. Drunk driving offenses occupy a category of their own; they result in fines but not demerit points as such, because they almost always trigger a license suspension that is applied independently of the demerit point scheme.

Offenses are detected through a combination of direct observation by police and fixed and mobile traffic enforcement cameras.[47] Cameras are used in the detection of only two types of offenses: speeding and red-light violations.[48] For directly-observed offenses, police issue tickets at the roadside.[49] Camera-detected offenses are notified by mail. Over the time period of our study in Queensland, about half of all speeding and red-light offenses were detected by direct observation and half were detected by camera.

4.5.3 Data and variables

DTMR routinely collects details of both traffic offenses and crashes. Accurate tracking of offenses and penalties is essential for the operation of the state's demerit point system. All crashes that cause death, injury or substantial property damage are recorded, provided they are reported to the Queensland Police Service.[50]

DTMR provided us with de-identified offense and crash data spanning the period 2 November 1996 to 31 December 2010. It also provided de-identified license histories for all drivers in Queensland over the same period; for each driver, this included dates when the driver was licensed and, if applicable, dates when the license was suspended or disqualified.

Using de-identified numbers unique to each licensee, we linked the infringement, crash and license history data to create the study dataset. The dataset included variables describing drivers (age, sex), crashes (severity, fault), and offenses (type, number of demerit points). Our offense typology was based on categories set forth in the Australian and New Zealand Offense Classification.[51] We also constructed measures of the cumulative number of demerit points each licensed driver had at successive points in time.

[47] State Penalties and Enforcement Act 1999 (Queensland, Australia).

[48] Transport Operations (Road Use Management) Act 1995 (Queensland, Australia), s 13, 158.

[49] Colloquially, these are referred to as "on-the-spot" tickets.

[50] The "substantial" property damage threshold is met if at least one vehicle is towed away, the cost of damage to all property exceeds $2,500 (before December 1999), or the cost of damage to property other than vehicles exceeds $2,500 (from December 1999). See Queensland Department of Transport and Main Roads. Data Analysis: Road crash glossary. Available at https://www.webcrash.transport.qld.gov.au/webcrash2/external/daupage/docs/glossary.pdf (accessed 10 Sept 2014).

[51] Australian Bureau of Statistics. Australian and New Zealand Offense Classification 2011 (3rd edition).

DTMR uses five mutually-exclusive categories to describe crash severity: (1) fatality; (2) injury requiring hospitalization; (3) injury requiring medical treatment but not hospitalization; (4) injury not requiring medical treatment; and (5) property damage only. We collapsed these into a binary variable indicating "serious or fatal injury" (first three categories) and "minor or no injury" (last two categories). Of course, many crashes involve multiple injuries and property damage. The variable we used pertains to the most serious outcome in each crash.

Determinations of fault for each crash, including single vehicle collisions, are made by DTMR on the basis of the police report. The "at fault" designation is applied to the person judged to be most at fault, and to any persons issued with traffic citations in connection with the crash.

4.5.4 Study design

We used a regression discontinuity design to compare the two outcome variables of interest—crash rates and recidivism rates—across two time periods.[52] This is a quasi-experimental design that permits causal inference in wide variety of situations. The design requires that experimental units are assigned to treatment on the basis of threshold defined by a covariate that takes values on a continuum. If the only difference between experimental units immediately on either side of the threshold is the fact of the treatment, then any discontinuity arising at the threshold must be linked causally to treatment.[53] In our analysis, the continuous measure is time and the threshold is defined by the moment at which offending drivers were notified that their violation had been caught on camera and that penalties were being imposed.[54]

[52]On one view, it is somewhat unorthodox to conceive of time as the continuous measure for assortment in a regression discontinuity design. However, previous studies with designs in the regression discontinuity family have used attainment of a certain age to "switch on" treatment status (see studies reviewed in David S. Lee and Thomas Lemieux, Regression discontinuity designs in economics. NBER Working Paper No. 14723 (February 2009). It is also true that our approach also has many of the basic features of an interrupted time series design (see Penfold RB, Zhang F. Use of interrupted time series in evaluating health care quality improvements. Acta Pediatr. 2013;13(6 Suppl):S38-44). We believe these are largely distinctions in labels, which do not have a substantive bearing on the nature or appropriateness of the design we implemented.

[53]Thistlethwaite DA, Campbell DT. Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology 1960;51(6):309-317; Imbens G, Lemieux, T. Regression discontinuity designs: a guide to practice. Journal of Econometrics 2008;142:615–635

[54]If notification and infringement occur on the same day then causal inference from a regression discontinuity design is not possible because drivers who experience crashes are much less likely to go on to commit offenses (because, for example, their car is off the road being repaired), so any discontinuity must be interpreted as the sum of the extent to which crashes prevent subsequent offenses and the deterrent of effect of the penalty associated with the offense.

Specific deterrence theory suggests that drivers' risks of both offenses and crashes should decrease from that moment forward.

Construction of the two time periods requires further explanation. The first time period consisted of an interval running from the day a driver committed a camera-detected offense to a day near the moment when the driver received notification of the offense and the applicable penalty. We refer to this interval as the "pre-notification period" . We set an upper limit of 21 days on the pre-notification period. The second time period, the "post-notification period" , ran for 90 days following notification.

One complication with construction of the discontinuity in our study is that we did not know the exact date drivers became aware of their offense. Our data included a variable indicating the date the offense was registered by DTMR; as a matter of routine practice, this date is one business day before the DTMR generates the notification letter.[55] The letter is sent by regular post, which typically takes 1-2 days, but the letter may not be opened and read by the offender immediately. Further, although licensees are obligated by law to notify DTMR within 14 days of any change of name or address, not all do, and some letters are sent to the wrong address or wrong person.

To address this fuzziness around the day of the offender's first awareness, we created a "notification week." Specifically, for each offense we calculated a best-estimate date, defined as the date three business days (i.e. exclusive of weekends and public holidays) after the date DTMR registered the offense. The notification week was calculated by counting three days forward and three days back from the best-estimate day. Thus, the pre-notification period actually ended on the fourth day before the best-estimate day and the post-notification period began on the fourth day after the best-estimate day. Some misclassification across time periods is inevitable, and its effect would be to bias any true differences between the periods to the null. However, we believe if it safe to assume that the vast majority of drivers in our sample would have become aware of their offense and penalty during the notification week.

4.5.5 Study sample

Drivers who committed camera-detected offenses entered the study sample, provided they were at least 20 years of age[56] and their license remained active throughout the

[55]Telephone conversation of July 23, 2014 with Dr. Nerida Leal, Principal Behavioural Scientist, Queensland Department of Transport and Main Roads.

[56]Drivers with learner and provisional licenses must comply with special rules that do not apply to other drivers. For example, learners must drive under supervision; provisional license holders have certain restrictions on carrying passengers, driving high-powered vehicles, and driving at night, and

study period. We observed this "cohort" for 112 days—21 days pre-notification, the day of notification, plus 90 days post-notification.[57] However, the cohort's membership was not completely fixed; it changed in two ways over the observation window.

First, we allowed for different lag periods between the offense and notification. Any driver whose lag period was 22 days or more entered on the first day of the pre-notification period; the rest joined on the day after their index offense.

Second, drivers who crashed or re-offended were censored from the cohort immediately after counting those events in the daily rate. There was a strong reason to do this for drivers who crashed: the crash reduced the likelihood they would continue driving for some or all of the remaining observation period, so retaining them would be likely to bias downwards the crash and offense rates calculated for subsequent days.[58] Re-offenders were censored because their response to any subsequent offense may have overlapped in time with their response to the index offense, thus blurring the deterrent effect under study. A countervailing consideration in the treatment of re-offenders is that, because offenses are such frequent events, our censoring rule served to eliminate nearly 15% of the cohort by the end of the post-notification period. Such attrition should not have affected the discontinuity effects, but it may have introduced some more general biases.[59]

4.5.6 Statistical analysis

The main goal of the statistical analysis was to estimate the size and significance of the two discontinuities of interest—differences in crash risk and recidivism, respectively—between the pre- and post-notification periods. We fit a generalized additive regression model from the Poisson family with an offset to permit the estimation of rates rather than counts. Time trends in the crash and recidivism rates were modelled

---

the demerit point threshold for license disqualification is lower than for drivers with open licenses. Hence, to avoid complications in the analysis and in interpretation of findings, we sought to exclude learners and provisional license holders from the study sample. The difficulty was that we did not have information that allowed us to identify these drivers directly. Therefore, we used age as a proxy, excluding all drivers under 20 years of age. This undoubtedly produced errors: Queenslanders as young as 19 obtain unrestricted open licenses, and some learners and provisional drivers are over 20 years of age. However, the overwhelming majority of learner and provisional drivers are under 20, and most drivers under 20 fall into one of these two license categories, so the age-based exclusion rule was a reasonable work-around.

[57] Hereafter, when it is necessary to distinguish the camera-detected speeding or red light offense that brought drivers into the sample from subsequent offenses drivers committed, we refer to the former as the "index" or "notified" offense.

[58] On the other hand, the fact that so few drivers crashed, relative to the size of the baseline population, means this bias would probably not be noticeable.

[59] We describe this issue in more detail in the Discussion section.

using thin plate splines. Other covariates were added to the model as parametric terms.[60]

In the primary crash model, the outcome variable was crashes per 100,000 drivers per day; in the primary recidivism model, the outcome variable was offenses per 100,000 drivers per day. The covariates specified in these two models were essentially the same. The predictor of interest was a dummy variable distinguishing observations in the pre-notification period from observations in the post-notification period. The thin plate splines controlled for time trends within the pre- and post-notification periods. The models also included dummy variables marking the number of days from notification modulo 7 to account for a day-of-the-week effect.[61] In addition, we adjusted for several covariates known to have an independent association with the outcomes of interest: driver age, offense type (red light, minor speeding, moderate speeding, major speeding), and cumulative demerit points. Inclusion of these baseline risk factors strengthened our ability to interpret the magnitude of the discontinuities as average effects at the population level.

Results of the primary crash and recidivism models are presented in two ways. We drew adjusted trend lines on either side of the discontinuity, and superimposed them on scatter plots of the raw daily rates. We also present coefficients and 95% confidence intervals from the models in tabular form.

The second step in our analysis was to determine whether there were subgroup differences in the size of the discontinuity effects estimated in the primary models. To do this we conducted a series of stratified analyses. The models had identical specifications to the primary models, but were run within defined subgroups of drivers.

For crash risk, we analysed the magnitude of the discontinuity within the following subgroups: (1) drivers for whom the notified offense brought their cumulative demerit point total near to the point of license disqualification (9-11 points) *versus* drivers who remained at a low point count after the notification (1-5 points); (2) crashes in which the driver was judged to be at fault *versus* not at fault; and (3) crashes that resulted in serious injury *versus* minor injury.

For recidivism risk, we estimated discontinuity effects within the following strata: (1) the same high-*versus*-low cumulative demerit point totals as described above for the crash risk analyses; (2) subsequent offenses detected by camera *versus* direct

[60]Wood SN. Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society (B) 2000;62(2):413-428; Hastie TJ, Tibshirani RJ. Generalized additive models. Chapman & Hall/CRC, 1990.

[61]We could not include day of the week directly because notification may have been made on any non-holiday weekday. Thus, a day that was 8 days from notification would assume the same value in this day-of-the-week variable as would the day that was one day from notification.

observation; (3) high-risk *versus* low-risk offenses;[62] and (4) concordance between the type of index offense and the type of subsequent offense *versus* discordance in offense types.

All analyses were conducted in R (version 3.1.1).[63] Only a small fraction of the data was missing (<0.25% for all variables analysed).

4.5.7 Ethics

The Human Research Ethics Committee at the University of Melbourne approved the study.

4.6 Results

4.6.1 Sample characteristics

Table 1 profiles characteristics of the drivers (n=2,880,763) in the study sample, together with the type of camera-detected offense that brought them into the sample. Sixty percent of the drivers were male and 64% were aged between 31 and 60 years. Ninety-one percent were caught speeding; the rest ran red lights. Two thirds of drivers had acquired fewer than six demerit points, inclusive of the points associated with the notified offense; at the other end of the spectrum, the notified offense brought 14% of drivers up to nine or more demerit points.

Table 2 shows the severity of crashes (n=15,317) that occurred during the study period. One percent caused at least one fatality, 22% caused injuries serious enough to require hospitalization, and 27% caused injuries that were treated outside hospital. Drivers in our sample were judged to be at fault in 62% of the crashes in which they were involved.

Table 3 describes the offenses (n=184,544) drivers committed during the 112-day observation period, following their index camera-detected offense.[64] Seventy-one

[62]These two strata were created by separating offenses clearly indicative of risk driving behaviour (e.g. dangerous/careless driving, speeding, red-light running, line crossing, drunk driving, failure to wear a seatbelt or helmet, failure to give way, etc) from offenses not clearly related to risky driving (e.g. administrative non-compliance, unsafe carriage of goods, defective vehicle, public nuisance etc.).

[63]R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[64]These figures do not represent a full accounting of the reoffending by drivers in our cohort during the 112-day period. Because we treated re-offending as a censoring event, we did not analyze and do not report in Table 3 any further offenses (third, fourth, fifth, etc.) by drivers who committed more than two offenses.

percent of the re-offending involved speeding. The next most prevalent types of offenses were red light violations (4%), driving while uninsured or unregistered (3%), and using a mobile phone while driving (2%). Fifty-three percent of the offenses were detected by camera and 47% by direct police observation.

4.6.2 Effects of notification on crashes

Notification did not lead to a significant change in drivers' risks of crashing (Rate Ratio [RR], 0.94; 95% Confidence Interval [CI], 0.86-1.02) (**Figure 2**).

Table 4 shows the full set of estimates from the multivariate regression model. Drivers' age, gender, and cumulative demerit point total were all significant predictors of crash risk, as was the type of camera-detected offense committed, but notification was not.

This result was robust across all subgroups examined (**Figure 3**). Drivers for whom the notified offense took their cumulative demerit point tally into the range of 9-11 points, the level at which one further offense would likely result in license disqualification, did not exhibit a significant change in crash risk following notification (RR, 1.31; 95% CI, 0.91-1.87). Nor did we detect a significant change in crash risk when the analysis was restricted to drivers at fault in the crash (RR=0.93; 95% CI, 0.87-1.01), or to crashes that caused fatal or serious injury (RR, 0.97; 95% CI, 0.85-1.10).

4.6.3 Effects of notification on recidivism

The rate at which drivers committed offenses decreased by 25% immediately after notification (RR, 0.75; 95% CI, 0.73-0.78) (**Figure 4**), and remained relatively low for the remainder of the post-notification period.

Table 5 shows the full regression results. The baseline predictors for recidivism ran in the same direction and were of a similar magnitude as in the crash model, except for offense type. Speeders reoffended at a higher rate than red light runners, but they crashed at a lower rate.

The decrease in recidivism following notification varied within several of the subgroups of drivers examined (**Figure 5**). The decrease in rates of offenses clearly indicative of risky driving (RR, 0.70; 95% CI, 0.69–0.72) was larger than the decrease observed in rates of low-risk offenses (RR, 0.78; 95% CI, 0.73–0.83).[65] There was a larger decrease in camera-detected offenses following notification (RR, 0.68;

[65]For a description of which offenses went into which categories, and the basis for the classifications, see supra note 62.

95% CI, 0.66-0.70) than there was in offenses detected by direct police observation (RR, 0.75; 95% CI, 0.73-0.77).

Drivers whose index offense was a camera-detected red light violation had rates of red light violation after notification that were 41% lower (RR=0.59, 95% CI, 0.49-0.72). Drivers whose notified offense was speeding had rates of speeding re-offending that were 31% lower (RR=0.69, 95% CI, 0.67-0.71). By contrast, rates of reoffending by offenses of a different type to the notified offense were "only" 23% lower after notification (RR=0.77, 95% CI, 0.74-0.80).

On the other hand, the magnitude of the decline in recidivism was insensitive to drivers' demerit point tallies. Drivers whose notified offense took them into the 9-11 point range decreased their offending rates (RR=0.69, 95% CI, 0.65-0.73) by about the same amount as drivers whose notified offense did not take their cumulative demerit point above 5 points (RR=0.71, 95% CI, 0.69-0.73).

4.7 Discussion

This study followed a cohort of drivers who had broken road rules and been caught. We followed them from the time the committed a speeding or red-light offense, through the time they were informed that they had been caught and would be penalized, and then for three months afterwards. We found that notification did not reduce the likelihood of subsequent crashes, even among drivers for whom another offense probably spelled license suspension.

The notification did, however, result in a substantial reduction in drivers' risks of committing additional offenses. Subgroup analyses shed further light on this effect. There was an especially large reduction in offenses of the same type as the notified offense, suggesting a kind of specific-specific deterrence at work. Similarly, notification reduced the incidence of offenses indicative of risky driving choices more than it did the incidence of offenses that were more administrative or technical in nature. In other words, behaviours that were relatively dangerous and which ostensibly fell more directly under the drivers' control decreased by a larger amount. This is a plausible result within a deterrence framework.

Set alongside each other, the effects we observed on crash risk and recidivism raise two obvious questions. First, if deterrence theory rests on the assumption that curbing offending prevents accidents, how does one happen without the other? Second, should one infer from our findings that traffic laws in Queensland are producing an effective form of specific deterrence? We consider each of these questions in turn.

4.7.1 The bifurcation of specific deterrence

Pure avoidance behaviour

One way in which recidivism could decline without moving the needle on crash rates is through avoidance behaviour. Notification may have prompted drivers to alter their driving behaviour in ways that substantially reduced their risk of incurring additional penalties but which did not materially affect their crash risk. Imagine a driver who is motivated to steer clear of intersections he knows have cameras, or avoid stretches of road he knows are a common speed trap. These moves could be made without necessarily changing the care with which he drove.

One of the subgroup analyses provides some indirect support for an avoidance explanation. The rate of camera-detected offenses dropped more sharply after notification than did the rate of offenses ticketed at the roadside.[66] Mobile enforcement by police has a stochastic dimension that probably makes it more difficult than cameras to thwart, in the absence of authentic changes to the safety with which one drives. In sum, behavioural responses focused on pure enforcement evasion could drive a wedge between risks of recidivism and risks of crashing.

Relationship between offending and crash risk

A more fundamental and damning explanation for the bifurcation is that offending behaviour—or more precisely, offending behaviour that law enforcement catches—has a weak connection to crash risk at the population level. Regulators and safety experts like to focus on the strength of the connection, emphasizing, for example, that the behaviours traffic law regimes sanction are associated with crash risk, and that multi-offenders have higher crash risks than occasional or never offenders.[67] But widening the frame, the reality is that the strength of the relationship is diluted at several critical nodes. Cameras and police capture only a small fraction of offenders. Crashes are rare events (much rarer than penalties.) And many crashes are not attributable to unlawful behaviours, although a sizeable proportion appears to be.[68]

[66] Recall that all cohort members entered on camera-detected offenses, but the daily rates reported in the recidivism model relate to both camera detected and directly observed offenses.

[67] See, for e.g., Gebers MA. An inventory of California driver accident risk factors. Technical report; California Department of Motor Vehicles: 2003

[68] It is difficult to find statistics to quantify this split. Data from the National Highway Transportation Safety Agency indicate that in 2012 31 percent of all road fatalities (n=10,322) were due to alcohol-impaired driving and 31 percent of all fatalities (n=10,219) were due to speeding. Red light runners are apparently responsible for 2% of fatalities and 7% of injuries. But there is undoubtedly overlap between these various groups, and the representation of other types of offending behaviours in crashes is not readily available.

Each of these factors weakens the relationship between the incidence of offending behaviour and crashes at the population level. [69] The sinews are probably loose enough to permit offending rates within a defined population to decrease or increase across a fairly large range without altering crash rates. This theory does not explain why, in our study, recidivism was deterred and crashes were not, but it explains how the two effects can coexist.

4.7.2 Is specific deterrence working in Queensland?

Unlike some other realms of law—much of criminal law, for example—traffic laws face a separation between the bad act and the bad outcome. The regime is oriented almost entirely to certain bad acts—specifically, it penalizes bad acts known or believed to be associated with bad outcomes (i.e. crashes, injury, and property damage). The bad outcome need not materialize in a given case for the penalty to apply—indeed it rarely does.[70] Our results throw this separation between act and outcome into sharp relief. This poses an interesting quandary for specific deterrence.

Tort scholars tend to focus primarily on behaviour change as the principal target for deterrence. At one level, this is sensible. The causes of crashes are multifactorial and the driver carelessness is just one factor. So surely the law should not be judged by its capacity to curtail dangers over which it has little or no control. Nonetheless, honouring the focus on behavioural change might lead one to conclude that Queensland's traffic laws are performing admirably as a specific deterrent.

From a broader public health and policy standpoint, however, the suggestion that a legal regime could be regarded as performing successfully for having curbed a surrogate of the bad outcome, without having any measurable effect on the outcome itself, is somewhat absurd. Traffic safety regulators do not proclaim to sanction unlawful driving as an end in itself; they sanction it to prevent harm. We did not find evidence of such prevention.

One caveat to that conclusion is that that there may have been some true reduction in crashes that we could not detect. The risk ratio for the notification variable in the primary crash risk analysis was 0.94. Although the estimate was not statistically

[69] The weakness of the relationship also helps to explain why it has proven infeasible to predict the incidence of crashes at the population level on the basis of drivers' offense records. Some studies have shown a significant relationship between infringement and crash histories (see, for e.g., Geber MA, Peck RC. Using traffic conviction correlates to identify high accident-risk drivers. Accident Analysis and Prevention 2003;35:903-912). But correlation and prediction are different beasts. No study has shown offense data can predict crashes with levels of accuracy that are high enough to justify aggressive interventions in high risk populations of drivers.

[70] The obvious exception is citations triggered by the behaviour of a driver who has crashed, but these represent only a small fraction of all citations issued.

significant at conventional levels, the relative rarity of crashes meant this analysis was not nearly as highly powered as the recidivism analysis. If it had been statistically significant, a reduction in crash risk of this magnitude is not trivial. Depending on how much it cost, a community might very happily embrace a safety intervention that reduced crash risk by 6 percent.[71]

4.7.3 Other trends in levels risk over time

Although our analysis was designed to examine changes in crash and offending rates immediately after notification, two other patterns appeared repeatedly in the pre- and post-notification periods, and warrant discussion.

Increase in recidivism in the pre-notification period

The primary recidivism plot and most of the stratified recidivism analyses show a steep rise in penalty risk in the pre-notification period, peaking at or shortly before notification. What accounts for this uptick?

Recall that the period of time between the camera detected offense that marked drivers' entry into the cohort and the time at which they received notification of their offense varied considerably. Drivers with notification periods of three weeks or longer contributed to the daily rate calculations beginning on the first day of the pre-notification period. Others had notification periods of less than a week, which meant they would have contributed only to the rates in the few days before the notification week began.

If riskier drivers were more likely to have short notification periods, this might explain the upward-sloping curves we observed in the pre-notification period. We found some evidence to support this theory. For example, daily recidivism rates increased slightly less in the pre-notification period among drivers with characteristics normally associated with lower crash risks (e.g. female drivers, city dwellers, low-level offenders)

Why would high-risk drivers have systematically shorter notification periods? Two reasons seem plausible. First, DTMR may have pursued a practice of faster notification of drivers who appeared to be riskier, based either on the nature of the offense they had committed or their demographic profile. Second, offenders who, on average, took longer to reach may have had lower risks of reoffending than those who were reached more quickly. Drivers living in rural areas, for example, generally have

[71] In the United States, a reduction of that size would equate to over 2,000 deaths and 100,000 injuries per year.

slower postal service than city dwellers, and intensity of enforcement in rural areas, particularly by cameras, is lower than in cities.

Decrease in crashes and recidivism in the post-notification period

Another noteworthy trend, not obviously related to the discontinuity of interest, is the steady decline in rates of both recidivism and crashes observed over the 90-day post-notification period. This is evident in nearly all of plots. One explanation is that the specific deterrent effect gains momentum gradually, possibly instigating a "learning curve" along which drivers move toward safer driving. This is a hopeful scenario, but a doubtful one. It cuts against several other studies that have found the opposite—namely, deterrent effects from traffic penalties tend to decay over a period of a few months.[72]

The steady decreases in the post-notification period are more likely to be an artefact of a methodological limitation of our study. One problem that is more-or-less intrinsic to population-level road safety studies was discussed at length earlier—namely, lack of direct adjustment for driving behaviour. Drivers in our cohort were known to have been driving at the date of their offense,[73] but the extent of their presence on the road becomes less clear the further one moves in time past the offense date. Consequently, the declines in penalty and crash risk observed in the post-notification periods probably reflect a reduced exposure to driving.

There are a couple of further points to be made here. First, this exposure measurement problem should not materially affect our main findings.[74] The daily rates that matter most to our estimates are those close to the discontinuity. Second, if drivers *are* driving less in the post notification period, part of their motivation for doing so may be a desire to avoid further penalties. Reductions in risk that stem

[72]See, for e.g., Homel 1988; Redelmeier et al 2003.

[73]This is not strictly true. The citation is normally sent to the vehicle's registered owner, and someone else may be been at the wheel the vehicle at the time of the camera detected offense. DTMR has a routine process for accepting challenges on "other driver" grounds. If the challenge is upheld, the penalty, together with any associated demerit points, are re-assigned to the actual offender. The offenders in our data are those to whom the penalty was ultimately assigned. Thus, our confidence that the offender was driver can be reasonably high, but some scope for misclassification still exists (e.g. a challenge is wrongly dismissed, a registered owner who did not offend does not challenge and "takes the hit" for someone else, drivers collude in some other way to shuffle penalties and demerit points between each other, etc.).

[74]The same cannot be said of measurement of road safety risks using the case-crossover design, where this phenomenon creates, what epidemiologists refer to as, "confounding by indication" . This is a substantial and largely intractable problem in studies of this kind. See, Redelmeier et al 2003; Simon J. Walter and David M. Studdert, Relationship between penalties for road traffic infringements and crash risk in Queensland, Australia: A case-crossover study (in press, 2015).

from such a response might legitimately be counted as part of a specific deterrent effect, not confounding.

4.7.4 Limitations

Our study has several limitations worth noting. First, the generalizability of our findings outside Queensland is unknown. Even within Queensland, it is unclear whether the cohort entry criterion—camera-detected speeding or red-light offenses—produced a sample of drivers whose subsequent behaviour differs from the universe of offenders.

Second, although we could guess with a fair degree of confidence at a date range within which offenders became aware of their penalty, we did not know with certainty. A non-trivial number of drivers in our sample of nearly three million will have learned of their penalty after the notification week. Penalty letters are not infrequently sent to the wrong address or the wrong person; people are away from their home address for extended periods; and some people defer opening their mail—perhaps especially if it has the hallmarks of bad news from the government!

Relatively few drivers are likely to have received their letters before the notification week. However, some may have known they offended and suspected they were caught at the time of the offense—alerted, for example, by the flash of a camera at night.[75]

Whether actual awareness occurred before or after the notification week we specified, the effect on our results is probably the same: a bias to the null. Incidentally, using a notification week instead of a precise date should bias our results in the same direction.

Third, our decision to censor drivers who crashed or reoffended from the cohort was not ideal, but preferable to the alternative. Many crashes force drivers off the road for a period of time, so retaining them inflates the "at risk" population for an unknown number of subsequent days. Retaining offenders is problematic for a different reason: deterrent effects of the subsequent offense could become entangled with deterrent effects associated with the offense of interest.

There is no realistic possibility that removal of drivers who crashed affected our findings: there were far too few of them. By contrast, 2,000-3,000 drivers reoffended each day, and by the end of the observation period in the recidivism analysis, 14% of the sample had been censored. To explore whether this censoring affected our

[75]Some offenders photographed at night will have experienced camera flashes; others will not have, because many of the cameras in use during the study period were fitted with infrared technology.

results, we re-ran the recidivism model without censoring recidivists. The results changed very little.

Finally, it would be inappropriate to infer from our findings that penalizing traffic offenses does not reduce crashes in Queensland. We did not consider the effects of general deterrence, and it may succeed where specific deterrence fails.

4.8 Conclusion

Traffic laws exist primarily to promote road safety. In a broad sense, "safety" refers to the care and competence with which people behave on and around the road. The two standard ways of measuring safety at the population level are rates of offending and rates of crashes; both are regularly used as proxies for unsafe driving.

Our results call into question some of the assumptions embedded in this practice. We found that offending rates dropped in Queensland following notification of an offense while crash rates were unmoved. This is a form of specific deterrence, but a hollow one.

The split finding also raises fundamental questions about what specific deterrence means and what traffic laws are accomplishing. Should recidivism be de-emphasized as a proxy for safety, and avoided as a measure of deterrence? Is Queensland's regime penalizing the wrong behaviours, or the right behaviours in the wrong way? Could better enforcement bridge the gap and produce real reductions in both recidivism and crash risk? And if true safety effects flow only from general deterrence, should specific deterrence goals be abandoned? Perhaps Queensland would be better off diverting resources from catching and penalizing drivers to publicity campaigns and showy displays of enforcement, in the interests of pursuing a form of deterrence that may work. Future research should address these questions in Queensland and consider them elsewhere. Important aspects of the logic and efficacy of traffic laws hinge on the answers.

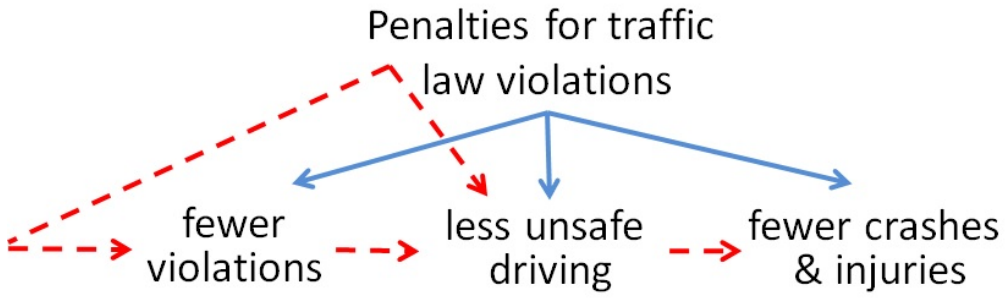
4.8.1 Acknowledgements

The authors gratefully acknowledge the assistance of Nerida Leal, Pam Palmer, and other staff of the Queensland Department of Transport and Main Roads for providing the study data and assisting with interpreting variables and understanding data collection methods.

4.8.2 Funding

Australian Research Council (Laureate Fellowship FL110100102 to Dr. Studdert). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Figure 1. Conceptual model of general and specific deterrent effects of traffic laws



Note: Solid blue lines indicate pathways for general deterrent effects. Dotted red lines indicate pathways for specific deterrent effects.

Table 1. Characteristics of drivers and index offenses in the study sample (n=2,880,763)

	n	%
Sex		
Male	1,736,971	60%
Female	1,143,792	40%
Age		
20-25 years	412,947	14%
26-30 years	364,480	13%
31-60 years	1,842,321	64%
>60 years	261,015	9%
Cumulative demerit point total		
<6 points	1,897,749	66%
6-8 points	572,804	20%
9-11 points	237,512	8%
>12 points	172,698	6%
Type of camera-detected offense		
Speeding	2,617,107	91%
Minor (0 or 1 points)∗	1,417,839	49%
Moderate (3 points)	1,067,386	37%
Major (4+ points)	131,882	5%
Red light violation (3 points)	263,656	9%

∗ 823 speeding violations (0.03% of the sample) resulted in no demerit points.

Table 2. Characteristics of crashes that occurred during the study period (*n*=15,317)

	n	%
Severity		
Fatal	147	1%
Injury requiring hospitalization	3,300	22%
Injury requiring medical treatment outside hospital	4,068	27%
Injury not requiring medical treatment	2,295	15%
Property damage only	5,507	36%
Fault		
At fault	9,548	62%
Not at fault	5,769	38%

Table 3. Characteristics of offenses that occurred during study period, following the index camera-detected offense (n=184,544)*

Type of offense	***n***	%
Speeding	131,922	71%
Minor	54,054	29%
Moderate	62,413	34%
Major	15,455	8%
Red light violation	7,966	4%
Uninsured or unregistered driving	4,860	3%
Use of mobile phone while driving	4,609	2%
Driving under influence of alcohol or drugs	4,571	2%
Illegal stop or park	3,956	2%
Failure to wear seatbelt or helmet	3,316	2%
Administrative non-compliance	2,425	1%
Defective vehicle	2,325	1%
Illegal turn	2,293	1%
Unlicensed driving	2,271	1%
Other failure to stop	2,155	1%
Violation of probationary driving rules	1,232	1%
Other	10,643	6%
Demerit points		
0	25,048	14%
1	57,150	31%
2	2,518	1%
3	84,373	46%
4+	15,455	8%
Mode of detection		
Camera	98,426	53%
Police observation	86,118	47%

*This table does not include the index camera-detected offenses that brought drivers into the study sample

Figure 2. Discontinuity in the crash rate

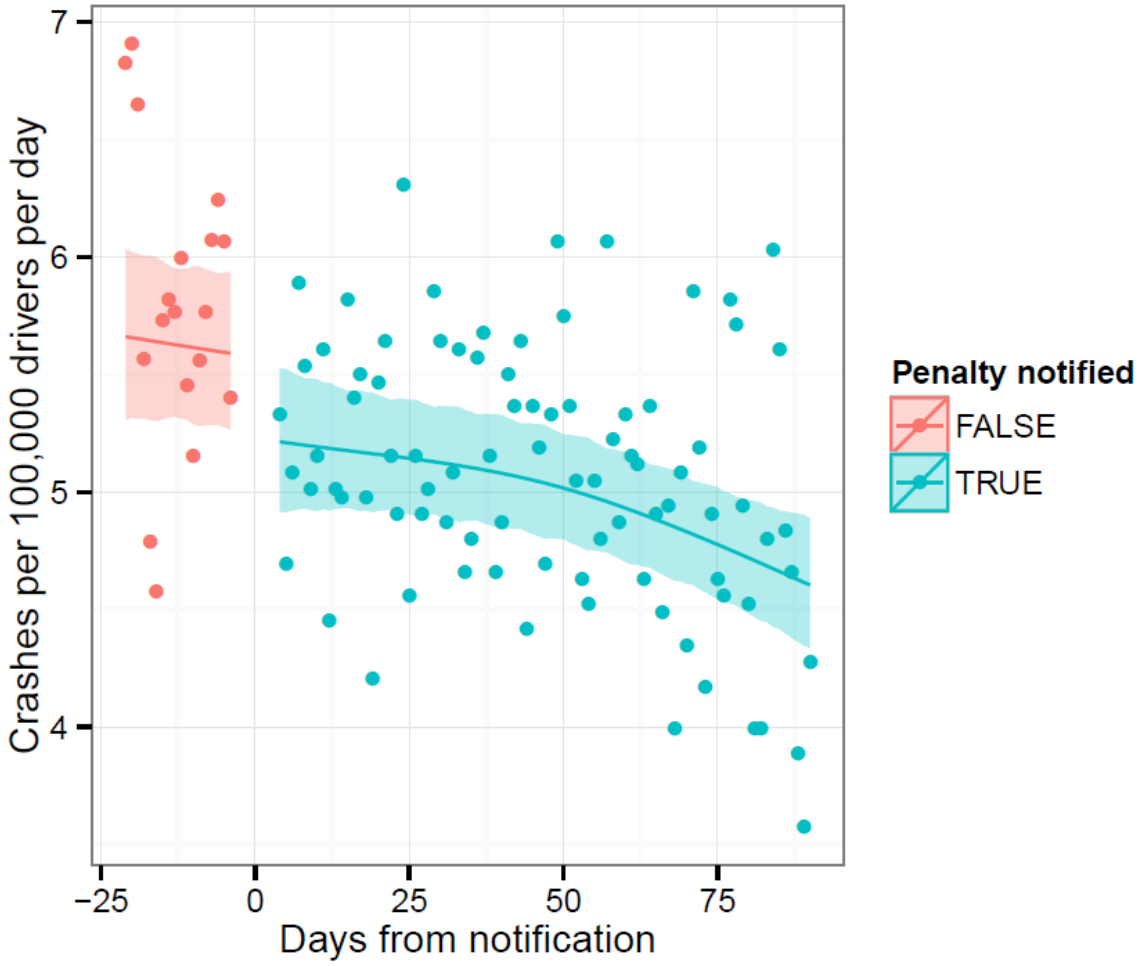


Table 4. Multivariate predictors of crashes*

	Rate ratio (95% CI)	***p*-value**
Notification	0.94 (0.86–1.02)	0.1384
Driver male (ref: Female)	1.31 (1.27–1.36)	<0.001
Driver age (ref: 20–25 years)		
26–30 years	0.73 (0.69–0.77)	<0.001
31–60 years	0.57 (0.55–0.59)	<0.001
>60 years	0.47 (0.44–0.51)	<0.001
Cumulative demerit points (ref: <5)		
5–8 points	1.37 (1.31–1.43)	<0.001
9–11 points	1.73 (1.64–1.82)	<0.001
>11 points	2.60 (2.47–2.74)	<0.001
Type of offense (ref: Red light violation)		
Speeding – minor	0.78 (0.74–0.82)	<0.001
Speeding – moderate	0.77 (0.73–0.81)	<0.001
Speeding – major	0.87 (0.81–0.94)	<0.001

* The model also included a smooth term to adjust for crash risk over time, and dummy variables indicating the number of days from notification modulo 7 to adjust for a day-of-week effect.

Figure 3. Discontinuities in crash rates within defined subgroups

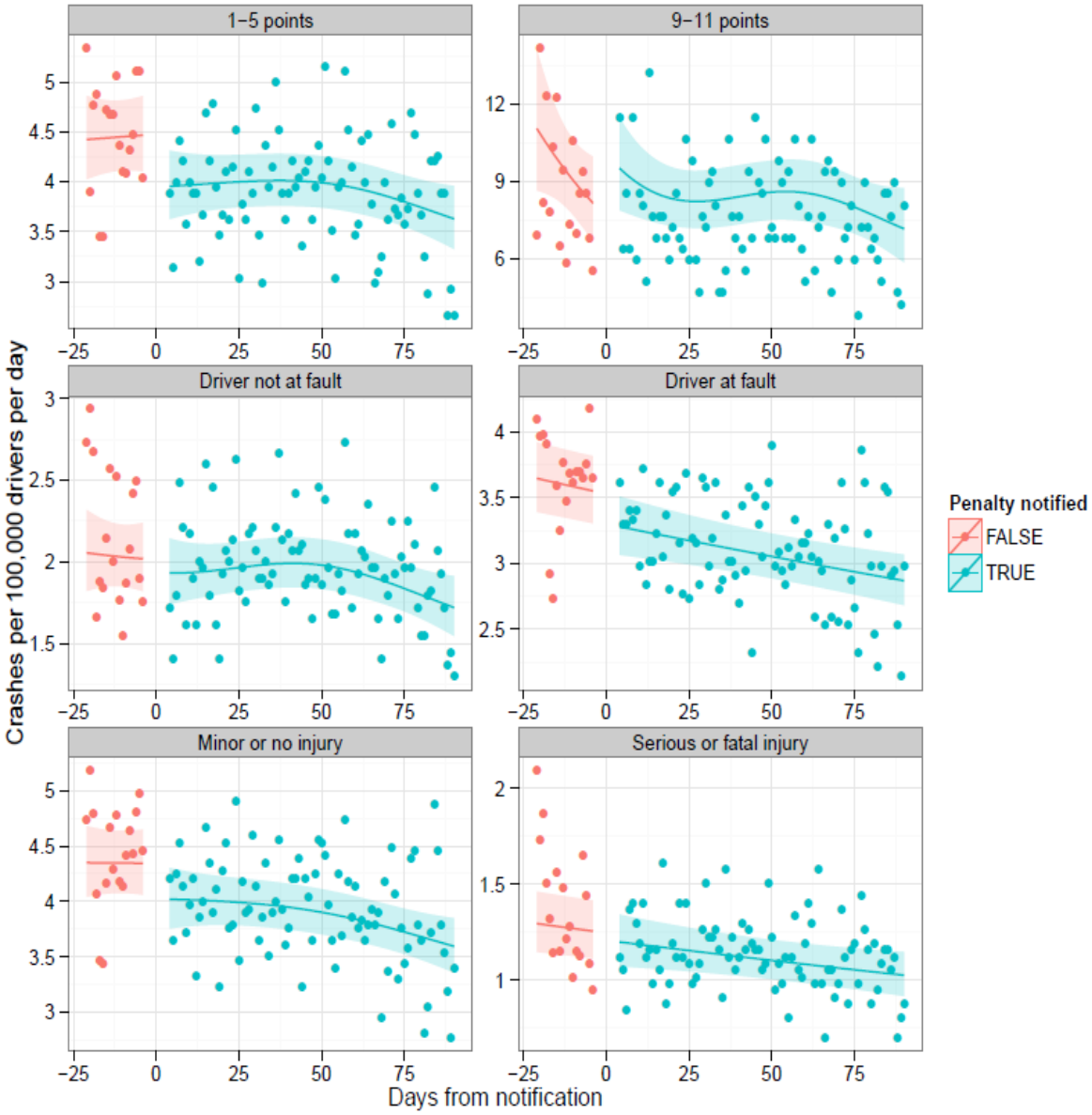


Figure 4. Discontinuity in the recidivism rate

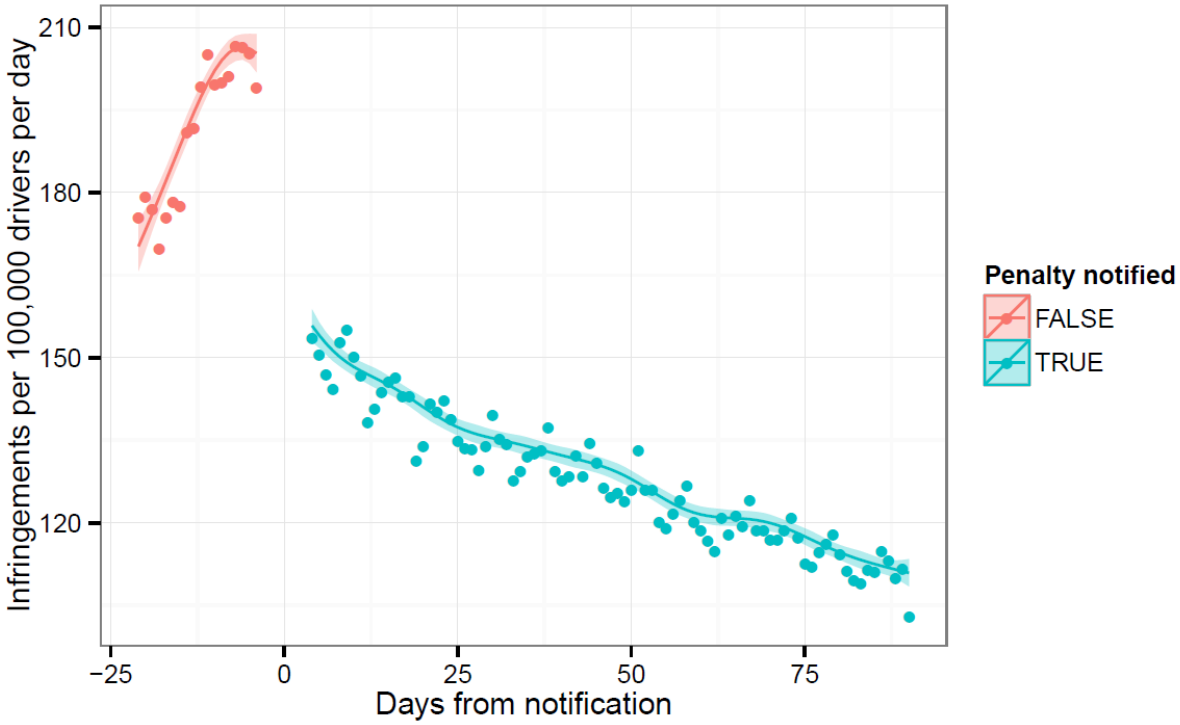
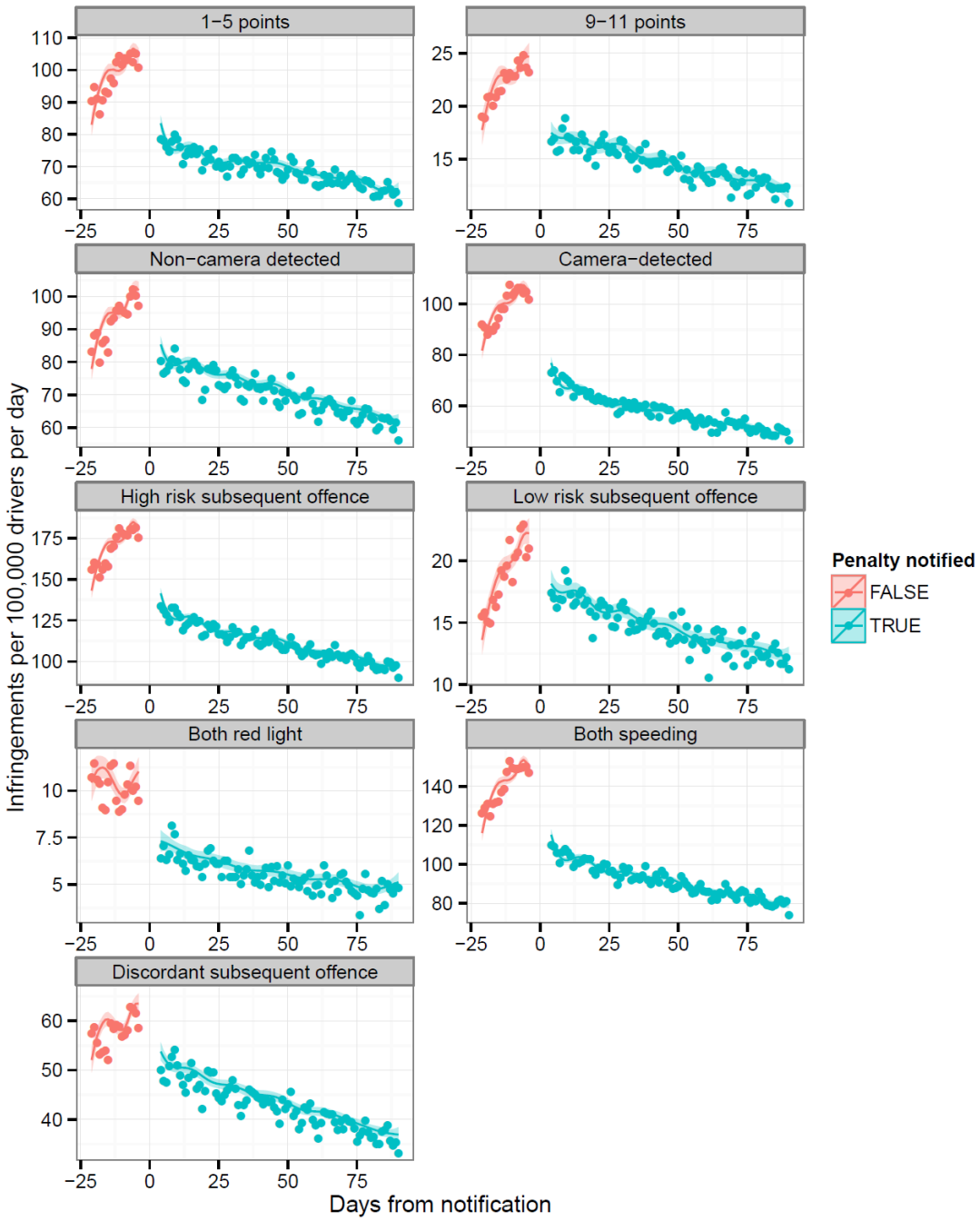


Table 5. Multivariate predictors of recidivism*

	Rate ratio (95% CI)		*p*-value
Notification	0.75	(0.73–0.78)	<0.001
Driver male (ref: Female)	1.31	(1.3–1.32)	<0.001
Driver age (ref: 20–25 years)			
26–30 years	0.87	(0.87–0.88)	<0.001
31–60 years	0.68	(0.67–0.68)	<0.001
>60 years	0.41	(0.41–0.42)	<0.001
Cumulative demerit points (ref: <5)			
5–8 points	1.35	(1.34–1.36)	<0.001
9–11 points	1.58	(1.56–1.6)	<0.001
>11 points	2.26	(2.23–2.28)	<0.001
Type of offense (ref: Red light violation)			
Speeding – minor	1.16	(1.15–1.17)	<0.001
Speeding – moderate	1.08	(1.06–1.09)	<0.001
Speeding – major	1.12	(1.11–1.14)	<0.001

* The model also included a smooth term to adjust for re-offending risk over time, and dummy variables indicating the number of days from notification modulo 7 to adjust for a day-of-week effect.

Figure 5. Discontinuities in recidivism rates within defined subgroups



Bibliography

John Aitchison. Statistical problems of treatment allocation. *Journal of the Royal Statistical Society. Series A*, 133(2):206–239, 1970.

Donald W. K. Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405, 2000.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Rudolf Beran. Estimated sampling distributions: the bootstrap and competitors. *The Annals of Statistics*, pages 212–225, 1982.

Christoph Bergmeir, Robert J. Hyndman, and Jos M. Benitez. Bagging exponential smoothing methods using stl decomposition and box-cox transformation. *International Journal of Forecasting*, 32(2):303 – 312, 2016.

Claude Bernard. *Introduction à l'étude de la médecine expérimentale*. 1865.

Gérard Biau, Frédéric Cérou, and Arnaud Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *Journal of Machine Learning Research*, 11 (Feb):687–712, 2010.

Peter J. Bickel and David A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, pages 1196–1217, 1981.

Peter J. Bickel, Elizaveta Levina, et al. Some theory for Fisher's linear discriminant function, naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

Jeff A. Bilmes and Katrin Kirchhoff. Generalized rules for combination and joint training of classifiers. *Pattern Analysis & Applications*, 6(3):201–211, 2003.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

David P Byar and Donald K Corle. Selecting optimal treatment in clinical trials using covariate information. *Journal of chronic diseases*, 30(7):445–459, 1977.

T. Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

T. Tony Cai, Cun-Hui Zhang, Harrison H Zhou, et al. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

T. Tony Cai, Harrison H Zhou, et al. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, 2012.

Edward Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, 14(3):1171–1179, 1986.

Yilun Chen, Ami Wiesel, and Alfred O. Hero. Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*, 59(9): 4097–4107, 2011.

Guillem Collell, Drazen Prelec, and Kaustubh R. Patil. A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing*, 275:330–340, 2018.

Timothy F. Cootes, Andrew Hill, Christopher J. Taylor, and Jane Haslam. The use of active shape models for locating structures in medical images. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 33–47. Springer, 1993.

Alan R. Dabney. Classification of microarrays to nearest centroids. *Bioinformatics*, 21(22):4148–4154, 2005.

Alan R. Dabney and John D. Storey. Optimality driven nearest centroid classification from genomic data. *PLoS One*, 2(10):e1002, 2007.

Persi Diaconis and Bradley Efron. Computer intensive methods in statistics. *Scientific American*, 248(5), 1983.

Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

Thomas J Fisher and Xiaoqian Sun. Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Computational Statistics & Data Analysis*, 55(5):1909–1918, 2011.

Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, 2011.

Hector Franco-Lopez, Alan R Ek, and Marvin E Bauer. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sensing of Environment*, 77(3):251–274, 2001.

Reinhard Furrer and Thomas Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.

Peter Hall. Resampling a coverage pattern. *Stochastic Processes and their Applications*, 20(2):231–246, 1985.

Peter Hall, Hugh Miller, et al. Using the bootstrap to quantify the authority of an empirical ranking. *The Annals of Statistics*, 37(6B):3929–3959, 2009.

Peter Hall, Tung Pham, et al. Optimal properties of centroid-based classifiers for very high-dimensional data. *The Annals of Statistics*, 38(2):1071–1093, 2010.

Eric Hillebrand and Marcelo C Medeiros. The benefits of bagging for forecast models of realized volatility. *Econometric Reviews*, 29(5-6):571–593, 2010.

Jianhua Z Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93 (1):85–98, 2006.

Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

Noureddine El Karoui and Elizabeth Purdom. The bootstrap, covariance matrices and PCA in moderate and high-dimensions, 2016.

Noureddine El Karoui and Elizabeth Purdom. Can we trust the bootstrap in high-dimensions? The case of linear models. *Journal of Machine Learning Research*, 19 (5):1–66, 2018.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

Sören Künzel, Jasjeet Sekhon, Peter Bickel, and Bin Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv preprint arXiv:1706.03461*, 2017.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001, 2009.

Alexander R Luedtke and Mark J van der Laan. Super-learning of an optimal dynamic treatment rule. *The International Journal of Biostatistics*, 12(1):305–332, 2016.

Brett A McKinney, David M Reif, Michael T Rock, Kathryn M Edwards, Stephen F Kingsmore, Jason H Moore, and James E Crowe. Cytokine expression patterns associated with systemic adverse events following smallpox immunization. *The Journal of Infectious Diseases*, 194(4):444–453, 2006.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 06 2006.

Rupert G Miller. Least squares regression with censored data. *Biometrika*, 63(3): 449–464, 1976.

Sahand Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 04 2011.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.

Dimitris N. Politis and Joseph P. Romano. A circular block-resampling procedure for stationary data. In *Exploring the Limits of Bootstrap (East Lansing, MI, 1990)*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., pages 263–270. Wiley, New York, 1992.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):pp. 1303–1313, 1994.

Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 04 2011.

Daniel Rubin and Mark J van der Laan. A doubly robust censoring unbiased transformation. *The International Journal of Biostatistics*, 3(1), 2007.

Jon R. Schoonover, Rob Marx, and Shuliang L. Zhang. Multivariate curve resolution in the analysis of vibrational spectroscopy data files. *Applied Spectroscopy*, 57(5): 154A–170A, May 2003.

A. Sharma and K. K. Paliwal. Improved nearest centroid classifier with shrunken distance measure for null lda method on cancer classification problem. *Electronics Letters*, 46:1251–1252(1), September 2010.

James E Signorovitch. Estimation and evaluation of regression models for patient-specic treatment efficacy. *Harvard Biostatistics PhD Thesis*, (4):69–93, 2007.

RICHARD Simon. Patient subsets and variation in therapeutic efficacy. *British Journal of Clinical Pharmacology*, 14(4):473–482, 1982.

Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158, 2009.

Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

Sijian Wang and Ji Zhu. Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics*, 23(8):972–979, 03 2007.

Wei Biao Wu and Mohsen Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 12 2003.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.

Appendix A

Technical results for the asynchronous bootstrap

A.1 Detailed proofs

There is some overlap in text in this appendix with the outline of proofs given in section 5 of the main paper.

A.1.1 Proof of Theorem 1

As in the main paper we consider only part (b) of the theorem, since part (a) be derived in a similar fashion to the proof of Theorem 2. Assume for notational simplicity that $p = bk$, where b, denoting block length, and k, the number of blocks, are both positive integers. Writing $\mathbb{E}_{\mathcal{X}}$ for expectation conditional on $\mathcal{X}$, we have:

$$S^* = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} (1 - \mathbb{E}_{\mathcal{X}}) g \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ X_{ij}^*(k) - \bar{X}_j \right\} \right] = \frac{1}{\sqrt{k}} \sum_{j=1}^{k} T_j^*,$$

where

$$T_j^* = \frac{1}{b^{1/2}} \sum_{r=1}^{b} (1 - \mathbb{E}_{\mathcal{X}}) g(V_{jr}^*), \quad V_{jr}^* = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left\{ X_{i,(j-1)b+r}^*(k) - \bar{X}_{(j-1)b+r} \right\},$$

and, by the definition of the block-bagging algorithm, if block-bagging is used then the variables T_j^*, for $1 \leq j \leq k$, are independent conditional on $\mathcal{X}$. Let $\mathrm{var}_{\mathcal{X}}$ and

$\mathrm{cov}_{\mathcal{X}}$ denote variance and covariance, respectively, conditional on $\mathcal{X}$, and note that

$$b\,\mathrm{var}_{\mathcal{X}}(T_j^*) = \sum_{r_1=1}^{b}\sum_{r_2=1}^{b} \mathrm{cov}_{\mathcal{X}}\left\{g(V_{jr_1}^*), g(V_{jr_2}^*)\right\}. \tag{A.1}$$

If (2.9) holds then, defining $\hat{\sigma}_j(r_\ell)^2 = \mathrm{var}_{\mathcal{X}}\{X_{i,(j-1)b+r_\ell}^*(k)\}$ and

$$\hat{\theta}_j(r_1,r_2) = \frac{1}{\hat{\sigma}_j(r_1),\hat{\sigma}_j(r_2)}\left|\mathrm{cov}_{\mathcal{X}}\left\{X_{i,(j-1)b+r_1}^*(k), X_{i,(j-1)b+r_2}^*(k)\middle|\mathcal{X}\right\}\right|$$

we have:

$$\begin{aligned}
&\left|\,\mathrm{cov}_{\mathcal{X}}\{g(V_{jr_1}^*), g(V_{jr_2}^*)\}\right| \\
&\quad\le C_1\Bigg[\left\{1-\hat{\theta}_j(r_1,r_2)\right\}^{-3/2}\frac{1}{n^{3/2}}\sum_{\ell=1}^{2}\frac{1}{\hat{\sigma}_j(r_\ell)^3}\sum_{i=1}^{n}\mathbb{E}\left\{\left|\left(X_{i,(j-1)b+r_\ell}^*(k)\right.\right.\right. \\
&\qquad\left.\left.\left.-\bar{X}_{(j-1)b+r_\ell}\right)\right|^3\middle|\mathcal{X}\right\} + \hat{\theta}_j(r_1,r_2)\Bigg]
\end{aligned} \tag{A.2}$$

Next we use (A.2) to derive an upper bound to $\mathbb{E}\left|\mathrm{cov}_{\mathcal{X}}\left\{g(V_{jr_1}^*), g(V_{jr_2}^*)\right\}\right|$. By the m-dependence property,

$$\mathbb{E}\left|\,\mathrm{cov}_{\mathcal{X}}(X_{i,(j-1)b+r_1}^*, X_{i,(j-1)b+r_2}^*)\right| = \mathcal{O}(n^{-1/2}), \tag{A.3}$$

uniformly in $1\le i\le n$, $1\le j\le p$, $1\le r_1,r_2\le b$ and $|r_1-r_2|>m$. (In fact, the left-hand side of (A.3) does not depend on i,j,r_1 or r_2 in this range.) Hence for each $\varepsilon>0$,

$$\mathbb{P}\left\{\left|\,\mathrm{cov}_{\mathcal{X}}(X_{i,(j-1)b+r_1}^*, X_{i,(j-1)b+r_2}^*)\right|>\varepsilon\right\} = \mathcal{O}(n^{-1/2})) \tag{A.4}$$

uniformly in the same sense. Defining σ^2 to be the common variance of the X_{ij}s, and using the finiteness of third moments of the process X_0, we have for each r and each $\varepsilon>0$:

$$\mathbb{P}\{|\hat{\sigma}_j(r)^2-\sigma^2|>\varepsilon\} \le \varepsilon^{-3/2}\,\mathbb{E}\,|\hat{\sigma}_j(r)^2-\sigma^2|^{3/2} = \mathcal{O}(n^{-1/2}) \tag{A.5}$$

Therefore, taking ε small and noting that $\sup|g|\le C_2$, we obtain:

$$\begin{aligned}
&\mathbb{E}\left|\mathrm{cov}_{\mathcal{X}}\left\{g(V_{jr_1}^*), g(V_{jr_2}^*)\right\}\right| \\
&\quad\le C_3\Bigg[n^{-1/2}\,\mathbb{E}(|X_{ij}^*-\bar{X}_j|^3) + \mathbb{E}\left|\,\mathrm{cov}_{\mathcal{X}}(X_{i,(j-1)b+r_1}^*, X_{i,(j-1)b+r_2}^*)\right| \\
&\qquad + \mathbb{P}\left\{\left|\,\mathrm{cov}_{\mathcal{X}}(X_{i,(j-1)b+r_1}^*, X_{i,(j-1)b+r_2}^*)\right|>\varepsilon\right\} + \sum_{\ell=1}^{2}\mathbb{P}\{|\hat{\sigma}_j(r_\ell)^2-\sigma^2|>\varepsilon\}\Bigg] \\
&\quad= \mathcal{O}(n^{-1/2})
\end{aligned} \tag{A.6}$$

uniformly in $1 \leq i \leq n, 1 \leq j \leq p, 1 \leq r_1, r_2 \leq b$ and $|r_1 - r_2| > m$, where the inequality follows from (A.1) and the equality from (A.3)-(A.5).

Result (A.6), and the condition $b^2/n \to 0$ as p diverges (which follows from the assumption $b^{2+\delta}/n = \mathcal{O}(1)$ in the theorem), imply that

$$\sum_{r_1=1}^{b} \sum_{1 \leq r_2 \leq b, |r_1 - r_2| > m} \mathbb{E}\left[\left|\mathrm{cov}_{\mathcal{X}}\{g(V^*_{jr_1}), g(V^*_{jr_2})\}\right|\right] = \mathcal{O}(n^{-1/2}b^2) = o(b) \tag{A.7}$$

Again using the m-dependence property,

$$\mathrm{cov}\left[\mathrm{cov}_{\mathcal{X}}\left\{g(V^*_{jr_1}), g(V^*_{jr_2})\right\}, \mathrm{cov}_{\mathcal{X}}\left\{g(V^*_{jr_3}), g(V^*_{jr_4})\right\}\right] = 0$$

if neither of r_1 and r_2 is nearer than m to either of r_3 and r_4. Therefore,

$$\begin{aligned}
&\mathrm{var}\left[\sum_{r_1=1}^{b} \sum_{1 \leq r_2 \leq b, |r_1 - r_2| > m} \mathrm{cov}_{\mathcal{X}}\{g(V^*_{jr_1}), g(V^*_{jr_2})\}\right] \\
&\quad = \sum_{r_1=1}^{b} \sum_{1 \leq r_2 \leq b, |r_1 - r_2| > m} \sum_{r_3=1}^{b} \sum_{1 \leq r_4 \leq b, |r_3 - r_4| > m} \\
&\qquad\qquad \mathrm{cov}\left[\mathrm{cov}_{\mathcal{X}}\left\{g(V^*_{jr_1}), g(V^*_{jr_2})\right\}, \mathrm{cov}_{\mathcal{X}}\left\{g(V^*_{jr_3}), g(V^*_{jr_4})\right\}\right] \\
&\quad = o(b^2).
\end{aligned} \tag{A.8}$$

Combining (A.1), (A.7) and (A.8) we deduce that

$$b\,\mathrm{var}_{\mathcal{X}}(T^*_j) = \sum_{r_1=1}^{b} \sum_{r_2=\max(1, r_1-m)}^{\min(b, r_1+m)} \mathbb{E}\left[\mathrm{cov}_{\mathcal{X}}\left\{g(V^*_{jr_1}), g(V^*_{jr_2})\right\}\right] + o_p(b). \tag{A.9}$$

Standard arguments show that

$$\mathbb{E}\left[\mathrm{cov}_{\mathcal{X}}\left\{g(V^*_{jr_1}), g(V^*_{jr_2})\right\}\right] = \mathrm{cov}\{g(Q_{r_1}), g(Q_{r_2})\} + o(1) \tag{A.10}$$

Neither the left-hand side, nor the covariance on the right-hand side, depends on b, j or p, and they depend only r_1 and r_2 only through the value of $r_1 - r_2$. Together, (A.9) and (A.10) imply that

$$\mathrm{var}_{\mathcal{X}}(T^*_j) = \sum_{r=1}^{2m} \mathrm{cov}\{g(Q_r), g(Q_m)\} + R_j, \tag{A.11}$$

where the random variables R_j are identically distributed and satisfy $R_j = o_p(1)$. Similar arguments show that, for either choice of the $\pm$ signs,

$$\mathrm{cov}_{\mathcal{X}}(T_j^*, T_{j\pm1}^*) = R_{j,\pm}, \tag{A.12}$$

where the variables $R_{j,+}$ are identically distributed, as too are the variables $R_{j,-}$, and $R_{j,\pm} = o_p(1)$. Combining (A.11) and (A.12) we deduce that, in the case of the asynchronous bootstrap,

$$\begin{aligned} k\,\mathrm{var}(S^*|\mathcal{X}) &= \sum_{j_1=1}^{k}\sum_{j_2=1}^{k}\mathrm{cov}(T_{j_1}^*, T_{j_2}^*|\mathcal{X}) \\ &= \sum_{j_1=1}^{k}\sum_{j_2=j_1-1,j_1,j_1+1}\mathrm{cov}(T_{j_1}^*, T_{j_2}^*|\mathcal{X}) \\ &= \sum_{j=1}^{k}\mathrm{var}_{\mathcal{X}}(T_j^*) + o_p(k) = k\sum_{r=1}^{2m}\mathrm{cov}\{g(Q_r), g(Q_m)\} + o_p(k) \end{aligned}$$

That is,

$$\mathrm{var}(S^*|\mathcal{X}) = \sum_{r=1}^{2m}\mathrm{cov}\{g(Q_r), g(Q_m)\} + o_p(1). \tag{A.13}$$

By (2.5) and the m-dependence property,

$$\begin{aligned} p\,\mathrm{var}(S) &= \sum_{j=1}^{p}\sum_{k=1}^{p}\mathrm{cov}\{g(Q_j), g(Q_k)\} = \sum_{j=1}^{p}\sum_{1\le k\le p, |k-j|\le m}\mathrm{cov}\{g(Q_j), g(Q_k)\} \\ &= p\sum_{r=1}^{2m}\{g(Q_r), g(Q_m)\} + o(p) \end{aligned} \tag{A.14}$$

Combining (A.13) and (A.14) we deduce that

$$\mathrm{var}(S^*|X) = \mathrm{var}(S) + o_p(1). \tag{A.15}$$

This is equivalent to the first part of (2.7).

In view of the first part of (2.7), to establish the second part it suffices to note that the variables

$$U_j^* = (1 - \mathbb{E}_{\mathcal{X}})g\left[\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left\{X_{ij}^*(k) - \bar{X}_j\right\}\right],$$

in the formula

$$S^* = \frac{1}{k^{1/2}} \sum_{j=1}^{k} T_j^* = \frac{1}{p^{1/2}} \sum_{j=1}^{p} U_j^*$$

are b-dependent conditional on $\mathcal{X}$, and to check that that the sufficient conditions given by Berk (1973), for a central limit theorem for b-dependent random variables, hold. Those conditions reduce here to

(a) $\sup_j \mathbb{E}(|U_j^*|^{2+\varepsilon}|\mathcal{X}) \le C_4 < \infty$

(b)

$$\lim_{C\to\infty} \liminf_{p\to\infty} \mathbb{P}\biggl\{ \mathrm{var}_{\mathcal{X}} \biggl(\sum_{j=j_1+1}^{j_2} U_j^* \biggr) \le C(j_2 - j_1) \text{ for all } 1 \le j_1 < j_2 \le p \biggr\} \to 1$$

(c) the in-probability limit of $p^{-1} \mathrm{var}_{\mathcal{X}}(\sum_j U_j^*)$ exists and is finite and nonzero, and

(d) $b^{2+(2/\varepsilon)}/p \to 0$.

Indeed, (a) follows from the boundedness of g, (b) can be proved using the argument leading to (A.15), (c) is a consequence of the first part of (2.7), and since (a) holds with $\varepsilon < \infty$ then for (d) it suffices to have $b^{2+\delta}/p \to 0$ for some $\delta > 0$; the latter property follows from the assumptions $b^{2+\delta}/n = \mathcal{O}(1)$ and $n/p \to 0$ imposed in the theorem.

A.1.2 Proof of Theorem 2.

Step 1: Decomposition of $\mathrm{var}(S^*|\mathcal{X})$ In the case of independent marginals, and when $g(x) = \mathbb{I}(x \le y)$, we have $\pi \equiv P(Q_1 \le y) \to \pi_0$, say, and $v^2 = \pi_0(1-\pi_0)$ is the limit of $\mathbb{E}(S^2)$. Define $\hat{\pi}_j = \mathbb{P}(Q_j^* \le y|\mathcal{X})$, an estimator of π, and put

$$\Delta_{jk} = \mathbb{P}(Q_j^* \le y, Q_k^* \le y|\mathcal{X}) - \mathbb{P}(Q_j^* \le y|\mathcal{X})\mathbb{P}(Q_k^* \le y|\mathcal{X}).$$

In this notation,

$$\mathrm{var}(S^*|\mathcal{X}) = \frac{1}{p} \sum_{j=1}^{p} \sum_{k=1}^{p} \Delta_{jk} = A_1 + A_2 \tag{A.16}$$

where

$$A_1 = \frac{1}{p} \sum_{j=1}^{p} \Delta_{jj} = \frac{1}{p} \sum_{j=1}^{p} \hat{\pi}_j(1-\hat{\pi}_j), \qquad A_2 = \frac{1}{p} \sum_{j,k:j\ne k} \sum \Delta_{jk}$$

Now, $\mathbb{E}(\hat{\pi}_1 - \pi)^2 \to 0$ as $n \to \infty$, whence it follows, since the variables $\hat{\pi}_j$ are identically distributed, that $A_1 = \pi(1-\pi) + o_p(1)$. From this property, (2.29), a central limit theorem and the fact that $\pi(1-\pi) \to v^2$, it follows that either part of (2.7) holds, and in particular that $\mathbb{E}(S^{*2}|X) \to v^2$, if and only if

$$A_2 \to 0 \text{ in probability.} \tag{A.17}$$

as $p \to \infty$.

Our proof of the equivalence of (2.10) and $\mathbb{E}(S^{*2}|\mathcal{X})/\mathbb{E}(S^2) \to 1$ in probability is completed by showing that (2.10) is necessary and sufficient for (A.17).

Step 2: Proof of Equivalence of (2.10) *and* (A.17). The proof is in four parts, given in sections A.2.1-A.2.4 respectively.

Expansion of bivariate normal distribution

To simplify exposition we take $\text{var}(X_{01}) = 1$. Recall the definition of Q_j in section 4.2. The random variables X_{i1} and X_{i2} are, under assumption (2.6), independent and identically distributed, but in this paragraph it is convenient to permit X_{i1} and X_{i2} to have a small but not necessarily zero correlation, $\rho = \rho(n)$, say, which converges to zero as $p \to \infty$. If the common distribution of the components X_ij has sufficiently many finite moments, and if the joint distribution of (X_{i1}, X_{i2}) is sufficiently smooth, then the probability $\mathbb{P}(Q_1 \le y, Q_2 \le y)$ can be developed in an Edgeworth expansion, in which the leading term is $\mathbb{P}_\rho(\xi_1 \le y, \xi_2 \le y|X)$, where $\mathbb{P}_\rho$ denotes probability measure for a two-vector (ξ_1, ξ_2) that has a joint normal distribution with zero means, unit variances and correlation ρ. (Without loss of generality the data X_{ij} have unit variances.) By Taylor expansion of the bivariate normal distribution function it can be proved that there exists an absolute constant $A > 0$ such that

$$\left|\mathbb{P}_\rho(\xi_1 \le y, \xi_2 \le y) - \{a_0(y) + a_1(y))\rho + a_2(y)\rho^2\}\right| \le A|\rho|^3 \tag{A.18}$$

whenever $|\rho| \le \frac{1}{2}$, where the nonzero constants $a_\ell(y)$ depend only on y and, for example,

$$a_0(y) = \mathbb{P}(\xi_1 \le y)^2, \quad a_1(y) = \frac{1}{2\pi} e^{-y^2} \tag{A.19}$$

Define $\hat{\rho}_{jk} = \hat{\gamma}_{jk}/(\hat{\sigma}_j \hat{\sigma}_k)$ where

$$\hat{\sigma}_j = \text{var}(Q_j^*|\mathcal{X}) = \frac{1}{n}\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2,$$

$$\hat{\gamma}_{jk} = \text{cov}(Q_j^*, Q_k^*|\mathcal{X}) = \frac{1}{n}\sum_{i=1}^n (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k).$$

Analogously to the expansions discussed in the previous paragraph, an Edgeworth expansion of $\mathbb{P}(Q_j^* \le y\hat\sigma_j, Q_k^* \le y\hat\sigma_k|\mathcal{X})$ has, as its first term, the probability $\mathbb{P}_\rho(\xi_1 \le y, \xi_2 \le y)$ in (A.18), with $\rho = \hat\rho_j k$. (We take ξ_1 and ξ_2 to be independent of $\mathcal{X}$, and preserve the interpretation that $\hat\rho_{jk}$ is a constant by treating the probability $\mathbb{P}_\rho(\xi_1 \le y, \xi_2 \le y)$ as $\mathbb{P}_\rho(\xi_1 \le y, \xi_2 \le y|\mathcal{X})$.) Therefore, in view of (A.18), the first term in an Edgeworth expansion of $\mathbb{P}(Q_j^* \le y\hat\sigma_j|\mathcal{X})\mathbb{P}(Q_k^* \le y\hat\sigma_k|\mathcal{X})$ can be written as

$$a_0(y) + a_1(y)\hat\rho_{jk} + a_2(y)\hat\rho_{jk}^2 + A\Theta_{jk}|\hat\rho_{jk}|^3,$$

where, for all $j \ne k$

$$\mathbb{P}(|\Theta_{jk}| > 1, |\hat\rho_{jk}| \le \tfrac{1}{2}) = 0 \tag{A.20}$$

First term in an Edgeworth expansion

Similarly, the first term in the Edgeworth expansions of each of $\mathbb{P}(Q_j^* \le y\hat\sigma_j|\mathcal{X})$ and $\mathbb{P}(Q_k^* \le y\hat\sigma_k|\mathcal{X})$ equals $\mathbb{P}(\xi_1 \le y)$, and so the first term in the analogous expansion of $\mathbb{P}(Q_j^* \le y\hat\sigma_k|\mathcal{X})$ equals $\mathbb{P}(\xi_1 \le y)^2 = a_0(y)$; see (A.19). Hence, the first term in the expansion of

$$\Delta \equiv \sum\sum_{j,k:j\ne k} \big\{\mathbb{P}(Q_j^* \le y\hat\sigma_j, Q_k^* \le y\hat\sigma_k|\mathcal{X}) - \mathbb{P}(Q_j^* \le y\hat\sigma_j|\mathcal{X})\mathbb{P}(Q_k^* \le y\hat\sigma_k|\mathcal{X})\big\} \tag{A.21}$$

equals

$$\begin{aligned}&\sum\sum_{j,k:j\ne k}\big\{a_0(y) + a_1(y)\hat\rho_{jk} + a_2(y)\hat\rho_{jk}^2 + A\Theta_{jk}|\hat\rho_{jk}|^3\big\}\\ &\qquad = a_0(y)\sum\sum_{j,k:j\ne k}\hat\rho_{jk} + a_1(y)\sum\sum_{j,k:j\ne k}\hat\rho_{jk}^2 + a_2(y)\sum\sum_{j,k:j\ne k}A\Theta_{jk}|\hat\rho_{jk}|^3.\end{aligned} \tag{A.22}$$

If $\varepsilon > 0$ is given, then, by taking $C_2 > 0$ sufficiently large in the moment condition $\mathbb{E}\,|Z|^{C_2} < \infty$ in (4.7), it can be proved that

$$\mathbb{P}\left(\max_{j,k:j\ne k}|\hat\rho_{jk}| > n^{\varepsilon-\frac{1}{2}}\right) \to 0$$

Therefore, by (A.20),

$$\sum\sum_{j,k:j\ne k}\Theta_{jk}|\hat\rho_{jk}|^3 = \mathcal{O}_p\left(p^2 n^{3\varepsilon-\frac{3}{2}}\right) \tag{A.23}$$

Furthermore, since $\hat\rho_{jj} = 1$ for each j,

$$\sum_{j,k:j\neq k}\sum \hat\rho_{jk} = \sum_{j=1}^p\sum_{k=1}^p \hat\rho_{jk} - \sum_{j=1}^p \hat\rho_{jj}^2 = \frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^p \frac{X_{ij}-\bar X_j}{\hat\sigma_j}\right)^2 - \sum_{j=1}^p \hat\rho_{jj}^2$$
$$= p + o_p(p) - p = o_p(p) \tag{A.24}$$

and more simply, since $w = 1$,

$$\sum_{j,k:j\neq k}\sum \hat\rho_{jk}^2 = \{1+o_p(1)\}\sum_{j,k:j\neq k}\sum \hat\gamma_{jk}^2$$
$$= \{1+o_p(1)\}n^{-1}\sum_{j,k:j\neq k}\sum 1 = \{1+o_p(1)\}\frac{p(p-1)}{n}.$$

Hence,

$$\sum_{j,k:j\neq k}\sum \hat\rho_{jk} = \{1+o_p(1)\}\frac{p(p-1)}{n}. \tag{A.25}$$

Combining (A.22)–(A.25), and choosing $\varepsilon < \frac{1}{6}$ in (A.23), we deduce that the first term in an Edgeworth expansion of the quantity Δ defined at (A.21) equals $a_2(y)n^{-1}p(p-1) + o_p(p+n^{-1}p^2)$. A similar analysis of higher-order terms shows that their net contribution equals $o_p(p+n^{-1}p^2)$; see the paragraph below. Moreover, the expansions are valid uniformly in real y. Therefore, again uniformly in y,

$$\Delta = a_2(y)n^{-1}p(p-1) + o_p(p+n^{-1}p^2) \tag{A.26}$$

Completion of proof of (A.26)

We complete the proof by establishing the claim made in the previous paragraph about high-order terms in an Edgeworth expansion of Δ. Note that

$$\mathbb{P}(Q_j^* \le \hat\sigma_j y\big|\mathcal{X}) = \mathbb{P}(\xi_1 \le y) + n^{-1/2}(\hat a_j + \hat b_j y^2)\phi(y)$$
$$+ n^{-1}(\hat c_j y + \hat d_j y^3 + \hat e_j y^5)\phi(y)$$
$$+ \mathcal{O}_p\left(n^{\varepsilon-\frac{3}{2}}\right) \tag{A.27}$$

$$\mathbb{P}(Q_j^* \le \hat\sigma_j y, Q_j^* \le \hat\sigma_k y|\mathcal{X}) = \mathbb{P}_{\hat\rho_{jk}}(\xi_1 \le y, \xi_2 \le y) + n^{-1/2}(\hat a_{jk} + \hat b_{jk} y^2)\phi(y)^2$$
$$+ n^{-1}(\hat c_{jk} y = \hat d_{jk} y^3 + \hat e_{jk} y^5)\phi(y)^2$$
$$+ \mathcal{O}_p\left(n^{\varepsilon-\frac{3}{2}}\right) \tag{A.28}$$

where ρ denotes the standard normal density, and, for any $\varepsilon > 0$, the $\mathcal{O}_p(n^{\varepsilon} - \frac{3}{2})$ remainders are of the stated orders uniformly in $1 \le j, k \le p$, provided the constant C_2, in (4.7), is sufficiently large (how large depends on ε). Moreover, the independent and identically distributed random variables $\hat{a}_{jk}$ (likewise, $\hat{b}_{jk}$, $\hat{c}_{jk}$, $\hat{d}_{jk}$, $\hat{e}_{jk}$, $\hat{a}_j$, $\hat{b}_j$, $\hat{c}_j$, $\hat{d}_j$, $\hat{e}_j$) satisfy $\mathbb{E}\,|\hat{a}_{jk}|^{C_3} \le C_4$ for all $1 \le j, k \le p$ and any given $C_3 > 0$ (provided C_2 is sufficiently large, depending on C_3) and C_4 does not depend on n. (Ensuring the *existence* of the moments $\mathbb{E}\,|\hat{a}_{jk}|^{C_3}$ requires Taylor expansion of denominators.)

Combining (A.27) and (A.28) we deduce that:

$$\begin{aligned}\Delta_{jk} &\equiv \mathbb{P}(Q_j^* \le y\hat{\sigma}_j, Q_k^* \le y\hat{\sigma}_k \big| \mathcal{X}) - \mathbb{P}(Q_j^* \le y\hat{\sigma}_j \big| \mathcal{X})\mathbb{P}(Q_k^* \le y\hat{\sigma}_k | \mathcal{X}) \\ &= \hat{G}_{0,jk}(y) + n^{-1/2}\hat{G}_{1,jk}(y) + n^{-1}\hat{G}_{2,jk}(y) + \mathcal{O}\left(n^{\varepsilon - \frac{3}{2}}\right) \qquad \text{(A.29)}\end{aligned}$$

where the remainder is of the stated size uniformly in the sense given in the previous paragraph, and

$$\begin{aligned}\hat{G}_{0,jk}(y) &= \mathbb{P}_{\hat{\rho}_{jk}}(\xi_1 \le y, \xi_2 \le y) - \mathbb{P}(\xi_j \le y)\mathbb{P}(\xi_k \le y), \\ \hat{G}_{1,jk}(y) &= (\hat{a}_{jk} + \hat{b}_{jk}y^2)\phi(y)^2 - (\hat{a}_j + \hat{b}_j y^2 + \hat{a}_k + \hat{b}_k y^2)\mathbb{P}(\xi_1 \le y)\phi(y), \\ \hat{G}_{2,jk}(y) &= \big\{\hat{c}_{jk}y + \hat{d}_{jk}y^3 + \hat{e}_{jk}y^5 - (\hat{a}_j + \hat{b}_j y^2)(\hat{a}_k + \hat{b}_k y^2)\big\}\phi(y)^2 \\ &\qquad - (\hat{c}_j y + \hat{d}_j y^3 + \hat{e}_j y^5 + \hat{c}_k y + \hat{d}_k y^3 + \hat{e}_k y^5)\mathbb{P}(\xi_1 \le y)\phi(y)\end{aligned}$$

An identical expansion holds for $\mathbb{P}(Q_j \le y, Q_k \le y) - \mathbb{P}(Q_j \le y)\mathbb{P}(Q_k \le y)$, except of course that the analogues of the coefficients a, b, c, d and e in the formulae for $\hat{G}_{0,jk}(y)$, $\hat{G}_{1,jk}(y)$, and $\hat{G}_{2,jk}(y)$) now all vanish identically, since Q_j and Q_k are independent. Comparing the coefficients it can be deduced that:

(i) The second moments $\mathbb{E}(\hat{a}^2)$, $\mathbb{E}(\hat{b}^2)$, $\mathbb{E}(\hat{c}^2)$, $\mathbb{E}(\hat{d}^2)$, and $\mathbb{E}(\hat{e}^2)$, with either single or (unequal) double subscripts, equal $\mathcal{O}(n^{-1})$ uniformly in the subscripts, and first moments also equal $\mathcal{O}(n^{-1})$, uniformly in the same sense.

(ii) Similarly, $\mathbb{E}(\hat{a}_{12}\hat{a}_{34}) = \mathcal{O}(n^{-2})$ and $\mathbb{E}(\hat{a}_{12}\hat{a}_{13}) = O(n^{-2})$, with analogous results holding for $\hat{b}$ instead of $\hat{a}$ and uniformly in other subscripts bearing the same relationships.

(A.30)

In view of part (i) of (A.30) the term $\hat{G}_{2,jk}(y)$ in (A.29) can be absorbed into the remainder, giving:

$$\Delta_{jk} = \hat{G}_{0,jk}(y) + n^{-1/2}\hat{G}_{1,jk}(y) + \mathcal{O}_p\left(n^{\varepsilon - \frac{3}{2}}\right) \qquad \text{(A.31)}$$

uniformly as before. Since the $\hat{a}_j$s, for $1 \le j \le p$, are independent and identically distributed, as too are the $\hat{b}_j$s, then by part (i) of (A.30),

$$\mathbb{E}\left|\sum_{j=1}^{p} \hat{a}_j\right| + \mathbb{E}\left|\sum_{j=1}^{p} \hat{b}_j\right| = \mathcal{O}_p(n^{-1}p + n^{-1/2}p^{1/2}) \tag{A.32}$$

Since $\hat{a}_{jk}$ is independent of $\hat{a}_{\ell m}$ if the integers j, k, ℓ, m are distinct then, by parts (i) and (ii) of (A.30),

$$\mathbb{E}\left|\sum\sum_{j,k:j\neq k} \hat{a}_{jk}\right|^2 + \mathbb{E}\left|\sum\sum_{j,k:j\neq k} \hat{b}_{jk}\right|^2 = \mathcal{O}(n^{-1}p^2 + b^{-2}p^4) \tag{A.33}$$

Combining (A.31)–(A.33) we deduce that if $\varepsilon \in (0, \frac{1}{2})$,

$$\begin{aligned}\sum\sum_{j,k:j\neq k} \Delta_{jk} &= \sum\sum_{j,k:j\neq k} \hat{G}_{0,jk}(y) + \mathcal{O}_p\Big\{n^{-1/2}(n^{-1/2}p + n^{-1}p) + n^{\varepsilon-\frac{3}{2}}\Big\} \\ &= \sum\sum_{j,k:j\neq k} \hat{G}_{1,jk}(y) + o_p(p + n^{-1}p^2)\end{aligned} \tag{A.34}$$

The first term on the second right-hand side of (A.34) represents the first term in an Edgeworth expansion of the quantity Δ at (A.21), and equals the first term on the righthand side of (A.26), i.e. $a_2(y)n^{-1}p(p-1) + o_p(p + n^{-1}p^2)$. Likewise, Δ is identical to the left-hand side of (A.34). Therefore we have established (A.26).

Removing the terms $\hat{\sigma}_j$ and $\hat{\sigma}_k$ in (A.21)

We describe their removal from the first term in the Edgeworth expansion; removal from subsequent terms is similar. As before, let $\mathbb{P}_\rho$ denote probability measure for a two-vector (ξ_1, ξ_2) that has a joint normal distribution with zero means, unit variances and correlation ρ, and interpret $\Psi(y_1, y_2) = \mathbb{P}_\rho(\xi_1 \le y_1, \xi_2 \le y_2)$ as $\mathbb{P}_\rho(\xi_1 \le y_1, \xi_2 \le y_2|X)$, where it is assumed that ξ_1 and ξ_2 are independent of $\mathcal{X}$. Let $\psi(y) = P(\xi_1 \le y)$. Writing $\hat{\sigma}_1 = 1 + \delta_j$ and $\hat{\sigma}_2 = 1 + \delta_k$ we have, by Taylor

expansion:

$$\begin{aligned}\Psi\left\{y(1+\delta_1), y(1+\delta_2)\right\} &= \Psi(y,y) + y\left\{\delta_1\Psi_{10}(y,y) + \delta_2\Psi_{01}(y,y)\right\} \\ &\quad + \tfrac{1}{2}y^2\left\{\delta_1\Psi_{20}(y,y) + \delta_2^2\Psi_{02}(y,y) + 2\delta_1\delta_2\Psi_{11}(y,y)\right\} \\ &\quad + \mathcal{O}_p\left(n^{\varepsilon-\frac{3}{2}}\right) \\ \psi\left\{y(1+\delta_1)\right\}\psi\left\{y(1+\delta_2)\right\} &= \psi(y)^2 + y(\delta_1+\delta_2)\psi(y)\psi'(y) \\ &\quad + \tfrac{1}{2}y^2\left\{(\delta_1^2+\delta_2^2)\phi(y)\phi''(y) + 2\delta_1\delta_2\psi'(y)^2\right\} \\ &\quad + \mathcal{O}_p\left(n^{\varepsilon-\frac{3}{2}}\right)\end{aligned}$$

and so

$$\begin{aligned}&\Psi\{y(1+\delta_1), y(1+\delta_2)\} - \psi\{y(1+\delta_1)\}\psi\{y(1+\delta_2)\} \\ &\quad = \Psi(y,y) - \psi(y)^2 + y(\delta_1+\delta_2)\{\Psi_{10}(y,y) - \psi(y)\psi'(y)\} \\ &\qquad + \tfrac{1}{2}y^2\left[(\delta_1^2+\delta_2^2)\{\Psi_{20}(y,y) - \psi(y)\psi''(y)\} + 2\delta_1\delta_2\{\Psi_{11}(y,y) - \psi'(y)^2\}\right] \\ &\qquad + \mathcal{O}_p\left(n^{\varepsilon-\frac{3}{2}}\right)\end{aligned}$$

where (here and below) the remainders) are of that size uniformly in unequal values of j, k in the range from 1 to p, and we have used the fact that $\Psi_{10}(y,y) = \Psi_{01}(y,y)$ and $\Psi_{20}(y,y) = \Psi_{02}(y,y)$. Now, $\Psi(y_1,y_2) = \psi(y_1)\psi(y_2) + \rho\psi'(y_1)\psi'(y_2) + \mathcal{O}(\rho^2)$, $\Psi_{10}(y,y) = \psi(y)\psi'(y) + \rho\psi'(y)\psi''(y) + \mathcal{O}(\rho^2)$, $\Psi_{20}(y,y) = \psi(y)\psi''(y) + \rho\psi'(y)\psi'''(y) + \mathcal{O}(\rho^2)$ and $\Psi_{11}(y,y) = \psi'(y)^2 + \rho\psi''(y)^2 + \mathcal{O}(\rho^2)$, whence it follows that:

$$\begin{aligned}&\Psi\{y(1+\delta_1),y(1+\delta_2)\} - \psi\{y(1+\delta_1)\}\psi\{y(1+\delta_2)\} \\ &\qquad = \Psi(y,y) - \phi(y)^2 + y(\delta_1+\delta_2)\rho\psi'(y)\psi''(y) \\ &\qquad\quad + \tfrac{1}{2}y^2\left\{(\delta_1^2+\delta_2)^2\rho\psi'(y)\psi'''(y) + 2\delta_1\delta_2\rho\psi''(y)^2\right\} + \mathcal{O}_p\left(n^{\varepsilon-\frac{3}{2}}\right) \\ &\qquad = \Psi(y,y) - \psi(y)^2 + y(\delta_1+\delta_2)\rho\psi'(y)\psi''(y) + \mathcal{O}_p\left(n^{\varepsilon-\frac{3}{2}}\right)\end{aligned}$$

Therefore, since $\hat{\sigma}_j - 1 = \frac{1}{2}(\hat{\sigma}_j^2 - 1) + \mathcal{O}_p(n^{\varepsilon-1})$, uniformly in j, then if we replace $\hat{\sigma}_j$ and $\hat{\sigma}_k$, in the definition of Δ at (A.21), by 1, the only change necessary to the expansion of Δ at (A.26) is to add a term equal to $\frac{1}{2}y\psi'(y)\psi''(y)$ multiplied by H_1, where

$$H_1 = \sum_{j,k:j\neq k}\sum(\hat{\sigma}_j^2 + \hat{\sigma}_k^2 - 2)\hat{\rho}_{jk}, \qquad H_2 = \sum_{j,k:j\neq k}\sum(\hat{\sigma}_j - 1)\hat{\rho}_{jk}$$

In order to prove that the added term is negligible it suffices to show that

$$H_1 = o_p(p + n^{-1}p^2), \tag{A.35}$$

and for this it is enough to prove the same property for H_2, and hence for

$$\begin{aligned} H_3 &= \sum_{j,k:j\neq k}\sum \left\{ \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2 - 1 \right\} \left\{ \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k) \right\} \\ &= H_4 + o_p(p + n^{-1}p^2), \end{aligned}$$

where

$$H_4 = \sum_{j,k:j\neq k}\sum H_{jk}, \qquad H_{jk} = \left\{ \frac{1}{n}\sum_{i=1}^{n}(U_{ij}^2 - 1) \right\} \left\{ \frac{1}{n}\sum_{i=1}^{n}(U_{ij}U_{ik}) \right\}$$

and $U_{ij} = (1 - \mathbb{E})X_{ij}$. By assumption the random variables U_{ij} and $U_{ij}^2 - 1$ have zero means, and the U_{ij}s are independent; and, in view of (4.7), we may assume that the U_{ij}s have eight finite moments. Using these properties it can be proved that $\mathbb{E}(V_4^2) = o(p^2 + n^{-2}p^4)$, and hence that (A.35) holds.

A.1.3 Proof of Theorem 3

Step 1: Preliminaries. Recall that $n = n^{(0)} + n^{(1)}$, and let $W_j = \sum_{r=0,1}(-1)^r W_j^{(r)}$ where

$$W_j^{(r)} = (\bar{X}_j^{(r)} - \mu_j^{(r)})^2 + 2(\mu_j^{(r)} - V_j)(\bar{X}_j^{(r)} - \mu_j^{(r)}) + \left(\mu_j^{(r)}\right)^2 - 2V_j\mu_j^{(r)}$$

In this notation, $D(V) = \sum_j W_j$; see (2.12). Minor changes to the proof of Berk–'s (1973) central limit theorem enable it to be shown that, if there exist constants $C, \varepsilon > 0$, not depending on p and such that

$$\mathbb{E}\,|(1 - \mathbb{E})n^{1/2}W_j|^{2+\varepsilon} \leq C, \quad \mathbb{E}\left\{ (1 - \mathbb{E})n^{1/2}\sum_{j=j_1+1}^{j_2} W_j \right\}^2 \leq C(j_2 - j_1), \tag{A.36}$$

for all $1 \leq j \leq p$ and $0 \leq j_1 < j_2 \leq p$; if, as assumed in Theorem 3, the quantity $m = m(p)$, introduced in (2.13), satisfies $m^{2+(2/\varepsilon)}/p \to 0$, where is as in (2.37); and if, as $p \to \infty$,

$$v_n^{(r)} \equiv \frac{1}{p}\left\{ (1 - \mathbb{E})(n^{(r)})^{1/2}\sum_{j=1}^{p} W_j^{(r)} \right\}^2 \to 4\tau^2, \tag{A.37}$$

for $r = 0, 1$, where τ is as in (2.14); then, defining $v_n = (n^{(0)})^{-1}v_n^{(0)} + (n^{(1)})^{-1}v_n^{(1)}$, $D(V)$ satisfies:

$$\sup_{-\infty<x<\infty} \left| \mathbb{P}\{D(V) \le x\} - \mathbb{P}\left[(pv_n)^{1/2}\mathcal{N} + \mathbb{E}\{D(V)\} \le x \right] \right| \tag{A.38}$$

where $\mathcal{N}$ denotes a random variable with the standard normal distribution. Result (2.39) is equivalent to (2.17) and so implies part (b) of the theorem. We shall complete our proof of Theorem 3 by establishing respectively, in steps 2 and 3 below, properties (2.37) and

$$\operatorname{var}\left(\sum_{j=1}^{p} W_j^{(r)} \right) \sim 4p\tau^2 (n^{(r)})^{-1}, \text{ for } r = 0, 1. \tag{A.39}$$

Note that (A.39) implies both (2.16) and (2.38).

Step 2: Proof of (2.37). The first part of (2.37) follows by direct calculation. Note that in (2.13) we assumed uniformly bounded $(4 + 2\varepsilon)$th moments of the variables $X_{ij}^{(r)} - \mathbb{E}(X_{ij}^{(r)})$.

Next we derive the second part of (2.37). Put $\xi_{jk} = \operatorname{cov}(X_{0j}, X_{0k})$ and observe that

$$\begin{aligned}
&\operatorname{cov}\left\{ (\bar{X}_j^{(r)} - \mu_j^{(r)}), (\bar{X}_k^{(r)} - \mu_k^{(r)}) \right\} \\
&\qquad = (n^{(r)})^{-4} \Big\{ n^{(r)}\, \mathbb{E}(X_{0j}^2 X_{0k}^2) + n^{(r)}(n^{(r)} - 1)\, \mathbb{E}(X_{0j}^2)\, \mathbb{E}(X_{0k}^2) \\
&\qquad\quad + 2n^{(r)}(n^{(r)} - 1)\xi_{jk}^2 - (n^{(r)})^2\, \mathbb{E}(X_{0j}^2)\, \mathbb{E}(X_{0k}^2) \Big\} \\
&\qquad = (n^{(r)})^{-3} \operatorname{cov}(X_{0j}^2, X_{0k}^2) + 2(n^{(r)})^{-2}\xi_{jk}^2
\end{aligned} \tag{A.40}$$

Therefore,

$$\begin{aligned}
\operatorname{var}\left\{ \sum_{j=j_1+1}^{j_2} (\bar{X}_j^{(r)} - \mu_j^{(r)})^2 \right\} &= \sum_{j=j_1+1}^{j_2} \sum_{k=j_1+1}^{j_2} \Big\{ \left(n^{(r)}\right)^{-3} \operatorname{cov}(X_{0j}^2, X_{0k}^2) \\
&\qquad + 2(n^{(r)})^{-2} \operatorname{cov}(X_{0j} m X_{0k}) \Big\} \\
&\le C_1(j_2 - j_1)
\end{aligned} \tag{A.41}$$

where the inequality follows from (2.13)(c).

Note too that, defining $\lambda_j = \mathbb{E}(V_j)$, we have:

$$\frac{1}{\eta}\sum_{j=j_1+1}^{j_2}|\xi_{jk}||\lambda_j-\mu_j^{(r)}| \le \left(\sum_{j=j_1+1}^{j_2}|\xi_{jk}|^q\right)^{1/q}\left(\sum_{j=j_1+1}^{j_2}\mathbb{I}(\lambda_j\ne\mu_j^{(r)})\right)^{(q-1)/q} = \mathcal{O}(\nu^{(q-1)/q})$$

where the inequality is Hölder's, and the identity is a consequence of the fact that, for either choice of r, $\mathbb{E}(V_j)$ and $\mathbb{E}(X_{ij}^{(r)})$ differ for at most ν values of j. Therefore, if $q \in [1,2]$ is as in (2.14)(c), then

$$\begin{aligned}
&\mathrm{var}\left\{\sum_{j=j_1+1}^{j_2}\left(V_j-\mu_j^{(r)}\right)\left(\bar{X}_j^{(r)}-\mu_j^{(r)}\right)\right\}\\
&\qquad=\sum_{j=j_1+1}^{j_2}\sum_{k=j_1+1}^{j_2}\left\{\left(\lambda_j\mu_j^{(r)}\right)\left(\lambda_k-\mu_k^{(r)}\right)+\xi_{jk}\right\}\xi_{jk}\\
&\qquad=\sum_{j=j_1+1}^{j_2}|\lambda_j\mu_j^{(r)}||\xi_{jk}|+\sum_{j=j_1+1}^{j_2}\sum_{k=j_1+1}^{j_2}\xi_{jk}^2\\
&\qquad\le C_2\left\{\nu^{(q-1)/q}\eta^2\sum_{j=j_1+1}^{j_2}\mathbb{I}(\lambda_j\ne\mu_j^{(r)})+\sum_{j=j_1+1}^{j_2}\sum_{k=j_1+1}^{j_2}|\xi_{jk}|^q\right\}\\
&\qquad\le C_3\{\eta^{(q-1)/q}\eta^2(j_2-j_1)+j_2-j_1\}\le C_4(j_2-j_1) \qquad\text{(A.42)}
\end{aligned}$$

where the second-last inequality follows from (2.15)(a) and (2.15)(b), and the last is a consequence of (2.15)(c). Similarly,

$$\mathrm{var}\left(\sum_{j=j_1+1}^{j_2}V_j\mu_j^{(r)}\right)\le\sum_{j=j_1+1}^{j_2}|\mu_j^{(r)}|\sum_{k=j_1+1}^{j_2}\mu_k^{(r)}||\xi_{jk}|\le C_5(j_2-j_1) \qquad\text{(A.43)}$$

The second part of (2.37) follows from (A.41)–(A.43).

Step 3: Proof of (A.39). Define $t^2 = \mathrm{var}(\sum_j W_j^{(r)})$,

$$t_1^2 = \mathrm{var}\left[\sum_{j=1}^{p}\left\{(\mu_j^{(r)} - V_j)(\bar{X}_j^{(r)} - \mu_j^{(r)}) - (V_j - \mathbb{E}\,V_j)\mu_j^{(r)}\right\}\right],$$

$$t_2^2 = \mathrm{var}\left\{\sum_{i=1}^{p}(\bar{X}_j^{(r)} - \mu_j^{(r)})^2\right\}, \qquad t_3^2 = \mathrm{var}\left\{\sum_{j=1}^{p}(\mu_j^{(r)} - V_j)(\bar{X}_j^{(r)} - \mu_j^{(r)})\right\}.$$

$$t_4^2 = \mathrm{var}\{\sum_j (V_j - \mathbb{E}\,V_j)\mu_j^{(r)}\}$$

Then $|t - 2t_1| \le t_2$ and $|t_1 - t_3| \le t_4$, so

$$|t - 2t_3| \le t_2 + t_4. \tag{A.44}$$

By (A.40),

$$\begin{aligned} t_2^2 &= \sum_{j=1}^{p}\sum_{k=1}^{p}\mathrm{cov}\left\{(\bar{X}_j^{(r)} - \mu_j^{(r)})^2, (\bar{X}_k^{(r)} - \mu_k^{(r)})^2\right\} \\ &= (n^{(r)})^{-3}\,\mathrm{var}\left(\sum_{j=1}^{p} X_{0j}^2\right) + 2(n^{(r)})^{-2}\sum_{j=1}^{p}\sum_{k=1}^{p}\xi_{jk}^2 \end{aligned}$$

The first and second parts of (2.14) imply, respectively, that

$$\sum_{j=1}^{p}\sum_{k=1}^{p}\xi_{jk}^2 = \mathcal{O}(p), \quad \mathrm{var}\left(\sum_{j=1}^{p} X_{0j}^2\right)$$

Therefore,

$$t_2^2 = \mathcal{O}(n^{-2}p) = o(n^{-1}p). \tag{A.45}$$

We claim too that, for $r = 0$ and 1,

$$s_1^2 \equiv \sum_{j=1}^{p}\sum_{k=1}^{p}\xi_{jk}\mu_j^{(r)}\mu_k^{(r)} = o(p/n) \tag{A.46}$$

Indeed, if $r = 0$ then (2.15)(a) implies $s_1^2 = 0$, and if $r = 1$ then the argument leading to (A.42) can be used to prove that:

$$\begin{aligned} s_1^2 &\le \eta^2 \sum_{j=1}^{p} \mathbb{I}(\mu_j^{(1)} \neq 0) \sum_{k=j_1+1}^{j_2} |\xi_{jk}| \mathbb{I}(\mu_k^{(1)} \neq 0) \\ &\le \eta^2 \sum_{j=1}^{p} \mathbb{I}(\mu_j^{(1)} \neq 0) \left(\sum_{k=j_1+1}^{j_2} |\xi_{jk}|^q \right)^{1/q} \left\{ \sum_{k=j_1+1}^{j_2} \mathbb{I}\left(\mu_k^{(1)} \neq 0\right) \right\}^{(q-1)/q} \\ &= \mathcal{O}(\eta^{2-(1/q)} \eta^2) = o(p/n), \end{aligned}$$

where the first inequality follows from (2.15)(b), the second from Hölder's inequality, and the first and second identities are consequences of (2.14)(c) and (2.15)(c), respectively. Define $\lambda_j = \mathbb{E}(V_j)$ and, which, in view of the first part of (2.14), equals $p\tau^2 + o(p)$. Combining the latter property and (A.46) we deduce that:

$$\begin{aligned} n^{(r)} t_3^2 &= \sum_{j=1}^{p} \sum_{k=1}^{p} \mathbb{E}\left\{ (V_j - \mu_j^{(r)})(V_k - \mu_k^{(r)}) \right\} \xi_{jk} \\ &= \sum_{j=1}^{p} \sum_{k=1}^{p} \left\{ (V_j - \mu_j^{(r)})(V_k - \mu_k^{(r)}) + \xi_{jk} \right\} \xi_{jk} \\ &= s_1^2 + s_2^2 = p\tau^2 + o(p) \end{aligned} \tag{A.47}$$

Moreover $t_4^2 = \mathrm{var}(\sum_j X_{0j} \mu_j^{(r)})$. From the latter result, (A.44), (A.45) and (A.47) we deduce that

$$t = 2t_3 + \mathcal{O}(t_2 + t_4) = 2\tau (p/n^{(r)})^{1/2} = o\big\{(p/n)^{1/2}\big\},$$

which implies (A.39).

A.1.4 Proof of Theorem 4

Observe that

$$D^*(V) = D(V) + D_1^*(V) \tag{A.48}$$

where $D(V)$ is as at (2.12),

$$D_1^*(V) = \sum_{j=1}^{p} \left[\left\{ (\bar{U}_j^{*(0)})^2 + 2\bar{U}_j^{*(0)}(\bar{X}_j^{(0)} - V_j) \right\} - \left\{ (\bar{U}_j^{*(1)})^2 + 2\bar{U}_j^{*(1)}(\bar{X}_j^{(1)} - V_j) \right\} \right] \tag{A.49}$$

and $\bar{U}_j^{*(r)} = (n^{(r)})^{-1} \sum_{1 \le i \le n^{(r)}} (\bar{X}_j^{*(0)} - V_j)$. (Note that $\mathbb{E}(\bar{U}_j^{*(r)} | \mathcal{X}) = 0$.) Since the resamples $\mathcal{X}^*(0)$ and $\mathcal{X}^*(1)$ were drawn independently, conditional on $\mathcal{X}$, hence:

$$\begin{aligned} \operatorname{var}\{D_1^*(V) | \mathcal{X}, V\} &= \sum_{r=1}^{1} \sum_{j=1}^{p} \sum_{k=1}^{p} \operatorname{cov} \left\{ (\bar{U}_j^{*(r)})^2 + 2\bar{U}_j^{*(r)}(\bar{X}_j^{(r)} - V_j), \right. \\ &\qquad \left. (\bar{U}_k^{*(r)})^2 + 2\bar{U}_k^{*(r)}(\bar{X}_k^{(r)} - V_k) \right\} \\ &= \sum_{r=1}^{1} \sum_{j=1}^{p} \sum_{k=1}^{p} \left[\operatorname{cov} \left\{ (\bar{U}_j^{*(r)})^2, (\bar{U}_k^{*(r)})^2 | \mathcal{X} \right\} \right. \\ &\quad + (\bar{X}_j^{(r)} - V_j) \operatorname{cov} \left\{ \bar{U}_j^{*(r)}, (\bar{U}_k^{*(r)})^2 \right\} \\ &\quad \left. + 4(\bar{X}_j^{(r)} - V_j)(\bar{X}_k^{(r)} - V_k) \operatorname{cov}(\bar{U}_j^{*(r)}, \bar{U}_k^{*(r)} | \mathcal{X}) \right] \end{aligned} \tag{A.50}$$

Note that

$$\hat{\rho}_{jk}^{(r)} \equiv n^{(r)} \operatorname{cov}(\bar{U}_j^{*(r)}, \bar{U}_k^{*(r)} | \mathcal{X}) = \frac{1}{n^{(r)}} \sum_{i=1}^{n^{(r)}} (X_{ij}^{(r)} - \bar{X}_j^{(r)})(X_{ik}^{(r)} - \bar{X}_k^{(r)})$$

$$\hat{\beta}_{jk}^{(r)} \equiv (n^{(r)})^2 \operatorname{cov}\left\{ \bar{U}_j^{*(r)}, \left(\bar{U}_k^{*(r)}\right)^2 \middle| \mathcal{X} \right\} = \frac{1}{n^{(r)}} \sum_{i=1}^{n^{(r)}} (X_{ij}^{(r)} - \bar{X}_j^{(r)})(X_{ik}^{(r)} - \bar{X}_k^{(r)})^2$$

$$\begin{aligned}
\hat{\gamma}_{jk}^{(r)} &\equiv \operatorname{cov}\left\{ \left(\bar{U}_j^{*(r)}\right)^2, \left(\bar{U}_k^{*(r)}\right)^2 \middle| \mathcal{X} \right\} \\
&= \mathbb{E}\left\{ \frac{1}{(n^{(r)})^4} \sum_{i_1} \cdots \sum_{i_4} (X_{i_1 j}^{(r)} - \bar{X}_j^{(r)})(X_{i_2 j}^{(r)} - \bar{X}_j^{(r)}) \right. \\
&\qquad \left. \times (X_{i_3 j}^{(r)} - \bar{X}_j^{(r)})(X_{i_4 j}^{(r)} - \bar{X}_j^{(r)}) \middle| \mathcal{X} \right\} \\
&\qquad - \mathbb{E}\left\{ \left(\bar{U}_j^{*(r)}\right)^2 \middle| \mathcal{X} \right\} \mathbb{E}\left\{ \left(\bar{U}_k^{*(r)}\right)^2 \middle| \mathcal{X} \right\} \\
&= \frac{1}{(n^{(r)})^4} \sum_{i1} \sum_{i2} \left\{ (X_{i_1 j}^{(r)} - \bar{X}_j^{(r)})^2 (X_{i_2 j}^{(r)} - \bar{X}_j^{(r)})^2 \right. \\
&\quad \left. + 2(X_{i_1 j}^{(r)} - \bar{X}_j^{(r)})(X_{i_1 k}^{(r)} - \bar{X}_k^{(r)}) \cdot (X_{i_2 j}^{(r)} - \bar{X}_j^{(r)})(X_{i_2 k}^{(r)} - \bar{X}_k^{(r)}) \right\} \\
&\quad - \frac{2}{(n^{(r)})^4} \sum_{i=1}^{n^{(r)}} (X_{ij}^{(r)} - \bar{X}_j^{(r)})^2 (X_{ik}^{(r)} - \bar{X}_k^{(r)})^2 - (n^{(r)})^{-2} (\hat{\sigma}_j^{(r)} \hat{\sigma}_k^{(r)})^2 \\
&= 2(n^{(r)})^{-2} \hat{\rho}_{jk}^2 - 2(n^{(r)})^{-3} \hat{\alpha}_{jk}^{(r)}
\end{aligned} \tag{A.51}$$

where $(\hat{\sigma}_j^{(r)})^2 = \hat{\rho}_{jj}^{(r)}$ and

$$\hat{\alpha}_{jk}^{(r)} = \frac{1}{n^{(r)}} \sum_{i=1}^{n^{(r)}} (X_{ij}^{(r)} - \bar{X}_j^{(r)})^2 (X_{ik}^{(r)} - \bar{X}_k^{(r)})^2$$

Observe too that

$$(n^{(r)})^4 \sum_{j=1}^{p}\sum_{k=1}^{p}\left\{(n^{(r)})^{-2}\hat{\rho}_{jk}^2 - (n^{(r)})^{-3}\hat{\alpha}_{jk}^{(r)}\right\}$$
$$= \sum_{i_1\neq i_2}\sum_{j=1}^{p}\sum_{k=1}^{p}(X_{i_1 j}^{(r)} - \bar{X}_j^{(r)})(X_{i_1 k}^{(r)} - \bar{X}_k^{(r)})(X_{i_2 j}^{(r)} - \bar{X}_j^{(r)})(X_{i_2 k}^{(r)} - \bar{X}_k^{(r)})$$
$$= \sum_{i_1\neq i_2}\sum\left\{\sum_{j=1}^{p}(X_{i_1 j}^{(r)} - \bar{X}_j^{(r)})(X_{i_2 k}^{(r)} - \bar{X}_k^{(r)})\right\}^2 \tag{A.52}$$

Combining (A.51) and (A.52) we deduce that

$$\sum_{j=1}^{p}\sum_{k=1}^{p}\hat{\gamma}_{jk}^{(r)} = \frac{2}{(n^{(r)})^4}\sum_{i_1\neq i_2}\sum\left(\hat{G}_{i_1 i_2}^{(r)}\right)^2, \tag{A.53}$$

where

$$\hat{G}_{i_1 i_2}^{(r)} = \sum_{j=1}^{p}(X_{i_1 j}^{(r)} - \bar{X}_j^{(r)})(X_{i_2 j}^{(r)} - \bar{X}_j^{(r)}), \tag{A.54}$$

It follows from (A.48) and (A.50) that

$$\mathrm{var}\{D^*(V)|\mathcal{X}, V\} = \mathrm{var}\left\{D_1^*(V)|\mathcal{X}, V\right\}$$
$$= \sum_{r=0}^{1}\sum_{j=1}^{p}\sum_{k=1}^{p}\left\{\hat{\gamma}_{jk}^{(r)} + 4\left(n^{(r)}\right)^{-2}(\bar{X}_j^{(r)} - V_j)\hat{\beta}_{jk}^{(r)}\right.$$
$$\left. + 4\left(n^{(r)}\right)^{-1}\left(\bar{X}_j^{(r)} - V_j\right)\left(\bar{X}_k^{(r)} - V_k\right)\hat{\rho}_{jk}^{(r)}\right\} \tag{A.55}$$

For $s = 1, 2$ define

$$\hat{H}_{is}^{(r)} = \sum_{j=1}^{p}(\bar{X}_j^{(r)} - V_j)^{2-s}(X_{ij}^{(r)} - \bar{X}_j^{(r)})^s. \tag{A.56}$$

Then,

$$\sum_{j=1}^{p}\sum_{k=1}^{p}(\bar{X}_j^{(r)} - V_j)\hat{\beta}_{jk}^{(r)} = \frac{1}{n^{(r)}}\sum_{i=1}^{n^{(r)}}\hat{H}_{i1}^{(r)}\hat{H}_{i2}^{(r)} \tag{A.57}$$

$$\sum_{j=1}^{p}\sum_{k=1}^{p}(\bar{X}_j^{(r)} - V_j)(\bar{X}_k^{(r)} - V_k)\hat{\rho}_{jk} = \frac{1}{n^{(r)}}\sum_{i=1}^{n^{(r)}}\left(\hat{H}_{i1}^{(r)}\right)^2. \tag{A.58}$$

Combining (A.53) and (A.55)–(A.58) we deduce that

$$\mathrm{var}(D^*(V)|\mathcal{X}, V) = \sum_{r=1}^{1}\left[\frac{2}{(n^{(r)})^4}\sum\sum_{i_1\neq i_2}(\hat{G}^{(r)}_{i_1 i_2})^2 + \frac{4}{(n^{(r)})^2}\sum_{i=1}^{n^{(r)}}\left\{\frac{1}{n^{(r)}}\hat{H}^{(r)}_{i1}\hat{H}^{(r)}_{i2} + (\hat{H}^{(r)}_{i1})^2\right\}\right] \tag{A.59}$$

Let $Z^{(r)}_{ij} = X^{(r)}_{ij} - \mathbb{E}(X^{(r)}_{ij})$, $\bar{Z}^{(r)}_j = (n^{(r)})^{-1}\sum_i Z^{(r)}_{ij}$ and $\sigma_j^2 = \mathrm{var}(X_{0j})$. Noting the definition of $\hat{H}^{(r)}_{is}$ at (A.56) we deduce that, for each j in the range $1 \le j \le p$,

$$\mathbb{E}(\hat{H}^{(r)}_{i1}) = \sum_{j=1}^{p}\mathbb{E}\{\bar{Z}^{(r)}_j(Z^{(r)}_{ij} - \bar{Z}^{(r)}_j)\} = \sum_{j=1}^{p}\left\{\sigma_j^2(n^{(r)})^{-1} - \sigma_j^2(n^{(r)})^{-1}\right\} = 0, \tag{A.60}$$

and so, using the m-dependence property,

$$\mathbb{E}\{(\hat{H}^{(r)}_{i1})^2\} = \mathrm{var}(\hat{H}^{(r)}_{i1}) = \mathcal{O}(mp). \tag{A.61}$$

Similarly, since by assumption in Theorem 4 the components of X_0 have eight finite moments, $\mathbb{E}\{(\hat{H}^{(r)}_{i1})^4\} = \mathcal{O}\{(mp)^2\}$. Therefore,

$$\begin{aligned}\mathrm{var}\left\{\frac{1}{n^{(r)}}\sum_{i=1}^{n^{(r)}}\left(\hat{H}^{(r)}_{i1}\right)^2\right\} &= \frac{1}{n^{(r)}}\sum_{i=1}^{n^{(r)}}\mathrm{var}\left\{\left(\hat{H}^{(r)}_{i1}\right)^2\right\} \\ &\le \frac{1}{(n^{(r)})^2}\sum_{i=1}^{n^{(r)}}\mathbb{E}\left\{\left(\hat{H}^{(r)}_{i1}\right)^4\right\} \\ &= \mathcal{O}\left\{(n^{(r)})^{-1}(mp)^2\right\}\end{aligned} \tag{A.62}$$

Combining (A.61) and (A.62)

$$\frac{1}{(n^{(r)})^2}\sum_{i=1}^{n^{(r)}}(\hat{H}^{(r)}_{i1})^2 = \mathcal{O}_p(mp/n) \tag{A.63}$$

Analogously using the property that $\mathbb{E}\{(\bar{X}^{(r)}_j - V_j)(X^{(r)}_{ij} - \bar{X}^{(r)}_j)\} = 0$ (see (A.60)) and the m-dependence of X_0, it can be proved that $\mathbb{E}(\hat{H}^{(r)}_{i1}\hat{H}^{(r)}_{i2}) = \mathrm{cov}(\hat{H}^{(r)}_{i1}\hat{H}^{(r)}_{i2}) =$

$\mathcal{O}(mp)$ and similarly,

$$\begin{aligned}\mathbb{E}\{(\hat{H}_{i1}^{(r)}\hat{H}_{i2}^{(r)})^2\} &= \sum_{j_1=1}^{p}\cdots\sum_{j_4=1}^{p}\left(X_{ij_1}^{(r)}-\bar{X}_{j_1}\right)^2\left(X_{ij_2}^{(r)}-\bar{X}_{j_2}\right)^2 \\ &\qquad\times\left(\bar{X}_{j_3}-V_{j_3}\right)\left(X_{ij_3}^{(r)}-\bar{X}_{j_3}\right)\left(\bar{X}_{j_4}-V_{j_4}\right)\left(X_{ij_4}^{(r)}-\bar{X}_{j_4}\right) \\ &= \mathcal{O}(mp^3)\end{aligned}$$

Therefore,

$$\frac{1}{(n^{(r)})^3}\sum_{i=1}^{n^{(r)}}\hat{H}_{i1}^{(r)}\hat{H}_{i2}^{(r)} = \mathcal{O}_p\big\{(mp/n^2)+(mp^3/n^5)^{1/2}\big\}. \tag{A.64}$$

Combining (A.63) and (A.64) we find that

$$\frac{1}{n^{(r)}}\sum_{i=1}^{n^{(r)}}\left\{\hat{H}_{i1}^{(r)}\hat{H}_{i2}^{(r)}+(\hat{H}_{i1}^{(r)})^2\right\} = \mathcal{O}_p\left\{(mp/n)+(mp^3/n^5)^{1/2}\right\} \tag{A.65}$$

Noting the definitions of $\hat{G}_{i_1i_2}^{(r)}$ at (A.54) and recalling the definitions of $Z_{ij}^{(r)}$ and $\bar{Z}_j^{(r)}$ we see that if $i_1 \neq i_2$ then

$$\mathbb{E}(\hat{G}_{i_1i_2}^{(r)}) = -\sum_{j=1}^{p}\mathbb{E}\left\{Z_{i_1j}^{(r)}\bar{Z}_j^{(r)}+Z_{i_2j}^{(r)}\bar{Z}_j^{(r)}-(\bar{Z}_j^{(r)})^2\right\} = -\frac{1}{n^{(r)}}\sum_{j=1}^{p}\mathbb{E}(X_{0j}^2). \tag{A.66}$$

Since the process X_0 in (2.13) is m-dependent then $\mathrm{var}(\hat{G}_{i_1i_2}^{(r)}) = \mathcal{O}(mp)$. Therefore,

$$\begin{aligned}\sum_{j,k:j\neq k}\sum\mathbb{E}(\hat{G}_{i_1i_2}^{(r)})^2 &= \sum_{j,k:j\neq k}\sum\left\{(\mathbb{E}\,\hat{G}_{i_1i_2}^{(r)})^2+\mathrm{var}(\hat{G}_{i_1i_2}^{(r)})\right\} \\ &= \sum_{j,k:j\neq k}\sum\left[\frac{1}{(n^{(r)})^2}\left(\sum_{j=1}^{p}\mathbb{E}\,X_{0j}^2\right)^2+\mathcal{O}(mp)\right] \\ &= \{1+o_p(1)\}\left(\sum_{j=1}^{p}\mathbb{E}\,X_{0j}^2\right)^2+\mathcal{O}\big\{m(n^{(r)})^2p\big\}. \end{aligned}\tag{A.67}$$

Furthermore, noting that, for $i_1 \neq i_2$, $-n^{(r)}p^{-1}\,\mathbb{E}(\hat{G}_{i_1i_2}^{(r)}) = p^{-1}\sum_j\mathbb{E}(X_{01}^2)$ (see (A.66)), which is bounded away from zero and is therefore denoted below by $C(p)$

and writing $(1-\mathbb{E})R = R - \mathbb{E}(R)$ for any random variable R, we have:

$$\begin{aligned}
\operatorname{var}\left\{\left(\hat{G}^{(r)}_{i_1 i_2}\right)^2\right\} &= \left(\frac{p}{n^{(r)}}\right)^2 \operatorname{var}\left[\left\{1 - \frac{n^{(r)}}{C(p)p}\sum_{j=1}^{p}(1-\mathbb{E})\left(X^{(r)}_{i_1 j} - \bar{X}^{(r)}_j\right)\left(X^{(r)}_{i_2 j} - \bar{X}^{(r)}_j\right)\right\}^2\right] \\
&= \mathcal{O}\Bigg\{\left(\frac{p}{n^{(r)}}\right)^2 \Bigg(\left(\frac{n^{(r)}}{p}\right)^2 \operatorname{var}\left\{\sum_{j=1}^{p}\left(X^{(r)}_{i_1 j} - \bar{X}^{(r)}_j\right)\left(X^{(r)}_{i_2 j} - \bar{X}^{(r)}_j\right)\right\} \\
&\qquad + \left(\frac{n^{(r)}}{p}\right)^4 \mathbb{E}\left[\left\{(1-\mathbb{E})\sum_{j=1}^{p}\left(X^{(r)}_{i_1 j} - \bar{X}^{(r)}_j\right)\left(X^{(r)}_{i_2 j} - \bar{X}^{(r)}_j\right)\right\}^4\right]\Bigg)\Bigg\} \\
&= \mathcal{O}\left[\left(\frac{p}{n^{(r)}}\right)^2 \left\{\left(\frac{n^{(r)}}{p}\right)^2 mp + \left(\frac{n^{(r)}}{p}\right)^4 (mp)^2\right\}\right] \\
&= \mathcal{O}\left\{mp + (mn^{(r)})^2\right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\operatorname{var}\left\{\sum\sum_{i_1 \neq i_2}(\hat{G}^{(r)}_{i_1 i_2})^2\right\} &= \mathcal{O}\left[(n^{(r)})^2 \operatorname{var}\left\{(\hat{G}^{(r)}_{i_1 i_2})^2\right\}\right] \\
&= \mathcal{O}\left[(n^{(r)})^2\left\{mp + (mn^{(r)})^2\right\}\right]. \qquad \text{(A.68)}
\end{aligned}$$

Together (A.67), (A.68) and the fact that $mn^2/p \to 0$ (a consequence of the assumption that $mn^3/p \to 0$), imply that

$$\begin{aligned}
\sum\sum_{j,k:j\neq k}(\hat{G}^{(r)}_{i_1 i_2})^2 &= \{1 + o_p(1)\}\left(\sum_{j-1}^{p}\mathbb{E}\, X^2_{0j}\right) + \mathcal{O}_p\left\{mp(n^{(r)})^2\right\} \\
&= \{1 + o_p(1)\}\left(\sum_{j-1}^{p}\mathbb{E}\, X^2_{0j}\right) \qquad \text{(A.69)}
\end{aligned}$$

From (A.59), (A.65) and (A.69) we deduce that, since $mn^3/p \to 0$,

$$\operatorname{var}\{D^*(V)|\mathcal{X}, V\} = \{1 + o_p(1)\}2\left(\sum_{j-1}^{p}\mathbb{E}\, X^2_{0j}\right)\sum_{r=0}^{1}(n^{(r)})^{-4} \qquad \text{(A.70)}$$

Result (2.18) follows from (2.16) and (A.70).

A.1.5 Proof of Theorem 5.

Step 1: Proof of (2.19). As in the proof of Theorem 1 we assume, for simplicity, that $p = bk$, where b and k are positive integers. Property (A.48) holds as before, for the definition of at (A.49), but we rewrite that formula to express the block structure of $D_1^*(V)$:

$$D_1^*(V) = \sum_{j=1}^{k} A_j^*(V), \tag{A.71}$$

where

$$A_j^*(V) = \sum_{t=1}^{b} \Bigg[\left(\bar{U}^{*(0)}_{(j-1)b+t} \right)^2 + 2\bar{U}^{*(0)}_{(j-1)b+t} \left(X^{(0)}_{(j-1)b+t} - V_{(j-1)b+t} \right) - \left\{ \left(\bar{U}^{*(1)}_{(j-1)b+t} \right)^2 + 2\bar{U}^{*(1)}_{(j-1)b+t} \left(X^{(1)}_{(j-1)b+t} - V_{(j-1)b+t} \right) \right\} \Bigg] \tag{A.72}$$

The independence of blocks, conditional on $\mathcal{X}$ and on V ensures that the variables $A_j^*(V)$ are independent, conditional on $\mathcal{X}$ and V. Therefore, instead of (A.59),

$$\mathrm{var}\{D_1^*(V)|\mathcal{X}, V\} = \sum_{r=0}^{1} \sum_{j=1}^{k} \mathrm{var}\{A_j^{*(r)}(V)|\mathcal{X}, V\}, \tag{A.73}$$

where

$$\mathrm{var}\{A_j^{*(r)}(V)|\mathcal{X}, V\} = \sum_{t_1}^{b} \sum_{t_2}^{b} \mathrm{cov} \left\{ \left(\bar{U}^{*(r)}_{(j-1)b+t_1} \right)^2 + 2\bar{U}^{*(r)}_{(j-1)b+t_1} \left(\bar{X}^{(r)}_{(j-1)b+t_1} - V_{(j-1)b+t_1} \right), \left(\bar{U}^{*(r)}_{(j-1)b+t_2} \right)^2 + 2\bar{U}^{*(r)}_{(j-1)b+t_2} \left(\bar{X}^{(r)}_{(j-1)b+t_2} - V_{(j-1)b+t_2} \right) \Bigg| \mathcal{X} \right\}$$

$$
\begin{aligned}
&= \sum_{t_1}^{b}\sum_{t_2}^{b} \operatorname{cov}\left\{ \left(\bar{U}^{*(r)}_{(j-1)b+t_1}\right)^2, \left(\bar{U}^{*(r)}_{(j-1)b+t_2}\right)^2 \middle| \mathcal{X} \right\} \\
&\quad + 4\left(\bar{X}^{(r)}_{(j-1)b+t_1} - V_{(j-1)b+t_1}\right) \operatorname{cov}\left\{ \bar{U}^{*(r)}_{(j-1)b+t_1}, \left(\bar{U}^{*(r)}_{(j-1)b+t_2}\right)^2 \middle| \mathcal{X} \right\} \\
&\quad + 4\left(\bar{X}^{(r)}_{(j-1)b+t_1} - V_{(j-1)b+t_1}\right)\left(\bar{X}^{(r)}_{(j-1)b+t_2} - V_{(j-1)b+t_2}\right) \\
&\qquad \times \operatorname{cov}\left\{ \bar{U}^{*(r)}_{(j-1)b+t_1}, \bar{U}^{*(r)}_{(j-1)b+t_2} \middle| \mathcal{X} \right\} \\
&= \sum_{r=0}^{1}\left[\frac{2}{(n^{(r)})^4} \sum_{i_1\neq i_2}\sum (\hat{G}^{(r)}_{i_1 i_2})^2 + \frac{4}{(n^{(r)})^2}\sum_{i=1}^{n^{(r)}} \left\{ \frac{1}{n^{(r)}} (\hat{H}^{(r)}_{ji_1}\hat{H}^{(r)}_{ji_2} + (\hat{H}^{(r)}_{ji_1})^2 \right\} \right]
\end{aligned}
\tag{A.74}
$$

with

$$
\begin{aligned}
\hat{G}^{(r)}_{i_1 i_2} &= \sum_{t=1}^{b} (X^{(r)}_{i_1,(j-1)b+1} - \bar{X}^{(r)}_{(j-1)b+1})(X^{(r)}_{i_2,(j-1)b+t} - \bar{X}^{(r)}_{(j-1)b+t}), \\
\hat{H}^{(r)}_{ji_s} &= \sum_{j=1}^{p} (\bar{X}^{(r)}_{j} - V_{(j-1)b+t})^{2-s} (X^{(r)}_{i,(j-1)b+t} - \bar{X}^{(r)}_{(j-1)b+t})^{s}.
\end{aligned}
$$

Define $Y_{ij} = X_{ij} - \mathbb{E}(X_{ij})$ and $Z_j = V_j - \mathbb{E}(V_j)$. Then,

$$
\mathbb{E}(\hat{G}^{(r)}_{i_1 i_2})^2 = \operatorname{var}\left(Y^{(r)}_{1,(j-1)b+t} Y^{(r)}_{2,(j-1)b+t} \right) + o(b) = \mathcal{O}(b),
$$

and therefore

$$
\frac{1}{(n^{(r)})^4} \sum_{i_1\neq i_2}\sum (\hat{G}^{(r)}_{i_1 i_2})^2 = \mathcal{O}_p(b/n^2). \tag{A.75}
$$

Note too that

$$
\mathbb{E}\left\{ \left(\hat{H}^{(r)}_{ji_1}\right)^2 \right\} = \mathbb{E}\left(Z_{j-1)b+t} Y^{(r)}_{i,(j-1)b+t} \right)^2 + o(b) = b\tau^2 + o(b),
$$

and

$$
\operatorname{var}\left\{ \sum_{i=1}^{n^{(r)}} \left(\hat{H}^{(r)}_{ji_1}\right)^2 \right\} = o(b^2 n^2),
$$

whence it follows that

$$
\frac{1}{(n^{(r)})^2} \sum_{i=1}^{n^{(r)}} \left(\hat{H}^{(r)}_{ji_1}\right)^2 = \frac{b\tau^2}{n^{(r)}} + o_p(b/n^{(r)}). \tag{A.76}
$$

More simply,

$$\frac{4}{(n^{(r)})^3}\sum_{i=1}^{n^{(r)}}\hat{H}^{(r)}_{ji_1}\hat{H}^{(r)}_{ji_2} = o_p(b/n) \tag{A.77}$$

Together, (A.74)–(A.77) imply that

$$\mathrm{var}\{D_1^*(V)|\mathcal{X}, V\} = 4p\tau^2\sum_{r=1}^{1}(n^{(r)})^{-1} + o(p/n). \tag{A.78}$$

Properties (2.16) and (A.78) together imply (2.19).

Step 2: Proof of (2.20). Note first that, by (A.48), (A.71) and (A.72),

$$(1-\mathbb{E}_{\mathcal{X}})D^*(V) = \sum_{j=1}^{k}(1-\mathbb{E}_{\mathcal{X}})A_j^*(V) = \sum_{r=0}^{1}(-1)^r\sum_{j=1}^{k}B_j^{*(r)}, \tag{A.79}$$

where the random variables

$$B_j^{*(r)} = \sum_{t=1}^{b}\Big\{(1-\mathbb{E}_{\mathcal{X}})\big(\bar{U}^{*(r)}_{(j-1)b+t}\big)^2 + 2\left(\bar{X}^{(r)}_{(j-1)b+t} - V_{(j-1)b+t}\right)(1-\mathbb{E}_{\mathcal{X},V})\bar{U}^{*(r)}_{(j-1)b+t}\Big\},$$

for $1 \le j \le k$ and $r = 0, 1$, are independent and have zero mean, both statements holding conditional on $\mathcal{X}$, and $\mathbb{E}_{\mathcal{X}}$ and $\mathbb{E}_{\mathcal{X},V}$ denote expectation conditional on $\mathcal{X}$, and expectation conditional on both $\mathcal{X}$ and V , respectively. The central limit theorem that asserts that the distribution function $\hat{F}$, say, of

$$\frac{(1-\mathbb{E}_{\mathcal{X}})D^*(V)}{\sqrt{\mathrm{var}\{D^*(V)|\mathcal{X}, V\}}},$$

conditional on $\mathcal{X}$, satisfies $\sup_x|\hat{F}(x) - \Phi(x)| \to 0$ in probability, where Φ is the standard normal distribution, follows from (A.78) and (A.79) if it is proved that

$$(p/n)^{-2}\sum_{j=1}^{k}\mathbb{E}\left\{(B_j^{*(r)})^4|\mathcal{X}, V\right\} \to 0 \tag{A.80}$$

in probability as $p \to \infty$. In view of (2.17) and (2.19) this is enough to give (2.20). Property (A.80) can be derived using arguments from the proof of Theorem 4.