Improved Performance and Stability of the Knockoff Filter and an Approach to Mixed Effects Modeling of Sequentially Randomized Trials

by

Brook Luers

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2020

Doctoral Committee:

Professor Kerby Shedden, Chair
Associate Professor Daniel Almirall
Assistant Professor Yuekai Sun
Assistant Professor Zhenke Wu

Brook Luers

luers@umich.edu

ORCID iD: 0000-0002-7847-5044

ACKNOWLEDGMENTS

Chapter IV is adapted from Luers et al. (2019). The authors of Luers et al. (2019) thank Donald Hedeker for helpful comments. Research for Luers et al. (2019) was supported by the following National Institutes of Health grants: R01DA039901 (Nahum-Shani and Almirall), P50DA039838 (Almirall), R01HD073975 (Kasari and Almirall), R01MH114203 (Almirall), and R01DA047279 (Almirall).

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

The knockoff filter is a variable selection technique for linear regression with finite-sample control of the regression false discovery rate (FDR). The regression FDR is the expected proportion of selected variables which, in fact, have no effect in the regression model. The knockoff filter constructs a set of synthetic variables which are known to be irrelevant to the regression and, by serving as negative controls, help identify relevant variables. The first two thirds of this thesis describe tradeoffs between power and collinearity due to tuning choices in the knockoff filter and provides a stabilization method to reduce variance and improve replicability of the selected variable set using the knockoff filter. The final third of this thesis develops an approach for mixed modeling and estimation for sequential multiple assignment randomized trials (SMARTs). SMARTs are an important data collection tool for informing the construction of dynamic treatment regimens (DTRs), which use cumulative patient information to recommend specific treatments during the course of an intervention. A common primary aim in a SMART is the marginal mean comparison between two or more of the DTRs embedded in the trial, and the mixed modeling approach is developed for these primary aim comparisons based on a continuous, longitudinal outcome. The method is illustrated using data from a SMART in autism research.

CHAPTER I

Introduction

The knockoff filter is a variable selection technique for linear regression with finite-sample control of the regression false discovery rate (FDR), the expected proportion of selected variables which, in fact, have no effect in the regression model. The knockoff filter constructs a set of synthetic variables which are known to be irrelevant to the regression and, by serving as negative controls, help identify relevant variables. By fitting an augmented regression model to the set of observed and knockoff variables, the estimated effect of each observed variable can be compared to that of a knockoff variable whose true effect is known to be zero. Variables are selected based on this measure of variable importance, which will be small when the effect of an observed variable is indistinguishable from noise (the estimated effect of a knockoff variable). The first two thirds of this thesis focuses on tuning parameter choices and a stabilization technique for the knockoff filter in the low-dimensional, design-based regression setting, in which there are fewer variables than observations and the design matrix is treated as fixed.

Chapter II focuses on improvements in a a set of tuning parameters in the knockoff filter which control correlations between each paired knockoff and original covariate. The existing method of choosing these tuning parameters seeks to minimize the correlations between each original covariate and its knockoff. However, unless the original

covariates are nearly orthogonal, this tuning approach leads to an augmented design matrix (the matrix of original and knockoff variables) with less than full rank, preventing the use of ordinary least squares regression in the knockoff selection procedure. Chapter II proposes that these tuning parameters be chosen by maximizing the determinant of the Gram matrix for the augmented design matrix. This directly attempts to improve conditioning in the matrix of original and knockoff variables and permits the use of ordinary least squares when comparing the estimated effect of each observed covariate to its knockoff counterpart. In some moderate-dimensional regression scenarios this determinant-based tuning is shown to improve statistical power to detect truly non-null variables, and in most situations this tuning does not reduce statistical power.

Chapter III describes a form of non-replicability and variance inflation in the knockoff filter and proposes a stabilized knockoff filter to mitigate these issues. The construction of knockoff variables for any fixed set of observed covariates involves arbitrary algebraic choices which can lead to substantial variation in the set of selected variables, even with a fixed design matrix and response vector. For any fixed design matrix, there are infinitely many sets of valid knockoff variables, any one of which can be used to select variables in the knockoff filter. In the knockoff filter, selecting variables and controlling the FDR is based on the conditional distribution of the response vector, given the fixed design matrix. In this setting, the non-uniqueness of knockoff variables for a given design matrix does not affect the distributional properties of the knockoff filter which lead to FDR control. However, this means that for a fixed design matrix and response vector, repeatedly applying the knockoff filter can lead to instability in the resulting inferences. This instability can reduce power, since the computation of knockoff variables can, by luck or arbitrary computational choices, lead to very few selected variables. This instability is also a form of non-replicability, in which the same statistical analysis repeatedly applied to a single, fixed set of ob-

servations can lead to non-negligible differences in the obtained inferences.

The stabilized knockoff filter in Chapter III takes advantage of this non-uniqueness of the knockoff variables, and the resulting variation in the knockoff-based estimates of variable importance, to reduce variance and improve power. This is achieved by computing a low-variance estimate of the FDR based on repeatedly generating sets of knockoff variables for a given design matrix. The low-variance FDR estimate leads to a low-variance threshold for the knockoff estimates of variable importance. A stabilized set of selected variables is obtained by selecting those variables whose importance statistics are most likely to exceed the stabilized threshold. This stabilized knockoff filter is shown to control the FDR in a wide array of simulation scenarios while reducing the standard deviation in the number of selected variables by as much as a factor of two or three. In nearly all simulations, the power of the stabilized knockoff filter is at least as large as that of the knockoff filter. In many simulations, it is necessary to use a modified version of the knockoff filter which only approximately controls the FDR in order to obtain power similar to that of the stabilized knockoff filter. Using the version of the knockoff filter which always controls the FDR leads to very low power in moderate-dimensional settings compared to the stabilized knockoff filter.

Finally, Chapter IV develops a linear mixed effects model for sequential multiple assignment randomized trials (SMARTs) with a continuous, longitudinal outcome. SMARTs are an important data collection tool for informing the construction of dynamic treatment regimens (DTRs), which use cumulative patient information to recommend specific treatments during the course of an intervention. Primary scientific questions in a SMART can involve the comparison of DTRs based on the marginal mean of a longitudinal outcome. Existing statistical methods for SMARTs are similar to generalized estimating equations. Chapter IV develops a linear mixed modeling and estimation approach appropriate for estimating the marginal mean of

a longitudinal outcome for each DTR in a SMART.

In this mixed model, the counterfactual potential outcomes under each DTR are modeled as a linear combination of the marginal mean and subject-specific random effects. The model is marginal over the interim variables used to define treatment decisions over the course of the intervention. Parameters in the model are estimated based on a weighted pseudo-likelihood whose weights permit estimation of the marginal means of interest. Unlike existing approaches based on generalized estimating equations, the mixed model distinguishes between-person and within-person variation and allows for flexible marginal covariance structures in the longitudinal outcome. As in other modeling approaches based on generalized estimating equations, the estimator for the marginal mean is consistent and asymptotically Gaussian even when the marginal variance-covariance of the longitudinal potential outcomes is misspecified, in this case due to misspecification of the random effects structure. Simulation studies confirm these theoretical results, and an illustrative analysis is provided using data from a SMART in autism research.

CHAPTER II

The Role of Collinearity and Tuning Parameters in the Knockoff Filter

2.1 Introduction

Consider the regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{Y}$ is an n-dimensional response vector, $\mathbf{X}$ is an $n \times p$ design matrix with columns $\mathbf{X}_1, \ldots, \mathbf{X}_p$, $\mathbb{E}(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\epsilon} \mid \mathbf{X}) = \sigma^2\mathbf{I}$. One goal of a regression analysis may be the identification of elements of $\boldsymbol{\beta}$ that are nonzero, that is, covariates $\mathbf{X}_j$ that are associated with the response conditional on the other $p-1$ covariates. In this context, the knockoff filter (Barber and Candès 2015; Candès et al. 2018; Barber, Candès, and Samworth 2018) controls the regression false discovery rate (FDR) in finite samples, with arbitrary covariate dependence, when $n \geq p$ and $\boldsymbol{\epsilon} \mid \mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. The regression FDR is the expected proportion of selected covariates that are, in fact, not associated with the response, conditional on the other covariates. In practice, "controlling" the FDR means that when a set of relevant variables is identified, we expect that a small fraction of those variables were identified erroneously.

The knockoff filter constructs a set of p synthetic covariates which mimic the correlation structure of the observed covariates and are constructed without knowledge of the response. Crucially, these synthetic covariates are correlated with the observed

covariates. A regression model is fit to this augmented set of $2p$ variables. By construction, the synthetic variables will behave similarly to the real variables with null effects in the regression model. If an original variable cannot be distinguished from its knockoff version in the regression model, it is identified as having a null effect; variables estimated as having much larger effects than their knockoff counterparts are selected as non-null.

This chapter examines the effect of collinearity on the statistical power of the knockoff filter in moderate-dimension regression problems. Specifically, I will describe how the construction of knockoff variables can exacerbate existing collinearity among the covariates and lead to a singular design matrix in the augmented regression model used to compare the original covariates to their knockoffs. I will then suggest an alternative tuning method for the knockoff filter which alleviates collinearity in the augmented model. As currently defined, the knockoff filter includes a p-dimensional tuning parameter $\mathbf{s} = (s_1, \ldots, s_p)$, and Barber and Candès (2015) provide a single criterion for choosing this tuning parameter. Using this criterion, the knockoff filter produces singular augmented design matrices (the set of $2p$ original and knockoff variables) when the original design matrix is not nearly orthogonal. This chapter describes how this tuning parameter can induce high collinearity or can be used to reduce collinearity in the augmented model, thereby allowing the knockoff filter to be applied with non-penalized regression methods when design matrices are not close to orthogonal. This alternative tuning method maintains FDR control and improves statistical power in some moderate-p settings.

An alternative, "model-$\mathbf{X}$" knockoff filter (Candès et al. 2018) has been developed which allows the distribution of $\mathbf{Y} \mid \mathbf{X}$ to be arbitrary (for example, when a nonlinear link function relates $\mathbf{X}$ to $\mathbb{E}(\mathbf{Y} \mid \mathbf{X})$) and provides FDR control in the $p > n$ setting. This model-$\mathbf{X}$ knockoff filter requires that the joint distribution of $\mathbf{X}$ be known. In this thesis I focus on "fixed-$\mathbf{X}$" knockoffs, where a linear model for $\mathbf{Y}$ is proposed

conditionally on $\mathbf{X}$.

2.1.1 FDR and multiple testing in regression

Before describing the knockoff filter, I will review the false discovery rate (FDR) in the linear regression context. Given a set of D null hypotheses, suppose that R of them are rejected on the basis of observed data. Of these R rejected hypotheses or "discoveries", V are in fact true (incorrectly rejected) and $R - V$ are false. The false discovery proportion (FDP) is $V/\max\{R, 1\}$, the proportion of rejected hypotheses that are actually true. The FDR is the expected false discovery proportion, while the family-wise error rate (FWER) is the probability that *any* of the hypotheses are incorrectly rejected (Efron 2010). For a desired error rate q, a multiple testing procedure "controls" the FDR (FWER) if FDR (FWER) $\leq q$.

The Bonferroni correction (Shaffer 1995, e.g.) uses P-values for the D hypotheses to control the FWER (with dependent or independent test statistics). Note that any FWER-controlling procedure also controls the FDR The FDR-controlling procedure of Benjamini and Hochberg (1995) proceeds as follows. Given ordered p-values $p_{(1)} \leq \cdots \leq p_{(D)}$ for the D null hypotheses and a desired FDR $q \in (0, 1)$, the hypotheses $H_{(1)}, \ldots, H_{(k)}$ are rejected, where

$$k = \max_i \left\{ p_{(i)} \leq \frac{i}{D} q \right\} \tag{2.1}$$

This procedure has FDR equal to $\frac{D_0}{D} q$, where D_0 is the number of false hypotheses, but requires that the set of test statistics from the D hypotheses are either independent or have "positive regression dependence on a subset" (Benjamini and Yekutieli 2001). Both the Bonferroni and Benjamini-Hochberg procedures require P-values, which may be difficult to obtain based on penalized regression techniques such as the lasso.

Here, we are concerned with p null hypotheses in a regression problem, namely

$H_{0j} : \beta_j = 0$, $j = 1, \ldots, p$, and the FDR is the expectation of the false discovery proportion taken over the distribution of the response vector conditional on the covariates. In this case, a discovery is a variable identified as non-null, that is, a variable $\mathbf{X}_j$ (the jth column of $\mathbf{X}$) such that $\mathbf{Y}$ depends on $\mathbf{X}_j$ conditional on $\{\mathbf{X}_k : k \neq j\}$. In many cases, the covariates are correlated, which means that the standard test statistics for each regression coefficient are not independent. The knockoff filter controls the FDR for the p hypotheses $H_{0j} : \beta_j = 0$ in the linear regression model $\mathbf{Y} \,|\, \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and does not require independent test statistics for each hypothesis, unlike the Benjamini-Hochberg procedure.

Testing the hypotheses $H_{0j} : \beta_j = 0$ is related to the task of variable selection in regression, but adds to it rigorous control of the error rate through the FDR. Note that this FDR is different from that of a marginal screening problem in which one hopes to discover a subset of covariates which all have marginal associations with the response. Intuitively, control of the regression FDR is a more difficult task than controlling the FDR for marginal relationships. If two covariates are highly correlated but only one of them has an effect in the regression, then they both are "true discoveries" in a marginal screening task whereas one of them is a false discovery in the regression.

The Bonferroni correction and other methods for controlling the FWER were developed in settings where few hypotheses are being tested, as in agricultural experiments with six or ten pairwise comparisons between treatment groups. Modern scientific applications, such as observational studies using data from electronic health records, may require testing thousands of hypotheses (potentially relevant variables). In this context, guaranteeing that a low fraction of hypothesis rejections are in error may be more a desirable goal than preventing even a single false rejection, which could lead to very few rejected hypotheses (Benjamini and Hochberg 1995; Efron 2010).

Wu, Boos, and Stefanski (2007) also proposed a procedure based on synthetic variables and present simulation studies suggesting it has approximate FDR control.

In their work, the synthetic variables are uncorrelated with the original variables and bootstrap-like replications are used to estimate the FDR for a given set of selected variables. Other work on false discovery rates in regression typically relies on near independence of test statistics (or of covariates) or provides asymptotic FDR control (Meinshausen, Meier, and Bühlmann 2009; Storey, Taylor, and Siegmund 2004; Storey 2002; Bogdan et al. 2015). The knockoff filter has a theoretical guarantee of FDR control in finite samples, with arbitrary covariate dependence, when $\mathbf{Y} \mid \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

2.1.2 Motivating application

Electronic health records (EHR) and medical claims databases provide detailed, longitudinal data on many variables recorded during encounters with the healthcare system. These databases potentially contain millions of individuals with high-dimensional measurements collected over time. The data can be sparse and error-prone, reflecting variation in data entry by individual providers, ambiguous relationships between recorded variables and clinical outcomes of interest, and the primary use of EHR data for billing and other non-research purposes. These issues can limit the ability of researchers to exploit the high-volume longitudinal data contained in EHRs (Weiskopf and Weng 2013; Jensen, Jensen, and Brunak 2012).

While many researchers have focused on predicting clinical outcomes such as the onset of heart failure (e.g. Wu, Roy, and Stewart 2010; Austin et al. 2013; Choi et al. 2017), the exploratory goal of identifying relevant variables in claims and EHR data has been underdeveloped. In this context, statistical methods could be applied to discover a small set of variables which are relevant to a clinical outcome among hundreds or thousands of variables contained in medical records or insurance claims. Although claims and EHR data are high-dimensional, massive numbers of patients permit the use of methods such as the knockoff filter in which the sample size is

typically larger than the number of variables. This kind of feature identification could be useful in the analysis of large claims databases such as MarketScan (IBM Watson Health 2018).

2.2 The knockoff filter

Next I will describe the fixed-$\mathbf{X}$ knockoff filter of Barber and Candès (2015) and the tuning parameter choices suggested by these authors. Section 2.4 proposes alternative tuning choices for the knockoff filter.

2.2.1 Definition

Following Barber and Candès (2015), let $\mathbf{X}$ be the $n \times p$ matrix of mean-centered covariates with columns denoted by $\mathbf{X}_j$. The columns of $\mathbf{X}$ are normalized so that $\|\mathbf{X}_j\| = 1$ and denote $\boldsymbol{\Sigma} = \mathbf{X}^\intercal\mathbf{X}$, which is assumed to be nonsingular. With $\mathbf{Y} = (Y_1, \ldots, Y_n)^\intercal$, the assumed population model is $\mathbf{Y} \mid \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. The fixed-$\mathbf{X}$ knockoff filter consists of three steps:

1. Without involving $\mathbf{Y}$, construct the $n \times p$ "knockoff" design matrix $\tilde{\mathbf{X}}$ so that

$$\mathbf{G} := \left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^\intercal \left[\mathbf{X}\ \tilde{\mathbf{X}}\right] = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathbf{S} \\ \boldsymbol{\Sigma} - \mathbf{S} & \boldsymbol{\Sigma} \end{bmatrix} \tag{2.2}$$

   is positive semidefinite, where $\mathbf{S} = \text{diag}(\mathbf{s})$ and $\mathbf{s} = (s_1, \ldots, s_p)$ has nonnegative entries. The vector $\mathbf{s}$ is a tuning parameter. The knockoff variables (columns of $\tilde{\mathbf{X}}$) serve as negative controls for the original variables.

2. Given the augmented design matrix $\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]$, compute "importance statistics" W_j, $j = 1, \ldots, p$ so that large, positive values of W_j provide evidence that β_j is nonzero. These importance statistics must be a function of $\mathbf{G}$ and $[\mathbf{X}\ \tilde{\mathbf{X}}]^\intercal\mathbf{Y}$

and must have the property that swapping $\tilde{\mathbf{X}}_j$ and $\mathbf{X}_j$ in the augmented design matrix changes the sign of the corresponding W_j.

3. Select variables j such that $W_j \geq T$, where T is a data-dependent threshold defined below in equation (2.4).

First note that for any nonsingular $\boldsymbol{\Sigma}$, there exists $\mathbf{s}$ such that $\mathbf{G}$ is positive semidefinite. As noted in Barber and Candès (2015, Section 2.1), $\mathbf{G}$ is positive semidefinite if and only if $\mathbf{S} \succeq 0$ and $2\boldsymbol{\Sigma} - \mathbf{S} \succeq 0$. For any $\mathbf{s}$ such that $0 < \min_j s_j$ and $\max_j s_j < 2\lambda_{\min}(\boldsymbol{\Sigma})$, where $\lambda_{\min}(\boldsymbol{\Sigma})$ is the smallest eigenvalue of $\boldsymbol{\Sigma}$, we will have $\mathbf{S}$ and $2\boldsymbol{\Sigma} - \mathbf{S}$ positive definite (see Section 2.3.1 for details). Since $\boldsymbol{\Sigma}$ is nonsingular, $\lambda_{\min}(\boldsymbol{\Sigma}) > 0$, so it is possible to choose $\mathbf{s}$ satisfying these conditions.

By construction, $\mathbf{X}_j^\intercal \tilde{\mathbf{X}}_j = 1 - s_j$, so s_j is a tuning parameter which controls the degree of correlation between the jth covariate and its knockoff. Variable selection in the knockoff filter is performed by using the statistics W_j to compare the original variables to their knockoffs, which are known to be independent of the response given the true covariates. An original variable that cannot be distinguished from its knockoff using the importance statistic W_j is likely to be a null variable, i.e. a variable j with $\beta_j = 0$. Since $\tilde{\mathbf{X}}_j$ is known to have no effect on $\mathbf{Y}$ given the other covariates, reducing correlation between $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ prevents the estimated effect of a non-null variable from being attenuated due to correlation with its knockoff. By this reasoning, the statistic W_j is more likely to detect a true signal ($\beta_j \neq 0$) when s_j is close to 1. As will be explored in the next section, the choice of $\mathbf{s}$ also affects the degree of linear dependence in the augmented design matrix $\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]$.

The correlation structure of $\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]$, along with the computation of W_j as a function of $\mathbf{G}$ and $\left[\mathbf{X}\,\tilde{\mathbf{X}}\right]^\intercal \mathbf{Y}$, means that $\#\{j : \beta_j = 0, W_j \leq -t\}$ has the same distribution as $\#\{j : \beta_j = 0, W_j \geq t\}$ for any fixed $t > 0$; this allows the false discovery

proportion (FDP) at a threshold $t > 0$ to be estimated by

$$\widehat{\text{FDP}}(t) = \frac{\#\{j : W_j \leq -t\}}{\max\{1, \#\{j : W_j \geq t\}\}}, \tag{2.3}$$

leading to a selection threshold which controls FDR at level $q \in [0, 1]$:

$$T = \min_j \left\{ t = |W_j| : \frac{1 + \#\{k : W_k \leq -t\}}{\max\{1, \#\{k : W_k \geq t\}\}} \leq q \right\}. \tag{2.4}$$

Note that the numerator of (2.3) is an estimate of the number of false discoveries, since $\#\{j : W_j \leq -t\} \geq \#\{j : \beta_j = 0, W_j \leq -t\}$ and this latter quantity has the same distribution as $\#\{j : \beta_j = 0, W_j \geq t\}$, the true number of false discoveries at threshold t.

The knockoff estimate of FDP given in (2.3) is not monotonic, but will tend to zero as t increases to $\max_j |W_j|$. The FDR-controlling threshold (2.4) adds 1 to the numerator of (2.3) before comparing the FDP estimate to the nominal level q. This is referred to as the "knockoff+" in Barber and Candès (2015). Adding 1 to the numerator of (2.3) produces a different estimate of the FDP which increases to 1 as t increases. The threshold in (2.4) leads to FDR control of the knockoff filter, while computing a threshold based on (2.3) only approximately controls the FDR. See Barber and Candès (2015) for details.

Figure 2.1 compares these two knockoff estimates of the FDP as a function of candidate thresholds $t > 0$ in a synthetic dataset with $p = 50$ orthogonal, multivariate Gaussian covariates and 25 truly non-null variables. With either FDP estimate, the threshold is chosen as the smallest t such that the estimated FDP is below the desired FDR. Figure 2.1 illustrates the conservatism in (2.4), which over-estimates the true FDP for a given vector of importance statistics. By initially decreasing and then increasing toward 1, the knockoff+ threshold prevents either too many or too few discoveries; in fact, based on (2.4), the threshold T which controls the FDR at level

q will, by construction, satisfy

$$\#\{j : W_j \geq T\} \geq \frac{1}{q} + \frac{1}{q}\#\{j : W_j \leq -T\}, \tag{2.5}$$

so, based on (2.4), when at least one variable is selected, at least $\frac{1}{q}$ variables will be selected.

The knockoff design matrix has three key properties: the distribution of the knockoffs approximates the distribution of the observed variables in that they have the same correlation structure among themselves; the knockoffs are independent of $\mathbf{Y}$ given the observed covariates; and the knockoffs are (in general) correlated with the observed variables. This last property is critical, as suggested in Barber and Candès (2015, Section 3.1). Note that $\mathbf{X}_j^\intercal \mathbf{X}_k = \mathbf{X}_j^\intercal \tilde{\mathbf{X}}_k$ and $\mathbf{X}_j^\intercal \mathbf{X}_k = \tilde{\mathbf{X}}_j^\intercal \tilde{\mathbf{X}}_k$ for any $j \neq k$. When the estimated effect of a null variable $\mathbf{X}_j$ is large due to its correlation with some other non-null variable, then the estimated effect of $\tilde{\mathbf{X}}_j$ will also tend to be large because $\tilde{\mathbf{X}}_j$ is equally correlated with the same non-null original variable. In this way the correlation between the knockoff and original variables allows the knockoffs to serve as a negative controls.

In the model-$\mathbf{X}$ knockoff filter (Candès et al. 2018), $\mathbf{X}$ and $\tilde{\mathbf{X}}$ are considered random, and a joint distribution for $(\mathbf{X}, \tilde{\mathbf{X}})$ is specified to satisfy two conditions: the distribution of $(\mathbf{X}, \tilde{\mathbf{X}})$ is unchanged under pairwise interchanging of columns of $\mathbf{X}$ with their knockoff counterparts; and $\tilde{\mathbf{X}}$ is independent of $\mathbf{Y}$, conditional on $\mathbf{X}$. Then a single $\tilde{\mathbf{X}}$ matrix is sampled from the conditional distribution $\tilde{\mathbf{X}} \mid \mathbf{X}$. One example of a joint distribution with these properties is $(\mathbf{X}, \tilde{\mathbf{X}}) \sim N(\mathbf{0}, \mathbf{G})$, where $\mathbf{G}$ is the augmented Gram matrix defined previously. This random sampling of the knockoff design matrix will result in lower sample correlations between original variables and their knockoffs. However, as stated previously the model-$\mathbf{X}$ knockoff requires knowledge of the joint distribution of $\mathbf{X}$.

2.2.2 Implementation and tuning

Now we can turn to the implementation and tuning choices made by the analyst when using the knockoff filter. First, the elements of $\mathbf{s}$ need to be specified so that $\mathbf{G}$ is positive semidefinite. Then, given a valid choice of $\mathbf{s}$, the knockoff design matrix $\tilde{\mathbf{X}}$ can be constructed so that $\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^\intercal \left[\mathbf{X}\ \tilde{\mathbf{X}}\right] = \mathbf{G}$ by first defining $\mathbf{C}^\intercal\mathbf{C} = 2\mathbf{S} - \mathbf{S}\boldsymbol{\Sigma}^{-1}\mathbf{S}$ and then setting

$$\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\mathbf{S}) + \tilde{\mathbf{U}}\mathbf{C}, \tag{2.6}$$

where $\tilde{\mathbf{U}}$ is an $n \times p$ matrix with orthonormal columns such that $\tilde{\mathbf{U}}^\intercal\mathbf{X} = \mathbf{0}$. Since $\boldsymbol{\Sigma}$ is nonsingular, $\mathbf{G}$ is positive semidefinite if and only if $\mathbf{C}^\intercal\mathbf{C}$ is positive semidefinite; hence $\mathbf{C}$ exists for any valid $\mathbf{s}$. Note that the matrix $\tilde{\mathbf{U}}$ is not uniquely defined given $\mathbf{X}$ and $\mathbf{s}$ (e.g., for any orthogonal matrix $\mathbf{Q}$, $\tilde{\mathbf{U}}\mathbf{Q}$ can be used instead of $\tilde{\mathbf{U}}$ in (2.6)). The particular choice of $\tilde{\mathbf{U}}$ can lead to wide variation in the set of selected variables in the knockoff filter. The issue of choosing $\tilde{\mathbf{U}}$ to reduce variability in the knockoff filter is explored in Chapter III.

There are two methods for calculating $\mathbf{s}$ given in Barber and Candès (2015): SDP and equi-correlated. The only restrictions placed on $\mathbf{S} = \text{diag}(\mathbf{s})$ are that $\mathbf{G}$ is positive semidefinite, which occurs when $\mathbf{S} \succeq 0$ and $2\boldsymbol{\Sigma} - \mathbf{S} \succeq 0$. Following the heuristic argument given earlier, high power may be achieved when $1 - s_j$, the correlation between $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$, is close to zero. With this in mind, Barber and Candès (2015) both the SDP and equi-correlated constructions seek to minimize the correlations between the knockoff and original covariates, and, in fact, the equi-correlated construction is a special case of the SDP construction. In Section 2.3 I will describe how tuning choices for $\mathbf{s}$ can affect both collinearity and power when using the knockoff filter.

SDP knockoffs are constructed by solving

$$\begin{aligned} &\min \ \sum_j (1 - s_j) \\ &\text{subject to } 0 \leq s_j \leq 1, \ 2\boldsymbol{\Sigma} - \mathbf{S} \succeq 0, \end{aligned} \tag{2.7}$$

which is a semidefinite program.

In the equi-correlated construction, $s_j = \min\{1, 2\lambda_{\min}(\boldsymbol{\Sigma})\}$ for all j, where $\lambda_{\min}(\boldsymbol{\Sigma})$ is the smallest eigenvalue of $\boldsymbol{\Sigma}$. Barber and Candès (2015) point out that this equi-correlated construction minimizes the pairwise correlations between knockoffs and original covariates. The following proposition states this fact more precisely.

Proposition II.1. *Setting* $s_j = \min\{1, 2\lambda_{min}(\boldsymbol{\Sigma})\}$ *for all* j *is the solution to* (2.7) *with the restriction* $\mathbf{s} = s(1, \ldots, 1)^\intercal$.

Proof. When $\mathbf{s} = s(1, \ldots, 1)$, then (2.7) is equivalent to

$$\begin{aligned} \min p(1-s) \iff \min -ps& \\ \text{s.t } 0 \leq s \leq 1, 2\boldsymbol{\Sigma} - sI \succeq 0& \\ \iff 0 \leq s \leq 1, 2\lambda_j(\boldsymbol{\Sigma}) - s \geq 0 \text{ for all } j& \\ \iff 0 \leq s \leq 1, s \leq 2\lambda_{\min}(\boldsymbol{\Sigma}),& \end{aligned}$$

where $\lambda_j(\boldsymbol{\Sigma})$ is the jth eigenvalue of $\boldsymbol{\Sigma}$ and $\lambda_{\min}(\boldsymbol{\Sigma})$ is the smallest eigenvalue. Taking s as large as possible while remaining feasible, we obtain $s = \min\{2\lambda_{\min}(\boldsymbol{\Sigma}), 1\}$ □

In certain circumstances the SDP solution will have an equi-correlated form, $\mathbf{s} = s(1, \ldots, 1)$, even without restricting the optimization problem to equi-correlated solutions. To understand when this occurs, define

$$\mathbf{M} = \begin{bmatrix} \mathbf{I} - \mathbf{S} & & \\ & \mathbf{S} & \\ & & 2\boldsymbol{\Sigma} - \mathbf{S} \end{bmatrix}$$

and rewrite the SDP optimziation problem as

$$\begin{aligned} &\min f(\mathbf{s}) = -(1, \ldots, 1)\mathbf{s} = -\sum_j s_j \\ &\text{s.t. } \mathbf{M} \succeq 0 \end{aligned} \tag{2.8}$$

Following Boyd and Vandenberghe (2004, Exercise 4.4) and Bachoc et al. (2012, Section 1.1), solutions to this problem will have the form $s(1, \ldots, 1)$ when the objective function is invariant under permutations of $\mathbf{s}$ and, for any feasible $\mathbf{s}$, permutations of $\mathbf{s}$ are also feasible. It is clear that $f(\mathbf{s})$ is invariant under permutation of $\mathbf{s}$. The condition $\mathbf{M} \succeq 0$ is satisfied when $s_j \in [0, 1]$ for all j and $2\boldsymbol{\Sigma} - \mathbf{S}$ is positive semidefinite. If $s_j \in [0, 1]$ for all j, then this is also true for permutations of $\mathbf{s}$, so the SDP problem will have an equi-correlated solution $s(1, \ldots, 1)$ whenever $2\boldsymbol{\Sigma} - \mathbf{S}$ remains positive semidefinite under permutations of $\mathbf{s}$. Future work will attempt to characterize the class of covariance matrices $\boldsymbol{\Sigma}$ such that positive definiteness of $2\boldsymbol{\Sigma} - \mathbf{S}$ is unchanged under permutation of $\mathbf{s}$.

One example of a case in which positive definiteness of $2\boldsymbol{\Sigma} - \mathbf{S}$ is invariant under permutation is when $\boldsymbol{\Sigma}$ has an exchangeable structure, with all off-diagonal elements equal to τ. In this case,

$$2\boldsymbol{\Sigma} - \mathbf{S} = \begin{pmatrix} 2 - s_1 & \tau & \tau & \cdots \\ \tau & 2 - s_2 & \tau & \cdots \\ \vdots & & \ddots & \end{pmatrix} \tag{2.9}$$

If $2\boldsymbol{\Sigma} - \mathbf{S}$ is positive semidefinite, then for any $\mathbf{v} \in \mathbb{R}^p, \mathbf{v} \neq 0$, $\mathbf{v}^\intercal(2\boldsymbol{\Sigma} - \mathbf{S})\mathbf{v} \geq 0$, and

$$\mathbf{v}^\intercal(2\boldsymbol{\Sigma} - \mathbf{S})\mathbf{v} = \sum_{j=1}^{p} v_j^2(2 - s_j) + \tau \left(\sum_{j=1}^{p} v_j \sum_{k \neq j} v_k \right) \geq 0. \tag{2.10}$$

Let $\mathbf{s}_\pi$ be a permutation of $\mathbf{s}$ with $\mathbf{S}_\pi = \text{diag}(\mathbf{s}_\pi)$. Let $\mathbf{v}_\pi$ be the result of applying the same permutation to $\mathbf{v}$. Then from (2.10) we have

$$\mathbf{v}^\intercal(2\boldsymbol{\Sigma} - \mathbf{S}_\pi)\mathbf{v} = \mathbf{v}_\pi^\intercal(2\boldsymbol{\Sigma} - \mathbf{S})\mathbf{v}_\pi, \tag{2.11}$$

which is nonnegative since v_π is a nonzero element of $\mathbb{R}^p$. Thus when $\boldsymbol{\Sigma}$ is exchangeable, positivity of $2\boldsymbol{\Sigma} - \mathbf{S}$ is invariant under permutations of $\mathbf{s}$ and the SDP solution will have the equi-correlated form $s(1, \ldots, 1)$.

Given $\mathbf{s}$ and the resulting knockoffs $\tilde{\mathbf{X}}$, the next step of the knockoff filter is to compute importance statistics $\mathbf{W} = (W_1, \ldots, W_p)$, which compare each $\mathbf{X}_j$ to its knockoff. Large, positive values of W_j provide evidence against $H_{0j} : \beta_j = 0$, and the variable $\mathbf{X}_j$ is selected as a non-null variable (H_{0j} is rejected) when W_j exceeds the threshold (2.4). As described previously, $\mathbf{W}$ must be a function of $\mathbf{G}$ and $\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^\intercal \mathbf{Y}$, and must have the property that swapping $\mathbf{X}_j$ with $\tilde{\mathbf{X}}_j$ in the augmented design matrix $\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]$ changes the sign of W_j. These two properties defining the importance statistics, given in Barber and Candès (2015, Sec. 2.2), allow $\#\{j : W_j \leq -t\}$ to be used as an estimate of the number of false discoveries at threshold t. One example of a statistic of this type is the difference in coefficient magnitudes for each knockoff and original variable: $W_j = |\hat{\boldsymbol{\beta}}_j(\lambda)| - |\hat{\boldsymbol{\beta}}_{j+p}(\lambda)|$, where λ is the fixed value of a regularization parameter in a penalized least squares fit. Another example is $|\hat{\boldsymbol{\beta}}_j| - |\hat{\boldsymbol{\beta}}_{j+p}|$, the difference in coefficient magnitudes for an ordinary least squares (OLS) regression. However, as will be explored in Section 2.3, in many cases linear dependence in the set of original and knockoff variables will prevent direct use of OLS importance statistics. One final example of a valid W_j statistic is $W_j = |\mathbf{X}_j^\intercal \mathbf{Y}| - |\mathbf{X}_{j+p}^\intercal \mathbf{Y}|$, the difference in

sample correlation magnitudes between each original and knockoff variable.

2.3 Collinearity in the knockoff design matrix

As noted above, correlation between the knockoff and original variables allows the knockoffs to serve as negative controls and is necessary for FDR control. But this correlation between knockoffs and original variables can amplify existing linear dependence in the original design matrix. Note that the off-diagonal elements of $\mathbf{\Sigma} - \mathbf{S}$ are equal to the off-diagonal elements of $\mathbf{\Sigma}$, so, for any $j \neq k$, we have $\mathbf{X}_j^\intercal \tilde{\mathbf{X}}_k = \mathbf{X}_j^\intercal \mathbf{X}_k$. So even when the jth knockoff variable $\tilde{\mathbf{X}}_j$ can be constructed so that $\mathbf{X}_j^\intercal \tilde{\mathbf{X}}_j = 1 - s_j = 0$, collinearity in the augmented design matrix is increased due to correlation between $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_k$ for any $j \neq k$. Here I focus on how the tuning parameter $\mathbf{s}$ influences collinearity, but I do not explore whether correlations between $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_k, j \neq k$ could be reduced while maintaining FDR control. Specifically, I describe how the SDP objective of minimizing pairwise correlations between $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ often leads to a linearly dependent augmented design matrix, or, equivalently, a singular Gram matrix $\mathbf{G}$. In some cases, this collinearity in the augmented design matrix can reduce the power of the knockoff filter.

2.3.1 Feasible choices of s

To understand the tuning parameter $\mathbf{s}$ we can start by examining the constraints imposed on the choice of $\mathbf{s}$ by the requirement that the augmented Gram matrix $\mathbf{G}$ is positive semidefinite. In Section 2.3.2 I describe how these constraints, when combined with the SDP or equi-correlated tuning choices, can lead to a singular $\mathbf{G}$.

The knockoff filter requires that the $2p \times 2p$ Gram matrix $\mathbf{G}$, defined in (2.2), is positive semidefinite, which occurs when $\mathbf{S} \succeq 0$ and $2\mathbf{\Sigma} - \mathbf{S} \succeq 0$. Equivalently, $\mathbf{G} \succeq 0$ if and only if $2\mathbf{S} - \mathbf{S}\mathbf{\Sigma}^{-1}\mathbf{S} \succeq 0$. Thus, collinearity in the original design matrix $\mathbf{X}$ constrains the choice of $\mathbf{s}$ through the conditions $2\mathbf{\Sigma} - \mathbf{S} \succeq 0$ and $2\mathbf{S} - \mathbf{S}\mathbf{\Sigma}^{-1}\mathbf{S} \succeq 0$

(recall the notation $\mathbf{\Sigma} = \mathbf{X}^\intercal\mathbf{X}$). In addition, since the columns of $\mathbf{X}$ are standardized, the diagonal elements of $\mathbf{\Sigma}$ are equal to 1 and therefore it is necessary that $s_j \leq 1$ to satisfy $2\mathbf{\Sigma} - \mathbf{S} \succeq 0$.

Let $\lambda_j(\mathbf{\Sigma})$ denote the jth eigenvalue of $\mathbf{\Sigma}$ and let $s_{\min} = \min\{s_1, \ldots, s_p\}$. Using Weyl's inequalities (Bhatia 1997, Theorem III.2.1), we have

$$2\lambda_j(\mathbf{\Sigma}) + \lambda_{\min}(-\mathbf{S}) \leq \lambda_j(2\mathbf{\Sigma} - \mathbf{S}) \leq \lambda_j(2\mathbf{\Sigma}) + \lambda_{\max}(-\mathbf{S}) \tag{2.12}$$

$$\iff 2\lambda_j(\mathbf{\Sigma}) - s_{\max} \leq \lambda_j(2\mathbf{\Sigma} - \mathbf{S}) \leq 2\lambda_j(\mathbf{\Sigma}) - s_{\min} \tag{2.13}$$

for all j. These inequalities lead to the following proposition.

Proposition II.2.

If $s_{min} > 2\lambda_{min}(\mathbf{\Sigma})$, *then* $2\mathbf{\Sigma} - \mathbf{S}$ *(hence* $\mathbf{G}$*) is not positive semidefinite.* (2.14)

If $0 < s_{min} \leq s_{max} < 2\lambda_{min}(\mathbf{\Sigma})$, *then* $\mathbf{S} \succ 0$ *and* $2\mathbf{\Sigma} - \mathbf{S} \succ 0$. (2.15)

Proof. Follows from the Schur complement relationship between $\mathbf{G}$, $\mathbf{S}$, and $2\mathbf{\Sigma} - \mathbf{S}$, along with (2.12)–(2.13). □

Statement (2.14) provides an upper bound on $s_{\min}$ as $\lambda_{\min}(\mathbf{\Sigma})$ approaches zero, while (2.15) shows that for any $\mathbf{\Sigma}$ such that $\lambda_{\min}(\mathbf{\Sigma}) > 0$, there exists $\mathbf{s}$ so that $\mathbf{G} \succeq 0$.

2.3.2 Singular G in SDP and equi-correlated constructions

We can now relate the general constraints on $\mathbf{s}$ imposed by the requirement that $\mathbf{G} \succeq 0$ to the SDP and equi-correlated choices of $\mathbf{s}$. The affine objective functions in the SDP and equi-correlated tunings of $\mathbf{s}$ cannot have extreme values on the interior of the feasible set. Thus, by using an affine objective function, the SDP and equi-correlated tuning of $\mathbf{s}$ lead to solutions on the boundary of the feasible set for $\mathbf{s}$, which in many cases causes $\mathbf{G}$ to be singular.

As stated above, the constraints of the SDP problem can be written as $\mathbf{M} \succeq 0$ where

$$\mathbf{M} = \begin{bmatrix} \mathbf{I} - \mathbf{S} & & \\ & \mathbf{S} & \\ & & 2\boldsymbol{\Sigma} - \mathbf{S} \end{bmatrix}.$$

Since the SDP objective function is affine with convex constraints, the solution will be on the boundary of the feasible set. That is, at least one of the constraints $\mathbf{I} - \mathbf{S} \succeq 0$, $\mathbf{S} \succeq 0$ or $2\boldsymbol{\Sigma} - \mathbf{S} \succeq 0$ will be active at the solution. Since

$$\det(\mathbf{G}) = \det(\mathbf{S}) \det(2\boldsymbol{\Sigma} - \mathbf{S}), \tag{2.16}$$

we can relate the constraints in the SDP problem to conditioning of the Gram matrix $\mathbf{G}$. Specifically, we can attempt to find conditions on $\boldsymbol{\Sigma}$ which lead to each of the constraints being active at the SDP solution, and describe how each active constraint influences the conditioning of $\mathbf{G}$.

First, if $\mathbf{S} \succeq 0$ is active, then at least one $s_j = 0$, which by (2.16) means that $\mathbf{G}$ is singular. Knowledge that $\mathbf{I} - \mathbf{S} \succeq 0$ is active at the solution does not provide information about the conditioning of $\mathbf{G}$ unless $\mathbf{I} - \mathbf{S} \succeq 0$ is the *only* active constraint, in which case $2\boldsymbol{\Sigma} - \mathbf{S}$ has positive eigenvalues and $\mathbf{G}$ is guaranteed to be nonsingular by (2.16). If at the solution, $s_j \in (0, 1)$ for all j, then $2\boldsymbol{\Sigma} - \mathbf{S} \succeq 0$ is the only active constraint, meaning that at least one eigenvalue of $2\boldsymbol{\Sigma} - \mathbf{S}$ will be equal to zero. Then, by (2.16), $\mathbf{G}$ is singular in this case.

The constraints on $\mathbf{I} - \mathbf{S}$ and $\mathbf{S}$ do not involve the observed data; only through $2\boldsymbol{\Sigma} - \mathbf{S} \succeq 0$ does collinearity in the observed covariates restrain the SDP solution for $\mathbf{s}$. Proposition II.2 suggests that when $\lambda_{\min}(\boldsymbol{\Sigma}) > 1/2$, the SDP solution is $\mathbf{s} = (1, \ldots, 1)^\intercal$, even if we do not restrict to equi-correlated $\mathbf{s}$. In this case, collinearity in the original variables is low enough so that $\mathbf{s}$ is unconstrained by the requirement that $\mathbf{G} \succeq 0$, and we can set $s_j = 1$ for all j to achieve minimal correlation between

each knockoff and original variable.

With the equi-correlated construction of $\mathbf{s}$, a more definitive statement is possible about conditions under which $\mathbf{G}$ is singular.

Proposition II.3. *Solving* (2.7) *with the restriction that* $\mathbf{s} = s(1, \ldots, 1)^\intercal$ *(equi-correlated) produces singular* $\mathbf{G}$ *whenever* $\lambda_{min}(\boldsymbol{\Sigma}) \leq \frac{1}{2}$.

Proof. Rewriting (2.16) with $\mathbf{s} = s(1, \ldots, 1)^\intercal$, we have

$$\begin{aligned} \det(\mathbf{G}) &= \det(s\mathbf{I}) \det(2\boldsymbol{\Sigma} - s\mathbf{I}) && (2.17) \\ &= s^p \det(2\boldsymbol{\Sigma} - s\mathbf{I}) && (2.18) \\ &= s^p \det(\mathbf{B}) \det(2\Lambda - s\mathbf{I}) \det(\mathbf{B}), && (2.19) \end{aligned}$$

where $\boldsymbol{\Sigma} = \mathbf{B}\Delta\mathbf{B}^\intercal$ is the eigendecomposition of $\boldsymbol{\Sigma}$. Since $\mathbf{B}$ is orthogonal, its determinant is either 1 or -1. So

$$\det(\mathbf{G}) = s^p \prod_j (2\lambda_j(\boldsymbol{\Sigma}) - s). \tag{2.20}$$

The equi-correlated solution to (2.7) is $s = \min\{2\lambda_{\min}(\boldsymbol{\Sigma}), 1\}$. Plugging this solution into (2.20) shows that $\mathbf{G}$ is singular whenever $\lambda_{\min}(\boldsymbol{\Sigma}) \leq \frac{1}{2}$. □

Combining Propositions II.2 and II.3, we see that $\lambda_{\min}(\boldsymbol{\Sigma})$ determines how collinearity in $\mathbf{X}$ constrains the SDP and equi-correlated constructions of $\mathbf{s}$. When $\lambda_{\min} > \frac{1}{2}$, the SDP solution will be $s_j = 1$ for all 1 and $\mathbf{G}$ will be nonsingular. When $\lambda_{\min}(\boldsymbol{\Sigma}) \leq \frac{1}{2}$, then the equi-correlated and SDP solutions can differ, and when $s_j = 0$ for some j or $2\boldsymbol{\Sigma} - \mathbf{S}$ has an eigenvalue equal to zero, $\mathbf{G}$ will be singular. To better understand these conditions, Figure 2.2 displays $\mathbb{P}(\lambda_{\min}(\boldsymbol{\Sigma}) > 0.5)$ when $\mathbf{X} \sim N(0, \boldsymbol{\Gamma})$ with $n = 1000$ and $\boldsymbol{\Gamma}$ has either an autoregressive or exchangeable structure Even with p as low as 10 and moderate correlation, only rarely will $\lambda_{\min}(\boldsymbol{\Sigma})$ be larger than $\frac{1}{2}$, allowing SDP tuning of $\mathbf{s}$ without leading to a singular $\mathbf{G}$.

As a concrete example of how the SDP knockoff construction affects conditioning of $\mathbf{G}$, Figure 2.3 displays the log-determinant of $\mathbf{G}$ as a function of the correlation τ when $\mathbf{\Sigma} = \mathbf{X}^\intercal\mathbf{X}$ is fixed to have an exact autoregressive or exchangeable structure and $p = 25$. The SDP and equi-correlated constructions for $\mathbf{s}$ lead to the same log-determinant of $\mathbf{G}$ in these covariance structures. Even in this low-dimensional setting, moderate correlation of 0.5 will lead to a singular augmented design matrix. This suggests that near orthogonality of $\mathbf{X}$ is needed to prevent the SDP knockoff construction from inducing a singular augmented Gram matrix, $\mathbf{G}$.

2.3.3 Variance inflation factors

In addition to the determinant of $\mathbf{G}$, Variance inflation factors (VIFs) can serve as an measure of the degree of collinearity in the augmented design matrix in the knockoff filter. In ordinary least squares regression, the variance inflation factor for $\mathbf{X}_j$ measures how linear relationships between $\mathbf{X}_j$ and the other columns of $\mathbf{X}$ inflate the variance of the jth least squares coefficient, $\hat{\boldsymbol{\beta}}_j$.

The VIFs for a design matrix $\mathbf{X}$ can be derived as follows. Recall that in the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \mid \mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, we have $\text{Var}\left(\hat{\boldsymbol{\beta}}_j \mid \mathbf{X}\right) = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}_{jj}$. Without loss of generality, take $\mathbf{X}_j$ as the last column of $\mathbf{X}$ and partition $\mathbf{X} = [\mathbf{X}_{-j}\ \mathbf{X}_j]$, where $\mathbf{X}_{-j}$ is the $n \times (p-1)$ matrix which results from removing the jth column of $\mathbf{X}$. Then

$$(\mathbf{X}^\intercal\mathbf{X})^{-1} = \begin{bmatrix} \mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{B}\mathbf{F}^{-1}\mathbf{C}\mathbf{L}^{-1} & -\mathbf{L}^{-1}\mathbf{B}\mathbf{F}^{-1} \\ -\mathbf{F}^{-1}\mathbf{C}\mathbf{L}^{-1} & \mathbf{F}^{-1} \end{bmatrix} \tag{2.21}$$

where

$$\mathbf{L} = \mathbf{X}_{-j}^\intercal\mathbf{X}_{-j} \qquad \mathbf{B} = \mathbf{X}_{-j}^\intercal\mathbf{X}_j$$

$$\mathbf{C} = \mathbf{B}^\intercal \qquad \mathbf{F} = \mathbf{X}_j^\intercal\mathbf{X}_j - \mathbf{X}_j^\intercal\mathbf{X}_{-j}\mathbf{L}^{-1}\mathbf{X}_{-j}^\intercal\mathbf{X}_j = \|(\mathbf{I} - \mathbf{P}_{-j})\mathbf{X}_j\|^2$$

and $\mathbf{P}_{-j}$ is the projection matrix onto the column space of $\mathbf{X}_{-j}$. This shows that $(\mathbf{X}^\intercal\mathbf{X})^{-1}_{jj}$, the jth diagonal element of $(\mathbf{X}^\intercal\mathbf{X})^{-1}$, is equal to $\mathbf{F}^{-1}$, defined above.

The multiple R^2 for the regression of $\mathbf{X}_j$ on $\mathbf{X}_{-j}$ is

$$R_j^2 = 1 - \frac{\|(\mathbf{I} - \mathbf{P}_{-j})\mathbf{X}_j\|^2}{\|\mathbf{X}_j - \overline{\mathbf{X}}_j\|^2}. \tag{2.22}$$

Thus

$$\text{Var}\left(\hat{\boldsymbol{\beta}}_j \mid \mathbf{X}\right) = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}_{jj} \tag{2.23}$$

$$(2.21) \implies \text{Var}\left(\hat{\boldsymbol{\beta}}_j \mid \mathbf{X}\right) = \sigma^2\frac{1}{\|(\mathbf{I} - \mathbf{P}_{-j})\mathbf{X}_j\|^2} \tag{2.24}$$

$$(2.22) \implies \text{Var}\left(\hat{\boldsymbol{\beta}}_j \mid \mathbf{X}\right) = \sigma^2\frac{1}{\|\mathbf{X}_j - \overline{\mathbf{X}}_j\|^2}\frac{1}{1 - R_j^2}. \tag{2.25}$$

The jth VIF is defined as $\frac{1}{1-R_j^2} \in [1, \infty)$ and reflects the contribution to $\text{Var}\left(\hat{\boldsymbol{\beta}}_j \mid \mathbf{X}\right)$ of correlations between $\mathbf{X}_j$ and the other columns of $\mathbf{X}$. Thus when $\|\mathbf{X}_j - \overline{\mathbf{X}}_j\|^2 = 1$ (i.e. with standardized covariates), we have that the jth VIF is the jth diagonal element of the inverse Gram matrix in an ordinary least squares fit.

In the knockoff filter, the augmented Gram matrix is

$$\mathbf{G} = \left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^\intercal\left[\mathbf{X}\ \tilde{\mathbf{X}}\right] = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathbf{S} \\ \boldsymbol{\Sigma} - \mathbf{S} & \boldsymbol{\Sigma} \end{bmatrix}, \tag{2.26}$$

and, using blockwise inversion, the upper $p \times p$ block of $\mathbf{G}^{-1}$ is equal to $(2\mathbf{S} - \mathbf{S}\boldsymbol{\Sigma}^{-1}\mathbf{S})^{-1}$. By construction, $\mathbf{G}$ is unchanged when $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are swapped, so the VIF for the jth covariate is equal to the VIF of the jth knockoff. Thus, for the augmented design matrix in the knockoff filter,

$$\text{VIF}_j = \left[(2\mathbf{S} - \mathbf{S}\boldsymbol{\Sigma}^{-1}\mathbf{S})^{-1}\right]_{jj} \tag{2.27}$$

for both the jth original and knockoff variable.

As an example, Figure 2.4 displays the average VIF for equi-correlated and SDP tuning of $\mathbf{s}$ using the same autoregressive or exchangeable structures for $\mathbf{\Sigma}$ from Figure 2.3. Comparing Figure 2.4 with Figure 2.3, we see that the VIF and log-determinant of $\mathbf{G}$ contain similar information about the conditioning of the augmented design matrix with SDP knockoffs. The VIFs become infinite at the same values of τ at which the determinant of $\mathbf{G}$ becomes zero.

2.4 Collinearity-reducing knockoff constructions

Here I propose two alternative methods of constructing the knockoff features via the choice of $\mathbf{s}$ which explicitly reduce collinearity in the augmented knockoff design matrix.

2.4.1 Maximizing the determinant of G

As we have seen, using the SDP tuning to maximize s_j in order to reduce correlation between $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ often leads to poor conditioning in the augmented design matrix. One alternative tuning method is to choose $\mathbf{s}$ to maximize the determinant of $\mathbf{G}$, the augmented design matrix. Recall that

$$\mathbf{G} = \left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^{\intercal}\left[\mathbf{X}\ \tilde{\mathbf{X}}\right] = \begin{bmatrix} \mathbf{\Sigma} & \mathbf{\Sigma}-\mathbf{S} \\ \mathbf{\Sigma}-\mathbf{S} & \mathbf{\Sigma} \end{bmatrix}, \tag{2.28}$$

with

$$\det(\mathbf{G}) = \det(\mathbf{\Sigma})\det(2\mathbf{S}-\mathbf{S}\mathbf{\Sigma}^{-1}S) = \det(\mathbf{S})\det(2\mathbf{\Sigma}-\mathbf{S}). \tag{2.29}$$

Proposition II.4. $\log\det(\mathbf{G})$ *is convex in the vector* $\mathbf{s} \in \mathbb{R}^p$.

Proof. This follows a similar proof in Boyd and Vandenberghe (2004, Section 3.1.5).

First consider $f(\mathbf{s}) = \log\det(2\boldsymbol{\Sigma} - \mathbf{S})$ where $\mathbf{S} = \text{diag}(\mathbf{s})$. One can prove convexity of f by restricting f to a line. Let $\mathbf{W} = 2\boldsymbol{\Sigma} - \mathbf{S}$ and consider

$$\begin{aligned} g(t) &= \log\det(2\boldsymbol{\Sigma} - \mathbf{S} - t\mathbf{V}) \\ &= \log\det\mathbf{W} + \log\det\left(\mathbf{I} - t\mathbf{W}^{-1/2}\mathbf{V}\mathbf{W}^{-1/2}\right) \\ &= \log\det\mathbf{W} + \sum_j \log(1 + t\lambda_j) \end{aligned}$$

where λ_j is the jth eigenvalue of $-\mathbf{W}^{-1/2}\mathbf{V}\mathbf{W}^{-1/2}$.

Then $g'(t) = \sum_j \frac{\lambda_j}{1+t\lambda_j}$ and $g''(t) = -\sum_j \frac{\lambda_j^2}{(1+t\lambda_j)^2}$ so g is concave and hence $\log\det(2\boldsymbol{\Sigma} - \mathbf{S})$ is concave. Since $\log\det(\mathbf{G}) = \log\det(\mathbf{S}) + \log\det(\mathbf{W})$ is the sum of concave functions, it is concave. □

Starting with (2.29), the gradient of $\log\det(\mathbf{G})$ can be computed as

$$\log\det(\mathbf{G}) = \sum_{j=1}^{p} \log s_j + \log\det(\mathbf{W}) \tag{2.30}$$

$$\frac{\partial \log\det(\mathbf{G})}{\partial s_j} = \frac{1}{s_j} + \frac{\partial \log\det(\mathbf{W})}{\partial s_j} \tag{2.31}$$

$$\frac{\partial \log\det(\mathbf{W})}{\partial s_j} = \text{Tr}\left[\frac{\partial \log\det(\mathbf{W})}{\partial \mathbf{W}}\frac{\partial \mathbf{W}}{\partial s_j}\right] \tag{2.32}$$

$$= \text{Tr}\left[(\mathbf{W}^{-1})^\intercal(-\mathbf{J}^{jj})\right] \tag{2.33}$$

where $\mathbf{J}^{jj}$ is the single-entry matrix with a 1 in the (j, j) position and zeroes elsewhere. Thus

$$\frac{\partial \log\det(\mathbf{G})}{\partial s_j} = \frac{1}{s_j} + \text{Tr}\left[\mathbf{W}^{-\intercal}(-\mathbf{J}^{jj})\right] \tag{2.34}$$

$$= \frac{1}{s_j} - \left[\mathbf{W}^{-1}\right]_{j,j} \tag{2.35}$$

These gradient calculations allow numerical optimization of $\log\det(\mathbf{G})$, leading to choices of $\mathbf{s}$ which will reduce linear dependencies among the columns of the aug-

mented design matrix.

As an initial example of how this maximum-determinant tuning reduces collinearity in the augmented design matrix, Figure 2.6 displays the ratio of the OLS standard errors in the augmented design matrix to those of the original design matrix when $\mathbf{\Sigma}$ has an autoregressive correlation structure and we fix $\sigma^2 = 1$. These ratios are given by

$$\frac{[(2\mathbf{S} - \mathbf{S}\mathbf{\Sigma}^{-1}\mathbf{S})^{-1}]_{jj}}{\mathbf{\Sigma}^{-1}_{jj}} \tag{2.36}$$

In this case, setting $\sigma^2 = 1$ and with standardized covariates, the OLS standard errors are equal to the VIFs. So Figure 2.6 can also be regarded as the ratio of the VIFs with and without the knockoff augmentation. The log-determinant tuning nearly eliminates the penalty in the OLS standard errors (or VIFs) due to augmenting the design matrix with the knockoff variables. The equi-correlated and SDP constructions of $\mathbf{s}$ quickly inflate the VIFs (or OLS standard errors) of the original design matrix as feature correlation increases.

2.4.2 Example: exchangeable population with equi-correlated knockoffs

As a special case, suppose $\mathbf{\Sigma}$ is exchangeable with diagonal entries equal to 1 and all off-diagonal entries equal to τ. As discussed previously, minimizing the average correlation between each pair of knockoff and original covariates leads to equi-correlated solutions, $\mathbf{s} = s(1, \cdots, 1)$, when $\mathbf{\Sigma}$ is exchangeable. As derived in Proposition II.3, we have

$$\det(\mathbf{G}) = s^p \prod_{j=1}^{p} (2\lambda_j - s) \tag{2.37}$$

whenever $\mathbf{s} = s(1, \ldots, 1)^\intercal$. Seeking to maximize $\log\det(\mathbf{G})$, we obtain

$$\frac{\partial \log\det(\mathbf{G})}{\partial s} = \frac{p}{s} - \frac{p-1}{2(1-\tau)-s} - \frac{1}{2+2(p-1)\tau - s} \tag{2.38}$$

because the eigenvalues of $\boldsymbol{\Sigma}$ are $1-\tau$ with multiplicity $p-1$ and $1+(p-1)\tau$ with multiplicity one.

The only critical point which lies in the interval $[0, \min\{2\lambda_{\min}(\boldsymbol{\Sigma}), 1\}]$ (required so that $\mathbf{G}$ is positive semidefinite) is

$$s = \frac{1}{2}\left(-\sqrt{4p^2\tau^2 - 8p\tau^2 + 4p\tau + 8\tau^2 - 8\tau + 1} + 2p\tau - 4\tau + 3\right). \tag{2.39}$$

Figure 2.5 displays the value of $\log\det(\mathbf{G})$ as a function of s when $\boldsymbol{\Sigma}$ is exchangeable, comparing the maximum-determinant equi-correlated construction to the equi-correlated construction in Barber and Candès (2015), where $s_j = \min\{2\min\lambda(\boldsymbol{\Sigma}), 1\}$, along with the minimum-VIF solution for s (described in the following section). First, we can see that the maximum-determinant and minimum-VIF choices for s lead to the same solution (it remains to determine whether this equivalence holds in general). The SDP-based solution is always at least as large as the maximum-determinant solution, which should lead to greater power according to the logic of Barber and Candès (2015). However, the instability in the importance statistics due to collinearity in the augmented design as a result of SDP or equi-correlated tuning of $\mathbf{s}$ could lead to reduced power. In order to improve the overall conditioning in the augmented design matrix, the maximum-determinant tuning of $\mathbf{s}$ tends to permit greater correlation (smaller s_j) between $\mathbf{X}_j$ and its knockoff.

2.4.3 Minimizing variance inflation factors

Alternatively, the VIFs in the augmented design matrix could be minimized as a function of $\mathbf{s}$ to reduce dependence in the augmented design matrix. Here I describe

the gradient calculations necessary to implement this optimization.

Let $\mathbf{R} = 2\mathbf{S} - \mathbf{S}\boldsymbol{\Sigma}^{-1}\mathbf{S}$ and $g(\mathbf{R}) = \mathbf{R}^{-1}$. The jth VIF, $j = 1, \ldots, p$ is

$$g_{jj} := \text{VIF}_j = e_j^\intercal g(\mathbf{R}) e_j = \text{Tr}\left[g(\mathbf{R}) e_j e_j^\intercal\right] \tag{2.40}$$

where $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)^\intercal$ is the jth standard basis vector. The objective is

$$\min_{\mathbf{s}} \sum_{i=j}^{p} g_{jj} = \text{Tr}\left[\mathbf{R}^{-1}\right]. \tag{2.41}$$

We have

$$\frac{\partial \mathbf{R}^{-1}}{\partial s_{jj}} = -\mathbf{R}^{-1} \frac{\partial \mathbf{R}}{\partial s_{jj}} \mathbf{R}^{-1} \tag{2.42}$$

and

$$\frac{\partial \mathbf{R}}{\partial s_{jj}} = 2\mathbf{J}^{jj} - \left(\mathbf{S}\boldsymbol{\Sigma}^{-1}\mathbf{J}^{jj} + \mathbf{J}^{jj}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right), \tag{2.43}$$

where $\mathbf{J}^{jj}$ is the $p \times p$ matrix with 1 in the j, j position and zeroes elsewhere.

The derivative of the ith VIF with respect to the jth entry of $\mathbf{s}$ is

$$\frac{\partial g_{ii}}{\partial s_{jj}} = \text{Tr}\left[\frac{\partial g(\mathbf{R})}{\partial s_{jj}} e_i e_i^\intercal\right] = \left[\frac{\partial g(\mathbf{R})}{\partial s_{jj}}\right]_{ii} \tag{2.44}$$

$$= \left[-\mathbf{R}^{-1} \frac{\partial \mathbf{R}}{\partial s_{jj}} \mathbf{R}^{-1}\right]_{ii} \tag{2.45}$$

and

$$\frac{\partial}{\partial s_{jj}} \sum_{i=1}^{p} \text{VIF}_i = \frac{\partial}{\partial s_{jj}} \text{Tr}\left[\mathbf{R}^{-1}\right] \tag{2.46}$$

$$= \sum_{i=1}^{p} \frac{\partial g_{ii}}{\partial s_{jj}} = \sum_{i=1}^{p} \left[\frac{\partial g(\mathbf{R})}{\partial s_{jj}}\right]_{ii} = \sum_{i=1}^{p} \left\{-\mathbf{R}^{-1} \frac{\partial \mathbf{R}}{\partial s_{jj}} \mathbf{R}^{-1}\right\}_{ii} \tag{2.47}$$

$$= \text{Tr}\left[-\mathbf{R}^{-1} \frac{\partial \mathbf{R}}{\partial s_{jj}} \mathbf{R}^{-1}\right]. \tag{2.48}$$

Initial numerical experiments suggest that the minimum-VIF choice of $\mathbf{s}$ has similar behavior in the knockoff filter as the maximum-determinant $\mathbf{s}$. In the simulations that follow I will focus on the maximum-determinant tuning for $\mathbf{s}$.

2.5 Simulation results

These numerical simulations compare the power and FDR control of the SDP knockoff construction with that of the maximum-determinant construction defined in Section 2.4.1 under various population structures and choices of importance statistics.

Following Barber and Candès (2015), simulations of the knockoff filter were performed as follows. Fixing $(n, p) = (5000, 100)$ or $(n, p) = (3000, 1000)$, the rows of $\mathbf{X}$ were generated as n independent samples from $N(\mathbf{0}, \mathbf{\Gamma})$, where $\mathbf{\Gamma}$ had either an exchangeable or autoregressive (AR) structure, with correlation parameter $\tau > 0$:

$$\begin{array}{cc} \text{Exchangeable} & \text{Autoregressive (AR)} \\ \begin{pmatrix} 1 & \tau & \tau & \cdots \\ \tau & 1 & \tau & \cdots \\ \vdots & & \ddots & \end{pmatrix} & \begin{pmatrix} 1 & \tau & \tau^2 & \tau^3 & \cdots \\ \tau & 1 & \tau & \tau^2 & \cdots \\ \tau^2 & \tau & 1 & \tau & \tau^2 \end{pmatrix} \end{array} \tag{2.49}$$

The columns of $\mathbf{X}$ were then normalized. The k nonzero β_j had magnitude 3.5 and the indices j such that $|\beta_j| = 3.5$ were selected uniformly from $\{1, 2, \ldots, p\}$. This

signal magnitude is roughly equal to the expected maximum least-squares coefficient under an orthogonal design with all true coefficients equal to zero. The nominal FDR was set at 0.1 for $p = 100$ and 0.2 for $p = 1000$.

Recall that, given the knockoff threshold T defined in equation (2.4), variable j is selected by the knockoff filter when $W_j \geq T$. The false discovery rate (FDR) is

$$\text{FDR} = \mathbb{E}\left(\frac{\#\left\{j : W_j \geq T \text{ and } \beta_j = 0\right\}}{\max\left\{1, \#\left\{j : W_j \geq T\right\}\right\}}\right), \tag{2.50}$$

and power in this context is defined as the expected proportion of the k non-null variables that are selected by the knockoff filter.

The SDP, equi-correlated, and maximum-determinant tuning choices for $\mathbf{s}$ were combined with four different importance statistics: $W_j = |\mathbf{X}_j^\intercal \mathbf{Y}| - |\mathbf{X}_{j+p}^\intercal \mathbf{Y}|$, the cross product difference (difference in sample correlation magnitudes); $W_j = |\hat{\boldsymbol{\beta}}_j(\lambda)| - |\hat{\boldsymbol{\beta}}_{j+p}(\lambda)|$, the difference in magnitude of the lasso regression coefficients (degree of penalization λ chosen by cross validation); $W_j = |\hat{\boldsymbol{\beta}}_j| - |\hat{\boldsymbol{\beta}}_{j+p}|$, the difference in magnitude of ordinary least squares (OLS) coefficients; and the difference in ridge regression coefficients (degree of penalization chosen by cross validation). If the augmented design matrix was numerically singular, a minimum-norm solution was used to compute the OLS importance statistics. As described previously, without the log-determinant tuning for $\mathbf{s}$, it will often be the case that the OLS-based importance statistics can only be computed with a minimum-norm solution to the OLS normal equations due to to exact singularity in the augmented design matrix.

First, for illustration, Figure 2.7 displays the distribution of $s_{\min} = \min\left\{s_1, \ldots, s_p\right\}$ and $s_{\max} = \max\left\{s_1, \ldots, s_p\right\}$ over 500 simulation replicates for each correlation structure and five levels of feature correlation. The maximum-determinant choice of $\mathbf{s}$ tends to have smaller $s_{\min}$ and $s_{\max}$ than the other two constructions. This allows each original variable to be more correlated with its knockoff variable than in the

equi-correlated or SDP constructions. However, with autoregressive correlation of 0.6 or 0.8, the the maximum-determinant construction has larger $s_{\min}$ than the SDP construction, so in these cases there is less correlation between knockoff and original variables using the maximum-determinant construction. There is generally less variance in $s_{\min}$ and $s_{\max}$ when using the maximum-determinant construction compared to the equi-correlated or SDP constructions.

To understand the performance of each combination of tuning method and importance statistic, I display FDR and power as a function of feature correlation in Figures 2.8–2.10 for $p = 100$ and Figures 2.11–2.12 for $p = 1000$. All tuning choices for $\mathbf{s}$ control FDR at the nominal level, and greater correlation among columns of $\mathbf{X}$ is associated with more conservative FDR control (lower rate of false discovery). With $p = 100$ and larger signal magnitude (with $|\beta_j| = 4.5$ for all nonzero β_j) or in a less sparse setting ($k = 40$ non-null variables out of $p = 100$), the log-determinant tuning of $\mathbf{s}$ is less conservative in controlling the FDR. Here we can also observe that, in general, the lasso coefficient importance statistics have higher power than the other importance statistics. However, I will make comparisons among the tuning methods for $\mathbf{s}$ within a fixed choice of importance statistic (e.g. OLS or lasso).

In the $p = 100$, $k = 10$, $|\beta_j| = 3.5$ setting in Figure 2.8, there is little or no reduction in power due to tuning $\mathbf{s}$ with the log-determinant of the augmented Gram matrix. In this low-signal setting (maximum power of approximately 0.3 achieved by any method), the log-determinant tuning neither harms nor improves power to detect true signals. With $k = 40$ out of $p = 100$ (Figure 2.9), the log-determinant tuning for $\mathbf{s}$ leads to some gain in power using the ridge, lasso, or OLS coefficient differences as importance statistics. For example, with exchangeable population correlation of 0.5 among the features, using lasso coefficient differences as importance statistics, the log-determinant tuning has power of about 0.53, while the equi-correlated tuning has power of about 0.38. When $k = 10$ but the true signal strength is large ($|\beta_j| = 4.5$ in

Figure 2.10), there is a smaller power gain achieved by the log-determinant tuning.

For $p = 1000$ and $n = 3000$, Figures 2.11 and 2.12 display the FDR and power of the equi-correlated and log-determinant tuning of $\mathbf{s}$ with autoregressive feature correlation. In this setting, the log-determinant tuning improves power when using OLS coefficient differences as the importance statistics. However, the log-determinant tuning of $\mathbf{s}$ with lasso coefficient differences has lower power than the equi-correlated tuning across all levels of feature correlation. In addition, the lasso coefficient differences (with either tuning approach for $\mathbf{s}$) has much greater power than either the simple cross products or OLS importance statistics.

Overall, comparing the SDP or equi-correlated tuning to the log-determinant tuning for $\mathbf{s}$ for a fixed choice of importance statistic, the log-determinant construction allows larger correlations between $\mathbf{X}_j$ and its knockoff variable but only slight reductions in power when p is large or there are very few true signals. Little or no power reduction was observed with $p = 100$ and a slight or moderate reduction in power was observed with $p = 1000$. When $p = 1000$, the log-determinant tuning improves the power of OLS importance statistics. When k is large relative to p, at least in moderate-dimension settings, these simulations suggest the potential for power gains using the log-determinant tuning of $\mathbf{s}$. In most settings, the lasso importance statistics had greater power than the OLS, ridge or cross product importance statistics. If it is desired to use OLS importance statistics with the knockoff filter, the log-determinant tuning is recommended to avoid computing these statistics using a singular augmented design matrix.

2.6 Discussion

To achieve FDR control with synthetic variables in regression requires that these synthetic variables reproduce correlations among the observed variables and are themselves correlated with the observed variables. This chapter showed that this technique,

implemented in the knockoff filter, can amplify existing collinearity in a set of observed variables and that this collinearity can reduce statistical power in some settings. But one of the tuning choices in the knockoff filter, namely, the degree of correlation between each knockoff and observed variable, can be used to mitigate collinearity in the knockoff design matrix, which can improve statistical power in some circumstances. Specifically, these tuning parameters can be chosen to maximize the log-determinant of the augmented Gram matrix $\mathbf{G}$. This tuning maintains FDR control in the knockoff filter and improves statistical power in some moderate-dimension and dense regression problems. In many settings, statistical power is unchanged when using this alternative tuning. In large-dimension, sparse problems, the determinant tuning incurs minor reductions in statistical power compared to existing tuning approaches.

The SDP construction of the tuning parameter $\mathbf{s}$ focuses only on each of the pairwise correlations, $\mathbf{X}_j^\intercal \tilde{\mathbf{X}}_j = 1 - s_j$. This is not surprising, since the statistics which identify non-null variables compare the effect estimate for each $\mathbf{X}_j$ to the paired effect estimate for $\tilde{\mathbf{X}}_j$. However, those effect estimates are affected by all of the correlations among the augmented set of variables, not just $\mathbf{X}_j^\intercal \tilde{\mathbf{X}}_j = 1 - s_j$, unless the design is nearly orthogonal. Hence by permitting greater pairwise correlations $\mathbf{X}_j^\intercal \tilde{\mathbf{X}}_j$ while attempting to reduce instability in those effect estimates as measured by collinearity, we do not sacrifice, and sometimes improve, statistical power.

This is only one of the potential tuning parameters in the knockoff filter which may impact its performance with correlated covariates. For example, the correlations $\mathbf{X}_j^\intercal \tilde{\mathbf{X}}_k$ for $j \neq k$ also induce greater collinearity in the augmented set of variables. It may be possible to shrink the magnitude of these correlations while maintaining the exchangeability properties of the knockoff variables.

It is apparent in the simulation results from Section 2.5 that different importance statistics W_j in the knockoff filter can affect statistical power and the knockoff filter's sensitivity to choices of $\mathbf{s}$. These results suggest that lasso-based importance statis-

tics are less sensitive to the choice of $\mathbf{s}$ and have higher power than ridge or OLS statistics. As in the lasso, ridge regression importance statistics are also based on penalized regression coefficients, which should themselves be less sensitive to collinearity. However, broadly speaking, the ridge coefficients had lower power than the lasso coefficients when used as importance statistics. The simple cross products were not sensitive to the choice of tuning method for $\mathbf{s}$ and had lower power than other importance statistics, even with weak feature correlation.

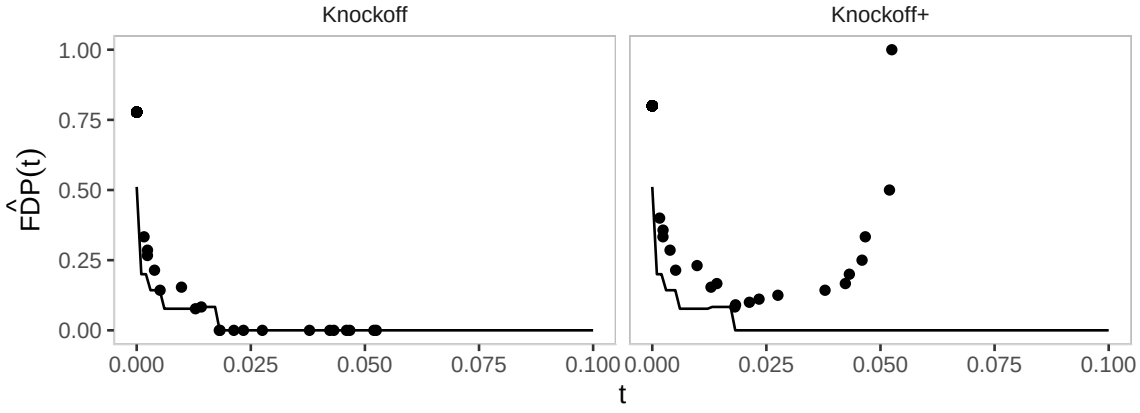


Figure 2.1: Example knockoff estimates of $\mathrm{FDP}(t)$, for a single simulated pair $(\mathbf{X}, \mathbf{Y})$ with $p = 50, n = 2000$ and 25 truly non-null variables. (See equations (2.3) and (2.4).) Solid line is the true false discovery proportion for this fixed vector $(W_1, \ldots, W_p)$ at a given threshold.

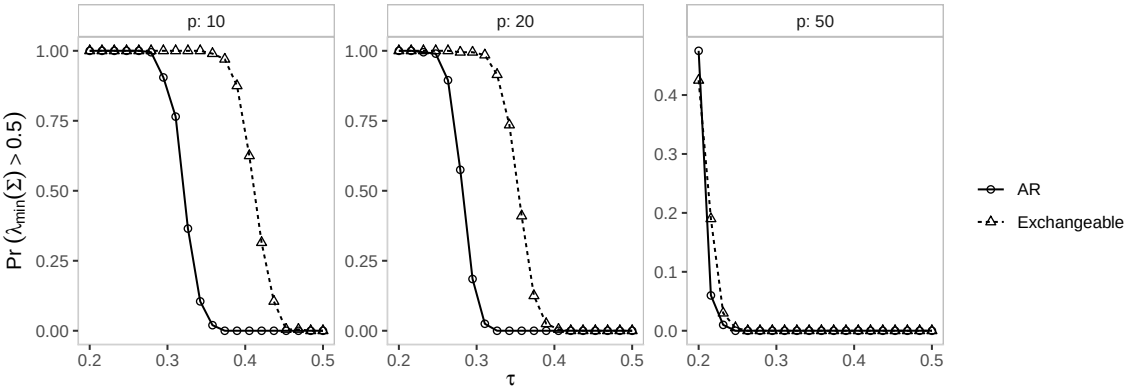


Figure 2.2: $\mathbb{P}\left(\lambda_{\min}(\mathbf{\Sigma}) \geq \frac{1}{2}\right)$ for Gaussian features, where $\mathbf{\Sigma} = \mathbf{X}^\intercal\mathbf{X}$. Features generated as $\mathbf{X} \sim N(0, \Gamma)$ when Γ has either an autoregressive or exchangeable structure (see equation (2.49) for definitions). Computed from 200 simulation replicates with $n = 1000$.

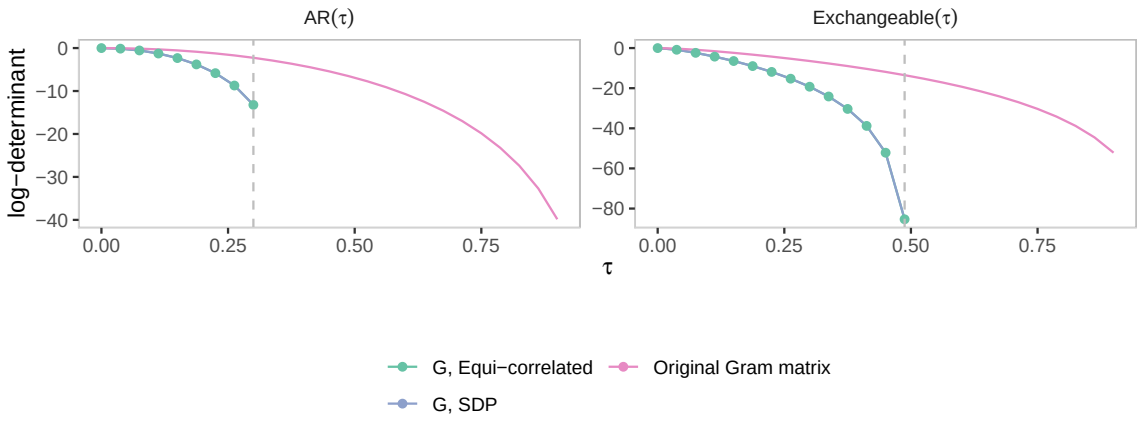


Figure 2.3: $\log\det(\mathbf{G})$ when $\mathbf{\Sigma}$ is autoregressive or exchangeable and $p = 25$. Vertical lines indicate the highest value of τ such that $\lambda_{\min}(\mathbf{\Sigma}) > \frac{1}{2}$.

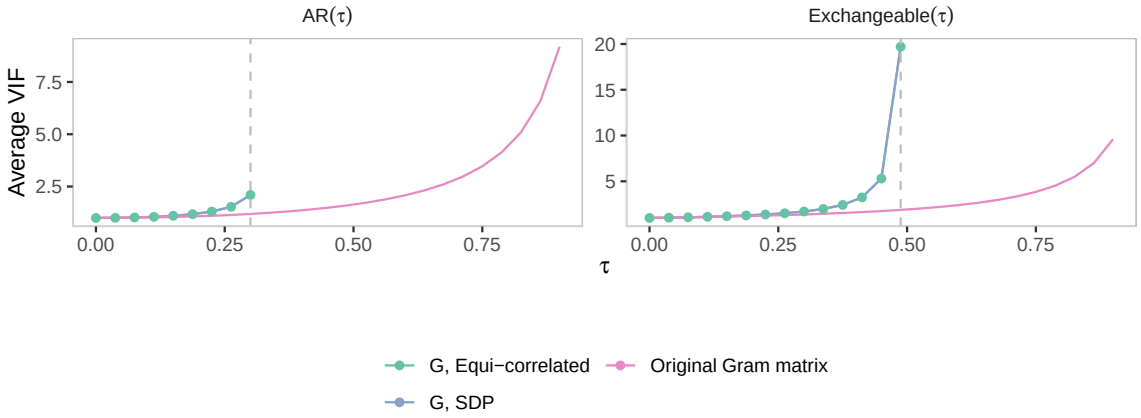


Figure 2.4: Average VIFs for the knockoff augmented design matrix. Plotted for $p = 25$ and a Gram matrix $\mathbf{\Sigma}$ with exact autoregressive or exchangeable structure (see equation (2.49)). Vertical lines indicate the highest value of τ such that $\lambda_{\min}(\mathbf{\Sigma}) > \frac{1}{2}$.

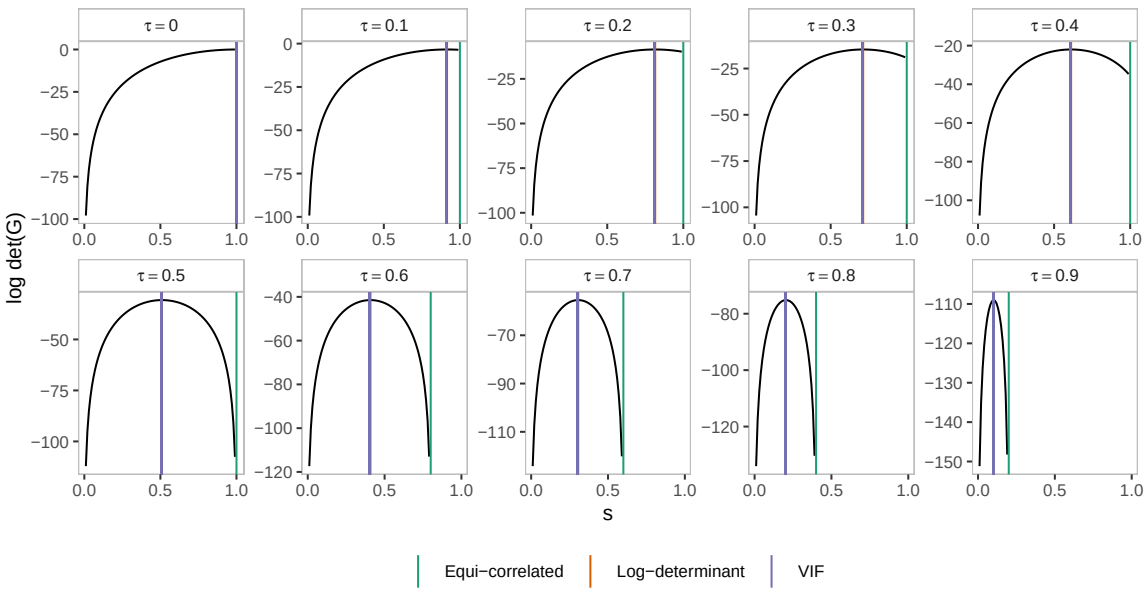


Figure 2.5: $\log\det(\mathbf{G})$ with exchangeable $\mathbf{\Sigma}$, $p = 25$, and $\mathbf{s} = s(1, \ldots, 1)$. Each panel corresponds to a single value for the exchangeable correlation parameter τ. The minimum-VIF and maximum-det($\mathbf{G}$) solutions overlap.

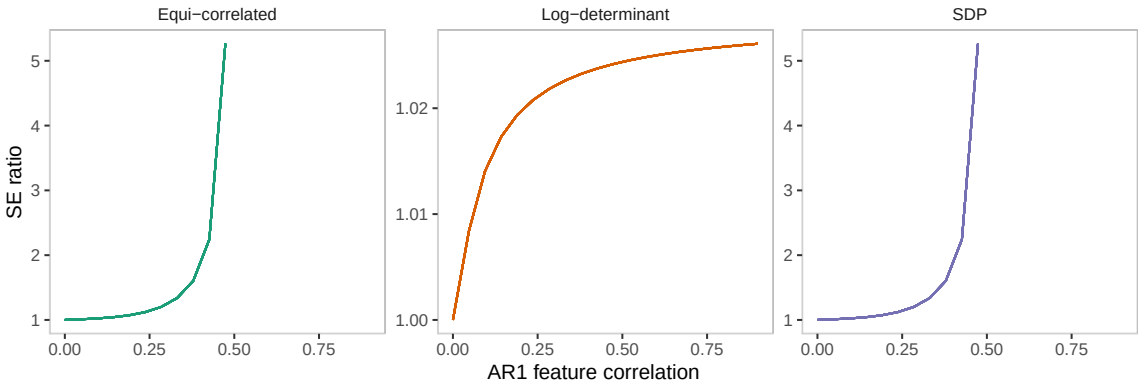


Figure 2.6: Ratio of augmented (knockoff) OLS standard errors to non-augmented OLS standard errors. Displayed for a Gram matrix $\mathbf{X}^\intercal\mathbf{X} = \mathbf{\Sigma}$ with an exact autoregressive structure and $p = 20$.

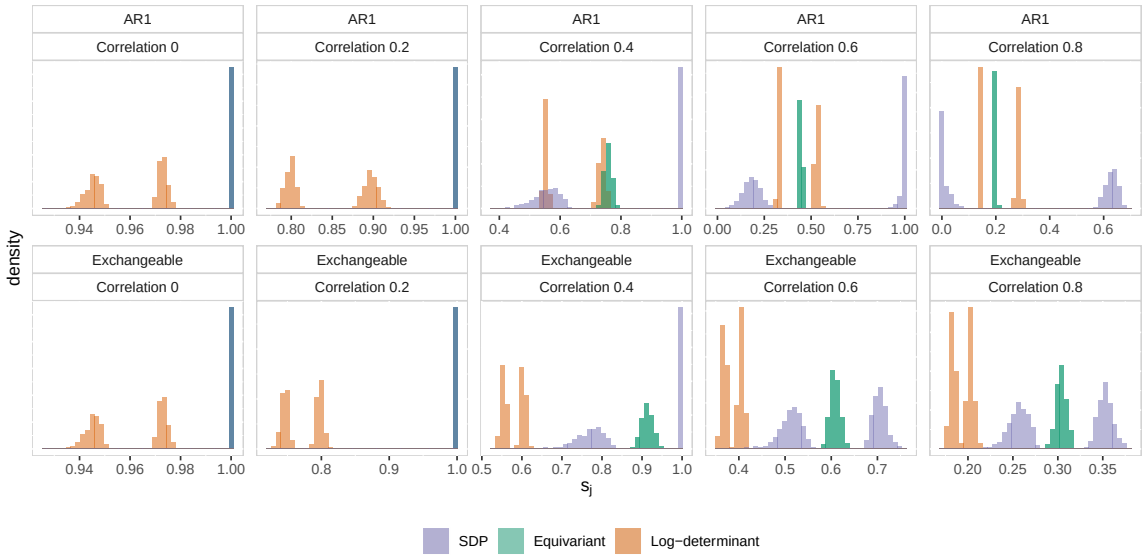
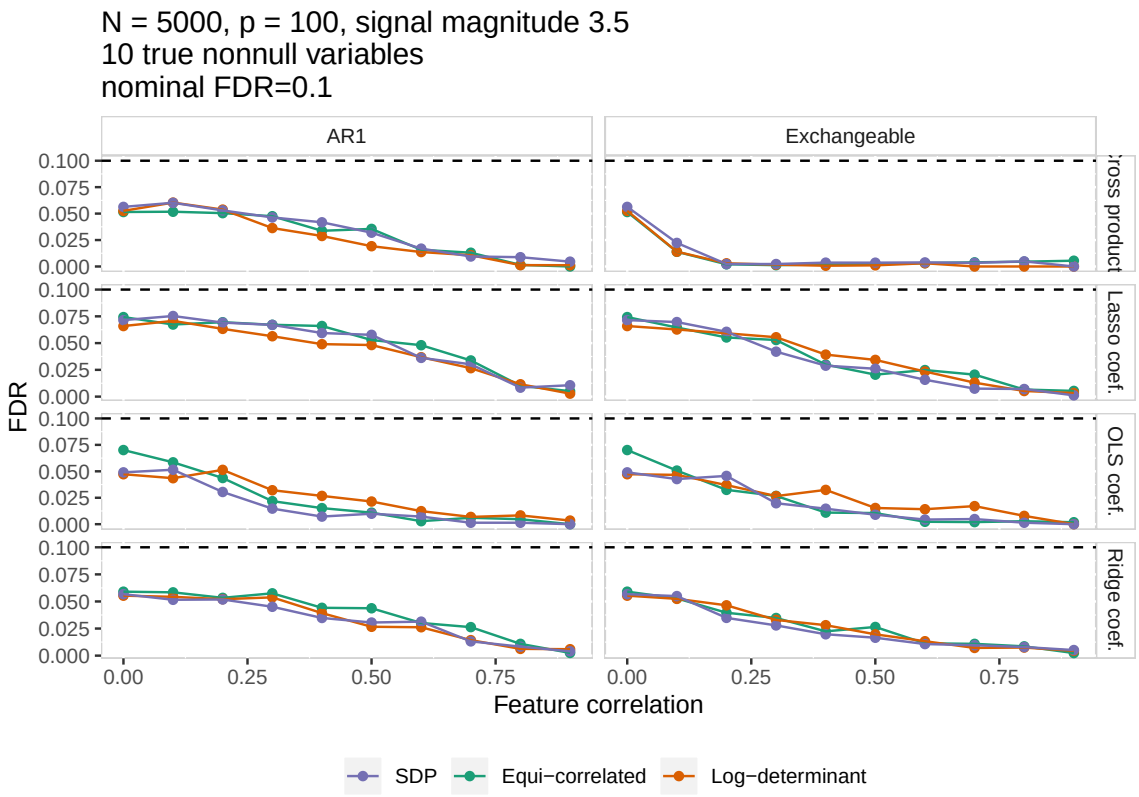


Figure 2.7: Empirical distribution of $s_{\min}, s_{\max}$ for Gaussian features with $p = 100$ and $n = 5000$. Each panel corresponds to the degree of feature correlation and contains 500 simulation replicates.



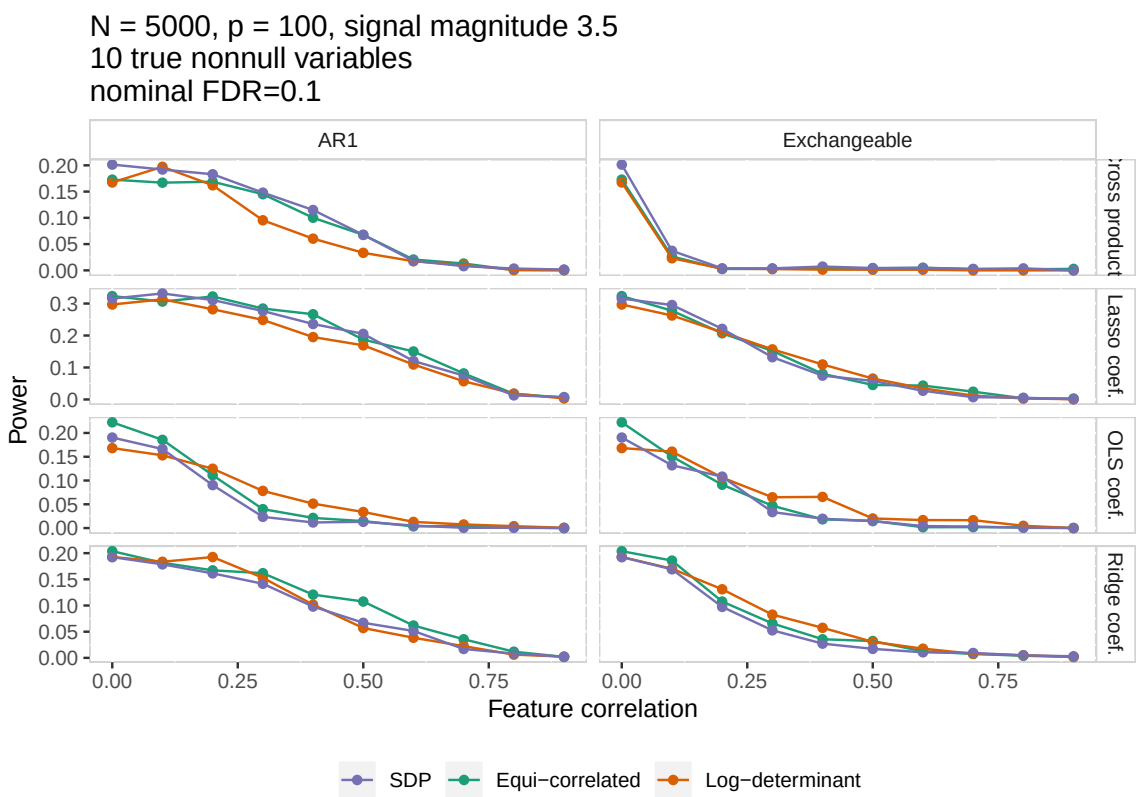
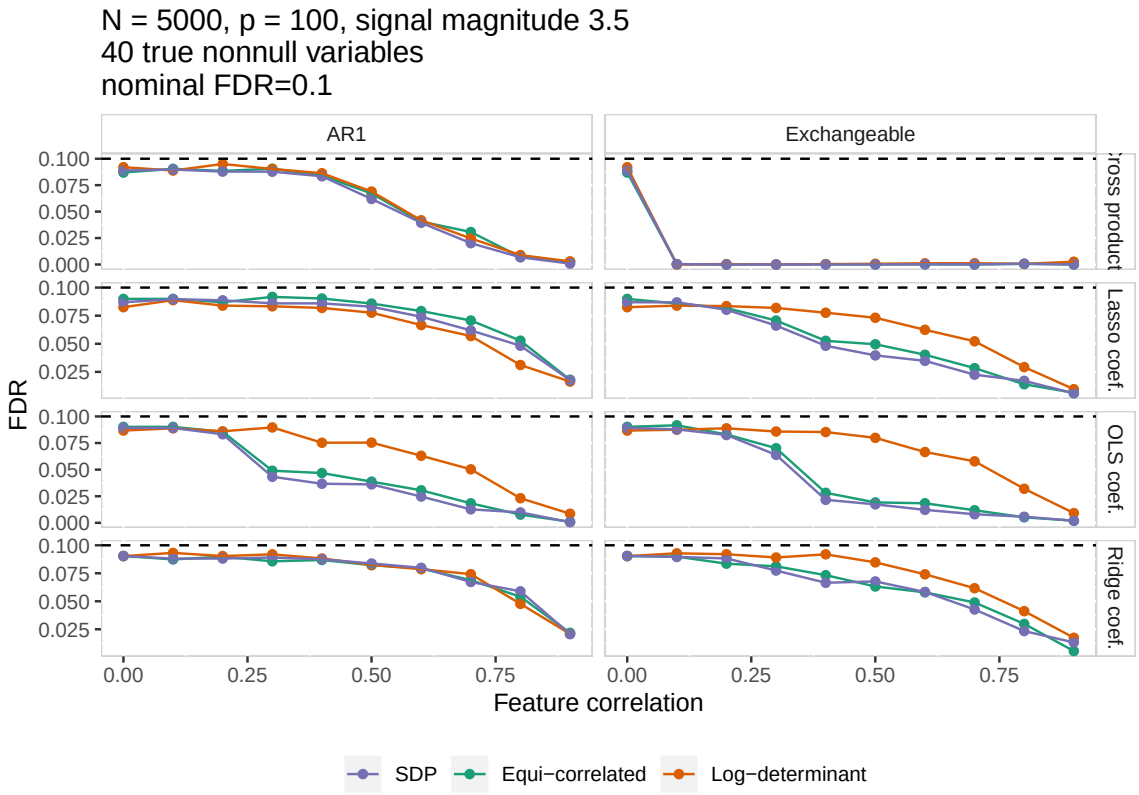


Figure 2.8: Power and FDR of each tuning method with $p = 100, k = 10$. Displayed as a function of feature correlation.



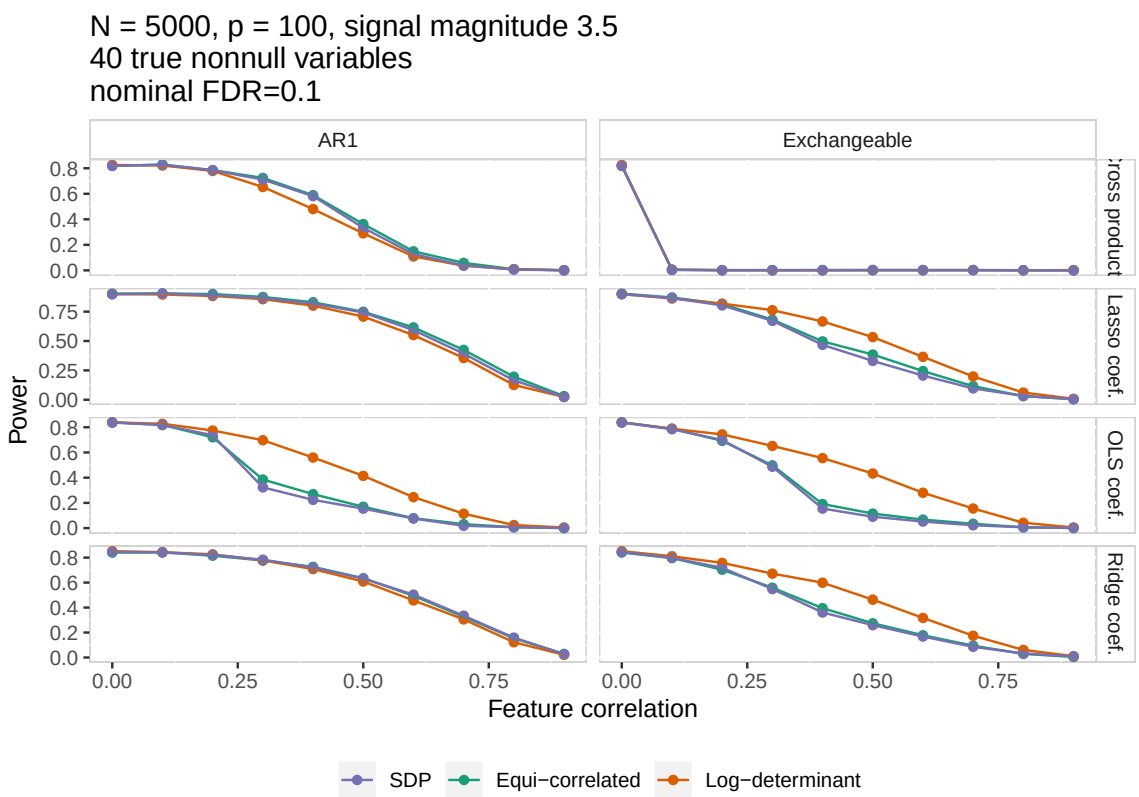


Figure 2.9: Power and FDR of each tuning method with $p = 100, k = 40$. Displayed as a function of feature correlation.

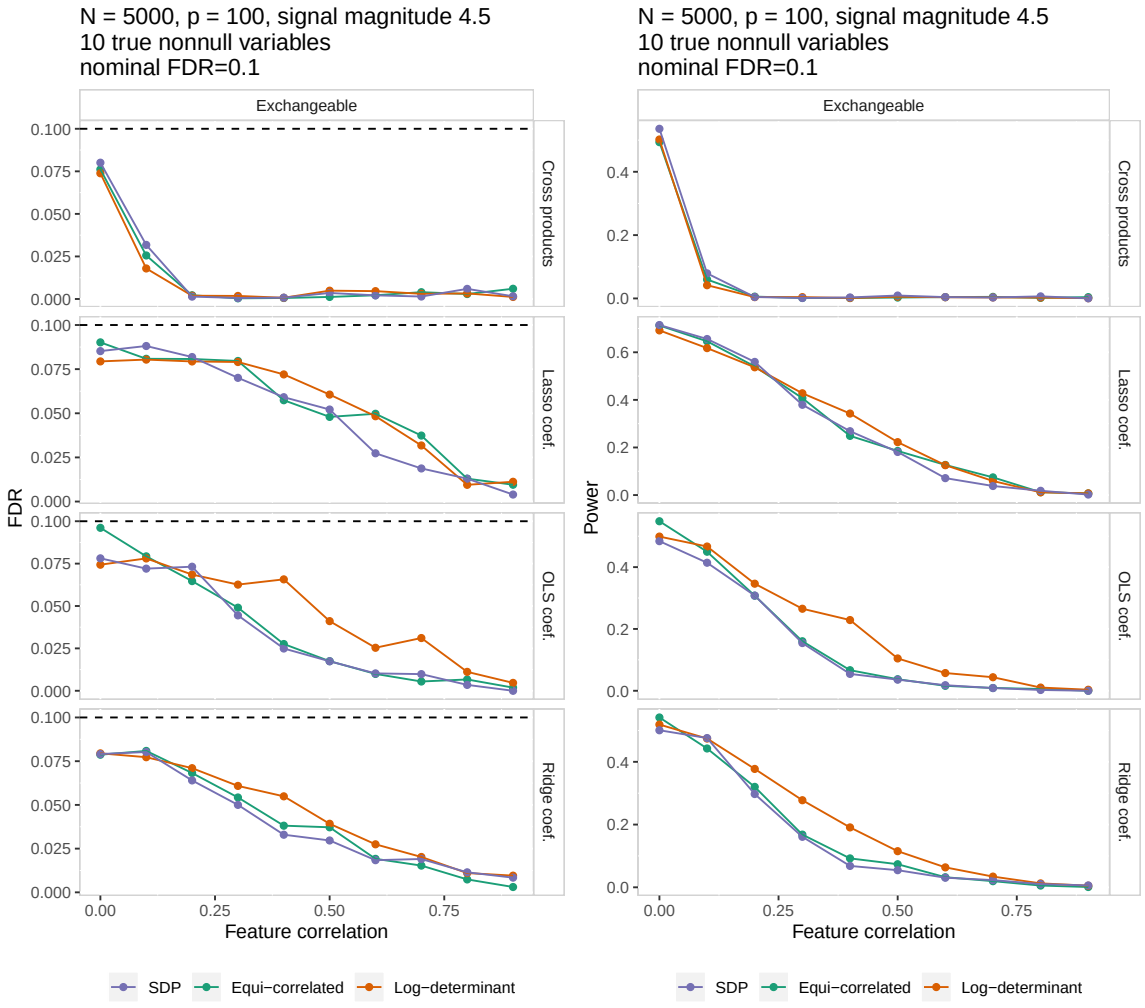


Figure 2.10: Power and FDR of each tuning method with $p = 100, k = 10, |\beta_j| = 4.5$. Displayed as a function of feature correlation.

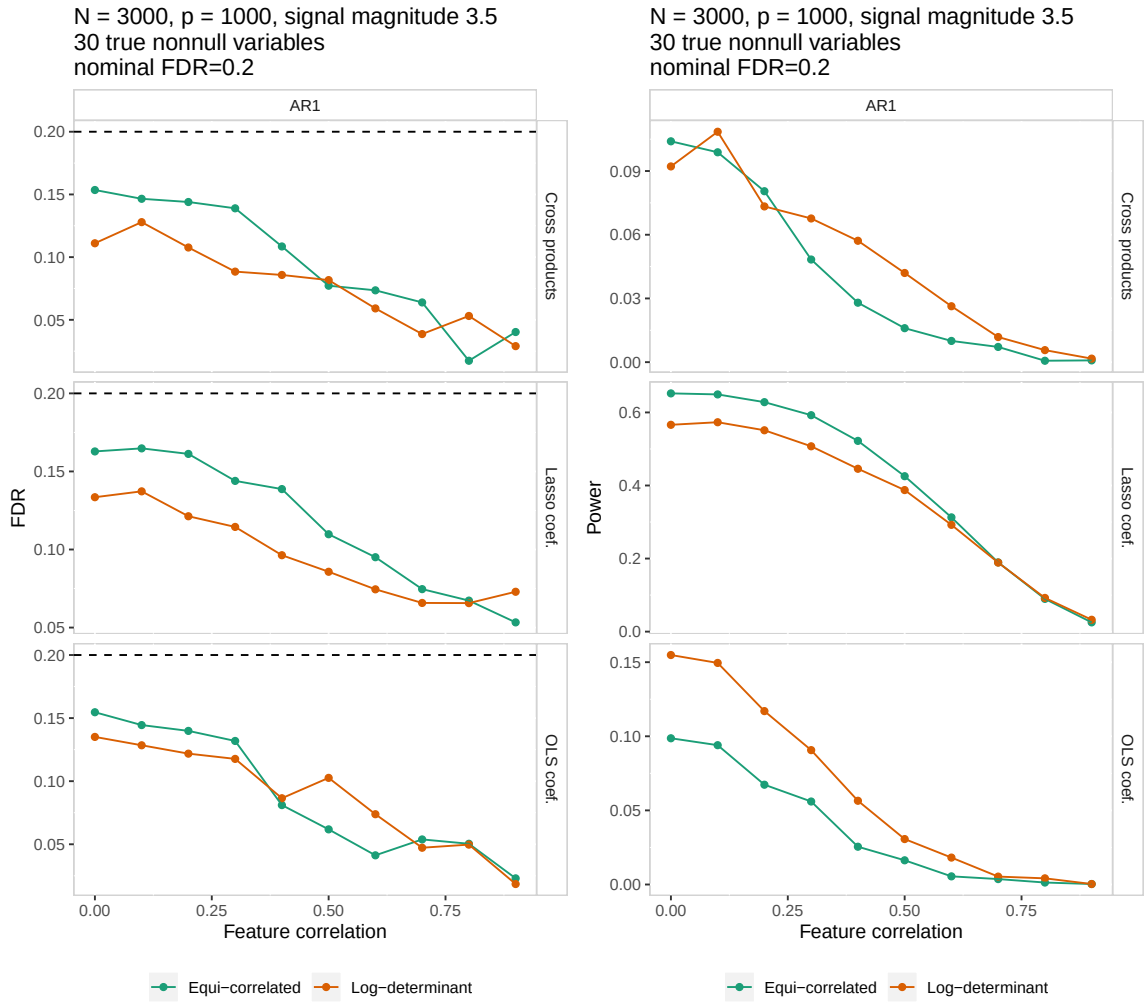


Figure 2.11: Power and FDR of each tuning method with $p = 1000, k = 30$. Displayed as a function of feature correlation.

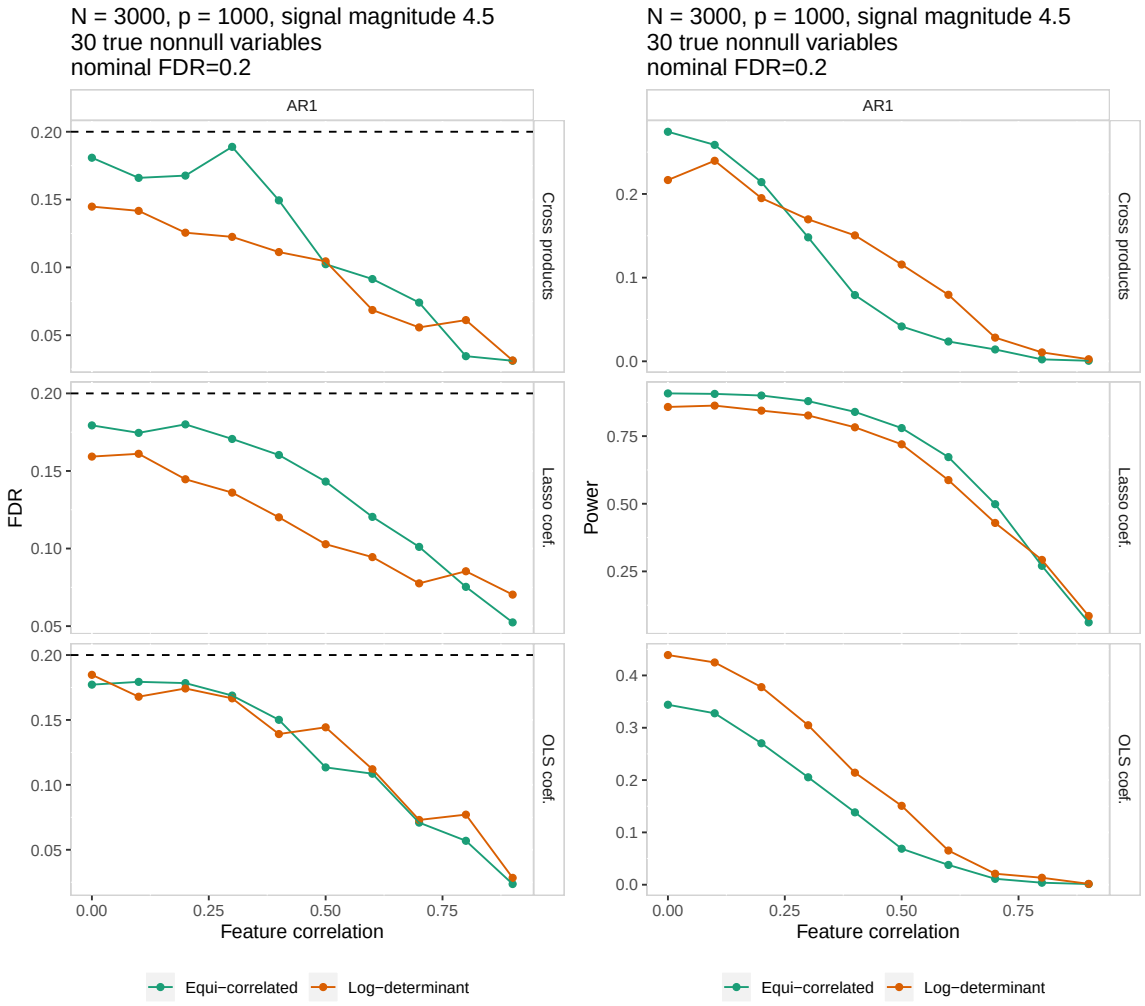


Figure 2.12: Power and FDR of each tuning method with $p = 1000, k = 30, |\beta_j| = 4.5$. Displayed as a function of feature correlation.

CHAPTER III

A Stabilized Knockoff Filter

3.1 Introduction

Performing variable selection while controlling the false discovery rate (FDR) can be seen as a way of ensuring replicability. In the context of linear regression with p covariates, an FDR-controlling variable selection method ensures that the subset of variables identified as having an association with the response will contain only a small proportion of variables whose population regression coefficient is, in fact, zero. This property exists on average over repeated sampling from the process generating the observed data. When repeatedly applying a FDR-controlling variable selection procedure to future studies of the same process, with measurements of the same candidate variables and response variable, one would hope for substantial overlap between the sets of selected variables across studies.

In the case of variable selection with the knockoff filter (Barber and Candès 2015), however, there exists an additional source of instability or non-replicability, above and beyond the inherent study-to-study variation due to repeated sampling from the same data-generating process. As described in Chapter II, the knockoff filter controls the FDR when selecting variables in the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is the n-dimensional response vector, $\mathbf{X}$ is the $n \times p$ fixed design matrix, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. A false discovery in this context is a selected variable j where, in fact,

$\beta_j = 0$. For a given, fixed data set, repeated applications of the knockoff filter can lead to wide variation in the number of variables selected and in the identity of the selected variables. The knockoff filter, restated in Algorithm III.1, constructs an $n \times p$ matrix of "knockoffs", $\tilde{\mathbf{X}}$, which are compared to the original covariates in order to select variables. Even with fixed $\mathbf{X}$ and $\mathbf{Y}$, there are many valid matrices of knockoffs variables which can be used in the knockoff filter, each of which can lead to a distinct set of selected variables.

A matrix of knockoffs for $\mathbf{X}$ involves an arbitrary choice in Step 3 of Algorithm III.1, in which $\tilde{\mathbf{U}}$ is chosen so that $\tilde{\mathbf{U}}^\intercal \mathbf{X} = \mathbf{0}$ and $\tilde{\mathbf{U}}^\intercal \tilde{\mathbf{U}} = \mathbf{I}$. The matrix of knockoffs is constructed via $\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\mathbf{S}) + \tilde{\mathbf{U}}\mathbf{C}$. The diagonal matrix $\mathbf{S}$, discussed in the previous chapter, is a p-dimensional tuning parameter fixed by a given tuning method (e.g. equivariant or log-determinant tuning) and design matrix. The matrix $\mathbf{C}$ is a function of $\mathbf{S}$ and $\mathbf{X}$. But the matrix of knockoffs $\tilde{\mathbf{X}}$, and hence the statistics $\mathbf{W}$, vary as functions of the arbitrary matrix $\tilde{\mathbf{U}}$. Variables are selected based on the vector $\mathbf{W}$, and this variation in $\mathbf{W}$ due to $\tilde{\mathbf{U}}$ leads to non-negligible instability in the set of selected variables.

To illustrate this instability, I generated a single design matrix $\mathbf{X}$ and a single response vector $\mathbf{Y}$ from the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $n = 5000$, $p = 100$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_n)$, and 10 non-null variables j with magnitude $|\beta_j| = 3.5$. The matrix $\mathbf{X}$ was generated with independent, mean-zero Gaussian rows with an autoregressive covariance structure (population correlation 0.4) in the columns. The matrix $\tilde{\mathbf{U}}$ was generated 500 times (as described below, in Algorithm III.3) to complete Step 3 of Algorithm III.1. For each $\tilde{\mathbf{U}}$, the knockoffs and importance statistics (lasso coefficient differences) were computed and variables were selected, as described in Steps 5–7 of Algorithm III.1, with target FDR $q = 0.1$. Figure 3.1 shows that for this fixed dataset, some truly null variables are selected in about 50 percent of trials while some truly non-null variables are selected in between 60 and 100 percent of trials. The

histogram in the right panel of Figure 3.1 shows that between zero and 16 variables were selected, with a mean of approximately 9.3 variables selected per trial (standard deviation approximately 3.3). This example shows how the arbitrary choice of $\tilde{\mathbf{U}}$ in the construction of the knockoff variables can lead to drastically different inferences when repeatedly applying the knockoff filter to a single fixed dataset.

Algorithm III.1 Fixed-design knockoff variable selection (Barber and Candès 2015)

Require: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{Y} \in \mathbb{R}^n$, $\|\mathbf{X}_j\| = 1$, $\mathbf{\Sigma} := \mathbf{X}^\intercal \mathbf{X}$, $q \in (0, 1)$

1: Choose $\mathbf{S} := \text{diag}(s_1, \ldots, s_p)$ so that all $s_j \geq 0$ and $2\mathbf{\Sigma} - \mathbf{S} \succeq 0$

2: Compute $\mathbf{C}$ satisfying $\mathbf{C}^\intercal \mathbf{C} = 2\mathbf{S} - \mathbf{S}\mathbf{\Sigma}^{-1}\mathbf{S}$

3: Choose $\tilde{\mathbf{U}} \in \mathbb{R}^{n \times p}$ so that $\tilde{\mathbf{U}}^\intercal \tilde{\mathbf{U}} = \mathbf{I}$ and $\tilde{\mathbf{U}}^\intercal \mathbf{X} = \mathbf{0}$.

4: $\tilde{\mathbf{X}} \leftarrow \mathbf{X}(\mathbf{I} - \mathbf{\Sigma}^{-1}\mathbf{S}) + \tilde{\mathbf{U}}\mathbf{C}$

5: Compute $\mathbf{W} = (W_1, \ldots, W_p)$ so that $W_j >> 0$ suggests $\beta_j \neq 0$, satisfying the following two properties:

(a) $\mathbf{W}$ is a function of $\left[\mathbf{X}\, \tilde{\mathbf{X}}\right]^\intercal \mathbf{Y}$ and $\left[\mathbf{X}\, \tilde{\mathbf{X}}\right]^\intercal \left[\mathbf{X}\, \tilde{\mathbf{X}}\right]$

(b) Swapping the columns $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ changes the sign of W_j

6: Compute threshold,

$$T_+ \leftarrow \min_j \left\{ t = |W_j| : \frac{1 + \#\{k : W_k \leq -t\}}{\max\{1, \#\{k : W_k \geq t\}\}} \leq q \right\} \text{ (knockoffs+)}$$

or

$$T \leftarrow \min_j \left\{ t = |W_j| : \frac{\#\{k : W_k \leq -t\}}{\max\{1, \#\{k : W_k \geq t\}\}} \leq q \right\} \text{ (knockoffs)}$$

7: Select variables $\{j : W_j \geq T\}$ (knockoffs) or $\{j : W_j \geq T_+\}$ (knockoffs+)

An FDR-controlling procedure such as that of Benjamini and Hochberg (1995) (referred to below as BH) does not have this additional source of indeterminacy for a fixed pair $(\mathbf{X}, \mathbf{Y})$. Given $(\mathbf{X}, \mathbf{Y})$ and a chosen test statistic, such as the Wald statistics for each ordinary least squares coefficient, the set of selected variables using the BH procedure is obtained from a deterministic function of the P-values computed from those test statistics. With the same design matrix and population parameters depicted in Figure 3.1, Figure 3.2 compares the distribution in the number of selected variables for the knockoff filter and BH procedure using ordinary least squares Wald tests across 500 samples of $\mathbf{Y} \mid \mathbf{X}$. The entries of $\mathbf{X}$ were chosen so that $\mathbf{X}^\intercal \mathbf{X} = \mathbf{I}$, so

the test statistics for BH were independent. The BH procedure selected an average of 9.1 variables, with a standard deviation of 2.0, while the knockoff filter selected an average of 9.7 variables, with a standard deviation of 3.4. For this design matrix there is nearly twice as much variation in the selected variables using the knockoff filter as there is using BH. With a single response vector $\mathbf{Y}$, there is no variation in the variables selected by BH, while the knockoff filter selects between one and 16 variables depending on the arbitrary choice of $\tilde{\mathbf{U}}$ (as depicted in Figure 3.1).

This chapter focuses on reducing this algorithmic instability in the knockoff filter so that there is little or no variation in the set of selected variables for a fixed dataset. There are at least three goals of reducing this instability in the knockoff filter: to improve power by reducing the likelihood of selecting very few variables due to an "unlucky" set of knockoffs; to enhance replicability; and to prevent misuses of the knockoff filter in which the procedure is performed repeatedly until a desired result is achieved. The rest of this chapter is organized as follows. After reviewing related work, I describe in Section 3.3 how instability in variable selections based on the knockoff filter results from indeterminacy in $\tilde{\mathbf{U}}$. Section 3.4 proposes a stabilized knockoff filter based on a collection of knockoffs generated for a single design matrix, where each set of knockoffs is constructed using a distinct matrix $\tilde{\mathbf{U}}$. Section 3.5.1 presents simulation results comparing the stabilized knockoff filter to the standard knockoff filter, and Section 3.5.2 presents simulations comparing the knockoff filter to variable selection with ordinary least squares P-values Benjamini-Hochberg or Bonferroni corrections for multiple comparisons.

3.2 Related work

In the high-dimensional setting with $p > n$, Barber and Candès (2019) extend the knockoff filter to control the "directional FDR", in which a false discovery is a selected variable j whose estimated sign is not equal to the true sign of β_j. This

definition of a false discovery includes type S errors (Gelman and Carlin 2014), in which a variable is identified as non-null but the estimated sign is incorrect, as well as the classical definition of a false discoveries. Su, Qian, and Liu (2015) aggregate knockoff statistics over a set of m independent linear models $\mathbf{Y}^{(i)} = \mathbf{X}^{(i)}\boldsymbol{\beta}^{(i)} + \boldsymbol{\epsilon}^{(i)}$, testing the hypotheses $H_{0j} : \beta_j^{(1)} = \cdots = \beta_j^{(m)}$ for $j = 1, \ldots, p$ and controlling a modified FDR using randomized decision rules. Here we focus on a single sample with population structure $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and $n > 2p$, and seek to reduce variance and maintain FDR control across repeated sampling of $\mathbf{Y} \mid \mathbf{X}$.

Other recent work on the knockoff filter has focused on the "model-$\mathbf{X}$" framework (Candès et al. 2018; Barber, Candès, and Samworth 2018), in which the rows of $\mathbf{X}$ are treated as independent random vectors whose probability distribution is known. In this framework, the knockoff variables $(\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_p)$ are constructed so that the distribution of $(\mathbf{X}, \tilde{\mathbf{X}})$ is unchanged when swapping original variables with their knockoff pairs and $\tilde{\mathbf{X}}$ is independent of $\mathbf{Y}$ given $\mathbf{X}$. This simple definition of model-$\mathbf{X}$ knockoffs belies the difficulty of actually computing these knockoffs for many covariate distributions, and ongoing research is tackling this problem (Sesia, Sabatti, and Candès 2018; Romano, Sesia, and Candès 2019, e.g.).

In this model-$\mathbf{X}$ framework, Roquero Gimenez and Zou (2018) discuss instability in the set of selected variables due to random sampling of the knockoff variables; the set of selected variables depends on a particular sample $\tilde{\mathbf{X}}$ (drawn from its population distribution), which is used to compute importance statistics and a threshold for those importance statistics. These authors define a set of "multi-knockoffs" $(\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_K)$, such that the joint distribution of $(\mathbf{X}, \tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_K)$ is exchangeable under permutations within a given feature, across the knockoff and original variables for that feature, and $(\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_K)$ is independent of $\mathbf{Y}$ given $\mathbf{X}$. Statistics estimating each variable's magnitude, such as the absolute lasso regression coefficients, are computed for each of the K knockoffs for each feature. These $K + 1$ statistics are used to generalize

the thresholding procedure for the importance statistics in the standard fixed-$\mathbf{X}$ and model-$\mathbf{X}$ knockoffs, thus obtaining a set of selected variables. As will be discussed below, this multi-knockoff thresholding cannot be directly applied in the fixed-$\mathbf{X}$ setting under discussion in this chapter.

The stabilized knockoff filter proposed in Section 3.4 involves aggregating vectors of test statistics, each of which could individually be used to select variables with FDR control. Pyne, Futcher, and Skiena (2006) study a related task in a different modeling context, describing how to aggregate independent P-values from several studies of a single set of genetic features. They focus on FDR control based on independent, identically distributed P-values from each experiment, which are screened to control the FDR within each experiment before being combined across experiments; the overall test statistic for each feature is the product of those P-values which met their experiment-specific cutoffs.

3.3 Unstable selection in fixed-X knockoffs

As noted previously, instability in the knockoff filter within a fixed dataset is a result of the arbitrary computation of $\tilde{\mathbf{U}}$ in Step 3 of Algorithm III.1. This matrix satisfies $\tilde{\mathbf{U}}^\intercal \mathbf{X} = \mathbf{0}$ and $\tilde{\mathbf{U}}^\intercal \tilde{\mathbf{U}} = \mathbf{I}$. These two properties of $\tilde{\mathbf{U}}$ are required to obtain the following structure in the $2p \times 2p$ Gram matrix

$$\mathbf{G} = \left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^\intercal \left[\mathbf{X}\ \tilde{\mathbf{X}}\right] = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathbf{S} \\ \boldsymbol{\Sigma} - \mathbf{S} & \boldsymbol{\Sigma} \end{bmatrix}. \tag{3.1}$$

Variable selection with the knockoff filter, for a single sample $(\mathbf{X}, \mathbf{Y})$, is performed using the importance statistics $\mathbf{W} = (W_1, \ldots, W_p)$, which are functions of $\mathbf{G}$ and the cross products $\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^\intercal \mathbf{Y}$. Instability in the selected variables due to $\tilde{\mathbf{U}}$ can therefore be understood by examining the effect of $\tilde{\mathbf{U}}$ on $\mathbf{W}$ via $\mathbf{G}$ and $\left[\mathbf{X}\,\tilde{\mathbf{X}}\right]^\intercal \mathbf{Y}$.

The requirements that $\tilde{\mathbf{U}}^\intercal\tilde{\mathbf{U}} = \mathbf{I}$, $\tilde{\mathbf{U}}^\intercal\mathbf{X} = \mathbf{0}$, and $\mathbf{C}^\intercal\mathbf{C} = 2\mathbf{S} - \mathbf{S}\boldsymbol{\Sigma}^{-1}\mathbf{S}$ can be seen as a solution to the following problem: find $n \times p$ matrices $\mathbf{A}, \mathbf{B}$ so that $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{A} + \mathbf{B}$ satisfies (3.1). Setting $\mathbf{A} = (\mathbf{I} - \boldsymbol{\Sigma}^{-1}\mathbf{S})$ and $\mathbf{B} = \tilde{\mathbf{U}}\mathbf{C}$ satisfies (3.1). With this construction, $\mathbf{G}$ does not depend on $\tilde{\mathbf{U}}$, and the structure of $\mathbf{G}$ is what permits the importance statistics to be used in estimating the number of false discoveries. Specifically, the knockoff filter is able to estimate the number of false discoveries at the threshold $t > 0$ because $\#\{j : \beta_j = 0, W_j \leq -t\}$ has the same distribution as $\#\{j : \beta_j = 0, W_j \geq t\}$ over repeated sampling of $\mathbf{Y} \mid \mathbf{X}$ (See Lemma 1 in Barber and Candès (2015) for details). The distribution of $\#\{j : \beta_j = 0, W_j \leq -t\}$ does not depend on $\tilde{\mathbf{U}}$ because the elements of $\left[\mathbf{X}\,\tilde{\mathbf{X}}\right]^\intercal\mathbf{Y}$ corresponding to $\beta_j = 0$ follow Gaussian distributions whose parameters are elements of $\mathbf{G}$ (See Lemmas 2 and 3 in Barber and Candès (2015)).

While $\mathbf{G}$ does not depend on $\tilde{\mathbf{U}}$, the cross products $\left[\mathbf{X}\;\tilde{\mathbf{X}}\right]^\intercal\mathbf{Y}$ do vary as a function of $\tilde{\mathbf{U}}$. Using the notation $\mathbf{A} = \mathbf{I} - \boldsymbol{\Sigma}^{-1}\mathbf{S}$, we can write

$$\left[\mathbf{X}\;\tilde{\mathbf{X}}\right]^\intercal\mathbf{Y} = \begin{pmatrix} \mathbf{X}^\intercal\mathbf{Y} \\ \mathbf{A}^\intercal\mathbf{X}^\intercal\mathbf{Y} + \mathbf{C}^\intercal\tilde{\mathbf{U}}^\intercal\boldsymbol{\epsilon} \end{pmatrix} \tag{3.2}$$

$$= \begin{pmatrix} \boldsymbol{\Sigma}\boldsymbol{\beta} + \mathbf{X}^\intercal\boldsymbol{\epsilon} \\ (\boldsymbol{\Sigma} - \mathbf{S})\boldsymbol{\beta} + ((\mathbf{I} - \mathbf{S}\boldsymbol{\Sigma}^{-1})\mathbf{X}^\intercal + \mathbf{C}^\intercal\tilde{\mathbf{U}}^\intercal)\boldsymbol{\epsilon} \end{pmatrix}. \tag{3.3}$$

For any fixed $\mathbf{X}$, we have that $\left[\mathbf{X}\;\tilde{\mathbf{X}}\right]^\intercal\mathbf{Y}$ is multivariate Gaussian and

$$\mathrm{Cov}\left(\mathbf{C}^\intercal\tilde{\mathbf{U}}^\intercal\boldsymbol{\epsilon}, \mathbf{X}^\intercal\mathbf{Y} \mid \mathbf{X}\right) = \mathrm{Cov}\left(\mathbf{C}^\intercal\tilde{\mathbf{U}}^\intercal\boldsymbol{\epsilon}, \mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} + \mathbf{A}^\intercal\mathbf{X}^\intercal\boldsymbol{\epsilon} \mid \mathbf{X}\right) \tag{3.4}$$

$$= \sigma^2\mathbf{C}^\intercal\tilde{\mathbf{U}}^\intercal\mathbf{X}\mathbf{A} = \mathbf{0} \tag{3.5}$$

Thus,

$$\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^{\intercal}\mathbf{Y} \overset{\text{distr.}}{=} \begin{pmatrix} \mathbf{X}^{\intercal}\mathbf{Y} \\ \mathbf{A}^{\intercal}\mathbf{X}^{\intercal}\mathbf{Y} + \mathbf{z} \end{pmatrix} \tag{3.6}$$

where $\mathbf{z} \sim N(\mathbf{0}_p, \sigma^2\mathbf{C}^{\intercal}\mathbf{C})$ is a Gaussian random vector generated independently of $\mathbf{X}^{\intercal}\mathbf{Y}$. Based on (3.3) or (3.6), one observes that for any fixed $(\mathbf{X}, \mathbf{Y})$, the importance statistics vary as a function of $\tilde{\mathbf{U}}$ through only the cross products $\left[\mathbf{X}\,\tilde{\mathbf{X}}\right]^{\intercal}\mathbf{Y}$. In addition, (3.6) shows that the *distribution* of $\left[\mathbf{X}\,\tilde{\mathbf{X}}\right]^{\intercal}\mathbf{Y}$, given $\mathbf{X}$, is the same for any $\tilde{\mathbf{U}}$. Whether the cross products are generated by sampling $\mathbf{z}$ in (3.6) or by calculating $\tilde{\mathbf{U}}$ in Algorithm III.1, the importance statistics $\mathbf{W}$ are random for any fixed $(\mathbf{X}, \mathbf{Y})$.

This discussion shows how the ability of the knockoff filter to control the FDR by estimating the number of false discoveries at a given threshold $t > 0$ is not affected by randomness in $\mathbf{W}$, for fixed $(\mathbf{X}, \mathbf{Y})$, due to the computation of $\left[\mathbf{X}\,\tilde{\mathbf{X}}\right]^{\intercal}\mathbf{Y}$ via $\tilde{\mathbf{U}}$ or (3.6). But since $\left[\mathbf{X}\,\tilde{\mathbf{X}}\right]^{\intercal}\mathbf{Y}$ is random, even for fixed $(\mathbf{X}, \mathbf{Y})$, this means that the knockoff filter outputs a random set of selected variables for any fixed dataset.

This chapter focuses on instability in the knockoff filter based on the computation of $\tilde{\mathbf{U}}$ in Algorithm III.1. Another strategy for stabilizing the knockoff filter, not pursued in this thesis, is to focus on the distribution of the cross products in (3.6). Before illustrating empirically instability in the knockoff filter due to the construction of $\tilde{\mathbf{U}}$, I will review specific numerical algorithms for computing $\tilde{\mathbf{U}}$.

In words, $\tilde{\mathbf{U}}$ consists of p orthonormal vectors from the left null space of $\mathbf{X}$. This means that $\tilde{\mathbf{U}}$ is an element of the Stiefel manifold, the set of $n \times p$ matrices $\mathbf{U}$ with $p \leq n$ and $\mathbf{U}^{\intercal}\mathbf{U} = \mathbf{I}$. In the knockoff filter, $\tilde{\mathbf{U}}$ is an element of the Stiefel manifold with the additional restriction that $\tilde{\mathbf{U}}^{\intercal}\mathbf{X} = \mathbf{0}$. Four possible methods of computing $\tilde{\mathbf{U}}$ are provided in III.2, III.3, III.4 and III.5.

In Algorithm III.2, $\tilde{\mathbf{U}}$ is equal to $\mathbf{Q}_0$ in the QR decomposition

$$[\mathbf{X}\ \mathbf{0}_{n\times p}] = [\mathbf{Q}_x\ \mathbf{Q}_0]\,\mathbf{R}, \tag{3.7}$$

where $[\mathbf{X}\ \mathbf{0}_{n\times p}]$ is the $n \times 2p$ column-wise concatenation of $\mathbf{X}$ and an $n \times p$ matrix of zeroes. With $\tilde{\mathbf{U}} = \mathbf{Q}_0$, we have $\tilde{\mathbf{U}}^\intercal\tilde{\mathbf{U}} = \mathbf{I}$ by definition of the QR decomposition, and $\tilde{\mathbf{U}}^\intercal\mathbf{X} = \mathbf{0}$ because

$$\begin{bmatrix} \mathbf{Q}_x^\intercal \\ \mathbf{Q}_0^\intercal \end{bmatrix} \begin{bmatrix} \mathbf{X} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_x^\intercal\mathbf{X} & \mathbf{Q}_x^\intercal\mathbf{0} \\ \mathbf{Q}_0^\intercal\mathbf{X} & \mathbf{Q}_0^\intercal\mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_x^\intercal\mathbf{X} & \mathbf{0} \\ \mathbf{Q}_0^\intercal\mathbf{X} & \mathbf{0} \end{bmatrix} = \mathbf{R}, \tag{3.8}$$

and $\mathbf{R}$ is upper triangular. This algorithm is deterministic in that for a fixed $\mathbf{X}$, a given numerical algorithm for the QR decomposition will return the same result when repeatedly computing (3.7). However, since $[\mathbf{X}\ \mathbf{0}]$ does not have full rank, its QR decomposition is not unique, so this algorithm provides a deterministic, but still arbitrary, choice of $\tilde{\mathbf{U}}$. In contrast, Algorithms III.3–III.5 compute $\tilde{\mathbf{U}}$ as a function of a randomly generated Gaussian matrix. Algorithm III.3 projects a random $n \times p$ Gaussian matrix away from the columns of $\mathbf{X}$ and then orthogonalizes the result. Algorithm III.4 performs the same computation as Algorithm III.2 and then rotates the result based on the QR decomposition of a random $p \times p$ Gaussian matrix. Finally, Algorithm III.5 draws a sample from the uniform distribution on the Stiefel manifold (Chikuse 2012, Theorem 2.2.1) before projecting away from $\mathbf{X}$ and orthogonalizing.

Algorithm III.2 Compute a deterministic $\tilde{\mathbf{U}}$

Require: $n \times p$ design matrix, $\mathbf{X}$

1: Perform the QR decomposition $[\mathbf{X}\ \mathbf{0}_{n\times p}] = [\mathbf{Q}_x\ \mathbf{Q}_0]\,\mathbf{R}$

2: $\tilde{\mathbf{U}} \leftarrow \mathbf{Q}_0$

Instability in the knockoff filter, that is, substantial variation in the set of selected variables, can be illustrated by fixing the design matrix and response vector and repeatedly applying the knockoff filter. The following simulation results fix $\mathbf{X}$ and $\mathbf{Y}$

Algorithm III.3 Compute random $\tilde{\mathbf{U}}$, method 1

Require: $n \times p$ design matrix, $\mathbf{X}$

1: Generate $\mathbf{Z}_{n\times p}$ with i.i.d. $N(0,1)$ entries.
2: Perform the QR decomposition $\mathbf{X} = \mathbf{Q}_x\mathbf{R}_x$
3: $\tilde{\mathbf{U}}^{(0)} \leftarrow (\mathbf{I} - \mathbf{Q}_x\mathbf{Q}_x^\intercal)\mathbf{Z}$
4: Perform the QR decomposition $\tilde{\mathbf{U}}^{(0)} = \mathbf{Q}_u\mathbf{R}_u$
5: $\tilde{\mathbf{U}} \leftarrow \mathbf{Q}_u$.

Algorithm III.4 Compute a random $\tilde{\mathbf{U}}$, method 2

Require: $n \times p$ design matrix, $\mathbf{X}$

1: Generate $\mathbf{Z}_{p\times p}$ with i.i.d. $N(0,1)$ entries.
2: Perform the QR decomposition $\mathbf{Z} = \mathbf{Q}_z\mathbf{R}_z$
3: Perform the QR decomposition $[\mathbf{X}\ \mathbf{0}_{n\times p}] = [\mathbf{Q}_x\ \mathbf{Q}_0]\,\mathbf{R}$
4: $\tilde{\mathbf{U}} \leftarrow \mathbf{Q}_0\mathbf{Q}_z$

Algorithm III.5 Compute a random $\tilde{\mathbf{U}}$, method 3

Require: $n \times p$ design matrix, $\mathbf{X}$

1: Generate $\mathbf{Z}_{n\times p}$ with i.i.d. $N(0,1)$ entries.
2: $\mathbf{U}_0 \leftarrow \mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z})^{-1/2}$
3: Perform the QR decomposition $\mathbf{X} = \mathbf{Q}_x\mathbf{R}_x$
4: Perform the QR decomposition $(\mathbf{I} - \mathbf{Q}_x\mathbf{Q}_x^\intercal)\mathbf{U}_0 = \mathbf{Q}_u\mathbf{R}_u$
5: $\tilde{\mathbf{U}} \leftarrow \mathbf{Q}_u$

and illustrate how the arbitrary choice of $\tilde{\mathbf{U}}$ leads to substantial variation in the set of variables selected by the knockoff filter. With $n = 5000$, $p = 100$, a single design matrix $\mathbf{X}$ was generated with mean-zero Gaussian rows. Given $\mathbf{X}$, a single response vector $\mathbf{Y}$ was drawn from $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\beta}$ had $k = 10$ nonzero elements, all of which had magnitude $|\beta_j| = 3.5$ or $|\beta_j| = 4.0$. In these simulations, $\mathbf{X}$ and $\mathbf{Y}$ are fixed and the knockoff filter is performed repeatedly. Since $\mathbf{X}$ and $\mathbf{Y}$ were fixed, $W_1, \ldots, W_p$, and therefore the set of selected variables, changed in each simulation replicate due only to the changing value of $\tilde{\mathbf{U}}$ in Algorithm III.1.

The third and fourth columns in Figure 3.3 display the distribution of the number of selected variables, for fixed $(\mathbf{X}, \mathbf{Y})$, over repeated generation of $\tilde{\mathbf{U}}$ using Algorithms III.3 and III.5. These histograms illustrate that repeated application of the knockoff filter to a single dataset can lead to wide variation in the resulting inferences (selected variable sets) due to the chosen value of $\tilde{\mathbf{U}}$. This variation exists across choices of importance statistics, although the specific degree of instability may depend on the chosen statistic (e.g. lasso coefficients or simple cross products). Figure 3.4 displays the frequency with which each null and non-null variable was selected across the 200 simulation replicates. Using a variable selection method without any indeterminacy for a fixed $(\mathbf{X}, \mathbf{Y})$, the set of selected variables would be exactly the same in each trial. Instead, when the effect magnitude is 4, ($|\beta_j| = 4$ for all non-null j), most of the non-null variables are selected approximately 40–50 percent of the time using Algorithms III.3 and III.5 to generate $\tilde{\mathbf{U}}$ in the knockoff filter. Some of the null variables are selected about 25 percent of the time. Figures 3.5 and 3.6 display similar histograms and variable-specific selection probabilities for a fixed design matrix whose rows were drawn from a Gaussian distribution with autoregressive covariance (population correlation 0.4). In this more challenging regression setting, similar within-dataset variation due to $\tilde{\mathbf{U}}$ is observed. The knockoff filter controls the FDR on average across sampling of $\mathbf{Y} \mid \mathbf{X}$, but in a given $(\mathbf{X}, \mathbf{Y})$ sample, one can

obtain an "unlucky" set of knockoffs due to the arbitrary choice of $\tilde{\mathbf{U}}$ and detect very few signals.

Figures 3.7–3.10 present similar simulation results, for fixed $(\mathbf{X}, \mathbf{Y})$, but with $k = 50$ nonzero β_j. In this non-sparse setting, many more features are selected, and the proportion of the time the non-null features are selected approaches one with stronger effect magnitude, uncorrelated features, and lasso importance statistics. However, there is still wide variation in the number of selected features and in the proportion of the time in which each null variable is selected, especially with correlated features (Figure 3.10). Figures 3.11–3.14 display similar results for $p = 1000$ and $n = 3000$ with $k = 30$ nonzero β_j. In this large-p, sparse setting, we can still observe wide variation in the probability of selecting a given variable and in the number of selected variables. In all of these simulations, we can also notice greater power to detect signals using the lasso importance statistics, and no difference in instability between Algorithms III.3 and III.5.

These simulated results show that repeatedly applying the knockoff filter to a single dataset can lead to wide variation in the set of selected variables, with an unknown probability that a given variable will be selected. That is, due only to randomness in the (arbitrary) choice of $\tilde{\mathbf{U}}$, the probability of selecting a given non-null variable is not equal to one or zero, as it would be with a deterministic procedure such as Benjamini-Hochberg. In its basic form, the knockoff filter requires us to choose a single, arbitrary $\tilde{\mathbf{U}}$ in order to control the FDR across repeated sampling of $\mathbf{Y}$ given $\mathbf{X}$. However, the desire to control FDR also suggests a desire to produce reliable or reproducible inferences, and this additional source of variation means that two analyses of the same data set using the knockoff filter can lead to very different conclusions.

The goal of this chapter is to maintain the average behavior of the knockoff filter, which controls FDR across sampling of $\mathbf{Y} \mid \mathbf{X}$, while reducing variance in the set of

selected variables. Secondarily, the potential to improve statistical power by reducing this source of variation will be explored. Specifically, the variation across repeated applications of the knockoff filter in the set of selected variables could provide additional information about which variables are likely to be null or non-null. Presumably, variables which appear more often in the selected sets, across repeated generation of $\tilde{\mathbf{U}}$, are more likely to be non-null variables.

Before developing the stabilized knockoff filter, described in the following section, I also explored whether it is possible to choose a single $\tilde{\mathbf{U}}$ for use in the knockoff filter in order to reduce variation in the selected variable set. These approaches included a deterministic (but arbitrary) computation of $\tilde{\mathbf{U}}$ (see Appendix A.1); a validation set approach in which a single $\tilde{\mathbf{U}}$ is chosen to maximize the number of selected variables on one half of the observations in $(\mathbf{X}, \mathbf{Y})$ (see Appendix A.2); and a choice of $\tilde{\mathbf{U}}$ designed to control the average geometric alignment between $\mathbf{Y}$ and a randomly computed $\tilde{\mathbf{U}}$ (see Appendix A.3). These approaches either did not reduce variation in the set of selected variables or reduced statistical power to near zero.

3.4 Stabilized knockoff filter

The previous section described how the knockoff filter selects a random set of variables with non-negligible variation even for a fixed dataset $(\mathbf{X}, \mathbf{Y})$. This randomness does not affect FDR control in the knockoff filter because the distribution of the importance statistics, conditional on $\mathbf{X}$, does not depend on the arbitrary computation of $\tilde{\mathbf{U}}$ in Algorithm III.1. The stabilized knockoff filter developed in this section takes advantage of the variation in the importance statistics across repeated applications of the knockoff filter in order to stabilize an estimated FDR-controlling threshold and to determine a consensus set of variable selections. Specifically, for fixed $(\mathbf{X}, \mathbf{Y})$, the stabilized knockoff filter generates $\tilde{\mathbf{U}}_b$, $b = 1, \ldots, B$ using Algorithm III.3, III.4 or III.5. For each $\tilde{\mathbf{U}}_b$, we compute the corresponding matrix of knockoffs, $\tilde{\mathbf{X}}_b = \mathbf{X}(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\mathbf{S}) + \tilde{\mathbf{U}}_b\mathbf{C}$, and the corresponding vector of importance statists, $\mathbf{W}_b$, as a function of $\left[\mathbf{X}\,\tilde{\mathbf{X}}_b\right]^\intercal \left[\mathbf{X}\,\tilde{\mathbf{X}}_b\right]$ and $\left[\mathbf{X}\,\tilde{\mathbf{X}}_b\right]^\intercal \mathbf{Y}$. The collection $\mathbf{W}_1, \ldots, \mathbf{W}_B$ is used to compute a stabilized FDR-controlling threshold and to output a set of variables j whose corresponding importance statistics W_j are estimated to exceed this threshold with high probability.

Before describing the stabilized knockoff procedure in greater detail, it is worth noting that simply averaging many $\tilde{\mathbf{U}}$ or $\tilde{\mathbf{X}}$ matrices for a single $(\mathbf{X}, \mathbf{Y})$, and then continuing with the knockoff filter as usual, will lead to a loss of FDR control. Similarly, averaging several vectors of importance statistics, each based on a distinct $\tilde{\mathbf{U}}$, also leads to a loss of FDR control. In addition, the multi-knockoff procedure (Roquero Gimenez and Zou 2018), developed for the model-$\mathbf{X}$ knockoff filter, cannot be directly applied in this fixed-$\mathbf{X}$ setting. For example, one fails to control the FDR by generating $\tilde{\mathbf{X}}^1, \ldots, \tilde{\mathbf{X}}^k$ from a set of k corresponding randomly generated $\tilde{\mathbf{U}}_1, \ldots, \tilde{\mathbf{U}}_k$ and then separately fitting each $2p$-dimensional regression in order to perform the multi-knockoff filter. This is likely because the multi-knockoff exchangeability results in Roquero Gimenez and Zou (2018) are not satisfied by $\left[\mathbf{X}, \tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_k\right]$. It may be

possible to define a fixed-$\mathbf{X}$ construction for $\tilde{\mathbf{X}}$ which leads to an analog of the Gaussian multi-knockoffs in Roquero Gimenez and Zou (2018) by constructing $\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_k$ so that the $(p+pk) \times (p+pk)$ Gram matrix $\left[\mathbf{X}\, \tilde{\mathbf{X}}_1 \ \ldots \ \tilde{\mathbf{X}}_k\right]^\intercal \left[\mathbf{X}\, \tilde{\mathbf{X}}_1 \ \ldots \ \tilde{\mathbf{X}}_k\right]$ has the following form:

$$\begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathbf{S} & \cdots & \boldsymbol{\Sigma} - \mathbf{S} \\ \boldsymbol{\Sigma} - \mathbf{S} & \ddots & & \\ \vdots & & & \\ \boldsymbol{\Sigma} - \mathbf{S} & \cdots & & \boldsymbol{\Sigma} \end{bmatrix} \tag{3.9}$$

This potential extension of the multi-knockoff filter to the fixed-$\mathbf{X}$ setting is not explored in this thesis.

The remainder of this chapter details the stabilized knockoff filter, which in simulation studies is shown to reduce variance in the set of selected variables and to improve statistical power in some settings. This stabilized knockoff filter involves two components: a stabilized threshold for the importance statistics and a stabilized set of selected variables given that threshold. This section adopts some of the notation and modeling ideas from Genovese and Wasserman (2004).

3.4.1 Stabilized FDR-controlling threshold

In Algorithm III.1, a single vector of importance statistics $(W_1, \ldots, W_p)$ is used to estimate the FDR, find a threshold T whose estimated FDR is below the nominal level q, and to select all variables j where $W_j \geq T$. As described previously, $W_1, \ldots, W_p$ will vary as a function of $\tilde{\mathbf{U}}$ for fixed $(\mathbf{X}, \mathbf{Y})$. So repeated sampling of $\tilde{\mathbf{U}}$ for a fixed dataset will lead to a different estimated FDR-controlling thresholds and different sets of selected variables. The proposed stabilized threshold for the knockoff filter takes advantage of this variation in $W_1, \ldots, W_p$ across repeated sampling of $\tilde{\mathbf{U}}$ to compute a low-variance threshold for the importance statistics.

For each variable $j = 1, \ldots, p$, let $H_j = \mathbb{I}[\beta_j \neq 0]$ be an indicator that the variable is a true signal. For a fixed threshold $t > 0$ and importance statistics $W_1, \ldots, W_p$, the knockoff filter selects variable j if $W_j \geq t$. So the FDR at threshold t is

$$\mathrm{FDR}(t) = \mathbb{E}\left(\frac{\sum_j \mathbb{I}[W_j \geq t](1 - H_j)}{\sum_j \mathbb{I}[W_j \geq t] + \mathbb{I}[\text{all } W_j < t]} \mid \mathbf{X}\right). \tag{3.10}$$

For a given $(\mathbf{X}, \mathbf{Y})$ sample, and a fixed value of $\tilde{\mathbf{U}}$, the knockoff filter uses

$$\widehat{\mathrm{FDP}}_{\mathrm{KO}}(t) = \frac{\sum_j \mathbb{I}[W_j \leq -t]}{\sum_j \mathbb{I}[W_j \geq t] + \mathbb{I}[\text{all } W_j < t]} \tag{3.11}$$

to estimate the FDP at threshold t. The knockoff filter threshold for the importance statistics is the smallest t such that $\widehat{\mathrm{FDP}}_{\mathrm{KO}}(t)$ is less than or equal to q, the desired FDR. The first step toward stabilizing this threshold is to use an alternative estimate for the FDR instead of (3.11) when computing a threshold Specifically, a conservative estimate (i.e. an upper bound) of the FDR is obtained as follows:

$$\begin{aligned}
\mathrm{FDR}(t) &= \mathbb{E}\left(\frac{\sum_j \mathbb{I}[W_j \geq t](1 - H_j)}{\sum_j \mathbb{I}[W_j \geq t] + \mathbb{I}[\text{all } W_j \leq t]} \mid \mathbf{X}\right) && (3.12)\\
&\approx \frac{\mathbb{E}\left(\sum_j \mathbb{I}[W_j \leq -t](1 - H_j) \mid \mathbf{X}\right)}{\mathbb{E}\left(\sum_j \mathbb{I}[W_j \geq t] \mid \mathbf{X}\right) + \mathbb{P}(\text{all } W_j \leq t \mid \mathbf{X})} && (3.13)\\
&\leq \frac{\mathbb{E}\left(\sum_j \mathbb{I}[W_j \leq -t](1 - H_j) \mid \mathbf{X}\right)}{\mathbb{E}\left(\sum_j \mathbb{I}[W_j \geq t] \mid \mathbf{X}\right)} && (3.14)\\
&\leq \frac{\mathbb{E}\left(\sum_j \mathbb{I}[W_j \leq -t] \mid \mathbf{X}\right)}{\mathbb{E}\left(\sum_j \mathbb{I}[W_j \geq t] \mid \mathbf{X}\right)} = \frac{\sum_j \mathbb{P}(W_j \leq -t \mid \mathbf{X})}{\sum_j \mathbb{P}(W_j \geq t \mid \mathbf{X})} && (3.15)
\end{aligned}$$

The first approximation is a result of computing the expectation of the numerator and denominator separately and applying the knockoff property that $\#\{j : W_j \leq -t\}$ estimates the number of false discoveries at threshold t. The final upper bound, (3.15), is obtained by noticing that H_j is either zero or one and $\mathbb{P}(\text{all } W_j \leq t \mid \mathbf{X})$ is

nonnegative, where

$$\mathbb{P}\left(\text{all } W_j \leq t \mid \mathbf{X}\right) := \mathbb{P}\left(W_1 \leq t, W_2 \leq t, \ldots, W_p \leq t \mid \mathbf{X}\right).$$

A stabilized threshold is found by using repeated samples of $\tilde{\mathbf{U}}$ to create a set of pseudo-samples of $W_1, \ldots, W_p$, which are then used to estimate (3.15). This stabilized estimate of an upper bound on the FDR is then used in place of $\widehat{\text{FDP}}_{\text{KO}}(t)$ when choosing a threshold to control the FDR. To obtain an estimate of (3.15), generate $\tilde{\mathbf{U}}_1, \ldots, \tilde{\mathbf{U}}_B$ using Algorithm III.3, III.4, or Algorithm III.5 and compute the corresponding vectors of importance statistics $\mathbf{W}_1, \ldots, \mathbf{W}_B$. Let W_{jb} be the importance statistic for the jth variable computed using $\tilde{\mathbf{U}}_b$. Define

$$\hat{\mathbb{P}}_B\left(W_j \geq t\right) := \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left[W_{jb} \geq t\right]. \tag{3.16}$$

Given $W_{1b}, \ldots, W_{pb}, b = 1, \ldots, B$ an estimate of (3.15) is

$$\frac{\sum_j \mathbb{P}\left(W_j \leq -t \mid \mathbf{X}\right)}{\sum_j \mathbb{P}\left(W_j \geq t \mid \mathbf{X}\right)} \approx \frac{\sum_j \hat{\mathbb{P}}_B\left(W_j \leq -t\right)}{\sum_j \hat{\mathbb{P}}_B\left(W_j \geq t\right)} = \frac{\frac{1}{B}\sum_{b=1}^{B}\sum_j \mathbb{I}\left[W_{jb} \leq -t\right]}{\frac{1}{B}\sum_{b=1}^{B}\sum_j \mathbb{I}\left[W_{jb} \geq t\right]} \tag{3.17}$$

$$:= \widehat{\text{FDR}}_B(t). \tag{3.18}$$

The stabilized threshold $\overline{T}$ is then

$$\overline{T} := \min\left\{t : \widehat{\text{FDR}}_B(t) \leq q\right\}. \tag{3.19}$$

Figure 3.15 displays $\widehat{\text{FDP}}_{\text{KO}}(t)$, $\widehat{\text{FDR}}_B(t)$ and $\text{FDR}(t)$ for four fixed $(\mathbf{X}, \mathbf{Y})$ pairs generated from the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ using the difference in lasso coefficient magnitudes as importance statistics. For either $\widehat{\text{FDP}}_{\text{KO}}(t)$ or $\widehat{\text{FDR}}_B(t)$, a selection threshold is chosen by finding the intersection of each curve with a horizontal line at the desired FDR (e.g. 0.1 or 0.2). The threshold $\overline{T}$ is stabilized to the extent that $\widehat{\text{FDR}}_B(t)$, its

corresponding estimate of the FDR, is less variable than $\widehat{\text{FDP}}_{\text{KO}}(t)$. These examples suggest that $\widehat{\text{FDR}}_B(t)$ is a smooth version of the individual $\widehat{\text{FDP}}_{\text{KO}}(t)$ computed from a single vector of importance statistics (based on a single $\tilde{\mathbf{U}}$).

Figure 3.16 compares the mean and pointwise variance of $\widehat{\text{FDR}}_B(t)$ and $\widehat{\text{FDP}}_{\text{KO}}(t)$ to $\frac{\sum_j \mathbb{P}(W_j \leq -t|\mathbf{X})}{\sum_j \mathbb{P}(W_j \geq t|\mathbf{X})}$ and $\text{FDR}(t)$. These were computed from 200 samples of $(\mathbf{X}, \mathbf{Y})$ with $\mathbf{X}$ having mean-zero Gaussian rows and $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I})$. Figure 3.16 illustrates that

$$\text{FDR}(t) \leq \mathbb{E}\left(\widehat{\text{FDR}}_B(t)\right) \approx \mathbb{E}\left(\widehat{\text{FDP}}_{\text{KO}}(t)\right) \approx \frac{\sum_j \mathbb{P}\left(W_j \leq -t \mid \mathbf{X}\right)}{\sum_j \mathbb{P}\left(W_j \geq t \mid \mathbf{X}\right)}, \tag{3.20}$$

and that $\widehat{\text{FDR}}_B(t)$ has lower variance, across repeated sampling of $(\mathbf{X}, \mathbf{Y})$, than $\widehat{\text{FDP}}_{\text{KO}}(t)$. These two properties allow $\widehat{\text{FDR}}_B(t)$ to be used as an FDR estimate which leads to a stabilized threshold for the importance statistics. First, the quantity $\frac{\sum_j \mathbb{P}(W_j \leq -t|\mathbf{X})}{\sum_j \mathbb{P}(W_j \geq t|\mathbf{X})}$ is an approximate upper bound for the true FDR, and $\mathbb{E}\left(\widehat{\text{FDR}}_B(t)\right)$ is approximately equal to this upper bound. Second, $\widehat{\text{FDR}}_B(t)$ has lower variance than $\widehat{\text{FDP}}_{\text{KO}}(t)$, meaning that a threshold derived from $\widehat{\text{FDR}}_B(t)$ will have lower variance than a threshold derived from $\widehat{\text{FDP}}_{\text{KO}}(t)$.

Averaged over datasets, $\widehat{\text{FDP}}_{\text{KO}}(t)$ and $\widehat{\text{FDR}}_B(t)$ provide conservative estimates of the true FDR and hence can be used to find an FDR-controlling threshold for the importance statistics. These empirical results suggest that $\widehat{\text{FDR}}_B(t)$ has the same population mean as $\widehat{\text{FDP}}_{\text{KO}}(t)$ but has lower variance. Both $\widehat{\text{FDP}}_{\text{KO}}(t)$ and $\widehat{\text{FDR}}_B(t)$ can be thought of as estimates of $\frac{\sum_j \mathbb{P}(W_j \leq -t|\mathbf{X})}{\sum_j \mathbb{P}(W_j \geq t|\mathbf{X})}$. Given a single W_j for each variable, $\widehat{\text{FDP}}_{\text{KO}}(t)$ estimates the marginal probability $\mathbb{P}\left(W_j \leq -t \mid \mathbf{X}\right)$ with an indicator function, while $\widehat{\text{FDR}}_B(t)$ those marginal probabilities using the empirical distribution function of the B pseudo-samples $W_{j1}, \ldots, W_{jB}$.

3.4.2 Stabilized variable selection

Given a threshold $t > 0$, the knockoff filter selects all variables j where $W_j \geq t$. This selection is based on the single vector $\mathbf{W}$ computed using a single matrix of knockoffs. As in the construction of a stabilized threshold, here I propose a way of stabilizing the set of selected variables at a fixed threshold t by taking advantage of variation in $\mathbf{W}$ across repeated sampling of $\tilde{\mathbf{U}}$.

Suppose that, based on (3.15), the threshold $t > 0$ controls the FDR. That is, we have

$$\frac{\sum_j \mathbb{P}\left(W_j \leq -t \mid \mathbf{X}\right)}{\sum_j \mathbb{P}\left(W_j \geq t \mid \mathbf{X}\right)} \leq q.$$

In this case, $\sum_j \mathbb{P}\left(W_j \geq t \mid \mathbf{X}\right)$, the expected number of selections at the threshold t, is calibrated to control the FDR. In a single application of the knockoff filter, the number of selected variables is $\sum_j \mathbb{I}\left[W_j \geq t\right]$. If we could reduce the variance of $\sum_j \mathbb{I}\left[W_j \geq t\right]$, without changing its expected value, the number of selected variables would be stabilized without losing FDR control.

The proposed stabilized variable selection is based on two separate steps: fixing $\hat{k}$, the number of selected variables, to be equal to an estimate of the expected number of selections at the FDR-controlling threshold t; and then identifying the $\hat{k}$ variables to be selected in order to maximize power.

The first step is to estimate

$$V(t) := \mathbb{E}\left(\sum_j \mathbb{I}\left[W_j \geq t\right] \mid \mathbf{X}\right) = \sum_j \mathbb{P}\left(W_j \geq t \mid \mathbf{X}\right), \tag{3.21}$$

the expected number of selections at the threshold t. As with the stabilized threshold, the marginal probabilities $\mathbb{P}\left(W_j \geq t \mid \mathbf{X}\right)$ are estimated using the B pseudo-samples $W_{j1}, \ldots, W_{jB}$, where each W_{jb} is computed from a different sample of $\tilde{\mathbf{U}}_b$. This

provides the following estimate for $V(t)$:

$$\hat{V}_B(t) := \frac{1}{B} \sum_b \sum_j \mathbb{I}\left[W_{jb} \geq t\right] \tag{3.22}$$

In the ideal case in which $V(t)$ were known, we could select exactly $V(t)$ variables in every sample. Since we have already stipulated that t is an FDR-controlling threshold, selecting $V(t)$ variables with probability one would maintain FDR control while reducing the variance of the number of selected variables to zero. If $\hat{V}_B(t)$ is a good estimate of $V(t)$, then we can potentially obtain significant reductions in the variance of the number of selected variables simply by increasing B, the number of samples of $\tilde{\mathbf{U}}$. This should maintain FDR control, since we are not changing the expected number of selections at a given threshold, and reduce variance.

Once the number of variables to be selected is fixed at $\hat{V}_B(t)$, the $\hat{V}_B(t)$ variables must be chosen among $\mathbf{X}_1, \ldots, \mathbf{X}_p$, the columns of $\mathbf{X}$. Suppose we had access to independent, identically distributed $\mathbf{W}_1, \mathbf{W}_2, \ldots$ based on repeated sampling from $\mathbf{Y} \mid \mathbf{X}$. Additionally, suppose that $t > 0$ satisfies (3.15) and $V(t)$ is known. How can we select $V(t)$ variables to maximize power? Consider again $H_j = \mathbb{I}\left[\beta_j \neq 0\right]$ and suppose there are prior probabilities (or frequencies of non-null and null variables) $\mathbb{P}\left(H_j = 1\right) = \pi_1$ and $\mathbb{P}\left(H_j = 0\right) = \pi_0$. We obtain high power to detect true signals when selecting variable j based on $\mathbb{I}\left[W_j \geq t\right]$ indicates $H_j = 1$ with high probability. In other words, when

$$\mathbb{P}\left(H_j = 1 \mid W_j \geq t, \mathbf{X}\right) \tag{3.23}$$

is close to one, the variable selection rule has high power. Maximizing this posterior

probability leads to

$$\arg\max_j \mathbb{P}\left(H_j = 1 \mid W_j \geq t, \mathbf{X}\right) = \arg\max_j (1 - \mathbb{P}\left(H_j = 0 \mid W_j \geq t, \mathbf{X}\right)), \tag{3.24}$$

$$\mathbb{P}\left(H_j = 0 \mid W_j \geq t, \mathbf{X}\right) = \frac{\mathbb{P}\left(W_j \geq t \mid H_j = 0, \mathbf{X}\right)\mathbb{P}\left(H_j = 0\right)}{\mathbb{P}\left(W_j \geq t \mid \mathbf{X}\right)}. \tag{3.25}$$

Suppose that $\pi_0 = \mathbb{P}\left(H_j = 0\right)$ does not depend on j (or simply that there is a fixed fraction of nulls) and that the distribution of W_j is the same for the null j, i.e. that $\mathbb{P}\left(W_j \geq t \mid H_j = 0, \mathbf{X}\right)$ does not depend on j. Then

$$\arg\max_j \mathbb{P}\left(H_j = 1 \mid W_j \geq t, \mathbf{X}\right) = \arg\min_j \frac{\mathbb{P}\left(W_j \geq t \mid H_j = 0, \mathbf{X}\right)\mathbb{P}\left(H_j = 0\right)}{\mathbb{P}\left(W_j \geq t \mid \mathbf{X}\right)} \tag{3.26}$$

$$= \arg\min_j \frac{1}{\mathbb{P}\left(W_j \geq t \mid \mathbf{X}\right)} \tag{3.27}$$

$$= \arg\max_j \mathbb{P}\left(W_j \geq t \mid \mathbf{X}\right). \tag{3.28}$$

This argument suggests that, for a fixed number of variables to be selected, we should select those variables with the largest values of $\mathbb{P}\left(W_j \geq t \mid \mathbf{X}\right)$. An analogy can be made here with the empirical Bayes interpretation of FDR and the Benjamini-Hochberg procedure (Efron 2010, Ch. 4). The probability $\mathbb{P}\left(H_j = 0 \mid W_j \geq t, \mathbf{X}\right)$ is called the *Bayes false discovery rate* in Efron (2010), Here, the W_j are analogous to the z statistics or P-values in the Benjamini-Hochberg procedure. Given $V(t)$, the fixed number of variables to be selected at threshold t, selecting variables based on (3.26)–(3.28) is equivalent to selecting variables with minimum posterior probability of being null, assuming that the test statistics W_j have the same marginal distribution when $\beta_j = 0$ (i.e. when the null hypothesis is true). ‘

The stabilized knockoff filter uses the estimate $\hat{V}_B(t)$ as the fixed number of variables to be selected, and the marginal probabilities $\mathbb{P}\left(W_j \geq t \mid \mathbf{X}\right)$ are again estimated

using $\hat{\mathbb{P}}_B(W_j \geq t)$. Let $\tau_1, \ldots, \tau_p$ be the permutation of $1, \ldots, p$ so that

$$\hat{\mathbb{P}}_B(W_{\tau_1} \geq t) \geq \hat{\mathbb{P}}_B(W_{\tau_2} \geq t) \geq \cdots \geq \hat{\mathbb{P}}_B\left(W_{\tau_p} \geq t\right), \tag{3.29}$$

where, as before, $\hat{\mathbb{P}}_B(W_j \geq t) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}[W_{jb} \geq t]$. Then the final variable selections are the first $\hat{V}_B(t)$ elements of $\tau_1, \ldots, \tau_p$ (where $\hat{V}_B(t)$ is rounded to the nearest integer).

The stabilized knockoff filter consists of combining the stabilized threshold $\overline{T}$ (Section 3.4.1, equation (3.19)) with the stabilized selection procedure outlined above. That is, we select $\hat{V}_B(\overline{T})$ variables, an estimate of the expected number of selections at the stabilized threshold $\overline{T}$. The $\hat{V}_B(\overline{T})$ variables selected are those with the largest values of $\hat{\mathbb{P}}_B\left(W_j \geq \overline{T}\right)$. This stabilized knockoff filter is summarized in Algorithm III.6.

Algorithm III.6 Stabilized knockoff filter

Require: $\mathbf{X}, \mathbf{Y}$, integer B, nominal FDR $q \in (0, 1)$

1: Compute $\tilde{\mathbf{U}}_1, \ldots, \tilde{\mathbf{U}}_B$ from B repetitions of Algorithm III.3, III.4, or III.5.

2: Compute knockoffs $\tilde{\mathbf{X}}_1, \ldots \tilde{\mathbf{X}}_B$ and importance statistics $\mathbf{W}_1, \ldots, \mathbf{W}_B$ based on $\tilde{\mathbf{U}}_1, \ldots, \tilde{\mathbf{U}}_B$ via steps 4–5 of Algorithm III.1.

3: Compute

$$\overline{T} = \min\left\{ t : \frac{\frac{1}{B}\sum_{b=1}^{B}\sum_j \mathbb{I}[W_{jb} \leq -t]}{\frac{1}{B}\sum_{b=1}^{B}\sum_j \mathbb{I}[W_{jb} \geq t]} \leq q \right\}$$

4: Compute $\hat{V}_B(\overline{T}) = \frac{1}{B}\sum_b \sum_j \mathbb{I}\left[W_{jb} \geq \overline{T}\right]$

5: Define

$$\hat{\mathbb{P}}_B(W_j \geq t) := \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}[W_{jb} \geq t]$$

and the permutation $\tau_1, \ldots, \tau_p$ of $1, \ldots, p$ so that $\hat{\mathbb{P}}_B\left(W_{\tau_1} \geq \overline{T}\right) \geq \hat{\mathbb{P}}_B\left(W_{\tau_2} \geq \overline{T}\right) \geq \cdots \geq \hat{\mathbb{P}}_B\left(W_{\tau_p} \geq \overline{T}\right)$

6: Output $\tau_1, \ldots, \tau_{\texttt{round}(\hat{V}_B(\overline{T}))}$

3.5 Simulation Results

3.5.1 Comparison with knockoff filter

Performance of the stabilized knockoff filter was assessed using similar simulation scenarios as in Barber and Candès (2015). Rows of the design matrix $\mathbf{X}$ were drawn from a mean-zero multivariate Gaussian distribution with population covariance equal to the identity (i.e. uncorrelated features) or an autoregressive covariance matrix with (i, j)th entry equal to $\rho^{|i-j|}$, $i, j \in \{1, \ldots, p\}$. Given $\mathbf{X}$, the response was generated from $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_n)$. The k nonzero elements of $\boldsymbol{\beta}$, all had magnitude $|\beta_j| = 3.5$ unless otherwise noted.

First, as in Section 3.3, we should examine whether the stabilized knockoff filter reduces instability in the set of selected variables for a fixed $(\mathbf{X}, \mathbf{Y})$ pair. In Figures 3.3–3.14, it is shown that the stabilized knockoff filter can substantially reduce variation in the selected variables for a fixed $(\mathbf{X}, \mathbf{Y})$ dataset even with only $B = 50$ or $B = 100$ pseudo-samples of $\mathbf{W}$. In Figure 3.6, for example, with correlated features, $k = 10$ truly non-null signals out of $p = 100$, and signal magnitude $|\beta_j| = 4$, the stabilized knockoff filter with lasso importance statistics selects nine out of ten of these true signals in all of the simulation replicates; with the standard knockoff filter, these true signals are selected in about 50 percent of the trials. Qualitatively similar results are observed with $k = 50$ and $p = 100$, or $p = 1000$ and $k = 30$. In all of these fixed-$(\mathbf{X}, \mathbf{Y})$ simulations, correlated features or weaker signal magnitude leads to fewer selections using any method, but the stabilized knockoff filter outperforms the standard knockoff filter—by selecting non-null variables in a larger proportion of trials and reducing the variance in the number of selected variables—in all scenarios.

Average behavior of the stabilized knockoff filter and the standard knockoff filter, over repeated sampling of $(\mathbf{X}, \mathbf{Y})$, is displayed in Figures 3.17–3.28. The knockoff+ and knockoff filter are displayed separately; recall that the knockoff+ threshold (see

Algorithm III.1) controls the FDR while the standard knockoff threshold controls an approximate FDR. In the $p = 100$ and $p = 1000$ simulation settings, the nominal FDR was set to $q = 0.1$ and $q = 0.2$, respectively. With $B = 50$ or $B = 100$, the stabilized knockoff filter controls the FDR in all simulation scenarios, and reduces the standard deviation in the number of selected variables by a factor of two or three in some scenarios. In nearly all simulation scenarios, the power of the stabilized knockoff filter is at least as large as that of the knockoff filter. In many simulation scenarios, it is necessary to use the knockoff+ threshold to control the FDR with the knockoff filter, but this threshold leads to very low power. The stabilized knockoff filter has at least as much power as the knockoff filter with its standard threshold, but this standard threshold often fails to control the FDR. Thus, the stabilized knockoff filter can provide similar statistical power as the knockoff filter with its standard threshold while controlling the FDR and reducing the standard deviation in the number of selected variables.

In the $n = 5000, p = 100$ setting with $k = 10$ non-null variables, the stabilized knockoff filter has at least twice as much power as the knockoff filter with the knockoff+ threshold with a small or moderate degree of feature correlation (Figure 3.17). In this setting, the knockoff filter with the standard threshold, shown in Figure 3.18, has similar power to the stabilized knockoff filter but with an FDR of about 0.15 (compared to a nominal rate of $q = 0.1$). As the number of non-null variables increases from $k = 10$, the power of the knockoff filter with knockoff+ threshold approaches that of the stabilized knockoff filter (Figure 3.19). With uncorrelated features, the stabilized knockoff filter has substantially larger power than the knockoff+ threshold at all signal magnitudes, while the standard knockoff filter again has comparable power to the stabilized knockoff filter but fails to control the FDR at the nominal level (Figures 3.21 and 3.22). In many situations, the knockoff filter and the stabilized knockoff filter have a similar average number of selected variables; however,

the standard deviation in the number of selected variables for the stabilized knockoff filter is always smaller than that of the standard knockoff filter, and is often smaller by a factor of two or three.

In the $p = 1000$, $n = 3000$ simulation scenarios (Figures 3.23–3.28), the stabilized knockoff filter with $B = 50$ or $B = 100$ again controls the FDR for any degree of feature correlation, signal magnitude, or sparsity. The stabilized knockoff filter does not improve on the power of the standard knockoff threshold as much as it did when $p = 100$, but the standard deviation in the number of selected features is again smaller than that of the standard knockoff filter across all simulation scenarios. For example, in Figure 3.23, the standard deviation is reduced by approximately one half compared to the knockoff+ threshold when feature correlation is smaller than about 0.5. The stabilized knockoff filter compares favorably with the knockoff+ threshold, in that it also controls the FDR but has similar power to the standard knockoff threshold while substantially reducing variation in the number of selected variables.

These simulation studies were also performed for two fixed design matrices while drawing samples from the distribution of $\mathbf{Y} \mid \mathbf{X}$. Figures 3.29–3.35 present the FDR, power and mean and standard deviation of the number of selected variables for the $p = 100, n = 5000$ and $p = 1000, n = 3000$ settings for two fixed design matrices in each setting. These results are similar to the results averaged over draws of $(\mathbf{X}, \mathbf{Y})$, with FDR control across a range of feature correlations and sparsity levels, and substantial reductions in the standard deviation of the number of selected variables using the stabilized knockoff filter.

3.5.2 Comparison with other multiple testing procedures

These simulation studies were also used to compare the performance of the knockoff filter and stabilized knockoff filter to alternative multiple testing procedures in the context of linear regression. Specifically, P-values from the Wald tests in ordi-

nary least squares were corrected for multiple testing using either the Bonferroni or Benjamini-Hochberg (BH) procedures. The Bonferroni correction controls the familywise error rate (FWER), the probability of one or more false discoveries. Note that any FWER-controlling procedure also controls the FDR.

With $p = 100$, $n = 5000$, and $k = 10$ truly non-null variables, the stabilized knockoff filter achieves FDR control at all levels of feature correlation (Figure 3.37) and has higher power than the knockoff+ threshold, Bonferroni-corrected P-values or BH-corrected P-values. The knockoff threshold has FDR of approximately 0.15, which is greater than the nominal level of $q = 0.1$. The knockoff+ threshold controls the FDR at $q = 0.1$ but has lower power than Bonferroni-corrected or BH-corrected P-values as well as the stabilized knockoff filter. Furthermore, the standard deviation in the number of selected variables for the stabilized knockoff filter is approximately half as large as that of the knockoff filter with the knockoff or knockoff+ thresholds.

With $p = 1000$ or greater number of non-null variables, selecting variables with Bonferroni-corrected P-values has lower statistical power than either the knockoff+ threshold or standard knockoff threshold. Using BH-corrected P-values leads to power and a degree of variation in the number of selected variables comparable to those of the stabilized knockoff filter when $p = 100$. When $p = 1000$ (Figures 3.40–3.42), the stabilized knockoff filter has slightly higher power than BH-corrected P-values for any degree of feature correlation, sparsity, or signal magnitude. However, BH-corrected P-values have a similarly low standard deviation in the number of selected variables as the stabilized knockoff filter.

3.6 Discussion

The stabilized knockoff filter reduces variance in two ways: by stabilizing the threshold used to screen the per-variable importance statistics and by stabilizing the set of selected variables at a given threshold. This is possible because for any

fixed dataset there is no unique vector importance statistics used to screen variables. Instead, there is one vector of importance statistics per arbitrary choice of $\tilde{\mathbf{U}}$, the $n \times p$ matrix used to construct knockoff variables to achieve a particular correlation structure across the $2p$ original and knockoff variables. This collection of importance statistics is used to compute a low-variance estimate of the FDR and hence a low-variance threshold at which the FDR is controlled. Given a fixed FDR-controlling threshold for the importance statistics a stabilized set of variable selections is obtained by selecting a number of variables equal to a low-variance estimate of the mean number of selections at that threshold. The expected number of selections at that threshold is calibrated (controls the FDR) since the threshold was chosen to control the FDR. Finally, for a fixed number of selections at a fixed threshold, the variables can be ranked according to the per-variable evidence against the null hypothesis, namely an estimate of $\mathbb{P}\left(W_j \geq t \mid \mathbf{X}\right)$ for each j.

Based on empirical study of the Gaussian linear model, ($\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$) for which the original knockoff filter of Barber and Candès (2015) is proven to control the FDR in finite samples, this stabilized knockoff filter seems to control the FDR with $B = 50$ or $B = 100$ pseudo-samples of the importance statistics. However, I have not presented any formal argument that the empirical distribution of the pseudo-samples of $(W_1, \ldots W_p)$ generated from a set of distinct $\tilde{\mathbf{U}}$ do, in fact, estimate the true marginal distributions $W_j \mid \mathbf{X}$. Further study of the distribution of W_j (with fixed $\mathbf{X}$ and $\mathbf{Y}$) as a function of the random matrix $\tilde{\mathbf{U}}$ (in Algorithms III.3–III.5) is required to understand the distribution of the pseudo-samples of W_j based on repeated sampling of $\tilde{\mathbf{U}}$.

Recall that FDR control for the knockoff filter was proved in Barber and Candès (2015) only for the knockoff+ threshold,

$$T_+ = \min_j \left\{ t = |W_j| : \frac{1 + \#\left\{k : W_k \leq -t\right\}}{\max\left\{1, \#\left\{k : W_k \geq t\right\}\right\}} \leq q \right\},$$

which estimates the FDP using

$$\frac{1 + \#\{k : W_k \leq -t\}}{\max\{1, \#\{k : W_k \geq t\}\}}.$$

The additional 1 in the numerator of this FDP estimate is necessary in order to prove FDR control in Barber and Candès (2015), and the empirical results in Section 3.5 suggest that this adjustment is required to control the FDR in practical regression scenarios. However, the knockoff+ threshold often has reduced power compared to the standard knockoff threshold, especially in moderate-p settings, but the standard knockoff threshold does not always control the FDR. In the stabilized knockoff filter, the approximate upper bound for the FDR given in equation (3.15) does not include a similar adjustment as in the knockoff+ threshold. The simulation results in Section 3.5 suggest that the stabilized knockoff filter enjoys similar power to the standard knockoff threshold while maintaining control of the FDR. Additional work could explore whether the stabilized knockoff filter enjoys theoretical control of the FDR even without the knockoff+ adjustment required to control FDR with the knockoff filter.

An alternative approach to stabilizing the knockoff filter is to choose a single $\tilde{\mathbf{U}}$ that stabilizes the computed importance statistics and the set of selected variables. Three such methods for computing $\tilde{\mathbf{U}}$ are described in Appendix A, but none of these methods are shown to reduce variation in the knockoff variable selections to the extent possible with the stabilized knockoff filter. Future work on stabilizing $\tilde{\mathbf{U}}$ directly could focus on computing the empirical average of a collection of $\tilde{\mathbf{U}}$ on the Stiefel manifold to which each $\tilde{\mathbf{U}}$ belongs (Kaneko, Fiori, and Tanaka 2012). Additionally, the characterization of the cross products $\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^\intercal \mathbf{Y}$ in (3.6) could suggest alternative methods of performing the knockoff filter without actually constructing the matrix $\tilde{\mathbf{X}}$: the importance statistics could be computed from the augmented Gram matrix and samples of the cross products based on (3.6). A sampling procedure for the

cross products based on (3.6) could be modified to reduce variance in the importance statistics.

The stabilized knockoff filter may be especially useful in moderate-p, low-signal scenarios such as those presented in Figures 3.17 and 3.18, in which the knockoff filter achieves power similar to that of the stabilized knockoff filter only when the FDR is not controlled at the nominal level (Figure 3.17).

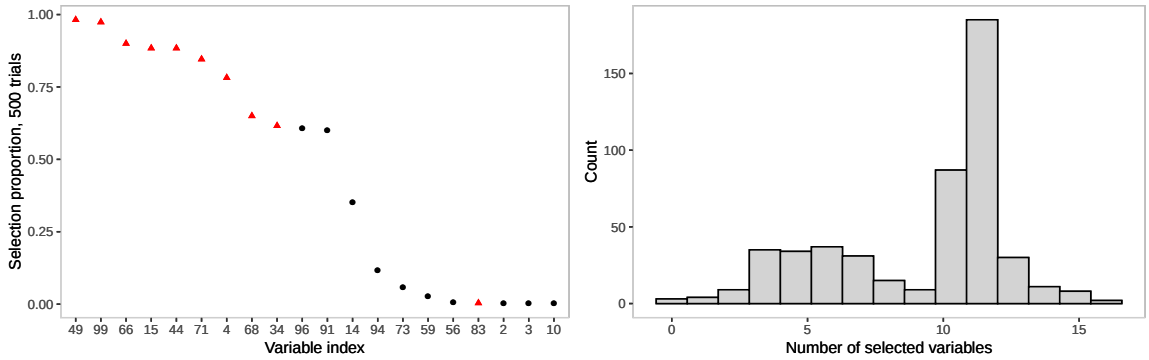


Figure 3.1: Example of knockoff instability for fixed $(\mathbf{X}, \mathbf{Y})$. Data generated from $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $n = 5000$ and $p = 100$. Left panel displays the number of times each variable was selected out of 500 repetitions of Step 3 in Algorithm III.1 (excluding null variables that were never selected) and the right panel displays the distribution of the number of variables selected over these trials. Red triangles indicate truly non-null variables.

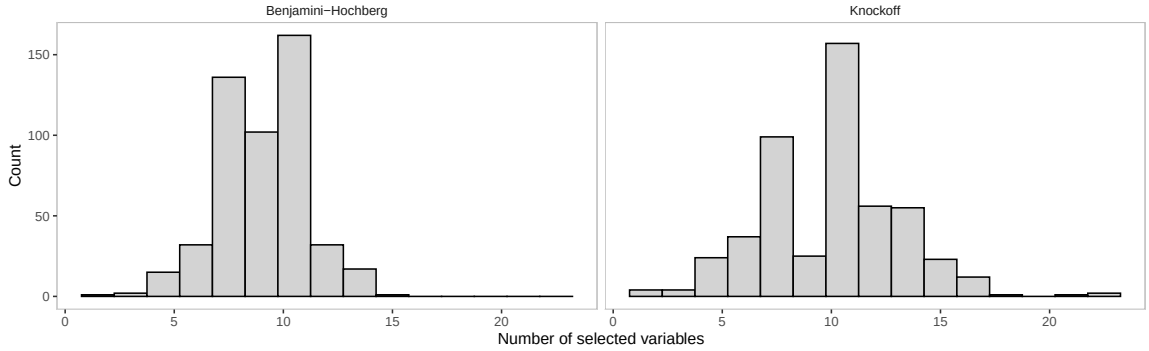


Figure 3.2: Number of selected variables, knockoff filter vs. Benjamini-Hochberg, with fixed $\mathbf{X}$. Data generated from $\mathbf{Y} \mid \mathbf{X}$, where $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $n = 5000$ and $p = 100$. Benjamini-Hochberg selection performed with least squares coefficient P-values.

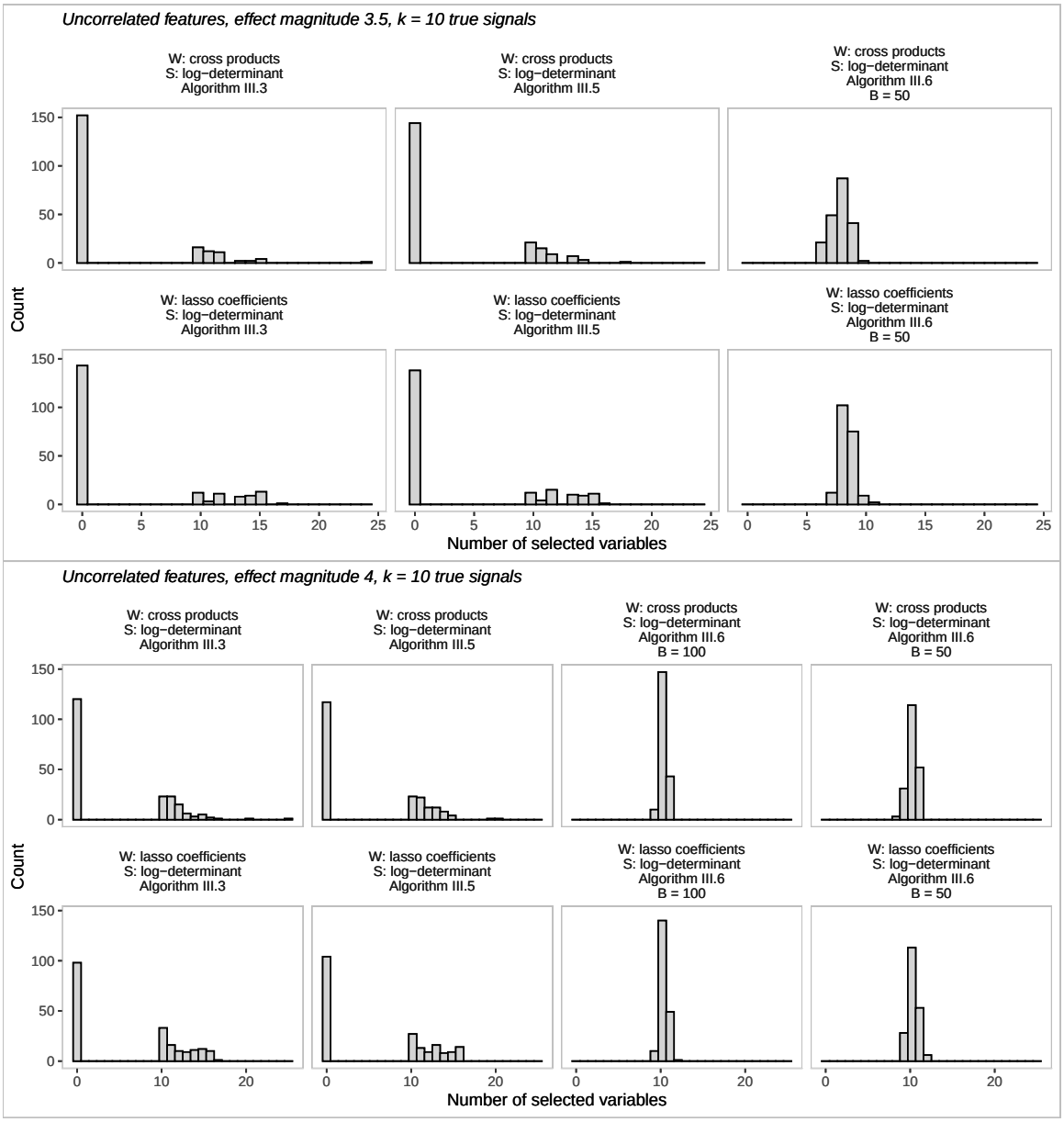


Figure 3.3: Number of selected variables for fixed $\mathbf{X}, \mathbf{Y}$ with $n = 5000, p = 100$. Based on 200 knockoff filter replicates, uncorrelated features, and $k = 10$ nonzero β_j.

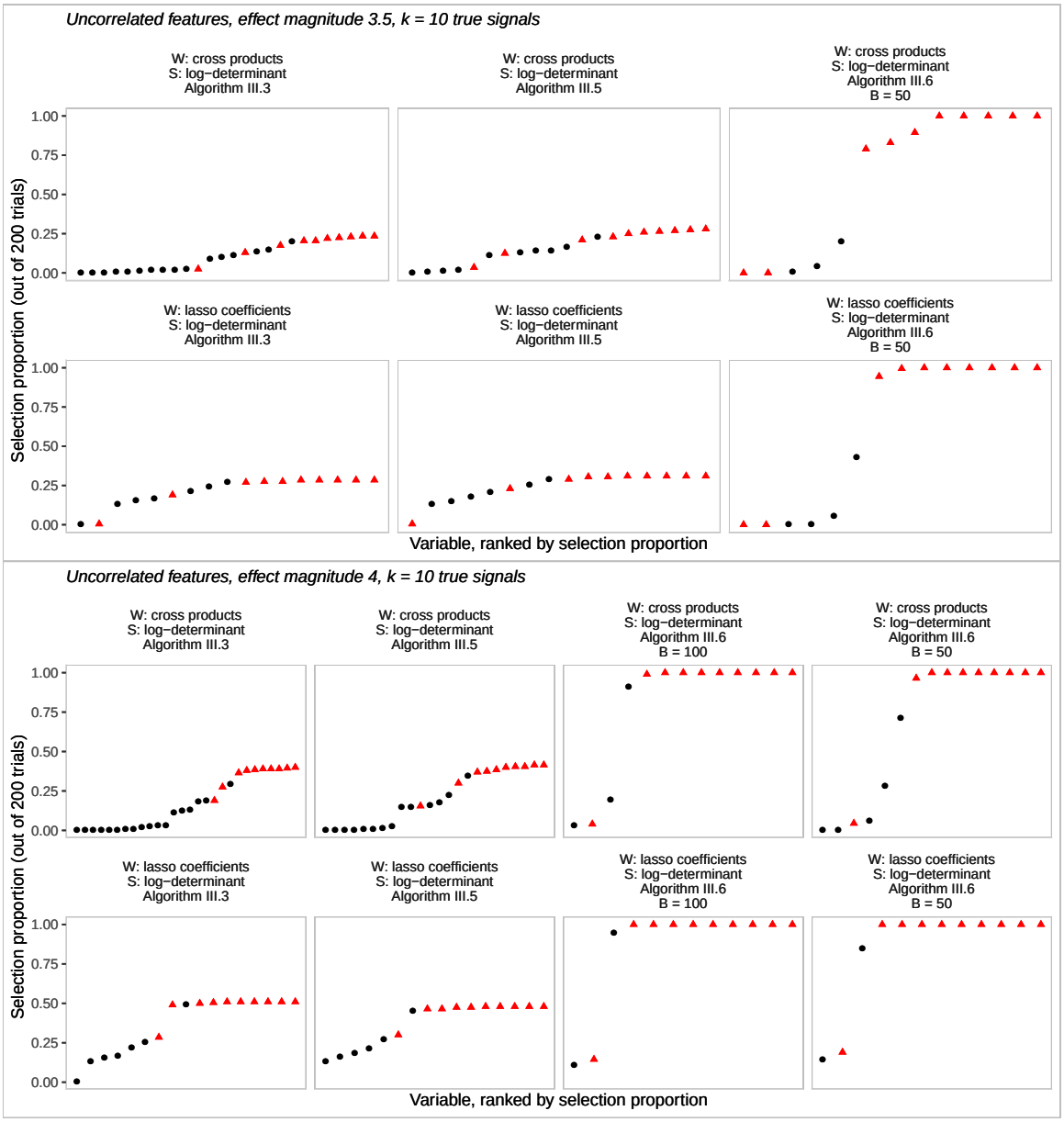


Figure 3.4: Variable-specific selection probability for fixed $\mathbf{X}, \mathbf{Y}$ with $n = 5000, p = 100$. Based on 200 knockoff filter replicates, uncorrelated features, and $k = 10$ nonzero β_j. Red triangles indicate truly non-null variables. Null variables which were never selected are not displayed.

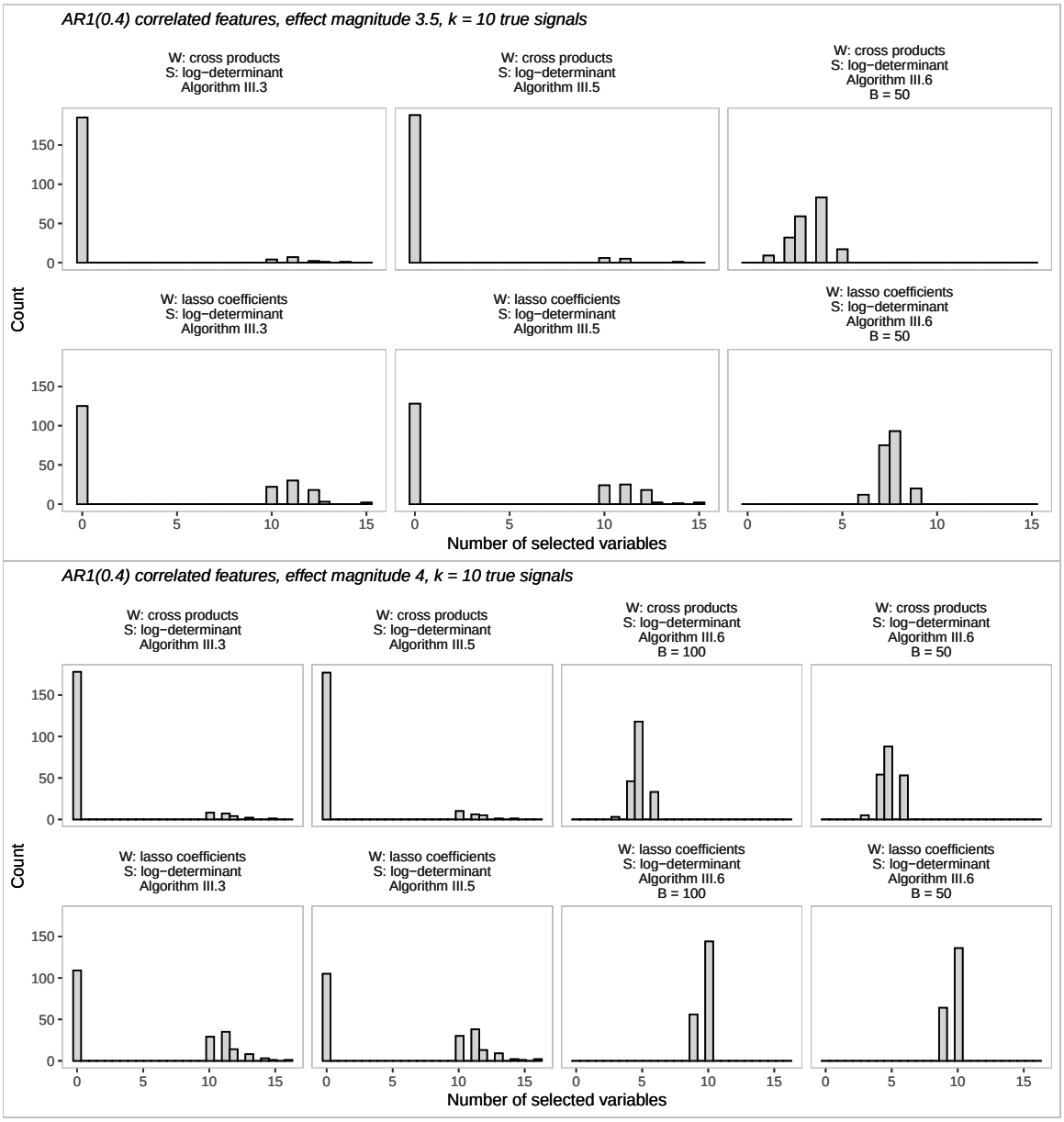


Figure 3.5: Number of selected variables for fixed $\mathbf{X}, \mathbf{Y}$ with $n = 5000, p = 100$ and correlated features. Based on 200 knockoff filter replicates, correlated features (autoregressive with population correlation 0.4), and $k = 10$ nonzero β_j.

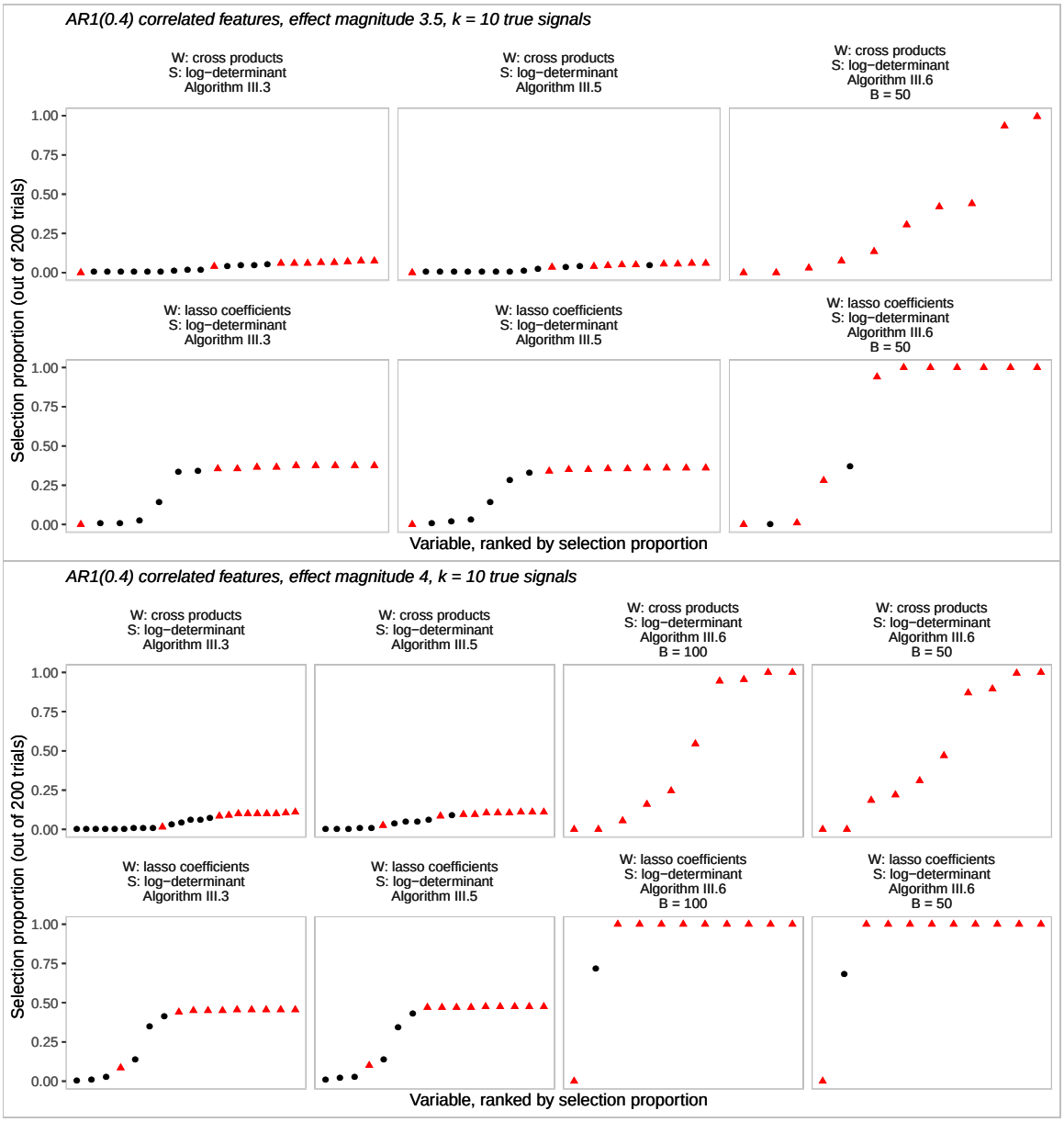


Figure 3.6: Variable-specific selection probability for fixed $\mathbf{X}, \mathbf{Y}$ with $n = 5000, p = 100$ and correlated features. Based on 200 knockoff filter replicates, correlated features (autoregressive with population correlation 0.4), and $k = 10$ nonzero β_j. Red triangles indicate truly non-null variables. Null variables which were never selected are not displayed.

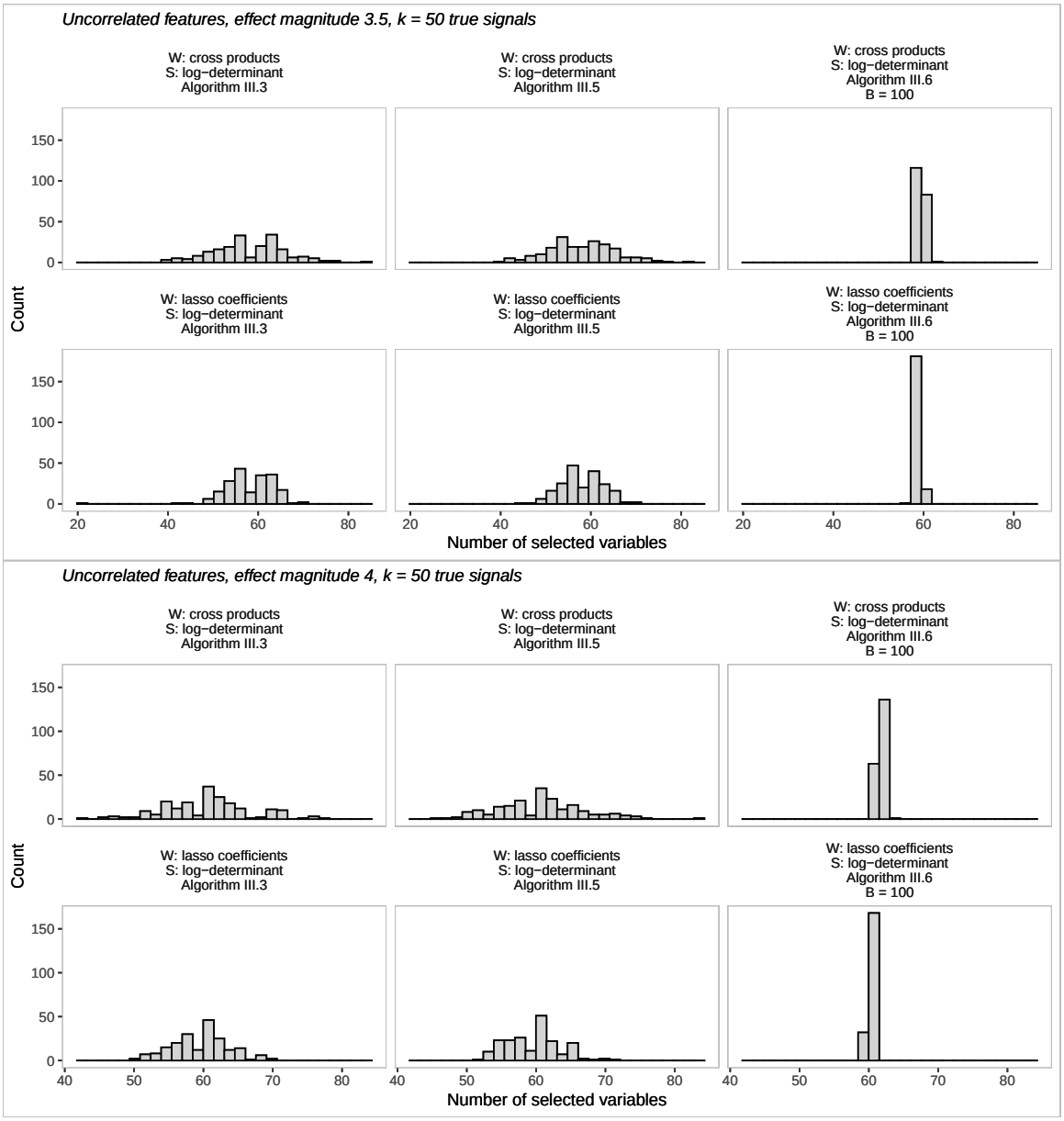


Figure 3.7: Number of selected variables for fixed $\mathbf{X}, \mathbf{Y}$ with $n = 5000, p = 100$, and $k = 50$ nonzero β_j. Based on 200 knockoff filter replicates and uncorrelated features.

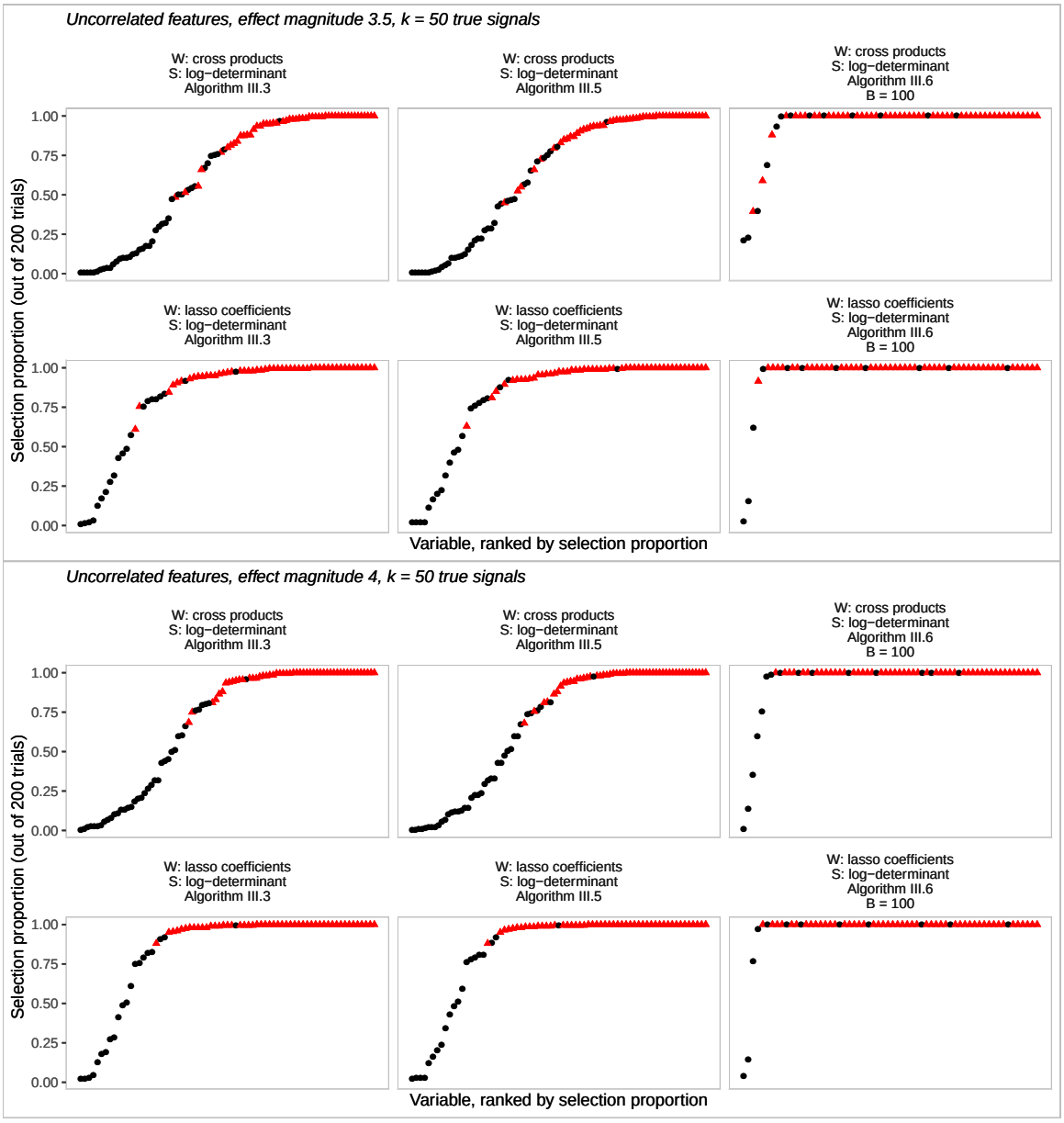


Figure 3.8: Variable-specific selection probability for fixed $\mathbf{X}, \mathbf{Y}$ with $n = 5000, p = 100$, and $k = 50$ nonzero β_j. Based on 200 knockoff filter replicates and uncorrelated features. Red triangles indicate truly non-null variables. Null variables which were never selected are not displayed.

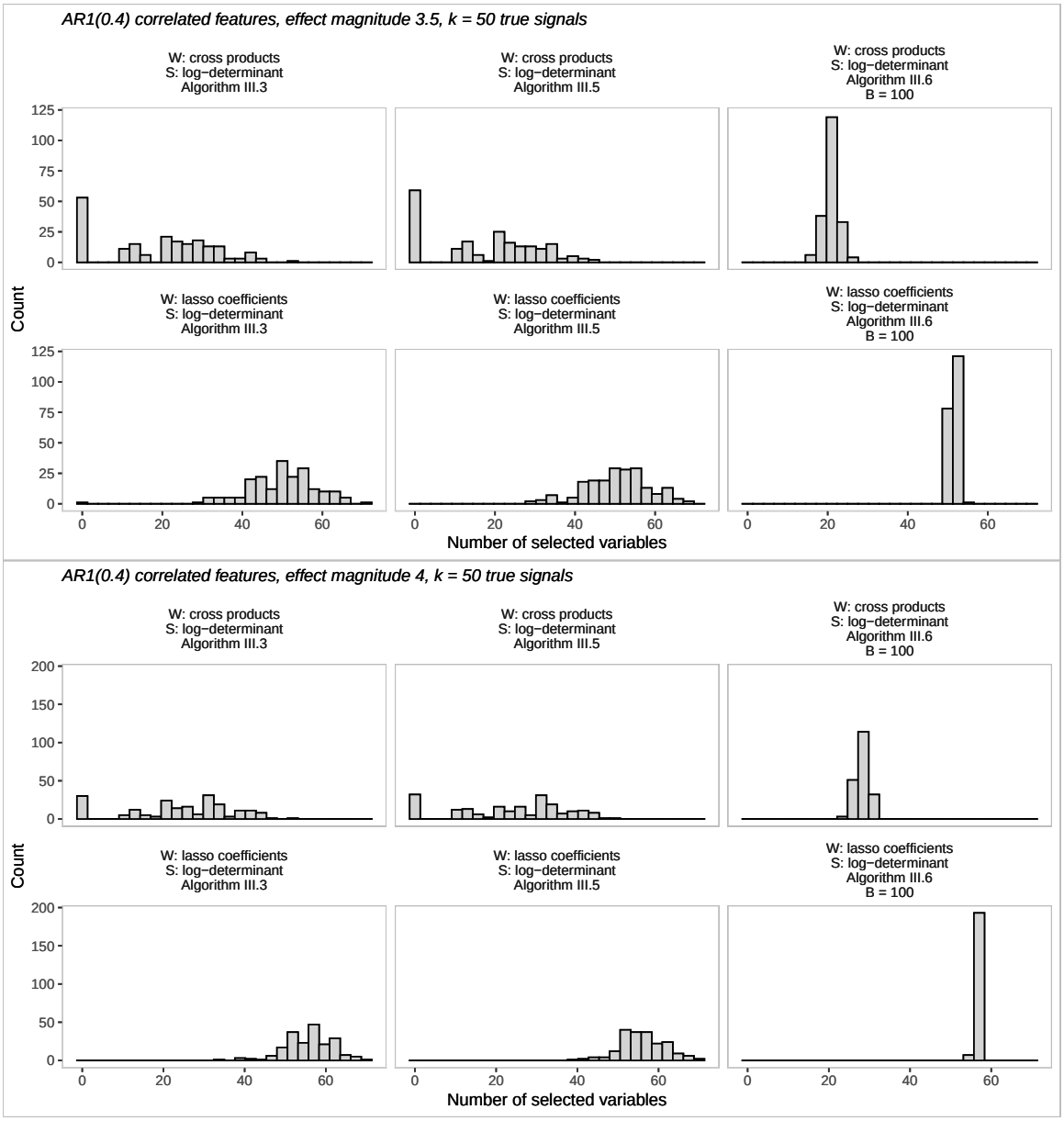


Figure 3.9: Number of selected variables for fixed $\mathbf{X}, \mathbf{Y}$ with $n = 5000, p = 100$, $k = 50$ nonzero β_j and correlated features. Based on 200 knockoff filter replicates and autoregressive features with population correlation 0.4.

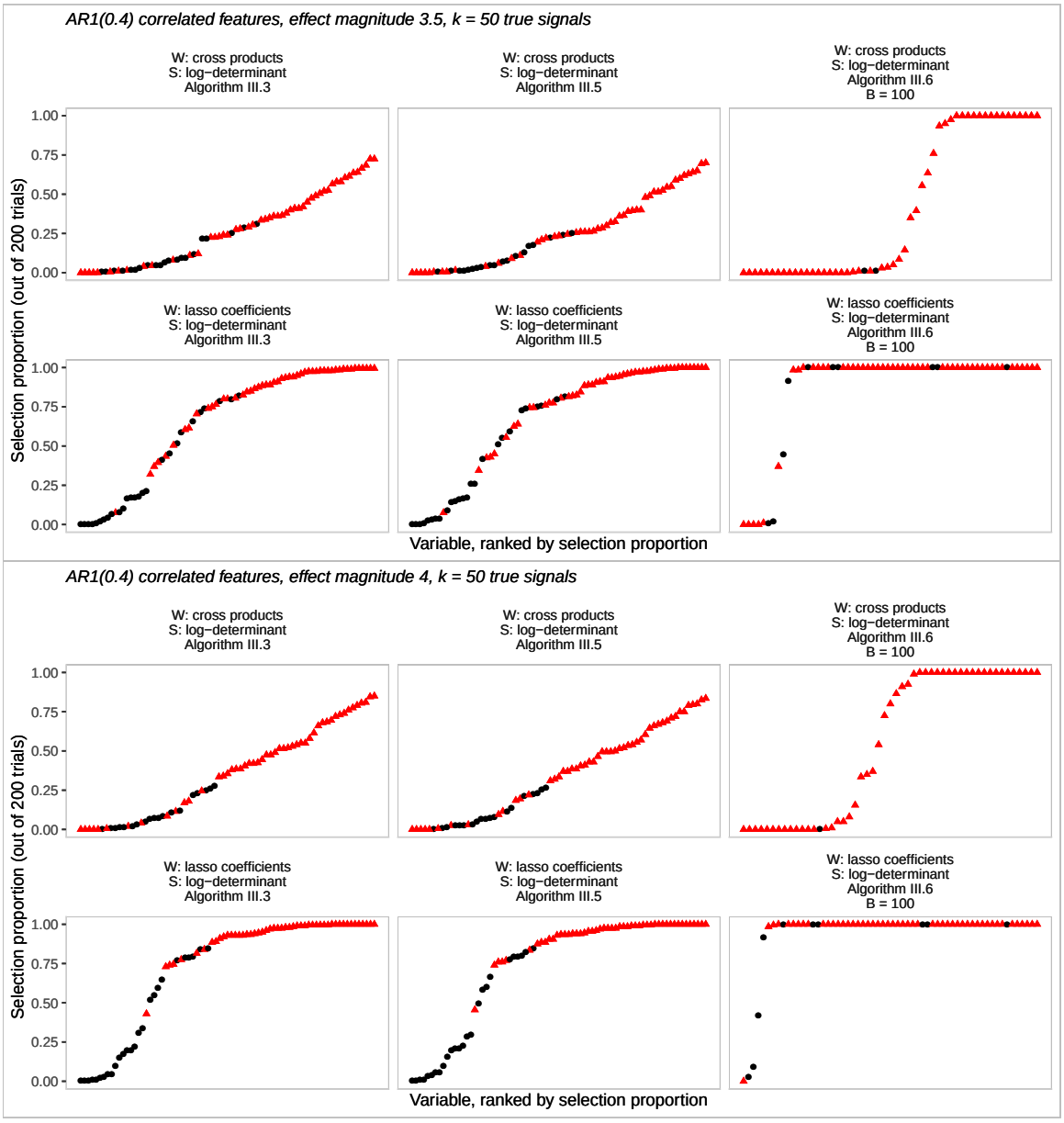


Figure 3.10: Variable-specific selection probability for fixed $\mathbf{X}, \mathbf{Y}$ with $n = 5000, p = 100$, correlated features, and $k = 50$ nonzero β_j. Based on 200 knockoff filter replicates with autoregressive features (population correlation 0.4). Red triangles indicate truly non-null variables. Null variables which were never selected are not displayed.

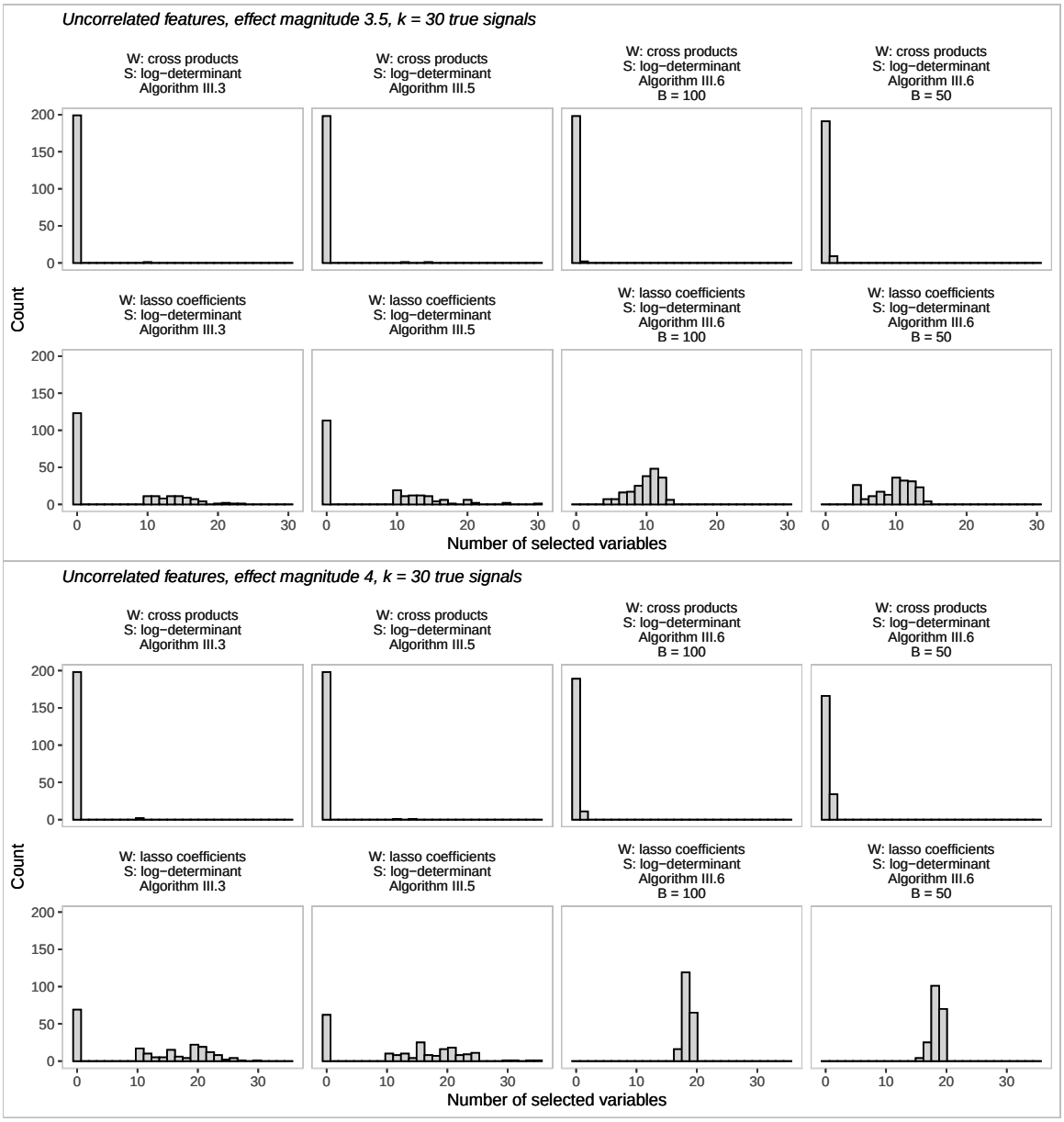


Figure 3.11: Number of selected variables for fixed $\mathbf{X}, \mathbf{Y}$ with $n = 3000, p = 1000$. Based on 200 knockoff filter replicates, uncorrelated features, and $k = 30$ nonzero β_j.

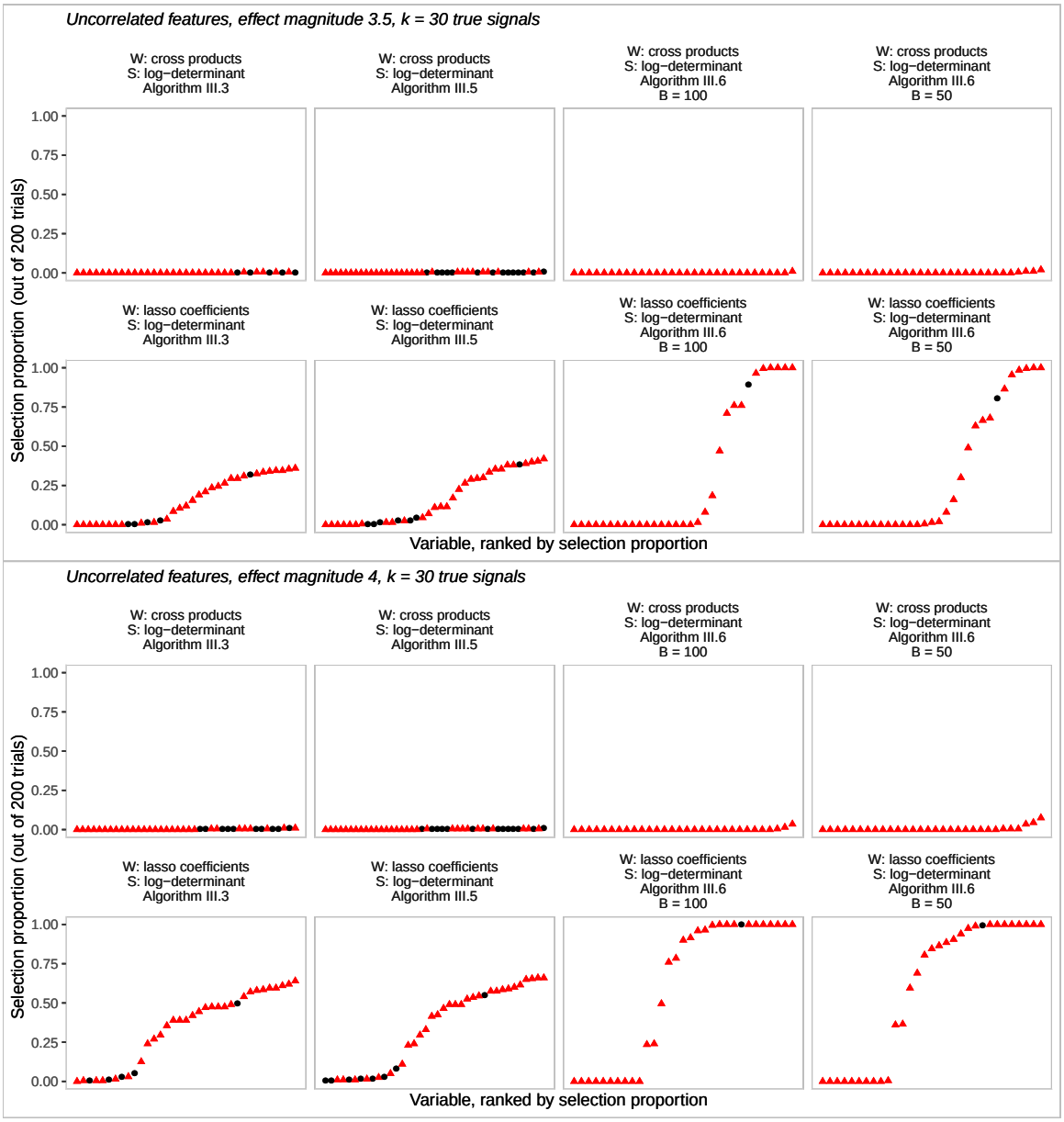


Figure 3.12: Variable-specific selection probability for fixed $\mathbf{X}, \mathbf{Y}$ with $n = 3000, p = 1000$. Based on 200 knockoff filter replicates, uncorrelated features, and $k = 30$ nonzero β_j. Red triangles indicate truly non-null variables. Null variables which were never selected are not displayed.

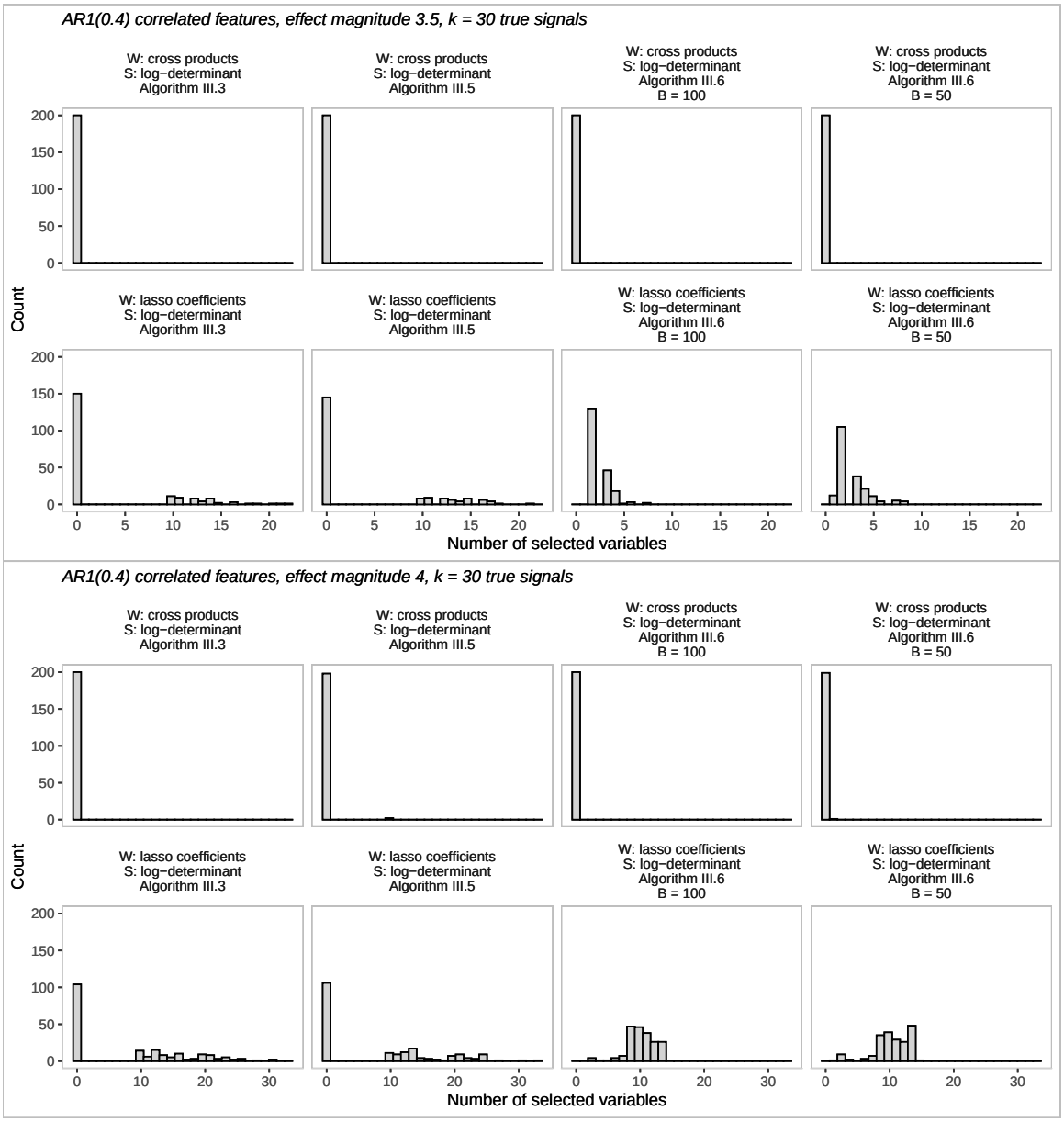


Figure 3.13: Number of selected variables for fixed $\mathbf{X}, \mathbf{Y}$ with $n = 3000, p = 1000$ and correlated features. Based on 200 knockoff filter replicates, autoregressive features (population correlation 0.4) and $k = 30$ nonzero β_j.

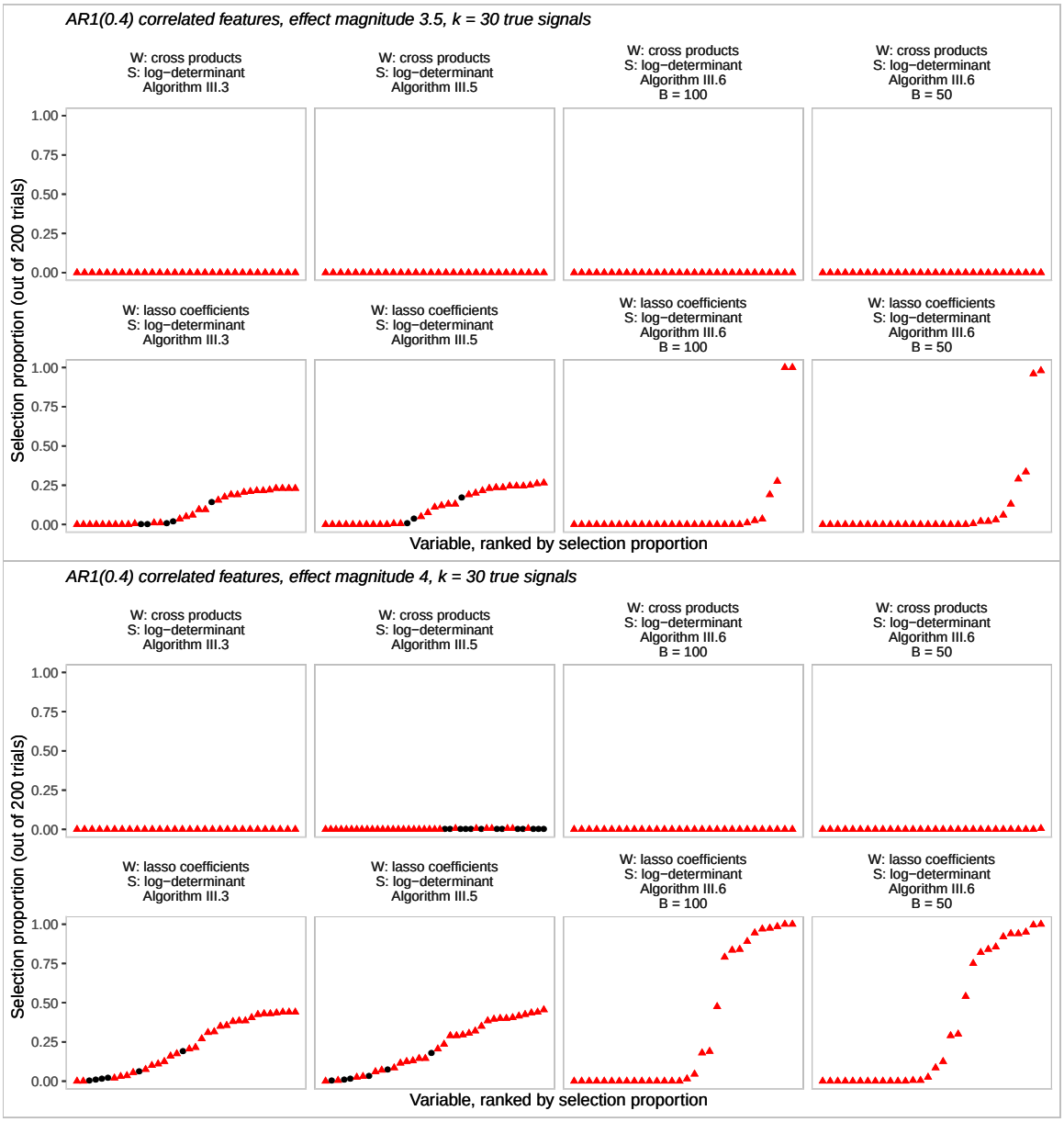


Figure 3.14: Variable-specific selection probability for fixed $\mathbf{X}, \mathbf{Y}$ with $n = 3000, p = 1000$, and correlated features. Based on 200 knockoff filter replicates, autoregressive features (population correlation 0.4), and $k = 30$ nonzero β_j. Red triangles indicate truly non-null variables. Null variables which were never selected are not displayed.

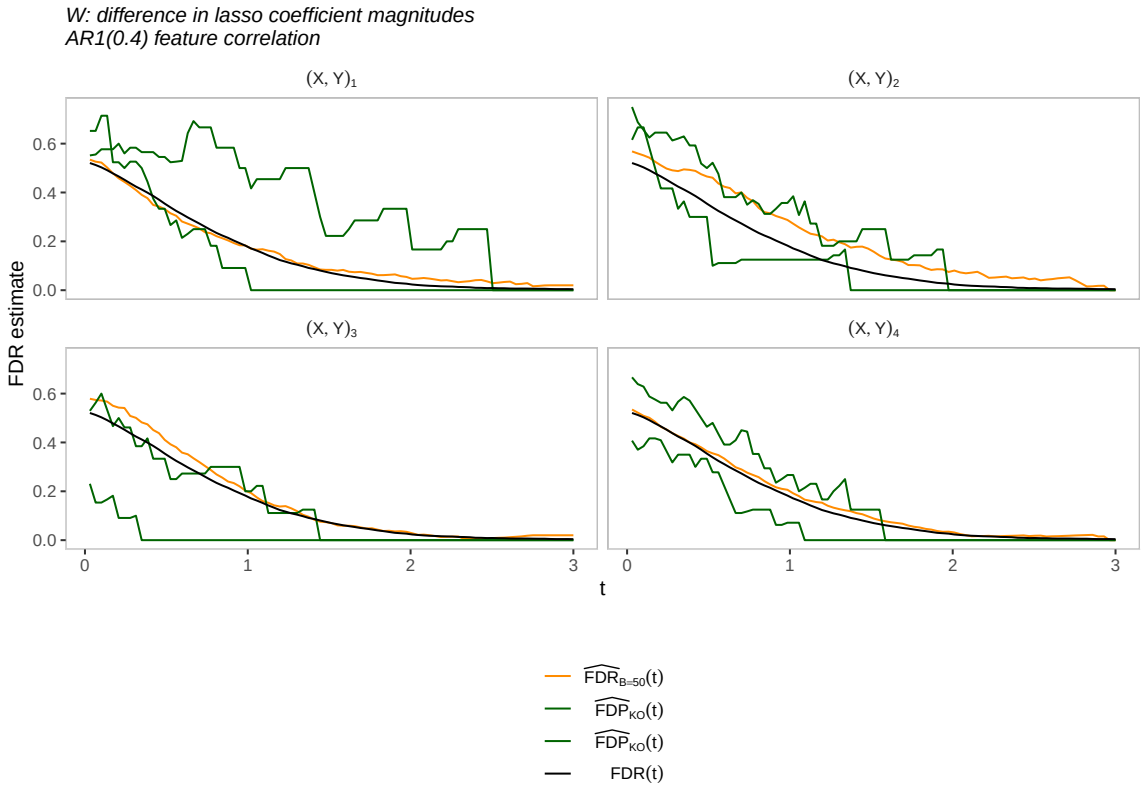


Figure 3.15: Knockoff FDR estimates in four fixed $(\mathbf{X}, \mathbf{Y})$ samples. Green lines represent two examples of $\widehat{\mathrm{FDP}}_{\mathrm{KO}}(t)$ based on two different matrices of knockoffs for a given, fixed design matrix.

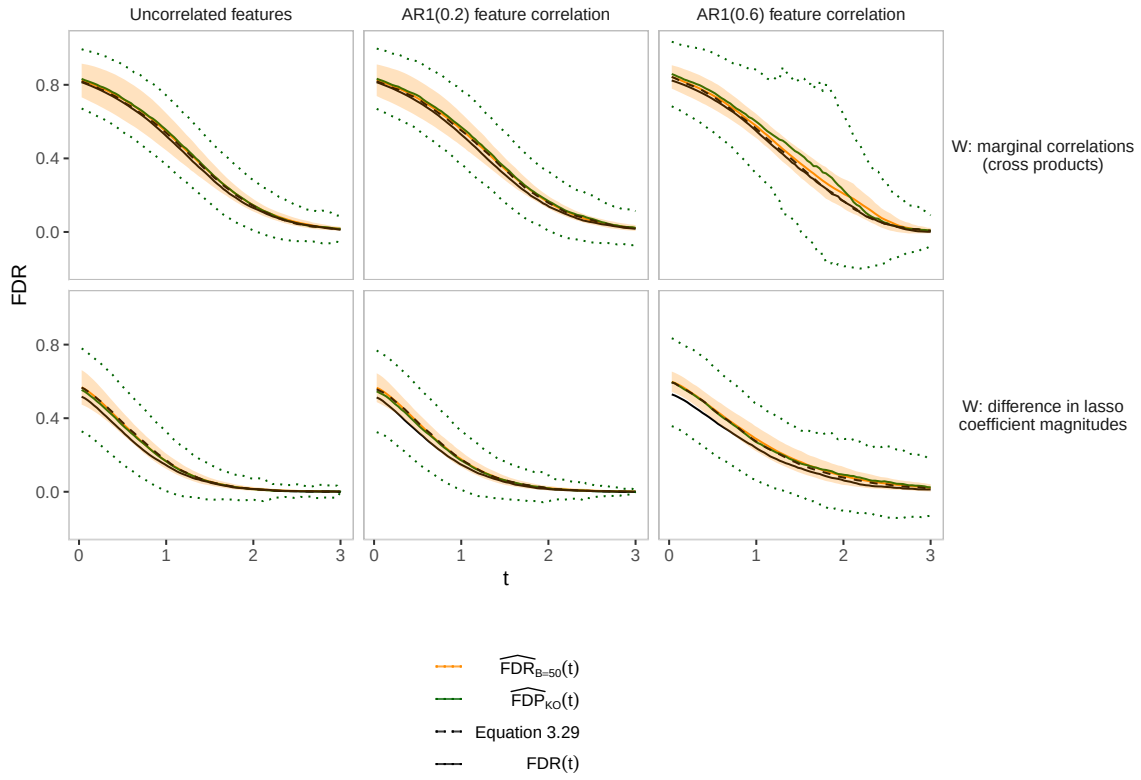


Figure 3.16: Variability in knockoff FDR estimates. Dotted green lines and shaded orange regions display the mean $\pm$ one standard deviation for $\widehat{\text{FDP}}_{\text{KO}}(t)$ and $\widehat{\text{FDR}}_B(t)$, respectively, based on 200 simulation replicates. With $n = 5000, p = 100$ and $k = 10$ variables β_j with $|\beta_j| = 3.5$.

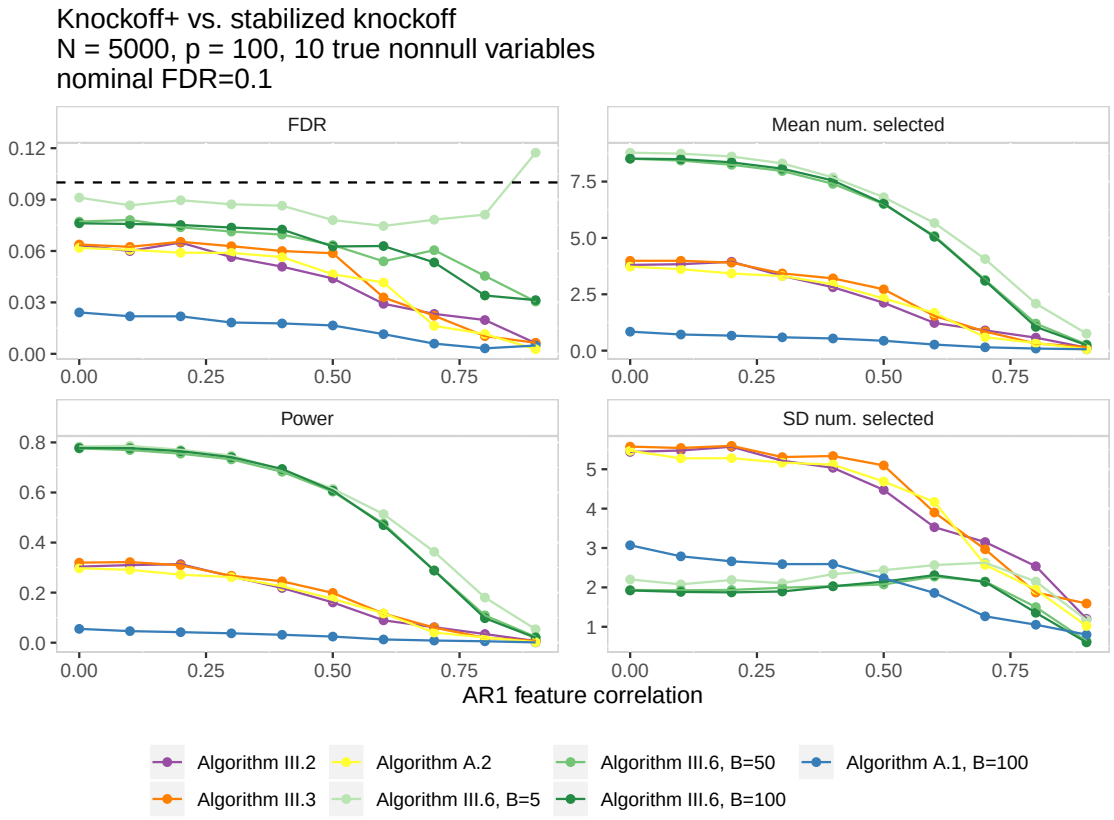


Figure 3.17: Power and FDR of knockoff+ and stabilized knockoff filter as a function of feature correlation with $n = 5000, p = 100$. Right column displays the mean and variance of the number of selected variables in a given sample. Averaged over 500 simulated $(\mathbf{X}, \mathbf{Y})$.

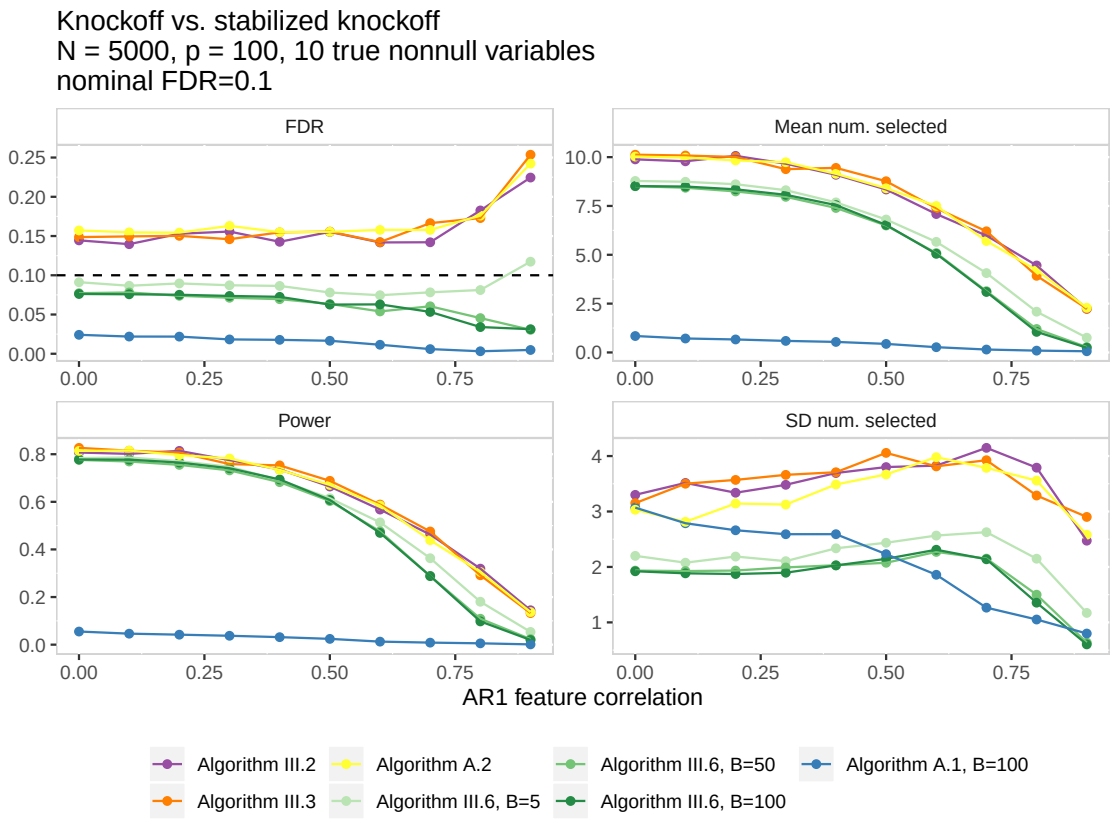


Figure 3.18: Power and FDR of knockoff and stabilized knockoff as a function of feature correlation with $n = 5000, p = 100$. Right column displays the mean and variance of the number of selected variables in a given sample. Averaged over 500 simulated $(\mathbf{X}, \mathbf{Y})$.

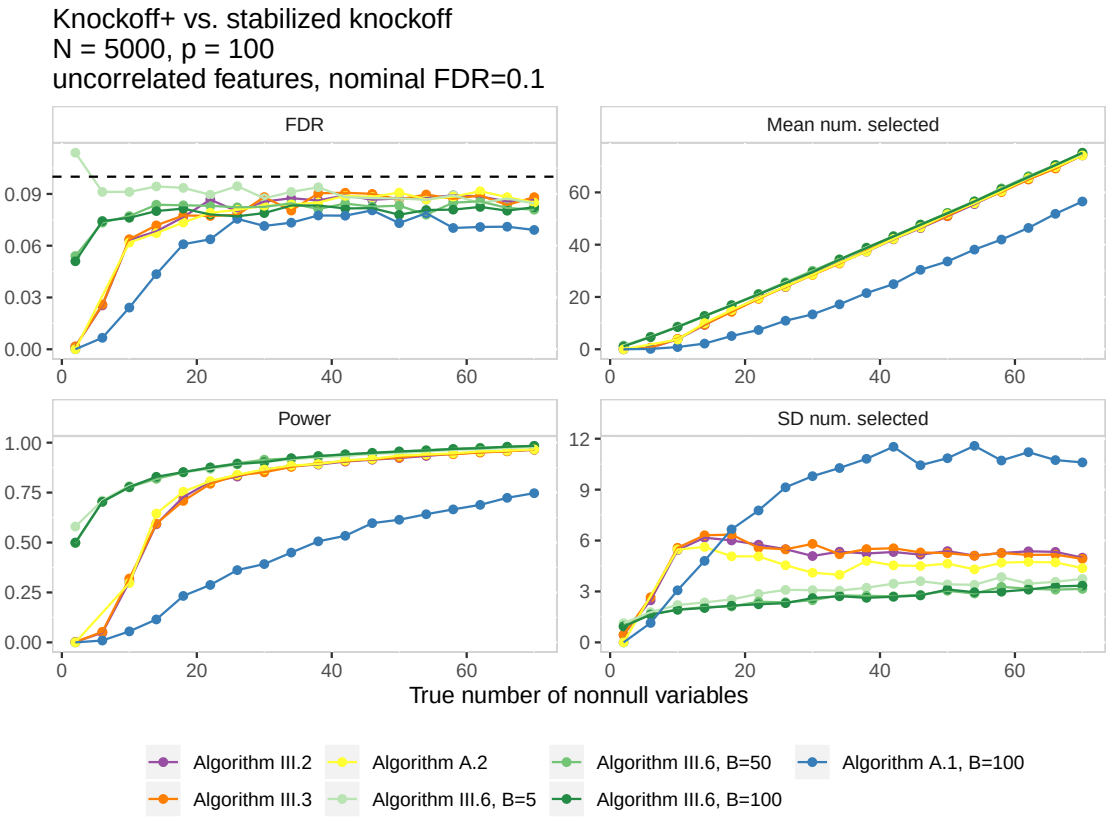


Figure 3.19: Power and FDR of knockoff+ and stabilized knockoff as a function of model sparsity with $n = 5000, p = 100$. Right column displays the mean and variance of the number of selected variables in a given sample. Averaged over 500 simulated $(\mathbf{X}, \mathbf{Y})$.

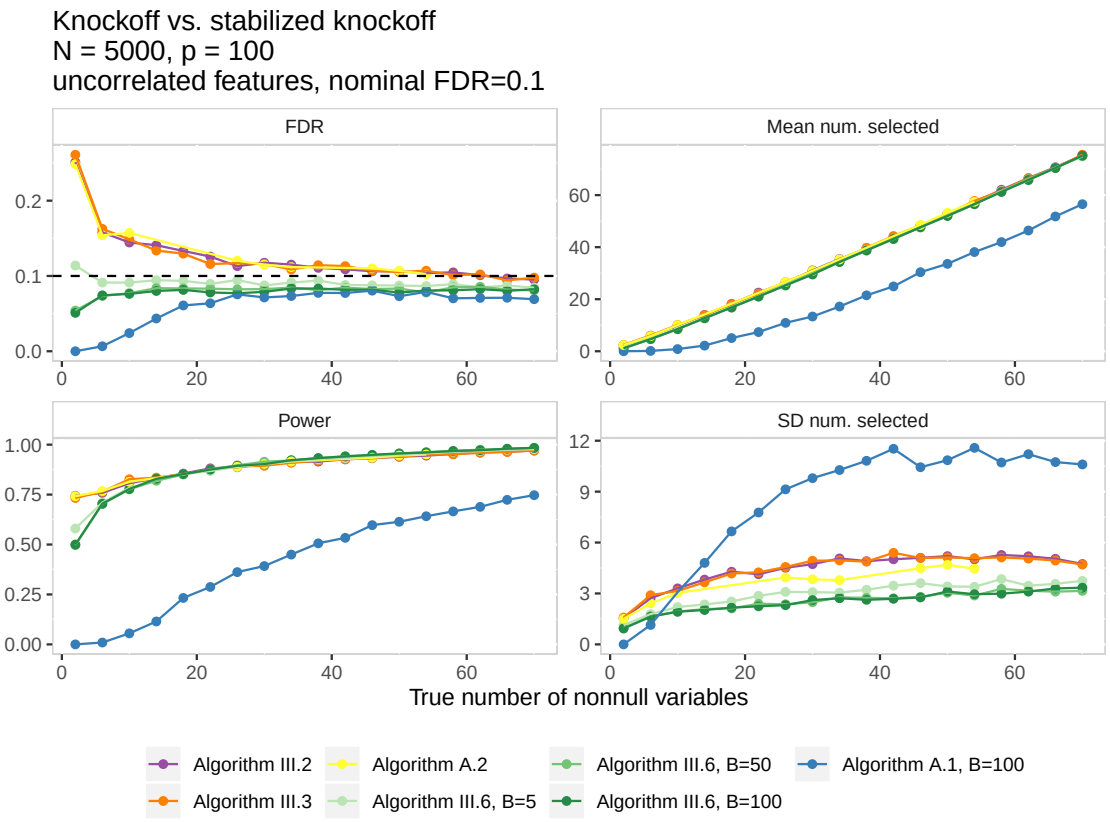


Figure 3.20: Power and FDR of knockoff and stabilized knockoff as a function of model sparsity with $n = 5000, p = 100$. Right column displays the mean and variance of the number of selected variables in a given sample. Averaged over 500 simulated $(\mathbf{X}, \mathbf{Y})$.

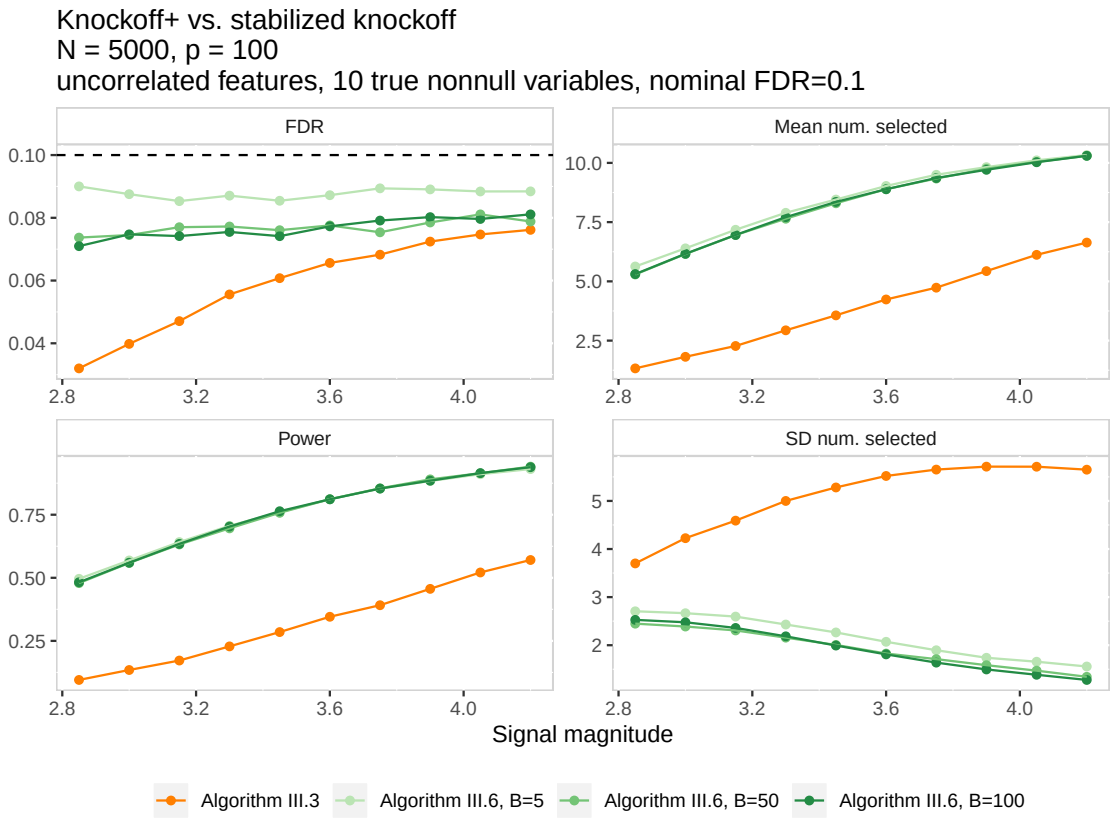


Figure 3.21: Power and FDR of knockoff+ and stabilized knockoff as a function of signal magnitude with $n = 5000, p = 100$. Right column displays the mean and variance of the number of selected variables in a given sample. There are $k = 10$ truly nonzero β_j. Averaged over 500 simulated $(\mathbf{X}, \mathbf{Y})$.

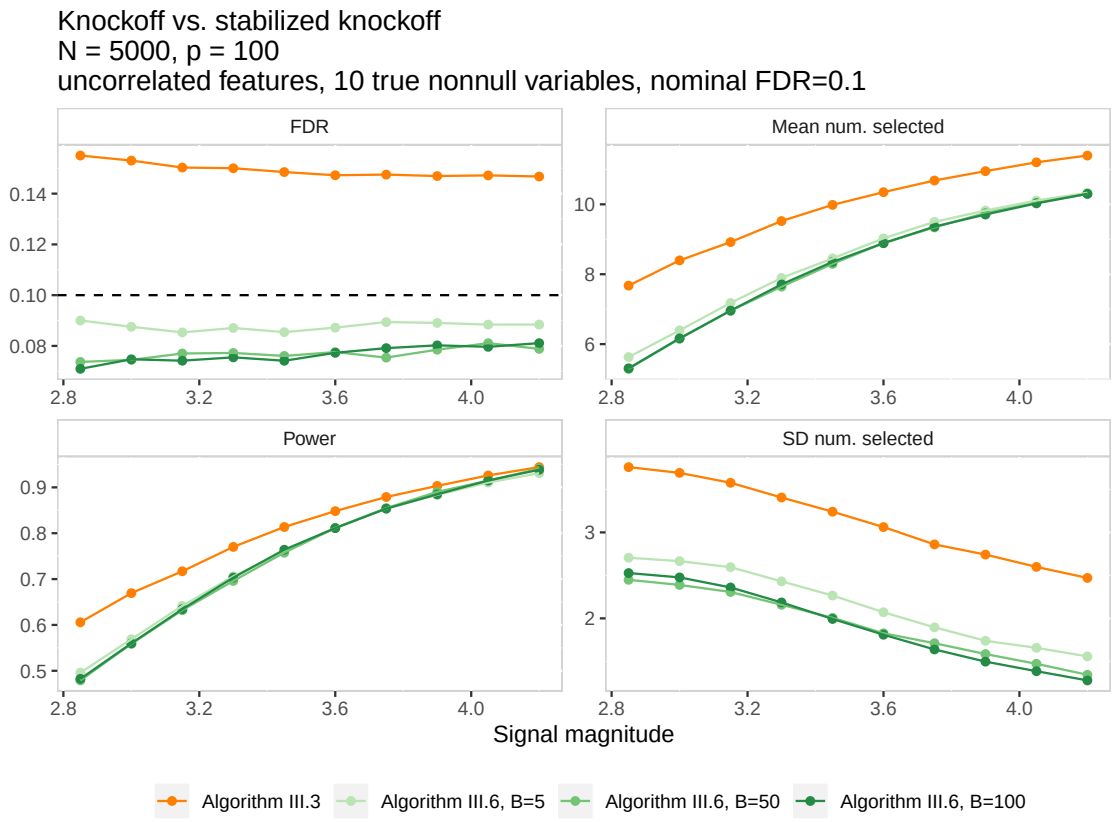


Figure 3.22: Power and FDR of knockoff and stabilized knockoff as a function of signal magnitude with $n = 5000, p = 100$. There are $k = 10$ truly nonzero β_j. Right column displays the mean and variance of the number of selected variables in a given sample. Averaged over 500 simulated $(\mathbf{X}, \mathbf{Y})$.

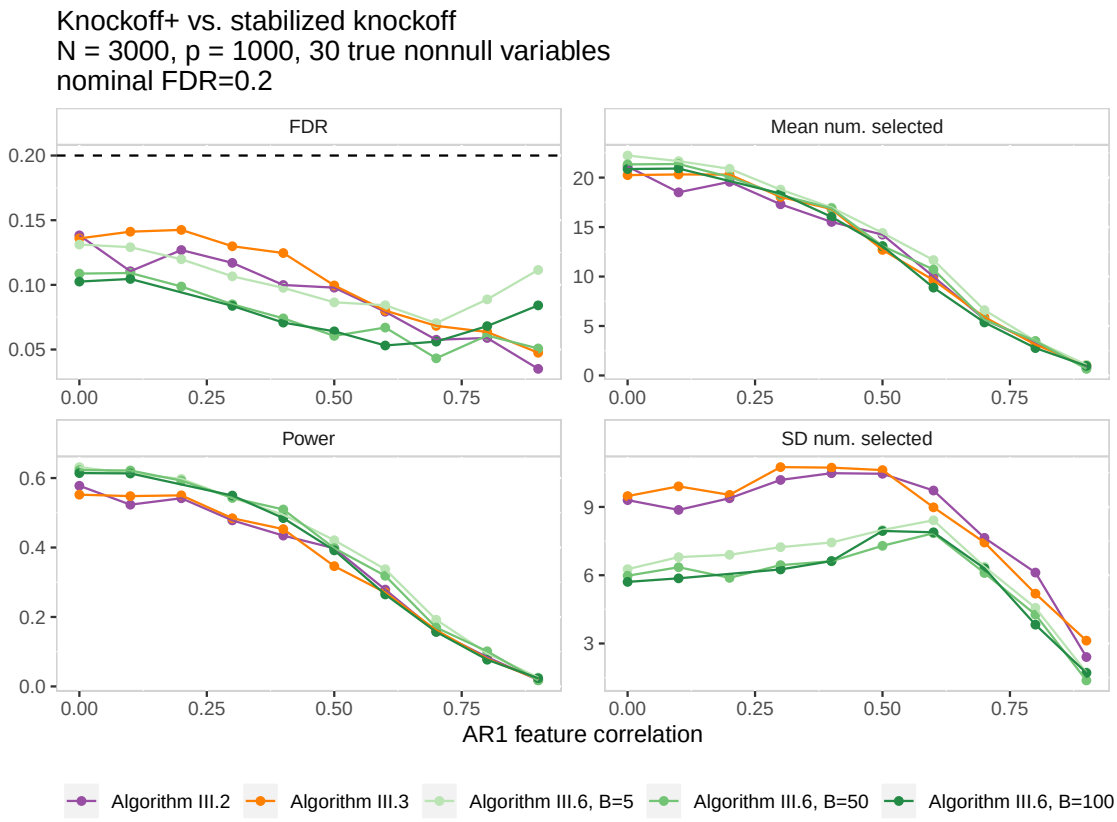


Figure 3.23: Power and FDR of knockoff+ and stabilized knockoff as a function of feature correlation with $n = 3000, p = 1000$. Right column displays the mean and variance of the number of selected variables in a given sample. Averaged over 200 simulated $(\mathbf{X}, \mathbf{Y})$.

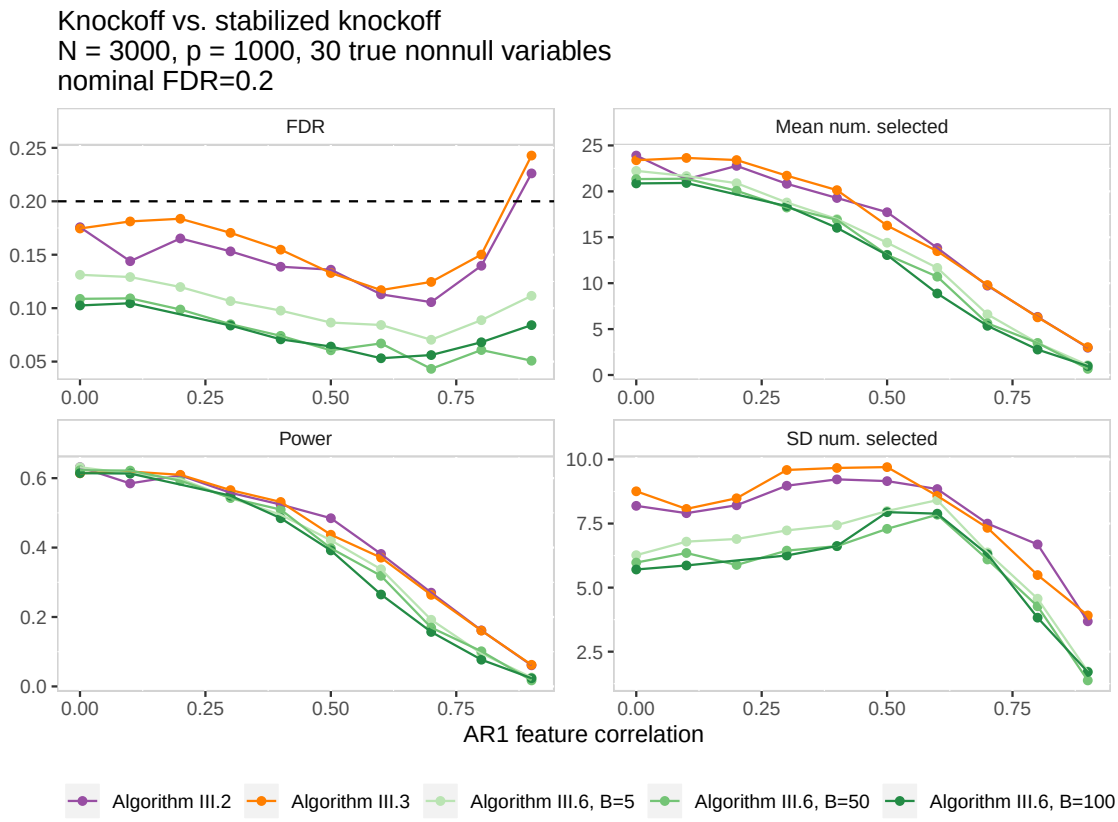


Figure 3.24: Power and FDR of knockoff and stabilized knockoff as a function of feature correlation with $n = 3000, p = 1000$. Right column displays the mean and variance of the number of selected variables in a given sample. Averaged over 200 simulated $(\mathbf{X}, \mathbf{Y})$.

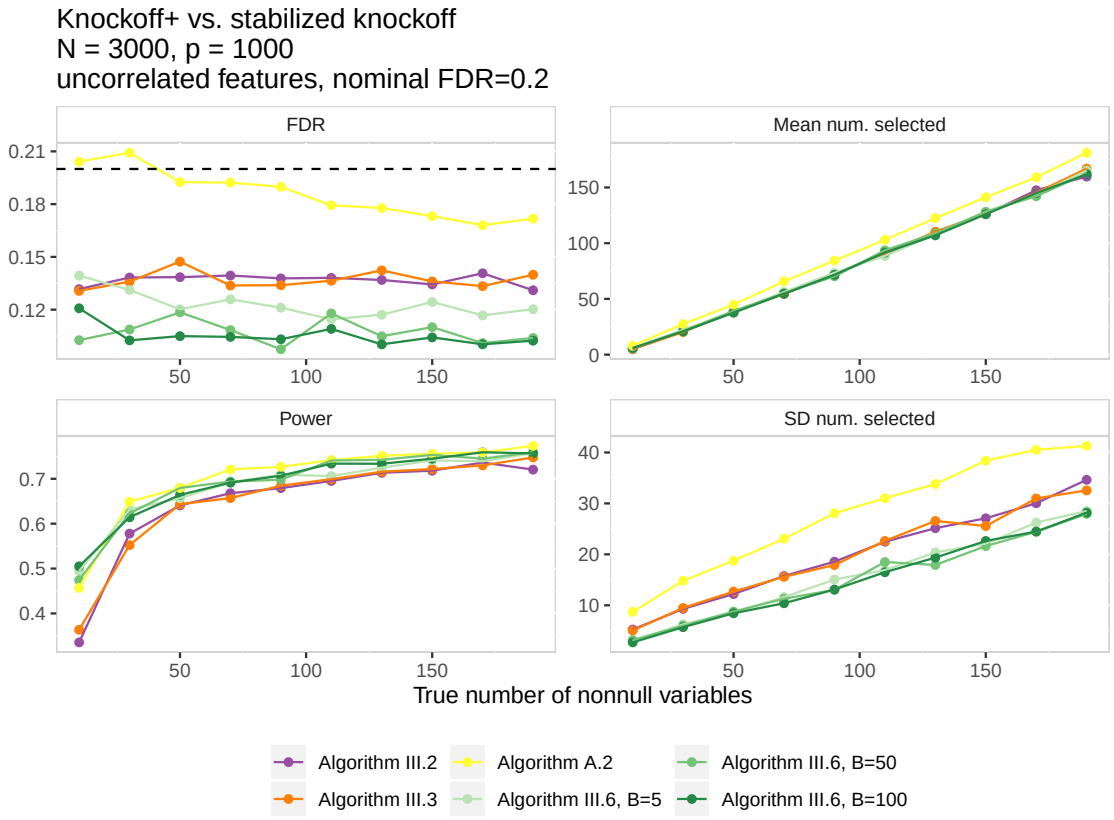


Figure 3.25: Power and FDR of knockoff+ and stabilized knockoff as a function of model sparsity with $n = 3000, p = 1000$. Right column displays the mean and variance of the number of selected variables in a given sample. Averaged over 200 simulated $(\mathbf{X}, \mathbf{Y})$.

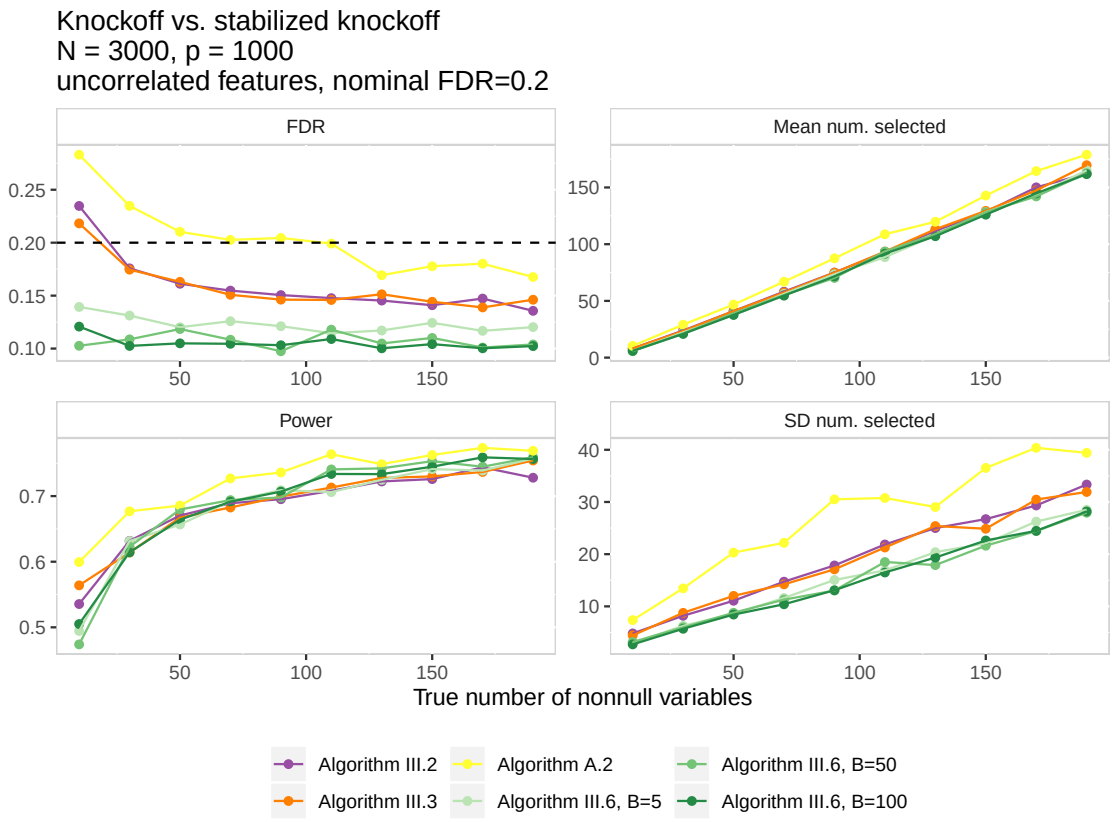


Figure 3.26: Power and FDR of knockoff and stabilized knockoff as a function of model sparsity with $n = 3000, p = 1000$. Right column displays the mean and variance of the number of selected variables in a given sample. Averaged over 200 simulated $(\mathbf{X}, \mathbf{Y})$.

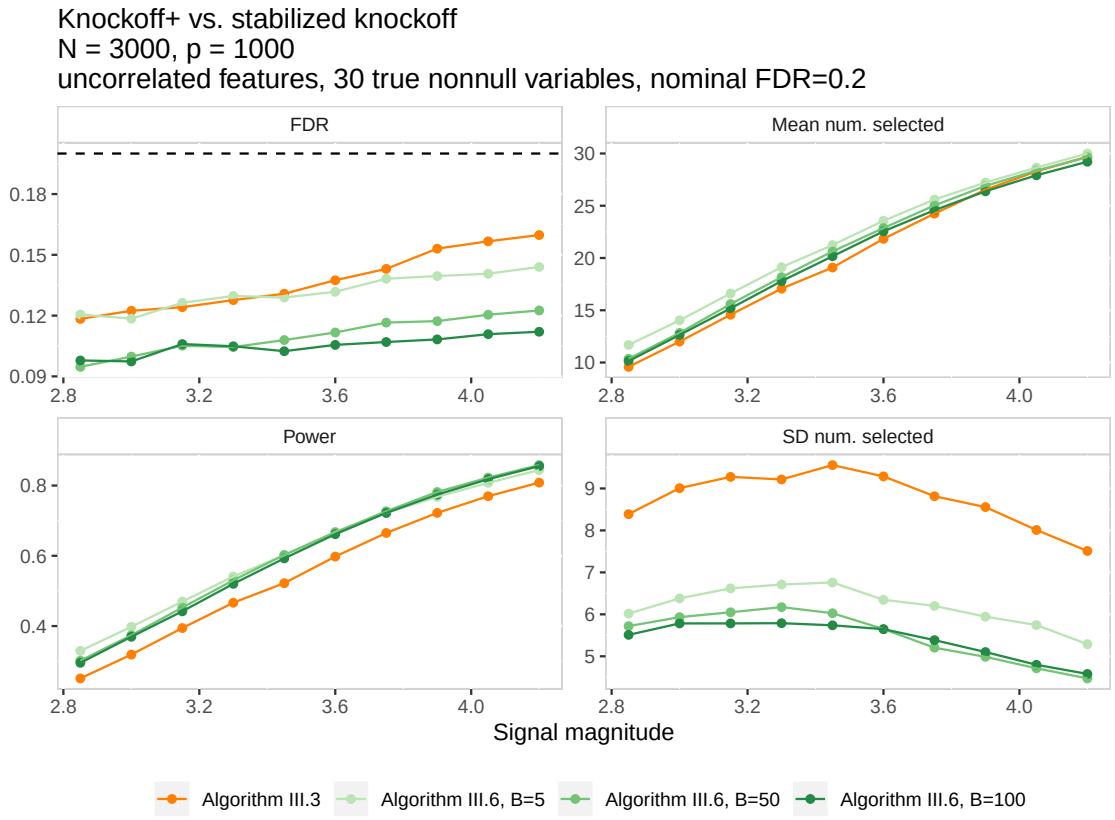


Figure 3.27: Power and FDR of knockoff+ and stabilized knockoff as a function of signal magnitude with $n = 3000, p = 1000$. Right column displays the mean and variance of the number of selected variables in a given sample. Averaged over 200 simulated $(\mathbf{X}, \mathbf{Y})$.

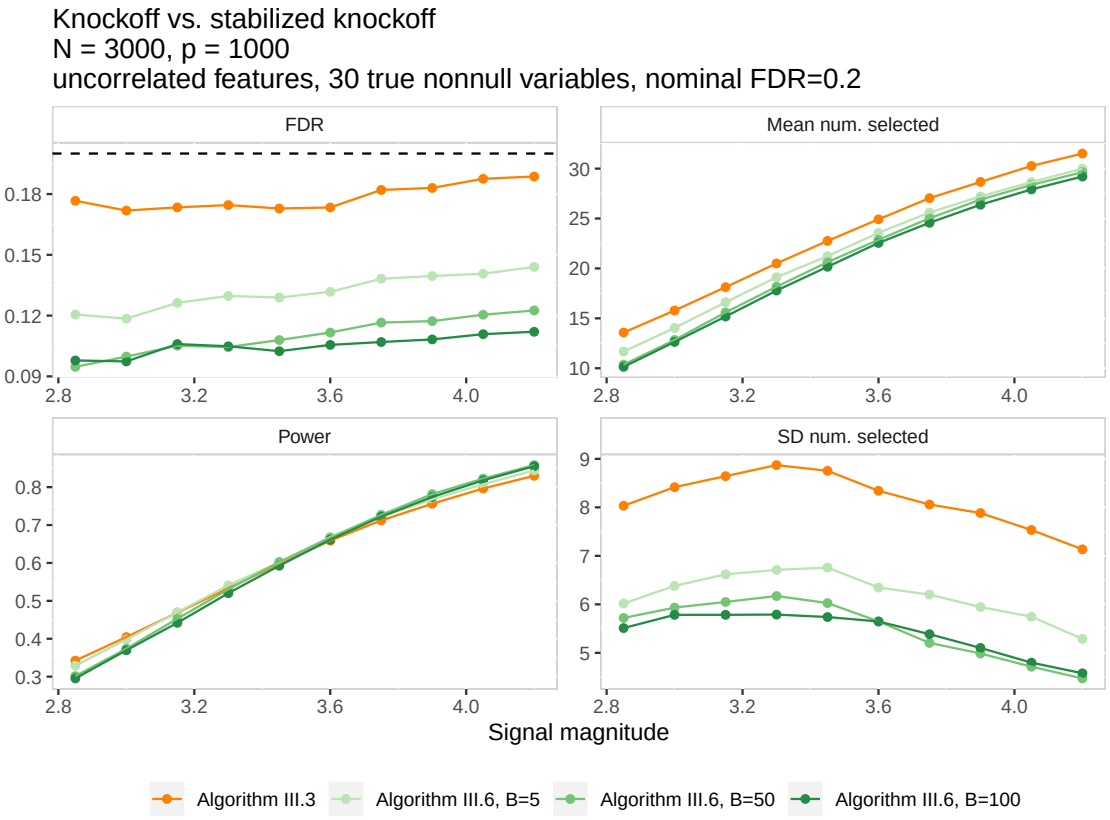


Figure 3.28: Power and FDR of knockoff and stabilized knockoff as a function of signal magnitude with $n = 3000, p = 1000$. Right column displays the mean and variance of the number of selected variables in a given sample. Averaged over 200 simulated $(\mathbf{X}, \mathbf{Y})$.

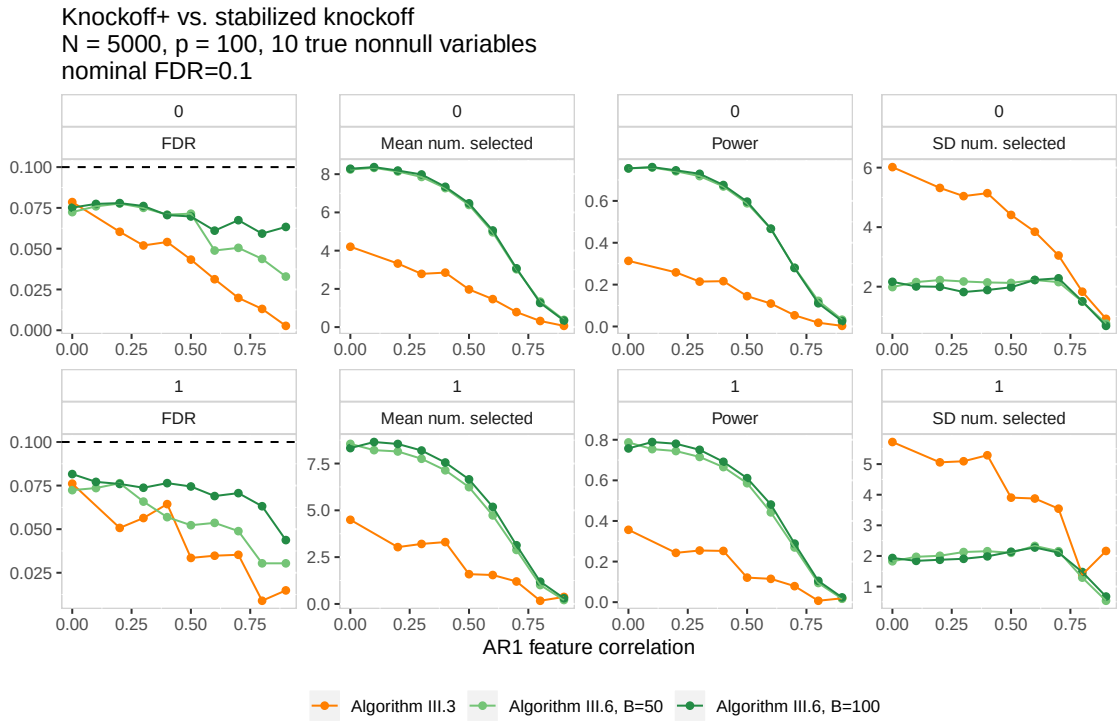


Figure 3.29: Power and FDR of knockoff+ and stabilized knockoff as a function of feature correlation with fixed $\mathbf{X}$, $n = 5000$, and $p = 100$. Second and fourth columns display the mean and variance of the number of selected variables in a given sample. Each row of plots corresponds to a single, fixed design matrix. Each point is an average over 200 replicates drawn from the $\mathbf{Y} \mid \mathbf{X}$ distribution.

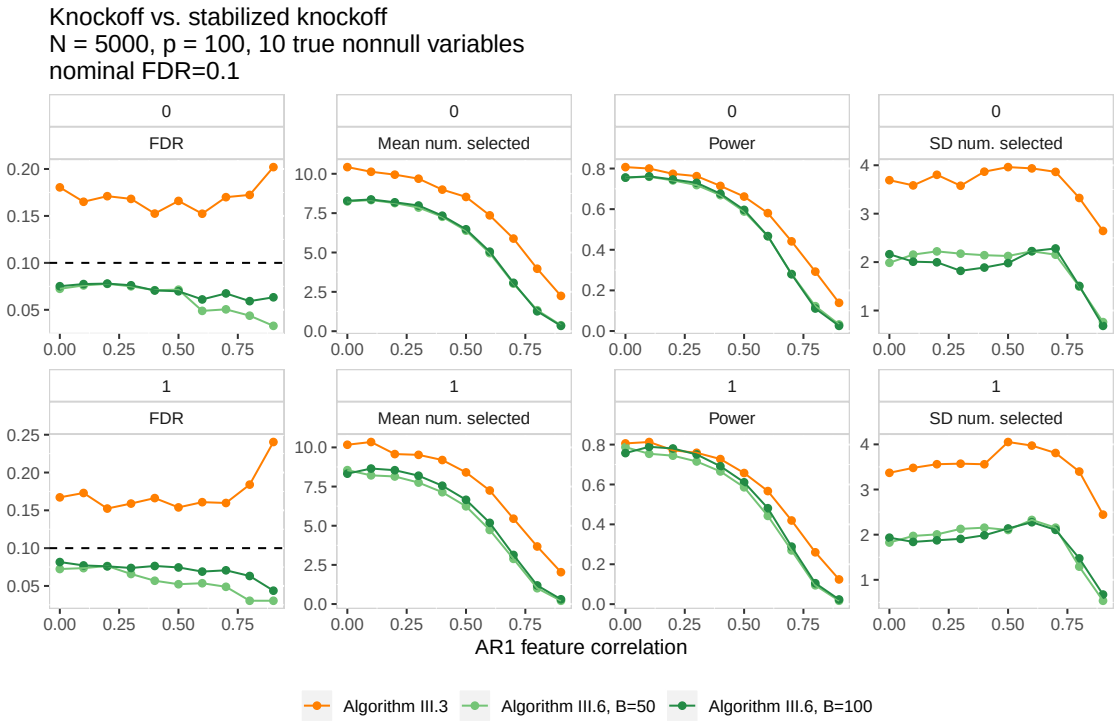


Figure 3.30: Power and FDR of knockoff and stabilized knockoff as a function of feature correlation with fixed $\mathbf{X}$, $n = 5000$, and $p = 100$. Second and fourth columns display the mean and variance of the number of selected variables in a given sample. Each row of plots corresponds to a single, fixed design matrix. Each point is an average over 200 replicates drawn from the $\mathbf{Y} \mid \mathbf{X}$ distribution.

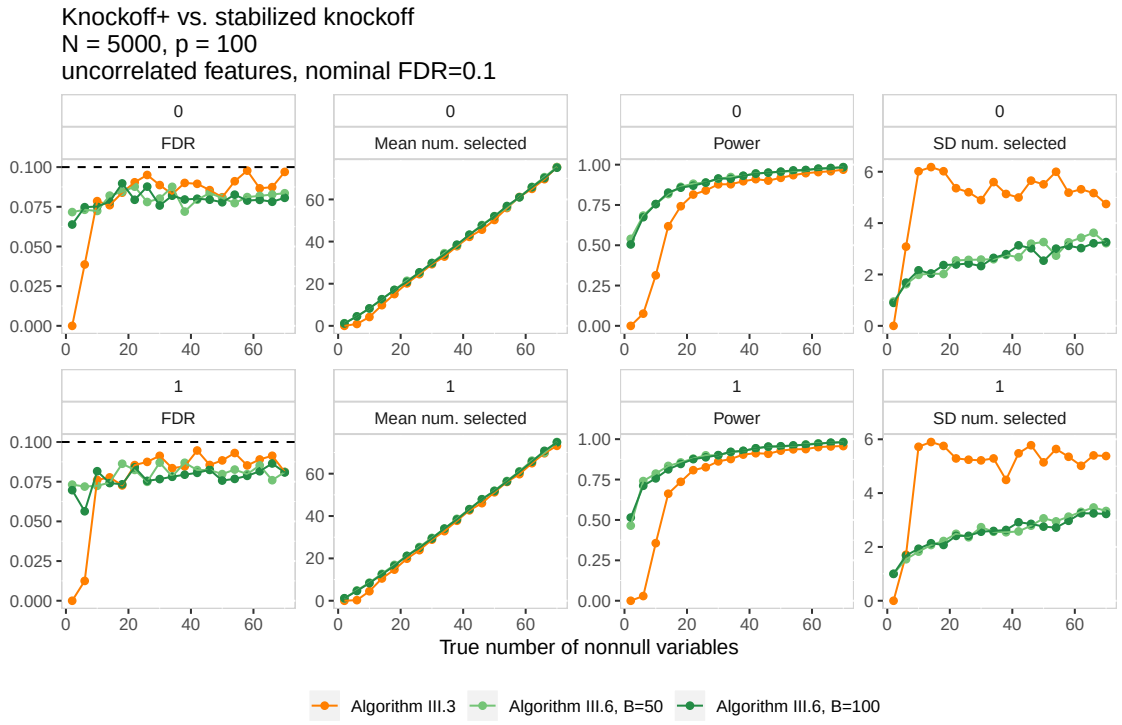


Figure 3.31: Power and FDR of knockoff+ and stabilized knockoff as a function of model sparsity with fixed $\mathbf{X}$, $n = 5000$, and $p = 100$. Second and fourth columns display the mean and variance of the number of selected variables in a given sample. Each row of plots corresponds to a single, fixed design matrix. Each point is an average over 200 replicates drawn from the $\mathbf{Y} \mid \mathbf{X}$ distribution.

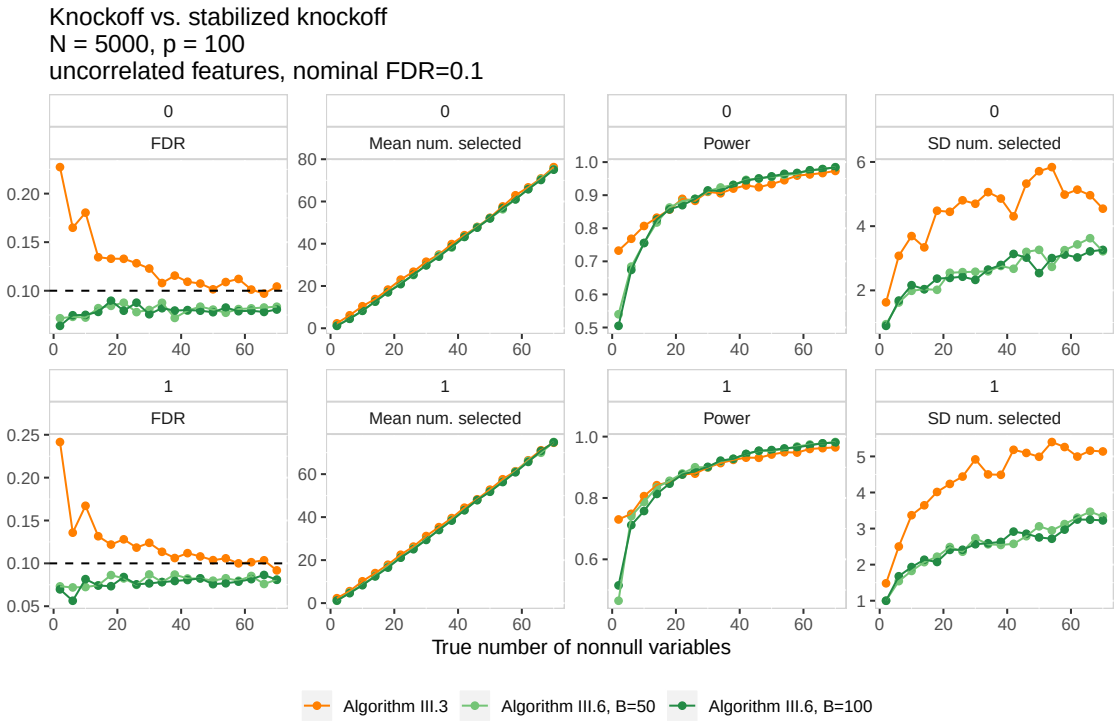


Figure 3.32: Power and FDR of knockoff and stabilized knockoff as a function of model sparsity with fixed $\mathbf{X}$, $n = 5000$, and $p = 100$. Second and fourth columns display the mean and variance of the number of selected variables in a given sample. Each row of plots corresponds to a single, fixed design matrix. Each point is an average over 200 replicates drawn from the $\mathbf{Y} \mid \mathbf{X}$ distribution.

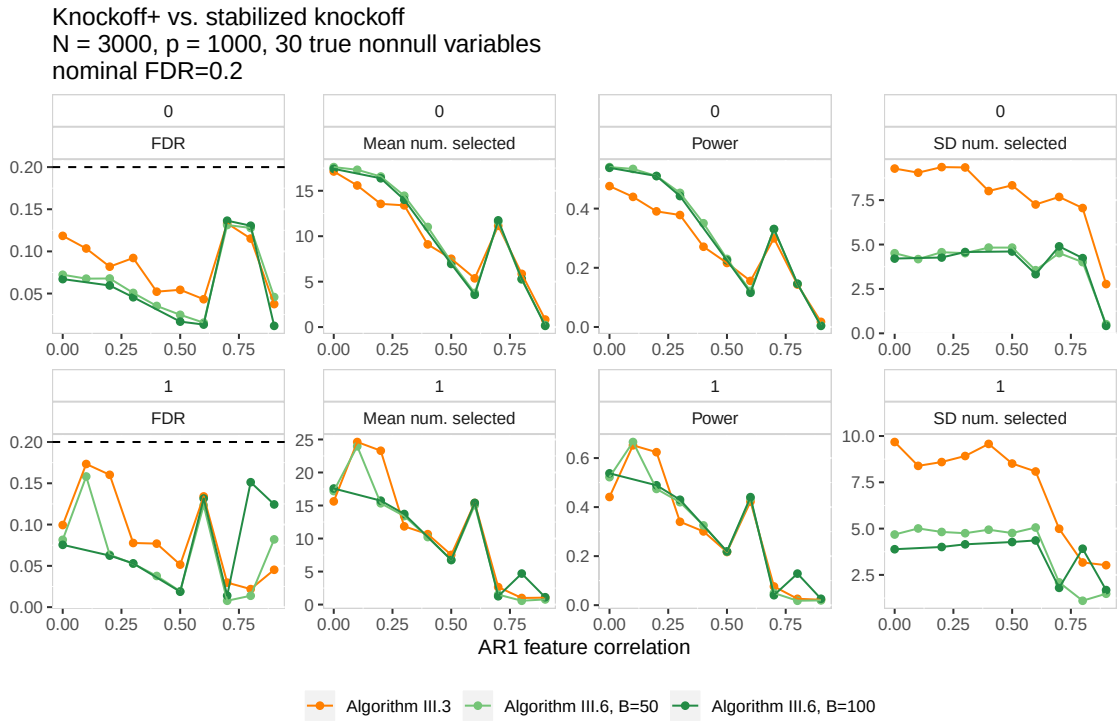


Figure 3.33: Power and FDR of knockoff+ and stabilized knockoff as a function of feature correlation with fixed $\mathbf{X}$, $n = 3000$, and $p = 1000$. Second and fourth columns display the mean and variance of the number of selected variables in a given sample. Each row of plots corresponds to a single, fixed design matrix. Each point is an average over 200 replicates drawn from the $\mathbf{Y} \mid \mathbf{X}$ distribution.

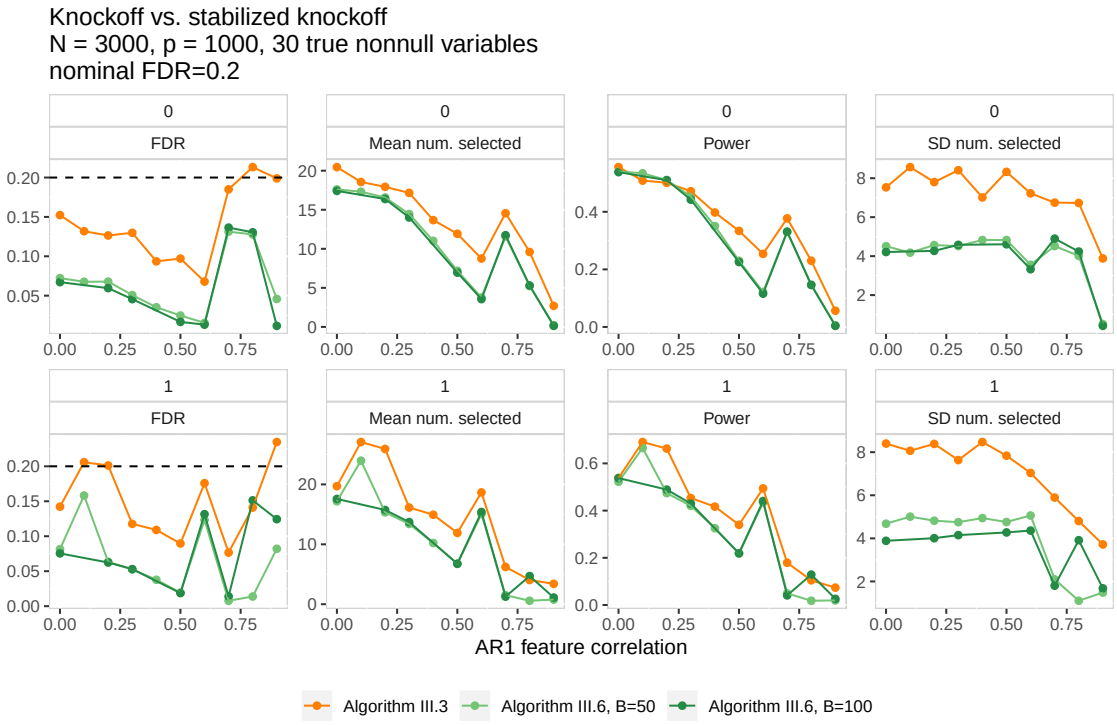


Figure 3.34: Power and FDR of knockoff and stabilized knockoff as a function of feature correlation with fixed $\mathbf{X}$, $n = 3000$, and $p = 1000$. Second and fourth columns display the mean and variance of the number of selected variables in a given sample. Each row of plots corresponds to a single, fixed design matrix. Each point is an average over 200 replicates drawn from the $\mathbf{Y} \mid \mathbf{X}$ distribution.

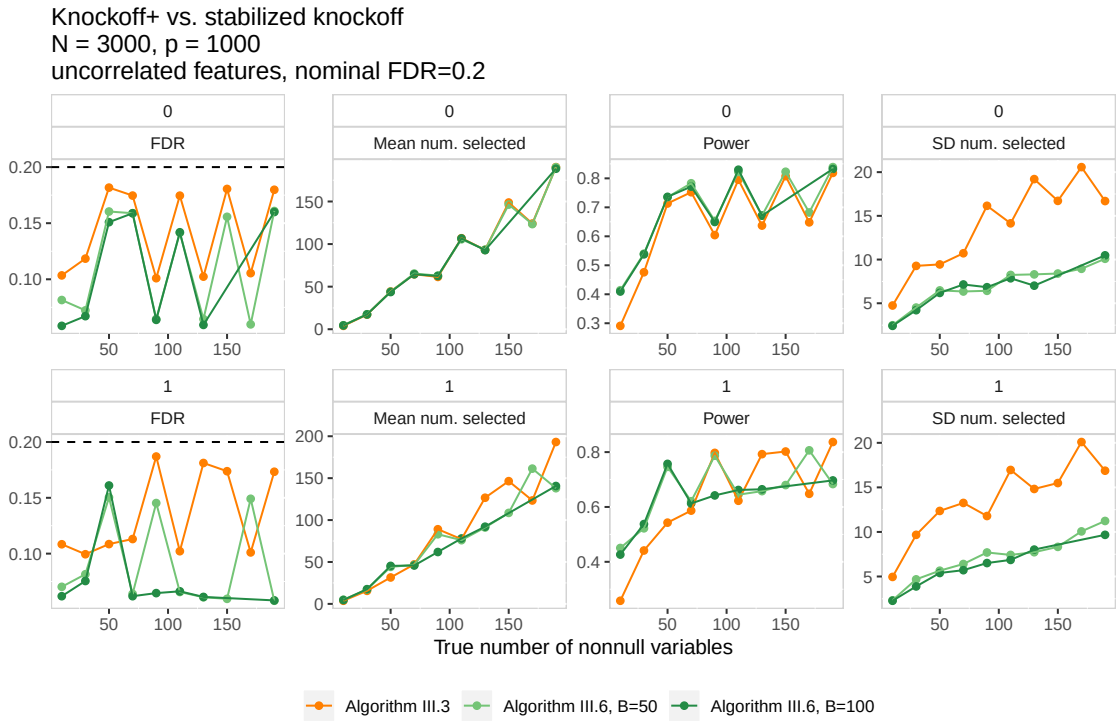


Figure 3.35: Power and FDR of knockoff+ and stabilized knockoff as a function of model sparsity with fixed $\mathbf{X}$, $n = 3000$, and $p = 1000$. Second and fourth columns display the mean and variance of the number of selected variables in a given sample. Each row of plots corresponds to a single, fixed design matrix. Each point is an average over 200 replicates drawn from the $\mathbf{Y} \mid \mathbf{X}$ distribution.

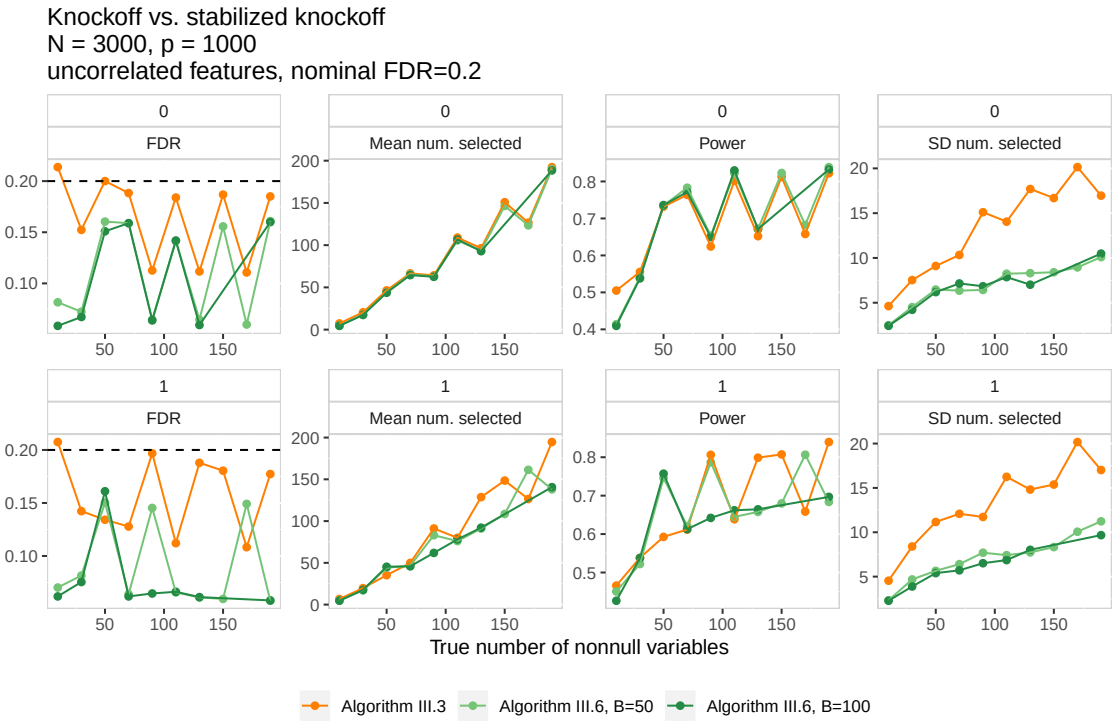


Figure 3.36: Power and FDR of knockoff and stabilized knockoff as a function of model sparsity with fixed $\mathbf{X}$, $n = 3000$, and $p = 1000$. Second and fourth columns display the mean and variance of the number of selected variables in a given sample. Each row of plots corresponds to a single, fixed design matrix. Each point is an average over 200 replicates drawn from the $\mathbf{Y} \mid \mathbf{X}$ distribution.

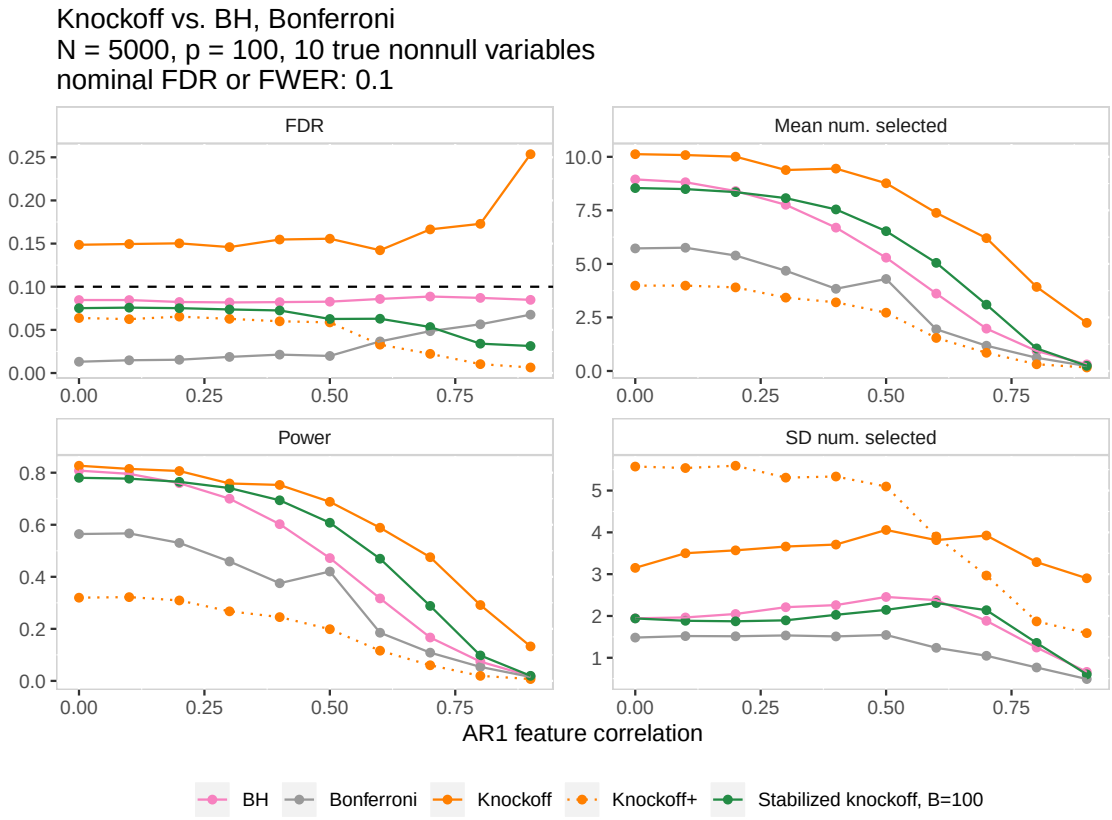


Figure 3.37: Power and FDR of stabilized knockoff and BH or Bonferroni adjusted OLS P-values as a function of feature correlation with $p = 100, n = 5000$. Right column displays the mean and variance of the number of selected variables in a given sample. Each point is an average of 500 simulation replicates.

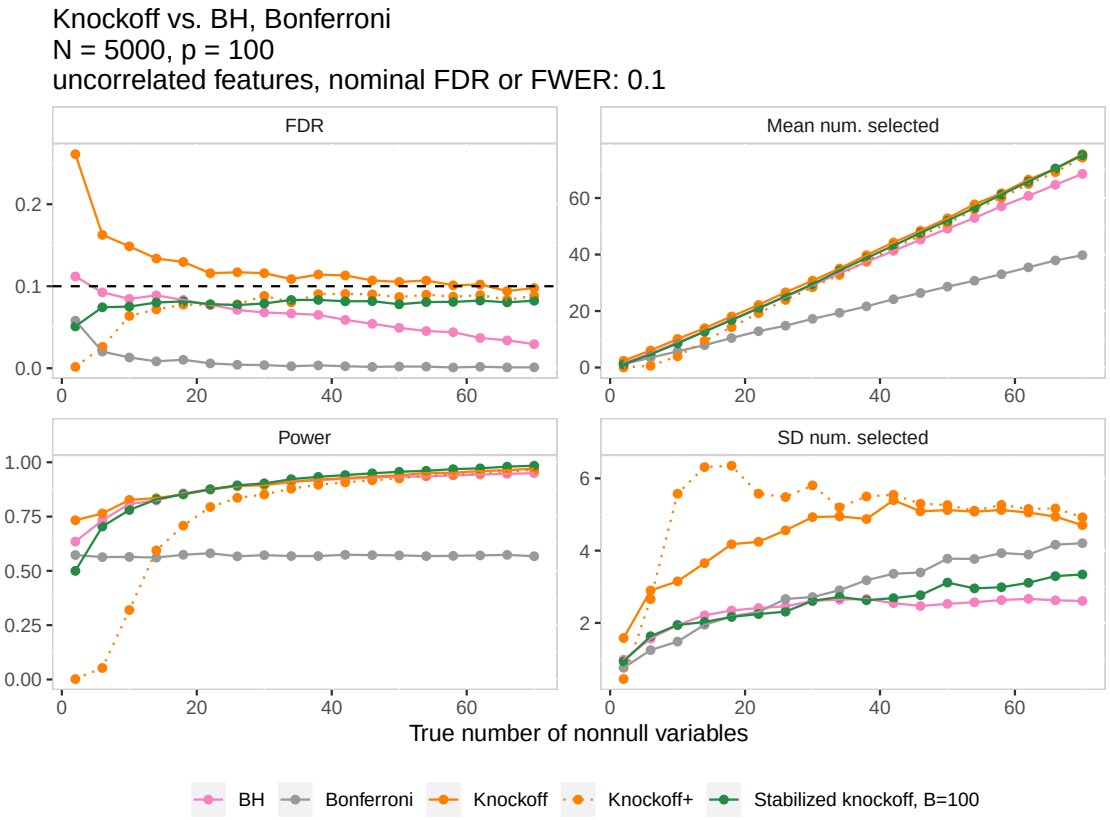


Figure 3.38: Power and FDR of stabilized knockoff and BH or Bonferroni adjusted OLS P-values as a function of model sparsity with $p = 100, n = 5000$. Right column displays the mean and variance of the number of selected variables in a given sample. Each point is an average of 500 simulation replicates.

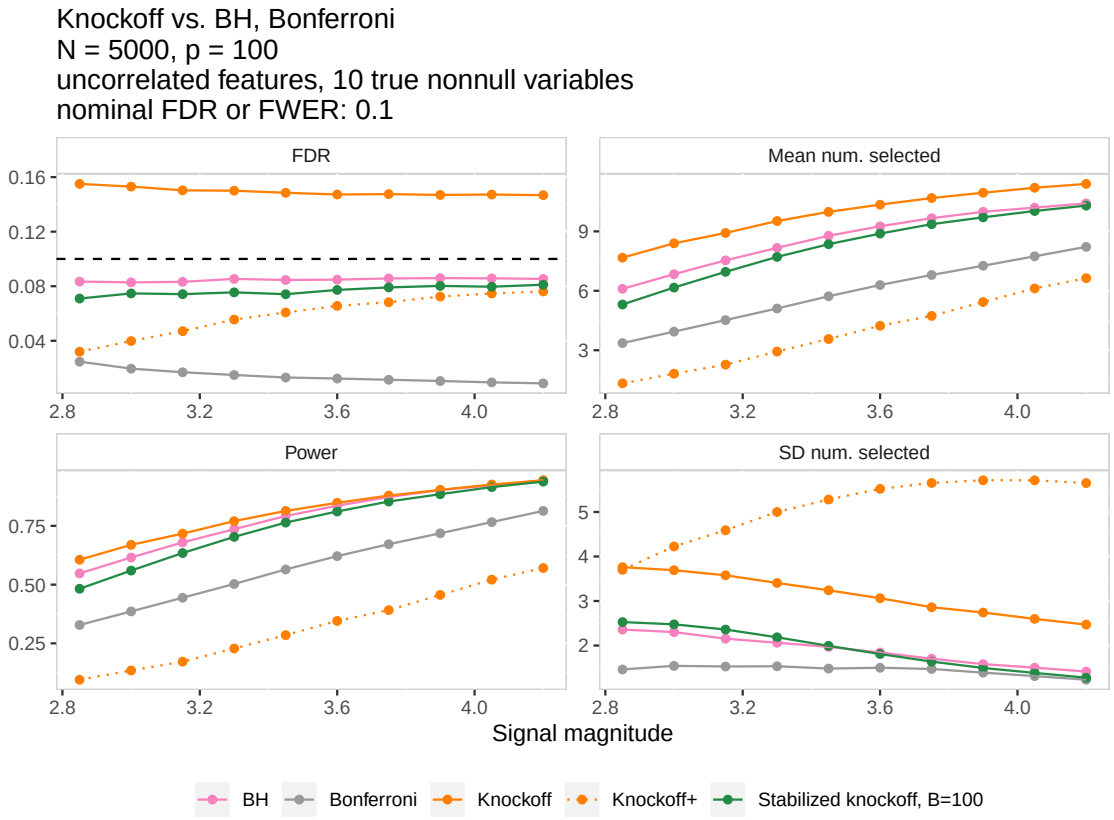


Figure 3.39: Power and FDR of stabilized knockoff and BH or Bonferroni adjusted OLS P-values as a function of signal magnitude with $p = 100, n = 5000$. Right column displays the mean and variance of the number of selected variables in a given sample. Each point is an average of 500 simulation replicates.

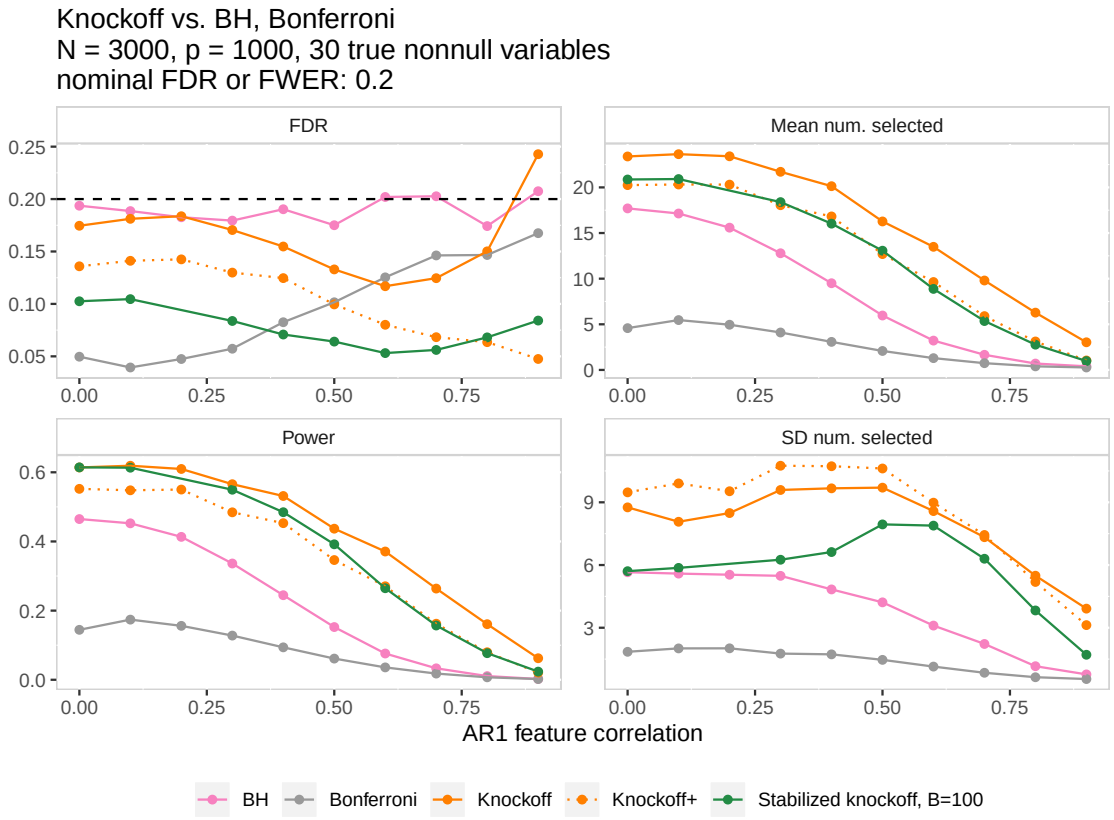


Figure 3.40: Power and FDR of stabilized knockoff and BH or Bonferroni adjusted OLS P-values as a function of feature correlation with $p = 1000, n = 3000$. Right column displays the mean and variance of the number of selected variables in a given sample. Each point is an average of 200 simulation replicates.

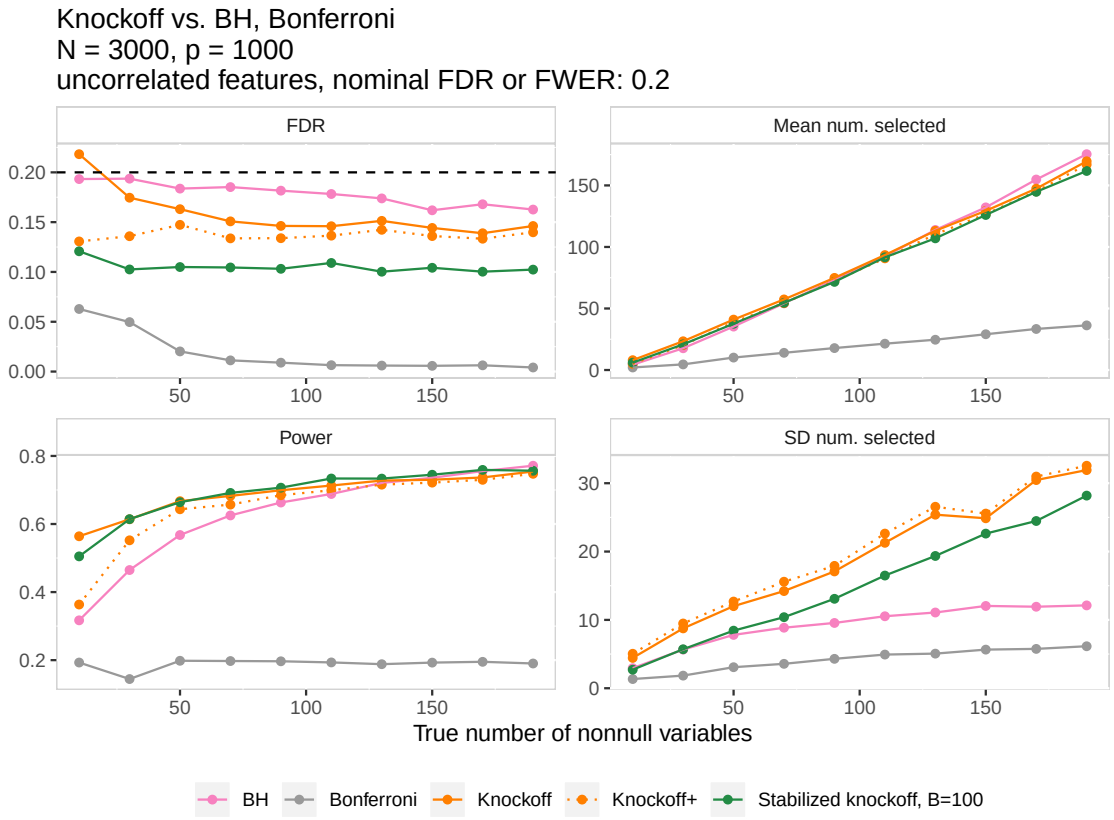


Figure 3.41: Power and FDR of stabilized knockoff and BH or Bonferroni adjusted OLS P-values as a function of model sparsity with $p = 1000, n = 3000$. Right column displays the mean and variance of the number of selected variables in a given sample. Each point is an average of 200 simulation replicates.

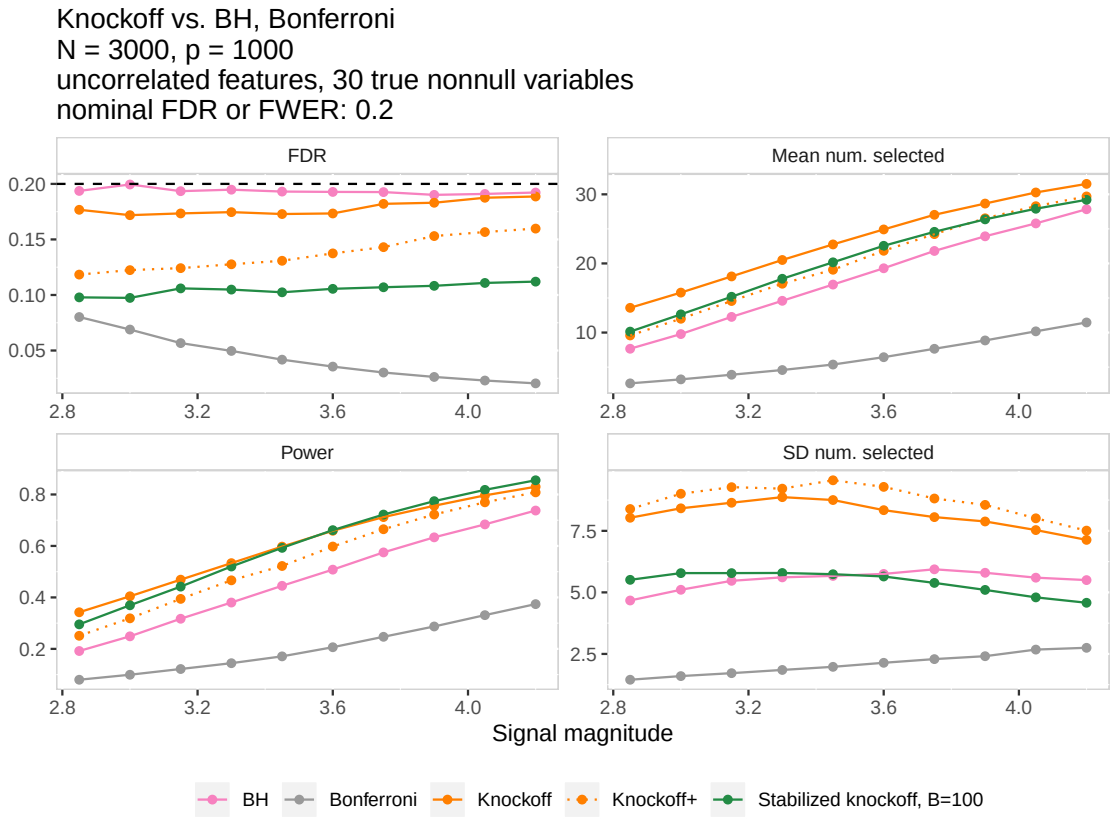


Figure 3.42: Power and FDR of stabilized knockoff and BH or Bonferroni adjusted OLS P-values as a function of signal magnitude with $p = 1000$, $n = 3000$. Right column displays the mean and variance of the number of selected variables in a given sample. Each point is an average of 200 simulation replicates.

CHAPTER IV

Linear Mixed Models for Comparing Dynamic Treatment Regimens on a Longitudinal Outcome in Sequentially Randomized Trials

4.1 Introduction

A dynamic treatment regimen (DTR) is a pre-specified sequence of decision rules which map baseline and time-varying measurements on an individual to a recommended set of interventions (Chakraborty and Moodie 2013; Orellana, Rotnitzky, and Robins 2010; Hernán et al. 2006; Murphy et al. 2001). DTRs are designed to assist clinicians with ongoing care decisions based on disease progress, treatment history, and other information collected during the course of treatment. DTRs are also known as adaptive treatment strategies (Kosorok and Moodie 2016; Murphy et al. 2007) or adaptive interventions (Almirall et al. 2014; Nahum-Shani et al. 2012).

A sequential, multiple assignment randomized trial (SMART) is a multi-stage trial design specifically created for comparing or constructing DTRs (Wallace, Moodie, and Stephens 2016; Kosorok and Moodie 2016; Lavori and Dawson 2014; Lei et al. 2012; Murphy 2005). Study participants in a SMART may experience multiple randomizations. These randomizations occur at decision points for which there is a question about which treatment to provide. By the end of the trial, specific groups

of study participants will have been subject to the sequence of treatment decisions corresponding to at least one of a pre-specified set of DTRs. SMARTs enable causal comparisons among these "embedded" DTRs.

This chapter focuses on scientific questions which involve comparing the embedded DTRs in a SMART based on the mean of a continuous, longitudinal outcome. Often this is a primary scientific aim in a SMART (Seewald et al. 2019). One way of answering these questions involves directly specifying a model for the marginal mean of the longitudinal outcome under each DTR and estimating the parameters in that model using weighted estimating equations (Lu et al. 2016; Seewald et al. 2019). Similar methods are available when the longitudinal outcome is binary (Dziak et al. 2019), for a survival outcome (Li and Murphy 2011), and for clustered SMARTs where the embedded DTRs are applied to clusters of people but outcomes are measured on individuals within each cluster (NeCamp, Kilbourne, and Almirall 2017).

The purpose of this chapter is to develop linear mixed effects models for primary aim comparisons of the embedded DTRs in a SMART with a continuous, longitudinal outcome. Mixed models were applied to longitudinal analysis of a specialized SMART design in Dai and Shete (2016), in which the duration of first-stage treatment is modeled as a survival outcome. In this case the binary tailoring variable used to define specific decision rules under the DTRs is a function of this survival outcome. These authors estimate the mean under each DTR conditional on the first-stage treatment duration, which is equivalent to estimating the mean conditional on the tailoring variable. This approach also assumes independence between longitudinal outcomes and DTR tailoring variables. In contrast, we specify a mixed model, and estimate the mean under each DTR, marginally over interim tailoring variables without assuming independence between those tailoring variables and the longitudinal primary outcome.

Other work on longitudinal analysis of two-stage trials has focused on alternative trial designs with a single randomization at the end of the first stage. In this context,

Hsu and Wahed (2017) estimate the marginal mean under each DTR using weighted estimating equations specified separately for each DTR. Miyahara and Wahed (2012) consider a single-randomization design in which the longitudinal outcomes are measured only during the second treatment stage. These authors estimate the marginal mean under each DTR by specifying a mixed model separately for each participant subgroup receiving a specific sequence of treatment decisions under one of the DTRs. By assuming that the longitudinal outcome is independent of the tailoring variables which define these participant subgroups, the group-specific models can be averaged to estimate the marginal mean under a given DTR.

Here we propose mixed models to analyze SMART designs which, by definition, have more than one randomization for some participants, and we model the marginal mean under each DTR without assuming independence between longitudinal outcomes and the interim variables defining treatment subgroups in a DTR. Mixed effects models are a well established tool for analyzing longitudinal, clustered, or multilevel data in the medical, social, and agricultural sciences (Fitzmaurice, Laird, and Ware 2011; Raudenbush and Bryk 2002; Snijders and Bosker 2012; Searle, Casella, and McCulloch 2006; Goldstein 2011; Hedeker and Gibbons 2006). The methods in this chapter provide a way for researchers to analyze data from SMARTs using these familiar statistical tools. We adapt the mixed modeling framework to the SMART trial design to consistently estimate scientific quantities of interest, namely, the marginal mean under each DTR, without conditioning on any interim tailoring variables used to define DTR decision rules. We also propose a weighted prediction for subject-specific random effects, which may be used to assess subject-to-subject heterogeneity in the primary outcomes under each DTR.

In addition to broadening the applicability of mixed models to SMART analysis, we see at least three reasons why scientists might prefer mixed effects models when analyzing SMARTs.

First, mixed models provide an intuitive, flexible way to model within-person correlations among longitudinal outcomes. Existing statistical methods for SMARTs with a continuous, longitudinal outcome (Lu et al. 2016; Seewald et al. 2019) involve directly specifying a working model for the marginal covariance matrix of the repeated measures, as in generalized estimating equations (GEE, Liang and Zeger (1986)). In contrast, our mixed effects model indirectly parameterizes the marginal covariance using random effects—latent random variables which describe subject-specific change over time. This specification distinguishes within-subject and between-subject variation and provides an intuitive and flexible way to model the marginal covariance as a function of time and other covariates.

Second, modeling the within-person correlation among longitudinal measurements can improve statistical efficiency in estimating regression parameters (e.g. Diggle et al. 2002, Section 4.6), and mixed models easily parameterize rich covariance functions using few parameters, regardless of the number or spacing of measurement occasions (Fitzmaurice, Laird, and Ware (2011, Chapter 8), Hedeker and Gibbons (2006, Chapter 8)).

Third, mixed models provide predictions of subject-specific outcome trajectories via prediction of the random effects (Skrondal and Rabe-Hesketh (2009), Hedeker and Gibbons (2006, Chapter 4), Searle, Casella, and McCulloch (2006, Chapter 7)). While such predictions do not constitute the primary aim of comparing embedded DTRs in a SMART, they may be useful in understanding the type and magnitude of heterogeneity in person-specific change with respect to the embedded DTRs and in identifying individuals with unusual response trajectories.

This chapter will refer to an example SMART designed to compare three DTRs for improving spoken language in children with autism. Section 4.2 introduces this study design and provides a general description of SMARTs and embedded DTRs. Section 4.3 introduces our proposed mixed model for comparing embedded DTRs

in a SMART and Section 4.4 describes how we estimate parameters and predict random effects in this model. In Section 4.6 we report the results of simulation experiments which investigate the operating characteristics of our estimation method, and in Section 4.7 we illustrate the method using data from the autism SMART introduced in Section 4.2.

4.2 Sequential, Multiple-Assignment Randomized Trials

Sequential, multiple assignment randomized trials (SMARTs) are multi-stage randomized trial designs which were developed explicitly for the purpose of building high-quality DTRs (Murphy 2005; Lavori and Dawson 2000; Dawson and Lavori 2008). Each participant in a SMART may move through multiple stages of treatment, and the defining feature of a SMART is that some or all participants are randomized at more than one decision point. At each decision point, the purpose of randomization is to address a question concerning the dosage intensity, type, or delivery of treatment at that decision point. A common primary aim in a SMART is the marginal mean comparison of two or more embedded DTRs on a longitudinal research outcome. The following example SMART illustrates these ideas.

4.2.1 An Example SMART in Autism

The SMART shown in Figure 4.1 (Kasari et al. 2014) involved $N = 61$ children, between five and eight years old, who had a previous diagnosis of autism spectrum disorder and were considered "minimally verbal" (used fewer than 20 spontaneous different words during a baseline 20-minute language test). All eligible children were initially randomized, with equal probability, to a behavioral treatment, called JASP, or to JASP together with a speech-generating device, called AAC (augmentative or alternative communication). Both of these first-stage treatment arms in the SMART involved twice-weekly sessions with a trained language therapist. The first-stage

JASP+AAC arm required that the AAC device was used at least 50 percent of the time during these sessions.

At the end of the first treatment stage, which lasted 12 weeks, all children were classified as "responders" or "slow responders". Response was defined, prior to the trial, as an improvement of at least 25 percent on seven or more language measures (e.g. words used per minute) by the end of week 12. Children who did not satisfy this criterion were considered slow responders.

The second-stage treatments were determined as follows. Responders to the initial intervention were continued on that intervention for an additional 12 weeks. Slow responders to JASP+AAC were offered intensified JASP+AAC, which involved increasing the number of weekly sessions from two to three. Slow responders to JASP were re-randomized, with equal probability, to either intensified JASP or to JASP+AAC. The status of "responder" or "slow responder" in this SMART is known as an *embedded tailoring variable*, since it is used to restrict subsequent randomizations and is therefore a part of the embedded DTRs. The primary research outcome in this SMART was the total number of spontaneous socially communicative utterances in a 20-minute language sample, measured by an independent evaluator who was blind to the assigned treatment sequence. This primary outcome was measured four times: prior to the initial randomization (baseline), prior to the second randomization (at week 12), at the end of treatment (week 24) and at a follow-up assessment (week 36).

4.2.2 Embedded Dynamic Treatment Regimens

A dynamic treatment regimen (DTR) is a sequence of decisions rules that, for all individuals in a population of interest, guides the provision of treatment at each decision point based on information known up to that decision point. In the case of the autism SMART, a DTR is a sequence of decision rules that guides the first and second treatment decisions for both responders and slow responders.

Specifically, the autism SMART has three DTRs embedded within it. These are listed in Table 4.1. The DTR labeled (AAC, AAC+) starts with JASP and AAC, continues this treatment for responders and intensifies this treatment for slow responders. The other two DTRs start with JASP only. For slow responders, (JASP, JASP+) intensifies JASP alone while (JASP, AAC) augments JASP with AAC. Many SMARTs use a two-stage design in which only slow responders are randomized at the start of the second stage (Kidwell 2014; Gunlicks-Stoessel et al. 2016, e.g.). In this SMART, however, second-stage randomization was restricted based on a combination of first-stage treatment and response status. We use (a_1, a_2) to index the DTRs embedded in the SMART, where a_j denotes the treatment provided at the jth decision point. Table 4.1 enumerates the values of (a_1, a_2) for each DTR in the autism SMART.

4.3 Linear Mixed Models for Comparing Embedded DTRs

We aim to develop a linear mixed model for primary aim comparisons based on a pre-specified summary of the mean outcome under each DTR in a SMART. To do this, we use the potential outcomes framework to describe the sequence of primary outcome measurements as a function of the embedded DTRs. For simplicity, we focus on two-stage designs. With slight changes in notation, the methodology presented here may be generalized to more complex SMART designs.

4.3.1 Potential outcomes and observed data

For each embedded DTR, indexed by (a_1, a_2) where $a_1, a_2 \in \{-1, +1\}$, and for the ith SMART participant, $i = 1, \ldots, N$, let

$$\mathbf{Y}_i(a_1, a_2) = (Y_{it_{i1}}(a_1, a_2), Y_{it_{i2}}(a_1, a_2), \ldots, Y_{it_{in_i}}(a_1, a_2))^\intercal$$

denote the vector of n_i time-ordered, potential outcome repeated measures. The vector $\mathbf{Y}_i(a_1, a_2)$ is simply the set of longitudinal potential outcomes for participant i under DTR (a_1, a_2). For example, in the case of the autism SMART, each participant has three potential values of $\mathbf{Y}_i(a_1, a_2)$, corresponding to the three values for (a_1, a_2) given in Table 4.1. Note that in the autism study, a_2 is undefined for the DTR beginning with JASP+AAC, since that DTR is fully characterized by $a_1 = -1$ (slow responders to JASP+AAC were not re-randomized). In the autism SMART and in many other common designs, the embedded tailoring variable is a binary summary of the data collected during the first stage and often represents early or delayed response to first-stage treatment. Let $R_i(a_1) \in \{0, 1\}$ be the potential outcome for the binary embedded tailoring variable under first-stage treatment a_1.

During the conduct of a SMART, we collect the following observed data: $Y_{it_{ij}}$, the observed primary outcome for participant i at time point t_{ij}; R_i, the ith participant's observed binary tailoring variable; $\mathbf{L}_i$, a pre-specified vector of baseline covariates collected prior to the first randomization; and A_{1i}, A_{2i}, the random treatment assignments in the first and second stage, respectively.

In the autism SMART, the primary outcome was collected for all children at each of $n_i = 4$ measurement occasions, occurring 12 weeks apart. So in this example we let $t_{ij} = t_j \in \{0, 12, 24, 36\}$ denote the time, in weeks, since baseline assessment. In the autism SMART, A_{1i} is equal to 1 or -1 with equal probability, indicating randomization to either JASP or JASP+AAC. Among slow responders to JASP, that is, among all subjects with $A_{1i} = 1$ and $R_i = 0$, A_{2i} equals 1 or -1 with equal probability, denoting randomization to receive either intensified JASP or the AAC device. In the autism study, A_{2i} is not defined for responders and participants randomized to $A_{1i} = -1$.

4.3.2 The model

For the ith participant and for a fixed DTR (a_1, a_2), consider the following linear mixed effects model:

$$\mathbf{Y}_i(a_1, a_2) = \mathbf{X}_i(a_1, a_2)\boldsymbol{\beta} + \mathbf{Z}_i(a_1, a_2)\mathbf{b}_i + \boldsymbol{\epsilon}_i(a_1, a_2), \tag{4.1}$$

where $\boldsymbol{\beta}$ is an unknown p-dimensional parameter, $\mathbf{b}_i$ is a q-dimensional ($q \leq p$) latent random vector (the random effects) with $\mathbb{E}(\mathbf{b}_i \mid \mathbf{L}_i) = \mathbf{0}$ and $\boldsymbol{\epsilon}_i(a_1, a_2)$ is the n_i-length vector of within-subject residual errors with $\mathbb{E}(\boldsymbol{\epsilon}_i(a_1, a_2) \mid \mathbf{L}_i) = \mathbf{0}$. We also assume that $\boldsymbol{\epsilon}_i(a_1, a_2)$ is independent of $\mathbf{b}_i$, given $\mathbf{L}_i$. The $n_i \times p$ design matrix $\mathbf{X}_i(a_1, a_2)$ depends on the SMART design and a chosen model for the mean, conditional on the baseline covariate vector $\mathbf{L}_i$. The $n_i \times q$ random effects design matrix $\mathbf{Z}_i(a_1, a_2)$ is a function of time, $\mathbf{X}_i(a_1, a_2)$, and (a_1, a_2) chosen so that $\mathbf{Z}_i(a_1, a_2)\mathbf{b}_i$ models subject-specific deviations from the mean over time. Since (a_1, a_2) indexes the embedded DTRs and is not a random variable, $\mathbf{X}_i(a_1, a_2)$ and $\mathbf{Z}_i(a_1, a_2)$ are random variables only as a function of $\mathbf{L}_i$. (We do not treat t_{ij} as a random variable.) Note that model (4.1) implies that $\mathbb{E}(\mathbf{Y}_i(a_1, a_2) \mid \mathbf{L}_i, \mathbf{b}_i) = \mathbf{X}_i(a_1, a_2)\boldsymbol{\beta} + \mathbf{Z}_i(a_1, a_2)\mathbf{b}_i$ and $\mathbb{E}(\mathbf{Y}_i(a_1, a_2) \mid \mathbf{L}_i) = \mathbf{X}_i(a_1, a_2)\boldsymbol{\beta}$.

With model (4.1), we make primary aim comparisons among embedded DTRs based on the linear, parametric marginal model for $\mathbb{E}(Y_{it}(a_1, a_2) \mid \mathbf{L}_i)$ given by $\boldsymbol{\beta}^\intercal \mathbf{X}_{it}(a_1, a_2)$, where $\mathbf{X}_{it}(a_1, a_2)^\intercal$ is the row of $\mathbf{X}_i(a_1, a_2)$ evaluated at $t_{ij} = t$, and $\boldsymbol{\beta}$ is a p-dimensional column-vector of unknown parameters. Recall that $\mathbf{L}_i$ is a vector of baseline covariates and that (a_1, a_2) indexes the embedded DTRs and is not a random variable. This is a *marginal* mean model in that $\mathbb{E}(Y_{it}(a_1, a_2) \mid \mathbf{L}_i)$ is marginal over both the embedded tailoring variable, $R_i(a_1)$, and any other intermediate random variables possibly impacted by a_1 or (a_1, a_2). For the autism SMART, an example marginal mean model used previously (Lu et al. 2016; Almirall et al. 2016) is a piecewise linear

model with a knot at week $t_j = 12$:

$$\begin{aligned}&\boldsymbol{\beta}^\intercal \mathbf{X}_{it}(a_1, a_2) \\ &= \beta_0 + t^{[0,12]}\left(\beta_1 + \beta_2 a_1\right) + t^{(12,36]}\left(\beta_3 + \beta_4 a_1 + \beta_5 \mathbb{I}\left[a_1 = 1\right] a_2\right) + \beta_6 \mathtt{age}_i,\end{aligned} \tag{4.2}$$

where $\mathbb{I}\left[\cdot\right]$ is the indicator function, $t^{[0,12]} = (t\mathbb{I}\left[t \leq 12\right] + 12\mathbb{I}\left[t > 12\right])$, $t^{(12,36]} = (t - 12)\mathbb{I}\left[t > 12\right]$, and $L_i = \mathtt{age}_i$ is the mean-centered age at baseline. In this case,

$$\mathbf{X}_{it}(a_1, a_2) = [1, t^{[0,12]}, t^{[0,12]}a_1, t^{(12,36]}, t^{(12,36]}a_1, t^{(12,36]}\mathbb{I}\left[a_1 = 1\right] a_2, \mathtt{age}_i]^\intercal.$$

In this example, the parameters β_2, β_4, and β_5 have a causal interpretation and can be used to specify the DTR effect estimands of primary interest. An example estimand of primary interest may be $\mathbb{E}\left(Y_{i24}(1,1)\right) - \mathbb{E}\left(Y_{i24}(-1,\cdot)\right) = 12(2(\beta_2+\beta_4)+\beta_5)$, an end-of-treatment comparison between the DTR with no AAC, $(a_1, a_2) = (1, 1)$, and the DTR with the highest dose of AAC, $(a_1, a_2) = (-1, \cdot)$. Other DTR effect estimands are similarly formed via linear combinations of β_2, β_4, and β_5.

In addition to specifying $\boldsymbol{\beta}^\intercal \mathbf{X}_{it}(a_1, a_2)$ as a model for $\mathbb{E}\left(Y_{it}(a_1, a_2) \mid \mathbf{L}_i\right)$, model (4.1) implicitly defines a working model for the marginal covariance $\mathbf{V}_i(a_1, a_2) := \mathrm{Var}\left(\mathbf{Y}_i(a_1, a_2) \mid \mathbf{L}_i\right)$. Since we assume $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i(a_1, a_2)$ are independent given $\mathbf{L}_i$, we have $\mathbf{V}_i(a_1, a_2) = \mathbf{Z}_i(a_1, a_2)\mathrm{Var}(\mathbf{b}_i \mid \mathbf{L}_i)\mathbf{Z}_i(a_1, a_2)^\intercal + \mathrm{Var}(\boldsymbol{\epsilon}_i(a_1, a_2) \mid \mathbf{L}_i)$. Previously, models for SMARTs with a longitudinal outcome involved directly specifying a working model for $\mathbf{V}_i(a_1, a_2)$ (Seewald et al. 2019; Almirall et al. 2016; Lu et al. 2016). In contrast, the working model for $\mathbf{V}_i(a_1, a_2)$ in (4.1) is a consequence of separately modeling within-subject and between-subject variation via $\mathbf{Z}_i(a_1, a_2)\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i(a_1, a_2)$. Together, the variance and covariance structures specified for $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i(a_1, a_2)$ imply a working model for $\mathbf{V}_i(a_1, a_2)$.

4.4 Estimation and prediction

To derive a set of estimating equations for $\boldsymbol{\beta}$, we initially consider the following two distributional assumptions, which are typical for a mixed model like (4.1):

$$\mathbf{b}_i \mid \mathbf{L}_i \sim N(\mathbf{0}, \mathbf{G}) \qquad \boldsymbol{\epsilon}_i(a_1, a_2) \mid \mathbf{L}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i}) \tag{4.3}$$

With the addition of the assumptions in (4.3), we have

$$\mathbf{Y}_i(a_1, a_2) \mid \mathbf{L}_i \sim N(\mathbf{X}_i(a_1, a_2)\boldsymbol{\beta}, \mathbf{V}_i(a_1, a_2)) \tag{4.4}$$

with $\mathbf{V}_i(a_1, a_2) = \mathbf{Z}_i(a_1, a_2)\mathbf{G}\mathbf{Z}_i(a_1, a_2)^\intercal + \sigma^2 \mathbf{I}_{n_i}$. Based on this distribution for $\mathbf{Y}_i(a_1, a_2) \mid \mathbf{L}_i$, the log-likelihood for a sample of N participants under DTR (a_1, a_2) is

$$\begin{aligned} &-\frac{1}{2}\sum_{i=1}^{N} \log\det\left[\mathbf{V}_i(a_1, a_2)\right] \\ &\quad -\frac{1}{2}\sum_{i=1}^{N} (\mathbf{Y}_i(a_1, a_2) - \mathbf{X}_i(a_1, a_2)\boldsymbol{\beta})^\intercal \mathbf{V}_i(a_1, a_2)^{-1} (\mathbf{Y}_i(a_1, a_2) - \mathbf{X}_i(a_1, a_2)\boldsymbol{\beta}), \end{aligned} \tag{4.5}$$

In practice, this log-likelihood cannot be maximized since the potential outcomes $\mathbf{Y}_i(a_1, a_2)$ are not observed for all participants under all DTRs in a SMART. Instead, we propose a weighted pseudo-likelihood based on the observed data collected in a SMART.

4.4.1 Pseudo-Likelihood Estimation

The log-likelihood (4.5) is a function of the following parameters: $\boldsymbol{\beta}$, σ^2 and the unique parameters in $\mathbf{G}$. We let $\boldsymbol{\alpha}$ denote the vector of unique variance parameters in $\mathbf{V}_i(a_1, a_2) = \mathbf{V}_i(a_1, a_2; \boldsymbol{\alpha})$, including σ^2. For example, if $\mathbf{b}_i$ is a scalar random variable and $Z_{it}(a_1, a_2) = 1$ for all a_1, a_2 and t, then $\boldsymbol{\alpha} = (\sigma^2, \mathrm{Var}(\mathbf{b}_i|\mathbf{L}_i))$. For brevity, we often

suppress notation indicating that $\mathbf{V}_i(a_1, a_2) = \mathbf{V}_i(a_1, a_2; \boldsymbol{\alpha})$ depends on $\boldsymbol{\alpha}$. Given the observed data in a SMART, defined in section 4.3.1, the pseudo-likelihood we use to estimate $\boldsymbol{\beta}$ is

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{a_1, a_2} \tilde{W}_i(a_1, a_2) \left(\log \det \left[\mathbf{V}_i(a_1, a_2) \right] + \mathbf{r}_i(a_1, a_2)^\intercal \mathbf{V}_i(a_1, a_2)^{-1} \mathbf{r}_i(a_1, a_2) \right), \tag{4.6}$$

where $\mathbf{r}_i(a_1, a_2) = \mathbf{r}_i(a_1, a_2; \boldsymbol{\beta}) = \mathbf{Y}_i - \mathbf{X}_i(a_1, a_2)\boldsymbol{\beta}$ and

$$\tilde{W}_i(a_1, a_2) = I_i^{(a_1, a_2)}(A_{1i}, R_i, A_{2i}) W_i^{(a_1, a_2)}(R_i).$$

The indicator $I_i^{(a_1, a_2)}(A_{1i}, R_i, A_{2i})$ is equal to one if and only if the sequence (A_{1i}, R_i, A_{2i}) is observable under DTR (a_1, a_2). For example, in the autism SMART, $I_i^{(a_1, a_2)}(A_{1i}, R_i, A_{2i}) = \mathbb{I}\left[A_{1i} = a_1\right] \left(R_i + (1 - R_i) \mathbb{I}\left[A_{2i} = a_2\right]\right)$, where $\mathbb{I}\left[v\right]$ equals 1 if the event v occurs and equals zero otherwise. The design-specific weight $W_i^{(a_1, a_2)}(R_i) := \mathbb{P}\left(A_{1i} = a_1, A_{2i} = a_2 \mid R_i\right)^{-1}$ is an inverse probability weight for the DTR (a_1, a_2) which depends on R_i because second-stage randomization is restricted according to this binary tailoring variable. In the autism SMART, and in many two-stage designs, only individuals with $R_i = 0$ are re-randomized, and

$$W_i^{(a_1, a_2)}(R_i) = \frac{1}{\mathbb{P}\left(A_{1i} = a_1\right)} \left[R_i + \frac{1}{\mathbb{P}\left(A_{2i} = a_2 \mid A_{1i} = a_1, R_i = 0\right)} (1 - R_i) \right]. \tag{4.7}$$

(When A_{2i} is not defined for a given value of a_1, we set $\mathbb{P}\left(A_{2i} = a_2 \mid A_{1i} = a_1, R_i = 0\right) = 1$.)

Differentiating (4.6) with respect to $\boldsymbol{\beta}$ leads to the following p-dimensional set of

estimating equations:

$$\sum_{i=1}^{N} \sum_{a_1,a_2} \tilde{W}_i(a_1,a_2)\mathbf{X}_i(a_1,a_2)^\intercal \mathbf{V}_i(a_1,a_2;\boldsymbol{\alpha})^{-1}\mathbf{r}_i(a_1,a_2;\boldsymbol{\beta}) = \mathbf{0}, \tag{4.8}$$

with the solution

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = &\left(\sum_{i=1}^{N} \sum_{a_1,a_2} \tilde{W}_i(a_1,a_2)\mathbf{X}_i(a_1,a_2)^\intercal \mathbf{V}_i(a_1,a_2;\boldsymbol{\alpha})^{-1}\mathbf{X}_i(a_1,a_2)\right)^{-1} \\ &\left(\sum_{i=1}^{N} \sum_{a_1,a_2} \tilde{W}_i(a_1,a_2)\mathbf{X}_i(a_1,a_2)^\intercal \mathbf{V}_i(a_1,a_2;\boldsymbol{\alpha})^{-1}\mathbf{Y}_i\right).\end{aligned} \tag{4.9}$$

Substituting $\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha})$ into (4.6), we can obtain estimates of $\boldsymbol{\beta}$ by first computing $\hat{\boldsymbol{\alpha}} = \arg\max_\alpha l(\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}),\boldsymbol{\alpha})$ and then estimating $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\alpha}})$. In the following theorem we derive the asymptotic properties of $\hat{\boldsymbol{\beta}}$.

Theorem IV.1. *Define* $\mathbf{U}_i(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{a_1,a_2} \tilde{W}_i(a_1,a_2)\mathbf{X}_i(a_1,a_2)^\intercal \mathbf{V}_i(a_1,a_2;\boldsymbol{\alpha})^{-1}(\mathbf{Y}_i - \mathbf{X}_i(a_1,a_2)\boldsymbol{\beta})$ *and let* $\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha})$ *be the solution to* $\sum_i \mathbf{U}_i(\boldsymbol{\beta},\boldsymbol{\alpha}) = \mathbf{0}$ *given in* (4.9). *Assume the following:*

i. Correctly specified marginal model: $\mathbb{E}\left(\mathbf{Y}_i(a_1,a_2) \mid \mathbf{L}_i\right) = \mathbf{X}_i(a_1,a_2)\boldsymbol{\beta}^*$

ii. Sequential randomization: $\mathbf{Y}_i(a_1,a_2)$ *is independent of* A_{1i} *given* $\mathbf{L}_i$*;* $R_i(a_1)$ *is independent of* A_{1i} *given* $\mathbf{L}_i$*; and* $\mathbf{Y}_i(a_1,a_2)$ *is independent of* A_{2i} *given* $(A_{1i}, R_i, \mathbf{L}_i)$.

iii. Consistency: $R_i = R_i(A_{1i}) = \sum_{a_1} \mathbb{I}\left[A_{1i} = a_1\right] R_i(a_1)$ *and*

$$\begin{aligned}\mathbf{Y}_i &= R_i\mathbf{Y}_i(A_{1i}) + (1-R_i)\mathbf{Y}_i(A_{1i},A_{2i}) \\ &= \sum_{a_1} \mathbb{I}\left[A_{1i}=a_1\right] R_i(a_1)\mathbf{Y}_i(a_1) + \sum_{a_1,a_2} \mathbb{I}\left[A_{1i}=a_1\right]\mathbb{I}\left[A_{2i}=a_2\right](1-R_i(a_1))\mathbf{Y}_i(a_1,a_2),\end{aligned}$$

where $R_i\mathbf{Y}_i(A_{1i}) := R_i\mathbf{Y}_i(A_{1i},a_2)$ *for all* a_2.

iv. Positivity: $\mathbb{P}(A_{1i} = a_1) > 0$ *and* $\mathbb{P}(A_{2i} = a_2 \mid A_{1i}, R_i = 0) > 0$ *for any* a_1, a_2.

v. Regularity conditions: For any given $\boldsymbol{\beta}$, $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})$ *converges to some* $\boldsymbol{\alpha}^*$ *at* $\sqrt{N}$ *rate, and*

$$\sup_{\beta} \left\| \frac{1}{N} \sum_i \mathbf{U}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})) - \lim_{N\to\infty} \frac{1}{N} \sum_i \mathbb{E}\left(\mathbf{U}_i(\boldsymbol{\beta}, \alpha^*(\boldsymbol{\beta})) \mid \mathbf{L}_i\right) \right\| \xrightarrow{P} 0.$$

Then $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\alpha}})$ *is consistent for* $\boldsymbol{\beta}^*$ *and* $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ *has an asymptotic* $N(0, \mathbf{B}^{-1}\mathbf{M}\mathbf{B}^{-1})$ *distribution, where* $\mathbf{M} = \lim_{N\to\infty} \frac{1}{N} \sum_i \mathbb{E}\left(\mathbf{U}_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*)\mathbf{U}_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*)^\intercal \mid \mathbf{L}_i\right)$ *and*

$$\mathbf{B} = \lim_{N\to\infty} \frac{1}{N} \sum_i \mathbb{E}\left(\sum_{a_1, a_2} \tilde{W}_i(a_1, a_2) \mathbf{X}_i(a_1, a_2)^\intercal \mathbf{V}_i(a_1, a_2; \boldsymbol{\alpha}^*)^{-1} \mathbf{X}_i(a_1, a_2) \mid \mathbf{L}_i \right).$$

The diagonal entries of $\frac{1}{N}\hat{\mathbf{B}}^{-1}\hat{\mathbf{M}}\hat{\mathbf{B}}^{-1}$ provide approximate standard errors for $\hat{\boldsymbol{\beta}}$, where

$$\hat{\mathbf{M}} = \frac{1}{N} \sum_i \hat{\mathbf{U}}_i \hat{\mathbf{U}}_i^\intercal,$$

$$\hat{\mathbf{U}}_i := \mathbf{U}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) = \sum_{(a_1, a_2)} \tilde{W}_i(a_1, a_2) \mathbf{X}_i(a_1, a_2)^\intercal \hat{\mathbf{V}}_i(a_1, a_2; \hat{\boldsymbol{\alpha}})^{-1} (\mathbf{Y}_i - \mathbf{X}_i(a_1, a_2)\hat{\boldsymbol{\beta}}),$$

and

$$\hat{\mathbf{B}} = \frac{1}{N} \sum_i \sum_{a_1, a_2} \tilde{W}_i(a_1, a_2) \mathbf{X}_i(a_1, a_2)^\intercal \hat{\mathbf{V}}_i(a_1, a_2; \hat{\boldsymbol{\alpha}})^{-1} \mathbf{X}_i(a_1, a_2).$$

The proof of Theorem IV.1 is given in Appendix B.1. Note that assumption (ii) and (iv), above, will be satisfied by design of the SMART, while assumption (iii) connects the observed data to the potential outcomes. Theorem IV.1 does not require the two assumptions in (4.3) to be true. These standard distributional assumptions were used only to motivate the pseudo-likelihood and set of estimating equations which led to an estimator for $\boldsymbol{\beta}$.

Given $\mathbf{V}_i(a_1, a_2)$, the estimating equation (4.8) is identical, with slight changes

in notation, to the estimating equation in Lu et al. (2016) for the parameters of the marginal mean model. Estimation of $\boldsymbol{\beta}$ in Lu et al. (2016) differs from our approach primarily in its modeling and estimation procedure for $\mathbf{V}_i(a_1, a_2) = \text{Var}(\mathbf{Y}_i(a_1, a_2) \mid \mathbf{L}_i)$. In Lu et al. (2016), the form of $\mathbf{V}_i(a_1, a_2)$ (e.g. autoregressive) is proposed by the data analyst and an estimate of $\mathbf{V}_i(a_1, a_2)$ is obtained via the method of moments. In our case the form of $\mathbf{V}_i(a_1, a_2)$ is a result of specifying $\mathbf{Z}_i(a_1, a_2)\mathbf{b}_i$ and the variance-covariance of $\boldsymbol{\epsilon}_i(a_1, a_2)$ and $\mathbf{b}_i$, while the estimate of $\mathbf{V}_i(a_1, a_2)$ is computed by maximizing a weighted pseudo-likelihood.

As in Lu et al. (2016), Theorem IV.1 implies that $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ and has an asymptotic Gaussian distribution, regardless of whether $\hat{\boldsymbol{\alpha}}$ converges to the true value of $\boldsymbol{\alpha}$ in model (4.1). This means that the random effects structure can be misspecified and the estimator $\hat{\boldsymbol{\beta}}$ will remain unbiased. However, the simulation results in Section 4.6 show that specifying a random effects structure which more closely models the true subject-to-subject variation in $\mathbf{Y}_i(a_1, a_2)$ can lead to greater efficiency in estimating $\boldsymbol{\beta}$. Before demonstrating the performance of our estimator in simulation studies, we propose a method for predicting the value of $\mathbf{b}_i$ in model (4.1) and hence predicting subject-specific trajectories for the primary outcome in a SMART.

4.4.2 Random Effects Prediction

The estimator for $\boldsymbol{\beta}$ derived above is all that is necessary for primary aim comparisons among the DTRs embedded in a SMART. Recall that a secondary motivation for using linear mixed models is the prediction of subject-specific outcome trajectories under specific DTRs. In this section we propose a method of predicting $\mathbf{b}_i$ in (4.1) using the weighted pseudo-likelihood in (4.6).

In Theorem IV.1 we do not require knowledge of $\mathbf{V}_i(a_1, a_2)$ or the distributions of $\boldsymbol{\epsilon}_i(a_1, a_2)$ and $\mathbf{b}_i$. To predict $\mathbf{b}_i$, however, we assume that the distributional assumptions in (4.3) are true in the population of potential outcomes. Specifically,

under model (4.1), assuming (4.3), $(\mathbf{Y}_i(a_1,a_2)^\intercal,\, \mathbf{b}_i^\intercal) \mid \mathbf{L}_i$ has a multivariate Gaussian distribution, which implies that

$$\mathbf{b}_i \mid \mathbf{Y}_i(a_1,a_2), \mathbf{L}_i \sim N\left(\mathbf{G}\mathbf{Z}_i(a_1,a_2)^\intercal \mathbf{V}_i(a_1,a_2)^{-1}(\mathbf{Y}_i(a_1,a_2) - \mathbf{X}_i(a_1,a_2)\boldsymbol{\beta}),\, \boldsymbol{\Sigma}_{\mathbf{b}|\mathbf{Y}(a_1,a_2)}\right), \tag{4.10}$$

where $\boldsymbol{\Sigma}_{\mathbf{b}|\mathbf{Y}(a_1,a_2)} = \mathrm{Var}\left(\mathbf{b}_i \mid \mathbf{Y}_i(a_1,a_2), \mathbf{L}_i\right)$. If all potential outcomes $\mathbf{Y}_i(a_1,a_2)$ were observed for each participant, a plug-in estimator of $\mathbb{E}\left(\mathbf{b}_i \mid \mathbf{Y}_i(a_1,a_2), \mathbf{L}_i\right)$ based on (4.10) would serve as a prediction of $\mathbf{b}_i$. Instead, motivated by (4.10), we propose the following:

$$\begin{aligned}
&\hat{\mathbf{b}}_i(\boldsymbol{\alpha},\boldsymbol{\beta}) = \\
&\arg\max_{\mathbf{b}_i} -\frac{1}{2}\sum_{a_1,a_2} \tilde{W}_i(a_1,a_2)\left(\mathbf{b}_i - \mathbf{G}\mathbf{Z}_i(a_1,a_2)^\intercal \mathbf{V}_i(a_1,a_2)^{-1}(\mathbf{Y}_i - \mathbf{X}_i(a_1,a_2)\boldsymbol{\beta})\right)^\intercal \\
&\qquad\qquad \Sigma_{b|Y}^{-1}\left(\mathbf{b}_i - \mathbf{G}\mathbf{Z}_i(a_1,a_2)^\intercal \mathbf{V}_i(a_1,a_2)^{-1}(\mathbf{Y}_i - \mathbf{X}_i(a_1,a_2)\boldsymbol{\beta})\right)
\end{aligned} \tag{4.11}$$

$$= \frac{\sum_{a_1,a_2} \tilde{W}_i(a_1,a_2)\mathbf{G}\mathbf{Z}_i(a_1,a_2)^\intercal \mathbf{V}_i(a_1,a_2);\boldsymbol{\alpha})^{-1}(\mathbf{Y}_i - \mathbf{X}_i(a_1,a_2)\boldsymbol{\beta})}{\sum_{a_1,a_2} \tilde{W}_i(a_1,a_2)}. \tag{4.12}$$

In practice, the predictions for each participant are obtained by substituting the estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ into (4.12), so that $\hat{\mathbf{b}}_i := \hat{\mathbf{b}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$. (Recall that estimates for the entries of $\mathbf{G}$ are given by some of the components of $\hat{\boldsymbol{\alpha}}$.) This predictor can be regarded as an empirical Bayes predictor for $\mathbf{b}_i$ (Skrondal and Rabe-Hesketh 2009; Carlin and Louis 2000) with weights that adjust for the probability of observing responders and slow responders under each embedded DTR.

4.5 Software implementation with integer-valued weights

Next we describe how the mixed model for longitudinal SMARTs can be implemented using standard mixed model software, such as Bates et al. (2015). This implementation is limited to analyses in which the weights $W_i^{(a_1,a_2)}(R_i)$ are integer-valued. When estimating these probability of treatment weights (Hirano, Imbens, and Ridder 2003; Brumback 2009) or when randomization probabilities are unequal across treatment options, the weights may not be integer-valued. Future work will develop software implementations for use in SMART designs beyond the special case of integer-valued weights.

Recall that $I^{(a_1,a_2)}(A_{1i}, R_i, A_{2i})$ is an indicator of whether subject i is observable under regimen (a_1, a_2). For example, in the autism study, $I^{(a_1,a_2)}(A_{1i}, R_i, A_{2i}) = \mathbb{I}[A_{1i} = a_1](R_i + (1 - R_i)\mathbb{I}[A_{2i} = a_2])$. Let $f(a_1, a_2, \mathbf{Y}_i, \mathbf{L}_i)$ be an arbitrary function of the observed response vector $\mathbf{Y}_i$, baseline covariates $\mathbf{L}_i$, and the DTR (a_1, a_2). In the autism SMART and other common designs, responders ($R_i = 1$) are observable under both of the DTRs $(A_{1i}, 1), (A_{1i}, -1)$. In this case,

$$\begin{aligned}
&\sum_{i=1}^{N} \sum_{a_1,a_2} I^{(a_1,a_2)}(A_{1i}, R_i, A_{2i}) W_i^{(a_1,a_2)}(R_i) f(a_1, a_2, \mathbf{Y}_i, \mathbf{L}_i) \\
&= \sum_{i:R_i=1} W_i^{(A_{1i},1)}(1) f(A_{1i}, 1, \mathbf{Y}_i, \mathbf{L}_i) + \sum_{i:R_i=1} W_i^{(A_{1i},-1)}(1) f(A_{1i}, -1, \mathbf{Y}_i, \mathbf{L}_i) \\
&\quad + \sum_{i:R_i=0} W_i^{(A_{1i},A_{2i})}(0) f(A_{1i}, A_{2i}, \mathbf{Y}_i, \mathbf{L}_i),
\end{aligned} \tag{4.13}$$

since when $R_i = 0$, subject i is observable under DTR (A_{1i}, A_{2i}) only.

The weighted pseudo-likelihood is

$$
\begin{aligned}
l(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= -\frac{1}{2} \sum_{i=1}^{N} \sum_{a_1, a_2} \tilde{W}_i(a_1, a_2) \log\det\left[\mathbf{V}_i(a_1, a_2)\right] \\
&\quad - \frac{1}{2} \sum_{i} \sum_{a_1, a_2} \tilde{W}_i(a_1, a_2) \mathbf{r}_i(a_1, a_2)^\intercal \mathbf{V}_i(a_1, a_2)^{-1} \mathbf{r}_i(a_1, a_2)
\end{aligned} \tag{4.14}
$$

$$
\begin{aligned}
&= -\frac{1}{2} \sum_{i:R_i=1} W_i^{(A_{1i},1)}(1) \log\det\left[\mathbf{V}_i(A_{1i}, 1)\right] \\
&\quad - \frac{1}{2} \sum_{i:R_i=1} W_i^{(A_{1i},-1)}(1) \log\det\left[\mathbf{V}_i(A_{1i}, -1)\right] \\
&\quad - \frac{1}{2} \sum_{i:R_i=0} W_i^{(A_{1i},A_{2i})}(0) \log\det\left[\mathbf{V}_i(A_{1i}, A_{2i})\right] \\
&\quad - \frac{1}{2} \sum_{i:R_i=1} W_i^{(A_{1i},1)}(1) (\mathbf{Y}_i - \mathbf{X}_i(A_{1i}, 1)\boldsymbol{\beta})^\intercal \mathbf{V}_i(A_{1i}, 1)^{-1} (\mathbf{Y}_i - \mathbf{X}_i(A_{1i}, 1)\boldsymbol{\beta}) \\
&\quad - \frac{1}{2} \sum_{i:R_i=1} W_i^{(A_{1i},-1)}(1) (\mathbf{Y}_i - \mathbf{X}_i(A_{1i}, -1)\boldsymbol{\beta})^\intercal \mathbf{V}_i(A_{1i}, -1)^{-1} (\mathbf{Y}_i - \mathbf{X}_i(A_{1i}, -1)\boldsymbol{\beta}) \\
&\quad - \frac{1}{2} \sum_{i:R_i=0} W_i^{(A_{1i},A_{2i})}(0) (\mathbf{Y}_i - \mathbf{X}_i(A_{1i}, A_{2i})\boldsymbol{\beta})^\intercal \mathbf{V}_i(A_{1i}, A_{2i})^{-1} (\mathbf{Y}_i - \mathbf{X}_i(A_{1i}, A_{2i})\boldsymbol{\beta})
\end{aligned} \tag{4.15}
$$

This objective function is equivalent to the log-likelihood in a linear mixed effects model based on an "augmented" data set constructed in the following manner. For all subjects i whose observed data are observable under more than one DTR, duplicate the baseline covariates $\mathbf{L}_i$ and response vectors $\mathbf{Y}_i$ once for each of those DTRs. In the autism SMART, subjects with $R_i = 1$ are observable under $(A_{1i}, 1)$ and $(A_{2i}, -1)$, so the baseline covariates and response vectors are duplicated twice. The design matrices $\mathbf{X}_i(a_1, a_2)$ and $\mathbf{Z}_i(a_1, a_2)$ for each replicate are formed by plugging in the values of (a_1, a_2) corresponding to the DTR under which that replicate is observable. The weights for these replicates are formed similarly. Thus, for a subject with $R_i = 1$, the

augmented data consist of

$$\left\{\mathbf{X}_i(A_{1i},1),\mathbf{Z}_i(A_{1i},1)\mathbf{L}_i,\mathbf{Y}_i,W_i^{(A_{1i},1)}(1)\right\}$$
$$\left\{\mathbf{X}_i(A_{1i},-1),\mathbf{Z}_i(A_{1i},-1),\mathbf{L}_i,\mathbf{Y}_i,W_i^{(A_{1i},-1)}(1)\right\}$$

For subjects whose observed data are observable only under the DTR (A_{1i}, A_{2i}), their observed data are unchanged and included in the augmented data set.

Indexing the artificial "subjects" in the augmented data set by $s = 1, \ldots, M$, we have, based on (4.15),

$$l(\boldsymbol{\beta},\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{s=1}^{M} W_s \log\det\left[\mathbf{V}_s\right] - \frac{1}{2}\sum_{s=1}^{M} W_s(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta})^{\intercal}\mathbf{V}_s^{-1}(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}), \quad (4.16)$$

where $W_s, \mathbf{X}_s$ and $\mathbf{V}_s$ are the values of $W_i^{(A_{1i},A_{2i})}(R_i)$, $\mathbf{X}_i(a_1,a_2)$ and $\mathbf{V}_i(a_1,a_2)$ evaluated under the DTR corresponding to replicate s in the augmented data. Thus, to find maximum likelihood estimates of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, we can use any software package which maximizes a weighted log-likelihood of the form (4.16). In particular, when W_s is an integer, we can maximize (4.16) by duplicating all of the terms indexed by s a total of W_s times and fitting the mixed model corresponding to (4.16) without the use of weights.

4.6 Simulation studies

Next we use simulation studies to evaluate the ability of our mixed effects model to estimate causal estimands of primary interest when comparing embedded DTRs in a SMART. We also compare our mixed model estimator to the GEE-like estimators discussed in Lu et al. (2016) and Seewald et al. (2019).

Data were generated from a hypothetical SMART with two treatment stages, two treatment options for all participants in stage one, and two treatment options

for slow responders in stage two, leading to four embedded DTRs with $a_1, a_2 \in \{1, -1\}$. This is a common SMART design (Naar-King et al. 2016; August, Piehler, and Bloomquist 2016, e.g.) and is different from the autism SMART in Figure 4.1, in which slow responders from only one of the stage-one treatment arms were randomized at the start of the second stage. In a given simulation replicate, potential outcomes were generated according to (B.10), below, and observed data were obtained from these potential outcomes via randomizations satisfying assumptions (ii) and (iii) in Theorem IV.1. All simulated participants were randomized with equal probability to either $A_{1i} = 1$ or $A_{1i} = -1$, and only slow responders were randomized to $A_{2i} = 1$ or $A_{2i} = -1$ with equal probability.

The potential outcomes in these simulation studies were generated from the following piecewise linear model:

$$\begin{aligned} Y_{it}(a_1, a_2) &= \theta_0 + \mathbb{I}\left[t \leq \kappa\right] t(\theta_1 + \theta_2 a_1) + \mathbb{I}\left[t > \kappa\right] \kappa(\theta_1 + \theta_2 a_1) \\ &+ \mathbb{I}\left[t > \kappa\right] (t - \kappa)(\theta_3 + \theta_4 a_1 + (\theta_5 a_2 + \theta_6 a_1 a_2)(1 - R_i(a_1))) \\ &+ \mathbb{I}\left[t > \kappa\right] (t - \kappa)(\psi^{(1)} \mathbb{I}\left[a_1 = 1\right] + \psi^{(-1)} \mathbb{I}\left[a_1 = -1\right]) \left[R_i(a_1) - \mathbb{P}\left(R_i(a_1) = 1 \mid L_i\right)\right] \\ &+ \theta_7 L_i + \gamma_{0i} + \gamma_{1i} t + \epsilon_{it}, \end{aligned} \tag{4.17}$$

where $R_i(a_1) = \mathbb{I}\left[Y_{i\kappa}(a_1) - \theta_7 L_i > c\right]$; $c = 1.1$; $(\gamma_{0i}, \gamma_{1i})^\intercal \mid L_i \sim N(0, \Gamma)$; $\epsilon_{it} \mid \mathbf{L}_i \sim N(0, \tau^2)$ with $\tau^2 = 1$; $t \in \{0, 0.5, 1.5, 2, 2.25, 2.5, 3\}$; and $\kappa = 2$.

The binary tailoring variable $R_i(a_1)$ is a function of the potential outcome at the end of the first treatment stage, and the fixed value of c means that $\mathbb{P}\left(R_i(a_1) = 1 \mid L_i\right)$ varies as a function of a_1. The parameters $\psi^{(1)}$ and $\psi^{(-1)}$ induce a marginal association between $R_i(a_1)$ and second-stage outcomes. The random intercepts and slopes, γ_{0i} and γ_{1i}, induce within-person correlation, and the residual errors ϵ_{it} were generated independently across i and t. The scalar random variable $L_i = \mathbf{L}_i$ is a binary baseline

covariate, and the knot κ represents the time when the first treatment stage ends. In all simulations, half of the participants were assigned $L_i = 1$ and half were assigned $L_i = -1$. Under (B.10), the marginal mean can be expressed as follows:

$$\begin{aligned}
&\mathbb{E}\left(Y_{it}(a_1, a_2) \mid \mathbf{L}_i\right) = \boldsymbol{\beta}^\intercal \mathbf{X}_{it}(a_1, a_2) \\
= \;& \beta_0 + \mathbb{I}\left[t \leq \kappa\right] t(\beta_1 + \beta_2 a_1) + \mathbb{I}\left[t > \kappa\right] \kappa(\beta_1 + \beta_2 a_1) \\
&+ \mathbb{I}\left[t > \kappa\right](t - \kappa)(\beta_3 + \beta_4 a_1 + \beta_5 a_2 + \beta_6 a_1 a_2) \\
&+ \beta_7 L_i,
\end{aligned} \tag{4.18}$$

where $\beta_j = \theta_j$ for $j \in \{0, 1, 2, 3, 4, 7\}$,

$$\begin{aligned}
\beta_5 &= \left\{\theta_5 \left(\frac{\pi^{(1)}}{2} + \frac{\pi^{(-1)}}{2}\right) + \theta_6 \left(\frac{\pi^{(1)}}{2} - \frac{\pi^{(-1)}}{2}\right)\right\}, \\
\beta_6 &= \left\{\theta_5 \left(\frac{\pi^{(1)}}{2} - \frac{\pi^{(-1)}}{2}\right) + \theta_6 \left(\frac{\pi^{(1)}}{2} + \frac{\pi^{(-1)}}{2}\right)\right\},
\end{aligned}$$

and $\pi^{(a_1)} := \mathbb{P}\left(R_i(a_1) = 0 \mid \mathbf{L}_i\right)$. Causal estimands of primary interest are expressed as functions of $\beta_2, \beta_4, \beta_5$, and β_6. Further details of this generative model are given in Appendix B.

Sections 4.6.1 and 4.6.2 present the results of two simulation studies which differ in whether or not they misspecify the marginal variance and distribution of $\mathbf{Y}_i(a_1, a_2)$. Section 4.6.3 presents a simulation study with ignorable missing data due to study dropout. In all simulation studies, the linear model for $\mathbb{E}\left(\mathbf{Y}_i(a_1, a_2) \mid \mathbf{L}_i\right)$ is correctly specified. We report estimation performance for the end-of-study contrast $\mathbb{E}\left(Y_{i3}(1, -1) \mid \mathbf{L}_i\right) - \mathbb{E}\left(Y_{i3}(-1, -1) \mid \mathbf{L}_i\right) = 2\kappa\beta_2 + 2(3 - \kappa)\beta_4 - 2(3 - \kappa)\beta_6$, and we chose simulation parameters so that this contrast had the largest magnitude among any pairwise contrast between embedded DTRs. Parameter values for the marginal

mean were chosen to achieve desired values of the standardized effect size

$$d = \frac{\mathbb{E}\left(Y_{i3}(1,-1) \mid \mathbf{L}_i\right) - \mathbb{E}\left(Y_{i3}(-1,-1) \mid \mathbf{L}_i\right)}{\sqrt{\frac{1}{2}\text{Var}\left(Y_{i3}(1,-1) \mid \mathbf{L}_i\right) + \frac{1}{2}\text{Var}\left(Y_{i3}(-1,-1) \mid \mathbf{L}_i\right)}}. \tag{4.19}$$

4.6.1 Simulation 1

The first simulation study verifies that our estimator $\hat{\boldsymbol{\beta}}$ is unbiased in large samples and that large-sample confidence interval coverage is attained with the standard errors based on Theorem IV.1. This is accomplished in the ideal setting in which the probability distribution of $\mathbf{Y}_i(a_1, a_2)|\mathbf{L}_i$ can be correctly specified using our proposed mixed model. In general, $\mathbf{Y}_i(a_1, a_2)|\mathbf{L}_i$ in (B.10) follows a Gaussian mixture distribution with mixing probability $\mathbb{P}\left(R_i(a_1) = 0 \mid \mathbf{L}_i\right)$. However, in this simulation study we choose $0 = \theta_5 = \theta_6 = \psi^{(1)} = \psi^{(-1)}$, so that $Y_{it}(a_1, a_2) - \mathbb{E}\left(Y_{it}(a_1, a_2) \mid \mathbf{L}_i\right) = \gamma_{0i} + \gamma_{1i}t + \epsilon_{it}$ and the distribution of $\mathbf{Y}_i(a_1, a_2)|\mathbf{L}_i$ is the same as the marginal distribution specified in the following mixed model:

$$Y_{it}(a_1, a_2) \mid b_{0i}, b_{1i}, \mathbf{L}_i \sim N(\boldsymbol{\beta}^\intercal \mathbf{X}_{it}(a_1, a_2) + b_{0i} + b_{1i}t, \sigma^2), \tag{4.20}$$

$$(b_{0i}, b_{1i})^\intercal \mid \mathbf{L}_i \sim N(\mathbf{0}, \mathbf{G}) \tag{4.21}$$

where $\boldsymbol{\beta}^\intercal \mathbf{X}_{it}(a_1, a_2)$ is the linear parametrization of the mean in equation (4.18). We compared this "slopes and intercepts" mixed model, in which the joint distribution of $\mathbf{Y}_i(a_1, a_2) \mid \mathbf{L}_i$ is correctly specified, to an "intercepts only" mixed model,

$$Y_{it}(a_1, a_2) \mid b_{0i}, \mathbf{L}_i \sim N(\boldsymbol{\beta}^\intercal \mathbf{X}_{it}(a_1, a_2) + b_{0i}, \sigma^2), \tag{4.22}$$

$$b_{0i} \mid \mathbf{L}_i \sim N(0, \text{Var}\left(b_{0i}\right)), \tag{4.23}$$

in which $\text{Var}(\mathbf{Y}_i(a_1, a_2) \mid \mathbf{L}_i)$ is misspecified. We use different notation for the random effects and variance parameters in (4.21)–(4.23) than we do in (B.10) to distinguish

models used for estimation from the true, data-generating probability distribution. In this simulation study we set $\text{Var}(\gamma_0) = 0.8, \text{Var}(\gamma_1) = 1$ and $\text{Cov}(\gamma_0, \gamma_1) = -0.2$.

Table 4.2 contains the bias and standard deviation of the point estimates, the mean of the approximate standard errors, the coverage probability for a 95-percent confidence interval, and the root mean squared error (RMSE) computed from 1,000 simulation replicates. In large samples, the bias is approximately two orders of magnitude smaller than the standard deviation of the point estimates, confirming that the mixed model estimator is unbiased for the linear mean parameters in (4.18). The standard errors based on Theorem IV.1 provide confidence interval coverage close to the nominal level in large samples. In addition, note that the intercepts only mixed model, which misspecifies $\text{Var}(\mathbf{Y}_i(a_1, a_2) \mid \mathbf{L}_i)$, does not introduce bias in large samples. Instead, the estimator is slightly less efficient than the slopes and intercepts model, in which both the mean and covariance of $\mathbf{Y}_i(a_1, a_2)$ are correctly specified.

4.6.2 Simulation 2

In this second simulation, we investigate whether the estimator $\hat{\boldsymbol{\beta}}$ is unbiased in large samples, and whether this estimator can provide efficiency gains relative to existing estimators, in a more realistic scenario in which it is not possible to correctly specify the distribution of $\mathbf{Y}_i(a_1, a_2) \,|\, \mathbf{L}_i$ using model (4.1). Data were again generated from model (B.10), but the coefficients $\theta_5, \theta_6, \psi^{(1)}$ and $\psi^{(-1)}$ were nonzero and therefore $\mathbf{Y}_i(a_1, a_2) \,|\, \mathbf{L}_i$ had a marginal Gaussian mixture distribution.

In addition to the mixed models (4.21)–(4.23), we also report estimation performance of the GEE-like estimator of Lu et al. (2016) and Seewald et al. (2019) in which only the marginal mean is assumed to be correctly specified and no further distributional assumptions are made about $\mathbf{Y}_i(a_1, a_2) \mid \mathbf{L}_i$. With these GEE estimators, a working model for $\mathbf{V}_i(a_1, a_2)$ (e.g. exchangeable) is specified directly and the method of moments is used to estimate the parameters in this working model.

These GEE estimators were implemented as follows. First, an initial least squares estimate is computed:

$$\hat{\boldsymbol{\beta}}^{(0)} = \left(\sum_i \sum_{a_1,a_2} \tilde{W}_i(a_1,a_2) \mathbf{X}_i(a_1,a_2)^\intercal \mathbf{X}_i(a_1,a_2) \right)^{-1} \left(\sum_i \sum_{a_1,a_2} \tilde{W}_i(a_1,a_2) \mathbf{X}_i(a_1,a_2)^\intercal \mathbf{Y}_i \right).$$

This initial estimate is used to compute the residual vectors

$$\mathbf{r}_i^{(0)}(a_1,a_2) = \mathbf{Y}_i - \mathbf{X}_i(a_1,a_2)\hat{\boldsymbol{\beta}}^{(0)}, \tag{4.24}$$

for all i and (a_1,a_2). Next we compute method of moments estimators for $\mathbf{V}_i(a_1,a_2)$. Let D be the number of embedded DTRs, i.e. $D = \sum_{a_1,a_2} 1$. For $t \neq s$, define the variance estimators

$$\hat{\sigma}_t^2(a_1,a_2) = \frac{\sum_i \tilde{W}_i(a_1,a_2) r_{it}^{(0)}(a_1,a_2)^2}{N_t}, \tag{4.25}$$

$$\hat{\sigma}^2(a_1,a_2) = \frac{N_t \hat{\sigma}_t^2(a_1,a_2)}{\sum_t N_t}, \tag{4.26}$$

$$\hat{\sigma}_t^2 = \frac{1}{D} \sum_{a_1,a_2} \hat{\sigma}_t^2(a_1,a_2) \tag{4.27}$$

and

$$\hat{\sigma}^2 = \frac{1}{D} \sum_{a_1,a_2} \hat{\sigma}^2(a_1,a_2), \tag{4.28}$$

and define the correlation estimators

$$\hat{\rho}_{ts}(a_1, a_2) = \frac{1}{N_{ts}} \sum_i \tilde{W}_i(a_1, a_2) \left[\frac{r_{it}^{(0)}(a_1, a_2) r_{is}^{(0)}(a_1, a_2)}{\hat{\sigma}_t(a_1, a_2)\hat{\sigma}_s(a_1, a_2)} \right], \tag{4.29}$$

$$\hat{\rho}_{ts} = \frac{1}{D} \sum_{a_1, a_2} \hat{\rho}_{ts}(a_1, a_2), \tag{4.30}$$

$$\hat{\rho}(a_1, a_2) = \frac{1}{N} \sum_i \frac{\tilde{W}_i(a_1, a_2)}{n_i(n_i - 1)/2} \sum_{s<t} \frac{r_{is}^{(0)}(a_1, a_2) r_{it}^{(0)}(a_1, a_2)}{\hat{\sigma}_s(a_1, a_2)\hat{\sigma}_t(a_1, a_2)}, \tag{4.31}$$

$$\hat{\rho} = \frac{1}{D} \sum_{a_1, a_2} \hat{\rho}(a_1, a_2), \tag{4.32}$$

$$\hat{\psi}(a_1, a_2) = \frac{1}{N} \sum_i \frac{\tilde{W}_i(a_1, a_2)}{n_i(n_i - 1)/2} \sum_{s<t} \frac{r_{is}^{(0)}(a_1, a_2) r_{it}^{(0)}(a_1, a_2)}{\hat{\sigma}^2(a_1, a_2)}, \tag{4.33}$$

and

$$\hat{\psi} = \frac{1}{D} \sum_{a_1, a_2} \hat{\psi}(a_1, a_2), \tag{4.34}$$

where N_t is the number of individuals with an observation at unique time point t and N_{ts} is the number of individuals with observations at both of the time points t and s. The estimators defined above are simply the method of moments estimators for correlation or variance parameters at each observation. They differ in whether the variances or correlations are assumed to be equal across DTRs and in whether the variance is assumed to be constant as a function of time. By combining these correlation and variance estimators, we can obtain various working models for $\mathbf{V}_i(a_1, a_2)$. For example, the unstructured and exchangeable estimates of $\mathbf{V}_i(a_1, a_2)$ have the following entries, $v_{ts}(a_1, a_2)$:

Unstructured	$v_{tt}(a_1, a_2) = \hat{\sigma}_t^2$	Exchangeable	$v_{tt}(a_1, a_2) = \hat{\sigma}^2$ for all t
	$v_{ts}(a_1, a_2) = \hat{\sigma}_s\hat{\sigma}_t\hat{\rho}_{ts}$		$v_{ts}(a_1, a_2) = \hat{\psi}\hat{\sigma}^2$
	for all a_1, a_2		for all a_1, a_2

The independence working model sets all off-diagonal entries of $\mathbf{V}_i(a_1, a_2)$ to zero and all diagonal entries to $\hat{\sigma}^2$. Autoregressive working models for $\mathbf{V}_i(a_1, a_2)$ are also possible using the correlation estimators

$$\hat{\tau}_t(a_1, a_2) = \frac{1}{N} \sum_i \frac{\tilde{W}_i(a_1, a_2)}{n_i - 1} \sum_{s=1}^{n_i - 1} \frac{r_{is}^{(0)}(a_1, a_2) r_{i(s+1)}^{(0)}(a_1, a_2)}{\hat{\sigma}_s(a_1, a_2) \hat{\sigma}_{s+1}(a_1, a_2)}, \tag{4.35}$$

$$\hat{\tau}_t = \frac{1}{D} \sum_{a_1, a_2} \hat{\tau}_t(a_1, a_2), \tag{4.36}$$

$$\hat{\tau}(a_1, a_2) = \frac{1}{N} \sum_i \frac{\tilde{W}_i(a_1, a_2)}{n_i - 1} \sum_{s=1}^{n_i - 1} \frac{r_{is}^{(0)}(a_1, a_2) r_{i(s+1)}^{(0)}(a_1, a_2)}{\hat{\sigma}^2(a_1, a_2)}, \tag{4.37}$$

and

$$\hat{\tau} = \frac{1}{D} \sum_{a_1, a_2} \hat{\tau}(a_1, a_2) \tag{4.38}$$

In this simulation study, all of the models used for estimation correctly specify the linear model $\boldsymbol{\beta}^\intercal \mathbf{X}_{it}(a_1, a_2)$, but none of them correctly specify the marginal covariance or the distribution of $\mathbf{Y}_i(a_1, a_2) \mid \mathbf{L}_i$.

Table 4.3 compares the two mixed models and the GEE estimator with exchangeable, unstructured, and independence working models for $\mathbf{V}_i(a_1, a_2)$ in their ability to estimate the end-of-study contrast with standardized effect size $d \approx 0.5$. The magnitude of the bias relative to the standard deviation in Table 4.3 indicates that all of these estimators are unbiased in large samples.

While none of the estimation models in this simulation study correctly specify $\mathbf{V}_i(a_1, a_2)$, we can see in Table 4.3 that efficiency (measured by RMSE) in estimating the end-of-study contrast is improved by a working model for $\mathbf{V}_i(a_1, a_2)$ which more closely resembles the true marginal covariance. Here we report RMSE as a fraction of the smallest RMSE for a fixed sample size. To measure the performance of each working model for $\mathbf{V}_i(a_1, a_2)$, we report $\frac{\left\| \mathbf{V}_{\text{true}} - \mathbb{E}(\hat{\mathbf{V}}) \right\|}{\left\| \mathbf{V}_{\text{true}} \right\|}$, the relative error in

the Frobenius norm between $\mathbf{V}_{\text{true}} := \frac{1}{D}\sum_{a_1,a_2} V_i(a_1,a_2)$, the true average covariance matrix of $\mathbf{Y}_i(a_1,a_2)$ according to the generative model, and the simulation-based estimate of $\mathbb{E}\left(\hat{\mathbf{V}}\right) := \frac{1}{D}\sum_{a_1,a_2}\mathbb{E}\left(\hat{\mathbf{V}}_i(a_1,a_2;\hat{\boldsymbol{\alpha}})\right)$, the large-sample average covariance matrix implied by the estimation model. The slopes and intercepts mixed model had both the lowest relative error in estimating $\mathbf{V}_{\text{true}}$ and the lowest RMSE for each fixed sample size. For these estimators, RMSE decreases as the working model for $\mathbf{V}_i(a_1,a_2)$ improves. This simulation study suggests that if the separate specification of between-person and within-person variation in a mixed effects model leads to improved modeling of the marginal covariance, we can expect efficiency gains over GEE-like approaches when comparing embedded DTRs.

4.6.3 Simulation 3

One potential benefit of mixed models is their ability to provide unbiased parameter estimates when data are missing at random, assuming that estimation and inference are based on a correctly specified likelihood for the observed data (Fitzmaurice, Laird, and Ware, 2011, Ch. 17; Hedeker and Gibbons, 2006, Ch. 14; Molenberghs and Kenward, 2007, Ch. 7; Gibbons, Hedeker, and DuToit, 2010). In the case of the mixed model we propose for longitudinal SMARTs, estimation and inference are based on a weighted pseudo-likelihood, not the true likelihood for the observed data, so it is not clear whether bias can be avoided with ignorable missing data in a SMART.

To help understand whether mixed models provide any protection against bias due to missing data in a longitudinal SMART, this additional simulation study describes the performance of our mixed model and the GEE-like estimators of Seewald et al. (2019) and Lu et al. (2016) when data are missing at random (ignorable) due to study dropout. In this scenario, if a participant's observed Y_{it} at $t = 2.25$ was less than -3.5, then all observations from that participant at time

points $t \geq 2.5$ were discarded. This results in about 20 percent dropout among participants with $(A_{1i}, A_{2i}, R_i) = (-1, -1, 0)$; about 17 percent dropout among participants with $(A_{1i}, A_{2i}, R_i) = (-1, 1, 0)$; about nine and ten percent dropout among participants with $(A_{1i}, A_{2i}, R_i) = (1, 1, 0)$ and $(1, -1, 0)$, respectively; and less than 0.1 percent dropout among participants with $R_i = 1$.

Table 4.4 compares the same estimators from Section 4.6.2 their ability to estimate an end-of-study contrast with effect size $d \approx 0.5$ in the presence of the dropout process described above. In this scenario, we see bias in large samples, although the degree of bias decreases as $\frac{\|\mathbf{V}_{\text{true}} - \mathbb{E}(\hat{\mathbf{V}})\|}{\|\mathbf{V}_{\text{true}}\|}$ decreases. This suggests that the ability of a mixed model to flexibly model $\mathbf{V}_i(a_1, a_2)$, and to efficiently estimate $\mathbf{V}_i(a_1, a_2)$, may provide some protection against bias due to ignorable missing data. Since we are not able to fit a mixed model using the true likelihood for the potential outcomes, the purported benefits of mixed models in the presence of ignorable missing data (compared to GEE approaches) might not exist when analyzing longitudinal SMARTs.

4.7 Application

Finally, we demonstrate our mixed model using the autism SMART of Kasari et al. (2014). Our goal here is to compare the three embedded DTRs based on changes in communication outcomes for the children receiving each DTR. Figure 4.2 displays the measured primary outcome, the number of socially communicative utterances, for each participant in this study at baseline and at weeks 12, 24, and 36. For the marginal mean, we specified the piecewise linear model (4.2), and we specified random intercepts as the random effects structure.

The parameter vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_6)$ was estimated as described in Section 4.4.1 using widely available software for linear mixed models (Bates et al. 2015) applied to the restructured version of the observed data described in Section 4.5. The estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ obtained in this manner were then used to compute estimated standard

errors as described in Section 4.4.1. Table 4.5 displays the estimated coefficients in this model with 95-percent confidence intervals, and Figure 4.3 displays the estimated marginal mean for each DTR at each time point.

To understand whether we have evidence that communication outcomes differ among children receiving each of these DTRs, we performed an "omnibus" test of whether the three DTRs differ at all. We tested the hypothesis that the area under the curve (AUC) for the marginal mean is the same across all three DTRs, which, in this case, is equivalent to testing $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ for a constant matrix $\mathbf{C}$. Based on Theorem IV.1, under H_0, the statistic $(\mathbf{C}\hat{\boldsymbol{\beta}})^\intercal \left(\mathbf{C}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\mathbf{C}^\intercal\right)^{-1} \mathbf{C}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, has a χ^2 distribution with two degrees of freedom in large samples. This test statistic was equal to 10.32 with a p-value of 0.006, suggesting differences in the AUCs among the three DTRs. Following this omnibus test, we examined pairwise contrasts between each DTR at each time point, given in Figure 4.4, which suggest that the DTR which starts with the AAC speech device is superior to the other two DTRs, at least during the first 12 weeks.

For demonstration, Figure 4.5 displays predicted person-specific trajectories,

$$\hat{\boldsymbol{\beta}}^\intercal \mathbf{X}_{it}(a_1, a_2) + \hat{\mathbf{b}}_i,$$

using the intercepts-only mixed model, along with the observed outcomes and the estimated mean outcome under each DTR. This display could be used to assess subject-to-subject variation relative to the estimated mean under each DTR or to identify individuals with outlying trajectories based on the fitted model. In this example, random intercepts lead to subject-specific trajectories which are parallel to the estimated mean under each DTR. The potential high outliers under the DTRs (JASP, AAC) and (AAC, AAC+) could be investigated to help characterize the variation in communication outcomes for these study participants.

4.8 Discussion

In Section 4.4.2 we proposed a method for predicting random effects based on a weighted pseudo-likelihood. The prediction method we propose is analogous to the "best linear unbiased predictors" commonly used in standard mixed effects analysis of longitudinal data (Robinson, 1991; Verbeke and Molenberghs, 2009, Section 7.4). However, our proposed predictor $\hat{\mathbf{b}}_i$ is a nonlinear function of (A_{1i}, R_i, A_{2i}) across all i, and it is unclear whether $\hat{\mathbf{b}}_i$ has minimum mean squared error (MSE) marginally over these random variables. Further work is needed to derive a minimum-MSE property for $\hat{\mathbf{b}}_i$ which is marginal over (A_{1i}, R_i, A_{2i}) and uses the same statistical and causal assumptions of Theorem IV.1.

The software implementation used for the analysis in Section 4.7 is limited to cases where the weights $W_i^{(a_1,a_2)}(R_i)$ are integers. Additional work is needed for a general implementation of the weighted pseudo-likelihood in cases where the weights are not integers, which may occur when the randomization probabilities are unequal across treatment options, or when the weights are estimated (e.g. Williamson, Forbes, and White 2014; Hirano, Imbens, and Ridder 2003).

Although we focused on SMARTs with a longitudinal outcome, the mixed model developed in this chapter could, in principle, be used to estimate the end-of-study marginal mean for cluster-level DTRs, as in NeCamp, Kilbourne, and Almirall (2017), by modifying the marginal mean model $\boldsymbol{\beta}^\intercal \mathbf{X}_i(a_1, a_2)$ to no longer be a function of time. An exchangeable correlation structure for clustered end-of-study outcomes could be modeled with a random intercept for each cluster. Another direction for future research is to develop a generalized linear mixed model for SMARTs with a longitudinal binary outcome.

This chapter focused on marginal mean models for the embedded DTRs that are conditional only on baseline covariates. This is analogous to primary aim analyses in standard randomized trials. An alternative approach would be to specify a mixed

model conditional on both the baseline covariates and the embedded tailoring variables. For example, in the autism SMART, one could propose a mixed effects model for $\mathbf{Y}_i$ conditional on A_{1i}, R_i, A_{2i}. Future work will investigate how to obtain consistent estimators for marginal estimands using this kind of conditional modeling of the observed longitudinal outcome.

In addition to the reasons given in Section 4.1, scientists might prefer mixed effects models because they may require less restrictive assumptions about missing data, at least when the true probability distribution for the observed data is correctly specified (Hedeker and Gibbons, 2006, Ch. 14; Fitzmaurice, Laird, and Ware, 2011, Ch. 17) Our marginal modeling and weighted, pseudo-likelihood estimation approach does not require a correct specification of the true probability distribution that generated the observed data. The simulation results in Section 4.6.3 suggest that our mixed model may offer some protection against bias in the presence of ignorable missing data, but additional work is needed to understand whether our marginal model for longitudinal SMARTs enjoys the purported benefits of standard mixed models in the presence of missing data.

DTR label (a_1, a_2)	First-stage treatment	Status at end of first-stage	Second-stage treatment	Cell in Figure 4.1	Known IPW
(JASP,JASP+) $(1, 1)$	JASP	Responder	Continue JASP	A	2
		Slow Responder	Intensify JASP	B	4
(JASP, AAC) $(1, -1)$	JASP	Responder	Continue JASP	A	2
		Slow Responder	Augment JASP+AAC	C	4
(AAC, AAC+) $(-1, \cdot)$	JASP+AAC	Responder	Continue JASP+AAC	D	2
		Slow Responder	Intensify JASP+AAC	E	2

Table 4.1: Embedded DTRs in the autism SMART. The last column provides the known inverse probability weight for subjects in each of the cells A–E in Figure 4.1.

Method	d	True value	N	Bias	Monte Carlo SD	SE Estimate	CI Coverage	RMSE
LMM slopes and intercepts	0.2	0.600	50	-0.102	1.222	1.119	0.911	1.226
			200	-0.018	0.659	0.629	0.931	0.659
			1000	0.010	0.296	0.290	0.945	0.296
			5000	-0.002	0.132	0.130	0.945	0.132
	0.8	2.480	50	0.071	1.228	1.117	0.905	1.229
			200	-0.002	0.661	0.623	0.932	0.661
			1000	-0.008	0.285	0.286	0.950	0.285
			5000	-0.007	0.129	0.128	0.948	0.129
LMM intercepts	0.2	0.600	50	-0.018	1.338	1.196	0.886	1.338
			200	0.001	0.748	0.694	0.917	0.748
			1000	0.010	0.336	0.323	0.938	0.336
			5000	0.002	0.146	0.145	0.951	0.146
	0.8	2.480	50	0.148	1.339	1.191	0.888	1.346
			200	0.011	0.728	0.682	0.922	0.727
			1000	-0.006	0.312	0.317	0.957	0.311
			5000	-0.005	0.143	0.142	0.957	0.143

Table 4.2: Mixed model estimation performance in Simulation 1. Reported for an end-of-study contrast with two mixed model specifications when the population of potential outcomes exactly follows the marginal distribution implied by the slopes and intercepts mixed model. The intercepts only model specifies the correct mean model but is otherwise misspecified. Values compuuted from 1,000 simulation replicates. The nominal confidence level was 95 percent.

N	Method	Bias	Monte Carlo SD	SE Estimate	CI Coverage	RMSE Inflation	$\frac{\lVert \mathbf{V}_{\text{true}} - \mathbb{E}(\hat{\mathbf{V}}) \rVert}{\lVert \mathbf{V}_{\text{true}} \rVert}$
50	LMM slopes and intercepts	0.013	1.655	1.505	0.910	1.000	0.051
	GEE Unstructured	0.064	1.761	1.452	0.854	1.064	0.107
	LMM intercepts only	0.115	1.922	1.656	0.871	1.163	0.640
	GEE Exchangeable	0.114	1.924	1.651	0.870	1.164	0.643
	GEE Independence	0.182	2.163	1.804	0.861	1.311	0.900
200	LMM slopes and intercepts	-0.041	0.842	0.839	0.938	1.000	0.015
	GEE Unstructured	-0.019	0.878	0.822	0.925	1.042	0.028
	LMM intercepts only	0.002	0.983	0.951	0.930	1.167	0.638
	GEE Exchangeable	0.002	0.984	0.950	0.931	1.167	0.638
	GEE Independence	0.016	1.099	1.055	0.936	1.305	0.900
1000	LMM slopes and intercepts	-0.006	0.396	0.385	0.948	1.000	0.009
	GEE Unstructured	-0.003	0.410	0.384	0.933	1.034	0.011
	LMM intercepts only	0.011	0.442	0.439	0.950	1.115	0.637
	GEE Exchangeable	0.011	0.442	0.439	0.949	1.115	0.637
	GEE Independence	0.021	0.483	0.489	0.950	1.220	0.900
5000	LMM slopes and intercepts	0.000	0.172	0.174	0.958	1.000	0.007
	GEE Unstructured	-0.001	0.178	0.174	0.947	1.039	0.007
	LMM intercepts only	0.005	0.204	0.198	0.936	1.191	0.637
	GEE Exchangeable	0.005	0.204	0.198	0.936	1.191	0.637
	GEE Independence	0.005	0.227	0.221	0.947	1.323	0.900

Table 4.3: Mixed model and GEE estimation performance in Simulation 2. Reported for an end-of-study contrast with true value 2.1197 and standardized effect size $d \approx 0.5$. Values computed from 1,000 simulation replicates. The nominal confidence level was 95 percent. RMSE inflation is the ratio of the RMSE to the smallest RMSE among the five methods for a fixed sample size.

N	Method	Bias	Monte Carlo SD	SE Estimate	CI Coverage	RMSE Inflation	$\frac{\|\mathbf{V}_{\text{true}} - \mathbb{E}(\hat{\mathbf{V}})\|}{\|\mathbf{V}_{\text{true}}\|}$
200	LMM slopes and intercepts	-0.066	0.887	0.836	0.929	1.000	0.046
	GEE Unstructured	-0.127	0.900	0.789	0.906	1.023	0.257
	LMM intercepts only	-0.401	0.950	0.891	0.900	1.161	0.660
	GEE Exchangeable	-0.402	0.951	0.890	0.900	1.161	0.663
	GEE Independence	-0.559	1.047	0.983	0.881	1.335	0.904
1000	LMM slopes and intercepts	-0.103	0.389	0.381	0.939	1.000	0.044
	GEE Unstructured	-0.158	0.396	0.366	0.902	1.059	0.246
	LMM intercepts only	-0.445	0.419	0.408	0.796	1.519	0.659
	GEE Exchangeable	-0.445	0.419	0.408	0.795	1.519	0.661
	GEE Independence	-0.612	0.464	0.451	0.699	1.907	0.904
5000	LMM slopes and intercepts	-0.080	0.177	0.171	0.921	1.000	0.042
	GEE Unstructured	-0.138	0.178	0.165	0.848	1.157	0.243
	LMM intercepts only	-0.437	0.187	0.184	0.349	2.448	0.658
	GEE Exchangeable	-0.437	0.187	0.184	0.350	2.448	0.660
	GEE Independence	-0.614	0.204	0.203	0.134	3.332	0.904

Table 4.4: Mixed model and GEE estimation performance in Simulation 3. Reported for an end-of-study contrast with true value 2.1197, corresponding to a standardized effect size of $d \approx 0.5$. Estimated in the presence of missing data due to simulated study dropout. Values computed from 1,000 simulation replicates. The nominal confidence level was 95 percent.

Coefficient	Estimate	SE	95% CI
β_0	28.885	3.763	$(21.509, 36.261)$
β_1	1.501	0.315	$(0.885, 2.118)$
β_2	-0.929	0.287	$(-1.492, -0.367)$
β_3	0.112	0.174	$(-0.229, 0.452)$
β_4	0.23	0.174	$(-0.111, 0.571)$
β_5	-0.111	0.137	$(-0.38, 0.158)$
β_6	-4.514	2.777	$(-9.957, 0.93)$

Table 4.5: Coefficient estimates for the autism SMART mixed model. Based on the random intercepts mixed model.

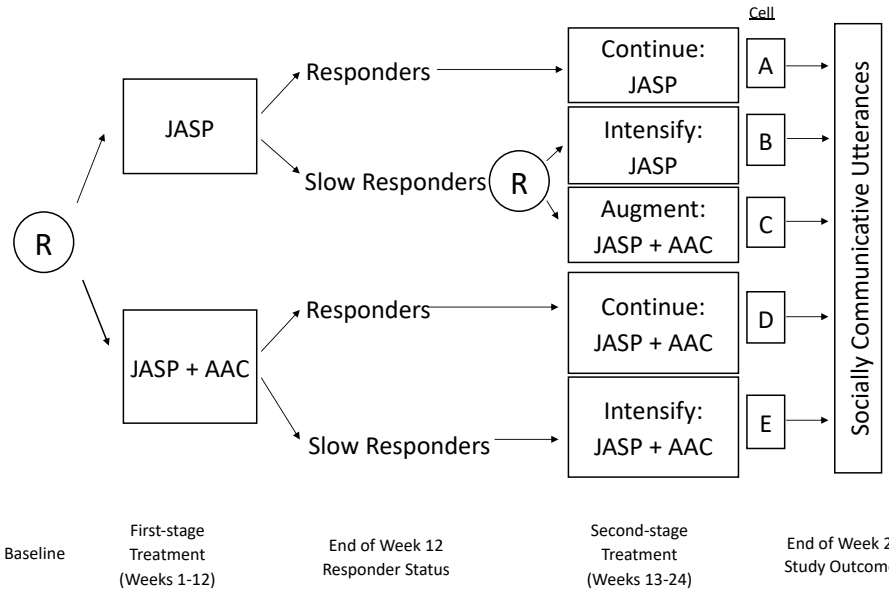


Figure 4.1: Schematic of an example SMART for children with ASD who are minimally verbal. JASP stands for joint attention social play intervention; AAC stands for alternative and augmentative communication. The encircled R signifies randomization; randomizations occurred at baseline and at the end of week 12 following identification of responder status. A child was considered a responder if there is a 25% or greater improvement on 7 or more (out of 14) language measures; otherwise, the child was labeled a slow responder.

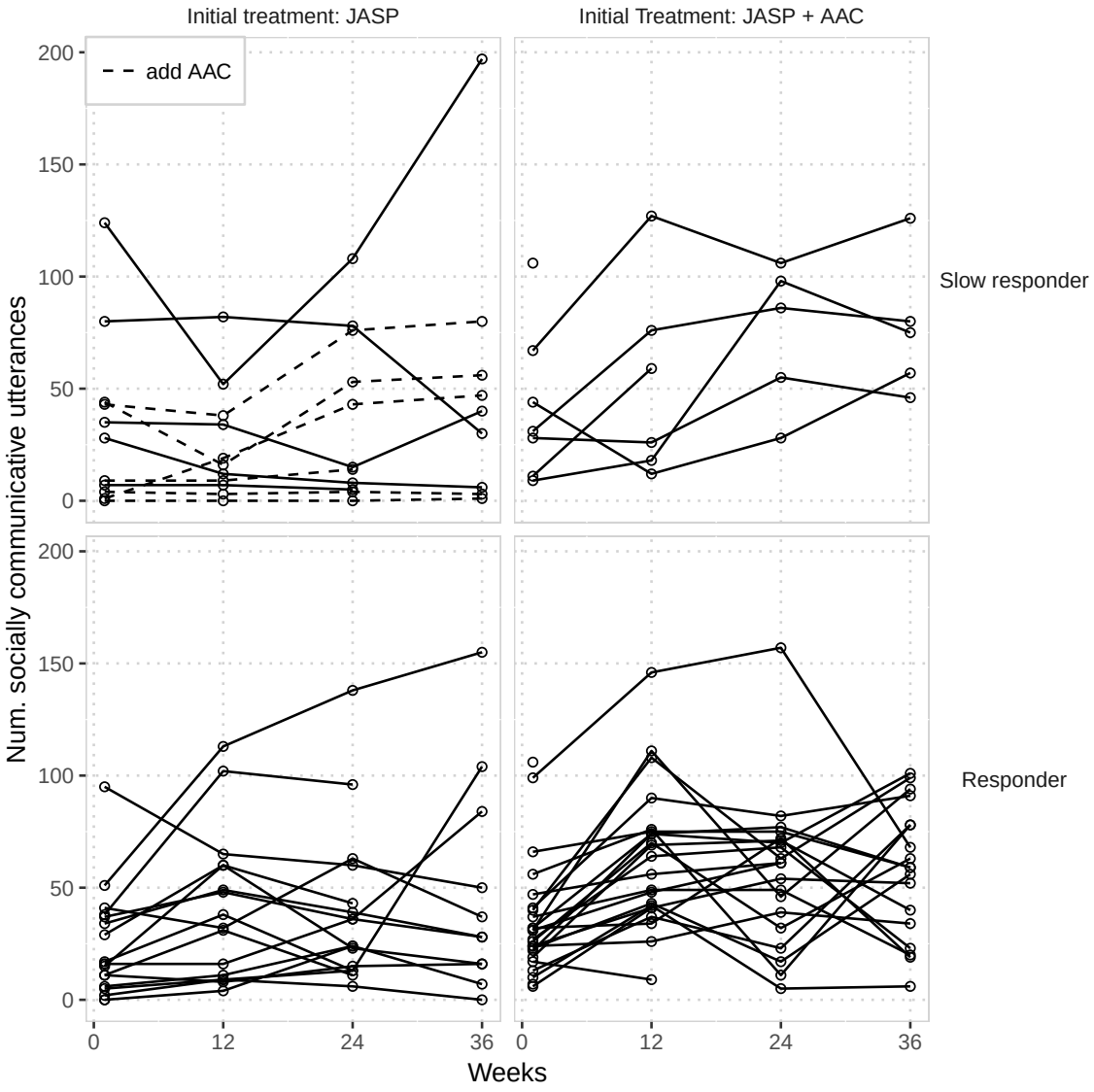


Figure 4.2: Observed number of socially communicative utterances in the autism SMART. There were $N = 61$ children in this SMART. Responders to either first-stage treatment continued that treatment. Dashed lines in the upper-left panel correspond to slow responders to initial JASP who were randomly assigned to receive JASP+AAC in the second stage. All other slow responders received an intensified version of the initial treatment.

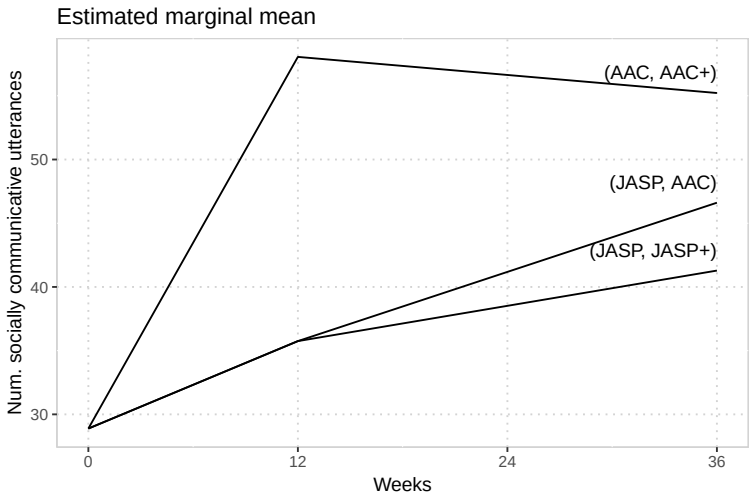


Figure 4.3: Estimated marginal mean under each DTR in the autism SMART. Estimated for the population of children at age 6.3 (the average age of participants in the study).

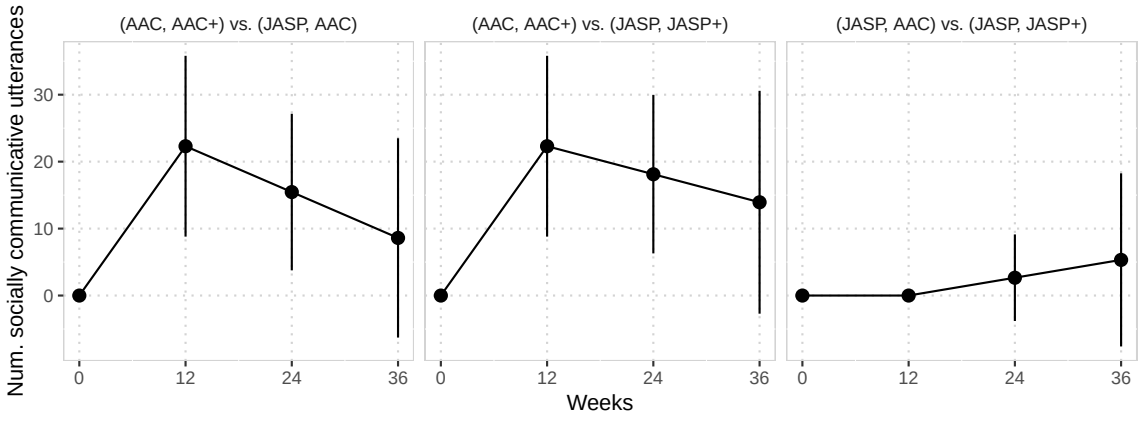


Figure 4.4: Pairwise DTR comparisons in the autism SMART. Vertical bars are 95-percent pointwise confidence intervals.

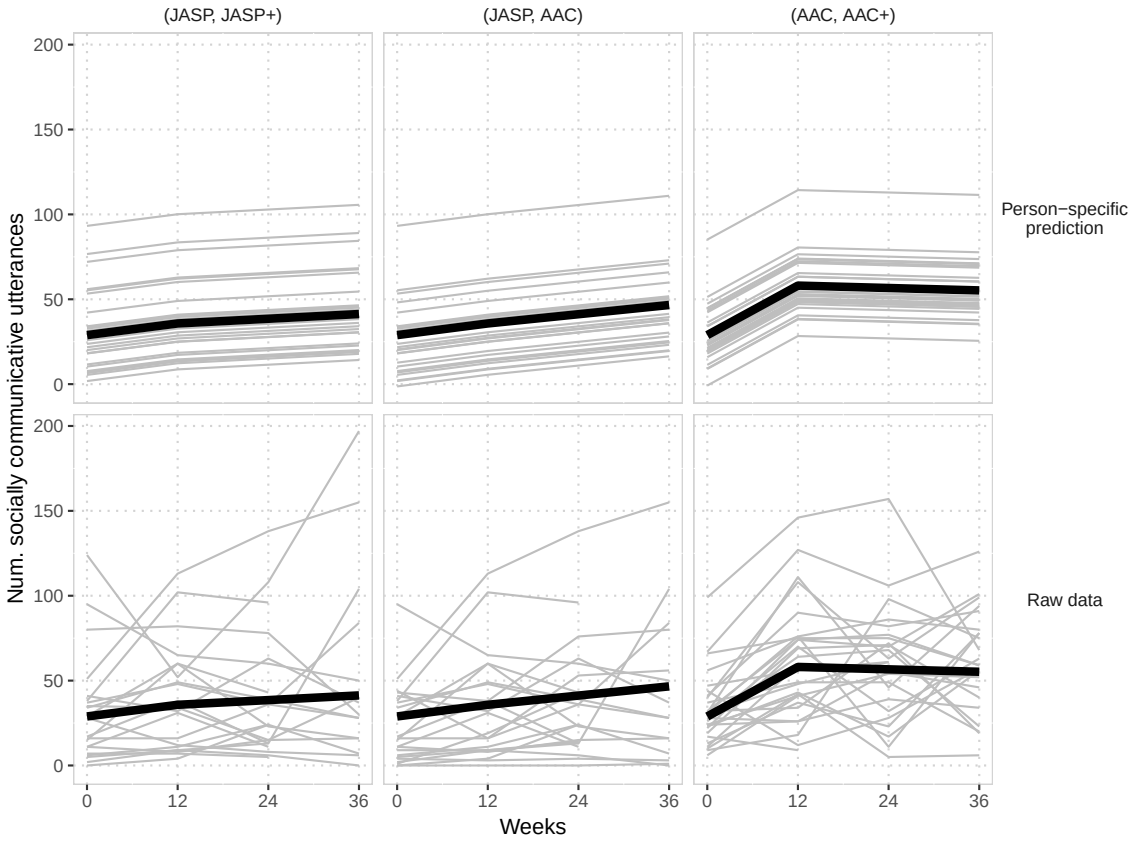


Figure 4.5: Person-specific predicted trajectories in the autism SMART. Top row contains the predicted trajectories based on the intercepts-only mixed model. Bottom row contains the observed number of socially communicative utterances. Bold lines are the estimated marginal mean trajectories under each DTR for children at age 6.3, the average age of the study participants. Responders to initial treatment with JASP are observable under both (JASP, JASP+) and (JASP, AAC), and the observed and predicted trajectories for these participants are displayed for both of these regimens.

APPENDICES

APPENDIX A

Alternative Constructions of $\tilde{\mathbf{U}}$ in the Knockoff Filter

A.1 A deterministic, arbitrary $\tilde{\mathbf{U}}$

First note that in Algorithms III.3–III.5, $\tilde{\mathbf{U}}$ is explicitly a function of a random matrix. In Algorithm III.2, however, $\tilde{\mathbf{U}}$ is deterministic, even though it is still arbitrary. With a given computational environment and numerical algorithm for the QR decomposition, Algorithm III.2 will return the same result every time the matrix $[\mathbf{X}\ \mathbf{0}]$ is decomposed. However, since $[\mathbf{X}\ \mathbf{0}]$ is not full rank, the QR decomposition of this matrix is not unique. Indeed, for any $p \times p$ orthogonal matrix $\mathbf{H}$, we can replace $\mathbf{Q}_0$ in Algorithm III.2 with $\mathbf{Q}_0\mathbf{H}$ and still obtain a valid QR decomposition. So Algorithm III.2 provides a deterministic (given $\mathbf{X}$) but arbitrary choice of $\tilde{\mathbf{U}}$ based on our chosen algorithm for the QR decomposition.

A.2 Validation set approach to $\tilde{\mathbf{U}}$

A validation set approach could provide guidance for selecting a single $\tilde{\mathbf{U}}$ with the largest number of estimated true signals. Algorithm A.1 suggests one way of using

a validation set approach in this manner. In this algorithm, the knockoff filter is repeatedly applied to one half of the training data. In each application of the knockoff filter, a single, random $\tilde{\mathbf{U}}$ is generated, $\mathbf{W}$ is computed, and variables are selected. In this case, $\tilde{\mathbf{U}}$ is an $n/2 \times p$ matrix constructed so that $\tilde{\mathbf{U}}^\intercal \mathbf{X}_1 = \tilde{\mathbf{U}}^\intercal \mathbf{X}_2 = \mathbf{0}$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are random row-wise partitions of $\mathbf{X}$. As previously demonstrated, the number of selected variables will vary across these repetitions, and for each repetition, we can record the number of selected variables. Finally, we use the $\tilde{\mathbf{U}}$ matrix corresponding to the largest number of selected variables to construct knockoffs and output a final set of selected variables using $\mathbf{X}_2$.

Note that this procedure for computing $\tilde{\mathbf{U}}$ seems to violate a guiding principle of the knockoff filter, which requires that the knockoff matrix $\tilde{\mathbf{X}}$ is constructed without knowledge of $\mathbf{Y}$. Choosing the single $\tilde{\mathbf{U}}$ which leads to the largest number of selections means that it is a function of the partitioned response vector $\mathbf{Y}_1$. However, the final variables selections are obtained by applying the knockoff filter to $\mathbf{X}_2, \mathbf{Y}_2$, the validation set, thereby ensuring that the knockoffs are constructed independently of the response.

The simulation results in Figures A.1–A.4 compare this validation set approach to Algorithms III.3 and III.5 for the fixed $(\mathbf{X}, \mathbf{Y})$ pair generated as described in Section 3.3. This validation set approach has very limited power to detect signals, and does not seem to reduce variation in the selected variable set except to the extent that it causes the knockoff filter to select zero variables with high probability. The average behavior of this validation set approach, over repeated sampling of $\mathbf{X}$ and $\mathbf{Y}$, was studied in the simulations presented in Section 3.5.1. In short, the validation set choice of $\tilde{\mathbf{U}}$ seems to reduce the power of the knockoff filter to an unacceptably low level. While Algorithm A.1 is only an initial attempt at implementing a validation set approach for selecting $\tilde{\mathbf{U}}$, this approach is not pursued further in this chapter.

Algorithm A.1 Choose a single $\tilde{\mathbf{U}}$ using holdout training set

Require: $n \times p$ design matrix, $\mathbf{X}$, response vector $\mathbf{Y}$, integer B

1: Partition $\mathbf{X}$ row-wise so that $\mathbf{X}_1$ and $\mathbf{X}_2$ each have $n/2$ randomly selected rows of $\mathbf{X}$
2: Partition $\mathbf{Y}$ into $\mathbf{Y}_1$ and $\mathbf{Y}_2$ corresponding to the same partitions $\mathbf{X}_1$ and $\mathbf{X}_2$
3: Decompose $[\mathbf{X}_1\ \mathbf{X}_2] = \mathbf{QR}$ (note $\mathbf{Q}$ is $n/2 \times 2p$)
4: Initialize $e \leftarrow -1$ and $\tilde{\mathbf{U}}_{\text{best}} \leftarrow$ `NULL`
5: **for** $j = 1, \ldots, B$ **do**
6: $\quad \mathbf{Z}_{\frac{n}{2} \times p} \sim N(0,1)$ with i.i.d. entries
7: $\quad$ Decompose $(\mathbf{I} - \mathbf{QQ}^\intercal)\mathbf{Z} = \mathbf{Q}_u\mathbf{R}_u$
8: $\quad \tilde{\mathbf{U}}_1 \leftarrow \mathbf{Q}_u$
9: $\quad$ Perform Algorithm III.1 (knockoff filter), using $\tilde{\mathbf{U}}_1$ in step 3, with $\mathbf{X}_1$, $\mathbf{Y}_1$ as inputs
10: $\quad$ **if** # selected variables $> e$ **then**
11: $\quad\quad e \leftarrow$ # selected variables
12: $\quad\quad \tilde{\mathbf{U}}_{\text{best}} \leftarrow \tilde{\mathbf{U}}_1$
13: $\quad$ **end if**
14: **end for**
15: Perform Algorithm III.1 (knockoff filter) with $\tilde{\mathbf{U}}_{\text{best}}$ in step 3 and $\mathbf{X}_2, \mathbf{Y}_2$ as inputs

A.3 Geometric alignment between $\tilde{\mathbf{U}}$ and $\mathbf{Y}$

Another principle to motivate a choice of a single valid $\tilde{\mathbf{U}}$ can be found via a geometric argument. Recall that in a fixed dataset $(\mathbf{X}, \mathbf{Y})$, the selected variables depend on $\tilde{\mathbf{U}}$ only through the cross products $\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^\intercal \mathbf{Y}$. Consider the decomposition

$$\mathbf{Y} = \mathbf{Y}_X + \mathbf{Y}_U + \mathbf{Y}_R, \tag{A.1}$$

where $\mathbf{Y}_X \in \text{Col}(\mathbf{X})$, $\mathbf{Y}_U \in \text{Col}(\tilde{\mathbf{U}})$ $\mathbf{Y}_R$ is the remainder, $\mathbf{Y}_R = \mathbf{Y} - \mathbf{Y}_X - \mathbf{Y}_U$, and $\text{Col}(\mathbf{A})$ denotes the column space of $\mathbf{A}$. (By construction, $\text{Col}(\mathbf{X})$ is orthogonal to $\text{Col}(\tilde{\mathbf{U}})$.) We can argue geometrically that the effect of $\tilde{\mathbf{U}}$ on the selected variables depends on the relative magnitudes of these three components of $\mathbf{Y}$. First recall that $\tilde{\mathbf{X}}$ can be written as $\tilde{\mathbf{X}} = \mathbf{XA} + \mathbf{B}$ where $\mathbf{B} = \text{Col}(\tilde{\mathbf{U}})$. If $\|\mathbf{Y}_U\|$ is large relative to $\|\mathbf{Y}\|$, then $\tilde{\mathbf{X}}$ contains more of the variation in $\mathbf{Y}$ that is not present in the columns of $\mathbf{X}$; If $\|\mathbf{Y}_U\| = 0$, then the component of $\mathbf{Y}$ that is contained in the column space

of $\tilde{\mathbf{X}}$ is fully contained in the column space of $\mathbf{X}$. In this latter case, the knockoff variables have no information about $\mathbf{Y}$ beyond that which is contained in $\mathbf{X}$, and therefore the estimated effects of $\tilde{\mathbf{X}}$ will be small compared to those of $\mathbf{X}$, leading to anti-conservative variable selection. In the former case, when a larger component of $\mathbf{Y}$ lies in the column space of $\tilde{\mathbf{X}}$, there is greater "confusion" between $\mathbf{X}$ and the knockoffs, leading to fewer detected signals. That is, the estimated partial effects of $\tilde{\mathbf{X}}$ on $\mathbf{Y}$, in the presence of $\mathbf{X}$, will be comparable to those of $\mathbf{X}$, leading to few selected variables.

More concretely, we can analyze

$$\frac{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|^2}{\|\mathbf{Y}\|^2} = \frac{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\boldsymbol{\epsilon}\|^2}{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|^2 + \|(\mathbf{I}-\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal)Y\|^2} \tag{A.2}$$

$$= \frac{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\boldsymbol{\epsilon}\|^2}{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\boldsymbol{\epsilon}\|^2 + \|\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}-\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal)\epsilon\|^2} \tag{A.3}$$

as a measure of the geometric overlap between $\tilde{\mathbf{U}}$ and $\mathbf{Y}$. Since $\tilde{\mathbf{U}}^\intercal\tilde{\mathbf{U}} = \mathbf{I}$, the orthogonal projection of $\mathbf{Y}$ onto $\mathrm{Col}(\tilde{\mathbf{U}})$ is given by $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}$. Thus $\frac{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|^2}{\|\mathbf{Y}\|^2}$ measures the relative magnitude of $\mathbf{Y}_U$ in the decomposition (A.1). Based on the previous argument, we should expect larger values of this fraction to correspond to very few selected variables, and near-zero values of this fraction to correspond with a large number of selected variables. To empirically assess this relationship between $\frac{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|}{\|\mathbf{Y}\|}$ and the number of selected variables, I generated 500 replicates of $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ has independent, mean-zero Gaussian rows, and $\mathbf{Y}$ is generated from $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I})$, with $n = 5000, p = 100$ and $|\beta_j| = 3.5$ for all nonzero β_j. In each replicate, $\tilde{\mathbf{U}}$ was computed using Algorithm III.3. Figures A.10 and A.11 demonstrate modest negative correlations between $\frac{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|}{\|\mathbf{Y}\|}$ and the number of selections, the false discovery proportion, and the true positive rate (power). This negative correlation is stronger with lower population correlation among the features and in less sparse settings.

This empirical evidence and heuristic argument associating higher values of $\frac{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|}{\|\mathbf{Y}\|}$ with fewer selected variables in the knockoff filter suggests that $\tilde{\mathbf{U}}$ should be selected to minimize this fraction. However, simply decreasing $\frac{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|}{\|\mathbf{Y}\|}$ as much as possible could lead to a loss of FDR control. One way to obtain a principled choice of $\tilde{\mathbf{U}}$ based on this metric, but without losing FDR control, is to choose a $\tilde{\mathbf{U}}$ so that $\frac{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|}{\|\mathbf{Y}\|}$ is close to its expected value.

An approximation to $\mathbb{E}\left(\frac{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|}{\|\mathbf{Y}\|}\right)$ can be derived as follows. Assume that $\mathbf{Y} \mid \mathbf{X}$ follows the population model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Generating $\tilde{\mathbf{U}}$ independently of $\mathbf{Y}$ (and conditional on $\mathbf{X}$) with Algorithm III.3, III.4, or III.5, we have that $\boldsymbol{\epsilon}$ and $\tilde{\mathbf{U}}$ are independent. Then

$$\begin{aligned}
\mathbb{E}\left(\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\boldsymbol{\epsilon}\|^2 \mid \mathbf{X}\right) &= \mathbb{E}\left(\mathbb{E}\left(\boldsymbol{\epsilon}^\intercal\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\boldsymbol{\epsilon} \mid \mathbf{X}, \tilde{\mathbf{U}}\right)\right) && \text{(A.4)}\\
&= \mathbb{E}\left(\operatorname{tr}\left[\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\operatorname{Cov}(\boldsymbol{\epsilon})\right]\right) + \mathbb{E}\left(\mathbb{E}(\boldsymbol{\epsilon})^\intercal\,\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbb{E}(\boldsymbol{\epsilon})\right) && \text{(A.5)}\\
&= \sigma^2\mathbb{E}\left(\operatorname{tr}(\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal)\right) && \text{(A.6)}\\
&= p\sigma^2, && \text{(A.7)}
\end{aligned}$$

since the trace of a projection matrix is its rank. In addition,

$$\begin{aligned}
&\|\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal)\epsilon\|^2 && \text{(A.8)}\\
&= \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\intercal\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^\intercal(\mathbf{I} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal)(\mathbf{I} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal)\boldsymbol{\epsilon} && \text{(A.9)}\\
&= \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\epsilon}^\intercal\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^\intercal(\mathbf{I} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal)\boldsymbol{\epsilon}, && \text{(A.10)}
\end{aligned}$$

so

$$\begin{aligned}
\mathbb{E}\left(\|\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal)\boldsymbol{\epsilon}\|^2 \mid \mathbf{X}\right) &= \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} + \mathbb{E}\left(\boldsymbol{\epsilon}^\intercal\boldsymbol{\epsilon} \mid \mathbf{X}\right) - \mathbb{E}\left(\boldsymbol{\epsilon}^\intercal\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\boldsymbol{\epsilon} \mid \mathbf{X}\right) && \text{(A.11)}\\
&= \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} + N\sigma^2 - p\sigma^2 && \text{(A.12)}
\end{aligned}$$

Then, making the approximation

$$\mathbb{E}\left(\frac{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|^2}{\|\mathbf{Y}\|^2} \mid \mathbf{X}\right) \approx \frac{\mathbb{E}\left(\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|^2 \mid \mathbf{X}\right)}{\mathbb{E}\left(\|\mathbf{Y}\|^2 \mid \mathbf{X}\right)}, \tag{A.13}$$

we have

$$\mathbb{E}\left(\frac{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|^2}{\|\mathbf{Y}\|^2} \mid \mathbf{X}\right) \approx \frac{\mathbb{E}\left(\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|^2 \mid \mathbf{X}\right)}{\mathbb{E}\left(\|\mathbf{Y}\|^2 \mid \mathbf{X}\right)} = \frac{p\sigma^2}{(N-p)\sigma^2 + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta}}. \tag{A.14}$$

This can be used to compute a single $\tilde{\mathbf{U}}$ so that the sample fraction $\frac{\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|^2}{\|\mathbf{Y}\|^2}$ is close to its approximated expected value in (A.14). Suppose there existed $\tilde{\mathbf{U}}_1$, $\tilde{\mathbf{U}}_2$ so that $\tilde{\mathbf{U}}_1^\intercal\tilde{\mathbf{U}}_1 = \tilde{\mathbf{U}}_2^\intercal\tilde{\mathbf{U}}_2 = \mathbf{I}$ and $\tilde{\mathbf{U}}_1^\intercal\tilde{\mathbf{U}}_2 = \tilde{\mathbf{U}}_1^\intercal X = \tilde{\mathbf{U}}_2^\intercal\mathbf{X} = \mathbf{0}$. Then, for any $\theta \in (0, \pi)$, define

$$\mathbf{U}_\theta := \sin\theta\tilde{\mathbf{U}}_1 + \cos\theta\tilde{\mathbf{U}}_2. \tag{A.15}$$

Then $\mathbf{U}_\theta^\intercal\mathbf{U}_\theta = \mathbf{I}$ and $\mathbf{U}_\theta^\intercal\mathbf{X} = \mathbf{0}$, so $\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\mathbf{S}) + \mathbf{U}_\theta\mathbf{C}$ is a valid matrix of knockoffs. Let $\mathbf{P}_X$ be the $n \times n$ projection matrix onto $\mathrm{Col}(\mathbf{X})$, and choose the $n \times p$ matrices $\tilde{\mathbf{U}}_1, \tilde{\mathbf{U}}_2, \tilde{\mathbf{U}}_3$ so that

$$\tilde{\mathbf{U}}_1^\intercal\tilde{\mathbf{U}}_1 = \mathbf{I},\ \tilde{\mathbf{U}}_1^\intercal\mathbf{X} = \mathbf{0} \tag{A.16}$$

$$\tilde{\mathbf{U}}_2^\intercal\tilde{\mathbf{U}}_2 = \mathbf{I},\ \tilde{\mathbf{U}}_1^\intercal\tilde{\mathbf{U}}_2 = \tilde{\mathbf{U}}_2^\intercal\mathbf{X} = \tilde{\mathbf{U}}_2^\intercal(\mathbf{I} - \mathbf{P}_X)\mathbf{Y} = \mathbf{0} \tag{A.17}$$

$$\tilde{\mathbf{U}}_3^\intercal\tilde{\mathbf{U}}_3 = \mathbf{I},\ \tilde{\mathbf{U}}_3^\intercal\mathbf{X} = \tilde{\mathbf{U}}_3^\intercal\tilde{\mathbf{U}}_1 = \mathbf{0}, (\mathbf{I} - \mathbf{P}_X)\mathbf{Y} \subset \mathrm{Col}(\tilde{\mathbf{U}}_3) \tag{A.18}$$

In other words, $\tilde{\mathbf{U}}_1$ is the usual, arbitrary choice of $\tilde{\mathbf{U}}$ in the knockoff construction; $\tilde{\mathbf{U}}_2$ is orthogonal to the complement of $\mathbf{Y}$ in $\mathrm{Col}(\mathbf{X})$; $\tilde{\mathbf{U}}_3$ *contains* the complement of $\mathbf{Y}$ in $\mathrm{Col}(\mathbf{X})$; and both $\tilde{\mathbf{U}}_2$ and $\tilde{\mathbf{U}}_3$ are orthogonal to $\tilde{\mathbf{U}}_1$. On the one hand, if we used $\tilde{\mathbf{U}}_2$ to construct $\tilde{\mathbf{X}}$, we will eliminate from $\tilde{\mathbf{X}}$ any component of $\mathbf{Y}$ that is not already captured by the columns of $\mathbf{X}$. This should induce less confusion between

$\tilde{\mathbf{X}}$ and $\mathbf{X}$ and hence more detected signals (which may be false positives). On the other hand, using $\tilde{\mathbf{U}}_3$ to construct $\tilde{\mathbf{X}}$ should increase the overlap between $\mathbf{Y}$ and $\tilde{\mathbf{X}}$, leading to greater confusion between $\tilde{\mathbf{X}}$ and $\mathbf{X}$ and hence fewer detected signals. The parameterization (A.15) allows for an intermediate choice between these two extremes. Letting

$$\mathbf{U}_{\theta_2} = \sin\theta_2\tilde{\mathbf{U}}_1 + \cos\theta_2\tilde{\mathbf{U}}_2 \tag{A.19}$$

$$\mathbf{U}_{\theta_3} = \sin\theta_3\tilde{\mathbf{U}}_1 + \cos\theta_3\tilde{\mathbf{U}}_3 \tag{A.20}$$

we can either increase (with $\mathbf{U}_{\theta_3}$) or decrease (with $\mathbf{U}_{\theta_2}$) the proportion of $\mathbf{Y}$ that overlaps with $\tilde{\mathbf{X}}$. For example, taking $\theta_2 = 0$ means that $\frac{\|\mathbf{U}_{\theta_2}\mathbf{U}_{\theta_2}^\intercal\mathbf{Y}\|^2}{\|\mathbf{Y}\|^2} = 0$.

Figure A.9 displays $\frac{\|\mathbf{U}_{\theta_2}\mathbf{U}_{\theta_2}^\intercal\mathbf{Y}\|^2}{\|\mathbf{Y}\|^2}$ and $\frac{\|\mathbf{U}_{\theta_3}\mathbf{U}_{\theta_3}^\intercal\mathbf{Y}\|^2}{\|\mathbf{Y}\|^2}$ as functions of θ_2 and θ_3, respectively, for two realizations of $(\mathbf{X}, \mathbf{Y})$. This illustrates how this measure of the overlap between $\tilde{\mathbf{U}}_1$ and $\mathbf{Y}$ can be increased, by interpolating between $\tilde{\mathbf{U}}_1$ and $\tilde{\mathbf{U}}_3$, or decreased, by interpolating between $\tilde{\mathbf{U}}_1$ and $\tilde{\mathbf{U}}_2$. The two realizations of $(\mathbf{X}, \mathbf{Y})$ in this example illustrate that the realized fraction can be greater than or less than the large-sample value given in (A.14) or the fraction we would expect based only on the dimensions of each corresponding subspace, i.e. $\frac{p}{N-p}$.

Based on these arguments, we can choose a desired value for $\frac{\|\mathbf{U}_\theta\mathbf{U}_\theta^\intercal\mathbf{Y}\|^2}{\|\mathbf{Y}\|^2}$ and achieve this value using a grid search over $\theta \in (0, \pi)$. Algorithm A.2 describes how to use a grid search over $\theta \in (0, \pi)$ to compute $\mathbf{U}_\theta$ so that $\frac{\|\mathbf{U}_\theta\mathbf{U}_\theta^\intercal\mathbf{Y}\|^2}{\|\mathbf{Y}\|^2}$ is equal to the plugin estimate of (A.14). This choice of $\tilde{\mathbf{U}}$ was compared with alternative methods in the fixed-$(\mathbf{X}, \mathbf{Y})$ simulation scenario with $n = 5000$ and $p = 100$; Figures 3.3–3.10 present these simulation results, with or without feature correlation, and in a sparse ($k = 10$ true signals out of $p = 100$) and non-sparse ($k = 50$ true signals out of $p = 100$) setting. These results show that computing $\mathbf{U}_\theta$ to control $\frac{\|\mathbf{U}_\theta\mathbf{U}_\theta^\intercal Y\|^2}{\|\mathbf{Y}\|^2}$ des not appreciably reduce variation in the number of selected variables or in the frequency

of selecting null and non-null variables for a fixed $(\mathbf{X}, \mathbf{Y})$ pair. The simulation studies presented below, in which $(\mathbf{X}, \mathbf{Y})$ are jointly sampled, this choice of $\tilde{\mathbf{U}}$ is shown to control FDR but to have otherwise similar operating characteristics (power, variation in the number of selected variables) as the standard knockoff filter with $\tilde{\mathbf{U}}$ computed using an algorithm like III.3.

Algorithm A.2 Control geometric alignment of $\tilde{\mathbf{U}}$ and $\mathbf{Y}$

Require: $n \times p$ design matrix, $\mathbf{X}$, response vector $\mathbf{Y}$

1: $\hat{\boldsymbol{\beta}} \leftarrow (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{Y}$

2: $\hat{\sigma}^2 \leftarrow \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(n-p)$

3: $c \leftarrow \frac{p\hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + \hat{\boldsymbol{\beta}}^\intercal \mathbf{X}^\intercal \mathbf{X} \hat{\boldsymbol{\beta}}}$

4: $\tilde{\mathbf{U}}_1 \leftarrow$result of Algorithm III.3

5: Construct $\mathbf{Z}_{n\times p}$ with i.i.d. $N(0,1)$ entries

6: Perform the QR decomposition $\left[\mathbf{X}\ \tilde{\mathbf{U}}_1\right] = \mathbf{Q}_{xu}\mathbf{R}_{xu}$

7: Perform the QR decomposition $[\mathbf{Q}_{xu},\ (\mathbf{I} - \mathbf{P}_X)\mathbf{Y}] = \mathbf{Q}_{xuy}\mathbf{R}_{xuy}$

8: Perform the QR decomposition $(\mathbf{I} - \mathbf{Q}_{xuy}\mathbf{Q}_{xuy}^\intercal)\mathbf{Z} = \mathbf{Q}_2\mathbf{R}_2$

9: $\tilde{\mathbf{U}}_2 \leftarrow \mathbf{Q}_2$

10: Perform the QR decomposition $(\mathbf{I} - \mathbf{Q}_{xu}\mathbf{Q}_{xu}^\intercal)\left[(\mathbf{I} - \mathbf{P}_X)\mathbf{Y},\ \mathbf{Z}_{[0:(p-1)]}\right] = \mathbf{Q}_3\mathbf{R}_3$

11: $\tilde{\mathbf{U}}_3 \leftarrow \mathbf{Q}_3$

12: Define $\mathbf{U}_{\theta_2} := \sin\theta_2 \tilde{\mathbf{U}}_1 + \cos\theta_2 \tilde{\mathbf{U}}_2$

13: Define $\mathbf{U}_{\theta_3} := \sin\theta_3 \tilde{\mathbf{U}}_1 + \cos\theta_3 \tilde{\mathbf{U}}_3$

14: Use a grid search to find $\theta_2, \theta_3 \in (0, \pi)$ to minimize $\left|\frac{\|\mathbf{U}_{\theta_2}\mathbf{U}_{\theta_2}^\intercal \mathbf{Y}\|^2}{\|\mathbf{Y}\|^2} - c\right|$ and $\left|\frac{\|\mathbf{U}_{\theta_3}\mathbf{U}_{\theta_3}^\intercal \mathbf{Y}\|^2}{\|\mathbf{Y}\|^2} - c\right|$, respectively.

15: Output $\mathbf{U}_{\theta_2}$ or $\mathbf{U}_{\theta_3}$ with the smallest minimizer from step 14.

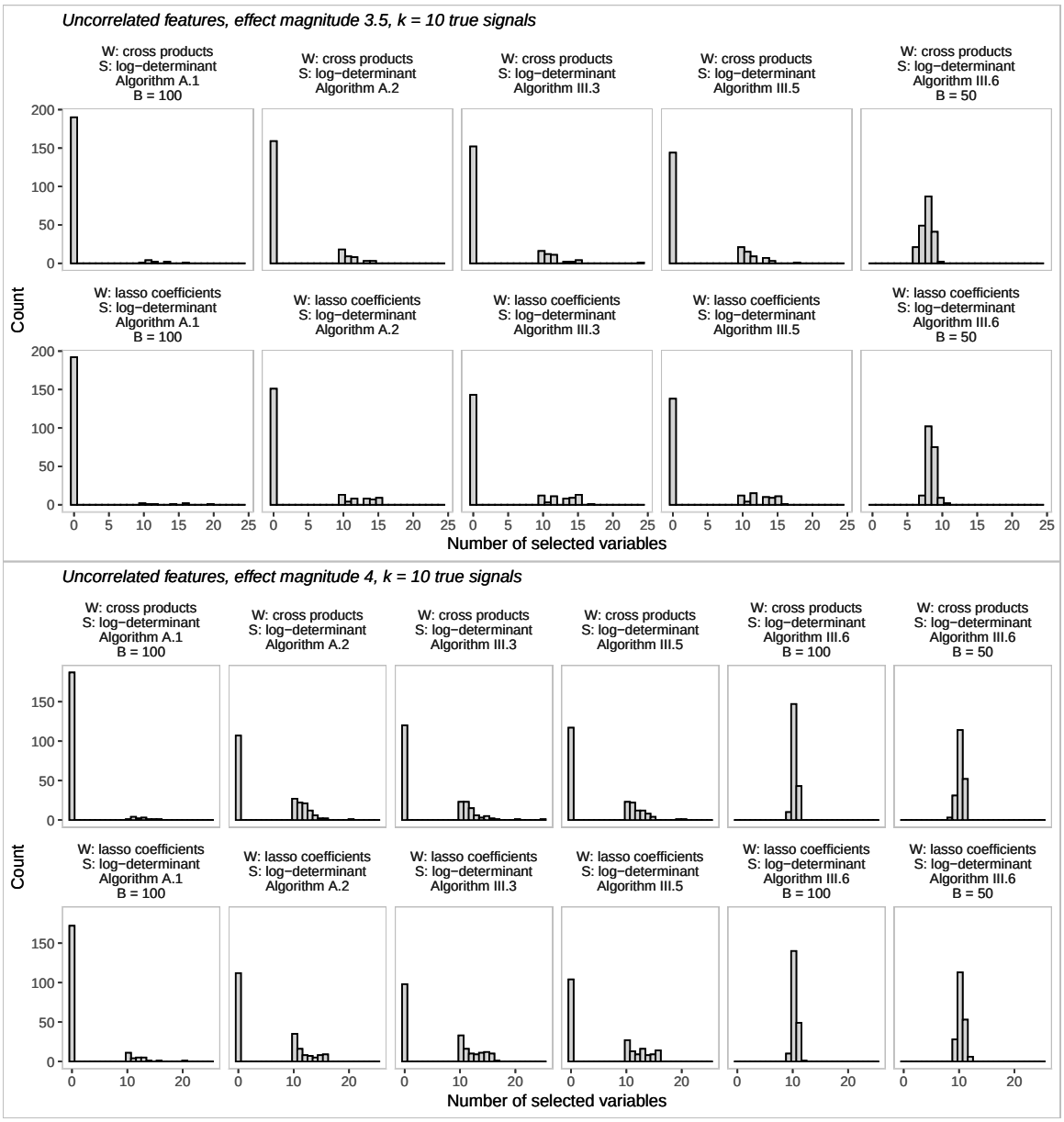


Figure A.1:
Number of selected variables for each method of generating $\tilde{\mathbf{U}}$ with fixed $\mathbf{X}, \mathbf{Y}$ and $n = 5000, p = 100$. Based on 200 knockoff filter replicates, uncorrelated features, and $k = 10$ nonzero β_j.

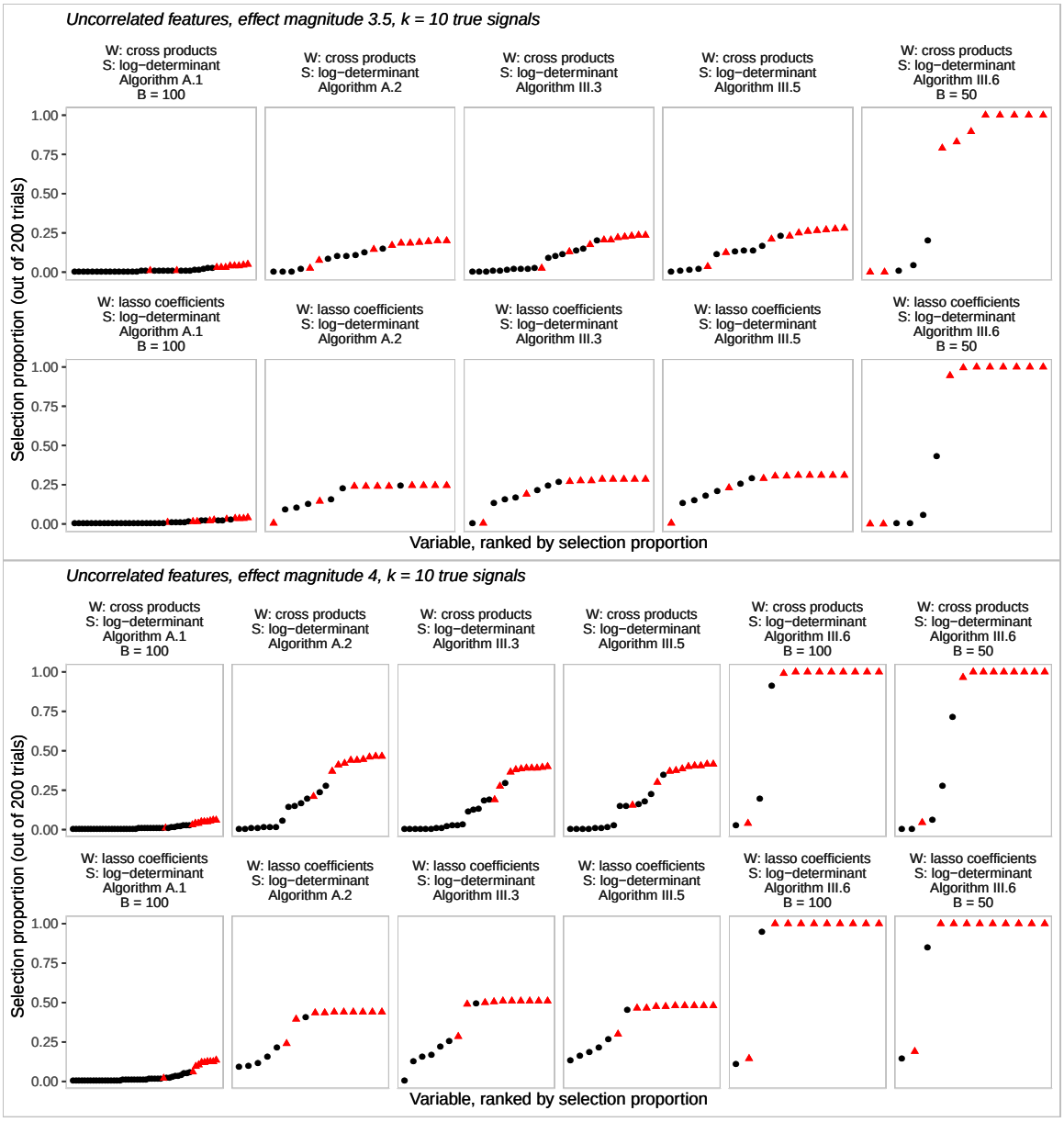


Figure A.2: Variable-specific selection probability for each method of generating $\tilde{\mathbf{U}}$ with fixed $\mathbf{X}, \mathbf{Y}$ and $n = 5000, p = 100$. Based on 200 knockoff filter replicates, uncorrelated features, and $k = 10$ nonzero β_j. Red triangles indicate truly non-null variables. Null variables which were never selected are not displayed.

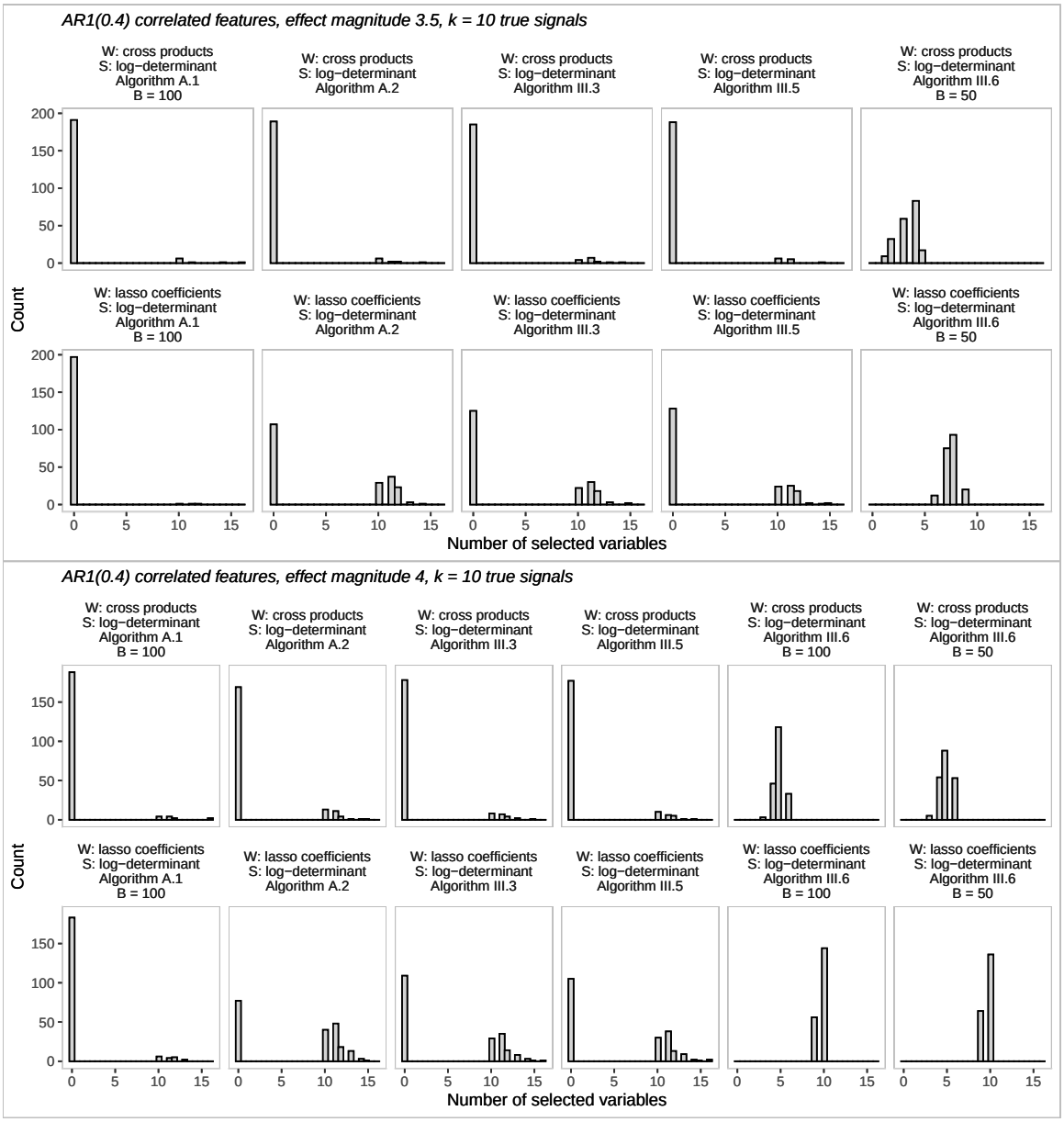


Figure A.3:
Number of selected variables for each method of generating $\tilde{\mathbf{U}}$ with fixed $\mathbf{X}, \mathbf{Y}$, correlated features, and $n = 5000, p = 100$. Based on 200 knockoff filter replicates, autoregressive feature correlation with population parameter 0.4, and $k = 10$ nonzero β_j.

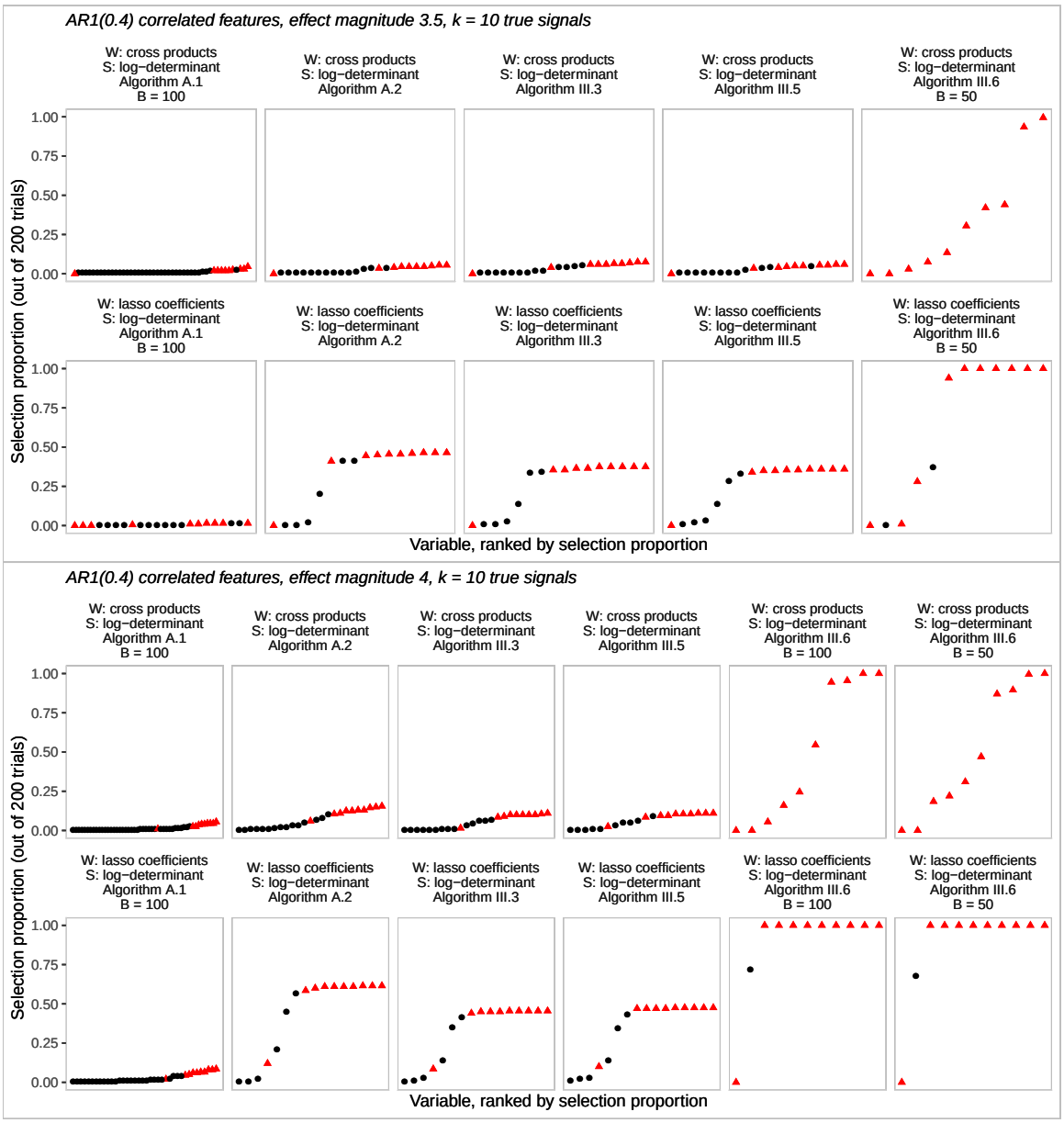


Figure A.4: Variable-specific selection probability for each method of generating $\tilde{\mathbf{U}}$ with fixed $\mathbf{X}, \mathbf{Y}$, correlated features, and $n = 5000, p = 100$. Based on 200 knockoff filter replicates, autoregressive feature correlation with population parameter 0.4, and $k = 10$ nonzero β_j. Red triangles indicate truly non-null variables. Null variables which were never selected are not displayed.

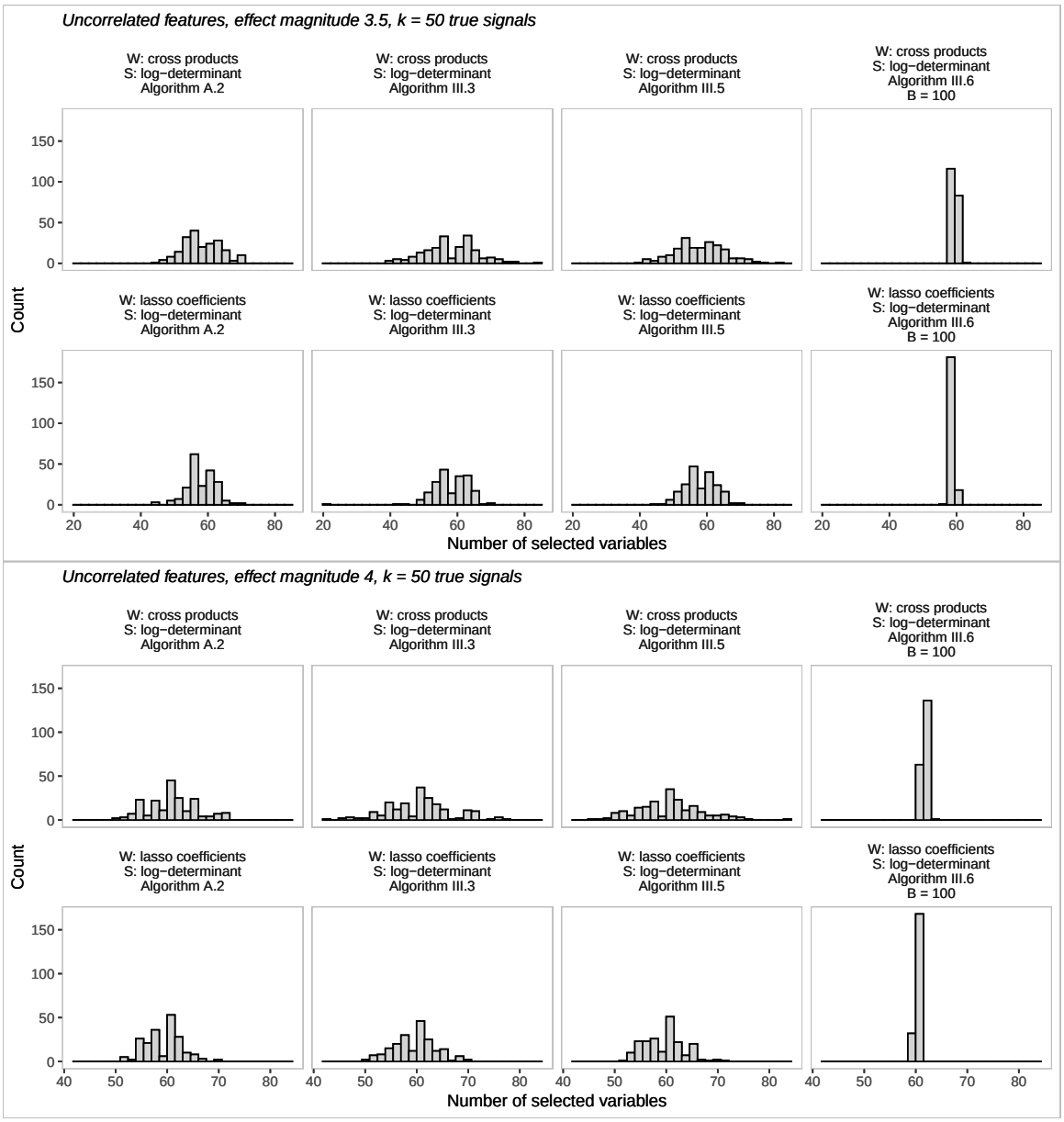


Figure A.5:
Number of selected variables for each method of generating $\tilde{\mathbf{U}}$ with fixed $\mathbf{X}, \mathbf{Y}$, and $n = 5000, p = 100$. Based on 200 knockoff filter replicates, uncorrelated features, and $k = 50$ nonzero β_j.

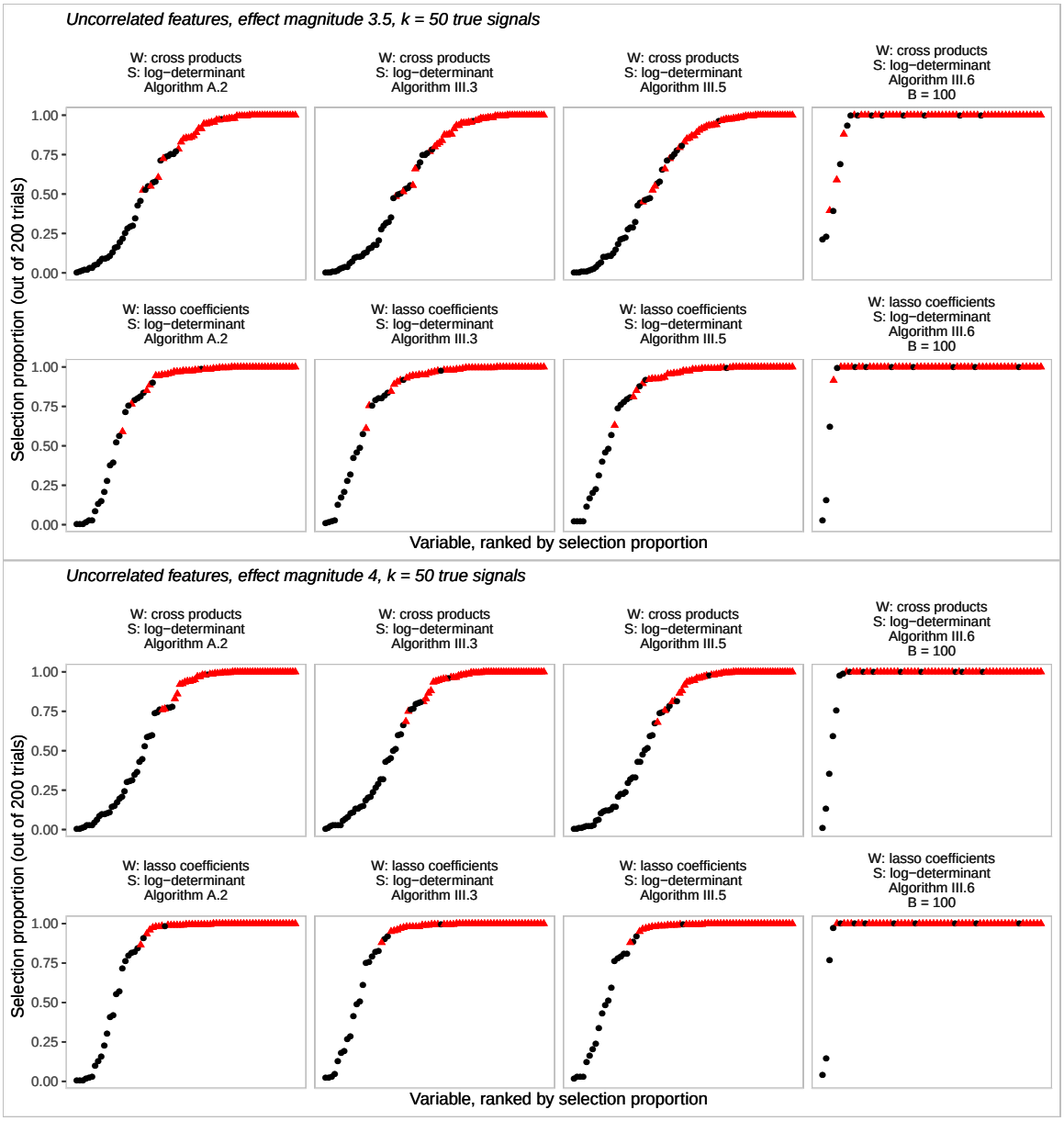


Figure A.6: Variable-specific selection probability for each method of generating $\tilde{\mathbf{U}}$ with fixed $\mathbf{X}, \mathbf{Y}$, and $n = 5000, p = 100$. Based on 200 knockoff filter replicates, uncorrelated features, and $k = 50$ nonzero β_j. Red triangles indicate truly non-null variables. Null variables which were never selected are not displayed.

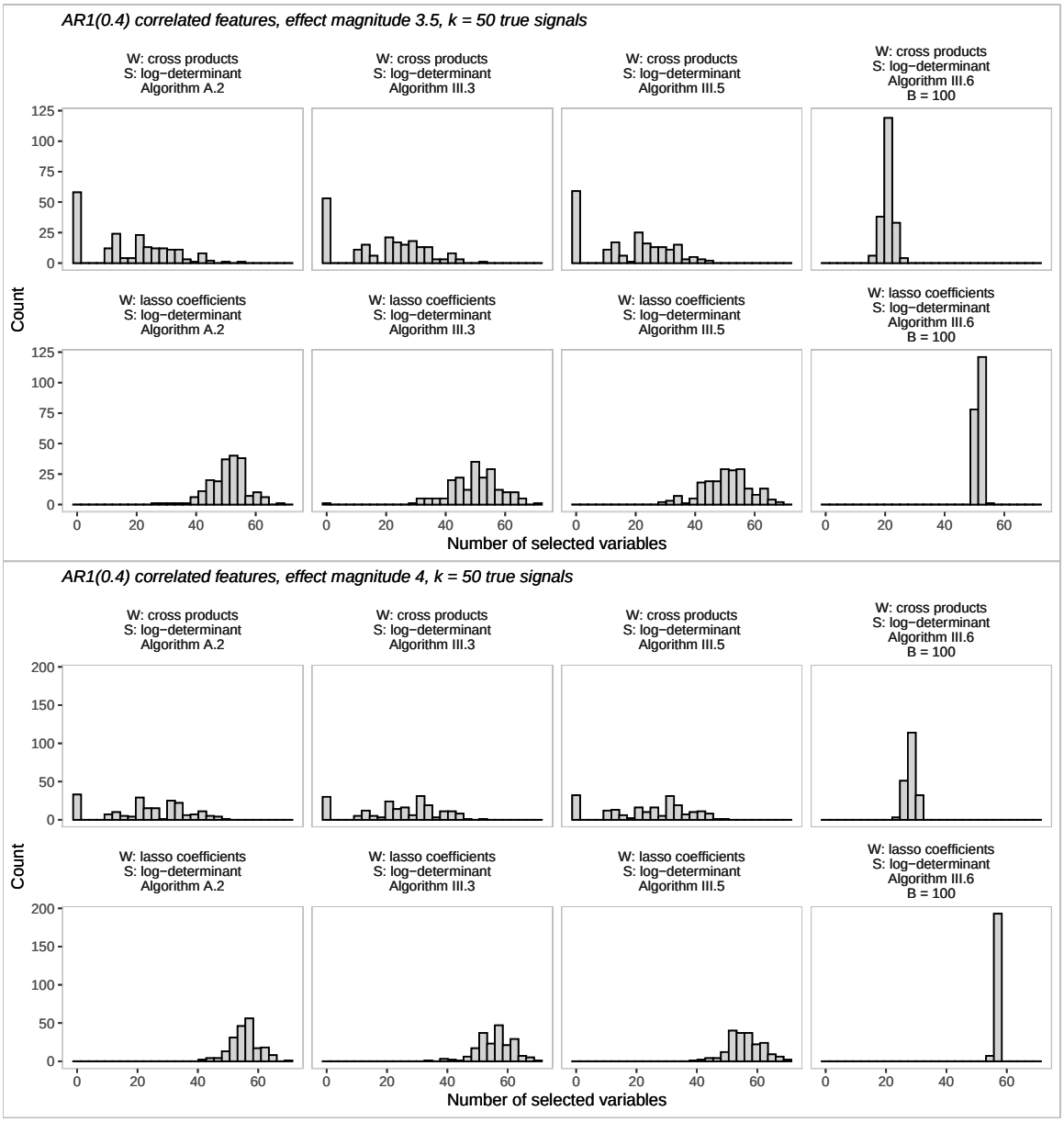


Figure A.7:
Number of selected variables for each method of generating $\tilde{\mathbf{U}}$ with fixed $\mathbf{X}, \mathbf{Y}$, correlated features, and $n = 5000, p = 100$. Based on 200 knockoff filter replicates, autoregressive feature correlation with population parameter 0.4, and $k = 50$ nonzero β_j.

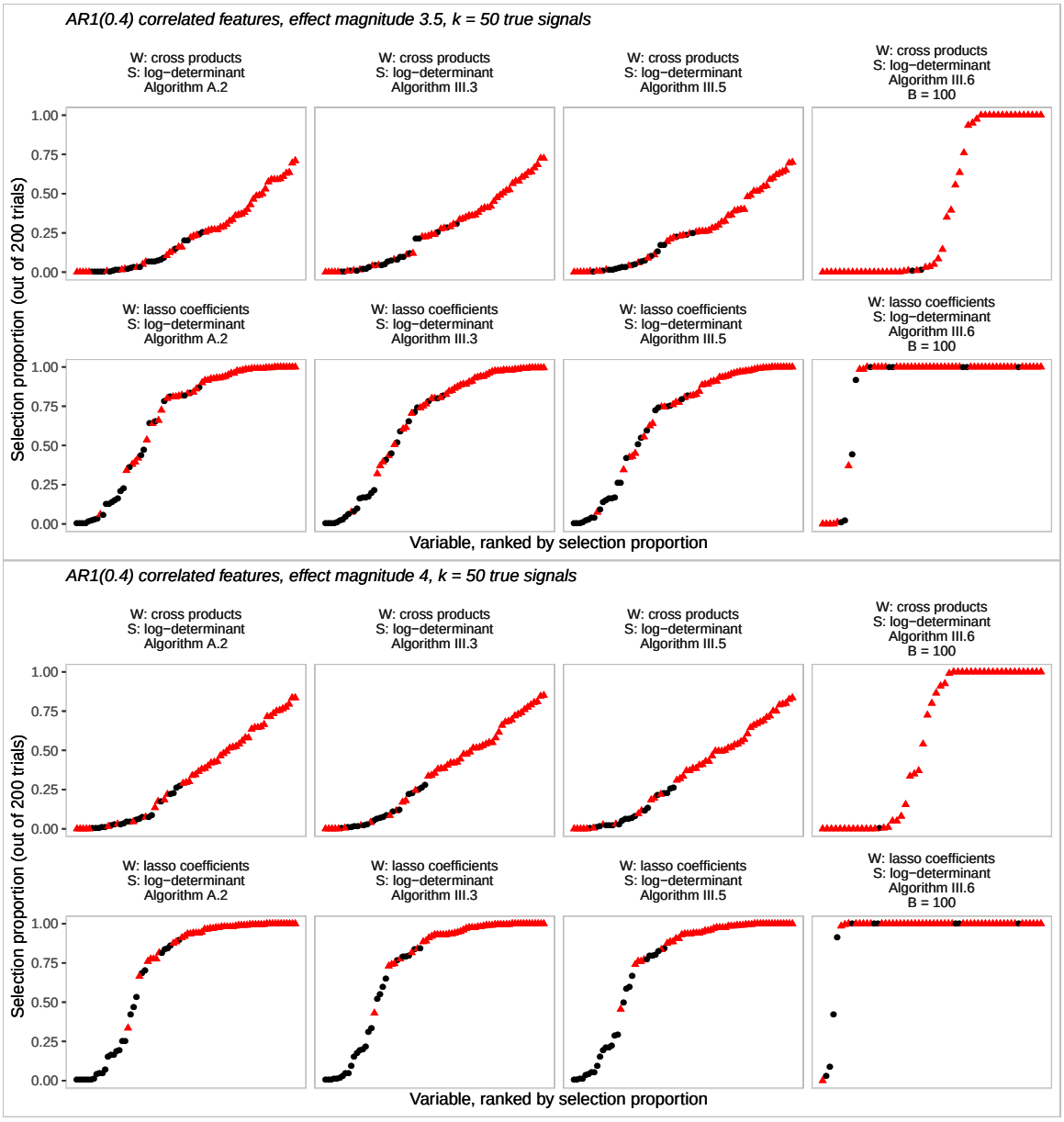


Figure A.8: Variable-specific selection probability for each method of generating $\tilde{\mathbf{U}}$ with fixed $\mathbf{X}, \mathbf{Y}$ and $n = 5000, p = 100$. Based on 200 knockoff filter replicates, autoregressive feature correlation with population parameter 0.4, and $k = 50$ nonzero β_j. Red triangles indicate truly non-null variables. Null variables which were never selected are not displayed.

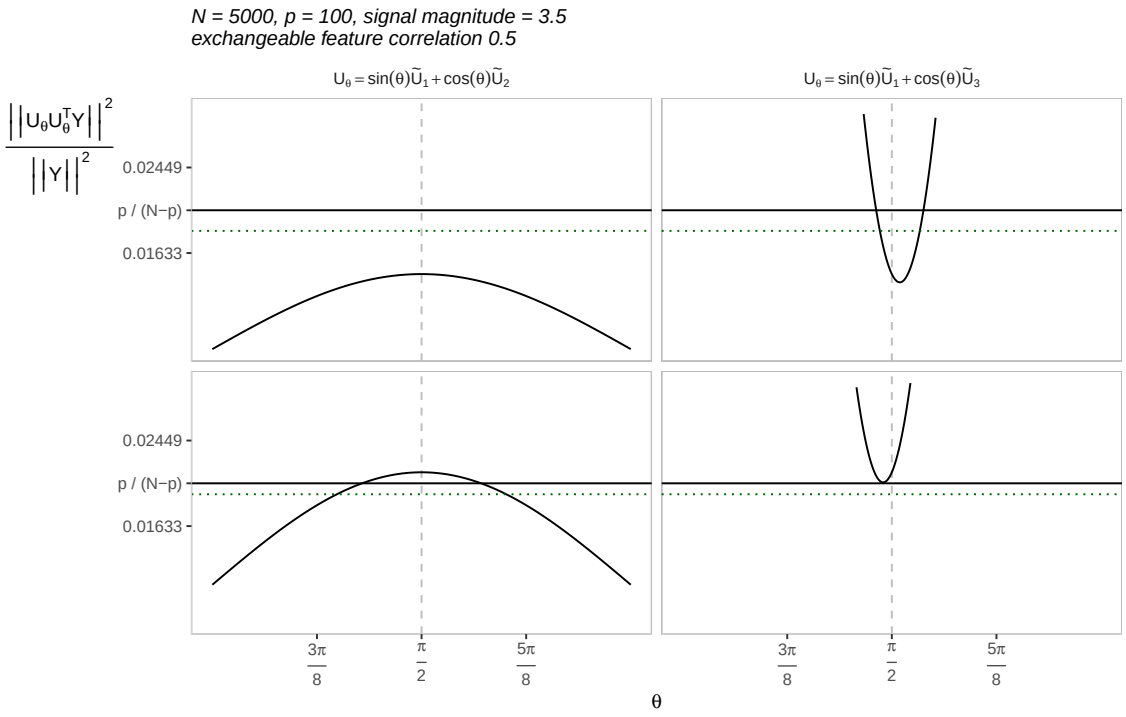


Figure A.9: Fraction of $\mathbf{Y}$ projected onto $\mathbf{U}_\theta$. Each column corresponds to one of the definitions of $\mathbf{U}_\theta$ defined in equations (A.19)–(A.20). Each row corresponds to a single $(\mathbf{X}, \mathbf{Y})$ realization. Dotted horizontal lines are equal to $\frac{p\hat{\sigma}^2}{(N-p)\hat{\sigma}^2 + \hat{\boldsymbol{\beta}}^\intercal \mathbf{X}^\intercal \mathbf{X} \hat{\boldsymbol{\beta}}}$.

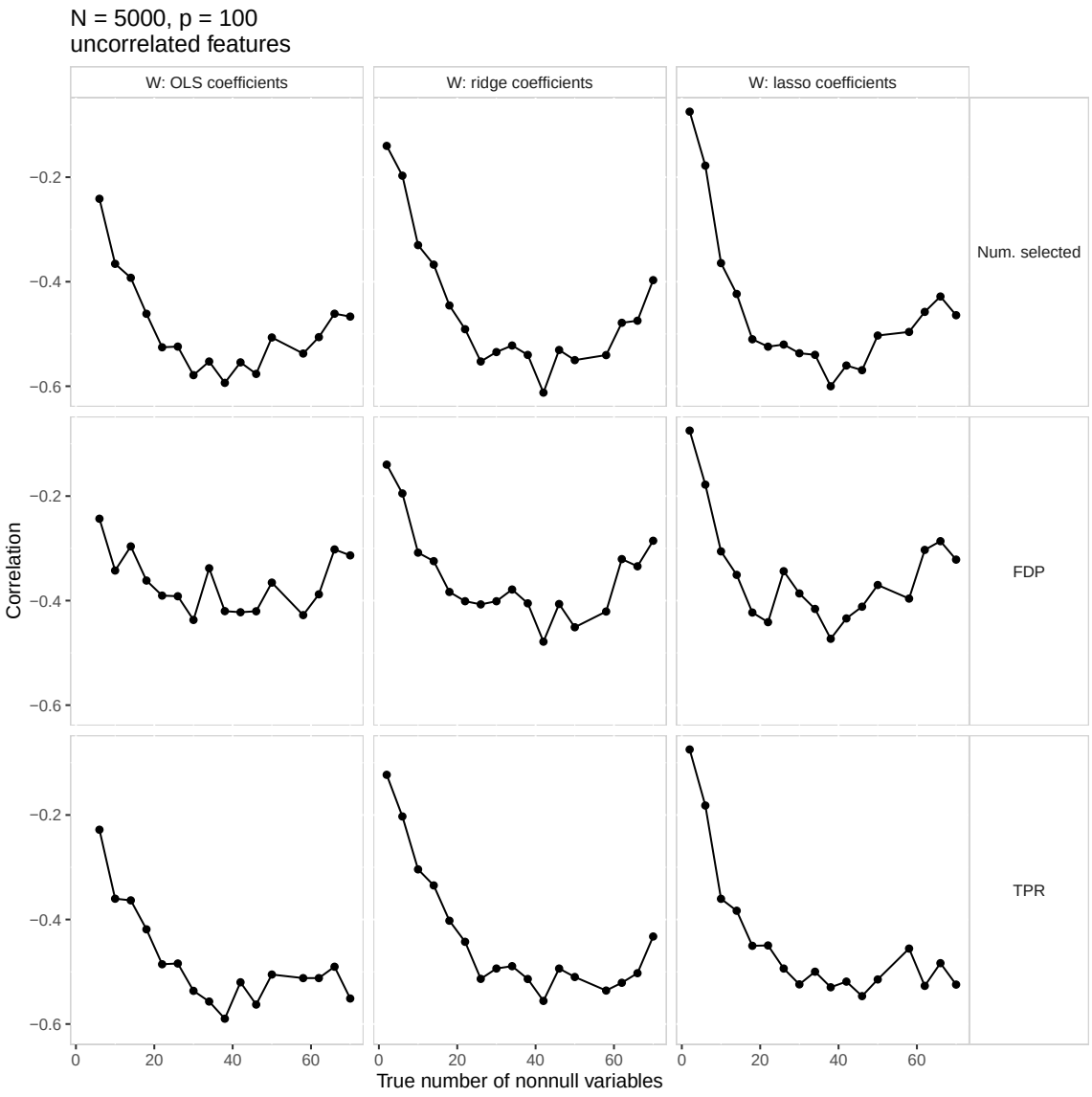


Figure A.10: Correlation between $\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|/\|\mathbf{Y}\|$ and knockoff filter performance metrics as a function of model sparsity.

Figure A.11: Correlation between $\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\intercal\mathbf{Y}\|/\|\mathbf{Y}\|$ and knockoff filter performance metrics as a function of feature correlation.

APPENDIX B

Proofs and Derivations for Chapter IV

B.1 Proof of Theorem IV.1

First note that $\sum_i \mathbf{U}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) = 0$, $R_i(1-R_i) = 0$, $R_i^2 = R_i$ and $(1-R_i)^2 = 1-R_i$. To simplify notation, we suppress dependence of $\mathbf{X}_i(a_1, a_2)$ and $\mathbf{V}_i(a_1, a_2; \boldsymbol{\alpha}^*)$ on (a_1, a_2). Then, by definition of $\tilde{W}_i(a_1, a_2)$,

$$\begin{aligned}
&\mathbb{E}\left(\tilde{W}_i(a_1,a_2)\mathbf{X}_i^\intercal \mathbf{V}_i(\boldsymbol{\alpha}^*)^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}^*) \mid \mathbf{L}_i\right) \\
&= \mathbb{E}\left[\frac{\mathbb{I}\left[A_{1i} = a_1\right]}{\mathbb{P}\left(A_{1i} = a_1\right)}\left(R_i + \frac{\mathbb{I}\left[A_{2i} = a_2\right]}{\mathbb{P}\left(A_{2i} = a_2 \mid A_{1i} = a_1, R_i = 0\right)}(1-R_i)\right)\right. \\
&\qquad \left. \times \mathbf{X}_i^\intercal \mathbf{V}_i(\boldsymbol{\alpha}^*)^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}^*) \mid \mathbf{L}_i\right],
\end{aligned} \tag{B.1}$$

and, using consistency assumption (ii)

$$\begin{aligned}
&= \mathbb{E}\left(\frac{\mathbb{I}\left[A_{1i} = a_1\right]}{\mathbb{P}\left(A_{1i} = a_1\right)} R_i \mathbf{X}_i^\intercal \mathbf{V}_i(\boldsymbol{\alpha}^*)^{-1}(R_i\mathbf{Y}_i(A_{1i}) - \mathbf{X}_i\boldsymbol{\beta}^*) \mid \mathbf{L}_i\right) \\
&\quad + \mathbb{E}\left[\frac{\mathbb{I}\left[A_{1i} = a_1\right]}{\mathbb{P}\left(A_{1i} = a_1\right)} \frac{\mathbb{I}\left[A_{2i} = a_2\right]}{\mathbb{P}\left(A_{2i} = a_2 | A_{1i} = a_1, R_i = 0\right)}(1-R_i)\right. \\
&\qquad \left. \times \mathbf{X}_i^\intercal \mathbf{V}_i(\boldsymbol{\alpha}^*)^{-1}(\mathbf{Y}_i(A_{1i}, A_{2i}) - \mathbf{X}_i\boldsymbol{\beta}^*) \mid \mathbf{L}_i\right].
\end{aligned} \tag{B.2}$$

Next note that $\mathbb{I}\left[A_{2i}=a_2\right]\mathbf{Y}_i(A_{1i},A_{2i})=\mathbb{I}\left[A_{2i}=a_2\right]\mathbf{Y}_i(A_{1i},a_2)$ and, by assumption (iii), $A_{2i}\perp\!\!\!\perp\mathbf{Y}_i(a_1,a_2)\mid A_{1i},R_i$ for any fixed regime (a_1,a_2). Let

$$Q=\mathbb{P}\left(A_{1i}=a_1\right)^{-1}\mathbb{P}\left(A_{2i}=a_2\mid A_{1i}=a_1,R_i=0\right)^{-1}$$

. Then

$$\mathbb{E}\left(Q\mathbb{I}\left[A_{1i}=a_1\right]\mathbb{I}\left[A_{2i}=a_2\right](1-R_i)\mathbf{X}_i^\intercal\mathbf{V}_i(\boldsymbol{\alpha}^*)^{-1}(\mathbf{Y}_i(A_{1i},A_{2i})-\mathbf{X}_i\boldsymbol{\beta}^*)\mid\mathbf{L}_i\right) \tag{B.3}$$

$$=\mathbb{E}\left\{Q\mathbb{I}\left[A_{1i}=a_1\right]\mathbb{E}\left(\mathbb{I}\left[A_{2i}=a_2\right](1-R_i)\mathbf{X}_i^\intercal\mathbf{V}_i(\boldsymbol{\alpha}^*)^{-1}(\mathbf{Y}_i(A_{1i},A_{2i})-\mathbf{X}_i\boldsymbol{\beta}^*)\mid A_{1i},R_i\right)\mid\mathbf{L}_i\right\} \tag{B.4}$$

$$=\mathbb{E}\left\{Q\mathbb{I}\left[A_{1i}=a_1\right]\mathbb{E}\left(\mathbb{I}\left[A_{2i}=a_2\right](1-R_i)\mathbf{X}_i^\intercal\mathbf{V}_i(\boldsymbol{\alpha}^*)^{-1}(\mathbf{Y}_i(A_{1i},a_2)-\mathbf{X}_i\boldsymbol{\beta}^*)\mid A_{1i},R_i=0\right)\mid\mathbf{L}_i\right\} \tag{B.5}$$

$$=\mathbb{E}\left(\frac{\mathbb{I}\left[A_{1i}=a_1\right]}{\mathbb{P}\left(A_{1i}=a_1\right)}\mathbb{E}\left((1-R_i)\mathbf{X}_i^\intercal\mathbf{V}_i(\boldsymbol{\alpha}^*)^{-1}(\mathbf{Y}_i(A_{1i},a_2)-\mathbf{X}_i\boldsymbol{\beta}^*)\mid A_{1i},R_i=0\right)\mid\mathbf{L}_i\right) \tag{B.6}$$

Substituting into equation (B.2),

$$\begin{aligned}(\text{B.2})=&\mathbb{E}\left(\frac{\mathbb{I}\left[A_{1i}=a_1\right]}{\mathbb{P}\left(A_{1i}=a_1\right)}R_i\mathbf{X}_i^\intercal\mathbf{V}_i(\boldsymbol{\alpha}^*)^{-1}(R_i\mathbf{Y}_i(A_{1i})-\mathbf{X}_i\boldsymbol{\beta}^*)\mid\mathbf{L}_i\right)\\&+\mathbb{E}\left(\frac{\mathbb{I}\left[A_{1i}=a_1\right]}{\mathbb{P}\left(A_{1i}=a_1\right)}(1-R_i)\mathbf{X}_i^\intercal\mathbf{V}_i(\boldsymbol{\alpha}^*)^{-1}(\mathbf{Y}_i(A_{1i},a_2)-\mathbf{X}_i\boldsymbol{\beta}^*)\mid\mathbf{L}_i\right)\end{aligned} \tag{B.7}$$

$$=\mathbf{X}_i^\intercal\mathbf{V}_i(\boldsymbol{\alpha}^*)^{-1}\mathbb{E}\left(R_i(a_1)\mathbf{Y}_i(a_1)+(1-R_i(a_1))\mathbf{Y}_i(a_1,a_2)-\mathbf{X}_i\boldsymbol{\beta}^*\mid\mathbf{L}_i\right) \tag{B.8}$$

$$=\mathbf{0} \tag{B.9}$$

where (B.8) is obtained from the consistency assumption on R_i and independence of A_{1i} and $R(a_1)$. Thus $\mathbb{E}\left(\sum_i\mathbf{U}_i(\boldsymbol{\beta}^*,\boldsymbol{\alpha}^*)\right)=\mathbf{0}$. Under Assumption (v), we have that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of β^*. To derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}$, note that

$$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)=-\left(\frac{1}{N}\frac{d\sum_i\mathbf{U}_i(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}))}{d\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}+o_P(1)\right)^{-1}\frac{1}{\sqrt{N}}\sum_i\mathbf{U}_i(\boldsymbol{\beta}^*,\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}^*))$$

The result follows using similar arguments as those in the proof of Theorem 2 in Liang and Zeger (1986).

B.2 Generative model for simulations in Section 4.6

This section provides more detail about the generative model used in the simulation studies in Section 4.6.

The potential outcomes were generated from the following piecewise linear model:

$$
\begin{aligned}
Y_{it}(a_1, a_2) = \\
&\theta_0 + \mathbb{I}\left[t \leq \kappa\right] t(\theta_1 + \theta_2 a_1) + \mathbb{I}\left[t > \kappa\right] \kappa(\theta_1 + \theta_2 a_1) \\
&+ \mathbb{I}\left[t > \kappa\right] (t - \kappa)(\theta_3 + \theta_4 a_1 + (\theta_5 a_2 + \theta_6 a_1 a_2)(1 - R_i(a_1))) \\
&+ \mathbb{I}\left[t > \kappa\right] (t - \kappa)(\psi^{(1)}\mathbb{I}\left[a_1 = 1\right] + \psi^{(-1)}\mathbb{I}\left[a_1 = -1\right]) \left[R_i(a_1) - \mathbb{P}\left(R_i(a_1) = 1 \mid \mathbf{L}_i\right)\right] \\
&+ \theta_7 L_i + \gamma_{0i} + \gamma_{1i} t + \epsilon_{it},
\end{aligned}
\tag{B.10}
$$

where $R_i(a_1) = \mathbb{I}\left[Y_{i\kappa}(a_1) - \theta_7 L_i > c\right]$; $L_i = \mathbf{L}_i \in \{1, -1\}$; $c = 1.1$; $\theta_7 = -0.2$; $(\gamma_{0i}, \gamma_{1i})^\intercal \sim N(\mathbf{0}, \mathbf{\Gamma})$; $\epsilon_{it} \sim N(0, \tau^2)$ with $\tau^2 = 1$; $t \in \{0, 0.5, 1.5, 2, 2.25, 2.5, 3\}$; and $\kappa = 2$.

Under model (B.10),

$$
\begin{aligned}
&Y_{it}(a_1, a_2) - \mathbb{E}\left(Y_{it}(a_1, a_2) \mid \mathbf{L}_i\right) \\
&\quad = \gamma_0 + \gamma_1 t + \epsilon_t \\
&\quad + \mathbb{I}\left[t > \kappa\right] (t - \kappa)\theta_5 a_2 \left[(1 - R_i(a_1)) - \mathbb{P}\left(R_i(a_1) = 0 \mid \mathbf{L}_i\right)\right] \\
&\quad + \mathbb{I}\left[t > \kappa\right] (t - \kappa)\theta_6 a_1 a_2 \left[(1 - R_i(a_1)) - \mathbb{P}\left(R_i(a_1) = 0 \mid \mathbf{L}_i\right)\right] \\
&\quad + \mathbb{I}\left[t > \kappa\right] (t - \kappa) \left(\psi^{(1)}\mathbb{I}\left[a_1 = 1\right] + \psi^{(-1)}\mathbb{I}\left[a_1 = -1\right]\right) \left[R_i(a_1) - \mathbb{P}\left(R_i(a_1) = 1 \mid \mathbf{L}_i\right)\right],
\end{aligned}
\tag{B.11}
$$

and we can parameterize this marginal mean model as follows:

$$\begin{aligned}\mathbb{E}\left(Y_{it}(a_1,a_2)\mid \mathbf{L}_i\right) = \boldsymbol{\beta}^\intercal \mathbf{X}_{it}(a_1,a_2) = \beta_0 &+ \mathbb{I}\left[t\le\kappa\right] t(\beta_1+\beta_2 a_1) + \mathbb{I}\left[t>\kappa\right]\kappa(\beta_1+\beta_2 a_1)\\ &+ \mathbb{I}\left[t>\kappa\right](t-\kappa)(\beta_3+\beta_4 a_1+\beta_5 a_2+\beta_6 a_1 a_2)\\ &+ \beta_7 L_i,\end{aligned} \tag{B.12}$$

where $\beta_j = \theta_j$ for $j \in \{0,1,2,3,4,7\}$,

$$\begin{aligned}\beta_5 &= \left\{\theta_5\left(\frac{\pi^{(1)}}{2}+\frac{\pi^{(-1)}}{2}\right)+\theta_6\left(\frac{\pi^{(1)}}{2}-\frac{\pi^{(-1)}}{2}\right)\right\},\\ \beta_6 &= \left\{\theta_5\left(\frac{\pi^{(1)}}{2}-\frac{\pi^{(-1)}}{2}\right)+\theta_6\left(\frac{\pi^{(1)}}{2}+\frac{\pi^{(-1)}}{2}\right)\right\},\end{aligned}$$

and $\pi^{(a_1)} := \mathbb{P}\left(R_i(a_1)=0\mid \mathbf{L}_i\right)$.

Next, we derive the marginal covariance and variance of the repeated measures outcomes under this generative model. These marginal covariances and variances are used to calculate the population standardized effect size

$$d = \frac{\mathbb{E}\left(Y_{i3}(1,-1)\mid \mathbf{L}_i\right) - \mathbb{E}\left(Y_{i3}(-1,-1)\mid \mathbf{L}_i\right)}{\sqrt{\frac{1}{2}\mathrm{Var}\left(Y_{i3}(1,-1)\mid \mathbf{L}_i\right)+\frac{1}{2}\mathrm{Var}\left(Y_{i3}(-1,-1)\mid \mathbf{L}_i\right)}}.$$

Let $W_{it} = \mathbf{z}_{it}^\intercal\boldsymbol{\gamma}+\epsilon_{it}$ and $\mathbf{z}_{it} = (1,t)^\intercal$. Then

$$\begin{aligned}&\mathrm{Cov}\left[Y_{it}(a_1,a_2), Y_{is}(a_1,a_2)\mid \mathbf{L}_i\right]\\ &= \mathbf{z}_{it}^\intercal\boldsymbol{\Gamma}\mathbf{z}_{is} + \tau^2\mathbb{I}\left[s=t\right]\\ &\quad - \left(C_1^{(a_1)}(s)+C_2^{(a_1,a_2)}(s)-C_3^{(a_1)}(s)\right)(1-\pi^{(a_1)})\mathbb{E}\left(W_{it}\mid R_i(a_1)=1,\mathbf{L}_i\right)\\ &\quad - \left(C_1^{(a_1)}(t)+C_2^{(a_1,a_2)}(t)-C_3^{(a_1)}(t)\right)(1-\pi^{(a_1)})\mathbb{E}\left(W_{is}\mid R_i(a_1)=1,\mathbf{L}_i\right)\\ &\quad + \left(C_1^{(a_1)}(t)+C_2^{(a_1,a_2)}(t)-C_3^{(a_1)}(t)\right)\left(C_1^{(a_1)}(s)+C_2^{(a_1,a_2)}(s)-C_3^{(a_1)}(s)\right)\pi^{(a_1)}(1-\pi^{(a_1)})\end{aligned} \tag{B.13}$$

where

$$
\begin{aligned}
C_1^{(a_2)}(t) &= \mathbb{I}\left[t > \kappa\right](t-\kappa)\theta_5 a_2 \\
C_2^{(a_1,a_2)}(t) &= \mathbb{I}\left[t > \kappa\right](t-\kappa)\theta_6 a_1 a_2 \\
C_3^{(a_1)}(t) &= \mathbb{I}\left[t > \kappa\right](t-\kappa)(\psi^{(1)}\mathbb{I}\left[a_1 = 1\right] + \psi^{(-1)}\mathbb{I}\left[a_1 = -1\right]).
\end{aligned}
$$

Note that

$$
\mathbb{E}\left(W_{it} \mid R_i(a_1) = 1, \mathbf{L}_i\right) = \mathbb{E}\left(W_{it} \mid W_{i\kappa} > c - \theta_0 - \kappa(\theta_1 + \theta_2 a_1), \mathbf{L}_i\right),
$$

$$
(W_{it}, W_{i\kappa})^\intercal \mid \mathbf{L}_i \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{z}_{it}^\intercal \mathbf{\Gamma} \mathbf{z}_{it} + \tau^2, & \mathbf{z}_{it}^\intercal \mathbf{\Gamma} \mathbf{z}_{i\kappa} + \mathrm{Cov}(\epsilon_{it}, \epsilon_{i\kappa} \mid \mathbf{L}_i) \\ \cdots & \mathbf{z}_{i\kappa}^\intercal \mathbf{\Gamma} \mathbf{z}_{i\kappa} + \tau^2 \end{bmatrix}\right),
$$

and since $(W_{it}, W_{i\kappa})^\intercal | \mathbf{L}_i$ is bivariate Gaussian, $\mathbb{E}\left(W_{it} \mid W_{i\kappa} > c - \theta_0 - \kappa(\theta_1 + \theta_2 a_1), \mathbf{L}_i\right)$ can be computed using the truncated multivariate Gaussian distribution.

BIBLIOGRAPHY

BIBLIOGRAPHY

Almirall, D., Nahum-Shani, I., Sherwood, N., and Murphy, S., 2014, "Introduction to SMART Designs for the Development of Adaptive Interventions: with application to weight loss research", *Translational Behavioral Medicine* 4:260–274.

Almirall, D., DiStefano, C., Chang, Y.-C., Shire, S., Kaiser, A., Lu, X., Nahum-Shani, I., Landa, R., Mathy, P., and Kasari, C., 2016, "Longitudinal effects of adaptive interventions with a speech-generating device in minimally verbal children with ASD", *Journal of Clinical Child and Adolescent Psychology* 45:442–456.

August, G. J., Piehler, T. F., and Bloomquist, M. L., 2016, "Being 'SMART' about adolescent conduct problems prevention: executing a SMART pilot study in a juvenile diversion agency", *Journal of Clinical Child and Adolescent Psychology* 45:495–509.

Austin, P. C., Tu, J. V., Ho, J. E., Levy, D., and Lee, D. S., 2013, "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes", *Journal of Clinical Epidemiology* 66:398–407.

Bachoc, C., Gijswijt, D. C., Schrijver, A., and Vallentin, F., 2012, "Invariant Semidefinite Programs". In *Handbook on Semidefinite, Conic and Polynomial Optimization*, 219–269, Boston, MA: Springer US.

Barber, R. F. and Candès, E. J., 2015, "Controlling the false discovery rate via knockoffs", *The Annals of Statistics* 43:2055–2085.

Barber, R. F. and Candès, E. J., 2019, "A knockoff filter for high-dimensional selective inference", *The Annals of Statistics* 47:2504–2537.

Barber, R. F., Candès, E. J., and Samworth, R. J., 2018, "Robust inference with knockoffs", *arXiv e-prints*, arXiv: `1801.03896`.

Bates, D., Mächler, M., Bolker, B., and Walker, S., 2015, "Fitting linear mixed-effects models using lme4", *Journal of Statistical Software* 67:1–48.

Benjamini, Y. and Hochberg, Y., 1995, "Controlling the false discovery rate: a practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society, Series B* 57:289–300.

Benjamini, Y. and Yekutieli, D., 2001, "The control of the false discovery rate in multiple testing under dependency", *The Annals of Statistics* 29:1165–1188.

Bhatia, R., 1997, *Matrix Analysis*, vol. 169, Graduate Texts in Mathematics, New York: Springer Science & Business Media.

Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., and Candès, E. J., 2015, "SLOPE—adaptive variable selection via convex optimization", *Annals of Applied Statistics* 9:1103.

Boyd, S. and Vandenberghe, L., 2004, *Convex Optimization*, Cambridge: Cambridge University Press.

Brumback, B. A., 2009, "A note on using the estimated versus the known propensity score to estimate the average treatment effect", *Statistics and Probability Letters* 79:537–542.

Candès, E. J., Fan, Y., Janson, L., and Lv, J., 2018, "Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection", *Journal of the Royal Statistical Society, Series B* 80:551–577.

Carlin, B. P. and Louis, T. A., 2000, "Empirical Bayes: Past, present and future", *Journal of the American Statistical Association* 95:1286–1289.

Chakraborty, B. and Moodie, E., 2013, *Statistical Methods for Dynamic Treatment Regimes*, Statistics for Biology and Health, New York: Springer.

Chikuse, Y., 2012, *Statistics on Special Manifolds*, vol. 174, New York: Springer Science & Business Media.

Choi, E., Schuetz, A., Stewart, W. F., and Sun, J., 2017, "Using recurrent neural network models for early detection of heart failure onset", *Journal of the American Medical Informatics Association* 24:361–370.

Dai, T. and Shete, S., 2016, "Time-varying SMART design and data analysis methods for evaluating adaptive intervention effects", *BMC Medical Research Methodology* 16:112.

Dawson, R. and Lavori, P. W., 2008, "Sequential causal inference: application to randomized trials of adaptive treatment strategies", *Statistics in Medicine* 27:1626–1645.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S., 2002, *Analysis of Longitudinal Data*, Oxford: Oxford University Press.

Dziak, J. J., Yap, J. R. T., Almirall, D., McKay, J. R., Lynch, K. G., and Nahum-Shani, I., 2019, "A Data Analysis Method for Using Longitudinal Binary Outcome Data from a SMART to Compare Adaptive Interventions", *Multivariate Behavioral Research* 0:1–24.

Efron, B., 2010, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Institute of Mathematical Statistics Monographs, Cambridge: Cambridge University Press.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H., 2011, *Applied Longitudinal Analysis*, Hoboken, New Jersey: John Wiley & Sons.

Gelman, A. and Carlin, J., 2014, "Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors", *Perspectives on Psychological Science* 9:641–651.

Genovese, C. and Wasserman, L., 2004, "A stochastic process approach to false discovery control", *The Annals of Statistics* 32:1035–1061.

Gibbons, R. D., Hedeker, D., and DuToit, S., 2010, "Advances in analysis of longitudinal data", *Annual Review of Clinical Psychology* 6:79–107.

Goldstein, H., 2011, *Multilevel Statistical Models*, Chichester, West Sussex: John Wiley & Sons.

Gunlicks-Stoessel, M., Mufson, L., Westervelt, A., Almirall, D., and Murphy, S., 2016, "A pilot SMART for developing an adaptive treatment strategy for adolescent depression", *Journal of Clinical Child and Adolescent Psychology* 45:480–494.

Hedeker, D. and Gibbons, R. D., 2006, *Longitudinal Data Analysis*, Hoboken, New Jersey: John Wiley & Sons.

Hernán, M., Lanoy, E., Costagliola, D., and Robins, J., 2006, "Comparison of Dynamic Treatment Regimes via Inverse Probability Weighting", *Basic and Clinical Pharmacology and Toxicology* 98:237–242.

Hirano, K., Imbens, G. W., and Ridder, G., 2003, "Efficient estimation of average treatment effects using the estimated propensity score", *Econometrica* 71:1161–1189.

Hsu, J. Y. and Wahed, A. S., 2017, "Weighted generalized estimating equations for response-adaptive treatment regimes in two-stage longitudinal studies", *Journal of Statistical Research* 51:79–100.

IBM Watson Health, 2018, *IBM MarketScan Research Databases for Health Services Researchers*, tech. rep., `https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=HPW03041USEN`, Accessed January 11, 2019, Somers, NY: IBM.

Jensen, P. B., Jensen, L. J., and Brunak, S., 2012, "Mining electronic health records: towards better research applications and clinical care", *Nature Reviews Genetics* 13:395.

Kaneko, T., Fiori, S., and Tanaka, T., 2012, "Empirical arithmetic averaging over the compact Stiefel manifold", *IEEE Transactions on Signal Processing* 61:883–894.

Kasari, C., Kaiser, A., Goods, K., Nietfeld, J., Mathy, P., Landa, R., Murphy, S., and Almirall, D., 2014, "Communication interventions for minimally verbal children with autism: a sequential multiple assignment randomized trial", *Journal of the American Academy of Child and Adolescent Psychiatry* 53:635–646.

Kidwell, K., 2014, "SMART designs in cancer research: past, present and future", *Clinical Trials* 11:445–456.

Kosorok, M. R. and Moodie, E. E., eds., 2016, *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*, ASA-SIAM Series on Statistics and Applied Probability, Philadelphia: SIAM.

Lavori, P. and Dawson, D., 2000, "A design for testing clinical strategies: biased individually tailored within-subject randomization", *Journal of the Royal Statistical Society, Series A* 163:29–38.

Lavori, P. W. and Dawson, R., 2014, "Introduction to dynamic treatment strategies and sequential multiple assignment randomization", *Clinical Trials* 11:393–399.

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., and Murphy, S., 2012, "A SMART design for building individualized treatment sequences", *Annual Review of Clinical Psychology* 8:21–48.

Li, Z. and Murphy, S. A., 2011, "Sample size formulae for two-stage randomized trials with survival outcomes", *Biometrika* 98:503–518.

Liang, K.-Y. and Zeger, S. L., 1986, "Longitudinal data analysis using generalized linear models", *Biometrika* 73:13–22.

Lu, X., Nahum-Shani, I., Kasari, C., Lynch, K. G., Oslin, D. W., Pelham, W. E., Fabiano, G., and Almirall, D., 2016, "Comparing dynamic treatment regimes using repeated-measures outcomes: modeling considerations in SMART studies", *Statistics in Medicine* 35:1595–1615.

Luers, B., Qian, M., Nahum-Shani, I., Kasari, C., and Almirall, D., 2019, "Linear Mixed Models for Comparing Dynamic Treatment Regimens on a Longitudinal Outcome in Sequentially Randomized Trials", *arXiv e-prints*, arXiv: `1910.10078`.

Meinshausen, N., Meier, L., and Bühlmann, P., 2009, "P-values for high-dimensional regression", *Journal of the American Statistical Association* 104:1671–1681.

Miyahara, S. and Wahed, A. S., 2012, "Assessing the effect of treatment regimes on longitudinal outcome data: Application to revamp study of depression", *Journal of Statistical Research* 46:233–254.

Molenberghs, G. and Kenward, M., 2007, *Missing Data in Clinical Studies*, Chichester, West Sussex: John Wiley & Sons.

Murphy, S., Lynch, K., Oslin, D., McKay, J., and Tenhave, T., 2007, "Developing adaptive treatment strategies in substance abuse research", *Drug and Alcohol Dependence* 88:S24–S30.

Murphy, S. A., 2005, "An experimental design for the development of adaptive treatment strategies", *Statistics in Medicine* 24:1455–1481.

Murphy, S. A., Laan, M. J. van der, Robins, J. M., and CPPRG, 2001, "Marginal Mean Models for Dynamic Regimes", *Journal of the American Statistical Association* 96:1410–1423.

Naar-King, S., Ellis, D. A., Idalski Carcone, A., Templin, T., Jacques-Tiura, A. J., Brogan Hartlieb, K., Cunningham, P., and Jen, K.-L. C., 2016, "Sequential multiple assignment randomized trial (SMART) to construct weight loss interventions for African American adolescents", *Journal of Clinical Child and Adolescent Psychology* 45:428–441.

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W., Gnagy, B., Fabiano, G., Waxmonsky, J., Yu, J., and Murphy, S., 2012, "Experimental design and primary data analysis methods for comparing adaptive interventions", *Psychological Methods* 17:457–477.

NeCamp, T., Kilbourne, A., and Almirall, D., 2017, "Comparing cluster-level dynamic treatment regimens using sequential, multiple assignment, randomized trials: regression estimation and sample size considerations", *Statistical Methods in Medical Research* 26:1572–1589.

Orellana, L., Rotnitzky, A., and Robins, J., 2010, "Dynamic regime marginal structural mean models for estimating optimal dynamic treatment regimes, Part I: Main content", *International Journal of Biostatistics* 6:Article 8.

Pyne, S., Futcher, B., and Skiena, S., 2006, "Meta-analysis based on control of false discovery rate: combining yeast ChIP-chip datasets", *Bioinformatics* 22:2516–2522.

Raudenbush, S. W. and Bryk, A. S., 2002, *Hierarchical Linear Models: Applications and Data Analysis Methods*, vol. 1, Advanced Quantitative Techniques in the Social Sciences, Thousand Oaks: Sage.

Robinson, G. K., 1991, "That BLUP is a good thing: the estimation of random effects", *Statistical Science* 6:15–32.

Romano, Y., Sesia, M., and Candès, E., 2019, "Deep Knockoffs", *Journal of the American Statistical Association* 0:1–12.

Roquero Gimenez, J. and Zou, J., 2018, "Improving the stability of the nockoff procedure: multiple simultaneous knockoffs and entropy maximization", *arXiv e-prints*, arXiv: 1810.11378.

Searle, S. R., Casella, G., and McCulloch, C. E., 2006, *Variance Components*, Hoboken, New Jersey: John Wiley & Sons.

Seewald, N. J., Kidwell, K. M., Nahum-Shani, I., Wu, T., McKay, J. R., and Almirall, D., 2019, "Sample size considerations for comparing dynamic treatment regimens in a sequential multiple-assignment randomized trial with a continuous longitudinal outcome", *Statistical Methods in Medical Research*, `https://doi.org/10.1177/0962280219877520`.

Sesia, M., Sabatti, C., and Candès, E. J., 2018, "Gene hunting with hidden Markov model knockoffs", *Biometrika* 106:1–18.

Shaffer, J. P., 1995, "Multiple hypothesis testing", *Annual review of psychology* 46:561–584.

Skrondal, A. and Rabe-Hesketh, S., 2009, "Prediction in multilevel generalized linear models", *Journal of the Royal Statistical Society, Series A* 172:659–687.

Snijders, T. and Bosker, R., 2012, *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, London: Sage.

Storey, J. D., 2002, "A direct approach to false discovery rates", *Journal of the Royal Statistical Society, Series B* 64:479–498.

Storey, J. D., Taylor, J. E., and Siegmund, D., 2004, "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach", *Journal of the Royal Statistical Society, Series B* 66:187–205.

Su, W., Qian, J., and Liu, L., 2015, "Communication-Efficient False Discovery Rate Control via Knockoff Aggregation", *arXiv e-prints*, arXiv: `1506.05446`.

Verbeke, G. and Molenberghs, G., 2009, *Linear mixed models for longitudinal data*, New York: Springer Science & Business Media.

Wallace, M. P., Moodie, E. E., and Stephens, D. A., 2016, "SMART thinking: a review of recent developments in sequential multiple assignment randomized trials", *Current Epidemiology Reports* 3:225–232.

Weiskopf, N. G. and Weng, C., 2013, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research", *Journal of the American Medical Informatics Association* 20:144–151.

Williamson, E., Forbes, A., and White, I., 2014, "Variance reduction in randomised trials by inverse probability weighting using the propensity score", *Statistics in Medicine* 33:721–737.

Wu, J., Roy, J., and Stewart, W. F., 2010, "Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches", *Medical care* 48:S106–S113.

Wu, Y., Boos, D. D., and Stefanski, L. A., 2007, "Controlling variable selection by the addition of pseudovariables", *Journal of the American Statistical Association* 102:235–243.