Model-free high-dimensional false discovery rate control with knockoffs

Emmanuel Candès, Yingying Fan, Lucas Janson, Jinchi Lv

Abstract

A recent work by Barber and Candès introduced the knockoffs variable selection framework for controlling the false discovery rate (FDR) in low-dimensional ($n \geq p$) linear regression. In the present paper we extend the framework to arbitrary (and unknown) regression models and dimensions, including $n < p$. This extension requires the design matrix be random (independent and identically distributed rows) with a covariate distribution that is known, for example from a large set of unlabeled covariate observations. The requirement of p-values as inputs to any FDR-controlling procedure combined with the challenges of obtaining p-values for general models, especially in high dimensions, makes the problem addressed in this paper a largely open one. In the few settings where competitors exist, we demonstrate the superior power of knockoffs through simulations. Also, through simulations using a real design matrix from genetics, we show that even when the covariate distribution is unknown and estimated in-sample, knockoffs still effectively controls the FDR. Finally, we apply the new procedure to a case-control study of Type-II diabetes, discovering mutations that have been replicated by outside experiments.

1 Introduction

This paper seeks to solve the problem, ubiquitous in modern statistical applications, of selecting from a large number of candidate variables the ones that are truly associated with some outcome of interest. These applications run the gamut from fundamental medical/biological problems of finding which traits (genetic or otherwise) are associated with a disease, to social science problems of finding which characteristics are associated with economic success or political views, to industry problems of finding what attributes are associated with certain user/client behaviors.

1.1 Problem Statement

We consider a very general regression setting where the responses y_i can depend in an arbitrary way on the covariate vectors $\boldsymbol{x}_i \in \mathbb{R}^p$. The only restriction we place on the model is that the observations $(y_i, \boldsymbol{x}_i)$ are independently and identically distributed (i.i.d.), which is often realistic in high-dimensional applications such as genetics, where subjects may be drawn randomly from some large population, or client behavioral modeling, where experiments on a service or user interface go out to a random subset of users. Therefore the model is simply

$$(y_i, \boldsymbol{x}_i) \overset{i.i.d.}{\sim} F, \tag{1.1}$$

for some arbitrary $(p+1)$-dimensional distribution F. For notational convenience we will often work with the vectorized versions of the data: $\boldsymbol{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p) \in \mathbb{R}^{n \times p}$.

We will assume *no knowledge* of the conditional distribution of $y_i|\boldsymbol{x}_i$, but we do assume the joint distribution of the covariates is known. This is clearly a strong assumption, and merits the following points:

- The typical setup for regression inference is to assume a strong parametric model for the response conditional on the covariates, such as a homoscedastic linear model, but to assume as little as possible about, or even condition on, the covariates. In practice such a model is nearly always just an approximation of the truth, with the quality of that approximation varying by domain. Analogously, we will show in Section 3 that a good approximation of the covariate distribution is often sufficient for accurate inference.
- We do not claim our assumptions will always be appropriate, but there are important cases when it is reasonable to think we know much more about the covariate distribution than about the conditional distribution of the response, including:
  - When we in fact know exactly the covariate distribution because we control it, such as in gene knockout experiments or sensitivity analysis of numerical models (for example climate models).

  - When we have a large amount of unsupervised data (covariate data without corresponding responses/labels) in addition to the n labeled observations $\boldsymbol{y}$, $\boldsymbol{X}$. This is not uncommon in genetic or economic studies, where many other studies will exist that have collected the same covariate information but different response variables.
  - When we simply have considerably more prior information about the covariates than about the response. Indeed, the point of many regression problems is to relate a poorly-understood response variable to a set of well-understood covariates. For instance, in genetic case-control studies, scientists seek to understand the causes of an extremely biologically-complex disease using many comparatively simple single nucleotide polymorphisms (SNPs) as covariates.

- There are substantial payoffs to our framework. Particularly in high dimensions, previous regression inference results rely not only on a parametric model that is often linear and homoscedastic, but also on the sparsity or ultra-sparsity of the parameters of that model in order to achieve some asymptotic guarantee. In contrast, our framework can accomodate *any* model for both the response and the covariates, and our guarantees are exact in finite samples (non-asymptotic).

Our setup clearly encompasses any regression or classification model, including any generalized linear model (GLM), but also allows for arbitrary nonlinearities and heteroscedasticity such as are found in many machine learning applications.

This paper is about the variable-selection problem, namely, answering the question: which covariates (columns of $\boldsymbol{X}$) are important for determining the response $\boldsymbol{y}$? Explicitly, a covariate is defined as unimportant if it is conditionally independent of the response given the other covariates. In particular, we aim to select a subset $S \subset \{1, \dots, p\}$ of important variables in a way that controls the false discovery rate (FDR) at some prespecified level q:

$$\text{FDR} := \mathbb{E}\left(\frac{|\{j \in S\colon \boldsymbol{X}_j \perp\!\!\!\perp \boldsymbol{y} | \boldsymbol{X}_{\text{-}j}\}|}{|S|}\right) \leq q, \tag{1.2}$$

where $\boldsymbol{X}_{\text{-}j}$ denotes all columns of $\boldsymbol{X}$ except the jth one. Note that in the common setting of a generalized linear model, $\boldsymbol{X}_j \perp\!\!\!\perp \boldsymbol{y} \mid \boldsymbol{X}_{\text{-}j}$ is equivalent to the jth entry of the coefficient vector $\boldsymbol{\beta}$ being equal to zero, so in this case our problem reduces to the usual variable selection problem of finding $j : \beta_j \neq 0$.

1.2 Related Work

This work builds on that of Barber and Candès (2015), who originally introduced the knockoffs framework. Their salient idea was to construct a set of so-called 'knockoff' variables which were not associated with the response, but whose structure mirrored that of the original covariates. These knockoff variables could then be used as controls for the real covariates, so that only real covariates which appeared considerably more associated with the response than their knockoff counterpart were selected. We will cover more of their technical details as preliminaries to our own results in Section 2, but their main result was achieving exact finite-sample FDR control, conditional on $\boldsymbol{X}$ (so no random design assumption), in the Gaussian linear regression model when $n \geq 2p$, along with a nearly-exact extension to when $p \leq n < 2p$. The random design assumption in our work allows us to extend Barber and Candès (2015) to arbitrary models and remove the low-dimensional constraint.

The concept of FDR was introduced in Benjamini and Hochberg (1995), along with the BHq procedure which can control the FDR among a set of hypothesis tests when given a set of independent p-values for those hypotheses. Benjamini and Yekutieli (2001) later showed that BHq controls the FDR even under a form of positive dependence called PRDS among the p-values, and also proposed a more conservative procedure that controls the FDR under arbitrary dependence. Various ideas including empirical Bayes procedures (see for example Efron and Tibshirani (2002)) and p-value weighting (see for example Genovese et al. (2006)) have been proposed to enhance power in certain settings. One thing these procedures have in common is that they all act on a set of p-values (or equivalent statistics). One exception is the SLOPE procedure (Bogdan et al., 2015), which acts like a regression analogue of BHq without ever computing p-values, but only controls the FDR in low-dimensional linear regression when the design matrix has orthogonal columns.

The requirement of computing valid p-values is quite constraining for general regression problems. In low-dimensional ($n \geq p$) Gaussian linear regression, p-values can be computed exactly even if the error variance is unknown, although the p-values will not in general have any simple dependence properties like independence or PRDS. Already for just the slightly-broader class of low-dimensional GLMs, one must resort to asymptotic p-values derived from maximum-likelihood theory, which we will show in Section 3.1 can be far from valid in practice. In high-dimensional ($n < p$) GLMs, it is not clear how to get p-values at all. Although some work (see for example van de Geer et al. (2014)) exists on computing asymptotic p-values under strong sparsity assumptions, these methods

also suffer from highly non-uniform null p-values in many finite-sample problems. Outside of linear models, there are feature importance measures, but not really any p-vals (RF p-vals are heuristic and do not satisfy the usual definition of a valid p-value). For binary treatment variables, the causal inference literature uses matching and propensity scores for approximately valid inference, but scaling these methods up to high dimensions is still a topic of current research similar in approximation and required assumptions to the aforementioned high-dimensional GLM literature.

However, as noted before, most previous works do not take the covariate distribution G to be known. With this assumption, a simple method for obtaining p-values for each $\boldsymbol{X}_j$, similar in spirit to both propensity scoring (where the conditional distribution of $\boldsymbol{X}_j$ given $\boldsymbol{X}_{-j}$ is estimated) and randomization/permutation tests (where $\boldsymbol{X}_j$ is either the only covariate or fully independent of $\boldsymbol{X}_{-j}$), exists. Explicitly, a conditional randomization test for the jth variable proceeds by first computing some feature importance statistic T_j for the jth variable. Then the null distribution of T_j can be computed through simulation by independently sampling $\boldsymbol{X}_j^*$'s from the *conditional* distribution of $\boldsymbol{X}_j$ given $\boldsymbol{X}_{-j}$ (derived from the known G) and recomputing the same statistic T_j^* with each resampled $\boldsymbol{X}_j^*$ instead of $\boldsymbol{X}_j$. Despite its simplicity, we have not seen this test proposed previously in the literature, although it nearly matches the usual randomization test when the covariates are independent. A disadvantage of this method is its computational cost, as in order to have power after any multiple-testing correction, the p-values must have resolution on the order of p^{-1}, requiring on the order of p samples for each of the p covariates, so that naïvely, the method scales computationally as p^2. However the most powerful feature importance statistics will take into account the full dimensionality of the model, e.g., $|\hat{\beta}_j|$, where $\hat{\boldsymbol{\beta}}$ is the coefficient vector resulting from an ℓ_1-penalized (lasso) maximum likelihood optimization. For Gaussian linear models, the lasso computation time scales as p^2, so that applying this conditional randomization test with the lasso coefficient statistic for all variables would take order of p^4 time, which will often be prohibitive in high dimensions. We will see in Section 3 that knockoffs achieves very similar power with drastically less required computation time, making it scalable to high dimensional problems.

We note here another line of work on *selective* inference for high-dimensional regression, wherein variables are considered null or not null only with respect to a (random) selected low-dimensional submodel. In particular, the work of Lockhart et al. (2014) selects a model sequentially, and at each step the null hypothesis for a given unselected variable is that it is uncorrelated with the response *conditional* on the selected variables. In Lee et al. (2016), the Lasso is used for selection and then for each selected variable, p-values for the null hypothesis that it has a zero coefficient in the *selected submodel* are computed. The knockoffs framework has also been applied to selective inference in Barber and Candes (2016), again with a screening step to first select $\leq n$ variables and then a selection step which controls the FDR with respect to the *screened submodel*. The key difference between these works and ours is that our inferential guarantees are given with respect to the (nonrandom) *full* model, and thus are not conditional on the quality or success of a random outside screening step. To elucidate the distinction, consider a GLM in which $\beta_1 = 1$ (perhaps $\boldsymbol{X}_1$ is an important causal variable) and $\beta_2 = 0$ but $\boldsymbol{X}_2$ is correlated to $\boldsymbol{X}_1$ (perhaps $\boldsymbol{X}_2$ is caused by $\boldsymbol{X}_1$ but is unrelated to the response). If a selective inference procedure's low-dimensional submodel accidentally selects β_2 instead of β_1 (a distinct possibility if the two are sufficiently correlated or n is not too large), then it would consider β_2 a non-null variable and hence its discovery would incur no penalty, while in the full model β_2 is always considered null and would be penalized accordingly in hopes of correctly selecting β_1.

1.3 Outline of the Paper

The remainder of the paper is structured as follows:

- Section 2 explains the knockoffs framework and our contribution to it, including proving FDR control and proposing a new knockoff construction and statistic.
- Section 3 demonstrates through simulations that knockoffs controls the FDR in a number of settings where no other procedure does, and then when competitors exist, knockoffs is more powerful. Using a design matrix from a real genetics data set, we also show that knockoffs' FDR control is surprisingly robust to random design distribution estimation error, making it useful even when the design distribution is unknown.
- Section 4 concludes the paper with extensions and potential lines of future research.

2 Methodology

...

2.1 Preliminaries

...

2.2 High-Dimensional Nonparametric Knockoffs

...

3 Numerical Experiments

In this section we demonstrate the importance, utility, and practicality of knockoffs for high-dimensional nonparametric regression. We start by showing that even in logistic regression when $n \gtrapprox p$, one cannot obtain valid p-values. Next we demonstrate the power of knockoffs as compared to what few alternatives exist in various settings. Lastly, we show on a real design matrix with simulated parameters that one does not necessarily need to know the exact distribution of $\boldsymbol{X}$ in order for knockoffs to control the FDR.

3.1 Logistic Regression P-Values

When $n \gg p$, asymptotic theory promises valid p-values for each coefficient in a GLM. However, these approximate p-values can always be computed as long as $n > p$, so a natural question arising from high-dimensional applications is whether such asymptotic p-values are valid when $n \gtrapprox p$ instead of $\gg p$. We simulated 10^4 independent design matrices ($n = 500$, $p = 200$) and binary responses from a logistic regression for the following two settings:

(1) $\boldsymbol{x}_i$ are each independent length-p AR(1) time series with AR coefficient 0.5, $y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(0.5)$

(2) $\boldsymbol{x}_i$ are each independent length-p AR(1) time series with AR coefficient 0.5, $y_i \overset{\perp\!\!\!\perp}{\sim} \text{Bernoulli}\left(\text{logit}\left(0.08(x_{i,2} + \cdots + x_{i,21})\right)\right)$

Histograms for the p-values for β_1 are shown in Figure 1. None of the histograms is anywhere close to uniform, and Table 1 shows each distribution's concentration near zero. We see that the small quantiles have extremely inflated probabilities—over 20 times nominal for $\mathbb{P}\{p\text{-value} \leq 0.1\%\}$ in setting (2). We also see that the exact null distribution depends on the (unknown) rest of the coefficient vector $\beta_2, \ldots, \beta_p$, since the probabilities between settings differ statistically significantly at all three cutoffs.

	(1)	(2)
$\mathbb{P}\{p\text{-value} \leq 5\%\}$	16.89% (0.37%)	19.17% (0.39%)
$\mathbb{P}\{p\text{-value} \leq 1\%\}$	6.78% (0.25%)	8.49% (0.28%)
$\mathbb{P}\{p\text{-value} \leq 0.1\%\}$	1.53% (0.12%)	2.27% (0.15%)

Table 1: Inflated p-value probabilities with estimated Monte Carlo standard errors in parentheses.

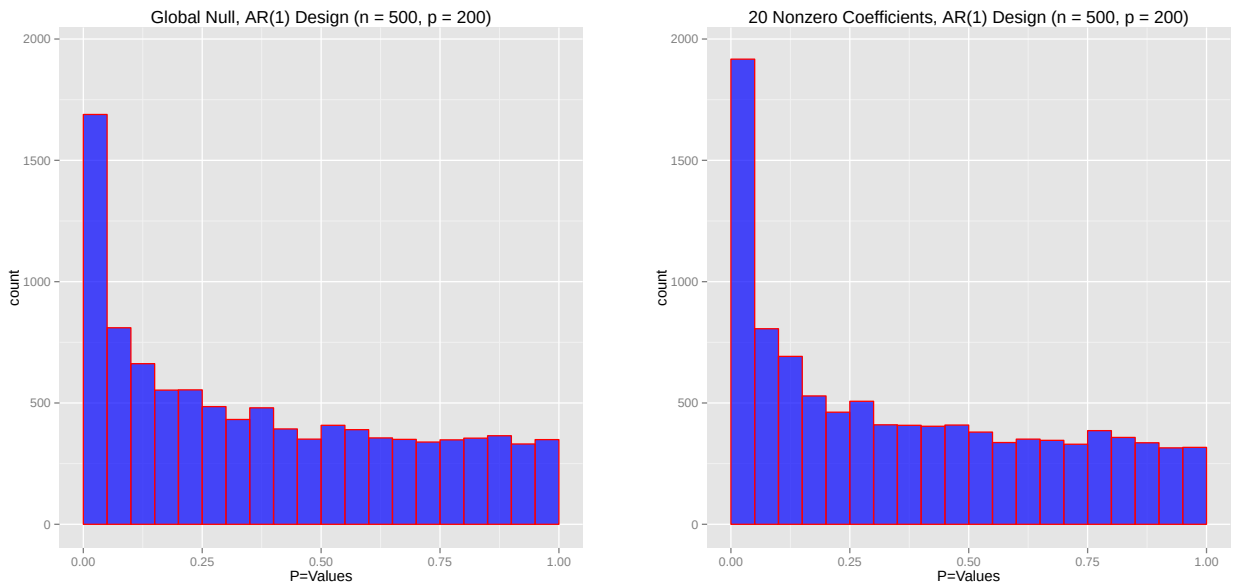
These results show that the usual logistic regression p-values one might use when $n \geq p$ can have null distributions that are quite far from uniform, and even if one wanted to try to correct that distribution, it depends in general on unknown problem parameters, further complicating matters. When $n < p$ the problem becomes even more challenging, with existing methods also asymptotic with stringent sparsity assumptions as well (van de Geer et al., 2014). Thus, despite the wealth of research on controlling FDR, without a way to obtain valid p-values, even the problem of controlling FDR in medium-to-high-dimensional GLMs remains unsolved.

3.2 Alternative Knockoff Statistics

As mentioned in Section 2, the new random knockoffs framework allows for a wider variety of W statistics to be used than in the original knockoffs. We discuss some appealing new options for statistics here.

3.2.1 General Feature Importance Statistics

With the sufficiency condition of the original knockoffs gone, we can come up with a generic formula for generating W statistics:



(a) (b)

Figure 1: Distribution of null logistic regression p-values with $n = 500$ and $p = 200$; 10,000 replications.

(1) Choose a feature importance measure Z and compute it on all original and knockoff variables together to generate $Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p$. Z_j and $\tilde{Z}_j$ must be invariant to swapping any set of pairs of original and knockoff variables except the jth pair.

(2) Choose a symmetric function w of two variables and set $W_j = w(Z_j, \tilde{Z}_j)$.

Choices of Z include well-studied statistical measures such as the correlation between $\boldsymbol{X}_j$ and $\boldsymbol{y}$ or the coefficient estimated in a linear model, but can also include much more ad-hoc/heuristic measures such as random forest bagging feature importances or sensitivity analysis measures like the Monte-Carlo-estimated total sensitivity index. There are also many choices for w, such as $|Z| - |\tilde{Z}|$, $\text{sign}(Z - \tilde{Z}) \cdot \max\{|Z|, |\tilde{Z}|\}$, or $\log(|Z|) - \log(|\tilde{Z}|)$. For instance, the default statistic suggested in Barber and Candès (2015) is the Lasso Signed Max (LSM), which corresponds to Z being the largest penalty parameter at which a variable enters the model in the lasso regression of $\boldsymbol{y}$ on $[\boldsymbol{X}\widetilde{\boldsymbol{X}}]$, and $w = \text{sign}(Z - \tilde{Z}) \cdot \max\{|Z|, |\tilde{Z}|\}$.

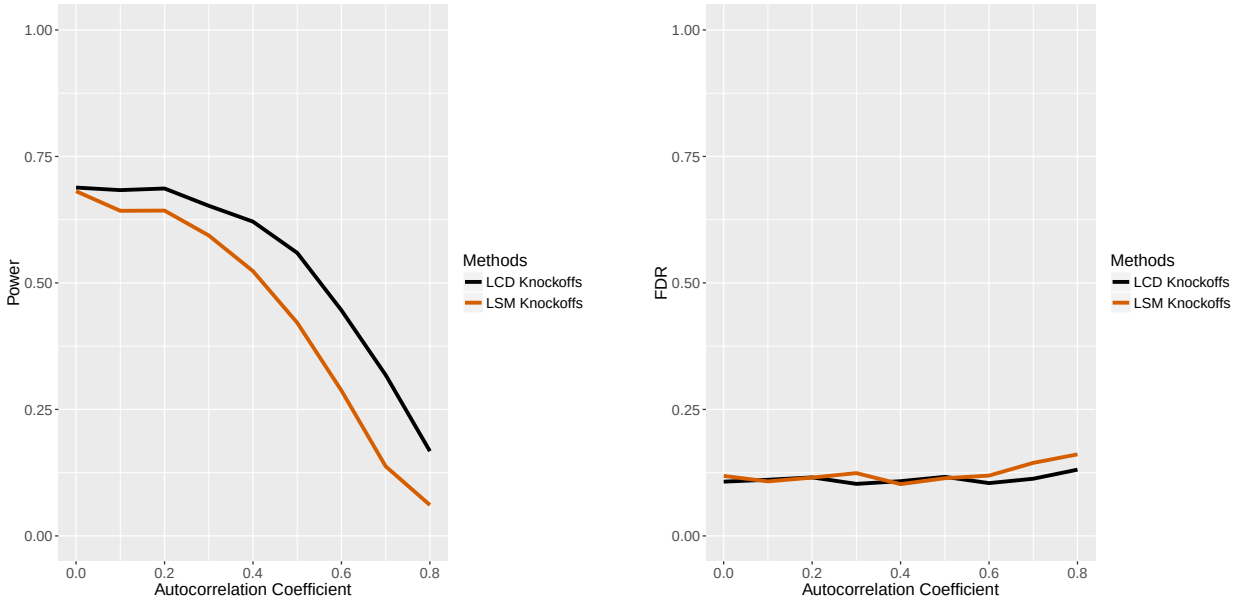
3.2.2 Adaptive Knockoff Statistics

In addition to the LSM statistic, Barber and Candès (2015) suggested alternatives such as the difference in absolute values of estimated coefficients in a model for a variable and its knockoff:

$$W_j = |\beta_j| - |\tilde{\beta}_j|,$$

where that model is estimated so that $\boldsymbol{W}$ obeys the sufficiency proprty required by the original knockoffs procedure, e.g. by ordinary least squares or lasso with a pre-specified tuning parameter. The removal of the sufficiency requirement for MF knockoffs allows us to improve this class of statistics by adaptively tuning the fitted model. The simplest example is to use cross-validation to choose the tuning parameter in the lasso, and we will call this statistic the Lasso Coefficient Difference (LCD) statistic. The key is that the tuning and cross-validation is done on the augmented design matrix $[\boldsymbol{X}\widetilde{\boldsymbol{X}}]$, so that $\boldsymbol{W}$ still obeys the antisymmetry property.

More generally, MF knockoffs allows us to construct statistics that are highly adaptive to the data, as long as that adaptivity is blind to which variables are knockoffs. For instance, we could compute the cross-validated error of the ordinary lasso (still of $\boldsymbol{y}$ on $[\boldsymbol{X}\widetilde{\boldsymbol{X}}]$) and compare it to that of a random forest, and choose Z to be a feature importance measure derived from whichever one is smaller. Since the lasso works best when the true model is close to linear, while random forests work best in non-smooth models, this approach gives us high-level adaptivity to the model smoothness without losing strict Type I error control.



(a) (b)

Figure 2: Power and FDR (target is 10%) for knockoffs with the LCD and LSM statistics. The design matrix has i.i.d. rows and AR(1) columns with autocorrelation coefficient specified by the x-axes of the plots, and the matrix is renormalized so that each column is marginally $N(0, 1/n)$. $n = 3000$, $p = 1000$, and $\boldsymbol{y}$ comes from a Gaussian linear model with $\|\boldsymbol{\beta}\|_0 = 60$, noise variance 1, and all nonzero entries of $\boldsymbol{\beta}$ having magnitude 3.5 and random signs; each point represents 200 replications.

Returning to the simpler example of adaptivity, we found the LCD statistic to be uniformly more powerful than the LSM statistic across a wide range of simulations, particularly under covariate dependence. We note, however, the importance of choosing the penalty parameter that minimizes the cross-validated error, as opposed to the default in some computational packages of the 1-σ rule, as the latter causes LCD to be underpowered compared to LSM in low-power settings. Figure 2 shows a simulation with $n = 3000$, $p = 1000$ of a linear model (with statistics computed from lasso linear regression) that is representative of the power difference between the two statistics (in high dimensions and in other GLMs as well, though not shown). In the remainder of this section, we will always use the LCD statistic. Explicitly, when the response variable is continuous, we use the standard lasso with Gaussian linear model likelihood, and when the response is binary, we use lasso-penalized logistic regression.

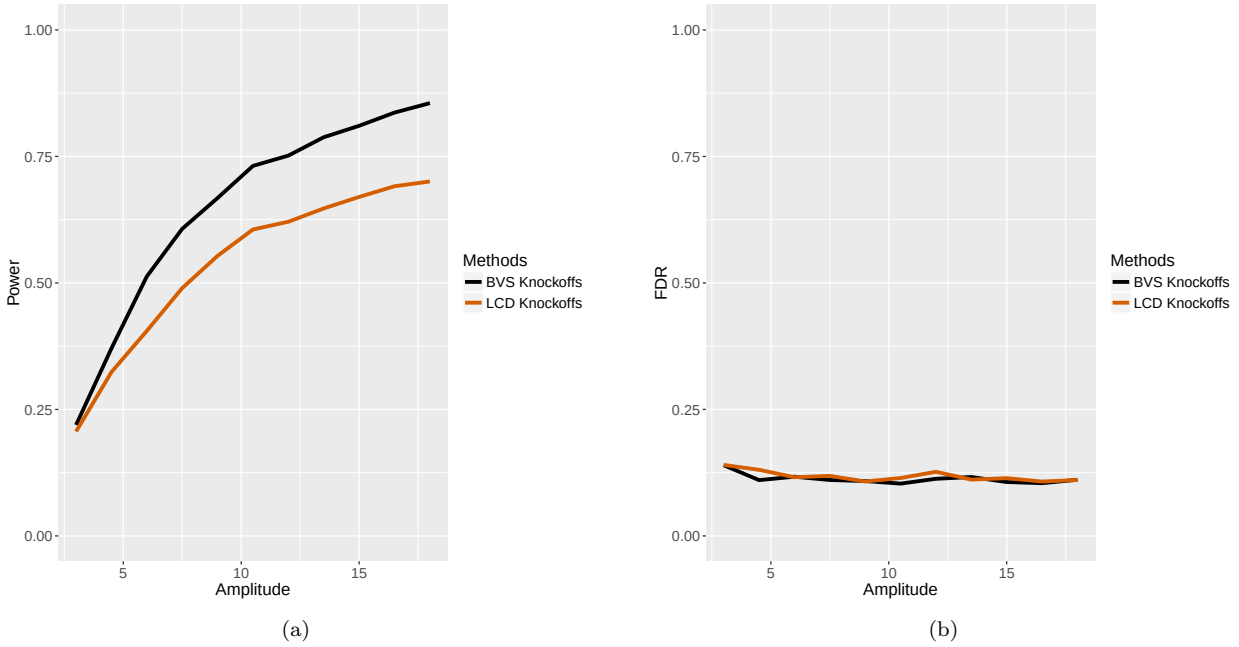
3.2.3 Bayesian Knockoff Statistics

Another very interesting source of knockoff statistics comes from Bayesian procedures. If a statistician has prior knowledge about the problem, he or she can encode it in a Bayesian model and use the resulting estimators to construct a statistic (e.g. difference of absolute posterior mean coefficients, or difference of posterior probabilities of nonzero coefficients with a sparse prior). What makes this especially appealing is that the statistician gets the power advantages of incorporating prior information, while maintaining a strict frequentist guarantee on the Type I error, *even if the prior is false*!

As an example, we ran knockoffs in an experiment with a Bayesian heirarchical regression model with $n = 300$, $p = 1000$, and $\mathbb{E}(\|\boldsymbol{\beta}\|_0) = 60$; see Appendix A for details. The statistics we used were the LCD and a Bayesian variable selection (BVS) statistic, namely the difference between each variable and its knockoff of the posterior probability that the coefficient is nonzero, computed by 500 Gibbs samples after a burn-in of 50 samples (George and McCulloch, 1997); again see Appendix A for details. Figure 3 shows that the accurate prior information supplied to the Bayesian knockoff statistic gives it improved power over LCD which lacks such information, but that they have the same FDR control (and they would even if the prior information were incorrect).

3.3 Alternative Procedures

The alternative procedures we consider, and the settings in which they are valid:



(a) (b)

Figure 3: Power and FDR (target is 10%) for knockoffs with the LCD and Bayesian variable selection (BVS) statistics. The design matrix is i.i.d. $N(0, 1/n)$, $n = 300$, $p = 1000$, and $\boldsymbol{y}$ comes from a Gaussian linear model with $\boldsymbol{\beta}$ and the noise variance randomly chosen, with $\mathbb{E}(\|\boldsymbol{\beta}\|_0) = 60$, expected noise variance 1, and nonzero entries of $\boldsymbol{\beta}$ Gaussian with mean zero and standard deviation given on the x-axis; see Appendix A for precise model. Each point represents 200 replications.

1. The original knockoffs procedure with setting recommended in Barber and Candès (2015). This method can only be applied in Gaussian linear regression when $n \geq p$.
2. BHq applied to asymptotic GLM p-values. This method can only be applied when $n \geq p$, and although for linear regression exact p-values can be computed (again when $n \geq p$), for any other GLM these p-values can be far from valid unless $n \gg p$, as shown in Section 3.1.
3. BHq applied to marginal test p-values. For each j, the correlation between $\boldsymbol{y}$ and $\boldsymbol{X}_j$ is computed and compared to its null distribution, which under certain Gaussian assumptions is closed-form, but in general can at least be simulated exactly by conditioning on $\boldsymbol{y}$ and using the known marginal distribution of $\boldsymbol{X}_j$. Although these tests are valid for testing hypotheses of *marginal* independence (regardless of n and p), such hypotheses only agree with the desired regression (*conditional* independence) hypotheses when the covariates are independent.
4. BHq applied to the p-values from the conditional randomization test described in Section 1.2.

Note that, even ignoring the marginal validity of the p-values in procedures 2–4, the joint distribution of the p-values will not in general satisfy the assumptions for BHq to control FDR. The alternative procedure proposed by Benjamini and Yekutieli (2001), which controls FDR under general p-value dependence, is quite conservative and had extremely noncompetitive power in every simulation we tried. Since in practice BHq tends to be fairly robust to p-value dependence, we only report results from BHq in procedures 2–4, despite the lack of a rigorous guarantee.

3.3.1 Comparison with Conditional Randomization

As mentioned in Section 1.2, procedure 4 is computationally-prohibitive in high dimensions. Therefore, in order to numerically compare it to knockoffs, we were forced to make two concessions. First, we simulated a fairly small logistic regression problem of $n = 300$, $p = 500$, $\|\boldsymbol{\beta}\|_0 = 40$ and only had 50 replications (as opposed to the 200 we use for all other simulations in this section). Second, all conditional randomization p-values corresponding to null variables ($\beta_j = 0$) were generated indendently by drawing from a uniform distribution, as opposed to computing each one with 1000 conditional randomizations as for the non-null p-values. This removes the statistical dependence that in reality would tie all the null variables together as well as the nulls and the non-null variables. While slightly approximate, this is in some sense a best-case scenario for the conditional randomization procedure, as it ensures the

conditions for FDR control of BHq are met exactly when in general they would not be, but the main motivation was that it saved a factor of $500/40 = 12.5$ in computation time (of course in practice the nulls are not known ahead of time, and this approximation could not be used as a computational speed-up).

Approximations aside, we chose the most powerful statistics for comparison of the two procedures, namely the LCD for knockoffs and its analogue the cross-validated absolute lasso coefficient $|\beta_j|$ for the conditional randomization procedure. Even with the aforementioned approximation/speedup, a single run of the conditional randomization procedure took 10 hours to run (on a single 4GB, 2.6GHz compute node in Matlab 2015b, using Glmnet for lasso computations), while knockoffs took just over 1 second under the same circumstances. Finally, Figure **??** shows that the two procedures perform extremely similarly in terms of power and FDR. For the remainder of this section, we only compare knockoffs to procedures 1-3 in simulations with more realistically-large n and p.

3.3.2 Effect of Signal Amplitude

Our first simulation comparing MF knockoffs to procedures 1–3 is by necessity in a Gaussian linear model with $n > p$ and independent covariates—the only setting in which all procedures control the FDR. Specifically, Figures 4(a) and 4(c) plot the power and FDR for the four procedures when $X_{ij} \overset{i.i.d.}{\sim} N(0, 1/n)$, $n = 3000$, $p = 1000$, $\|\boldsymbol{\beta}\|_0 = 60$, the noise variance $\sigma^2 = 1$, and the nonzero entries of $\boldsymbol{\beta}$ have random signs and equal magnitudes, varied along the x-axis. All methods indeed control the FDR, and MF knockoffs is the most powerful, with as much as 10% higher power than its nearest alternative. Figures 4(b) and 4(d) show the same setup but in high dimensions: $p = 6000$. In the high-dimensional regime, neither maximum likelihood p-values nor original knockoffs can even be computed, and MF knockoffs has considerably higher power than BHq applied to marginal p-values.

Next we move beyond the Gaussian linear model to a binomial linear model with logit link function, precluding the use of the original knockoffs procedure. Figure 5 shows the same simulations as Figure 4 but with $\boldsymbol{y}$ drawn from from the binomial model. The results are similar to in the Gaussian linear model, except that BHq applied to the asymptotic maxium likelihood p-values now has an FDR above 50% (rendering its high power meaningless), which can be understood as a manifestation of the phenomenon from Section 3.1. In summary, MF knockoffs continues to have the highest power among FDR-controlling procedures.

3.3.3 Effect of Covariate Dependence

To assess the relative power and FDR control of MF knockoffs as a function of covariate dependence, we ran similar simulations as the previous section, but with covariates that are AR(1) with varying autocorrelation coefficient (while the coefficient amplitude remains fixed). It is now relevant to specify that the locations of the nonzero coefficients are uniformly distributed on $\{1, \dots, p\}$. In the interest of space, we only show the low-dimensional ($p = 1000$) Gaussian setting (where all four procedures can be computed) and the high-dimensional ($p = 6000$) Binomial setting, as little new information is contained in the plots for the remaining two settings. Figures 6(c) and 6(d) show that, as expected, BHq with marginal testing quickly loses FDR control with increasing covariate dependence, as the marginal tests are testing the incorrect null hypothesis of *marginal* independence between covariate and response, and so falsely discover many variables that do not contribute to the model, but are correlated with true variables that do. Concentrating on the remaining methods and just the lefthand part of the BHq Marginal curves where FDR is controlled, Figures 6(a) and 6(b) show that MF knockoffs continues to be considerably more powerful than alternatives as covariate dependence is introduced, in low- and high-dimensional linear and nonlinear models.

3.4 Simulations from Real Data

To test the robustness and practicality of the new knockoffs procedure, we ran simulations using a real genetics design matrix but simulated coefficient vectors (so we know the ground truth) and noise processes. Explicitly, we use data from the Northern Finland Birth Cohort[1] study (NFBC) (Järvelin et al., 2004; Sabatti et al., 2009), which contains over 300,000 single nucleotide polymorphism (SNP) measurements (0 for homozygous major allele, 1 for heterozygous, 2 for homozygous minor allele) on 5,402 subjects. After removing non-SNPs, SNPs with minor allele frequence below 1%, SNPs with Hardy-Weinberg Equilibrium test p-value below 0.0001, and SNPs that were more than 5% missing (for the SNPs that remained, missing values were imputed with the mean of the non-missing values), 334,120 columns remained, 99% of which had at least 300 (out of 5402) non-zeros. Restricting to those with at least 300 non-zero values, we downsampled the columns by a factor of 300, leaving us with 1,100 columns in the end. The downsampling was necessary for computational reasons, as our simulations required us to estimate the unknown covariance matrix hundreds of times.

[1] obtained through the dbGaP database, accession number phs000276.v2.p1

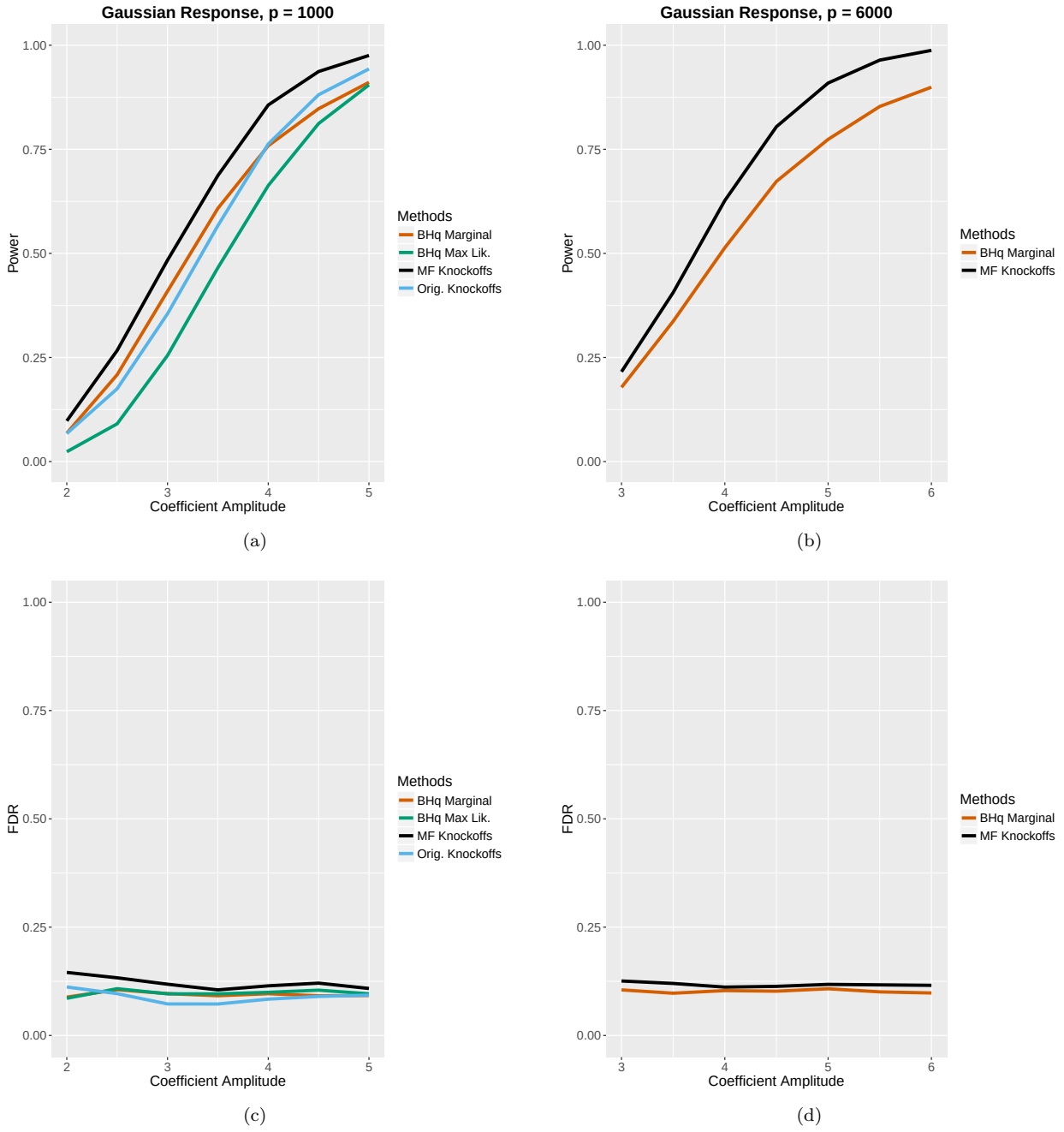


Figure 4: Power and FDR (target is 10%) for MF knockoffs and alternative procedures. The design matrix is i.i.d. $N(0, 1/n)$, $n = 3000$, $p =$ (a)/(c): 1000 and (b)/(d): 6000, and $\boldsymbol{y}$ comes from a Gaussian linear model with $\|\boldsymbol{\beta}\|_0 = 60$, noise variance 1, and all nonzero entries of $\boldsymbol{\beta}$ having equal magnitudes and random signs; each point represents 200 replications.

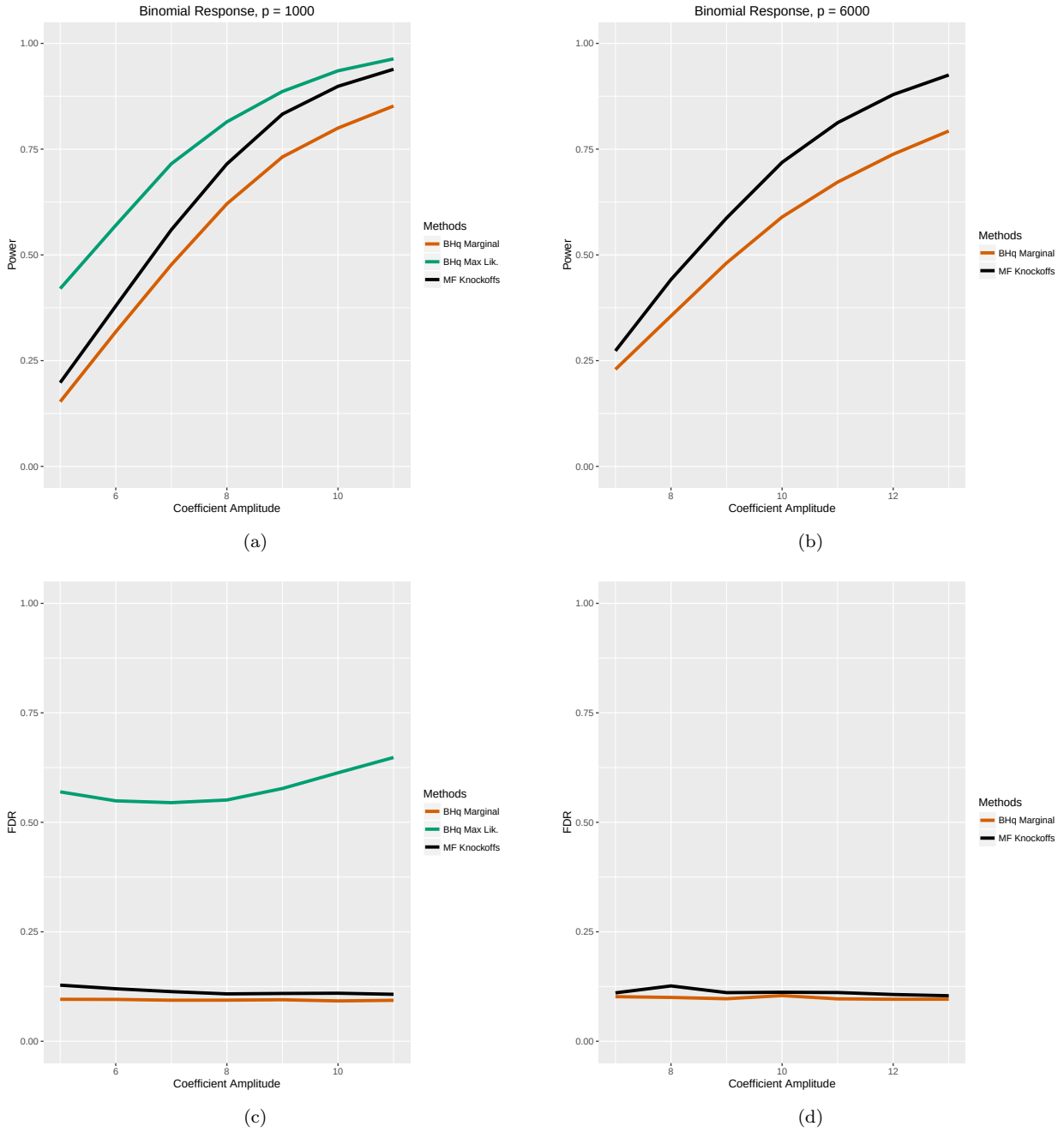


Figure 5: Power and FDR (target is 10%) for MF knockoffs and alternative procedures. The design matrix is i.i.d. $N(0, 1/n)$, $n = 3000$, $p =$ (a)/(c): 1000 and (b)/(d): 6000, and $\boldsymbol{y}$ comes from a binomial linear model with logit link function, with $\|\boldsymbol{\beta}\|_0 = 60$ and all nonzero entries of $\boldsymbol{\beta}$ having equal magnitudes and random signs; each point represents 200 replications.

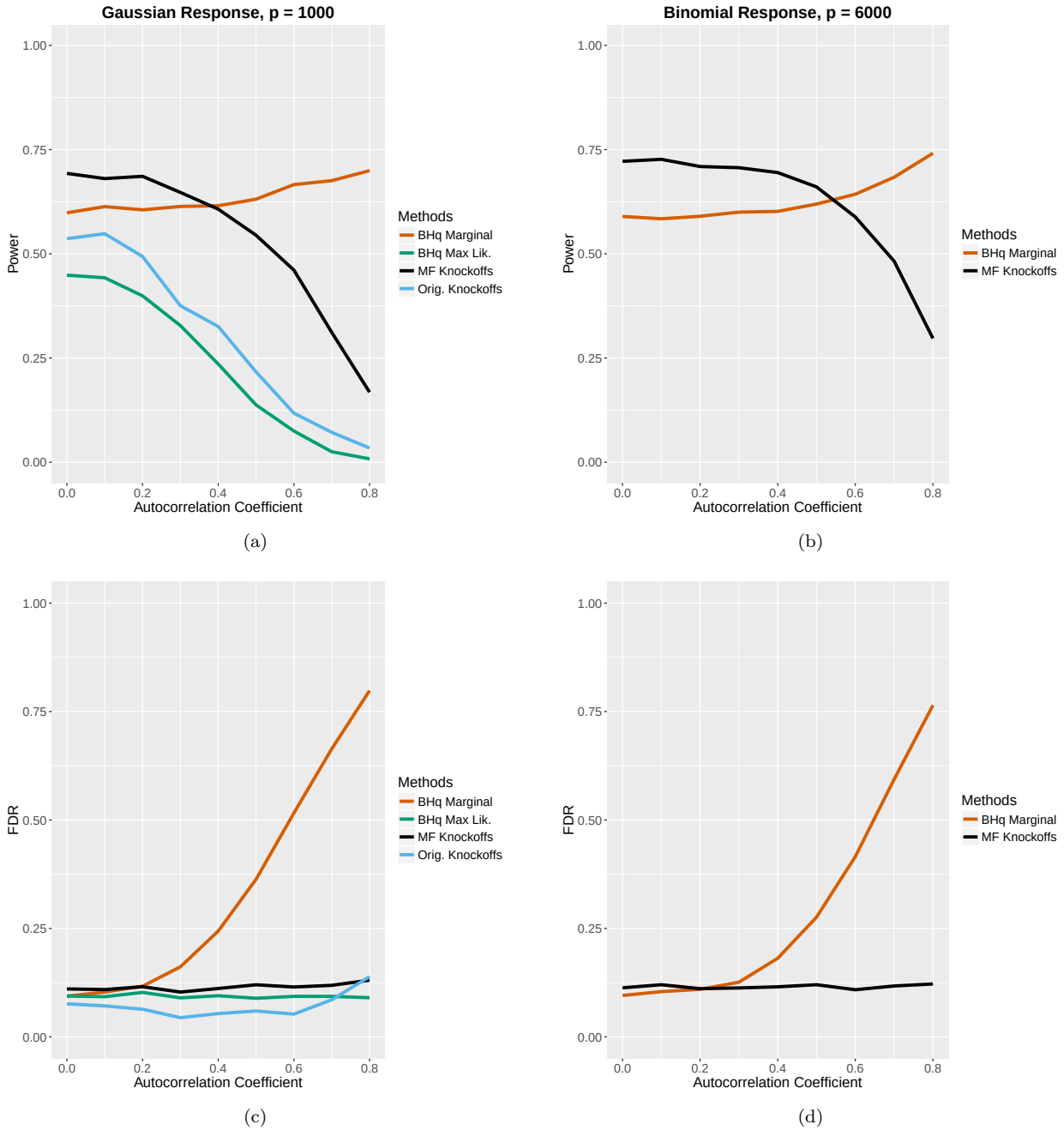
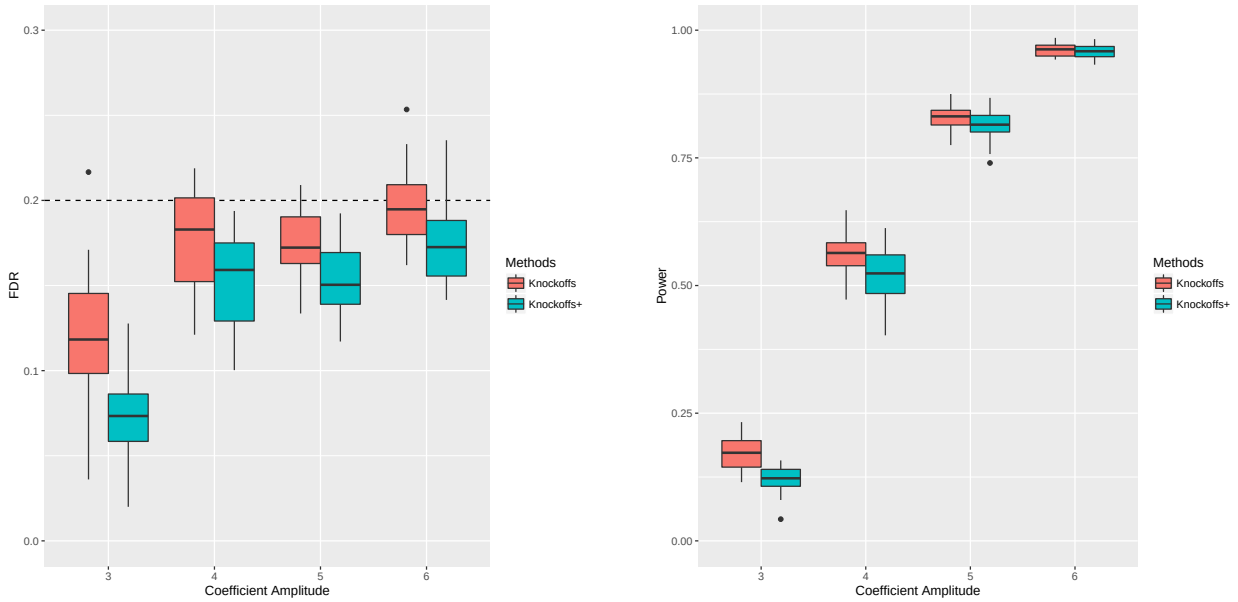


Figure 6: Power and FDR (target is 10%) for MF knockoffs and alternative procedures. The design matrix has i.i.d. rows and AR(1) columns with autocorrelation coefficient specified by the x-axes of the plots, and the matrix is renormalized so that each column is marginally $N(0, 1/n)$. (a)/(c): $p = 1000$, $\boldsymbol{y}$ follows a Gaussian linear model, and nonzero coefficients have amplitude 3.5. (b)/(d): $p = 6000$, $\boldsymbol{y}$ follows a binomial linear model with logit link function, and nonzero coefficients have amplitude 10. $n = 3000$, $\|\boldsymbol{\beta}\|_0 = 60$ and nonzero entries of $\boldsymbol{\beta}$ having equal magnitudes and random signs, and their locations are uniformly distributed; each point represents 200 replications.



(a) (b)

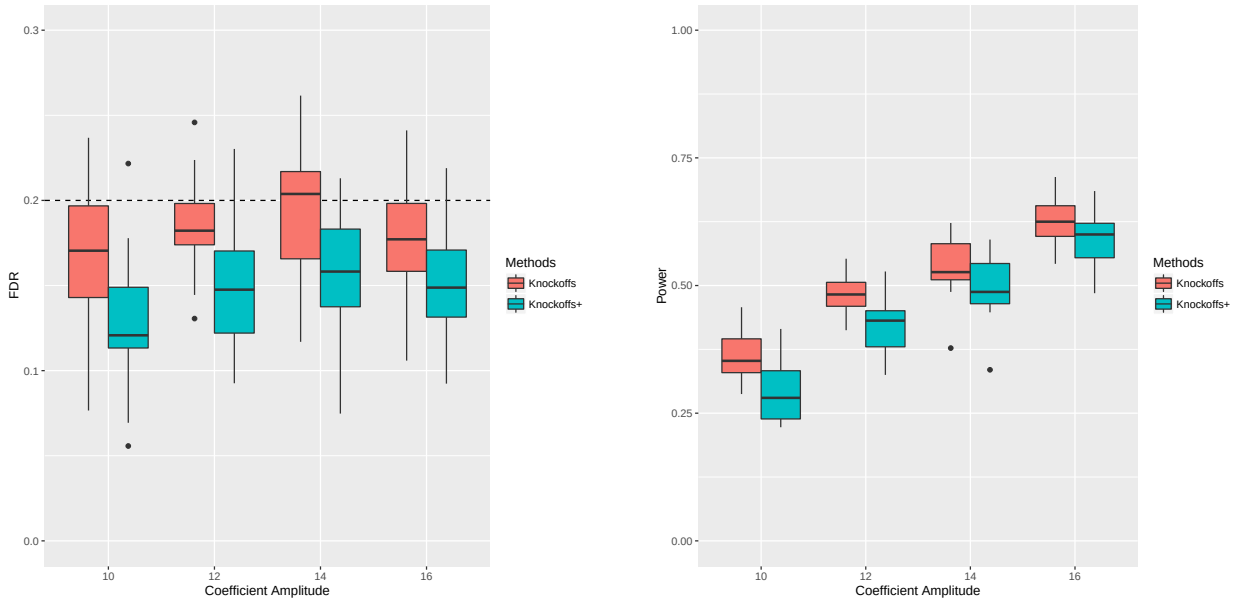
Figure 7: (a) FDR and (b) power in linear regression for knockoffs applied to the NFBC data split into 10 disjoint samples (each point in each boxplot is averaged over these samples), then repeated 20 times (giving the boxplot distribution). The covariance is estimated using the graphical lasso, and $n = 540$, $p = 1100$ with 40 nonzero coefficients.

In each experiment, we mimicked realistic sampling from a random design by throwing away the last two subjects and splitting the remaining 5400 subjects randomly into 10 blocks of 540. For each block, the graphical lasso with 2-fold cross-validation to choose the tuning parameter was used to estimate the covariance matrix among the 1,100 columns, and then knockoffs was run using this estimated covariance matrix (LCD statistic, equivariant construction). In the case of linear regression, Figure 7 shows the realized power and FDR (averaged only over the 10 blocks, so some variability is expected) of 20 independent such experiments, each with a different random block splitting and random choice of coefficient support (there were 40 nonzero coefficients). Figure 8 shows the same for logistic regression. Because the support of the coefficient vector could affect performance, each point is plotted as a boxplot summarizing the 20 trials.

It is comforting to see that FDR is effectively controlled in all trials. We emphasize that the covariance was unknown here, and the marginal distributions (not to mention joint distribution) of the covariates were far from Gaussian. This speaks to the robustness of our procedure, despite the stringent technical assumption that the covariate distribution be multivariate Gaussian with known covariance. This robustness can be at least partially understood by considering the statistic used, which is based on the Lasso. Ignoring for the moment the cross-validation for choosing the penalty parameter, the Lasso only uses certain sufficient statistics which are inner-products among the columns of X and between y and the columns of X. Thus, the knockoffs exchangeability property is really only needed for these sufficient statistics, which are all inner-products over many (in this case, 540) subjects, and the multivariate central limit theorem tells us that they will be approximately multivariate Gaussian under very weak assumptions on the joint distribution of covariates. This makes much of the approximation error in assuming the covariates are Gaussian irrelevant, as the covariates are only used through their inner-products, which turn out to be nearly multivariate Gaussian anyway.

4 Discussion

- understand robustness to random design distribution estimation error
- what are other more general ways of obtaining knockoff variables with pairwise exchangeability property
- extension to other error rates as in Janson and Su (2016)
- other statistics now with only antisymmetry requirement and no sufficiency requirement



(a) (b)

Figure 8: (a) FDR and (b) power in logistic regression for knockoffs applied to the NFBC data split into 10 disjoint samples (each point in each boxplot is averaged over these samples), then repeated 20 times (giving the boxplot distribution). The covariance is estimated using the graphical lasso, and $n = 540$, $p = 1100$ with 40 nonzero coefficients.

References

Affymetrix (2006). BRLMM: an improved genotype calling method for the genechip human mapping 500k array set. Technical report, Affymetrix.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5):2055–2085.

Barber, R. F. and Candes, E. J. (2016). A knockoff filter for high-dimensional selective inference. *arXiv preprint arXiv:1602.03574*.

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 (Methodological)*, 57(1):pp. 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). Slope—adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140.

Efron, B. and Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3):509–524.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica Sinica*, 7(2):339–373.

Janson, L. and Su, W. (2016). Familywise error rate control via knockoffs. *Electron. J. Statist.*, 10(1):960–975.

Järvelin, M.-R., Sovio, U., King, V., Lauren, L., Xu, B., McCarthy, M. I., Hartikainen, A.-L., Laitinen, J., Zitting, P., Rantakallio, P., and Elliott, P. (2004). Early life factors and blood pressure at age 31 years in the 1966 northern finland birth cohort. *Hypertension*, 44(6):838–846.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.*, 42(2):413–468.

Sabatti, C., Service, S. K., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., Sovio, U., Ruokonen, A., Laitinen, J., Jakkula, E., Coin, L., Hoggart, C., Collins, A., Turunen, H., Gabriel, S., Elliot, P., McCarthy, M. I., Daly, M. J., Jrvelin, M.-R., Freimer, N. B., and Peltonen, L. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35–46.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202.

A Bayesian Knockoff Statistics

The data for the simulation of Section 3.2.3 was drawn from:

$$\begin{aligned}
X_{i,j} &\overset{\text{iid}}{\sim} N(0, 1/n), \\
\boldsymbol{\beta} &\sim N(\mathbf{0}, \tau^2 \mathbf{I}_p), \\
\gamma_j &\overset{\text{iid}}{\sim} \text{Bernoulli}(\pi),\ j \in \{1, \dots, p\}, \\
\frac{1}{\sigma^2} &\sim \text{Gamma}(A, B) \qquad \text{(shape/scale parameterization, as opposed to shape/rate)}, \\
\boldsymbol{y} &\sim N(\boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 \mathbf{I}_n),
\end{aligned}$$

where $\boldsymbol{X}_{\boldsymbol{\gamma}}$ denotes the matrix composed of the columns of $\boldsymbol{X}$ for which $\boldsymbol{\gamma} = 1$, and similarly for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. The parameter values in the simulation were $n = 300$, $p = 1000$, $\pi = \frac{60}{1000}$, $A = 5$, $B = 4$, and τ varied along the x-axis of the plot.

To compute the Bayesian variable selection (BVS) knockoff statistic, we used a Gibbs sampler on the following model (treating $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$ as fixed):

$$\begin{aligned}
\boldsymbol{\beta} &\sim N(\mathbf{0}, \tau^2 \mathbf{I}_{2p}), \\
\lambda_j &\overset{\text{iid}}{\sim} \text{Bernoulli}(\pi),\ j \in \{1, \dots, p\}, \\
(\gamma_j, \gamma_{j+p}) &\overset{\text{iid}}{\sim} \left\{ \begin{array}{ll} (0,0) & \text{if } \lambda_j = 0 \\ (0,1) \text{ w.p. } 1/2 & \text{if } \lambda_j = 1 \\ (1,0) \text{ w.p. } 1/2 & \text{if } \lambda_j = 1 \end{array} \right\},\ j \in \{1, \dots, p\}, \\
\frac{1}{\sigma^2} &\sim \text{Gamma}(A, B) \qquad \text{(shape/scale parameterization, as opposed to shape/rate)}, \\
\boldsymbol{y} &\sim N([\boldsymbol{X}\widetilde{\boldsymbol{X}}]_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 \mathbf{I}_n),
\end{aligned}$$

which requires only a very slight modification of the procedure in George and McCulloch (1997). After computing the posterior probabilities $\hat{\gamma}_j$ with 500 Gibbs samples (after 50 burn-in samples), we computed the jth knockoff statistic as

$$W_j = |\hat{\gamma}_j| - |\hat{\gamma}_{j+p}|.$$

B WTCCC Data

The SNP arrays came from an Affymetrix 500K chip, with calls made by the BRLMM algorithm Affymetrix (2006). Values were considered missing if any of the following conditions were satisfied:

- BRLMM score > 0.5