Supplementary material to "Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection"

Emmanuel Candès[1], Yingying Fan[2], Lucas Janson[1] and Jinchi Lv[2]

Stanford University[1] and University of Southern California[2]

This Supplementary Material contains extended discussions, additional technical details on the sequential conditional independent pairs algorithm, an example of logistic regression producing invalid p-values, implementation details for Bayesian knockoff statistics, some computational speed-ups/shortcuts for conditional randomization, and more detailed results from genetic analysis of Crohn's disease summarized in Section 6.

A. Difference from work on inference after selection

The main differences between our work and those on inference after selection are as follows:

- First, our focus is on selecting the right variables, whereas the goal of this line of work is to adjust inference after some selection has taken place. In more detail, these works presuppose a selection procedure has been chosen (for reasons that may have nothing to do with controlling Type I error) and then compute p-values for the selected variables, taking into account the selection step. In contrast, MX knockoffs is by itself a selection procedure that controls Type I error.
- Second, inference after selection relies heavily on parametric assumptions about the conditional distribution, namely that

  $$Y|X_1, \dots, X_p \sim \mathcal{N}(\mu(X_1, \dots, X_p), \sigma^2),$$

  making it unclear how to extend it to the more general setting of the present paper.
- The third difference stems from their objects of inference. In the selection step, a subset of size $m \leq n$ of the original p covariates is selected, say $X_{j_1}, \dots, X_{j_m}$, and the objects of inference are the coefficients of the X_{j_k}'s in the projection of μ onto the linear subspace spanned by the $n \times m$ matrix of observed values of these X_{j_k}'s. That is, the kth null hypothesis is that the aforementioned coefficient on X_{j_k} is zero—note that whether or not inference on the jth variable is produced at all, and if it is, the object of that inference, both depend on the initial selection step. In contrast, if MX knockoffs were restricted to the homoscedastic Gaussian model above, the jth null hypothesis would be that μ does not depend on X_j, and there would be p such null hypotheses, one for each of the original variables.

B. Marginal testing

Some specific drawbacks of using marginal p-values are:

1. **Power loss** Even when the covariates are independent, so that the hypotheses of conditional and unconditional independence coincide, p-values resulting from marginal testing procedures may be less powerful than those from conditional testing procedures. This phenomenon has been reported previously, for example in statistical genetics by Hoggart et al. (2008) and many others. Intuitively, this is because a joint model in $X_1, \dots, X_p$ for Y will have less residual variance than a marginal

model in just X_j. There are exceptions, for instance if there is only one important variable, then its marginal model is the correct joint model and a conditional test will be less powerful due to the uncertainty in how the other variables are included in the joint model. But in general, marginal testing becomes increasingly underpowered relative to conditional testing as the *absolute* number of important covariates increases (Frommlet et al., 2012), suggesting particular advantage for conditional testing in modern applications with complex high-dimensional models.

There are also cases in which important variables are in fact fully marginally independent of the response. As a toy example, if $X_1, X_2 \overset{\text{iid}}{\sim} \text{Bernoulli}(0.5)$ and $Y = \mathbb{1}_{\{X_1+X_2=1\}}$, then Y is marginally independent of each of X_1 and X_2, even though together they determine Y perfectly. X_1 and X_2 are certainly *conditionally* dependent on Y, however, so a conditional test can have power to discover them.

2. Interpretability When the covariates are not independent, marginal and conditional independence do not correspond, and we end up asking the wrong question. For example, in a model with a few important covariates which determine the outcome, and many unimportant covariates which have no influence on the outcome but are correlated with the important covariates, marginal testing will treat such unimportant covariates as important. Thus, because marginal testing is testing the wrong hypotheses, there will be many "discoveries" which have no influence on the outcome.

The argument is often made, especially in genetics, that although discovering an unimportant covariate just because it was correlated with an important one is technically incorrect, it can still be useful as it suggests that there is an important covariate correlated with the discovered one. While this is indeed useful, especially in genetics where correlated SNPs tend to be very close to one another on the genome, this comes at a price since it significantly alters the meaning of the FDR. Indeed, if we adopt this viewpoint, the unit of inference is no longer a SNP but, rather, a region on the genome, yet FDR is still being tested at the SNP level. A consequence of this mismatch is that the result of the analysis may be completely misleading, as beautifully argued in Pacifico et al. (2004); Benjamini and Heller (2007); Siegmund et al. (2011), see also Chouldechova (2014) and Brzyski et al. (2016) for later references.

3. Dependent p-values Marginal p-values will, in general, have quite complicated joint dependence, so that BHq does not control FDR exactly. Although procedures for controlling FDR under arbitrary dependence exist, their increased generality tends to make them considerably more conservative than BHq. In practice, however, the FDR of BHq applied to dependent p-values is usually below its nominal level, but the problem is that it can have highly *variable* FDP. Recall that FDR is the expectation of FDP, the latter being the random quantity we actually care about but cannot control directly. Therefore, FDR control is only useful if the realized FDP is relatively concentrated around its expectation, and it is well-established (Efron, 2010, Chapter 4) that under correlations BHq can produce highly skewed FDP distributions. In such cases, with large probability, FDP $= 0$ perhaps because no discoveries are made, and when discoveries are made, the FDP may be much higher than the nominal FDR, making it a misleading error bound.

C. Relationship with the knockoffs of Barber and Candès

We can immediately see a key difference with the earlier framework of Barber and Candès (2015). In their work, the design matrix is viewed as being fixed and setting $\boldsymbol{\Sigma} = \boldsymbol{X}^\top \boldsymbol{X}$, knockoff variables are constructed as to obey

$$[\boldsymbol{X}, \tilde{\boldsymbol{X}}]^\top [\boldsymbol{X}, \tilde{\boldsymbol{X}}] = \boldsymbol{G}, \quad \boldsymbol{G} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \text{diag}\{s\} \\ \boldsymbol{\Sigma} - \text{diag}\{s\} & \boldsymbol{\Sigma} \end{bmatrix}. \tag{A.1}$$

Imagine the columns of $\boldsymbol{X}$ are centered, i.e., have vanishing means, so that we can think of $\boldsymbol{X}^\top \boldsymbol{X}$ as a sample covariance matrix. Then the construction of Barber and Candès is asking that the *sample* covariance matrix of the joint set of variables obeys the exchangeability property—i.e., swapping rows and columns leaves the covariance invariant—whereas in this paper, it is the *population* covariance that must be invariant. In particular, the MX knockoffs will be far from obeying the relationship $\boldsymbol{X}^\top \boldsymbol{X} = \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}$ required in (A.1). To drive this point home, assume $X \sim \mathcal{N}(0, \boldsymbol{I})$. Then we can choose $\tilde{X} \sim \mathcal{N}(0, \boldsymbol{I})$ independently from X, in which case $\boldsymbol{X}^\top \boldsymbol{X}/n$ and $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}/n$ are two independent Wishart variables. An important consequence of this is that in the MX approach, the sample correlation between the jth columns of $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ will typically be far smaller than that in the original framework of Barber and Candès. For example, take $n = 3000$ and $p = 1000$ and assume the equicorrelated construction from Barber and Candès (2015). Then the sample correlation between any variable $\boldsymbol{X}_j$ and its knockoff $\tilde{\boldsymbol{X}}_j$ will be about 0.65 while in the random case, the average magnitude of the correlation is about 0.015. This explains the power gain of our new method that is observed in Section 4.

D. Proof of exchangeability of null covariates and their knockoffs

PROOF (LEMMA 3.2). Since y's marginal distribution is the same on both sides of the equation, it is equivalent to show that $([\boldsymbol{X}, \tilde{\boldsymbol{X}}], y) \overset{d}{=} ([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\text{swap}(S)}, y)$, which is how we proceed. Assume without loss of generality that $S = \{1, 2, \ldots, m\}$. By row independence, it suffices to show that $((X, \tilde{X}), Y) \overset{d}{=} ((X, \tilde{X})_{\text{swap}(S)}, Y)$, where X (resp. Y) is a row of $\boldsymbol{X}$ (resp. y). Furthermore, since $(X, \tilde{X}) \overset{d}{=} (X, \tilde{X})_{\text{swap}(S)}$, we only need to establish that

$$Y \mid (X, \tilde{X})_{\text{swap}(S)} \overset{d}{=} Y \mid (X, \tilde{X}). \tag{A.1}$$

Letting $p_{Y|X}(y|x)$ be the conditional distribution of Y, observe that

$$\begin{aligned} p_{Y|(X,\tilde{X})_{\text{swap}(S)}}(y|(x, \tilde{x})) &= p_{Y|(X,\tilde{X})}(y|(x, \tilde{x})_{\text{swap}(S)}) \\ &= p_{Y|X}(y|x'), \end{aligned}$$

where $x'_i = \tilde{x}_i$ if $i \in S$ and $x'_i = x_i$ otherwise. The second equality above comes from the fact that Y is conditionally independent of $\tilde{X}$ by property (2) in the definition of MX knockoffs. Next, since Y and X_1 are independent conditional on $X_{2:p}$, we have

$$\begin{aligned} p_{Y|X_{1:p}}(y|\tilde{x}_1, x'_{2:p}) &= p_{Y|X_{2:p}}(y|x'_{2:p}) \\ &= p_{Y|X_{1:p}}(y|x_1, x'_{2:p}). \end{aligned}$$

This shows that

$$Y \mid (X, \tilde{X})_{\text{swap}(S)} \overset{d}{=} Y \mid (X, \tilde{X})_{\text{swap}(S \setminus \{1\})}.$$

We can repeat this argument with the second variable, the third, and so on until S is empty. This proves (A.1).

E. Sequential conditional independent pairs algorithm

We will prove by induction on j that Algorithm 1 produces knockoffs that satisfy the exchangeability property (3.1). We prove the result for the discrete case; the general case follows the same argument with a slightly more careful measure-theoretic treatment using Radon–Nikodym derivatives instead of probability mass functions. Below we denote the probability mass function (PMF) of $(X_{1:p}, \tilde{X}_{1:j-1})$ by $\mathcal{L}(X_{\text{-}j}, X_j, \tilde{X}_{1:j-1})$.

INDUCTION HYPOTHESIS 1. *After j steps, every pair $X_k, \tilde{X}_k$ is exchangeable in the joint distribution of $(X_{1:p}, \tilde{X}_{1:j})$ for $k = 1, \ldots, j$.*

By construction, the induction hypothesis is true after 1 step since X_1 and $\tilde{X}_1$ are conditionally independent and have the same marginal distribution (this implies conditional exchangeability). Assuming the induction hypothesis holds until $j-1$, we prove that it holds after j steps. Note that by by assumption, $\mathcal{L}$ is symmetric in $X_k, \tilde{X}_k$ (that is, the function value remains unchanged when the argument values $X_k, \tilde{X}_k$ are swapped) for $k = 1, \ldots, j-1$. Also, the conditional PMF of $\tilde{X}_j$ given $X_{1:p}, \tilde{X}_{1:j-1}$ is given by

$$\frac{\mathcal{L}(X_{\text{-}j}, \tilde{X}_j, \tilde{X}_{1:j-1})}{\sum_u \mathcal{L}(X_{\text{-}j}, u, \tilde{X}_{1:j-1})}.$$

Therefore, the joint PMF of $(X_{1:p}, \tilde{X}_{1:j})$ is given by the product of the aforementioned conditional PMF with the joint PMF of $(X_{1:p}, \tilde{X}_{1:j-1})$:

$$\frac{\mathcal{L}(X_{\text{-}j}, X_j, \tilde{X}_{1:j-1})\mathcal{L}(X_{\text{-}j}, \tilde{X}_j, \tilde{X}_{1:j-1})}{\sum_u \mathcal{L}(X_{\text{-}j}, u, \tilde{X}_{1:j-1})}. \tag{A.1}$$

Exchangeability of $X_j, \tilde{X}_j$ follows from the symmetry of (A.1) to those two values. For $k < j$, note that (A.1) only depends on $X_k, \tilde{X}_k$ through the function $\mathcal{L}$, and that $\mathcal{L}$ is symmetric in $X_k, \tilde{X}_k$. Therefore, (A.1) is also symmetric in $X_k, \tilde{X}_k$, and therefore the pair is exchangeable in the joint distribution of $(X_{1:p}, \tilde{X}_{1:j})$.

F. The conditional randomization test

This section presents an alternative approach to the controlled variable selection problem. To describe our approach, it may be best to consider an example. Assume we are in a regression setting and let $\hat{b}_j(\lambda)$ be the value of the Lasso estimate of the jth regression coefficient. We would like to use the statistic $\hat{b}_j(\lambda)$ to test whether Y is conditionally independent of X_j since large values of $|\hat{b}_j(\lambda)|$ provide evidence against the null. To construct a test, however, we would need to know the sampling distribution of $\hat{b}_j(\lambda)$ under the null hypothesis that Y and X_j are conditionally independent, and it is quite unclear how one would obtain such knowledge.

F.1. The test

A way out is to sample the covariate X_j conditional on all the other covariates (but not the response), where by "sample" we explicitly mean to draw a new sample from the conditional distribution of $X_j \,|\, X_{\text{-}j}$ using a random number generator. We then compute the Lasso statistic $\hat{b}^*_j(\lambda)$, where the * superscript indicates that the statistic is computed from the artificially sampled value of the covariate X_j. Now, under the null hypothesis of conditional independence between Y and X_j, it happens that $\hat{b}^*_j(\lambda)$ and $\hat{b}_j(\lambda)$ are identically distributed and that, furthermore, this statement holds true conditional on Y and all the other covariates. This claim is proved in Lemma F.1 below. A consequence of this is that by simulating a covariate conditional on the others, we can sample at will from the conditional distribution of any test statistic and compute p-values as described in Algorithm 1.

LEMMA F.1. *Let (Z_1, Z_2, Y) be a triple of random variables, and construct another triple (Z^*_1, Z_2, Y) as*

$$Z^*_1 \,|\, (Z_2, Y) \quad \overset{d}{=} \quad Z_1 \,|\, Z_2.$$

Then under the null hypothesis $Y \perp\!\!\!\perp Z_1 \,|\, Z_2$, any test statistic $T = t(Z_1, Z_2, Y)$ obeys

$$T \,|\, (Z_2, Y) \quad \overset{d}{=} \quad T^* \,|\, (Z_2, Y),$$

where $T^* = t(Z^*_1, Z_2, Y)$.

Algorithm 1 Conditional Randomization Test.

```
Input: A set of $n$ independent samples $(X_{i1}, \dots, X_{ip}, Y_i)_{1 \le i \le n}$ assembled in a data matrix $\boldsymbol{X}$ and a response vector $y$, a feature importance statistic $T_j(\boldsymbol{X}, y)$ to test whether $X_j$ and $Y$ are conditionally independent.

Loop: for $k = 1, 2, \dots, K$ do
    Create a new data matrix $\boldsymbol{X}^{(k)}$ by simulating the $j$th column of $\boldsymbol{X}$ from $\mathcal{L}(X_j | X_{\text{-}j})$ (and keeping the remaining columns the same). That is, $X_{ij}^{(k)}$ is sampled from the conditional distribution $X_{ij} \mid \{X_{i1}, \dots, X_{ip}\} \setminus \{X_{ij}\}$, and is (conditionally) independent of $X_{ij}$.

Output: A (one-sided) p-value
    $P_j = \frac{1}{K+1}\left[1 + \sum_{k=1}^{K} \mathbb{1}_{T_j(\boldsymbol{X}^{(k)}, y) \ge T_j(\boldsymbol{X}, y)}\right].$
As with permutation tests, adding one in both the numerator and the denominator makes sure that the null p-values are stochastically larger than uniform variables.
```

PROOF. To prove the claim, it suffices to show that Z_1 and Z_1^* have the same distribution conditionally on (Z_2, Y). This follows from

$$Z_1^* \mid (Y, Z_2) \quad \overset{d}{=} \quad Z_1 \mid Z_2 \quad \overset{d}{=} \quad Z_1 \mid (Z_2, Y).$$

The first equality comes from the definition of Z_1^* while the second follows from the conditional independence of Y and Z_1, which holds under the null.

The consequence of Lemma F.1 is that we can compute the 95% percentile, say, of the conditional distribution of T^* denoted by $t^*_{0.95}(Z_2, Y)$. Then by definition, under the null,

$$\mathbb{P}(T > t^*_{0.95}(Z_2, Y) \mid (Z_2, Y)) \le 0.05.$$

Since this equality holds conditionally, it also holds marginally.

F.2. Literature review

The conditional randomization test is most closely related to the propensity score (Rosenbaum and Rubin, 1983), which also uses the conditional distribution $X_j \mid X_{\text{-}j}$ to perform inference on the conditional relationship between Y and X_j given $X_{\text{-}j}$. However, propensity scores require X_j be binary, and the propensity score itself is normally estimated, although Rosenbaum (1984) shows that when all the covariates jointly take a small number of discrete values, propensity score analysis can be done exactly. Doran et al. (2014) also rely on the data containing repeated observations of $X_{\text{-}j}$ so that certain observations can be permuted nonparametrically while maintaining the null distribution. In fact, the exact term "conditional randomization test" has also been used in randomized controlled experiments to test for independence of Y and X_j conditioned more generally on some function of $X_{\text{-}j}$ (such as a measure of imbalance in $X_{\text{-}j}$ if X_j is binary), again relying on discreteness of the function so that there exist permutations of X_j which leave the function value unchanged. Despite the similar name, our conditional randomization test is quite distinct from these, as it does not rely on discreteness or experimental control in any of the covariates.

Another line of work exists within the linear model regime, whereby the null (without X_j) model is estimated and then the empirical residuals are permuted to produce a null distribution for Y (Freedman and Lane, 1983). Because this approach is only exact when the empirical residuals match the true residuals, it explicitly relies on a parametric model for $Y \mid X_{\text{-}j}$, as well as the ability to estimate it quite accurately.

F.3. Comparisons with knockoffs

One major limitation of the conditional randomization method is its computational cost. It requires computing randomization p-values for many covariates and to a high-enough resolution for multiple-comparisons correction. Clearly, this requires samples in the extreme tail of the p-value distribution. This means computing a very large number of feature importance statistics T_j, each of which can be expensive since for reasons outlined in the drawbacks associated with marginal testing, powerful T_j's will take into account the full dimensionality of the model, e.g., absolute value of the Lasso-estimated coefficient. In fact, the number of computations of T_j, tallied over all j, required by the conditional randomization method is $\Omega(p)$.[†] To see this, suppose for simplicity that all R rejected p-values take on the value of half the BHq cutoff equal to $\tau = qR/p$, and all we need to do is upper-bound them below τ. This means there are R p-values P_j for which plugging $K = \infty$ into Algorithm 1 would yield $P_j = \tau/2$. After $K < \infty$ samples, the approximate p-value (ignoring the $+1$ correction) is distributed as $K^{-1}\ \mathrm{Bin}(K, P_j)$. We could then use this binomial count to construct a confidence interval for P_j. A simple calculation shows that to be reasonably confident that $P_j \le \tau$, K must be on the order of at least $1/\tau$. Since there are R such p-values, this justifies the claim.

Note that for knockoffs, the analogous computation of T need only be done exactly once. If, for instance, each T_j requires a Lasso computation, then the conditional randomization test's computational burden is very challenging for medium-scale p in the thousands and prohibitive for large-scale (e.g., genetics) p in the hundreds of thousands or millions. We will see in Section 4.2.1 that there are power gains, along with huge computational costs, to be had by using conditional randomization in place of knockoffs, and Section 6 will show that the MX knockoff procedure easily scales to large data sets.

Another advantage of MX knockoffs is its guaranteed control of the FDR, whereas the BHq procedure does not offer strict control when applied to arbitrarily dependent p-values.

G. Logistic regression p-values

Asymptotic maximum likelihood theory promises valid p-values for each coefficient in a GLM only when $n \gg p$. However, these approximate p-values can usually 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 and p are both large with $p/n \ge 0.1$, for example. We simulated 10^4 independent design matrices ($n = 500$, $p = 200$) and binary responses from a logistic regression for the following two settings:

(1) $(X_1, \dots, X_p)$ is an AR(1) time series with AR coefficient 0.5 and

$$Y \mid X_1, \dots, X_p \sim \mathrm{Bernoulli}(0.5)$$

(2) $(X_1, \dots, X_p)$ is an AR(1) time series with AR coefficient 0.5 and

$$Y \mid X_1, \dots, X_p \sim \mathrm{Bernoulli}\left(\mathrm{logit}\left(0.08(X_2 + \cdots + X_{21})\right)\right)$$

Histograms for the p-values for β_1 (null in all cases) are shown in Figure 1. Both histograms are far from 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 $\mathbb{P}\left\{p\text{-value} \le 0.1\%\right\}$ in setting (2). We

[†] $a(N) \in \Omega(b(N))$ means that there exist N_0 and $C > 0$ such that $a(N) \ge Cb(N)$ for $N \ge N_0$.

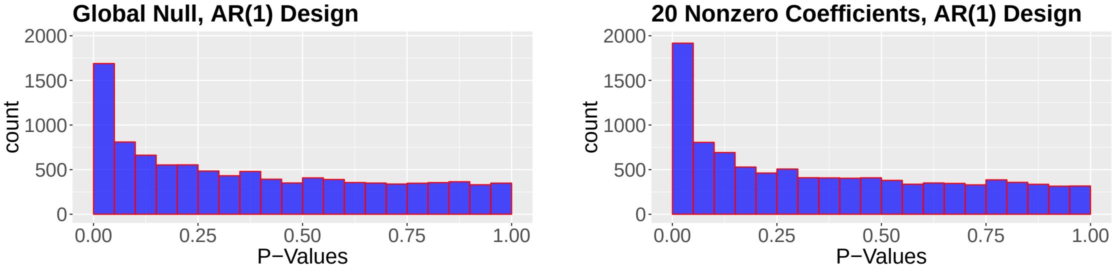

Fig. 1. Distribution of null logistic regression p-values with $n = 500$ and $p = 200$; 10,000 replications.

Table 1. Inflated p-value probabilities with estimated Monte Carlo standard errors in parentheses. See text for meanings of settings (1), (2), (3), (4).

	(1)	(2)	(3)	(4)
$\mathbb{P}\{p\text{-value} \le 5\%\}$	16.89% (0.37%)	19.17% (0.39%)	16.88% (0.37%)	16.78% (0.37%)
$\mathbb{P}\{p\text{-value} \le 1\%\}$	6.78% (0.25%)	8.49% (0.28%)	7.02% (0.26%)	7.03% (0.26%)
$\mathbb{P}\{p\text{-value} \le 0.1\%\}$	1.53% (0.12%)	2.27% (0.15%)	1.87% (0.14%)	2.04% (0.14%)

also see that the exact null distribution depends on the unknown coefficient sequence $\beta_2, \ldots, \beta_p$, since the probabilities between settings differ statistically significantly at all three cutoffs.

To confirm that this non-uniformity is not just a finite-sample effect, we also simulated 10^4 i.i.d. $\mathcal{N}(0,1)$ design matrices with independent Bernoulli(0.5) responses for $n = 500$, $p = 200$ and $n = 5000$, $p = 2000$ as settings (3) and (4), respectively. Table 1 shows that the distribution does not really change as n and p are increased with constant proportion.

These results show that the usual logistic regression p-values one might use when $n \ge p$ can have null distributions that are quite far from uniform, and even if one wanted 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 similarly asymptotic as well as requiring stringent sparsity assumptions (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.

H. Bayesian knockoff statistics

The data for the simulation of Section 4.1.2 was drawn from:

$$\begin{aligned}
&X_j \overset{\text{iid}}{\sim} \mathcal{N}(0, 1/n),\ j \in \{1, \ldots, p\},\\
&\beta_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2),\ j \in \{1, \ldots, p\},\\
&\delta_j \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi),\ j \in \{1, \ldots, p\},\\
&\frac{1}{\sigma^2} \sim \text{Gamma}(A, B) \qquad \text{(shape/scale parameterization, as opposed to shape/rate)},\\
&Y \sim \mathcal{N}\left(\sum_{j:\delta_j=1} X_j\beta_j,\ \sigma^2\right).
\end{aligned}$$

The simulation used $n = 300$, $p = 1000$, with parameter values $\pi = \frac{60}{1000}$, $A = 5$, $B = 4$, 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 $X_1, \dots, X_p$ and $\tilde{X}_1, \dots, \tilde{X}_p$ as fixed):

$$\begin{aligned}
&\beta_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2),\ j \in \{1, \dots, p\},\\
&\tilde{\beta}_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2),\ j \in \{1, \dots, p\},\\
&\lambda_j \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi),\ j \in \{1, \dots, p\},\\
(\delta_j, \tilde{\delta}_j) &\overset{\text{iid}}{\sim} \left\{ \begin{array}{ll} (0,0) & \text{if } \lambda_j = 0 \\ (0,1) \text{ w.p. } 1/2 & \text{if } \lambda_j = 1 \\ (1,0) \text{ w.p. } 1/2 & \text{if } \lambda_j = 1 \end{array} \right\},\ j \in \{1, \dots, p\},\\
&\frac{1}{\sigma^2} \sim \text{Gamma}(A, B) \qquad \text{(shape/scale parameterization, as opposed to shape/rate)},\\
&Y \sim \mathcal{N}\left(\sum_{j:\delta_j = 1} X_j \beta_j + \sum_{j:\tilde{\delta}_j = 1} \tilde{X}_j \tilde{\beta}_j,\ \sigma^2 \right),
\end{aligned}$$

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

$$W_j = \hat{\delta}_j - \hat{\tilde{\delta}}_j.$$

I. Conditional randomization speedups

In order to make the conditional randomization computation feasible for Figure 3, we had to apply a number of speed-ups/shortcuts. With these, the serial computation time for the figure was reduced to roughly three years.

- As mentioned, the LCD statistic is a powerful one for MX knockoffs, so we wanted to compare it to the analogue for the conditional randomization test, namely, the absolute value of the Lasso-estimated coefficient. However, choosing the penalty parameter by cross-validation turned out to be too expensive, so instead we chose a fixed value by simulating repeatedly from the known model (with amplitude 30), running cross-validation for each repetition, and choosing the median of the cross-validation-error-minimizing penalty parameters (the chosen value was 0.00053). For a fair comparison, we did the same for MX knockoffs (except the simulations included knockoff variables too—the chosen value was 0.00077). This speed-up is of course impossible in practice, as the true model is not known. It is not clear how one would choose the penalty parameter if not by cross-validation (at considerable extra computational expense), although a topic of current research is how to choose tuning parameters efficiently and without explicit reliance on a model for the response, e.g., (Lv and Liu, 2014).

- Because the statistic used was the (absolute value of the) estimated coefficient from a sparse estimation procedure, many of the observed statistics were exactly zero, and for these, the conditional randomization p-values can be set to one without further computation. This did not require any prior knowledge, although it will only work for statistics whose distribution has a point mass at the smallest possible value.

- Because the power was calibrated, we knew to expect at least around 10 discoveries, and thus could anticipate the BHq cutoff being at least $0.1 \times 10/600$. This cutoff gives a sense of the p-value resolution needed, and we chose the number of randomizations to be roughly 10 times the inverse of the BHq cutoff, namely, $10,000$. However, we made sure that all $10,000$ randomizations were only computed for very few covariates, both using the previous bullet point and also by checking after periodic numbers of randomizations whether we can reject the hypothesis that the true p-value is below the approximate BHq cutoff upper-bound of $0.1 \times 44/600$ (the 44 comes from 40 nonzeros with 10% false discoveries). For instance, after just 10 randomizations, if the approximate p-value so far is greater than 0.2, we can reject the hypothesis that the exact p-value is below $0.1 \times 44/600$ at significance 0.0001 (and thus compute no further randomizations for that covariate). Speed-ups like this are possible in practice, although they require having an idea of how many rejections will be made, which is not generally available ahead of time.

J. Robustness Simulations

In Figure 7, the points labeled "$100\alpha\%$ Emp. Cov" represents knockoffs run using the following convex combination of the true and empirical covariance matrices:

$$\mathbf{\Sigma}_{\text{EC}} = (1-\alpha)\mathbf{\Sigma} + \alpha\hat{\mathbf{\Sigma}},$$

where $\hat{\mathbf{\Sigma}}$ is the empirical covariance matrix. The point labeled "Graph. Lasso" represents knockoffs run using the following covariance estimate:

$$\mathbf{\Sigma}_{\text{GL}} = \text{diag}(r)\,\hat{\mathbf{\Theta}}^{-1}\,\text{diag}(r)$$

where the rescaling vector r has $r_j = \sqrt{\Sigma_{jj}/(\hat{\mathbf{\Theta}}^{-1})_{jj}}$, and $\hat{\mathbf{\Theta}}$ is the inverse covariance estimated by the graphical Lasso with penalty parameter chosen by 2-fold cross-validation on the log-likelihood.

K. Genetic analysis of Crohn's disease

The SNP arrays came from an Affymetrix 500K chip, with calls made by the BRLMM algorithm Affymetrix (2006). SNPs satisfying any of the following conditions were removed:

- minor allele frequency (MAF) $< 1\%$,
- Hardy–Weinberg equilibrium test p-value $< 0.01\%$,
- missing $> 5\%$ of values (values were considered missing if the BRLMM score was > 0.5),
- position was listed as same as another SNP (this occured just for the pair rs16969329 and rs4886982; the former had smaller MAF and was removed),
- position was not in genetic map.

Furthermore, subjects missing $> 5\%$ of values were removed. Following the original paper describing/using this data (WTCCC, 2007), we did not adjust for population structure. Any missing values that remained after the above preprocessing were replaced by the mean of the nonmissing values for their respective SNPs.

After preprocessing, there were $p = 377{,}749$ single nucleotide polymorphisms (SNPs) measured on $n = 4{,}913$ subjects (1,917 CD patients and 2,996 healthy controls). Although $p \gg n$, the joint dependence of SNPs has a strong spatial structure, and outside data can be used to improve estimation. In particular, we approximated the standardized joint distribution as multivariate Gaussian with covariance matrix estimated using the methodology of Wen and Stephens (2010), which shrinks the off-diagonal entries of the empirical covariance matrix using genetic distance information estimated from the HapMap CEU population. This approximation was used on each chromosome, and SNPs on different chromosomes were assumed to be independent. The statistic we use is the LCD. Although the data itself cannot be made available, all code is available at `http://web.stanford.edu/~msesia/software.html`.

One aspect of SNP data is that it contains some very high correlations, which presents two challenges to our methodology. The first is generic to the variable selection problem: it is very hard to choose between two or more nearly-identical (highly-correlated) variables if the data supports at least one of them being selected.‡ To alleviate this, we clustered the SNPs using the estimated correlations as a similarity measure with a single-linkage cutoff of 0.5, and settle for discovering important SNP clusters. To do so we choose one representative from each cluster and approximate the null hypothesis that a cluster is conditionally independent of the response given the other clusters by the null hypothesis that a cluster *representative* is conditionally independent of the response given the other cluster *representatives*. To choose the representatives, we could ignore the observed data altogether and do something like pick representatives with the highest minor allele frequency (computed from outside data), and then run knockoffs as described in the paper. Although this produces a powerful procedure (about 50% more powerful than the original analysis by WTCCC (2007)), a more powerful approach is to select cluster representatives using a fraction of the observations, including their responses, such as by marginal testing. Note that such a data-splitting approach appears to make our null hypotheses random, as in the work on inference after selection reviewed in Section 1.4. However, the approximation we are making is that each representative stands for its cluster, and each cluster has exactly one associated null hypothesis, no matter how selection is performed, even if it were nonrandom. That is, the *approximate* hypotheses being tested do not actually depend on the selection (unlike Brzyski et al. (2016) where clusters are selected and where the very definition of a cluster actually depends on the selection), and our approach remains free of a model for $Y \mid X$, which together should make it clear that it is still quite different from the literature on inference after selection.

Explicitly, we randomly chose 20% of our observations and on those observations only, we ran marginal t-tests between each SNP and the response. Then from each cluster we chose the SNP with smallest t-test p-value to be the single representative of that cluster. Because the observations used for selecting cluster representatives have had their representative covariate values selected for dependence on the outcome, if we constructed knockoff variables as usual and included them in our procedure, the required exchangeability established in Lemma 3.2 would be violated. However, taking a page from Barber and Candès (2016), we can still use these observations in our procedure by making their knockoff variables just identical copies of the original variables (just for these observations). It is easy to show (see Section L of the Supplementary Material) that constructing knockoffs in this way, as exact copies for the observations used to pick representatives and as usual for the remaining observations, the pairwise exchangeability of null covariates with their knockoffs is maintained. Of course, the observations with identical original and knockoff covariate values do not directly help the procedure distinguish between the original and knockoff variables, but including them improves the accuracy of the fitting procedure used to produce the feature importance statistics, so that power is improved indirectly because the Z_j become more accurate measures of feature importance.

Replacing clusters with single representatives reduces p down to $71,145$ (so the average cluster size was just over five SNPs, although there was substantial variance) and, by construction, upper-bounds pairwise SNP correlations by 0.5. Note that this is far from removing all dependence among the SNPs, so considering conditional independence instead of marginal dependence remains necessary for interpretation. Scientifically, we consider a selected SNP to be a true discovery if and only if it is the representative of a cluster containing a truly important SNP.

The second challenge is the one discussed in Section 3.4, and we use the approximate SDP knockoff construction proposed there. The approximate covariance matrix was just the estimated covariance matrix with zeros outside of the block diagonal, with the blocks chosen by single-linkage clustering on the estimated correlation matrix, aggregating clusters up the dendrogram as much as possible subject to a maximum cluster size of 999. In this case, even separating the problem by chromosome, the SDP

‡This is purely a problem of power and would not affect the Type I error control of knockoffs.

construction was computationally infeasible and the equicorrelated construction produced extremely small s: $\text{mean}(s^{\text{EQ}}) = 0.08$. The parallelized approximate SDP construction took just a matter of hours to run, and increased s on average by almost an order of magnitude, with $\text{mean}(s^{\text{ASDP}}) = 0.57$.

Although it incorporates strong prior information, our estimate of the joint distribution of the SNPs is still an approximation, so our first step is to test the robustness of knockoffs to this approximation.

K.1. Simulations with genetic design matrix

In Section 5, the second and third experiments check the robustness of knockoffs to this particular joint distribution approximation by taking a reasonable model for $Y|X_1, \ldots, X_p$ and simulating artificial response data using the real covariate data itself. We split the rows of our design matrix into 10 smaller data sets (re-estimating the joint covariate distribution each time), and for each conditional model we run knockoffs 10 times and compute the realized FDP each time. Then averaging the 10 FDPs gives an estimate of knockoffs' FDR for that conditional model. In an attempt to make each smaller data set have size more comparable to the $n \approx 5,000$ in our actual experiment, we combined the healthy and CD genetic information with that from 5 other diseases from the same data set:§ coronary artery disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes. This made for $14,708$ samples altogether. To further increase the effective sample size and match the actual experiment, we used a random subset of 1000 observations for choosing cluster representatives, but the *same* 1000 for each smaller data set, so that each subsampled data set contained $\approx 1,400$ unique samples, $+1000$ common samples (for each of these common samples $\tilde{X}_{ij} = X_{ij}$). For computational reasons, we used only the first chromosome in this experiment, so each of our simulations had (pre-clustering) $29,258$ covariates. The conditional model for the response was chosen to be a logistic regression model with 60 nonzero coefficients of random signs and locations uniformly chosen from among the original (not just representatives) SNPs. In the second experiment (Figure 8), the coefficient amplitude was varied, and for each amplitude value, 10 different conditional models (different random locations of the nonzero coefficients) were simulated. Each simulation ran the exact same covariance estimation, SNP clustering and representative selection, knockoff construction, and knockoff selection procedure as used for the real data. In the third experiment (Figure 9), the shrinkage was amplified or attenuated by varying the value m in Wen and Stephens (2010, Equation (2.7)) from its intended value, while the coefficient amplitude was fixed at 14.

K.2. Results on real data

Encouraged by the simulation results of the previous section, we proceeded to run knockoffs with a nominal FDR level of 10% on the full genetic design matrix and real Crohn's disease outcomes. Since knockoffs is a randomized procedure, we re-ran knockoffs 10 times (after choosing the representatives) and recorded the selected SNPs over all repetitions, summarized in Table 2. The serial computation time for a single run of knockoffs was about 6 hours, but the knockoff generation process is trivially parallelizable over chromosomes, so with 20 available computation nodes, the total parallelized computation time was about one hour. Although in this case we certainly do not know the ground truth, we can get some sort of confirmation by comparing to the results of studies with newer and much larger data sets than ours. In particular, we compared with the results of Franke et al. (2010), which used roughly 22,000 cases and 29,000 controls, or about 10 times the sample size of the WTCCC data. We also compare to the WTCCC (2007) results, where the p-value cutoff used was justified as controlling the Bayesian FDR at close to 10%—the same level we use. We consider discovered clusters in different studies to correspond ("Yes" in Table 2) if their position ranges overlap, and to nearly correspond

§Bipolar disorder was also part of this data set, but a formatting error in the data we were given prevented us from including it.

Table 2. SNP clusters discovered to be important for Crohn's disease over 10 repetitions of knockoffs. Clusters not found in Franke et al. (2010) represent promising sites for further investigation, especially rs6601764 and rs4692386, whose nearest genes have been independently linked to CD. See text for detailed description. SNP positions are as listed in the original data, which uses Human Genome Build 35.

Selection frequency	Cluster Representative (Cluster Size)	Chrom.	Position Range (Mb)	Confirmed in Franke et al. (2010)?	Selected in WTCCC (2007)?
100%	rs11805303 (16)	1	67.31–67.46	Yes	Yes
100%	rs11209026 (2)	1	67.31–67.42	Yes	Yes
100%	rs6431654 (20)	2	233.94–234.11	Yes	Yes
100%	rs6601764 (1)	10	3.85–3.85	No	No
100%	rs7095491 (18)	10	101.26–101.32	Yes	Yes
90%	rs6688532 (33)	1	169.4–169.65	Yes	No
90%	rs17234657 (1)	5	40.44–40.44	Yes	Yes
90%	rs3135503 (16)	16	49.28–49.36	Yes	Yes
80%	rs9783122 (234)	10	106.43–107.61	No	No
80%	rs11627513 (7)	14	96.61–96.63	No	No
60%	rs4437159 (4)	3	84.8–84.81	No	No
60%	rs7768538 (1145)	6	25.19–32.91	Yes	No
60%	rs6500315 (4)	16	49.03–49.07	Yes	Yes
60%	rs2738758 (5)	20	61.71–61.82	Yes	No
50%	rs7726744 (46)	5	40.35–40.71	Yes	Yes
50%	rs4246045 (46)	5	150.07–150.41	Yes	Yes
50%	rs2390248 (13)	7	19.8–19.89	No	No
50%	rs7186163 (6)	16	49.2–49.25	Yes	Yes
40%	rs10916631 (14)	1	220.87–221.08	No	No
40%	rs4692386 (1)	4	25.81–25.81	No	No
40%	rs7655059 (5)	4	89.5–89.53	No	No
40%	rs7759649 (2)	6	21.57–21.58	Yes*	No
40%	rs1345022 (44)	9	21.67–21.92	No	No
30%	rs6825958 (3)	4	55.73–55.77	No	No
30%	rs9469615 (2)	6	33.91–33.92	Yes*	No
30%	rs4263839 (23)	9	114.58–114.78	Yes	No
30%	rs2836753 (5)	21	39.21–39.23	No	No
10%	rs459160 (2)	1	44.75–44.75	No	No
10%	rs6743984 (23)	2	230.91–231.05	Yes	No
10%	rs2279980 (20)	5	57.95–58.07	No	No
10%	rs4959830 (11)	6	3.36–3.41	Yes	No
10%	rs13230911 (9)	7	1.9–2.06	No	No
10%	rs7807268 (5)	7	147.65–147.7	No	No
10%	rs2147240 (1)	9	71.83–71.83	No	No
10%	rs10761659 (53)	10	64.06–64.41	Yes	Yes
10%	rs4984405 (3)	15	93.06–93.08	No	No
10%	rs17694108 (1)	19	38.42–38.42	Yes	No
10%	rs3932489 (30)	20	15.01–15.09	No	No

("Yes*" in Table 2) if the distance from our discovered cluster to the nearest cluster in the other study was less than the width of that cluster in the other study.

One thing to notice in the table is that a small number of the discovered clusters actually overlap with other clusters, specifically the clusters represented by rs11805303 and rs11209026 on chromosome 1, and rs17234657 and rs7726744 on chromosome 5. And although they don't overlap one another, the three nearby clusters represented by rs3135503, rs6500315, and rs7186163 on chromosome 16 all overlap the same discovered region in Franke et al. (2010). Although puzzling at first, this phenomenon is readily explained by one of four possibilities:

- By construction, the representatives of overlapping clusters are not very highly-correlated (less than 0.5), so the fact that knockoffs chose multiple clusters in the same region may mean there are multiple important SNPs in this region, with one (or more) in each cluster. Focusing on the clusters on chromosome 1, the same region on the IL23R gene was reported in Franke et al. (2010) to have by far the strongest signal (estimated odds ratio of 2.66, next highest was 1.53) among all the regions they identified. If we conclude from this that the region or gene is fundamentally important for Crohn's disease, it stands to reason that mutations at multiple nearby but distinct loci could have important detrimental effects of their own.

- There could be an important SNP located between the two clusters, but which was not in the data. Then the two clusters on either side would both be conditionally dependent on the response, and knockoffs would be correct to reject them.

- One or more of the overlapping clusters could be mundane false discoveries caused by null covariates that happened to take values more conditionally related to the response than would be typical. This would be a facet of the data itself, and thus an unavoidable mistake.

- The tandem discoveries could also be due to a breakdown in our covariate distribution estimate. If the covariance between the two representatives were substantially underestimated and one had a large effect while the other was null, then the Lasso would have a (relatively) hard time distinguishing the two original variables, but a much easier time separating them from the knockoff of the null representative, since it is less correlated with the signal variable. As a result, the null variable and its knockoff would not be exchangeable as required by knockoffs, and a consistent error could be made. However, given that the empirical and estimated correlations between the two representatives on chromosome 1 were -0.1813 and -0.1811, respectively, and for the two on chromosome 5 were -0.2287 and -0.2286, respectively, it seems unlikely that we have made a gross estimation error. Also, note that the separation of nearly all the discovered regions, along with the simulations of the previous subsection, suggest this effect is at worst very small among our discoveries.

Overlapping clusters aside, the knockoffs results display a number of advantages over the original marginal analysis in WTCCC (2007):

- First, the power is much higher, with WTCCC (2007) making 9 discoveries, while knockoffs made 18 discoveries on average, doubling the power.

- Quite a few of the discoveries made by knockoffs that were confirmed by Franke et al. (2010) were not discovered in WTCCC (2007)'s original analysis.

- Knockoffs made a number of discoveries not found in either WTCCC (2007) or Franke et al. (2010). Of course we expect some (roughly 10%) of these to be false discoveries, particularly towards the bottom of the table. However, especially given the evidence from the simulations of the previous subsection suggesting the FDR is controlled, it is likely that many of these correspond to true discoveries. Indeed, evidence from independent studies about adjacent genes shows some of the top hits to be promising candidates. For example, the closest gene to rs6601764 is KLF6, which has been found to be associated with multiple forms of IBD, including CD and ulcerative colitis (Goodman et al., 2016); and the closest gene to rs4692386 is RBP-J, which has been linked to CD through its role in macrophage polarization (Barros et al., 2013).

Note that these benefits required relatively little customization of the knockoff procedure. For instance, WTCCC (2007) used marginal tests specifically tailored to SNP case-control data, while we simply used the LCD statistic. We conjecture that the careful use of knockoffs by domain experts would compound the advantages of knockoffs, as such users could devise more powerful statistics and better model/cluster the covariates for their particular application.

L. Knockoffs with selection

First, recall that the results of Theorem 3.4 hold if for any subset $S \subset \mathcal{H}_0$, we have

$$([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\text{swap}(S)}, y) \overset{d}{=} ([\boldsymbol{X}, \tilde{\boldsymbol{X}}], y). \tag{A.1}$$

In fact, MX knockoffs are defined in such a way that this property holds. Now, the procedure employed in Section 6 to construct knockoffs is slightly different from that described in the rest of the paper. Explicitly, the data looks like

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \end{bmatrix}, \quad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \end{bmatrix},$$

where $y^{(1)}$ is 983×1, $\boldsymbol{X}^{(1)}$ is $983 \times 71,145$, $y^{(2)}$ is 3930×1 and $\boldsymbol{X}^{(2)}$ is $3930 \times 71,145$; recall that the set of samples labeled (1) is used to select the cluster representatives, and that the two sets (1) and (2) are independent of each other. The knockoffs $\tilde{\boldsymbol{X}}^{(2)}$ for $\boldsymbol{X}^{(2)}$ are generated as described in the paper (using the ASDP construction) and for (1), we set $\tilde{\boldsymbol{X}}^{(1)} = \boldsymbol{X}^{(1)}$. To verify (A.1), it suffices to show that for any subset $S \in \mathcal{H}_0$ and each $i \in \{1, 2\}$,

$$([\boldsymbol{X}^{(i)}, \tilde{\boldsymbol{X}}^{(i)}], y^{(i)}) \overset{d}{=} ([\boldsymbol{X}^{(i)}, \tilde{\boldsymbol{X}}^{(i)}]_{\text{swap}(S)}, y^{(i)}) \tag{A.2}$$

since the sets (1) and (2) are independent. (A.2) holds for $i = 2$ because we are following the classical knockoffs construction. For $i = 1$, (A.2) actually holds with exact equality, and not just equality in distribution.

We can also argue from the perspective of MX knockoffs from Definition 3.1. By construction, it is clear that $\tilde{\boldsymbol{X}}^{(2)}$ are valid MX knockoffs for $\boldsymbol{X}^{(2)}$. We thus study $\tilde{\boldsymbol{X}}^{(1)}$: since $\tilde{\boldsymbol{X}}^{(1)} = \boldsymbol{X}^{(1)}$, the exchangeability property is trivial; also, $\tilde{\boldsymbol{X}}^{(1)} \perp\!\!\!\perp y^{(1)} \mid \boldsymbol{X}^{(1)}$ since $\boldsymbol{X}^{(1)}$ determines $\tilde{\boldsymbol{X}}^{(1)}$.

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 Candès, E. J. (2016). A knockoff filter for high-dimensional selective inference. *arXiv preprint arXiv:1602.03574*.

Barros, M. H. M., Hauck, F., Dreyer, J. H., Kempkes, B., and Niedobitek, G. (2013). Macrophage polarisation: an immunohistochemical approach for identifying m1 and m2 macrophages. *PLoS ONE*, 8(11).

Benjamini, Y. and Heller, R. (2007). False discovery rates for spatial signals. *Journal of the American Statistical Association*, 102(480):1272–1281.

Brzyski, D., Peterson, C. B., Sobczyk, P., Candès, E. J., Bogdan, M., and Sabatti, C. (2016). Controlling the rate of gwas false discoveries. *bioRxiv*.

Chouldechova, A. (2014). *False discovery rate control for spatial data*. PhD thesis, Stanford University.

Doran, G., Muandet, K., Zhang, K., and Schölkopf, B. (2014). A permutation-based kernel conditional independence test. *The 30th Conference on Uncertainty in Artificial Intelligence*.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics*, 42(12):1118–1125.

Freedman, D. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298.

Frommlet, F., Ruhaltinger, F., Twaróg, P., and Bogdan, M. (2012). Modified versions of bayesian information criterion for genome-wide association studies. *Computational Statistics & Data Analysis*, 56(5):1038 – 1051.

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

Goodman, W., Omenetti, S., Date, D., Di Martino, L., De Salvo, C., Kim, G., Chowdhry, S., Bamias, G., Cominelli, F., Pizarro, T., et al. (2016). Klf6 contributes to myeloid cell plasticity in the pathogenesis of intestinal inflammation. *Mucosal immunology*.

Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008). Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. *PLoS Genet*, 4(7):1–8.

Lv, J. and Liu, J. S. (2014). Model selection principles in misspecified models. *Journal of the Royal Statistical Society Series B*, 76:141–167.

Pacifico, M. P., Genovese, C., Verdinelli, I., and Wasserman, L. (2004). False discovery control for random fields. *Journal of the American Statistical Association*, 99(468):1002–1014.

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, (79):pp. 565–574.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Siegmund, D. O., Zhang, N. R., and Yakir, B. (2011). False discovery rate for scanning statistics. *Biometrika*, 98(4):979–985.

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.

Wen, X. and Stephens, M. (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Ann. Appl. Stat.*, 4(3):1158–1182.

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.