ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2018)
80, *Part* 3, *pp.* 551–577

Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection

Emmanuel Candès,

Stanford University, USA

Yingying Fan,

University of Southern California, Los Angeles, USA

Lucas Janson

Stanford University, USA

and Jinchi Lv

University of Southern California, Los Angeles, USA

[Received January 2017. Final revision November 2017]

Summary. Many contemporary large-scale applications involve building interpretable models linking a large set of potential covariates to a response in a non-linear fashion, such as when the response is binary. Although this modelling problem has been extensively studied, it remains unclear how to control the fraction of false discoveries effectively even in high dimensional logistic regression, not to mention general high dimensional non-linear models. To address such a practical problem, we propose a new framework of '*model-X*' knockoffs, which reads from a different perspective the knockoff procedure that was originally designed for controlling the false discovery rate in linear models. Whereas the knockoffs procedure is constrained to homoscedastic linear models with $n \geqslant p$, the key innovation here is that model-X knockoffs provide valid inference from finite samples in settings in which the conditional distribution of the response is arbitrary and completely unknown. Furthermore, this holds no matter the number of covariates. Correct inference in such a broad setting is achieved by constructing knockoff variables probabilistically instead of geometrically. To do this, our approach requires that the covariates are random (independent and identically distributed rows) with a distribution that is known, although we provide preliminary experimental evidence that our procedure is robust to unknown or estimated distributions. To our knowledge, no other procedure solves the *controlled* variable selection problem in such generality but, in the restricted settings where competitors exist, we demonstrate the superior power of knockoffs through simulations. Finally, we apply our procedure to data from a case–control study of Crohn's disease in the UK, making twice as many discoveries as the original analysis of the same data.

Keywords: False discovery rate; Generalized linear models; Genomewide association study; Knockoff filter; Logistic regression; Markov blanket; Testing for conditional independence in non-linear models

1. Introduction

1.1. Panning for gold

Certain diseases have a genetic basis, and an important biological problem is to find which

Address for correspondence: Lucas Janson, Department of Statistics, Harvard University, One Oxford Street Science Center, Cambridge, MA 02138, USA.
E-mail: ljanson@fas.harvard.edu

 1369–7412/18/80551

genetic features (e.g. gene expressions or single-nucleotide polymorphisms) are important for determining a given disease. In healthcare, researchers often want to know which electronic medical record entries determine future medical costs. Political scientists study which demographic or socio-economic variables determine political opinions. Economists are similarly interested in which demographic or socio-economic variables affect future income. Those in the technology industry seek specific software characteristics that they can change to increase user engagement. In the current data-driven science and engineering era, a list of such problems would go on and on. The common theme in all these instances is that we have a deluge of explanatory variables, often many more than the number of observations, knowing full well that the outcome that we wish to understand better actually depends on only a small fraction of them. Therefore, a primary goal in modern 'big data analysis' is to identify those important predictors in a sea of noise variables. Having said this, a reasonable question is why do we have so many covariates in the first place? The answer is twofold: first, because we can. To be sure, it may be fairly easy to measure thousands if not millions of attributes at the same time. For instance, it has become relatively inexpensive to genotype an individual, collecting hundreds of thousands of genetic variations at once. Second, even though we may believe that a trait or phenotype depends on a comparably small set of genetic variations, we have *a priori* no idea about which are relevant and therefore must include them all in our search for those nuggets of gold, so to speak. To complicate matters further, a common challenge in these big data problems, and a central focus of this paper, is that we often have little to no knowledge of how the outcome even depends on the few truly important variables.

To cast the ubiquitous *(variable) selection* problem in statistical terms, call Y the random variable representing the outcome whose determining factors we are interested in, and $X_1, \ldots, X_p$ the set of p potential determining factors or explanatory variables. The object of study is the *conditional* distribution of the outcome Y given the covariates $X = (X_1, \ldots, X_p)$, and we shall denote this conditional distribution function by $F_{Y|X}$. Ideally we would like to estimate $F_{Y|X}$, but in general this is effectively impossible from a finite sample. For instance, even knowing that the conditional density depends on 20 *known* covariates makes the problem impossible unless either the sample size n is astronomically large, and/or we are willing to impose a very restrictive model. However, in most problems $F_{Y|X}$ may realistically be assumed to depend on a small fraction of the p covariates, i.e. the function $F_{Y|X}(y|x_1, \ldots, x_p)$ depends only on a small number of co-ordinates x_i (or is well approximated by such a lower dimensional function). Although this assumption does not magically make the estimation of $F_{Y|X}$ easy, it does suggest consideration of the simpler problem: *which of the many variables does Y depend on?* Often, finding a few of the important covariates—in other words, teasing out the relevant factors from those which are not—is already scientifically extremely useful and can be considered a first step in understanding the dependence between an outcome and some interesting variables; we regard this as a crucial problem in modern data science.

1.2. A peek at our contribution

This paper addresses the selection problem by considering a very general conditional model, where the response Y can depend in an arbitrary fashion on the covariates $X_1, \ldots, X_p$. The only restriction that we place on the model is that the observations $(X_{i1}, \ldots, X_{ip}, Y_i)$ are independently and identically distributed (IID), which is often realistic in high dimensional applications such as genetics, where subjects may be drawn randomly from some large population, or client behavioural modelling, where experiments on a service or user interface go out to a random subset of users. Therefore, the model is simply

$$(X_{i1}, \ldots, X_{ip}, Y_i) \overset{\text{IID}}{\sim} F_{XY}, \qquad i = 1, \ldots, n, \tag{1.1}$$

for some arbitrary $(p+1)$-dimensional joint distribution F_{XY}. We shall assume *no knowledge* of the conditional distribution of $Y|X_1, \ldots, X_p$, but we do assume that the joint distribution of the covariates is known, and we shall denote it by F_X. As a concrete example, consider a case–control experiment to determine the genetic factors which contribute to a rare disease, with diseased subjects oversampled to 50% to increase power. Then the joint distribution of features and disease status obeys model (1.1), where F_X is a 50%–50% mixture of the genetic distributions of diseased and healthy subjects, and Y is the subjects' binary disease state.

In Section 1.3 we shall discuss the merits of this model but we immediately remark on an important benefit: namely, one can pose a meaningful problem. To do this, observe that when we say that the conditional distribution of Y actually depends on a (small) subset $\mathcal{S} \subset \{1, \ldots, p\}$ of the variables $X_1, \ldots, X_p$, which we would like to identify, we mean that we would like to find the 'smallest' subset $\mathcal{S}$ such that, conditionally on $\{X_j\}_{j \in \mathcal{S}}$, Y is independent of all other variables. Another way to say this is that the other variables do not provide additional information about Y. A minimal set $\mathcal{S}$ with this property is usually called a Markov blanket or Markov boundary for Y in the literature on graphical models (Pearl (1988), section 3.2.1). Under very mild conditions about the joint distribution F_{XY}, the Markov blanket is well defined and unique (see Section 2 for details) so that we have a cleanly stated selection problem. Note also that the Markov blanket can be defined purely in terms of $F_{Y|X}$ without any reference to F_X, so that in our case–control example the problem is defined in exactly the same way as if F_X were the true population genetic distribution instead of the oversampled mixture of diseased and healthy genetic distributions.

In most problems of interest, even with the knowledge of F_X, it is beyond hope to recover the blanket $\mathcal{S}$ with no error. Hence, we are naturally interested in procedures that control a type I error, i.e. we would like to find as many variables as possible while not having too many false positive results. In this paper, we focus on controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995), which we can define here as follows: letting $\hat{\mathcal{S}}$ be the outcome of a selection procedure operating on the sampled data (we have used a circumflex because $\hat{\mathcal{S}}$ is random), the FDR is

$$\text{FDR} := \mathbb{E}[\text{FDP}], \qquad \text{FDP} = \frac{\#\{j : j \in \hat{\mathcal{S}} \setminus \mathcal{S}\}}{\#\{j : j \in \hat{\mathcal{S}}\}}, \tag{1.2}$$

where FDP is the false discovery proportion, with the convention $0/0 = 0$. Procedures that control the FDR are interpretable, as they roughly bound what fraction of discoveries are false, and they can be quite powerful as well.

One achievement of this paper is to show that we can design quite powerful procedures that rigorously control the FDR (1.2) in finite samples. This holds no matter the unknown relationship between the explanatory variables X and the outcome Y. We achieve this by rethinking the conceptual framework of Barber and Candès (2015), who originally introduced the knockoff procedure (throughout this paper, we shall sometimes use 'knockoffs' as shorthand for the knockoff framework or procedure). Their salient idea was to construct a set of so-called 'knockoff' variables which were not (conditionally on the original variables) 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 to be considerably more associated with the response than their knockoff counterparts were selected. Their main result was achieving exact finite sample FDR control in the homoscedastic Gaussian linear model when $n \geqslant 2p$ (along with a nearly exact extension to when $p \leqslant n < 2p$). By reading the knockoffs framework from a new perspective, the present paper places *no restriction*

on p relative to n, in sharp constrast with the original knockoffs work which required the low dimensional setting of $n \geqslant p$. The conceptual difference is that the original knockoff procedure treats the X_{ij} as fixed and relies on specific stochastic properties of the linear model, precluding consideration of $p > n$ or non-linear models. By treating the X_{ij} as *random* and relying on that stochasticity instead, the 'model-X' (MX) perspective allows treatment of the high dimensional setting which is increasingly the norm in modern applications. We refer to the new approach as *MX* knockoffs, and by contrast we refer to the original knockoffs approach of Barber and Candès (2015) as '*fixed-X*' (FX) knockoffs. In a nutshell:

(a) we propose a new knockoff construction that is amenable to the random covariate setting (1.1);
(b) as in Barber and Candès (2015) and further reviewed in Section 3, we shall use the knockoff variables as controls in such a way that we can tease apart important variables from noise while controlling the FDR, and *we place no restriction on the dimensionality of the data or the conditional distribution of* $Y|X_1, \ldots, X_p$;
(c) we apply the new procedure to real data from a case–control study of Crohn's disease in the UK; see Section 6, where we show that the new knockoff method makes twice as many discoveries as the original analysis of the same data.

Before turning to the presentation of our method and results, we pause to discuss the merits and limitations of our model, the relationships between this work and others on selective inference and the larger problem of high dimensional statistical testing.

1.3. Relationship with the classical set-up for inference

It may seem to the statistician that our model appears rather different from what she is used to. Our framework is, however, not as exotic as it looks.

1.3.1. Classical set-up

The usual set-up for inference in conditional models 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 to condition on, the covariates. We do the exact opposite by assuming that we know *everything* about the covariate distribution but *nothing* about the conditional distribution $Y|X_1, \ldots, X_p$. Hence, we merely shift the *burden of knowledge*. Our philosophy is, therefore, to model X, not Y, whereas, classically, Y (given X) is modelled and X is not. In practice, the parametric model in the classical approach is just an approximation and does not need to hold exactly to produce useful inference. Analogously, we do not need to know the covariate distribution exactly for our method to be useful, as we shall demonstrate in Sections 5 and 6.

1.3.2. When are our assumptions useful?

We do not claim that our assumptions will always be appropriate, but there are important cases when it is reasonable to think that we know much more about the covariate distribution than about the conditional distribution of the response, including the following cases.

(a) One case is when we in fact know exactly the covariate distribution because we control it, such as in gene knockout experiments (Cong *et al.*, 2013; Peters *et al.*, 2016), genetic crossing experiments (Haldane and Waddington, 1931) or sensitivity analysis of numerical models (Saltelli *et al.*, 2008) (e.g. climate models). In some cases we may also essentially

know the covariate distribution even when we do not control it, such as in admixture mapping (Tang *et al.*, 2006).

(b) Another case is when we have a large amount of unsupervised data (covariate data without corresponding responses or labels) in addition to the n labelled observations. This is quite common in genetic or economic studies, where many other studies will exist that have collected the same covariate information but different response variables.

(c) A third case is when we simply have considerably more prior information about the covariates than about the response. Indeed, the point of many conditional modelling 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 genetic basis of an extremely biologically complex disease by using many comparatively simple single-nucleotide polymorphisms as covariates.

1.3.3. Pay-off

There are substantial pay-offs to our framework. Perhaps the main advantage is the ability to use the knockoff framework in high dimensions: a setting that was impossible by using the original approach. Even in high dimensions, previous inference results rely not only on a parametric model that is often linear and homoscedastic, but also on the sparsity or ultrasparsity of the parameters of that model to achieve some asymptotic guarantee. In contrast, our framework can accommodate *any* model for both the response and the covariates, and our guarantees are exact in finite samples (non-asymptotic). In particular, our set-up encompasses any regression, classification or survival model, including any generalized linear model (GLM), and allows for arbitrary non-linearities and heteroscedasticity, such as are found in many machine learning applications.

1.4. Relationship with work on inference after selection

There is a line of work on inference after selection, or post-selection inference, for high dimensional regression, the goal of which is first to perform selection to make the problem low dimensional, and then to produce p-values that are valid *conditionally* on the selection step (Berk *et al.*, 2013; Lockhart *et al.*, 2014; Lee *et al.*, 2016). These works differ from ours in various ways so we largely see them as complementary activities; see section A of the on-line supplementary material for more detailed explanation of the differences.

1.5. Obstacles to obtaining p-values

Our procedure does not follow the canonical approach to FDR control and multiple testing in general. The canonical approach is to plug p-values into the Benjamini–Hochberg (BH) procedure, which controls the FDR under p-value independence and certain forms of dependence (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). Although these works have seeded a wealth of methodological innovations over the past two decades (Benjamini, 2010), all these procedures act on a set of valid p-values (or equivalent statistics), which they assume can be computed. (Benjamini and Gavrilov (2009) and Bogdan *et al.* (2015) transformed the p-value cut-offs of common FDR controlling procedures into penalized regression analogues to avoid p-values altogether. They only provably control the FDR in homoscedastic linear regression when the design matrix has orthogonal columns (necessitating, importantly, that $n \geqslant p$) but Bogdan *et al.* (2015) empirically retained control more generally whenever the signal obeys sparsity constraints. In a very different setting with spatial hypotheses, Li *et al.* (2016) used

approximations from Gaussian random-field theory to control the FDR heuristically.) The requirement of having valid p-values is quite constraining for general conditional modelling problems.

1.5.1. Regression p-value approximations

In low dimensional ($n \geqslant p$) homoscedastic 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 positive regression dependence on a subset. Already for just the slightly broader class of low dimensional GLMs, we must resort to asymptotic p-values derived from maximum likelihood theory, which we show in section G of the on-line supplementary material can be far from valid in practice. In high dimensional ($n < p$) GLMs, it is not clear how to obtain 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 (usually the number of important variables must be $o\{\sqrt{n}/\log(p)\}$), like their low dimensional maximum likelihood counterparts, these methods suffer from highly non-uniform null p-values in many finite sample problems (see, for example, simulations in Dezeure *et al.* (2015)). For binary covariates, the causal inference literature uses matching and propensity scores for approximately valid inference, but extending these methods to high dimensions is still a topic of current research, requiring similar assumptions and asymptotic approximations to the aforementioned high dimensional GLM literature (Athey *et al.*, 2016). Moving beyond GLMs to the non-parametric setting, there are measures of feature importance, but no p-values. (In their on-line description of random forests (`http://www.math.usu.edu/~adele/forests/`), Leo Breiman and Adele Cutler proposed a way to obtain a 'z-score' for each variable, but without any theoretical distributional justification, and Strobl and Zeileis (2008) found 'that the suggested test is not appropriate for statements of significance'.)

1.5.2. Marginal testing

Faced with the inability to compute p-values for hypothesis tests of conditional independence, one solution is to use *marginal* p-values, i.e. p-values for testing *unconditional* (or marginal) independence between Y and X_j. This simplifies the problem considerably, and many options exist for obtaining valid p-values for such a test. However, marginal p-values are in general *invalid* for testing conditional independence, and replacing tests of conditional independence with tests of unconditional independence is often undesirable; see section B of the on-line supplementary material for a detailed discussion of the drawbacks of marginal testing. Indeed when $p \ll n$, so that classical (e.g. maximum likelihood) inference techniques for regression give valid p-values for parametric tests of conditional independence, it would be very unusual to resort to marginal testing to select important covariates, and we cannot think of a textbook that takes this route. Furthermore, the class of conditional test statistics is far richer than that of marginal statistics and includes the most powerful statistical inference and prediction methodology available. For example, in compressed sensing, the signal recovery guarantees for state of the art l_1-based (joint) algorithms are stronger than any guarantees that are possible with marginal methods. To constrain oneself to marginal testing is to ignore completely the vast modern literature on sparse regression that, although lacking finite sample type I error control, has had tremendous success establishing other useful inferential guarantees such as model selection consistency under high dimensional asymptotics in both parametric (e.g. lasso (Zhao and Yu, 2006; Candès and Plan, 2009)) and non-parametric (e.g. random forests (Wager and Athey, 2016)) settings. Realizing this, the statistical genetics community has worked on several multivariate approaches

to improve power in genomewide association studies by using both penalized (Wu *et al.*, 2009; He and Lin, 2011) and Bayesian regression (Guan and Stephens, 2011; Li *et al.*, 2011), but both approaches still suffer from a lack of type I error control (without making strong assumptions on parameter priors). We shall see that the MX knockoff procedure can leverage the power of any of these techniques while adding rigorous finite sample type I error control when the covariate distribution is known.

1.6. Obtaining valid p-values via conditional randomization testing

If we insist on obtaining p-values for each X_j, there is a simple method when the covariate distribution is assumed known, as it is in this paper. This method is similar in spirit to both propensity scoring (where the distribution of a binary X_j conditional on the other variables is often estimated) and randomization or permutation tests (where X_j is either the only covariate or fully independent of the other explanatory variables). 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 X_j^*s from the *conditional* distribution of X_j given the others (derived from the known F_X) and recomputing the same statistic T_j^* with each new X_j^* in place of X_j; see section F of the on-line supplementary material for details. 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 of one another.

1.7. Outline of the paper

The remainder of the paper is structured as follows: Section 2 frames the controlled selection problem in rigorous mathematical terms. Section 3 introduces the MX knockoff procedure, examines its relationship with the earlier proposal of Barber and Candès (2015), proposes knockoff constructions and feature statistics, and establishes FDR control. Section 4 demonstrates through simulations that the MX knockoff procedure controls the FDR in various settings where no other procedure does, and that, when competitors exist, the knockoff procedure is more powerful. Section 5 gives some preliminary simulations using artificial and real data regarding the robustness of MX knockoffs to unknown or misspecified covariate distributions. Section 6 applies our procedure to a case–control study of Crohn's disease in the UK. Section 7 concludes the paper with extensions and potential lines of future research.

MATLAB and R packages of the code that was used to analyse the data can be found at `https://web.stanford.edu/group/candes/knockoffs/software/knockoff/`.

2. Problem statement

To state the controlled variable selection problem carefully, suppose that we have n IID samples from a population, each of the form (X, Y), where $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. If the conditional distribution of Y actually depends on a smaller subset of these variables, we would like to classify each variable as relevant or not depending on whether it belongs to this subset or not. Mathematically speaking, we are looking for the Markov blanket $\mathcal{S}$, i.e. the 'smallest' subset $\mathcal{S}$ such that, conditionally on $\{X_j\}_{j \in \mathcal{S}}$, Y is independent of all other variables. For almost all joint distributions of (X, Y), there is a unique Markov blanket but there are pathological cases where it does not. An example is this: suppose that X_1 and X_2 are independent Gaussian variables and that $X_3 = X_1 - X_2$. Further assume that the distribution of Y depends on the

vector X only through $X_1 + X_2$, e.g. $Y|X \sim \mathcal{N}(X_1 + X_2, 1)$. Then the set of relevant variables—or, equivalently, the Markov blanket—is ill defined since we can say that the likelihood of Y depends on X through either (X_1, X_2), (X_1, X_3) or (X_2, X_3), all these subsets being equally good. To define a unique set of relevant variables, we shall work with the notion of conditional *pairwise* independence.

Definition 1. A variable X_j is said to be 'null' if and only if Y is independent of X_j conditionally on the other variables $X_{-j} = \{X_1, \ldots, X_p\} \setminus \{X_j\}$. The subset of null variables is denoted by $\mathcal{H}_0 \subset \{1, \ldots, p\}$ and we call a variable X_j 'non-null' or relevant if $j \notin \mathcal{H}_0$.

From now on, *our goal is to discover as many relevant (conditionally dependent) variables as possible while keeping the FDR under control*. (Using the methods of Janson and Su (2016), other error rates such as the k-familywise error rate can also be controlled by using MX knockoffs, but we focus on the FDR for this paper.) Formally, for a selection rule that selects a subset $\hat{S}$ of the covariates,

$$\text{FDR} := \mathbb{E}\left[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}|}\right]. \tag{2.1}$$

In this example, because of the perfect functional relationship $X_3 = X_2 - X_1$, all three variables X_1, X_2 and X_3 would be classified as nulls. Imagine, however, breaking this relationship by adding a little noise, e.g. $X_3 = X_2 - X_1 + Z$, where Z is Gaussian noise (independent of X_1 and X_2) however small. Then, according to our definition, X_1 and X_2 are both non-null whereas X_3 is null—and everything makes sense. Having said this, we should not let ourselves be distracted by such subtleties. In the literature on graphical models there, in fact, are weak regularity conditions that guarantee that the (unique) set of relevant variables defined by pairwise conditional independence exactly coincides with the Markov blanket so there is no ambiguity. In this field, researchers typically assume that these weak regularity conditions hold (examples would include the local and global Markov properties) and proceed from there. For example, Edwards (2000) described these properties on page 8 as holding 'under quite general conditions' and then assumed them for the rest of the book.

Our definition is very natural to anyone working with parametric GLMs. In a GLM, the response Y has a probability distribution taken from an exponential family, which depends on the covariates only through the linear combination $\eta = \beta_1 X_1 + \ldots + \beta_p X_p$. The relationship between Y and X is specified via a link function g such that $\mathbb{E}[Y|X] = g^{-1}(\eta)$. In such models and under broad conditions, $Y \perp\!\!\!\perp X_j | X_{-j}$ if and only if $\beta_j = 0$. In this context, testing the hypothesis that X_j is a null variable is the same as testing $H_j : \beta_j = 0$.

Proposition 1. Take a family of random variables $X_1, \ldots, X_p$ such that one cannot perfectly predict any of them from knowledge of the others. If the likelihood of Y follows a GLM, then $Y \perp\!\!\!\perp X_j | X_{-j}$ if and only if $\beta_j = 0$. Hence, $\mathcal{H}_0$ from definition 1 is exactly the set $\{j : \beta_j = 0\}$.

Proof. We prove proposition 1 in the case of the logistic regression model as the general case is similar. Here, the conditional distribution of Y is Bernoulli with

$$\mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X) = \frac{\exp(\eta)}{1 + \exp(\eta)} = g^{-1}(\eta), \qquad \eta = \beta_1 X_1 + \ldots + \beta_p X_p;$$

note that the assumption about the covariates implies that the model is identifiable. Now assume first that $\beta_j = 0$. Then

$$p_{Y, X_j | X_{-j}}(y, x_j | x_{-j}) = p_{Y | X_j, X_{-j}}(y | x_j, x_{-j}) \, p_{X_j | X_{-j}}(x_j | x_{-j}) \tag{2.2}$$

and, since the first factor on the right-hand side does not depend on X_j, we see that the conditional probability distribution function factorizes. This implies conditional independence. In the other direction, assume that Y and X_j are conditionally independent. Then the likelihood function

$$\frac{\exp\{Y(\beta_1 X_1 + \ldots + \beta_p X_p)\}}{1 + \exp(\beta_1 X_1 + \ldots + \beta_p X_p)}$$

must, conditionally on X_{-j}, factorize into a function of Y times a function of X_j. A consequence of this is that, conditionally on X_{-j}, the odds ratio must not depend on X_j (it must be constant). However, this ratio is equal to $\exp(\beta_j X_j)$ and is constant only if $\beta_j = 0$ since, by assumption, X_j is not determined by X_{-j}.

The assumption regarding the covariates is needed. Indeed, suppose that $X_1 \sim \mathcal{N}(0,1)$ $X_2 = \mathbf{1}\{X_1 > 0\}$ and Y follows a logistic model as above with $\eta = X_1 + X_2$. Then $Y \perp\!\!\!\perp X_2 | X_1$ even though $\beta_2 = 1$. In this example, the conditional distribution of Y depends on (X_1, X_2) only through X_1. Therefore, for identifying important variables (recall that our task is to find important variables and not to learn exactly how the likelihood function depends on these variables), we would like to find X_1 and do not care about X_2 since it provides no new information.

3. Methodology

3.1. Model-X knockoffs

3.1.1. Definition

Definition 2. *MX knockoffs* for the family of random variables $X = (X_1, \ldots, X_p)$ are a new family of random variables $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ constructed with the following two properties:

(a) for any subset $S \subset \{1, \ldots, p\}$ ($=^{\mathrm{d}}$ denotes equality in distribution, and the definition of the swapping operation is given just below),

$$(X, \tilde{X})_{\mathrm{swap}(S)} \stackrel{\mathrm{d}}{=} (X, \tilde{X}); \tag{3.1}$$

(b) $\tilde{X} \perp\!\!\!\perp Y | X$ if there is a response Y.

Property (b) is guaranteed if $\tilde{X}$ is constructed without looking at Y.

Above, the vector $(X, \tilde{X})_{\mathrm{swap}(S)}$ is obtained from $(X, \tilde{X})$ by swapping the entries X_j and $\tilde{X}_j$ for each $j \in S$; for example, with $p = 3$ and $S = \{2, 3\}$,

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)_{\mathrm{swap}(\{2,3\})} \stackrel{\mathrm{d}}{=} (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3).$$

We see from result (3.1) that original and knockoff variables are pairwise exchangeable: taking any subset of variables and swapping them with their knockoffs leaves the joint distribution invariant. Note that our exchangeability condition is on the covariates and thus bears little resemblance to exchangeability conditions for closed permutation testing (see, for example, Westfall and Troendle (2008)). To give an example of MX knockoffs, suppose that $X \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. Then a joint distribution obeying result (3.1) is

$$(X, \tilde{X}) \sim \mathcal{N}(0, \mathbf{G}), \qquad \text{where } \mathbf{G} = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diag}(s) \\ \boldsymbol{\Sigma} - \mathrm{diag}(s) & \boldsymbol{\Sigma} \end{pmatrix}; \tag{3.2}$$

here, $\mathrm{diag}(s)$ is any diagonal matrix selected in such a way that the joint covariance matrix $\mathbf{G}$

is positive semidefinite. Indeed, the distribution that is obtained by swapping variables with their knockoffs is Gaussian with a covariance given by **PGP**, where **P** is the permutation matrix encoding the swap. Since $\mathbf{PGP} = \mathbf{G}$ for any swapping operation, the distribution is invariant. For an interesting connection with the invariance condition in Barber and Candès (2015), see section C of the on-line supplementary material.

We shall soon be interested in the problem of constructing knockoff variables, having observed X. In the above example, a possibility is to sample the knockoff vector $\tilde{X}$ from the conditional distribution

$$\tilde{X} | X \stackrel{\mathrm{d}}{=} \mathcal{N}(\mu, \mathbf{V}),$$

where μ and $\mathbf{V}$ are given by classical regression formulae, namely

$$\mu = X - X\boldsymbol{\Sigma}^{-1} \operatorname{diag}(s),$$
$$\mathbf{V} = 2 \operatorname{diag}(s) - \operatorname{diag}(s)\boldsymbol{\Sigma}^{-1} \operatorname{diag}(s).$$

There are, of course, many other ways of constructing knockoff variables and, for the time being, we prefer to postpone the discussion of more general constructions.

In the setting of the paper, we are given IID pairs $(X_{i1}, \ldots, X_{ip}, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ of covariates and responses, which we can assemble in a data matrix $\mathbf{X}$ and a data vector y in such a way that the ith row of $\mathbf{X}$ is $(X_{i1}, \ldots, X_{ip})$ and the ith entry of y is Y_i. Then the MX knockoff matrix $\tilde{\mathbf{X}}$ is constructed in such a way that, for each observation label i, $(\tilde{X}_{i1}, \ldots, \tilde{X}_{ip})$ is a knockoff for $(X_{i1}, \ldots, X_{ip})$ as explained above; that is to say the joint vector $(X_{i1}, \ldots, X_{ip}, \tilde{X}_{i1}, \ldots, \tilde{X}_{ip})$ obeys the pairwise exchangeability property (3.1).

3.1.2. *Exchangeability of null covariates and their knockoffs*

A crucial property of MX knockoffs is that we can swap null covariates with their knockoffs without changing the joint distribution of the original covariates X and their knockoffs $\tilde{X}$, *conditionally* on the response Y. From now on, $X_{i:j}$ for $i \leqslant j$ is a shorthand for $(X_i, \ldots, X_j)$.

Lemma 1. Take any subset $S \subset \mathcal{H}_0$ of nulls. Then

$$(\mathbf{X}, \mathbf{X}) \mid y \stackrel{\mathrm{d}}{=} (\mathbf{X}, \mathbf{X})_{\mathrm{swap}(S)} \mid y.$$

Proof. The proof of lemma 1 can be found in section D of the on-line supplementary material.

3.2. Feature statistics

To find the relevant variables, we now compute statistics W_j for each $j \in \{1, \ldots, p\}$, a large positive value of W_j providing evidence against the hypothesis that X_j is null. This statistic depends on the response and the original variables but also on the knockoffs, i.e.

$$W_j = w_j\{(\mathbf{X}, \tilde{\mathbf{X}}), y\}$$

for some function w_j. As in Barber and Candès (2015), we impose a *flip sign property*, which says that swapping the jth variable with its knockoff has the effect of changing the sign of W_j. Formally, if $(\mathbf{X}, \tilde{\mathbf{X}})_{\mathrm{swap}(S)}$ is the matrix that is obtained by swapping columns in S,

$$w_j\{(\mathbf{X}, \tilde{\mathbf{X}})_{\mathrm{swap}(S)}, y\} = \begin{cases} w_j\{(\mathbf{X}, \tilde{\mathbf{X}}), y\}, & j \notin S, \\ -w_j\{(\mathbf{X}, \tilde{\mathbf{X}}), y\}, & j \in S. \end{cases} \tag{3.3}$$

In contrast with the aforementioned work, we do not require the sufficiency property that w_j depend on $\mathbf{X}$, $\tilde{\mathbf{X}}$ and y only through $(\mathbf{X}, \tilde{\mathbf{X}})^{\mathrm{T}}(\mathbf{X}, \tilde{\mathbf{X}})$ and $(\mathbf{X}, \tilde{\mathbf{X}})^{\mathrm{T}} y$.

At this point, it may help the reader who is unfamiliar with the knockoff framework to think about knockoff statistics $W = (W_1, \ldots, W_p)$ in two steps: first, consider a statistic T for each original and knockoff variable,

$$T \triangleq (Z, \tilde{Z}) = (Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p) = t\{(\mathbf{X}, \tilde{\mathbf{X}}), y\},$$

with the idea that Z_j and $\tilde{Z}_j$ respectively measure the importance of X_j and $\tilde{X}_j$. Assume the natural property that switching a variable with its knockoff simply switches the components of T in the same way, namely, for each $S \subset \{1, \ldots, p\}$,

$$(Z, \tilde{Z})_{\text{swap}(S)} = t\{(\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(S)}, y\}. \tag{3.4}$$

Then one can construct a W_j obeying the flip sign condition (3.3) by setting

$$W_j = f_j(Z_j, \tilde{Z}_j),$$

where f_j is any antisymmetric function. (An antisymmetric function f is such that $f(v, u) = -f(u, v)$.) (Conversely, any statistic W_j verifying the flip sign condition can be constructed in this fashion.) Adopting this approach, consider a regression problem and run the lasso on the original design augmented with knockoffs,

$$\min_{b \in \mathbb{R}^{2p}} \tfrac{1}{2}\|y - (\mathbf{X}, \tilde{\mathbf{X}})b\|_2^2 + \lambda \|b\|_1 \tag{3.5}$$

and denote the solution by $\hat{b}(\lambda)$ (the first p components are the coefficients of the original variables and the last p are for the knockoffs). Then the *lasso coefficient difference* (LCD) statistic sets $Z_j = |\hat{b}_j(\lambda)|$, $\tilde{Z}_j = |\hat{b}_{j+p}(\lambda)|$ and

$$W_j = Z_j - \tilde{Z}_j = |\hat{b}_j(\lambda)| - |\hat{b}_{j+p}(\lambda)|. \tag{3.6}$$

A large positive value of W_j provides some evidence that the distribution of Y depends on X_j, whereas under the null W_j has a symmetric distribution and, therefore, is equally likely to take positive and negative values, as we shall see next. Before moving on, however, carefully observe that the value of λ in equation (3.6) does not need to be fixed in advance and can be computed from y and $(\mathbf{X}, \tilde{\mathbf{X}})$ in any data-dependent fashion as long as permuting the columns of $\mathbf{X}$ does not change its value; for instance, it can be selected by cross-validation.

Lemma 2. Conditionally on $(|W_1|, \ldots, |W_p|)$, the signs of the null W_js, $j \in \mathcal{H}_0$, are IID coin flips.

Proof. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_p)$ be a sequence of independent random variables such that $\epsilon_j = \pm 1$ with probability $\frac{1}{2}$ if $j \in \mathcal{H}_0$, and $\epsilon_j = 1$ otherwise. To prove the claim, it suffices to establish that

$$W \overset{\text{d}}{=} \epsilon \odot W, \tag{3.7}$$

where '$\odot$' denotes pointwise multiplication, i.e. $\epsilon \odot W = (\epsilon_1 W_1, \ldots, \epsilon_p W_p)$. Now, take ϵ as above and put $S = \{j : \epsilon_j = -1\} \subset \mathcal{H}_0$. Consider swapping variables in S:

$$W_{\text{swap}(S)} \triangleq w\{(\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(S)}, y\}.$$

On the one hand, it follows from the flip sign property that $W_{\text{swap}(S)} = \epsilon \odot W$. On the other hand, lemma 1 implies that $W_{\text{swap}(S)} =^{\text{d}} W$ since $S \subset \mathcal{H}_0$. These last two properties give result (3.7).

In fact, since the pairwise exchangeability property of $(\mathbf{X}, \tilde{\mathbf{X}})$ holds conditionally on y according to lemma 1, the coin flipping property also holds conditionally on y.

3.3. False discovery rate control

From now on, our methodology follows that of Barber and Candès (2015) and we simply rehearse the main ingredients while referring to Barber and Candès (2015) for additional insights. It follows from lemma 2 that the null statistics W_j are symmetric and that, for any fixed threshold $t > 0$,

$$\#\{j : W_j \leqslant -t\} \geqslant \#\{\text{ null } j : W_j \leqslant -t\} \stackrel{\mathrm{d}}{=} \#\{\text{ null } j : W_j \geqslant t\}.$$

Imagine then selecting those variables such that W_j is sufficiently large, e.g. $W_j \geqslant t$; then the false discovery proportion FDP

$$\mathrm{FDP}(t) = \frac{\#\{\text{ null } j : W_j \geqslant t\}}{\#\{j : W_j \geqslant t\}} \tag{3.8}$$

can be estimated via the statistic

$$\widehat{\mathrm{FDP}}(t) = \frac{\#\{j : W_j \leqslant -t\}}{\#\{j : W_j \geqslant t\}}$$

since the numerator is an upward biased estimate of the unknown numerator in equation (3.8). The idea of the knockoff procedure is to choose a data-dependent threshold as liberal as possible while having an estimate of FDP under control. The following theorem shows that estimates of the FDR process can be inverted to give tight FDR control.

Theorem 1. Choose a threshold $\tau > 0$ by setting

$$\tau = \min\left\{t > 0 : \frac{\#\{j : W_j \leqslant -t\}}{\#\{j : W_j \geqslant t\}} \leqslant q\right\} \qquad (\textit{knockoffs}), \tag{3.9}$$

where q is the target FDR level (or $\tau = \infty$ if the set above is empty). (When we write $\min\{t > 0 : \ldots\}$, we abuse the notation since we actually mean $\min\{t \in \mathcal{W}_+ : \ldots\}$, where $\mathcal{W}_+ = \{|W_j| : |W_j| > 0\}$.) Then the procedure selecting the variables

$$\hat{S} = \{j : W_j \geqslant \tau\}$$

controls the modified FDR defined as

$$\mathrm{mFDR} = \mathbb{E}\left[\frac{|\{j \in \hat{S} \cap \mathcal{H}_0\}|}{|\hat{S}| + 1/q}\right] \leqslant q.$$

The slightly more conservative procedure, given by incrementing the number of negatives by 1,

$$\tau_+ = \min\left\{t > 0 : \frac{1 + \#\{j : W_j \leqslant -t\}}{\#\{j : W_j \geqslant t\}} \leqslant q\right\} \qquad (\textit{knockoffs}+), \tag{3.10}$$

and setting $\hat{S} = \{j : W_j \geqslant \tau_+\}$, controls the usual FDR,

$$\mathbb{E}\left[\frac{|\{j \in \hat{S} \cap \mathcal{H}_0\}|}{|\hat{S}| \vee 1}\right] \leqslant q.$$

These results are non-asymptotic and hold no matter the dependence between the response and the covariates—in fact, they hold *conditionally* on the response y.

Table 1. Algorithm 1: sequential conditional independent pairs

$j = 1$ *while* $j \leqslant p$ *do*
  sample $\tilde{X}_j$ from $\mathcal{L}(X_j | X_{-j}, \tilde{X}_{1:j-1})$
  $j = j + 1$
end

The proof is the same as that of theorems 1 and 2 in Barber and Candès (2015)—and, therefore, has been omitted—since all we need is that the null statistics have signs distributed as IID coin flips (even conditionally on y). Note that theorem 1 tells only one side of the story: type I error control; the other very important side is power, which leads us to spend most of the remainder of the paper considering how best to construct knockoff variables and statistics.

3.4. Constructing model-X knockoffs

3.4.1. Exact constructions

We have seen in Section 3.1.1 one way of constructing MX knockoffs in the case where the covariates are Gaussian. How should we proceed for non-Gaussian data? In this regard, the characterization below may be useful.

Proposition 2. The random variables $(\tilde{X}_1, \ldots, \tilde{X}_p)$ are MX knockoffs for $(X_1, \ldots, X_p)$ if and only if, for any $j \in \{1, \ldots, p\}$, the pair $(X_j, \tilde{X}_j)$ is exchangeable conditionally on all the other variables and their knockoffs (and, of course, $\tilde{X} \perp\!\!\!\perp Y | X$).

The proof consists of simple manipulations of the definition and is, therefore, omitted. Our problem can thus also be posed as constructing pairs that are conditionally exchangeable. If the components of the vector X are independent, then any independent copy of X would work, i.e. any vector $\tilde{X}$ independently sampled from the same joint distribution as X would work. With dependent co-ordinates, we may proceed as in algorithm 1 (Table 1).

In algorithm 1 $\mathcal{L}(X_j | X_{-j}, \tilde{X}_{1:j-1})$ is the conditional distribution of X_j given $(X_{-j}, \tilde{X}_{1:j-1})$. When $p = 3$, this would work as follows: sample $\tilde{X}_1$ from $\mathcal{L}(X_1 | X_{2:3})$. Once this has been done, $\mathcal{L}(X_{1:3}, \tilde{X}_1)$ is available and we, therefore, know $\mathcal{L}(X_2 | X_1, X_3, \tilde{X}_1)$. Hence, we can sample $\tilde{X}_2$ from this distribution. Continuing, $\mathcal{L}(X_{1:3}, \tilde{X}_{1:2})$ becomes known and we can sample $\tilde{X}_3$ from $\mathcal{L}(X_3 | X_{1:2}, \tilde{X}_{1:2})$.

It is not immediately clear why algorithm 1 yields a sequence of random variables obeying the exchangeability property (3.1), and we prove this fact in section E of the on-line supplementary material. There is, of course, nothing special about the ordering in which knockoffs are created and equally valid constructions may be obtained by looping through an arbitrary ordering of the variables. For example, in a data analysis application where we would need to build a knockoff copy for each row of the design, independent (random) orderings may be used.

To have power or, equivalently, to have a low type II error rate, it is intuitive that we would like to have original features X_j and their knockoff companions $\tilde{X}_j$ to be as 'independent' as possible.

We do not mean to imply that running algorithm 1 is a simple matter. In fact, it may prove rather complicated since we would have to recompute the conditional distribution at each step; this problem is left for future research. Instead, in this paper we shall work with approximate MX knockoffs and will demonstrate empirically that, for models of interest, such constructions yield FDR control.

3.4.2. Approximate constructions: second-order model-X knockoffs

Rather than asking that $(X, \tilde{X})_{\mathrm{swap}(S)}$ and $(X, \tilde{X})$ have the same distribution for any subset S, we can ask that they have the same first two moments, i.e. the same mean and covariance. Equality of means is a simple matter. As far as the covariances are concerned, equality is equivalent to

$$\mathrm{cov}(X, \tilde{X}) = \mathbf{G}, \qquad \mathbf{G} = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diag}(s) \\ \boldsymbol{\Sigma} - \mathrm{diag}(s) & \boldsymbol{\Sigma} \end{pmatrix}. \tag{3.11}$$

.

We, of course, recognize the same form as in expression (3.2) where the parameter s is chosen to yield a positive semidefinite covariance matrix. (When $(X, \tilde{X})$ is Gaussian, a matching of the first two moments implies a matching of the joint distributions so that we have an exact construction.) Furthermore, section C of the on-line supplementary material shows that the same problem was already solved in Barber and Candès (2015), as the same constraint on s applies but with the empirical covariance replacing the true covariance. This means that the same two constructions as proposed in Barber and Candès (2015) are just as applicable to *second-order MX knockoffs*.

For the remainder of this section, we shall assume that the covariates have each been translated and rescaled to have mean 0 and variance 1. To review, the *equicorrelated* construction uses

$$s_j^{\mathrm{EQ}} = 2\lambda_{\min}(\boldsymbol{\Sigma}) \wedge 1 \qquad \text{for all } j,$$

which minimizes the correlation between variable knockoff pairs subject to the constraint that all such pairs must have the same correlation. The *semidefinite programme* construction solves the convex programme

$$\begin{aligned} \text{minimize} \quad & \sum_j |1 - s_j^{\mathrm{SDP}}| \\ \text{subject to} \quad & s_j^{\mathrm{SDP}} \geqslant 0 \\ & \mathrm{diag}(s^{\mathrm{SDP}}) \preceq 2\boldsymbol{\Sigma}, \end{aligned} \tag{3.12}$$

which minimizes the sum of absolute values of variable knockoff correlations between all suitable s.

In applying these constructions to problems with large p, we run into some new difficulties.

(a) Except for very specially structured matrices like the identity matrix, $\lambda_{\min}(\boldsymbol{\Sigma})$ tends to be extremely small as p grows large. The result is that constructing equicorrelated knockoffs in high dimensions, although fairly computationally easy, will result in very low power, since all the original variables will be nearly indistinguishable from their knockoff counterparts.
(b) For large p, problem (3.12), although convex, is prohibitively computationally expensive. However, if it could be computed, it would produce much larger s_js than the equicorrelated construction and thus be considerably more powerful.

To address these difficulties, we first generalize the two knockoff constructions by the following two-step procedure, which we call the approximate semidefinite programme construction.

Step 1: choose an approximation $\boldsymbol{\Sigma}_{\mathrm{approx}}$ of $\boldsymbol{\Sigma}$ and solve

$$\begin{aligned} \text{minimize} \quad & \sum_j |1 - \hat{s}_j| \\ \text{subject to} \quad & \hat{s}_j \geqslant 0 \\ & \mathrm{diag}(\hat{s}) \preceq 2\boldsymbol{\Sigma}_{\mathrm{approx}}. \end{aligned} \tag{3.13}$$

Step 2: solve

$$\begin{aligned} &\text{maximize} \quad \gamma \\ &\text{subject to} \quad \text{diag}(\gamma \hat{s}) \preceq 2\boldsymbol{\Sigma}, \end{aligned} \tag{3.14}$$

and set $s^{\text{ASDP}} = \gamma \hat{s}$. This problem can be solved quickly by, for example, bisection search over $\gamma \in [0, 1]$.

Approximate semidefinite programming with $\boldsymbol{\Sigma}_{\text{approx}} = \mathbf{I}$ trivially gives $\hat{s}_j = 1$ and $\gamma = 2 \times \lambda_{\min}(\boldsymbol{\Sigma}) \wedge 1$, reproducing the equicorrelated construction. Approximate semidefinite programming with $\boldsymbol{\Sigma}_{\text{approx}} = \boldsymbol{\Sigma}$ clearly gives $\hat{s}_j = s^{\text{SDP}}$ and $\gamma = 1$, reproducing the semidefinite programme construction. Note that the approximate semidefinite programme step 2 is always fast, so the speed of the equicorrelated construction comes largely because the problem *separates* into p computationally independent semidefinite programme subproblems of $\min |1 - \hat{s}_j|$ subject to $0 \leqslant \hat{s}_j \leqslant 2$. However, power is lost because of the very naive approximation $\boldsymbol{\Sigma}_{\text{approx}} = \mathbf{I}$ which results in a very small γ.

In general, we can choose $\boldsymbol{\Sigma}_{\text{approx}}$ to be an m-block-diagonal approximation of $\boldsymbol{\Sigma}$, so that the approximate semidefinite programme from step 1 separates into m smaller, more computationally tractable, and trivially parallelizable semidefinite programme subproblems. If the approximation is fairly accurate, we may also find that γ remains large, so that the knockoffs are nearly as powerful as if we had used the semidefinite programme construction. We demonstrate the approximate semidefinite programme construction in Section 6 when we analyse the Crohn's disease data.

4. Numerical simulations

In this section we demonstrate the importance, utility and practicality of MX knockoffs for high dimensional non-parametric conditional modelling. To emphasize the need for a method like MX knockoffs, we show in section G of the on-line supplementary material that the usual logistic regression p-values that we might use when $n \geqslant p$ can have null distributions that are quite far from uniform.

4.1. Alternative knockoff statistics

As mentioned in Section 3.2, the new MX knockoffs framework allows for a wider variety of W-statistics to be used than in the FX framework. Choices of Z_j include well-studied statistical measures such as the coefficient estimated in a GLM but can also include much more *ad hoc* or heuristic measures such as random-forest bagging feature importances or sensitivity analysis measures such as the Monte-Carlo-estimated total sensitivity index. By providing variable selection with rigorous type I error control for general models and statistics, knockoffs can be used to improve the interpretability of complex black box supervised or machine learning models. There are also many available choices for the antisymmetric function f_j, such as $|Z_j| - |\tilde{Z}_j|$, $\text{sgn}(|Z_j| - |\tilde{Z}_j|) \max\{|Z_j|, |\tilde{Z}_j|\}$, or $\log(|Z_j|) - \log(|\tilde{Z}_j|)$.

The main point of this subsection is that knockoffs can be used as a wrapper around essentially *any* data fitting or prediction algorithm, and regardless of the chosen algorithm still provides rigorous error control for variable selection. We discuss here a few appealing new options for statistics W, but we defer full exploration of these very extensive possibilities to future work.

4.1.1. Adaptive knockoff statistics

The default statistic that was suggested in Barber and Candès (2015) is the lasso signed max

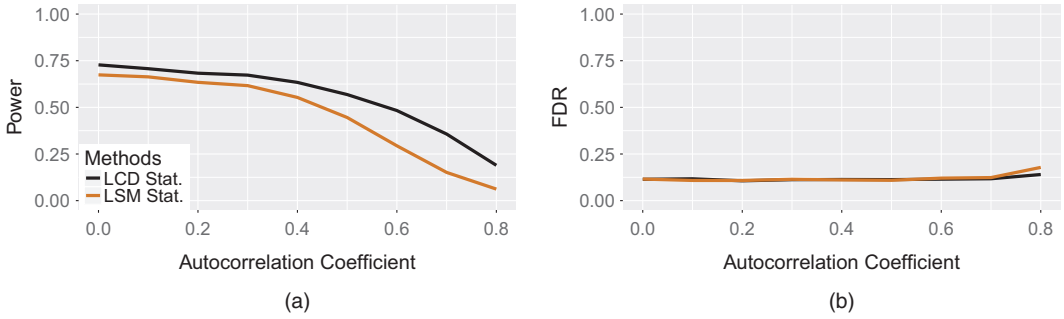

Fig. 1. (a) Power and (b) FDR (the target is 10%) for knockoffs with the LCD and LSM statistics: the design matrix has IID rows and auto-regressive AR(1) columns with auto-correlation coefficient specified by the x-axes of the plots, and marginally each $X_j \sim \mathcal{N}(0, 1/n)$; here $n = 3000$, $p = 1000$ and y comes from a binomial linear model with logit link function with 60 non-zero regression coefficients of magnitude 3.5 and random signs; each point represents 200 replications

(LSM) statistic, which corresponds to Z_j being the largest penalty parameter at which the jth variable enters the model in the lasso regression of y on $(\mathbf{X}, \tilde{\mathbf{X}})$, and $f_j = \operatorname{sgn}(|Z_j| - |\tilde{Z}_j|) \max\{|Z_j|, |\tilde{Z}_j|\}$. In addition to the LSM statistic, Barber and Candès (2015) suggested alternatives such as the difference in absolute values of estimated coefficients for a variable and its knockoff:

$$W_j = |\hat{b}_j| - |\hat{b}_{j+p}|,$$

where the $\hat{b}_j$ and $\hat{b}_{j+p}$ are estimated so that W obeys the sufficiency property that is required by the FX knockoff procedure, e.g. by ordinary least squares or the lasso with a prespecified tuning parameter. The removal of the sufficiency requirement for MX knockoffs enables us to improve this class of statistics by adaptively tuning the fitted model. The simplest example is the LCD statistic that was introduced in Section 3.2, which uses cross-validation to choose the tuning parameter in the lasso. Note that the LCD statistic can be easily extended to any GLM by replacing the first term in expression (3.5) by a non-Gaussian negative log-likelihood, such as in logistic regression; we shall refer to all such statistics generically as LCD. The key is that the tuning and cross-validation are done on the augmented design matrix $(\mathbf{X}, \tilde{\mathbf{X}})$, so that W still obeys the flip sign property.

More generally, MX knockoffs enable us to construct statistics that are highly adaptive to the data, as long as that adaptivity does not distinguish between original and knockoff variables. For instance, we could compute the cross-validated error of the ordinary lasso (still of y on $(\mathbf{X}, \tilde{\mathbf{X}})$) and compare it with that of a random forest and choose Z to be a feature importance measure derived from whichever one has smaller error. Since the lasso works best when the true model is close to linear, whereas random forests work best in non-smooth models, this approach gives us high level adaptivity to the model smoothness, whereas the MX knockoff framework ensures strict type I error control.

Returning to the simpler example of adaptivity, we found that the LCD statistic was uniformly more powerful than the LSM statistic across a wide range of simulations (linear and binomial GLMs, ranging covariate dependence, effect size, sparsity, sample size and total number of variables), 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 using the '1-standard-error' rule, as the latter causes LCD to be underpowered compared with the LSM statistic in low power settings. Fig. 1 shows a

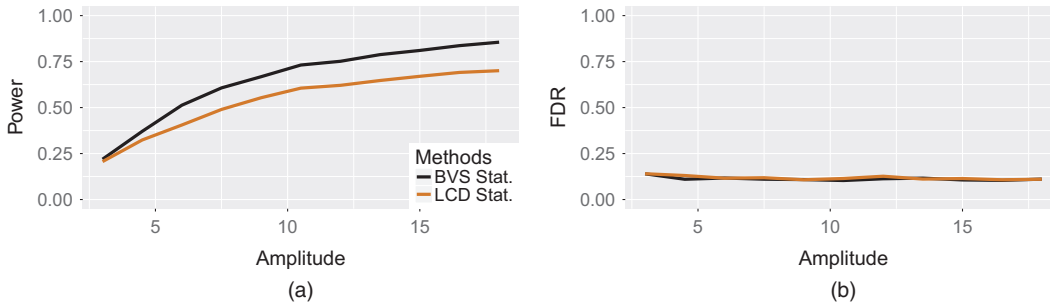

Fig. 2. (a) Power and (b) FDR (the target is 10%) for knockoffs with the LCD and BVS statistics: the design matrix is IID $\mathcal{N}(0, 1/n)$, $n = 300$, $p = 1000$ and y comes from a Gaussian linear model with β and the noise variance randomly chosen (see section H of the on-line supplementary material for the precise model): here, the non-zero entries of β are Gaussian with mean 0 and standard deviation given on the x-axis; the expected number of non-zero components is 60; the expected variance of the noise is 1; each point represents 200 replications

simulation with $n = 3000$ and $p = 1000$ of a binomial linear model (with statistics computed from lasso logistic regression) that is representative of the power difference between the two statistics. In all our simulations, unless otherwise specified, MX knockoffs are always run by using 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.

4.1.2. 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. the difference of absolute posterior mean coefficients, or the difference or log-ratio of posterior probabilities of non-zero coefficients with a sparse prior). What makes this especially appealing is that the statistician obtains 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 hierarchical regression model with $n = 300$ and $p = 1000$, and $\mathbb{E}(\|\beta\|_0) = 60$ ($\|\cdot\|_0$ denotes the l_0-norm, or the number of non-zero entries in a vector); see section H of the on-line supplementary material for details. We chose a simple canonical model with Gaussian response to demonstrate our point, but the same principle applies to more complex, non-linear and non-Gaussian Bayesian models as well. The statistics that we used were the LCD and a Bayesian variable selection statistic, namely $Z_j - \tilde{Z}_j$ where Z_j and $\tilde{Z}_j$ are the posterior probabilities that the jth original and knock-off coefficients are non-zero respectively (George and McCulloch, 1997); again see section H of the on-line supplementary material for details. Fig. 2 shows that the accurate prior information that is 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).

4.2. Alternative procedures

To assess the relative power of knockoffs, we compare with several alternatives in settings in which they are valid:

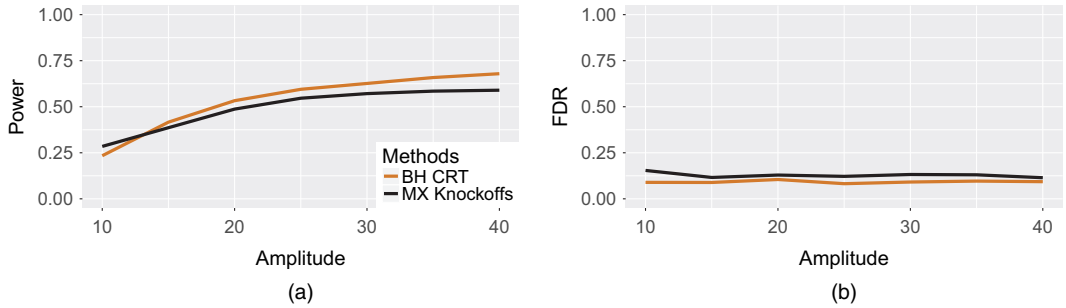

Fig. 3. (a) Power and (b) FDR (the target is 10%) for MX knockoffs and the BH procedure applied to conditional randomization test p-values: the design matrix has IID rows and AR(1) columns with auto-correlation 0.3, $n = 400$, $p = 600$ and y comes from a binomial linear model with logit link function with $\|\beta\|_0 = 40$, and all non-zero entries of β having equal magnitudes and random signs; each point represents 200 replications

(a) the FX knockoff procedure with settings recommended in Barber and Candès (2015) (this method can only be applied in homoscedastic Gaussian linear regression when $n \geqslant p$);
(b) the BH procedure applied to asymptotic GLM p-values—this method can only be applied when $n \geqslant p$ and, although for linear regression exact p-values can be computed (when the maximum likelihood estimator exists), for any other GLM these p-values can be far from valid unless $n \gg p$, as shown in section G of the on-line supplementary material;
(c) the BH procedure applied to marginal test p-values—the correlation between the response and each covariate is computed and compared with its null distribution, which under certain Gaussian assumptions is closed form but in general can at least be simulated exactly by conditioning on y and using the known marginal distribution of 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 *conditional* independence hypotheses when the covariates are exactly independent of one another;
(d) the BH procedure applied to the p-values from the conditional randomization test described in Section 1.6 and section F of the supplementary material.

Note that we are using knockoffs, not 'knockoffs+', in all simulations, and thus we are technically controlling a slightly modified version of FDR. The FDR is nevertheless effectively controlled in all simulations except in extremely low power settings, and even then the violations are small. We could have sacrificed a small amount of power and used 'knockoffs+' (both MX and FX) for exact FDR control, but then a more fair comparison in settings (b)–(d) would replace the BH procedure with the conservative procedure in Benjamini and Yekutieli (2001), since the joint distribution of the p-values will not in general satisfy the assumptions for the BH procedure to control the FDR exactly. However, that conservative procedure had extremely non-competitive power, so we prefer instead to compare knockoffs and the regular BH procedure, which are more powerful and still effectively control the FDR.

4.2.1. Comparison with conditional randomization

We start by comparing MX knockoffs with procedure (d), the BH procedure applied to conditional randomization test p-values, for computational reasons. We simulated $n = 400$ IID rows of $p = 600$ auto-regressive AR(1) covariates with auto-correlation 0.3, and response following a logistic regression model with 40 non-zero coefficients of random signs. Fig. 3 shows the power and FDR curves as the coefficient amplitude was varied. We see that the conditional randomization test gives higher power with similar FDR control, but this comes at a hugely increased

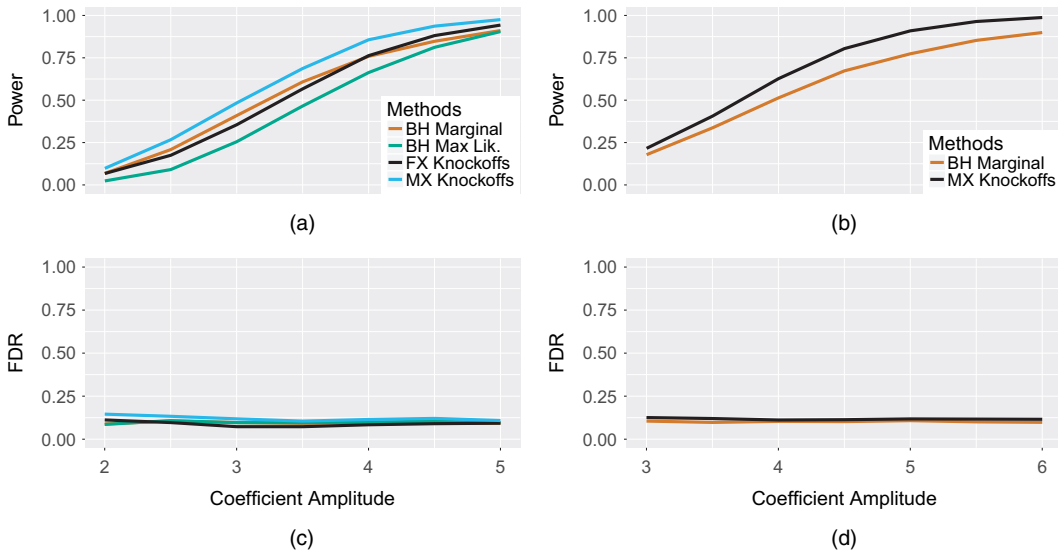

Fig. 4. (a), (b) Power and (c), (d) FDR (the target is 10%) for MX knockoffs and alternative procedures: the design matrix is IID $\mathcal{N}(0, 1/n)$, $n = 3000$, $p =$ ((a), (c)) 1000 and $p =$ ((b), (d)) 6000 and y comes from a Gaussian linear model with 60 non-zero regression coefficients having equal magnitudes and random signs; the noise variance is 1; each point represents 200 replications

computational cost. This simulation has considerably smaller n and p than any other simulation in the paper, and we still had to apply some computational speed-ups or short-cuts, described in section I of the on-line supplementary material, to keep the computation time within reason.

With these speed-ups, Fig. 3 took roughly 3 years of serial computation time, whereas the MX knockoffs component took only about 6 h, or about 1/5000 times as much (all computation was run in MATLAB 2015b (MATLAB, 2015), and both methods used `glmnet` to compute statistics). Because of the heavy computational burden, we could not include the conditional randomization test in our further, larger simulations—we show in section F.3 of the supplementary material that the number of T_j-computations scales optimistically linearly in p. To summarize, conditional randomization testing appears somewhat more powerful than MX knockoffs but is computationally infeasible for large data sets (like that in Section 6).

4.2.2. Effect of signal amplitude

Our first simulation comparing MX knockoffs with procedures (a)–(c) is by necessity in a Gaussian linear model with $n > p$ and independent covariates—the only setting in which all procedures approximately control the FDR. Specifically, Fig. 4(a) plots the power and FDR for the four procedures when $X_{ij} \sim^{\text{IID}} \mathcal{N}(0, 1/n)$, $n = 3000$, $p = 1000$, $\|\beta\|_0 = 60$ the noise variance $\sigma^2 = 1$ and the non-zero entries of β have random signs and equal magnitudes, varied along the x-axis. All methods indeed control the FDR, and the MX knockoff procedure is the most powerful, with as much as 10% higher power than its *nearest* alternative. Fig. 4(b) shows the same set-up but in high dimensions: $p = 6000$. In the high dimensional regime, neither maximum likelihood p-values nor FX knockoffs can even be computed, and the MX knockoff procedure has considerably higher power than the BH procedure 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 knockoff procedure. Fig. 5 shows the same simu-

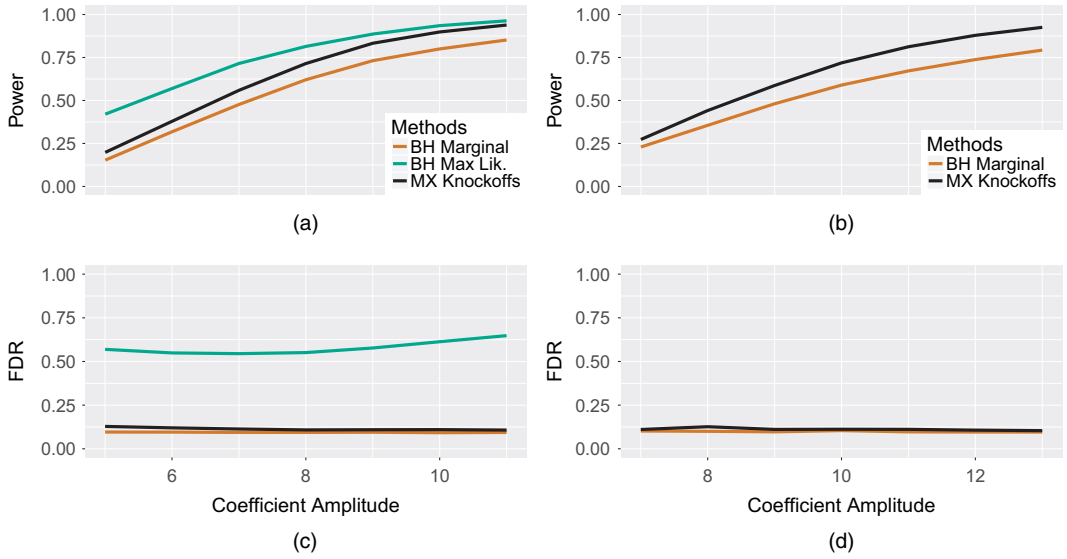

Fig. 5. (a), (b) Power and (c), (d) FDR (the target is 10%) for MX knockoffs and alternative procedures; the design matrix is IID $\mathcal{N}(0, 1/n)$, $n = 3000$, $p =$ ((a), (c)) 1000 and $p =$ ((b), (d)) 6000, and y comes from a binomial linear model with logit link function, and 60 non-zero regression coefficients having equal magnitudes and random signs; each point represents 200 replications

lations as Fig. 4 but with Y following the binomial model. The results are similar to those for the Gaussian linear model, except that the BH procedure applied to the asymptotic maximum 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 G of the supplementary material. In summary, MX knockoffs continue to have the highest power among FDR controlling procedures.

4.2.3. Effect of covariate dependence

To assess the relative power and FDR control of MX knockoffs as a function of covariate dependence, we ran similar simulations to those in the previous section, but with covariates that are AR(1) with varying auto-correlation coefficient (whereas the coefficient amplitude remains fixed). It is now relevant to specify that the locations of the non-zero coefficients are uniformly distributed on $\{1, \ldots, p\}$. For brevity, we show only 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. Fig. 6 shows that, as expected, the BH procedure with marginal testing quickly loses FDR control with increasing covariate dependence. This is because the marginal tests are testing the null hypothesis of *marginal* independence between covariate and response, whereas recall from definition 1 that all conditionally independent covariates are considered null, even if they are marginally dependent on the response. Concentrating on the remaining methods and just the left-hand part of the BH marginal curves where the FDR is controlled, Fig. 6 shows that MX knockoffs continue to be considerably more powerful than alternatives as covariate dependence is introduced, in low and high dimensional linear and non-linear models.

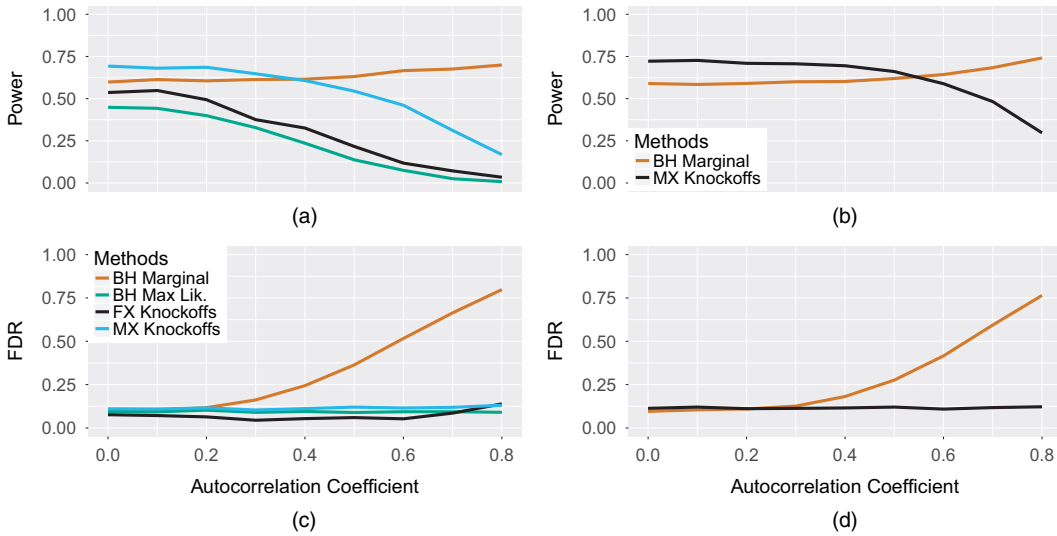

Fig. 6. (a), (b) Power and (c), (d) FDR (the target is 10%) for MX knock-offs and alternative procedures (the design matrix has IID rows and AR(1) columns with auto-correlation coefficient specified by the x-axes of the plots, and marginally each $X_j \sim \mathcal{N}(0, 1/n)$): (a), (c) $n = 3000$, $p = 1000$ and y follows a Gaussian linear model; (b), (d) $n = 3000$, $p = 6000$ and y follows a binomial linear model with logit link function (in both cases, there are 60 non-zero coefficients having magnitudes equal to (a), (c) 3.5 and (b), (d) 10 random signs and randomly selected locations; each point represents 200 replications

5. Robustness

In many real applications, the true joint covariate distribution may not be known exactly, forcing the user to estimate it from the available data. As already mentioned, this is a challenging problem by itself, but often we have considerable outside information or unsupervised data that can be brought to bear to improve estimation. This raises the important question of how robust MX knockoffs are to error in the joint covariate distribution. Theoretical guarantees of robustness are beyond the scope of this paper, but we present instead three compelling simulation studies to demonstrate robustness. The first study investigates error that biases that distribution towards the empirical covariate distribution, which is often referred to as overfitting error, on simulated data. We generated knockoffs for Gaussian variables but, instead of using the true covariance matrix, we used in-sample covariance estimates which ranged in overfitting error. Fig. 7 shows the power and FDR as the covariance that we use ranges from the true covariance matrix (AR(1) with auto-correlation 0.3), to a graphical lasso estimator, to convex combinations of the true and empirical covariance (see section J of the on-line supplementary material for explicit formulae for the estimators). The plot is indexed on the x-axis by the average relative Frobenius norm $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{Fro}} / \|\boldsymbol{\Sigma}\|_{\mathrm{Fro}}$ of the estimator $\hat{\boldsymbol{\Sigma}}$. Although the graphical lasso is well suited for this problem since the covariates have a sparse precision matrix, its covariance estimate is still off by nearly 50%, and yet surprisingly the resulting power and FDR are nearly indistinguishable from when the exact covariance is used. The covariance estimate worsens as the empirical covariance—a very poor estimate of the true covariance given the high dimensionality—is combined in increasing proportion with the truth. At 75% weight on the empirical covariance, the covariance estimate is nearly 100% off and yet the power and FDR of MX knockoffs are only slightly decreased. Beyond this point, MX knockoffs become quite conservative, with power and FDR approaching 0 as the estimated covariance approaches the empirical covariance. This behaviour

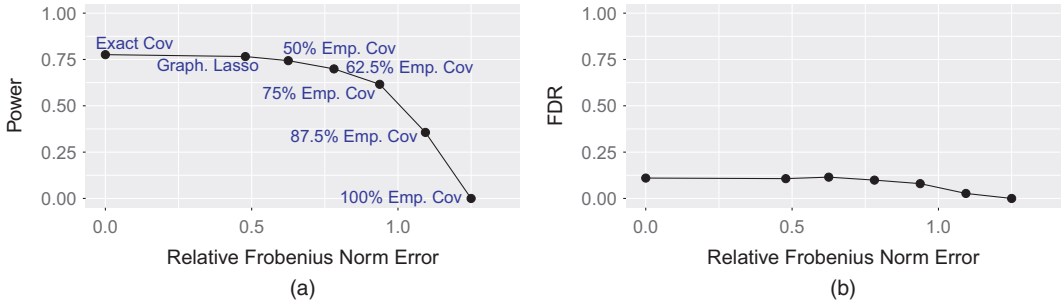

Fig. 7. (a) Power and (b) FDR (the target is 10%) for knockoffs with the LCD statistic as the covariance matrix used to generate knockoffs ranges from the truth to an estimated covariance (see the text for details): the design matrix has IID rows and AR(1) columns with auto-correlation coefficient 0.3, and the matrix (including knockoffs) is standardized so that each column has mean 0 and Euclidean norm 1; here, $n = 800$, $p = 1500$ and y comes from a binomial linear model with logit link function with 50 non-zero entries having magnitude 20 and random signs; each point represents 200 replications

at 100% weight on the empirical covariance is not surprising, since $p > n$ and thus the empirical covariance is rank deficient, forcing the knockoff variables to be exact replicas of their original counterparts. (When the knockoff variables are exact copies of the original variables we are guaranteed zero power and zero FDR since all $W_j = 0$. Although in principle we could break ties and assign signs by coin flips when $W_j = 0$, we prefer only to select X_j with $W_j > 0$, as $W_j = 0$ provides no evidence against the null hypothesis.) The main conclusions from this plot are that

(a) the nominal level of 10% FDR is never violated, even for covariance estimates that are very far from the truth, and
(b) the more overfitting done on the covariance, the more conservative the procedure is, although, even at almost 100% relative error, MX knockoffs had lost about only 20% of the power that they would have had if the covariance were known exactly.

Intuitively, instead of treating the variables as coming from their true joint distribution, MX knockoffs with an overfitted covariate distribution seem to treat them as coming from their true distribution 'conditionally' on being similar to their observed values. Thus FDR should be roughly controlled *conditionally*, which implies marginal FDR control, whereas power may be lost if the conditioning is too great, which matches what we see in the simulations.

Our second and third experiments use real covariate data from a genomewide association study, the details of which are given in Section 6 and section K of the on-line supplementary material. In brief, it is a high dimensional setting with $X_{ij} \in \{0, 1, 2\}$ and strong spatial structure, whose covariance we estimate in sample by using the genomewide association study tailored covariance estimator of Wen and Stephens (2010). We check the robustness of constructing second-order MX knockoffs by approximate semidefinite programming (from Section 3.4.2) by choosing a reasonable but artificial model for $Y|X_1, \ldots, X_p$ and simulating artificial response data by using the real covariate data. The exact details of the simulation are given in section K.1 of the supplementary material, but note that this simulation used the same covariance estimation, single-nucleotide polymorphism clustering and representative selection, knockoff construction and knockoff selection procedure as used for the real data analysis of the next section. Our second experiment varies the signal amplitude in a binomial linear model, and Fig. 8 shows the FDR and power. As hoped, the FDR is consistently controlled over a wide range of powers. Our third experiment, instead of varying the signal strength of our artificial model for $Y|X_1, \ldots, X_p$, deliberately corrupts the covariance estimate of Wen and Stephens (2010) by

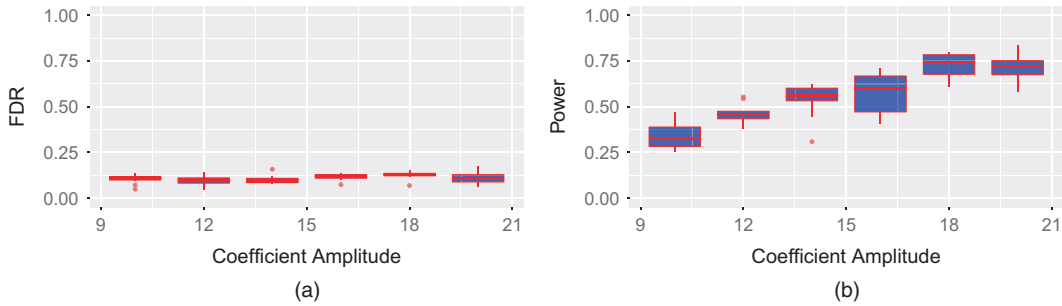

Fig. 8. (a) FDR and (b) power (the target is 10%) for knockoffs with the LCD statistic applied to subsamples of a real genetic design matrix: each boxplot represents 10 different logistic regression models with 60 non-zero coefficients with amplitudes given by the x-axis, and, for each model, 1000 common observations were used for picking cluster representatives, and the remaining 13708 observations were divided into 10 disjoint parts and knockoffs run on each part, with power and FDR for that model then computed by averaging the results over those 10 parts

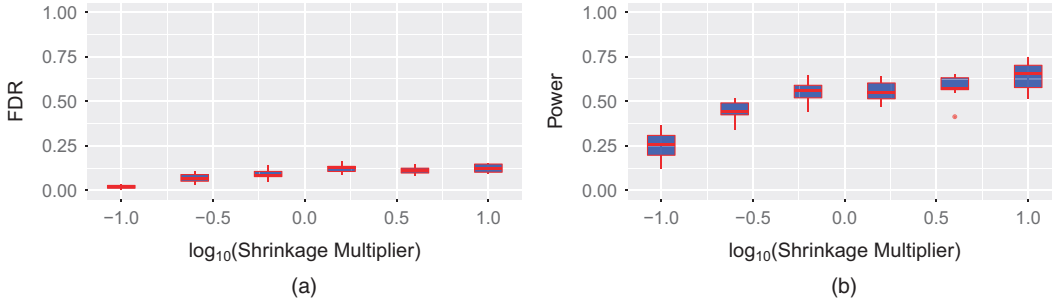

Fig. 9. Set-up the same as Fig. 8 except with the amplitude fixed at 14 and the amount of off-diagonal shrinkage in the covariance estimate varied on the x-axis

varying a shrinkage parameter that is not meant to be varied. That parameter is m in Wen and Stephens (2010), equation (2.7), and we vary it from 1/10 to 10 times its intended value. This variation has a huge effect on how much shrinkage is applied off the diagonal, with the average correlation varying by a factor of about 13 over the range of shrinkage. Fig. 9 shows that, even as we range from substantial undershrinkage to substantial overshrinkage, the MX knockoff procedure never significantly violates FDR control, with only a little conservativeness when the undershrinkage is most drastic (the same phenomenon as in Fig. 7(b)).

6. Genetic analysis of Crohn's disease

To test the robustness and practicality of the new knockoff procedure, we applied it to a data set containing genetic information on cases and controls for Crohn's disease. The data were provided by the Wellcome Trust Case Control Consortium and have been studied previously (Wellcome Trust Case Control Consortium, 2007). They contain $p \approx 400\,000$ single-nucleotide polymorphisms measured on $n \approx 5000$ subjects (approximately 2000 CD patients and approximately 3000 healthy controls). Details of the analysis, including preprocessing, generation of knockoffs, simulations confirming robustness and a full table of discoveries can be found in section K of the supplementary material.

In summary, we ran knockoffs with a nominal FDR level of 10% and the results display some advantages over the original marginal analysis in Wellcome Trust Case Control Consortium (2007), where the p-value cut-off that was used was justified as controlling the Bayesian FDR at close to the same level as we use: 10%.

(a) First, the power is much higher, with Wellcome Trust Case Control Consortium (2007) making nine discoveries, whereas knockoffs made 18 discoveries on average, doubling the power.
(b) Quite a few of the discoveries made by knockoffs were confirmed by a larger genomewide association study (Franke *et al.*, 2010) and were not discovered in the Wellcome Trust Case Control Consortium (2007) original analysis.
(c) Knockoffs made some discoveries that were not found in either Wellcome Trust Case Control Consortium (2007) or Franke *et al.* (2010). Of course we expect some (roughly 10%) of these to be false discoveries. However, especially given the evidence from the simulations of Section 5 suggesting that the FDR is controlled, it is likely that many of these correspond to true discoveries. Indeed, evidence from independent studies about adjacent genes shows that some of the top unconfirmed hits are promising candidates. For example, the closest gene to rs6601764 is KLF6, which has been found to be associated with multiple forms of irritable bowel disease, including Crohn's disease and ulcerative colitis (Goodman *et al.*, 2016), and the closest gene to rs4692386 is RBP-J, which has been linked to Crohn's disease through its role in macrophage polarization (Barros *et al.*, 2013).

7. Discussion

This paper has introduced a novel approach to variable selection in general non-parametric models, which teases apart important from irrelevant variables while guaranteeing type I error control. This approach is a significant rethinking of the knockoff filter from Barber and Candès (2015). A distinctive feature of our approach is that selection is achieved without ever constructing p-values. This is attractive since

(a) p-values are not needed and
(b) it is unclear how they could be efficiently constructed, in general.

(The conditional randomization approach that we proposed is one way of obtaining such p-values but it comes at a computational cost.)

7.1 Deployment in highly correlated settings

We posed a simple question: which variables does a response of interest depend on? In many problems, there may not be enough 'resolution' in the data to tell whether Y depends on X_1 or, instead, on X_2 when the two are strongly correlated. This issue is apparent in our genetic analysis of Crohn's disease from Section 6, where co-located single-nucleotide polymorphisms may be extremely correlated. In such examples, controlling the FDR may not be a fruitful question. A more meaningful question is whether the response appears to depend on a group of correlated variables while controlling for the effects of a number of other variables (e.g. from single-nucleotide polymorphisms in a certain region of the genome while controlling for the effects of single-nucleotide polymorphisms elsewhere on the chromosomes). In such problems, we envision applying our techniques to grouped variables: one possibility is to develop an MX group knockoff approach following Dai and Barber (2016). Another is to construct group representatives and to proceed as we have done in Section 6. It is likely that there are several other ways to formulate a meaningful problem and solution.

7.2. Open questions

Admittedly, this paper may pose more problems than it solves; we close our discussion with a few of them below.

7.2.1. How do we construct model-X knockoffs?

Even though we presented a general strategy for constructing knockoffs, we have essentially skirted this issue except for the important case of Gaussian covariates. It would be important to address this problem, and to write down concrete algorithms for some specific distributions of features of practical relevance.

7.2.2. Which model-X knockoffs?

Even in the case of Gaussian covariates, the question remains about how to choose $\text{corr}(X_j, \tilde{X}_j)$ or, equivalently, the parameter s_j from Section 3 since $\text{corr}(X_j, \tilde{X}_j) = 1 - s_j$. Should we make the marginal correlations small? Should we make the partial correlations small? Should we take an information theoretic approach and minimize a functional of the joint distribution such as the mutual information between X and $\tilde{X}$?

7.2.3. What would we do with multiple model-X knockoffs?

As suggested in Barber and Candès (2015), we could in principle construct multiple knockoff variables $(\tilde{X}^{(1)}, \ldots, \tilde{X}^{(d)})$ in such a way that the $(d+1)p$-dimensional family $(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(d)})$ obeys the following extended exchangeability property: for any variable X_j, any permutation in the list $(X_j, \tilde{X}_j^{(1)}, \ldots, \tilde{X}_j^{(d)})$ leaves the joint distribution invariant. On the one hand, such constructions would yield more accurate information since we could compute, among multiple knockoffs, the rank with which an original variable enters a model. On the other hand, this would constrain the construction of knockoffs a little more, perhaps making them less distinguishable from the original features. What is the right trade-off?

Another point of view is to construct several knockoff matrices exactly as described in the paper. Each knockoff matrix would yield a selection, with each selection providing FDR control as described in this paper. Now an important question is this: is it possible to combine or aggregate all these selections leading to an increase in power while still controlling the FDR?

7.2.4. Can we prove some form of robustness?

Although our theoretical guarantees rely on knowledge of the joint covariate distribution, Section 5 showed preliminary examples with remarkable robustness when this distribution is simply estimated from data. For instance, the estimation of the precision matrix for certain Gaussian designs seems to have rather secondary effects on the FDR and power levels. It would be interesting to provide some theoretical insights into this.

7.2.5. Which feature importance statistics should we use?

The knockoff framework can be seen as an inference machine: the statistician provides the test statistic W_j and the machine performs inference. It is of interest to understand which statistics yield high power, as well as to design new ones.

7.2.6. Can we speed up the conditional randomization testing procedure?

Conditional randomization provides a powerful alternative method for controlling the FDR in MX variable selection, but at a computational cost that is currently prohibitive for large

problems. However, there are several promising directions for speeding it up, including importance sampling to estimate small p-values with fewer randomizations, faster feature statistics T_j with comparable or higher power than the absolute value of lasso-estimated coefficients and efficient computation reuse and warm starts to take advantage of the fact that each randomization changes only a single column of the design matrix.

7.3. Conclusion

Much remains to be done. On the up side, though, we have shown how to select features in high dimensional non-linear models (e.g. GLMs) in a reliable way. This arguably is a fundamental problem, and it is really not clear how else it could be achieved.

Acknowledgements

The authors are listed alphabetically. EC was partially supported by the Office of Naval Research under grant N00014-16-1-2712, and by the Math + X award from the Simons Foundation. YF was partially supported by National Science Foundation Career award DMS-1150318. LJ was partially supported by National Institutes of Health training grant T32GM096982. JL was partially supported by a grant from the Simons Foundation. EC thanks Malgorzata Bogdan, Amir Dembo and Chiara Sabatti for helpful discussions regarding this project. EC also thanks Sabatti for superb feedback regarding an earlier version of the paper. LJ thanks Kaia Mattioli for her help in understanding certain genetic principles.

References

Athey, S., Imbens, G. W. and Wager, S. (2016) Efficient inference of average treatment effects in high dimensions via approximate residual balancing. *Preprint arXiv:1604.07125*. Stanford University, Stanford.

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

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**, no. 11, article e80908.

Benjamini, Y. (2010) Discovering the false discovery rate. *J. R. Statist. Soc.* B, **72**, 405–416.

Benjamini, Y. and Gavrilov, Y. (2009) A simple forward selection procedure based on false discovery rate control. *Ann. Appl. Statist.*, **3**, 179–198.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

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

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013) Valid post-selection inference. *Ann. Statist.*, **41**, 802–837.

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

Candès, E. J. and Plan, Y. (2009) Near-ideal model selection by l_1 minimization. *Ann. Statist.*, **37**, 2145–2177.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A. and Zhang, F. (2013) Multiplex genome engineering using crispr/cas systems. *Science*, **339**, 819–823.

Dai, R. and Barber, R. F. (2016) The knockoff filter for fdr control in group-sparse and multitask regression. *Preprint arXiv:1602.03589*. University of Chicago, Chicago.

Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2015) High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statist. Sci.*, **30**, 533–558.

Edwards, D. (2000) *Introduction to Graphical Modelling*. New York: Springer.

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., Anderson, C. A., Bis, J. C., Bumpstead, S., Ellinghaus, D., Festen, E. M., Georges, M., Green, T., Haritunians, T., Jostins, L., Latiano, A., Mathew, C. G., Montgomery, G. W., Prescott, N. J., Raychaudhuri, S., Rotter, J. I., Schumm, P., Sharma, Y., Simms, L. A., Taylor, K. D., Whiteman, D., Wijmenga, C., Baldassano, R. N., Barclay, M., Bayless, T. M., Brand, S., Buning, C., Cohen, A., Colombel, J.-F., Cottone, M., Stronati, L., Denson, T., De Vos, M., D'Inca, R., Dubinsky, M., Edwards, C., Florin, T., Franchimont, D.,

Gearry, R., Glas, J., Van Gossum, A., Guthery, S. L., Halfvarson, J., Verspaget, H. W., Hugot, J.-P., Karban, A. Laukens, D., Lawrance, I., Lemann, M., Levine, A., Libioulle, C., Louis, E., Mowat, C., Newman, W., Panés, J., Phillips, A., Proctor, D. D., Regueiro, M., Russell, R., Rutgeerts, P., Sanderson, J., Sans, M., Seibold, F., Steinhart, A. H., Stokkers, P. C. F., Torkvist, L., Kullak-Ublick, G., Wilson, D., Walters, T., Targan, S. R., Brant, S. R., Rioux, J. D., D'Arnato, M., Weersma, R. K., Kugathasan, S., Griffiths, A. M., Mansfield, J. C., Vermeire, S., Duerr, R. H., Silverberg, M. S., Satsangi, J., Schreiber, S., Cho, J. H., Annese, V., Hakonarson, H., Daly, M. J. and Parkes, M. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.*, **42**, 1118–1125.

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**, 1166–1202.

George, E. I. and McCulloch, R. E. (1997) Approaches for bayesian variable selection. *Statist. Sin.*, **7**, 339–373.

Goodman, W. A., Omenetti, S., Date, D., Di Martino, L., De Salvo, C., Kim, G.-D., Chowdhry, S., Bamias, G., Cominelli, F., Pizarro, T. and Mahabaleshwar, G. H. (2016) KLF6 contributes to myeloid cell plasticity in the pathogenesis of intestinal inflammation. *Mucsl Immun.*, **9**, 1250–1262.

Guan, Y. and Stephens, M. (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Statist.*, **5**, 1780–1815.

Haldane, J. B. S. and Waddington, C. H. (1931) Inbreeding and linkage. *Genetics*, **16**, 357–374.

He, Q. and Lin, D.-Y. (2011) A variable selection method for genome-wide association studies. *Bioinformatics*, **27**, 1–8.

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

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**, 907–927.

Li, J., Das, K., Fu, G., Li, R. and Wu, R. (2011) The bayesian lasso for genome-wide association studies. *Bioinformatics*, **27**, 516–523.

Li, J., Gahm, J. K., Shi, Y. and Toga, A. W. (2016) Topological false discovery rates for brain mapping based on signal height. *Neuroimage*, to be published.

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

MATLAB (2015) *MATLAB 2015b*. Natick: MathWorks.

Pearl, J. (1988) *Probabilistic Inference in Intelligent Systems*. San Mateo: Morgan Kaufmann.

Peters, J. M., Colavin, A., Shi, H., Czarny, T. L., Larson, M. H., Wong, S., Hawkins, J. S., Lu, C. H. S., Koo, B.-M., Marta, E., Shiver, A. L. and Whitehead, E. H. (2016) A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. *Cell*, **165**, 1493–1506.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S. (2008) *Global Sensitivity Analysis: the Primer*. New York: Wiley.

Strobl, C. and Zeileis, A. (2008) Danger: High power!? — exploring the statistical properties of a test for random forest variable importance . In *COMPSTAT 2008–Proceedings in Computational Statistics*, vol. II (ed P. Brito), pp. 59–66. Heidelberg: Physica.

Tang, H., Coram, M., Wang, P., Zhu, X. and Risch, N. (2006) Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.*, **79**, 1–12.

Wager, S. and Athey, S. (2016) Estimation and inference of heterogeneous treatment effects using random forests. *Preprint arXiv:1510.04342*. Stanford University, Stanford.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

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

Westfall, P. H. and Troendle, J. F. (2008) Multiple testing with minimal assumptions. *Biometr. J.*, **50**, 745–755.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. and Lange, K. (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

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