Biometrika (2019), **106**, 1, *pp.* 1–18
Printed in Great Britain
doi: 10.1093/biomet/asy033
Advance Access publication 4 August 2018

Gene hunting with hidden Markov model knockoffs

BY M. SESIA, C. SABATTI AND E. J. CANDÈS

Department of Statistics, Stanford University, 390 Serra Mall, Stanford, California 94305, U.S.A.

msesia@stanford.edu sabatti@stanford.edu candes@stanford.edu

SUMMARY

Modern scientific studies often require the identification of a subset of explanatory variables. Several statistical methods have been developed to automate this task, and the framework of knockoffs has been proposed as a general solution for variable selection under rigorous Type I error control, without relying on strong modelling assumptions. In this paper, we extend the methodology of knockoffs to problems where the distribution of the covariates can be described by a hidden Markov model. We develop an exact and efficient algorithm to sample knockoff variables in this setting and then argue that, combined with the existing selective framework, this provides a natural and powerful tool for inference in genome-wide association studies with guaranteed false discovery rate control. We apply our method to datasets on Crohn's disease and some continuous phenotypes.

Some key words: False discovery rate; Genome-wide association study; Knockoff; Variable selection.

1. INTRODUCTION

1.1. *The need for controlled variable selection*

Automatic variable selection is a fundamental challenge in statistics, the urgency of which is induced by the growing reliance of many fields of science on the analysis of large amounts of data. As researchers strive to understand increasingly complex phenomena, the technology of high-throughput experiments allows them to measure and simultaneously examine millions of covariates. However, despite the abundance of variables available, often only a fraction of these are expected to be relevant to the question of interest. By discovering which variables are important, scientists can design a more targeted follow-up investigation and hope to understand how certain factors influence an outcome. A compelling example is offered by genome-wide association studies, whose goal is to identify which markers of genetic variation influence the risk of a particular disease or a trait, choosing from up to millions of single-nucleotide polymorphisms. A good selection algorithm should be able to detect as many relevant variables as possible using only a small number of samples, since these tend to be expensive to acquire. It should also ensure that the findings are replicable. Several statistical techniques have been proposed in an effort to address and balance these conflicting needs. The standard approach in genome-wide association studies is to separately compute a p-value for the null hypothesis of no association between the outcome of interest and each polymorphism, using a generalized linear model with one fixed effect and possibly random effects capturing the contribution of all other variables. To identify significant associations, the p-values may be compared to a threshold that guarantees approximate control of the familywise error rate at the 0.05 level, i.e., the probability of committing at least

one Type I error, across all tests. This approach is very conservative and the selected variables, while apparently reproducibly associated with the response, can typically only explain a small portion of the genetic variance in the phenotype of interest (Manolio et al., 2009).

An alternative criterion for evaluating significance is the false discovery rate (Benjamini & Hochberg, 1995). This is attractive when one expects a multiplicity of true discoveries and it has been adopted in studies involving gene expression and many other genomic measurements (Storey & Tibshirani, 2003), including the study of expression quantitative trait loci. A broader adoption of the false discovery rate has been advocated as a natural strategy for improving the power of association studies for complex traits (Sabatti et al., 2003; Storey & Tibshirani, 2003; Brzyski et al., 2017).

Controlled variable selection is inherently difficult in high dimensions, but genome-wide association studies present at least two specific challenges. First, many phenotypes depend on the genetic variants through mechanisms that are mostly unknown (Zuk et al., 2012) and may involve interactions (Carlborg & Haley, 2004). Unfortunately, methods based on marginal testing are ill-equipped to detect interactions and the few current approaches that simultaneously analyse the role of multiple variants rely on linearity assumptions. The second prominent obstacle arises from the presence of correlations between the explanatory variables, as polymorphisms that occupy nearby positions in the genome are tightly linked. This results from the process by which the DNA is transmitted in humans and, as a fundamental characteristic of association studies, it cannot be neglected by methods aiming for valid inference.

These issues motivate the need for methods that can identify important variables for complex phenomena, while providing rigorous guarantees of Type I error control under milder and well-justified assumptions. In the following, we will present our solution and its detailed application to a few studies, after a brief summary of related previous work. Since a few technical terms from genetics appear in this paper, a glossary is included in the Supplementary Material.

1.2. *Model-X knockoffs*

Knockoffs (Candès et al., 2018) partially address the aforementioned issues by taking a radically different path from the traditional literature on high-dimensional variable selection. They provide a powerful and versatile method that rigorously controls the false discovery rate, under no modelling assumptions on the conditional distribution $F_{Y|X}$ of the response Y given the covariates X. In fact, $F_{Y|X}$ may remain completely unspecified. This result is achieved by considering a setting in which the distribution F_X of the covariates is presumed to be known. When this is the case, the latter can be used to generate a new set of artificial variables, the knockoff copies, that serves as a negative control for the original variables. It thus becomes possible to estimate and control the false discovery rate. Since this procedure takes the somewhat unusual path of modelling the covariates instead of the response, we sometimes refer to it as model-X knockoffs. In many circumstances, the premise of model-X knockoffs is arguably more principled than those of its traditional counterparts. In general, it is reasonable to shift the central burden of assumptions from $F_{Y|X}$ to F_X, since the former is the object of inference. In a genome-wide association study, an agnostic approach to the conditional distribution of the response is especially valuable, due to the possibly complex nature of the relations between genetic variants and phenotypes. Moreover, the presumption of knowing F_X is well-grounded, since geneticists have at their disposal a rich set of models for how DNA variants arise and spread across human populations over time. Genetic variation has been assessed in large collections of individuals: the UK Biobank (Sudlow et al., 2015) contains the genotypes of 500 000 subjects, while hundreds of thousands of additional samples are available from the National Center for Biotechnology Information (Mailman et al., 2007). This combination of theoretical knowledge and data gives us a good understanding of F_X.

Since knockoffs require knowledge of the underlying distribution F_X of the original variables, which may not be accessible exactly, in practice some approximation is needed. However, even if the true F_X is known, creating the knockoff copies is in general very difficult. To this date, the only special case for which an algorithm has been developed is that of multivariate Gaussian covariates (Candès et al., 2018). In this sense, knockoffs have not yet fully resolved the second crucial difficulty of association studies mentioned earlier, because a multivariate Gaussian approximation cannot fully take advantage of our prior information on the sequential structure of DNA (Wall & Pritchard, 2003). It thus seems important to develop new techniques that can benefit from advances in the study of population genetics and exploit more accurate parametric models for F_X.

1.3. *Our contributions*

In this paper, we introduce a new algorithm to sample knockoff copies of variables distributed as a hidden Markov model. To the best of our knowledge, this result is the first extension of model-X knockoffs beyond the special case of a Gaussian design, and it involves a class of covariate distributions that is of great practical interest. In fact, hidden Markov models are widely employed to describe sequential data with complex correlations.

While many applications of hidden Markov models are found in the context of speech processing (Juang & Rabiner, 1991) and video segmentation (Boreczky & Wilcox, 1998), their presence has also become nearly ubiquitous in the statistical analysis of biological sequences. Important instances include protein modelling (Krogh et al., 1994), sequence alignment (Hughey & Krogh, 1996), gene prediction (Krogh, 1997), copy number reconstruction (Wang et al., 2007), segmentation of the genome into diverse functional elements (Ernst & Kellis, 2012) and identification of ancestral DNA segments (Falush et al., 2003; Tang et al., 2006; Li & Durbin, 2011). Of special interest to us, following the empirical observation that variation along the human genome could be described by blocks of limited diversity (Patil et al., 2001), hidden Markov models have been broadly adopted to describe haplotypes, i.e., the sequence of alleles at a series of markers along one chromosome. The literature is too extensive to recapitulate: starting from some initial formulations (Stephens et al., 2001; Zhang et al., 2002; Qin et al., 2002; Li & Stephens, 2003), a vast set of models and algorithms is used routinely and effectively to reconstruct haplotypes and to impute missing genotype values. Software implementations include fastPHASE (Scheet & Stephens, 2006), Impute (Marchini et al., 2007; Marchini & Howie, 2010), Beagle (Browning & Browning, 2007, 2011), Bimbam (Guan & Stephens, 2008) and MaCH (Li et al., 2010). The success of these algorithms in reconstructing partially observed genotypes can be tested empirically, and their realized accuracy is a testament to the fact that hidden Markov models offer a good phenomenological description of the dependence between the explanatory variables in genome-wide association studies.

By developing a suitable construction for the knockoffs, we incorporate the prior knowledge on patterns of genetic variation and obtain a new variable selection method that addresses all the critical issues of association studies discussed in § 1.1.

1.4. *Related work*

This paper is most closely related to Candès et al. (2018), which introduced the framework of model-X knockoffs and considered the special case of multivariate Gaussian variables. Earlier work (Barber & Candès, 2015) developed a closely related methodology specific to linear regression with a fixed design matrix, i.e., fixed-X knockoffs. In the interest of simplicity, in the rest of this paper we will refer to model-X knockoffs simply as knockoffs.

Traditional multivariate variable selection techniques have been applied in genome-wide association studies on numerous occasions. Some works have employed penalized regression, but they either lack Type I error control (Hoggart et al., 2008; Wu et al., 2009) or require very restrictive modelling assumptions (Brzyski et al., 2017). Similarly, their Bayesian alternatives (Li et al., 2011; Guan & Stephens, 2011) do not provide finite-sample guarantees. Some have tried to control the Type I errors of standard penalized regression methods through stability selection (Alexander & Lange, 2011), but the resulting procedure does not correctly account for variable correlations and is less powerful than marginal testing. Others have employed machine learning tools (Bureau et al., 2005) that can produce variable importance measures but no valid inference. In theory, some inferential guarantees have been obtained for the lasso (Zhao & Yu, 2006; Candès & Plan, 2009), generalized linear models (van de Geer et al., 2014) and even random forests (Wager & Athey, 2018), but they only hold under rather stringent sparsity assumptions.

Hidden Markov models have appeared before as part of a variable selection procedure for association studies, in order to combine marginal tests of association from correlated polymorphisms (Sun & Cai, 2009; Wei et al., 2009). However, this approach is fundamentally different from ours, since it is not multivariate and makes very different modelling assumptions.

2. CONTROLLED VARIABLE SELECTION VIA KNOCKOFFS

2.1. *Problem statement*

The controlled variable selection problem can be stated in formal terms by adopting the general setting of Candès et al. (2018). Suppose that we can observe a response $Y \in \mathbb{R}$ and a vector of covariates $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$. Given n such samples $(X^{(i)}, Y^{(i)})_{i=1}^n$ drawn from a population, we would like to know which variables are associated with the response. This can be made more precise by assuming that the observations are sampled independently from

$$(X^{(i)}, Y^{(i)}) \sim F_{XY}, \qquad i \in \{1, \ldots, n\},$$

for some joint distribution F_{XY}. The concept of a relevant variable can be understood by first defining its opposite. We say that X_j is null if and only if Y is independent of X_j, conditionally on all other variables $X_{-j} = \{X_1, \ldots, X_p\} \setminus \{X_j\}$. This uniquely defines the set of null covariates $\mathcal{H}_0 = \{j : X_j \text{ is null}\}$ and the complement $\mathcal{S} = \{j : X_j \text{ is relevant}\} = \{1, \ldots, p\} \setminus \mathcal{H}_0$. Our goal is to obtain an estimate $\hat{\mathcal{S}}$ of $\mathcal{S}$ while controlling the false discovery rate, the expected value of the false discovery proportion,

$$\text{FDR} = E\left(\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| \vee 1}\right).$$

We emphasize the logic of this definition: a variable is null if it has no predictive power once we take into account all the other variables, i.e., it does not influence the response in any way. To relate this to traditional inference, in a generalized linear model, being null is equivalent to having a vanishing regression coefficient, under an extremely mild condition (Candès et al., 2018).

2.2. *The limitations of marginal testing*

Although by far the most common data analysis strategy in genome-wide association studies, marginal inference is not necessarily a principled choice, but rather one of convenience. Indeed, the scientific goal is to uncover the genetic basis of complex traits, those that are expected to be

influenced by a large number of possibly interacting genetic variants. In this framework, the most natural model for relating a trait to genetic polymorphisms includes many such DNA variations. Adopting the simplifying additive assumption that is pervasive in genetics, one might be interested in estimating a generalized linear model that relates the trait value to a linear combination of the allele counts at many polymorphisms. Indeed, the statistical genetics literature documents many contributions in this direction, both in the Bayesian (Hoggart et al., 2008; Guan & Stephens, 2011) and in the frequentist (Wu et al., 2009) setting, as more comprehensively reviewed in Sabatti (2013). Yet, approaches that study the effects of many variants jointly, and try to identify the contribution of each one conditional on the rest, have not become part of the standard analysis pipeline for genome-wide data, even if they are the prevalent approach for variants prioritization and follow-up studies (Hormozdiari et al., 2014). This is due to difficulties encountered in articulating an effective genome-wide search for variants that influence the phenotype given every other polymorphism. These range from considerations of computational and data manipulation convenience, e.g., handling of missing data, to the challenge of distinguishing the contribution of highly correlated neighbouring variants, to the fact that, until recently, high-dimensional model selection strategies lacked finite-sample guarantees on the quality of the selected set. The contribution of this paper stems from the observation that this latest impasse can now in principle be overcome by deploying the knockoffs framework (Candès et al., 2018). We will describe how we handle the other difficulties in § 6.1 and § 7. For an up-to-date discussion of the advantages of investigating the effects of a variant in the context of all other recorded polymorphisms, see Brzyski et al. (2017).

2.3. *The method of knockoffs*

The main idea in Candès et al. (2018) is to generate a set of artificial covariates that have the same structure as the original ones but are known to be null. These are called the knockoff copies of X and they can be used as negative controls to estimate the false discovery rate with almost any existing variable selection algorithm. In this paper, we develop new methods for sampling the knockoff copies, but we do not alter other aspects of the variable selection procedure of Candès et al. (2018). Therefore, we only present a short summary below, leaving a more detailed description for the Supplementary Material.

For each variable X_j, we need to construct a knockoff copy $\tilde{X}_j$ in such a way that $X = (X_1, \ldots, X_p)$ and $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ satisfy the following conditions:

$$\tilde{X} \perp\!\!\!\perp Y | X, \tag{1}$$

$$(X, \tilde{X})_{\text{swap}(S)} \stackrel{\text{d}}{=} (X, \tilde{X}), \qquad S \subseteq \{1, \ldots, p\}. \tag{2}$$

Above, the symbol $\stackrel{\text{d}}{=}$ indicates equality in distribution, while $(X, \tilde{X})_{\text{swap}(S)}$ denotes the vector obtained by swapping the entries X_j and $\tilde{X}_j$ for each $j \in S$. The pairwise exchangeability condition (2) requires the distribution of $(X, \tilde{X})$ to be invariant under this transformation. As we discuss later, (2) is essential and it is not always easy to produce a nontrivial, i.e., different from X itself, vector $\tilde{X}$ that satisfies it. We refer to (1) as the nullity condition, since it implies that all knockoff copies are null variables in the augmented model that includes both X and $\tilde{X}$. This clearly holds whenever $\tilde{X}$ is constructed without looking at Y.

Once we have the knockoff copies $\tilde{X}$, we can perform controlled variable selection in two steps. First, we compute feature importance statistics T and $\tilde{T}$, such that T_j and $\tilde{T}_j$ measure the importance of X_j and $\tilde{X}_j$ in predicting Y, for each $j \in \{1, \ldots, p\}$. For example, we can think of T_j

and $\tilde{T}_j$ as the magnitudes of the lasso coefficients for X_j and $\tilde{X}_j$, obtained by regressing Y on X and $\tilde{X}$ jointly, although many other options are available. Then, we combine them into a vector W with p entries defined as $W_j = |T_j| - |\tilde{T}_j|$. Intuitively, a positive and large value of W_j indicates that the jth variable is truly important. More precisely, the knockoff filter of Barber & Candès (2015) is used to compute a data-dependent significance threshold W in such a way as to select important variables with provable control of the false dicovery rate. In summary, knockoffs can be seen as a versatile wrapper that makes it possible to extend rigorous statistical guarantees, under very mild assumptions, to powerful practical methods that would otherwise be too complex for a direct theoretical analysis.

2.4. *Constructing knockoffs*

In § 2.3 we have said that the knockoff variables need to satisfy the nullity and pairwise exchangeability properties, (1) and (2). We now develop exact and computationally efficient procedures for the case in which F_X corresponds to a Markov chain or a hidden Markov model, inspired by following result.

PROPOSITION 1 (Appendix B in Candès et al., 2018). *Let X be a vector of p covariates with some known distribution F_X. Suppose that, with a single iteration over $j = 1, \ldots, p$, we sequentially sample $\tilde{X}_j$ from $p(X_j \mid X_{-j}, \tilde{X}_{1:(j-1)})$, independently of the observed value of X_j. Then, the vector $\tilde{X}$ that we obtain is a knockoff copy of X.*

The conditional distribution above of X_j given all the other variables X_{-j} and $\tilde{X}_{1:(j-1)} = (\tilde{X}_1, \ldots, \tilde{X}_{j-1})$ depends on the knockoff copies generated during the previous iterations, and it can be very difficult to compute in general, even though the distribution of X is known. Therefore, Proposition 1 suggests a general recipe, but obtaining a practical algorithm is not always straightforward.

3. KNOCKOFFS FOR MARKOV CHAINS

We begin by focusing our attention on discrete Markov chains. Formally, we say that a vector of random variables $X = (X_1, \ldots, X_p)$, each taking values in a finite state space $\mathcal{X}$, is distributed as a discrete Markov chain if its joint probability mass function can be written as

$$\mathrm{pr}\big(X_1 = x_1, \ldots, X_p = x_p\big) = q_1(x_1) \prod_{j=2}^{p} Q_j(x_j \mid x_{j-1}), \tag{3}$$

where $q_1(x_1)$ denotes the marginal distribution of the first element of the chain and the transition matrices between consecutive variables are $Q_j(x_j \mid x_{j-1}) = \mathrm{pr}\big(X_j = x_j \mid X_{j-1} = x_{j-1}\big)$.

Our first result, whose proof can be found in the Supplementary Material, provides a way of sampling exact knockoff copies of a discrete Markov chain.

PROPOSITION 2. *Suppose that X is distributed as the Markov chain in (3), with known parameters (q_1, Q). Then, a knockoff copy $\tilde{X}$ can be obtained by sequentially sampling, with a single*

iteration over $j = 1, \ldots, p$, the jth knockoff variable $\tilde{X}_j$ from

$$\mathrm{pr}\big(\tilde{X}_j = \tilde{x}_j \mid x_{-j}, \tilde{x}_{1:(j-1)}\big) = \begin{cases} \dfrac{q_1(\tilde{x}_1)\, Q_2(x_2 \mid \tilde{x}_1)}{\mathcal{N}_1(x_2)}, & j = 1, \\ \dfrac{Q_j(\tilde{x}_j \mid x_{j-1})\, Q_j(\tilde{x}_j \mid \tilde{x}_{j-1})\, Q_{j+1}(x_{j+1} \mid \tilde{x}_j)}{\mathcal{N}_{j-1}(\tilde{x}_j)\, \mathcal{N}_j(x_{j+1})}, & 1 < j < p, \\ \dfrac{Q_p(\tilde{x}_p \mid x_{p-1})\, Q_p(\tilde{x}_p \mid \tilde{x}_{p-1})}{\mathcal{N}_{p-1}(\tilde{x}_p)\, \mathcal{N}_p(1)}, & j = p, \end{cases} \tag{4}$$

with the normalization functions $\mathcal{N}_j : \mathcal{X} \mapsto \mathbb{R}_+$ defined recursively as

$$\mathcal{N}_j(k) = \begin{cases} \displaystyle\sum_{l \in \mathcal{X}} q_1(l)\, Q_2(k \mid l), & j = 1, \\ \displaystyle\sum_{l \in \mathcal{X}} \frac{Q_j(l \mid x_{j-1})\, Q_j(l \mid \tilde{x}_{j-1})\, Q_{j+1}(k \mid l)}{\mathcal{N}_{j-1}(l)}, & 1 < j < p, \\ \displaystyle\sum_{l \in \mathcal{X}} \frac{Q_p(l \mid x_{p-1})\, Q_p(l \mid \tilde{x}_{p-1})}{\mathcal{N}_{p-1}(l)}, & j = p. \end{cases} \tag{5}$$

Therefore, Algorithm 1 is an exact procedure for sampling knockoff copies of a Markov chain.

Algorithm 1. Knockoff copies of a discrete Markov chain.
For $j = 1$ to $j = p$:
  For k in $\mathcal{X}$:
    Compute $\mathcal{N}_j(k)$ according to (5).
  Sample $\tilde{X}_j$ according to (4).

At each step j of Algorithm 1, the evaluation of the normalization function $\mathcal{N}_j(k)$ involves a sum over all elements of the finite state space $\mathcal{X}$ and depends only on the previous $\mathcal{N}_{j-1}(\cdot)$. Since this operation must be repeated for all values of k, sampling the jth knockoff variable requires $O(|\mathcal{X}|^2)$ time, where $|\mathcal{X}|$ is the number of possible states of the Markov chain. This procedure is sequential, generating one knockoff variable at a time. Therefore, the total computation time is $O(p|\mathcal{X}|^2)$, while the required memory is $O(|\mathcal{X}|)$. It is also trivially parallelizable if one wishes to construct a knockoff copy for each of n independent Markov chains. These features make Algorithm 1 efficient and suitable for high-dimensional applications.

4. Knockoffs for hidden Markov models

4.1. *Hidden Markov models*

A hidden Markov model assumes the presence of a latent Markov chain, whose states are not directly visible but conditional on which the observations are independently sampled. Formally, we say that $X = (X_1, \ldots, X_p)$, taking values in a finite state space $\mathcal{X}$, is distributed as a hidden Markov model with K hidden states if there exists a vector $Z = (Z_1, \ldots, Z_p)$ such that

$$\begin{cases} Z \sim \mathrm{MC}\,(q_1, Q) & \text{(latent discrete Markov chain)}, \\ X_j \mid Z \sim X_j \mid Z_j \sim f_j(X_j \mid Z_j) & \text{(emission distribution)}, \end{cases} \tag{6}$$

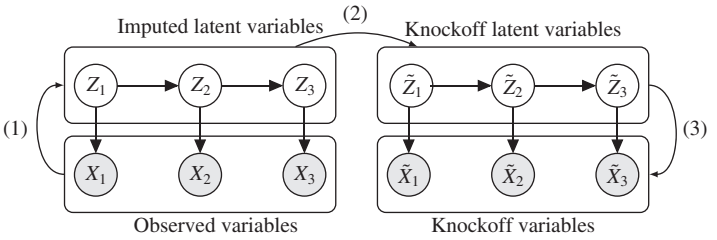

Fig. 1. Sketch of Algorithm 2 for knockoff copies of a hidden Markov model, in the case $p = 3$.

where MC (q_1, Q) indicates the law of a discrete Markov chain as in (3), with each element X_j taking values in $\{1, \ldots, K\}$. Conditional on Z, each X_j is sampled independently from the emission distribution $f_j(X_j \mid Z_j)$. We emphasize that we are restricting our attention to these discrete distributions solely for simplicity. At the price of slightly more involved notation, the knockoff construction can easily be extended to continuous emission distributions.

4.2. *Generating knockoffs for hidden Markov models*

The observed variables X in the hidden Markov model (6) do not satisfy the Markov property. In fact, computing the conditional distributions $p(X_j \mid X_{-j}, \tilde{X}_{1:(j-1)})$ from Proposition 1 would involve a sum over all possible configurations of Z. The complexity of this operation is exponential in p, thus making the naïve approach unfeasible even for moderately large datasets. Our solution is inspired by the traditional forward-backward methods for hidden Markov models. Having observed X, we propose to construct a knockoff copy $\tilde{X}$ according to Algorithm 2.

Algorithm 2. Knockoff copies of a hidden Markov model.
Sample $Z = (Z_1, \ldots, Z_p)$ from $\text{pr}\big(Z \mid X = x\big)$ using Algorithm 3.
Sample a knockoff copy $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_p)$ of $Z = (Z_1, \ldots, Z_p)$ using Algorithm 1.
Sample $\tilde{X}$ from $\text{pr}\big(X \mid Z = \tilde{z}\big)$, which is easy by conditional independence.

A graphical representation of Algorithm 2 is shown in Fig. 1. In the first stage, the latent Markov chain is imputed by sampling from the conditional distribution of Z given X. This is done efficiently with Algorithm 3, a forward-backward iteration discussed in the Supplementary Material and similar to the Viterbi algorithm. Once Z has been sampled, a knockoff copy $\tilde{Z}$ can be obtained with Algorithm 1. Finally, we sample $\tilde{X}$ from $\text{pr}\big(X \mid Z = \tilde{z}\big)$, which is easy because of the conditional independence between the emission distributions in the hidden Markov model.

Algorithm 3. Forward-backward sampling for a hidden Markov model.
Initialize $\alpha_0(k) = 1$, $Q_1(k \mid l) = q_1(k)$ and $Q_{p+1}(k \mid l) = 1$ for all $1 \leqslant k, l \leqslant K$.
For $j = 1$ to $j = p$ (forward pass):
  For $k = 1$ to $k = K$:
    $\alpha_j(k) = f_j(x_j \mid k) \sum_{l=1}^{K} Q_j(k \mid l)\, \alpha_{j-1}(l)$.
For $j = p$ to $j = 1$ (backward pass):
  Sample z_j according to $\pi_j(z_j) = \frac{Q_{j+1}(z_{j+1} \mid z_j)\, \alpha_j(z_j)}{\sum_{k=1}^{K} Q_{j+1}(z_{j+1} \mid k)\, \alpha_j(k)}$.
Return $(z_1, \ldots, z_p)$.

The computation time required by Algorithms 1 and 3 is $O(pK^2)$, while the complexity of the final stage is simply $O(p|\mathcal{X}|)$ because the emission distributions are independent conditional on

the latent Markov chain. Therefore, Algorithm 2 runs in $O\{p(K^2|\mathcal{X}|)\}$ time. The following two results establish the correctness of this approach.

PROPOSITION 3. *Suppose that $X = (X_1, \ldots, X_p)$ is observed from the hidden Markov model in (6), with known parameters (q_1, Q, f). Then, Algorithm 3 produces an exact sample from the conditional distribution of its latent Markov chain $Z = (Z_1, \ldots, Z_p)$ given $X = (X_1, \ldots, X_p)$.*

THEOREM 1. *Suppose that $X = (X_1, \ldots, X_p)$ is observed from the hidden Markov model in (6), with known parameters (q_1, Q, f). Then $(\tilde{X}, \tilde{Z})$ generated by Algorithm 2 is a knockoff copy of (X, Z). That is, for any subset $S \subseteq \{1, \ldots, p\}$,*

$$\left\{(X, \tilde{X})_{\mathrm{swap}(S)}, (Z, \tilde{Z})_{\mathrm{swap}(S)}\right\} \stackrel{\mathrm{d}}{=} \left\{(X, \tilde{X}), (Z, \tilde{Z})\right\}. \tag{7}$$

In particular, this implies that $\tilde{X}$ is a knockoff copy of X.

Proof. It suffices to prove (7), since marginalizing over $(Z, \tilde{Z})$ implies that $(X, \tilde{X})_{\mathrm{swap}(S)}$ has the same distribution as $(X, \tilde{X})$. Conditioning on the values of the latent variables, one can write

$$\begin{aligned}
&\mathrm{pr}\left\{(X, \tilde{X}) = (x, \tilde{x})_{\mathrm{swap}(S)}, (Z, \tilde{Z}) = (z, \tilde{z})_{\mathrm{swap}(S)}\right\} \\
&= \mathrm{pr}\left\{(X, \tilde{X}) = (x, \tilde{x})_{\mathrm{swap}(S)} \mid (Z, \tilde{Z}) = (z, \tilde{z})_{\mathrm{swap}(S)}\right\} \mathrm{pr}\left\{(Z, \tilde{Z}) = (z, \tilde{z})_{\mathrm{swap}(S)}\right\} \\
&= \mathrm{pr}\left\{(X, \tilde{X}) = (x, \tilde{x}) \mid (Z, \tilde{Z}) = (z, \tilde{z})\right\} \mathrm{pr}\left\{(Z, \tilde{Z}) = (z, \tilde{z})_{\mathrm{swap}(S)}\right\} \\
&= \mathrm{pr}\left\{(X, \tilde{X}) = (x, \tilde{x}) \mid (Z, \tilde{Z}) = (z, \tilde{z})\right\} \mathrm{pr}\left\{(Z, \tilde{Z}) = (z, \tilde{z})\right\}.
\end{aligned}$$

The first equality above follows from line 1 of Algorithm 2 and Proposition 3, whose proof can be found in the Supplementary Material. The second equality follows from the conditional independence of the emission distributions in a hidden Markov model. The third equality follows from $\tilde{Z}$ being a knockoff copy of Z, as established in Proposition 2. □

5. HIDDEN MARKOV MODELS IN GENOME-WIDE ASSOCIATION STUDIES

5.1. *Modelling single-nucleotide polymorphisms*

In a genome-wide association study, the response Y is the status of a disease or a quantitative trait of interest, while each sample of X consists of the genotype for a set of single-nucleotide polymorphisms. In particular, we consider the case in which $X \in \{0, 1, 2\}^p$ collects unphased genotypes. For simplicity, in this section we restrict our attention to a single chromosome, since distinct ones are typically assumed to be independent. Different hidden Markov models have been proposed to describe the block-like patterns observed in the distribution of the alleles at adjacent markers, but in this paper we adopt the model implemented in fastPHASE (Scheet & Stephens, 2006) and outlined below. We opt for this model because we find that it has both an intuitive interpretation and remarkable computational efficiency. However, our knockoff construction from § 4 can easily be implemented with other parameterizations.

The unphased genotype of an individual can be seen as the componentwise sum of two unobserved sequences, called haplotypes $H = (H_1, \ldots, H_p)$, where $H_i \in \{0, 1\}$ is a binary variable representing the allele on the ith marker. The main modelling assumption is that the two haplotypes are independent and identically distributed as hidden Markov models. This idea is sketched

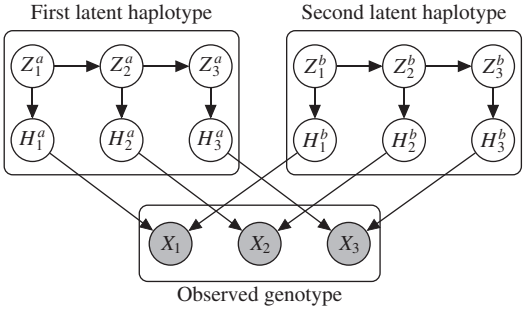

Fig. 2. Sequence of $p = 3$ genotype polymorphisms (shaded) as the componentwise sum of two hidden Markov model haplotypes (white).

in Fig. 2 for $p = 3$. In order to precisely describe this model, we begin by focusing on a single sequence H. Its distribution is in the same form as the model defined earlier in (6),

$$\begin{cases} Z \sim \text{MC}\left(q_1^{\text{h}}, Q^{\text{h}}\right) & \text{(latent Markov chain for one haplotype)}, \\ H_j \mid Z \sim H_j \mid Z_j \sim f_j^{\text{h}}(H_j \mid Z_j) & \text{(haplotype emission distribution)}, \end{cases}$$

with a latent Markov chain $Z = (Z_1, \ldots, Z_p)$ whose elements indicate membership in one of K groups of closely related haplotypes. These groups are characterized by specific allele frequencies at the various markers, so that one can see H as a mosaic of segments, each originating from one of K distinct motifs that can be loosely taken as representing the genome of the population founders. This model provides a good description of the local patterns of correlation, but it is phenomenological in nature and should not be interpreted as an accurate representation of the real sequence of mutations and recombinations that originate the population haplotypes.

The marginal distribution of the first element of the hidden Markov chain Z is

$$q_1^{\text{h}}(k) = \alpha_{1,k}, \qquad k \in \{1, \ldots, K\},$$

while the transition matrices are

$$Q_j^{\text{h}}(k' \mid k) = \begin{cases} \exp(-r_j) + \{1 - \exp(-r_j)\}\,\alpha_{j,k'}, & k' = k, \\ \{1 - \exp(-r_j)\}\,\alpha_{j,k'}, & k' \neq k. \end{cases}$$

The parameters $\alpha = (\alpha_{j,k})_{k \in [K], j \in [p]}$ describe the propensity of different motifs to succeed each other. The occurrence of a transition is regulated by the values of $r = (r_1, \ldots, r_p)$, which are intuitively related to the genetic recombination rates. Once a sequence of ancestral segments is fixed, the allele H_j in position j is sampled from the emission distribution

$$f_j^{\text{h}}(h_j; z_j, \theta) = \begin{cases} 1 - \theta_{j,z_j}, & h_j = 0, \\ \theta_{j,z_j}, & h_j = 1. \end{cases}$$

The parameters $\theta = (\theta_{j,k})_{k \in [K], j \in [p]}$ represent the probabilities of the alleles being equal to 1, for each of the p polymorphisms and the K ancestral haplotype motifs. These can be estimated along with α and r.

Having defined the distribution of H, we return our attention to the observed genotype vector. By definition, the genotype X of an individual is obtained by pairing, marker by marker, the

alleles on each haplotype and discarding information on the haplotype of origin, i.e., the phase. Then, under standard assumptions such as the Hardy–Weinberg equilibrium, the population from which the genotype vector of a subject is randomly sampled can be described as the elementwise sum of two independent and identical haplotype distributions described by the above model. Consequently, its distribution is also a hidden Markov model. The latent Markov chain has bivariate states, corresponding to unordered pairs of haplotype latent states. It is easy to verify that these can take $K(K+1)/2$ possible values. By this construction, it follows that the initial-state probabilities for the genotype model are

$$q_1^{\mathrm{g}}(\{k_a, k_b\}) = \begin{cases} (\alpha_{1,k_a})^2, & k_a = k_b, \\ 2\alpha_{1,k_a}\alpha_{1,k_b}, & k_a \neq k_b, \end{cases} \tag{8}$$

and the transition matrices are

$$Q_j^{\mathrm{g}}(\{k'_a, k'_b\} \mid \{k_a, k_b\}) = \begin{cases} Q_j^{\mathrm{h}}(k'_a \mid k_a)\, Q_j^{\mathrm{h}}(k'_b \mid k_b) + Q_j^{\mathrm{h}}(k'_b \mid k_a)\, Q_j^{\mathrm{h}}(k'_a \mid k_b), & k'_a \neq k'_b, \\ Q_j^{\mathrm{h}}(k'_a \mid k_a)\, Q_j^{\mathrm{h}}(k'_b \mid k_b), & \text{otherwise.} \end{cases} \tag{9}$$

Similarly, the emission probabilities for X_j are

$$f_j(x_j; \{k_a, k_b\}, \theta) = \begin{cases} (1-\theta_{j,k_a})(1-\theta_{j,k_b}), & x_j = 0, \\ \theta_{j,k_a}(1-\theta_{j,k_b}) + (1-\theta_{j,k_a})\theta_{j,k_b}, & x_j = 1, \\ \theta_{j,k_a}\theta_{j,k_b}, & x_j = 2. \end{cases} \tag{10}$$

5.2. *Parameter estimation*

The construction of knockoff copies requires knowing the distribution of the covariates, as discussed in § 2.3. However, exact knowledge is unrealistic in practical applications and some degree of approximation is ultimately unavoidable. Since we have argued that the model in (8)–(10) offers a sensible and tractable description of real genotypes, it makes sense to estimate the $p(2K+1)$ parameters in (r, α, θ) from the data. In the usual setting for genome-wide association studies, one has available $n \gg 2K+1$ observations for each of the p sites, so this task is not unreasonable. Moreover, the validity of this approach is empirically verified in our simulations with real genetic covariates, as discussed in the next section. Alternatively, if additional unsupervised observations, i.e., including only the covariates, from the same population are available, one could include them to improve the estimation.

All parameters can be efficiently estimated with a standard expectation-maximization technique in $O(npK^2)$ time, as already implemented in the freely available imputation software fastPHASE. This fits the model described above, for the original purpose of recovering missing observations, and it conveniently provides the estimates $(\hat{r}, \hat{\alpha}, \hat{\theta})$. An important advantage of the hidden Markov model is that the number of parameters only grows linearly in p, thus greatly reducing the risk of overfitting compared to a multivariate Gaussian approximation. The complexity of this model is controlled by the number K of haplotype motifs, whose typical recommended values are in the range of 10 (Scheet & Stephens, 2006) and can be fine-tuned with crossvalidation. Even though the theoretical guarantee of false discovery rate control with knockoffs requires F_X to be known, we have observed that our procedure is robust with respect to estimation, by performing several numerical experiments discussed in the next section and in the Supplementary Material.

6. Numerical simulations

6.1. *Numerical simulation with real genetic covariates*

We now verify the power and robustness of our procedure with real covariates obtained from a genome-wide association study. We consider 29 258 polymorphisms on chromosome 1, genotyped in 14 708 individuals from WTCCC (2007). Following Candès et al. (2018), we simulate the response from a conditional logistic regression model of $Y \mid X$ with 60 nonzero coefficients, as described in the Supplementary Material.

Before applying our procedure, we reduce the number of covariates by pruning. This is desirable due to the presence of extremely high correlations between neighbouring sites, which makes it fundamentally impossible to distinguish nearly identical variables with a limited amount of data. Our solution uses hierarchical clustering to identify groups of sites in such a way that no two polymorphisms in different clusters have correlation greater than 0.5. Then, within each group we identify a single representative that is most strongly associated with the phenotype in a hold-out set of 1000 observations, described in more detail in the Supplementary Material. At this point, we will use knockoffs to perform variable selection on the cluster representatives, thus effectively interpreting these groups as the basic units of inference among which we search for important variables. Far from removing all correlations and making variable selection trivial, by pruning we acknowledge that a limited amount of data only allows limited resolution. Had we more data, we would prune less. This approach is also consistent with the common practice in genome-wide association studies of interpreting findings as identifying regions in the genome rather than as individual polymorphisms.

Having reduced the number of variables to 5260 by pruning, we split the samples, i.e., the rows of X, into 10 subsets and separately fit the model of § 5.1 with fastPHASE, using the default settings and assuming the presence of 12 latent haplotype clusters. Once the parameters are estimated, we construct the knockoff copies using Algorithm 2. With our implementation, this takes approximatively 0.1 seconds on a single core of a 2.60 GHz Intel Xeon CPU for each individual. We run the knockoffs procedure on each split, adopting as variable importance measures the magnitudes of the logistic regression coefficients fitted with a ℓ_1-norm penalty tuned by crossvalidation. The knockoff filter is then applied at level $\alpha = 0.1$ and with offset equal to 0. The power and proportion of false discoveries are estimated by comparing our selections to the true logistic model, counting a finding as true if and only if any of the polymorphisms in the selected cluster has a nonzero coefficient. The entire experiment is repeated 10 times, starting with the choice of the logistic model. This yields 100 point estimates for the power and false discovery rate, whose empirical distribution is shown in Fig. 3 and Table 1, for different values of the signal amplitude. We have also applied the knockoff filter with offset, i.e., its slightly more conservative version, as explained in the Supplementary Material. As shown in Table 1, the value of the offset is of little practical consequence, except when very few discoveries are made, i.e., for a weak signal.

The results show that the false discovery rate is controlled and suggest that one can safely apply our method to a genome-wide association study. Our confidence derives from the fact that our procedure enjoys the rigorous robustness of knockoffs for any conditional distribution of the phenotype. As far as Type I error control is concerned, it does not seem consequential that in this experiment we have chosen to simulate the response from a generalized linear model. In fact, the false discovery rate is provably controlled for any $F_{Y|X}$, provided that F_X is well-specified. Since we have not artificially simulated the covariates but used real genotypes, we can see no reason why our procedure should not similarly enjoy the same control on a real association study.

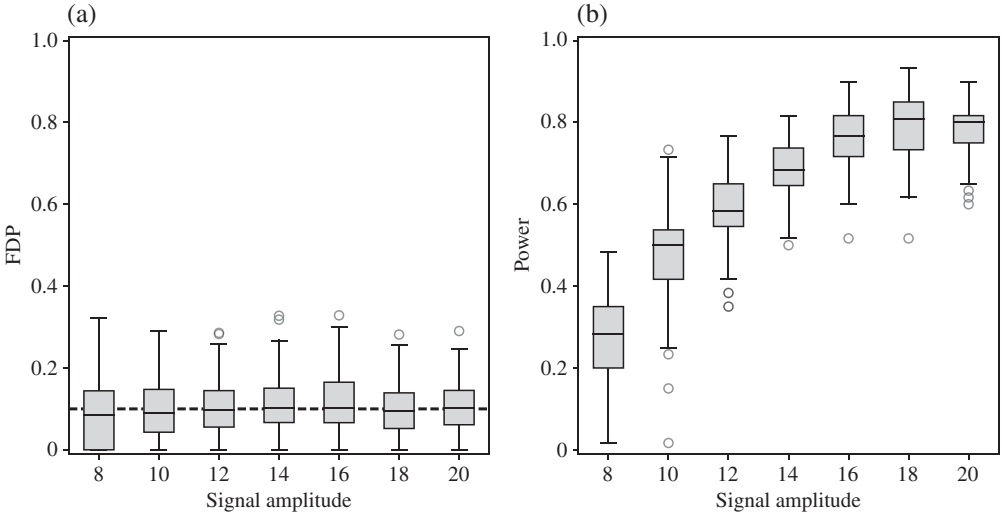

Fig. 3. (a) False discovery proportion, FDP, and (b) power of our procedure with real genetic variables. A box represents 100 experiments. The dashed black line in (a) indicates the target level $\alpha = 0.1$. The offset of the knockoff filter is set equal to 1.

Table 1. *False discovery rate and power, in percentage, for the experiment of Fig. 3 with* 95% *normal confidence intervals, i.e., standard errors multiplied by* 1.96*, with and without offset*

Signal amplitude	FDR (95% c.i.)		Power (95% c.i.)	
	Offset 0	Offset 1	Offset 0	Offset 1
8	9.3 ± 1.7	4.7 ± 1.4	27.9 ± 2.0	17.3 ± 3.0
10	10.3 ± 1.4	7.6 ± 1.3	47.9 ± 2.1	42.2 ± 3.0
12	10.6 ± 1.3	8.2 ± 1.3	59.1 ± 1.7	55.8 ± 2.2
14	11.1 ± 1.3	9.1 ± 1.2	68.4 ± 1.4	66.7 ± 1.6
16	11.8 ± 1.4	9.7 ± 1.4	76.0 ± 1.4	74.3 ± 1.6
18	10.1 ± 1.2	8.0 ± 1.1	79.1 ± 1.5	77.9 ± 1.7
20	10.5 ± 1.2	8.7 ± 1.1	78.3 ± 1.2	77.6 ± 1.3

c.i., confidence interval.

7. Applications to genome-wide association studies

7.1. *Analysis of genome-wide association data*

We apply our procedure to data from two association studies: the Northern Finland 1966 Birth Cohort study of metabolic syndrome (Sabatti et al., 2009), accession number phs000276.v2.p1, and the Wellcome Trust Case Control Consortium study of Crohn's disease (WTCCC, 2007).

The metabolic syndrome study comprises observations on 5402 individuals from northern Finland, including genotypes for approximately 300 000 polymorphisms and nine phenotypes. We focus on measurements of cholesterol, triglyceride levels and height, as there is a rich literature on their genetic bases that we can rely upon for comparison. Since not all outcome measurements are available for every subject, the effective values of n are different for each phenotype and a little lower than 5402. From the Crohn's disease study, we analyse 2996 control and 1917 disease samples typed at $p = 377\,749$ polymorphisms.

We pre-process the data as described in the Supplementary Material and reduce the number of variables by pruning with the same method used in the numerical simulation of § 6.1. Then, we perform variable selection using our knockoff procedure. Before applying Algorithm 2 to construct

the knockoff copies, we estimate the parameters $(\hat{r}, \hat{\alpha}, \hat{\theta})$ of the hidden Markov model from § 5.1 using fastPHASE, separately for each of the first 22 chromosomes. Since the estimation of the covariate distribution does not make use of the response, we compute a single set of estimates for all phenotypes in the metabolic syndrome study, and a separate one for the Crohn's disease study. In both cases, we run fastPHASE with a prespecified number of latent haplotype clusters $K = 12$. With its default settings, the imputation software estimates $\hat{\alpha}$ with the additional constraint that $\alpha_{j,k}$ can only depend on the first index j. For simplicity, we do not modify this setting.

Having sampled the knockoff copies, we assess variable importance as in § 6.1, by performing a lasso regression of Y on the standardized knockoff-augmented matrix of covariates $[X, \tilde{X}] \in \{0, 1, 2\}^{n \times 2p}$, with a regularization parameter λ chosen through ten-fold crossvalidation. For the Crohn's disease study the response is binary and we use logistic regression with an ℓ_1-norm penalty instead of the lasso. Relevant polymorphisms are then selected by applying the knockoff filter with target level $\alpha = 0.1$ and offset equal to 0.

7.2. *Results*

We performed the analysis described above on the four datasets. Since our method is not deterministic, in each case the selections depend on the realization of $\tilde{X}$. Repeating the procedure multiple times and cherry-picking the results would obviously violate the control of the false discovery rate, so we instead display all findings that are selected at least 10 times over 100 independent repeats of the knockoffs procedure. This is only supposed to provide the reader with an impression of the variability of our method, since in principle control of the false discovery rate does not necessarily hold if one aggregates selections obtained with different realizations of $\tilde{X}$. Finding a good way of combining these selections remains an open research problem.

While we do not have sufficient experimental evidence to assess which of our discoveries are true, we can compare our results to those of studies carried out on much larger samples and consider these as the only available approximation of the truth. For lipids we rely on Global Lipids Genetics Consortium (2013), for height on Wood et al. (2014) and Marouli et al. (2017), and for Crohn's disease on Franke et al. (2010). Since different studies include slightly different sets of polymorphisms and our analysis involves a pruning phase, some care has to be taken in deciding when findings match. Each of our clusters spans a genomic locus that can be described by the positions of the first and last polymorphisms. We consider one of our findings to be replicated if the larger study reports as significant a variable whose position is within the region spanned by the cluster we discover.

Our procedure identifies a larger number of potentially significant loci than traditional methods based on marginal testing, except in the case of triglycerides, for which very few findings are obtained with either approach. In Fig. 4(a), the distribution of the number of discoveries over 100 independent realizations of our knockoff variables is compared to the corresponding fixed quantity from the standard genomic analysis on the same dataset, as performed in the earlier works cited above. We can thus verify that, while our procedure is not deterministic, we consistently select more variables. In Fig. 4(b), we show the proportion of our discoveries that is confirmed by the corresponding meta-analyses, separately for each dataset. If we tried to naïvely estimate the false discovery rate from these plots, we would obtain a value much larger than the target level $\alpha = 0.1$, but this would not be very meaningful because none of the meta-analyses is believed to have correctly identified all relevant associations. Instead, some perspective can be gained by comparing our proportion of confirmed discoveries to that obtained with marginal testing on the same data. In the case of one type of cholesterol and triglycerides, our confirmed proportion is appreciably higher, even though one may have intuitively expected a better agreement between studies relying on the same testing framework.

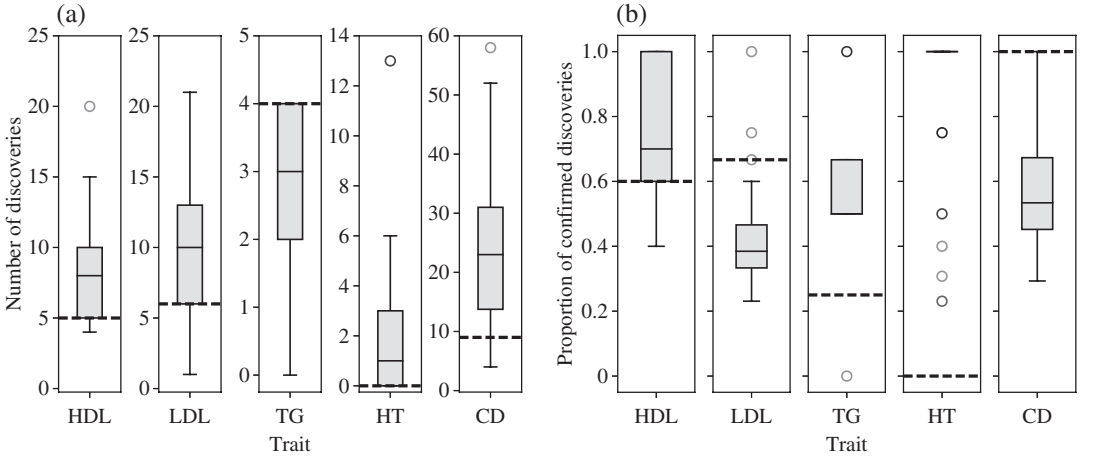

Fig. 4. Discoveries made on different datasets: (a) Total number of discoveries; (b) proportion of the discoveries that are confirmed by the meta-analysis. The boxplots refer to our method, while the dashed black lines represent the standard genomic analysis with the same data. The phenotypes are cholesterol, HDL, LDL; triglycerides, TG; height, HT, and Crohn's disease, CT.

It should not be surprising that our results are at least partially consistent with those of previous studies. In spite of the fact that our method relies on fundamentally different principles, we have selected relevant variables after computing importance measures based on sparse generalized linear regression. The robustness of our Type I error control is completely unaffected by the validity of such a model, but a bias towards the discovery of additive linear effects naturally arises. In future studies, one may discover additional associations by easily deploying our procedure with more complex nonlinear measures of feature importance.

8. Discussion

Conditionally on X and Y, the selections depend on the specific realization of the knockoffs $\tilde{X}$. Different repetitions of our procedure provide reasonably consistent answers on the same data, but at this point it is not clear how to best aggregate the different results.

In our analysis of genetic data, we have pruned the variables during the pre-processing phase and restricted the inference to the representatives for each group. Alternatively, one could try to adapt the idea of group knockoffs in Dai & Barber (2016) to our method.

Different parameterizations of the hidden Markov model have been developed within the genotype imputation community and they can be easily exploited by our procedure. For example, if a collection of known haplotypes is available, it is possible to include them in the description of F_X used to generate the knockoff copies. It would be interesting to investigate from an applied perspective the relative advantages of one choice over another.

Since we have computed variable importance measures based on generalized linear models, even though our false discovery rate control does not rely on any assumptions of linearity, the power may be negatively affected if the true likelihood is far from linear. In order to fully exploit the flexibility and robustness of knockoffs, it would be interesting to explore the use of alternative statistics that can better capture interactions and nonlinearities, e.g., trees.

At this point we know how to perform controlled variable selection with knockoffs in the special cases where the variables can be described by either a hidden Markov model or a multivariate normal distribution. It would be interesting to extend this to other classes of covariates, such as more general graphical models.

ACKNOWLEDGEMENT

Candès was partially supported by the U.S. Office of Naval Research and by a Math+X Award from the Simons Foundation. Sesia was partially supported by the U.S. National Institutes of Health and the Simons Foundation. We thank Lucas Janson for inspiring discussions and for sharing his computer code.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of the theoretical results, a summary of existing knockoff methodology, further methodological details related to the numerical simulation and the data analysis, and a glossary of relevant technical terms from genetics.

REFERENCES

ALEXANDER, D. H. & LANGE, K. (2011). Stability selection for genome-wide association. *Genet. Epidemiol.* **35**, 722–8.
BARBER, R. F. & CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43**, 2055–85.
BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.
BORECZKY, J. S. & WILCOX, L. D. (1998). A hidden Markov model framework for video segmentation using audio and image features. In *Proc. 1998 IEEE Int. Conf. Acoust. Speech Sig. Proces.*, vol. 6. IEEE.
BROWNING, S. & BROWNING, B. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–97.
BROWNING, S. R. & BROWNING, B. L. (2011). Haplotype phasing: Existing methods and new developments. *Nature Rev. Genet.* **12**, 703–14.
BRZYSKI, D., PETERSON, C. B., SOBCZYK, P., CANDÈS, E. J., BOGDAN, M. & SABATTI, C. (2017). Controlling the rate of GWAS false discoveries. *Genetics* **205**, 61–75.
BUREAU, A., DUPUIS, J., FALLS, K., LUNETTA, K. L., HAYWARD, B., KEITH, T. P. & VAN EERDEWEGH, P. (2005). Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.* **28**, 171–82.
CANDÈS, E. J., FAN, Y., JANSON, L. & LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Statist. Soc.* B **80**, 551–77.
CANDÈS, E. J. & PLAN, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* **37**, 2145–77.
CARLBORG, O. & HALEY, C. S. (2004). Epistasis: Too often neglected in complex trait studies? *Nature Rev. Genet.* **5**, 618–25.
DAI, R. & BARBER, R. (2016). The knockoff filter for FDR control in group-sparse and multitask regression. In *Proc. 33rd Int. Conf. Mach. Learn.*, M. F. Balcan & K. Q. Weinberger, eds., vol. 48 of *Proceedings of Machine Learning Research*. New York: Association for Computing Machinery.
ERNST, J. & KELLIS, M. (2012). ChromHMM: Automating chromatin-state discovery and characterization. *Nature Meth.* **9**, 215–6.
FALUSH, D., STEPHENS, M. & PRITCHARD, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**, 1567–87.
FRANKE, A., MCGOVERN, D. P. B., BARRETT, J. C., WANG, K., RADFORD-SMITH, G. L., AHMAD, T., LEES, C. W., BALSCHUN, T., LEE, J., ROBERTS, R. ET AL. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genet.* **42**, 1118–25.
GLOBAL LIPIDS GENETICS CONSORTIUM (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genet.* **45**, 1274–83.
GUAN, Y. & STEPHENS, M. (2008). Practical issues in imputation-based association mapping. *PLOS Genet.* **4**, 1–11.
GUAN, Y. & STEPHENS, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Statist.* **5**, 1780–815.
HOGGART, C. J., WHITTAKER, J. C., DE IORIO, M. & BALDING, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLOS Genet.* **4**, 1–8.
HORMOZDIARI, F., KOSTEM, E., KANG, E. Y., PASANIUC, B. & ESKIN, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508.
HUGHEY, R. & KROGH, A. (1996). Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *Bioinformatics* **12**, 95–107.
JUANG, B. H. & RABINER, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics* **33**, 251–72.

KROGH, A. (1997). Two methods for improving performance of a HMM and their application for gene finding. In *Proc. 5th Int. Conf. on Intelligent Systems for Molecular Biology*, T. Gaasterland, P. Karp, K. Karplus, G. Ouzounis, C. Sander & A. Valencia, eds. Menlo Park, California: AAAI Press, pp. 179–86.

KROGH, A., BROWN, M., MIAN, I., SJÖLANDER, K. & HAUSSLER, D. (1994). Hidden Markov models in computational biology. *J. Molec. Biol.* **235**, 1501–31.

LI, H. & DURBIN, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–6.

LI, J., DAS, K., FU, G., LI, R. & WU, R. (2011). The Bayesian lasso for genome-wide association studies. *Bioinformatics* **27**, 516–23.

LI, N. & STEPHENS, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–33.

LI, Y., WILLER, C. J., DING, J., SCHEET, P. & ABECASIS, G. R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–34.

MAILMAN, M. D., FEOLO, M., JIN, Y., KIMURA, M., TRYKA, K., BAGOUTDINOV, R., HAO, L., KIANG, A., PASCHALL, J., PHAN, L. et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nature Genet.* **39**, 1181–6.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A. et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461**, 747–53.

MARCHINI, J. & HOWIE, B. (2010). Genotype imputation for genome-wide association studies. *Nature Rev. Genet.* **11**, 499–511.

MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. & DONNELLY, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genet.* **39**, 906–13.

MAROULI, E., GRAFF, M., MEDINA-GOMEZ, C., LO, K. S., WOOD, A. R., KJAER, T. R., FINE, R. S., LU, Y., SCHURMANN, C., HIGHLAND, H. M. et al. (2017). Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–90.

PATIL, N., BERNO, A. J., HINDS, D. A., BARRETT, W. A., DOSHI, J. M., HACKER, C. R., KAUTZER, C. R., LEE, D. H., MARJORIBANKS, C., MCDONOUGH, D. P. et al. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–23.

QIN, Z. S., NIU, T. & LIU, J. S. (2002). Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **71**, 1242–7.

SABATTI, C. (2013). Multivariate linear models for GWAS. In *Advances in Statistical Bioinformatics*, K.-A. Do, Z. Qin & M. Vannucci, eds. Cambridge: Cambridge University Press, pp. 188–207.

SABATTI, C., HARTIKAINEN, A.-L., POUTA, A., RIPATTI, S., BRODSKY, J., JONES, C. G., ZAITLEN, N. A., VARILO, T., KAAKINEN, M., SOVIO, U. et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genet.* **41**, 35–46.

SABATTI, C., SERVICE, S. & FREIMER, N. (2003). False discovery rate in linkage and association genome screens for complex disorders. *Genetics* **164**, 829–33.

SCHEET, P. & STEPHENS, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–44.

STEPHENS, M., SMITH, N. J. & DONNELLY, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–89.

STOREY, J. D. & TIBSHIRANI, R. J. (2003). Statistical significance for genomewide studies. *Proc. Nat. Acad. Sci.* **100**, 9440–5.

SUDLOW, C., GALLACHER, J., ALLEN, N., BERAL, V., BURTON, P., DANESH, J., DOWNEY, P., ELLIOTT, P., GREEN, J., LANDRAY, M. et al. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779.

SUN, W. & CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Statist. Soc.* B **71**, 393–424.

TANG, H., CORAM, M., WANG, P., ZHU, X. & RISCH, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.* **79**, 1–12.

VAN DE GEER, S., BUHLMANN, P., RITOV, Y. & DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42**, 1166–202.

WAGER, S. & ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Statist. Assoc.* DOI: 10.1080/01621459.2017.1319839.

WALL, J. D. & PRITCHARD, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Rev. Genet.* **4**, 587–97.

WANG, K., LI, M., HADLEY, D., LIU, R., GLESSNER, J., GRANT, S. F., HAKONARSON, H. & BUCAN, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–74.

WEI, Z., SUN, W., WANG, K. & HAKONARSON, H. (2009). Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics* **25**, 2802–8.

WOOD, A. R., ESKO, T., YANG, J., VEDANTAM, S., PERS, T. H., GUSTAFSSON, S., CHU, A. Y., ESTRADA, K., LUAN, J., KUTALIK, Z. et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genet.* **46**, 1173–86.

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–78.

WU, T. T., CHEN, Y. F., HASTIE, T. J., SOBEL, E. & LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714–21.

ZHANG, K., DENG, M., CHEN, T., WATERMAN, M. S. & SUN, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proc. Nat. Acad. Sci.* **99**, 7335–9.

ZHAO, P. & YU, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–63.

ZUK, O., HECHTER, E., SUNYAEV, S. R. & LANDER, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Nat. Acad. Sci.* **109**, 1193–8.

[*Received on* 11 *July* 2017. *Editorial decision on* 17 *March* 2018]