Biometrika (?), **?**, ?, *pp*. 1–19
Printed in Great Britain Advance Access publication on ? ? ?

Searching for robust associations with a multi-environment knockoff filter

BY S. LI*
Department of Statistics, Stanford University, Stanford, California 94305, USA
lsn@stanford.edu

M. SESIA*
Department of Data Sciences and Operations, University of Southern California, Los Angeles, California 90089, USA
sesia@marshall.usc.edu

Y. ROMANO
Departments of Electrical Engineering and of Computer Science, Technion, Haifa, Israel
yromano@technion.ac.il

E. CANDÈS, C. SABATTI
Department of Statistics, Stanford University, Stanford, California 94305, USA
candes@stanford.edu sabatti@stanford.edu

SUMMARY

This paper develops a method based on model-X knockoffs to find conditional associations that are consistent across environments, controlling the false discovery rate. The motivation for this problem is that large data sets may contain numerous associations that are statistically significant and yet misleading, as they are induced by confounders or sampling imperfections. However, associations replicated under different conditions may be more interesting. In fact, consistency sometimes provably leads to valid causal inferences even if conditional associations do not. While the proposed method is widely applicable, this paper highlights its relevance to genome-wide association studies, in which robustness across populations with diverse ancestries mitigates confounding due to unmeasured variants. The effectiveness of this approach is demonstrated by simulations and applications to the UK Biobank data.

Some key words: Conditional independence; causality; false discovery rate; genome-wide association studies.

1. INTRODUCTION

One goal of statistics is to discover which among many variables are meaningfully associated with an outcome of interest. Associations can have different meanings, ranging from marginal, a tendency of two variables to covary, to causal, a relation ensuring interventions on one variable affect another. For example, a genome-wide association study may reveal some genetic variants are more frequent among diabetic patients. This marginal association may occur simply because

* Equal contribution

the discovered variants are shared, due to common ancestry, by a population following a less healthy diet (Devlin & Roeder, 1999). Of course, it would be more actionable to identify variants involved in biological processes which, if modified by a drug, could influence the disease. Marginal associations are the easiest to recognize but also the least informative, while causal associations are more elusive, although they better lend themselves to scientific interpretations.

Between marginal and causal associations one finds conditional association: the tendency of two variables to covary as others are fixed. Conditional testing often relies on parametric models, but these require assumptions that are not always justified. The alternative "model-X" approach of Candès et al. (2018) requires no model for the conditional distribution of the outcome, assuming instead that the joint distribution of the predictors is known, at least within a reasonable approximation in practice. This framework is widely relevant, and especially so in genetic studies because reliable knowledge is available about the distribution of the genotypes (Sesia et al., 2018). Two types of model-X methods have been developed: a version of knockoffs (Barber & Candès, 2015), and the conditional randomization test, both of which can harness the power of any machine learning algorithm while controlling type-I errors in finite-samples.

Despite the advantages, conditional testing is not fully satisfactory. First, it neglects confounders: unobserved variables that would explain the association (Pearl, 2009). For instance, a genotyped variant may be associated with a disease only because physically close to an unseen causal one (Pritchard & Przeworski, 2001). Second, data may not be collected exactly from the target population, due to either sampling bias (Heckman, 1979) or convenience (Harford, 2014). Third, unknown sample dependencies may lead to spurious associations (Lee & Ogburn, 2020).

This paper mitigates the above limitations by analyzing data from many environments, corresponding to different populations, experimental settings, or sampling strategies. The motivating conjecture is that the most informative associations are those consistently reproducible, as these enable robust predictions and may even reflect causal relations. This old idea (Hume, 1739) is translated into a method for testing hypotheses of consistent conditional association. Our solution utilizes knockoffs because they scale well to high dimensions and control the false discovery rate (Benjamini & Hochberg, 1995). However, an alternative approach based on the conditional randomization test is possible; see Section S1 and Figure S1 in the Supplementary Material.

Related work

We take inspiration from Peters et al. (2016) and Heinze-Deml et al. (2018), which advanced invariance across environments as a framework for causal inference. Departing from their work, we do not assume homogeneous effects; indeed, we can obtain meaningful inferences without any reference to a causal model. Further, we seek different guarantees, controlling the false discovery rate, and we can handle hundreds of thousands of variables. The invariance of Peters et al. (2016) has been utilized to make predictions robust to covariate shift (Rojas-Carulla et al., 2018; Rothenhäusler et al., 2021)—changes in the distribution of explanatory variables. Although prediction is a related problem, our paper concentrates on testing. This work is related to causal discovery (Pearl, 2009; Mooij et al., 2020), which aims to learn the full graph describing the relations between many variables, without specifying one outcome. Causal discovery can uncover causal directions, unlike our method, but it requires parametric assumptions and asymptotic rather than finite-sample guarantees. Invariance also appears in the feature selection literature (Yu et al., 2020), though typically without finite-sample inferences. Compared to earlier works on knockoffs (Barber & Candès, 2015; Candès et al., 2018), this paper differs because those mostly focused on the construction of the knockoffs (Gimenez et al., 2019; Romano et al., 2019; Bates et al., 2020a), robustness to model misspecifications (Barber et al., 2020), and power (Katsevich & Ramdas, 2020; Wang & Janson, 2020). Our applications are in genetics, exploit-

ing knockoff constructions that account for dependencies among genotypes (Sesia et al., 2018, 2020), population structure (Sesia et al., 2021), and other confounders (Bates et al., 2020b), but did not deal with missing variants. Knockoffs are however applicable to many other fields (Shen et al., 2019; Chia et al., 2021; Fan et al., 2020) for which our methods may also be helpful.

2. Conditional associations that hold across environments

Consider observations (X, Y) consisting of p variables, $X \in \mathcal{X}^p$, and an outcome, $Y \in \mathcal{Y}$, sampled from E environments. Here, $\mathcal{X}$ and $\mathcal{Y}$ may be discrete or continuous. Assume the joint distribution of X within any environment $e \in [E] = \{1, \ldots, E\}$, P_X^e, is known. For simplicity, imagine different samples as mutually independent, although known dependencies can be accommodated (Sesia et al., 2021). The model-X framework (Candès et al., 2018) provides methods to test the hypothesis that Y is independent of the j-th variable conditional on the other ones,

$$\mathcal{H}_j^{\mathrm{ci},e} : Y^e \perp\!\!\!\perp X_j^e \mid X_{-j}^e, \tag{1}$$

for all $j \in [p] = \{1, \ldots, p\}$. Here, X_{-j} denotes all explanatory variables except X_j, and the superscript e clarifies we are focusing on the distribution of the data in the e-th environment. Our goal is to powerfully test the following consistent conditional independence hypothesis,

$$\mathcal{H}_j^{\mathrm{cst}} : \text{there exists } e \in \{1, \ldots, E\} \text{ such that the null } \mathcal{H}_j^{\mathrm{ci},e} \text{ in (1) is true,} \tag{2}$$

controlling the false discovery rate. Intuitively, we would interpret any findings by noting that, if $\mathcal{H}_j^{\mathrm{cst}}$ (2) is false, the conditional association of X_j with Y must hold across all environments.

Note that rejecting $\mathcal{H}_j^{\mathrm{cst}}$ does not necessarily suggest that X_j has a constant effect on Y. For example, X_j may be associated to Y through interactions with other environment-specific variables and, in a parametric analysis, this would require environment-specific models. However, we simply seek variables for which there is evidence of some conditional association with Y across environments, leaving the modeling task to other statistical methods or follow-up studies.

At first sight, it may be tempting to test $\mathcal{H}_j^{\mathrm{ci},e}$ (1) environment by environment and report the common discoveries. Unfortunately, this intersection heuristic does not control the false discovery rate for $\mathcal{H}_j^{\mathrm{cst}}$ (2) even if the tests of $\mathcal{H}_j^{\mathrm{ci},e}$ (1) control it for all e (Katsevich et al., 2021). Alternatively, one may analyze the pooled data from all environments, controlling the false discovery rate for $\mathcal{H}_j^{\mathrm{ci}}$ (1) in the broader population defined by the union of all environments, but not testing consistency. Indeed, the problem is non-trivial, as illustrated by the simulations in Figure 1, which will be explained in Section 6.2. This gives a preview of our method, which provably controls the false discovery rate for $\mathcal{H}_j^{\mathrm{cst}}$ (2) and achieves good power.

A limitation of $\mathcal{H}_j^{\mathrm{cst}}$ (2) is that it becomes harder to reject if E grows but the total number of samples remains constant, because every environment must provide sufficient evidence. However, consistency across most environments might be satisfactory, especially if some have smaller sample sizes. This motivates partial consistency testing, or partial conjunction (Benjamini & Heller, 2008). For $j \in [p]$ and $r \in [E]$, define the null hypothesis

$$\mathcal{H}_j^{\mathrm{pcst},r} : \left| \left\{ e \in \{1, \ldots, E\} : \mathcal{H}_j^{\mathrm{ci},e} \text{ is true} \right\} \right| > E - r. \tag{3}$$

A rejection of $\mathcal{H}_j^{\mathrm{pcst},r}$ suggests X_j is associated with Y in at least r environments. This generalizes $\mathcal{H}_j^{\mathrm{cst}}$ (2), which is recovered if $r = E$. Note that, unlike Benjamini & Heller (2008), we will not account for multiplicity over different r, which we take instead as fixed.

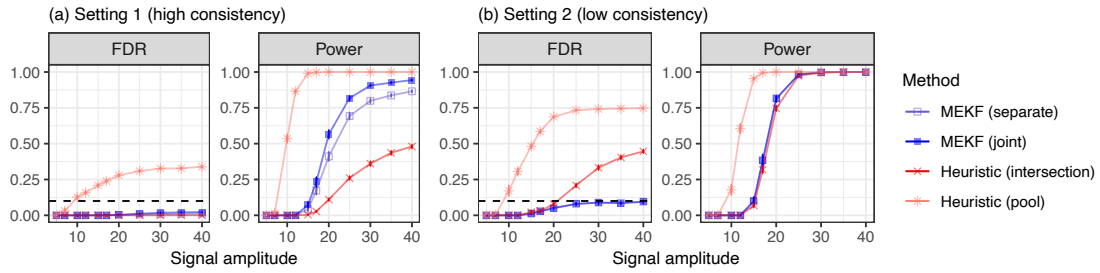

Fig. 1: Performance of the multi-environment knockoff filter (MEKF), implemented with two alternative statistics, and two heuristics for consistent conditional testing on simulated data from many environments. The nominal false discovery rate is 0.1. (a) Data in which most conditional associations are consistent. (b) Data in which most conditional associations are not consistent.

The next section establishes a link between $\mathcal{H}_j^{\text{cst}}$ (2) and causal inference under specific assumptions, but consistent associations can be informative even beyond any causal framework, including in situations where the explanatory variables do not predate the outcome (Castro et al., 2020) or there is no clear notion of interventions (Hernán & Taubman, 2008). In particular, conditional associations are useful to make reliable predictions, and predictive robustness across environments is a well-known challenge in many fields (Heckman, 1979; Duncan et al., 2019). Imagine that Y measures competences acquired by students, and X collects some explanatory variables, e.g., school attendance, family size. Even without hoping to estimate constant causal effects, as variables may interact differently across environments (e.g., low vs. high income neighborhoods), consistent associations may help predict academic difficulties or design useful interventions. While "one size fits all" is certainly wishful thinking, considerations of practicality, transparency, and fairness make it preferable to focus on policies that have good chances of being widely effective. Moreover, consistent associations are more likely to be robust to changes in environments or covariate shifts, such as those that may be expected with the passing of time.

Searching for patterns among a large number of variables can lead to fitted models relating X to Y with a precision that cannot be replicated in other data sets. We typically rely on "cross-validation" to mitigate this problem, but that only guarantees internal consistency. For example, liking curly fries may predict IQ well among some Facebook users (Kosinski et al., 2013) but this association is probably not universal. Testing for $\mathcal{H}_j^{\text{cst}}$ (2) can be seen as "external cross-validation" (Waldron et al., 2014), thereby identifying less ephemeral relations (Efron, 2020).

3. FROM CONSISTENT ASSOCIATIONS TO CAUSAL INFERENCES

3.1. *A constant causal model*

This section assumes a constant causal model across all environments, which is unnecessary for the interest or validity of the proposed tests, but helps illustrate how consistent associations may facilitate the discovery of variables with causal effects in some settings.

Assume a structural equation model (Boolen, 1989) to link p_{z} explanatory variables (Z), of which p are observed (X) and p_{c} are unobserved (C), to the outcome (Y). Note that we will write $Z = (X, C) \in \mathcal{X}^{p_{\text{z}}}$, with $p_{\text{z}} = p + p_{\text{c}}$. In this constant model, the i-th individual outcome, $Y^{(i)}$, is given by $Y^{(i)} = \bar{f}\left(Z^{(i)}, V^{(i)}\right)$, where $\bar{f}$ is unknown and V is exogenous noise from a standard uniform distribution, for example. Assume the causal direction is known: Y is caused by some combination of Z and V. This simplification is appropriate in genetic studies, for exam-

ple, because the genotypes predate the phenotype. Consider then the goal of discovering which variables have a direct effect on Y and do not satisfy the following sharp causal null

$$\mathcal{H}_j^{\mathrm{cs}} : \bar{f}\left(z_1, \ldots, z_{j-1}, z_j, z_{j+1}, \ldots, v\right) = \bar{f}\left(z_1, \ldots, z_{j-1}, z'_j, z_{j+1}, \ldots, v\right), \ \forall z, z'_j, v. \quad (4)$$

Intuitively, this says that intervening on Z_j holding all other variables fixed would have no effect at all on Y. Alternatively, one may think of $\mathcal{H}_j^{\mathrm{cs}}$ as saying the potential outcomes are identical under all Z_j (Rubin, 2005). Sharp hypotheses do not allow the estimation of heterogeneous treatment effects (Neyman & Iwaszkiewicz, 1935), but they are helpful to discover which variables are more likely to be causal, especially for a preliminary exploratory analysis (Imbens & Rubin, 2015). Further, our non-parametric causal model is very flexible and does not exclude that a variable may appear to have different linear effects on the outcome across environments with covariate shifts. Indeed, the sole purpose of this model is to concretely define causality.

Suppose we have data from E environments, each corresponding to a distribution P_Z^e of explanatory variables, for $e \in [E]$. Conditional on Z, the outcome Y is generated by the above model. For simplicity we rewrite this model as a function only of the causal variables, listed as $\mathrm{Pa}(Y) = \{j \in [p_{\mathrm{z}}] : \mathcal{H}_j^{\mathrm{cs}} \text{ in (4) is false}\}$, where $\mathrm{Pa}(Y)$ stands for parents of Y. In particular,

$$Y^{(i)} = f\left(Z^{(i)}_{\mathrm{Pa}(Y)}, V^{(i)}\right), \quad (5)$$

where f is the restriction of $\bar{f}$ on $\mathrm{Pa}(Y)$. For convenience, we split $\mathrm{Pa}(Y)$ into observed $\mathrm{Pa}_{\mathrm{x}}(Y) = \{j \in [p] : j \in \mathrm{Pa}(Y)\}$ and unobserved $\mathrm{Pa}_{\mathrm{c}}(Y) = \mathrm{Pa}(Y) \setminus \mathrm{Pa}_{\mathrm{x}}(Y)$ causal variables. Of course, we only inquire about $\mathrm{Pa}_{\mathrm{x}}(Y)$ as we have no data about $\mathrm{Pa}_{\mathrm{c}}(Y)$.

3.2. The gap between conditional testing and causal inference

Rejecting $\mathcal{H}_j^{\mathrm{ci},e}$ (1) yields a causal inference if there is no confounding, i.e., if the unobserved causal variables are non-existent or conditionally independent of the observed ones.

PROPOSITION 1. *Fix e and j. Under the causal model and the data sampling scheme in Section 3.1, assuming (i) $\mathrm{Pa}_{\mathrm{c}}(Y) = \emptyset$ or (ii) $X_j^e \perp\!\!\!\perp C^e_{\mathrm{Pa}_{\mathrm{c}}(Y)} \mid X^e_{-j}$, then $\mathcal{H}_j^{\mathrm{cs}}$ (4) implies $\mathcal{H}_j^{\mathrm{ci},e}$ (1).*

The first assumption above would require measuring every relevant variable, which seems unrealistic. The second holds in randomized experiments and in certain observational studies such as those involving genetic parents-child trio data (Bates et al., 2020b). However, it is unclear why the missing variables should generally be conditionally independent of the observed ones. Therefore, causal inference remains challenging even if the model in Section 3.1 is acceptable. Note that more formal statements of our theoretical results, along with all mathematical proofs, can be found in Section S2, Supplementary Material.

3.3. Consistency improves robustness to missing variables

The assumptions necessary for causal inferences can be relaxed if the associations are consistent, because shifts in P_Z^e may induce different variables to pick up spurious associations in different environments while causal associations tend to remain unchanged. Figure 2 visualizes this idea with a toy example involving two causal variables, one of which is unmeasured. Here the environments differ in P_Z^e to a sufficient extent that their spurious associations have no overlap. Section 5 will explain the relevance of this idea to genome-wide association studies.

PROPOSITION 2. *Fix any j and consider E environments. In the setting of Proposition 1, if (i) $\mathrm{Pa}_{\mathrm{c}}(Y) = \emptyset$ or (ii) there exists $e \in [E] : X_j^e \perp\!\!\!\perp C^e_{\mathrm{Pa}_{\mathrm{c}}(Y)} \mid X^e_{-j}$, the $\mathcal{H}_j^{\mathrm{cs}}$ (4) implies $\mathcal{H}_j^{\mathrm{cst}}$ (2).*

3.4. Consistency improves robustness to sampling imperfections

Perfect random samples from the target population are not always available, as real data sampling may involve unknown biases (Heckman, 1979; Hargittai, 2015) or network effects (Shalizi

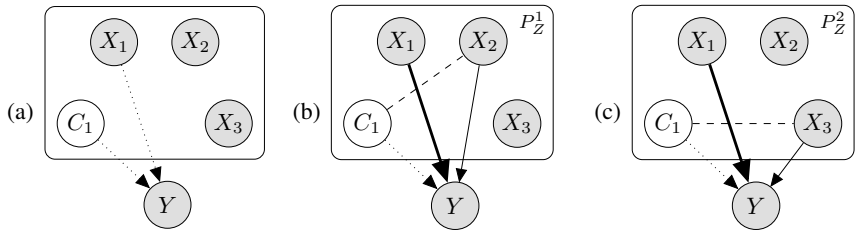

Fig. 2: Consistency improves robustness to missing variables. (a) Causal model for $Y \mid X, C$. (b,c) Conditional associations in two environments. Shaded nodes: outcome or observed variables. White nodes: unobserved variables. Dotted arrows: causal links. The dashed segments represent P_Z^e, which differs across environments: C_1 is associated with X_2 in the first one, and with X_3 in the second one, while X_1 and X_2 are conditionally independent of C_1, X_3 and of each other. Solid arrows: observable conditional associations with Y (thick if causal, thin if spurious).

& Thomas, 2011; Lee & Ogburn, 2020). Consistency will not fully resolve these difficulties, but it can mitigate them, as explained in Section S3, Supplementary Material. These results make our method relevant in many fields, including the social sciences, where collecting random samples is difficult, and where it is increasingly common to find large data sets with many associations, for which controlling the false discovery rate is desirable.

4. METHODS

4.1. Review: the methodology of model-X knockoffs

Knockoffs enable testing $\mathcal{H}_j^{\mathrm{ci},e}$ (1), for all $j \in [p]$ and any $e \in [E]$. The idea is to augment the data for each of the n individuals with p knockoffs, based on the joint distribution of X, which is assumed to be known (Candès et al., 2018). The knockoffs $\tilde{X}$ are created without looking at Y, so $Y \perp\!\!\!\perp \tilde{X} \mid X$, and they are pairwise exchangeable with the original variables. If $[X^e, \tilde{X}^e] \in \mathbb{R}^{n \times 2p}$ is obtained by concatenating $X^e \in \mathbb{R}^{n \times p}$ with the corresponding $\tilde{X}^e \in \mathbb{R}^{n \times p}$, this has the same distribution as $[X^e, \tilde{X}^e]_{\mathrm{swap}(j)}$, the matrix obtained from the latter by swapping the j-th column of X^e with the j-th column of $\tilde{X}^e$, for all $j \in [p]$. Therefore, swaps of a variable with its knockoff cannot be detected without looking at Y: the only significant difference between X_j and $\tilde{X}_j$ may be the lack of conditional association of $\tilde{X}_j$ with Y. As our contribution does not concern this aspect of the analysis, we assume valid knockoffs are available based on the known P_X^e, referring to the prior works mentioned in Section 1 for specific algorithms to construct them.

The second step is to fit a model predicting Y from $X, \tilde{X}$, computing importance measures T_j^e and $\tilde{T}_j^e$ for each X_j and $\tilde{X}_j$, respectively. Any model can be employed, as long as swapping X_j with $\tilde{X}_j$ results in T_j^e being swapped with $\tilde{T}_j^e$. A typical choice is to fit a sparse generalized linear model, e.g., the lasso (Tibshirani, 1996), tuning its regularization via cross-validation; then, the absolute values of the (scaled) regression coefficients are powerful importance measures. For any true $\mathcal{H}_j^{\mathrm{ci},e}$ (1), the sign of $W_j^e = T_j^e - \tilde{T}_j^e$ is a coin flip, while a large positive value is evidence against the null. Further, if a random $\epsilon^e \in \{-1, +1\}^p$ is such that $\epsilon_j^e = +1$ if $\mathcal{H}_j^{\mathrm{ci},e}$ is false and otherwise $\mathrm{pr}(\epsilon_j^e = +1) = 1/2$ independently of everything else, then W^e has the same distribution as $W^e \odot \epsilon^e$, where $\odot$ indicates element-wise multiplication. Thus, the signs of W^e are independent one-bit conservative p-values for $\mathcal{H}_j^{\mathrm{ci},e}$ (1), if transformed as $p_j^e = 1/2$ if $W_j^e > 0$ and $p_j^e = 1$ otherwise. The ordering provided by the absolute values of W^e allows

one to powerfully test the above hypotheses sequentially. Concretely, the knockoff filter (Barber & Candès, 2015) computes a significance threshold such that rejecting $\mathcal{H}_j^{\mathrm{ci},e}$ (1) for all j with larger W_j^e controls the false discovery rate below a desired level α. Equivalently, this can be seen as applying the selective SeqStep+ test (Barber & Candès, 2015) to the above ordered one-bit p-values. We will extend this method to analyze data from many environments, testing $\mathcal{H}_j^{\mathrm{cst}}$ (2).

4.2. Definition of multi-environment knockoff statistics

Consider E environments, each corresponding to data $Y^e \in \mathbb{R}^n, X^e \in \mathbb{R}^{n\times p}$, and knockoffs $\tilde{X}^e \in \mathbb{R}^{n\times p}$; all environments have n samples here, although this is unnecessary. The first ingredients for testing $\mathcal{H}_j^{\mathrm{cst}}$ are the following statistics, generalizing those of Candès et al. (2018).

DEFINITION 1. *$W \in \mathbb{R}^{E\times p}$ are multi-environment knockoff statistics if W has the same distribution as $W \odot \epsilon$, where $\epsilon \in \{\pm 1\}^{E\times p}$ has independent entries and rows ϵ^e such that $\epsilon_j^e = \pm 1$ with probability $1/2$ if $\mathcal{H}_j^{\mathrm{ci},e}$ (1) is true and $\epsilon_j^e = +1$ otherwise, for $j \in [p]$ and $e \in [E]$.*

One way to compute the matrix W is to *separately* analyze the data from each environment and stack the output W^e row by row. This approach is fast but not necessarily very data efficient. For example, if all environments are identical, statistics obtained by separate analyses can only utilize a fraction $1/E$ of the samples compared to a pooled analysis, even though pooling already tests the correct target hypotheses $\mathcal{H}_j^{\mathrm{cst}}$ (2) in this special case. Therefore, we also consider a more general and potentially more powerful *joint* analysis of all environments, as explained next.

Let $Y \in \mathbb{R}^{En}, X \in \mathbb{R}^{En\times p}, \tilde{X} \in \mathbb{R}^{En\times p}$ be the matrices obtained by stacking all observations or knockoffs. We define $[T, \tilde{T}] \in \mathbb{R}^{E\times 2p}$ as a matrix of multi-environment importance measures computed by a randomized function τ: $[T, \tilde{T}] = \tau(Y, [X, \tilde{X}])$. The function τ may involve any machine learning algorithms (examples are in the next section), as long as swapping variables and knockoffs in one environment has the effect of swapping the corresponding importance measures in that environment, leaving all other elements of τ unchanged. This only needs to hold in distribution conditional on $X, Y, \tilde{X}$, as τ may be randomized. Formally, conditional on $Y, [X, \tilde{X}]$,

$$\tau\left(Y, [X, \tilde{X}]_{\mathrm{swap}(\mathcal{S})}\right) \overset{d}{=} \left[\tau\left(Y, [X, \tilde{X}]\right)\right]_{\mathrm{swap}(\mathcal{S})}, \tag{6}$$

for any $\mathcal{S} \subseteq [E] \times [p]$, where $\overset{d}{=}$ indicates equality in distribution and $[X, \tilde{X}]_{\mathrm{swap}(\mathcal{S})}$ is obtained from $[X, \tilde{X}]$ by swapping the column X_j^e for environment e with the corresponding $\tilde{X}_j^e$, for all $(e, j) \in \mathcal{S}$. The definition of $[\tau(Y, [X, \tilde{X}])]_{\mathrm{swap}(\mathcal{S})}$ is analogous. Note the slight change of notation compared to Section 4.1: there, swapping was defined only for one environment.

The next section discusses how to compute importance measures satisfying (6) through a joint analysis of the data from all environments, with relatively lower power loss compared to pooling. Here, we note that the property in (6) is sufficient to obtain valid multi-environment statistics.

PROPOSITION 3. *Let $[T, \tilde{T}] \in \mathbb{R}^{E\times 2p}$ be importance measures satisfying* (6)*, and define $W \in \mathbb{R}^{E\times p}$ through $W_j^e = T_j^e - \tilde{T}_j^e$ for all $e \in [E]$ and $j \in [p]$. Then, W satisfies Definition 1.*

4.3. Computation of multi-environment knockoff statistics

Some care must be exercised to ensure statistics computed by a joint multi-environment analysis satisfy (6), as standard techniques (Candès et al., 2018) naively applied to the pooled data would not be valid; see Figure S3, Section S4. A solution that works well is to allow each row of W^e to leverage the data from other environments after perturbing the latter through random column swaps. This perturbation acts on a copy of $[X, \tilde{X}]$, so the order in which the environments are processed is irrelevant, and consists of swapping each observation of one variable with its knockoff based on independent coin flips. Then, we estimate importance weights that are sym-

metric with respect to variables and knockoffs; these will serve as prior information for the next step, capturing the importance of a variable or its knockoff based on the information from all other environments. The influence of this prior information on the final statistics is controlled by a scalar parameter $\gamma^e \in [0, 1]$ designed to summarize the overall relevance of the data from other environments to the prediction task in that of interest. The value of γ^e is tuned by cross-validation within the e-th environment, following the intuition that the most important features are those enabling the most accurate predictions (Candès et al., 2018). Figure 3 provides a schematic of this procedure, while the details are below. As it will become clear, if the data distributions in all environments are identical, this joint analysis is comparable to pooling with 1/2 of all samples, and this is more efficient than the separate analysis described in the previous section if E is large.

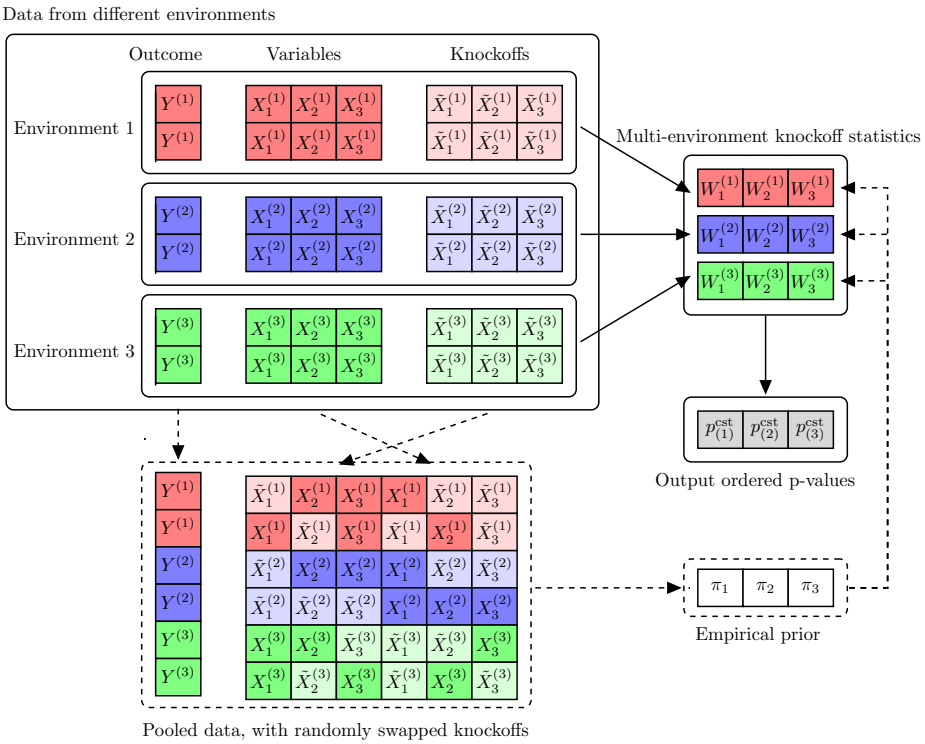

Fig. 3: Schematics for a multi-environment knockoff analysis. In this example, there are 3 variables, 3 environments, and 2 observations per environment. The solid arrows represent separate environment-by-environment analyses before combining the resulting knockoff statistics. The dashed arrows represent the additional steps corresponding to a joint analysis based on empirical cross-prior statistics. The darker blocks indicate data, while the lighter ones indicate knockoffs.

Let $V \in \{0, 1\}^{En \times p}$ be independent coin flips, and $[X, \tilde{X}]_{\text{swap}(V)}$ be the matrix obtained by swapping the i-th observation of X_j with the corresponding $\tilde{X}_j$ if and only if $V_{ij} = 1$. Prior importance measures T^{prior} (resp. $\tilde{T}^{\text{prior}}$) are computed for all variables (resp. knockoffs) as the absolute values of the regression coefficients estimated by fitting a sparse generalized linear model to predict Y given $[X, \tilde{X}]_{\text{swap}(V)}$, using the data from all environments and tuning the regularization parameter by cross-validation. The results are combined symmetrically into an *empirical* prior weight π_j for each j, e.g., $\pi_j = \zeta(T_j^{\text{prior}} + \tilde{T}_j^{\text{prior}})$ for some positive and decreasing function ζ, such as $\zeta(t) = 1/(0.05 + t)$. Finally, we compute the importance measures T^e and $\tilde{T}^e$

based on the unperturbed data in the e-th environment with an approach similar to the separate analysis of Section 4.2. The difference is that now the regularization is feature-specific and depends on two scalar hyper-parameters, $\lambda^e > 0$ and $\gamma^e \in [0, 1]$, both tuned by cross-validation, as well as on the prior weights, which are fixed. In particular, the regularization parameter for $X_j, \tilde{X}_j$ is $\lambda_j^e = \lambda^e(1 - \gamma^e) + \gamma^e \pi_j$. This recovers the separate analysis statistics if $\gamma^e = 0$, but a larger γ^e may improve power by making it less likely to select spurious variables. If $\gamma^e = 1$, the data from the target environment are ignored when deciding which variables to include in the sparse model; this may be reasonable if the data from other environments are overwhelmingly more informative, perhaps due to larger sample sizes. As it is unclear how informative the prior may be in general, the value of γ^e is determined adaptively via cross-validation. To avoid a two-dimensional search, in practice we first tune λ^e and then γ^e.

PROPOSITION 4. *The statistics W obtained by applying $W_j^e = T_j^e - \tilde{T}_j^e$ to the empirical cross-prior importance measures described above satisfy Definition 1.*

The random data perturbation may dilute some of the potentially useful information from other environments, decreasing power compared to pooling, but it is necessary to guarantee false discovery rate control; see the proof of Proposition 4 and the example of Figure S3. Fortunately, random swapping does not prevent the empirical prior from learning which pairs $(X_j, \tilde{X}_j)$ may be important, although it makes approximately half of the samples uninformative. In any case, the role of the prior is modulated by γ^e, which is adaptively tuned via cross-validation. Thus, we expect at worst to select $\gamma^e \approx 0$ and thus approximately recover the separate analysis solution from Section 4.2, meaning that this joint analysis typically is at least as powerful as the latter.

4.4. The multi-environment knockoff filter

The rows of a matrix of multi-environment knockoff statistics W can be combined to obtain one-bit conservative p-values for testing $\mathcal{H}_j^{\mathrm{cst}}$ in (2). Precisely, for each $j \in [p]$, we compute

$$p_j^{\mathrm{cst}} = \begin{cases} 1/2, & \text{if } \min\{\mathrm{sign}(W_j^e)\}_{e=1}^E = +1, \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

The order in which these hypotheses will be tested depends on the absolute values of W, which we combine column-wise with some symmetric function $\bar{w}$ to obtain invariant statistics $|W_j^{\mathrm{cst}}| = \bar{w}(|W_j^1|, \ldots, |W_j^E|)$. Concretely, we will adopt $\bar{w}(|W_j^1|, \ldots, |W_j^E|) = \prod_{e=1}^E |W_j^e|$. In general, the one-bit p-values in (7) are valid for $\mathcal{H}_j^{\mathrm{cst}}$ (2) even conditional on $|W_j^{\mathrm{cst}}|$.

PROPOSITION 5. *If W satisfies Definition 1, the p-values p_j^{cst} in* (7) *for true $\mathcal{H}_j^{\mathrm{cst}}$* (2) *are stochastically larger than uniform conditional on $|W^{\mathrm{cst}}| = (|W_1^{\mathrm{cst}}|, \ldots, |W_p^{\mathrm{cst}}|)$, and "almost independent"; that is, $pr(p_j^{\mathrm{cst}} \leq \alpha \mid |W^{\mathrm{cst}}|, p_{-j}^{\mathrm{cst}}) \leq \alpha$, for all $\alpha \in [0, 1]$ if $\mathcal{H}_j^{\mathrm{cst}}$ is true.*

Proposition 5 suggests sequential testing, e.g., selective SeqStep+. However, this is not obviously valid because the null p_j^{cst} are not independent as required by Barber & Candès (2015). In fact, each of them is also affected by the signs of W corresponding to non-null environments; $\mathcal{H}_j^{\mathrm{cst}}$ only says there is a null in the j-th column, but p_j^{cst} also counts the other signs. Fortunately, our p-values are at worst conservative, provably retaining false discovery rate control.

THEOREM 1. *Selective SeqStep+ applied to p_j^{cst}* (7) *ordered by $|W_j^{\mathrm{cst}}|$ controls the false discovery rate if $pr(p_j^{\mathrm{cst}} \leq \alpha \mid |W^{\mathrm{cst}}|, p_{-j}^{\mathrm{cst}}) \leq \alpha$ for any $\alpha \in [0, 1]$.*

In addition to being useful here to test consistency, the above results indicates broader applicability of selective SeqStep+ than previously known, and suggests possible interesting connections to Fithian & Lei (2020), where false discovery rate control under known dependency is studied for the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995).

4.5. *Testing partially consistent conditional associations*

Fix any $r \in [E]$. For each $j \in [p]$, let n_j^- count the negative signs in the j-th column of the multi-environment statistics W, and let n_j^0 be the number of zeros. Then, compute

$$p_j^{\mathrm{pcst},r} = \Psi(n_j^- - 1, E - r + 1 - n_j^0, 1/2) + U_j \cdot \psi(n_j^-, E - r + 1 - n_j^0, 1/2), \qquad (8)$$

where $\Psi(x, m, \pi)$ is the binomial cumulative distribution function at x and $\psi(x, m, \pi)$ is the corresponding probability mass (m is truncated at 0 if negative), while U_j is uniform on [0,1], independently of all else. Intuitively, $p_j^{\mathrm{pcst},r}$ is a randomized binomial p-value based on the observed number of negative signs among the non-zero statistics, which we can compute because we know there must be at least $E - r + 1$ null environments under $\mathcal{H}_j^{\mathrm{pcst},r}$ (3). As in Section 4.4, we will filter these p-values in decreasing order of $|W_j^{\mathrm{pcst},r}| = \bar{w}(|W_j^1|, \ldots, |W_j^E|)$, for a symmetric function $\bar{w}$. For example, $\bar{w}(|W_j^1|, \ldots, |W_j^E|) = \prod_{e=1}^{r} |W_j|^{(E-e+1)}$, where $|W_j|^{(e)}$ are the order statistics for the absolute values in the j-th column of W. The next result states these ordered p-values are conservative for $\mathcal{H}_j^{\mathrm{pcst},r}$ (3), and "almost independent" as in Proposition 5. Combined with Theorem 1, this guarantees false discovery rate control with selective SeqStep+.

PROPOSITION 6. *If W satisfies Definition 1, for any fixed $r \in [E]$, the p_j^{cst} (8) corresponding to true $\mathcal{H}_j^{\mathrm{pcst},r}$ (3) satisfy $pr(p_j^{\mathrm{pcst},r} \le \alpha \mid |W^{\mathrm{pcst},r}|, p_{-j}^{\mathrm{pcst},r}) \le \alpha$, for all $\alpha \in [0, 1]$.*

Note that, if $r = E$, the $p_j^{\mathrm{pcst},r}$ (8) is not identical to p_j^{cst} (7) since the latter is always 1 when $n_j^0 > 0$. This discrepancy is practically irrelevant because it would not make sense to reject $\mathcal{H}_j^{\mathrm{cst}}$ (2) if $n_j^0 > 0$. While randomization here allows us to deal powerfully with the case in which $n_j^0 > 0$, it would have not helped earlier as the p_j^{cst} in (7) have one bit of information, and selective SeqStep+ only looks at whether they are above $1/2$. Note also that a pooled analysis generally tests $\mathcal{H}_j^{\mathrm{pcst},1}$ (3), possibly more powerfully compared to our method because it allows additional flexibility in the statistics. Thus, our method should only be applied with $r > 1$.

A limitation of selective SeqStep+ applied to the p-values $p_j^{\mathrm{pcst},r}$ (8), which have more than one bit of information, is that it involves a parameter c, the baseline rejection threshold (Barber & Candès, 2015), whose choice can affect power. This problem did not arise in Section 4.4 because $c = 1/2$ is the only option for one-bit p-values such as p_j^{cst} (7). To avoid having to guess a good value of c, we also consider filtering $p_j^{\mathrm{pcst},r}$ (8) with the accumulation test of Li & Barber (2017). If the p-values are independent, this test controls a modified false discovery rate, $\mathrm{mFDR}_q = E[|\{j : \mathcal{H}_j^{\mathrm{ci},e} \text{ is rejected}\} \cap \{j : \mathcal{H}_j^{\mathrm{ci},e} \text{ is true}\}|/(|\{j : \mathcal{H}_j^{\mathrm{ci},e} \text{ is rejected}\}| + q)]$, for some constant q specified below. If q is small and the discoveries are numerous, the above is close to the false discovery rate. Although the $p_j^{\mathrm{pcst},r}$ (8) are dependent, this method remains valid in our context if we randomize the p-values slightly, as explained next.

Starting from multi-environment statistics W, randomly assign imaginary positive or negative signs to any zero entry by flipping independent coins. Then, define n_j^- as the number of negative entries in the j-th column of the resulting tie-breaking W, and compute, with the notation of (8),

$$p_j^{\mathrm{pcst},r} = \Psi(n_j^- - 1, E - r + 1, 1/2) + U_j \cdot \psi(n_j^-, E - r + 1, 1/2). \qquad (9)$$

THEOREM 2. *The accumulation test with increasing accumulation function (e.g., HingeExp with parameter $C = 2$) applied to the p-values in* (9) *controls the mFDR_q defined above (e.g., with $q = C/\alpha$). That is, Theorem 1 of Li & Barber (2017) holds for the p-values in* (9).

Note that Theorem 1 accommodates the p-values in (9), but Theorem 2 would not hold with those in (8) because n_j^0 may vary across j, breaking the symmetry needed by our martingale proof. Nonetheless, simulations suggest this may not be a problem in practice; see Figure S7.

5. CONSISTENT GENOME-WIDE ASSOCIATIONS ACROSS DIVERSE ANCESTRIES

5.1. Missing variants and knockoffs in genome-wide association studies

Genome-wide association studies search for variants with biological effects on a phenotype. These studies are exploratory, aiming to prioritize genetic loci for follow-up investigations, and nowadays largely concentrate on polygenic phenotypes using large samples, resulting in numerous associations and making it desirable to control the false discovery rate (Storey & Tibshirani, 2003). As the DNA is fixed prior to any phenotypes, it is relatively easy to deduce causal relations. It is also reasonable that all humans share the same biology: modulo some genetic diversity across populations, we can imagine a common causal mechanism and attempt to uncover which variants are involved in it by testing $\mathcal{H}_j^{\mathrm{cst}}$ (2) or $\mathcal{H}_j^{\mathrm{pcst},r}$ (3).

In practice, only a few hundred thousands variants across the genome are measured (genotyped), as sequencing is expensive. Such relatively few markers can capture most genetic variation because nearby alleles are in linkage disequilibrium (Slatkin, 1994): they have strong dependencies and can be accurately inferred from one another. This facilitates the localization of broad regions, loci, containing associations with the phenotype, but it complicates the attribution of distinct signals to specific variants. Indeed, many genotypes can be marginally associated with the phenotype simply because they are in linkage disequilibrium with the same causal variant; conditional testing alleviates this issue but does not fully account for missing variants.

The traditional analysis imputes the missing variants using linkage disequilibrium models fitted on fully-sequenced reference samples (Marchini & Howie, 2010). The imputed variants are analyzed alongside the typed ones, through either genome-wide marginal tests or multivariate linear models within narrow regions (Schaid et al., 2018). However, this is not fully satisfactory because imputation is not as informative as a direct measurement; in fact, imputed variants carry no information in addition to that contained in the typed ones, as they are a function of the latter, conditionally independent of the phenotype. This may pass unnoticed if one fully trusts a multivariate linear model for the outcome—imputed variants may be significant within such models because their dependence with the measured variants is nonlinear—but it makes it impossible to find evidence that a missing variant is causal within our model-X perspective (Sesia et al., 2020). Therefore, we will leverage consistency to gather indirect evidence of causal associations. First though, we must briefly recall some details of the existing methodology.

The current knockoff analysis partitions the genome into contiguous segments and tests whether any of these contain conditional associations (Sesia et al., 2020). Letting $G \subset [p]$ index the genotypes in one segment, knockoffs can test a slightly generalized version of $\mathcal{H}_j^{\mathrm{ci},e}$ (1):

$$\mathcal{H}_G^{\mathrm{ci},e} : Y^e \perp\!\!\!\perp X_G^e \mid X_{-G}^e, \tag{10}$$

where $X_G = \{X_j : j \in G\}$, $X_{-G} = \{X_j : j \notin G\}$. Under $\mathcal{H}_G^{\mathrm{ci},e}$ (10), the variants in G are independent of the phenotype conditional on all other genotypes. This analysis can be performed at different resolutions, separately controlling the false discovery rate for increasingly refined genomic partitions to balance between power and the value of each discovery (Sesia et al., 2020). Hypotheses involving smaller groups (higher resolution) are harder to reject because the variables have strong local dependencies, making it difficult to distinguish the signal of one variant from those of its neighbors. Higher-resolution hypotheses are more specific and informative. Although $\mathcal{H}_G^{\mathrm{ci},e}$ (10) is not asking whether a genetic segment contains causal variants, it is a reasonable proxy. Intuitively, we will verify that low-resolution hypotheses are more robust to missing variants. The robustness of $\mathcal{H}_G^{\mathrm{ci},e}$ (10) is less clear at high resolution because there each tested segment contains few measured genotypes; this is where consistency will be most useful.

5.2. *Linkage disequilibrium in populations with different ancestries*

Linkage disequilibrium is explained by the inheritance of long and randomly cut genetic segments from parents to offspring, with occasional mutations. Generation after generation, the genotypes thus come to resemble an imperfect mosaic of ancestral motifs, which can be encoded as a hidden Markov model (Li & Stephens, 2003); this is at the heart of imputation (Marchini & Howie, 2010) and knockoff generation (Sesia et al., 2018). The block-like patterns of linkage disequilibrium vary across populations because these share different recent ancestors, and so their mosaics involve different patterns, and possibly different transition (recombination) rates (Laan & Pääbo, 1997). In other words, different populations differ by covariate shift. This heterogeneity has already been factored into the generation of knockoffs for pooled analyses (Sesia et al., 2021), and it will be leveraged here to help highlight causal variants.

Figure 4 visualizes why covariate shift helps localize causal variants within an example involving two populations, four observed and five missing variables, one of which is causal. This toy genome is partitioned into segments containing one or two typed variants each; three segments are highlighted. Linkage disequilibrium is described by hidden Markov models yielding genetic blocks separated by high-recombination spots (Wall & Pritchard, 2003). If the alleles across these spots are independent, the only consistent association is that of the segment containing the causal variant. Of course, in reality linkage disequilibrium is not perfectly organized into blocks, although this is a common simplification (Berisa & Pickrell, 2016). Further, we can only study a limited number of populations, so not all spurious associations may be removed. Nonetheless, consistency enables a useful step forward in a challenging problem.

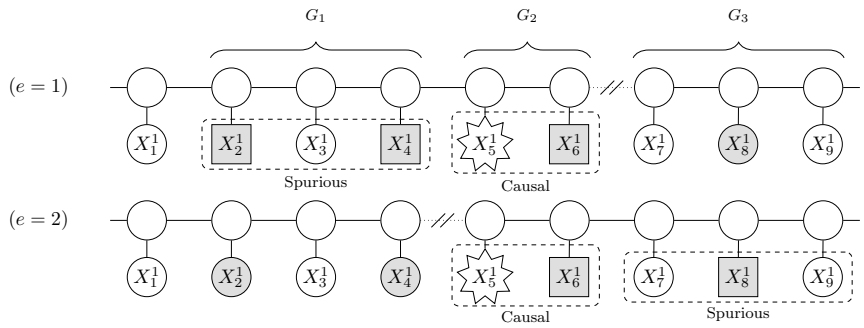

Fig. 4: Consistency in genome-wide association data from two populations. The unobserved causal variant (star-shaped node) induces different spurious associations depending on the patterns of linkage disequilibrium, described by population-specific hidden Markov models. Shaded nodes: measured variables. Squares: variables that are not conditionally independent of the causal one. The broken segments symbolize the boundaries of linkage disequilibrium blocks.

6. NUMERICAL EXPERIMENTS

6.1. *Setup*

Our method is applied with the separate and joint analysis statistics describe above. Intersection and pooling are taken as benchmarks. All statistics are computed with the R package `glmnet` (Friedman et al., 2010), or `bigstatsr` (Privé et al., 2019) for genetic data. Our computer code is available from `https://github.com/lsn235711/MEKF_code`. Several additional experiments are in Section S5 due to lack of space.

6.2. Testing for full consistency with synthetic data

In each environment, p variables X are generated from an autoregressive model of order one with correlation 0.2. We leverage the known P_X to construct semi-definite Gaussian knockoffs (Candès et al., 2018). The distribution of $Y^e \mid X^e$ in the e-th environment is given by a logistic model with logit $\text{pr}\,(Y^e = 1 \mid X^e) = X^e\beta^e$, where $\beta^e \in \mathbb{R}^p$ are environment-specific effects. We consider two settings corresponding to different E, p, and β^e.

In setting one, $E = 4$, $p = 500$, and there are $n = 1000$ observations per environment. First, 100 effects are randomly chosen to be non-zero in all environments; then, for each e, 10 of the remaining ones are non-zero in all but the e-th environment, and these 40 associations are thus not consistent. See Figure S8 (a) for a sketch of this setup. The absolute values of the 100 consistent effects are $a/\sqrt{n}$, where a is a signal amplitude parameter which we will vary; the remaining non-zero values are $0.5a/\sqrt{n}$. The effect signs are independent coin flips. In setting two, $E = 3$, $p = 200$, and $n = 2000$, while β^e is determined as follows. First, 50 effects are randomly chosen to be non-zero in all environments; then, for each e, 50 of the remaining ones are non-zero in all but the e-th environment; see Figure S8 (b). The 100 consistent effects are $a/\sqrt{n}$ in absolute value, and the remaining ones are $0.5a/\sqrt{n}$. All effect signs are independent coin flips.

Our goal is to discover the subset of consistent effects, controlling the false discovery rate below 10%. Figure 1, previewed earlier, compares our method to the benchmarks, averaging over 100 experiments. Here, our p-values (7) are filtered with selective SeqStep+, using $c = 1/2$. The results confirm our method controls the false discovery rate, as predicted by the theory. The joint analysis statistics are more powerful than the separate analysis ones in the first setting, in which most associations are consistent, and equivalent to the latter in the second setting, where most associations are not consistent. Pooling yields too many false discoveries because it reports all associations regardless of whether they are consistent, while the intersection heuristic may lead to either low power (first setting) or high false discovery rate (second setting).

6.3. Causal inference in a simulated genome-wide association study

We analyze simulated yet realistic genetic data involving different populations, based on the haplotypes in the 1000 Genomes Project (Consortium et al., 2015). This resource contains phased haplotypes from five populations: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). We utilize these haplotypes to simulate genetic data from a hidden Markov model for 50,000 individuals (10,000 per population); see Section S6. This approach ensures the genotype distribution is known exactly, allowing us to concentrate on missing variants, as we can simply apply the algorithm from Sesia et al. (2020) to generate knockoffs separately for each population. We construct knockoffs for testing $\mathcal{H}_G^{\text{ci},e}$ (10) at two resolutions, with genetic segments of median lengths 233 kb or 15 kb. In the interest of time, we only analyze 359,811 biallelic single-nucleotide polymorphisms on chromosome 22.

Conditional on the genotypes, we simulate a continuous trait for all 50,000 individuals from a constant linear model with independent Gaussian errors and 50 causal variants. The causal variables are randomly chosen such that each population has at least 10 with minor allele frequency above 0.1. The effect signs are independent coin flips, and their sizes are inversely proportional to the standard deviation of the allele count, so that rarer variants have larger effects. The total heritability of the trait is varied. All causal variants are unmeasured, so that their exact identification is impossible; however, we can localize genetic segments likely to contain them. The proportion of typed variants is varied between 1% and 10%. Knockoffs are constructed only for the measured variants. This setup is particularly challenging because our genotyping is random, while real studies often preferentially type potentially interesting variants (Bycroft et al., 2018). Thus, confounding may be a less severe problem in practice.

We carry out conditional independence tests at the 10% false discovery rate level but measure performance in stricter terms, based on the causal false discovery rate and power: a discovery is counted as true if and only if it reports a genetic segment containing a causal variant. The power is defined as the average proportion of segments containing causal variants that are discovered. All results are averaged over 100 experiments with independent traits. In theory, the genotypes should also be resampled to ensure false discovery rate control because the knockoffs treat them as random; however, that would be computationally expensive with such large data.

Figure 5 (a) summarizes the results of separate population-specific analyses with 1% genotyping density. These analyses correctly test conditional association but do not yield valid causal inferences, demonstrating the need for consistency, especially at high resolution. Our method is applied with $r = 3$, and with separate analysis statistics due to the large size of the data. Statistical significance is determined with the accumulation test or with selective SeqStep+, using $c = 1/2$ for the latter. Our method controls the causal false discovery rate and, when applied with the accumulation test, is only slightly less powerful than pooling at low resolution. The selective SeqStep+ tends to yield lower power, plausibly because it extracts less information from the p-values. At higher resolution, our method is not as powerful as pooling while the causal false discovery rate inflation of the latter becomes more severe. Unsurprisingly, all methods are less powerful at high resolution. The causal false discovery rate violation for the intersection is smaller but noticeable.

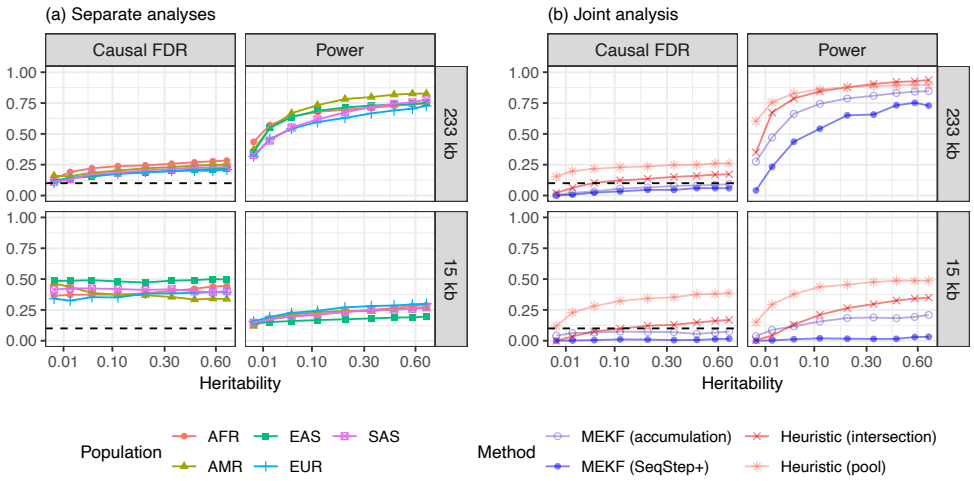

Fig. 5: Analysis of a simulated multi-population genome-wide association study with missing causal variants. Top: low-resolution analysis (233 kb); bottom: high-resolution analysis (15 kb). The empirical performance is evaluated in a strict causal sense. Our method seeks associations supported by the data from at least 3 populations. The nominal false discovery rate is 10%.

Figure S10 shows confounding decreases as the genotyping density increases, unsurprisingly. As the density increases, our method becomes very conservative compared to pooling; this may seem unavoidable but Figure 1 suggests our method may perform better with the cross-prior statistics. Figure S11 shows similar results corresponding to analyses at the 20% false discovery rate level, emphasizing the causal type-I error inflation incurred by the heuristics. Figure S12 shows our method performs similarly regardless of whether the accumulation test is applied to the p-values computed with random tie breaking (9) or without it (8).

7. Analysis of UK Biobank genome-wide association data

7.1. Data pre-processing

We study four continuous traits (body mass index, height, platelet count, and systolic blood pressure) and four diseases (cardiovascular disease, diabetes, hypothyroidism, and respiratory disease) using the UK Biobank (Bycroft et al., 2018) data; see Table S1 for more details. This analysis is based on the same pre-processing and knockoffs for 486,975 genotyped and phased subjects in the UK Biobank (application 27837) as in Sesia et al. (2021). The knockoffs preserve the population structure, including familial relatedness; this accounts for most possible confounders except missing variants. Our goal is to address this remaining limitation. As in Sesia et al. (2021), we only analyze 591,513 biallelic single nucleotide polymorphisms with minor allele frequency above 0.1% and in Hardy-Weinberg equilibrium (10^{-6}) among the subset of 350,119 unrelated British individuals previously studied by Sesia et al. (2020). The genome is partitioned at 7 levels of resolution, ranging from that of single polymorphisms to that of 425 kb-wide groups. The resolution of each partition is defined as its median segment width.

The UK Biobank subjects are divided into five populations based on their self-reported ancestry (African: 7,635; Asian: 3,284; British: 429,934; Non-British European: 28,994; and Indian: 7,628). We exclude subjects with unreported ancestry, as well as those outside these five categories; this leaves us with a total of 477,475 individuals; see Table S2 for more details.

7.2. Searching for consistent associations

We apply our method to search for associations consistent in at least r environments, with $r = 2, 3, 4, 5$. Significance is computed by applying the accumulation test to the p-values in (8) because this test without the random tie breaking (9) tends to be more powerful than selective SeqStep+ (Section 6), and tie breaking seems practically unnecessary; see Figure S7. The analysis is performed at the 10% false discovery rate level, separately for each resolution (Sesia et al., 2020). We apply separate analysis statistics because the data set is very large. The intersection heuristic and the pooled analysis from Sesia et al. (2021) will serve as benchmarks.

All tests are repeated for 100 independent realizations of U_j (8); this allows some understanding and a reduction of the variability of any findings, as our method is randomized. Alternatively, one may repeat the entire analysis starting from the knockoff generation (Ren et al., 2021); however, that would be impractical for such large data. In comparison, resampling U_j is computationally negligible. Table S3 reports the numbers of discoveries for height and platelet count obtained in at least 51% of the randomizations. The results for other phenotypes are in Table S4. Unfortunately, there are fewer consistent associations for the other phenotypes, consistently with previous observations that height and platelet count display the strongest signals (Sesia et al., 2020, 2021). This reporting rule is not guaranteed to control the false discovery rate, but the simulations in Figure S13 empirically confirm it to be conservative. The variability of the findings over different p-value randomizations is summarized in Figure S14.

Several consistent associations are discovered, although these are relatively few compared to those obtained by pooling because our power is limited by the paucity of non-British samples. Fortunately, the awareness that genetic studies should increase the representation of different ancestries (Duncan et al., 2019) suggests promising future opportunities, especially as some large diverse studies already exist (Gaziano et al., 2016). Some of our discoveries for platelet count are visualized in Figure 6 through a Chicago plot (Sesia et al., 2020). It is not guaranteed that all discoveries corresponding to a fixed r are also found with $r' < r$, although this occurs often; see Figure S15. The findings obtained with selective SeqStep+ instead of the accumulation test, as well those obtained with the intersection heuristic, are summarized in Table S5.

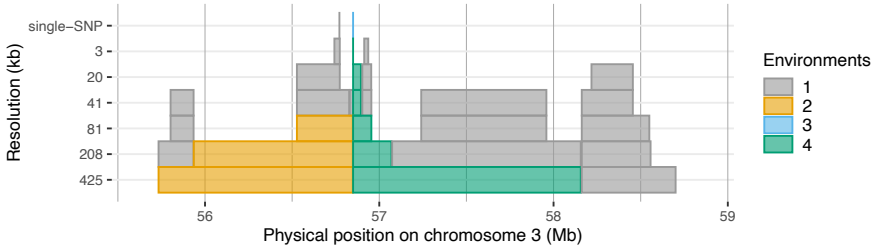

Fig. 6: Some discoveries for platelet count based on UK Biobank data from five populations (environments). Each block represents a genetic segment containing distinct associations; the colors indicate the numbers of environments across which they are consistent. The vertical position denotes the resolution of the discovery measured in millions of base pairs (Mb).

7.3. Validation of genetic findings

Table S6 demonstrates almost all of our consistent discoveries for height and platelet count are confirmed by the NHGRI-EBI GWAS Catalog (Buniello et al., 2019) (accessed on April 15, 2021). We say that a discovered genetic segment is confirmed if it spans a region containing reported associations for the same phenotype. Relatively fewer discoveries obtained by pooling are thus confirmed. Of course, this is not fully conclusive because the GWAS Catalog may include spurious associations and is likely to miss many causal ones, although it is a standard reference. Table S6 also summarizes the numbers of findings obtained with the intersection heuristic, as well as the proportions of those which are confirmed by the GWAS Catalog. This shows the intersection heuristic yields either fewer discoveries, or a (slightly) lower validation rate. This is consistent with our simulations suggesting the intersection heuristic is often either underpowered or excessively liberal. Analogous information for the other phenotypes is in Table S7. Table S8 reports the names and associated genes of the genetic variants identified by our method at the single-nucleotide resolution. These results indicate all but two of our high-resolution consistent discoveries correspond to variants with known biological consequences, which are located on genes previously reported to be associated with the phenotypes of interest. The full list of discoveries is available online at `https://msesia.github.io/knockoffgwas/`.

ACKNOWLEDGEMENTS

We thank Stefan Wager, the associate editor and two referees for their insightful comments, as well as CARC at the University of Southern California and SRCC at Stanford University for computing resources. The authors acknowledge support from the following grants: NIH grant R56 HG010812, R01 MH113078, R01 MH123157; NSF grants OAC 1934578, DMS 2032014; ONR grant N00014-20-12157; Simons Foundation award 814641; and Technion Career Advancement Fellowship. We are grateful to the participants and investigators of the UK Biobank.

SUPPLEMENTARY MATERIAL

Supplementary Material available at Biometrika online includes mathematical proofs, a discussion of conditional randomization, further simulations, and additional tables and figures.

REFERENCES

BARBER, R. F. & CANDÈS, E. (2015). Controlling the false discovery rate via knockoffs. *Ann. Stat.* **43**, 2055–2085.

BARBER, R. F., CANDÈS, E. & SAMWORTH, R. (2020). Robust inference with knockoffs. *Ann. Stat.* **48**, 1409–1431.

BATES, S., CANDÈS, E., JANSON, L. & WANG, W. (2020a). Metropolized knockoff sampling. *J. Am. Stat. Assoc.* , 1–15.

BATES, S., SESIA, M., SABATTI, C. & CANDÈS, E. (2020b). Causal inference in genetic trio studies. *Proc. Natl. Acad. Sci. U.S.A* **117**, 24117–24126.

BENJAMINI, Y. & HELLER, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64**, 1215–1222.

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

BERISA, T. & PICKRELL, J. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283.

BOOLEN, K. (1989). *Structural Equations with Latent Variables*. Wiley, New York.

BUNIELLO, A., MACARTHUR, J., CEREZO, M., HARRIS, L., HAYHURST, J., MALANGONE, C., MCMAHON, A., MORALES, J., MOUNTJOY, E., SOLLIS, E. et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012.

BYCROFT, C., FREEMAN, C., PETKOVA, D., BAND, G., ELLIOTT, L. T., SHARP, K., MOTYER, A., VUKCEVIC, D., DELANEAU, O., O'CONNELL, J., CORTES, A., WELSH, S., YOUNG, A., EFFINGHAM, M., MCVEAN, G., LESLIE, S., ALLEN, N., DONNELLY, P. & MARCHINI, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209.

CANDÈS, E., FAN, Y., JANSON, L. & LV, J. (2018). Panning for gold: "model-x" knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. B* **80**, 551–577.

CASTRO, D., WALKER, I. & GLOCKER, B. (2020). Causality matters in medical imaging. *Nat. Commun.* **11**, 1–10.

CHIA, C., SESIA, M., HO, C.-S., JEFFREY, S. S., DIONNE, J. A., CANDES, E. & HOWE, R. T. (2021). Interpretable classification of bacterial raman spectra with knockoff wavelets. *IEEE J. Biomed. Health Inform.* , 1–1.

CONSORTIUM, . G. P. et al. (2015). A global reference for human genetic variation. *Nature* **526**, 68.

DEVLIN, B. & ROEDER, K. (1999). Genomic control for association studies. *Biometrics* **55**, 997–1004.

DUNCAN, L., SHEN, H., GELAYE, B., MEIJSEN, J., RESSLER, K., FELDMAN, M., PETERSON, R. & DOMINGUE, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328.

EFRON, B. (2020). Prediction, estimation, and attribution. *J. Am. Stat. Assoc* **115**, 636–655.

FAN, Y., LV, J., SHARIFVAGHEFI, M. & UEMATSU, Y. (2020). IPAD: stable interpretable forecasting with knockoffs inference. *J. Am. Stat. Assoc* **115**, 1822–1834.

FITHIAN, W. & LEI, L. (2020). Conditional calibration for false discovery rate control under dependence. *arXiv preprint arXiv:2007.10438* .

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1.

GAZIANO, J., CONCATO, J., BROPHY, M., FIORE, L., PYARAJAN, S., BREELING, J., WHITBOURNE, S., DEEN, J., SHANNON, C., HUMPHRIES, D. et al. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223.

GIMENEZ, J. R., GHORBANI, A. & ZOU, J. (2019). Knockoffs for the mass: new feature importance statistics with false discovery guarantees. In *22nd International Conference on Artificial Intelligence and Statistics*. PMLR.

HARFORD, T. (2014). Big data: A big mistake? *Significance* **11**, 14–19.

HARGITTAI, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *Ann. Am. Acad. Pol. Soc. Sci.* **659**, 63–76.

HECKMAN, J. J. (1979). Sample selection bias as a specification error. *Econometrica* , 153–161.

HEINZE-DEML, C., PETERS, J. & MEINSHAUSEN, N. (2018). Invariant causal prediction for nonlinear models. *Journal of Causal Inference* **6**.

HERNÁN, M. A. & TAUBMAN, S. L. (2008). Does obesity shorten life? the importance of well-defined interventions to answer causal questions. *Int. J. Obes.* **32**, S8–S14.

HUME, D. (1739). *A Treatise of Human Nature: A Critical Edition*. London: John Noon.

IMBENS, G. W. & RUBIN, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. CUP.

KATSEVICH, E. & RAMDAS, A. (2020). A theoretical treatment of conditional independence testing under model-X. *preprint at arXiv:2005.05506* .

KATSEVICH, E., SABATTI, C. & BOGOMOLOV, M. (2021). Filtering the rejection set while preserving false discovery rate control. *J. Am. Stat. Assoc* , 1–27.

KOSINSKI, M., STILLWELL, D. & GRAEPEL, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 5802–5805.

LAAN, M. & PÄÄBO, S. (1997). Demographic history and linkage disequilibrium in human populations. *Nat. Genet.* **17**, 435–438.

LEE, Y. & OGBURN, E. L. (2020). Network dependence can lead to spurious associations and invalid inference. *J. Am. Stat. Assoc* , 1–15.

LI, A. & BARBER, R. F. (2017). Accumulation tests for FDR control in ordered hypothesis testing. *J. Am. Stat. Assoc* **112**, 837–849.

LI, N. & STEPHENS, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233.

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

MOOIJ, J. M., MAGLIACANE, S. & CLAASSEN, T. (2020). Joint causal inference from multiple contexts. *J. Mach. Learn. Res.* **21**, 1–108.

NEYMAN, J. & IWASZKIEWICZ, K. (1935). Statistical problems in agricultural experimentation. *Supplement to to J. R. Stat. Soc.* **2**, 107–180.

PEARL, J. (2009). *Causality*. Cambridge university press.

PETERS, J., BÜHLMANN, P. & MEINSHAUSEN, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. B* , 947–1012.

PRITCHARD, J. & PRZEWORSKI, M. (2001). Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet* **69**, 1–14.

PRIVÉ, F., ASCHARD, H. & BLUM, M. G. (2019). Efficient implementation of penalized regression for genetic risk prediction. *Genetics* **212**, 65–74.

REN, Z., WEI, Y. & CANDÈS, E. (2021). Derandomizing knockoffs. *J. Am. Stat. Assoc* **0**, 1–11.

ROJAS-CARULLA, M., SCHÖLKOPF, B., TURNER, R. & PETERS, J. (2018). Invariant models for causal transfer learning. *J. Mach. Learn. Res.* **19**, 1309–1342.

ROMANO, Y., SESIA, M. & CANDÈS, E. (2019). Deep knockoffs. *J. Am. Stat. Assoc.* **0**, 1–27.

ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P. & PETERS, J. (2021). Anchor regression: Heterogeneous data meet causality. *J. R. Stat. Soc. B* .

RUBIN, D. B. (2005). Causal inference using potential outcomes. *J. Am. Stat. Assoc* **100**, 322–331.

SCHAID, D. J., CHEN, W. & LARSON, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504.

SESIA, M., BATES, S., CANDÈS, E., MARCHINI, J. & SABATTI, C. (2021). False discovery rate control in genome-wide association studies with population structure. *Proc. Natl. Acad. Sci. U.S.A.* **118**.

SESIA, M., KATSEVICH, E., BATES, S., CANDÈS, E. & SABATTI, C. (2020). Multi-resolution localization of causal variants across the genome. *Nat. Commun.* **11**, 1–10.

SESIA, M., SABATTI, C. & CANDÈS, E. (2018). Gene hunting with hidden Markov model knockoffs. *Biometrika* **106**, 1–18.

SHALIZI, C. R. & THOMAS, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociol. Methods Res.* **40**, 211–239.

SHEN, A., FU, H., HE, K. & JIANG, H. (2019). False discovery rate control in cancer biomarker selection using knockoffs. *Cancers* **11**, 744.

SLATKIN, M. (1994). Linkage disequilibrium in growing and stable populations. *Genetics* **137**, 331–336.

STOREY, J. D. & TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9440–9445.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288.

WALDRON, L., HAIBE-KAINS, B., CULHANE, A., RIESTER, M., DING, J., WANG, X., AHMADIFAR, M., TYEKUCHEVA, S., BERNAU, C., RISCH, T., GANZFRIED, B., HUTTENHOWER, C., BIRRER, M. & PARMIGIANI, G. (2014). Comparative Meta-analysis of Prognostic Gene Signatures for Late-Stage Ovarian Cancer. *J. Natl. Cancer Inst.* **106**. Dju049.

WALL, J. & PRITCHARD, J. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **4**, 587–597.

WANG, W. & JANSON, L. (2020). A power analysis of the conditional randomization test and knockoffs. *preprint at arXiv:2010.02304* .

YU, K., GUO, X., LIU, L., LI, J., WANG, H., LING, Z. & WU, X. (2020). Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys (CSUR)* **53**, 1–36.

[*Received on ? ? ?. Editorial decision on ? ? ?*]