Biometrika (2018), **yd**, y, *pp.* 1–5
Printed in Great Britain yd

Discussion of Sesia et al. and the Knockoff Framework

BY AND JONATHAN D. ROSENBLATT
Dept. of Industrial Engineering and Management,
Ben Gurion University of the Negev, Israel.
johnros@bgu.ac.il

YA'ACOV RITOV
Department of Statistics, University of Michigan, USA
yritov@umich.edu

JELLE J. GOEMAN
Department of Biomedical Data Sciences, Leiden University Medical Center, The Netherlands.
J.J.Goeman@lumc.nl

1. ON THE MOTIVATION

The authors of Sesia et al. (2018) set out to design a procedure for variable selection with provable finite sample statistical guarantees. The *knockoff* algorithm proposed by Sesia et al. (2018), provably controls the FDR of conditionally independent variables. Denoting with $X := (X_j)_{j=1}^p$ and Y the predictor and outcome variables, respectively. The *false discovery proportion*, a.k.a. the *false selection proportion*, is defined as $\mathrm{FDP} := \mathrm{V}/\mathrm{R}$ where R is the number of variables selected, and V is the number of falsely selected. A false selection defined by Sesia et al. (2018) to be a selected X_j where $Y|X_{-j}$ is independent of X_j. The knockoff algorithm of Sesia et al. (2018) provably controls the $\mathrm{FDR} := \mathbb{E}[\mathrm{FDP}]$, at some user selected magnitude.

The fundamental idea of the method is to generate another set of variables, the *knockoffs*, that are exchangeable with the original X_js, only that conditionally on X they are uncorrelated with Y. The method then proceeds to compute a test statistic that captures the difference between the apparent dependence between Y and X_j and to that of Y with the knockoff $\tilde{X}_j$.

Crucially for our discussion: (1) The FDR is an expectation with respect to variability in X and Y, i.e., it is based on a *random design* guarantee. (2) The procedure is *model-free*, i.e. *non-parametric*, with respect to prediction, in that nothing is assumed on the form of $Y|X$. (3) The framework assumes full knowledge of F_X, i.e., the joint distribution of predictors, marginalized over Y. In application it is based on the assumption that F_X can be estimated well enough to be assumed known. (4) The method aims at good variable selection and not prediction, estimation nor ranking.

We think of the method in Sesia et al. (2018) as an adaptation of Candès et al. (2018) to genome-wide association studies (GWAS). The differences between the two: (1) The non-null variables in Candès et al. (2018) are those that belong to the minimal set that renders all others independent, i.e., the non-null is the Markov-Blanket of X on Y. In Sesia et al. (2018), the non-null variables are those with non-null partial correlations. Under some minimal conditions these two sets are the same. (2) Candès et al. (2018) discusses a multivariate Gaussian model, while Sesia et al. (2018) a hidden Markov model (HMM). An important contribution in each paper is an algorithm for sampling knockoffs from the assumed model.

The method of Sesia et al. (2018) is also similar in flavor to Barber & Candès (2015), with the following differences: (1) Barber & Candès (2015) assume a linear generative model, so that the null is simply $H_j : \beta_j = 0$. (2) Barber & Candès (2015) crucially assume $n > p$, and Gaussian distributed errors. (3)

Barber & Candès (2015) control the fixed-design FDR, which is stronger, and implies, random-design control.

2. ON THE PROBLEM SETUP

The problem setup in Candès et al. (2018) and Sesia et al. (2018) deals with random design inference, in a non-parametric model. We find this to be a very useful setup for screening problems: It is consistent with the random designs typically found in observational studies, and it avoids the very-useful-yet-controversial linearity assumption.

A more surprising component of the problem setup is the knowledge of F_X, i.e., the joint distribution of predictors, marginalized over Y. Many authors would consider this an Oracle assumption, and given the difficulty of estimating joint distributions, an unrealistic one. The GWAS application, however, represents an ideal situation in which much is known about F_X, e.g. from the HapMap project (International HapMap Consortium, 2003). Other areas of potential application include Semi-Supervised Machine Learning in which F_X may be inferred from extra observations of X. In general high-dimensional data settings, however, F_X has so many parameters that estimating it may turn out be a more difficult problem than estimating the conditional distribution of $Y|X$. In all cases, robustness to errors in F_X is crucial for the practical usefulness of the method, and we are happy to see the promising preliminary results of Candès et al. (2018). On the other hand, the pruning step in the GWAS example of Sesia et al. (2018) suggests that strong dependencies may adversely affect the performance of the method; both with respect to power, and with respect to the tails of the FDP distribution. We expect this matter to be further investigated in the future.

3. KNOCKOFFS AS PSEUDO-VARIABLES

The idea of augmenting design matrices with random variables is not new. It has been suggested many times, for the purposes of prediction, variable ranking, large-sample support recovery, etc. Tusher et al. (2001) have already proposed the idea of permuting the original variables for FDR control on selected variables. While intuitive and elegant, their algorithm did not have any provable guarantees, and implies marginal nulls and not conditional. Some more algorithms adding "fake", "phony", "probes" or "pseudo variables", are reviewed in Guyon & Elisseeff (2003).

Perhaps the most similar work is that of Wu et al. (2007), which not only propose adding "pseudo-variables" for the purpose of estimating the variable selection FDR, but also require two conditions very similar to the knockoff conditions. Wu et al. (2007) require that: "(A1) real unimportant variables and phony unimportant variables have the same probability of being selected on average", and "(A2) real important variables have the same probability of being selected whether or not phony variables are present". These two conditions cannot be satisfied according to Wu et al. (2007), but they are clearly related to the *pairwise exchangeability* and *nullity condition* in Sesia et al. (2018) and Candès et al. (2018). One may thus view the two knockoff conditions as a satisfiable version of Wu's A1 and A2. To the credit of Wu et al. (2007) we quote their insights, which already hint at what will be later formalized in the knockoff conditions: "Permutation produces pseudovariables that when appended to the real data create what are essentially matched pairs. To each real variable there corresponds a pseudo variable with identical sample moments and also with preservation of correlations".

4. CONDITIONAL RESAMPLING

The invariance property in Sesia et al. (2018) describes so-called *null-invariant transformations* (Goeman & Solari, 2010): the joint distribution of the augmented data $(Y, X, \tilde{X})$ is invariant under the transformation swap(S) provided that S is a set of true nulls. Known null-invariant transformations, e.g. permutations and rotations (Langsrud, 2005), existed so far only for null hypotheses about marginal association. E.g. Tusher et al. (2001) use a knockoff-like framework to test marginal, not conditional, nulls. This

method was recently proven to control tail probabilities of the False Discovery Proportion (Hemerik & Goeman, 2018). Which null hypothesis is to be preferred? Conditional of marginal? Think of two correlated SNPs. One causal and the other is not. In a screening experiment such as GWAS, we would like both to be selected. It thus seems to us that for the purpose of screening, marginal nulls, not conditional, are desirable.

Knockoffs represent, as far as we know, the first null-invariant transformations with respect to conditional nulls rather than marginal nulls. Seeing knockoffs as null-invariants opens the way for their more classical use, e.g. using multiple random knockoffs for controlling familywise error in the manner of Westfall & Young (1993). Since familywise error control is the norm in the field of GWAS, the latter would be a worthwhile extension. An alternative resampling approach for testing for dependence between Y and X_j would be to sample from the conditional distribution of X_j given X_{-j}. This approach was found to be powerful in Candès et al. (2018), but dismissed there for computational reasons; it was not considered again in Sesia et al. (2018). This may be a missed opportunity. Conditonal resampling can be done easily in the context of hidden Markov models, where, for example, if $Y \in \{0, 1\}$, one can easily test for conditional independence in the $2 \times L$ table of Y and X_j. Resampling-based methods such as Westfall & Young (1993) are computationally quite efficient, requiring only a small multiple of α^{-1} resampled data sets for powerful familywise error control at level α. Alternatively, applying early stopping (e.g. Jiang & Salzman, 2012), and tail resampling (e.g. Yu et al., 2011) can be used for additional computational efficiency.

5. Power Considerations and Counter Examples

There are many examples were the knockoffs work excellently well and a few of them are given by the authors. We give below counter examples. We find it useful if the authors advise us when to use the knockoff method and when other methods will be more practical.

In the set up we consider the number of variables is ultra high (i.e., in the $p \gg n$ situation), most of them are true nulls. To be more precise, the asymptotics we consider are such that the proportion of false hypotheses $\dagger_1$ goes to 0. Moreover, we assume that the p test statistics have sub-Gaussian distribution. We argue that in our models, if the the familywise-error-rate (FWER) is q, then the Bonferroni correction procedure has more power than the knockoff methods. The argument stands on two legs. Firstly, an optimal FDR procedure gives only a small power gain relative to an FWER procedure. Secondly, the knockoff methods is losing considerable power due to the built-in randomization.

The main justification for preferring the FDR criterion over the FWER one is that the FWER seems to be too conservative and if the number of the true alternatives, $|S|$ is large, we can permit some false detections. However under some standard assumptions, if we consider p test statistics each of them with (marginally) $N(0, 1)$ distribution under the null, rejecting $q|S|$ true nulls on the average means rejecting variables with test statistics greater than $\sqrt{2\log(p/q|S|)}$ and not $\Phi^{-1}(1 - q/p) \approx \sqrt{2\log(p/q)}$ as would be implied by a standard Bonferroni correction . That implies the effect size of an alternative can be $\left\{1 + \log(|S|)/\log(p/q)\right\}^{1/2}$ smaller while keeping its power level constant. However, this ratio is very close to 1 when $p \gg q|S|$ as it is the case in most $p \gg n$ models of interest. This leaves a small margin of inefficiency for a method that its main justification is a proven FDR. We argue now that knockoff methods, being randomized procedures, are strictly inefficient in this situation.

Knockoff methods reject all variables such that $W_j \geq \hat{t}$, where W_j is the difference in the criterion applied to X_j and $\tilde{X}_j$, and

$$\hat{t} = \inf\Big\{t : \frac{c + |\{j : W_j \leq -t\}|}{1 \vee |\{j : W_j \geq t\}|} \leq q\Big\},$$

where $c \in \{0, 1\}$. Setting $c = 0$ yields the standard knockoff method, which provably controls only $mFDR = \mathbb{E}\Big[|\hat{S} \cap \mathcal{H}_0|/(|\hat{S}| + 1/q)\Big] \leq q$. Setting $c = 1$ yields the *knockoff+* method, which provably control the usual FDR.

What happens when the number of true alternatives is small? If $|\hat{S}| \not\gg q^{-1}$ then $mFDR \ll FDR$. In fact, the actual FDR under the null of the knockoff method with $c = 0$ is 1/2 (i.e. $P(\max W_j > -\min W_j)$), and not the P-value as it is supposed to be. The reason that $mFDR > FDR$ is that $E|\{W_j > \hat{t}\}| = 1 + E|W_j < -\hat{t}|$. Thus if the number of detected true alternatives is not much larger than $1/q$, we are likely to reject one null too many. To correct it is suggested to use $c = 1$, which has a proven FDR, but the knockoff+ method does not make sense unless it is known that $q|\mathcal{H}_1| \geq 1$. Otherwise, no matter how large is the effect size of these alternative the power is approximately $2^{q^{-1}-|S|}$, since we need that the rejected set will be at least greater than $1/q$, or that (essentially) all of $\{W_j : |W_j| > \hat{t}\}$, will be positive. If $q|S| \gg 1$ then both methods have a reasonable and provable FDR performance.

We come now to the potential loss of efficiency of the knockoff method as described. Suppose, for example, that the distribution of T_j has much heavier far away tails for null variables. The selection cutoff will be driven by the few selected null variables with heavier tails, at the risk of masking non-nulls. The knockoff method thus does not save the need for very careful selection of the test statistics, and using it with difficult to analyze large dimensional methods like the lasso is questionable, where the standard error of each variable is different.

But the main problem seems to be a direct power loss due to the randomization. Here is a toy problem. Consider the very simple model where we observe $Y = \beta^T X + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, $X \in R^{2p}$, $(X_1, X_2), \ldots, (X_{2p-1}, X_{2p})$ are i.i.d., $(X_{2j-1}, X_{2j})^T \sim N_2(0, \Sigma)$ where $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Candès et al. (2018) show that the knockoffs are given by

$$\mathcal{L}\Big\{ \begin{pmatrix} \tilde{X}_{2j-1} \\ \tilde{X}_{2j} \end{pmatrix} | X = x \Big\} = N_2\Big((I + \lambda\Sigma^{-1}) \begin{pmatrix} x_{2j-1} \\ x_{2j} \end{pmatrix}, \quad 2\lambda I - \lambda^2 \Sigma^{-1} \Big),$$

where $\lambda < 2(1 - \rho)$. However, tedious calculations show the effect size needed to be detected with reasonable power is larger by at least a factor of $(1 - \rho^2)^{-1/2}$ (for the optimal selected λ which is the minimal). As we mentioned, the main motivation to use FDR is that it reduces the needed effect size by a factor of $(1 + \log(|S|)/\log(p/q))^{1/2}$. This may be a negligible gain relative to the above loss due to randomization. It seems that keeping the family-wise error-rate at q using a Bonferroni method will yield more power (and hence, in effect, better detection) than the randomized knockoff method.

6. ON SAMPLING FROM HMMS

In Sesia et al. (2018) the explanatory variables are considered as HMM: there is a latent Markov process Z, such that the observations $X_1, \ldots, X_p$ are independent given Z and $\mathcal{L}(X_j|Z) = \mathcal{L}(X_j|Z_j)$. The knockoff construction algorithm is based on drawing a single sample from the Markov process $X \to Z \to \tilde{Z} \to \tilde{X}$, where $(\tilde{Z}, \tilde{X})$ is a knockoff of (Z, X). It is not clear to us why the extra step of a random sample of $\tilde{Z}$ is needed, and why we cannot sample directly according to the Markov process $X \to Z \to \tilde{X}$ where (Z, X) and $(Z, \tilde{X})$ are identically distributed.

Is knowledge of F_X realistic for HMMs? Now, any process can be represented as a weak limit of *non-stationary* HMMs. The difficulty is, of course, with the number of needed parameters. If you consider K hidden states and L possible output values then the number of parameters is approximately $pK(K + L - 2)$. With the n and p as in the examples, the demand that the estimated model will be considered as "known", enforces quite as small K, and hence the HMM assumption is restrictive. An HMM is not Markovian, but it is exponentially mixing. Practically speaking, K small does not enable medial range non-monotone dependency, as one may find in GWAS due to Linkage-Disequilibriumm. If n forces K small, so that dependencies are local, one may be better of considering a test under local dependencies. In such a test, marginal and conditional nulls are the same except locally. One may thus test for independence between Y and X_j and some of its neighbors, while ignoring distant j's.

7. FUTURE RESEARCH

We find the knockoff framework to be quite exciting. Not because it offers solutions to all possible difficulties, but on the contrary: because it sets the stage for many important research questions. Some of the following questions are already acknowledged in Candès et al. (2018, Sec.7.2): Are error guarantees **robust** to misspecification of F_X? What test statistics have more **power**? Knockoffs are not uniquely defined so how to best generate them? Do methods defined for other null-invariants, such as permutations, generalize to knockoffs easily? How to sample knockoffs efficiently? Does screening with knockoffs have more power than the linear model (even if miss-specified)? What other algorithms are possible assuming F_X known? What can we borrow from the semi-supervised learning literature to the knockoff setup? Are the conditional nulls of Sesia et al. (2018) preferable over marginal nulls such as in Tusher et al. (2001)?

Dai & Barber (2016), Janson & Su (2016), Chen et al. (2017b), Chen et al. (2017a), and others, have already started to explore and extend the knockoff framework of Barber & Candès (2015). We expect many such explorations in the upcoming future.

ACKNOWLEDGEMENT

The authors thank Dr. Aldo Solari, Dr. Livio Finos, Prof. Yuekai Sun, and the students in the University of Michigan Stats 710 class for fruitful discussions leading to this manuscript, while not necessarily agreeing with its conclusions.

REFERENCES

BARBER, R. F. & CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43**, 2055–2085. 00123.

CANDÈS, E., FAN, Y., JANSON, L. & LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 551–577. 00000.

CHEN, J., HOU, A. & HOU, T. Y. (2017a). A Pseudo Knockoff Filter for Correlated Features. *arXiv:1708.09305 [math, stat]* 00000.

CHEN, J., HOU, A. & HOU, T. Y. (2017b). Some Analysis of the Knockoff Filter and its Variants. *arXiv:1706.03400 [math, stat]* 00001.

DAI, R. & BARBER, R. (2016). The knockoff filter for FDR control in group-sparse and multitask regression. In *International Conference on Machine Learning*. 00008.

GOEMAN, J. J. & SOLARI, A. (2010). The sequential rejection principle of familywise error control. *The Annals of Statistics* , 3782–3810.

GUYON, I. & ELISSEEFF, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* **3**, 1157–1182. 11418.

HEMERIK, J. & GOEMAN, J. J. (2018). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 137–155.

INTERNATIONAL HAPMAP CONSORTIUM (2003). The international hapmap project. *Nature* **426**, 789.

JANSON, L. & SU, W. (2016). Familywise error rate control via knockoffs. *Electronic Journal of Statistics* **10**, 960–975. 00011.

JIANG, H. & SALZMAN, J. (2012). Statistical properties of an early stopping rule for resampling-based multiple testing. *Biometrika* **99**, 973–980.

LANGSRUD, Ø. (2005). Rotation tests. *Statistics and computing* **15**, 53–60.

SESIA, M., SABATTI, C. & CANDES, E. J. (2018). Gene hunting with hidden markov model knockoffs. *Biometrika* 00000.

TUSHER, V. G., TIBSHIRANI, R. & CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98**, 5116–5121. 12117.

WESTFALL, P. & YOUNG, S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Wiley-Interscience.

WU, Y., BOOS, D. D. & STEFANSKI, L. A. (2007). Controlling Variable Selection by the Addition of Pseudovariables. *Journal of the American Statistical Association* **102**, 235–243.

YU, K., LIANG, F., CIAMPA, J. & CHATTERJEE, N. (2011). Efficient p-value evaluation for resampling-based tests. *Biostatistics* **12**, 582–593.

[*Received* y yd YandY. *Editorial decision on* y yd YandY]