Biometrika (2018), **yyy**, y, *pp*. 1–4
Printed in Great Britain

Discussion of Sesia et al. and the Knockoff Framework

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

AND 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 statistical guarantees. The *knockoff* algorithm proposed by Sesia et al. (2018), provably controls the FDR of conditionally independent variables. Denoting with x and y the predictor and outcome variables, respectively. The *false discovery proportion*, a.k.a. the *false selection proportion*, is defined as $FDP := V/R$ 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 control the $FDR := \mathbb{E}[FDP]$, at some user selected magnitude.

[TODO: consistent notation x or X? JDR:x for random predictors, and X for the observed design.]

The fundamental idea of the method is to generate variables that have all the properties of the original x_j, only that they are conditionally uncorrelated to y. These are termed *knockoff* variables. The method then proceeds to compute a test statistic that captures the difference between the dependence of y to x_j and to its knockoff. [TODO: how is a variable actually selected?]

Crucially for our discussion: (1) The FDR is an expectation with respect to variability in x and y, i.e., a *random design* guarantee. (2) The procedure is *model-free*, i.e. *non-parametric*, in that nothing is assumed on the form of $y|x$. (3) The proofs assume full knowledge of F_x, i.e., the joint distribution of predictors, marginalized over y. (4) The method aims at good variable selection, not prediction.

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. (2) Candès et al. (2018) discusses a multivariate Gaussian model, while Sesia et al. (2018) a hidden Markov model. 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) infer conditional on $X_{n\times p} := (x_1, \ldots, x_n)'$. They thus control a fixed-design FDR, and not the random-design.

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 generative 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 (Consortium et al., 2003). Other areas of potential application include semi-supervised settings 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). The pruning step in the GWAS example of Sesia et al. (2018) suggests that high collinearity may adversely affect the performance of the method. We suspect that the higher the correlations in F_x, the more precisely F_x should be known. 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, consistent 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: $H_j : Cov[y, x_j] = 0$, and not $H_j : Cov[y, x_j|x_{-j}] = 0$. 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 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. ON NULL HYPOTHESES AND INVARIANTS

The problem of variable selection with error guarantees is not new. Previously proposed algorithms include, for instance, Stability Selection (Meinshausen & Bühlmann, 2010), SURE Screening (Fan & Lv, 2008), BOLASSO (Bach, 2008), Benjamini-Gavrilov (Benjamini & Gavrilov, 2009), and many more. These procedures propose varying algorithms, with varying statistical guarantees in varying scenarios. We do not review this literature for the sake of brevity. We do, however, wish to discuss the matter of identifiability and estimability. I.e., is the parameter well defined, and is the estimation problem well-posed?

When doing variable selection, one will always require some assumption to ensure that "a good" selection is well defined. For this purpose a linear generative model is typically assumed. In the linear generative case multicollinearity will render the problem non-identifiable. To ensure identifiability in the fixed design, authors have proposed various conditions such as *Sparse Eigenvalue*, *Sparse Riesz Condition*, *Neighbourhood Stability*, *Irrepresentable Condition*, and *Exact Recovery Criterion*. See Meinshausen &

Bühlmann (2010, Sec 3.1.1) for a review. In the random-design literature, identifiability is typically ensured with some restriction on the condition number of $Cov[x]$, or with some strong-convexity assumption on the risk surface.

How is identifiability ensured in the knockoff framework? The crucial assumption is the knowledge of the distribution F_x, which avoids difficult to check sparsity conditions. [TODO Jelle: I argue identifiability is neither ensured nor required in this setup] The resulting 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 the 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. As far as we know, knockoffs are the first null-invariant transformations with respect to conditional nulls, and not marginal nulls. Seeing knockoffs as null-invariants opens the way for their more classical use, e.g. using multiple random knockoffs (Hemerik & Goeman, 2018), or 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 worthwile extension.

5. The Question of causality

The random-design knockoff framework does not deal with the linear generative case, so that null and alternative cannot be stated in terms of β coefficients. On the other hand, support recovery, unlike consistent estimation, does not require an identifiable model. Errors, and thus FDR, can thus be defined without identifiable parameters.

Lacking a parametric generative model, null and alternative have a more information-theoretic flavor, and are stated in terms of dependence. Candès et al. (2018) define the null using a Markov-Blanket, and Sesia et al. (2018) use conditional independence. This raises several questions: Are these hypotheses consistent with the motivating problems? Why this change in hypotheses?

We find that the conditional independence nulls is not quite consistent with screening problems such as GWAS. To see this, consider two perfectly correlated SNPs. One causal and the other not. In a screening study, we want to discover both. Using the conditional independence hypotheses in Sesia et al. (2018), however, they both belong to the null. Using the Markov-Blanket hypotheses in Candès et al. (2018), they will also belong to the null. A similar concern was raised by J.T. Kent in the discussion of Meinshausen & Bühlmann (2010). It seems to us that for the purpose of screening, marginal hypotheses, such as the ones in Tusher et al. (2001), are more appropriate. If the purpose were not screening for associations, but rather, causal inference, then clearly the definition of hypotheses, and errors, would have required a different language altogether. Maybe the language of Do-Calculus (Pearl, 1995) would have been more appropriate The authors, however, postpone the causal question: "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." [TODO: do we have an opinion about this?]

6. 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. 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 them efficiently? Does screening with knockoffs have more power than the linear model (even if miss-specified)?

The assumption of knowing F_x, even without the knockoff variables, paves the way for interesting research. For instance, in an observational study assuming a linear generative model, why not estimate $Var[\hat{\beta}]$ with $\mathbb{E}[(xx')^{-1}]\sigma^2$ instead of $(X'X)^{-1}\sigma^2$? [TODO: Yanki- does your on LSE example fit here?]

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 Prof. Yaakov Ritov, Dr. Aldo Solari, Dr. Livio Finos, and ... for fruitful discussions leading to this manuscript.

[TO DISCUSS; in relation to identifiab

REFERENCES

BACH, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*. ACM.

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

BENJAMINI, Y. & GAVRILOV, Y. (2009). A simple forward selection procedure based on false discovery rate control. *The Annals of Applied Statistics* **3**, 179–198. 00049.

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.

CONSORTIUM, I. H. et al. (2003). The international hapmap project. *Nature* **426**, 789.

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

FAN, J. & LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911.

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.

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

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

MEINSHAUSEN, N. & BÜHLMANN, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 417–473.

PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82**, 669–688.

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.

[*Received* y *yyy* yyyy. *Editorial decision on* y *yyy* yyyy]