Biometrika (2018), **yyy**, y, *pp.* 1–3
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
TOOD: Jelle
J.J.Goeman@lumc.nl

1. ON THE PROBLEM

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 detection proportion*, a.k.a. the *false selection proportion*, is defined as V/R where R is the number of variables selected, and V is the number of such variables 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.

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 in the strength of the dependence of the knockoff, and the original variable. This statistic is then compared to it resampling distribution: the distribution over resampled knockoffs.

Crucially for our comments: (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*, or *non-parametric* in that nothing is assumed on the parametric 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. The non-null variables in Sesia et al. (2018) are those with non-null partial correlation. (2) Candès et al. (2018) discusses a multivariate Gaussian model, while Sesia et al. (2018) a hidden Markov model. Each paper offers an sampling algorithm for sampling knockoffs from the assumed model.

The method of Sesia et al. (2018) is similar in flavor to Barber & Candès (2015), but Sesia et al. (2018) is quite more general: (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.

2. OTHER KNOCKOFFS

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. Some notable examples include the authors' own Candès et al. (2006). 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. 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, but they are clearly related to the *pairwise exchangeability* and *nullity condition* in Sesia et al. (2018) and Candès et al. (2018).

The impossibility to satisfy A1 and A2 was already observed by Wu et al. (2007). 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".

3. OTHER VARIABLE SELECTION METHODS

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.

4. PERMUTATION TESTING AND SYMMETRIES

5. FUTURE RESEARCH

6. ON THE HYPOTHESES

7. ON THE PROBLEM SETUP

ACKNOWLEDGEMENT

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

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.

CANDÈS, E. J., ROMBERG, J. & TAO, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory* **52**, 489–509.

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.

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

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

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.

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]