Biometrika (2018), **yyy**, y, *pp*. 1–2
Printed in Great Britain

Discussion of Sesia et al. 2018 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.

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. ON THE METHOD

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.

ACKNOWLEDGEMENT

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

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.

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

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