Biometrika (2019), **106**, 1, *pp.* 19–22
Printed in Great Britain
doi: 10.1093/biomet/asy063

Discussion of 'Gene hunting with hidden Markov model knockoffs'

By L. BOTTOLO

Department of Medical Genetics, University of Cambridge, J. J. Thomson Avenue, Cambridge CB2 0QQ, U.K.
lb664@cam.ac.uk

and S. RICHARDSON

MRC Biostatistics Unit, University of Cambridge, Robinson Way, Cambridge CB2 0SR, U.K.
sylvia.richardson@mrc-bsu.cam.ac.uk

1. Introduction

We congratulate the authors on this interesting paper that tackles the difficult problem of extending the work of Candès et al. (2018) to structured predictors possessing a particular type of dependence. Motivated by the aim of making the knockoff construction applicable to genome-wide association studies, Sesia et al. (2019) propose using a hidden Markov model to generate knockoffs that capture patterns of DNA variation, following the work of Scheet & Stephens (2006).

The principle underlying the knockoff method is to generate mock predictors $\tilde{X}$, which should be representative of, and exchangeable with, the true X but, by construction, not linked to the response Y. In order to claim robustness with respect to the conditional model of $Y \mid X$, the knockoff principle moves the goal post for false discovery rate-controlling variable selection to that of generating mock $\tilde{X}$. The framework thus strongly hinges on how plausible and computationally feasible it is to generate such $\tilde{X}$ for each particular class of problem and predictor structure. Generating faithful patterns of DNA variation has been at the heart of genetic research aimed at devising computer simulation programs to reproduce realistic DNA sequence data under complex demographic patterns and genetic features such as population bottleneck and expansion, natural selection, mutation and recombination (Peng et al., 2015).

Ideas derived from the coalescent theory have been used in a series of papers by Stephens and colleagues (Stephens et al., 2001; Stephens & Scheet, 2005) to develop methods that can perform haplotype inference and imputation of missing values in a population of unrelated individuals. Unfortunately, these methods only allowed analysis of short DNA sequences since they were very computationally intensive, especially for large numbers of individuals. A breakthrough came from Scheet & Stephens (2006), who, instead of using the computationally expensive coalescent-based product of approximate conditionals likelihood (Li & Stephens, 2003; Stephens & Scheet, 2005), assumed that haplotypes are generated from a cluster model where each cluster consists of a group of closely related haplotypes and the component membership changes continuously along the genome. However, their work was restricted to estimation of the haplotype phase in a homogeneous population, or with known subpopulations, an important point to which we will return later.

2. GENERAL STRATEGIES FOR CALIBRATING TEST STATISTICS AND FALSE DISCOVERIES

Genetic X generated by computer simulation programs have commonly been used to assess the suitability of new methods in a variety of demographic and genetic contexts. However, the jump to using simulators to create mock $\tilde{X}$ that are assumed to be exchangeable with observed X had, up to now, not been made. Instead, statistical genetic analysis turned to using permutation tests (Good, 2013), where the predictors X are kept fixed and a suitable permutation is applied to the response Y, effectively breaking the genotype-phenotype relationship of the dataset. A permutation-based approach has also been advocated in the Bayesian analysis of genome-wide association studies by Stephens & Balding (2009), and was used in Bottolo et al. (2013) to evaluate decision rules and calibrate cut-offs for Bayes factors.

Furthermore, in the context of sparse genome-wide association studies, where very few associations are expected with regard to the vast number of features, even without permutation, it can be assumed that most of the test statistics come from the null and hence a mixture distribution can be used for the purpose of false discovery rate control (Müller et al., 2006). Control of the false discovery rate via permutation-based methods is not yet fully understood, but some solutions have been proposed (Xie et al., 2005).

It is surprising that no connection has been made by Sesia et al. (2019) to either the literature on permutation-based methods in genome-wide association studies or the literature on false discovery rate control through mixture distributions. We would be interested to see what the authors think of the pros and cons of these approaches with respect to implementing the knockoffs in terms of efficiency and empirical control of the false discovery rate.

3. CONFOUNDERS IN GENETIC ASSOCIATION STUDIES

The method of Scheet & Stephens (2006) is suitable for haplotype inference and imputation of missing values in a homogeneous population of nominally unrelated individuals, or with known subpopulations. It cannot be used for more complex demographic patterns, in particular for admixed populations (Falush et al., 2003) or for explicitly related samples (O'Connell et al., 2014).

Sesia et al. (2019) claim that their knockoff method provides a new tool for making more discoveries in genome-wide association studies. However, one important point that they seem to have overlooked is the presence of population structure. This has been recognized as one of the main confounding effects that needs to be accounted for in any association study (Pritchard et al., 2000). It is not clear to us how the proposed hidden Markov model generating framework could be extended in an easy way, since the estimated parameters and the generative hidden Markov model have to be conditioned on the latent population structure. The same problem arises in the presence of another important type of genetic confounding, cryptic relatedness (Astle & Balding, 2009). In their real data analysis Sesia et al. (2019) removed the effect of strong population stratification on the responses by using the first five principal components, but they did not account for it in the generation of the knockoffs, casting doubts on the crucial assumption of exchangeability, which is at the heart of their approach to false discovery rate control. It would be interesting to read their thoughts on how the proposed generative hidden Markov model should be modified to account for the near-ubiquitous presence of confounding factors in genome-wide association studies.

4. COMPUTATIONAL EFFICIENCY AND DIMENSION REDUCTION

There is clearly a cost in generating the mock $\tilde{X}$, and the procedure would be more efficient if a single pass were enough. Nevertheless, Sesia et al. (2019) recommend multiple passes, since the knockoff method is intrinsically stochastic, but then the difficulty arises of not knowing how to combine the results of the multiple passes and thus losing the theoretical guarantees (Candès et al., 2018). Hence the strong theoretical guarantees which motivated the work are ultimately lost for the recommended procedure, somewhat defeating the purpose. Moreover, the implementation of the method might be computationally infeasible for a large population such as the UK Biobank (Sudlow et al., 2015). We therefore feel that the scope of the hidden Markov model's extension to gene hunting is more limited than the authors claim, and

not yet extendable to realistic large-scale genome-wide association studies analysis on cohorts such as the UK Biobank where cryptic relatedness is also present (O'Connell et al., 2016).

Finally, the prefiltering strategy seems a shortcut to decrease the computational cost of any multivariate learning procedure that needs to be performed on the augmented predictor space $(X, \tilde{X})$, with the number of models growing exponentially as 2^{2p}. The chosen prefiltering cut-off of 0.5 is usually not recommended since an expensive downstream fine-mapping analysis is then required to identify putative causal variants. A higher tagging value of 0.8 is the standard rule, which achieves a good compromise between narrowing down the number of genetic markers and identifying important associations with sufficient precision. Irrespective of the pruning strategy, with the availability of large cohorts that have millions of genotyped and imputed variants (Nelson et al., 2017), performing any kind of multivariate analysis on the augmented predictor space $(X, \tilde{X})$ might not be computationally feasible.

5. Generalizability to other -omics datasets

Our final comments express some concern regarding the generalizability of generating knockoffs for false discovery rate control to the analysis of a range of -omics datasets. Each particular data type will have its own dependence structure inherited from a succession of complex biological processes, and reproducing such structure faithfully will require an involved modelling effort (Chen et al., 2015), which will be difficult to check. Permutation-based false discovery rate control does not require such generative models and seems a promising alternative.

References

Astle, W. & Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statist. Sci.* **24**, 451–71.

Bottolo, L., Chadeau-Hyam, M., Hastie, D. I., Zeller, T., Liquet, B., Newcombe, P., Yengo, L., Wild, P. S., Schillert, A., Ziegler, A. et al. (2013). GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS Genet.* **9**, e1003657.

Candès, E., Fan, Y., Janson, L. & Lv, J. (2018). Panning for gold: Model-X knockoffs for high dimensional controlled variable selection. *J. R. Statist. Soc.* B **80**, 551–77.

Chen, H.-S., Hutter, C. M., Mechanic, L. E., Amos, C. I., Bafna, V., Hauser, E. R., Hernandez, R. D., Li, C., Liberles, D. A., McAllister, K. et al. (2015). Genetic simulation tools for post-genome wide association studies of complex diseases. *Genet. Epidemiol.* **39**, 11–19.

Falush, D., Stephens, M. & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**, 1567–87.

Good, P. (2013). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer.

Li, N. & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–33.

Müller, P., Parmigiani, G. & Rice, K. (2006). FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics 8: Proceedings of the Valencia/ISBA 8th World Meeting on Bayesian Statistics (Benidorm, Alicante, Spain, June 1–6, 2006)*, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith & M. West, eds. Oxford: Oxford Univiversity Press, pp. 349–70.

Nelson, C. P., Goel, A., Butterworth, A. S., Kanoni, S., Webb, T. R., Marouli, E., Zeng, L., Ntalla, I., Lai, F. Y., Hopewell, J. C. et al. (2017). Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nature Genet.* **49**, 1385–91.

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J. E., Rudan, I. et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10**, e1004234.

O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., Zagury, J.-F., Delaneau, O. & Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. *Nature Genet.* **48**, 817–20.

Peng, B., Chen, H.-S., Mechanic, L. E., Racine, B., Clarke, J., Gillanders, E. & Feuer, E. J. (2015). Genetic data simulators and their applications: An overview. *Genet. Epidemiol.* **39**, 2–10.

Pritchard, J. K., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–59.

Scheet, P. & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–44.

SESIA, M., SABATTI, C. & CANDÈS, E. (2019). Gene hunting with hidden Markov model knockoffs. *Biometrika* **106**, 1–18.

STEPHENS, M. & BALDING, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Rev. Genet.* **10**, 681–90.

STEPHENS, M. & SCHEET, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76**, 449–62.

STEPHENS, M., SMITH, N. J. & DONNELLY, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–89.

SUDLOW, C., GALLACHER, J., ALLEN, N., BERAL, V., BURTON, P., DANESH, J., DOWNEY, P., ELLIOTT, P., GREEN, J., LANDRAY, M. et al. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779.

XIE, Y., PAN, W. & KHODURSKY, A. B. (2005). A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics* **21**, 4280–8.

[*Received on* 1 *October* 2018. *Editorial decision on* 2 *October* 2018]