Daniel YEKUTIELI

Department of Statistics and OR
School of Mathematical Sciences
Tel Aviv University

Research Statement – February 2015

1 Overview

In my research I develop, theoretically study, and implement methodology for providing statistical inference for large or complicated data. In my MSc. and PhD. theses, supervised by Prof. Yoav Benjamini, and in the beginning of my research career I worked on frequentist False Discovery Rate (FDR) controlling methodology. But in the last few years the focus of my work has changed to Bayesian methodology.

In Section 2 I discuss my important contributions to frequentist FDR methodology. I describe my work on Bayesian methodology in Section 3. I outline the projects I am currently working on in Section 4.

2 Contributions to FDR methodology

Yekutieli and Benjamini (1999) introduced a resampling based approach to control the FDR and used it to discover northern hemisphere regions whose atmospheric pressure is correlated to the precipitation in Israel; this methodology is applied to microarray analysis in the very highly cited paper Reiner, Benjamini and Yekutieli (2003). However, I think that the main significance of Yekutieli and Benjamini (1999) is that it expressed control over the FDR in a form that is much much more amenable to statistical investigation that was the basis for the seminal work on Bayesian FDR of Bradley Efron and John Storey.

Benjamini and Yekutieli (2001) is a very highly cited paper that derives an explicit expression for the FDR that is used in many paper on FDR methodology, and uses this expression to prove that the BH procedure controls the FDR for the more general case of positively dependent test statistics and to produce a general upper bound for the FDR of the BH procedure.

Benjamini, Krieger and Yekutieli (2006) is a highly cited paper that introduces an adaptive procedure in which the BH procedure is applied twice: once at the nominal level q to estimate the proportion of true null hypotheses $\hat{\pi}_0$, and

a second time at higher level $q/\hat{\pi}_0$ to decide which hypotheses to reject. The paper provides exact FDR computations and proves FDR control of adaptive procedure for independently distributed test statistics, and shows in simulations that their procedure controls the FDR for dependent test statistics, while other, more powerful adaptive procedures, fail to control the FDR for dependent test statistics.

The BH procedure can only be applied to control the FDR for a single family of null hypotheses. In large-scale studies it is usually necessary to consider several families of hypotheses and since a positive answer to one question usually leads to followup questions, the hypotheses are often considered and tested hierarchically. Yekutieli (2008) derives a bound for the FDR of the entire study in cases where the BH procedure is separately applied to several families of hypotheses and introduces a hierarchical testing approach that controls the FDR.

The multiplicity problem is identified in the statistical literature with providing simultaneous and selective inference. Simultaneous inference is inference that is simultaneously valid for all the parameters in the study and it is solved by controlling for the Family-wise Error rate. Benjamini and Yekutieli (2005) argues that selective inference is the distinct problem of providing inference for parameters that are selected after viewing the data. Benjamini and Yekutieli (2005) suggest control over the false coverage-statement rate (FCR) – a generalization of the FDR, defined as the expected proportion of non-covering confidence intervals out of the total number of confidence intervals – as a the measure for the validity of confidence intervals constructed for selected parameters. Benjamini and Yekutieli (2005) also presents a general method for adjusting confidence intervals for selection and show that applying this method to tests yields the BH procedure.

3 Bayesian methodology

Even though I think that Benjamini and Yekutieli (2005) was very important conceptually, the frequentist selective inference framework it presented is very limited and the method for controlling the FCR has many shortcomings: the selection adjustment can only be applied to marginal confidence intervals for a set of a-priori considered parameters, not to confidence intervals for functions of several parameters, nor to point estimators for the parameters; and the selection adjustment is the same for all parameters, regardless of the parameter's estimate

and the selection rule used.

I had initially started working on Bayesian methodology to answer a question that has puzzled me for some time: Bayesian inference is generally assumed to be unaffected by selection, why is it necessary to correct frequentist inference for selection and not necessary to correct Bayesian inference for selection? I was also hoping that if I could understand why Bayesian inference has to be corrected for selection then I could develop a framework for providing selective inference that doesn't have the limitations of the FCR approach. While working on this problem I became more and more interested in Bayesian methoology, I taught two courses on Bayesian inference and eventually decided to change the focus of my research and my collaborative efforts from frequentist to Bayesian methodology.

Yekutieli (2012) introduces a general framework for providing comprehensive Bayesian inference for selected parameters. In Bayesian selective inference for each potential object of inference, that can be any function of the parameter, there is a selection event in the data sample space such that inference is provided for the object inference only if the selection event is observed. The data used for providing selective inference is a realisation of the joint distribution of the parameters and the data truncated by the selection event, and thus the average risk incurred in providing selective inference is the expectation of the loss over the joint truncated distribution of the parameter and the data. Yekutieli (2012) defines selection-adjusted Bayesian inferences as the Bayes rules that minimize the truncated average risk. Yekutieli (2012) shows that the way that truncation acts on the parameter determines whether and how the Bayesian inference is affected by selection, and demonstrates how to implement this approach in a simple data example. Yekutieli (2012) also shows that Bayesian FDR procedures are empirical Bayes selection rules that produce calibrated posterior error rates because the prior distribution on which they are based is the empirical parameter distribution, and explains how to generalize this approach to increase its scope of application.

I had also realized that in addition to offering a general comprehensive solution to the problem of selective inference, the Bayesian paradigm can also yield optimal calibrated tests in cases where frequentist methods fail either because the data is very high dimensional and it is not clear how to construct good tests or because the tested null hypothesis is composite and frequentist tests that have to be valid for every null parameter value are inherently underpowered. Thus in big data applications Bayesian methods can produce incomparably better statis-

tical tests than frequentist methods. Heller and Yekutieli (2014) illustrates this point: it presents an empirical Bayes algorithm for deriving optimal tests that are considerably more powerful than frequentist FDR controlling procedures for discovering genetic loci with replicated associations in several Genome-Wide Association Scans (GWAS).

Yekutieli (2014), presents methodology for constructing significance tests for "difficult" composite alternative hypotheses that have no natural test statistic. The tests are Bayesian extensions of the likelihood ratio test, they are optimal with respect to the prior distribution, and are also closely related to Bayes factors and Bayesian FDR controlling testing procedures. Yekutieli (2014) applies this methodology for constructing exact tests for cross tabulated data, where the motivating example is constructing a test for discovering Simpson's Paradox.

Amar et al. (2014) presents an algorithm for finding coherent and flexible modules in 3-way data. The algorithm is based on a hierarchical Bayesian data model and Gibbs sampling. It outperforms extant methods on simulated data. For gene expression time series measurements of patients after septic shock, the method was able to dissect key components of the response and detected patient-specific module augmentations that are informative for disease outcome. In analyzing brain fMRI time series of subjects at rest, the method detected the pertinent interacting brain regions.

4 Ongoing research

My primary research interest is a collaboration with Dr. Amit Saad, a physician from Shalvata Hospital in Israel, to develop an inferential framework, based on the optimal Bayesian testing approach of Yekutieli (2014), for predicting treatment effects from the Cochrane Collaboration reviews. The Cochrane Collaboration is an independent (non-profit and non-governmental) organization that conducts very systematic and extensive reviews of health-care interventions that are meant to help health-care providers, patients, and policy makers to make informed evidence-based medical decisions. We use the Cochrane Collaboration review, for a given outcome, to provide a confidence statement regarding the distribution of the treatment effect of the same outcome in a new treatment group. For the case where data from a small pilot study assessing the treatment efficacy in the new treatment group is available, we extend this framework to provide an updated and much tighter confidence statement for the distribution of the new group treatment effect.

Dr. Ruth Heller and I are beginning a new research project that extends our work on replicability (Heller and Yekutieli, 2014). We plan to develop a general framework for replicability analysis for multiple high-dimensional studies, specifically sequencing data from the ENCODE project. I am renewing my collaboration with Prof. Edward George from the University of Pennsylvania on developing a Bayesian framework for providing comprehensive (point estimators, credible sets and predictive distributions) post model-selection inference. I am also beginning a collaboration with several other researchers on developing eBayes methods for model selection.

References

[1] Amar D., Yekutieli D., Maron-Katz A., Hendler T., Shamir R., (2015) "A hierarchical Bayesian model for flexible module discovery in three-way time series data" Sumitted to ISMB/ECCB 2015.

[2] Benjamini Y., Hochberg Y., (1995). "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society B*, **57**, 289-300.

[3] Benjamini, Y., Krieger, A.M., Yekutieli, D. (2006) "Adaptive Linear Step-up False Discovery Rate controlling procedures" *Biometrika* (3): 491-507.

[4] Benjamini Y., Yekutieli D., (2001) "The Control of the False Discovery Rate in Multiple Testing under Dependency" , *The Annals of Statistics*, **29**, 4, 1165-1188.

[5] Benjamini Y., Yekutieli D., (2005) "False Discovery Rate-Adjusted Multiple Confidence Intervals for Selected Parameters" *Journal of the American Statistical Association*, **100**, 71.

[6] Y. Benjamini, D. Yekutieli "Quantitative traits loci analysis using the False Discovery Rate" Genetics 171(2) (2005), 783-789.

[7] Heller R., Yekutieli D., "Bayesian FDR procedure for discovering replicability in Genome-Wide Association Scans" (2014), *The Annals of Applied Statistics* **8** (1) 481-498.

[8] Reiner A., Yekutieli D., Benjamini Y., (2003) "Identifying differentially expressed genes using false discovery rate controlling procedures" *Bioinformatics*, **19**, 368-75.

[9] Yekutieli, D., "Hierarchical False Discovery Rate controlling methodology" *Journal of the American Statistical Association*, 2008, 103 (481) 309-316

[10] Yekutieli D. "Adjusted Bayesian inference for selected parameters" *Journal of the Royal Statistical Society: Series B*, 2012, **74** (3) 515 - 541.

[11] Yekutieli D., "Bayesian tests for composite alternative hypotheses in cross-tabulated data" (2014), *TEST* doi 10.1007/s11749-014-0407-1.

[12] Yekutieli D., Benjamini Y. (1999), "Resampling based false discovery rate controlling procedure for dependent test statistics", *Journal of Statistical Planning and Inference*, **82** (1-2), 171-196.