Daniel YEKUTIELI

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

Research Statement: June 2012

1 Overview

In the period since my previous research statement, I completed my research on hierarchical FDR controlling procedures (Yekutieli, 2008) and then decided to change the focus of my research to Bayesian methodology. I will discuss my work on Bayesian methodology in the next section. The emphasis in my collaborative efforts also shifted from frequentist analysis (Rubinshtein et al., 2008; Kafkafi et al., 2008; Kafkafi et al., 2009; Benjamini et al., 2009) towards Bayesian analysis. I am collaborating with Prof. Ziv Shkedy from the Center for Statistics in Hasselt University, co-editor and co-author of the monograph on analysis of dose response microarray experiments (Lin et al., 2012), on implementation of Bayesian FDR in analysis of dose response microarray experiments. I am collaborating with Dr. Ruth Heller from TAU on developing Bayesian FDR methodology for discovering replicability in Genome-Wide Association Scans and we also plan on implementing similar ideas for analyzing ChIP-sequencing data. I am collaborating with Prof. Edward George from the University of Pennsylvania on providing Bayesian post-model selection inference, we also plan to develop a Bayesian FDR framework for model selection. I am collaborating with Dr. Amit Saad on developing a Bayesian approach for integration of results from randomized multi-center clinical trials. I am also collaborating with Asaf Weinstein, a PhD student at the University of Pennsylvania, on frequentist methodology for making directional discoveries and constructing confidence intervals for the discovered effects.

2 My work on Bayesian methodology

Bayesian inference is generally assumed to be unaffected by selection (Senn, 2008). In Yekutieli (2012) I show that this is not necessarily the case. The observation that selection may affect Bayesian inference carries the important implication that in almost all Bayesian analyses (especially analyses of large

data sets in which there are many potential parameters to choose from), it is necessary to specify explicitly, for each potential parameter, a selection rule that determines when inference is provided for this parameter and to adjust the inference accordingly. In Yekutieli (2012) I present a framework for providing Bayesian inference for selected parameters. The second contribution is methodology for selecting parameters. I will discuss this aspect of my work in Section 2.2.

2.1 Bayesian adjusted inference for selected parameters

From a Bayesian perspective selective inference can be expressed as follows. θ is the parameter, $Y \in \Omega$ is the data. The parameters that may or may not be selected are functions of θ: $h_1(\theta), h_2(\theta), \ldots$, and for each $h_i(\theta)$ there is a subset $S^i_\Omega \subseteq \Omega$, such that inference is provided for $h_i(\theta)$ only if $y \in S^i_\Omega$ is observed. This means that providing selective inference is a truncation problem in which only realizations of (θ, y) with $y \in S^i_\Omega$ are used for providing inference on $h_i(\theta)$. Notice that selective inference for each parameter involves a separate truncation problem, and that the truncation involves both the data and the parameter. I show that the way that truncation acts on the parameter determines whether and how the Bayesian inference is affected by selection.

I call θ a fixed parameter in cases where only the conditional distribution of Y given θ is truncated, θ is a random parameter in cases where the joint distribution of θ and Y is truncated. The fixed parameters are generally fixed unknown constants whose value can be thought to be generated once from a prior distribution and remain unchanged, while the random parameters are usually the random effects whose values are generated, and thus truncated, concurrently with the data. For example, in an experiment comparing m groups of laboratory yields, where $\theta = (\theta_1 \cdots \theta_m)$ is the vector of expected yields and $Y = (Y_1 \cdots Y_m)$ is the vector of mean yields. θ is a fixed parameter when the groups correspond to different methods of making the particular chemical product. But when the groups correspond to different batches made by the same method θ is a random parameter.

Bayesian selective inference for $h_i(\theta)$ is based on the truncated conditional distribution of θ given $Y = y$, which I call the selection-adjusted posterior distribution. I show that if θ is a random parameter than the selection-adjusted posterior distribution is the same as the posterior distribution of θ and thus the Bayesian inference does not have to be corrected for selection. However, if θ is

a fixed parameter, or if θ is elicited a non-informative prior, then the selection-adjusted posterior distribution is different than the posterior distribution and thus the Bayesian inference must be corrected for selection.

2.2 Bayesian FDR controlling selection rules

The more widely studied aspect of selective inference is how to select interesting parameters. The frequentist approach to selecting parameters identifies selecting a parameter, an action called making a statistical discovery, with the rejection of a null hypothesis. A true discovery is rejecting a false null hypothesis and a false discovery is rejecting a true null hypothesis, i.e. committing a type-I error. Thus the decision whether to select multiple parameters is phrased as a multiple testing problem. In post-hoc analysis (Scheffe, 1953) the set of parameters that may be selected is the set of all contrasts of a vector of effects, and the selection rule that is applied a Family-Wise Error rate controlling multiple testing procedure that ensures that the probability of making at least one type-I error is less than α. Benjamini and Hochberg (1995) suggest a new paradigm for selecting interesting parameters: selection via multiple testing procedures that control the FDR at level α, which can be thought of as a frequentist mechanism for ensuring that the marginal conditional probability of committing a type-I error given selection is less than α.

Scott and Berger (2006) present a Bayesian approach for discovering active genes in a microarray experiment that declare a gene active if the posterior expected loss of this action is smaller than the posterior expected loss of declaring the gene inactive. However, the more common form of Bayesian parameter selection procedures are Bayesian FDR controlling methods that select a parameter if the probability of making a type-I error, for this parameter, is less than α. In Efron et al. (2001) a parameter is selected if the posterior probability of making a type-I error (the local FDR) is less than α, while Storey (2002, 2003) suggests specifying selection rules for which the conditional distribution of making a type-I error given selection (the pFDR) s less than α. Despite its great conceptual importance, the practical implication of Bayesian FDR methodology has been small. The only difference between the BH procedure and a pFDR controlling procedure is calibration – the level q the BH procedure controls the pFDR at level $m_0 q/m$ instead of q. In fact, the pFDR controlling suggested in Storey (2002) is equivalent to an adaptive BH procedure (Benjamini et al., 2006) in which the BH procedure is applied $mq/\hat{m}_0$ to ensure level q FDR control,

with $\hat{m}_0$ being an estimate of m_0.

In Yekutieli (2012) I study the relation between Bayesian selective inference and Bayesian FDR methodology and generalize Bayesian FDR methodology to increase its scope of application. I show that parameter selection can be expressed as a Bayesian selective inference problem in which θ is treated as a random parameter (regardless of whether it is indeed a random parameter) and that the local FDR is the posterior expected loss and pFDR is the average risk for the $0-1$ loss corresponding to to the event that null hypothesis is true. I generalize Bayesian FDR methodology by allowing the discovery event associated with selecting a parameter to be any subset of the parameter space not just the rejection of a null hypothesis. Note that this with this generalization I can apply the FDR paradigm for selecting parameters in more complex situations. For example, to discover that θ_i is the largest component in $\theta = (\theta_1 \cdots \theta_m)$ I only need to verify whether $\Pr\{\theta_i = max(\theta_1 \cdots \theta_m) | \ Data\} \geq 1 - \alpha$, whereas previously discovering that θ_i is the largest component involved testing the pairwise comparisons between θ_i and the other components of θ. Generalizing Storey (2007), I show that for any discovery event the optimal Bayesian FDR controlling selection rule is specified by the local FDR. I also present an eBayes approach for controlling the FDR for selecting parameters with different prior distributions.

References

[1] Benjamini Y., Heller R., Yekutieli D., (2009) "Selective inference in complex research", *Philosophical Transactions of the Royal Society A - Mathematical and Engineering Sciences*, **367** 4255-4271.

[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, Series B*; **57** (1): 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] Kafkafi N., Yarowsky P., Yekutieli D. (2008) "Data mining in a behavioral test detects early symptoms in a model of amyotrophic lateral sclerosis," *Behavioral Neuroscience*, 122 (4) 777-787

[5] Kafkafi N., Yekutieli D., Elmer G. I. (2009) "A Data Mining Approach to In Vivo Classification of Psychopharmacological Drugs," *Neuropsychopharmocology*, **34** 607-623.

[6] Lin D., Shkedy Z., Yekutieli D., Amaratunga, D. Bijnens, L. (Eds.) (2012) "Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R," Use R! Series, Springer.

[7] Rubinstein N.D., Mayrose I., Halperin D., Yekutieli D., Gershoni J.M., Pupko T., "Computational characterization of B-cell epitopes," Molecular Immunology, 2008, 45 (12) 3477-3489

[8] Scheffe' H. (1953) "A method for judging all contrasts in the analysis of variance," *Biometrika*, **40** 87-104.

[9] Scott J. G., Berger J. O. (2006) "An exploration of aspects of Bayesian multiple testing," *Journal of Statistical Planning and Inference*, **136** 2144-2162.

[10] Senn S. (2008) "A Note Concerning a Selection Paradox of Dawid's," *The American Statistician*, **62**, 206-210.

[11] Storey J. D. (2002) "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B*, **64** 479-498.

[12] Storey J. D., (2003) "The positive false discovery rate: A Bayesian interpretation and the q-value," *Annals of Statistics*, **31**, 2013-2035.

[13] Storey, J. D. (2007) "The optimal discovery procedure: a new approach to simultaneous significance testing," *Journal of the Royal Statistical Society, Series B*; **69** 347-368.

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

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