Bayesian Multiplicity Control

Jim Berger

Duke University

B.G. Greenberg Distinguished Lectures

Department of Biostatistics

University of North Carolina at Chapel Hill

May 13, 2016

Outline

- I. Introduction to multiplicity control
- II. Types of multiplicity control
- III. Variable selection
- IV. Subgroup analysis

I. Introduction to Multiplicity Control

An example of the need for multiplicity control:

In a recent talk about the drug discovery process, the following numbers were given in illustration.

- 10,000 relevant compounds were screened for biological activity.
- 500 passed the initial screen and were studied in vitro.
- 25 passed this screening and were studied in Phase I animal trials.
- 1 passed this screening and was studied in a Phase II human trial.

This could be nothing but noise, if screening was done based on 'significance at the 0.05 level.'

If no compound had any effect,

- about $10,000 \times 0.05 = 500$ would initially be significant at the 0.05 level;
- about $500 \times 0.05 = 25$ of those would next be significant at the 0.05 level;
- about $25 \times 0.05 = 1.25$ of those would next be significant at the 0.05 level
- the 1 that went to Phase II would fail with probability 0.95.

Multiplicity control is lacking in science:

- The tradition in epidemiology is to ignore multiple testing
  - usually arguing that the purpose is to find anomalies for further study.
    - But that distinction is rarely understood by readers of the article.
- Even fields that profess to care about such things as multiplicity adjustment often fail:
  - Rami Cohen studied a sample of 100 papers from NEJM from 2002-2010 that required treatment of multiplicity; only 13 even addressed the issue (and those inadequately).
- The tradition in psychology is to ignore optional stopping; if one is close to $p = 0.05$, go get more data to try to get there (with no adjustment).
  - *Example:* Suppose one has $p = 0.08$ on a sample of size n. If one takes up to four additional samples of size $\frac{n}{4}$, the probability of reaching $p = 0.05$, at some stage, is $\frac{2}{3}$.

- Multiple statistical analyses
  - Data selection "Torture the data long enough and they will confess to anything."
  - Removing 'outliers' (that don't seem 'reasonable')
  - Removing unfavorable data (see the report of the Staple committee):
    - From a senior social psychologist: "The goal is to show where the hypothesis holds, so it is of no use to include data or subjects where the agreement is not found, for any reason. Would they want a physicist to use data that went against their theories or that showed errors? No of course not, yet when we omit such data, we are accused of sloppy science. This is unfair."
  - *Verification bias:* Repeat an experiment with "improvements" until results agree with one's new theory (and report only that "working" experiment).
  - Trying out multiple models until 'one works.'
  - Trying out multiple statistical procedures until 'one reveals the signal.'
  - The Staple committee report gives 9 other variants on these.
    - From the report: "... several [social psychologists] ... defended the serious and less serious violations of proper scientific method with the words: that is what I have learned in practice; everyone in my research environment does the same, and so does everyone we talk to at international conferences."

The Two Main Approaches to Multiplicity Control

- *Frequentist approach:* A collection of techniques for penalizing or assessing the impact of multiplicity, so as to preserve an overall frequentist accuracy assessment.
  - The most basic (and general) frequentist approach is to repeatedly simulate the multiple testing scenario, under the assumption of 'no signal,' and estimate the probability of a false discovery.
    * The problem is that this can be computationally infeasible in modern big data problems.
  - Another common approach is to ignore the issue, assuming strict standards (e.g. 5-sigma in physics) will cover up such sins.
- *Bayesian approach:* If a multiplicity adjustment is necessary, it is accommodated through prior probabilities associated with the multiplicities. Typically, the more possible hypotheses there are, the lower prior probabilities they each receive.
  - Recall the exclusive hypothesis testing example from talk 2.

Bayesian prior probability assignments do not automatically provide multiplicity control

- Suppose $X_i \sim N(x_i \mid \mu_i, 1)$, $i = 1, \ldots, m$, are observed.
- It is desired to test $H_i^0 : \mu_i = 0$ versus $H_i^1 : \mu_i \neq 0$, $i = 1, \ldots, m$, but any test could be true or false regardless of the others.
- The simplest probability assignment is $Pr(H_i^0) = Pr(H_i^1) = 0.5$, independently, for all i.
- This does *not* control for multiplicity; indeed, each test is then done completely independently of the others. Thus H_1^0 is accepted or rejected whether $m = 1$ or $m = 1,000,000$.

1. The same holds in many other model selection problems such as variable selection: use of equal probabilities for all models does not induce any multiplicity control.

2. The above is a proper prior probability assignment. Thus, if these are one's real prior probabilities, no multiplicity adjustment is needed.

Inducing multiplicity control in this simultaneous testing situation (Scott and Berger, 2006 JSPI; other, more sophisticated full Bayesian analyses are in Gönen et. al. (03), Do, Müller, and Tang (02), Newton et all. (01), Newton and Kendziorski (03), Müller et al. (03), Guindani, M., Zhang, S. and Mueller, P.M. (2007), . . .; many empirical Bayes such as Efron and Tibshirani (2002), Storey, J.D., Dai, J.Y and Leek, J.T. (2007), Efron (2010))

- Suppose $x_i \sim N(x_i \mid \mu_i, \sigma^2)$, $i = 1, \ldots, m$, are observed, σ^2 known, and test $H_i^0 : \mu_i = 0$ versus $H_i^1 : \mu_i \neq 0$.
- If the hypotheses are viewed as exchangeable, let p denote the common prior probability of H_i^1, and *assume p is unknown* with a uniform prior distribution. *This does provide multiplicity control.*
- Complete the prior specification, e.g.
  - Assume that the nonzero μ_i follow a $N(0, V)$ distribution, with V unknown.
  - Assign V the objective (proper) prior density $\pi(V) = \sigma^2/(\sigma^2 + V)^2$.

- Then the posterior probability that $\mu_i \neq 0$ is

$$p_i = 1 - \frac{\int_0^1 \int_0^1 p \prod_{j \neq i} \left(p + (1-p)\sqrt{1-w}\, e^{w {x_j}^2/(2\sigma^2)} \right) dpdw}{\int_0^1 \int_0^1 \prod_{j=1}^m \left(p + (1-p)\sqrt{1-w}\, e^{w {x_j}^2/(2\sigma^2)} \right) dpdw}.$$

- $(p_1, p_2, \ldots, p_m)$ can be computed numerically; for large m, it is most efficient to use importance sampling, with a common importance sample for all p_i.

Example: Consider the following ten 'signal' observations:
-8.48, -5.43, -4.81, -2.64, -2.40, 3.32, 4.07, 4.81, 5.81, 6.24

- Generate $n = 10, 50, 500,$ and 5000 $N(0,1)$ noise observations.
- Mix them together and try to identify the signals.

n	The ten 'signal' observations										#noise
	-8.5	-5.4	-4.8	-2.6	-2.4	3.3	4.1	4.8	5.8	6.2	$p_i > .6$
10	1	1	1	.94	.89	.99	1	1	1	1	1
50	1	1	1	.71	.59	.94	1	1	1	1	0
500	1	1	1	.26	.17	.67	.96	1	1	1	2
5000	1	1.0	.98	.03	.02	.16	.67	.98	1	1	1

Table 1: The posterior probabilities of being nonzero for the ten 'signal' means.

Note 1: The penalty for multiple comparisons is automatic.

Note 2: **Theorem:** $E[\#i : p_i > .6 \mid \text{all } \mu_j = 0] \to 0$ as $m \to \infty$, so the Bayesian procedure exerts very strong control over false positives. (In comparison, $E[\#i :$ Bonferroni rejects $\mid$ all $\mu_j = 0] = \alpha$.)

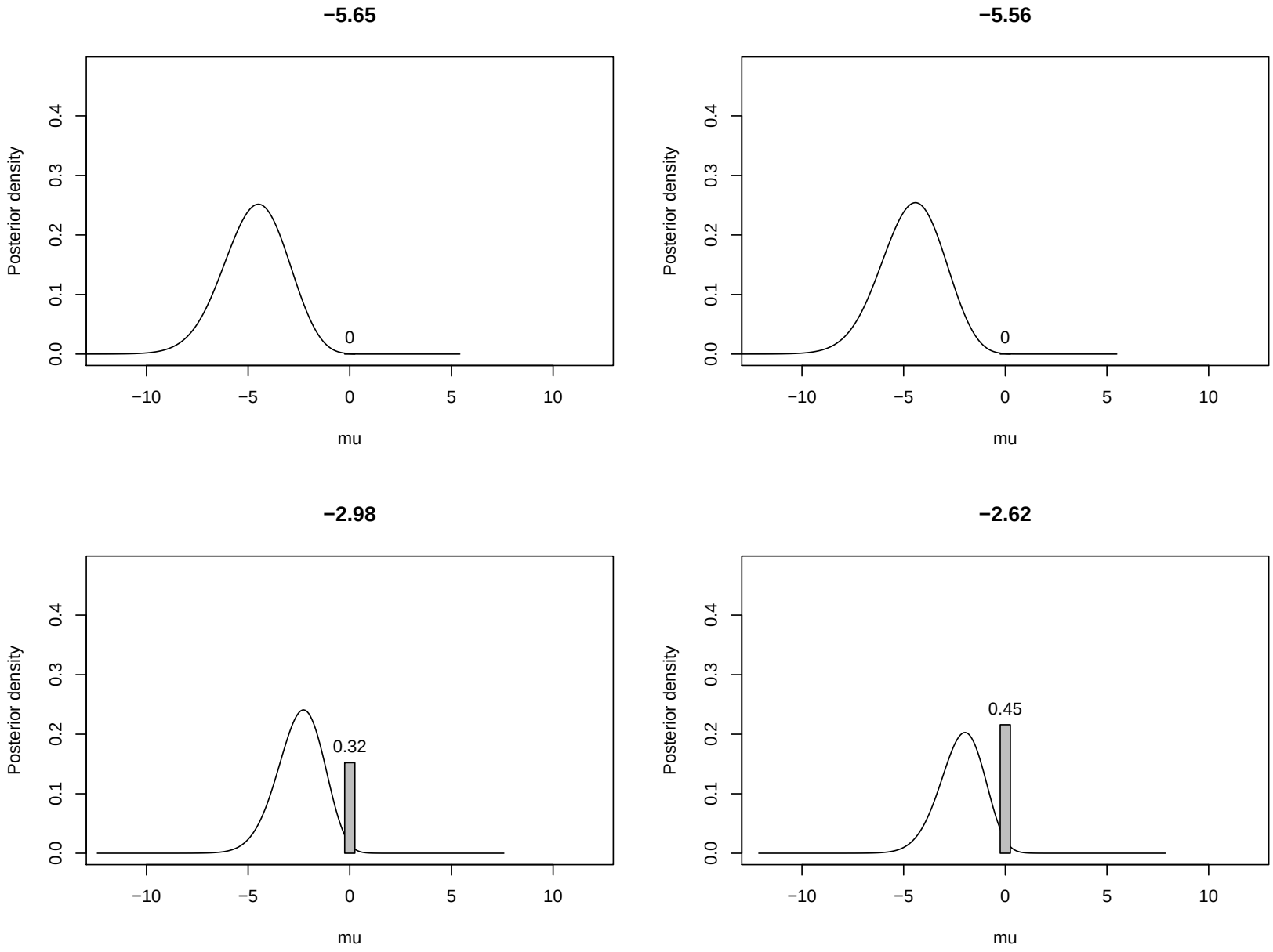

Figure 1: For four of the observations, $1 - p_i = \Pr(\mu_i = 0 \,|\, \boldsymbol{y})$ (the vertical bar), and the posterior densities for $\mu_i \neq 0$.

Interim Summary

- Bayesian multiplicity control is implemented through the assignment of prior probabilities of models. Because it enters through the prior probabilities there is never a loss in power when dealing with dependent testing situations, a problem with many frequentist methods of control.
- Bayesian probability assignments can fail to provide multiplicity control. In particular, assigning all models equal prior probability fails in many situations (testing of exclusive hypotheses being an exception).
- A key technique for Bayesian multiplicity control is to think hierarchically by specifying *unknown* inclusion probabilities for hypotheses or variables, and assigning them a prior distribution.
  - Assigning probability 1/2 to 'no signal' also provides null control, with little cost in power.
  - Any *pre-experimental* prior probability assignment is allowed.

II. Types of Multiplicities

Class 1: Multiplicities Not Affecting the Likelihood

- Consideration of multiple test statistics
  - *Example:* Doing a test of fit, and trying both a Kolmogorov-Smirnov test and a Chi-squared test.
  - Frequentists should either report all tests, or adjust; e.g., if p_i is the p-value of test i, base testing on the *statistic* $p_{min} = \min p_i$.
- Consideration of multiple priors: a Bayesian must either
  - have principled reasons for settling on a particular prior, or
  - implement a hierarchical or robustness analysis over the priors.
- Interim analysis (also called optional stopping and sequential analysis)
  - Bayesians do not adjust, as the posterior is unaffected.
  - Frequentists should adjust: 'spending α' for interim looks at the data with analysis.

Class 2: Multiplicities Affecting the Likelihood

- Choice of transformation/model
- Multiple testing (mentioned)
- Variable selection (later section)
- Subgroup analysis (later section)

Multiple testing

- Multiple hypothesis testing (earlier Bayesian example)
- Multiple multiple testing

  Example: In a pharmaceutical company, plasma samples are sent to
  - a metabolic lab, where an association is sought with any of 200 metabolites;
  - a proteomic lab, where an association is sought with any of 2000 proteins;
  - a genomics lab, where an association is sought with any of 2,000,000 genes.

  The company should do a joint multiplicity analysis.

  A Bayesian analysis could (very roughly) give each lab 1/3 of the prior probability of a discovery, with each third to be divided within the lab.
- Serial tests

  Example: All 16 large Phase III Alzheimer's trials have failed. If the next one is a success, should we believe it?

III. Variable Selection

Example: a retrospective study of a data-base investigates the relationship between 200 foods and 25 health conditions. It is reported that eating broccoli reduces lung cancer (p-value=0.02).

- Not adjusting for multiplicity (5000 tests) in this type of situation is a leading cause of 'junk science.'
- There are other contributing problems here, such as the use of p-values.

Frequentist solutions:

- Bonferonni could be used: to achieve an overall level of 0.05 with 5000 tests, one would need to use a per-test rejection level of $\alpha = 0.05/5000 = 0.00001$.
  - This is likely much too conservative because of the probably high dependence in the 5000 tests.
- Some type of bootstrap could be used, but this is difficult when faced, as here, with 2^{5000} models.

Bayesian solution:

- Assign prior variable inclusion probabilities.
- Implement Bayesian model averaging or variable selection.

- Options in choosing prior variable inclusion probabilities:
  - Objective Bayesian choices:
    - *Option 1:* each variable has unknown common probability p_i of having no effect on health condition i.
    - *Option 2:* variable j has common probability p_j of having no effect on each health condition.
    - *Option 3:* some combination.
  - Main effects may have a common unknown prior inclusion probability p_1; second order interactions prior inclusion probability p_2; etc.
  - An oversight committee for a prospective study might judge that at most one effect might be found, and so could prescribe that a protocol be submitted in which
    - prior probability 1/2 be assigned to 'no effect;'
    - the remaining probability of 1/2 could be divided among possible effects as desired pre-experimentally. (Bonferonni adjustments can also be unequally divided pre-experimentally.)

II. Bayesian subgroup analysis

(with Lei Shen and Xiaojing Wang)

Some previous papers on Bayesian subgroup analysis: Berry (1990), Dixon and Simon (1991), Simon (2002), Gopalan and Berry (1998), Jones et. al. (2011), Sivaganesan, Laud, and Müller (2011), Laud, Sivaganesan and Müller (2013)

Our Guiding Principles for Subgroup Analysis

- Null control (allowing for the possibility of 'no effect') and control for multiple testing need to be present.
- To maximize power to detect subgroup effects,
  - the subgroups and allowed population partitions need to be restricted to those that are scientifically plausible (not discussed today);
  - allowance for 'scientifically favored' subgroups should be made.
- Full Bayesian analysis is sought.
  - because it naturally allows the favoring of subgroups through choice of prior probabilities;
  - because it is fully powered even in the presence of highly dependent test statistics (as is the situation of subgroup analysis);
  - because it yields an answer to any question asked – e.g., what is the probability that a specific individual will demonstrate a treatment effect?

Factors Defining the Subgroups

Suppose we have m factors $X_1, X_2, \ldots, X_m$;

- in our examples, these will be values of m SNPs for an individual.

We only consider the dichotomous case, where each X_j is either 0 or 1.

Subgroups are defined by specification of some (or all) of the values of the X_i.

Here we only consider the case where subgroups are defined by a single factor, e.g. $S = \{\text{all individuals with } X_9 = 1\}$.

Including baseline (prognostic) effects

The factors may have baseline (prognostic) effects separate from or together with treatment (predictive) effects and these must be modeled to prevent confounding.

Statistical models if only one-factor subgroups are allowed:

Response of an individual is

$$\begin{aligned} Y &= B_k + T_j + error, \\ B_k &= \text{one of several baseline (prognostic) models involving factor } k \\ T_j &= \text{one of several treatment (predictive) models involving factor } j \end{aligned}$$

- Either B_k or T_j or both could be absent.
- The models will have unknown parameters (e.g., treatment effect size).
- There are typically many thousands of possible models.

Specifying Prior Probabilities of Models

Interpretable prior inputs:

- o_i, the *effect odds* of Factor i to Factor 1, defined as the prior relative odds that Factor i has an effect compared to Factor 1. (Default: $o_i = 1$.)
- *Null control:* specify p_0 and q_0, the prior probability that an individual has no treatment (predictive) effect and no baseline (prognostic) effect, respectively. (Default: $p_0 = q_0 = 0.5$.)
- r_i is the ratio of the prior probability of the overall treatment model to the sum of the prior probabilities of the treatment models with i factor splits. (Default: $r_i = 1$.)

These inputs determine the prior probability $\mathrm{P}(M)$ of a model M.

Objective prior distributions are also specified for the parameters of each model (e.g., treatment effect size).

Bayes theorem then yields the posterior probabilities, $\mathrm{P}(M \mid \text{data})$, of each model, given the data, as well as the posterior distributions of parameters.

Posterior Inferences for Subgroup Analyses

For individual i (i.e., a specification of the values of all of the factors), let $\mathcal{M}_i$ be all the models under which there is a predictive effect for that individual,

Individual treatment effect probability (*personalized medicine*) is given, for individual i, by

$$P_i = \sum_{\text{all } M_l \text{ in } \mathcal{M}_i} \mathrm{P}(M_l \mid \text{data}) \,.$$

Individual treatment effect size can similarly be defined as the appropriate posterior average of the estimated effect sizes for each model.

Subgroup treatment effect probability is then given by the average of the P_i for all the individuals in the specified subgroup.

Subgroup treatment effect size is similarly defined.

Illustrations

Two synthetic datasets generated at Eli Lilly and Company:

- created to mimic data from clinical trials concerning the development of new therapies for schizophrenia, with the response variable being the amount of reduction in PANSS Total Score;
- factors are 32 single-nucleotide polymorphisms (SNPs), coded as $X_j = 0$ or $X_j = 1$, and having a two-level correlation structure – this leads to a total of 3236 different possible models;
- 200 subjects are randomized equally to treatment and placebo;
- Two datasets, y_a and y_b, were generated from the predictors as follows:
  - y_a: there was an overall nonzero treatment effect but no predictive biomarker (i.e., no biomarker corresponding to a treatment effect); also Factor 31 was a prognostic biomarker (i.e., corresponds to a baseline effect).
  - y_b: Factor 13 is a predictive biomarker and Factor 7 is a prognostic biomarker.

Table 2: Models with posterior probability > 0.03 for dataset y_a.

(The true model generating the data is the 5^{th} listed below.)

	Predictive Factor j	Treatment Submodel	Prognostic Factor k	Baseline Submodel	Prior Probability	Posterior Probability
1	–	null model	–	null model	0.2	0.041
2	–	full model	–	null model	0.15	0.034
3	31	treatment effect	–	null model	0.00156	0.046
4	–	null model	31	baseline effect	0.00625	0.097
5	–	full model	31	baseline effect	0.00469	0.214
6	16	treatment effect	31	baseline effect	0.00005	0.107
7	31	treatment effect	15	baseline effect	0.00005	0.062

Note that the posterior to prior odds *(of interest in flagging models for possible further study)* for the (correct) 5th model are $0.214/0.00469 = 45.6$.

Note that the posterior to prior odds for the (incorrect) 6th model are $0.10689/.00005 = 2138$ (equivalent to a p-value of about 0.00001).

Table 3: Models with posterior probability > 0.03 for dataset y_b.

(The true model generating the data is the 5^{th} listed below.)

	Predictive Factor j	Treatment Submodel	Prognostic Factor k	Baseline Submodel	Prior Probability	Posterior Probability
1	–	null model	–	null model	0.2	0.173
2	–	full model	–	null model	0.15	0.176
3	13	treatment effect	–	null model	0.00156	0.152
4	14	treatment effect	–	null model	0.00156	0.055
5	13	treatment effect	7	baseline effect	0.00005	0.030

Posterior to prior odds:

- line 5: 593.6
- line 3: 97.6
- line 2: 1.17

Table 4: For data set y_b (for which Factor 13 is the correct predictive factor), the effect of first changing the effect odds of Factors 13 and 14 to $o_j = 5$. (True model is the 5^{th}.)

	Predictive Factor j	Treatment Submodel	Prognostic Factor k	Baseline Submodel	Prior Probability	Posterior Probability
1	–	null model	–	null model	0.2	0.173
2	–	full model	–	null model	0.15	0.176
3	13	treatment effect	–	null model	0.00156	0.152
4	14	treatment effect	–	null model	0.00156	0.055
5	13	treatment effect	7	baseline effect	0.00005	0.030
1	–	null model	–	null model	0.2	0.067
2	–	full model	–	null model	0.15	0.068
3	13	treatment effect	–	null model	0.00852	0.319
4	14	treatment effect	–	null model	0.00852	0.115
5	13	treatment effect	7	baseline effect	0.00029	0.068

Table 5: For each of the data sets and first 13 individuals, P_i is the posterior probability of a nonzero treatment effect and Λ_i is the posterior expected effect.

	y_a		y_b		y_c		y_d	
	P_i	Λ_i	P_i	Λ_i	P_i	Λ_i	P_i	Λ_i
1	0.612	8.550	0.699	9.465	0.996	15.343	0.613	8.972
2	0.737	9.521	0.706	9.483	0.223	-2.760	0.389	5.573
3	0.425	5.834	0.394	5.514	0.993	15.346	0.375	5.355
4	0.438	5.938	0.729	9.533	0.992	15.344	0.411	5.830
5	0.599	8.448	0.412	5.715	0.997	15.339	0.394	5.685
6	0.579	7.971	0.395	5.514	0.220	-3.041	0.598	8.861
7	0.619	8.549	0.405	5.613	0.223	-2.769	0.367	5.253
8	0.714	9.493	0.695	9.465	0.220	-3.022	0.604	8.925
9	0.570	7.906	0.400	5.572	0.994	15.343	0.637	9.039
10	0.738	9.532	0.400	5.564	0.223	-2.765	0.619	8.935
11	0.579	7.962	0.402	5.615	0.220	-3.026	0.618	8.915
12	0.590	8.397	0.411	5.688	0.993	15.343	0.364	5.199
13	0.579	7.975	0.478	6.881	0.996	15.341	0.639	9.034

Table 6: For data y_b (where Factor 13 is the correct predictive factor), posterior probabilities of nonzero treatment effects for subgroups formed by splits of the 32 factors; if a factor is not listed, it's probabilities are similar to factor 1.

	Default Prior		With Prior Info. I		With Prior Info. II	
Predictive Factor j	$X_j = 1$	$X_j = 0$	$X_j = 1$	$X_j = 0$	$X_j = 1$	$X_j = 0$
1	0.561	0.544	0.567	0.545	0.583	0.550
13	0.706	0.412	0.749	0.380	0.880	0.279
14	0.678	0.407	0.712	0.374	0.820	0.271

Summary: Bayesian approach to subgroup analysis

- It only depends on prior model probabilities and is hence fully powered even with highly dependent test statistics.
- It allows pre-experimental favoring of certain factors or submodels.
- It provides fully interpretable probabilistic answers and is both exploratory and confirmatory.
- It provides not only subgroup predictive and prognostic effects but also individual effect assessments for personalized medicine.

Thanks

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, B* **85**, 289–300.

Bickel, D.R. (2003). Error-rate and decision-theoretic methods of multiple testing. Alternatives to controlling conventional false discovery rates, with an application to microarrays. *Tech. Rep*, Office of Biostatistics and Bioinformatics, Medical College of Georgia.

Do, K.A., Müller, P., and Tang, F. (2002). *Tech. Rep*, Department of Bioestatistics, University of Texas.

Dudoit, S., Shaffer, J.P., and Boldrick, J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71–103.

DuMouchel, W.H. (1988). A Bayesian model and graphical elicitation model for multi[le comparison. In *Bayesian Statistics 3* (J.M. Bernardo, M.H. DeGroot, D.V. Lindey and A.F.M. Smith, eds.) 127–146. Oxford University Press.

Duncan, D.B. (1965). A Bayesian approach to multiple comparisons. *Technometrics* **7**, 171-222.

Efron, B. (2010). Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics Monographs.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23** 70–86.

Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001). Empirical Bayes analysis of

a microarray experiment. *Journal of the American Statistical Association* **96** 1151–1160

Finner, H., and Roters, M. (2001). On the false discovery rate and expected Type I errors. *Biometrical Journal***43**, 895–1005

Genovese, C.R. and Wasserman, L. (2002a). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society* **64** 499–518.

Genovese, C.R. and Wasserman, L. (2002b). Bayesian and frequentist multiple testing. *Tech. Rep.* Department of Statistics, Carnegie-Mellon University.

Morris, J.S., Baggerly, K.A., and Coombes, K.R. (2003). Bayesian shrinkage estimation of the relative abundance of mRNA transcripts using SAGE. *Biometrics*, **59**, 476–486.

Müller, P., Parmigiani, G., Robert, C., and Rouseau, J. (2002), "Optimal Sample Size for Multiple Testing: the Case of Gene Expression Microarrays," Tech. rep., University of Texas, M.D. Anderson Cancer Center.

Newton, M.A., and Kendziorski, C.M. (2003). Parametric empirical Bayes methods for microarrays. In *The analysis of gene expression data: methods and software*, Springer.

Newton, M.A., Kendziorski, C.M., Richmon, C.S., Blattner, F.R., and Tsui, K.W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8** 37–52.

Scott, G. and Berger, J.O. (2003). An exploration of Bayesian multiple testing. To appear in Volume in honour of Shanti Gupta.

Shaffer, J.P. (1995). Multiple hypothesis testing: a review. *Annual Review of Psychology* **46**, 561–584. Also *Technical Report* # 23, National Institute of Statistical Sciences.

Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754.

Sivaganesan, S., Laud, P.W., Mueller, P. A. (2011). Bayesian subgroup analysis with a zero-enriched Polya urn scheme. Statistics in Medicine, 30(4), 312–323.

Storey J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B* **64** 479–498.

Storey J.D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics*

Storey J.D., Taylor, J.E., and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society B* **66** 187–205.

Storey J.D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *PNAS* **100** 9440–9445.

Waller, R.A. and Duncan, D.B. (1969). A Bayes rule for the symmetric multiple comparison problem. *Journal of the American Statistical Association* **64**, 1484-–1503.