A Two-Level Toeplitz Model for Large-Scale Simultaneous Hypothesis Testing

Dan Cervone
Advisor: Carl Morris

December 10, 2012

Efron's fdr[3]

Suppose we have M test statistics (assumed to be z scores):

- $z_i|\mu_i \overset{ind}{\sim} N(\mu_i, 1)$ for $i = 1, ..., M$.
- $\mu_i \overset{iid}{\sim} p_0\delta_0 + (1 - p_0)g(\mu_i)$
- $z_i \overset{iid}{\sim} f(z_i)$ (marginally)
- Define
    - $\text{fdr}(z) = P(\mu_i = 0|z_i = z) = \frac{p_0\phi(z_i)}{f(z_i)}$
    - $\widehat{\text{fdr}}(z) = \frac{p_0\phi(z_i)}{\hat{f}(z_i)}$

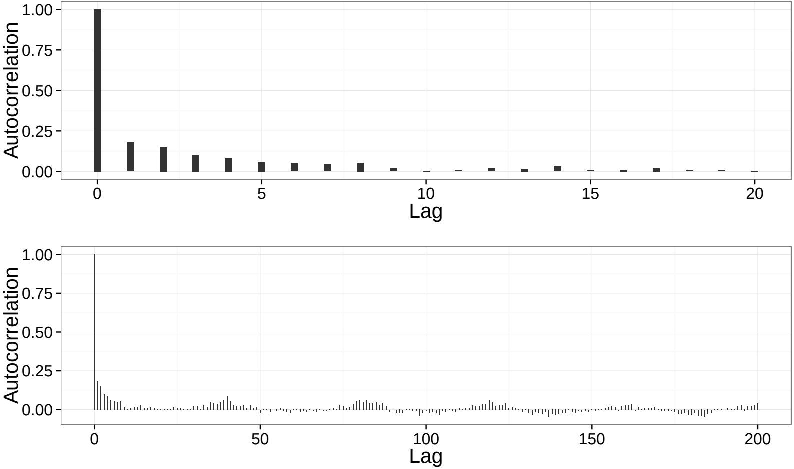
  where $\hat{f}$ is an estimate of the density function f.
- Declare the i^{th} test statistic nonnull if: $\widehat{\text{fdr}}(z_i) \leq q$.

Bayesian posterior probability interpretation relies on independence!

Graphical summary of HIV data z scores

z scores obtained by transforming test statistics for 2-sample t-tests (4 HIV patients, 4 non-HIV patients)[5].

Autocorrelation of HIV z scores



Alternative two-level model

We assume the following two-level model for z scores:

- $z_i | \mu_i \overset{ind}{\sim} N(\mu_i, V = 1)$ for $i = 1, ..., M$.
- $\boldsymbol{\mu} \sim N_M(0, \boldsymbol{\Sigma})$

where $\boldsymbol{\Sigma}$ is assumed to be of symmetric Toeplitz form:

$$\begin{pmatrix} \sigma_0 & \sigma_1 & \sigma_2 & \cdots & \cdots & \sigma_{M-1} \\ \sigma_1 & \sigma_0 & \sigma_1 & \ddots & & \vdots \\ \sigma_2 & \sigma_1 & \sigma_0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \sigma_2 \\ \vdots & & \ddots & \ddots & \ddots & \sigma_1 \\ \sigma_{M-1} & \cdots & \cdots & \sigma_2 & \sigma_1 & \sigma_0 \end{pmatrix}$$

Inference

Inferential model:

- $\boldsymbol{\mu}|\mathbf{z}, \boldsymbol{\Sigma} \sim N_M(\mathbf{B}\mathbf{z}/V, \mathbf{B})$
- $\mathbf{B} = (\boldsymbol{\Sigma}^{-1} + V^{-1}\mathbf{I}_M)^{-1}$
- $\mathbf{z} \sim N_M(0, \boldsymbol{\Sigma} + V\mathbf{I}_M)$

Empirical Bayes approach:
can estimate $\boldsymbol{\Sigma}$ from marginal likelihood and plug into the posterior.

Inference

Inferential model:

- $\boldsymbol{\mu}|\mathbf{z}, \boldsymbol{\Sigma} \sim N_M(\mathbf{Bz}/V, \mathbf{B})$
- $\mathbf{B} = (\boldsymbol{\Sigma}^{-1} + V^{-1}\mathbf{I}_M)^{-1}$
- $\mathbf{z} \sim N_M(0, \boldsymbol{\Sigma} + V\mathbf{I}_M)$

Empirical Bayes approach:
can estimate $\boldsymbol{\Sigma}$ from marginal likelihood and plug into the posterior.

Compared to existing approaches handling dependent test statistics, ours has the following benefits:

- Decision rule is not monotonic in the size of test statistic.
- Generic covariance structure, but comes at the assumption of normality.

Data augmentation

Consider the following data augmentation[2]:

- Let $\mathbf{y}^T = (\mathbf{z}^T \ \ \mathbf{z}_{mis}^T)$, where $\mathbf{z}_{mis}$ is a $(M-1) \times 1$ vector of missing observations.
- Assume $\mathbf{y} \sim N_L(0, \mathbf{\Sigma}_C + V\mathbf{I}_L)$ with $\mathbf{\Sigma}_C$ (symmetric) circulant and $L = 2M - 1$.

Example: Assume $M = 4$, so $L = 7$. $\mathbf{\Sigma}_C$ has the the form:

$$\begin{pmatrix} \sigma_0 & \sigma_1 & \sigma_2 & \sigma_3 & \sigma_3 & \sigma_2 & \sigma_1 \\ \sigma_1 & \sigma_0 & \sigma_1 & \sigma_2 & \sigma_3 & \sigma_3 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma_0 & \sigma_1 & \sigma_2 & \sigma_3 & \sigma_3 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma_0 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_3 & \sigma_3 & \sigma_2 & \sigma_1 & \sigma_0 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_3 & \sigma_3 & \sigma_2 & \sigma_1 & \sigma_0 & \sigma_1 \\ \sigma_1 & \sigma_2 & \sigma_3 & \sigma_3 & \sigma_2 & \sigma_1 & \sigma_0 \end{pmatrix}$$

Data augmentation

Consider the following data augmentation[2]:

- Let $\mathbf{y}^T = (\mathbf{z}^T \ \ \mathbf{z}_{mis}^T)$, where $\mathbf{z}_{mis}$ is a $(M-1) \times 1$ vector of missing observations.
- Assume $\mathbf{y} \sim N_L(0, \mathbf{\Sigma}_C + V\mathbf{I}_L)$ with $\mathbf{\Sigma}_C$ (symmetric) circulant and $L = 2M - 1$.

Example: Assume $M = 4$, so $L = 7$. $\mathbf{\Sigma}_C$ has the the form:

$$\begin{pmatrix} \sigma_0 & \sigma_1 & \sigma_2 & \sigma_3 & \sigma_3 & \sigma_2 & \sigma_1 \\ \sigma_1 & \sigma_0 & \sigma_1 & \sigma_2 & \sigma_3 & \sigma_3 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma_0 & \sigma_1 & \sigma_2 & \sigma_3 & \sigma_3 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma_0 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_3 & \sigma_3 & \sigma_2 & \sigma_1 & \sigma_0 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_3 & \sigma_3 & \sigma_2 & \sigma_1 & \sigma_0 & \sigma_1 \\ \sigma_1 & \sigma_2 & \sigma_3 & \sigma_3 & \sigma_2 & \sigma_1 & \sigma_0 \end{pmatrix}$$

Upper left $M \times M$ block is (unconstrained) symmetric Toeplitz.

EM algorithm

Using the augmented (complete) data $\mathbf{y}$, we use the EM algorithm to estimate $\hat{\mathbf{\Sigma}}_C$.

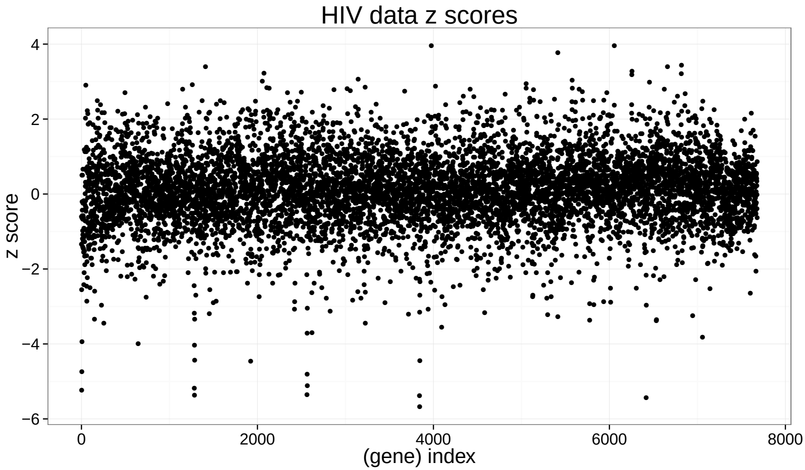
- E-step: $Q(\mathbf{\Sigma}_C|\mathbf{z}, \hat{\mathbf{\Sigma}}_C^{(k)}) = -\log(|\mathbf{\Sigma}_C|) - \mathrm{Tr}(\mathbf{\Sigma}_C^{-1} S^{(k)})$
  - $S^{(k)}$ derived from $\mathbf{z}, \hat{\mathbf{\Sigma}}_C^{(k)}$ using MVN properties.
- M-step: $\hat{\mathbf{\Sigma}}_C^{(k+1)} = \mathrm{argmax} Q(\mathbf{\Sigma}_C|\mathbf{z}, \hat{\mathbf{\Sigma}}_C^{(k)})$
  - Has closed-form solution, since all $\mathbf{\Sigma}_C$ have constant, known eigenvectors (entries consist of powers of complex roots of unity).
- Some (minor) technical considerations needed to ensure convergence of $\hat{\mathbf{\Sigma}}^{(k)}$ to local maximum.[4]

Large sample properties

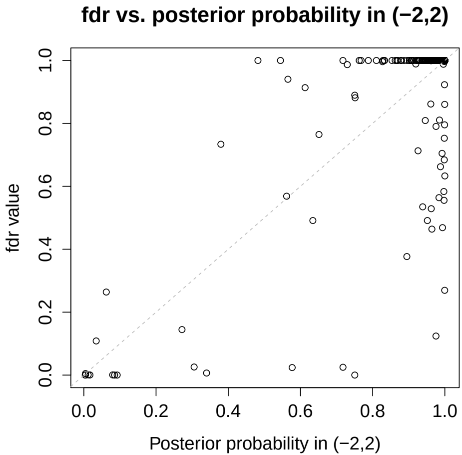
The MLE $\hat{\mathbf{\Sigma}}$ is not consistent in the usual sense as the number of parameters is the same as the number of observations (M). However,

- Information for all unknown parameters increases with each new observation.
- Geometric constraints of Toeplitz form, positive definiteness.
- Simulation results show good approximation of EB posterior to oracle posterior.
- If the autocovariances form a convergent sum, their estimates have variance $\mathcal{O}(1/L)$.
- Related results from the literature involving conditions of sparsity, or smoothness of spectral density.[1][6]

HIV data

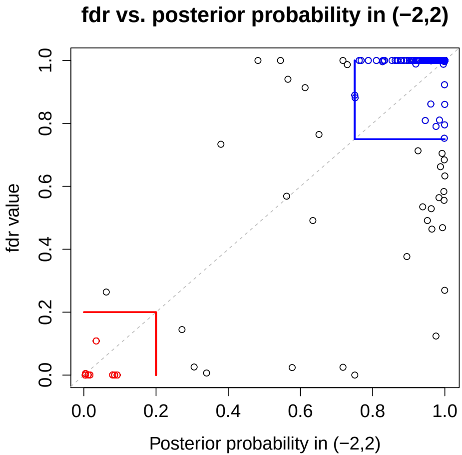

HIV data z scores
z score
(gene) index
4
2
0
−2
−4
−6
0
2000
4000
6000
8000

HIV data



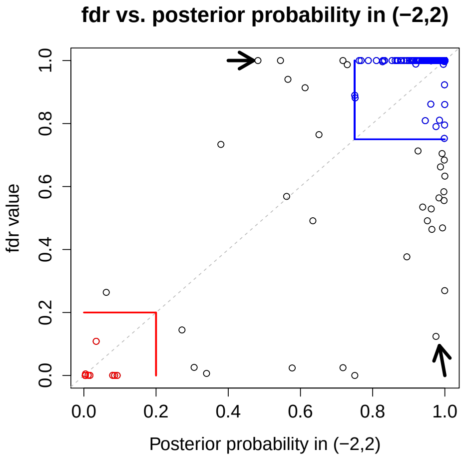
fdr vs. posterior probability in (−2,2)
fdr value
Posterior probability in (−2,2)
0.0
0.2
0.4
0.6
0.8
1.0

HIV data



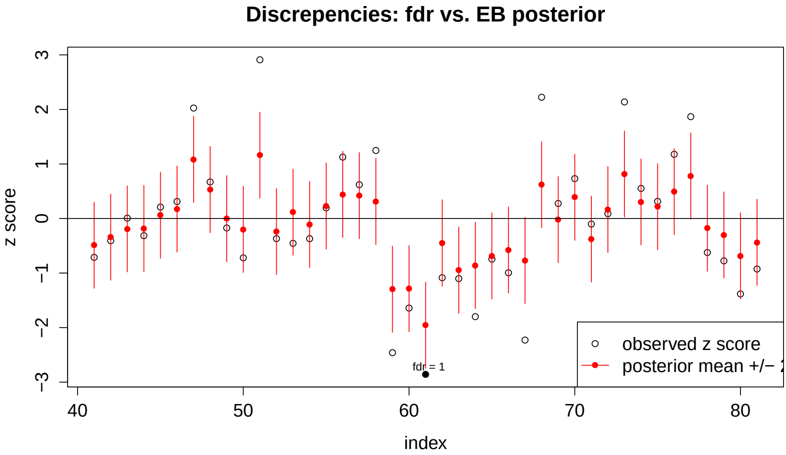
9 cases identified as non-null by both; 99.4% cases identified as null by both

HIV data



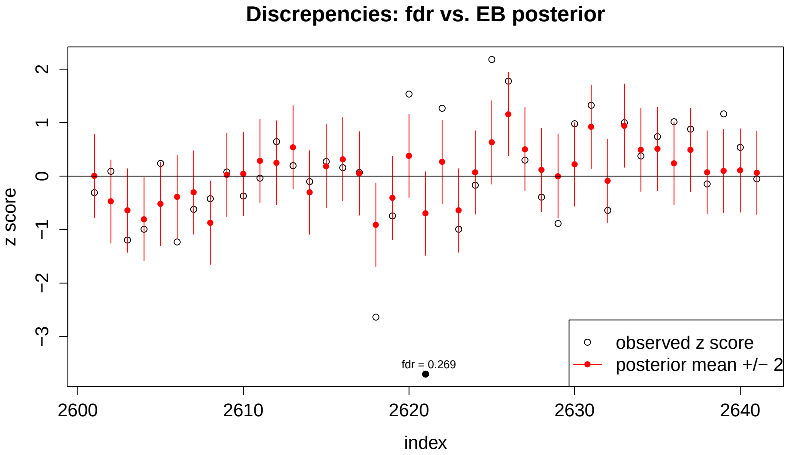
9 cases identified as non-null by both; 99.4% cases identified as null by both

HIV data



Discrepencies: fdr vs. EB posterior
z score
3
2
1
0
−1
−2
−3
40
50
60
70
80
index
fdr = 1
observed z score
posterior mean +/− 2

HIV data



Discrepencies: fdr vs. EB posterior
z score
index
2600
2610
2620
2630
2640
2
1
0
−1
−2
−3
fdr = 0.269
observed z score
posterior mean +/− 2

Areas of further work

- Theoretical conditions for consistency of Toeplitz MLE.
- Full Bayes: prior on matrix parameters.
- What can this model tell us if the Toeplitz covariance is correctly specified but without normality?
- Scalability: $\mathcal{O}(L^3)$ implementation can be dramatically improved theoretically.

- R. Dahlhaus.
  Efficient parameter estimation for self-similar processes.
  The Annals of Statistics, pages 1749–1766, 1989.

- A. Dembo, C.L. Mallows, and L.A. Shepp.
  Embedding nonnegative definite toeplitz matrices in nonnegative definite circulant matrices, with application to covariance estimation.
  Information Theory, IEEE Transactions on, 35(6):1206–1212, 1989.

- B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher.
  Empirical bayes analysis of a microarray experiment.
  Journal of the American Statistical Association, 96(456):1151–1160, 2001.

D.R. Fuhrmann and M.I. Miller.
On the existence of positive-definite maximum-likelihood estimates of structured covariance matrices.
Information Theory, IEEE Transactions on, 34(4):722–729, 1988.

A.B. Van't Wout, G.K. Lehrman, S.A. Mikheeva, G.C. O'Keeffe, M.G. Katze, R.E. Bumgarner, G.K. Geiss, and J.I. Mullins.
Cellular gene expression upon human immunodeficiency virus type 1 infection of cd4+-t-cell lines.
Journal of Virology, 77(2):1392–1402, 2003.

H. Xiao and W.B. Wu.
Covariance matrix estimation for stationary time series.
The Annals of Statistics, 40(1):466–493, 2012.