Bayesian Nonparametric Variable Selection

Babak Shahbaba and Wesley Johnson

July 30, 2012
The Joint Statistical Meetings

Introduction

- Large-scale genomic studies examine thousands of genes simultaneously
- Objective is to identify a small number of genes for *follow-up studies*
- Our method uses a nonparametric Bayesian approach based on Dirichlet process mixture models
- We divide the set of genes into several subgroups according to their degrees of "relevance," or potential effect, in relation to the outcome of interest (e.g., disease status)
- This could lead to a better identification of the underlying structure in our data and ultimately, genes that "matter"

Data

	Case				Control			
Subjects	1	2	3	...	1	2	3	...
Gene 1	-1.2	-1.1	0.1	...	2.2	0.7	1.8	...
Gene 2	-0.7	1.7	1.5	...	0.4	-2.1	1.5	...
Gene 3	3.2	-0.7	-2.5	...	2.2	1.9	-2.0	...
Gene 4	0.2	3.1	0.6	...	-3.0	-0.3	-1.3	...
⋮								

Background

- Most current methods applied to high-throughput experiments are extensions of the classical hypothesis testing approach (i.e., when there is a single hypothesis).

- For each gene $\mathcal{G}_i$, where $i = 1, \ldots, N$, there is a corresponding [null] hypothesis, H_i, stating that there is no change in gene expression between two biological conditions (i.e., diseased vs. healthy).

- The observed expression values $\{Y_{ijk} : j = 1, ..., n_{ik}, k = 0, 1\}$ are used to compute a simple test statistic T_i for gene i

- Statistics above a certain cutoff are deemed significant, after adjustment to control the family-wise Type I error rate or false discovery rate (FDR); also called the false positive rate in diagnostic testing literature

FDR

- FDR is one of the most widely used measures for coping with multiplicity
- Suppose we observe values for $T_1, T_2, \ldots, T_N$ and obtain the corresponding p-values:

$$p_j = P(T_j \geq t_j | H_j)$$

- Reject H_j if $\quad p_j > \lambda$

$$FDR(\lambda) \quad = \quad E(\text{Proportion of true } H_j \mid \text{ rejected })$$

FDR

- Instead of p-values, it is convenient to work with

$$z_j = \Phi^{-1}[P(T_j \geq t_j | H_j)]$$

- Under H_j, $\qquad z_j \sim N(0,1)$

- Large-scale testing situations however permit estimation of the null distribution

- The following mixture density is assumed for the transformed p-values:

$$f(z) \quad = \quad p_0 f_0(z) + (1 - p_0) f_1(z)$$

  where p_0 is the proportion of true null hypotheses, and f_0 and f_1 are the distributions of z values when null hypotheses are true and false respectively

FDR

- Under this model, if all inputs were known, then the Bayesian approach based on zero/one loss for just a single hypothesis rejects H_j if

$$\mathrm{fdr}(z_j) \equiv p_0 f_0(z_j)/f(z_j) < \lambda$$

- Efron et. al. (2001) use empirical Bayes approach to estimate fdr(z)
- Their approach is referred to as *locFDR*

Optimal discovery procedure

- More recently, Storey (2007) has proposed a related method called the *optimal discovery procedure* (ODP), which is approximately equivalent to minimizing *missed discovery rate* (false negative) for each fixed FDR (false positive rate)
- Suppose $z_j \sim f(z_j; \mu_j)$, where f is some distribution indexed by an unknown parameter μ_j.
- The ODP for testing $H_j : \mu_j \in A$ is then based on a single significance thresholding statistic,

$$S_{ODP}(z_j) \quad = \quad \frac{\sum_{\mu_j \notin A} f(z_j; \mu_j)}{\sum_{\mu_j \in A} f(z_j; \mu_j)}$$

- We reject the null hypothesis H_j if $S_{ODP}(z_j) \geq \lambda$ for some $0 \leq \lambda < \infty$.

Bayesian discovery procedure

- Guindani et. al.(2009) showed that the ODP could be interpreted as approximate Bayes rule under a semiparametric model.

- They proposed a Bayesian discovery procedure (BDP) that improves the approximation and allows for multiple shrinkage in clusters implied by a Dirichlet process mixture model.

Bayesian discovery procedure

- Their model has the following form:

$$\begin{aligned} z_i|\mu_i &\sim f(z_i \mid \mu_i), \qquad i = 1, \ldots, N \\ \mu_i|G &\sim G \\ G &\sim \mathcal{D}(G_0, \gamma) \\ G_0 &= p_0 h_{\{0\}}(.) + (1 - p_0) h_{\{0\}^c}(.) \end{aligned}$$

- Here, $f(z_i \mid \mu_i)$ is typically considered to be a normal distribution, $N(z_i \mid \mu_i, \sigma^2)$.

- G has a Dirichlet process prior with a baseline distribution G_0 that itself is a mixture of two terms.

- The distribution $h_{\{0\}}$ is point mass at zero

Bayesian discovery procedure

- Also, $h_{\{0\}^c}$ is set to a continuous distribution such as $N(0, \sigma^2)$.

- Latent cluster membership indicators, s_i, partition the observations into clusters such that

$$s_i = s_k \qquad \text{if } \mu_i = \mu_k$$

- The label $s_i = 1$ is reserved for the null distribution; that is, $s_i = 1$ when $\mu_i = 0$.

- Guindani et al. (2009) showed that thresholding based on the measure

$$v_i \;\; = \;\; 1 - \sum_{b=1}^{B} I(s_i^{(b)} = 1)/B$$

can be approximated by $\hat{S}_{ODP}$

Nonparametric relevance determination

- Our model is similar

$$\begin{aligned} z_j \mid \tau_j^2 &\sim N(0, \tau_j^2) \\ \tau_j^2 \mid G &\sim G \\ G &\sim \mathcal{D}(G_0, \gamma) \end{aligned}$$

- We refer to our final model as *BRD*.

Nonparametric relevance determination based on the observed data

- Let y_{ijk} denote the j^{th} observed gene expression value in group k for gene i

$$y_{ijk} \mid \alpha_i, \beta_i \quad \overset{ind}{\sim} \quad N(\alpha_i + \beta_i x_{ijk}, \sigma_i^2)$$

- Our model for the regression coefficients is hierarchical where the first level assigns independent normal priors to the β_is with distinct variances, namely

$$\beta_i \mid \tau_i^2 \quad \overset{ind}{\sim} \quad N(0, \tau_i^2)$$

- We assume a Dirichlet Process prior for τ_i^2:

$$\tau_i^2 \mid G \overset{ind}{\sim} G \qquad G \sim \mathcal{D}(G_0, \gamma)$$

BRD

- We use Markov chain Monte Carlo (MCMC) methods to simulate samples from posterior distributions

- The number of clusters and the rank of each gene, R_i, change from one iteration to another

- We obtain the posterior mean $\bar{R}_i$ and mode $\hat{R}_i$ of these ranks and use them as measures of relevance

BRD

- Or we could define a relevance measure similar to that of Guindani et. al.(2009)

- To this end, we denote $\min_j\{\tau_j^2\}$ at each iteration as ϕ_0^2

- For gene i, we create a binary indicator, s_i, which is set to 1 when $\tau_i^2 = \phi_0^2$, and zero otherwise

- Similar to the measure proposed by Guindani et. al.(2009), we can use B posterior Monte Carlo samples to calculate

$$v_i = 1 - \sum_{b=1}^{B} I(s_i^{(b)} = 1)/B$$

Comparing BRD to BDP

- Both methods use a Dirichlet process mixture of normals for modeling gene expression data

- For BDP, the DP prior is assumed for the means of the normal distributions (all mixture components share the same variance)

- An alternative variation of BDP mixes on the means and the variances

- We use the DP prior on the variances, τ^2, and fix means at zero

- Our model provides a natural framework for ranking mixture components, and in turn, for ranking the genes assigned to each component with respect to their potential importance

Comparing BRD to BDP

- Our model allows for dividing genes into several groups with different degrees of relevance

- We could select one or more top ranking groups as potentially important

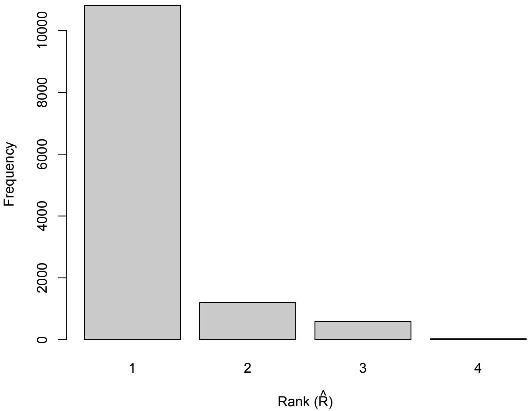
- By focusing on variances, we can detect both location shift and scale change in the distribution of z

- Our method allows for implicit specification of the distribution for genes that are irrelevant; the group with the lowest degree of relevance is regarded as irrelevant

Results for HCMV

- Our first example involves identifying differentially expressed genes due to infection by human cytomegalovirus (Chan et al., 2008)
- Out of 12,626 genes, they identified 1,204 genes as statistically and biologically significant.
- By using local false discovery rate and setting the cutoff q-value to 0.05, the number of selected genes is reduced to 361.
- Using our model, the most relevant group, $\hat{R} = 4$, includes 27 genes.

Results for HCMV



Results for HCMV

- We used the Functional Annotation tool in DAVID Bioinformatics Resources to learn more about the selected genes.

- We found that 7 of these genes are involved in rheumatoid arthritis, 7 are involved in colorectal cancer, 6 of them in breast cancer, and 5 of them are involved in inflammatory bowel disease.

- HCMV is in fact known to be related to inflammatory diseases (e.g., rheumatoid arthritis) and several cancers. Naucler (2008) reviews the evidences for involvement of HCMV microinfections in inflammatory diseases and cancer.

Results for leukemia

- We applied BRD to real data based a study aimed at identifying differentially expressed genes between two types of leukemia (Armstrong et al., 2002).

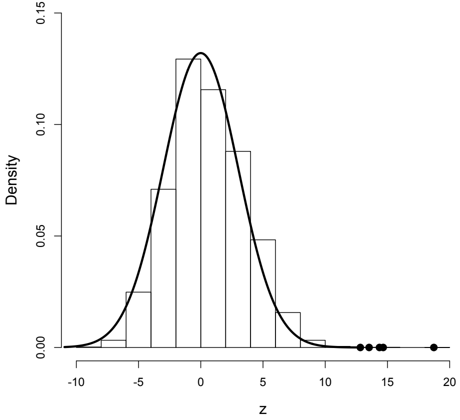
- There are five genes whose posterior mode of rank is $\hat{R} = 2$.

- These genes (in the descending order of $\bar{R}$) are TCL1A, DNTT, CD24, TOP2B, and PSMA6.

- The values of v for these genes are 0.011, 0.048, 0.06, 0.107, and 0.19 respectively.

- The identified genes all are known to be associated with leukemia.

Results for leukemia



Simulation

AUC%	locFDR	BODP	BDP	BRD	
				v	$\bar{R}$
Simulation 1	75.8 (0.7)	-	75.8 (0.4)	76.8 (0.4)	77.0 (0.4)
Simulation 2	90.9 (1.5)	-	88.9 (0.4)	93.5 (0.3)	94.8 (0.2)
Simulation 3	64.5 (2.3)	-	82.2 (1.3)	90.3 (0.7)	94.4 (0.4)
Simulation 4	57.9 (1.8)	-	88.5 (1.1)	86.8 (0.8)	91.8 (0.6)
Simulation 5	73.7 (1.1)	76.2 (1.2)	65.1 (1.0)	79.6 (1.0)	77.7 (1.0)
Simulation 6	75.5 (1.1)	72.8 (1.2)	63.9 (1.0)	77.6 (1.1)	78.3 (1.1)

Conclusion

- We have proposed a new approach for analyzing large-scale studies, where the objective is to identify factors that are relevant to an outcome of interest.

- Our method uses the Dirichlet process to introduce a random grouping on factors (e.g., genes).

- Future directions could involve incorporating additional knowledge about the underlying structure of data.

- Another possible research direction involves extending our model to allow for incorporating more information on subjects.

- Finally, there is much scope for theoretical investigation of the DPM based approaches.

Acknowledgments

- Peter Müller (UT Austin)
- Michele Guindani (MD Anderson)
- Catherine Shachaf (Stanford University)
- ICTS - UCI