A Modified Mixture Model Approach to the Large Scale Multiple Testing Problem

Paramita Chakraborty, John Grego, James Lynch and Chong Ma

University of South Carolina, Columbia

Latent Variables Conference 2016

October 14, 2016

Multiple testing and reproducibility problem

- In modern big data situations, such as in microarray analysis, one source of lack of reproducibility is the voluminous number of false positives/false discoveries that occur.
- This leads to later experiments that do not confirm the earlier findings, resulting in much skepticism towards some of the big data analytic being used.
- We propose an approach that can be used in a study with a large number of observations but where only a small number of real "discoveries" or "significant cases" are expected to be found.
- This method is based on a simple two point mixture contamination model where one component corresponds to the baseline/background information and the second to the sources which are the real discoveries (the contamination).

The Mixture Model

- Start with a basic model for the density of the population under study, viz.

$$f(x) = p_0 f_0(x) + p_1 f_1(x)$$

  with $p_0 + p_1 = 1$. Where f_0 corresponds to the background density and f_1 is the contamination density or the density of the signal that one wants to find.
- Consider the mixture contamination model with the general testing problem $H_0 : p_1 = p_1^*$ *vs.* $H_1 : p_1 > p_1^*$.
- In many problems, the weight p_1^* is very small ($p_1^* \approx 0$) and the problem may be euphemistically referred to as a needle in a haystack (NIHP), while, if $p_1^* > 0$, we refer to it as looking for needles (LFN). For the present discussion we will use the LFN model to develop a method to tackle the large scale multiple testing problem

Mixture model in multiple testing

- In multiple testing scenario the entire group of observation can be viewed as a mixture of null (baseline) and non-null (contamination) distributions.
- Vera et al. [8] advocates using the *LMP* test for

  $$H_0 : p_1 = p_1^* \ vs.\ H_1 : p_1 > p_1^*$$

  in order to screen for contamination (extreme values).
- Efron [2] uses Bayesian *two-class* (null and non-null) model with:
  Prior probabilities:

  $$p_0 = P(\text{null}) \text{ and } p_1 = P(\text{non-null})$$

  Population densities:

  $$f_0 = \text{null density and } f_1 = \text{non-null density.}$$

  Then used posterior probabilities to identify extreme cases.

Bayes Approach

- For the identification of significant cases one can use the assignment functions:

$$A_0(x) = \frac{\hat{p}_0 f_0(x)}{\hat{f}(x)}$$

$$A_1(x) = \frac{\hat{p}_1 f_1(x)}{\hat{f}(x)}$$

  where $\hat{p}_0$ and $\hat{p}_1$ are estimates of p_0 and p_1 respectively, with $\hat{f}(x) = \hat{p}_0 f_0(x) + \hat{p}_1 f_1(x)$.

- For a two class Baysian model assignment function $A_0(x)$ gives the *local false discovery rate* (*fdr*) (Efron [2, 3, 4]).

$$fdr(x) \equiv P(\text{null}|x) = \frac{p_0 f_0(x)}{f(x)}$$

Tail-area FDR

- Recently, the tail area “False Discovery Rate” (FDR) has been promoted as an useful tool for multiple testing problems.
- For example for a left-tail region with observed value x,

$$FDR(x) = P(null|X \leq x) = \frac{p_0 F_0(x)}{F(x)}.$$

- In general for a tail region is given by set B, the tail area FDR is given by

$$FDR(B) \equiv P(null|X \in B) = \frac{p_0 \int_B dF_0}{\int_B dF}$$

- If F is estimated by the empirical cdf then it can be shown that controlling tail-area FDR is equivalent to Benjamini-Hochberg procedure of controlling over-all false discovery rate in the entire study.

Relation between FDR and fdr

- It can be shown that the relationship between the tail area false discovery rate $FDR(A)$, for a set A, to local $fdr(x)$ is

$$FDR(A) = E(fdr(X)|X \in A)$$

- For a Baysian two-class mixture model, cases with *fdr* (or *FDR*) below a pre-determined cutoff point can be considered as true discoveries (Efron).

Our Approach

We propose a modification of Vera et al. and Efron’s approach.

- In most studies of multiple testing situations with a large data, the entire data set is first used to fit a model. Then the same data is used to detect significant cases based on that fitted model.
- We postulate that, using a data for model fitting and then using the very same data for identifying significant cases, may distort the real picture.
- We propose a mixture-model based method using a cross validation type data partitioning at the beginning. Where one part of the data is used for model building and the other part is for anomaly detection using an updated form of *FDR*.
- This new approach not only avoids over-fitting, but also provides some insight into the inter-relation between various significant observations.

Proposed Analysis Method

The analysis is done in the following the stages:

(i) The data is first divided into two (equal) parts, viz. training data and verification data sets. **Using only the training data** we fit a mixture contamination model

$$f(x) = p_0^* f_0(x) + p_1^* f_1(x)$$

that we think captures the baseline and the extreme values best (empirical null).

Then we proceed to identify significant cases based on the **verification data set alone** as follows:

Proposed Analysis Method (contd.)

(ii) We use the LMP test for $H_0 : p_1 = p_1^* \ vs. \ H_1 : p_1 > p_1^*$ as a screening test on the verification data. Given the observed LMP test from the verification data, we then ‘update’ the fitted model obtained from the training data.

(iii) Finally we use the updated model to calculate the FDR associated with each observation in the verification data. The observations, with FDR below a pre-determined cutoff point, are identified as the significant cases.

(iv) The entire process (stages i, ii, iii) is repeated several times with different partitioning of the training and the verification subsets. For each repetition a set of significant cases are identified. The most frequently identified significant cases are considered as potential “true discoveries”.

LMP Test for screening

- The “update” in the stage (ii) of the analysis is based on a conditional asymptotic distribution of the observations; where the condition is given by the observed LMP test statistic.
- We start with setup: let $X_1, \dots, X_n$ be i.i.d. with density $f(x) = p_0 f_0(x) + p_1 f_1(x)$, with $p_1 \in (0, 1)$ and $p_0 + p_1 = 1$. For testing

  $$H_0 : p_1 = p_1^* \ vs \ H_1 : p_1 > p_1^*,$$

  the generalized Neyman-Pearson Lemma shows that the LMP test statistic is

  $$T_n = \sum_{i=1}^{n} \frac{f_1(X_i) - f_0(X_i)}{f_{H_0}(X_i)} \tag{1}$$

  Here f_{H_0} is the common pdf of $X_1, \dots, X_n$ under H_0; i.e. $f_{H_0}(x) = p_0^* f_0(x) + p_1^* f_1(x)$ with $p_0^* + p_1^* = 1$.

Exponential tilting

- Define $Y_i = \frac{f_1(X_i) - f_0(X_i)}{f_{H_0}(X_i)} \equiv h(X_i)$.
- Since $T_n \equiv \sum_{i=1}^n Y_i \equiv \sum_{i=1}^n h(X_i)$ is a sum of i.i.d. mean 0 r.v's under H_0, when $p_1^* > 0$, $T_n/n \overset{a.s.}{\to} 0$. Positive values of T_n/n are the relevant values for rejecting H_0.
- Suppose $S = \{\theta \geq 0 : E\left(e^{\theta Y}\right) < \infty\}$ and S^0 its interior.
- For $\theta \in S^0$, we define families with the following distribution functions:

$$F_\theta(y') = \int_{-\infty}^{y'} e^{\theta y} f_Y(y) dy / E\left(e^{\theta Y}\right) \tag{2}$$

$$G_\theta(x') = \int_{-\infty}^{x'} e^{\theta h(x)} f_{H_0}(x) dx / E\left(e^{\theta h(X)}\right). \tag{3}$$

These are the families generated by exponentially tilting f_Y and f_{H_0}, respectively.

“Update” by exponentially tilting

Define $\bar{m}(\theta) = \dfrac{E\left(Ye^{\theta Y}\right)}{E\left(e^{\theta Y}\right)}$ and the set $M = \{\bar{m}(\theta) : \theta \in S^0\}$.

Proposition

Let $S^0 \neq \emptyset$. Suppose, $I = (c, d)$ with $0 < c < d$ where $c \in M$. If θ_c is chosen such that $c = \bar{m}(\theta_c)$, then, as $n \to \infty$,

$$P\left(Y_i \leq y_i, i = 1, \ldots, m \mid \frac{T_n}{n} \in I\right) \xrightarrow{d} \prod_{i=1}^{m} F_{\theta_c}(y_i) \tag{4}$$

$$P\left(X_i \leq x_i, i = 1, \ldots, m \mid \frac{T_n}{n} \in I\right) \xrightarrow{d} \prod_{i=1}^{m} G_{\theta_c}(x_i) \tag{5}$$

We first fit f_{H_0} (empirical null) using the training data. Then use the verification data to get the LMP test $\hat{T}_n$ and update the empirical null f_{H_0} by exponentially tilting it with tilt parameter $\hat{\theta}$ s.t $\bar{m}(\hat{\theta}) = \hat{T}_n/n$.

Updated FDR

We update the baseline and contamination distribution by exponentially tilting:

$$f^*_{\theta_c,0}(x) = \frac{e^{\theta_c \cdot h(x)} f_0(x)}{E_0\left(e^{\theta_c \cdot h(X)}\right)}, \quad f^*_{\theta_c,1}(x) = \frac{e^{\theta_c \cdot h(x)} f_1(x)}{E_1\left(e^{\theta_c \cdot h(X)}\right)}, \tag{6a}$$

$$p_1^*(\theta_c) = \frac{p_1^* E_1\left(e^{\theta_c \cdot h(X)}\right)}{E_{H_0}\left(e^{\theta_c \cdot h(X)}\right)}, \quad p_0^*(\theta_c) = \frac{p_0^* E_0\left(e^{\theta_c \cdot h(X)}\right)}{E_{H_0}\left(e^{\theta_c \cdot h(X)}\right)}. \tag{6b}$$

Thus, $FDR(A) \mid \frac{T_n}{n} \in (c, d)$ is given by:

Proposition

$$E\left[fdr(X) \mid X \in A, \frac{T_n}{n} \in (c, d)\right] \approx \int_A \frac{p_0^*(\theta_c) f^*_{\theta_c,0}(x) dx}{G_{\theta_c}(A)}$$

We use this updated FDR to identify extreme cases from the verification data set.

An application

- Our proposed analysis method is specially useful for microarray studies where a large number of genes are studied but only a handful of the genes are expected to be significantly differentially expressed and scientists look for a regulator gene associated with a specific disease or an interactome of genes that may control the disease.
- We used our method to a prostate cancer study data where 52 prostate cancer patients and 50 healthy people were subjects and 6033 gene expressions were studied. The main goal is to identify potential regulator genes.
- We considered two sample t-test statistics for each gene and looked at $x_i = P(t < t_i)$ where t_i is the observed t-test statistics from the i^{th} gene. Our main model is $f(x_i) = p_0 f_0(x_i) + p_1 f_1(x_i)$, where f_0 is baseline ($Uniform(0, 1)$ or some version of it) and f_1 is the contamination (*Beta*-distribution or some version of it).

Fitting f_{H_0}

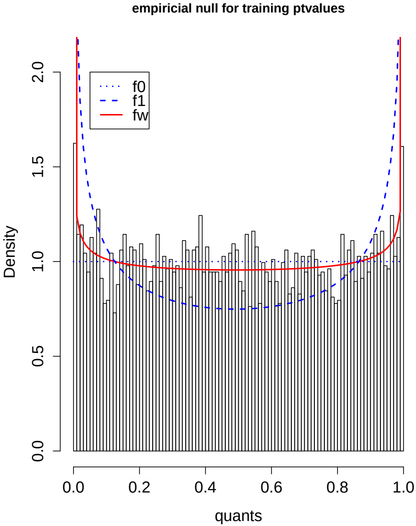

Figure 1: A two-point mixture model fitting the left tail probability x obtained using a training data.

Empirical Null

- First we fitted a mixture of $U(0,1)$ and $Beta(\alpha,\beta)$ distributions $f(x) = \tilde{p}_0\tilde{f}_0(x) + \tilde{p}_1\tilde{f}_1(x)$ to the t-statistics tail area data.
- In order to capture the tail behavior for the contamination distribution we re-wrote the fit as: $\tilde{f}_1(x) = f_{0,1}(x) + f_{1,1}(x)$, where

$$f_{0,1}(x) = \tilde{f}_1(x)I\left[\tilde{f}_1(x) \leq 1\right] + 1 \cdot 1I\left[\tilde{f}_1(x) > 1\right]$$

$$f_{1,1}(x) = 0 \cdot I\left[\tilde{f}_1(x) \leq 1\right] + (1 - \tilde{f}_1(x)) \cdot I\left[\tilde{f}(x)_1 > 1\right]$$

- Let $A_{11} = \int_{\mathbb{R}} f_{1,1}(x)dx$. Define $f_1^*(x) = \frac{1}{A_{11}}f_{1,1}(x)$.
- Define $p_1^* = \tilde{p}_1 \cdot A_{11}$ and $p_0^* = 1 - p_1^*$.
- And $f_0^*(x) = \frac{\tilde{p}_0}{p_0^*}\tilde{f}_0(x) + \frac{\tilde{p}_1}{p_0^*}f_{0,1}(x)$
- Then the fitted model can be written as:

$$f_{H_0}(x) = p_0^* f_0^*(x) + p_1^* f_1^*(x)$$

Here f_0^* can be thought as an “empirical null”.

LMP, update and anomaly detection

- From the verification data set we got $\hat{T}_n = \sum_{i=1}^{n} \frac{f_1^*(x_i) - f_0^*(x_i)}{f_{H_0}(x_i)}$.
- Then chose the tilt parameter $\hat{\theta}$ such that $\bar{m}(\hat{\theta}) = \frac{\hat{T}_n}{n}$.
- We used this $\hat{\theta}$ to exponentially tilt f_0^* and f_1^* and update p_0^* and p_1^*.
- Then calculated updated *FDR* for each gene using the tilted distributions and updated weights.
- Genes with $FDR < 0.1$ were considered extreme or significant or true discoveries.
- The steps were repeated for 100 cross-validation type partitions. 69 partitions showed significant genes (H_0 was rejected). 31 partitions did not produce any significant cases (H_0 accepted). These 69 partitions identified total 73 significant genes.

Significant genes from 100 cross-validation

ID	Gene	freq	med.tailp	avg.tailp	sd.tailp
1	610	14	0.99985	0.99876	0.00393
2	1720	12	0.99964	0.99754	0.00628
3	4331	10	0.00122	0.00680	0.01322
4	914	10	0.99929	0.99698	0.00594
5	579	6	0.99864	0.99296	0.01154
6	1089	5	0.99849	0.99147	0.02054
7	1068	4	0.99848	0.99557	0.00808
8	332	4	0.99913	0.99661	0.00829
9	4546	4	0.00104	0.00574	0.01273
10	2856	3	0.00705	0.02427	0.04575
11	1077	2	0.99754	0.98979	0.02306
12	1130	2	0.99702	0.98905	0.01955
13	1314	2	0.99446	0.98702	0.01870
14	1458	2	0.03666	0.06608	0.07405
15	2945	2	0.00662	0.01907	0.03301
16	3017	2	0.00652	0.02089	0.03409
17	3505	2	0.00690	0.01882	0.02736
18	364	2	0.00070	0.00343	0.00735
19	3647	2	0.99711	0.99087	0.01916
20	3940	2	0.00110	0.00679	0.01516

significant genes

ID	Gene	freq	med.tailp	avg.tailp	sd.tailp
21	4000	2	0.00518	0.01938	0.04030
22	4316	2	0.00307	0.00907	0.01501
23	4518	2	0.99613	0.98757	0.02382
24	921	2	0.00427	0.01553	0.02844
25	1019	1	0.02609	0.05919	0.08251
26	1097	1	0.98372	0.95464	0.06977
27	1254	1	0.05608	0.10606	0.13399
28	1304	1	0.45729	0.44219	0.25522
29	1329	1	0.98314	0.94902	0.08370
30	1346	1	0.00623	0.01732	0.02315
31	1376	1	0.94766	0.90139	0.11893
32	1507	1	0.98338	0.96269	0.05127
33	1557	1	0.99700	0.99364	0.00971
34	1572	1	0.98078	0.96464	0.04936
35	1589	1	0.00465	0.01368	0.02841
36	2196	1	0.87675	0.83352	0.15338
37	2211	1	0.93297	0.89519	0.11501
38	2562	1	0.94549	0.91661	0.10483
39	2621	1	0.94060	0.90535	0.10385
40	2785	1	0.03247	0.06530	0.08392

significant genes

ID	Gene	freq	med.tailp	avg.tailp	sd.tailp
41	2852	1	0.98477	0.96116	0.05833
42	2923	1	0.98263	0.94807	0.07953
43	3200	1	0.98383	0.97065	0.03308
44	324	1	0.07255	0.10029	0.10782
45	3250	1	0.11034	0.18107	0.16665
46	3269	1	0.01113	0.02749	0.04406
47	3282	1	0.99231	0.98366	0.02078
48	3375	1	0.99625	0.98876	0.01913
49	3665	1	0.00334	0.01259	0.02587
50	3746	1	0.02420	0.05999	0.08984
51	3913	1	0.04891	0.08153	0.09197
52	3991	1	0.00386	0.01089	0.01700
53	4013	1	0.99163	0.96972	0.05307
54	4040	1	0.00873	0.02553	0.04489
55	4088	1	0.00244	0.01099	0.02712
56	4104	1	0.00551	0.01658	0.02696
57	4396	1	0.01437	0.03665	0.05615
58	4405	1	0.04284	0.08130	0.09194
59	4417	1	0.06078	0.09663	0.10167
60	4496	1	0.01351	0.03074	0.04737

significant genes

ID	Gene	freq	med.tailp	avg.tailp	sd.tailp
61	4500	1	0.01956	0.04209	0.05434
62	4515	1	0.01350	0.03210	0.05167
63	4541	1	0.04966	0.08819	0.10353
64	478	1	0.01699	0.03281	0.04496
65	4997	1	0.97501	0.94558	0.08181
66	5287	1	0.01858	0.03591	0.05126
67	5746	1	0.86325	0.81823	0.17506
68	594	1	0.97485	0.93502	0.09160
69	676	1	0.03004	0.05106	0.06270
70	690	1	0.11505	0.16742	0.14933
71	694	1	0.00514	0.01530	0.02568
72	735	1	0.00318	0.01201	0.02058
73	987	1	0.97621	0.94328	0.08751

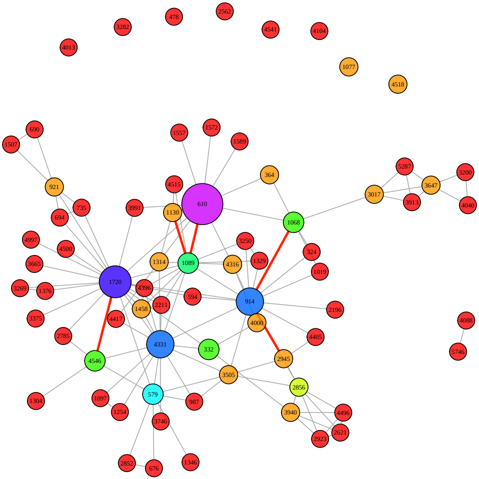

Figure 2: There are 73 significant genes by using the threshold FDR ≤ 0.1 from 100 cross-validations. The node denotes the significant genes and edges denotes the occurrence of two genes at the same time in a cross-validation. The node size indicates the frequency of occurrence for that gene and the edge width indicates the frequency of occurrence of the pair of significant genes at the same time. The genes with the same color have the same frequency of occurrence.

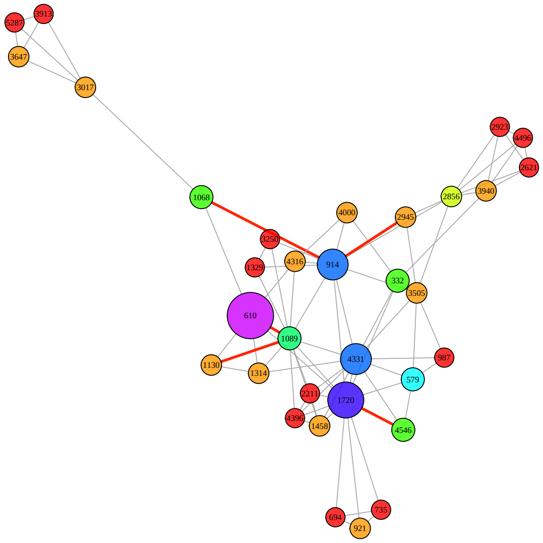

Figure 3: The simplified gene network compare to figure 2. It is obtained by deleting the significant genes in figure 2 with less than 3 edges. 33 significant genes shows in this simplified gene network.

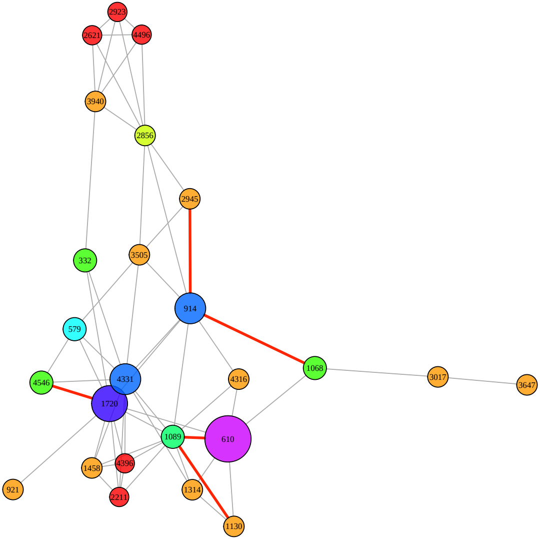

Figure 4: The sparse gene network compare to figure 2. It is obtained by deleting the significant genes in figure 2 with less than 4 edges. 25 significant genes shows in this simplified gene network.

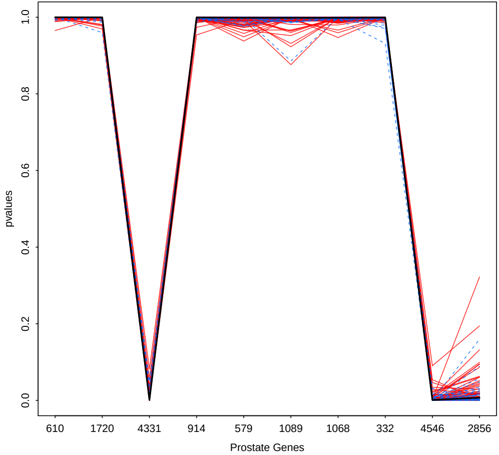

Figure 5: Parallel Coordinate Plot of the left tail probabilities of *t*-statistics for the top 10 significant genes 610, 1720, 4331, 914, 579, 1089, 1068, 332, 4546 and 2856 discovered from 100 cross-validation data. The red solid lines indicate the 69 cross-validations with significant cases screened by FDR ≤ 0.1 and the blue dashed lines indicate the 31 cross-validations without significant cases. The black solid line represents the median of left tail probabilities.

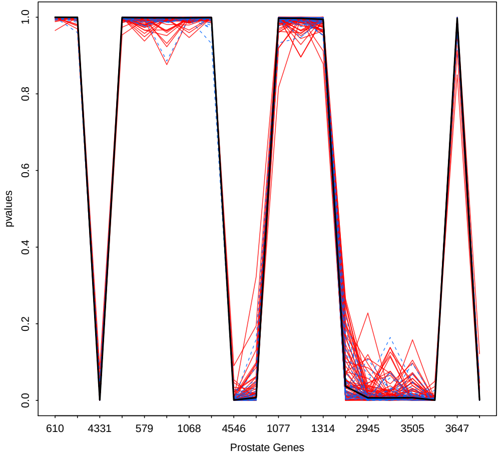

Figure 6: Parallel Coordinate Plot of the left tail probabilities of *t*-statistics for the top 20 significant genes discovered from the 100 cross-validation data. The red solid lines indicate the 69 cross-validations with significant cases screened by FDR ≤ 0.1 and the blue dashed lines indicate the 31 cross-validations without significant cases. The black solid line represents the median of left tail probabilities.

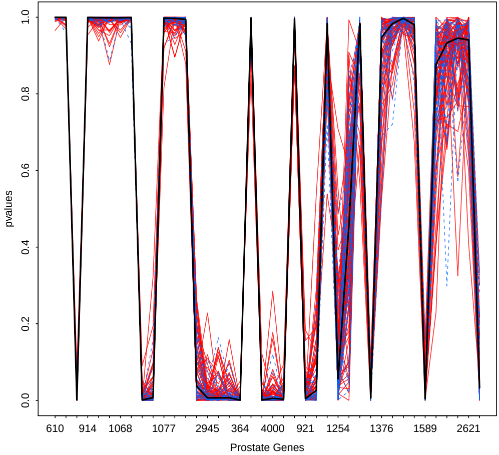

Figure 7: Parallel Coordinate Plot of the left tail probabilities of *t*-statistics for the top 40 significant genes discovered from the 100 cross-validation data. The red solid lines indicate the 69 cross-validations with significant cases screened by FDR ≤ 0.1 and the blue dashed lines indicate the 31 cross-validations without significant cases. The black solid line represents the median of left tail probabilities.

Concluding Remarks

- This approach circumvents over-fitting.
- Using different subsets as training and verification data for each repetition, we balance out other sources of variation in the data.
- The observation(s), that turns out as significant frequently, can give us an idea about the “regulator(s)” associated with extreme behavior in the data.
- If the same sets of observations get identified as significant cases again and again, we can get an idea about a network between them or some hierarchical regulation pattern that may control the signal.
- By using half of the data for model fitting and other half for anomaly detection we are loosing some power of the test that can be achieved by the full model.
- The screening test and the FDR update are large sample results. For a handful of multiple tests this approach will not work well.

Future Work

- We are planning to use similar methods with a discrete mixture model that can be used for count data.
- We need to look into the network of genes revealed by the repeated cross-validation and try to incorporate the inter-relation in the model.

Thank You.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testinfg. *J. Roy. Statist. Soc. Ser. B* **57(1)**. 289-300.

Efron, B. (2007). Size, power and false discovery rates. *Annals of Statistics*. **35 (4)**. 1351–1377.

Efron, B. (2008). Microarrays, empirical Bayes, and the two-groups model. *Statist. Sci.* **23**. 1–22.

Efron, B. (2010). Large-Scale Inference: Empirical Bayes Methods for Estimation. *Testing and Prediction.* Cambridge University Press, Cambridge, UK.

Efron, B. and Tibshirani, R. (2002). *Empirical Bayes methods and false discovery rates for microarrays.* Genetic Epidemiology 23 7086.

Robbins, H. (1956). *An empirical Bayes approach to statistics.* Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. **I**. University of California Press, Berkeley, CA, 157–163.

Van Campenhout, J. M. and Cover, T. M. (1981). Maximum Entropy and Conditional Probability, *IEEE Trans on Info. Th.* **27**. 483–489.

Vera, F., Dickey, D. and Lynch, J. (2010). Asymptotic distribution theory for contamination models, unpublished draft, *available at http://www.stat.sc.edu/ lynch/SpuriousObservations4-19-10.pdf.*

Storey, J. (2002). *A direct approach to false discovery rates.* J. R. Stat. Soc. Ser. B Stat. Methodol. 64

479498.