Distribution Agreement

Signature:

______________________________ _______________
Wanqi Chen Date

Approval Sheet

Detecting differentially expressed genes when there is no replicate: a Bayesian inference with historical data-based informative priors

By

Wanqi Chen
Degree to be awarded: MSPH

Department of Biostatistics and Bioinformatics

____ [Chair’s signature]
Zhaohui "Steve" Qin
Committee Chair

____ [Member’s signature]
Howard Chang
Committee Member

Detecting differentially expressed genes when there is no replicate: a Bayesian inference with historical data-based informative priors

By

Wanqi Chen

Bachelor of Science
Sun Yat-Sen University
2016

Thesis Committee Chair: Zhaohui "Steve" Qin, Ph.D

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2019

Abstract

Detecting differentially expressed genes when there is no replicate: a Bayesian inference with historical data-based informative priors

By

Wanqi Chen

Background: As modern high-throughput sequencing technologies such as microarray have become essential in biological studies, the number of publicly assessible datasets has also dramatical increased. The next generation sequencing technologies lead to the '*large p, small n'* problem, we focus on the extreme case that there is no replicate in the sample when detecting differentially expressed genes. To combine historical data and current studies, hierarchical models serve as the ideal tools.

Methods: The key idea of our method is to borrow information from highly correlated "relative" genes when conducting inference on a single gene. We utilize historical data to identify the correlation structure and specify an informative prior distribution, followed by Bayesian inference using the informative prior. We use the posterior distribution to make statistical inference, and also rank the probability of differential expressed genes.

Results: In simulation studies, our proposed strategy make accurate and robust inference on gene expression levels. It also outperforms GFOLD in differentially expressed genes detection with lower false discovery rate and larger area under the receiver operating characteristic curve.

Conclusion: We illustrated the feasibility and effectiveness of using informative priors from historical data to help detect differentially expressed genes when there is no replicate.

Detecting differentially expressed genes when there is no replicate: a Bayesian inference with historical data-based informative priors

By

Wanqi Chen

Bachelor of Science
Sun Yat-Sen University
2016

Thesis Committee Chair: Zhaohui "Steve" Qin, Ph.D

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics
2019

Table of Contents

1. Introduction

Modern high-throughput sequencing technologies such as microarray have become essential in biological studies. These developing technologies have the capability of sequencing multiple DNA molecules in parallel, generating comprehensive insights into molecular biology, genetics, and genomics (Churko, Mantalas, Snyder, & Wu, 2013). The wider application of modern sequencing technologies also leads to statistical challenges.

In conventional statistical problems, the number of observations *n* is usually much larger than the number of features *p* to be analyzed. However, in modern sequencing applications, the difficulty of sample preparation and its relatively high cost, result in the number of featured variables being larger than the number of samples, which is often termed as the '*large p, small n'* problem (Fan & Lv, 2010). With the aim to detect the difference in gene expression level between two groups, traditional statistical methods faced great challenges. Extensions of these procedures have been used to improve the accuracy of statistical inferences for '*large p, small n'* problem. Nevertheless, all these methods have been developed only based on current data in the analysis, ignoring the information from previous experiments. Given the dramatical increase of publicly accessible genomic datasets, existing data and information can be considered and utilized.

To combine historical data with current study, hierarchical models are ideal options which have been shown effective in detecting gene expression changes from high throughput genomics data (Newton, Kendziorski, Richmond, Blattner, & Tsui, 2001) and other bioinformatics areas (Ji & Liu, 2010). In a hierarchical model framework, all features parameters distributions are assumed to be randomly drawn from a higher level distribution, which provides a reasonable solution to 'borrow' information from historical data an improve the inference results.

In this study, we focused on the extreme case where there is no replicate in case and control samples in a study. In this case, variances within case and control samples cannot be estimated from the current data directly. Our proposed strategy is based on Bayesian hierarchical modeling, using historical data to provide informative priors. We hypothesized that with reasonable normalization and assumed conditions (such as control), the correlation between the expression values of any pair of genes in the genome remain the same, and hence the distributions of such correlation coefficients can be estimated from historical data. We used Gibbs Sampler method (George, 1992) to make statistical inference, and then conducted simulation study to test the effectiveness of this novel model-based method.

2. Method

2.1 Motivation

The majority of high throughput experiments have limited sample size due to practical constraints, which present great challenges for statistical inference. Sophisticated statistical approaches such as DESeq-2(Love, Huber, & Anders, 2014) and Limma (Smyth, 2005) have been developed to handle these difficulties. However, these methods no longer work in the most extreme case when there is no replicate in cases and controls.

2.2 Model

Taking advantage of the fact that many genes in the genome are highly correlated, the key idea of our method is to borrow information from highly correlated "relative" genes when conducting inference on a single gene. To achieve our strategy, we utilize historical data to identify the correlation structure and specify an informative prior distribution, followed by Bayesian inference using the informative prior.

Let $(X_1, X_2, \cdots, X_n)$ denote the gene expression values of a group of genes in the sample which have similar biological functions. The full model is:

$$(X_1, X_2, \cdots, X_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mu_i \sim N(\mu_0, \sigma_0^2)$$

$$\sigma_i^2 \sim InvGamma(\alpha_0, \beta_0)$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)$, $\boldsymbol{\Sigma} = \boldsymbol{S}^{\frac{1}{2}}\boldsymbol{R}\boldsymbol{S}^{\frac{1}{2}}, \boldsymbol{S} = Diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$, and the correlation matrix $\boldsymbol{R}$ is assumed known (derived based on historical data). $\boldsymbol{\mu}$ is the location parameter for the sample, and $\boldsymbol{\Sigma}$ is the covariance matrix.

There are four prior parameters $\mu_0, \sigma_0, \alpha_0, \beta_0$ for one sample. μ_0, σ_0^2 can be interpreted as location and scale parameters for μ_i while α_0, β_0 are shape and scale parameters for σ_i^2. And we also assumed that the $\mu_i's$ and $\sigma_i^2\,'s$ in one sample were slightly different with each other.

2.3 Statistical Inference and Testing

Since we consider the situation with no replicate, the data likelihood is the data distribution itself.

With the aim to detect the expression difference for each gene between two groups, the data distribution can be conditionally written as a one-dimensional distribution:

$$X_1 | \boldsymbol{X_{-1}} \sim N\left(\mu_1^*, {\sigma_1^*}^2\right)$$

where $\mu_1^*, {\sigma_1^*}^2$ are the parameters for the conditional data distribution. Here we chose μ_1 as the example, and the inference for other $\mu_i's$ are similar.

$$\mu_1^* = \mu_1 + \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}}(\boldsymbol{X_{-1}} - \boldsymbol{\mu_{-1}})$$

$${\sigma_1^*}^2 = \sigma_1^2 - \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{bmatrix}$$

$$\boldsymbol{X_{-1}} = (X_2, \cdots, X_n), \qquad \boldsymbol{\mu_{-1}} = (\mu_2, \mu_3, \cdots, \mu_n), \quad \boldsymbol{\sigma_{-1}^2} = (\sigma_2^2, \sigma_3^2, \cdots, \sigma_n^2)$$

Due to conjugacy, the conditional posterior distributions for μ_1 and σ_1^2 are also normal and inverse- gamma distribution:

$$\mu_1 \mid \boldsymbol{X}, \boldsymbol{\mu_{-1}}, \boldsymbol{S} \sim N\left(\mu_p, \sigma_p^2\right)$$

$$\sigma_1^2 | \boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\sigma_{-1}^2} \sim InvGamma(\alpha_p, \beta_p)$$

where $\mu_p, \sigma_p, \alpha_p, \beta_p$ are parameters for the posterior distributions conditional on all X's:

$$\mu_p = \frac{\frac{\mu_0 + c}{\sigma_0^2} + \frac{x_1}{{\sigma_1^*}^2}}{\frac{1}{\sigma_0^2} + \frac{1}{{\sigma_1^*}^2}} - c, \; c = \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}}(\boldsymbol{X_{-1}} - \boldsymbol{\mu_{-1}})$$

$$\sigma_p^2 = \frac{1}{\sigma_0^2} + \frac{1}{\left(\sigma_1^2 - \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}\right)^2}$$

$$\alpha_p = \alpha_0 + \frac{1}{2}$$

$$\beta_p = \beta_0 + \frac{(X_1 - \mu_1 - c\,)^2}{2}$$

We have two known parameters in our model: μ and σ^2. Since we would like to use this model-based strategy to detect the difference in expression level, we focus more on the expression level so μ is the parameter of interest. For the ith gene in two samples (i.e. X_i, Y_i), the hypotheses are:

$$H_0: \mu_{Xi} = \mu_{Yi}$$

$$H_A: \mu_{Xi} \neq \mu_{Yi}$$

We calculated the probability of observing Y_i given that H_0 is true:

$$\Pr(observed\ Y_i \mid \mu_{Xi} = \mu_{Yi}) = \Pr\,(Y_i | \mu_{Xi}, \sigma_{Xi}^2, \boldsymbol{X})$$

We used this probability to rank differentially expressed (DE) genes.

3. Results

3.1 Statistical Inference using Gibbs Sampler

We conducted simulation studies to test the performance of the proposed method in estimating model parameters. To show the generality of this model-based strategy, we tried various settings of the correlation matrix, dimension, as well as the initial values of the MCMC. Our method performed well in inferring mean (μ_i) under all conditions.

We randomly generated 30 gene sets from the same multivariate normal distribution and each gene set contained 20 genes. Then we applied our model-based strategy to make inference on. Prior parameters were $\mu_0 = 2, \sigma_0 = 0.15, \alpha_0 = 3, \beta_0 = 0.5$, and random correlations between 0 to 1. We named this informative prior Bayesian model-based method IPBM.

The boxplot (Figure 1) summarized the difference between observations/estimator and true means for all 20 genes in 30 repeated simulations. Generally, we can see the estimators are accurate and robust.

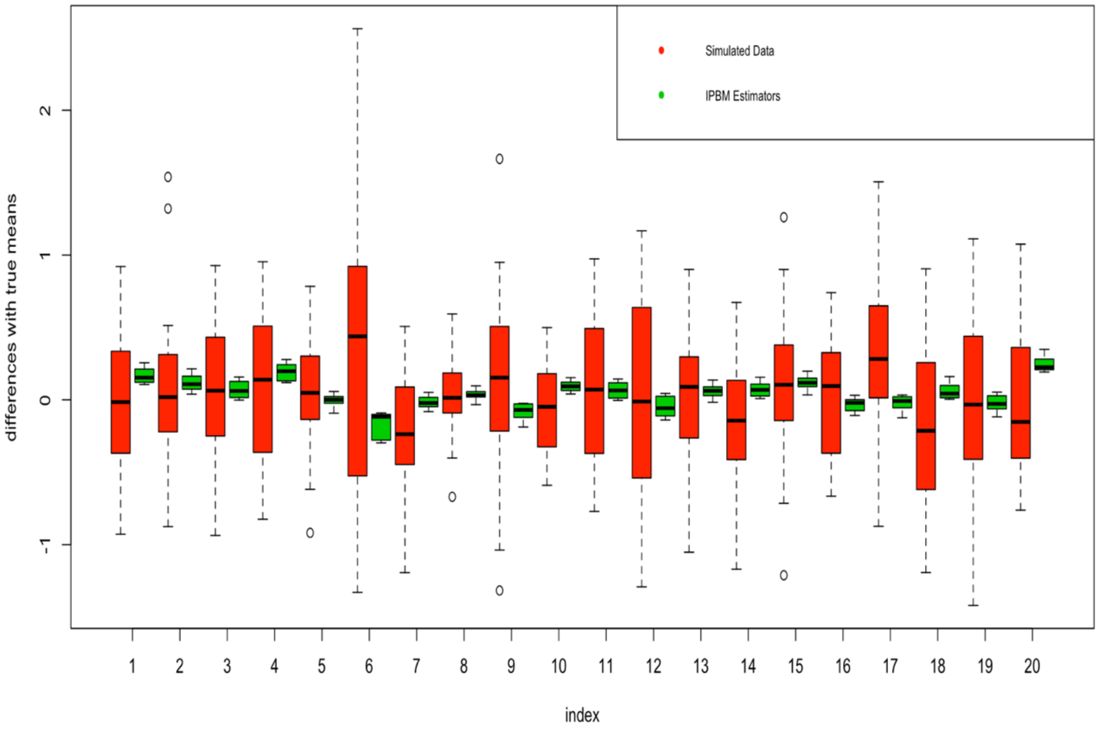

Figure 1 The differences between simulated data or IPBM estimators and true means

3.2 Differentially Expressed Genes Detection

We used simulation to test the performance of IPBM. The control group contained 50 gene sets, and each gene set had 20 genes. We created two treatment samples, both of them contained 50*20 genes and 20% of them were differentially expressed. We examined two scenarios: the differences in gene expressions between treatment and control groups were either uniformly distributed between 0.15 and 0.30, or uniformly distributed between 0.30 and 0.60. Prior parameters were $\mu_0 = 2, \sigma_0 = 0.15, \alpha_0 = 3, \beta_0 = 0.5$, with random correlations between 0 to 1 assigned to each gene set.

Among existing methods, only GFOLD (Meyer et al., 2012) can handle the situation that there is no replicate. We repeated the simulations 500 times and calculated the empirical false discovery rate (FDR) (Benjamini & Hochberg, 1995). Our method performed better than GFOLD with lower FDRs under both situations (Figure 2).

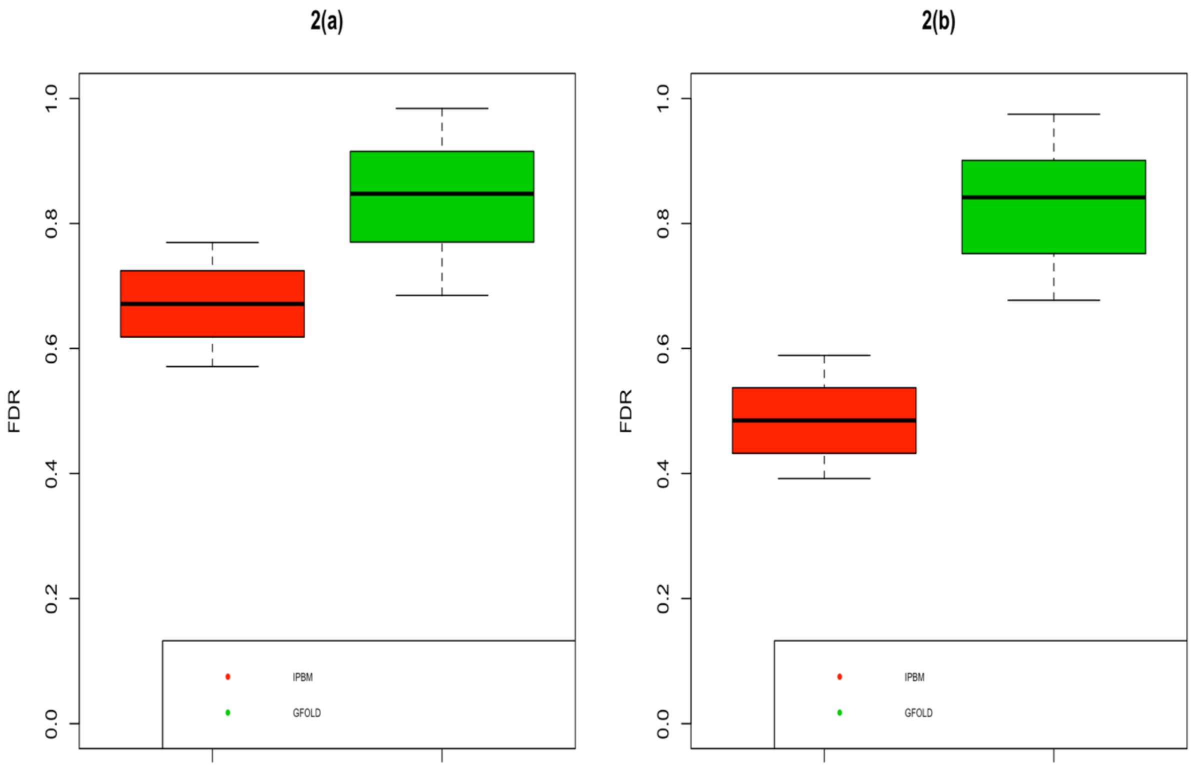

Figure 2 FDR for IPBM and GFOLD methods. (a) Differences between two groups of data were uniformly distributed between 0.15 to 0.30. (b) Differences between two groups of data were uniformly distributed between 0.30 to 0.60.

We used Receiver Operating Characteristic (ROC) (Hanley & McNeil, 1982) curves to further evaluate the performance of these two methods. Figure 3 showed typical ROC curves for one single simulation. Our method clearly outperformed the GFOLD method regardless of the different settings of DE genes.

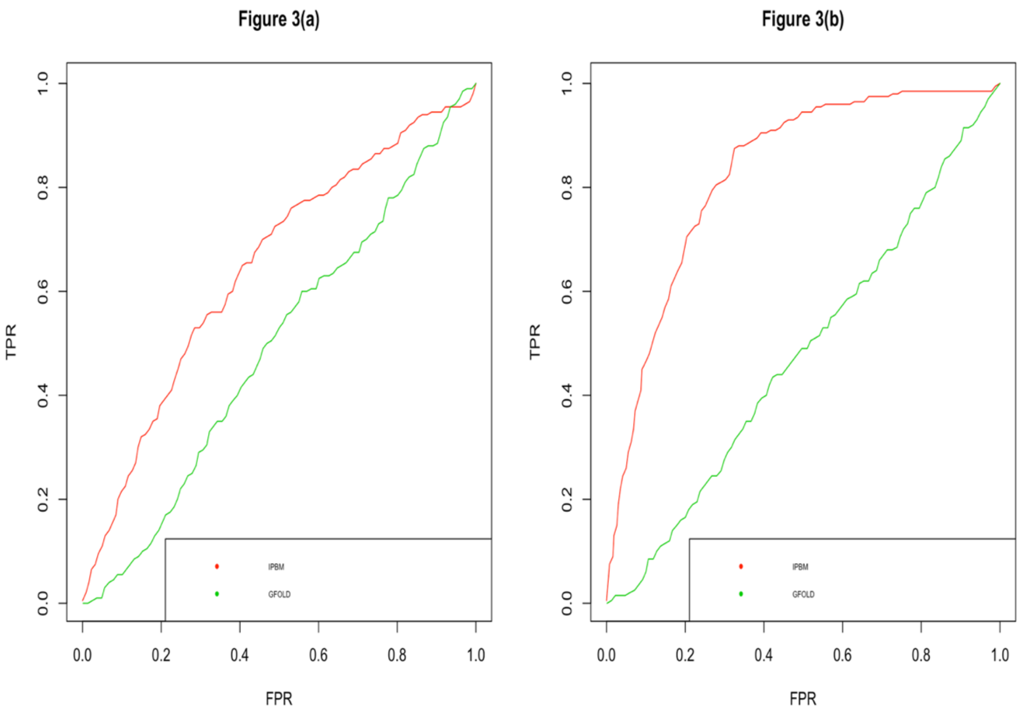

Figure 3 ROC curves for IPBM and GFOLD methods. (a) Differences between two groups of data were uniformly distributed between 0.15 to 0.30. (b) Differences between two groups of data were uniformly distributed between 0.30 to 0.60.

4. Discussion

We presented a novel method called IPBM based on Bayesian hierarchical model with informative priors in this thesis. Simulation studies showed IPBM outperformed GFOLD in detecting differentially expressed genes when there is no replicate in our sample, with better accuracy, sensitivity, and specificity. However, all results were based on simulation studies. We plan to conduct real data analysis to further test the performance of IPBM.

IPBM focused on the extreme case that there is no replicate, but theoretically it can handle the common cases in bioinformatics that there are small number of replicates as well. And many widely used DE genes detecting methods were designed for similar situations, such as DESeq2 (Love et al., 2014) and Limma (Smyth, 2005).

The current IPBM method uses multivariate normal distribution to model the log transformed expression level of gene sets for mathematical convenience. The application of other non-normal distributions may help to improve the robustness of our method in inference and prediction (Ganjali, Baghfalaki, & Berridge, 2015). We used IPBM to detect different expression level gene by gene. But since we considered the correlations between genes in the model, another possible improvement of our method is to detect the different expression level for whole gene sets. This will potentially improve IPBM's performance with greater flexibility.

5. Conclusion

We illustrated the feasibility and effectiveness of using informative priors from historical data to help detect DE genes when there is no replicate. In simulation studies, our proposed strategy showed its better performance than GFOLD. With the increasingly available genomics data, we presented a promising method to make statistical inference and to detect differentially expressed genes.

We simulated microarray data to test IPBM, since microarray is more popular than RNA-seq due to its relatively lower cost. But RNA-seq can provide more information about the transcriptome, our proposed method can be applied on it as well. Additionally, cross-platform models may also be feasible.

Reference

Benjamini, Y., & Hochberg, Y. J. J. o. t. R. s. s. s. B. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *57*(1), 289-300.

Churko, J. M., Mantalas, G. L., Snyder, M. P., & Wu, J. C. (2013). Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. *Circ Res, 112*(12), 1613-1623. doi:10.1161/CIRCRESAHA.113.300939

Fan, J., & Lv, J. (2010). A Selective Overview of Variable Selection in High Dimensional Feature Space. *Statistica Sinica, 20*(1), 101-148.

Ganjali, M., Baghfalaki, T., & Berridge, D. J. P. o. (2015). Robust modeling of differential gene expression data using normal/independent distributions: a Bayesian approach. *10*(4), e0123791.

George, E. I. (1992). Explaining the Gibbs Sampler AU - Casella, George. *The American Statistician, 46*(3), 167-174. doi:10.1080/00031305.1992.10475878

Hanley, J. A., & McNeil, B. J. J. R. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *143*(1), 29-36.

Ji, H., & Liu, X. S. (2010). Analyzing 'omics data using hierarchical models. *Nat Biotechnol, 28*(4), 337-340. doi:10.1038/nbt.1619

Love, M. I., Huber, W., & Anders, S. J. G. b. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *15*(12), 550.

Meyer, C. A., Feng, J., Liu, J. S., Wang, Q., Shirley Liu, X., & Zhang, Y. (2012). GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics, 28*(21), 2782-2788. doi:10.1093/bioinformatics/bts515 %J Bioinformatics

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., & 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*(1), 37-52. doi:10.1089/106652701300099074

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor* (pp. 397-420): Springer.