Latent Process Decomposition of High-Dimensional Count Data

By

Sanvesh Srivastava and R.W. Doerge

Technical Report #13-04

Department of Statistics
Purdue University

June, 2013

Biometrics 000, 1–25 DOI: 10.1111/j.1541-0420.2005.00454.x
May 0000

Latent Process Decomposition Of High-Dimensional Count Data

Sanvesh Srivastava* and R.W. Doerge**

Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907, USA.

**email:* srivasta@purdue.edu

***email:* doerge@purdue.edu

Summary: Next-generation sequencing (NGS) technologies have become the preferred way of exploring a genome. These data are high-dimensional discrete counts with latent structure that, once revealed, will reduce the dimensions and will lead to the subsets of genes that are suitable for further exploration. Latent Process Decomposition of high-dimensional count data (LPD-C) is presented as a two stage approach that is based on the assumption that genes work in groups or processes. The first stage uses a variational empirical Bayesian approach that adapts the Latent Dirichlet Allocation algorithm and extends the Latent Process Decomposition algorithm for high-dimensional Gaussian data. The second stage of LPD-C selects gene-subsets using empirical Bayes hypothesis testing. The performance of LPD-C is explored using simulated and publicly available NGS data, compared with existing approaches, and shown to be a useful and extensible framework for identifying genes suitable for further exploration. Although we apply LPD-C in a genomic context, it can be used for any high-dimensional count data.

Key words: Empirical Bayes; hypothesis testing; hierarchical Bayesian modelling; mixed-membership models; next-generation sequencing data; variational inference.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Most genomic data are collected and analyzed for the purpose of associating genomic covariates (e.g., gene expression, microRNA expression, DNA copy number, and single nucleotide polymorphisms) with individuals' phenotypes (e.g., disease status and survival time). Recent advances in high-throughput technologies such as microarrays and next-generation sequencing (NGS) have enabled measurements of complex biological and genomic activities at an extremely high resolution (Marioni et al., 2008). Typically, the number of available genomic covariates is much larger than the number of individuals sampled; therefore, these data are considered high-dimensional. One of the major challenges in genomic data analysis is to model the statistical dependence of phenotypes on the genomic covariates, while accounting for their high-dimensionality and interactions. The aim is to identify the covariates that are most predictive of the phenotypes, and thereby suitable for further exploration. Low signal strength and presence of confounding variables only complicate the analyses.

NGS technologies have emerged as the preferred approach for exploring a genome because their data are highly replicable with little technical variation, and they facilitate novel genomic discoveries (Marioni et al., 2008). Compared to microarrays, the issues in NGS data analysis are magnified simply due to the non-Gaussian, discrete, and overdispersed nature and increased complexity and size of the data (Marioni et al., 2008). We present the Latent Process Decomposition of high-dimensional count data (LPD-C), a two stage approach for NGS data analysis that models the generative mechanism of NGS data and selects a pre-specified number of gene-subsets that have desirable properties. The proposed framework is generic and computationally efficient, is well-suited to handle the increased complexity and size of genomic data, and can be easily used by practitioners.

Historically, microarray investigations and research has addressed many of the aforementioned challenges posed by high-dimensional data. It has led to significant advancements in

the applications and theory of multiple hypothesis testing (Efron, 2010), high-dimensional variable selection, and the use of penalized likelihood approaches, especially the Lasso, for high-dimensional data analysis (Friedman et al., 2010). Of these approaches, two major themes arise for selecting candidate genes suitable for further exploration. The first theme frames the problem as a gene-wise multiple hypothesis testing problem, with the rejected hypotheses corresponding to the candidate genes (Efron, 2010). The second theme proposes modeling the exchangeability of genes either using two level generative Bayesian models or using penalized likelihood approaches (Friedman et al., 2010). The Bayesian approach uses posterior distributions and the penalized likelihood approach chooses appropriate tuning parameters to select candidate genes. Both of these themes borrow information across genes as recommended by Efron (2010), and work well, both theoretically and practically, for high-dimensional Gaussian data such as microarray data.

Limited results exist for the analysis of high-dimensional non-Gaussian data, such as NGS data. Multiple hypothesis testing for discrete data is nontrivial because there are no equivalents of the t- or F-test statistics (Robinson et al., 2010). Further, the sampling distributions of test statistics are hard to justify in the current small sample size setting and even worse if interacting genes are considered (Efron, 2010). Luckily, Bayesian modeling approaches for high-dimensional non-Gaussian data analysis do not suffer from these problems. However, the convergence and diagnostic procedures for, and the scalability of, Markov chain Monte Carlo (MCMC) based approaches in high dimensions are fairly complex for generic applications. Few penalized likelihood approaches address statistical significance in high-dimensional models and the analysis of genomic data from complex experimental designs (e.g., time series of gene expression data; Meier and Bühlmann (2007)). Young et al. (2012) provide an excellent overview of existing hypothesis testing based methods for NGS data analysis; most of these approaches model NGS data using a negative binomial distribution.

Witten (2011) proposes sparse Poisson Linear Discriminant Analysis (SPLDA), a penalized likelihood approach using Lasso penalty, that performs much better than many existing methods for classifying and clustering NGS data. Currently, no Bayesian approach exists that selects genes while modeling the generative mechanism of NGS data along with the biological hypothesis that genes work in networks or pathways.

LPD-C is a special case of the Latent Process Decomposition (LPD) framework. A specific case of LPD was first proposed for high-dimensional Gaussian data (LPD-G) in the context of microarray data (Rogers et al., 2005). Rogers et al. (2005) proposed LPD-G as a more flexible approach than classical unsupervised approaches (hierarchical or K-means clustering) to model the biological hypothesis that genes work in groups or networks. Because their main objective was to find clusters of genes in microarray data, Rogers et al. (2005) do not select candidate genes suitable for further exploration. The proposed LPD framework, of which LPD-G and LPD-C are special cases, amends the original generative Bayesian model via a second stage that selects candidate gene-subsets. Selected genes have two properties: they are a small fraction of the total number of genes, and they are associated to their respective subsets with high probabilities. Since the generative Bayesian model of LPD is an example of Bayesian Latent Factor model (BLFM), the processes in LPD correspond to factors in BLFM. West (2003) and Carvalho et al. (2008) present applications of BLFM to microarray data. Their model is similar to LPD-G's modeling approach. Dunson and Herring (2005) model discrete outcomes, including count data, using BLFM. An extension of their model to high-dimensional count data, which accounts for the fact that genes act in networks, is similar to LPD-C's generative Bayesian model. That said, there is a key difference between BLFM and LPD. In BLFM, latent factors (processes) are of main interest, whereas in LPD the focus is on estimating mean genomic effects.

Motivated by the need for a Bayesian approach that selects genes, the methodology and

associated computations for LPD-C are developed in the context of NGS data. LPD-C is a two stage approach that adapts to the underlying latent structure of the data and helps in understanding complex systems. When applied to NGS data, LPD-C's first stage is an unsupervised approach to model the biological hypothesis that genes work in groups, or processes, and is a special case of the mixed membership modeling framework (Airoldi et al., 2005). This stage uses a variational empirical Bayesian approach that adapts the Latent Dirichlet Allocation algorithm (LDA) (Blei et al., 2003) to model the generative mechanism of NGS data. In doing so the framework becomes highly extensible and facilitates computationally efficient estimation of the parameters and hyperparameters. The second stage uses the parameter estimates from the first stage to select candidate genes, organized as gene-subsets, using empirical Bayes hypothesis testing framework (Efron, 2010). The second stage has few assumptions and controls the number of false discoveries. In real data analysis, LPD-C's results agree closely with those of hypothesis testing and penalized likelihood based approaches. LPD-C's distinguishing features are that it selects gene-subsets in NGS data, and it can be easily extended to model data from more complicated experimental designs.

2. Latent process decomposition of high-dimensional count data

NGS data can be represented as a matrix N of gene counts with S rows and G columns that represent samples and genes, respectively. The gene counts for s-th sample are denoted as $\mathbf{n}_s$ (i.e., the s-th row of N), and n_{sg} is the count for gene g in sample s. There are K latent processes (hereafter processes) associated with each sample. Any gene in a sample can belong to one of the K processes. Due to the unsupervised nature of the analysis, we ignore any covariate information associated with the samples.

2.1 *First stage of LPD-C: Hierarchical Bayesian model*

Consider a three level generative Bayesian model for $\mathbf{n}_s$. The first (population) level of the sampling model generates the probability vector $\boldsymbol{\pi}_s = (\pi_{s1}, \ldots, \pi_{sK})$ of process memberships for genes in sample s from a Dirichlet distribution with parameters, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, such that

$$\boldsymbol{\pi}_s \mid \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \quad \text{(Level 1)} \tag{1}$$

and $\boldsymbol{\pi}_s$ is a latent variable specific to sample s. Model (1) implies that the genes in sample s belong to K sub-populations called processes, and the probability of a gene belonging to a process depends on the sample.

In the second level, the model generates the process membership k of gene g in sample s as the latent Multinomial random vector $\mathbf{z}_{sg}$ of length K, with all zeros except 1 at the k-th position

$$\mathbf{z}_{sg} \mid \boldsymbol{\pi}_s \sim \text{Multinomial}(1; \boldsymbol{\pi}_s), \text{ for } g = 1, \ldots, G, \quad \text{(Level 2)} \tag{2}$$

$$\mathbf{z}_{sg} = (z_{sg1}, \ldots, z_{sgk}),\ k \text{ is such that } z_{sgk} = 1 \text{ and } z_{sgj} = 0 \text{ for } j \neq k.$$

Z_s is a latent indicator matrix specific to sample s with $\mathbf{z}_{sg}$ as its rows. It has G rows and K columns representing genes and processes, respectively. The column with the non-zero entry in the g-th row of Z_s indicates the latent process membership of gene g.

Finally, the third level generates the count n_{sg} for gene g in sample s based on its process membership k as

$$n_{sg} \mid \lambda_{gk} \sim \text{Poisson}(\lambda_{gk}), \quad \text{(Level 3)} \tag{3}$$

where λ_{gk} is an element of the gene- and process-specific mean ("loadings") matrix Λ with G rows and K columns that represent genes and processes, respectively. The gene counts for all the samples are generated following (1) – (3). Hereafter $\boldsymbol{\alpha}$ and Λ are assumed to be fixed parameters. The generative model of LPD-C adapts the sampling models of LDA (Blei

et al., 2003) and LPD-G (Rogers et al., 2005) for NGS data, which implies that the genes and processes are exchangeable within a sample. All of these models are examples of BLFMs that have been successfully used for microarray data analysis (West, 2003; Carvalho et al., 2008).

The generative model (1) – (3) makes LPD-C more flexible than classical unsupervised approaches, such as hierarchical and K-means clustering (Blei et al., 2003; Rogers et al., 2005). Specifically, (2) associates genes in sample s to different processes chosen from the K processes using Multinomial$(1; \boldsymbol{\pi}_s)$. This level gives rise to two major advantages of LPD-C. First, (2) enables LPD-C to both model the biological hypothesis that genes work in groups (processes) and to associate genes to processes that can vary depending on their functions in a sample. Second, due to its greater flexibility than classical clustering models, LPD-C can better adapt to the latent structure of NGS data.

2.1.1 *Estimation of posterior distributions of parameters.* Each gene-subset corresponds to a process and $Z_{1:S}$ relate samples, genes, and processes. LPD-C selects K gene-subsets using test statistics obtained from the posterior density for $Z_{1:S}$, $p(Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$. The joint density of the latent variables $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_S$ (hereafter $\boldsymbol{\pi}_{1:S}$) and $Z_1, \ldots, Z_S$ (hereafter $Z_{1:S}$) and NGS data $\mathbf{n}_1, \ldots, \mathbf{n}_S$ (hereafter $\mathbf{n}_{1:S}$) given $\boldsymbol{\alpha}$ and Λ, $p(\boldsymbol{\pi}_{1:S}, Z_{1:S}, \mathbf{n}_{1:S}\,|\,\boldsymbol{\alpha}, \Lambda)$ is analytically intractable (Blei et al., 2003); therefore, the posterior density $p(Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$ is also analytically intractable. There are a host of techniques that can be used to approximate $p(Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$, including MCMC.

We employ Poisson variational Bayes methods from machine learning and obtain analytically tractable variational density $q(\boldsymbol{\pi}_{1:S}, Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$ that approximates analytically intractable true posterior density $p(\boldsymbol{\pi}_{1:S}, Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$ (Bishop, 2006). This choice is important for the computational efficiency and practical applicability of LPD-C. The variational approach minimizes the Kullback-Liebler (KL) divergence between $q(\boldsymbol{\pi}_{1:S}, Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$

and $p(\boldsymbol{\pi}_{1:S}, Z_{1:S} |\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$ $(KL(q||p))$. This approximation achieves analytic tractability by assuming that latent variables $\boldsymbol{\pi}_{1:S}$ and $Z_{1:S}$ are independent under the variational posterior density, so that

$$q(\boldsymbol{\pi}_{1:S}, Z_{1:S} |\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda) = \prod_{s=1}^{S} q(\boldsymbol{\pi}_s) q(Z_s) = \prod_{s=1}^{S} q(\boldsymbol{\pi}_s) \Big(\prod_{g=1}^{G} q(\mathbf{z}_{sg}) \Big), \tag{4}$$

where the conditioning on data and hyperparameters is suppressed. The variational posterior densities of $\boldsymbol{\pi}_s$ and Z_s are $q(\boldsymbol{\pi}_s) = q(\boldsymbol{\pi}_s \,|\, \mathbf{n}_s, \boldsymbol{\alpha}, \Lambda)$ and $q(Z_s) = q(Z_s |\, \mathbf{n}_s, \boldsymbol{\alpha}, \Lambda)$. The factorization (4) alone guarantees the analytic tractability of $q(\boldsymbol{\pi}_{1:S}, Z_{1:S} |\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$, and there are no further distributional assumptions for q's. Using Section 1 of Supplementary Material, the variational approximation introduces variational parameters, $\boldsymbol{\gamma}_s = (\gamma_{s1}, \ldots, \gamma_{sK})$ and $\{\Phi_{sg} = (\phi_{sg1}, \ldots, \phi_{sgK})\}_{g=1}^{G}$, which are estimated using $\mathbf{n}_s, \boldsymbol{\alpha}$, and Λ, so that

$$\begin{aligned} q(\boldsymbol{\pi}_s \,|\, \boldsymbol{\gamma}_s) &= \text{Dirichlet}(\gamma_{s1}, \ldots, \gamma_{sK}),\ \gamma_{sk} = \alpha_k + \sum_{g=1}^{G} \phi_{sgk}, \\ q(\mathbf{z}_{sg} \,|\, \Phi_{sg}) &= \text{Multinomial}(1; \phi_{sg1}, \ldots, \phi_{sgK}),\ \phi_{sgk} = \frac{\mathcal{P}(n_{sg}|\lambda_{gk}) \exp[\Psi(\gamma_{sk})]}{\sum_{k'=1}^{K} \mathcal{P}(n_{sg}|\lambda_{gk'}) \exp[\Psi(\gamma_{sk'})]}, \end{aligned} \tag{5}$$

where $\mathcal{P}(n_{sg}|\lambda_{gk})$ denotes the Poisson density with mean λ_{gk} evaluated at n_{sg}.

Because in real data analysis $\boldsymbol{\alpha}$ and Λ are rarely known, we choose an empirical Bayesian approach and estimate $\boldsymbol{\alpha}$ and Λ based on $\mathbf{n}_{1:S}$. Following Blei et al. (2003), instead of maximizing $\log p(\mathbf{n}_{1:S} \,|\, \boldsymbol{\alpha}, \Lambda)$, its lower bound from variational inference $\log q(\mathbf{n}_{1:S} \,|\, \boldsymbol{\alpha}, \Lambda)$ is maximized for estimating $\boldsymbol{\alpha}$ and Λ. The log variational lower bound (ELBO) $\log q(\mathbf{n}_{1:S} \,|\, \boldsymbol{\alpha}, \Lambda)$ has the advantage of being analytically tractable. ELBO is obtained by replacing the functions of latent variables $\boldsymbol{\pi}_{1:S}$ and $Z_{1:S}$ in $\log p(\boldsymbol{\pi}_{1:S}, Z_{1:S}, \mathbf{n}_{1:S} \,|\, \boldsymbol{\alpha}, \Lambda)$ by their conditional expectations with respect to $q(\boldsymbol{\pi}_s \,|\, \boldsymbol{\gamma}_s)$ and $q(Z_s |\, \Phi_s)$ for $s = 1, \ldots, S$. This observation motivates iterative estimation of $\boldsymbol{\alpha}$ and Λ based on $\mathbf{n}_{1:S}$ similar to EM algorithm (Dempster et al., 1977). Specifically, if $\phi_{sgk}^{(t)}, \gamma_{sk}^{(t)}, \lambda_{gk}^{(t)}, \alpha_k^{(t)}$ represent the parameter estimates at the t-th

iteration, then the $(t+1)$-th updates for these parameters are obtained as

$$\phi_{sgk}^{(t+1)} = \frac{\mathcal{P}(n_{sg}|\lambda_{gk}^{(t)}) \exp[\Psi(\gamma_{sk}^{(t)})]}{\sum_{k=1}^{K} \mathcal{P}(n_{sg}|\lambda_{gk}^{(t)}) \exp[\Psi(\gamma_{sk}^{(t)})]}, \qquad \gamma_{sk}^{(t+1)} = \alpha_k^{(t)} + \sum_{g=1}^{G} \phi_{sgk}^{(t+1)},$$

$$\lambda_{gk}^{(t+1)} = \frac{\sum_{s=1}^{S} \phi_{sgk}^{(t+1)} n_{sg}}{\sum_{s=1}^{S} \phi_{sgk}^{(t+1)}}, \qquad \boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} - \mathbf{H}(\boldsymbol{\alpha}^{(t)})^{-1}\, \mathbf{g}(\boldsymbol{\alpha}^{(t)}), \tag{6}$$

where $\Psi(.)$ is the digamma function (Abramowitz and Stegun, 1970) and $\mathbf{H}$ and $\mathbf{g}$ are the Hessian and gradient for $\boldsymbol{\alpha}$ update (see Sections 1 and 2 of Supplementary Material; and Bishop (2006) for details). We start the iterations using $\phi_{sgk} = \frac{1}{K}$ for all samples, genes, and processes, and $\alpha_k = 1$ for all processes. Later, we recommend two practical approaches for choosing K (see Sections 2.1 – 2.4 of Supplementary Material for details).

2.1.2 *Interpretation Of Parameter Estimates.* The interpretation of parameters in Section 2.1.1, and the relation between them are described using (6). The probability that gene g in sample s belongs to the process k is ϕ_{sgk}; therefore, $\sum_{k=1}^{K} \phi_{sgk} = 1$ and $\sum_{g=1}^{G} \phi_{sgk}$ is the expected number of genes in sample s that belong to process k. The probability that sample s belongs to the process k is proportional to γ_{sk}. The prior probability that a gene in any NGS experiment belongs to process k is proportional to α_k. The expected number of genes in process k in sample s is $\gamma_{sk} - \alpha_k$, which equals $\sum_{g=1}^{G} \phi_{gsk}$. This relation can be used for checking the convergence of iterative updates in (6). The expected value of the count for gene g, when it belongs to the process k, is λ_{gk}.

2.2 *Second stage of LPD-C: Selection of gene-subsets*

The second stage of LPD-C employs $\widehat{\Phi}_1, \ldots, \widehat{\Phi}_S$, which are estimated using (6), to select genes grouped in K subsets. This stage depends on the local false discovery rate (locfdr) cutoff for each subset (Efron, 2007, 2010). The selected genes are a small fraction of the

total number of genes and are associated to their respective subsets with high probabilities. Most importantly, this stage extends the original LPD framework of Rogers et al. (2005) and makes it more useful for high-dimensional genomic data analysis by selecting genes, grouped in subsets, that are suitable for further exploration, while controlling the number of false discoveries. Based on the locfdr procedure, the second stage of LPD-C has the following advantages. It does not require modeling of full error structure of the original data set, has few assumptions, and is easy to implement (Efron, 2007). The trade-off for these advantages is the loss of statistical efficiency (Efron, 2007). The advantages, however, are of primary importance in practical data analysis.

Because gene-subsets correspond to processes, genes in subset k are selected using test statistics based on $\widehat{\phi}_{sgk}$ for G genes across S samples. If z_{sgk}'s are known for all the samples and genes, then $p_{gk} = \frac{\sum_{s=1}^{S} z_{sgk}}{S}$ represents the probability that gene g belongs to process k. Motivated from EM algorithm (Dempster et al., 1977), modified test statistics $\widehat{p}_{gk}$ are defined by replacing the latent variables z_{sgk}'s in p_{gk} by their conditional expectations with respect to $q(Z_s|\widehat{\Phi}_s)$ for $s = 1, \ldots, S$ and

$$\widehat{p}_{gk} = \frac{\sum_{s=1}^{S} \mathbb{E}_{z_{sgk}}[z_{sgk}|\, \mathbf{n}_s]}{S} \approx \frac{\sum_{s=1}^{S} \mathbb{E}_{q(Z_s|\widehat{\Phi}_s)}[z_{sgk}]}{S} = \frac{\sum_{s=1}^{S} \widehat{\phi}_{sgk}}{S}. \tag{7}$$

The test statistic (7) represents the approximate posterior probability of gene g belonging to process k. Because Efron (2007) recommends using the test statistics for genes that have the same range as the normal distribution, $\widehat{p}_{gk}$ is transformed to the corresponding quantile of the central t-distribution with ν degrees of freedom, t_{gk}, using its cumulative distribution function $\mathcal{F}_{t_\nu}$, and $t_{gk} = \mathcal{F}^{-1}_{t_\nu}(\widehat{p}_{gk})$. The t-distribution is chosen due to its heavy tails; in real data analysis, we choose $\nu = 3$. Assuming that T represents the matrix of test statistics with G rows and K columns, genes with high posterior probabilities of belonging to process k are in the right tail of $\mathbf{t}_k$, k-th column of T; therefore, $\mathbf{t}_k$ is used as the vector of test statistics in an empirical Bayes testing framework to select genes in subset k that are non-

null, that lie in the right tail of $\mathbf{t}_k$, and when locfdr is controlled at a small pre-specified value. This procedure selects a small fraction of genes that are associated with subset or process k with high probabilities. We select K gene-subsets based on the columns of T and separately control locfdr for each column. For the NGS data applications presented later, the R package `locfdr` (Efron et al., 2008) is employed.

3. Applications of LPD-C

We apply LPD-C to simulated and real NGS data, and compare its performance to both SPLDA (Witten, 2011) and a negative binomial model (EdgeR; Robinson et al. (2010)). These methods are chosen because Witten (2011) shows that SPLDA performs significantly better than current approaches (except EdgeR) for classifying and clustering NGS data. The simulated data are generated using the hierarchical model (1) – (3). Two publicly available NGS datasets are used: human cervical cancer data (hereafter cervical cancer data; Witten et al. (2010)) and human gene expression data from liver and kidney (hereafter human data; Marioni et al. (2008)). These real data are chosen because Witten (2011) shows that both EdgeR and SPLDA perform well for the human data, but that the cervical cancer data are challenging for both of these methods. It is important to remember that LPD-C models the generative mechanism of NGS data, and that it is fundamentally different from the hypothesis testing based approach of EdgeR, and from the penalized likelihood based approach of SPLDA. However, the comparisons illustrate the similarities and differences in these methods. The novel feature that distinguishes LPD-C from existing approaches for NGS data analysis is that it groups selected genes into a pre-specified number of gene-subsets, and these gene-subsets have desirable properties.

3.1 *Simulation*

We simulated 10 NGS datasets such that each dataset contains 12 samples (S) with 2 processes (K) for different settings of G and λ_{gk}'s. For each setting, the simulated data have the following process membership for the genes. In samples 1 to 10, the first 100 genes (hereafter group 1 genes) belong to the first process and the last 100 genes (hereafter group 2 genes) belong to the second process. For a particular G, the first 10 samples have the following five settings of gene- and process-specific means λ_{gk}'s depending on Δ,

$$\lambda_{g1} = \begin{cases} \exp(z_{g1}),\ z_{g1} \sim \text{Normal}(\Delta, 1) & \text{for } g = 1, \ldots, 100, \\ \exp(z_{g1}),\ z_{g1} \sim \text{Normal}(0, 0.25) & \text{for } g = 101, \ldots, G, \end{cases}$$
$$\lambda_{g2} = \begin{cases} \exp(z_{g2}),\ z_{g2} \sim \text{Normal}(0, 0.25) & \text{for } g = 1, \ldots, G - 100, \\ \exp(z_{g2}),\ z_{g2} \sim \text{Normal}(\Delta, 1) & \text{for } g = G - 99, \ldots, G, \end{cases} \tag{8}$$

where Δ is varied as $1, 2, 3, 4$, and 5. These values of Δ represent the difference between the log-means of the "null" (i.e., genes that are not in group 1 and 2) and "non-null" genes (i.e., group 1 and 2 genes) in the two processes. The number of genes (G) is varied as 2000 and 20,000 genes, respectively, while the number of non-null genes is 200 in both cases. For samples 11 and 12, the process memberships of group 1 and 2 genes are reversed. The remaining genes belong to the two processes with 0.5 probability across all samples. These parameter values are motivated from Efron et al. (2008) and Witten (2011). NGS data are simulated using these parameter values and LPD-C's generative model (1) – (3). The simulated data are similar to those observed in practice, with a large fraction of small counts and a small fraction of large counts.

3.1.1 *Application of EdgeR, LPD-C, and SPLDA to simulated data.*

[Figure 1 about here.]

We applied the first stage of LPD-C to 50 replications of the simulated data. For each application of LPD-C, we chose $K = 2$ to facilitate comparison with the truth and estimated $\boldsymbol{\alpha}$, Λ, Φ's, and $\boldsymbol{\gamma}$'s (see (6) for their definition). The results of variational approximation are known to be sensitive to the starting points, which in LPD-C's case depend on $\boldsymbol{\alpha}$ and Φ's (Bishop, 2006). We used multiple starting points until convergence to the posterior mode was stable. We observed that the final parameter estimates were most sensitive to the starting values of Φ's and were fairly robust to the starting values of $\boldsymbol{\alpha}$. The process numbers are identified based on the ascending order of $\widehat{\alpha}_k$'s such that $\widehat{\alpha}_{(1)}$ and $\widehat{\alpha}_{(2)}$ correspond to processes 1 and 2, respectively.

[Figure 2 about here.]

Figure 1 compares $\log \widehat{\lambda}_{gk}$'s with their true values in processes 1 and 2 given $\Delta = 1, \ldots, 5$. It shows that as Δ increases, LPD-C's estimates accurately capture the true bimodal density of $\log \lambda_{gk}$'s. Despite being an approximate method, LPD-C's estimates are fairly close to the truth even at low values of Δ. Figure 2 shows the ELBO that variational inference maximizes for estimating the variational posterior densities across LPD-C's iterations for $\Delta = 1, \ldots, 5$. Similar to the EM algorithm, variational updates monotonically increase the ELBO guaranteeing convergence to the local mode of the objective function for determining the variational posterior densities.

After estimating $\widehat{p}_{gk}$'s from $\widehat{\phi}_{sgk}$'s, we obtain $t_{gk} = \mathcal{F}_{t_3}^{-1}(\widehat{p}_{gk})$, where $\mathcal{F}_{t_3}$ is the cumulative distribution function of the central t-distribution with 3 degrees of freedom (see Section 2.2). We apply empirical Bayes hypothesis testing to the columns of T, which correspond to processes, and select genes that are non-null, that are in the right tail, and that have a locfdr of 0. The selected genes belong to their processes, and hence to the corresponding gene-subsets, with high probability. Figure 2 shows the proportion of true positive genes, grouped as subsets, selected by LPD-C depending on the number of genes G and Δ. It shows that,

for $G = 2000$ and $20,000$, as Δ increases, the proportion of true positive genes respectively selected in the two gene-subsets by LPD-C increases to 1.

3.1.2 *Comparison of results obtained using EdgeR, LPD-C, and SPLDA.*

[Figure 3 about here.]

LPD-C selects genes grouped in subsets, but EdgeR and SPLDA do not; therefore, we compare overall gene selection of LPD-C with that of EdgeR and SPLDA. Unlike LPD-C, both EdgeR and SPLDA select genes based on a response variable. We define a response variable (Y) that is 1 for the first ten samples and is 2 for samples 11 and 12, and EdgeR selects genes that are differentially expressed between samples with $Y = 1$ and $Y = 2$. Similarly, SPLDA finds a sparse list of genes that can classify samples as $Y = 1$ or $Y = 2$ based on their expression while minimizing the cross-validation (CV) error for classification. We use `edgeR` package (Robinson et al., 2010) for EdgeR and `PoiClaClu` package (Witten, 2011) for SPLDA. We obtain the gene-wise p-values for differential expression using `edgeR`, correct for multiple comparisons using the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995), and choose the genes that corresponded to 200 smallest BH corrected p-values. Because `PoiClaClu` uses CV to select the tuning parameters and requires minimum 4 samples, which is infeasible in our simulation, we instead choose the tuning parameters depending on G and Δ that select 200 genes. These modified gene selection criteria for EdgeR and SPLDA facilitate better comparison of their results with LPD-C.

Figure 3 shows the true positive and false discovery proportions for EdgeR, LPD-C, and SPLDA at different values of G and Δ. The proportion of true positives selected by LPD-C increases with Δ when $G = 2000$ and 20,000; however, the proportion of false discoveries are much higher than their expected value when $G = 20,000$. This observation is expected because the number of non-null genes are 200 for both $G = 2000$ and 20,000, and the apparent increase in true positives when $G = 20,000$ comes at the cost of increased false discoveries.

The true positive proportions for both EdgeR and SPLDA behave similar to that in LPD-C, but their values are lower than that of LPD-C. The false discovery proportions of SPLDA and LPD-C are much higher than that of EdgeR when $G = 20,000$. This observation for EdgeR is an artifact our gene selection procedure that only selects the first 200 genes based on the ascending order of p-values; however, the proportion of false discoveries of EdgeR when $G = 20,000$ is also much higher than those shown in Figure 3 when genes are selected based on the standard FDR cutoff 0.05.

Biological knowledge suggests that genes work in networks. Our simulation study illustrates that LPD-C leads to better results than EdgeR and SPLDA; both do not model the dependence among genes. LPD-C also provides interpretable parameters, and its distinguishing feature is that it selects genes grouped as subsets. Further, because LPD-C uses variational inference, it is computationally tractable and more efficient than other sampling-based Bayesian approaches.

3.2 *Real data examples*

We apply LPD-C to two publicly available NGS datasets. The cervical cancer data provide measurements of the digital expression for 714 small RNAs (miRNAs; hereafter miRNAs and genes are used interchangeably) in 29 tumor and 29 normal cervical tissue samples from humans (Witten et al., 2010). The human data provide measurements of the digital expression for 22,925 genes in 14 samples from a single human male, which consists of seven technical replicates from liver and kidney, respectively (Marioni et al., 2008). The cervical cancer data were collected for the purpose of discovering miRNAs associated with human cervical cancer. The human data were collected in order to compare microarray and NGS technologies.

3.2.1 *Two approaches for selecting the number of processes K in real data.* We suggest two practical approaches based on n-fold CV for selecting the number of processes K in

real data. The problem of selecting K is similar to that of selecting the number of clusters, which is known to be a notoriously difficult problem; therefore, we suggest fitting LPD-C for a range of K's and selecting those K's which lead to results that agree closely with the biological knowledge.

The first approach chooses K based on the held-out log likelihood in n-fold CV (Rogers et al., 2005). It is well-suited for small sample sizes, such as in the human data. This approach randomly splits the samples into n partitions, fits LPD-C using the samples in $n-1$ partitions (training data), calculates the held-out log likelihood for the samples that are not in training data (test data) using parameter estimates obtained from the training data, and repeats this process n times by separately using each partition as the test data. It yields n held-out log likelihoods for a particular K. These n log likelihoods are calculated when K is varied from 1 to a large integer. The chosen number of processes corresponds to the K that maximizes the medians of the held-out log likelihoods.

The second approach chooses K using the true positive proportion (TPP) and false discovery proportion (FDP) determined using training and test data (Friedman et al., 2010). It is more suitable for relatively large sample size data such as found for the cervical cancer data. Similar to the first approach, the second approach randomly splits the samples into training and test data, and separately selects genes in both sets using LPD-C for a particular value of K. Assuming that the genes selected in the training data represent the truth, this approach calculates the proportion of true positives and false discoveries in the genes selected by LPD-C in the test data. This process is repeated n times to yield n TPPs and FDPs for each K. The values of K that have large TPPs and small FDPs represent good choices of K (see Section 2.4 in Supplementary Material for greater details about the CV-based approaches).

[Figure 4 about here.]

Figures 4a and 4b illustrate the determination of K for both the cervical cancer and human

data using 5-fold CV. We choose $K = 5$ for the human data because it has the maximum held-out log likelihood with a small median absolute deviation estimate compared to other values of K. For the cervical cancer data, both $K = 5$ and 3 are reasonable choices. As such, we selected genes in the cervical cancer data using LPD-C for both $K = 5$ and 3 and found that the results obtained using $K = 5$ agree closely with biological knowledge, as well as with other approaches for selecting genes in NGS data; therefore, all our subsequent analyses for the cervical cancer data are based on $K = 5$.

3.2.2 *Application of EdgeR, LPD-C, and SPLDA to real data.*

[Figure 5 about here.]

The first stage of LPD-C estimates $\boldsymbol{\alpha}$, Λ, Φ's, and $\boldsymbol{\gamma}$'s for both the cervical cancer and human data using $K = 5$ (see (6)). Similar to the simulation study, we tried various starting points for $\boldsymbol{\alpha}$ and Φ's until convergence to the posterior mode was stable; identified the process numbers based on the ascending order of $\widehat{\alpha}_k$'s; after estimating $\widehat{p}_{gk}$'s from $\widehat{\Phi}$'s, obtained $t_{gk} = \mathcal{F}_{t_3}^{-1}(\widehat{p}_{gk})$; and used columns of T to select genes using a locfdr cutoff of 0.2 for each subset. The two features of empirical Bayes hypothesis testing that are useful here are its mild distributional assumptions on, and no requirement for modeling the full error structure of t_{gk}'s (Efron, 2007, 2010).

We also apply EdgeR and SPLDA to the cervical cancer data using tumor status as the response variable. Similarly, EdgeR and SPLDA are applied to the human data using liver and kidney as values of the response variable. We obtain gene-wise p-values for differential expression using `edgeR`, correct for multiple comparisons using the BH procedure, and select genes using 0.05 as the cutoff for the BH corrected p-values. We used 5-fold CV for both the cancer and human data in SPLDA. While the choice of tuning parameter using CV is fairly stable for the human data, multiple tuning parameters lead to the same CV classification error for the cervical cancer data, which results in unstable gene-selection. For example, at the

same value of CV error for classification, SPLDA selects 2 genes when the tuning parameter is 8.23 and selects 499 genes when the tuning parameter is 0.57. The reason for these unstable results is that a large range of tuning parameters yield the same classification error, 0.172, which corresponds to 10 out of 58 errors. Classification error is a very coarse measure (unlike, for example, the mean square error for regression), so many tuning parameter values are tied in terms of CV error for classification. We choose the tuning parameter as the mean of all the tuning parameter values that correspond to the minimum CV classification error (personal communication, D. Witten).

3.2.3 *Results obtained using EdgeR, LPD-C, and SPLDA.* Figures 6a and 6b summarize the number of genes selected in both the cervical cancer and the human data using EdgeR, LPD-C, and SPLDA, respectively. The total number of genes selected by EdgeR, LPD-C, and SPLDA are 267, 265, and 39 for the cervical cancer data, and 12746, 14029, and 7 for the human data. LPD-C selected 103 unique miRNAs (genes) in the cervical cancer data that are related to different types of cancers, including cervical cancer, and that are not selected by EdgeR or by SPLDA. Specifically, some of these 103 miRNAs are known to be in the *let-7*, *mir-7*, *mir-17*, *mir-24*,*mir-26*, *mir-27*, *mir-29*, *mir-124*, *mir-127*, *mir-192*, and *miR-744* families, and include miRNAs that play important roles in regulating different forms of cancers, such as metastasis, tumorigenesis, and tumor suppression. These miRNAs also have clinical applications in cancer diagnosis and therapy. Because the human data were collected for comparing microarray and NGS technology, we did not investigate the biological annotation of the genes selected by LPD-C.

LPD-C's results agree closely with those of EdgeR in that both of these methods select most of the genes chosen by SPLDA. When compared to EdgeR, about 61% and 70% of the genes selected by LPD-C in the cervical cancer data and the human data are also declared as differentially expressed. We also notice that the number of genes selected by SPLDA in

both datasets is much smaller than the number of genes selected by either EdgeR or LPD-C. Further, unlike SPLDA and EdgeR, LPD-C groups the selected genes into subsets with desirable properties.

[Figure 6 about here.]

Marioni et al. (2008) employed both microarray and NGS technologies to compare the differentially expressed genes. They used two sample t tests for the microarray data analysis using a Gaussian model and likelihood ratio tests for NGS data analysis using a negative binomial model. Figure 6c compares the 4105 genes selected only by LPD-C in the human data (Figure 6b) with the differentially expressed genes in microarray or NGS data reported by Marioni et al. (2008), excluding the 9924 genes that lie in the intersection of LPD-C and EdgeR (Figure 6b). We observe that almost half of the genes that are selected solely by LPD-C are also reported as differentially expressed in the microarray or NGS data results of Marioni et al. (2008). Among the remaining differentially expressed genes, about 88% of the genes are in the 2822 differentially expressed detected only by EdgeR (see Figure 6b). This observation suggests that the results of LPD-C do contain genes that are potentially differentially expressed, and that are not selected by EdgeR.

The distinguishing feature of LPD-C is that it selects genes grouped as subsets having desirable properties. Table 1 summarizes the number of genes in each of the five gene-subsets as selected by LPD-C for both the cervical cancer data, and the human data. It also illustrates the number of genes that are in common when any two gene-subsets are compared, as well as the proportion of genes that are differentially expressed. For the human data, where SPLDA and EdgeR results agree, LPD-C results are similar. For the cervical cancer data, where SPLDA and EdgeR results do not agree, LPD-C leads to results that are close to those of EdgeR.

[Table 1 about here.]

3.3 *Summary of data analysis*

We have demonstrated that LPD-C selects genes that compare favorably with existing approaches, such as EdgeR and SPLDA, even though SPLDA consistently selects a smaller number of genes than EdgeR and LPD-C. We note that this under-selection issue has been observed in applications of Lasso for variable selection to high-dimensional data that have dependence among the variables (Friedman et al., 2010). Since most high-dimensional biological data, including NGS data, have dependence among their variables, this could be a potential reason for SPLDA selecting a smaller number of genes. Furthermore, since the tuning parameters is not identifiable in the cervical cancer data it leads to an unstable selection of genes which in turn makes SPLDA undesirable if used for gene selection. As an alternative suggestion for situations like these, we recommend using the glmnet algorithm for variable selection (Friedman et al., 2010).

LPD-C's generative model is a three level Bayesian model that includes two level hierarchical models for NGS data (e.g., EdgeR) as special cases. Therefore, LPD-C's modeling results are more flexible, but similar to those of EdgeR. Witten (2011) shows that both EdgeR and SPLDA perform well for the human data. We in turn demonstrate that the results of LPD-C for the human data closely agree with that of EdgeR and SPLDA. Witten (2011) went on to illustrate that the cervical cancer data are a challenge for both EdgeR and SPLDA. Our application of LPD-C to the cervical cancer data successfully discovers subsets of miRNAs that are known to be biologically associated with various types of cancer, and that are not selected by EdgeR or by SPLDA.

4. Discussion

Due to the decreasing cost of using high-throughput technologies and the potential impact of large-scale genome-wide epidemiological and clinical projects, such as ENCODE project

(Birney et al., 2007) and 1000 Genomes Project (Siva, 2008), genomic data are becoming increasingly complex and large (Stein, 2010). Bayesian generative models offer an attractive approach to model the latent structure of genomic data by combining the hierarchy in the sampled population, uncertainty about the underlying model and unknown parameters, and prior experimental knowledge. This type of modeling framework naturally facilities extensions that can capture complex data-specific patterns. Although many approaches exist for analyzing NGS data using multiple hypothesis testing or using penalized likelihood based approaches, few Bayesian approaches exist for NGS data that model the data generative mechanism, and that acknowledge the biological concept that genes work in groups to perform biological functions. To this end, we have presented an application of Latent Process Decomposition (LPD) framework to NGS data, LPD-C, that addresses the aforementioned issues in two stages. LPD-C's first stage extends the generative Bayesian model of LPD for modeling microarray data to NGS data. Its second stage uses the parameter estimates from first stage to select genes, organized as gene-subsets, that are a small fraction of total number of genes and that belong their respective subsets with high probability. To achieve computationally tractable Bayesian inference, we have applied variational techniques from machine learning, and combined the results of variational inference with empirical Bayes hypothesis testing to select gene-subsets that control the number of false discoveries at a certain level. We have explored LPD-C's application in the context of simulated and real NGS data, and demonstrated that LPD-C discovers genes with known biological significance that competing approaches cannot. Furthermore, LPD can be applied to any high-dimensional data by simply modifying the distributional assumptions.

An additional benefit of LPD-C is that it can be easily extended to a supervised model. Typically, a variety of covariate information is available for the experimental units (e.g., disease status, survival time, treatment information, etc.), and about the genes (e.g., depen-

dence between the genes due to known functional associations, pathway information, etc.). The supervised extension of LPD-C modifies (3) as

$$n_{sg}|\lambda_{gk}, \boldsymbol{\beta} \sim \text{Poisson}(\lambda_{gk}\, \mathbf{x}_{sg}^T\, \boldsymbol{\beta}),$$

where $\mathbf{x}_{sg}$ are the covariates specific to gene g in sample s and $\boldsymbol{\beta}$ are the corresponding mean covariate effects. Dunson and Herring (2005) use a similar BLFM for analyzing discrete outcomes in complex health conditions and mixtures of discrete outcomes. They employ a fully Bayesian approach using the Metropolis-Hastings algorithm for sampling from the posterior density of $\boldsymbol{\beta}$ under a multivariate normal prior. Using the same prior for $\boldsymbol{\beta}$, LPD-C's supervised extension yields an approximate multivariate t variational posterior density for $\boldsymbol{\beta}$.

The generative Bayesian model in the first stage of LPD, which is adapted from LDA, naturally facilitates extensions that capture patterns specific to the data (Gelman et al., 2003). LDA has been studied extensively in text mining literature, therefore its extensions motivate the development of methods for modeling genomic data from complex experimental designs. Two extensions of LPD that follow immediately by adapting the extensions of LDA are as follows. Using Dynamic Topic Models (Blei and Lafferty, 2006), LPD can be extended to model a time series of high-dimensional genomic data. This extension models the exchangeability of genes and processes at a time point, but not across time points. Using Correlated Topic Models (Blei and Lafferty, 2007), we can account for correlation between the process memberships of genes, which cannot be modeled through Dirichlet($\boldsymbol{\alpha}$) in (1). The apriori choice of the number of processes, K, facilitates efficient parameter estimation. Sometimes, however, the apriori knowledge about K is unavailable or K is unidentifiable from approaches recommended in Section 3.2.1. In these scenarios it is desirable to adaptively select K based on the genomic data using applications of Bayesian Nonparametrics in genetics, signal processing, and text mining (Hjort et al., 2010).

The second stage of the LPD uses the parameter estimates from the Bayesian model and selects gene-subsets with desirable properties. Similar ideas about finding groups of differentially expressed gene-subsets have been explored starting with gene-set enrichment analysis (Subramanian et al., 2005), and then generalized to gene-set analysis (GSA) (Efron and Tibshirani, 2007). We propose to investigate the relationship between the enriched gene-subsets obtained from GSA and gene-subsets obtained from the LPD. Further, we propose to incorporate sparsity in the second stage of the LPD using appropriate priors on Λ from Bayesian variable selection literature, such as Bayesian Lassos, g-priors, and horse-shoe prior (Liang et al., 2008; Carvalho et al., 2010; Bhattacharya and Dunson, 2011).

Finally, it is also desirable to develop MCMC algorithms tuned for LPD to estimate uncertainty in parameters of interest by sampling from their posterior densities. Following Griffiths and Steyvers (2004), collapsed Gibbs samplers can be developed for LPD. For LPD-C, we first extend its generative model and impose conjugate Gamma priors on λ_{gk}'s. A collapsed Gibbs sampler marginalizes over $\boldsymbol{\pi}$'s and Λ, samples $\mathbf{z}_{sg} = (z_{sg1}, \ldots, z_{sgK})$ given $Z_1, \ldots, Z_s \setminus \mathbf{z}_{sg}, \ldots, Z_S$, $\mathbf{n}_{1:S}$, and $\boldsymbol{\alpha}$ for all the samples and genes, and finally updates the Gamma distributions of λ_{gk}'s given $Z_1, \ldots, Z_S$ and $\mathbf{n}_{1:S}$ and the Dirichlet distribution of $\boldsymbol{\pi}_s$ given Z_s and $\boldsymbol{\alpha}$ for all the samples.

Acknowledgements

This work is funded in part by a National Science Foundation (DBI-0733857) grant to RWD and her colleagues. SS benefited from discussions with Professors J.K. Ghosh and S. Kirshner.

References

Abramowitz, M. and Stegun, I. (1970). *Handbook of Mathematical Functions.* Dover Publications.

Airoldi, E., Blei, D., Xing, E., and Fienberg, S. (2005). A latent mixed membership model

for relational data. In *Proceedings of the 3rd international workshop on Link discovery*, pages 82–89. ACM.

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. Series B (Methodological)* **57,** 289–300.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse bayesian infinite factor models. *Biometrika* **98,** 291–306.

Birney, E., Stamatoyannopoulos, J., Dutta, A., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature* **447,** 799–816.

Bishop, C. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer New York.

Blei, D. and Lafferty, J. (2007). A correlated topic model of science. *The Annals of Applied Statistics* pages 17–35.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research* **3,** 993–1022.

Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association* **103,** 1438–1456.

Carvalho, C., Polson, N., and Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97,** 465–480.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series*

B (Methodological) **39,** 1–38.

Dunson, D. and Herring, A. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics* **6,** 11–25.

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

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

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics* **1,** 107–129.

Efron, B., Turnbull, B., and Narasimhan, B. (2008). locfdr: Computes local false discovery rates. *R package* page 195.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33,** 1.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, Florida, 2 edition.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 5228.

Hjort, N., Holmes, C., Müller, P., and Walker, S., editors (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association* **103,** 410–423.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* **18,** 1509.

Meier, L. and Bühlmann, P. (2007). Smoothing l1-penalized estimators for high-dimensional time-course data. *Electronic Journal of Statistics* **1,** 597–615.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139.

Rogers, S., Girolami, M., Campbell, C., and Breitling, R. (2005). The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* pages 143–156.

Siva, N. (2008). 1000 genomes project. *Nature biotechnology* **26,** 256–256.

Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biology* **11,** 207.

Subramanian, A., Tamayo, P., Mootha, V. K., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 15545.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian statistics* **7,** 723–732.

Witten, D. (2011). Classification and clustering of sequencing data using a poisson model. *The Annals of Applied Statistics* **5,** 2493–2518.

Witten, D., Tibshirani, R., Gu, S., Fire, A., and Lui, W. (2010). Ultra-high throughput sequencing-based small rna discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC biology* **8,** 58.

Young, M., McCarthy, D., Wakefield, M., Smyth, G., Oshlack, A., and Robinson, M. (2012). Differential expression for rna sequencing (rna-seq) data: Mapping, summarization, statistical analysis, and experimental design. *Bioinformatics for High Throughput Sequencing* pages 169–190.

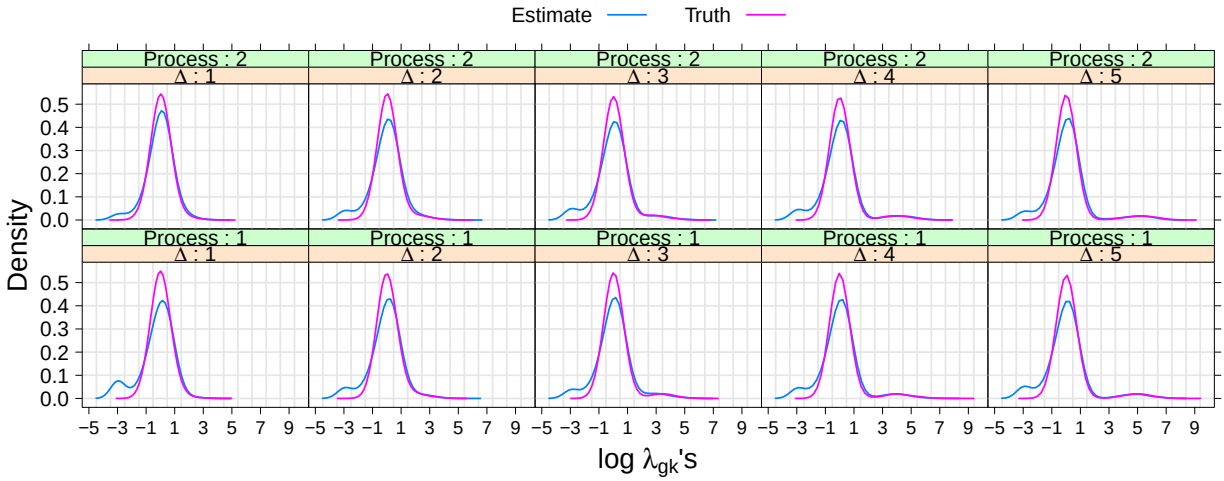


Figure 1: Estimated gene- and process-specific means Λ for $\Delta = 1, 2, 3, 4$, and 5. The bottom and top rows respectively compare estimated (blue) and true (red) log λ_{g1}'s and log λ_{g2}'s conditioned on the difference between the log-means of null and non-null genes Δ using kernel density estimates. In both processes, as Δ increases from low to high (left to right), LPD-C's estimates accurately capture the bimodality of the true density. Although variational inference is an approximate procedure, the estimated means are fairly close to the truth. There are boundary effects in the left tails that are due to the truncation of estimated means to achieve numerical stability.

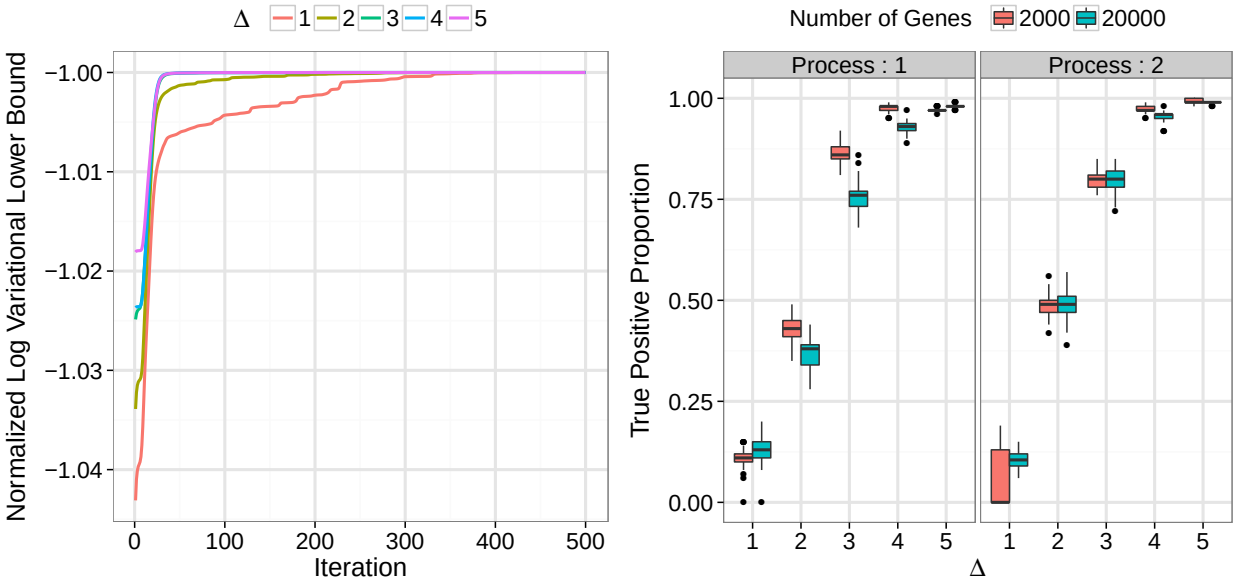


Figure 2: Log variational lower bound (ELBO) during parameter estimation in LPD-C (left) and the proportion of true positive genes selected by LPD-C (right). Variational inference iteratively maximizes ELBO, which is similar to the EM algorithm. The left plot shows the ascent of ELBO normalized by its absolute maximum (y-axis) during the iterations of LPD-C (x-axis) for values of the difference between the log-means of the null and non-null genes in the two processes (Δ; see (8)). The right plot shows the performance of LPD-C in selecting genes for 50 replications of simulated data. The x-axis represents Δ and the y-axis represents the true positive proportion (TPP) for the genes selected in the two processes represented by the panels respectively when the number of genes (G) is 2000 (red) and 20,000 (blue). The TPP increases with Δ in both the processes when $G = 2000$ and 20,000. TPPs for $G = 2000$ are greater than those for $G = 20,000$. This trend is expected because the number of non-null genes is 200 in both the cases, and it easier to detect true positives when their proportion is large (i.e., when $G = 2000$).

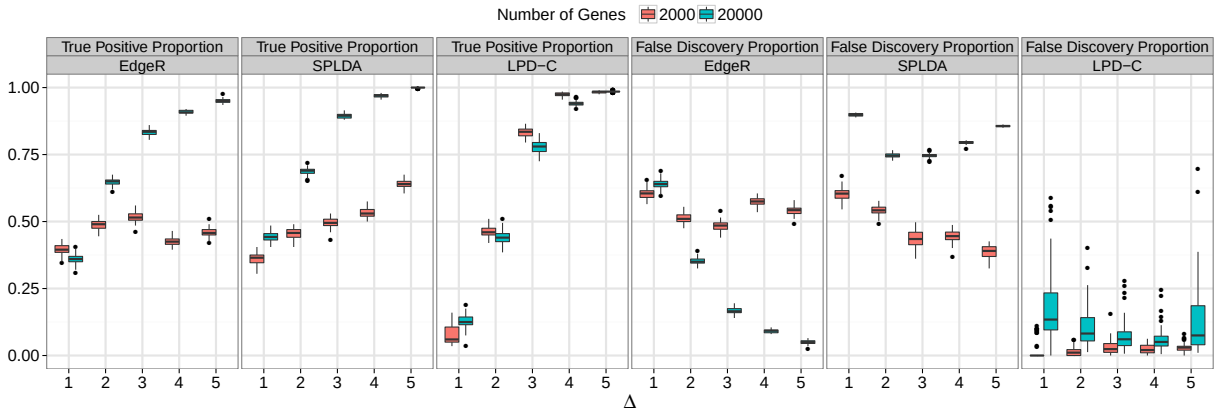
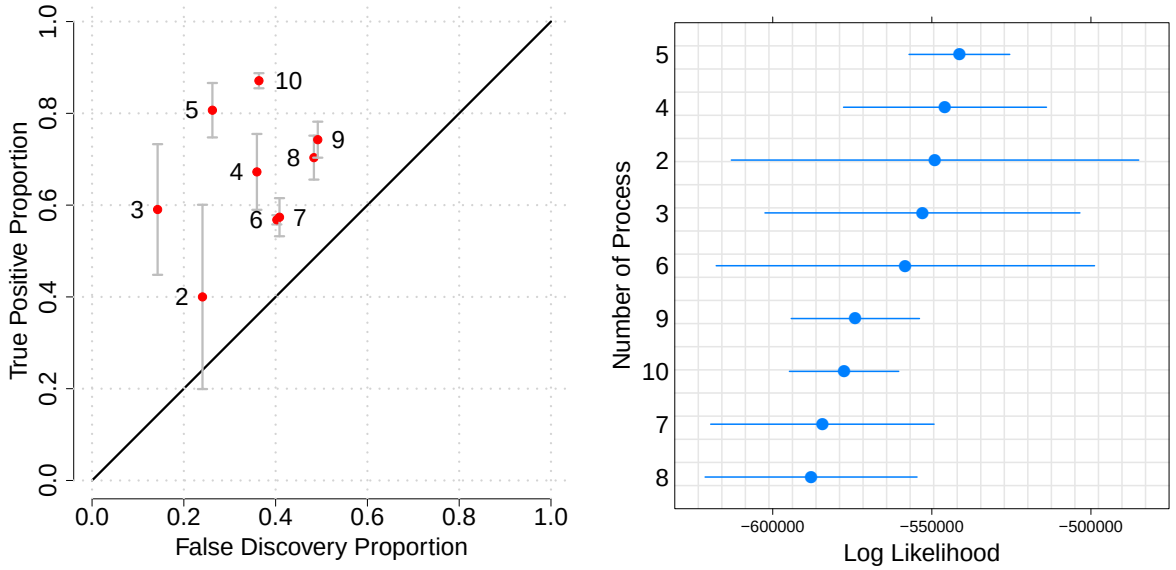
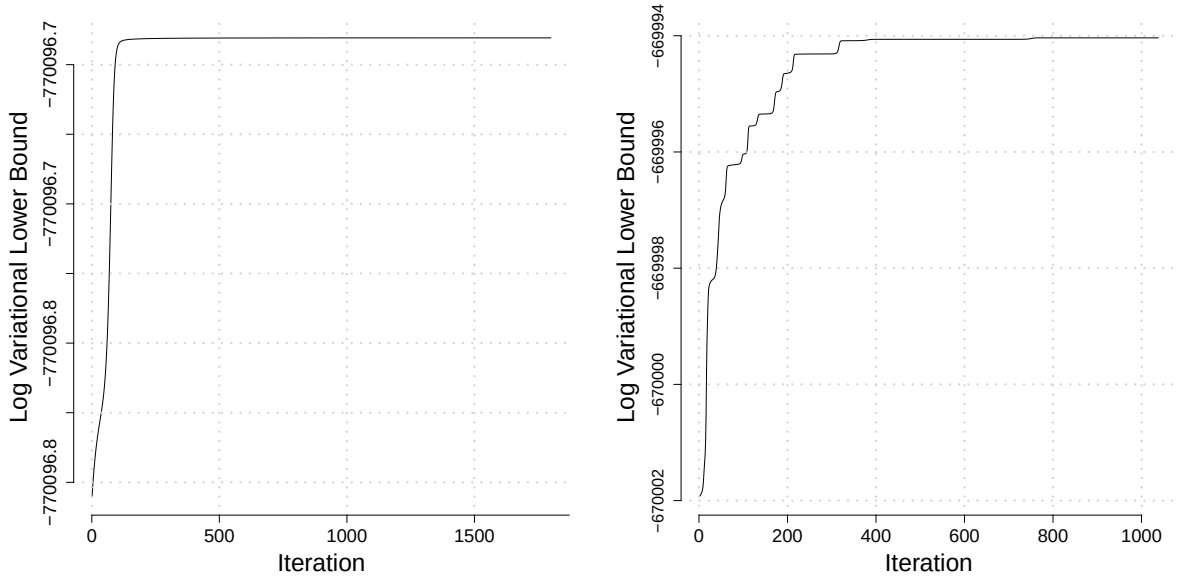


Figure 3: Comparison of true positives and false discoveries in the genes selected by EdgeR, LPD-C, and SPLDA in 50 replications of simulated data analysis. The x-axis represents the difference between the log-means of the null and non-null genes in the two processes (Δ; see (8)) and y-axis represents the true positive (TPP) and false discovery (FDP) proportions. Panels one through three represent TPPs, while panels four through six represent FDPs, for EdgeR, LPD-C, and SPLDA, respectively, when the number of genes (G) is 2000 (red) and 20,000 (blue). For both LPD-C and SPLDA, TPP increases and FDP decreases as Δ increases when $G = 2000$ and $20,000$; EdgeR also has a similar pattern except when $G = 2000$, where the TPP and FDP oscillates around 0.5 for all the Δ's. Although TPPs appear to increase with Δ for EdgeR, LPD-C, and SPLDA, the FDPs are much higher than their expected values. This pattern is expected because the number of non-null genes is same for $G = 2000$ and 20,000, and the power increase is accompanied by an increase in FDPs.



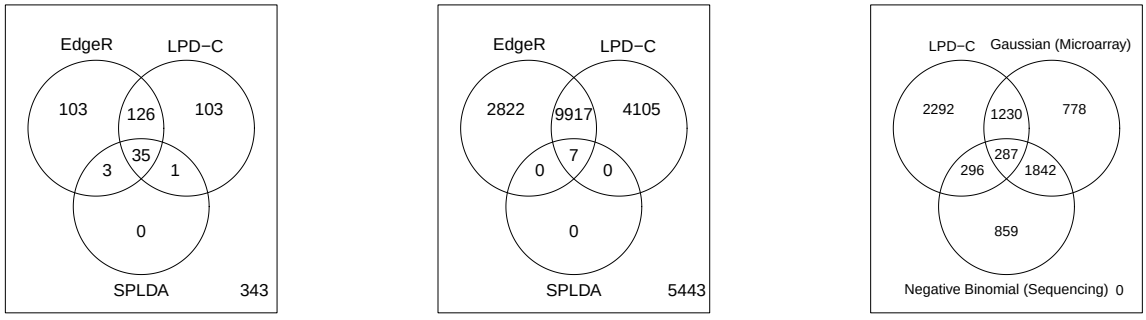
(a) (b)

Figure 4: Plots for determining the number of processes K in LPD-C for cervical cancer data (a) and human data (b) using 5-fold CV. (a) Plot of median false discovery proportion (FDP; on x-axis) versus median true positive proportion (TPP; on y-axis) determined from 5-fold CV when LPD-C is applied to cervical cancer data using $K = 2, \ldots, 10$ (red points), respectively. The vertical lines (grey) show 1 median absolute deviation (MAD) intervals for the TPPs. Based on this plot, 3 and 5 are good candidates for K in cervical cancer data due to their relatively low FDPs and high TPPs. Further data analysis shows that $K = 5$ is a better candidate than $K = 3$. (b) Dot plot with y-axis showing the number of processes and the x-axis showing the medians and 1 MAD intervals of the held-out log likelihoods determined from 5-fold CV when LPD-C is applied to human data using $K = 2, \ldots, 10$; K's are ordered so that the medians of the held-out log likelihoods increase from bottom to top. Based on this plot, we choose $K = 5$ for human data.



(a) (b)

Figure 5: Log variational lower bound (ELBO) for LPD-C as applied to cervical cancer data (a) and to human data (b) with $K = 5$. The x-axes in (a) and (b) represent the iterations in the application of LPD-C, and the y-axes represent the ELBO at each iteration. Both figures show that LPD-C monotonically increases the ELBO at every iteration. The monotonic ascent property of ELBO in variational inference is similar to the ascent property of the log likelihood in EM algorithm, and guarantees convergence of variational updates in (6) to the local mode of the objective function for determining variational posterior densities.



(a) (b) (c)

Figure 6: (a) Comparison of the miRNAs selected by EdgeR, LPD-C, and SPLDA in the cervical cancer data. All the miRNAs selected by SPLDA are selected by EdgeR or LPD-C. The miRNAs selected by LPD-C and EdgeR overlap significantly. (b) Comparison of the genes selected by EdgeR, LPD-C, and SPLDA in the human data. All genes selected by SPLDA are selected by EdgeR, and by LPD-C. The genes selected by EdgeR and LPD-C agree closely. (c) Comparison of the 4105 genes, selected solely by LPD-C in the human data, with the differentially expressed genes found in the microarray and NGS data analyses reported by Marioni et al. (2008). It excludes the 9924 genes that are in the intersection of EdgeR and LPD-C. The microarray and NGS data analysis, respectively, use a Gaussian and a negative binomial model. Approximately 44% of the genes that are selected by LPD-C in the human data are differentially expressed in either the microarray or the NGS data analysis of Marioni et al. (2008).

Table 1: Number of genes, and fraction of differentially expressed genes, selected by LPD-C in the five processes (gene-subsets) for both the cervical cancer data and the human data. The diagonal elements in columns 1-5 represent the total number of genes selected by LPD-C in the respective gene-subsets. The upper off-diagonal elements in columns 1-5 represent the number of genes that are in common between two gene-subsets, while columns 6-10 represent the fraction of differentially expressed genes among the processes as determined by `edgeR` (Robinson et al., 2010).

	Number of Selected Genes					**Differentially Expressed Genes**				
Cervical Cancer Data										
Process	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	5	1	2	0	0	0.60	1	0.5	0	0
2		44	13	4	3		0.59	0.62	0.5	0.33
3			61	8	6			0.59	0.75	0.67
4				103	25				0.67	0.80
5					113					0.62
Human Data										
Process	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	4128	806	935	1150	1773	0.72	0.74	0.72	0.71	0.79
2		3188	695	985	13		0.78	0.78	0.79	0.46
3			4107	469	1163			0.70	0.84	0.79
4				5476	1391				0.68	0.77
5					5090					0.77