Context Aware Group Nearest Shrunken Centroids in Large-Scale Genomic Studies

Juemin Yang
Department of Biostatistics, Johns Hopkins University

Fang Han
Department of Biostatistics, Johns Hopkins University

Rafael A Irizarry
Department of Biostatistics, Harvard University

Han Liu
Operations Research and Financial Engineering, Princeton University

Abstract

Recent genomic studies have identified genes related to specific phenotypes. In addition to marginal association analysis for individual genes, analyzing gene pathways (functionally related sets of genes) may yield additional valuable insights. We have devised an approach to phenotype classification from gene expression profiling. Our method named "group Nearest Shrunken Centroids (gNSC)" is an enhancement of the Nearest Shrunken Centroids (NSC) (Tibshirani, Hastie, Narasimhan and Chu 2002) which is a popular and scalable method to analyze big data. While fully utilizing the variable structure of gene pathways, gNSC shares comparable computational speed as NSC if the group size is small. Comparing with NSC, gNSC improves the power of classification by utilizing the gene pathway information. In practice, we investigate the performance of gNSC on one of the largest microarray datasets aggregated from the internet. We show the effectiveness of our method by comparing the misclassification rate of gNSC with that of NSC. Additionally, we present a novel application of NSC/gNSC on context analysis of association between pathways and certain medical words. Some newest biological findings are rediscovered.

1 Introduction

Recent advances in DNA microarray experiment are generating data sets of the expression levels of large number of genes simultaneously. The aggregation of these data sets across experiments provides better representation of the overall population and contains more information which allows better insights into certain diseases and their causing genes. The aggregated data, however, is often of large scale, high-dimensional (*number of variables* > *number of observations*), with non-Gaussian structure, and thus beyond the ability of typical analysis. This kind of data is so called "big data" (Manyika, Chui, Brown, Bughin, Dobbs, Roxburgh and Byers 2011). It was only recently that people have begun to develop methods of analyzing big data deriving from microarray experiments. One of the aims of such data is to identify a small subset of functional genes which discriminate between certain phenotypes such as the tumor and the normal tissues. Traditional discriminant analysis methods such as linear discriminant analysis (LDA), support vector machine (SVM), and logistic regression are either restricted to relatively small data set or not consistent under the high-dimensional situation. Take the standard LDA as an example. The standard LDA, which uses a linear combination of features as the criterion for classification, has been shown to perform well and enjoy certain optimality as the sample size tends to infinity while the dimension is fixed. In the high-dimensional settings, however, Bickel and Levina (2004) show that the classical LDA is asymptotically equivalent to random guessing when $p/(n_1 + n_2) \to \infty$, even if a Gaussian assumption is made. To handle this problem, known as "the curse of dimensionality," a sparsity condition has to be added, which leads to a variety of works: Cai and Liu (2011) made a sparsity assumption on the precision matrix and proposed a direct estimation method for sparse LDA by estimating $\Omega\delta$ (the product of the precision matrix and the difference of the means) through a constrained l_1 minimization method; Ravikumar, Wainwright and Lafferty (2010) presented a sparse logistic regression method which involves performing l_1-regularized logistic regression of each variable on the remaining variables and then using the sparsity pattern of the regression vector to infer the underlying neighborhood structure; Zhu, Rosset, Hastie and Tibshirani (2004) considered the l_1 norm SVM to accomplish the goal of automatic feature selection in the SVM and Friedman, Hastie, Rosset, Tibshirani and Zhu (2004) shows that the l_1 norm is preferred if the underlying true model is sparse.

Although significant process has been made in this direction, the Nearest Shrunken Centroids (NSC) pro-

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

posed by Tibshirani et al. (2002) is still one of the most scalable method in the field of large-scale data analysis. Comparing with sparse LDA, l_1 norm SVM and sparse logistic regression, the NSC is appealing in the sense that the algorithm is much faster, able to deal with big data, and easy to implement. Moreover good empirical performances have been constantly verified in recent years (Kobayashi, Absher, Gulzar, Young, McKenney, Peehl, Brooks, Myers and Sherlock 2011). Particularly useful is the NSC's ability to simultaneously conduct feature selection and classification via shrinking the marginal centroids. Theoretically speaking, the NSC works the best in a Naive Bayes situation (Fan and Fan 2008), where variables are supposed to be independent of each other; its robustness is such that a high-degree of efficiency is maintained even under more complicated high dimensional models (Fan and Fan 2008).

In this paper, we propose a new high dimensional discriminant analysis method of group Nearest Shrunken Centroids (gNSC). Our new method is one of the earliest attempts to deal with big data of microarray expressions with context information. Also, this is the first paper provide theoretical justification for NSC like methods. Similar with the NSC, gNSC can simultaneously perform sample classification and feature selection. The non-gaussianity of the data is overcome by conducting a normal score transformation in data prepocessing. This has been discovered to work well when the true data are coming from the Nonparanormal (Liu, Han, Yuan, Lafferty and Wasserman 2012) and lose little when the true data are indeed Gaussian. Moreover, in addition to marginal association analysis for individual genes, gNSC enables us to use gene pathway information. Genes work independently and interactively to perform various biological functions. A gene pathway refers to a set of genes that work together to finish a specific biological function. Utilizing the variable structure information of gene pathways may lead to valuable insight into the disease etiology or treatment effect and could inform clinical decisions concerning disease prevention or therapeutic maneuvers. Furthermore, when multiple genes from a same pathway show concerted signals, there may be enhanced the power of sample classification, which is convinced in our experiment. We test the effectiveness of gNSC in analyzing big data against GPL96, a large-scale microarray dataset aggregated from the internet (McCall, Bolstad and Irizarry 2010). Pathway information for the genes is extracted from Molecular Signature Database (MSigDB) (Subramanian, Tamayo, Mootha, Mukherjee, Ebert, Gillette, Paulovich, Pomeroy, Golub, Lander et al. 2005). We compare our gNSC with NSC to show that gNSC improve the power of sample classification by utilizing the pathway information. We also apply gNSC to a context analysis where we combine the sample text information into the GPL96 data (McCall et al. 2010). Our results are consistent with the newest biological finding: the expression of MYC target genes is correlated with B cell lymphomas and Wilms tumor (Ji, Wu, Zhan, Nolan, Koh, De Marzo, Doan, Fan, Cheadle, Fallahi et al. 2011).

We arrange the rest of the paper as follows. In Section 2, sees the introduction of the Nearest Shrunken Centroids (NSC) proposed by Tibshirani et al. (2002) and normal score transformation (Liu, Lafferty and Wasserman 2009). In Section 3, the theoretical body, sees our group Nearest Shrunken Centroids method. We prove some theoretical properties of gNSC. Notably, we prove that (i) under certain regularity conditions, the sparsity pattern can be recovered in an exponential rate; (ii) under certain conditions, we prove that $\mathcal{C}(g) - \mathcal{C}(g^*) = O_P(n^{-1})$, where we denote by $\mathcal{C}(g^*)$ and $\mathcal{C}(g)$ the Bayes risk and the gNSC misclassification rate. We also show the semiparametric efficiency of performing normal score transformation in data preprocessing. In Section 4, we apply both our gNSC method and our context analysis algorithm to the GPL96 microarray dataset.

2 Background

2.1 Nearest Shrunken Centroids (NSC)

Tibshirani et al. (2002) proposed the Nearest Shrunken Centroid method for sample classification in DNA microarray studies. They use shrunken centroids as prototypes for each class and identify subsets of genes that best characterize each class. The NSC shrinks each of the class centroids toward the overall centroid for all classes by a threshold and makes the classifier more accurate by eliminating the effect of noisy genes. As a result it also has an internal gene selection facility (Zou and Hastie 2005). In detail, given x_{ij} for variable j and sample i where $j = 1, .., d$ and $i = 1, .., n$, we have M classes, each with n_m samples. $i \in C_m$ means that the i-th sample is in class m. The NSC utilizes the simple two sample t-test statistic between $\{x_{ij}, i \in C_m\}$ and $\{x_{ij}, i = 1, .., n\}$, and define the classification score d_{mj} as:

$$d_{mj} := \frac{\bar{x}_{mj} - \bar{x}_{\cdot j}}{\eta_m \cdot (s_j + s_0)}, \tag{2.1}$$

where $s_j^2 = \frac{1}{n-M} \sum_m \sum_{i \in C_m} (x_{ij} - \bar{x}_{mj})^2$, $\bar{x}_{mj} = \frac{\sum_{i \in C_m} x_{ij}}{n_m}$, $\bar{x}_{\cdot j} = \frac{\sum_{i=1}^n x_{ij}}{n}$, where $\eta_m := (1/n_m + 1/n)^{-1/2}$ and s_0 is chosen as a global constant to control the variance term. In practice, Tibshirani et al. (2002) suggest setting s_0 equal to the median of the s_j over all genes, i.e., $s_0 := \text{median}\{s_1, \ldots, s_d\}$. Tibshirani et al. (2002) recommend using a soft thresholding function to balance the estimation bias and the model complexity: $\widehat{d}_{mj} = \text{sign}(d_{mj})(|d_{mj}| - \lambda)_+$, where for any $x \in \mathbb{R}$,

$$(x)_+ := \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{otherwise,} \end{cases}$$

and consider j-th variable to be nonfunctional to the m-th class if $\widehat{d}_{mj} = 0$. With regard to classification, given a new sample $\boldsymbol{x}' = (x'_1, \ldots, x'_d)^T$, the discriminant score for class m is defined as:

$$\delta_m(\boldsymbol{x}') = \sum_{j=1}^{d} \frac{(x'_j - \widehat{x}_{mj})^2}{s_j^2} - 2\log(\frac{n_m}{n}),$$

with $\widehat{x}_{mj} = \bar{x}_{\cdot j} + (\eta_m(s_j + s_0))\widehat{d}_{mj}$ and $\widehat{d}_{mj}$ defined in Equation (2.1). λ is accordingly chosen by 10-fold cross validation procedure.

It is further shown in Wang and Zhu (2007) and Hastie, Tibshirani and Friedman (2009) that the Nearest Shrunken Centroids can be explained as a solution to an optimization problem, provided that the data has a certain type of structure. In detail, suppose that $x_{ij} \sim N(\mu_j + \mu_{mj}, \sigma_j^2)$ for $i \in C_m$ with $\sum_{m=1}^{M} \mu_{mj} = 0$ to make the model identifiable, then

$$(\bar{x}_{\cdot j}, \widehat{d}_{mj}) = \underset{\mu_j, \mu_{mj}}{\arg\min} \frac{1}{2} \sum_{j=1}^{d} \sum_{m=1}^{M} \sum_{i \in C_m} \frac{(x_{ij} - \mu_j - \mu_{mj})^2}{s_j^2} + \lambda \sum_{m=1}^{M} \sqrt{n_m} \sum_{j=1}^{d} \frac{|\mu_{mj}|}{s_j}, \tag{2.2}$$

where $\widehat{d}_{mj}$ is shown in Equation (2.1) with $s_0 = 0$ and $\eta_m = \sqrt{1/n_m}$.

2.2 Normal Score Transformation

More recently, the Gaussian assumption commonly adopted by almost all high dimensional discriminant analysis methods is weakened by Liu et al. (2009). They generalize the Gaussian distribution family to a strictly larger Nonparanormal (Gaussian Copula) family. A random variable $X := (X_1, ..., X_d)^T \in \mathbb{R}^d$ belongs to a nonparanormal family if and only if there exist a set of univariate monotone functions $\{f_j\}_{j=1}^d$ such that $(f_1(X_1), ..., f_d(X_d))^T$ is multivariate Gaussian. Liu et al. (2009) utilize the normal score transformation to infer the variable structure, which is proved to be semiparametric efficiency in low dimensional settings by Klaassen and Wellner (1997). Moreover, they analyze its theoretical performance in high dimensional settings (Liu et al. 2012). We refer to their papers for further discussions.

In detail, given n data points $x_1, \ldots, x_n \in \mathbb{R}$, we define

$$\widetilde{F}(t; x_1, \ldots, x_n) := \frac{1}{n+1} \sum_{i=1}^{n} I(x_i \leq t), \tag{2.3}$$

to be the skewed empirical cumulative distribution function. Let $\Phi^{-1}(\cdot)$ be the quantile function of standard Gaussian, we define $\widetilde{f}(t) = \Phi^{-1}\big(\widetilde{F}(t; x_1, \ldots, x_n)\big)$. The normal score transformed data points $\{z_1, \ldots, z_n\}$ are then defined to be:

$$z_i := \widehat{\mu} + \widehat{\sigma} \cdot \widetilde{f}(x_i), \quad i = 1, \ldots, n,$$

where $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\widehat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2}$, are the sample mean and standard deviation.

3 Method

We begin by establishing some notations. Let $\boldsymbol{M} = [M_{jk}] \in \mathbb{R}^{d \times d}$ and $v = (v_1, ..., v_d)^T \in \mathbb{R}^d$. Let v's subvector with entries indexed by I be denoted by v_I, $\boldsymbol{M}$'s submatrix with rows indexed by I and columns indexed by J be denoted by M_{IJ}, $\boldsymbol{M}$'s submatrix with all rows and columns indexed by J is denoted by $\boldsymbol{M}_{\cdot J}$. We define $\|v\|_2 = (\sum_{i=1}^{d} |v_i|^2)^{1/2}$ and $\|v\|_\infty = \max_{1 \leq i \leq d} |v_i|$. We define the matrix $\ell_{\max}$ norm as the elementwise maximum value: $\|\boldsymbol{M}\|_{\max} = \max\{|M_{ij}|\}$ and the ℓ_∞ norm as $\|\boldsymbol{M}\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |M_{ij}|$. $\Lambda_{\min}(\boldsymbol{M})$ and $\Lambda_{\max}(\boldsymbol{M})$ are the smallest and largest eigenvalues of M. We further define the matrix operator norm as $\|M\| = \lambda_{\max}(\boldsymbol{M})$.

3.1 Model

Let $\boldsymbol{X} = [x_{ij}]$ be the dataset we are interested in, with $i = 1, \ldots, n$ and $j = 1, \ldots, d$ representing the n samples and d variables. We assume that there are d variables belonging to K groups, and collect the set of indices of the d_k variables in the k-th group in the set G_k, $k = 1, \ldots, K$. Similarly, we assume that there are n samples belonging to M classes, and that C_m is equal to the set of indices of the n_m samples in the m-th class, $m = 1, \ldots, M$. For simplicity, we rearrange the variables such that $\boldsymbol{X}_{i\cdot} = (\boldsymbol{X}_{iG_1}^T, \ldots, \boldsymbol{X}_{iG_K}^T)^T$.

We consider the data matrix $\boldsymbol{X}$ and denote by: $\boldsymbol{x}_{ik} = (x_{ij}, j \in G_k)^T \in \mathbb{R}^{d_k}$, $\boldsymbol{x}_{ik}^* = (x_{ij} - \bar{x}_{\cdot j}, j \in G_k)^T$. We suppose that for $k = 1, \ldots, K$,

$$\boldsymbol{x}_{ik} \sim^{i.i.d} N(\boldsymbol{\mu}_k + \boldsymbol{\mu}_{mk}, \Sigma_k), \quad \forall\ i \in C_m, \tag{3.1}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\mu}_{mk}$ are both unknown vectors with $\sum_{m=1}^{M} \boldsymbol{\mu}_{mk} = \boldsymbol{0}$, to make the model identifiable.

3.2 Group Nearest Shrunken Centroids

Let $\widetilde{\Sigma}_k$ be an arbitrary estimator of Σ_k. We propose a loss function with a similar version as the NSC's shown in Equation (2.2), but with a group penalty:

$$L := \tfrac{1}{2}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i \in C_m} \|\boldsymbol{x}_{ik} - \boldsymbol{\mu}_k - \boldsymbol{\mu}_{mk}\|_k^2 + \lambda \sum_{k=1}^{K}\sum_{m=1}^{M}(n_m \omega_{mk})\|\boldsymbol{\mu}_{mk}\|_k, \tag{3.2}$$

where for any $v \in \mathbb{R}^{d_k}$, $\|v\|_k$ is defined as:

$$\|\boldsymbol{v}\|_k := (\boldsymbol{v}^T \widetilde{\Sigma}_k^{-1} \boldsymbol{v})^{1/2}. \tag{3.3}$$

The following theorem, whose proof we defer to Appendix A, provides the closed form of the minimizers to Equation (3.2):

Theorem 3.1. *We denote by* $\{\widetilde{\boldsymbol{\mu}}_k\}_{k=1}^K$ *and* $\{\widetilde{\boldsymbol{\mu}}_{mk}\}_{m=1,k=1}^{m=M,k=K}$ *the optima to Equation* (3.2)*:*

$$\{\{\widetilde{\boldsymbol{\mu}}_k\}_{k=1}^K, \{\widetilde{\boldsymbol{\mu}}_{mk}\}_{m=1,k=1}^{m=M,k=K}\} := \arg\min L. \quad (3.4)$$

Then for all $k \in \{1,\ldots,K\}$ *and* $m \in \{1,\ldots,M\}$,

$$\widetilde{\boldsymbol{\mu}}_k = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_{ik} \quad \text{and} \quad \widetilde{\boldsymbol{\mu}}_{mk} = (1 - \frac{\lambda\omega_{mk}}{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k})_+ \widehat{\boldsymbol{\mu}}_{mk}, \quad (3.5)$$

where $\widehat{\boldsymbol{\mu}}_{mk} := \frac{1}{n_m}\sum_{i\in C_m} \boldsymbol{x}^*_{ik}$.

Remark 3.1. *In practice, while defining* $\|\cdot\|_k$ *in Equation* (3.3)*, we adopt a similar idea as Tibshirani et al. (2002) and choose*

$$\widetilde{\Sigma}_k = \widehat{\Sigma}_k + s_0^2 I_{d_k\times d_k}, \quad (3.6)$$

where $\widehat{\Sigma}_k$ *is the sample covariance matrix of the* k*-th group of variables using the whole samples,* $I_{d_k\times d_k}$ *is the* $d_k \times d_k$ *identity matrix and* $s_0^2 = \text{median}(((\text{diag}(\widehat{\Sigma}_1))^T,\ldots,(\text{diag}(\widehat{\Sigma}_K))^T)^T)$*, is the median of all marginal sample variances.*

Given a new data point $x \in \mathbb{R}^d$, the discriminant score for class m is defined as:

$$\delta_m(x) = \sum_{k=1}^K \|x_{G_k} - \widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_{mk}\|_k^2 - 2\log(\frac{n_m}{n}). \quad (3.7)$$

3.3 Theoretical Properties of gNSC

For simplicity, we analyze the theoretical performance of a slightly simpler version of the model proposed in Section 3.1, where for any $k \in \{1,\ldots,K\}$, $\boldsymbol{\mu}_k := \mathbf{0}$. In this way, the estimators in Equation (3.5) can be reduced to:

$$\widetilde{\boldsymbol{\mu}}_k = \mathbf{0} \quad \text{and} \quad \widetilde{\boldsymbol{\mu}}_{mk} = (1 - \frac{\lambda\omega_{mk}}{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k})_+ \widehat{\boldsymbol{\mu}}_{mk}, \quad (3.8)$$

where $\widehat{\boldsymbol{\mu}}_{mk} = \frac{1}{n_m}\sum_{i\in C_m} \boldsymbol{x}_{ik}$ and $\|v\|_k := v^T\widehat{\Sigma}_k^{-1}v$. Remind that $\widehat{\Sigma}_k$ is the sample covariance matrix of the k-th group of variables using the whole samples. Furthermore, to achieve a better theoretical performance, we define the sparse set of $\{\boldsymbol{\mu}_{m1},\ldots,\boldsymbol{\mu}_{mK}\}$ with respect to the sample class C_m to be $S_m := \{k \in \{1,\ldots,K\} : \boldsymbol{\mu}_{mk} \neq \mathbf{0}\}$, and the corresponding estimated sparse set with respect to the m-th sample class to be $\widehat{S}_m := \{k \in \{1,\ldots,K\} : \widetilde{\boldsymbol{\mu}}_{mk} \neq \mathbf{0}\}$.

3.3.1 *Estimation Consistency.*

To achieve the estimation consistency result, we need the following three "boundedness" assumptions:

(A1) There exist two finite constants $c_1, c_2 \in (0,\infty)$ such that
$$c_1 < \min_{1\le m\le M}\min_{k\in S_m}\left\{(\boldsymbol{\mu}_{mk}^T\Sigma_k^{-1}\boldsymbol{\mu}_{mk})^{\frac{1}{2}}, \|\boldsymbol{\mu}_{mk}\|\right\} \le \max_{1\le m\le M}\max_{k\in S_m}\left\{(\boldsymbol{\mu}_{mk}^T\Sigma_k^{-1}\boldsymbol{\mu}_{mk})^{\frac{1}{2}}, \|\boldsymbol{\mu}_{mk}\|\right\} < c_2;$$

(A2) There exists $0 < c_3 = \min\{\Lambda_{\min}(\Sigma_k^{-1}), k = 1,...,K\} < \infty$.

(A3) $\omega_{mk} \propto \sqrt{d_k}$ is upper bounded by $\omega_0 = O((\min_{1\le m\le M} n_m)^{\gamma_0/2})$ for some $0 \le \gamma_0 < 1$.

Theorem 3.2. *(Estimation Consistency) Under assumption (A1)-(A3), for any* $m \in \{1,\ldots,M\}$*, for large enough* n_m*, if we further suppose that* $\lambda \to 0$*,* $n_m^{1/2}\lambda \to \infty$*, and* $\epsilon = \gamma\lambda^2$ *with* $\gamma > \omega_0^2c_2^2/c_1^2$*, we have* $\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk} - \boldsymbol{\mu}_{mk}\|_2^2 > 2\epsilon) = O(\exp(-C_k n_m^2\lambda^4)) \to 0$*, where* $C_k = \min(\frac{3\gamma^2c_3^2}{16d_k}, \frac{3\omega_k^4}{64d_k})$.

3.3.2 *Sparsity Recovery and Misclassification Consistency.*

For sparsity recovery and misclassification consistency, we only consider the situation when there are only two groups of samples, indexed by C_1 and C_2. We denote by y_i the label of i-th sample: $y_i = 0$ if $i \in C_1$ and $y_i = 1$ if $i \in C_2$. Suppose that $(\boldsymbol{X}_{i\cdot}, y_i)$ are i.i.d drawn from the joint distribution of (X, Y), where $X \in \mathbb{R}^d$ and $Y \in \{0,1\}$. The target of the classification is to determine the value of Y given a new data point $\boldsymbol{x} \in \mathbb{R}^d$. Here we further suppose that

$$(X|Y=0) \sim N(\boldsymbol{\mu}_1, \Sigma) \quad \text{and} \quad (X|Y=1) \sim N(\boldsymbol{\mu}_2, \Sigma),$$

where $\boldsymbol{\mu}_1 = (\mu_{11},\ldots,\mu_{1d})^T := (\boldsymbol{\mu}_{11}^T,\ldots,\boldsymbol{\mu}_{1K}^T)^T$, $\boldsymbol{\mu}_2 = (\mu_{21},\ldots,\mu_{2d})^T := (\boldsymbol{\mu}_{21}^T,\ldots,\boldsymbol{\mu}_{2K}^T)^T$, and $\Sigma_{G_kG_k} = \Sigma_k$, $\forall k \in \{1,\ldots,K\}$. Define the prior probabilities $\pi_1 = \mathbb{P}(Y=0)$, $\pi_2 = \mathbb{P}(Y=1)$, and we assume that $\pi_1 = \pi_2$. It is easy to extend it to the case where $\pi_1 \neq \pi_2$ (Hastie et al. 2009). In this way, the Bayes rule is given by:

$$g^*(x) = \begin{cases} 1, & \text{if } \langle \Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1),\ x - \boldsymbol{\mu}_a\rangle > 0 \\ 0, & \text{otherwise} \end{cases} \quad (3.9)$$

where $\langle a, b\rangle$ is the inner product of a and b and $\boldsymbol{\mu}_a := (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$. Define

$$\begin{aligned} \mathcal{S} &:= \{k : \|\boldsymbol{\mu}_{1k} - \boldsymbol{\mu}_{2k}\|_2 > 0\}, \\ \widehat{\mathcal{S}} &:= \{k : \|\widetilde{\boldsymbol{\mu}}_{1k} - \widetilde{\boldsymbol{\mu}}_{2k}\|_2 > 0\}, \end{aligned} \quad (3.10)$$

where $\widetilde{\boldsymbol{\mu}}_1 = (\widetilde{\mu}_{11},\ldots,\widetilde{\mu}_{1d})^T := (\widetilde{\boldsymbol{\mu}}_{11}^T,\ldots,\widetilde{\boldsymbol{\mu}}_{1K}^T)^T$, and $\widetilde{\boldsymbol{\mu}}_2 = (\widetilde{\mu}_{21},\ldots,\widetilde{\mu}_{2d})^T := (\widetilde{\boldsymbol{\mu}}_{21}^T,\ldots,\widetilde{\boldsymbol{\mu}}_{2K}^T)^T$. Remind that $\widetilde{\boldsymbol{\mu}}_{mk}$ is defined in Equation (3.8). Define $s = \text{card}(\mathcal{S})$. It is easy to see that

$$\mathcal{S} \subset S_1 \cup S_2 \quad \text{and} \quad \widehat{\mathcal{S}} = \widehat{S}_1 \cup \widehat{S}_2,\ a.s.. \quad (3.11)$$

Given $\pi_1 = \pi_2$, the gNSC classification procedure g proposed in Equation (3.7):

$$g(x) = \begin{cases} 1, & \text{if} \quad \delta_2(x) - \delta_1(x) > 0, \\ 0, & \text{otherwise}, \end{cases} \quad (3.12)$$

can be further written as:

$$g(x) = \begin{cases} 1, & \text{if} \left\langle \widehat{\Sigma}^{-1}(\widetilde{\boldsymbol{\mu}}_2 - \widetilde{\boldsymbol{\mu}}_1),\ x - \widetilde{\boldsymbol{\mu}}_a \right\rangle > 0, \\ 0, & \text{otherwise}, \end{cases} \quad (3.13)$$

where $\widetilde{\boldsymbol{\mu}}_a = \frac{\widetilde{\boldsymbol{\mu}}_1 + \widetilde{\boldsymbol{\mu}}_2}{2}$, $\widehat{\Sigma}$ and $\widehat{\Sigma}^{-1}$ are defined to be:

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \widehat{\Sigma}_2 & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \cdot & \cdot & \ldots & \cdot \\ \mathbf{0} & \mathbf{0} & \ldots & \widehat{\Sigma}_K \end{pmatrix}. \quad (3.14)$$

The corresponding misclassification errors are

$$\mathcal{C}(g^*) = \bar{\Phi}(\Psi_\Sigma(\Delta, \xi)) \quad \text{and} \quad \mathcal{C}(g) = \bar{\Phi}(\Psi_\Sigma(\widetilde{\Delta}, \widehat{\xi})),$$

where $\bar{\Phi}$ is the survival probability of the standard Gaussian distribution, i.e. $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$, and $\Delta := \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, $\widetilde{\Delta} := \widetilde{\boldsymbol{\mu}}_2 - \widetilde{\boldsymbol{\mu}}_1$, $\xi := \Sigma^{-1}\Delta$, $\widehat{\xi} := \widehat{\Sigma}^{-1}\widetilde{\Delta}$ $\Psi_\Sigma(\boldsymbol{a}, \boldsymbol{b}) := \frac{\boldsymbol{a}^T\boldsymbol{b}}{2\sqrt{\boldsymbol{b}^T\Sigma\boldsymbol{b}}}$.

To obtain a fast rate on the misclassification consistency, we need the following assumption:

(A4) For any $k \in S_1 \cap S_2$, we have $k \in \mathcal{S}$. In other words, $S_1 \cup S_2 = \mathcal{S}$.

The next theorem states that the sparsity pattern can be recovered consistently:

Theorem 3.3. *(Sparsity Recovery) Under assumptions (A1)-(A4), if we further suppose that $\epsilon = \gamma\lambda^2$, $\gamma > c_2^2\omega_0^2/c_1^2$, $\lambda \to 0$ and $\lambda n^{1/2} \to \infty$. Then if $K = o(e^{Cn^2\lambda^4})$, $C > 0$ is a sufficient small constant, we have:*

$$\mathbb{P}(\widehat{S} \neq S) \to 0.$$

Finally, we define $\mathcal{M} := \{j \in \{1, \ldots, d\} : j \in G_k \text{ for some } k \in \mathcal{S}\}$.

Theorem 3.4. *(Misclassification Consistency) Under assumptions (A1)-(A4), if we define $\|(\Sigma_{\mathcal{M}\mathcal{M}})^{-1} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\| := O(a_{n,d})$, where $a_{n,d}$ scales with (n, d) and further suppose $\lambda = O(n^{-\frac{1}{2}}[\log(n)\log(K)]^{\frac{1}{4}})$, then we have*

$$\mathcal{C}(g) - \mathcal{C}(g^*) = O_P(c_s s^2\omega_0^4 \cdot \frac{\log s \log \omega_0}{n}) + c_s O(a_{n,d}^2),$$

where $c_s := \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_2^2$.

As alluded to previously, we defer the proofs of the above theoretical results to Appendix A.

3.4 Nonparanormal and Normal Score Transformation

To the extent that both gNSC and NSC are only well-justified when data are Gaussian, their applications to the "real" data, e.g. gene expression data, is highly limited. To attack this problem, Liu et al. (2009) weaken the Gaussian assumption via introducing the Nonparanormal distribution family. In detail, a random variable $X = (X_1, ..., X_d)^T$ is said to follow a *nonparanormal* distribution if and only if there exist a set of univariate monotone transformations $f = \{f_j\}_{j=1}^d$ such that: $f(X) = (f_1(X_1), ..., f_d(X_d))^T := Z \sim N(\mu, \Sigma)$, where $\mu = (\mu_1, ..., \mu_d)^T$, $\Sigma = [\Sigma_{jk}]$ are the mean and covariance matrix of the Gaussian distribution Z. $\{\sigma_j^2 := \Sigma_{jj}\}_{j=1}^d$ are the corresponding marginal variances. To make the model identifiable, we assume, for $1 \leq j \leq d$, $\mathbb{E}(X_j) = \mathbb{E}(f_j(X_j)) = \mu_j$ and $\mathrm{Var}(X_j) = \mathrm{Var}(f_j(X_j)) = \sigma_j^2$. For notational convenience, we denote such X by $X \sim NPN(\mu, \Sigma, f)$. Liu et al. (2009) prove that if the transformation functions are monotone, the nonparanormal family is equivalent to the Gaussian Copula.

In practice, a parallel model to Equation (3.1) can be constructed: $\boldsymbol{x}_{ik} \sim^{i.i.d.} NPN(\boldsymbol{\mu}_k + \boldsymbol{\mu}_{mk}, \Sigma_k, \boldsymbol{f}_k)$, $\forall i \in C_m$, $k \in \{1, \ldots, K\}$, where $\boldsymbol{f}_k = \{f_k^j\}_{j=1}^{d_k}$ is a set of univariate monotone functions common across difference classes. Under this model, a natural data preprocessing approach is to do normal score transformation first on the data and achieve $\boldsymbol{Z} = [z_{ij}] \in \mathbb{R}^{n\times d}$ such that for all $m \in \{1, \ldots, M\}$: $z_{ij} = \widehat{\mu}_{mj} + \widehat{\sigma}_j \cdot \Phi^{-1}(\widetilde{F}(x_{ij}; \{x_{i'j}\}_{i' \in C_m}))$, $\forall i \in C_m$, where $\widetilde{F}(\cdot;\cdot)$ is defined in Equation (2.3), $\widehat{\mu}_{mj} = \frac{1}{n_m}\sum_{i\in C_m} x_{ij}$ and $\widehat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^n (x_{ij} - \frac{1}{n}\sum x_{ij})^2$. Its theoretical performance has been deeply studied by Klaassen and Wellner (1997) and Bickel (1998). Its theoretical and empirical performance in high dimensional settings has been further verified by Liu et al. (2012). We therefore recommend conducting normal score transformation while preprocessing the data and use $\boldsymbol{Z}$ as the input to the gNSC algorithm. With regard to classification, given a new data point $\boldsymbol{x} \in \mathbb{R}^d$, we transform it to a new data $\boldsymbol{z} = (z_1, \ldots, z_d)^T$ by:

$$z_j = \frac{1}{M}\sum_{i=1}^M (\widehat{\mu}_{mj} + \widehat{\sigma}_j \cdot \Phi^{-1}(\widetilde{F}(x_j; \{x_{i'j}\}_{i' \in C_m}))).$$

We then apply $\boldsymbol{z}$ to Equation (3.7) to obtain the discriminant score for classification.

4 Application

Gene expression is the process by which information encoded *within* a gene is used in the synthesis of a functional gene *product*, such as proteins. After genome sequencing, microarray analysis has become one of the indispensable tools that many biologists use to monitor genome-wide expression levels of genes in a given organism. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously. To the extent that the quantity of data collected from such experiments is overwhelming, efficiently synthesizing these expression levels, via microarray analysis, is an essential part of current methodology. The GPL96 set (Affymetrix GeneChip Human Genome U133 Array Set HG-U133A) (McCall et al. 2010), a collection of publicly available microarray data from hundreds of different experiments, is among the highest accessible microarray datasets. This set includes over 1,000,000 unique oligonucleotide features covering more than 39,000 transcript variants, which in turn represent greater than 33,000 of the best characterized human genes. Sequences were selected from GenBank, dbEST, and RefSeq. Sequence clusters were created from Build 133 of UniGene and refined by analysis and comparison with a number of other publicly available databases including the Washington University EST trace repository and the University of California, Santa Cruz golden-path human genome database.

We will apply our above-discussed technique (cf. Sec-

tion 3), to 20,248 genes and 8,124 microarray samples from Affymetrixs HG-U133A platform. Each sample belong to a certain tissue type (e.g., lung cancers, brain tumor etc.), of which we have 312 types total. We also are interested in certain gene pathways extracted from the one of the largest pathway databases, Molecular Signature Database (MSigDB) (Subramanian et al. 2005). This database consists of 12,713 genes, notably including information concerning biological pathways and responses to a drug treatment. The pathway information is extracted from the MSigDB. A total of 6,769 pathways are obtained. The main purpose of our experiment is to test the association between gene pathways and certain diseases or tissue types. To demonstrate the effectiveness of our new approach for high dimensional discriminant analysis, we look at the performance of both classification and feature selection of group Nearest Shrunken Centroids. The task of sample classification is to classify and predict the diagnostic category of a sample on the basis of its gene expression profile. We show our effectiveness of sample classification by comparing the performance of our new method with Nearest Shrunken Centroids. The misclassification rates of both of the two methods are calculated in Section 4.1. The performance of gNSC on feature selection is shown in Section 4.2, where we apply gNSC on context analysis. We show that the power of feature selection is improved by utilizing the text information.

4.1 gNSC for Classifcation

The raw data of GPL96 contain 20,248 genes and 8,124 samples belonging to 312 tissue types. Since the tissue types with too few samples may not follow the asymptotic properties, we exclude from consideration tissue types with fewer than 30 samples. 5,510 samples, belonging to 24 tissue types, form our data set. To explore the association between the gene pathways and the tissues, we utilize the gene structure information extracted from the Molecular Signature Database (MSigDB), which contains 6,769 pathways and 12,713 genes. To preserve efficiency, we exercise data-rich gene pathways – those with more than 50 genes. We finally have 4,005 pathways, containing 10,990 different genes. Consequently, the final dataset we use contains 10,990 genes belonging to 4,005 pathways, 5,510 samples belonging to 24 tissue types. Finally, we arrange the genes by pathways, giving us a matrix with dimension $5,510 \times 88,396$. gNSC is then applied to this data for detecting the association between gene pathways and tissue types. Note that there are 88,396 columns instead of 10,990. This is because the genes can belong to more than one pathway.

4.1.1 *Procedures.*

In Section 3.3, we have shown that the asymptotic variable selection and misclassification consistency results of gNSC hold under the assumption of normality of the data. Therefore we first test the normality of the dataset. For each gene in each sample class we present the Quantile-to-Quantile plot (QQ plot) to visualize the normality. Three of them are shown in Appendix C. It can be observed that all the three marginal distributions are severely away from Gaussian. Accordingly, we utilize the idea of normal score transformation (NST) to generalize the model to nonparanormal (Liu et al. 2012).

We calculate $\widetilde{\boldsymbol{\mu}}_k$ and $\widetilde{\boldsymbol{\mu}}_{mk}$ by using the Equation (3.5), where $\omega_{mk} = \sqrt{d_k/n_m}$, $\widetilde{\Sigma}_k$ is calculated using Equation (3.6), and λ is a tuning parameter. We use two types of cross validation methods to tune the parameter λ. 10-fold cross-validation is used by (Tibshirani et al. 2002) to find the λ with the lowest average misclassification error. In practice, however, the new data points usually come from a new experiment. We therefore propose an alternative way, "leave experiment out" cross-validation, to select λ. In detail, we isolate all samples from a single experiment as our "testing data;" remaining data is used as "training data." We calculate the discriminant scores of all classes for each data point in the testing data using Equation (3.7). The estimated class for each data point is the one that achieves the lowest discriminant score. Then, for each λ, we calculate an average misclassification error by summing up the number of misclassified data points for all experiments and dividing it by the total number of samples. The parameter λ is chosen to be the one with the lowest average misclassification error.

4.1.2 *Results.*

We say that the pathway k is significantly associated with the sample class m if $\widetilde{\mu}_{mk}$ is greater than 0. We call the combination of one certain pathway and one certain tissue type a block. There are then $M \times K$ blocks.

We compare our method of group Nearest Shrunken Centroids with the Nearest Shrunken Centroids. The averaged misclassification error for each λ from 0.1 to 10 is calculated using both "leave fold out" and "leave experiment out" cross-validation, where "leave fold out" represents the commonly used 10-fold cross validation. The λ with the lowest averaged misclassification error is picked up using these two cross validation methods. The corresponding averaged misclassification errors with their standard deviations are calculated. Moreover, we present the corresponding averaged significant numbers of unique genes across different tissue types with their standard deviations. All the results are illustrated in Table 1. It can be observed that group Nearest Shrunken Centroids has – on average – lower misclassification errors than Nearest Shrunken Centroids, as the gNSC requires much fewer genes to obtain a better prediction result. We also provide two tables to show the general trend of the

averaged misclassification errors and the corresponding genes with increasing of λ in Appendix B.

Table 1: Leave Fold Out v.s. Leave Experiment Out Cross Validation.

		gNSC	sd	NSC	sd
leave experiment out	gene number	3901	22.50	4495	17.26
	error	0.089	0.0505	0.149	0.0479
leave fold out	gene number	5502	21.33	5670	11.14
	error	0.139	0.0866	0.136	0.0906

By using "leave experiment out" cross validation, the tuning parameter with lowest averaged misclassification error is around $\lambda = 6.5$. To illustrate our result more clearly, we randomly pick 12 tissue types and 50 gene pathways for visualization. Figure 1 presents the significant associations, i.e. the threshold term $(1-\frac{\lambda\omega_{mk}}{\|\hat{\boldsymbol{\mu}}_{mk}\|_k})_+$, between gene pathways and tissue types: red color suggests that the corresponding pathway and tissue type are estimated to be associated. The heatplots of the negative "shrinkage amount," i.e. $(1-\frac{\lambda\omega_{mk}}{\|\hat{\boldsymbol{\mu}}_{mk}\|_k})$ and the biological expression levels are shown in Appendix D.

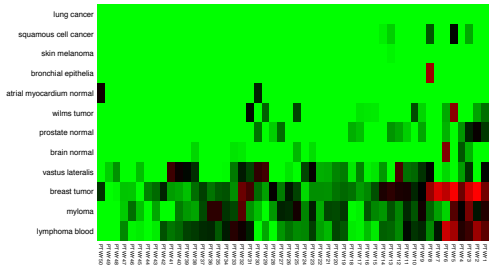


Figure 1: Significant Association Between Pathways and Tissue Types

Using the parameters extracted from our data, we can reclassify the samples associated with different tissue types and compare them with the true labels. The result is shown in Figure 2, with the y-axis as true labels and x-axis as predictive labels. The integer with the

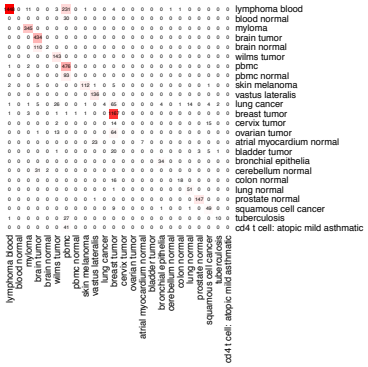


Figure 2: The True Tissue Types v.s. the Predictive Tissue Types.

coordinates (A,B) represents the number of the tissues that is truly B and is predicted to be A. For example, the value 1448 in the left top corner means that there are 1448 samples that are truly acute lymphoma blood and are successfully predicted to be so. We succeed in accurately predicting over 80% of the samples. Moreover, our errors are largely due to similarity in tissue types, which are also hard to differentiate biologically. For example, we fail to differentiate between wilms tumor that is relapse and the wilms tumor that is non-relapse.

We find 3,220 significant relations in all and 174 significant relations based on the tissue types and pathways we present in Figure 5(a). Note that each relation involves one tissue type and one pathway which includes a number of genes. One gene can be involved in several significant relations and one relation involves all the genes in the corresponding pathway. Even for the subset with only 174 significant relations, many have been found to be biologically meaningful. Part of the results are showed in Table 2. For example, "CHESLER BRAIN D6MIT150 QTL CIS" is a Cis-regulatory quantitative trait loci found at the D6Mit150 region. It is believed to regulate the central nervous system. Therefore, this pathway is considered to be highly related to certain brain diseases (Chesler, Lu, Shou, Qu, Gu, Wang, Hsu, Mountz, Baldwin, Langston et al. 2005) and is successfully identified by our techniques.

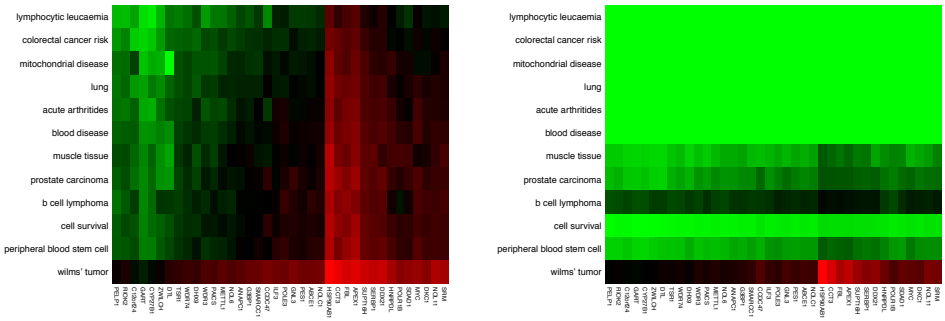
Table 2: True relations learnt from the GPL96 data.

(Pathway Name) Disease Related to the Class
(GAUSSMANN MLL AF4 FUSION TARGETS B DN) leukemia
(OZANNE AP1 TARGETS UP) breast tumor
(GOLUB ALL VS AML DN) blasts and mononuclear cells: leukemia
(CHESLER BRAIN D6MIT150 QTL CIS) barin: glioblastoma
(PODAR RESPONSE TO ADAPHOSTIN DN) breast tumor
(HAHTOLA MYCOSIS FUNGOIDES UP) b cell: lymphoma
(HEDVAT ELF4 TARGETS UP) blasts and mononuclear cells: leukemia
(STEIN ESTROGEN RESPONSE NOT VIA ESRRA) breast tumor
(MACLACHLAN BRCA1 TARGETS DN) breast tumor
(VETTER TARGETS OF PRKCA AND ETS1 DN) breast tumor
(NIKOLSKY BREAST CANCER 22Q13 AMPLICON) breast tumor
(DONATO CELL CYCLE TRETINOIN) breast tumor
(CAFFAREL RESPONSE TO THC 8HR 3 DN) b cell: lymphoma
(SINGH NFE2L2 TARGETS) breast tumor
(WANG RESPONSE TO ANDROGEN UP) prostate tumor
(WAGNER APO2 SENSITIVITY) breast tumor
(DE YY1 TARGETS UP) breast tumor
(SABATES COLORECTAL ADENOMA SIZE UP) breast tumor
(MYLLYKANGAS AMPLIFICATION HOT SPOT 11) breast tumor
(The information used here are from http://www.broadinstitute.org/.)

4.2 Context Analysis of Myc pathway

Myc is a regulator gene that codes for a transcription factor. A mutated version of Myc is found in many cancers. Translocation involving Myc is critical to certain kinds of B-cell lymphoma (Lovec, Grzeschiczek, Kowalski and Möröy 1994). A very recent result obtained by Ji et al. (2011) concludes that microarray samples enriched in Wilms tumor have low Myc. A list of 51 genes are believed to be highly positively correlated with Myc in Myc pathways, 37 of which are included in GPL96. To show the effectiveness of feature selection, we use context analysis to identify most related genes of medical terms including "wilms tumor" and "b-cell lymphoma". Since it is not the main part of this paper, we put the detailed procedure of conducting context analysis in the Appendix G. We will discuss more about it in our future papers.

The 37 positively related genes in Myc pathways are



(a) (b)

Figure 3: gNSC Results of Keywords v.s. Myc pathway. (a) The mean relevance levels of synonym word groups with the 37 genes in Myc pathway; (b) The figure illustrates $\widetilde{\mu}_{mk}$ calculated by Equation (3.5).

believed to be related to both "Wilms tumor" and "b-cell lymphoma". Both NSC and gNSC are used in our analysis. By including the text information, we show that both "Wilms tumor" and "b-cell lymphoma" are predicted to be significantly related to the Myc pathway which coincides with the finding of Ji et al. (2011). For the sake of comparison, we used both NSC and gNSC without including the text information. To the extent that the results of these "control" experiments were insignificant, we conjecture that text information is necessary to the discernative power of context analysis in feature selection. We show the results of context analysis using NSC and gNSC separately below.

4.2.1 *Context Analysis using NSC*

Similar to Section 4.1, we consider only tissue types with more than 30 samples. After further screening out the samples without document information, there remain 5,484 samples for study, leading to an expression matrix with a dimension of $5,484 \times 12,713$. We extract 11,220 meaningful single terms from the text information of GPL96, from which 4,308 terms are included in the 5,484 sample documents we have in the gene expression matrix. The dictionary we use consists of 1,048,576 words and phrases in total. Among them we only need nouns, resulting in 565,308 words and phrases left. Each word and phrase has been indexed to a specific synonym cluster. We exclude those with no terms belonging to any sample document. There are 4,560 synonym clusters left with different indices. In summary, we end up with an index-doc matrix with a dimension of $4,560 \times 5,484$ (see Appendix G for more details). Based on the expression matrix and the index-doc matrix, we can construct the index-gene relevance matrix, which has a dimension of $5,484 \times 4,560 \times 12,713$. We then implement NSC to analyze the associations between synonym clusters and the genes.

Remark 4.1. *Although the dimension of the index-gene relevance matrix is over* 3×10^{11}*, the clean encoding of the data in R allows for efficient analysis of this large-scale information: by calculating sufficient statistics of the original microarray data, we were able to finish our whole procedure in minutes.*

Since it is impossible to obtain all true relations between genes and words , we use a simpler algorithm to choose the amount of shrinkage instead of doing cross validation. We choose λ to be the 95% quantile of $|d_{mj}|$ with $m = 1, 2, \cdots, M$ and $j = 1, 2, \cdots, d$. Here we have $M = 4,560$ and $d = 12,713$. Therefore 5% of the index-gene relations are considered to be significant. This gives us a λ around 0.00348.

To show the effectiveness of our method, we count the number of genes in the list that are significantly related to the word "wilms tumor" and "b-cell lymphoma". All the 37 genes are significant related with "wilms tumor" and 32 of them are significant related to "b-cell lymphoma". Both words are significantly related to Myc. The heatplot of the relevance of certain words, including "wilms tumor" and "b-cell lymphoma", with the 37 genes is shown in Appendix E.

4.2.2 *Context Analysis using gNSC*

We can also include the pathway information into the context analysis and use gNSC to identify the most related pathways of certain words. Similar to Section 4.1, we screen out the gene pathways with more than 50 genes and sort the matrix of expression levels by pathways. We end up with a matrix with a dimension of $5,484 \times 88,396$. As above, we can construct the index-gene relevance matrix, which has a dimension of $5,484 \times 4,560 \times 88,396$. The mean relevance levels in one synonym block of the relevance matrix are defined to be the means of all relevance values restricted to the pertaining block. The result can be visualized in Figure 3(a). Figure 3(b) shows the relevance of certain words and the 37 genes in Myc pathway. The red block shows high relevance of the word and the genes. As we can see, the Myc gene pathway is significantly related to both "wilms tumor" and "b-cell lymphoma".

Acknowledgement

The authors are supported by NSF Grants III-1116730 and NSF III-1332109, NIH R01MH102339, NIH R01GM083084, and NIH R01HG06841, and FDA HHSF223201000072C. Fang is also supported by a fellowship from Google.

References

Bickel, P. (1998), *Efficient and adaptive estimation for semiparametric models,* Springer Verlag.

Bickel, P., and Levina, E. (2004), "Some theory for Fisher's linear discriminant function,naive Bayes', and some alternatives when there are many more variables than observations," *Bernoulli*, 10(6), 989–1010.

Bickel, P., and Levina, E. (2008), "Regularized estimation of large covariance matrices," *The Annals of Statistics*, 36(1), 199–227.

Cai, T., and Liu, W. (2011), "A direct estimation approach to sparse linear discriminant analysis," *Journal of the American Statistical Association*, 106(496), 1566–1577.

Chesler, E., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H., Mountz, J., Baldwin, N., Langston, M. et al. (2005), "Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function," *Nature genetics*, 37(3), 233–242.

Fan, J., and Fan, Y. (2008), "High dimensional classification using features annealed independence rules," *Annals of statistics*, 36(6), 2605–2637.

Friedman, J., Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004), "[Consistency in Boosting]: Discussion," *The Annals of Statistics*, 32(1), 102–107.

Hastie, T., Tibshirani, R., and Friedman, J. J. H. (2009), *The elements of statistical learning,* Springer.

Ji, H., Wu, G., Zhan, X., Nolan, A., Koh, C., De Marzo, A., Doan, H., Fan, J., Cheadle, C., Fallahi, M. et al. (2011), "Cell-Type Independent MYC Target Genes Reveal a Primordial Signature Involved in Biomass Accumulation," *PloS one*, 6(10), e26057.

Klaassen, C., and Wellner, J. (1997), "Efficient estimation in the bivariate normal copula model: normal margins are least favourable," *Bernoulli*, 3(1), 55–77.

Kobayashi, Y., Absher, D., Gulzar, Z., Young, S., McKenney, J., Peehl, D., Brooks, J., Myers, R., and Sherlock, G. (2011), "DNA methylation profiling reveals novel biomarkers and important roles for DNA methyltransferases in prostate cancer," *Genome research*, 21(7), 1017–1027.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012), "High Dimensional Semiparametric Gaussian Copula Graphical Models," *The Annals of Statistics*, 40(4), 1935–2357.

Liu, H., Lafferty, J., and Wasserman, L. (2009), "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs," *The Journal of Machine Learning Research*, 10, 2295–2328.

Lovec, H., Grzeschiczek, A., Kowalski, M., and Möröy, T. (1994), "Cyclin D1/bcl-1 cooperates with myc genes in the generation of B-cell lymphoma in transgenic mice.," *The EMBO journal*, 13(15), 3487.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. (2011), *Big data: The next frontier for innovation, competition and productivity,* McKinsey Global Institute.

McCall, M., Bolstad, B., and Irizarry, R. (2010), "Frozen robust multiarray analysis (fRMA)," *Biostatistics*, 11(2), 242–253.

Ravikumar, P., Wainwright, M., and Lafferty, J. (2010), "High-dimensional Ising model selection using ?1-regularized logistic regression," *The Annals of Statistics*, 38(3), 1287–1319.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. 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(43), 15545.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002), "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences*, 99(10), 6567—6572.

Wang, S., and Zhu, J. (2007), "Improved centroids estimation for the nearest shrunken centroid classifier," *Bioinformatics*, 23(8), 972–979.

Wu, H., Luk, R., Wong, K., and Kwok, K. (2008), "Interpreting TF-IDF term weights as making relevance decisions," *ACM Transactions on Information Systems (TOIS)*, 26(3), 13.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2004), "1-norm support vector machines," *Advances in neural information processing systems*, 16(1), 49–56.

Zou, H., and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.