Published in final edited form as:
Comput Biol Chem. 2012 August ; 39: 14–19. doi:10.1016/j.compbiolchem.2012.06.001.

Sparse regularized discriminant analysis with application to microarrays

Ran Li and **Baolin Wu***
Division of Biostatistics, School of Public Health University of Minnesota, Minneapolis, MN 55455, USA

Abstract

For cancer prediction using large-scale gene expression data, it often helps to incorporate gene interactions in the model. However it is not straightforward to simultaneously select important genes while modeling gene interactions. Some heuristic approaches have been proposed in the literature. In this paper, we study a unified modeling approach based on the ℓ_1 penalized likelihood estimation that can simultaneously select important genes and model gene interactions. We will illustrate its competitive performance through simulation studies and applications to public microarray data.

1 Introduction

One of the main research questions in microarray data analysis is cancer prediction using gene expressions. Typical microarray data often has small sample size compared to the large number of genes, which called for development of special statistical methods. Many prediction methods have been proposed. Some are based on directly modeling the cancer status using gene expressions, e.g., logistic regression based models (see Friedman *et al.*, 2010, e.g.) and commonly used machine learning methods including SVM and classification tree based methods (see Vapnik, 1998; Brown *et al.*, 2000; Breiman, 2001, e.g.) etc. While others are based on modeling gene expressions conditional on the cancer status to indirectly produce prediction models, e.g., linear or quadratic discriminant analyses based methods (see Tibshirani *et al.*, 2003; Guo *et al.*, 2007, e.g.), which are often computationally efficient and have very good model interpretability.

One key issue in the conditional modeling approach is the modeling of gene interactions. Some approaches completely ignored the dependence among genes and built a simple model. The intuition is that simple model is very stable and could provide better prediction especially for relatively small sample size microarray data. Selecting important genes for improved prediction is straightforward for independence based models and has been well studied. For example, Tibshirani *et al.* (2003) proposed the use of soft-thresholding for simultaneous gene selection and prediction model estimation, which can be linked to the lasso penalized linear regression model (Tibshirani, 1996; Wu, 2006). These independence prediction models often performed well in practice and were widely used. While others tried to model the complete dependence among genes by using, e.g., regularized covariance or

*To whom correspondence should be addressed. baolin@umn.edu. Phone: 6126240647. Fax: 6126260660.

precision matrix (see Guo *et al.*, 2007; Witten and Tibshirani, 2008, e.g.). When gene dependence is modeled, how to simultaneously select important genes is not straightforward. Often some two-step approaches are used: important genes are selected individually before interaction modeling; or estimated gene covariance matrix is first used to obtain prediction coefficients for all genes, which are then subject to soft-thresholding for gene selection as if they are independent. These dependence prediction models have shown some improvement over independence models for analyzing large-scale gene expression data. In this paper we study a unified dependence prediction model that concisely models gene interactions and simultaneously selects important genes. We develop very efficient computational algorithms and illustrate its competitive performance through simulation and application studies.

In Section 2, we discuss the proposed method. And we develop efficient computational algorithms in Section 3. Simulation studies are provided in Section 4 to study the performance of the proposed method. We analyze two microarray data for illustration in Section 5. Concluding remarks are provided in Section 6.

2 Sparse regularized discriminant analysis

Consider a two-class gene expression data. Denote the class indicator as $Y \in \{0, 1\}$, and expressions of m genes as X. When assuming the expressions of m genes follow the multivariate normal distribution $X|Y \sim N(\mu_0 + Y(\mu_1 - \mu_0), \mathbf{\Sigma})$, we can check that

$$\log\frac{\Pr(Y=1|X)/\Pr(Y=1)}{\Pr(Y=0|X)/\Pr(Y=0)}=(X-\frac{\mu_1+\mu_0}{2})^T\sum^{-1}(\mu_1-\mu_0)=\eta. \quad (1)$$

Therefore

$$\Pr(Y=1|X)=\frac{1}{1+\Pr(Y=0)/\Pr(Y=1)\exp(-\eta)}.$$

Given the gene expressions of n observed samples, the prediction coefficients $\mathbf{\Sigma}^{-1}(\mu_1 - \mu_0)$ can be estimated, for example, based on some estimated covariance matrix $\hat{\mathbf{\Sigma}}^{-1}$ (more details later) and sample averages $(\hat{\mu}_0, \hat{\mu}_1)$ of class 0 and 1

$$\Delta=\widehat{\sum}^{-1}Z, \quad Z=\widehat{\mu}_1-\widehat{\mu}_0.$$

To build a prediction model based on only a subset of important genes, there are two widely used variable selection approaches. In the first approach, gene dependence is ignored (i.e., assuming a diagonal matrix for $\mathbf{\Sigma}$). And variable selection is applied to individual genes separately (see Tibshirani *et al.*, 2003, e.g.). In the second approach, gene dependence is modeled by a non-diagonal matrix $\mathbf{\Sigma}$. And variable selection is directly applied to Δ. For example, Guo *et al.* (2007) proposed treating each component of Δ independently and shrinking them separately based on soft-thresholding. Note that due to gene dependence, in general Δ are not independent, and their covariance matrix can be roughly approximated by

$$\widehat{\sum}^{-1}\widehat{\mathrm{Cov}}(\widehat{\mu}_1-\widehat{\mu}_0)\widehat{\sum}^{-1}=\frac{n}{n_0n_1}\widehat{\sum}^{-1},$$

where n_i are the sample sizes for class 0/1, and $n = n_0 + n_1$. To select important genes and incorporate gene dependence simultaneously, we propose an estimating equation model (modeling gene dependence) with the lasso penalty that can automatically select important genes (Tibshirani, 1996, 2011)

$$\min_{\beta}\frac{n_0n_1}{2n}(\Delta-\beta)^T\widehat{\sum}(\Delta-\beta)+\lambda\sum_{j=1}^{m}\tau_j|\beta_j|, \tag{2}$$

where τ_j is some pre-selected positive weight (typically set as 1). Note that we can also derive model (2) from the perspective of maximum penalized log likelihood estimation (see Appendix for details).

Gene covariance matrix estimation based on the large scale expression data with small sample size has been a very challenging problem. Typically some type of regularization is often required to obtain a nonsingular covariance matrix estimate. Variety of methods have been proposed and studied in the statistics and machine learning field (e.g., Hansen *et al.*, 1992, Zou *et al.*, 2006, Bickel *et al.*, 2008, Jenatton *et al.*, 2010). Many of them are motivated by and developed for estimating a sparse covariance or precision matrix, and typically involve very intensive computations. Another commonly used approach is to use a weighted average of sample covariance matrix and some diagonal matrix as studied at Guo *et al.* (2007). In this paper, our main purpose is to incorporate gene dependence through some covariance matrix into the prediction model. We adopt a principal component analysis (PCA) based approach proposed by Tipping and Bishop (1999) mainly for its computation efficiency. In addition it also leads to very efficient computational algorithms for the proposed prediction model as described in the next section. Specifically it assumed a structured covariance matrix that can be derived from a mixed model representation

$$\mathbf{R}\mathbf{R}^T+\sigma_0^2\mathbf{I}_m=\mathrm{Cov}(\mathbf{R}u+\varepsilon),\quad u\sim N(0,\mathbf{I}_q),\quad \varepsilon\sim N(0,\sigma_0^2\mathbf{I}_m),$$

where the random effects design matrix $\mathbf{R}$ is of dimension $m \times q$ and constrained to be column independent, $R_j^TR_k=0,\forall j\neq k$. Tipping and Bishop (1999) derived a very efficient algorithm for maximum likelihood estimation of $\mathbf{R}$ and σ_0^2 based on the eigen value decomposition of the sample covariance matrix. Let $\tilde{\mathbf{X}}$ be the observed gene expressions of dimension $n \times m$ with class mean subtracted from each gene. Denote its singular value decomposition as

$$\frac{\tilde{\mathbf{X}}}{\sqrt{n-2}}=\mathbf{U}_0\mathbf{E}_0\mathbf{V}_0^T,$$

where $\mathbf{V}_0$ is an $m \times (n - 2)$ sub-orthogonal matrix, $\mathbf{V}_0^T\mathbf{V}_0=\mathbf{I}_{n-2}$, and $\mathbf{E}_0$ is a diagonal matrix of dimension $n - 2$, $\mathbf{E}_0 = \mathrm{diag}(e_1, \cdots, e_{n-2})$ with $|e_1| \geq \cdots \geq |e_{n-2}|$ The maximum likelihood estimate of q-dimensional PCA covariance matrix is then

$$\widehat{\sum}=\mathbf{V}\mathbf{D}\mathbf{V}^T+\sigma_0^2\mathbf{I}_q=\sum_{j=1}^{q}d_j v_j v_j^T+\sigma_0^2\mathbf{I}_q,\quad \mathbf{D}=\mathrm{diag}(d_1,\ldots,d_q),$$

where $\mathbf{V}$ consists of the first q columns of $\mathbf{V}_0$, v_j is the j-th column of $\mathbf{V}_0$, and

$$\sigma_0^2=\frac{\sum_{j=q+1}^{n-2}e_j^2}{m-q},\quad d_j=e_j^2-\sigma_0^2,\quad j=1,\cdots,q.$$

We can further check that

$$\widehat{\sum}^{-1}=\mathbf{V}\mathbf{D}_1\mathbf{V}^T+\sigma_0^{-2}\mathbf{I}_q,\quad \mathbf{D}_1=\mathrm{diag}\left\{\frac{-d_j}{\sigma_0^2(d_j+\sigma_0^2)}\right\}_{j=1}^{q}.$$

In a closely related paper, Witten and Tibshirani (2008) proposed estimating gene covariance matrix by penalizing, e.g., the ℓ_2 norm of inverse covariance matrix $\mathbf{\Sigma}^{-1}$, and built prediction model by selecting important genes before covariance modeling or, penalizing the ℓ_1 norm of $\mathbf{\Sigma}^{-1}\mu_0$ and $\mathbf{\Sigma}^{-1}\mu_1$ for sample prediction.

In the following section, we develop very efficient computational algorithms for estimating the proposed model based on the iterative coordinate descent approach (Friedman *et al.*, 2010).

3 Computational algorithm for efficient model estimation

We consider minimizing a more general ℓ_1 penalized estimating equation model as follows

$$\ell=\tfrac{1}{2}(\Delta-\beta)^T(\mathbf{W}^T\mathbf{W}+\mathbf{U})(\Delta-\beta)+\sum_{j=1}^{m}\lambda\tau_j|\beta_j|,$$
$$=\tfrac{1}{2}\sum_{i=1}^{q}(s_i-W_i^T\beta)^2+\tfrac{1}{2}\sum_{j=1}^{m}u_j(\delta_j-\beta_j)^2+\sum_{j=1}^{m}\lambda\tau_j|\beta_j|,\quad s_i=W_i^T\Delta,$$

where $\Delta=(\delta_1,\cdots,\delta_m)^T$, $\mathbf{W}$ is an $q\times m$ matrix and $\mathbf{U}=\mathrm{diag}(u_1,\cdots,u_m)$. We have

$$\frac{\partial\ell}{\partial\beta_j}=-\sum_{i=1}^{q}w_{ij}(s_i-W_i^T\beta)-u_j(\delta_j-\beta_j)+\lambda\tau_j\mathrm{sign}(\beta_j).$$

Therefore in the iterative coordinate descent update, we have

$$\beta_j\leftarrow\frac{L_1(\sum_{i=1}^{q}w_{ij}r_i+u_j\delta_j+\beta_j\sum_{i=1}^{q}w_{ij}^2;\lambda\tau_j)}{\sum_{i=1}^{q}w_{ij}^2+u_j},\quad r_i=s_i-W_i^T\beta,$$

where L_1 is the lasso soft-thresholding operator (Tibshirani, 1996). In the update formula, u_j and $\sum_{i=1}^{q} w_{ij}^2$ are fixed and can be pre-computed. Whenever β_j is changed, we update the residual vector r_i of length q

$$r_i \leftarrow r_i - w_{ij}(\beta_j^{(1)} - \beta_j^{(0)}), \quad i = 1, \cdots, q,$$

where $\beta_j^{(0)}/\beta_j^{(1)}$ are the previous/current estimates respectively. Computationally this is equivalent to a lasso regression with q observations and m covariates, which can be solved very efficiently based on the previous iterative coordinate descent approach (q is typically much smaller than sample size n).

Specifically for the previous model, we have

$$\mathbf{W} = \sqrt{\mathbf{D}}\mathbf{V}^T, \mathbf{U} = \sigma_0^2\mathbf{I}_q, \quad \mathbf{S} = \mathbf{W}\Delta = \mathbf{W}\widehat{\sum}^{-1} Z = \sqrt{\mathbf{D}}(\mathbf{D}_1 + \sigma_0^{-2}\mathbf{I}_q)\mathbf{V}^T Z.$$

With some large λ, all β_j will be shrunken down to zero with the gradient vector being

$$-\mathbf{W}^T\mathbf{S} - \mathbf{U}\Delta = -(\mathbf{W}^T\mathbf{W} + \mathbf{U})\Delta = -Z.$$

Therefore we have for any $\lambda \geq \lambda_0$, $\hat{\beta}_j = 0$, $\forall j$, where $\lambda_0 = \max_j |Z_j/\tau_j|$ When estimating models for a sequence of λ, we use the solution from larger λ as a warm starting point for the next smaller λ, which often can improve the algorithm convergence and dramatically reduce the computation time. For the proposed algorithm, we empirically find that the computation time scales linearly with the sample size n and number of genes m, and exponentially with the covariance matrix dimension q (see the Application section for more details).

In the next section, we conduct simulation studies to compare the performance of the proposed method (denoted as SRDA for `sparse regularized discriminant analysis') to two representative methods: the 'nearest shrunken centroid' classifier proposed by Tibshirani *et al.* (2003) (denoted as NSC) assuming gene independence and the `shrunken centroid regularized discriminant analysis' method proposed by Guo *et al.* (2007) (denoted as SCRDA) modeling gene dependence with a regularized sample covariance matrix.

4 Simulation study

We simulated the expressions of $m = 5000$ genes for 25 case and 25 control samples with 10% of genes being differentially associated with the sample class. We simulated those nonzero δ_j from a normal distribution N(0,0.3). The covariance matrix $\mathbf{\Sigma} = \mathbf{DRD}$ has a compound symmetry correlation matrix $\mathbf{R}$ with correlation parameter ρ and variance $\mathbf{D} = \text{diag}\{\sigma_1, \cdots, \sigma_m\}$. The variance of each gene σ_j^2 was simulated from the chi-square distribution with 3 degrees of freedom. Then we obtain the effect size by multiplying $\mathbf{\Sigma}$ and $\Delta = (\delta_1, \cdots, \delta_m)^T$.

We used 5-fold cross validation to select an optimal λ based on validation error rate from 100 evenly spaced points from $\lambda_{\max}$ to 0 on the log scale, where $\lambda_{\max}$ is the value of λ that

would penalize all β_j to zero. The optimal value of q was chosen from (1, 11, 21, 30, 40, 50). 1000 case and 1000 control samples were simulated separately to compute prediction errors. We conducted simulations for ρ = (0.25, 0.5, 0.75) representing weak to strong gene interactions. The average prediction errors over 30 simulations are shown in Table 1. Overall we can see that the proposed SRDA has competitive performance. By appropriately modeling the gene dependence, the proposed SRDA could genuinely benefit from commonly observed gene interactions: it shows better predictions for stronger gene dependence. NSC in general is adversely affected by strong gene dependence due to its independence assumption. SCRDA improved upon NSC by partially incorporating gene interactions. However it shows increased prediction errors with increasing gene dependence partly due to its ad hoc gene interaction modeling.

In the next section we analyze two microarray data with relatively weak and strong gene interactions respectively to illustrate the performance of the proposed method.

5 Application to public microarray data

We analyze the breast cancer (West *et al.*, 2001) and prostate cancer (Singh *et al.*, 2002) microarray data. The breast cancer data consists of 49 tumor samples with genes measured using the Affymetrix hu6800 genechip (including around 7000 probes). The tumors are divided into two groups based on estrogen receptor status: 24 positive and 25 negative. The prostate cancer data consists of 50 normal and 52 tumor prostate tissue samples with genes measured using the Affymetrix hgu95av2 genechip (including around 12000 probes). For each data, we used all genes for sample prediction. In addition, we also tried sample prediction only using those genes with annotated molecular functions in the Gene Ontology (GO, Ashburner *et al.*, 2000), which leads to around 3500 and 5500 probes respectively for the breast cancer and prostate cancer data. This gives us a range of medium to large-scale prediction problems. We compared the performance of the proposed SRDA to SCRDA and NSC.

Figure 1 shows the histogram of all gene pairwise correlations for the breast cancer and prostate cancer microarray data. Similar patterns are observed for pairwise correlations only from those genes with annotated molecular functions in GO (data not shown). The prostate cancer data has very strong gene interactions with many nonzero correlations. For the breast cancer data, the gene dependence is relatively weak with many correlations clustering around zero. Let $\{r_i\}_{i=1}^K$ be the computed gene pairwise correlations. Assuming normal distribution for gene expressions, we can check that $t_i = \sqrt{n-3}\, r_i / \sqrt{1-r_i^2}$ follows the t-distribution with $n - 3$ degrees of freedom, where n is the total sample size (see Kutner *et al.*, 2004, e.g.). We apply the normal transformation $z_i = \Phi^{-1}(T_{n-3}(t_i))$, where $\Phi = T_{n-3}$ are the normal and t-distribution functions. When correlation is truly zero, z_i follows the standard normal distribution. However the computed pairwise gene correlations are strongly correlated with each other, the standard normal distribution might not fit the null data very well. Since majority of the correlations are close to zero for the breast cancer microarray data, we can fit an empirical null distribution (Efron, 2007). The estimated nonzero correlation proportion is 4% for breast cancer data. For prostate cancer data, majority of the correlations are not zero and the empirical null estimation approach does not work well. Instead we roughly estimate the nonzero correlation proportion as follows. We compute the t-statistic t_i and corresponding p-value p_i based on the t-distribution with 99 degrees of freedom. We then estimate the nonzero correlation proportion as $1-2\sum_{i=1}^K I(p_i>0.5)/K$, which is 85% for prostate cancer data. For each gene, we then count the number of its significant correlations with other genes that are ranked in the top 4%/85% for the breast and

prostate cancer data. The number of significant correlations ranges from 0 to 1500 for the breast cancer data and 1000 to 11000 for the prostate cancer data. Figure 2 shows the histogram of the number of significant gene correlations. For the breast cancer data, majority of genes have very small number of interactions with other genes, while for the prostate cancer data, majority of genes have interactions with many other genes.

We use cross validation to estimate the classification errors for three methods. Specifically each time we randomly select one third of the samples as a testing set and apply 5-fold cross validation on the remaining samples to build the prediction model, which is then used to predict the testing set. For all methods, we use the same testing set and repeat the cross validation 50 times to estimate the average classification errors.

Table 2 summarizes the cross validation results. For each data, we reported the classification errors for using all genes (denoted as “all”) and only using those genes with annotated molecular functions in GO (denoted as “GO.mf”). In addition, we also reported the computed significance values for testing the proposed method performed better than the other two methods using the paired sign test based on the binomial distribution. Prostate cancer data has very strong gene dependence. Both SRDA and SCRDA explicitly incorporate gene interactions in the model and have much smaller prediction error than NSC that treated genes independently. Breast cancer data shows relatively weak gene dependence, and we observed smaller difference between SRDA/SCRDA and NSC. Overall SCRDA performed well especially under strong gene interactions. And the proposed SRDA performed favorably compared to SCRDA and NSC.

In addition we also empirically evaluate the computing time of the proposed algorithm based on the prostate cancer data. We implemented the algorithm in C that is called from R, and timed all computations using the R ‘system.time’ function in a Linux workstation with 3 GHz Intel CPU and 8 GB memory. Firstly using all 102 samples and 12000 genes, we evaluate the time used to compute all solutions for 100 equally spaced λ values as a function of the covariance matrix dimension q. The total computing time ranges from 1 to 36 seconds for $q = 9$ to 99. The first plot in figure 3 shows the computing time on the log scale. The superimposed dashed gray line is the fitted linear model of log time versus q. Using all genes, we then randomly sample N prostate samples, $3 \times q = N - 3$, and evaluate the time used to compute all solutions for 100 equally spaced λ values. The total computing time ranges from 10 to 35 seconds for $N = 20$ to 100. The second plot in figure 3 shows the computing time (on the original scale). The superimposed dashed gray line is the fitted linear model of time versus N. Finally using all 102 samples, we randomly sample M genes, and evaluate the time used to compute all solutions for 100 equally spaced λ values. The total computing time ranges from 6 to 30 seconds for $M = 2000$ to 10000. The third plot in figure 3 shows the computing time (on the original scale). The superimposed dashed gray line is the fitted linear model of time versus M. We have repeated the random sampling many times and similar patterns have been observed. Overall we can see that the computation time scales linearly with the sample size and number of genes, and exponentially with the covariance matrix dimension.

6 Discussion

For large-scale gene expression data, how to model gene interactions and simultaneously select important genes pose unique challenges for prediction model building. When treating genes independently, we can apply soft-thresholding to estimate prediction model and select important genes simultaneously. However when modeling gene interactions, selecting important genes has been often performed separately after prediction model estimation. In this paper, we propose modeling gene interactions concisely with a PCA structured

covariance matrix, which provides a principled approach to general large-scale dependence modeling. We adopt the lasso penalized likelihood estimation approach for unified prediction model estimation and important gene selection. Very efficient computational algorithms are also developed based on the iterative coordinate descent approach. The proposed method performed competitively in our simulation and application studies. Currently we have included all pairwise gene interactions in the proposed prediction model. Potentially we could improve the prediction by simplifying the gene covariance structure using some prior information. For example, public gene pathway or annotations often contain curated biological knowledge regarding gene molecular function and interaction information, which could be incorporated into statistical models for improved inference. We are currently exploring this approach and will report the results elsewhere in the future. In the current approach, we have adopted the PCA based covariance matrix for gene dependence modeling mainly for its computation efficiency. Genes in a system often exhibit some sparse interaction patterns, and it will be worthwhile to empirically compare the relative prediction performance of incorporating various regularized/sparse covariance/precision matrix estimation approaches into the proposed method through extensive simulation and application studies. We will report the results elsewhere in the future.

Acknowledgments

This research was supported in part by NIH grant GM083345 and CA134848. We would like to thank two anonymous referees for their constructive comments that have dramatically improved the presentation of the paper.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. Nat Genet. 2000; 25 (1):25–29. [PubMed: 10802651]

Bickel PJ, Levina E. Regularized estimation of large covariance matrices. The Annals of Statistics. 2008; 36 (1):199–227.

Breiman L. Random forests. Machine Learning. 2001; 45:5–32.

Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS. 2000; 97 (1):262–267. [PubMed: 10618406]

Efron B. Correlation and Large-Scale Simultaneous Significance Testing. Journal of the American Statistical Association. 2007; 102:93–103.

Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010; 33:1. [PubMed: 20808728]

Guo Y, Hastie T, Tibshirani R. Regularized linear discriminant analysis and its application in microarrays. Biostat. 2007; 8 (1):86–100.

Hansen PC, Sekii T, Shibahashi H. The modified truncated SVD method for regularization in general form. SIAM Journal on Scientific and Statistical Computing. 1992; 13 (5):1142–1150.

Jenatton, R.; Obozinski, G.; Bach, F. Structured sparse principal component analysis. Proc. 13th Int. Conf. Artificial Intelligence and Statistics; 2010.

Kutner, M.; Nachtsheim, C.; Neter, J. Applied Linear Regression Models. 4. McGraw-Hill/Irwin; 2004.

Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell. 2002; 1:203–209. [PubMed: 12086878]

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B. 1996; 58:267–288.

Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective (with discussion). Journal of the Royal Statistical Society, Series B. 2011; 73 (3):273–282.

Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with application to DNA microarrays. Statistical Science. 2003; 18:104–117.

Tipping ME, Bishop CM. Probabilistics Principle Component Analysis. Journal of the Royal Statistical Society, Series B. 1999; 61 (3):611–622.

Vapnik, VN. Statistical Learning Theory. New York, NY: Wiley-Interscience; 1998.

West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson John AJ, Marks JR, Nevins JR. Predicting the clinical status of human breast cancer by using gene expression profiles. PNAS. 2001; 98:11462–11467. [PubMed: 11562467]

Witten DM, Tibshirani R. Covariance-regularized regression and classification for high-dimensional problems. Journal of the Royal Statistical Society, Series B. 2008; 71 (3):615–636.

Wu B. Differential gene expression detection and sample classification using penalized linear regression models. Bioinformatics. 2006; 22:472–476. [PubMed: 16352654]

Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. Journal of Computational and Graphical Statistics. 2006; 15 (2):265–286.

Appendix

Maximum penalized log likelihood estimation

Denote x_{ij} as the expressions of sample $i = 1, \cdots, n$ and gene $j = 1, \cdots, m$. Let $y_i \in \{0, 1\}$ be the sample class. Denote the prediction coefficients as $\beta = \Sigma^{-1}(\mu_1 - \mu_0)$, and hence we have $\mu_1 = \mu_0 + \Sigma\beta$. The penalized log likelihood is proportional to

$$-\frac{1}{2}\sum_{i=1}^{n}(x_{ij}-\mu_0-y_i\sum\beta)^T\sum^{-1}(x_{ij}-\mu_0-y_i\sum\beta)-\lambda\sum_{j=1}^{m}\tau_j|\beta_j|,$$

which can be checked proportional to

$$-\frac{n_0}{2}(\bar{x}_{j0}-\mu_0)^T\sum^{-1}(\bar{x}_{j0}-\mu_0)-\frac{n_1}{2}(\bar{x}_{j1}-\mu_0-\sum\beta)^T\sum^{-1}(\bar{x}_{j1}-\mu_0-\sum\beta)-\lambda\sum_{j=1}^{m}\tau_j|\beta_j|,$$

where $n_0=\sum_{i=1}^{n}(1-y_i)$, $n_1 = n - n_0$ and

$$\bar{x}_{j0}=\frac{\sum_{i=1}^{n}x_{ij}(1-y_i)}{n_0},\quad \bar{x}_{j1}=\frac{\sum_{i=1}^{n}x_{ij}y_i}{n_1}.$$

We can easily check that

$$\widehat{\mu_0}=\frac{n_0}{n}\bar{x}_{j0}+\frac{n_1}{n}(\bar{x}_{j1}-\sum\beta),$$

and maximizing penalized log likelihood is equivalent to

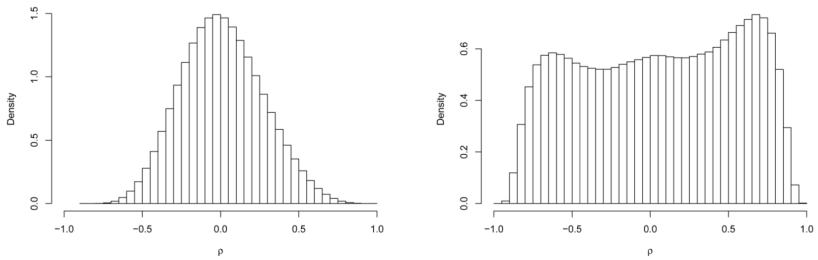
$$\min_{\beta} \frac{n_0 n_1}{2n} (\Delta - \beta)^T \sum (\Delta - \beta) + \lambda \sum_{j=1}^{m} \tau_j |\beta_j|,$$

where

$$\Delta = \sum^{-1} (\bar{x}_{j1} - \bar{x}_{j0}).$$

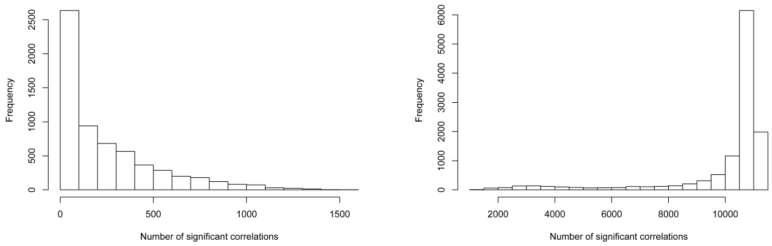
Highlights

1. A unified modeling approach that simultaenously analyzes gene interactions and selects important genes for improved prediction of microarrays.
2. Very efficient computational algorithms developed for model estimation and selection.
3. demonstrate the very competitive performance of the proposed method.



(a) Breast cancer microarray data (b) Prostate cancer microarray data

Figure 1.
Pairwise gene correlations



(a) Breast cancer microarray data (b) Prostate cancer microarray data

Figure 2.
Number of significant gene correlations

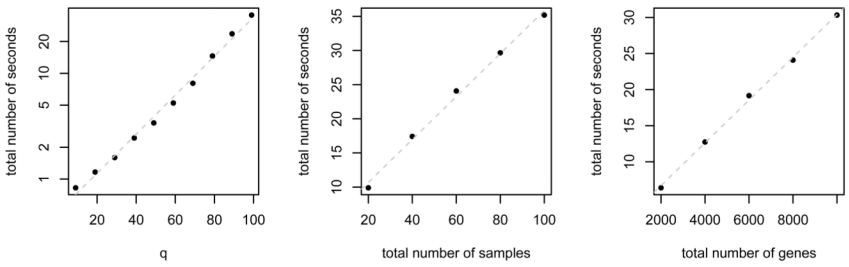


Figure 3.
Computing time with respect to covariance matrix dimension *q*, number of samples/genes

Table 1

Average prediction errors (%) over 30 simulations (listed within parenthesis are standard errors).

ρ	0.25	0.5	0.75
SRDA	9.2 (1.5)	8.5 (2.1)	8.4 (2.4)
NSC	15.6 (2.8)	25.3 (3.9)	28.4 (4.1)
SCRDA	9.9 (1.3)	13.9 (2.8)	20.9 (5.2)

Table 2

Average classification errors (%) over 50 cross validations (listed within parenthesis are the standard errors) and p-values testing SRDA performs better than the other methods.

	SRDA	NSC		SCRDA	
	err	err	SRDA>NSC p-value	err	SRDA>SCRDA p-value
prostate cancer (all)	7.82 (0.55)	13.35 (1.17)	2E-6	9.24 (0.95)	0.44
prostate cancer (GO.mf)	8.53 (0.65)	21.41 (1.56)	6E-13	9.88 (0.75)	0.03
breast cancer (all)	12.75 (1.11)	14.88 (1.08)	0.004	15.13 (1.06)	0.004
breast cancer (GO.mf)	10.50 (1.06)	13.38 (1.01)	0.002	13.25 (1.02)	0.0005

Comput Biol Chem. Author manuscript; available in PMC 2013 August 01.