Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms

Feng Tai and Wei Pan*
Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building (MMC 303), Minneapolis, MN 55455-0378, USA
Received on January 5, 2007; revised on April 4, 2007; accepted on April 26, 2007
Advance Access publication May 5, 2007
Associate Editor: Olga Troyanskaya

ABSTRACT
Motivation: In the context of sample (e.g. tumor) classifications with microarray gene expression data, many methods have been proposed. However, almost all the methods ignore existing biological knowledge and treat all the genes equally a priori. On the other hand, because some genes have been identified by previous studies to have biological functions or to be involved in pathways related to the outcome (e.g. cancer), incorporating this type of prior knowledge into a classifier can potentially improve both the predictive performance and interpretability of the resulting model.

Results: We propose a simple and general framework to incorporate such prior knowledge into building a penalized classifier. As two concrete examples, we apply the idea to two penalized classifiers, nearest shrunken centroids (also called PAM) and penalized partial least squares (PPLS). Instead of treating all the genes equally a priori as in standard penalized methods, we group the genes according to their functional associations based on existing biological knowledge or data, and adopt group-specific penalty terms and penalization parameters. Simulated and real data examples demonstrate that, if prior knowledge on gene grouping is indeed informative, our new methods perform better than the two standard penalized methods, yielding higher predictive accuracy and screening out more irrelevant genes.

Contact: weip@biostat.umn.edu

1 INTRODUCTION
In recent years, tumor classification based on microarray gene expression data has become one of the most active research topics in bioinformatics. Numerous studies on various types of cancers have appeared in the literature, such as breast cancer (Huang et al., 2003; Wang et al., 2005), prostate cancer (Singh et al., 2002; Welsh et al., 2001), lung cancer (Bhattacharjee et al., 2001), leukemia (Golub et al., 1999), etc. Many classification methods have been newly developed or adapted for these applications. In addition to predicting some outcomes related to the cancer, a major goal is to identify genes that are related to the cancer. At the same time, biological functions of genes have been explored intensively by the biological research community. A large amount of biological information gathered so far is stored in databases, such as those with the Gene Ontology (GO) annotations (Ashburner et al., 2000) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, 1996). In addition, prior experiments with similar biological objectives may have generated data that are relevant to the current study. Hence, borrowing information from prior data or biological knowledge seems natural for the current study; it opens up an opportunity of, and at the same time poses a challenge to, further improving over standard analysis methods that ignore prior knowledge and data.

From a methodological point, in the context of sample classification, in order to achieve a high predictive performance and effectively select a few relevant predictors for high-dimensional data like microarray gene expression data, many statistical learning methods have been adapted or developed, such as SVM (Vapnik, 1998), random forest (Breiman, 2001) and PAM (Tibshirani et al., 2003). These methods have gained much popularity because of their superior performance in practice. However, almost all the existing methods treat all the genes equally a priori in the process of model building, ignoring biological knowledge of gene functions, which may result in a loss of their effectiveness. For example, some genes have been identified or hypothesized to be related to cancer by previous studies; others may be known to have the same function of or be involved in a pathway with some known/putative cancer-related genes, hence we may want to treat these genes differently from other genes a priori when choosing genes to predict cancer-related outcomes. To take advantages of such prior information, Lottaz and Spang (2005) proposed a structured analysis of microarray data (SfAM), while Wei and Li (2006) proposed a modified boosting method called non-parametric pathway-based regression (NPR). The SfAM is based on the GO hierarchical structure, in which biological functions of genes are organized as a directed acyclic graph: each node in the graph represents a biological function; a child node has a more specific function while its parent node has a more general one. SfAM works by first building a separate classifier for each leaf node based on an existing method...
(e.g. PAM), then propagating their classification results by a weighted sum to their parent nodes, where the weights are related to the performance of the classifiers; a shrinkage scheme is used to shrink the weights towards zero so that a sparse representation is possible; and the process is repeated until the results are propagated to the root node. Because the final classifier is built based on the GO tree, it greatly facilitates the interpretation of a final result in terms of identifying biological processes that are related to the outcome. However, a downside is that only the genes annotated in the leaf nodes (i.e. with most detailed biological functions) are used as predictors; because of incomplete knowledge, other relevant genes that are not annotated yet cannot be used, which may in turn result in missing important new genes and losing predictive performance of the final model. In NPR, it is assumed that the genes can be first partitioned into several groups or pathways, then in boosting, only pathway-specific new classifiers (i.e. using only the genes in each of the pathways) were built. Our idea is similar to NPR with regard to grouping genes, but ours applies to any penalized method through the use of group-specific penalty terms while NPR only applies to boosting. More recently, Pang et al. (2006) proposed using random forests to rank biological pathways in regression and classification.

In this article, we propose a simple and flexible framework to incorporate prior knowledge of genes into penalized methods. In the Bayesian inference, it is standard to incorporate prior information by specifying a prior distribution. Penalized methods have a close connection to the Bayesian inference: a penalty term is related to a prior distribution for the genes involved in the penalty term (Hastie et al., 2001). However, most penalized methods only have a global penalty term involving all the genes, which essentially specifies the same prior distribution for all the genes. In order to incorporate prior information about different functional groups of the genes into analysis, we adopt group-specific penalty terms in a penalized method, thus allowing genes from different groups to have different prior distributions (e.g. different prior probabilities of being related to the cancer). As two concrete examples, we apply the idea to two penalized methods, the nearest shrunken centroids (also called PAM, Tibshirani et al., 2002) and penalized partial least squares (PPLS, Huang and Pan, 2003).

The rest of the article is organized as follows. We first review the standard PPLS and PAM. Then we introduce two new methods, mPPLS and mPAM, two modifications to PPLS and PAM, respectively; they have multiple penalty terms with multiple penalization parameters; the choice of the penalty terms is guided by prior knowledge. To reduce the computing demand in searching for multiple penalization parameters in mPPLS and mPAM, we present a weighting method that effectively reduces multiple unknown penalization parameters to only one. Simulation studies and analyses of two breast cancer datasets and a prostate cancer dataset are used to evaluate the proposed methods, and in particular illustrate the advantage of the new methods over the standard ones. We end with a short summary and discussion.

2 METHODS

2.1 Notation

Let \( x_i \) be the expression level of gene \( i \) in sample \( j \), and \( y_j \) be the cancer type for sample \( j \), \( i = 1, \ldots, p \) and \( j = 1, \ldots, n \) and \( y_j \in \{1, \ldots, K\} \). Denote \( Y = (y_1, \ldots, y_n)' \) and \( X = (x_{ij}, \ldots, x_{ij})' \). Here, we only consider two-class classification (\( K = 2 \)) where \( y_j \) is binary. Suppose we have \( n_1 \) tumor (Class I, \( C_I \)) samples and \( n_2 \) controls (Class II, \( C_{II} \)) such that \( n = n_1 + n_2 \). The mean of the expression levels of Class I samples, Class II samples and all \( n \) samples for gene \( i \) are \( \bar{x}_{i1}, \bar{x}_{i2} \) and \( \bar{x}_i \), respectively.

2.2 Nearest shrunken centroids

The ideal of nearest shrunken centroids is to shrink the class centroids \( \bar{x}_k, k = 1, 2 \) toward the overall centroid \( \bar{x}_i \). Let

\[
d_{ak} = \frac{\bar{x}_k - \bar{x}_i}{s_k}
\]

where \( s_k \) is the pooled within-class SD defined by

\[
s^2_k = \frac{1}{n} - \frac{1}{2} \left( \sum_{j=1}^{n} (x_{ij} - \bar{x}_k) + \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2 \right)
\]

with \( s_0 \) being a positive constant, usually set as the median of \( s_k \). Basically, \( d_{ak} \) is a modified t-statistic for gene \( i \). We can rewrite

\[
\tilde{x}_k = \bar{x}_i + m_k(s_k + s_0)d_{ak}
\]

and shrink \( d_{ak} \) toward zero by an amount \( \lambda \geq 0 \) by soft thresholding:

\[
d'_{ak} = \text{sign}(d_{ak}) |d_{ak}| - \lambda
\]

where \( \lambda \) has to be decided, usually by cross-validation (CV). We thus obtain a new shrunken centroid

\[
\tilde{x}_k = \bar{x}_i + m_k(s_k + s_0)d'_{ak}
\]

and define discriminant score for class \( k \) as

\[
\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \tilde{x}_k)^2}{s^2_i} - 2 \log(\pi_k)
\]

where \( x^* = (x_1^*, \ldots, x_p^*) \) is a new test sample and \( \pi_k = n_k/n \) is the class prior probability. The new test sample is classified as Class I if \( \delta_1(x^*) < \delta_2(x^*) \); otherwise, as Class II.

2.3 Penalized partial least squares regression

Partial least square (PLS) was first introduced by Wold (1966) and has been heavily promoted in the chemometrics literature as an alternative to ordinary least squares. It is often used in situations where the predictors are highly collinear, and/or the number of predictors \( p \) is large relative to the sample size \( n \), as encountered in microarray data (Nguyen and Rocke, 2002). PLS forms a sequence of uncorrelated linear components, which are linear combinations of the original predictors (i.e. gene expression levels), to predict the outcome.

To construct PLS components, we first center the \( Y \) and \( X \), to give \( U_1 = Y - \bar{Y}1 \) and \( V_1 = X - \bar{X}1, i = 1, \ldots, p \), where \( 1 = (1, \ldots, 1)' \) is the \( p \)-dimensional unit vector. \( U_1 \) is regressed against each \( V_i \) separately. Since the mean of \( U_1 \) and \( V_i \) are 0, for \( i = 1, \ldots, p \), the resulting least squares regression equations are

\[
\hat{U}_{1|i} = b_{1i}V_{1i}, \quad b_{1i} = (V_{1i}'V_{1i})^{-1}V_{1i}'u_{1i},
\]

1776
where \( u_i \) and \( v_{ij} \) are realized values of random variables \( U_1 \) and \( V_{1j} \), respectively. The first PLS components \( T_1 \) is constructed to be the average of \( \hat{U}_{1(i)} \),

\[
T_1 = \frac{1}{p} \sum_{i=1}^{p} u_i v_{1i}.
\]

Let \( U_2 \) be the residual from regressing \( U_1 \) on \( T_1 \) and let \( V_2 \) be the residual of regressing \( V_{1j} \) on \( T_1 \) for \( i = 1, \ldots, p \); we can repeat the above procedure to construct \( T_2 \) and thus iterate the process to get \( T_3, \ldots, T_p \).

To be specific, suppose that \( T_k = 0 \) (\( k \geq 2 \)) (where \( T_k \) is orthogonal to \( T_1, \ldots, T_{k-1} \) (therefore, \( T_k \) is orthogonal to \( T_1, \ldots, T_{k-1} \) (therefore, \( T_k \)).)

The residuals \( U_k \) and \( V_{k,j} \) are orthogonal to \( T_1, \ldots, T_k-1 \) (therefore, \( T_k \)). Thus, regressing \( U_k \) against \( V_{k,1}, \ldots, V_{k,p} \) in turn. The \( i \)th regression gives

\[
\hat{U}_{k(i)} = b_{ki} V_{ki}, b_{ki} = (\mathbf{v}_i \mathbf{x}_k)^{-1} \mathbf{v}_i \mathbf{u}_k
\]

and

\[
T_k = \frac{1}{p} \sum_{i=1}^{p} b_{ki} V_{ki}.
\]

Repeat this procedure until \( T_p \) is obtained, where the number of components \( q \) is pre-specified.

The final model is obtained by regressing \( Y \) on \( T_1, \ldots, T_q \) and has form

\[
Y = \beta_0 + \beta_1 T_1 + \cdots + \beta_q T_q
\]

where each \( T_k \) is a linear combination of \( X_k \) and the correlation between each pair of \( T_k \) is 0. The parameters \( \beta_0, \ldots, \beta_q \) are estimated by OLS. Since each \( T_k \) is a linear combination of \( X_k \), we can rewrite the model as

\[
Y = \beta_0 + \sum_{i=1}^{p} a_i X_i
\]

Penalized PLS (PPLS) was proposed by Huang and Pan (2003) to penalize \( a_i \), the coefficient for gene \( i \) in the PLS model, towards zero to result in a simpler model: using soft thresholding (2)

\[
d_i' = \text{sign}(a_i) |a_i| - \lambda_j, j
\]

where the penalty parameter \( \lambda \) has to be determined in practice, we construct a new component \( T = \sum_{i=1}^{p} d_i' X_i \) and regress \( Y \) against \( T \) using OLS

\[
Y = \beta_0 + \sum_{i=1}^{p} d_i' X_i
\]

To predict for a new sample with predictors \( X' \), we use and dichotomize \( Y' \) accordingly.

Note that in practice, one has to choose two tuning parameters, the number of components \( q \) in PLS and the shrinkage parameter \( \lambda \). We use crossvalidation to select the parameters, \( \theta = (q, \lambda) \), such that a minimum CV error is achieved. In the situation that there are multiple \( \theta \) giving a minimum CV error, we choose \( \theta \) with the smallest \( q \) and for this \( q \), choose the \( \lambda \)s such that the number of genes in the model is smallest.

### 2.4 New methods

In either of PAM and PPLS, the magnitude of shrinkage is determined by one shrinkage parameter \( \lambda \). In order to incorporate prior knowledge of different gene groups into the model, we propose using group-specific parameters \( \lambda_j \)’s. Specifically, we assume that, based on prior data or biological knowledge, the genes can be partitioned into \( J \leq 1 \) groups, \( G_1, \ldots, G_J \), and we shrink the genes from different groups by possibly different magnitudes; we replace expression (1) in PAM by

\[
d_k = \text{sign}(d_k)(|d_k| - \lambda_j), j \in G_j \]

yielding a modified PAM with multiple shrinkage parameters called mPAM; for PPLS, we replace \( 2 \) by

\[
d_i' = \text{sign}(a_i)(|a_i| - \lambda_j), i \in G_j
\]

leading to a new method called mPPLS. The basic idea is to treat the genes from the same group to be equal a priori while those from different groups unequal a priori. From a Bayesian perspective, it is equivalent to specifying separate prior distributions for the gene groups, thus allowing different priors for different gene groups if the data dictate such a requirement.

In practice, each of the shrinkage parameter is unknown and has to be determined; we used a grid search and CV to tune shrinkage parameters. For a large \( J \), it may be computationally too demanding to determine \( \lambda_1, \ldots, \lambda_J \) separately. Hence, we propose the following weighted method: we assume that \( \lambda_j = \lambda_j / w_j \) for \( j = 1, \ldots, J \), where

\[
w_j = \sum_{i \in G_j} |d_k| / |G_j|
\]

for a weighted PAM (wPAM), and

\[
w_j = \sum_{i \in G_j} |a_i| / |G_j|
\]

for a weighted PPLS (wPPLS), and \( |G_j| \) is the number of the genes in \( G_j \). Simply put, the shrinkage parameter for a group is inversely weighted by its group mean of the original unshrunken coefficients; see (1) and (2). In this way, we only need to tune one parameter \( \lambda \) as in the standard methods, though multiple shrinkage parameters are used. Note that the weight in (5) and thus the shrinkage parameter depend on class \( k \); because we only considered two-class classifications, we fixed \( k = 1 \) in (5).

Compared to treating the multiple shrinkage parameters separately, the weighted method, albeit less flexible, largely reduces the computational demand; furthermore, the weighted method penalizes less on the genes in a group with a larger mean parameter estimate: when a group contains a larger proportion of non-zero coefficients or a few large non-zero coefficients, indicating the existence of potentially useful genes in the group, the coefficients of the genes in the group will be shrunk less and thus less likely to be zero, leading to both higher chance of identifying important genes and in general smaller biases of shrinkage estimates, the latter of which may in turn improve predictive performance (Dabney, 2005). Given that the genes in a pathway or from the same functional group tend to work together, this is biologically reasonable: if the prior knowledge on grouping genes is informative, then the genes in the same group are more likely to be either involved or not involved in the biology; i.e. it is more likely that the genes from the same group have either zero or non-zero parameters simultaneously, hence borrowing information across the genes from the same group not only reduces the variability of the resulting estimate for a gene, but also leads to bias reduction for those non-zero parameters (i.e. relevant genes) as opposed to their being shrunk by a constant as in soft-thresholding.

Treating each gene as a separate group, the above weighting methods become

\[
d_k = \text{sign}(d_k)(|d_k| - \lambda_j), j \in G_j
\]

which have the same form as the non-negative garrote (NG) estimate in linear regression with orthogonal designs (Breiman, 1995; Yuan and Lin, 2007), and are closely related to the adaptive Lasso (Zou, 2006).
Again the intuition is to penalize more on a smaller parameter estimate while penalize less on a larger estimate, avoiding the bias problem of the soft-thresholding (or Lasso) while maintaining the continuity advantage of the soft-thresholding. As a special case of our proposed weighted methods, the NG estimate will be shown to improve over the standard methods with soft-thresholding, lending support for the idea of using the initial estimates for weighting, as suggested by Zou (2006). Of course, there is a key difference between our proposed methods and the NG estimate or adaptive Lasso: our weighting scheme is applied to groups of the genes based on prior knowledge, which should be more efficient through borrowing information across the genes within the same group if the grouping is reasonable.

3 RESULTS

To evaluate our methods, we applied our methods to both simulated and real data to compare our methods with the standard ones.

3.1 Simulation

To mimic real data, we used gene expression profiles from a breast cancer study (Huang et al., 2003) as the predictors in our simulation study. The original breast cancer data consisted of a total of 89 breast cancer patients. The microarray platform used was Affymetrix HG U95Av2, each containing 12,625 probe sets (also called genes for convenience). We used the observed expression levels as predictors $X$ to simulate the outcome $Y$. We restricted the number of genes to be $p = 1000$ or $3000$ in order to limit the computational time: we used only top $p$ genes with the largest sample variances across all 89 samples. The binary outcome $Y$ was generated by two steps: we first generated a continuous response $Z$ as

$$Z = X' \beta + \epsilon, \quad \epsilon \sim N(0, SI)$$

then dichotomized $Z$ to obtain

$$Y = I(Z > \text{mean}(X'\beta)),$$

where $I$ is an $89 \times 89$ identity matrix. After obtaining a simulated dataset, we randomly partitioned it into test and training data with 29 and 60 samples, respectively. For each simulation setup, 1000 independent replications were generated.

Two sets of regression coefficients $\beta$ were used: the first set, $\beta_{1, \ldots, \beta_{100}}$, for the 100 genes, were randomly drawn from $N(0, 10)$ while in the second set each of the remaining $p - 100$ $\beta$s was set to 0, representing two gene functional groups: informative and non-informative ones. To investigate the sensitivity of the proposed methods to the misspecification of the gene functional groups, we partitioned the whole set of genes as the following.

(1) Perfect specification: we correctly set the 100 informative genes as one group and the remaining ones as another group.

(2) Misspecification: we randomly chose $m$ genes from the informative and non-informative groups, respectively, then exchanged their group memberships; the first group contained $100 - m$ informative and $m$ non-informative genes, while the other group contained all other genes.

We tried $m = 20$ and $m = 80$, corresponding to ‘light’ and ‘heavy’ misspecifications, respectively.

Seven approaches were considered: standard PLS, standard PPLS with only one penalization parameter, our new PPLS with multiple penalization parameters (mPPLS) and weighted shrinkage parameter (wPPLS), standard PAM with one penalization parameter, our new PAM with multiple shrinkage parameters (mPAM) and weighted shrinkage parameter (wPAM). As a bench mark, we also considered PLS with only informative genes, the ideal (but not practical) case where we knew the truth about which genes were relevant.

Table 1 showed the mean classification errors and the mean numbers of genes selected over 1000 replications. As expected, the PLS with only informative genes gave lowest misclassification errors (7.17). Generally, the corresponding PPLS- and PAM-based methods performed very similarly. The PPLS-based methods tended to select more informative genes.

<table>
<thead>
<tr>
<th>$p$</th>
<th>Classifiers</th>
<th>Errors</th>
<th>$P$-value</th>
<th>Signal</th>
<th>Noise</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>PLS (100)</td>
<td>7.171</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td></td>
<td>wPAM (NG)</td>
<td>8.652</td>
<td>&lt;0.001</td>
<td>22.208</td>
<td>65.064</td>
<td>87.272</td>
</tr>
<tr>
<td></td>
<td>wPAM</td>
<td>8.362</td>
<td>&lt;0.001</td>
<td>37.182</td>
<td>99.376</td>
<td>136.558</td>
</tr>
<tr>
<td></td>
<td>wPAM (m = 20)</td>
<td>8.405</td>
<td>&lt;0.001</td>
<td>33.500</td>
<td>104.221</td>
<td>137.721</td>
</tr>
<tr>
<td></td>
<td>wPAM (m = 80)</td>
<td>8.681</td>
<td>0.363</td>
<td>28.705</td>
<td>124.163</td>
<td>152.868</td>
</tr>
<tr>
<td></td>
<td>mPAM</td>
<td>8.269</td>
<td>&lt;0.001</td>
<td>41.787</td>
<td>53.741</td>
<td>95.528</td>
</tr>
<tr>
<td></td>
<td>mPAM (m = 20)</td>
<td>8.321</td>
<td>&lt;0.001</td>
<td>34.164</td>
<td>54.082</td>
<td>88.246</td>
</tr>
<tr>
<td></td>
<td>mPAM (m = 80)</td>
<td>8.716</td>
<td>0.302</td>
<td>19.664</td>
<td>62.609</td>
<td>82.273</td>
</tr>
<tr>
<td></td>
<td>PPLS</td>
<td>8.759</td>
<td>0.001</td>
<td>33.519</td>
<td>217.448</td>
<td>250.967</td>
</tr>
<tr>
<td></td>
<td>wPPLS(NG)</td>
<td>8.512</td>
<td>0.002</td>
<td>35.074</td>
<td>202.46</td>
<td>237.534</td>
</tr>
<tr>
<td></td>
<td>wPPLS</td>
<td>8.409</td>
<td>&lt;0.001</td>
<td>37.201</td>
<td>169.652</td>
<td>206.855</td>
</tr>
<tr>
<td></td>
<td>wPPLS (m = 20)</td>
<td>8.513</td>
<td>&lt;0.001</td>
<td>35.185</td>
<td>182.527</td>
<td>217.712</td>
</tr>
<tr>
<td></td>
<td>wPPLS (m = 80)</td>
<td>8.722</td>
<td>0.379</td>
<td>33.608</td>
<td>218.140</td>
<td>251.748</td>
</tr>
<tr>
<td></td>
<td>mPPLS</td>
<td>7.835</td>
<td>&lt;0.001</td>
<td>50.906</td>
<td>80.877</td>
<td>131.783</td>
</tr>
<tr>
<td></td>
<td>mPPLS (m = 20)</td>
<td>8.149</td>
<td>&lt;0.001</td>
<td>41.335</td>
<td>90.309</td>
<td>131.644</td>
</tr>
<tr>
<td></td>
<td>mPPLS (m = 80)</td>
<td>8.926</td>
<td>0.013</td>
<td>23.644</td>
<td>131.237</td>
<td>154.881</td>
</tr>
<tr>
<td></td>
<td>PPLS (100)b</td>
<td>7.171</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td></td>
<td>wPPLS(NG)</td>
<td>8.604</td>
<td>&lt;0.001</td>
<td>18.758</td>
<td>89.263</td>
<td>108.021</td>
</tr>
<tr>
<td></td>
<td>wPPLS</td>
<td>8.407</td>
<td>&lt;0.001</td>
<td>34.565</td>
<td>158.517</td>
<td>193.082</td>
</tr>
<tr>
<td></td>
<td>wPPLS (m = 20)</td>
<td>8.532</td>
<td>&lt;0.001</td>
<td>29.749</td>
<td>166.058</td>
<td>195.807</td>
</tr>
<tr>
<td></td>
<td>wPPLS (m = 80)</td>
<td>8.894</td>
<td>0.505</td>
<td>23.942</td>
<td>70.718</td>
<td>94.636</td>
</tr>
<tr>
<td></td>
<td>mPAM</td>
<td>8.234</td>
<td>&lt;0.001</td>
<td>41.994</td>
<td>68.845</td>
<td>110.839</td>
</tr>
<tr>
<td></td>
<td>mPAM (m = 20)</td>
<td>8.328</td>
<td>&lt;0.001</td>
<td>34.522</td>
<td>70.718</td>
<td>105.240</td>
</tr>
<tr>
<td></td>
<td>mPAM (m = 80)</td>
<td>8.786</td>
<td>0.234</td>
<td>17.116</td>
<td>86.536</td>
<td>103.652</td>
</tr>
<tr>
<td></td>
<td>PPLS</td>
<td>9.056</td>
<td>0.001</td>
<td>27.474</td>
<td>478.805</td>
<td>506.279</td>
</tr>
<tr>
<td></td>
<td>wPPLS(NG)</td>
<td>8.760</td>
<td>&lt;0.001</td>
<td>32.041</td>
<td>446.414</td>
<td>478.455</td>
</tr>
<tr>
<td></td>
<td>wPPLS</td>
<td>8.486</td>
<td>&lt;0.001</td>
<td>32.972</td>
<td>296.905</td>
<td>329.877</td>
</tr>
<tr>
<td></td>
<td>wPPLS (m = 20)</td>
<td>8.679</td>
<td>&lt;0.001</td>
<td>29.677</td>
<td>322.565</td>
<td>352.242</td>
</tr>
<tr>
<td></td>
<td>wPPLS (m = 80)</td>
<td>9.017</td>
<td>0.346</td>
<td>27.444</td>
<td>464.522</td>
<td>491.966</td>
</tr>
<tr>
<td></td>
<td>mPPLS</td>
<td>7.980</td>
<td>&lt;0.001</td>
<td>48.610</td>
<td>123.536</td>
<td>172.046</td>
</tr>
<tr>
<td></td>
<td>mPPLS (m = 20)</td>
<td>8.224</td>
<td>&lt;0.001</td>
<td>38.878</td>
<td>132.678</td>
<td>171.556</td>
</tr>
<tr>
<td></td>
<td>mPPLS (m = 80)</td>
<td>8.944</td>
<td>0.103</td>
<td>19.011</td>
<td>214.851</td>
<td>233.862</td>
</tr>
</tbody>
</table>

*a Two-sided P-value for a paired t-test for an equal mean error between a new classifier and a standard classifier.

b PLS(100) is the ideal PLS method with only 100 informative genes.
Specifically, (1) we randomly partitioned a dataset into 10 parts and recorded the number of the genes included in a final model. The multiple shrinkage parameter methods based on ‘light’ misspecification \( m = 20 \) also performed better than the methods with no or only one shrinkage parameter. Even the new methods with a ‘severe’ misspecification \( m = 80 \) performed similarly as the standard PPLS and PAM for \( p = 1000 \), and strikingly, it might perform slightly better than PPLS and PAM as the total number of the genes \( p \) increased. The multiple shrinkage parameter methods, without regard to perfect specification or mis-specification of the gene groups, were much less sensitive to the total number of the genes included in a starting model: their performance went down much slower than other methods. On the other hand, in all cases, the multiple shrinkage parameter methods used not only a much higher percentage of informative genes but also much fewer genes in total than the standard methods. The weighted penalized methods (i.e., wPPLS and wPAM) provided a good approximation to multiple penalized methods in terms of prediction and gene selection performance. We can also observe that the weighted methods with each gene being treated as an individual group, corresponding to using the NG estimates, performed slightly better than the standard method (with the soft-thresholding) while using fewer genes; however, they were not as good as the weighted methods (with two groups).

In summary, the proposed methods, by including more informative genes and less non-informative genes, gave better predictive performance and better interpretability than the standard methods.

### 3.2 Examples

We applied our new methods to three public datasets. The first one was the breast cancer data from Huang et al. (2003), denoted as BCH and here we only focused on the recurrence outcome. There were in total \( n = 52 \) samples (18 with recurrence of tumor and 34 without) and \( p = 12,625 \) probe sets from Affymetrix HG U95Av2 genechips. The second one was the breast cancer data from Wang et al. (2005), denoted as BCW, containing expression profiles for \( n = 286 \) patients with lymph-node-negative primary breast cancer, of whom 107 patients developed distant metastasis during the 5-year follow-up while 179 were relapse free. The genechips used were Affymetrix HG U133A, each containing \( p = 22,283 \) probe sets. The third one was on prostate cancer (Welsh et al., 2001), denoted as PSW. There were in total \( n = 34 \) samples in PSW: 25 tumors and 9 normal tissues arrayed by Affymetrix HG U95A chips.

A double 10-fold CV was used to estimate the classification errors and the numbers of the genes included in a final model. Specifically, (1) we randomly partitioned a dataset into 10 parts of almost equal size, denoted as \( D_1, \ldots, D_{10} \); (2) for \( k = 1, \ldots, 10 \), we left out \( D_k \) as the test data and used \( D_\setminus k = \cup_{j \neq k} D_j \) as training data: (a) a 10-fold CV was conducted on \( D_\setminus k \) to select appropriate tuning parameters (i.e., \( \lambda \)'s and \( g \)); (b) the model with the selected tuning parameters was fitted using \( D_k \), and we recorded the number of the genes in the fitted model and (c) the number of classification errors was recorded when the fitted model was applied to \( D_k \). In short, we conducted honest CV in which any test data were never used in any aspect of model building (Ambroise and McLachlan, 2002).

#### 3.2.1 Breast cancer data

The SuperArray cancer arrays provided a list of 113 genes known or hypothesized to be related to tumor metastasis (www.superarray.com). We identified 223 probe sets in an Affymetrix HG U95Av2 genechip corresponding to a subset of those 113 genes. These 223 probe sets were treated as the informative group while the remaining 12,402 ones as the non-informative groups. Table 2 provides a comparison between our new multiple shrinkage parameter methods and the standard penalized methods based on a double 10-fold CV. The mPPLS and mPAM performed much better than PPLS and PAM: the former two had less CV errors, while including much fewer genes, among which higher proportions came from the informative group. The wPPLS and wPAM performed similarly to PPLS and PAM.

Using the same set of the 113 genes related to tumor metastasis, we obtained 275 probe sets as the informative group while the remaining 22,008 as the non-informative group for the BCW data. Table 3 shows the performance of the methods: mPPLS selected fewer genes than PPLS but gave more errors; on the other hand, mPAM selected much fewer genes and performed better than PAM. Again, The wPPLS and wPAM performed similar to PPLS and PAM, but selected fewer genes.

Those 113 genes can be further partitioned into seven groups according to their biological functions: cell adhesion, extracellular matrix proteins, cell cycle, cell growth and proliferation, apoptosis, transcription factors and regulators and other genes involved in metastasis; we treated the remaining genes as another group, resulting in a total of eight groups. We applied the weighted methods to the eight groups of the genes. Results are shown in Tables 2 and 3: there was no or only slight improvement.

### Table 2. Breast cancer data: BCH

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Errors</th>
<th>Informative ((p_1 = 223))</th>
<th>Non-informative ((p_2 = 12,402))</th>
<th>Total ((p = 12,625))</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPLS</td>
<td>13</td>
<td>39.2</td>
<td>2585.6</td>
<td>2624.8</td>
</tr>
<tr>
<td>mPPLS</td>
<td>9</td>
<td>28.9</td>
<td>72</td>
<td>36.1</td>
</tr>
<tr>
<td>wPPLS</td>
<td>13</td>
<td>36.9</td>
<td>2585.4</td>
<td>2622.3</td>
</tr>
<tr>
<td>wPAM</td>
<td>11</td>
<td>27.2</td>
<td>1710.1</td>
<td>1737.3</td>
</tr>
<tr>
<td>(8 groups)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAM</td>
<td>12</td>
<td>4.0</td>
<td>299.9</td>
<td>303.9</td>
</tr>
<tr>
<td>mPAM</td>
<td>11</td>
<td>24.1</td>
<td>78.8</td>
<td>102.9</td>
</tr>
<tr>
<td>wPAM</td>
<td>12</td>
<td>3.2</td>
<td>299.8</td>
<td>303.0</td>
</tr>
<tr>
<td>(8 groups)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We also applied wPAM and wPPLS to the BCW data based on the cancer pathway information as described in Wei and Li (2006). 245 genes from 33 cancer-related sub-pathways and 188 cancer-related genes yielded a total of 433 genes in 34 gene groups. The misclassification rates based on 10-fold CV for wPAM and wPPLS were 99/286 = 34.6% and 110/286 = 38.5%, respectively, which were competitive with other methods as shown in Wei and Li (2006): e.g. the random forest gave 33% while SVM 42%. Nevertheless, our main purpose here is not to compare our results with those of other non-PAM or non-PPLS classifiers, because there may be inherent differences among the classifiers, e.g. between PAM and SVM, implying that it may be unfair to compare wPAM with SVM; rather, because of the generality of our proposal, we may pursue in the future to compare SVM with its modified versions, say mSVM and wSVM, that have multiple penalization parameters (Hastie et al., 2001 for a formulation of SVM as a penalized method with only a single penalization parameter).

3.2.2 Prostate cancer data The SuperArray cancer arrays also provided a list of 263 genes useful as molecular markers for the prognosis and diagnosis of prostate cancer, which corresponded to 411 probe sets in an Affymetrix HG U95A chip. As before, we used this list of the probe sets as the informative group while the remaining ones as the non-informative group, and the classification results were shown in Table 4. Even though all the methods gave good predictive accuracy rates, the new methods mPPLS and mPAM used much fewer genes, most of which came from the informative group.

3.2.3 Using BCH data to generate prior for analyzing BCW data Since the two breast cancer datasets had the same clinical outcome, recurrence of tumor, would like to combine them into an analysis. We chose using the BCH data to generate prior information because the BCH study was actually conducted prior to BCW (year 2003 versus 2005). The goal was to find out whether we could gain some improvement on prediction for the BCW data by incorporating gene information drawn from the BCH data. We applied PAM to the BCH data: the final model contained 234 genes, some of which were likely to be related to the recurrence of breast cancer. Those 234 genes corresponded to 452 probe sets on a HGU133A chip for the BCW data. We used those 452 probe sets as the informative group and the remaining 21831 ones as the non-informative group for the BCW data. Table 5 shows the performance of the methods. The mPAM performed much better than PAM: mPAM had 104 samples misclassified with on average only 439.8 genes, while PAM had 116 samples misclassified with 2410.8 genes. In addition, mPAM used a higher proportion of the genes from the informative group. The wPPLS and PPLS performed the same: both performed better than mPAM and PAM with fewer classification errors, but used much more genes than the two PAM-based methods; though giving five more errors, mPPLS selected fewer genes than PPLS and wPPLS.

### Table 3. Breast cancer data: BCW

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Errors</th>
<th>Informative</th>
<th>Non-informative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPLS</td>
<td>88</td>
<td>149.2</td>
<td>13 153.6</td>
<td>13 302.8</td>
</tr>
<tr>
<td>mPPLS</td>
<td>98</td>
<td>137.5</td>
<td>11 143.6</td>
<td>11 281.1</td>
</tr>
<tr>
<td>wPPLS</td>
<td>88</td>
<td>148.5</td>
<td>13 153.6</td>
<td>13 302.1</td>
</tr>
<tr>
<td>wPPLS (8 groups)</td>
<td>87</td>
<td>142.8</td>
<td>12 265.7</td>
<td>12 408.5</td>
</tr>
<tr>
<td>PAM</td>
<td>116</td>
<td>29.9</td>
<td>2380.9</td>
<td>2410.8</td>
</tr>
<tr>
<td>mPAM</td>
<td>106</td>
<td>10.2</td>
<td>23.5</td>
<td>33.7</td>
</tr>
<tr>
<td>wPAM</td>
<td>112</td>
<td>15.3</td>
<td>691.7</td>
<td>707.0</td>
</tr>
<tr>
<td>wPAM (8 groups)</td>
<td>107</td>
<td>12.9</td>
<td>391.6</td>
<td>404.5</td>
</tr>
</tbody>
</table>

### Table 4. Prostate cancer data

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Errors</th>
<th>Mean number of genes selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPLS</td>
<td>4</td>
<td>43.7 1102.0 1145.7</td>
</tr>
<tr>
<td>mPPLS</td>
<td>4</td>
<td>182.2 22.3 204.5</td>
</tr>
<tr>
<td>wPPLS</td>
<td>4</td>
<td>49.6 1105.7 1155.3</td>
</tr>
<tr>
<td>PAM</td>
<td>2</td>
<td>58.3 1524.6 1582.9</td>
</tr>
<tr>
<td>mPAM</td>
<td>1</td>
<td>106.3 14.4 120.7</td>
</tr>
<tr>
<td>wPAM</td>
<td>3</td>
<td>73.9 1772.1 1846.0</td>
</tr>
</tbody>
</table>

### Table 5. Breast cancer data: BCH as prior for BCW

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Errors</th>
<th>Mean number of genes selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPLS</td>
<td>88</td>
<td>323.6 12979.2 1302.8</td>
</tr>
<tr>
<td>mPPLS</td>
<td>93</td>
<td>285.7 8302.0 8587.7</td>
</tr>
<tr>
<td>wPPLS</td>
<td>88</td>
<td>323.6 12979.2 1302.8</td>
</tr>
<tr>
<td>PAM</td>
<td>116</td>
<td>63.3 2347.5 2410.8</td>
</tr>
<tr>
<td>mPAM</td>
<td>104</td>
<td>74.6 365.2 439.8</td>
</tr>
<tr>
<td>wPAM</td>
<td>114</td>
<td>75.0 2447.0 2522.0</td>
</tr>
</tbody>
</table>
in any pathway as individual groups with group size one. To deal with those genes belonging to multiple pathways, we considered two approaches for the weighted methods. The first, called random assignment, was to randomly assign a gene into one of the several pathways in which it was annotated. Another approach, called non-random assignment, was to first consider it in each of the multiple pathways in which it was annotated, such as in calculating group weights \( w_j \); then we selected the minimum of these \( w_j \)’s, say \( w_j^0 \), and used \( \lambda / w_j^0 \) as the shrinkage parameter for the gene; other genes in only one or none of the pathways were treated the same as before.

For comparison, the standard PAM/PPLS and wPAM/wPPLS with either KEGG pathways or individual genes as groups (NG) were applied. Results obtained from 10 independent runs of a double 10-fold CV are shown in Table 6. First, it is reassuring to see that for the weighted methods the two approaches to handling genes in multiple groups yielded similar results. Second, although the wPPLS with the KEGG grouping had a slightly larger misclassification error rate than PPLS, it did use much fewer genes; on the other hand, wPAM based on the KEGG pathways had a smaller error rate and used much fewer genes than PAM. Finally, the NG gave either a higher or similar misclassification error as compared to the standard methods, but selected fewer genes. In summary, in terms of both predictive accuracy and model parsimony, the weighted methods using the KEGG pathways seemed to be the winner.

### 4 DISCUSSION

Here, we have proposed a simple and flexible framework to incorporate various sources of prior knowledge on gene functions into building more effective penalized classifiers. In contrast to standard methods of treating all the genes equally a priori, we propose to partition the genes into various groups based on prior data or biological knowledge such that the genes in the same group are more likely to function similarly, then we use group-specific penalty terms and associated penalty parameters to account for possibly varying degrees of relevance of the gene groups to the outcome of interest. Implemented in PAM and PPLS, the proposed methods were shown to have better predictive performance while containing fewer genes as compared to the standard PAM and PPLS with simulated data and several real datasets. We also investigated the robustness of the new methods in a simulation study; even when the gene groups were not completely correctly specified, the new methods worked either better than or at least as well as the standard methods. Nevertheless, in general, the performance of the proposed methods depend on the degree of informativeness of gene grouping, as expected.

The basic idea of our proposal parallels that of Pan (2005) in the context of detecting differential gene expression and that of Pan (2006) in clustering gene expression profiles for gene function discovery. The importance of incorporating biological knowledge into analysis has been increasingly recognized (Dopazo, 2006), but most applications are in clustering analysis (e.g. Al-Shahrour et al., 2005; Cheng et al., 2004; Fang et al., 2005; Huang and Pan, 2006), while there seems to be fewer studies in classification with only a few exceptions (Lottaz and Spang, 2005; Pang et al., 2006 and Wei and Li, 2006).

In the real data examples, we have shown how to incorporate biological knowledge, extracted from either an existing database or a previous study, into the current analysis; for the latter case, the previous study and the current study used two different microarray platforms for gene expression profiling. This demonstrates the flexibility of our proposed methods. As biological knowledge as well as data from relevant experimental studies accumulate over time, the proposed framework provides a general way to incorporate such ever-increasing amount of prior knowledge into analysis and thus also a potential to further improve the predictive performance.

Our use of multiple penalization parameters or terms for multiple gene groups is related to block thresholding in wavelets (Cai, 1999). However, a major difference is that our groups (or blocks) of the genes are formed based on prior knowledge while they are data-driven in the latter; nonetheless, theoretical optimum properties of block thresholding as compared to term-by-term thresholding (corresponding to a single shrinkage parameter in standard penalized methods) in wavelets may provide theoretical support for our proposal in the current context. Likewise, the theory of adaptive Lasso (Zou, 2006) may also help justify and explain the good performance of our proposed weighted methods. At this moment, we do not have a rigorous theory for our proposed methods; however, in addition to empirical evidence shown in numerical examples, the connection of our methods to the NG estimator, as opposed to soft-thresholding, along with the common wisdom of averaging over groups to reduce noise, suggests an intuitive argument favoring our proposed methods. Finally, although we have only focused on PAM and PPLS as concrete examples, our idea can be equally applied to many penalized methods for regression and classification, such as Lasso (Tibshirani, 1996) and SVM (Vapnik, 1998), or for other outcome variables, such as survival times (Broet et al., 2006; Gui and Li 2005). These are all interesting topics to be studied in the future.

### ACKNOWLEDGEMENTS

This research was partially supported by NIH grant HL65462 and a UM AHC Faculty Research Development grant.
The authors thank the reviewers for helpful and constructive comments.

Conflict of Interest: none declared.

REFERENCES


