Borrowing information from relevant microarray studies for sample classification using weighted partial least squares

Xiaohong Huang\textsuperscript{1}, Wei Pan\textsuperscript{1,*}, Xinqiang Han\textsuperscript{2}, Yingjie Chen\textsuperscript{2}, Leslie W. Miller\textsuperscript{2}, Jennifer Hall\textsuperscript{2}

December 23, 2004

\textsuperscript{1} Division of Biostatistics, School of Public Health, University of Minnesota

\textsuperscript{2} Cardiovascular Division, Department of Medicine, Medical School, University of Minnesota

* Corresponding author. A460 Mayo Building (MMC 303), A460 Mayo Building (MMC 303), Minneapolis, MN 55455-0378, USA. Email: weip@biostat.umn.edu, phone: (612)626-2705, fax: (612)626-0660.
Borrowing information from relevant microarray studies for sample classification using weighted partial least squares

ABSTRACT
With an increasing number of publicly available microarray datasets, it becomes attractive to borrow information from other relevant studies to have more reliable and powerful analysis of a given dataset. We do not assume that subjects in the current study and other relevant studies are drawn from the same population as assumed by meta-analysis. In particular, the set of parameters in the current study may be different from that of the other studies. We consider sample classification based on gene expression profiles in this context. We propose two new methods, a weighted partial least squares (WPLS) method and a weighted penalized partial least squares (WPPLS) method, to build a classifier by a combined use of multiple datasets. The methods can weight the individual datasets depending on their relevance to the current study. A more standard approach is first to build a classifier using each of the individual datasets, then to combine the outputs of the multiple classifiers using a weighted voting. Using two quite different datasets on human heart failure, we show first that WPLS/WPPLS, by borrowing information from the other dataset, can improve the performance of PLS/PPLS built on only a single dataset. Second, WPLS/WPPLS performs better than the standard approach of combining multiple classifiers. Third, WPPLS can improve over WPLS, just as PPLS does over PLS for a single dataset.

Key words: Meta-analysis; Partial least squares; Penalized partial least squares; Gradient directed path; Squared error loss.
1. Introduction

DNA microarray technologies allow the measurement of expression levels of thousands of genes simultaneously. Microarray experiments are more and more widely used in classification of tumor samples, prediction of clinical outcomes, or detecting differential gene expressions. With a rapidly increasing number of publicly available microarray datasets addressing various biological questions for various organisms, there is potential to gain more information by a combined analysis of multiple studies. For example, it has become popular to take a meta-analysis approach to combining data from multiple studies to detect differential gene expression (Rhodes et al., 2002; Xin et al., 2003; Choi et al., 2004; Ghosh et al., 2004; Wang et al, 2004) or for sample classification (Jiang et al, 2004; Parmigiani et al, 2004; Shen et al, 2004). A technical issue is how to combine microarray data measured using different microarray techniques or platforms, such as cDNA vs Affymetrix arrays, or different versions of Affymetrix arrays, because of possibly different gene identities and possibly incomparable expression measurements across different platforms (e.g. Morris et al., 2003; Robb et al., 2003; Hu et al., 2003; Lin et al., 2003).

Here we consider a related but different problem. Our goal is to analyze a given dataset drawn from a current study. To increase the statistical power, we would like to borrow information from other relevant studies. A key difference from meta-analysis is that we do not assume that the current study shares a common set of parameters with other studies. For example, we might be interested in identifying genes associated with ventilator-associated lung injury (VALI) based on a human study. On the other hand, there are studies on the same subject using animal models, such as rat, mouse and dog (Grigoryev et al 2004). We would like to borrow information from these animal studies to address the scientific question, identifying the human genes associated with VALI. This is different from a meta-analysis (Grigoryev et al 2004) with its goal to identify the genes associated with VALI that are conserved across the species over the evolutionary history, perhaps only a subset of the human genes of interest. In other words, meta-analysis has to assume that we have a common set of parameters across the human and animal studies, on which a statistical inference is to
be drawn. In our analysis, we would not assume such a common set of parameters; rather, we are only interested in inference on a set of parameters specific for humans. Although the human-specific set of parameters is in general different from that of animal models, it is reasonable to assume a priori that they are likely to be close. Hence, we may borrow information from the animal studies to improve the estimation on the parameters for humans; the animal studies will be called secondary or auxiliary as compared to the primary human study. The statistical motivation of our proposal is similar to that of the weighted likelihood theory, which seeks to reduce the variance of an estimator (with a possible price of increasing its bias) and thus to result in a smaller mean squared error or prediction error (Newton and Raftery, 1994; Rao, 1991; Hu and Zidek, 2002; Wang et al., 2004; Ghosh et al, 2004 and references therein). As a concept-of-proof, we consider two studies on human heart failure, the LVAD study and the PGA study; more details on the studies are presented later. Our goal is to use gene expression profiles to distinguish etiologies of heart failure for LVAD patients, and we treat the PGA data as secondary or auxiliary. To be specific, we consider comparing ischemic (IS) group with idiopathic (ID) group. There are 10/13 IS,ID samples and 11/13 IS,ID samples in the LVAD and PGA data respectively. Intuitively, due to the small sample size and the relevance of the two studies, we would like to borrow information from the PGA data to build a model for LVAD patients. Although both the LVAD and PGA studies are on humans, they own some features shared by other more typical and less relevant studies for which and from which we would like to borrow information: due to the population heterogeneity (i.e. unobserved differences in patient characteristics), different study protocols and different microarray platforms, the data from the two studies are quite different. In particular, it is much harder to distinguish IS,ID patients using gene expression profiles in the LVAD study than in the PGA study. For example, using the penalized partial least squares (PPLS) method with varying numbers of genes and of components in a starting partial least squares (PLS) model (Huang and Pan, 2003), i) the leave-one-out-cross-validation (LOOCV) misclassification errors range from 5 to 11 for the LVAD data; ii) the LOOCV errors range only from 1 to 3 for the PGA data; iii) the minimum test error on
the LVAD data with the PPLS model built using the PGA data is 8. These results highlight
some existing differences underlying the two datasets; in particular, appropriate models for
the LVAD data and the PGA data may be different. Nevertheless, it is desirable to take
account possible differences between the two datasets while borrowing information from the
PGA to build a better classifier for the LVAD study.

We propose a weighted partial least squares (WPLS) method and a weighted penalized
partial least squares (WPPLS) method, which account for possibly different relevances of the
studies by assigning them possibly different weights. PLS method is considered especially
useful for constructing linear models when there are many covariates and a relatively small
sample size, as is typical with microarray data. There has been an increasing application
of PLS/PPLS to microarray data (e.g., among others, Nguyen and Rocke 2002; Hawkins et
is an extension of the standard PLS method by giving samples different weights (based on
their relevance to the current study); it is facilitated by formulating PLS solutions as a
gradient directed path in minimizing a loss function. The WPPLS method, which penalizes
or regularizes the coefficients of a WPLS model, aims to facilitate model interpretation
and further reduce noise effects of microarray data on the model and thus to improve the
performance over the WPLS method.

2. Methods

Suppose that $x_{ij}$ is the expression level of gene $i$ in sample (array) $j$, and the random
variable $y_j$ is the response variable for sample $j$, where $i = 1, \ldots, p$ and $j = 1, \ldots, n$. In
the current context, $y_j = 1$ or $-1$, representing one of the two types of the sample. Denote
column vectors $x_i = (x_{i1}, \ldots, x_{in})^T$ and $y = (y_1, \ldots, y_n)^T$. Given the data, the goal is to
construct a linear model for the sample type $Y$; that is, in the linear model

$$F(X, a) = a_0 + \sum_{i=1}^{p} a_i x_{ij}$$

we are to estimate the parameters $a = (a_0, a_1, \ldots, a_p)$. Ideally, we'd like to minimize the
expected loss/risk

\[ R(a) = E_Y L(Y, F(X, a)), \]

where \( L(Y, F(X, a)) \) is the loss of predicting the response variable \( Y \) by its predicted value of \( F(X, a) \). The optimal values of \( a \) are those that minimize the expected loss. Since in practice the distribution of \( Y \) is unknown, we estimate the expected loss by the empirical loss based on the observed data

\[ \hat{R}(a) = \frac{1}{n} \sum_{j=1}^{n} L(y_j, a_0 + \sum_{i=1}^{p} a_i x_{ij}). \]

The most common estimator of \( a \) is the one that minimizes the empirical loss

\[ \hat{a} = \arg\min_a \hat{R}(a). \]

**Partial least squares with a gradient directed path**

Taking the loss criterion as the squared error loss

\[ L(Y, F(X, a)) = (Y - F(X, a))^2 / 2, \]

the above minimizer of the empirical loss becomes the ordinary least squares (OLS) estimator. However, if \( p > n \), as in the current context, the OLS estimator is not unique and has large variability, whereas the partial least squares (PLS) estimator is taken as a suitable alternative. It is known that the PLS estimator corresponds to a conjugate gradient directed path to minimize the empirical squared error loss (Wold et al., 1984; Friedman and Popescu, 2004). Specifically, the PLS estimators \( \hat{a}_k \) are an ordered sequence of conjugate gradient steps (Gill, Murray and Wright, 1981):

\[ \hat{a}_{k+1} = \hat{a}_k + \rho_k s_k, \]

with

\[ s_k = g_k + \frac{g_k^T g_k}{g_{k-1}^T g_{k-1}} s_{k-1}, \]
\[ g_k = -\frac{\partial}{\partial a} \tilde{R}(a) \bigg|_{a = \hat{a}_k}, \]

and \( \rho_k \) is the step size defined as

\[ \rho_k = \text{argmin}_\rho \tilde{R}(\hat{a}_k + \rho \mathbf{s}_k). \]

The initial value \( \mathbf{s}_0 \) is defined as the \((p + 1)\)-dim zero vector and \( \hat{a}_0 \) is defined as the \((p + 1)\)-vector \((\bar{y}, 0, \ldots, 0)\), where \( \bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j \). Note that \( k \) corresponds to the number of the (latent) linear components in the usual PLS formulation.

The covariates \( x_{ij} \) are often scaled to have sample variance one (i.e. \( \text{var}(x_{i1}, \ldots, x_{in}) = 1 \) for any \( i \)) so that they have equal a priori influence as predictors.

**Weighted partial least squares method with a gradient directed path**

Following the idea of Friedman and Popescu (2004), we generalize the PLS to a weighted PLS (WPLS) with an empirical weighted squared error loss. Suppose \( w_j \) is the weight assigned to sample \( j \) \((j = 1, \ldots, n)\) with \( 0 \leq w_j \leq 1 \) and \( \sum_{j=1}^{n} w_j = 1 \). Define the empirical weighted squared error loss

\[ \tilde{R}_w(a) = \sum_{j=1}^{n} w_j L(y_j, a_0 + \sum_{i=1}^{p} a_i x_{ij}), \]

where the loss function \( L \) is defined to be the squared error loss, which is the same as that in the usual PLS. Then the WPLS estimates \( \{\hat{a}_k\} \ (k = 0, 1, \ldots, \text{min}(n, p)) \) correspond to a conjugate gradient path in minimizing \( \tilde{R}_w \), as described for PLS. The initial value \( \mathbf{s}_0 \) is defined as the \( p + 1 \) zero vector, and \( \hat{a}_0 \) is defined as the \( p + 1 \) vector \((\bar{y}^w, 0, \ldots, 0)\), where \( \bar{y}^w = \frac{1}{n} \sum_{j=1}^{n} w_j y_j \). The derivation for the negative gradient \( g_k \) and the step size \( \rho_k \) is given below.

First, we define an \( n \times n \) diagonal matrix \( W = \text{diag}(w_1, w_2, \ldots, w_n) \), and an \( n \times (p + 1) \)
matrix $Z = (1, x_1, ..., x_p)$. Then

$$g_k = - \frac{\partial}{\partial a} \tilde{R}_w(a) \bigg|_{a=\hat{a}_k}$$
$$= - \frac{\partial}{\partial a} \sum_{j=1}^{n} w_j L(y_j, a_0 + \sum_{i=1}^{p} a_i x_{ij}) \bigg|_{a=\hat{a}_k}$$
$$= - \frac{\partial}{\partial a} \sum_{j=1}^{n} w_j (y_j - (a_0 + \sum_{i=1}^{p} a_i x_{ij}))^2 / 2 \bigg|_{a=\hat{a}_k}$$
$$= - \frac{\partial}{\partial a} (y - Za)^T W (y - Za) / 2 \bigg|_{a=\hat{a}_k}$$
$$= Z^T W (y - Z\hat{a}_k).$$

We use the second derivative test to estimate the step size $\rho_k$, which is the minimum value of the weighted average risk $\tilde{R}_w(\hat{a}_k + \rho \hat{s}_k)$. We take the first and the second derivatives of the weighted average risk:

$$\frac{\partial}{\partial \rho} \tilde{R}_w(\hat{a}_k + \rho \hat{s}_k) = \frac{\partial}{\partial \rho} \left( (y - Z(\hat{a}_k + \rho \hat{s}_k))^T W (y - Z(\hat{a}_k + \rho \hat{s}_k)) / 2 \right)$$
$$= - (Z \hat{s}_k)^T W (y - Z(\hat{a}_k + \rho \hat{s}_k))$$

and

$$\frac{\partial^2}{\partial \rho^2} \tilde{R}_w(\hat{a}_k + \rho \hat{s}_k) = \frac{\partial}{\partial \rho} \left( (Z \hat{s}_k)^T W (y - Z(\hat{a}_k + \rho \hat{s}_k)) \right)$$
$$= (Z \hat{s}_k)^T W (Z \hat{s}_k).$$

Then we consider two cases with $Z \hat{s}_k = 0$ and $Z \hat{s}_k \neq 0$ respectively:

1. If $Z \hat{s}_k = 0$, then $\tilde{R}_w(\hat{a}_k + \rho \hat{s}_k)$ does not depend on $\rho$ and we define $\rho_k = 1$.

2. If $Z \hat{s}_k \neq 0$, then $\frac{\partial^2}{\partial \rho^2} \tilde{R}(\hat{a}_k + \rho \hat{s}_k) > 0$ and we can obtain $\rho_k = \frac{(Z \hat{s}_k)^T W (Y - Z \hat{a}_k)}{(Z \hat{s}_k)^T W (Z \hat{s}_k)}$.

Finally we note that the PLS method is a special case of the weighted PLS: if we take $w_1 = \ldots = w_n = \frac{1}{n}$, then the weighted average risk is exactly the empirical risk and the weighted PLS reduces to the usual PLS method.

**Weighted penalized partial least squares method**
As pointed out in Huang and Pan (2003), microarray data are usually very noisy, and with a large number of genes, some regularization or penalization on the parameter estimates $a_i$ may be productive. This motivates constructing a PPLS model over a PLS model. In
parallel, the main idea of WPPLS is to penalize the coefficients in a WPLS model to remove some noises. Next we will only briefly introduce the WPPLS procedure, which is in the same spirit of PPLS as discussed in Huang and Pan (2003).

We first center each covariate: define $x_{ij}^c = x_{ij} - \bar{x}_i^w \mathbf{1}$, where $\bar{x}_i^w = \frac{1}{n} \sum_{j=1}^n w_j x_{ij}$ is the weighted average of $x_i$, and $\mathbf{1} = (1, \ldots, 1)^T_{n \times 1}$. Suppose we have a fitted WPLS model

$$\hat{Y}_j = a_0 + \sum_{i=1}^p a_i x_{ij}^c.$$  

We use soft-thresholding to penalize $a_i$, the coefficient for gene $i$. We let $a_i' = \text{sign}(a_i)(|a_i| - \lambda)_+$, where $f_+ = \max(0, f)$ and $\lambda$ is the shrinkage parameter to be determined by an independent tuning dataset or by cross-validation. Then we construct a new vector $T = (T_1, \ldots, T_n)^T$ with $T_j = \sum_{i=1}^p a_i' x_{ij}^c$ and regress $Y$ against $T$ using weighted least squares. That is, we want to minimize

$$\sum_{j=1}^n w_j (Y_j - \gamma_0 - \gamma_1 T_j)^2 = (Y - \gamma_0 \mathbf{1} - \gamma_1 T)^T W (Y - \gamma_0 \mathbf{1} - \gamma_1 T),$$

where the superscript $T$ means matrix transpose. Since $\mathbf{1}^T W X_i^c = 0$ for $i = 1, \ldots, p$ and $T$ is a linear combination of $X_i^c$, we have $\mathbf{1}^T W T = 0$. Under the above condition, it is easy to show that

$$\hat{\gamma}_0 = \hat{y}^w,$$

$$\hat{\gamma}_1 = (T^T W T)^{-1} T^T W \mathbf{y}$$

where $\hat{y}^w$ was defined as the weighted average of the response in constructing the weighted PLS. Therefore the final weighted PLS model can be expressed as

$$\hat{Y}_j = \hat{\gamma}_0 + \hat{\gamma}_1 \sum_{i=1}^p a_i' x_{ij}^c.$$  

Note that if $a_i' = 0$, then the coefficient $a_i$ is penalized to 0 and gene $i$ is not used in the model.

To classify a new sample with gene expression intensities $X^* = (X^*_1, \ldots, X^*_p)$, we use the sign of the predicted $Y$ value:

$$\hat{Y}^* = \hat{\gamma}_0 + \hat{\gamma}_1 \sum_{i=1}^p a_i'(X^*_i - \bar{x}_i^w),$$
if \( \hat{Y}^* > 0 \), then the new sample is classified in class I; otherwise, in class II.

Note that WPPLS reduces to the usual PPLS when we simply take all the weights in WPPLS to be \( w_j = 1/n \) for all \( j = 1, 2, \ldots, n. \)

3. Results

Data
In this section, we present the results of applying WPLS/WPPLS to data taken from two heart failure studies. The first study, being conducted at the University of Minnesota, aims to investigate the response of the failing heart to the mechanical unloading of the left ventricular assist device (LVAD). The LVAD is a mechanical device that replaces the pumping of the ventricle in patients with severe refractory heart failure and has been shown to significantly decrease mortality (Altemose et al., 1997; Levin et al., 1996; McCarthy and Smith, 2002; Rose et al., 2001). In the LVAD study, core samples were collected from the left ventricular apex of patients with refractory heart failure at the time of implantation of the mechanical assist device. The transcriptional profile of the mRNA in these samples was measured with Affymetrix U133A chips containing 22,283 probe sets. The gene expression levels were summarized in MAS 5.0. More details on the experiment can be found in Hall et al. (2004).

Patients in the LVAD study were divided into three classes based on the underlying etiology of heart failure: 1) ischemic; 2) acute myocardial infarction within 48 hours of LVAD implant; or 3) idiopathic. In the current LVAD dataset, there are 10, 7 and 13 patients in the above three classes respectively.

The CardioGenomics project at the Harvard Medical School, a part of the NHLBI-funded Programs for Genomic Applications (PGA), also carried out a similar heart failure study. They collected myocardial samples from patients undergoing heart transplantation whose failure arises from different etiologies (e.g. idiopathic, ischemic, alcoholic, valvular, and hypertrophic) and from normal organ donors whose hearts cannot be used for transplants. The transcriptional profiles of the samples were measured with Affymetrix U133 Plus 2 chips
containing 54,675 probe sets, and expression levels were summarized in MAS 5.0. In the PGA dataset, there are 12 normal samples, 11 ischemic samples and 13 idiopathic samples. The PGA data were downloaded in May 2004 at the site http://www.cardiogenomics.org.

Our primary goal is to use the microarray gene expression data to build a classifier that can differentiate between ischemic and idiopathic samples for the LVAD patients. Partly due to a small sample size (i.e. 10 and 13 samples in the two classes), our preliminary analysis shows that classifiers built with LVAD data alone cannot distinguish the ischemic and idiopathic samples of the LVAD patients very well. Noticing the obvious similarity and relevance of the PGA study to the LVAD study, we hope to build a classifier based on combining the data from the two studies to improve the prediction performance.

The first problem is how to combine (i.e. align the genes in) the LVAD data with the PGA data. Although these two studies used different Affymetrix chips, combining the two datasets is not challenging because all the probe sets present on a U133A chip are identically replicated on a U133 Plus 2 chip and both studies used MAS 5.0 to summarize expression levels. In the experiment, 22,277 probe sets in the LVAD data were successfully mapped to those of the PGA data by probe set ID (6 probe sets ID could not be found in the PGA data and were not used in the analysis). The two data sets were combined together according to these 22,277 common probe sets. The new data contain the LVAD data with (10+7+13) ischemic/acute MI/idiopathic samples and the PGA data with (11+11+13) normal/ischemic/idiopathic samples. Note that the LVAD data are our primary data and the PGA data are the secondary data. To make the arrays comparable, we normalized each array by subtracting its median expression level and then dividing by the quartile range (i.e. difference between its 75th-percentile and 25th-percentile).

**Gene ranking**

To explore the effect of the number of genes on the classification performance, we have a preliminary gene ranking using a F-statistic. This F-statistic can be used when there are more than two classes of samples, and when there are only two classes, it is equivalent to the usual t-statistic. This univariate ranking is used throughout, and obviously is by no means
to be optimal.

Suppose that the first $n_1$ samples are in the first class, and the remaining $n_2 = n - n_1$ samples are in the second class. Let $\bar{x}_{ik}$ be the sample mean of the expression levels of gene $i$ in class $k$, $k = 1$ and 2; let $\bar{x}_i$ be the overall mean expression level of gene $i$ in the $n$ samples. We construct an F-statistic for gene $i$ as the ratio of its mean sums of squares for between class and within class variations:

$$F_i = \frac{MS_{class}}{MS_{error}} = \frac{\sum_{k=1}^{2} n_k (\bar{x}_{ik} - \bar{x}_i)^2}{\left(\sum_{j=1}^{n_1} (x_{ij} - \bar{x}_{i1})^2 + \sum_{j=n_1+1}^{n} (x_{ij} - \bar{x}_{i2})^2\right)/(n-2)}.$$

We rank the genes based on their F-statistics: a gene with a larger F-statistic indicates a stronger relationship between its expression level and the class membership, and therefore has a higher rank for its potential to discriminate the classes.

In the current application, we have two classes, ischemic and idiopathic samples. We can calculate a gene’s F-statistic using only the LVAD samples, or only the PGA samples, or a weighted average of the above two. It was found that using only the LVAD data to calculate the F-statistics and rank the genes yields the best performance, and thus will be used throughout.

**Experiment 1: Combining two separate PPLS classifiers**

We first consider a standard approach: as usual we use each of the two data sets separately to build a PLS/PPLS to discriminate ischemic/idiopathic samples; to predict on a new sample, we combine the outputs of the two PLS/PPLS models using a weighted average to classify the new sample.

The relative weight for the LVAD classifier is fixed at 1 and the relative weight for the PGA classifier is taken to be 0, 1/4, 1/2, 3/4, and 1 respective. We tried various PPLS models that starts with top 50, 200, or 22277 ranked genes within CV. Various shrinkage parameters were also tried to shrink various percentages of regression coefficients in a PLS model to be 0. For example, imposing a 90% shrinkage on a PLS model means that 90% of the genes’ coefficients are shrunken to 0 and the resulting PPLS model contains only the
remaining 10% genes. Hence a 0% shrinkage parameter corresponds to a PLS model and non-zero shrinkage parameters to PPLS models. We consider $k = 1, \ldots, 8$ components in a PLS model. To save space, we only present results for PLS models containing top 200 genes and PPLS models that shrink 80% of the 200 genes to 0; other results are similar.

Table 1 shows the LOOCV misclassification errors. The first column represents the number of top ranked genes that a PLS classifier starts with; the second column represents the percent of genes whose coefficients are shrunk to 0; $w_{PGA}$ represents the relative weight for the PGA classifier.

If we compare the misclassification errors horizontally, we can see that when giving higher weights to the PGA classifiers, the combined PLS or PPLS classifiers may lead to slightly improved classification performances over that of the single PLS or PPLS for a given number of PLS components. However, in general, the performance of the combined PPLS classifiers is only comparable to that of using the LVAD data alone.

**Experiment 2: Weighted PPLS**

The experiment set-up is the same as before except that we now apply WPLS/WPPLS
methods to the combined data. The LVAD samples are given weight 1 while the PGA samples are given weights 0, 1/4, 1/2 and 3/4 respectively. Table 2 shows the LOOCV misclassification errors.

The shrinkage in WPPLS can improve the classification performance over the WPLS method. More interestingly, the WPLS/WPPLS models built using the combined data can improve the predictive performance over the PLS/PPLS models built using the LVAD data alone: the minimum number of LOOCV errors is 5 for using the LVAD data alone whereas it is 2 when any of the four non-zero weights is used to combine the two datasets.

Strikingly, if we compare the results here (Table 2) with those of Experiment 1 (Tables 1), we can clearly see that WPLS/WPPLS performs much better than the standard approach of combining two separate PLS/PPLS models.

4. Discussion

We have proposed a weighted PLS/PPLS methodology to build a classifier for a study of current interest by borrowing information from other more or less relevant studies. The
main idea of the methodology is to account for possibly different relevances of the other studies by weighting. Using data from two human heart studies, which are quite different, we have demonstrated its feasibility and potential to improve the predictive performance over the current practice of building a classifier using data from a single study. In addition, the proposed method is shown to perform better than the standard approach of combining classifiers, each built using only data from a single study. The improved performance of the proposed weighted method is not really surprising in light of the recently developed weighted likelihood theory (Hu and Zidek, 2002).

Although we have focused on the use of WPLS/WPPLS, our proposed weighted methodology is general and flexible. First, any other weighted classifiers can be applied. Second, the scope of the weighting scheme is quite broad. For example, suppose that the PGA data only contain ischemic and normal groups, not idiopathic group. Then, obviously, we cannot combine the classifier of the LVAD data for ischemic vs idiopathic groups with that of the PGA data for ischemic vs normal groups. In contrast, in our weighted method, we can still use the information in the PGA data by including only the data of the ischemic group in the loss function to build a model. Furthermore, the idea can be extended to cases where the primary and secondary experiments were conducted under different (but relevant) conditions, or on different organisms, such as humans of the primary study and mouse models of the secondary studies. Finally, it is also possible to extend WPPLS/WPLS to other types of data, such as microarray data with a survival end point (Li and Gui, 2004; Gui and Li 2004), or other loss functions for other classifiers (e.g. Ghosh 2003; Shen et al 2003; Li et al 2004).

In practice, the choice of the weights may be important. It can be based on prior knowledge. More generally, cross-validation can be used to select appropriate weights as was done in our example. This is a topic to be further evaluated in the future.

**Acknowledgements**

XH and WP were supported by an NIH grant. JH was supported by an AHA grant, the Lillehei Heart Institute and the Minnesota Medical Foundation.
References


