Regularized biomarker selection in microarray meta-analysis

Shuangge Ma [a] and Jian Huang [b]

[a]Department of Epidemiology and Public Health, Yale University, New Haven, CT, USA
[b]Departments of Statistics and Act Sci, and Biostatistics, University of Iowa, Iowa City, IA, USA

Department of Statistics and Actuarial Science
University of Iowa
Iowa City, IA 52242
Technical Report No. 383
November 2007

ABSTRACT

Motivation: In pharmacogenetic studies, it is common that multiple microarray experiments are conducted to investigate the relationship between the same clinical outcome and gene expressions. An important goal of such experiments is to discover a small number of influential genes out of thousands being measured. To increase statistical power and achieve consistent biomarker selection across multiple experiments, meta analysis techniques should be used. However, it is difficult to apply standard meta analysis approaches because of high dimensionality of microarray data and because different microarray experimental settings in different experiments may not be comparable.

Results: We propose the Meta Threshold Gradient Descent Regularization (MTGDR) approach for regularized microarray meta analysis. The proposed approach is model-based, allows for experiment-specific estimates and is capable of selecting the same sets of influential genes across different experiments. We demonstrate the MTGDR method for biomarker selection in pancreatic and liver cancer studies, where the logistic regression models are assumed.

Availability: Research R code is available at

http://publichealth.yale.edu/faculty/labs/ma/.

1 INTRODUCTION

Microarrays are capable of profiling human tissues on a genome-wide scale and have been extensively used in pharmacogenetic studies, where expression levels of thousands of genes are measured along with certain clinical outcomes. A major goal of such studies is to identify a small number of "interesting" genes that can be used as biomarkers for disease diagnosis and prognosis prediction and as targets for therapy. There are usually multiple independent microarray experiments with the same gene sets and the same phenotype measured in the same species. Meta analysis is needed to combine data from different experiments and analyze them (Choi et al. 2004; Ghosh et al. 2003; Wang et al. 2004; and Warnat et al. 2005).

Microarray meta analysis is challenging because (1) microarray experiments usually measure a huge number of genes from a small number of samples (i.e., the "*large d, small n*" setting), while only a small number of those genes are important. Any model-based approach needs a corresponding biomarker selection mechanism (Li 2007); and (2) different experiments may have different setups. Especially, different platforms may be used. Arrays that hybridize one sample at a time (e.g. synthesized oligonucleotide arrays) measure gene expression based directly on the signal intensity of each probe set. Spotted cDNA arrays hybridized with fluorescent labeled targets, in contrast, typically measure the ratio of the signal from a test sample to the signal of a co-hybridized reference sample. Thus one unit increase in the expression levels measured in a study using cDNA arrays is usually not directly comparable to one unit increase in a study using oligonucleotide arrays. For example, it has been shown that data from Affymetrix GeneChip oligonucleotide microarrays correlate poorly with the data from custom-printed cDNA microarrays (Kuo et al. 2002). Thus, data from different platforms cannot be directly combined.

Several approaches have been proposed to deal with the problem of detecting differentially expressed genes based on multiple datasets. Examples include using proper transformations to directly integrate raw gene expression data (Warnat et al. 2005); a Lasso based method (Ghosh et al. 2003); a random effects model based method (Stevens and Doerge 2005); a robust gene ranking approach (Hong et al. 2006) and a Bayesian approach (Jung et al. 2006), among others. The aforementioned approaches are only applicable to the simple "disease-versus-normal" setting and can be inefficient.

There are also studies that consider constructing predictive models from multiple microarray experiments. A majority voting with impact factors algorithm is proposed in Fung and Ng (2004). Gene shaving methods based on random forrest and Fisher's linear discrimination are applied in Jiang et al. (2004). A computationally intensive Bayesian approach is proposed in Conlon et al. (2007). Statistical software, including the R packages metaArray (Ghosh and Choi 2006), MergeMaid (*http://astor.som.jhmi.edu/MergeMaid/*) and RankProd (Hong et al. 2006), has been developed for microarray meta analysis. The aforementioned studies investigate predictive model building. However biomarker selection, which is of critical interest, is either ignored or carried out with relative ineffective approaches.

On the other hand, for microarray data generated under a single experimental setting, simultaneous biomarker selection and estimation has been extensively investigated. Examples include the Lasso method (Ghosh and Chinnaiyan 2004; Gui and Li 2005a), the Threshold Gradient Directed Regularization–TGDR (Gui and Li 2005b; Ma and Huang 2005, 2007), and the support vector machine with SCAD penalty (Zhang et al. 2006). We refer to Li (2007) for a thorough review of existing methods. Regularized approaches are capable of selecting a small number of influential genes along with model construction. In terms of biomarker selection, they are usually much more efficient than simply detecting differential genes. Although great successes have been demonstrated by these approaches, they cannot be used directly in meta analysis, since different platforms and experimental settings are not directly comparable.

In this paper, we propose a model-based method, Meta Threshold Gradient Descent Regularization (MTGDR), for biomarker selection in microarray meta analysis. The MTGDR takes advantage of recent development in regularized biomarker selection methods (with a single microarray dataset) and is capable of analyzing several datasets generated under different experimental settings. It thus fills the gap between the available meta analysis methods and single-dataset regularized biomarker selection methods. Compared to the available meta analysis methods, the MTGDR is model-based, allows for experiment-specific estimates and can be more efficient. Compared to the single-dataset regularized biomarker selection methods, the MTGDR has the desired flexibility of allowing for different estimates/models for different experiments and thus can accommodate different experimental setups.

Notations and data settings are first introduced in Section 2. The MTGDR algorithm is described in Section 3. We demonstrate the proposed MTDGR on microarray data when binary clinical outcomes are available, although it is also applicable to other quantitative outcomes. We analyze four pancreatic cancer experiments in Section 4 and four liver cancer experiments in Section 5. Discussions are provided in Section 6.

2 DATA AND MODEL

For simplicity of notation, we assume that the same set of d genes are measured in all M different experiments with $M > 1$. We postpone discussions of possibly different sets of genes from different experiments to the Discussions section. Let $Y^1, \ldots, Y^M$ be the clinical outcomes and let $Z^1, \ldots, Z^M$ represent the gene expressions measured. For $m = 1, \ldots, M$, we assume Y^m is associated with Z^m via the model $Y^m \sim \phi(Z^{m\prime}\beta^m)$, where β^m is the regression coefficient, $Z^{m\prime}$ denotes the transpose of Z^m, and ϕ is the known link function.

We assume the same link function across different experiments. This assumption has been generally made in meta analysis. However, we allow for different regression coefficients β^m, and hence different models for different experiments. The rationale is that one unit gene expression change in experiment 1 (say for example a cDNA study) may not be equivalent to one unit change in experiment 2 (say for example an Affymetrix study). The regression coefficients, which measure the degree of associations, should be different. This assumption shares the same spirits as the fixed effect models in standard meta analysis (Stevens and George 2005).

We choose data with binary outcomes as an example. We note that the proposed MTGDR is generally applicable, regardless of clinical outcome types and statistical models. For experiment m, let $Y^m = 1$ denote the presence and $Y^m = 0$ denote the absence of disease. We assume the commonly used logistic regression model, where the logit of the conditional probability is $logit(P(Y^m = 1|Z^m)) = \alpha^m + Z^{m\prime}\beta^m$. Here α^m is the unknown intercept.

Suppose that there are n_m iid observations in experiment m. The log-likelihood is:

$$R^m(\alpha^m, \beta^m) = \sum_{j=1}^{n_m} Y_j^m \log \frac{\exp(\alpha^m + \beta^{m\prime} Z_j^m)}{1 + \exp(\alpha^m + \beta^{m\prime} Z_j^m)} + (1 - Y_j^m) \log \frac{1}{1 + \exp(\alpha^m + \beta^{m\prime} Z_j^m)} \quad . \qquad (1)$$

Since the intercept α^m will not be subject to regularization, for simplicity, we rewrite $R^m(\alpha^m, \beta^m)$ as $R^m(\beta^m)$.

3 MTGDR METHOD

3.1 Regularized microarray biomarker selection

Although thousands of genes are surveyed in microarray studies, usually there are only a portion of them that are actually associated with the clinical outcome of interest. This is the basis for biomarker selection and the sparsity assumption, which states that most components of the regression coefficient β^m are zero.

The proposed MTGDR is related to the TGDR, which is introduced by Friedman and Popescu (2004) in the context of linear regression and has been used for biomarker selection in microarray classification (Ma and Huang 2005) and survival analysis (Gui and Li 2005b). Compared with other methods such as the Lasso, the TGDR approach is a non-linear boosting-like approach. It can be used to analyze a single dataset, or pooled dataset by simply merging different datasets. However, it is not a meta analysis method.

3.2 MTGDR algorithm

We propose the MTGDR approach for regularized biomarker selection in microarray meta analysis. We first make the following two basic assumptions:

(S1) the sets of genes with nonzero coefficients (i.e., the identified genes) are the same across experiments.

(S2) although the same logistic regression model holds, the nonzero components of the regression coefficients β^m may be not equal across experiments;

Assumption (S1) assumes that although different experiments are not directly comparable, the biological conclusions should be comparable, i.e, we should conclude the same sets of genes to be significantly associated with the outcome across different experiments. Assumption (S2) is mainly due to the concern of different experimental setups, especially platforms;

Let $\beta = (\beta^1, \ldots, \beta^M)$ and $R(\beta) = R^1(\beta^1) + \ldots + R^M(\beta^M)$. Here β is a $d \times M$ matrix. Denote $\Delta\nu$ as the small positive increment as in ordinary gradient descent searching. In the implementation of this algorithm, we choose $\Delta\nu = 10^{-3}$, Let $\beta^m(\nu)$ denote the parameter estimate of β^m corresponding to ν. For a fixed threshold $0 \leq \tau \leq 1$, the MTGDR algorithm can be summarized as follows.

1. Initialize $\beta = 0$ (component-wise) and $\nu = 0$.
2. With current estimate β, compute the $d \times M$ negative gradient matrix $g(\nu) = -\partial R(\beta)/\partial\beta$, where the (j, m) element of g is $g_{j,m}(\nu) = -\partial R^m(\beta^m)/\partial\beta_j^m$.
3. Compute the length d vector of meta gradient G, where the j^{th} component of G is $G_j(\nu) = \sum_{m=1}^M g_{j,m}(\nu)$.

4. Compute the meta threshold vector $F(\nu)$ of length d, where the j^{th} component of $F(\nu)$:

$$F_j(\nu) = I(|G_j(\nu)| \geq \tau \times max_l |G_l(\nu)|).$$

5. Update the (j, m) element of β: $\beta_{j,m}(\nu + \Delta\nu) = \beta_{j,m}(\nu) - \Delta\nu g_{j,m}(\nu) F(\nu)$ and update ν by $\nu + \Delta\nu$.
6. Steps 2-5 are iterated k times, where k is determined by cross validation.

The MTGDR algorithm shares some similarities with the TGDR: it starts with the zero estimate and coefficients for important genes (defined as those with large meta gradients) are updated at each iteration. The tuning parameters τ and k jointly determine the property of β and hence the property of biomarker selection. When $\tau \approx 0$, β is dense even for small values of k. When $\tau \approx 1$, β is sparse for small k and remains so for a relatively large number of iterations, but will become dense eventually. At the extreme when $\tau = 1$, the MTGDR usually updates estimate for a single gene at each iteration, which is similar to the Lasso approach. When τ is in the middle range, the characteristics of β are between those for $\tau = 0$ and $\tau = 1$. For $\tau \neq 0$, biomarker selection can be achieved with cross validated, finite k, by having certain components of β exactly zero.

In step 3, the meta gradient, which is defined as the sum across different experiments, is computed. A meta threshold vector is computed in step 4. By doing so, we force the threshold vector to be the same for each gene across experiments. So when a gene is selected, it is selected in all models across experiments, which corresponds to assumption (S1). In steps 2 and 5, the gradients are computed for each experiment (dataset) and estimates are updated accordingly. By doing so, we allow different estimates (and hence different models) for different experiments, which corresponds to assumption (S2).

The meta gradient in step 3 is the most straightforward definition that considers the common effects across all experiments. Consider for example gene 1 only shows significant effect in experiment 1; whereas gene 2 shows moderate negative effects in all experiments. Then the sum of gradients (combined effects) for gene 2 may be larger than that for gene 1. Gene 2 is thus more likely to be selected since consistent effects are demonstrated across experiments, whereas gene 1 may demonstrate significant effect in experiment 1 purely by chance. If a gene shows significant effects in all experiments but the gradients have both positive and negative signs, then the sum may be small and hence this gene may not be selected. The rationale is that if a gene is selected, it is supposed to show similar biological effects across experiments (for example, up-regulation of this gene is positively associated with the clinical outcome). If both positive and negative associations are observed, then inconsistent biological conclusions are reached in different experiments. The corresponding gene thus should not be selected. It is worth pointing out that with the proposed MTGDR, it is still possible that a gene selected has coefficients with different signs in different experiments. For example, if a gene demonstrates dramatically large positive effect in one study but no or small negative effect in other studies, this gene may still be selected.

3.3 Tuning parameter selection

As in Ma and Huang (2005), we use the V-fold cross validation to select the optimal k and τ. For $\tau = 0, 0.05, \ldots, 0.95, 1$, we search over integer k to maximize the V-fold cross validation objective function, which can be defined following Ma and Huang (2005) and is omitted here. With the V-fold cross validation, the tuning parameters with the best predictive power are selected. Partial protection against over-fitting is also provided. In our study, we set $V = 3$ mainly due to the small sample sizes consideration.

3.4 Evaluation

The proposed MTGDR is a statistical method for biomarker selection using multiple microarray data sets. In general, biomarker selection results ultimately should be validated biologically and by use of independent data (Grutzmann et al. 2005; Choi et al. 2004). Statistically, with the available data, we consider the following Leave-One-Out (LOO) cross validation approach which provides an unbiased evaluation of the prediction performance.

We first remove one subject from the data. With the reduced data, we first carry out cross validation and MTGDR estimation and then use this estimate to make prediction for the one removed record. With the logistic model, the predicted probability can be computed. We use 0.5 as the cutoff and predict the class label. We repeat this procedure over all subjects and calculate the prediction error.

In our numerical study, gene screening is carried out before the MTGDR estimation. As pointed out in Ma (2006), gene screening can be classified as un-supervised (where associations between genes and clinical outcome are not considered) and supervised (where associations between genes and clinical outcome are considered). As pointed in Simon et al. (2003) and Ma (2006), supervised gene screening uses information on the clinical outcome and may lead to over-optimistic prediction evaluation. On the other hand, unsupervised screening is mainly due to technical concerns and usually does not cause biased evaluation. In our study, we first carry out the un-supervised screening. Then in the LOO evaluation, we carry out the supervised screening for each reduced data (with one subject removed) separately. So for each reduced data, the sets of genes that pass the screening may be different.

3.5 A graphic demonstration

We use the following small numerical example to demonstrate the MTGDR parameter paths. For $m = 1, 2$ and 3, we generate data from $logit(P(Y^m = 1|Z^m) = \beta_1^m Z_1^m + \beta_2^m Z_2^m + \beta_3^m Z_3^m + \beta_4^m Z_4^m$. In this simulated meta analysis, there are three independent experiments and four genes per experiment. Z_i^js are generated independently and $N(0, 0.5)$ distributed. We set $\beta^1 = (2.0, 2.0, 0, 0)$, $\beta^2 = (1.5, 1.5, 0, 0)$ and $\beta^3 = (1.0, 1.0, 0, 0)$. In all three experiments, only the first two genes are associated with the binary outcomes, and their corresponding coefficients are different. We simulate 50 observations under each experiment.

The 3-fold cross validation select $\tau = 1.0$ and $k = 620$. We show in Figure 1 the parameter paths as a function of k for $\tau = 1.0$. Individual parameter paths are similar to Lasso paths. Similar properties have been demonstrated for the TGDR in Friedman and Popescu (2004). We can see that for any k the estimated coefficients for one gene are either all zero or all nonzero across experiments. This corresponds to assumption (S1). We also note that for a specific gene with nonzero coefficients, the estimated coefficients are different across experiments, as required in assumption (S2).

Table 1. Pancreatic cancer study. PDAC: number of PDAC samples. Normal: number of normal samples. Array: type of array used. UG: number of unique UniGene clusters.

Dataset	P1	P2	P3	P4
Author	Logsdon	Friess	Iacobuzio-Donahue	Crnogorac-Jurcevic
PDAC	10	8	9	8
N	5	3	8	5
CP	5	8	–	–
Array	Affy. HuGeneFL	Affy. HuGeneFL	cDNA Stanford	cDNA Sanger
UG	5521	5521	29621	5794

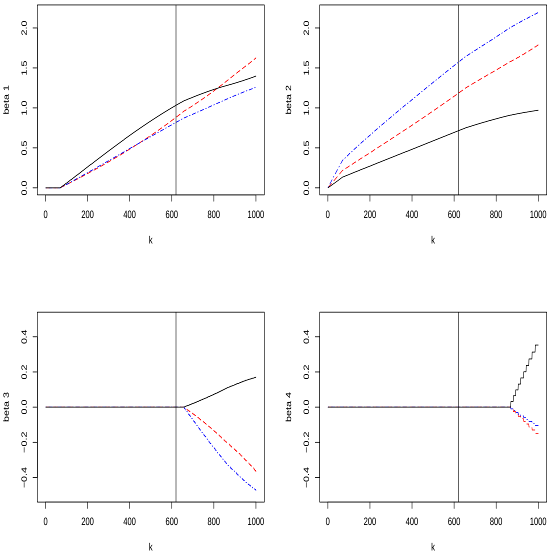

Fig. 1. Parameter path as a function of k. Dashed red line: simulated experiment 1; Dash-dotted blue line: simulated experiment 2; Solid black line: simulated experiment 3. Vertical lines: cross-validated k.

4 PANCREATIC CANCER STUDY

4.1 Data

Pancreatic ductal adenocarcinoma (PDAC) is a major cause of malignancy-related death. Apart from surgery, there is still no effective therapy and even resected patients usually die within one year postoperatively. Several experiments have applied microarray technology to pancreatic cancer, targeting to identifying predictive pancreatic cancer biomarkers. We use four datasets in our study: Iacobuzio-Donahue et al. (2003), Logsdon et al. (2003), Crnogorac-Jurcevic et al. (2003) and Friess et al. (2003). These four datasets have been selected and used in the meta analysis of Grutzmann et al. (2005). We show data descriptions in Table 1. Two of the four studies use cDNA arrays and two use oligonucleotide arrays. Cluster ID and gene names are assigned to all of the cDNA clones and Affymetrix probes based on UniGene Build 161. The two sample groups considered in our analysis are PDAC and normal pancreatic tissues. Data on chronic pancreatitis are available for Logsdon et al. (2003) and Friess et al. (2003). As in Grutzmann et al. (2005), those CP samples are not used in our study.

Table 2. Pancreatic cancer datasets: estimations for genes with nonzero coefficients.

UniGene	P1	P2	P3	P4
Hs.107	-0.078	-0.074	-0.096	-0.062
Hs.12068	-0.265	-0.387	-0.189	-0.250
Hs.16269	0.038	0.055	0.060	0.017
Hs.169900	-0.879	-0.992	-0.693	-0.775
Hs.180920	-0.144	-0.244	-0.223	-0.189
Hs.241257	0.096	0.128	0.124	0.062
Hs.287820	1.051	1.157	1.055	0.736
Hs.317432	-0.023	-0.012	-0.053	-0.022
Hs.5591	-0.082	-0.170	-0.149	-0.149
Hs.62	0.111	0.100	0.104	0.126
Hs.66581	-0.024	-0.028	-0.034	-0.013
Hs.75335	-0.270	-0.259	-0.250	-0.250
Hs.76307	0.435	0.303	0.616	0.416
Hs.78225	0.011	0.010	0.018	0.010
Hs.83383	-0.074	-0.094	-0.066	-0.085

Data pre-processing, including normalization, is carried out for each experiment separately. We refer to the original articles for details on data preparation. Grutzmann et al. (2005) identifies a consensus set of 2984 UniGene IDs. Our dataset is provided by Dr. Grutzmann and contains the same set of 2984 genes. We carry out un-supervised screening and remove genes with more than 30% missingness in any of the four datasets. There are 1204 genes that pass this screening.

For Affymetrix expression measurements, we add a floor of 10 and make log2 transformations. We fill in missing values with medians across samples (for each dataset separately), and then standardize each gene to zero mean and unit variance. For cDNA studies, we fill in missing values with sample medians for each dataset separately, and then standardize each gene to zero mean and unit variance.

4.2 Alternative approaches

Individual TGDR. The TGDR itself is not a meta analysis method. To illustrate this point, we analyze each dataset separately using the TGDR, following Ma and Huang (2007). For the four datasets, 7 (P1), 10 (P2), 6 (P3) and 1 (P4) genes are identified, respectively. There is only one common gene identified in both P2 and P3. Otherwise there is no overlap between the four sets of selected genes. Therefore, straightforward application of the original TGDR method to each individual data set will not yield a set of genes with common effect across different datasets. Similar phenomenon has been observed with other single-dataset biomarker selection methods.

A simple method. We compute the two-sample t-statistic for each gene and each dataset. We then assign a rank for each gene and each dataset, based on the t-statistic. The overall rank for one gene is defined as the sum of ranks across four datasets. The five genes with the largest ranks are identified. Only five genes are selected because of the small sample sizes, especially for data P2. With this simple method, assumption (S1) is satisfied. We then construct experiment-specific logistic models (assumption S2). With the LOO approach, a total of 4 subjects cannot be properly predicted.

Pooled TGDR analysis. We ignore the fact that the four datasets are from different studies and different platforms, and simply pool them together. The sample size of the pooled dataset is 56. We apply the TGDR to the pooled dataset. A total of 22 genes are identified in the final model. The LOO approach misclassifies 2 subjects. As has been noted in Grutzmann et al. (2005), the four selected datasets are relatively easy to classify. This partly explains the satisfactory prediction performance of the simple pooling.

4.3 MTGDR analysis

We employ the MTGDR. Tuning parameters are chosen via the 3-fold cross validation. 15 genes are identified. We show the gene IDs and corresponding estimates in Table 2. We can observe from Table 2 that (1) if a gene has nonzero coefficient for one dataset, then it has nonzero coefficients for all datasets, as required by (S1); (2) the estimated coefficients for one gene can be different across experiments; this corresponds to (S2) and is the extra flexibility allowed by the MTGDR compared with the pooled analysis; and (3) although the estimated coefficients may be different for one gene across experiments, their signs are the same. The same signs lead to similar biological conclusions – i.e., whether up-regulations of this gene are positively or negatively associated with the clinical outcome of interest. Prediction error is computed using the LOO approach described in Section 3.4. There are 2 subjects that cannot be properly predicted.

Individual TGDR (and other single-dataset methods) cannot achieve consistent biomarker identification across experiments and will not be further discussed here. The simple method achives inferior prediction. In addition, the number of biomarkers identified is limited by the smallest sample size across experiments. The MTGDR identifies 15 genes, which is fewer than the pooled analysis. Previous microarray studies (Ghosh and Chinnaiyan 2005; Li 2007; Ma and Huang 2005) have been targeting to constructing parsimonious models, which can lead to a more focused hypothesis for further investigation (for example a shorter list of genes to be studied). The MTGDR outperforms the pooled analysis in this sense.

5 LIVER CANCER STUDY

5.1 Data

Gene expression profiling studies have been conducted on hepatocellular carcinoma (HCC), which is among the leading causes of cancer death in the world. A microarray meta analysis is carried out in Choi et al. (2004), where the main goal is to detect differentially expressed genes. We study datasets D1–D4 in Choi et al. (2004). Data information is provided in Table 3. Datasets D1–D4 were generated in three different hospitals in South Korea. Although the studies were conducted in a controlled setting, Choi et al. (2004) "failed to directly merge the data even after normalization of each dataset."

In studies D1–D3, expressions of 10336 genes are measured. In study D4, expressions of 9984 genes are measured. We focus on the 9984 genes that are measured in all four studies. For each dataset, the within-print-tip-group normalization is first carried out (Choi et al. 2004). We then pre-process the data as follows.

(1) Un-supervised screening:
(1.1) if a gene has more than 30% of missing in any dataset, then this gene is removed from downstream analysis. 3122 out of 9984 genes pass this screening.
(1.2) if a subject has more than 30% missing expressions for the 3122 genes, then this subject is removed. 8 subjects are removed, leading to an effective sample size of 125. We show the number of subjects actually used in the analysis in Table 3.

(2) For each dataset, we then fill in missing expression values with medians across samples.

(3) Supervised screening: we compute the two-sample t-statistic for each gene and each dataset. We then assign a rank for each gene and each dataset, based on the t-statistic. The overall rank for one gene is defined as the sum of ranks for all four datasets. The 1000 genes with the highest ranks are selected for downstream analysis. This rank based screening shares similar spirits as Hong et al. (2006).

(4) Normalize each gene expression to zero mean and unit variance.

Gene screening is conducted to exclude genes which are not likely to be influential. Similar procedure has been used in Ma and Huang (2005) and references therein. The proposed MTGDR has no limitation on the number of genes that can be used in the analysis.

5.2 Alternative approaches

Individual TGDR. With the individual TGDR, 27 (D1), 10 (D2), 20 (D3) and 6 (D4) genes are identified, respectively. The gene sets identified are quite different. For example, the gene sets identified from datasets D1 and D2 have no overlap. Assumption (S1) is seriously violated.

A simple method. We employ the simple method described in Section 4.2. With larger sample sizes, we select the top ten genes. The LOO prediction error is 0.26.

Pooled TGDR analysis. We simply pool the four datasets together after data pre-processing. The four liver datasets are generated in similar experimental settings and are expected to behave similarly. Using the TGDR for regularization, a total of 34 genes are selected in the final model. There are 36 subjects that are misclassified. The LOO prediction error is 0.29.

5.3 MTGDR analysis

We analyze the liver cancer data using the proposed MTGDR, with optimal tuning parameters selected using the 3-fold cross validation. 34 genes are identified. We provide the gene information and corresponding estimates in Table 4. We can see that Table 4 has similar characteristics as Table 2. However, for a small number of genes, the signs of the four estimates can be different. For example, for gene 15.4.E1/Rab9 effector p40, three out of four estimated coefficients

Table 3. Liver cancer study. Tumor: number of tumor samples. Normal: number of normal samples. Numbers in the "()" are the number of subjects used in the analysis. Ver. 2 chips have different spot locations from Ver. 1 chips.

Dataset	D1	D2	D3	D4
Experimenter	Hospital A	Hospital B	Hospital C	Hospital C
# tumor	16 (14)	23	29	12 (10)
# normal	16 (14)	23	5	9(7)
Chip type	cDNA(Ver.1)	cDNA(Ver.1)	cDNA(Ver.1)	cDNA(Ver.2)
(Cy5:Cy3)	sample:normal liver	sample:placenta	sample:placenta	sample:sample

Table 4. Liver cancer datasets: estimations for genes with nonzero coefficients.

Gene Information	D1	D2	D3	D4
1.2.F.7/noseq/	-0.076	-0.100	-0.078	-0.035
1.3.A.8/clone MGC:5207 IMAGE:2901089	0.147	0.199	0.030	0.054
10.1.B.9/cDNA FLJ20844 fis, clone ADKA01904	-0.020	-0.016	-0.002	-0.002
11.3.F.6/noseq/	-0.275	-0.519	-0.225	-0.170
15.1.G.7/Cyt19 protein (Cyt19), mRNA	0.023	0.019	-0.001	0.009
15.2.D.10/EST387826 cDNA	-0.041	-0.031	-0.003	-0.015
15.3.E.9/hypothetical protein MGC11287	0.016	0.034	0.015	0.014
15.4.E.1/Rab9 effector p40 (RAB9P40), mRNA	0.166	0.243	-0.012	0.083
17.2.B.11/ATPase, H+ transporting, lysosomal 9kD	0.145	0.258	0.108	0.020
18.3.F.6/nomatch/	0.072	0.073	0.070	0.045
19.1.G.5/Ras association (RalGDS/	0.168	0.176	-0.036	0.042
2.2.E.11/triosephosphate isomerase 1 (TPI1), mRNA	0.012	0.012	0.004	0.011
2.2.G.10/UDP-glucose pyrophosphorylase 2 (UGP2)	-0.296	-0.274	-0.043	-0.178
21.3.A.4/noseq/	0.016	0.011	0.002	0.001
23.3.H.1/thioredoxin-like, 32kD (TXNL)	0.285	0.226	0.066	0.033
25.2.A.5/noseq/	0.016	0.014	0.001	0.009
26.2.D.2/adipose differentiation-related protein (ADFP)	-0.169	-0.114	-0.219	-0.118
26.4.B.5/Human zyxin related protein ZRP-1 mRNA	0.161	0.127	0.042	0.070
3.2.E.10/Human G protein-coupled receptor V28 mRNA	-0.707	-0.589	-0.359	-0.375
4.1.D.1/multiple endocrine neoplasia I (MEN1), mRNA	-0.086	-0.075	-0.130	-0.090
4.2.H.5/solute carrier family 22, member 1	-0.014	-0.120	-0.144	-0.092
4.3.C.1/noseq/	-0.058	-0.020	-0.008	0.007
4.4.B.9/noseq/	-0.438	-0.670	-0.460	-0.502
5.1.A.9/noseq/	-0.001	-0.007	-0.002	-0.001
5.1.D.1/malate dehydrogenase 2, NAD (mitochondrial)	0.135	0.043	0.063	0.060
6.2.E.3/tubulin, beta polypeptide (TUBB), mRNA /	0.024	0.012	0.004	0.011
6.3.B.3/noseq/	0.104	0.104	-0.023	0.015
6.4.D.11/non-metastatic cells 2, protein expressed NME2	0.053	0.072	0.020	0.025
6.4.F.5/H2A histone family, member Z (H2AFZ), mRNA	0.047	0.062	-0.001	0.042
7.3.A.5/nomatch/	-0.329	-0.432	-0.297	-0.222
7.3.G.9/guanine nucleotide binding protein, q polypeptide	0.073	0.019	0.049	0.029
8.2.B.11/cystatin B (stefin B) (CSTB), mRNA	0.040	0.112	0.051	0.046
8.2.D.8/RNA helicase-related protein (RNAHP), mRNA	-0.739	-1.369	-1.002	-1.140
8.3.A.7/proline-rich Gla polypeptide 2	-0.001	-0.019	-0.024	-0.026

are positive, and one is negative. As discussed above, different signs of estimates may indicate conflicting biological conclusions. However, we observe that the negative coefficient is very small, which may be caused by random variations.

Prediction performance is evaluated using the LOO approach. There are 20 subjects that are misclassified, leading to a prediction error of 0.16, which is a significant improvement over the simple approach and pooled TGDR.

6 DISCUSSIONS

Multiple microarray experiments have been conducted to investigate relationship between the same clinical outcomes and gene expressions. Individual experiments usually have small sample sizes and may not have sufficient statistical power. It is thus critical to develop microarray meta analysis methods that can effectively combine multiple datasets. In this article, we propose the MTGDR method for biomarker selection in microarray meta analysis. The MTGDR is a novel extension of the TGDR. The MTGDR is designed for microarray meta-analysis and is computationally affordable for even extremely high dimensional data.

An important feature of the MTGDR is that it can accommodate different platforms and experimental settings, and is capable of finding genes that have consistent effects across different experiments. The MTGDR achieves this task in an efficient way by combining all the available data sets in the analysis, while properly taking into account possible heterogeneity among different experiments. Our simulation and real data example demonstrate the superiority of this combined approach for finding genes with consistent effects than individual dataset-based analysis. It is clear that the genes that are found to have consistent effects in multiple data sets should be the priority targets for verification based on different platforms or using independent data.

In our data analysis, the same sets of genes across experiments are considered. When different sets of genes are included in different experiments, the MTGDR is still applicable by setting gradients for missing genes zero. However meta analysis will be less powerful, if the sets of genes measured differ greatly.

We have considered studies with binary outcome and the logistic regression model only. The MTGDR method is generally applicable, as long as the objective function $R(\beta)$ is well defined and differentiable. For example, the MTGDR can be applied to find a common set of genes that influence a continuous outcomes including censored survival data based on multiple microarray gene expression data sets.

ACKNOWLEDGMENT

We would like to thank two anonymous reviewers for very insightful comments, which have led to a significant improvement of this article. The work of Huang is partially supported by the NIH/NCI grant P30 CA 086862.

REFERENCES

Choi, J., Choi, J., Kim, D., Choi, D., Kim, B., Lee, K., Yeom, Y. Yoo, H., Yoo, O. and Kim, S. (2004) Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Letters* **565**, 93-100.

Conlon, E.M., Song, J.J. and Liu, A. (2007) Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics* **8**: 80.

Crnogorac-Jurcevic, T., Missiaglia, E., Blaveri, E., Gangeswaran, R., Jones, M., Terris, B., Costello, E., Neoptolemos, J.P. and Lemoine, N.R. (2003) Molecular alterations in pancreatic carcinoma: expression profiling shows that dysregulated expression of S100 genes is highly prevalent. *Journal of Pathology* **201**, 63-74.

Dettling, M. and Buhlmann, P. (2003) Boosting for tumor classification with gene expression data. *Bioinformatics* **9**, 1061-1069.

Friedman, J. and Popescu, B.E. (2004) Gradient directed regularization. *Technical Report, Stanford Department of Statistics.*

Friess, H., Ding, J., Kleeff, J., Fenkell, L., Rosinski, J.A., Guweidhi, A., Reidhaar-Olson, J.F., Korc, M., Hammer, J. and Buchler, M.W. (2003) Microarray-based identification of differentially expressed growth-and metastasis-associated genes in pancreatic cancer. *Cellular and Molecular Life Sciences* **60**, 1180-1199.

Fung, B. and Ng, V. (2004) Meta-classification of multi-type cancer gene expression data. *Proceeding of 4th Workshop on Data Mining in Bioinformatics*, 31-39.

Ghosh, D. and Choi, H. (2006) metaArray package for meta analysis of microarray data. *R package. http://www.r-project.org.*

Ghosh, D. and Chinnaiyan, A. (2004) Classification and selection of biomarkers in genomic data using LASSO. *Journal of Biomedicine and Biotechnology* **2**, 147-154.

Ghosh, D., Barette, T.R., Rhodes, D. and Chinnaiyan, A.M. (2003) Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Functional and Integrative Genomics* **3**, 180-188.

Grutzmann, R., Boriss, H., Ammerpoh, O., Luttges, J., Kalthoff, H., Schackert, H., Kloppel, G., Saeger, H. and Pilarsky, C. (2005) Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene*, 1-10.

Gui, J. and Li, H. (2005a) Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* **21**, 3001-3008.

Gui, J. and Li, H. (2005b) Threshold gradient descent method for censored data regression, with applications in pharmacogenomics. *Pacific Symposium on Biocomputing* **10**, 272-283.

Hong, F., Breitling, R., McEntee, C.W., Witter, B.S., Nemhauser, J.L. and Chory, J. (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* **22**, 2825-2827.

Iacobuzio-Donahue, C.A., Ashfaq, R., Maitra, A., Adsay, N.V., Shen-Ong, G.L., Berg, K., Hollingsworth, M.A., Cameron, J.L., Yeo, C.J., Kern, S.E., Goggins, M. and Hruban, R.H. (2003) Highly expressed genes in pancreatic ductal adenocarcinomas: a comprehensive characterization and comparison of the transcription profiles obtained from three major technologies. *Cancer Research* **63**, 8614-8622.

Jiang, H., Deng, Y., Chen, H., Tao, L., Sha, Q., Chen, J., Tsai, C. and Zhang, S. (2004) Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* **5**: 81.

Jung, Y., Oh, M., Shin, D., Kang, S. and Oh, H. (2006) Identifying differentially expressed genes in meta-analysis via Bayesian model-based clustering. *Biometrical Journal* **48**, 435-450.

Kuo, W.P, Jenssen, T-K, Butte, A.J., Ohno-Machado, L. and Kohane, I.S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405-412.

Li, H. (2007) Censored data regression in high-dimension and low sample size settings for genomic applications. *Statistical Advances in Biomedical Sciences: State of Art and Future Directions. Edited by A. Biswas, S. Datta, J. Fine and M. Segal, in press.*

Logsdon, C.D., Simeone, D.M., Binkley, C., Arumugam, T., Greenson, J., Giordano, T.J., Misek, D. and Hanash, S. (2003) Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentially regulated in pancreatic cancer. *Cancer Research* **63**, 2649-2657.

Ma, S. (2006) Empirical study of supervised gene screening. *BMC Bioinformatics* **7**:537.

Ma, S. and Huang, J. (2005) Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics* **21**, 4356-4362.

Ma, S. and Huang, J. (2007) Clustering threshold gradient descent regularization: with applications to microarray studies. *Bioinformatics* **23**, 466-472.

Ma, S., Song, X. and Huang, J. (2007) Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics* **8**:60.

Stevens, J.R. and George, R.W. (2005) Meta-analysis combines Affymetrix microarray results across laboratories. *Comparative and Functional Genomics* **6**, 116-122.

Simon, R., Radmacher, M.D., Dobbin, K. and McShane, L.M. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *JNCI* **95**, 14-18.

Wang, J., Coombes, K.R., Highsmith, W.E., Keating, M.J. and Abruzzo, L.V. (2004) Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics* **17**, 3166-3178.

Warnet, P., Eils, R. and Brors, B. (2005) Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics* **6**: 265.

Zhang, H., Ahn, J., Lin, X. and Park, C. (2006) Gene selection using support vector machines with non-convex penalty. *Bioinformatics* **22**, 88-95.