Integrative Analysis of Multiple Cancer Prognosis Datasets Under the Heterogeneity Model

Jin Liu, Jian Huang, and Shuangge Ma

Abstract In cancer research, genomic studies have been extensively conducted, searching for markers associated with prognosis. Because of the "large d, small n" characteristic, results generated from the analysis of a single dataset can be unsatisfactory. Integrative analysis simultaneously analyzes multiple datasets and can be more effective than the analysis of single datasets and classic meta-analysis. In many existing integrative analyses, the homogeneity model has been assumed, which postulates that different datasets share the same set of markers. In practice, datasets may have been generated in studies that differ in patient selection criteria, profiling techniques, and many other aspects. Such differences may make the homogeneity model too restricted. Here we explore the heterogeneity model, which assumes that different datasets may have different sets of markers. With multiple cancer prognosis datasets, we adopt the AFT (accelerated failure time) models to describe survival. A weighted least squares approach is adopted for estimation. For marker selection, penalization-based methods are examined. These methods have intuitive formulations and can be computed using effective group

J. Liu
Department of Biostatistics, School of Public Health, Yale University,
New Haven, CT 06520, USA
e-mail: jin.liu.jl2329@yale.edu

J. Huang
Department of Statistics and Actuarial Science, University of Iowa,
2 W Jefferson St, Iowa City, IA 52246, USA
e-mail: jian-huang@uiowa.edu

S. Ma (✉)
Department of Biostatistics, School of Public Health, Yale University,
New Haven, CT 06520, USA

VA Cooperative Studies Program Coordinating Center, West Haven, CT, USA
e-mail: shuangge.ma@yale.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_21, © Springer Science+Business Media New York 2013

coordinate descent algorithms. Analysis of three lung cancer prognosis datasets with gene expression measurements demonstrates the merit of heterogeneity model and proposed methods.

1 Introduction

Genomic studies have been extensively conducted, searching for markers associated with the prognosis of cancer. Data generated in such studies have the "large d, small n" characteristic, with the number of genes profiled d much larger than sample size n. In addition, in whole-genome studies, only a subset of the profiled genes are expected to be associated with prognosis. Thus, the analysis of cancer prognosis data with genomic measurements demands regularized estimation and selection.

In practical data analysis, genomic markers identified from the analysis of single datasets are often unsatisfactory. Multiple factors contribute to the unsatisfactory performance, including the highly noisy nature of cancer genomic data, technical variations of profiling techniques and, more importantly, the small sample sizes of individual studies. Recent studies have shown that pooling and analyzing multiple studies may effectively increase sample size and improve properties of the identified markers (Guerra and Goldsterin 2009; Ma et al. 2009, and references therein). Multi-dataset methods include meta-analysis and integrative analysis methods. Integrative analysis pools and analyzes raw data from multiple studies and can be more informative than classic meta-analysis, which analyzes multiple studies separately and then pools summary statistics (lists of identified genes, p-values, effect sizes, etc.).

In studies such as Ma et al. (2011b), the homogeneity model has been assumed. Under this model, multiple datasets share the same set of markers. This model has also been adopted with cancer diagnosis studies and categorical responses (Ma et al. 2011a and references therein). In practical data analysis, when multiple datasets are generated in independent studies, heterogeneity (in patients' characteristics, technical aspects such as profiling protocols, etc) inevitably exists. Such heterogeneity may make the homogeneity model too restricted. In addition, data analyses in Ma et al. (2011a;b) show that for some of the identified genes, the magnitudes of estimated regression coefficients may vary significantly across datasets. It is possible that the very small regression coefficients are actually zero. Such an observation further suggests the necessity of relaxing the homogeneity model assumption.

In this study, we describe cancer survival using AFT (accelerated failure time) models. Compared with alternatives such as the Cox model, the AFT model has a significantly simpler objective function and lower computational cost, which is especially desirable with high-dimensional data. In addition, its regression coefficients may have more lucid interpretations. As an alternative to the homogeneity model, we consider the heterogeneity model. It includes the homogeneity model as a special case and can be more flexible. For marker selection, we adopt

penalization. The proposed penalization methods are intuitively reasonable and can be computationally realized using the group coordinate descent algorithms. This study complements the existing ones by conducting integrative analysis under the more flexible heterogeneity model and by adopting penalization methods tailored to this model.

2 Integrative Analysis of Cancer Prognosis Studies

2.1 Data and Model Settings

Assume M independent studies and n^m iid observations in study $m(=1,\ldots,M)$. The total sample size is $n=\sum_{m=1}^{M} n^m$. In study m, denote T^m as the logarithm (or another known monotone transformation) of failure time. Denote X^m as the length-d vector of gene expressions. Although gene expression data is used as an example in this study, it should be noted that the proposed methods are also applicable to studies with other types of genomic measurements. For simplicity of notation, assume that the same set of genes are measured in all M studies. For the ith subject, the AFT model assumes that

$$T_i^m = \beta_0^m + X_i^{m\prime}\beta^m + \epsilon_i^m, i=1,\ldots,n^m. \tag{1}$$

where β_0^m is the intercept, $\beta^m \in \mathbb{R}^d$ is the length-d vector of regression coefficients, and ϵ_i^m is the error term. When T_i^m is subject to right censoring, we observe $(Y_i^m, \delta_i^m, X_i^m)$, where $Y_i^m = \min\{T_i^m, C_i^m\}$, C_i^m is the logarithm of censoring time, and $\delta_i^m = I\{T_i^m \le C_i^m\}$ is the event indicator.

When the distribution of ϵ_i^m is known, the parametric likelihood function can be easily constructed. Here we consider the more flexible case where this distribution is unknown. In the literature, multiple estimation approaches have been developed, including, for example, the Buckley-James and rank-based approaches (Buckley and James 1979; Jin et al. 2003). In this study, we adopt the weighted least squares estimator (Stute 1996), which to the best of our knowledge, has the lowest computational cost. This property is especially desirable with high-dimensional data.

Let $\hat{F}^m$ be the Kaplan–Meier estimator of the distribution function F^m of T^m. $\hat{F}^m(y) = \sum_{i=1}^{n^m} \omega_i^m I\{Y_{(i)}^m \le y\}$, where ω_i^ms are the jumps in the Kaplan–Meier estimator and can be computed as

$$\omega_1^m = \frac{\delta_{(1)}^m}{n^m},\ \omega_i^m = \frac{\delta_{(i)}^m}{n^m - i + 1}\prod_{j=1}^{i-1}\left(\frac{n^m - j}{n^m - j + 1}\right)^{\delta_{(j)}^m},\ i=2,\ldots,n^m.$$

Here $Y_{(1)}^m \le \cdots \le Y_{(n^m)}^m$ are the order statistics of Y_i^ms, and $\delta_{(1)}^m,\ldots,\delta_{(n^m)}^m$ are the associated event indicators. Similarly, let $X_{(1)}^m,\ldots,X_{(n^m)}^m$ be the associated gene

expressions of the ordered Y_i^ms. Stute (1996) proposed the weighted least squares estimator $(\hat{\beta}_0^m, \hat{\beta}^m)$ that minimizes

$$\frac{1}{2}\sum_{i=1}^{n^m} \omega_i^m (Y_{(i)}^m - \beta_0^m - X_{(i)}^{m\prime}\beta^m)^2. \tag{2}$$

We center $X_{(i)}^m$ and $Y_{(i)}^m$ using their ω_i^m-weighted means, respectively. Define

$$\bar{X}_w^m = \sum_{i=1}^{n^m} \omega_i^m X_{(i)}^m / \sum_{i=1}^{n^m} \omega_i^m, \bar{Y}_w^m = \sum_{i=1}^{n^m} \omega_i^m Y_{(i)}^m / \sum_{i=1}^{n^m} \omega_i^m.$$

Let $X_{\omega(i)}^m = \sqrt{\omega_i^m}(X_{(i)}^m - \bar{X}_w^m)$ and $Y_{\omega(i)}^m = \sqrt{\omega_i^m}(Y_{(i)}^m - \bar{Y}_w^m)$, respectively. With the weighted centered values, the intercept is zero. The weighted least squares objective function can be written as

$$L^m(\beta^m) = \frac{1}{2}\sum_{i=1}^{n^m} (Y_{\omega(i)}^m - X_{\omega(i)}^{m\prime}\beta^m)^2. \tag{3}$$

Denote $Y^m = (Y_{\omega(1)}^m, \ldots, Y_{\omega(n^m)}^m)'$ and $X^m = (X_{\omega(1)}^m, \ldots, X_{\omega(n^m)}^m)'$. Further denote $Y = (Y^{1\prime}, \ldots, Y^{M\prime})'$, $X = \text{diag}(X^1, \ldots, X^M)$, and $\beta = (\beta^{1\prime}, \ldots, \beta^{M\prime})'$.

Consider the overall objective function $L(\beta) = \frac{1}{n}\sum_{m=1}^M L^m(\beta^m)$. With this objective function, larger datasets have more contributions, which is intuitively reasonable. When desirable, normalization by sample size can be applied.

2.2 Homogeneity Model and Penalized Selection

In Huang et al. (2012) and Ma et al. (2011a;b), the homogeneity model is adopted to describe the genomic basis of M datasets. Denote β_j^m as the jth component of β^m. Then $\beta_j = (\beta_j^1, \ldots, \beta_j^M)'$ is the length-M vector of regression coefficients representing the effects of gene j in M studies. Under the homogeneity model, for any $j(= 1, \ldots, d)$,

$$I(\beta_j^1 = 0) = \ldots = I(\beta_j^M = 0).$$

That is, if a gene is identified as associated with prognosis in one dataset, it is identified in all of the M datasets. Thus, the M datasets have the same sparsity structure. This is a sensible model when multiple datasets have been generated under the same protocol. With multiple datasets generated independently, if the analysis of individual datasets and examination of the protocols suggest a high degree of similarity, then this model can be adopted.

For marker selection, Ma et al. (2011b) adopts penalization and proposes using the group MCP (gMCP) approach, where the estimate is defined as

$$\hat{\beta} = argmin\left\{L(\beta) + P_{gMCP}(\beta)\right\},$$

with

$$P_{gMCP}(\beta) = \sum_{j=1}^{d} \rho(||\beta_j||_{\Sigma_j}; \sqrt{d_j}\lambda_1, \gamma). \tag{4}$$

$\rho(t;\lambda,\gamma) = \lambda \int_0^{|t|} \left(1 - \frac{x}{\lambda\gamma}\right)_+ dx$ is the MCP penalty (Zhang 2010). $||\beta_j||_{\Sigma_j} = ||\Sigma_j^{1/2}\beta_j||_2$, $||\cdot||_2$ is the L_2 norm, $\Sigma_j = n^{-1}X[\cdot,j]'X[\cdot,j]$, and $X[\cdot,j]$ is the $n \times d_j$ submatrix of X that corresponds to β_j. In (4), d_j is the size of the coefficient group corresponding to gene j. When the M datasets have exactly matched gene sets, $d_j \equiv M$. We keep d_j so that this formulation can accommodate partially matched gene sets. When gene j is not measured in dataset k, we take the convention $\beta_j^k \equiv 0$. $\lambda_1 > 0$ is the tuning parameter, and $\gamma > 0$ is the regularization parameter (Zhang 2010).

Penalty function (4) has been motivated by the following considerations. In this analysis, genes are the functional units. The overall penalty is the sum over d individual penalties, with one for each gene. For gene selection, the MCP penalization is adopted. In single-dataset analysis, MCP has been shown to have performance comparable to or better than some of the alternative penalties. For a specific gene, its effects in the M studies are represented by a "group" of M regression coefficients. Under the homogeneity model, the M studies are expected to identify the same set of genes. Thus, within a group, no further selection is needed, and so the L_2 norm is adopted. Note that here we adopt the $||\cdot||_{\Sigma_j}$ norm, which rescales the regression coefficient vector by the covariance matrix Σ_j, so that the computation can be less ad hoc.

3 Heterogeneity Model and Penalized Selection

3.1 Heterogeneity Model

When multiple datasets are generated in independent studies, heterogeneity inevitably exists (Knudsen 2006). The degree of heterogeneity depends on the differences in study protocols, profiling techniques, and many other factors. In cancer prognosis studies, the effort to unify the sets of identified markers across independent studies has not been very successful (Cheang et al. 2008; Knudsen 2006). This can also be partly seen from the data analysis in Ma et al. (2011b). Such observations raise the question whether the homogeneity model is too restricted and motivates the heterogeneity model. Under the heterogeneity model, one gene can be associated with prognosis in some studies but not others. This model includes the homogeneity model as a special case and can be more flexible.

In addition, there are scenarios under which the homogeneity model is conceptually not sensible, but the heterogeneity model is. The first is where different

studies are on different types of cancers (Ma et al. 2009). On the one hand, different cancers have different prognosis patterns and different sets of markers. On the other hand, multiple pathways, such as apoptosis, DNA repair, cell cycle, and signaling, are associated with the prognosis of multiple cancers. The second scenario is the analysis of different subtypes of the same cancer. Different subtypes have different risks of occurrence and progression, and it is not sensible to reinforce the same genomic basis. The third scenario is where subjects in different studies have different demographic measurements, clinical risk factors, environmental exposures, and treatment regimens. For genes not intervened with those "additional" variables, their importance remains consistent across multiple studies. However, for other genes, they may be important in some studies but not others.

3.2 Penalized Marker Selection

Under the heterogeneity model, the model and regression coefficients have two dimensions. The first is the gene dimension as in other marker selection studies. The second, which is unique to integrative analysis, is the study (dataset) dimension. In marker selection, we need to determine *whether a gene is associated with prognosis in any study at all* as well as *in which studies it is associated with prognosis*. Such an objective demands two-way selection.

3.2.1 A Two-Step Approach

Since integrative analysis under the heterogeneity model calls for two-way selection, a natural strategy is to achieve the selection in two steps, with one step for each way. The first step is to determine whether a gene is associated with prognosis in any study. As $\beta_j = (\beta_j^1, \ldots, \beta_j^M)'$ represent the effects of gene j $(= 1, \ldots, d)$ in M studies, this step of selection amounts to determining whether $\|\beta_j\|_2 = 0$. We propose achieving this step of selection using the *gMCP penalization approach*. With genes selected in the first step (i.e., $\{j : ||\beta_j||_2 \neq 0\}$), in the second step, we determine which prognosis responses (studies) they are associated with. For this step, we propose applying the *MCP approach* to each dataset separately. This step conducts standard single-dataset analysis. Note that although both steps employ the MCP penalties, they may have different tuning and regularization parameters.

3.2.2 Composite Penalization

In the analysis of a single dataset, when there is a grouping structure among covariates, two-way selection can be achieved using composite penalization. The idea is to use a group penalty for variable selection at the group level and a second penalty for variable selection at the within-group level. The composite of the two penalties will then be able to conduct variable selection at both levels.

In integrative analysis with multiple datasets, we adopt a similar strategy. First consider the *1-norm gMCP* approach, where the penalty takes the form

$$P_{1-norm\ gMCP}(\beta) = \sum_{j=1}^{d} \rho(\|\beta_j\|_1; \sqrt{d_j}\lambda, \gamma). \tag{5}$$

$\|\beta_j\|_1 = \sum_{m=1}^{M} |\beta_j^m|$ is the L_1 norm of β_j, which can also be viewed as a Lasso penalty. Other notations have similar implications as with gMCP. With this composite penalty, the outer penalty has a gMCP form. In integrative analysis, it conducts selection at the gene level. The inner penalty is Lasso. For a gene with nonzero effects, it identifies in which study(ies) the gene is associated with prognosis.

In (5), the Lasso penalty is adopted mainly because of its computational simplicity. In single-dataset analysis, it has been shown that MCP can have better properties (for example, more accurate selection) than Lasso. Motivated by such a consideration, we propose the *composite MCP (cMCP)* approach, where the penalty takes the form

$$P_{cMCP}(\beta) = \sum_{j=1}^{d} \rho\left(\sum_{m=1}^{M} \rho(\beta_j^m; \lambda_2, b); \lambda_1, a\right). \tag{6}$$

Here λ_1, λ_2 are the tuning parameters, and a, b are the regularization parameters.

Computational algorithm for cMCP . Below we describe the computational algorithm for cMCP. The two-step and 1-norm gMCP estimates can be computed in a similar manner.

Consider the group coordinate descent algorithm. This algorithm is iterative and optimizes over the regression coefficients of one gene at a time. It cycles through all genes, and the overall iteration is repeated multiple times until convergence. Here the key is update of estimates for a single group (gene). Unfortunately, the cMCP approach does not have a simple form for updating individual groups. To tackle this problem, we adopt an approximation approach. Consider update with the jth group. By taking the first order Taylor series approximation about β_j and evaluating at $\tilde{\beta}_j$ (the current estimate), the penalty as a function of β_j^k is approximately proportional to $\tilde{\lambda}_{jk}|\beta_j^k|$ where

$$\tilde{\lambda}_{jk} = \rho'\left(\sum_{m=1}^{M} \rho(|\tilde{\beta}_j^m|; \lambda_2, b); \lambda_1, a\right) \rho'(|\tilde{\beta}_j^k|; \lambda_2, b). \tag{7}$$

For update with each β_j^k, we have an explicit solution:

$$\hat{\beta}_j^k = f_{cMCP}(z; \lambda) = S_1(z, \lambda), \tag{8}$$

with $S_1(z, \lambda) = sgn(z)(|z| - \lambda)_+$, and z and λ to be defined below.

Consider the following algorithm. With fixed tuning and regularization parameters,

1. Initialize $s = 0$, the estimate $\beta^{(0)} = (\beta_1^{(0)\prime}, \ldots, \beta_d^{(0)\prime})' = (0, \ldots, 0)'$, and the vector of residuals $r = Y - X\beta^{(0)}$;
2. For $j = 1, \ldots, d$,
   (a) Calculate $\tilde{\lambda}_{jk}$ according to expression (7).
   (b) Calculate $z_j = n^{-1}X[\cdot, j]'r + \beta_j^{(s)}$. $X[\cdot, j]$ is the $n \times d_j$ submatrix of X that corresponds to β_j.
   (c) For $k = 1, \ldots, M$, update $\beta_j^{k(s+1)} \leftarrow f_{cMCP}(z_j^k; \tilde{\lambda}_{jk})$, where z_j^k is the kth element of z_j.
   (d) Update $r \leftarrow r - X[\cdot, j](\beta_j^{(s+1)} - \beta_j^{(s)})$.

   Update $s \leftarrow s + 1$.
3. Repeat Step 2 until convergence.

We use the L_2 norm of the difference between two consecutive estimates smaller than 0.001 as the convergence criterion. Convergence is achieved for the lung cancer datasets within twenty iterations. For the proposed methods, in the objective function, the first term is continuously differentiable and regular in the sense of Tseng (2001). The penalty term is separable. Thus the coordinate descent estimate converges to a coordinate-wise minimum of the first term, which is also a stationary point. Our limited experience suggests that the proposed computational algorithms are affordable. Among the three approaches, cMCP has the highest computational cost. With fixed tunings, the analysis of the lung cancer datasets (Sect. 4) takes about 40 s using a regular desktop PC.

3.2.3 Tuning Parameter Selection

The proposed methods involve the following tuning/regularization parameters: two-step approach: (λ, γ) for gMCP and possibly different (λ, γ) for MCP, 1-norm gMCP: (λ, γ), and cMCP: $(\lambda_1, \lambda_2, a, b)$.

Properties of the estimates are jointly determined by the tuning/regularization parameters. Generally speaking, smaller values of a and b (γ in MCP and gMCP) are better at retaining the unbiasedness of the MCP penalty for large coefficients, but they also have the risk of creating objective functions with a nonconvexity problem that are difficult to optimize and yield solutions that are discontinuous with respect to λ_1 and λ_2 (λ). It is therefore advisable to choose values of a and b (γ) that are big enough to avoid this problem but not too big. As suggested in Breheny and Huang (2011) and Zhang (2010), we have experimented with a few values for a and b (γ), particularly including 1.8, 3, 6, and 10.

In our numerical study, we select tuning parameters via V-fold cross validation with $V = 5$. Our limited unpublished simulation suggests that $a = 6$, $b = 6$ and $\gamma = 6$ lead to the best performance. We note that such a result does not indicate the

universal superiority of those values. In practice, searching over multiple possible values is still needed. With λ (λ_1, λ_2), one may expect that its value cannot go down to very small values since there are regions not locally convex (Breheny and Huang 2009; 2011). The criteria over non-locally convex regions may go up and down. To avoid the unexpectedness of such regions, we select λ (λ_1, λ_2) where the criterion first goes up (see Breheny and Huang 2011 for related discussions).

4 Analysis of Lung Cancer Prognosis Studies

Lung cancer is the leading cause of death from cancer for both men and women in the USA and in most other parts of the world. Non-small-cell lung cancer (NSCLC) is the most common cause of lung cancer death, accounting for up to 85% of such deaths (Tsuboi et al. 2007). Gene profiling studies have been extensively conducted on lung cancer, searching for markers associated with prognosis. Three studies are described in Xie et al. (2011). The UM (University of Michigan Cancer Center) study had a total of 175 patients, among whom 102 died during follow-up. The median follow-up was 53 months. The HLM (Moffitt Cancer Center) study had a total of 79 subjects, among whom 60 died during follow-up. The median follow-up was 39 months. The CAN/DF (Dana-Farber Cancer Institute) study had a total of 82 patients, among whom 35 died during follow-up. The median follow-up was 51 months. We refer to Xie et al. (2011) and references therein for more details on study designs, subjects' characteristics, and profiling protocols; 22,283 genes were profiled in all three studies.

In previous studies such as Xie et al. (2011), the three datasets were combined and analyzed. Such a strategy corresponds to a special case of the homogeneity model in the present study. As the three datasets were generated in three independent studies, heterogeneity is expected to exist across datasets. This can be partly seen from the summary survival data and profiling protocols. Here we assume the heterogeneity model and analyze using the two-step method (Table 1), 1-norm gMCP (Table 2), and cMCP (Table 3). Note that with all methods, the small magnitudes of regression coefficients are caused by the "clustered" log survival times. The estimates suggest that different datasets may have different prognosis-associated genes. This partly explains why published studies have failed to unify the identified markers across different lung cancer prognosis datasets. As described in Sect. 1, multiple factors may contribute to this heterogeneity. Without having access to all the experiment details, we are not able to determine the exact cause of heterogeneity. Although there are overlaps, different approaches identify different sets of genes. Such an observation is not surprising and has been made in published studies such as Ma et al. (2011b).

To provide a more comprehensive description of the three datasets and various methods, we also conduct the evaluation of prediction performance. Although in principle marker identification and prediction are two distinct objectives, evaluation of prediction performance can be informative for marker identification. In particular,

Table 1 Two-step method: identified genes and their estimates.

Probe	Gene	UM	HLM	CAN/DF
200041_s_at	DDX39B	0.004		0.027
200642_at	SOD1		0.016	0.016
200650_s_at	LDHA	−0.007	0.044	
200694_s_at	DDX24		−0.031	
200747_s_at	NUMA1	−0.016	−0.022	
200772_x_at	PTMA	−0.029	0.034	−0.025
201021_s_at	DSTN		−0.050	
201033_x_at	RPLP0	−0.006		
201508_at	IGFBP4	0.0004	−0.050	
201523_x_at	UBE2N	0.017		
201568_at	UQCRQ	0.016	0.000	0.002
201789_at	Hs.59719		−0.058	
201875_s_at	MPZL1	−0.015		0.001
202081_at	IER2	0.0002	−0.026	
202162_s_at	CNOT8	0.006	0.017	0.001
202176_at	ERCC3	0.006		

Table 2 1-norm gMCP: identified genes and their estimates.

Probe	Gene	UM	HLM	CAN/DF
200041_s_at	DDX39B			0.005
200633_at	UBB		−0.001	
200642_at	SOD1		0.0004	2.8E-05
200674_s_at	RPL32		0.002	
200693_at	YWHAQ		−0.003	
200694_s_at	DDX24		−0.002	
200724_at	RPL10	0.0005		
200772_x_at	PTMA	−0.0002		−0.002
200804_at	TMBIM6	−0.001		
200972_at	TSPAN3	−0.003		
200973_s_at	TSPAN3			0.003
201021_s_at	DSTN		−0.016	
201033_x_at	RPLP0		−0.0005	
201173_x_at	NUDC		0.003	
201201_at	CSTB		0.005	
201508_at	IGFBP4		−0.001	
201611_s_at	ICMT			−0.001
201645_at	TNC	0.003		
201729_s_at	KIAA0100	5.0E-05		
201789_at	Hs.59719		−0.012	
202081_at	IER2		−0.002	
202146_at	IFRD1			−0.0001
202176_at	ERCC3		−0.002	
202183_s_at	KIF22		0.002	
202413_s_at	USP1			−0.001

Table 3 cMCP: identified genes and their estimates.

Probe	Gene	UM	HLM	CAN/DF
200041_s_at	DDX39B			0.005
200633_at	UBB		−0.001	
200674_s_at	RPL32		0.002	
200693_at	YWHAQ		−0.005	
200772_x_at	PTMA			−0.001
200972_at	TSPAN3	−0.002		
200973_s_at	TSPAN3			0.002
201021_s_at	DSTN		−0.017	
201033_x_at	RPLP0		−0.0003	
201173_x_at	NUDC		0.001	
201201_at	CSTB		0.004	
201645_at	TNC	0.001		
201789_at	Hs.59719		−0.011	
202176_at	ERCC3		−0.0003	
202183_s_at	KIF22		0.001	
202413_s_at	USP1			−0.0002

if prediction is more accurate, then the identified markers are expected to be more meaningful. For prediction evaluation, we adopt a random sampling approach as in Ma et al. (2009). More specifically, we generate training sets and corresponding testing sets by random splitting data (with sizes 3:1). Estimates are generated using the training sets only. We then make prediction for subjects in the testing sets. We dichotomize the predicted linear risk scores $X\hat{\beta}$ at the median, create two risk groups, and compute the logrank statistic, which measures the difference in survival between the two groups. To avoid extreme splits, this procedure is repeated 100 times. The average logrank statistics are calculated as 2.17 (two-step), 4.77 (1-norm gMCP), and 3.70 (cMCP). 1-norm gMCP is the only approach that can separate subjects into groups with significantly different survival risks (p-value = 0.029). Based on this prediction evaluation, genes and estimates presented in Table 2 are suggested as the final results for these three datasets.

5 Discussion

In cancer genomic research, multi-dataset analysis provides an effective way to overcome certain drawbacks of single-dataset analysis. In most published studies, it has been reinforced that multiple datasets share the same set of prognosis-associated genes, that is, the homogeneity model. In this study, for multiple cancer prognosis datasets, we consider the heterogeneity model, which includes the homogeneity model as a special case and can be less restricted. This model may provide a way to explain the failure to unify cancer prognosis markers across independent studies (Knudsen 2006; Cheang et al. 2008). Under the heterogeneity model, we propose

three penalization methods for marker identification. Such methods are intuitively reasonable and computationally feasible. Analysis of three lung cancer studies demonstrates the practical feasibility of proposed methods.

Under the heterogeneity model, marker selection needs to be conducted in two dimensions. Methods beyond penalization, for example thresholding and boosting, may also be able to achieve such selection. Comprehensive investigation and comparison of different approaches are beyond the scope of this article. The proposed methods are based on the MCP penalty, which has been shown to have satisfactory performance in single-dataset analysis. We suspect that it is possible to develop similar approaches based on, for example, bridge and SCAD penalties. As in single-dataset analysis there is no evidence that such penalties are superior to MCP, such a development is not pursued.

Acknowledgments We would like to thank the attendants of the ICSA 2012 Applied Statistics Symposium for valuable comments and organizers of the proceedings. The authors have been supported by the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development, awards CA142774, CA165923, CA152301 from NIH, DMS0904181 from NSF, and award 2012LD001 from National Bureau of Statistics of China.

References

Breheny, P. and Huang, J. (2009) Penalized methods for bi-level variable selection. Statistics and Its Interface. 2: 369–380.

Breheny, P. and Huang, J. (2011) Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Annals of Applied Statistics. 5: 232–253.

Buckley, I.V. and James, I. (1979) Linear regression with censored data. Biometrika. 66: 429–436.

Cheang, M.C.U., van de Rijn, M. and Nielsen, T.O. (2008) Gene expression profiling of breast cancer. Annual Reviews of Pathology: Mechanisms of Disease. 3, 67–97.

Guerra, R. and Goldsterin, D.R. (2009). Meta-Analysis and Combining Information in Genetics and Genomics. Chapman and Hall/CRC, 1st edition.

Huang, Y., Huang, J., Shia, B.C. and Ma, S. (2012) Identification of cancer genomic markers via integrative sparse boosting. Biostatistics. 13: 509–522.

Jin, Z., Lin, D.Y., Wei, L.J. and Ying, Z. (2003) Rank-based inference for the accelerated failure time model. Biometrika. 90: 341–353.

Knudsen, S. (2006) Cancer Diagnostics with DNA Microarrays. Wiley.

Ma, S., Huang, J. and Moran, M. (2009) Identification of genes associated with multiple cancers via integrative analysis. BMC Genomics, 10: 535.

Ma, S., Huang, J. and Song, X. (2011a) Integrative analysis and variable selection with multiple high-dimensional datasets. Biostatistics, 12: 763–775.

Ma, S., Huang, J., Wei, F., Xie, Y. and Fang, K. (2011b) Integrative analysis of multiple cancer prognosis studies with gene expression measurements. Statistics in Medicine. 30: 3361–3371.

Stute, W. (1996) Distributional convergence under random censorship when covariables are present. Scandinavian Journal of Statistics. 23: 461–471.

Tseng, P. (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of Optimization Theory and Applications. 109: 475–494.

Tsuboi, M., Ohira, T., Saji, H., Hiyajima, K., Kajiwara, N., Uchida, O. et al. (2007) The present status of postoperative adjuvant chemotherapy for completely resected non-small cell lung cancer. Ann Thorac Cardiovasc Surg. 13:73–77.

Xie, Y., Xiao, G., Coombes, K., Behrens, C., Solis, L., Raso, G., Girard, L., Erickson, H., Roth, J., Heymach, J., Moran, C., Danenberg, K., Minna, J. and Wistuba, I. (2011) Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non small cell lung cancer patients. Clin Cancer Res, 17(17): 5705–5714.

Zhang, C.H. (2010) Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics. 38: 894–942.