Genet. Res., Camb. (2013), **95**, *pp.* 68–77. © Cambridge University Press 2013
doi:10.1017/S0016672313000086

Sparse group penalized integrative analysis of multiple cancer prognosis datasets

JIN LIU[1], JIAN HUANG[2], YANG XIE[3] AND SHUANGGE MA[4]*

[1] *Division of Epidemiology and Biostatistics, UIC School of Public Health 1603 W. Taylor Street (MC 923), Chicago, IL 60612-4394, USA*
[2] *Department of Statistics and Actuarial Science, and Department of Biostatistics, University of Iowa, Iowa, USA*
[3] *Department of Clinical Sciences, UT Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390, USA*
[4] *Department of Biostatistics, Yale University; and VA Cooperative Studies Program Coordinating Center, West Haven, CT, USA*

(*Received 11 April 2013; revised 20 May 2013*)

Summary

In cancer research, high-throughput profiling studies have been extensively conducted, searching for markers associated with prognosis. Owing to the 'large *d*, small *n*' characteristic, results generated from the analysis of a single dataset can be unsatisfactory. Recent studies have shown that integrative analysis, which simultaneously analyses multiple datasets, can be more effective than single-dataset analysis and classic meta-analysis. In most of existing integrative analysis, the homogeneity model has been assumed, which postulates that different datasets share the same set of markers. Several approaches have been designed to reinforce this assumption. In practice, different datasets may differ in terms of patient selection criteria, profiling techniques, and many other aspects. Such differences may make the homogeneity model too restricted. In this study, we assume the heterogeneity model, under which different datasets are allowed to have different sets of markers. With multiple cancer prognosis datasets, we adopt the accelerated failure time model to describe survival. This model may have the lowest computational cost among popular semiparametric survival models. For marker selection, we adopt a sparse group minimax concave penalty approach. This approach has an intuitive formulation and can be computed using an effective group coordinate descent algorithm. Simulation study shows that it outperforms the existing approaches under both the homogeneity and heterogeneity models. Data analysis further demonstrates the merit of heterogeneity model and proposed approach.

1. Introduction

High-throughput genomic studies have been extensively conducted, searching for markers associated with the risk and prognosis of cancer. In this paper, we focus on cancer prognosis studies with gene expression measurements. Data generated in high-throughput studies have the 'large *d*, small *n*' characteristic, with the number of genes profiled *d* much larger than the sample size *n*. In addition, in whole-genome studies, only a subset of the profiled genes is associated with prognosis. Thus, analysis of cancer prognosis data with high-throughput genomic measurements demands simultaneous regularized estimation and selection.

It has been recognized that genomic markers identified from the analysis of single datasets can be unsatisfactory. Multiple factors may contribute to the unsatisfactory performance, including the highly noisy nature of cancer genomic data, technical variations of profiling techniques and, more importantly, small sample sizes of individual studies. Recent studies have shown that pooling and analysing multiple studies may effectively increase sample size and improve properties of the identified markers (Guerra & Goldsterin, 2009). For example, simulation and data analysis have shown that markers identified in multi-dataset analysis may have more true positives, fewer false positives and better prediction performance (Ma *et al.*, 2009, 2011*a*, *b*). Multi-dataset methods include meta-analysis and integrative analysis methods. Integrative analysis pools and analyses raw data from multiple studies, and differs significantly from classic meta-analysis, which analyses multiple studies

* Corresponding author: Department of Biostatistics, School of Public Health, Yale University, 60 College ST, LEPH 206, New Haven, CT 06520, USA. Tel: 203-785-3119. Fax: 203-785-6912. E-mail: shuangge.ma@yale.edu

separately and then pools summary statistics (lists of identified genes, *P*-values, effect sizes, etc.).

In this paper, we conduct integrative analysis of multiple cancer prognosis studies, with note that the proposed method may also be applicable to other types of data (e.g. aetiology studies with categorical responses and treatment studies with continuous markers). In studies such as Ma *et al.* (2011*b*), the homogeneity model has been assumed. Under this model, multiple datasets have exactly the same set of markers. This model has also been adopted with cancer diagnosis studies (Ma *et al.*, 2011*a* and references therein). The existing approaches have been designed to reinforce this homogeneity assumption (more details available in Section 2). In practice, different datasets may differ in terms of patient selection criteria, platforms and protocols used for profiling, normalization methods, and other aspects. Such differences may make the homogeneity assumption too restricted. In addition, data analyses in Ma *et al.* (2011*a*, *b*) show that for some 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. This can also be seen from data analysis presented in Section 4. Such an observation further suggests the necessity of relaxing the homogeneity assumption.

In what follows, we describe cancer survival using the accelerated failure time (AFT) model. Among the popular semiparametric survival models, the AFT model may have the lowest computational cost, which is especially desirable with high-dimensional data. In addition, its regression coefficients 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 is more flexible. For marker selection, we adopt a sparse group penalization method. The proposed penalization is intuitively reasonable and computationally feasible. *This study complements the existing prognosis studies by conducting integrative survival analysis under the more flexible heterogeneity model and by adopting a penalization approach that has satisfactory performance under both the homogeneity and heterogeneity models.* In a recent paper (Liu *et al.*, 2013), we have studied the heterogeneity model and sparse penalization for binary data and logistic regression models. As the present data and model settings are significantly different from Liu *et al.* (2013), the performance of the sparse group penalization needs to be 're-examined', the computational algorithm needs to be adjusted, and so the present study is warranted. In addition, this study may be the first to observe that for prognosis data under the homogeneity model, sparse penalization outperforms group penalization.

The rest of the paper is organized as follows. Data and model settings are described in Section 2. For integrity of this paper, we also briefly describe the homogeneity model and an existing penalization approach designed to reinforce the model assumption. The heterogeneity model and sparse group penalization approach are described in Section 3. The proposed estimate can be computed using an effective group coordinate descent (GCD) algorithm. Numerical studies, including simulation and data analysis, are conducted in Section 4. The paper concludes with the discussion in Section 5. Some additional analysis results are presented in the Appendix.

2. Integrative analysis of cancer prognosis studies

(i) *Data and model settings*

Assume that there are M-independent studies, and there are 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 the failure time. Denote X^m as the length-d vector of gene expressions. For simplicity of notation, assume that the same set of genes is measured in all M studies. For the ith subject, the AFT model assumes that

$$T_i^m = \beta_0^m + X_i^{m'} \beta^m + \varepsilon_i^m, \quad i=1, \ldots, n^m. \tag{1}$$

where β_0^m is the intercept, β^m 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 $\delta_i^m = \mathrm{I}\{T_i^m \leqslant C_i^m\}$ is the event indicator.

When the distribution of ε_i^m is known, a parametric likelihood function can be constructed. Here we consider the more flexible case where this distribution is unknown (i.e., a semiparametric model). In the literature, multiple estimation approaches have been developed, including for example the Buckley–James and rank-based approaches. In this study, we adopt the weighted least-squares approach (Stute, 1996), which 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. $\hat{F}^m$ can be written as $\hat{F}^m(y) = \sum_{i=1}^{n^m} \omega_i^m \mathrm{I}\{Y_{(i)}^m \leqslant y\}$, where ω_i^ms can be computed as

$$\omega_1^m = \frac{\delta_{(1)}^m}{n^m},$$

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

Here $Y^m_{(1)} \leqslant \cdots \leqslant Y^m_{(n^m)}$ are the order statistics of Y^m_is, and $\delta^m_{(1)}, \ldots, \delta^m_{(n^m)}$ are the associated event indicators. Similarly, let $X^m_{(1)}, \ldots, X^m_{(n^m)}$ be the associated gene expressions of the ordered Y^m_is. Stute (1996) proposed the weighted least-squares estimator $(\hat{\beta}^m_0, \hat{\beta}^m)$ that minimizes

$$\frac{1}{2n} \sum_{i=1}^{n^m} \omega^m_i (Y^m_{(i)} - \beta^m_0 - X^{m'}_{(i)} \beta^m)^2. \tag{2}$$

Define $\bar{X}^m_\omega = \sum_{i=1}^{n^m} \omega^m_i X^m_{(i)} / \sum_{i=1}^{n^m} \omega^m_i$, $\bar{Y}^m_\omega = \sum_{i=1}^{n^m} \omega^m_i Y^m_{(i)} / \sum_{i=1}^{n^m} \omega^m_i$. Let $X^m_{\omega(i)} = \sqrt{\omega^m_i}(X^m_{(i)} - \bar{X}^m_\omega)$ and $Y^m_{\omega(i)} = \sqrt{\omega^m_i}(Y^m_{(i)} - \bar{Y}^m_\omega)$, respectively. With the weighted centred values, the intercept is zero. The weighted least-squares objective function can be written as

$$L^m(\beta^m) = \frac{1}{2n} \sum_{i=1}^{n^m} (Y^m_{\omega(i)} - X^{m'}_{\omega(i)} \beta^m)^2. \tag{3}$$

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

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

(ii) *Homogeneity model and penalized selection*

In Huang *et al.* (2012*c*) and Ma *et al.* (2011*a*, *b*), the homogeneity model is adopted to describe the genomic basis of M datasets. We briefly describe this model here for integrity of this paper. Denote β^m_j as the jth component of β^m. Then $\beta_j = (\beta^1_j, \ldots, \beta^M_j)'$ is the length-M vector of regression coefficients representing the effects of gene j in M studies. Under the homogeneity model, for any $j (= 1, \ldots, d)$, $I(\beta^1_j = 0) = \ldots = I(\beta^M_j = 0)$. That is, the M datasets have the same set of cancer-associated genes. This is a sensible model when multiple datasets have been generated under the same protocol.

For marker selection, Ma *et al.* (2011*b*) propose using the group minimax concave penalty (GMCP) approach, where the estimate is defined as

$$\hat{\beta} = \arg\min\{L(\beta) + P_{\lambda_1}(\beta)\},$$

with

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

$\rho(t; \lambda, \gamma) = \lambda \int_0^{|t|} (1 - (x/\lambda\gamma)) + \mathrm{d}x$ is the MCP penalty (Zhang, 2010). $\|\beta_j\|_{\Sigma_j} = \|\Sigma_j^{1/2}\beta_j\|_2$, $\|\cdot\|_2$ is the L_2 norm, $\Sigma_j = n^{-1} X_j' X_j$, and X_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 matched gene sets, $d_j \equiv M$. We keep d_j so that this formulation can easily accommodate partially matched gene sets. When gene j is not measured in dataset k, we take the convention $\beta^k_j \equiv 0$. λ_1 is a tuning parameter, and γ is a regularization parameter.

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, MCP penalization is adopted. 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 hence the L_2 norm is adopted. Note that here we adopt the $\|\cdot\|_{\Sigma_j}$ norm, which rescales the regression coefficient vector by covariance matrix Σ_j, so that computation is less *ad hoc*. This differs from the approach in Ma *et al.* (2011*b*).

3. Heterogeneity model and penalized selection

(i) *Heterogeneity model*

When multiple datasets are generated in independent studies, heterogeneity inevitably exists (Knudsen, 2006). The degree of heterogeneity depends on 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 (Knudsen, 2006; Cheang *et al.*, 2008). This can also be partly seen from Ma *et al.* (2011*b*) and our data analysis in Section 4. 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.

There are also 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 usually have different sets of markers. On the other hand, multiple pathways, such as apoptosis, cell cycle and signalling, are associated with 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 patterns, 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

https://doi.org/10.1017/S0016672313000086 Published online by Cambridge University Press

consistent across multiple studies. However, for other genes, they may be important in some studies but not others.

(ii) *Penalized selection*

Consider the penalized estimate

$$\hat{\beta} = \arg\min\{L(\beta) + P_{\lambda_1,\lambda_2}(\beta)\},$$

where

$$P_{\lambda_1,\lambda_2} = \sum_{j=1}^{d} \rho\left(\|\beta_j\|_{\Sigma_j}; \sqrt{d_j}\lambda_1, \gamma\right) + \sum_{j=1}^{d}\sum_{k=1}^{M} \rho\left(\left|\beta_j^k\right|; \lambda_2, \gamma\right). \tag{5}$$

Here, notations have similar definitions as in Section (ii). λ_2 is another tuning parameter.

Under the heterogeneity model, marker selection needs to be achieved in two dimensions. The first is to determine whether a gene is associated with prognosis in any dataset at all. This step of selection is achieved using a GMCP penalty. For an important gene, the second dimension is to determine in which dataset(s) it is associated with prognosis. This step of selection is achieved using an MCP penalty. The sum of two penalties can achieve two-dimensional selection and fully recover the associations between multiple genes and prognosis in multiple studies.

In the single-dataset analysis with a continuous response and linear regression model, Friedman *et al.* (2010) proposed a sparse group Lasso (SGLasso) penalty. The penalty in (5) has been partly motivated by that study. Owing to the analogy, we refer to penalty (5) as 'sparse group MCP' or SGMCP. Unlike in Friedman *et al.* (2010), multiple heterogeneous datasets are analysed in this study, and a group corresponds to one gene in multiple datasets, as opposed to multiple variables. In addition, the MCP penalty is adopted to replace the Lasso penalty, as in single-dataset analysis, MCP has been shown to outperform Lasso (Zhang, 2010; Huang *et al.*, 2012*a*, *b*). In addition, censored survival data under the AFT model are analysed, which differ from the continuous response and linear model in Friedman *et al.* (2010). Note that MCP becomes Lasso when $\gamma = \infty$. Thus, SGMCP includes SGLasso as a special case.

(iii) *Computation*

For dataset $m(=1, \ldots, M)$, we standardize the gene expressions to have marginal means zero and satisfy $X_j^{m\prime} X_j^m = n$. Thus, $n^{-1}X_j'X_j = \mathrm{I}_{d_j}$, where I_{d_j} is the $d_j \times d_j$ identity matrix. With slight abuse of notation, we denote $\|\beta_j\|$ for $\|\beta_j\|_{\mathrm{I}_{d_j}}$. For computation, we consider a GCD algorithm. It optimizes the objective function with respect to one gene at a time, and iteratively cycles through all genes. The overall cycling is repeated multiple times until convergence.

Consider the overall objective function

$$\tilde{L}(\beta, \lambda_1, \lambda_2, \gamma) = \frac{1}{2n}\left\| Y - \sum_{j=1}^{d} X_j\beta_j \right\|^2 + \sum_{j=1}^{d} \rho\left(\|\beta_j\|_{\Sigma_j}; \sqrt{d_j}\lambda_1, \gamma\right) + \sum_{j=1}^{d}\sum_{k=1}^{M} \rho\left(\left|\beta_j^k\right|; \lambda_2, \gamma\right). \tag{6}$$

For $j = 1, \ldots, d$, given β_k $(k \neq j)$ fixed at their current estimates $\tilde{\beta}_k^{(s)}$, we seek to minimize $\tilde{L}(\beta, \lambda_1, \lambda_2, \gamma)$ with respect to β_j. Here, only the terms involving β_j in $\tilde{L}(\beta, \lambda_1, \lambda_2, \gamma)$ matter. This is equivalent to minimizing

$$R(\beta_j) = \frac{1}{2n}\|r_{-j} - X_j\beta_j\|^2 + \rho\left(\|\beta_j\|; \sqrt{d_j}\lambda_1, \lambda\right) + \sum_{k=1}^{M} \rho\left(\left|\beta_j^k\right|; \lambda_2, \gamma\right), \tag{7}$$

where $r_{-j} = Y - \sum_{k \neq j} X_j \tilde{\beta}_k^{(s)}$. The first-order derivative of (7) is:

$$\frac{\partial R(\beta_j)}{\partial \beta_j} = -\frac{1}{n}X_j' r_{-j} + \frac{1}{n}X_j' X_j \beta_j + \frac{\beta_j}{\|\beta_j\|} \times \begin{cases} \sqrt{d_j}\lambda_1 - \frac{\|\beta_j\|}{\gamma}, & \text{if } \|\beta_j\| \leqslant \gamma\sqrt{d_j}\lambda_1 \\ 0, & \text{if } \|\beta_j\| > \gamma\sqrt{d_j}\lambda_1 \end{cases} + t, \tag{8}$$

where

$$t = \left(\operatorname{sgn}(\beta_j^1) \begin{cases} \lambda_2 - \frac{|\beta_j^1|}{\gamma}, & \text{if } \left|\beta_j^1\right| \leqslant \gamma\lambda_2 \\ 0, & \text{if } \left|\beta_j^1\right| > \gamma\lambda_2 \end{cases}, \ldots, \operatorname{sgn}(\beta_j^M) \begin{cases} \lambda_2 - \frac{|\beta_j^M|}{\gamma}, & \text{if } \left|\beta_j^M\right| \leqslant \gamma\lambda_2 \\ 0, & \text{if } \left|\beta_j^M\right| > \gamma\lambda_2 \end{cases} \right)'.$$

With normalization, $n^{-1}X_j'X_j = \mathrm{I}_{d_j}$. By setting expression (8) to be zero, we have:

$$-z_j + g\beta_j + t = 0, \tag{9}$$

where

$$z_j = n^{-1}X_j{}' r_{-j} = (z_j^1, \ldots, z_j^M)',$$

$$g = \left(1 + \frac{1}{\|\beta_j\|} \times \begin{cases} \sqrt{d_j}\lambda_1 - \frac{\|\beta_j\|}{\gamma}, & \text{if } \|\beta_j\| \leqslant \gamma\sqrt{d_j}\lambda_1 \\ 0, & \text{if } \|\beta_j\| > \gamma\sqrt{d_j}\lambda_1 \end{cases}\right).$$

https://doi.org/10.1017/S0016672313000086 Published online by Cambridge University Press

We first take g fixed at the current estimate of β_j. The kth element in (9) can be rewritten as:

$$-\frac{z_j^k}{g}+\beta_j^k+\text{sgn}(\beta_j^k)\begin{cases}\frac{\lambda_2}{g}-\frac{\left|\beta_j^k\right|}{\gamma g}, & \text{if } \left|\beta_j^k\right|\leqslant\gamma\lambda_2\\ 0, & \text{if } \left|\beta_j^k\right|\leqslant\gamma\lambda_2\end{cases}=0. \tag{10}$$

The solution to equation (10) is

$$\widehat{g\beta_j^k}=\begin{cases}\frac{\text{S}_1(z_j^k,\lambda_2)}{1-1/(\gamma g)}, & \text{if } \left|z_j^k\right|\leqslant\gamma\lambda_2 g\\ z_j^k, & \text{if } \left|z_j^k\right|>\gamma\lambda_2 g\end{cases},\ \ \text{S}_1=\text{sgn}(u)(|u|-\lambda)_+.$$

For $k=1, \ldots M$, we set $u_k=\widehat{g\beta_j^k}$ and $u=(u_1, \ldots, u_M)'$. Taking u back into its definition, we have

$$\beta_j+\frac{\beta_j}{\|\beta_j\|}\begin{cases}\sqrt{d_j}\lambda_1-\frac{\|\beta_j\|}{\gamma}, & \text{if } \|\beta_j\|\leqslant\gamma\sqrt{d_j}\lambda_1\\ 0, & \text{if } \|\beta_j\|>\gamma\sqrt{d_j}\lambda_1\end{cases}=u. \tag{11}$$

Solving equation (11) leads to

$$\hat{\beta}_j=\begin{cases}\frac{\gamma}{\gamma-1}\text{S}_2(u,\sqrt{d_j}\lambda_1), & \text{if } \|u\|\leqslant\gamma\sqrt{d_j}\lambda_1\\ u, & \text{if } \|u\|>\gamma\sqrt{d_j}\lambda_1\end{cases}, \tag{12}$$

where $\text{S}_2(u,t)=\left(1-\frac{t}{\|u\|}\right)_+ u$.

Solving equation (9) amounts to iteratively calculating equation (10)–(12) until convergence. Overall, with fixed tuning parameters,

1. Initialize $s=0$, the estimate $\tilde{\beta}^{(0)}=(0, \ldots, 0)'$, and the vector of residuals $r=Y-\sum_{j=1}^{d}X_j\tilde{\beta}_j^{(0)}$;
2. For $j=1, \ldots, d$,
   (a) Update $\tilde{\beta}_j^{(s+1)}$. This is achieved by solving equation (9)–(12) iteratively until convergence. In our numerical study, convergence is achieved for all datasets within ten iterations;
   (b) Update $r\leftarrow r-X_j(\tilde{\beta}_j^{(s+1)}-\tilde{\beta}_j^{(s)})$;
3. Update $s\leftarrow s+1$;
4. Repeat Steps 2–3 until convergence.

This algorithm starts with a null model. In each iteration, it cycles through all d genes. For each gene, it only involves simple calculations, and the update can be accomplished easily. We use the L_2 norm of the difference between two consecutive estimates smaller than 10^{-5} as the convergence criterion. Convergence is achieved with all our simulated and real data.

(iv) *Tuning parameter selection*

With the MCP penalty, there is one tuning parameter λ and one regularization parameter γ. Generally speaking, smaller values of γ 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 non-convexity problem that are difficult to optimize and yield solutions that are discontinuous with respect to λ. It is therefore advisable to choose a γ value that is big enough to avoid this problem but not too big (Zhang, 2010; Breheny & Huang, 2011). Such results can be extended to the SGMCP.

In principle, it is possible to have different γ values for the two penalties. In practice, adopting the same γ value may reduce computational cost without affecting performance of the estimate much. In numerical study, we have experimented with a few values for γ, particularly including 1·8, 3, 6 and 10 and ∞, as suggested in Zhang (2010) and Breheny & Huang (2011). λ_1 and λ_2 control shrinkage at the group and individual-coefficient levels, respectively. Let $\lambda_{2\text{max}}$ be the largest λ_2 for which all regression coefficients are shrunk to zero. Let $\lambda_{1\text{max}}(\lambda_2)$ be the largest λ_1 under a fixed λ_2 for which all regression coefficients are shrunk to zero. It can be shown that $\lambda_{2\,\text{max}}=n^{-1}\max_{j=1,\ldots,d,k=1,\ldots,M}|X_j^k Y^k|$, where X_j^k is the component of X_j that corresponds to the kth dataset. $\lambda_{1\,\text{max}}(\lambda_2)=\max_{j=1,\ldots,d}\left\|\text{S}_1(n^{-1}X_j{}'Y,\lambda_2)\right\|/\sqrt{d_j}$. This result puts constraints on the range of tuning parameters.

In this study, we search for γ, λ_1 and λ_2 values jointly using V-fold cross-validation ($V=5$ in numerical study). This is computationally feasible as the GCD algorithm has low computational cost. It is expected that λ_1, λ_2 cannot go down to very small values that correspond to regions not locally convex. The cross-validation criteria in such non-locally convex regions may not be monotonous. To avoid such regions, we select λ_1 and λ_2 where the cross-validation criterion first goes up.

4. Numerical study

(i) *Simulation*

We simulate three datasets, each with 100 subjects. For each subject, we simulate the expressions of 1000 genes. The gene expressions are jointly normally distributed, with marginal means equal to zero and variances equal to one. We consider two correlation structures. The first is the auto-regressive (AR) correlation, where expressions of genes j and k have correlation coefficient $\rho^{|j-k|}$. We consider two scenarios with $\rho=0·2$ and $0·7$, corresponding to weak and strong correlations, respectively. The second is the banded correlation. Here, two scenarios are considered. Under the first scenario, genes j and k have correlation coefficient 0·33 if $|j-k|=1$ and 0 otherwise; under the second scenario, genes j and k have correlation coefficient 0·6 if $|j-k|=1$, 0·33 if $|j-k|=2$, and 0 otherwise. We consider the heterogeneity model, under which each dataset has 10 prognosis-associated genes. The three datasets share five

https://doi.org/10.1017/S0016672313000086 Published online by Cambridge University Press

Table 1. *Simulation under the homogeneity model. In each cell, the first/second row is the mean number* (SD) *of true/false positives. When* $\gamma = \infty$, *MCP simplifies to Lasso*

Correlation	$\gamma = 1·8$	$\gamma = 3$	$\gamma = 6$	$\gamma = \infty$
MCP				
AR $\rho = 0·2$	24·02 (3·09)	25·18 (3·49)	26·36 (2·97)	19·74 (5·26)
	25·38 (21·76)	44·80 (24·16)	67·90 (10·11)	179·48 (67·80)
AR $\rho = 0·7$	15·12 (2·62)	15·66 (2·60)	16·88 (2·38)	23·84 (1·40)
	9·74 (6·80)	24·64 (7·86)	49·28 (17·30)	97·76 (40·83)
Banded 1	22·14 (3·47)	23·54 (3·23)	24·40 (2·88)	20·30 (3·20)
	24·66 (23·86)	44·86 (23·16)	66·96 (10·69)	173·56 (64·82)
Banded 2	16·20 (2·14)	16·60 (3·19)	18·16 (2·57)	21·66 (1·87)
	12·42 (7·12)	29·12 (12·30)	60·54 (11·63)	134·90 (57·59)
GMCP				
AR $\rho = 0·2$	29·92 (0·44)	30·00 (0·00)	30·00 (0·00)	29·82 (0·72)
	2·36 (3·88)	16·50 (13·43)	93·30 (29·26)	222·54 (75·33)
AR $\rho = 0·7$	25·74 (3·79)	27·60 (3·74)	26·88 (4·65)	27·12 (1·04)
	7·56 (7·23)	42·24 (17·68)	104·94 (46·98)	88·74 (59·42)
Banded 1	29·94 (0·42)	30·00 (0·00)	30·00 (0·00)	29·70 (0·91)
	2·22 (3·25)	20·34 (10·69)	99·48 (24·99)	198·12 (86·65)
Banded 2	26·22 (4·69)	27·36 (4·23)	27·48 (3·85)	26·76 (1·90)
	7·50 (8·23)	43·44 (17·71)	108·60 (48·77)	117·72 (54·09)
SGMCP				
AR $\rho = 0·2$	29·34 (0·92)	29·42 (1·14)	29·68 (0·51)	28·46 (1·62)
	0·58 (1·36)	0·94 (1·46)	19·56 (9·37)	143·18 (49·32)
AR $\rho = 0·7$	20·88 (3·70)	20·96 (4·92)	22·80 (5·04)	26·42 (0·91)
	1·98 (3·01)	4·40 (4·60)	31·92 (12·69)	67·46 (36·97)
Banded 1	28·96 (0·99)	29·26 (0·83)	29·60 (0·61)	27·02 (2·33)
	0·56 (1·55)	1·34 (1·98)	23·96 (12·09)	111·96 (50·98)
Banded 2	21·36 (3·92)	21·60 (4·62)	24·64 (4·28)	25·42 (1·60)
	1·24 (2·17)	4·22 (4·50)	31·72 (14·72)	76·76 (49·50)

common important genes, and the rest prognosis-associated genes are dataset-specific. As a special case of the heterogeneity model, we also consider the homogeneity model, under which all three datasets share the same 10 prognosis-associated genes. Under both models, across the three datasets, there are a total of 30 true positives. For the *m*th dataset, we generate the logarithm of event time from the AFT model: $T^m = \beta_0^m + X^{m'}\beta^m + \varepsilon^m$, where $\beta_0^m = 0·5$ and $\epsilon^m \sim N(0, 0·25)$. The non-zero regression coefficients for β^1, β^2 and β^3 are (0·4, 0·4, 0·6, −0·5, 0·3, 0·3, 0·6, 0·5, 0·5, 0·2), (0·5, 0·2, 0·3, −0·5, 0·4, 0·4, 0·3, 0·2, 0·6, 0·5) and (0·6, 0·3, 0·7, −0·4, 0·5, 0·3, 0·5, 0·7, 0·4, 0·3), respectively. The logarithm of censoring time is generated as uniformly distributed and independent of the event time. We adjust the censoring distributions so that the overall censoring rate is about 30% under all simulation scenarios. In the Appendix, we also present simulation results under $\epsilon^1 \sim 0·2t(60)$, $\epsilon^2 \sim 0·6t(30)$, $\epsilon^3 \sim t(20)$, that is, the three datasets have different error distributions with different variances.

For a simulated dataset, we analyse using SGMCP as well as the following alternatives: (i) MCP and Lasso (which is MCP with $\gamma = \infty$). Here, we analyse the three datasets separately and then combine analysis results across datasets. This is a meta-analysis approach. (ii) GMCP and GLasso (which is GMCP with $\gamma = \infty$). These two approaches reinforce that the same set of genes are identified across datasets, are sensible under the homogeneity model, but can be too restricted under the heterogeneity model; and (iii) SGLasso (which is SGMCP with $\gamma = \infty$). With the above approaches, we consider three different γ values for the MCP and GMCP penalties. For all approaches, tuning parameters are chosen using 5-fold cross-validation.

Summary statistics, including the number of true positives, number of false positives, and their standard deviations based on 100 replicates, are shown in Table 1 (homogeneity model, homogeneous errors) and 2 (heterogeneity model, homogeneous errors) and 5 (Appendix, homogeneity model, heterogeneous errors) and 6 (Appendix, heterogeneity model, heterogeneous errors), respectively. We observe that SGMCP has the best performance under all simulated scenarios. It can significantly outperform GMCP under the homogeneity model. SGMCP is able to identify most or all of the true positives. Under some simulated scenarios, there are a number of false positives. This is partly caused by correlations among genes. The false positive rate increases as genes

https://doi.org/10.1017/S0016672313000086 Published online by Cambridge University Press

Table 2. *Simulation under the heterogeneity model. In each cell, the first/second row is the mean number* (SD) *of true/false positives. When* $\gamma = \infty$, *MCP simplifies to Lasso*

Correlation	$\gamma = 1·8$	$\gamma = 3$	$\gamma = 6$	$\gamma = \infty$
MCP				
AR $\rho = 0·2$	24·30 (3·26)	25·50 (2·82)	24·96 (2·91)	15·44 (4·98)
	31·58 (37·28)	51·86 (33·81)	68·96 (13·42)	87·02 (39·50)
AR $\rho = 0·7$	15·06 (2·19)	15·32 (2·68)	16·82 (2·46)	23·28 (1·71)
	8·94 (4·63)	24·04 (9·06)	51·30 (17·30)	84·40 (30·32)
Banded 1	23·08 (3·40)	23·48 (3·64)	24·80 (3·18)	18·22 (3·80)
	39·94 (50·17)	45·16 (31·55)	69·16 (13·07)	99·60 (40·52)
Banded 2	16·10 (2·65)	16·40 (2·72)	17·64 (2·71)	21·50 (1·99)
	11·38 (8·23)	26·26 (8·22)	57·32 (13·42)	88·38 (28·67)
GMCP				
AR $\rho = 0·2$	25·38 (2·03)	26·44 (1·73)	25·74 (5·18)	23·68 (4·02)
	32·82 (9·46)	59·60 (14·28)	124·44 (25·75)	209·72 (96·92)
AR $\rho = 0·7$	18·64 (4·89)	18·86 (4·92)	19·00 (4·17)	26·04 (1·09)
	27·86 (10·98)	57·70 (15·30)	83·30 (44·74)	163·86 (65·25)
Banded 1	25·10 (1·56)	25·96 (1·54)	25·28 (4·28)	22·84 (3·75)
	32·50 (9·78)	64·58 (11·96)	125·44 (23·75)	173·90 (74·02)
Banded 2	19·78 (5·63)	19·48 (4·94)	18·46 (4·41)	24·96 (1·89)
	28·46 (10·55)	58·40 (17·36)	103·64 (40·09)	198·12 (72·69)
SGMCP				
AR $\rho = 0·2$	26·62 (1·74)	25·96 (2·64)	27·00 (1·98)	22·86 (5·74)
	7·78 (5·60)	11·20 (7·64)	39·70 (13·59)	141·62 (72·83)
AR $\rho = 0·7$	18·72 (4·21)	17·58 (4·36)	17·48 (4·28)	25·44 (1·11)
	5·72 (4·84)	8·92 (5·78)	34·88 (15·24)	109·44 (52·82)
Banded 1	24·36 (3·45)	25·42 (2·47)	26·64 (2·35)	22·70 (3·65)
	5·48 (3·25)	12·94 (7·99)	40·04 (10·44)	130·88 (58·28)
Banded 2	18·96 (4·36)	16·88 (5·46)	17·96 (4·28)	23·58 (2·20)
	5·56 (4·81)	9·62 (8·54)	38·48 (13·88)	113·78 (45·67)

become 'more correlated'. When γ increases, then the number of false positives increases. This is as expected, as in single-dataset analysis, it has been suggested that Lasso-type penalization ($\gamma = \infty$) tends to overselect (Huang *et al.*, 2012*a*). GMCP and MCP may identify a satisfactory number of true positives, however, at the price of a much larger number of false positives (Table 2).

(ii) *Analysis of breast cancer studies*

Worldwide, breast cancer is the commonest cancer death among women. In 2008, breast cancer caused 458 503 deaths worldwide (13·7% of cancer deaths in women). Multiple high-throughput profiling studies have been conducted, searching for genomic markers associated with breast cancer prognosis. We collect data from three independent breast cancer prognosis studies, which were originally reported in Huang *et al.* (2003), Sotiriou *et al.* (2003) and van't Veer *et al.* (2002), respectively, and referred to as D1, D2 and D3 hereafter. In D1, Affymetrix chips were used to profile the expressions of 12 625 genes. There were a total of 71 subjects, among whom 35 died during follow-up. The median follow-up was 39 months. In D2, cDNA chips were used to profile the expressions of 7650 genes. There were a total of 98 subjects, among whom 45 died during follow-up. The median follow-up was 67·9 months. In D3, oligonucleotide arrays were used to profile 24 481 genes. There were a total of 78 patients, among whom 34 died during follow-up. The median follow-up was 64·2 months. We process each dataset separately as follows. With Affymetrix data, a floor and a ceiling are added, and then measurements are log 2 transformed. With both Affymetrix and cDNA data, we fill in missing expressions with means across samples. We then standardize each gene expression to have zero mean and unit variance.

In previous studies such as Ma *et al.* (2011*b*), the homogeneity model is assumed. 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, the profiling protocols, as well as results from analysis of each individual dataset using MCP (Table 7). We analyse the three datasets using the proposed approach as well as MCP, GMCP and SGLasso. The identified genes and corresponding estimates are shown in Tables 3 and 7–9 in the Appendix. Note that with all approaches, the small magnitudes of regression coefficients are caused by 'clustered' log survival times. SGMCP identifies seven, eight and five

https://doi.org/10.1017/S0016672313000086 Published online by Cambridge University Press

Table 3. *Analysis of breast cancer data using SGMCP: identified genes and their estimates*

UniGene	D1	D2	D3
Hs.159142		0·0049	−0·0019
Hs.168075	0·0359		
Hs.23311		0·0015	
Hs.240534	0·0001	0·0001	−0·0004
Hs.41587		0·0037	0·0044
Hs.50282	0·0073		0·0082
Hs.646		0·0003	
Hs.75372	0·0006		
Hs.75890	−0·0011	0·0153	
Hs.78225	−0·039	0·0008	
Hs.80768		0·0168	
Hs.98658	0·0067		−0·0023

Table 4. *Analysis of lung cancer data using SGMCP: identified genes and their estimates*

Probe	Gene	UM	HLM	CAN/DF
200786_at	PSMB7		−0·0002	−0·0002
203544_s_at	STAM		−0·002	
207814_at	DEFA6	−0·0001	−0·001	−0·0003
207850_at	CXCL3		−0·004	
208933_s_at	LGALS8	0·004	0·004	−0·009
214261_s_at	ADH6		−0·014	−0·001
214374_s_at	PPFIBP1		−0·0005	
217299_s_at	NBN		−0·012	
217583_at	PAH	−0·0004	−0·016	

genes for D1–D3, respectively. Estimates in Table 3 suggest that different datasets may have different prognosis-associated genes. This partly explains why it fails to unify the identified markers across different breast cancer prognosis studies (Cheang *et al*., 2008). As described in Section 1, multiple factors may contribute to this heterogeneity. Without having access to all the experimental details, we are not able to determine the exact cause of heterogeneity. Although there are overlaps, different approaches identify significantly different sets of genes. This observation is not surprising given the simulation results. We mine the published literature and find that genes identified by SGMCP may have important biological implications. More details are provided in the Appendix. Of note, as there is no objective way to determine which set of genes are 'biologically more meaningful', we do not discuss the biological implications of genes identified by the alternative methods.

To provide a more comprehensive description of the three datasets and various approaches, we also conduct 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, 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 statistics, which measure 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 7·30 (SGMCP), 3·43 (MCP), 2·77 (GMCP) and 3·33 (SGLasso). SGMCP is the only approach that can separate subjects into groups with significantly different survival risks (P-value = 0·007). Such a result suggests that allowing for heterogeneity in markers across datasets can lead to better prediction performance.

(iii) *Analysis of lung cancer studies*

Lung cancer is the leading cause of death from cancer for both men and women in 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. Data were collected from three independent experiments (Shedden *et al*., 2008; Jeong *et al*., 2010; 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. 22 283 genes were profiled in all three studies.

In studies such as Xie *et al*. (2011), the three datasets were combined and analysed. Such a strategy corresponds to the homogeneity model. Here, we analyse using SGMCP (Table 4), MCP (Table 10), GMCP (Table 11) and SGLasso (Table 12). The observed estimation patterns are similar to those for the breast cancer data. The SGMCP estimates again suggest that different datasets may have different prognosis-associated genes. In the Appendix, we show that the SGMCP identified genes may have important biological implications. We also evaluate prediction performance using the resampling approach described in the last subsection. The average logrank statistics are calculated as 10·93 (SGMCP), 2·83 (MCP), 1·73 (GMCP) and 3·07 (SGLasso), respectively. The observation is again similar to that for the breast cancer data.

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 is less restricted. This model may provide a way to accommodate the failure to unify cancer prognosis markers across independent studies (Knudsen, 2006; Cheang, 2008). Under the heterogeneity model, we use sparse group penalization for marker identification. This penalization approach is intuitively reasonable and computationally feasible. Simulation study shows that it has satisfactory performance. It is interesting to note that SGMCP outperforms GMCP even under the homogeneity model. Thus, in practical data analysis, the heterogeneity model and SGMCP can be a 'safer' choice. With the breast cancer and lung cancer prognosis datasets, existing analyses assume the same set of markers across datasets. Our analysis suggests that it may be more reasonable to allow for different sets of markers.

Under the heterogeneity model, marker identification needs to be conducted in two dimensions. Beyond sparse group penalization, composite penalization and multi-step approaches may also be able to achieve such identification. In this paper, we focus on sparse group penalization. Comprehensive investigation and comparison of different approaches are beyond the scope of this paper. The proposed approach is 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. In our numerical study, we observe superior performance of the proposed approach. A limitation of this study is lack of theoretical support. Pursuit of theoretical properties is postponed to future studies.

We thank the editor, associate editor and two reviewers for careful review and insightful comments. Part of the study was conducted when the first author was at Yale University. The authors have been supported by awards CA142774, CA165923 and CA152301 from NIH, DMS0904181 and DMS1208225 from NSF, 2012LD001 from the National Bureau of Statistics of China, and the VA Cooperative Studies Programme of the Department of Veterans Affairs, Office of Research and Development.

The appendix for this paper is available online at the journal's website (http://journals.cambridge.org/grh).

References

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

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

Friedman, J., Hastie, T. & Tibshirani, R. (2010). A note on the group Lasso and a sparse group Lasso. arXiv: 1001.0736.

Guerra, R. & Goldsterin, D. R. (2009). *Meta-Analysis and Combining Information in Genetics and Genomics*, 1st edn. New York: Chapman and Hall/CRC.

Huang, E., Cheng, S. H., Dressman, H., Pittman, J., Tsou, M. H., Horng, C. F., Bild, A., Iversen, E. S., Liao, M., Chen, C. M., West, M., Nevins, J. R. & Huang, A. T. (2003). Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1590–1596.

Huang, J., Breheney, P. & Ma, S. (2012*a*). A selective review of group selection in high dimensional models. *Statistical Science* **27**, 481–499.

Huang, J., Wei, F. & Ma, S. (2012*b*). Semiparametric reregression pursuit. *Statistica Sinica* **22**, 1403–1426.

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

Jeong, Y., Xie, Y., Xiao, G., Behrens, C., Girard, L., Wistuba, I. I., Minna, J. D. & Mangelsdorf, D. J. (2010). Nuclear receptor expression defines a set of prognostic biomarkers for lung cancer. *PLoS Medicine* **7**, e1000378.

Knudsen, S. (2006). *Cancer Diagnostics with DNA Microarrays*. New York: John Wiley and Sons.

Liu, J., Huang, J. & Ma, S. (2013). Integrative analysis of multiple cancer genomic datasets under the heterogeneity model. *Statistics in Medicine*, In press.

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

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

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

Shedden, K., Taylor, J. M., Enkemann, S. A., Tsao, M. S., Yeatman, T. J., Gerald, W. L., Eschrich, S., Jurisica, I., Giordano, T. J., Misek, D. E., Chang, A. C., Zhu, C. Q., Strumpf, D., Hanash, S., Shepherd, F. A., Ding, K., Seymour, L., Naoki, K., Pennell, N., Weir, B., Verhaak, R., Ladd-Acosta, C., Golub, T., Gruidl, M., Sharma, A., Szoke, J., Zakowski, M., Rusch, V., Kris, M., Viale, A., Motoi, N., Travis, W., Conley, B., Seshan, V. E., Meyerson, M., Kuick, R., Dobbin, K. K., Lively, T., Jacobson, J. W. & Beer, D. G. (2008). Gene

https://doi.org/10.1017/S0016672313000086 Published online by Cambridge University Press

expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine* **14**, 822–827.

Sotiriou, C., Neo, S. Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L. & Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population based study. *Proceedings of National Academy of Sciences USA* **100**, 10393–10398.

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

Tsuboi, M., Ohira, T., Saji, H., Hiyajima, K., Kajiwara, N., Uchida, O., Usuda, J. & Kato, H. (2007). The present status of postoperative adjuvant chemotherapy for completely resected non-small cell lung cancer. *Annals of Thoracic and Cardiovascular Surgery* **13**, 73–77.

Van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

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. & Wistuba, I. (2011). Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small cell lung cancer patients. *Clinical Cancer Research* **17**, 5705–5714.

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

https://doi.org/10.1017/S0016672313000086 Published online by Cambridge University Press