Published in final edited form as:

Stat Med. 2013 September 10; 32(20): 3509–3521. doi:10.1002/sim.5780.

Integrative analysis of multiple cancer genomic datasets under the heterogeneity model

Jin Liu^a, Jian Huang^b, and Shuangge Ma^{a,*}

^aDepartment of Biostatistics, School of Public Health, Yale University, 60 College Street, New Haven, CT 06520

^bDepartment of Statistics & Actuarial Science, Department of Biostatistics, University of Iowa, Iowa City, IA 52242

Abstract

In the analysis of cancer studies with high-dimensional genomic measurements, integrative analysis provides an effective way of pooling information across multiple heterogeneous datasets. The genomic basis of multiple independent datasets, which can be characterized by the sets of genomic markers, can be described using the homogeneity model or heterogeneity model. Under the homogeneity model, all datasets share the same set of markers associated with responses. In contrast, under the heterogeneity model, different studies have overlapping but possibly different sets of markers. The heterogeneity model contains the homogeneity model as a special case and can be much more flexible. Marker selection under the heterogeneity model calls for bi-level selection to determine whether a covariate is associated with response in any study at all as well as in which studies it is associated with responses. In this study, we consider two minimax concave penalty (MCP) based penalization approaches for marker selection under the heterogeneity model. For each approach, we describe its rationale and an effective computational algorithm. We conduct simulation to investigate their performance and compare with the existing alternatives. We also apply the proposed approaches to the analysis of gene expression data on multiple cancers.

Keywords

Integrative analysis; Heterogeneity model; Marker selection

1. Introduction

In cancer research, high-throughput profiling has been extensively conducted, searching for markers associated with traits and clinical outcomes. Cancer genomic data has the *large d*, *small n* characteristic, where the sample size n is much smaller than d , the number of profiled genomic measurements (also referred to as “covariates” hereafter). With such data, marker selection is needed along with regularized estimation, as it is expected that only a subset of the profiled covariates are associated with cancer response variables. Among the existing analysis approaches, penalization has attracted much attention. In the analysis of a single dataset, the most popular penalization approach is Lasso. Penalization approaches that may have better selection properties include adaptive Lasso, elastic net, bridge, smoothly clipped absolute deviation (SCAD), minimax concave penalty (MCP) and others. When

there exists a grouping structure among covariates, the group counterparts of the aforementioned penalties have been developed.

In practical data analysis, it has been noted that results obtained from the analysis of a single dataset may be unsatisfactory. In particular, the identified markers may suffer a lack of reproducibility [1, 2]. Multiple factors may contribute to the unsatisfactory performance, among which the most important one is the small sample sizes of individual studies. A series of recent studies have shown that pooling and analyzing multiple studies may effectively increase sample size and so improve performance of the identified markers. “Classic” meta-analysis, where multiple datasets are analyzed separately and then the analysis results are pooled across datasets, has been extensively conducted. Recent research shows that integrative analysis, which is termed as “type III meta-analysis” in epidemiologic studies [3], pools and analyzes raw data from multiple studies and can be more informative.

Regularization methods have been developed for marker selection in the integrative analysis of multiple cancer genomic datasets. Ma *et al.* [4] investigate the integrative analysis of multiple diagnosis studies where the response variables are binary. A composite penalty, where the outer penalty is bridge and the inner penalty is ridge, is developed for marker selection. Huang *et al.* [2] also analyze multiple diagnosis studies. A sparse boosting approach is developed. Here the loss function is not differentiable and may incur high computational cost. Ma *et al.* [5] analyze multiple prognosis studies with censored survival responses. The proposed marker selection approach adopts the composite of MCP (outer) and ridge (inner) penalties. In the aforementioned studies, it is reinforced that multiple studies have the same set of markers associated with cancer responses. Such a model is referred to as the homogeneity model. An alternative is the heterogeneity model, under which different studies have possibly different sets of markers. In [6], a gradient thresholding approach is proposed for cancer marker selection under the heterogeneity model. Drawbacks of the thresholding approach include a lack of well-defined statistical framework and high computational cost.

In this article, we consider the integrative analysis of multiple cancer diagnosis studies with binary response variables. We focus on the heterogeneity model, which includes the homogeneity model as a special case and can be more flexible. We consider two MCP-based penalization approaches. For each approach, we describe its rationale and develop an effective computational algorithm. This study may advance from the existing ones along the following directions. First, it provides a more careful study of the heterogeneity model, which is more challenging than the commonly assumed homogeneity model. Second, the penalization approaches have a more lucid statistical framework than the thresholding approach in [6]. Third, the study on MCP penalization approaches may serve as prototype for other types of penalties. Fourth, it provides a practically useful way of analyzing heterogeneous data from multiple cancer genomic studies.

Analysis of multiple datasets is inevitably more complicated than single-dataset analysis. In integrative analysis, multiple datasets should have a certain degree of comparability. For example, if the overlapped markers are of interest, different studies should have comparable definitions for the outcomes. In addition, the types of genomic measurements in different studies should be comparable. In our data analysis, all datasets have microarray gene expression measurements, although the platforms are different. It may be not sensible to analyze datasets with, for example, gene expression and SNP measurements together. The proposed model and methods can accommodate some, but not all of the heterogeneity across multiple datasets. We fully acknowledge the importance and difficulty of the aforementioned issues. In this article, we focus on the development of two penalized marker

selection methods, and refer to published studies such as [1, 5, 6] for more relevant discussions.

The rest of the article is organized as follows. The data and model settings are described in Section 2. MCP penalized marker selection approaches are described in Section 3. Numerical studies, including simulation in Section 4 and data analysis in Section 5, are conducted to investigate empirical performance. The article concludes with discussion in Section 6.

2. Integrative analysis of multiple cancer diagnosis studies

To better describe the context of the heterogeneity model, consider the integrative analysis of multiple cancer gene expression studies. Although most existing multi-dataset studies focus on the scenario with the same cancer outcome in multiple datasets, studies such as [7] and [6] show that the integrative analysis of multiple datasets on different types of cancers may identify more essential features of cancer pathogenesis. Such effort has been motivated by the observation that despite their seemingly significant differences, all cancer cells share two essential characteristics: uncontrolled growth and local tissue invasion or metastasis. In addition, there is strong evidence that certain cancers share common susceptibility genes. Specific examples include the BRCA1 and BRCA2 tumor suppressor genes, whose mutations are associated with the inherited forms of both breast and ovarian cancers. Over-expression of the HER-2 oncogene has been reported in 10–40% of primary breast and ovarian tumors, and is strongly associated with a poor clinical prognosis. Gene WWOX is a tumor suppressor gene mutated in both breast and prostate cancers. Gene ADH is associated with the development of lung cancer and head/neck cancer. The wound response signature, which is a breast cancer prognosis gene signature, also has predictive power for prognosis of lung cancer and prostate cancer. Many more examples are available in the literature.

Assume that there are M independent studies, and n^m iid observations in study $m (= 1, \dots, M)$. The total sample size is $n = \sum_{m=1}^M n^m$. In study m , denote the response variable as y^m . Consider diagnosis studies where the response variables are binary. Denote x^m as the length d covariates. For simplicity of notation, assume that the same set of covariates are measured in all M studies. Assume the model $y^m \sim \varphi(x^m \beta^m)$, where φ is the known link function and β^m is the length d vector of regression coefficients. Denote β_j^m as the j th component of β^m . Then $\beta_j = (\beta_j^1, \dots, \beta_j^M)'$ is the length M vector of regression coefficients representing the effects of covariate j in M studies. Published studies suggest that it is reasonable to assume the same link function in different studies. However, because of the heterogeneity across studies, for a covariate, its strengths of association with response variables, which are measured with regression coefficients, can differ across studies.

In study $m (= 1, \dots, M)$, denote Y^m as the length n^m vector of response variables and X^m as the $n^m \times d$ covariate design matrix. The M models can be unified into a single model $Y \sim \varphi(X\beta)$, where $Y = (Y^1', \dots, Y^M)'$, $X = \text{diag}(X^1, \dots, X^M)$, and $\beta = (\beta^1', \dots, \beta^M)'$. With binary responses, assume the logistic regression models. Denote $\ell(\beta)$ as the negative log-likelihood function for the unified model. With multiple studies, different marker selection scenarios can be described with the following two models.

Homogeneity model—Under this model, for any $j (= 1, \dots, d)$, it is assumed that

$$I(\beta_j^1=0)=\dots=I(\beta_j^M=0),$$

where $I(\cdot)$ is the indicator function. Or equivalently, the M studies have the same set of markers associated with responses. The homogeneity model is sensible when multiple datasets are generated under comparable protocols (the same outcome variable, similar subject selection criteria, similar profiling approaches, etc). In practice, the homogeneity model may be too restricted.

Heterogeneity model—Under the heterogeneity model, a covariate can be associated with response variables in some but not other studies. Here different studies may have possibly different sets of markers associated with responses. Of note, the heterogeneity model allows for the extreme case where different studies have completely different sets of markers. The heterogeneity model is sensible when different studies are on different cancers or different subtypes of the same cancer. Another possible scenario is where subjects in different studies have different confounders. For covariates not intervened with confounders, their importance is consistent across multiple studies. However, for other covariates, after adjusting for confounders, they may be important in some studies but not others.

Because of the simplicity of homogeneity model, it has been assumed in multiple studies. The homogeneity model is a special case of the heterogeneity model. In practical data analysis, unless there is a strong reason for the homogeneity model, it is sensible to first consider the heterogeneity model. If it is found that the selection indicators are mostly consistent across studies, then a homogeneity model can be adopted.

3. Penalized marker selection

Under the heterogeneity model, the model and regression coefficients have two dimensions. The first is the covariate 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 covariate is associated with response in any study at all* as well as *in which studies it is associated with responses*.

3.1. MCP

Integrative analysis approaches investigated in this study are based on the MCP penalty, which is proposed in [8] for the analysis of a single dataset. It belongs to the family of quadratic spline penalties. In this section, consider the analysis of a single dataset where there is a length d vector of regression coefficient β . Denote β_j as the j th element of β . The MCP penalty is defined as

$$P_{MCP}(\beta; \lambda, \gamma) = \sum_{j=1}^d \rho(\beta_j; \lambda, \gamma) \quad (1)$$

where

$$\rho(t; \lambda, \gamma) = \lambda \int_0^{|t|} (1 - x/(\gamma\lambda))_+ dx. \quad (2)$$

Here λ is the tuning parameter, γ is the regularization parameter that controls the concavity of ρ , and $x_+ = xI(x \geq 0)$. The MCP can be understood by considering its derivative, which is

$\dot{\rho}(t; \lambda, \gamma) = \lambda(1 - |t|/(\gamma\lambda))_+ \text{sgn}(t)$, where $\text{sgn}(t) = -1, 0$, or 1 if $t < 0, = 0$, or > 0 , respectively. As $|t|$ increases from zero, MCP begins by applying the same rate of penalization as Lasso, but continuously relaxes that penalization until $|t| > \gamma\lambda$, a condition under which the rate of penalization drops to zero. It provides a continuum of penalties where the Lasso penalty corresponds to $\gamma = \infty$ and the hard-thresholding penalty corresponds to $\gamma \rightarrow 1+$. Compared with other nonconvex penalties, MCP may be preferred because of its computational simplicity and satisfactory empirical properties [8, 9].

3.2. Sparse group MCP

In integrative analysis, the effect of one covariate in multiple datasets is represented by a *group* of regression coefficients with size equal to the number of studies. Under the heterogeneity model, marker selection demands determining which groups of regression coefficients are nonzero (the first level of selection), as well as within specific groups, which coefficients are nonzero (the second level of selection). Such selection can be achieved by the combination of two penalties, with one for group-level selection and the other for within-group selection.

In particular, consider the *sparse group MCP* (sparse gMCP) penalty defined as

$$\sum_{j=1}^d \left\{ \rho(\|\beta_j\|_{\Sigma_j}; \sqrt{d_j}\lambda_1, a) + \sum_{k=1}^M \rho(|\beta_j^k|; \lambda_2, b) \right\}. \quad (3)$$

Here ρ is defined in (2), $\|\beta_j\|_{\Sigma_j} = \left\| \sum_j^{1/2} \beta_j \right\|_2$, $\|\cdot\|_2$ is the L_2 norm, $\Sigma_j = X[\cdot, j]^\top X[\cdot, j]$, $X[\cdot, j]$ is the $n \times d_j$ submatrix of X that corresponds to β_j , and d_j is the size of group j . When covariate sets in multiple studies are fully matched, $d_j \equiv M$, and this rescaling parameter can be omitted. We keep d_j so that this approach can be directly applied to the scenario with partially matched covariate sets across studies. For each dataset, we standardize the covariates to have marginal means zero and satisfy $n^{-1} X[\cdot, j]^\top X[\cdot, j] = I_{d_j}$, where I_{d_j} is the $d_j \times d_j$ identity matrix.

In (3), the first term is a group MCP (gMCP) penalty. The gMCP has been developed for single-dataset analysis in [10] and others. It is the group counterpart of MCP. Here it selects covariates that are associated with responses in at least one study. For a selected covariate, the second penalty, which is a MCP penalty, identifies in which studies it is associated with responses. By varying the tuning parameters, the sparse gMCP penalty includes the MCP and gMCP as special cases. The sparse gMCP penalty has been partly motivated by the sparse group Lasso penalty in [11], which is developed for the analysis of a single dataset with a continuous response and linear regression model. We adopt MCP based penalties because of their better performance over Lasso [8, 12], and analyze multiple heterogeneous datasets with categorical responses under logistic models.

Computational algorithm—With a sparse group Lasso penalty in the analysis of a single dataset under a linear model, an effective algorithm based on coordinate descent is presented in [11]. Unfortunately, because of the significant differences in penalties and data and model settings, that algorithm is not directly applicable under the present setup.

We first consider the integrative analysis of multiple studies with continuous responses and linear regression models. Using notations similar to those in Section 2, we have a unified model $Y = X\beta + \varepsilon$. Here the objective function is

$$R(\beta) = \frac{1}{2n} \left\| Y - \sum_{j=1}^d X[\cdot, j] \beta_j \right\|_2^2 + \sum_{j=1}^d \left\{ \rho(\|\beta_j\|_2; \sqrt{d_j} \lambda_1, a) + \sum_{k=1}^M \rho(\|\beta_j^k\|_2; \lambda_2, b) \right\}. \quad (4)$$

In a coordinate descent algorithm, consider update of estimate for a single group. Given the group parameter vectors $\beta_l (l \neq j)$ fixed at their current estimates $\tilde{\beta}_l$, we seek to minimize the objective function $R(\beta)$ with respect to the j th group parameter β_j . Consider

$$R(\beta_j) = \frac{1}{2n} \|r_{-j} - X[\cdot, j] \beta_j\|_2^2 + \rho(\|\beta_j\|_2; \sqrt{d_j} \lambda_1, a) + \sum_{k=1}^M \rho(\|\beta_j^k\|_2; \lambda_2, b), \quad (5)$$

where $r_{-j} = Y - \sum_{l \neq j} X[\cdot, l] \tilde{\beta}_l$. The first order derivative of expression (5) is:

$$\frac{\partial R(\beta_j)}{\partial \beta_j} = -\frac{1}{n} X[\cdot, j]' r_{-j} + \frac{1}{n} X[\cdot, j]' X[\cdot, j] \beta_j + \frac{\beta_j}{\|\beta_j\|_2} \begin{cases} \sqrt{d_j} \lambda_1 - \frac{\|\beta_j\|_2}{a}, & \text{if } \|\beta_j\|_2 \leq a \sqrt{d_j} \lambda_1 \\ 0, & \text{if } \|\beta_j\|_2 > a \sqrt{d_j} \lambda_1 \end{cases} + t \quad (6)$$

where

$$t = \left(\text{sgn}(\beta_j^1) \begin{cases} \lambda_2 - \frac{\|\beta_j^1\|_2}{b}, & \text{if } \|\beta_j^1\|_2 \leq b \lambda_2 \\ 0, & \text{if } \|\beta_j^1\|_2 > b \lambda_2 \end{cases}, \dots, \text{sgn}(\beta_j^M) \begin{cases} \lambda_2 - \frac{\|\beta_j^M\|_2}{b}, & \text{if } \|\beta_j^M\|_2 \leq b \lambda_2 \\ 0, & \text{if } \|\beta_j^M\|_2 > b \lambda_2 \end{cases} \right)'.$$

By setting expression (6) as zero, we have:

$$-z_j + g(\beta_j) \beta_j + t = 0, \quad (7)$$

where

$$z_j = n^{-1} X[\cdot, j]' r_{-j} - (z_j^1, \dots, z_j^M)', \quad g(\beta_j) = \left(1 + \frac{1}{\|\beta_j\|_2} \right) \begin{cases} \sqrt{M_j} \lambda_1 - \frac{\|\beta_j\|_2}{a}, & \text{if } \|\beta_j\|_2 \leq a \sqrt{M_j} \lambda_1 \\ 0, & \text{if } \|\beta_j\|_2 > a \sqrt{M_j} \lambda_1 \end{cases}.$$

Denote z_j^k as the k th element of z_j . First, fix $g(\beta_j)$ at the current estimate $\tilde{\beta}_j$. We use g short for $g(\tilde{\beta}_j)$. The k th element in equation (7) can be rewritten as:

$$-\frac{z_j^k}{g} + \beta_j^k + \text{sgn}(\beta_j^k) \begin{cases} \lambda_2 - \frac{\|\beta_j^k\|_2}{b}, & \text{if } \|\beta_j^k\|_2 \leq b \lambda_2 \\ 0, & \text{if } \|\beta_j^k\|_2 > b \lambda_2 \end{cases} = 0. \quad (8)$$

The solution to equation (8) is

$$\widehat{g\beta_j^k} = \begin{cases} \frac{S_1(z_j^k, \lambda_2)}{1 - \frac{1}{b_g}}, & \text{if } |z_j^k| \leq b \lambda_2 g \\ z_j^k, & \text{if } |z_j^k| > b \lambda_2 g. \end{cases}$$

Here $S_1(z, \lambda) = \text{sgn}(z)(|z| - \lambda)_+$. For $k = 1, \dots, M$, set $u_k = \widehat{g\beta_j^k}$ and $u = (u_1, \dots, u_M)'$. Taking u back into its definition,

$$\beta_j + \frac{\beta_j}{\|\beta_j\|_2} \begin{cases} \sqrt{d_j} \lambda_1 - \frac{\|\beta_j\|_2}{a}, & \text{if } \|\beta_j\|_2 \leq a \sqrt{d_j} \lambda_1 \\ 0, & \text{if } \|\beta_j\|_2 > a \sqrt{d_j} \lambda_1 \end{cases} = u. \quad (9)$$

Expression (9) can be solved in a similar manner as with the gMCP, leading to

$$\widehat{\beta}_j = \begin{cases} \frac{a}{a-1} S_2(u, \sqrt{d_j} \lambda_1), & \text{if } \|u\|_2 \leq a \sqrt{d_j} \lambda_1 \\ u, & \text{if } \|u\|_2 > a \sqrt{d_j} \lambda_1 \end{cases}, \quad (10)$$

where $S_2(u, \lambda) = (1 - \lambda/\|u\|_2)_+ u$.

Solving equation (7) amounts to iteratively calculating equations (8) to (10) until convergence. Overall, consider the following algorithm. With fixed tuning parameters,

1. Initialize $s = 0$, the estimate $\beta^{(0)} = (0, \dots, 0)'$, and the vector of residuals $r = Y - X\beta^{(0)}$;
2. For $j = 1, \dots, d$,
  - a. Compute $\beta_j^{(s+1)}$. This is achieved by solving equation (7), which amounts to iterating equations (8) to (10) until convergence;
  - b. Update $r \leftarrow r - X[:, j](\beta_j^{(s+1)} - \beta_j^{(s)})$;
Update $s \leftarrow s + 1$;
3. Repeat Step 2 until convergence.

Solution path—To better understand the sparse gMCP approach, we simulate four datasets under the scenario described in row 1 of Table 1. We fix $a = 6$ and $b = 6$. For four different $\alpha (= \lambda_1/(\lambda_1 + \lambda_2))$ values, we show the estimates as a function of $\lambda (= \lambda_1 + \lambda_2)$ in Figures 1–3 (Appendix). For a better view, we only plot a covariate associated with responses in all four studies (Figure 1), a covariate associated with responses in two out of four studies (Figure 2), and a covariate not associated with any response (Figure 3).

We can see that for the covariate associated with responses in all studies, the four estimated regression coefficients are nonzero for small to moderate λ values. For the covariate associated with responses in only two studies, two of the four estimates are shrunk to zero with relatively small λ values. α plays an important role in determining the shape of solution path. For example with $\alpha = 0.6$, two estimates are shrunk to zero with $\lambda < 0.2$, one estimate is shrunk to zero around $\lambda = 0.25$, and the fourth one is shrunk to zero around $\lambda = 0.35$. However with $\alpha = 0.9$, all four estimates are shrunk to zero around $\lambda = 0.28$. For the covariate not associated with any response, the estimates are zero unless with very small λ values.

3.3. Composite MCP

In the analysis of a single dataset, when there is a grouping structure among covariates, bi-level selection can also be achieved using composite penalization. The intuition is similar to that with sparse gMCP, which is to use a group penalty for marker selection at the group level and a second penalty for marker selection at the within-group level. The composite of the two penalties will then be able to conduct selection at both levels. Here the way that two penalties are combined differs from that with sparse gMCP.

Consider the *composite MCP (cMCP)* penalty, where the penalty takes the form

$$P_{cMCP}(\beta; \lambda_1, \lambda_2, a, b) = \sum_{j=1}^d \rho \left(\sum_{k=1}^M \rho(\beta_j^k; \lambda_2, b); \lambda_1, a \right). \quad (11)$$

Here notations have similar implications as described in the previous sections. With cMCP, the outer penalty is MCP, and the inner penalty is MCP.

Computational algorithm—Consider the integrative analysis of multiple datasets with continuous responses and linear models. In a coordinate descent algorithm, the key is update of estimate for a single group (covariate). Unfortunately, even under this simple setup, the cMCP approach does not have a simple form for updating individual groups. To tackle this problem, we adopt a local coordinate descent approach. Consider update with the j th group. By taking the first order Taylor series approximation about $\tilde{\beta}_j$ and evaluating at $\tilde{\beta}_j$, the penalty as a function of β_j^k is approximately proportional to $\tilde{\lambda}_{jk} |\beta_j^k|$ where

$$\tilde{\lambda}_{jk} = \rho' \left(\sum_{m=1}^M \rho(\tilde{\beta}_j^m; \lambda_2, b); \lambda_1, a \right) \rho'(\tilde{\beta}_j^k; \lambda_2, b). \quad (12)$$

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

$$\widehat{\beta}_j^k = S_1(z, \lambda),$$

with $S_1(z, \lambda) = \text{sgn}(z)(|z| - \lambda)_+$ and z and λ to be defined below. Consider the following coordinate descent algorithm. With fixed tuning parameters,

1. Initialize $s = 0$, the estimate $\beta^{(0)} = (\beta_1^{(0)}, \dots, \beta_d^{(0)})' = (0, \dots, 0)'$, and the vector of residuals $r = Y - X\beta^{(0)}$;
2. For $j = 1, \dots, d$,
  - a. Calculate $\tilde{\lambda}_{jk}$ according to expression (12).
  - b. Calculate $z_j = n^{-1} X[\cdot, j]' r + \beta_j^{(s)}$.
  - c. For $k = 1, \dots, M$, update $\beta_j^{k(s+1)} \leftarrow S_1(z_j^k; \tilde{\lambda}_{jk})$, where z_j^k is the k th 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.

Plots of solution path are similar to those in Figures 1–3 (Appendix) and available from the authors.

3.4. Penalized logistic regression

In Section 3.2 and 3.3, we describe the computational algorithms under linear regression models. Now consider logistic regression model. In study $m (= 1 \dots M)$, consider

$\{(y_i^m, x_i^m), i=1 \dots n^m\}$, the n^m iid copies of (y^m, x^m) . Under the logistic regression model, $p(x_i^m) = Pr(y_i^m = 1 | x_i^m) = 1 / (1 + \exp(-(\beta_0^m + \sum_{j=1}^d x_{ij}^m \beta_j^m)))$. The negative log-likelihood function is

$$\ell(\beta_0^m, \beta^m) = - \sum_{i=1}^{n^m} \left[y_i^m \cdot (\beta_0^m + \sum_{j=1}^d x_{ij}^m \beta_j^m) - \log(1 + e^{(\beta_0^m + \sum_{j=1}^d x_{ij}^m \beta_j^m)}) \right]. \quad (13)$$

The overall objective function is $\frac{1}{n} \sum_{m=1}^M \ell(\beta_0^m, \beta^m)$.

Under the logistic regression model, there is no simple, closed-form solution for update with a single group. To tackle this problem, we resort to a majorization-minimization (MM) approach [13]. Note that the negative log-likelihood function is convex. With the MM approach, when the current estimate is $(\tilde{\beta}_0^m, \tilde{\beta}^m)$, we majorize the negative log-likelihood by a quadratic loss given by

$$\ell_Q^m(\beta_0^m, \beta^m | \tilde{\beta}_0^m, \tilde{\beta}^m) = \frac{1}{8} \sum_{i=1}^{n^m} (z_i^m - \beta_0^m - \sum_{j=1}^d x_{ij}^m \beta_j^m)^2 + C(\tilde{\beta}_0^m, \tilde{\beta}^m),$$

where $z_i^m = \tilde{\beta}_0^m + \sum_{j=1}^d x_{ij}^m \tilde{\beta}_j^m + \frac{y_i^m - \tilde{p}(x_i^m)}{\tilde{p}(x_i^m)(1 - \tilde{p}(x_i^m))}$ and $\tilde{p}(x_i^m) = \frac{1}{1 + \exp[-(\tilde{\beta}_0^m + \sum_{j=1}^d x_{ij}^m \tilde{\beta}_j^m)]}$, and $C(\tilde{\beta}_0^m, \tilde{\beta}^m)$ is free of (β_0^m, β^m) . Note that the above approximation can be conducted for all M studies separately. The approximated objective function has a least squares form. With fixed tuning parameters, the computational algorithm consists of a sequence of nested loops:

Outer loop: For $m = 1, \dots, M$, update the majorized quadratic function ℓ_Q^m using the current estimate $(\tilde{\beta}_0^m, \tilde{\beta}^m)$.

Inner loop: Run the algorithm developed for the penalized least-squares problem with the objective function $\sum_{m=1}^M \ell_Q^m$.

For both of the proposed approaches, we use coordinate descent for computation. The coordinate descent algorithm starts with null models. In each iteration, it cycles through all d covariates (groups). For each covariate, it only involves simple computations, and the update can be accomplished easily. We use the L_2 norm of the difference between two consecutive estimates smaller than 0.01 as the convergence criterion. Convergence is achieved in all of our simulation and data analysis. For each of the proposed approach, in the objective function, the first term (negative log-likelihood function) is continuously differentiable and regular in the sense of [14]. The penalty term is separable. Thus, the coordinate descent algorithm converges to a coordinate-wise minimum of the first term, which is also a stationary point [14]. Research code written in R is available at <http://works.bepress.com/shuangge/>.

3.5. Tuning parameter

Both the sparse gMCP and cMCP approaches involve four tuning parameters: $\lambda_1, \lambda_2, a, b$. Properties of the estimates are jointly determined by the tuning parameters. Generally speaking, smaller values of a and b 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 . It is therefore advisable to choose the values of a and b that are big enough to avoid this problem but not too big. As suggested in published studies, we have experimented with multiple values for a and b , particularly including 1.8, 3, 6 and 10.

First consider linear regression models. For sparse gMCP, the effect of λ_1 is similar to that with gMCP, shrinking the estimated coefficients at the group level. Meanwhile, the effect of λ_2 is similar to that with MCP, shrinking the estimated coefficients at the within-group level. Let $\lambda_{2\max}$ be the smallest λ_2 for which all regression coefficients are shrunk to zero. Let $\lambda_{1\max}(\lambda_2)$ be the smallest λ_1 under a fixed λ_2 for which all regression coefficients are

shrunk to zero. Calculations show that $\lambda_{2\max} = n^{-1} \max_{j=1, \dots, d; k=1, \dots, M} |\sum_{i=1}^{n^k} x_{ij}^k y_i^k|$ and

$\lambda_{1\max}(\lambda_2) = \max_{j=1, \dots, d} \|S_1(n^{-1} X[\cdot, j]' Y, \lambda_2)\|_2 / \sqrt{d_j}$. As $\lambda_{1\max}$ depends on the value of λ_2 , in general, λ_1 and λ_2 cannot vary freely. In numerical study, λ_1 and λ_2 can be chosen via searching a two-dimensional grid. We suggest the reparameterization $\lambda = \lambda_1 + \lambda_2$ and $\alpha = \lambda_1 / \lambda$. To reduce computational cost, one may consider only a few values for α or even fix its value. For cMCP, the product of λ_1 and λ_2 controls sparsity. Let $\tau = \lambda_1 \lambda_2$. Denote $\tau_{\max}$ as the smallest τ that shrinks all estimated coefficients to zero. We have

$\tau_{\max} = n^{-1} \max_{j=1, \dots, d; k=1, \dots, M} |\sum_{i=1}^{n^k} x_{ij}^k y_i^k|$. With logistic regression models, properties of λ_1 , λ_2 are similar to those described above. The approximated values of $\lambda(\tau)_{\max}$ can be derived by conducting weighted least squares expansion of the log-likelihood functions, however, do not have simple analytic forms.

In numerical study, we select tuning parameters via V-fold cross validation with $V = 5$. We have experimented with multiple values using simulated data and found that $a = 1.8$ and $b = 1.8$ may 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_1, λ_2) , one may expect that their values cannot go down to very small values since there are regions not locally convex. The criteria over non-locally convex regions may go up and down. To avoid such unexpected regions, we select (λ_1, λ_2) where the criterion first goes up [15].

4. Simulation study

We conduct simulation to better gauge performance of the proposed approaches. We simulate four datasets, each with 50, 100 or 200 subjects. For each subject, we simulate 1,000 covariates. The covariates are jointly normally distributed, with marginal means equal to zero and variances equal to one. We consider two correlation structures. The first is the autoregressive correlation, where covariates j and l have correlation coefficient $\rho^{|j-l|}$. 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, covariates j and l have correlation coefficient 0.33 if $|j-l| = 1$ and 0 otherwise; Under the second scenario, covariates j and l have correlation coefficient 0.6 if $|j-l| = 1$, 0.33 if $|j-l| = 2$, and 0 otherwise. Consider the heterogeneity model, where studies 1 and 2 share the same ten response-associated covariates, and studies 3 and 4 share the same ten response-associated covariates. The two sets have five overlapped covariates. Thus, across the four datasets, there are a total of forty true positives, corresponding to fifteen covariates. Note that here the degree of overlap among different marker sets can be higher than what is expected in practice (as can be partly seen from data analysis in Section 5). The nonzero regression coefficients in study 1–4 are (0.7, 1.1, 1.3, 1.1, 0.9, 0.9, 1.3, 0.9, 1.7, 0.3), (0.5, 1.2, 1.1, 0.9, 0.8, 1.1, 0.4, 1.3, 0.5, 1.2), (1.1, 0.8, 0.9,

1.2, 1.2, 1.2, 0.7, 1.1, 0.7, 1.3), and (1.2, 0.9, 0.6, 1.1, 1.3, 0.8, 1.5, 1.9, 0.6, 0.8), respectively. We generate the binary response variables from the logistic regression models with zero intercepts and Bernoulli distributions.

We analyze the simulated data using the proposed approaches, MCP and gMCP. MCP analyzes each dataset separately and is not an integrative analysis approach. gMCP conducts one-level selection and conceptually may not be sensible under the heterogeneity model. These two approaches are included here because they have been investigated in published literature particularly in the context of multi-dataset analysis and also because of their tight connections with the proposed approaches. With all approaches, tuning parameters are selected using 5-fold cross validation. Summary statistics based on 100 replicates, including the number of covariates identified and number of true positives, are shown in Table 1. It can be seen that as sample size increases, all approaches identify more true positives, possibly at the price of more false positives. Selection performance also depends on the correlation structure. For example with $n^m = 200$ and autoregressive correlation, cMCP identifies 51.42 positives with 30.64 true positives on average when $\rho = 0.2$. In comparison, when $\rho = 0.7$, it identifies 14.24 positives with 13.00 true positives on average. Comparing different approaches suggests the following. MCP identifies a smaller number of true positives than the alternatives. It is interesting to note the reasonable performance of gMCP. This is not surprising, since the degree of overlaps is high in simulation, which may favor one-level selection. Comparison among the four two-level selection approaches (cMCP and sparse gMCP with three different α values) does not lead to very definitive results. Sparse gMCP with $\alpha = 0.5$ seems to have better performance, particularly when $n^m = 100$ and 200. It can identify more true positives than cMCP and sparse gMCP with $\alpha = 0.2$, while having a much lower number of false positives than sparse gMCP with $\alpha = 0.8$. Note that there are other ways of presenting marker identification results, for example using false discovery rate. With the number of true positives and model size, other accuracy measures can be easily computed.

5. Analysis of multiple cancer gene expression studies

An important application of the proposed approaches is the analysis of microarray gene expression data on multiple types of cancers. Such analysis may reveal the common genomic basis of seemingly unrelated cancers. The common susceptibility genes may suggest pathogenetic connections of multiple cancers, and are likely to play a more fundamental role in cancer development and progression. The biological importance of such a problem has been discussed in great detail in [7], [6] and others. In this study, we focus on the implementation and comparison of different approaches and refer to other publications for the biological aspect of this problem. We collect four gene expression studies. The first study is on liver cancer [16] and contains 76 normal and 76 tumor subjects. cDNA arrays were used for profiling. The second study is on prostate cancer [17] and contains 50 normal and 52 tumor subjects. Affymetrix U95A arrays were used. The third study is on lymphoma [18] and contains 31 normal and 31 tumor subjects. cDNA arrays were used. The fourth study is on colon cancer [19] and contains 22 normal and 22 tumor subjects. Hu6800 arrays were used. The four studies used four different platforms for gene profiling. We refer to the original publications for more details. The four datasets and a few others have been considered in [20] and shown to have a certain degree of comparability. But the analysis in [20] is single-marker based and has a strategy significantly different from the present one. We note that the proposed approaches as well as alternatives can directly accommodate partially matched covariate sets. However, to be able to make conclusions on gene effects across multiple types of cancers, we focus on the analysis of 703 genes that are measured in all four studies. As it is expected that the number of cancer-associated genes to be far smaller than 703, focusing on this set of genes is expected to have small impact on marker

identification. Prior to analysis, we conduct microarray normalization with each dataset separately. The gene expressions are then standardized as described in Section 3.

We apply the proposed approaches and two alternatives. The identified genes and their corresponding estimates are shown in Table 2 (sparse gMCP), 3 (cMCP), 4 (MCP, Appendix) and 5 (gMCP, Appendix), respectively. The sparse gMCP approach identifies a total of 14 genes. Liver cancer and colon cancer share five common genes; Prostate cancer and liver cancer share one common gene; Prostate cancer and colon cancer share two common genes; And lymphoma only has one gene not shared by other types of cancer. Among the identified genes, gene AIMP1 is identified as associated with all four cancers. AIMP1 is a non-catalytic component of the multisynthase complex that stimulates the catalytic activity of cytoplasmic arginyl-tRNA synthase, binds tRNA, possesses inflammatory cytokine activity, negatively regulates TGF-beta signaling through stabilization of SMURF2 by binding to SMURF2 and inhibiting its SMAD7-mediated degradation, and has many other functions. In [21], Kim and others have shown that the expression of AIMP members (AIMP1, AIMP2 and AIMP3) is downregulated in gastric cancer (GC) and colorectal cancer (CRC) tissues. Three other genes are identified as associated with three types of cancers. Hayward et al. [23] found that NEK2 protein is significantly up-regulated in preinvasive *in situ* ductal carcinomas of the breast as well as in invasive breast carcinomas. Desmots et al. [24] demonstrated that the expression of GSTA4 is differently expressed in liver cancer. D'Souza et al. [25] showed that null mutations in a member of fermitin family result in increased risk of cancer. RAN is a member of the RAS oncogene family. Ferrante et al. [26] showed that mutations in the RAS oncogene occur with great frequency in a number of human cancers including lung, pancreas, and colon cancer. The cMCP approach identifies a total of 15 genes. However, only one gene is identified as associated with two cancers. The MCP approach identifies a total of 27 genes, with no gene identified as associated with multiple cancers. The gMCP approach reinforces the homogeneity model and identifies 8 genes to be associated with all four types of cancers. Although there exist overlaps, the sets of genes identified by different approaches are hugely different. Such an observation is not surprising considering what is observed in simulation and the fact that practical data can be noisier than simulated data.

As it is difficult to objectively evaluate the identification accuracy, we evaluate prediction performance, which may provide an indirect evaluation of the identification result. In particular, it is expected that if a set of genes has better prediction performance, it may have more important implications. We adopt a resampling approach. In particular, we first randomly partition the data into a training and a testing sets, with sizes 4:1. We conduct marker selection and generate prediction models using the training set, make prediction for subjects in the testing set, and compute prediction errors. To avoid an extreme partition, the process is repeated 100 times. The average numbers of mistakenly predicted subjects are 3 (sparse gMCP), 6 (cMCP), 5 (MCP) and 7 (gMCP), respectively. The sparse gMCP approach has the best prediction performance, although it should be noted that the differences among different approaches are small. We have also examined specificity and sensitivity of prediction and reached similar conclusions.

With the four public datasets, information on clinical risk factors and environmental exposures is not available. With other datasets, if such variables are present, they need to be properly accounted for. Since they are usually of low dimensionality, selection may not be needed.

6. Discussion

In the analysis of high dimensional cancer genomic data, integrative analysis provides an effective way of pooling and analyzing multiple datasets. The heterogeneity model is more flexible than the homogeneity model but has not been carefully studied in the literature. In this article, we develop two MCP-based penalized marker selection approaches for the analysis of heterogeneity model. We choose MCP because of its satisfactory performance in single-dataset analysis. We suspect that it is possible to follow a similar strategy and replace the MCP penalties with bridge, SCAD and other penalties. We expect such an extension to be nontrivial and postpone it to future studies. Our simulation study suggests that the performance of cMCP and sparse gMCP is data-dependent. Such an observation highlights the necessity of developing and applying multiple approaches in practice. We have explored the possibility of identifying a data-driven optimal approach but failed. The proposed approaches are used to analyze gene expression data. We conjecture that with minor or no modification, they can be used to analyze other types of (e.g, SNP) data.

In this study, we have focused on methodological development and numerical study. We expect the theoretical study of proposed approaches to be highly challenging. With a single dataset under the linear regression model, asymptotic properties of the MCP approach are established in [8]. Huang *et al.* [10] establish properties of the gMCP under a very different data/model setup. In the aforementioned two studies, the selection consistency properties are established under the $d \gg n$ condition. The cMCP approach has gMCP as its inner penalty and MCP as its outer penalty. Considering the results in [27] and [8], we suspect that cMCP may be able to achieve asymptotic selection consistency at both group and within-group levels. Of note, all the aforementioned theoretical results are established for the analysis of a single dataset under linear regression model. Because of heterogeneity, integrative analysis can be considerably more complicated than the analysis of a single dataset. In addition, logistic regression can be more complicated than linear regression model. Properties of the sparse gMCP are less lucid as there is a lack of related study in the literature. The theoretical aspect of the proposed approaches is worth investigation in future study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Associate Editor and two reviewers for careful review and insightful comments, which have led to a significant improvement of this article. This study has been partly supported by CA142774 and CA165923 from NCI, NIH.

References

1. Guerra, R.; Goldstein, DR. Meta-Analysis and Combining Information in Genetics and Genomics. 1. Chapman and Hall/CRC; 2009.
2. Huang Y, Huang J, Shia BC, Ma S. Identification of cancer genomic markers via integrative sparse boosting. Biostatistics. 2011 In press.
3. Blettner M, Sauerbrei W, Schlehofer B, Scheuchensplig T, Friedenreich C. tradiational reviews, meta-analyses and pooled analyses in epidemiology. International Journal of Epidemiology. 1999; 28:1–9. [PubMed: 10195657]
4. Ma S, Huang J, Song X. Integrative analysis and variable selection with multiple high-dimensional datasets. Biostatistics. In press.
5. Ma S, Huang J, Wei F, Xie Y, Fang K. Integrative analysis of multiple cancer prognosis studies with gene expression measurements. Statistics in Medicine. 2011; 30:3361–3371. [PubMed: 22105693]

6. Ma S, Huang J, Moran M. Identification of genes associated with multiple cancers via integrative analysis. *BMC Genomics*. 2009; 10:535. [PubMed: 19919702]
7. Rhodes DR, Yu J, Shanker K, Deshpande N, Ghosh R, Barrette T, Pandey A, Chinnaiyan AM. Large-scale meta-analysis of cancer micorarray data identifies common transcriptional profiles of neoplastic tranformation and progression. *PNAS*. 2004; 101:9309–9314. [PubMed: 15184677]
8. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Ann Statist*. 2010; 38:894–942.
9. Mazumder R, Friedman J, Hastie T. SparseNet:Coordinate descent with non-convex penalties. *JASA*. 2011; 106:1125–1138.
10. Huang J, Wei F, Ma S. Semiparametric reregression pursuit. Accepted for publication by *Statistica Sinica*. 201110.5705/ss.2010.29
11. Friedman J, Hastie T, Tibshirani R. A note on the group Lasso and a sparse group Lasso. 2010 arXiv:1001.0736.
12. Zhang CH, Huang J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann Statist*. 2008; 36:1567–1594.
13. Ortega, J.; Rheinboldt, W. *Classics in Applied Mathematics*. 4. SIAM; Philadelphia, PA: 2000. Iterative solution of nonlinear equations in several variables.
14. Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*. 2001; 109:475–494.
15. Breheny P, Huang J. Penalized methods for bi-level variable selection. *Statist and Its Interface*. 2009; 2:369–380.
16. Chen X, Cheung S, So S, Fan S, Barry C, Higgins J, Lai K, Ji J, Dudoit S, Ng I, Rijn M, Botsetein D, Brown P. Gene expression patterns in human liver cancers. *Molecular Biology of the Cell*. 2002; 13:1929–1939. [PubMed: 12058060]
17. Singh D, Febbo P, Ross K, Jackson D, Manola J, Ladd C, Tamayo P, Renshaw A, D'Amico A, Richie J, Lander E, Loda M, Kantoff P, Golub T, Sellers W. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*. 2002; 1:203–209. [PubMed: 12086878]
18. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JJ, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000; 403:503–511. [PubMed: 10676951]
19. Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res*. 2001; 61:3124–3130. [PubMed: 11306497]
20. Choi JK, Yu U, Yoo OJ, Kim S. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*. 2005; 21:4348–4355. [PubMed: 16234317]
21. Kim SS, Hur SY, Kim YR, Yoo NJ, Lee SH. Expression of AIMP1, 2 and 3, the scaffolds for the multi-tRNA synthetase complex, is downregulated in gastric and colorectal cancer. *Tumori*. 2011; 97(3):380–385.10.1700/912.10038 [PubMed: 21789020]
22. Alaminos M, Davalos V, Ropero S, Setien S, Paz MF, Herranz M, Fraga MF, Mora J, Cheung NK, Gerald WL, Esteller M. EMP3, a Myelin-Related Gene Located in the Critical 19q13.3 Region, Is Epigenetically Silenced and Exhibits Features of a Candidate Tumor Suppressor in Glioma and Neuroblastoma. *Cancer Res*. 2005; 65:2565.10.1158/0008-5472 [PubMed: 15805250]
23. Hayward DG, Clarke RB, Faragher AJ, Pillai MR, Hagan IM, Fry AM. The Centrosomal Kinase Nek2 Displays Elevated Levels of Protein Expression in Human Breast Cancer. *Cancer Res*. 2004; 64:7370.10.1158/0008-5472 [PubMed: 15492258]
24. Desmots F, Rissel M, Loyer P, Turlin B, Guillouzo A. Immunohistological analysis of glutathione transferase A4 distribution in several human tissues using a specific polyclonal antibody. *J Histochem Cytochem*. 2001; 49(12):1573–1579.10.1177/002215540104901211 [PubMed: 11724905]

25. D'Souza MA, Kimble RM, McMillan JR. Kindler Syndrome Pathogenesis and Fermitin Family Homologue 1 (Kindlin-1) Function. *Dermatologic Clinics*. 2010; 28(1):115–118.10.1016/j.det.2009.10.012 [PubMed: 19945623]
26. Ferrante K, Winograd B, Canetta R. Promising new developments in cancer chemotherapy. *Cancer Chemother Pharmacol*. 1999; 43(Suppl):S61–8. [PubMed: 10357561]
27. Huang J, Ma S, Xie H, Zhang CH. A group bridge approach for variable selection. *Biometrika*. 2009; 96:339–355. [PubMed: 20037673]

Table 1

Simulation: in each cell, the first row is number of true positives (standard deviation), and the second row is model size (standard deviation).

n^m	MCP	gMCP	cMCP	Sparse gMCP		
				$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$
Auto-regressive $\rho = 0.2$						
50	1.70(1.34)	11.88(4.32)	1.54(2.02)	0.58(1.11)	1.50(2.77)	2.78(3.53)
	4.36(2.13)	20.80(10.09)	3.42(4.25)	1.06(1.95)	2.00(3.75)	4.10(5.32)
100	9.38(2.74)	26.88(3.82)	14.98(3.98)	8.08(4.92)	15.70(5.81)	17.22(5.69)
	13.02(4.40)	40.96(8.38)	26.18(8.18)	9.20(6.03)	19.40(8.67)	24.38(10.30)
200	24.08(2.39)	36.48(2.38)	30.64(2.12)	27.64(3.56)	34.06(2.40)	36.26(2.35)
	28.12(3.85)	54.72(5.33)	51.42(7.33)	33.20(6.83)	43.16(6.79)	63.86(21.67)
Auto-regressive $\rho = 0.7$						
50	5.30(1.98)	11.68(2.87)	6.80(1.73)	6.74(2.15)	8.72(2.21)	8.62(1.66)
	6.84(2.69)	16.64(3.46)	8.52(2.72)	8.08(2.95)	10.90(3.69)	13.86(5.23)
100	10.76(1.55)	12.88(2.22)	9.44(1.16)	11.02(2.05)	13.38(2.60)	13.50(2.48)
	11.60(1.95)	18.32(3.14)	10.16(1.74)	13.38(4.12)	16.82(5.15)	24.34(6.88)
200	13.34(1.60)	16.80(3.16)	13.00(2.24)	16.10(2.26)	18.36(2.66)	18.48(2.60)
	14.04(2.19)	24.48(4.96)	14.24(3.04)	20.22(5.38)	23.88(7.24)	48.86(18.24)
Banded scenario 1						
50	2.30(1.57)	11.04(4.36)	1.96(1.98)	0.90(2.13)	3.00(4.09)	4.04(4.04)
	4.72(2.35)	19.04(10.46)	3.86(3.80)	1.12(2.50)	4.00(5.57)	6.18(7.09)
100	10.74(2.83)	24.08(4.52)	15.72(2.48)	10.14(4.44)	15.34(4.47)	16.40(4.09)
	13.62(3.78)	37.28(9.64)	27.34(5.80)	11.84(6.45)	19.80(8.84)	24.26(9.35)
200	23.28(2.30)	33.72(2.94)	27.92(2.41)	23.42(3.09)	31.68(4.30)	32.06(4.02)
	25.68(2.87)	49.28(4.81)	41.82(6.39)	28.96(6.54)	44.06(11.16)	49.96(11.63)
Banded scenario 2						
50	4.78(1.76)	11.32(3.37)	5.90(2.59)	5.70(2.49)	8.98(2.61)	8.74(3.26)
	7.12(2.54)	16.64(5.50)	8.96(4.06)	6.92(3.34)	10.94(3.65)	13.70(6.67)
100	10.82(2.16)	14.72(3.37)	10.50(1.62)	11.14(2.33)	13.50(2.92)	12.80(2.63)
	12.00(2.48)	21.04(4.84)	13.12(3.30)	13.70(3.89)	16.78(5.27)	22.26(7.06)

n^m	MCP	gMCP	cMCP	Sparse gMCP		
				$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$
200	15.18(1.85)	19.32(3.69)	14.08(2.02)	17.70(2.67)	20.56(2.98)	21.08(2.82)
	15.76(2.00)	28.00(5.30)	17.88(4.01)	23.00(5.65)	27.82(8.21)	37.46(13.25)

Table 2
Parameter estimates for the analysis of multiple cancer datasets using sparse gMCP.

UniGene	Gene name	Liver	Prostate	Lymphoma	Colon
Hs.105656	AIMP1	0.67	3.51	-1.93	2.25
Hs.111779	SPARC	1.48		8.60	
Hs.111903	FCGRT	-0.42			-1.09
Hs.153704	NEK2	3.03		2.83	-1.93
Hs.169907	GSTA4	2.03		-1.82	-1.36
Hs.270411	FERMT2	-2.50			-2.83
Hs.446504	FGFR4	6.72		-1.68	
Hs.74615	PDGFRA	-3.10			
Hs.10247	ALCAM		1.32	-4.60	
Hs.10842	RAN		0.92	2.20	4.28
Hs.169470	ASS1		11.50		
Hs.367437	ATM			2.36	
Hs.126608	NR3C1				8.42
Hs.388681	HDAC3				-2.94

Table 3

Parameter estimates for the analysis of multiple cancer datasets using cMCP.

UniGene	Gene name	Liver	Prostate	Lymphoma	Colon
Hs.111779	SPARC	1.49		4.69	
Hs.270411	FERMT2	-2.67			
Hs.272458	PPP3CA	-3.21			
Hs.2780	JUND	-1.54			
Hs.409934	HLA-DQB1	-2.49			
Hs.446504	FGFR4	3.76			
Hs.74615	PDGFRA	-3.43			
Hs.169470	ASS1		5.79		
Hs.58593	Transcribed locus		5.95		
Hs.367437	ATM			2.33	
Hs.385766	HNRNPM			-3.33	
Hs.436836	ZNF536			-2.29	
Hs.93207	NF1			1.40	
Hs.386078	MYLK				-4.68
Hs.407830	MYB				-3.21