Sparse Multivariate Factor Analysis Regression Models and Its Applications to Integrative Genomics Analysis

Yan Zhou

Merck & Co., PA 19454, USA

**email:* yan.zhou1@merck.com

and

Pei Wang

Icahn School of Medicine at Mount Sinai, New York, NY 10026, USA

**email:* pei.wang@mssm.edu

and

Xianlong Wang

Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

**email:* xianlong.wang@gmail.com

and

Ji Zhu, Peter X.-K. Song

University of Michigan, MI 48109, USA

**email:* jizhu@umich.edu, pxsong@umich.edu

Summary: The multivariate regression model is a useful tool to explore complex associations between two kinds of molecular markers, which enables the understanding of the biological pathways underlying disease etiology. For a set of correlated response variables, accounting for such dependency can increase statistical power. Motivated by integrative genomic data analyses, we propose a new methodology – sparse multivariate factor analysis regression model (smFARM), in which correlations of response variables are assumed to follow a factor analysis model with latent factors. This proposed method not only allows us to address the challenge that the number of association parameters is

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/gepi.22018

larger than the sample size, but also to adjust for unobserved genetic and/or non-genetic factors that potentially conceal the underlying response-predictor associations. The proposed smFARM is implemented by the EM algorithm and the blockwise coordinate descent algorithm. The proposed methodology is evaluated and compared to the existing methods through extensive simulation studies. Our results show that accounting for latent factors through the proposed smFARM can improve sensitivity of signal detection and accuracy of sparse association map estimation. We illustrate smFARM by two integrative genomics analysis examples, a breast cancer dataset and an ovarian cancer dataset, to assess the relationship between DNA copy numbers and gene expression arrays to understand genetic regulatory patterns relevant to the disease. We identify two trans-hub regions: one in cytoband 17q12 whose amplification influences the RNA expression levels of important breast cancer genes, and the other in cytoband 9q21.32-33, which is associated with chemoresistance in ovarian cancer.

Key words: EM-blockwise coordinate descent; High dimensional data; Latent factors; Regularization.

1. Introduction

Unveiling regulatory patterns between genetic variants and gene expressions is of great importance to a broad range of biological studies, in the hope to improve our understanding of complex disease pathogenesis. As reported in many recent genetic studies, high-throughput gene expression array experiments and genotype or DNA copy number array experiments are carried out on the same set of subjects. This provides the unique opportunity to assess regulatory relationships among DNAs and RNAs via an integrative genomic analysis. Copy number alterations (CNAs), including both germline variants and somatic copy number aberrations, are found to be largely associated with disease mechanisms in many studies; see for example, **?**. In particular, somatic aberrations are discovered to be important for tumorigenesis. For instance, oncogene activation by gene amplification or the loss of a tumor suppressor by gene deletion can cause transcriptional errors, which contributes to cancer pathogenesis (**?**). On the other hand, gene expression can be related to copy number alterations in proximal genes within a a window of several megabase pairs (cis-acting), as well as remote alterations throughout the genome (trans-acting). It has been regarded as a difficult task to detect genomewide cis- and trans-acting effects simultaneously due to the fact that numerous passenger genes amidst the limited set of drivers may contribute to tumor progression. Recent studies (**???**) have focused on the cis-acting effects of copy number on gene expressions and there are few studies that have considered trans-acting effects on a genomewide scale. To address these challenges require new analytic tools suitable for well-powered genomic studies.

The construction of genome-wide regulatory map by exploiting genomic and transcriptomic data typically involves in a large number of gene expressions as response variables and high-dimensional genetic variants (e.g. DNA copy number alterations) as predictors. This analytic task can be primarily formulated by a multivariate regression analysis (e.g. **??**). Usually,

the genetic regulatory relationships are intrinsically sparse, in the sense that one genetic variant may regulate only a small proportion of gene expressions, rather than the majority of them. It is also reported that some genetic variants, known as master regulators, play more important roles than other variants in the regulatory network, in terms of their ability of influencing many gene expressions simultaneously (**??**). Thus, it is of great interest to develop proper multivariate regression models that account for both the sparsity in the regulatory relationships and the existence of master regulators in the mapping of genetic associations. Towards this goal, sparse penalty functions such as LASSO (**?**), elastic net (**?**), and group LASSO (**?**) have been introduced to the multivariate regression framework (e.g. **?**, **?** and **?**). Readers can find more details about the comparison of our work with the existing methods in Section 5.

Some researchers have pointed out (e.g. **?** and **?**) that gene expressions are influenced by many biological and non-biological factors. Biological factors could include, for example, genotype polymorphisms/mutations, DNA copy number variations, DNA methylation, microRNA regulations, protein regulations and others. Non-biological factors include sample collection noise, instrumental errors, and batch effects. In addition, population admixtures or kinships in a study population may also influence data generation mechanism of gene expression profiles. Because of these complications, quite often only a small portion of variations in gene expressions can be explained by one type of genetic markers under investigation. Moreover, it is reported that gene expression heterogeneity is presented strongly in many studies but it is not yet properly taken into account in statistical analysis. For example, **?** and **?** have showed that gene expression heterogeneity not only leads to the reduction of statistical power but also produces spurious association signals when studying the regulatory relationships between genotypes and gene expressions. This motivates us to develop a new method that employs the factor analysis model to account for such heterogeneity attributed

to some unobserved genetic and/or non-genetic variabilities. As a result, we can improve both statistical power and accuracy of identifying significant associations between genes and genetic markers.

In this article, we plan to achieve three objectives via a sparse multivariate factor analysis regression model (smFARM): (i) to identify both trans-acting and cis-acting effects in one modeling framework; (ii) to regularize the association map by encouraging the selection of important predictors (or regulators); and (iii) to estimate the covariance matrix of the response variables via the means of multivariate factor analysis. The smFARM is specified in a similar spirit of the seemingly unrelated regression (SUR) model (**?**), which aims to improve the estimation efficiency of association in the detection of important signals by utilizing the residual correlations of gene expressions among genes. The factor analysis model enables us to understand and interpret additional association features beyond what expression-genetic variant associations describe. The mean model component of smFARM is parameterized by a matrix of regression coefficients that are supposed to contain many zeros because of sparse genetic regulatory relationships. This part of modeling relates closely to the remMap method proposed by **?** for the identification of genetic regulatory relationships and master predictors using a regularized multivariate regression model. Compared to remMap, our proposed smFARM further extends their model and is able to capture residual correlations of the responses using latent factors. As discussed earlier, when studying the regulatory relationships between gene expressions and DNA copy numbers, gene expression levels could be often confounded by unobserved genetic and/or non-genetic factors. Thus, incorporating latent factors in smFARM leads to a more efficient method to extract important features of the regulatory network than remMap. This advantage is shown in both the analysis of breast cancer data set and the analysis of ovarian cancer data set. As shown, smFARM

identifies several new novel regulatory relationships between gene expressions and copy number alternation intervals (CNAIs).

2. Model

2.1 *Multivariate regression model*

Multivariate regression model plays an important role in multivariate data analysis. Such model extends the classical one-dimensional regression model, which is widely used to deal with correlated response variables. Following the common notations in multivariate regression model, for subject i, we assume that the conditional distribution of a $Q \times 1$ random vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iQ})^T$ given P-element explanatory vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{iP})^T$ is a multivariate normal distribution. And its expectation is specified by the following linear equations:

$$E(\mathbf{y}_i|\mathbf{x}_i) = \mathbf{\Theta}\mathbf{x}_i, \ i = 1, \ldots, N, \tag{1}$$

where $\mathbf{\Theta} = \{\theta_{qp}\}$ is a $Q \times P$ matrix of unknown regression coefficients, and its covariance is $\mathrm{Var}(\mathbf{y}_i|\mathbf{x}_i) = \mathbf{\Sigma}$, which is an unknown $Q \times Q$ positive definite covariance matrix independent of $\mathbf{x}_i$. Obviously, if $Q = 1$, model (1) becomes the classical one-dimensional regression model, where $\mathbf{\Theta}$ is a P-dimensional regression coefficient vector. In matrix $\mathbf{\Theta}$, the q-th row represents the vector of regression coefficients corresponding to the q-th regression model, i.e. $E(y_{iq}|\mathbf{x}_i) = \sum_{p=1}^{P} \theta_{qp} x_{ip}$, which is a linear model of the q-th response variable y_{iq} on all P predictors. Clearly, the ordinary least square method (or equivalently the maximum likelihood method under the normally distributed errors) yields an estimator of $\mathbf{\Theta}$ as $\hat{\mathbf{\Theta}}^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. This implies that each row of $\mathbf{\Theta}$ can be estimated separately by regressing each of Q responses on the P predictors without accounting for any dependence across the Q responses. This is because in this estimation there are no common coefficients and/or common parameters in $\mathbf{\Sigma}$ shared across Q individual one-dimensional regression models. In contrast, when some common features are present in the mean models and/or

covariance matrices, borrowing data information across different margins will be beneficial to improve statistical power, and consequently, joint estimation involving all Q rows is the focus of methodology development in this paper.

2.2 *Factor analysis model*

In this paper, we propose to model the covariance $\mathbf{\Sigma}$ by the following factor analysis model:

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^T + \mathbf{\Psi}, \tag{2}$$

where $\mathbf{B}$ is a $Q \times K$ matrix of factor loadings pertinent to communalities for K ($\leqslant Q$) latent factors and $\mathbf{\Psi}$ is a $Q \times Q$ diagonal matrix of uniqueness. Clearly, the mean model (1) does not involve the K latent factors, while the covariance model (2) is determined by loadings $\mathbf{B}$ and uniqueness $\mathbf{\Psi}$. Factor analysis is one of the popular dimension reduction techniques that represents variations of correlated variables by a low number of latent factors. See for example, **?**, **?**, **?** and **?**, among others, in which the factor analysis model has been employed to deal with heterogeneity in functional gene expression profiles.

2.3 *Multivariate factor analysis regression model*

Combining models (1) and (2), with P predictors $\mathbf{x}_i$ and K unobserved latent factors $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})^T$, we propose the following multivariate factor analysis regression model (mFARM):

$$\mathbf{y}_i = \mathbf{\Theta}\mathbf{x}_i + \mathbf{B}\mathbf{z}_i + \boldsymbol{\epsilon}_i, \ i = 1, \ldots, N, \tag{3}$$

where $\mathbf{z}_i$'s are i.i.d. K-variate vectors of latent factors following multivariate normal distribution $\text{MVN}_K(\mathbf{0}, \mathbf{I})$, and $\boldsymbol{\epsilon}_i$'s are i.i.d. measurement errors with $\text{MVN}_Q(\mathbf{0}, \mathbf{\Psi})$ and are independent of the latent factors $z_{i1}, \ldots, z_{iK}$. In matrix notation, model (3) may be rewritten as follows:

$$\mathbf{Y} = \mathbf{X}\mathbf{\Theta}^T + \mathbf{Z}\mathbf{B}^T + \mathbf{E}, \tag{4}$$

where $\mathbf{Y}_{Q\times N}^T = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, $\mathbf{X}_{P\times N}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, $\mathbf{Z}_{K\times N}^T = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$ and $\mathbf{E}_{Q\times N}^T = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N)$. For simplicity, we assume that all Q responses and all P predictors are standardized to have zero mean and thus the intercept terms are removed from (4).

Our proposed mFARM model (4) will improve the capacity of statistical analysis for the construction of genetic regulatory maps with high-throughput array data, because it accounts for unobserved factors that better capture variabilities in the residuals.

3. Regularized Estimation

To achieve sparsity in the estimation of parameter matrix $\boldsymbol{\Theta}$, which characterizes the association map of interest, and to encourage the detection of master predictors (i.e. master regulators) in a similar spirit to the remMap method (**?**), we propose the following doubly penalized loss function:

$$\begin{aligned} L(\boldsymbol{\Theta}, \boldsymbol{\Psi}, \mathbf{B}) =& \frac{1}{2N}\sum_{i=1}^{N}(\mathbf{y}_i - \boldsymbol{\Theta}\mathbf{x}_i)^T(\mathbf{B}\mathbf{B}^T + \boldsymbol{\Psi})^{-1}(\mathbf{y}_i - \boldsymbol{\Theta}\mathbf{x}_i) \\ &+ \lambda_1 \sum_{q=1}^{Q}\sum_{p=1}^{P}|\theta_{qp}| + \lambda_2 \sum_{p=1}^{P}\sqrt{\theta_{1p}^2 + \cdots + \theta_{Qp}^2}, \end{aligned} \tag{5}$$

where λ_1 and λ_2 are two nonnegative tuning parameters. The first penalty term in (5) is the L_1 norm penalty that controls the overall sparsity in $\boldsymbol{\Theta}$ by tuning parameter λ_1, while the second penalty is the L_2 norm penalty that controls the column sparsity in $\boldsymbol{\Theta}$ via tuning parameter λ_2. The use of the two penalties facilitates the selection of important predictors, at both individual and group levels, that affect multiple responses simultaneously.

If there is some *a priori* knowledge about the known relationship between a predictor X_p and a response Y_q, such information may be incorporated into the estimation procedure via (5) in a similar way suggested in **?**. That is, consider a pre-specified $Q \times P$ matrix $\mathbf{C}^*$ whose

(q, p)-th element is given by:

$$C^*_{qp} = \begin{cases} 2, \text{ if } X_p \text{ is independent of } Y_q; \\ 0, \text{ if } X_p \text{ is associated with} Y_q; \\ 1, \text{ if there is no prior information.} \end{cases} \tag{6}$$

According to (5), given an unknown matrix $\mathbf{\Theta}^*$, the (q, p)-th entry θ^*_{qp} will be set as 0 in advance if $C^*_{qp} = 2$; otherwise, θ^*_{qp} will or will not be penalized by a flag value $C^*_{qp} = 1$ or $C^*_{qp} = 0$. After setting matrix $\mathbf{\Theta} = \mathbf{\Theta}^*$ according to $\mathbf{C}^*$, the modified objective function is given by

$$\begin{aligned} L(\mathbf{\Theta}, \mathbf{\Psi}, \mathbf{B}) = & \frac{1}{2N} \sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{\Theta}\mathbf{x}_i)^T (\mathbf{B}\mathbf{B}^T + \mathbf{\Psi})^{-1} (\mathbf{y}_i - \mathbf{\Theta}\mathbf{x}_i) \\ & + \lambda_1 \sum_{q=1}^{Q} \sum_{p=1}^{P} |C_{qp}\theta_{qp}| + \lambda_2 \sum_{p=1}^{P} \sqrt{C_{1_p}\theta^2_{1p} + \cdots + C_{Qp}\theta^2_{Qp}}, \end{aligned} \tag{7}$$

where a $Q \times P$ matrix $\mathbf{C} = \{C_{qp}\}$ is defined as $C_{qp} = \mathbf{1}\{C^*_{qp} = 1\}$.

Without loss of generality, we assume that both λ_1 and λ_2 are positive, and if one of them is zero, we can modify our methodology with little effort. Also, the proposed smFARM may be used to deal with the case of high-dimensional measurements with $min(P, Q) \gg N$, which is pervasive in biological studies, such as microarray data that contain thousands of biological markers measured from typically dozens to hundreds of subjects.

4. Algorithm

4.1 *EM-blockwise coordinate descent algorithm*

In this paper, we estimate three unknown parameter matrices, $(\mathbf{\Theta}, \mathbf{B}, \mathbf{\Psi})$, through minimizing the doubly penalized loss function (7), where $\mathbf{\Theta}$ and $(\mathbf{B}, \mathbf{\Psi})$ are involved in the mean model and the covariance model, respectively. A two-step iterative approach is used to estimate these three matrices. Given the current estimates of the factor model terms, $(\mathbf{B}^{(t)}, \mathbf{\Psi}^{(t)})$, updating the association matrix, $\mathbf{\Theta}^{(t+1)}$, is done by minimizing the doubly penalized loss function (7) by the blockwise coordinate descent algorithm proposed by**?**, while updating

the factor model terms $(\mathbf{B}^{(t+1)}, \mathbf{\Psi}^{(t+1)})$ is carried out through the EM algorithm after $\mathbf{\Theta}^{(t+1)}$ being given. Repeating these two-step procedures iteratively till algorithmic convergence, we obtain estimates $(\hat{\mathbf{\Theta}}, \hat{\mathbf{B}}, \hat{\mathbf{\Psi}})$ at the end of the algorithm operation. The computational complexity of the above algorithm may be assessed separately for the operation of the EM algorithm to estimate the loading coefficients $\mathbf{B}$ and the uniqueness $\mathbf{\Psi} = \sigma^2\mathbf{I}$, and the operation of blockwise coordinate descent algorithm to obtain sparse group lasso estimation for the association matrix $\mathbf{\Theta}$. The computational complexity of the former is in the order of $O(NQK)$ per iteration, and that of the latter is in the order of $O(NPQ)$. Refer to the Supplementary Material where actual computation times in simulation studies are reported.

4.2 *Tuning parameter selection*

We consider the selection of the tuning parameters (λ_1, λ_2) with a given $K = K_0$. Following **?**, we adopt the M-fold cross-validation method to choose the tuning parameters $(\lambda_1,\ \lambda_2)$. Since the true model is believed to be sparse, as suggested by **?** we utilize the ordinary least squares (OLS) estimates instead of the shrunken estimates to calculate the cross-validation score. This is because, when there are many potential poor predictors, the cross-validation score based on shrunken estimates often leads to severe false positive rates (**??**). In contrast, using the OLS estimates seems to make a reasonable remedy for such a problem, which is also observed in our simulation studies. It is worth pointing out that Bayesian information criterion (BIC), another popular tuning selection method, is not considered here, mainly because estimating the degrees of freedom required by the BIC is difficult under a nonorthogonal design matrix of predictors.

In this paper smFARM is run at a prespecified number of latent factors K. In practice, K may be estimated from the data, and there exists a large amount of the literature concerning consistent estimation of K, including the widely used AIC **?** and BIC **?**, as well as other methods proposed by **?**, **?**, and **?**, among others.

5. Simulation

5.1 *Simulation Setup*

We conduct two simulation experiments to assess the performance of the proposed model and optimization method. To specify simulation settings, we mimic a microarray dataset with $N = 200$ subjects, $Q = 400$ gene expressions and $P = 400$ variables of copy number alterations (CNAs). For each simulation, we consider a specific association map between genes and CNAs, which is specified as being sparse in groups. The graphic presentations of the association maps are given, respectively, in panels (a) and (b) of Figure 1. In simulation experiment I, we begin with a simple association map shown in Figure 1(a), in which 5 CNAs (i.e. black nodes) are set as master regulators (or hubs). These master CNAs are designed to so strong that they link to a total of 114 genes (i.e. circles), on average each CNA regulating 20 to 30 gene expressions. The total number of nonzero associations in this map is 125. Simulation experiment II concerns a more practical situation, where the topology of the given association map appears to be neither group dominated nor individual dominated. As shown in Figure 1(b), such association map includes 5 strong master CNAI regulators, each influencing 24 to 37 genes, 5 weak master CNAI regulators, each influencing 3 to 7 genes, and 20 CNAIs linking to only 1 or 2 genes. The total number of nonzero associations is 192.

[Figure 1 about here.]

In the first simulation experiment, P categorical CNAs $\mathbf{x} = (x_1, \ldots, x_P)^T$ are generated as predictors from $x_p \sim \text{Binomial}(2, 0.2) - 1$, with values -1, 0, or 1, representing copy number deletion, normal and amplification. In the second simulation study, continuous copy number alternation intervals (CNAIs) are generated to mimic the true predictor characteristics discussed in Section 6. Based on the real breast cancer data and ovarian cancer data, we find that there exits the heterogeneity within CNAIs, characterized by certain chromosome-specific structures, occurring in the forms of both within-chromosome and

between-chromosome differences. Here we assume that these P continuous CNAIs belong to 23 distinct chromosomes, where the number of CNAIs on the i-th chromosome (i.e. $P_i, i = 1, \ldots, 23$) is proportional to the size of that chromosome obtained from the real data. Within the i-th chromosome, any pair of CNAIs, say, CNAI_m and CNAI_n, is set to be positively correlated and such correlation decreases when their genetic distance increases according to $0.9^{|m-n|/2}$ for $m,\ n = 1, \ldots, P_i$. If two CNAIs come from different chromosomes, a much weaker correlation is randomly drawn from $\{0.25, 0.25^2, \ldots, 0.25^{23}\}$ together with a randomly generated positive or negative sign. Finally we compute the nearest positive definite symmetric matrix $\mathbf{\Xi}$ based on the above correlations using the algorithm in **?**, and P continuous CNAs are generated from $\mathbf{x} \sim \text{MVN}_P(\mathbf{0}, \mathbf{\Xi})$.

To specify the $Q \times P$ association map of $\mathbf{\Theta} = \{\theta_{qp}\}$, we first specify a sparse indicator matrix $\mathbf{\Delta} = \{\delta_{qp}\}$ which defines the connectivity in a genetic association mapping between Q genes and P CNAs. If $\delta_{qp} = 1$, we generate θ_{qp} from $\text{Unif}([-5, -1] \bigcup [1, 5])$; otherwise, $\theta_{qp} = 0$. To specify the $Q \times K$ factor loadings matrix $\mathbf{B}$, we start with an initial matrix $\mathbf{B}^* = \{b^*_{qk}\}$, with $b^*_{qk} \overset{i.i.d.}{\sim} \text{Unif}([0, \tau])$ and τ is a given positive constant. Then, we specify a matrix $\mathbf{B}$ as of the form $\mathbf{B} = \mathbf{U}\mathbf{V}^{\frac{1}{2}}$, where $\mathbf{V}$ is a diagonal matrix with diagonal entries being the eigenvalues of $\mathbf{B}^*\mathbf{B}^{*T}$, and the column vectors of $\mathbf{U}$ are the orthonormal eigenvectors of $\mathbf{B}^*\mathbf{B}^{*T}$. In other words, matrix $\mathbf{B}$ is specified by an orthogonal rotation of the initial matrix $\mathbf{B}^*$. Note that the factor loadings have an "indeterminacy" problem, which means both $\mathbf{B}$ and $\mathbf{BT}$ give rise to the same covariance matrix $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^T + \mathbf{\Psi}$, where $\mathbf{T}$ is an arbitrary orthogonal matrix. To ensure a unique solution, we impose a constraint on $\mathbf{B}$, according to **?**, to enforce that $\mathbf{B}^T\mathbf{B}$ is a diagonal matrix , which is accounted for in our procedure of generating the values of factor loadings for matrix $\mathbf{B}$. Given $\mathbf{\Theta}$ and $\mathbf{B}$, for each subject, we generate K latent factors $\mathbf{z} = (z_1, \ldots, z_K)^T$ by $z_k \sim \text{Normal}(0, 1)$ and Q measurement errors $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_Q)^T \sim \text{MVN}_Q(\mathbf{0}, \mathbf{\Psi})$, where the uniqueness $\mathbf{\Psi}$ is set as $\mathbf{\Psi} = \sigma^2\mathbf{I}_Q$ in the

simulation studies. Recall that τ and σ^2 are two variance parameters that control the size of communality and that of uniqueness, respectively. The choice of τ and σ^2 is based on a pre-specified scale of signal-to-noise ratio, according to SNR_1 of regression mean effects and SNR_2 of latent factor's effects; they are, $\text{SNR}_1 = \text{avg}\left[\frac{\text{diag}(\text{Cov}(\mathbf{\Theta}x))}{\text{diag}(\text{Cov}(\epsilon))}\right]$ and $\text{SNR}_2 = \text{avg}\left[\frac{\text{diag}(\text{Cov}(\mathbf{B}z))}{\text{diag}(\text{Cov}(\varepsilon))}\right]$, respectively. Finally, Q gene expressions $\mathbf{y} = (y_1, \ldots, y_Q)^T$ are generated from model (3) by $\mathbf{y}|\mathbf{x}, \mathbf{z} \sim \text{MVN}_Q(\mathbf{\Theta x} + \mathbf{Bz}, \mathbf{\Psi})$. Hereafter, a dataset of N i.i.d. $(\mathbf{y}, \mathbf{x})$ pairs is generated for each simulation round.

For convenience, the response variables and predictors are all centered to have mean zero, and the prior knowledge matrix $\mathbf{C} = \{C_{qp}\}$ is set as all entries being 1; in this case, all predictors are subject to shrinkage. Our primary evaluation criterion is the total number of false discoveries, TF = FP + FN, where FP and FN are the respective numbers of false positives and false negatives. Here, a "positive" (or a "negative") refers to a nonzero (or a zero) entry of $\mathbf{\Theta}$. Following **?**, additional criteria used in the evaluation include sensitivity (Sen), and Matthews correlation coefficient (MCC) score defined respectively, by $\text{Sen} = \text{TP}/(\text{TP} + \text{FN})$, and $\text{MCC} = \frac{(\text{TP}\times\text{TN}-\text{FP}\times\text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$.

To assess the performance of our smFARM, we mainly compare it with remMap ($K = 0$) by varying SNR_1, SNR_2 and K. It is worth noting that Peng et al.'s (**?**) remMap approach, which is established for the classic multivariate regression models (i.e. $K_{\text{true}} = 0$), has been compared with two popular existing methods, single lasso penalty (i.e. $\lambda_2 = 0$) and Q separate individual lasso regressions, and its superiority has been showed in **?**. So the comparisons to the latter two methods are not reported in our comparison. Here we set the true number of latent factors as $K_{\text{true}} = 2$, and focus on comparing three scenarios with $K = 0$ (i.e. remMap), $K = K_{\text{true}}$ (i.e. 2), and $K = 3$. The tuning parameters (λ_1, λ_2) are determined through 5-fold cross validation. And a total of 50 independently replicated datasets is used

in the evaluation of our method. Results of method comparisons are summarized in Table 1. Additional simulation results may be found in the Supplementary Material.

5.2 *Findings from Simulation Studies*

The results given in Table 1 concern simulation studies I and II. These results show that the proposed smFARM performs very well in all key aspects of regulator detection and group selection. Let us first focus on simulation study I, including two cases I.1 and I.2, with the corresponding numerical results being reported in the top part of Table 1. In Simulation I.1, when the true model contains no latent factors, subject to rounding errors, the proposed smFARM and the existing remMap perform equally well in terms of MCC. With no surprise, we find that, in both smFARM and remMap, larger SNR_1 leads to better performance in terms of lower TF, higher sensitivity and higher MCC in the comparison between SNR=1:0:3 and SNR=1:0:5. This outperformance of the smFARM repeats in the comparison between SNR=1:1:3 and SNR=1:1:5 with $K_{\text{true}} = 2$ in Simulation I.2. When the ratio of SNR_1 to SNR_2 is fixed at 1:1, smaller variation in the measurement errors (i.e. larger SNR_1) will lead to better performances. Moreover, an encouraging finding in Simulation I.2 is that, comparing our method accounting for the latent factors to the remMap that ignores latent factors, the smFARM approach is clearly more effective to identify true signals than the remMap when the data are from a multivariate model with correlated residuals or $K_{\text{true}} \neq 0$. With fixed SNR_1, in a comparison of $(\text{SNR}, K_{\text{true}}) = (1{:}0{:}3, 0)$ in Simulation I.1 with $(\text{SNR}, K_{\text{true}}) = (1{:}1{:}3, 2)$ in Simulation I.2, or in another comparison of $(\text{SNR}, K_{\text{true}}) = (1{:}0{:}5, 0)$ in Simulation I.1 with $(\text{SNR}, K_{\text{true}}) = (1{:}1{:}5, 2)$ in Simulation I.2, very similar findings are obtained from the smFARM that accounts for latent factors. We also find that SNR_2 has a strong influence on the reconstruction of the association map, when the dependency of latent factors is ignored in the analysis.

It is interesting to note that results of group selection in simulation study I are rather

stable and accurate across the four cases in the top part of Table 1. This is probably because identifying clusters in these settings is not hard due to group-dominant topology designed in the association maps (see Figure 1(a)). In other words, relative to the L_1-penalty, the L_2-penalty is more effective to remove irrelevant groups or clusters.

In addition, all the above conclusions have repeated consistently in the more realistic simulation study II with continuous predictors. To examine the robustness of the proposed method, we simulated 50 replicates under the Simulation II setup from a model $\mathbf{y}_i = \mathbf{\Theta}\mathbf{x}_i + \mathbf{u}_i,\ i = 1, \ldots, N$, where the errors $\mathbf{u}_i$ are drawn directly from a multivariate normal distribution $\text{MVN}_Q(\mathbf{0}, \mathbf{B}\mathbf{B}^T + \mathbf{\Psi})$ with a certain non-diagonal covariance matrix used in the data simulation. In this case, we again found that the proposed smFARM model with $K = 2$ performed better in identifying the true signals than the remMap (or smFARM model with $K = 0$). The detail of this simulation is included in the Supplementary Material. To sum up, our proposed method has demonstrated clearly as being a very effective tool to achieve desirable statistical power by accounting for latent factors in the regulatory map reconstruction with high-dimensional complex data.

[Table 1 about here.]

6. Application

In this section we apply the proposed smFARM to analyze TCGA (The Cancer Genome Atlas) breast and ovarian cancer data sets. We are interested in detecting DNA copy number alterations (CNA) that have large impact on transcript activities (i.e. trans-regulate many RNA expressions). Such trans-hub CNAs often play important roles in tumor initiation and progression. Information on the regulatory pattern between these trans-hub CNAs and their downstream genes deems to shed important light on disease etiology.

6.1 *Data preparation*

Level-three RNAseq data and level-three segmented DNA copy number data of breast and ovarian cancer tumor samples were obtained from the TCGA website. We focus on subsets of samples (77 breast tumors and 71 ovarian tumors), which are also subjected to deep protein-profiling by CPTAC (Clinical Proteomic Tumor Analysis Consortium). Thus findings from our analysis may lead to a further investigation and knowledge generation through the corresponding protein profiles in the future.

We preprocess the breast and ovarian cancer data separately. For breast cancer data, based on level-three segmented DNA copy number profiles, we first break the genome using the union of the break-points detected in all tumor samples and filter the small regions with less than 10k base pairs. This result in 17482 regions. Then for each region of each sample, we record its copy number based on the inferred DNA copy number of the corresponding segment in the sample, with tail values truncated at ± 1.5. Due to the high spatial correlation in DNA copy number profiles, we further condense these 17482 regions into 1730 copy number alteration intervals (CNAI) by applying the fixed order clustering (FOC) (**?**), so that DNAs in the same interval tend to have similar CNA patterns in one sample. The copy number of one CNAI in a given sample is then calculated as the mean of the copy number of all regions within the interval in that sample. We exclude CNAI with no variation across the 77 samples, which results in 1571 CNAIs. For RNAseq data, we first set zeros to be missing values and take log transformation. We then standardize each sample to have median 0 and MAD (median absolute deviance) 1. We exclude genes with more than 10% missing, and select the top 15% genes with largest interquartile ranges across samples. The resulting data matrix consists of 1466 gene expressions.

We preprocess the ovarian cancer data set in the same manner as described above. Specifically, we derive 1617 CNAIs by applying FOC on merged level-three segmented DNA copy

number profiles. By further eliminating CNAIs with little variation, we end up with 1300 CNAIs that are actually used in the analyses in this paper. For RNAseq data, we select 2437 genes after applying the same normalization and filtering criteria as those applied in the breast cancer data above.

6.2 *smFARM analysis*

We apply smFARM to analyze the preprocessed breast cancer data and ovarian cancer data, separately. Our primary goal is to construct the regulatory map between copy number alterations and RNA expressions in each cancer dataset, adjusting for potential latent factors. Specifically, for each cancer type, we fit the following model:

$$\mathbf{Y}_{\text{RNA}} = \mathbf{X}_{\text{CNAI}}\mathbf{\Theta}^T + \mathbf{Z}\mathbf{B}^T + \mathbf{E}, \tag{8}$$

where $\mathbf{Y}_{\text{RNA}}$ is the RNA expression matrix, $\mathbf{X}_{\text{CNAI}}$ is the CNAIs copy number matrix, $\mathbf{\Theta}$ is the regression coefficient matrix with respect to CNAIs. In the above model, Q responses ($\mathbf{Y}_{\text{RNA}}$) and P predictors ($\mathbf{X}_{\text{CNAI}}$) are all standardized to have mean 0 and standard deviation 1. Note that $Q = 1466$, $P = 1571$ in the breast cancer data, while $Q = 2437$, $P = 1300$ in the ovarian cancer data. The estimated latent factors ($\mathbf{B}$) help to account for additional genetic and/or non-genetic features beyond the observed CNAI genetic markers, $\mathbf{X}_{\text{CNAI}}$.

In addition, we classify a CNAI×RNA pair to be a *cis* pair, if the RNA gene falls in the genome region of the CNAI; or otherwise the pair is referred to as a *trans* pair. There are in total 1172 cis pairs in the breast and 1862 cis pairs in the ovarian cancer data set, respectively. Since we are particularly interested in identifying trans-hub CNAIs, we do not impose shrinkage on the coefficients of these cis pairs. As pointed above, this choice can be managed by setting $C_{qp} = 0$ given that the p-th CNAI and the q-th gene form a cis pair; and $C_{qp} = 1$, otherwise in equation (7). We apply the proposed model fitting procedure and select the tuning parameters (λ_1, λ_2) using 10-fold cross validation on a 25×25 grid. We

vary the number of latent factors K from 0 to 20, and explore how the regulatory map varies accordingly as K increases.

6.3 *Results*

Some interesting trans-hub CNAIs are revealed by the application of smFARM for both the breast cancer and the ovarian cancer.

Figure 2 shows that with an increase in the number of latent factors, the detected number of tran edges decreases. When fully ignoring latent factors in the analysis, we detect a total of 2429 trans edges from the breast cancer data and a total of 318 trans edges from the ovarian cancer data. However, most of these detected edges are deemed false positive and are not biologically meaningful. Note that in either the breast cancer dataset or the ovarian dataset only about 70 subjects are measured, each being observed with thousands of genes and CNAIs. Indeed, both give rise to an ultra high-dimensional estimation problem, for which it is not easy to select the optimal number of latent factors. In this analysis, we choose $K = 2$, because this choice leads to the association maps that achieve a desirable balance between sparsity and discovery of important biological signals.

[Figure 2 about here.]

For the breast cancer data, at $K = 2$, the proposed smFARM detected 190 trans-regulation edges between 10 CNAIs and 134 transcripts. The detailed CNAI-RNA regulatory map is illustrated in Figure 3. The biggest trans-hub CNAIs are all from chromosome arm 5q. Deletions on chromosome arm 5q are key characteristics for basal-like breast cancer. Our findings that the DNA copy number alterations in 5q have big impact on a large number of transcripts is consistent with previous observations in the literature (**?**). Besides the tran-hub CNAIs on 5q, another major trans-hub is from 17q12. This CNAI is known as the harbor of the famous oncogene ERBB2, whose amplification is a trigger event for HER2 subtype of breast cancer (**?**). In addition to ERBB2, the 17q12 amplicon also harbors many

other important cancer genes and transcript factors (**?**), thus it is expected that this region serves as a tran-hub in the CNAI-RNA regulatory map. Among the transcripts regulated by these major trans-hub CNAIs, one transcript, TNFSF10, is regulated by all CNAIs in 17q12, 5q34, and 5q35.3. TNFSF10 is a member of the tumor necrosis factor superfamily. It has been shown to mediate p53-dependent cell death (**?**) and can be used as therapeutic targets to improve the treatment of triple-negative breast cancer patients (**?**). Our analysis suggests that the DNA copy number alterations in ERBB2 amplicon and 5q34-35.3 region could act as upper-stream regulator for TNFSF10 during tumor initiation and progression. These intriguing results help to cast light on the regulatory mechanism of these important disease genes.

On the other hand, the analysis of the ovarian cancer data reveals a different set of CNAIs trans-hubs, suggesting these two types of cancers are driven by distinct tumor mechanisms. Specifically, we find that the CNAI-RNA regulatory map consists 77 trans-regulation edges between 5 CNAIs and 77 transcripts. The CNAI with the largest number of trans-edges locates in 9q21.32-33. Copy number gain in this region is reported to be associated with chemoresistance in ovarian cancer patients (**?**). The transcripts regulated by this CNAI include two known cancer genes, GREB1 and NODAL. Gene GREB1 regulated by estrogen in breast cancer 1 was first identified as a hormone-responsive gene in the breast cancer cell line. Recently, this gene has also been found to be up-regulated by E2 (Exogenous 17β-estradiol) in ovarian tumors, and thus could serve as a novel gene target for therapeutic intervention (**?**). Gene NODAL encodes a protein belonging to the TGF-beta superfamily, which is an important regulator of embryonic stem cell and possibly cancer stem cells (**?**). The signaling of NODAL promotes a tumorigenic phenotype in human breast cancer through activating MAPK signaling pathway and could serve as a promising target for treating triple-negative breast cancer (**?**). Our analyses suggest potential regulatory relationships among

these known cancer related alterations and genes in the current literature. Such findings could lead to useful biological hypotheses to be tested in future studies.

[Figure 3 about here.]

7. Discussion

We developed a new methodology, sparse multivariate factor analysis regression model, to reconstruct a sparse genetic association map. The proposed smFARM extended the classic multivariate regression model, allowing a low-dimensional set of latent factors to account for the dependence among response variables instead of assuming residuals being independent noise. We developed an effective and flexible EM-blockwise coordinate descent algorithm to obtain regularized estimation and variable selection in the smFARM.

We have shown that by accounting for latent factors, the proposed smFARM can effectively identify response-predictor associations from high dimensional data with improved sensitivity and accuracy. The numerical results have indicated that the proposed smFARM works well to derive the underlying sparse association relationship. Furthermore, both real breast cancer data and ovarian cancer examples have also shown that our proposed smFARM provides richer and biologically relevant discoveries to facilitate transcriptomic analyses. The sparse genetic association map between CNAIs and gene expressions helped us understand and interpret genetic regulation mechanisms and generate useful biological hypotheses on those detected signals given in this paper.

To our knowledge, there are some other methods that can characterize the variability in the gene expressions such as singular value decomposition (SVD) or principle component analysis (PCA). There is a direct relationship between PCA and SVD in the case where principal components are calculated from the covariance (**?**). Furthermore, the essential difference between SVD/PCA and factor analysis lies whether or not a covariance model is used for

the residuals. Refer to **?**, **?** and **?** for more details. We find that unlike PCA/SVD using superficial labeling such as "eigengenes", "supergenes", or "meta-genes" without clear biological entity (**?**), the number of latent factors can provide a biologically relevant parameter in the reconstruction of association map, which is appealing in practice.

Besides the gene-CNA association analysis illustrated in this paper, our proposed method may be applied in a broad range of problems. For instance, it may be applied to systematically explore the relationship between gene expression levels and genotypes as to, for example, whether a gene is differentially expressed with different genotypes (or alleles) at a specific locus. The loci that are associated with gene expression levels are known as expression quantitative loci (eQTL). For a given gene, an eQTL data analysis aims to identify genetic loci or single nucleotide polymorphisms (SNPs) that are linked or associated with expression levels of a common gene. Moreover, in eQTL analysis, SNPs may be naturally grouped according to their functionality or biological pathways based on some prior knowledge. When we are interested in associations of multiple SNPs simultaneously within a biological pathway, incorporating genetic or non-genetic latent factors would help us to achieve a more powerful and richer analysis, leading to better understanding of the underlying biological mechanisms.

Acknowledgements

The authors are grateful to two annoymous reviewers for their constructive comments that led to an improvement of this paper. Pei Wang and Xianlong Wang were partially supported by National Institutes of Health grant U24CA160034. Pei Wang was also supported by National Institutes of Health grants R01GM082802, R01GM108711, and P01CA53996. Peter Song was partially supported by National Science Foundation grant DMS1513595, and by National Institutes of Health grants R01ES024732, and NIH P01ES022844.

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81,** 1203–1227.

Akaike, H. (1992). *Information theory and an extension of the maximum likelihood principle*, pages 610–624. Springer.

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America* **97,** 10101–10106.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. *In Proceedings of the third Berkeley symposium on mathematical statistics and probability* **5,** 111–150.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70,** 191–221.

Bedrick, E. J. and Tsai, C. L. (1994). Model selection for multivariate regression in small samples. *Biometrics* **50,** 226–231.

Bergamaschi, A., Kim, Y. H., Wang, P., Sørlie, T., Hernandez-Boussard, T., Lonning, P. E., Tibshirani, R., Børresen-Dale, A.-L., and Pollack, J. R. (2006). Distinct patterns of dna copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes, Chromosomes and Cancer* **45,** 1033–1040.

Blum, Y., Le Mignon, G., Lagarrigue, S., and Causeur, D. (2010). A factor model to analyze heterogeneity in gene expression. *Bmc Bioinformatics* **11,**.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–352.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32,** 407–451.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The annals of applied statistics* **3,** 521.

Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association* **104,** 1406–1415.

Gardner, T. S., di Bernardo, D., Lorenz, D., and Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301,** 102–105.

Gibson, G. (2008). The environmental contribution to gene expression profiles. *Nature Reviews Genetics* **9,** 575–581.

Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and Its Applications* **103,** 103–118.

Horlings, H. M., Lai, C., Nuyten, D. S. A., Halfwerk, H., Kristel, P., van Beers, E., Joosse, S. A., Klijn, C., Nederlof, P. M., Reinders, M. J. T., Wessels, L. F. A., and van de Vijver, M. J. (2010). Integration of dna copy number alterations and prognostic gene expression signatures in breast cancer patients. *Clinical Cancer Research* **16,** 651–663.

Hunter, D., Edson, L., and Coleman, W. (2014). Loss of tumor necrosis factor superfamily genes in breast cancer cell lines (1047.8). *The FASEB Journal* **28,** 1047–8.

Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411,** 41–42.

Kirsammer, G., Strizzi, L., Margaryan, N. V., Gilgur, A., Hyser, M., Atkinson, J., Kirschmann, D. A., Seftor, E. A., and Hendrix, M. J. (2014). Nodal signaling promotes a tumorigenic phenotype in human breast cancer. In *Seminars in cancer biology*, volume 29, pages 40–50. Elsevier.

Kuribayashi, K., Krigsfeld, G., Wang, W., Xu, J., Mayes, P. A., Dicker, D. T., Wu, G. S., and El-Deiry, W. S.

(2008). Tnfsf10 (trail), a p53 target gene that mediates p53-dependent cell death. *Cancer biology & therapy* **7,** 2034–2038.

Kustra, R., Shioda, R., and Zhu, M. (2006). A factor analysis model for functional genomics. *Bmc Bioinformatics* **7,**.

Lahti, L., Schafer, M., Klein, H. U., Bicciato, S., and Dugas, M. (2013). Cancer gene prioritization by integrative analysis of mrna expression and dna copy number data: a comparative review. *Briefings in Bioinformatics* **14,** 27–35.

Lamy, P.-J., Fina, F., Bascoul-Mollevi, C., Laberenne, A.-C., Martin, P.-M., Ouafik, L., and Jacot, W. (2011). Quantification and clinical relevance of gene amplification at chromosome 17q12-q21 in human epidermal growth factor receptor 2-amplified breast cancers. *Breast Cancer Res* **13,** R15.

Laviolette, L. A., Hodgkinson, K. M., Minhas, N., Perez-Iratxeta, C., and Vanderhyden, B. C. (2014). 17β-estradiol upregulates greb1 and accelerates ovarian tumor progression in vivo. *International Journal of Cancer* **135,** 1072–1084.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *Plos Genetics* **3,** 1724–1735.

Lonardo, E., Hermann, P. C., Mueller, M.-T., Huber, S., Balic, A., Miranda-Lorenzo, I., Zagorac, S., Alcala, S., Rodriguez-Arabaolaza, I., Ramirez, J. C., et al. (2011). Nodal/activin signaling drives self-renewal and tumorigenicity of pancreatic cancer stem cells and provides a target for combined drug therapy. *Cell stem cell* **9,** 433–446.

Lutz, R. W. and Buhlmann, P. (2006). Boosting for high-multivariate responses in high-dimensional linear regression. *Statistica Sinica* **16,** 471–494.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica* **77,** 1447–1479.

Österberg, L., Levan, K., Partheen, K., Delle, U., Olsson, B., Sundfeldt, K., and Horvath, G. (2010). Specific copy number alterations associated with docetaxel/carboplatin response in ovarian carcinomas. *Anticancer research* **30,** 4451–4458.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R., and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics* **4,** 53–77.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999). Genome-wide analysis of dna copy-number changes using cdna microarrays. *Nature Genetics* **23,** 41–46.

Pollack, J. R., Sorlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Borresen-Dale, A. L., and Brown, P. O. (2002). Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 12963–12968.

Schneeweiss, H. and Mathes, H. (1995). Factor-analysis and principal components. *Journal of Multivariate Analysis* **55,** 105–124.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics* **6,** 461–464.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22,** 231–245.

Stegle, O., Kannan, A., Durbin, R., and Winn, J. (2008). *Accounting for non-genetic factors improves the power of eQTL studies*, volume 4955 of *Lecture Notes in Bioinformatics*, pages 411–422.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series*

B (Methodological) pages 267–288.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **61,** 611–622.

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005). Simultaneous variable selection. *Technometrics* **47,** 349–363.

Van Wieringen, W. N. and Van De Wiel, M. A. (2011). Exploratory factor analysis of pathway copy number data with an application towards the integration with gene expression data. *Journal of Computational Biology* **18,** 729–741.

Wall, M., Rechtsteiner, A., and Rocha, L. (2003). Singular value decomposition and principal component analysis. *A practical approach to microarray data analysis* pages 91–109.

Wang, P. (2010). *Statistical Methods for CGH Array Analysis*. VDM Verlag Dr.MIIer.

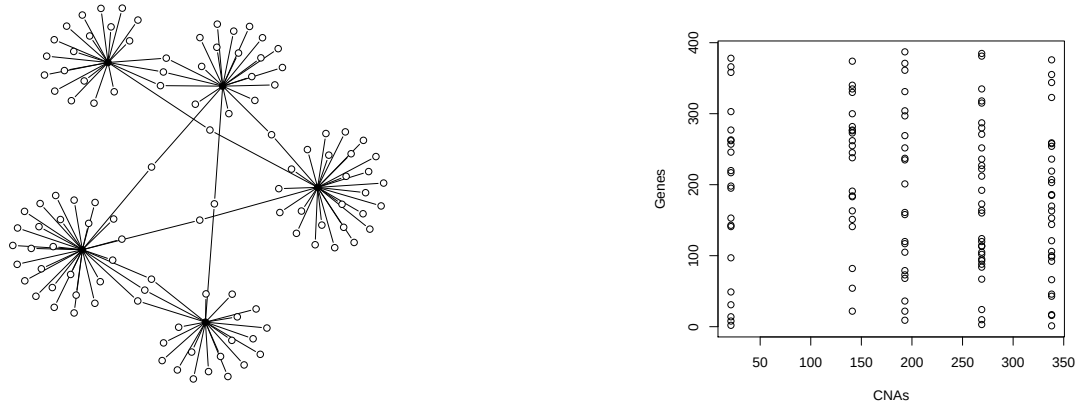
Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **68,** 49–67.

Yuan, Y. Y., Curtis, C., Caldas, C., and Markowetz, F. (2012). A sparse regulatory network of copy-number driven gene expression reveals putative breast cancer oncogenes. *Ieee-Acm Transactions on Computational Biology and Bioinformatics* **9,** 947–954.

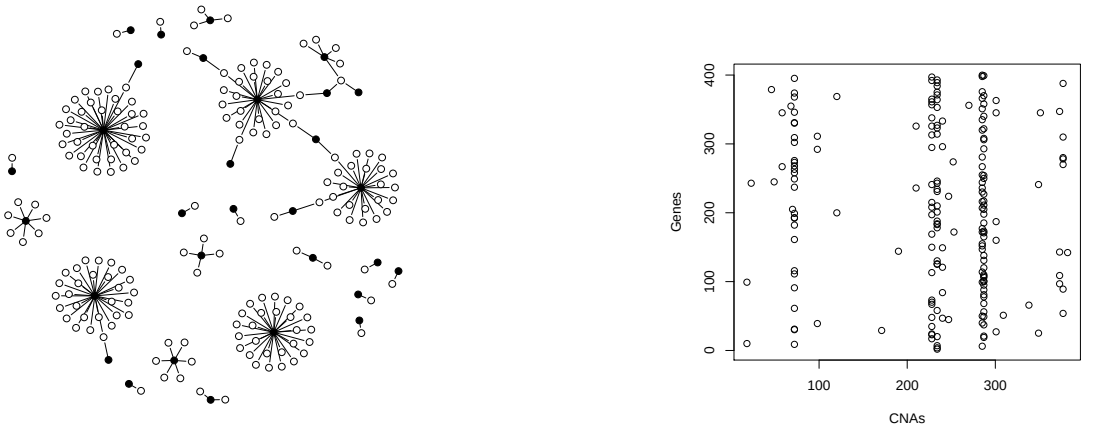
Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association* **57,** 348–368.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **67,** 301–320.

Figure 1: True association maps of $\mathbf{\Theta}$ (connectivity vs. heatmap) for Simulation I, and II. (**LHS**: connectivity maps of $\mathbf{\Theta}$ between genes (white) and biomarkers (black); **RHS**: corresponding heatmap of $\mathbf{\Theta}$.)



(a) Simulation I



(b) Simulation II

Figure 2: The number of detected trans edges under different K.

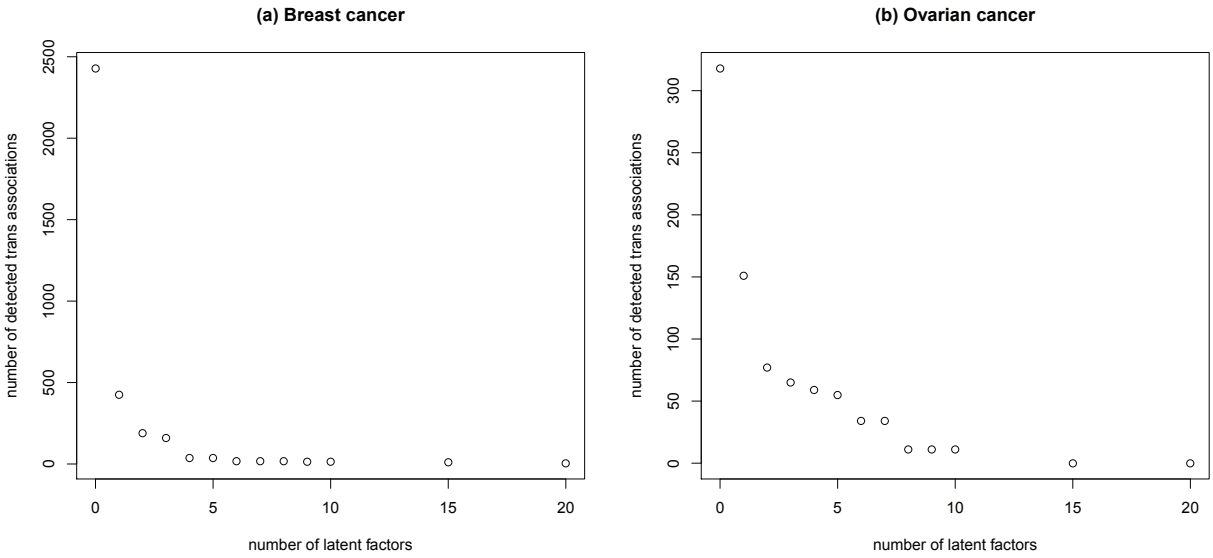


Figure 3: Detected association maps under $K = 2$. (**Top**: Breast cancer; **Bottom**: Ovarian cancer.

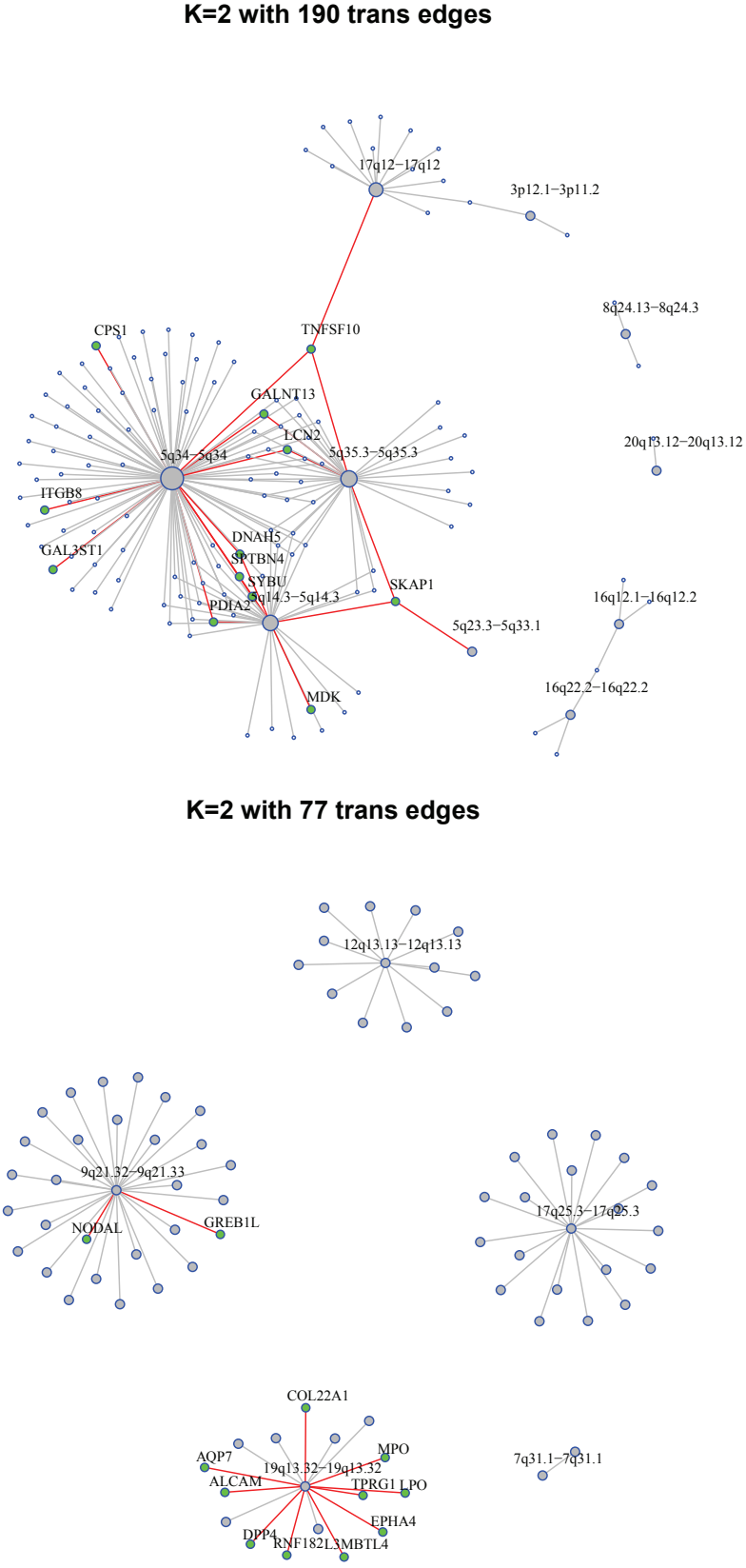


Table 1: Impact of different number of latent factors K and different SNR levels on Regulator Selection and Group Selection

			Regulator Selection			Group Selection		
SNR	K_{true}	Method	TF	Sen	MCC	TF	Sen	MCC
Simulation I.1								
1:0:3	0	smFARM$_{K=0}$	18.90(6.02)	0.89(0.04)	0.92(0.02)	0.06(0.24)	1(0)	0.99(0.02)
		remMap	21.88(6.61)	0.93(0.02)	0.92(0.02)	0.02(0.14)	1(0)	1(0.01)
1:0:5	0	smFARM$_{K=0}$	27.24(3.51)	0.81(0.03)	0.88(0.01)	0(0)	1(0)	1(0)
		remMap	34.10(5.17)	0.88(0.03)	0.87(0.02)	0(0)	1(0)	1(0)
Simulation I.2								
1:1:3	2	smFARM$_{K=2}$	18.24(3.46)	0.87(0.03)	0.92(0.01)	0(0)	1(0)	1(0)
		remMap	25.68(11.32)	0.83(0.04)	0.89(0.04)	0.02(0.14)	1(0)	1(0.01)
1:1:5	2	smFARM$_{K=2}$	28.51(4.26)	0.80(0.03)	0.88(0.02)	0(0)	1(0)	1(0)
		remMap	33.40(4.92)	0.76(0.04)	0.86(0.02)	0(0)	1(0)	1(0)
Simulation II								
1:3:5	2	smFARM$_{K=2}$	48.89(11.54)	0.82(0.05)	0.87(0.03)	10.89(2.53)	0.66(0.06)	0.79(0.05)
		smFARM$_{K=0}$	79.80(16.76)	0.77(0.02)	0.79(0.04)	12.10(1.25)	0.62(0.04)	0.76(0.03)
		remMap	87.46(20.67)	0.79(0.03)	0.77(0.05)	12.46(1.35)	0.62(0.05)	0.75(0.03)

Note: For each Total False (TF), Sensitivity (Sen), or Matthews correlation coefficient (MCC) measurement, we report mean values together with their standard errors on 50 replicates. smFARM$_{K=K_0}$ represents fitting the smFARM on a given number of latent factors K_0.