HHS Public Access

Author manuscript

J Multivar Anal. Author manuscript; available in PMC 2019 November 01.

Published in final edited form as:

J Multivar Anal. 2018 November ; 168: 119–130. doi:10.1016/j.jmva.2018.06.009.

Robust network-based analysis of the associations between (epi)genetic measurements

Cen Wu[#a], **Qingzhao Zhang**[#b], **Yu Jiang**[c], and **Shuangge Ma**[d,*]

[a]Department of Statistics, Kansas State University, Manhattan, KS, 66506, USA

[b]School of Economics and the Wang Yanan Institute for Studies in Economics, Xiamen University

[c]Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN, 38111, USA

[d]Department of Biostatistics, Yale University, New Haven, CT, 06510, USA

[#] These authors contributed equally to this work.

Abstract

With its important biological implications, modeling the associations of gene expression (GE) and copy number variation (CNV) has been extensively conducted. Such analysis is challenging because of the high data dimensionality, lack of knowledge regulating CNVs for a specific GE, different behaviors of the *cis*-acting and *trans*-acting CNVs, possible long-tailed distributions and contamination of GE measurements, and correlations between CNVs. The existing methods fail to address one or more of these challenges. In this study, a new method is developed to model more effectively the GE-CNV associations. Specifically, for each GE, a partially linear model, with a nonlinear *cis*-acting CNV effect, is assumed. A robust loss function is adopted to accommodate long-tailed distributions and data contamination. We adopt penalization to accommodate the high dimensionality and identify relevant CNVs. A network structure is introduced to accommodate the correlations among CNVs. The proposed method comprehensively accommodates multiple challenging characteristics of GE-CNV modeling and effectively overcomes the limitations of existing methods. We develop an effective computational algorithm and rigorously establish the consistency properties. Simulation shows the superiority of the proposed method over alternatives. The TCGA (The Cancer Genome Atlas) data on the PCD (programmed cell death) pathway are analyzed, and the proposed method has improved prediction and stability and biologically plausible findings.

Keywords

Copy number variation; Gene expression; Network structure; Partially linear model; Penalization; Robust estimation

[*]Corresponding author shuangge.ma@yale.edu (Shuangge Ma).

1. Introduction

For complex diseases, profiling studies have been extensively conducted, collecting data on multiple types of omics measurements. Different types of omics measurements are interconnected, for example with molecular changes at the DNA/epigenetic level and microRNAs regulating gene expressions. Studying the associations between different types of omics measurements can lead to a better understanding of disease biology and clinically useful models. In this study, we analyze the association between GE (gene expression) and CNV (copy number variation), which have one of the first known regulation relationships and have attracted extensive attention. The proposed approach is also potentially applicable to other types of omics measurements and other high-dimensional problems.

Modeling the GE-CNV association is challenging. Both GE and CNV measurements are high-dimensional. The expression level of a specific gene can be affected by both the *cis*-acting and *trans*-acting CNVs [6], with the set of relevant CNVs usually unknown. Some studies, such as [17], are limited by conducting the "one GE versus one CNV" analysis on only the *cis*-acting CNVs. *Cis*-acting and *trans*-acting CNVs behave differently, with the *cis*-acting CNVs usually having dominant effects. Further Figure A.1 in the Online Supplement and those alike suggest that the effects of *cis*-acting CNVs can be nonlinear. Many of the existing studies, such as [9, 16], are limited by assuming linear effects. There are a few studies that consider nonlinear modeling [12] and suggest the limitations of linear modeling. However, they are under different contexts, and there is still a lack of large-scale regression analysis.

In data analysis, many GEs have been observed to have long-tailed distributions. In addition, various technical problems can cause data contamination [21]. In Figure A.2 (Online Supplement), the long-tailed characteristic of GE distributions is clearly seen. Most of the existing studies adopt non-robust estimation, which, with such distribution, may result in biased model parameter estimation and false variable selection. The existing studies are also limited by lacking attention to the correlations among CNVs, which are commonly observed in data analysis. Studies in other contexts have shown that effectively accommodating the correlations among variables leads to more accurate selection and estimation. However, this has not been pursued in the GE-CNV analysis.

Although multiple studies have been conducted on modeling the GE-CNV associations, they fail to address one or more of the aforementioned challenges. In this article, we develop a new and more effective analysis approach, which can directly overcome the limitations of existing analyses. The primary analysis goal is to identify CNVs that are relevant for GEs and estimate their effects. For a specific GE, we jointly model the effects of multiple candidate CNVs and data-dependently search for the important ones. A partially linear model is adopted, with a nonlinear effect for the *cis*-acting CNV and linear effects for the *trans*-acting CNVs. A robust loss function is adopted to accommodate long-tailed distributions and data contamination. Penalization is used for estimation and variable selection. We describe the correlations among CNVs using a network structure and accommodate it in estimation. It is remarkable that with multiple advancements over the existing alternatives, the proposed method is still statistically and numerically manageable.

In Section 2, we introduce the data and model settings. The proposed method is described in Section 3. We develop an effective computational algorithm and rigorously establish the consistency properties. Simulation is conducted in Section 4. We analyze the TCGA (The Cancer Genome Atlas) data on the PCD (programmed cell death) pathway in Section 5. Additional technical details and numerical results are provided in an online supplement.

2. Data and model settings

Consider a random sample of size n, each with p GE and k CNV measurements. Let $Y = (y^1, \ldots, y^p)$ denote the $n \times p$ matrix of GEs with $y^m = (y_1^m, \ldots, y_n^m)^\top$, for all $m \in \{1, ..., p\}$, and let $X = \left(X_1, \ldots, X_k, X_{k+1}\right)$ be the $n \times (k+1)$ matrix of CNVs with $X_t = \left(X_{t1}, \ldots, X_{tn}\right)^\top$, for all $t \in \{1, \ldots, k\}$ and the $n \times 1$ vector $X_{k+1} = (1, \ldots, 1)^\top$ for intercept. For simplicity of notation, consider data with matched GE and CNV measurements. For the ith sample and mth GE, we consider

$$y_i^m = f_m\left(X_{mi}\right) + \sum_{t=1}^{k+1} X_{ti}\alpha_t^m \mathbf{1}(t \neq m) + \varepsilon_i^m. \quad (1)$$

where the f_ms are unknown smooth functions, with the consideration that most biological processes are continuous. In practice, some datasets may have not fully matched GEs and CNVs. This can be easily accommodated by making minor modifications to the proposed method and will not be further discussed.

For identifiability, we impose the constraint $f_m\left(X_{m1}\right) + \cdots + f_m\left(X_{mn}\right) = 0$ in estimation. Let $\alpha^m = \left(\alpha_1^m, \ldots, \alpha_{m-1}^m, \alpha_{m+1}^m, \ldots, \alpha_{k+1}^m\right)^\top$ denote the $k \times 1$ regression coefficient vector, and ε_i^m be the random error. Also denote $X = \left(x_1, \ldots, x_n\right)^\top$, where $x_i^\top$ is the ith row of X. For the random error, assume that $\Pr\left(\varepsilon_i^m \leq 0 | x_i\right) = 1/2$. Note that strict moment assumptions, which are commonly needed in the existing studies, are not made on ε_i^m.

Under the smoothness condition, we approximate f_m using the basis expansion

$$f_m\left(X_{mi}\right) \approx \sum_{\ell=1}^{L} \gamma_{m\ell} B_{m\ell}\left(X_{mi}\right),$$

where L is the number of basis functions, $\gamma_m = \left(\gamma_{m1}, \ldots, \gamma_{mL}\right)^\top$ is the spline coefficient vector, and $B_m\left(X_{mi}\right) = \left(B_{m1}\left(X_{mi}\right), \ldots, B_{mL}\left(X_{mi}\right)\right)^\top$ is the set of normalized B-spline basis. Denote $\alpha = \left(\alpha^{1\top}, \ldots, \alpha^{p\top}\right)^\top$ and $\gamma = \left(\gamma_1^\top, \ldots, \gamma_p^\top\right)^\top$.

In the literature, there are multiple strategies for modeling *trans*-acting CNV effects. Some describe their effects through their regulating genes and model across-gene effects. Some

studies suggest that such a strategy is “more direct” and may have a strong biological basis. However, this strategy can be mathematically challenging, since GEs are on both sides of the regression and dynamic network analysis may be needed. In this study, we directly regress GEs on CNVs. This strategy has been adopted in multiple recent studies, including [20, 26, 30], and generated scientifically and statistically interesting results. It is beyond the scope of this article to compare different modeling strategies. Model (1) jointly describes the effects of all candidate CNVs, includes that with a single CNV as a special case, and is more flexible. It is a partially linear model and allows the *cis*-acting CNV to contribute to GE in a nonlinear way. It is possible to also allow nonlinear effects for *trans*-acting CNVs. However, their effects are comparatively smaller, and it is reasonable to “pay more attention” to the *cis*-acting CNV with a larger effect. In addition, excessive nonlinear modeling leads to high computational cost and unstable estimation. Model (1) also includes the multivariate linear model as a special case.

3. Robust network-based penalized estimation

Robust estimation is needed when the random error has a long-tailed distribution or contamination. Consider the LAD (least absolute deviation) loss function

$$Q(\alpha,\gamma)=\frac{1}{n}\sum_{m=1}^{p}\sum_{i=1}^{n}\left|y_i^m-\sum_{\ell=1}^{L}\gamma_{m\ell}B_{m\ell}(X_{mi})-\sum_{t=1}^{k+1}X_{ti}\alpha_t^m\mathbf{1}(t\neq m)\right|. \quad (2)$$

The LAD loss is a special case of the popular quantile regression.

To accommodate the high data dimensionality, and to select relevant CNVs, we consider the penalized estimate

$$(\hat{\alpha},\hat{\gamma}) = \arg\min_{\alpha,\gamma}\{Q(\alpha,\gamma)+P(\alpha,\gamma;\lambda,\zeta)\}, \quad (3)$$

where $\alpha=\left(\alpha^{1\top},\ldots,\alpha^{p\top}\right)^{\top}$, $\gamma=\left(\gamma_1^{\top},\ldots,\gamma_p^{\top}\right)^{\top}$, and λ and ζ are tuning parameters. A nonzero component of the estimate suggests an association between the corresponding GE and CNV. For penalty, we first consider

$$P_A(\alpha,\gamma;\lambda_1,\lambda_2,\zeta)=\sum_{m=1}^{p}\sum_{t=1}^{k+1}\phi(\alpha_t^m;\lambda_1,\zeta)\mathbf{1}(t\neq m)+\sum_{m=1}^{p}\phi\left(\|\gamma_m\|_1;\lambda_2,\zeta\right)\equiv P_1+P_2, \quad (4)$$

where $\phi(s;\lambda,\zeta)=\lambda\int_0^{|s|}(1-x/(\lambda\zeta))_+dx$ is the Minimax Concave Penalty (MCP, [29]) with tuning parameter λ and regularization parameter ζ. For *trans*-acting CNVs, penalties are imposed on their regression coefficients. For *cis*-acting CNVs , their effects are represented by vectors of regression coefficients, and penalties are imposed on the group norms of these vectors. We adopt the ℓ_1 group norm, which leads to similar statistical properties as the ℓ_2

group norm but simplifies computation. The proposed analysis imposes the same tuning parameters across all GEs and CNVs, which ensures that they are analyzed on the same ground. This is reasonable as no GE/CNV is "special" compared to others. In addition, having a smaller number of tunings (compared to allowing for GE-/CNV-specific tunings) significantly reduces computational cost.

A limitation of P_A is that it does not accommodate the (sometimes high) correlations among CNVs. To solve this problem, we first adopt a network structure to describe the correlations among CNVs. In the CNV network, a node corresponds to a CNV, and two nodes are connected if the corresponding CNVs are correlated. To construct adjacency, which quantifies the network connection between any two nodes, we consider the following approach [28]. For nodes i and j, let r_{ij} be their correlation coefficient. Let A = (a_{ij} : 1 ≤ i, j ≤ k) be the adjacency matrix. Consider $a_{ij} = r_{ij}^{\varrho}\mathbf{1}\{|r_{ij}| > r\}$. In subsequent numerical studies, we set ϱ = 5, which retains the sign of r_{ij}, down-weighs weak correlations (which are possibly noises), and keeps the strong ones. Choosing the power transformation and the specific value of ϱ follows the published literature [8, 28]. The cutoff r leads to a sparse network and is calculated from the Fisher transformation [4] as $r = \{\exp(2c/\sqrt{n-3}) - 1\}/\{\exp(2c/\sqrt{n-3}) + 1\}$, where c = 1.96 is determined from the standard normal distribution. We refer to [8, 28] for more discussions on network construction. There are other ways of defining network adjacency. For example, some approaches use biological (such as pathway) information. It is expected that they are equally applicable. As our goal is not to compare different network constructions, we focus on this specific network structure without further discussing others.

To accommodate the CNV network, we propose the penalty

$$P_B(\alpha,\gamma;\tilde{\lambda},\zeta) = P_A(\alpha,\gamma;\lambda_1,\lambda_2,\zeta) + \lambda_3\sum_{m=1}^{P}\sum_{1\le j<t\le k}|a_{jt}|\times|\alpha_j^m - \mathrm{sgn}(a_{jt})\alpha_t^m|\mathbf{1}(j,t\neq m) \quad (5)$$

$$\equiv P_1 + P_2 + P_3,$$

where $\tilde{\lambda} = (\lambda_1,\lambda_2,\lambda_3)$, and sgn is the sign function. The newly added P_3 has been motivated by the following considerations. When two CNVs are highly correlated, it encourages their regression coefficients to have similar magnitudes. The "directions" of estimates and "degree of encouragement" are adjusted by a_{jt}. A similar strategy has been developed in the literature under different contexts [14]. The ℓ_1 penalty is adopted to be "consistent" with the loss function, which significantly simplifies computation. The tuning parameters are data-dependently adjusted to effectively avoid over shrinkage.

The proposed P_3 shares some similarity with the Laplacian penalty [8], contrasted penalty [19], and others in shrinking the differences of regression coefficients but also has notable differences. First it is noted that the application context is significantly different. In addition, different from the Laplacian penalty, it is based on the ℓ_1 norm. The fused Lasso is defined on the differences of consecutive coefficients and demands a "spatial lining-up", which

differs from the considered network structure. Also different from the existing methods, penalties are only imposed on *trans*-acting CNVs, not all CNVs. *Cis*- and *trans*-acting CNVs behave differently, and shrinking the difference between a linear and a nonlinear effect is insensible. In the literature, there are also methods that incorporate the networks of both GEs and CNVs [20]. The numerical study in [20] suggests that, after incorporating the CNV network structure, further accommodating the GE network leads to incremental improvements but significant computational cost. In addition, the goal of penalization is to select and estimate the coefficients of CNV effects. It thus seems "more direct" to build the penalty on the correlations (network) of CNVs.

3.1. Computation

With fixed tunings, optimization with different GEs can be conducted separately in a parallel manner. For each GE, the proposed penalty has an "MCP+ℓ_1" form, both of which have been well studied in the literature. Overall, optimization can be achieved by adopting and modifying existing techniques. Specifically, we develop an effective computational algorithm based on the MM (majorize minimization) and CD (coordinate descent) techniques.

The MM step: Denote $\theta_{s_0}(s)$ as the majorization function of $\phi(s;\lambda,\zeta)$ at s_0. In the proposed iterative algorithm, we use the superscript "(d)" to denote the dth iteration. Denote $\phi'\left(\left|\alpha_t^m\right|+;\lambda,\zeta\right)$ as the limit of $\phi'(|s|;\lambda,\zeta)$ as $s\to\alpha_t^m$ from above. Then we have

$$\theta_{\alpha_t^{m(d-1)}}\left(\left|\alpha_t^m\right|\right)=\phi'\left(\left|\alpha_t^{m(d-1)}\right|+;\lambda,\zeta\right)\left[\left|\alpha_t^m\right|-\left|\alpha_t^{m(d-1)}\right|\right]+\phi\left(\left|\alpha_t^{m(d-1)}\right|;\lambda,\zeta\right) \quad (6)$$

as the majorization function of $\phi\left(\left|\alpha_t^m\right|;\lambda,\zeta\right)$.

The CD step: With the assistance of MM, minimization can be solved using the CD approach. In the dth iteration, for $m\in\{1,\ldots,p\}$ and $t\in\{1,\ldots,k+1\}$.

Step 1. Minimize the majorized penalized objective function with respect to γ_m, with the other parameters fixed at their current estimates. Consider only terms relevant to γ_m:

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i^m-\sum_{t=1}^{k+1}X_{ti}\alpha_t^m 1(t\neq m)-\sum_{\ell=1}^{L}\gamma_{m\ell}B_{m\ell}\left(X_{mi}\right)\right|+\phi'\left(\left\|\gamma_m^{d-1}\right\|_1+;\lambda_2,\zeta\right)\left\|\gamma_m\right\|_1.$$

With the assistance of slack variables, this optimization problem can be casted as a linear programming problem, viz.

$$\underset{\xi,\gamma_m}{\text{minimize}}\quad \frac{1}{n}\sum_{i=1}^{n}\left(\xi_i^+ +\xi_i^-\right)+\phi'\left(\left\|\gamma_m^{(d-1)}\right\|+;\lambda_2,\zeta\right)\sum_{\ell=1}^{L}\left(\gamma_{m\ell}^+ +\gamma_{m\ell}^-\right)$$

$$\text{subject to}\ \xi_i^+ - \xi_i^- = y_i^m - \sum_{t=1}^{k+1} X_{ti}\alpha_t^m \mathbf{1}(t \neq m) - \sum_{\ell=1}^{L} \gamma_{m\ell} B_{m\ell}(X_{mi}); \quad (7)$$

$$\xi_i^+ \geq 0, \xi_i^- \geq 0 \text{ for all } i \in \{1, \ldots, n\},$$

where $a^+ = a\mathbf{1}(a \geq 0)$ and $a^- = -a\mathbf{1}(a \leq 0)$. This can be solved using existing software.

Step 2. Minimize the majorized penalized objective function with respect to α_j^m, with the other parameters fixed at their current estimates. It is equivalent to minimizing

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left|y_i^m - \sum_{\ell=1}^{L}\gamma_{m\ell}B_{m\ell}(X_{mi}) - \sum_{t=1}^{k+1}X_{ti}\alpha_t^m\mathbf{1}(t \neq j, m) - X_{ji}\alpha_j^m\right| \\
&+ \phi'\left(\left|\alpha_j^{m(d-1)}\right| + ; \lambda_1, \zeta\right)\left|\alpha_j^m\right| + \lambda_3 \sum_{t=j+1}^{k}\left|a_{jt}\right| \times \left|\alpha_j^m - \operatorname{sgn}(a_{jt})\alpha_t^m\right|\mathbf{1}(j, t \neq m) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left|X_{ji}\right| \times \left|v_{ij}^m\right| + \phi'\left(\left|\alpha_j^{m(d-1)}\right| + ; \lambda_1, \zeta\right)\left|\alpha_j^m\right| + \lambda_3 \sum_{t=j+1}^{k}\left|a_{jt}\right| \times \left|\alpha_j^m - \operatorname{sgn}(a_{jt})\alpha_t^m\right|\mathbf{1}(j, t \neq m),
\end{aligned}$$

where

$$v_{ij}^m = \frac{1}{X_{ji}}\left\{y_i^m - \sum_{\ell=1}^{L}\gamma_{m\ell}B_{m\ell}(X_{mi}) - \sum_{t=1}^{k+1}X_{ti}\alpha_t^m\mathbf{1}(t \neq j, m)\right\} - \alpha_j^m,$$

for all $i \in \{1, \ldots, n\}$. Minimizing the above objective function can be further formulated as a weighted median regression with $n + k - j + 1$ pseudo-observations

$$\arg\min_{\alpha_j^m}\left\{\frac{1}{n+k-j+1}\sum_{i=1}^{n+k-j+1} w_{ij}\left|\tilde{v}_{ij}^m\right|\right\}, \quad (8)$$

Where

$$\tilde{v}_{ij}^m = \begin{cases} v_{ij}^m & \text{if i} \in \{1, \ldots, n\}, \\ \alpha_j^m & \text{if i = n+1} \\ \alpha_j^m - \operatorname{sgn}(a_{jt})\alpha_t^m & \text{if i} \in \{n+2, \ldots, n+k-j+1\}, \end{cases}$$

and

$$w_{ij}^{m} = \begin{cases} \left|X_{ji}\right|/n & \text{if i } \in \{1,\ldots,n\}, \\ \phi'\left(\left|\alpha_j^{m(d-1)}\right| + ;\lambda_1,\zeta\right) & \text{if i = n+1} \\ \lambda_3\left|a_{jt}\right| & \text{if i } \in \{n+2,\ldots,n+k-j+1\}. \end{cases}$$

The minimizer of (8) is the weighted median of the $n + k - j + 1$ pseudo-observations.

The overall algorithm can be summarized as follows.

Algorithm: The two-step coordinate descent algorithm

Initialize $d = 0$, $\gamma^{(d)} = 0$, and $\alpha^{(d)} = 0$

Repeat

for $m \in \{1, \ldots, p\}$ do

for $j \in \{1, \ldots, k+1\}$ do

Update $\gamma_m^{(d+1)} (j = m)$ via (7);

Update $\alpha_j^{m(d+1)} (j \neq m)$ via (8);

end

end

$d = d + 1$;

until the difference between two consecutive estimates is less than a cutoff (set as 10^{-4} in our numerical study); Return the estimate of (α, γ) at convergence.

The LAD approach is a special case of quantile regression. For the nonconvex penalized quantile regression, Peng and Wang [15] established convergence to a stationary point. Convergence properties for fused Lasso type penalties have been examined in Friedman et al. [5]. It should be noted that, in the aforementioned studies, the penalties have much simpler forms. The newly added P_3 does not have a separable form, whereas most studies that are able to establish convergence to global optimizers have separable penalties. With the proposed algorithm, the value of the penalized objective function decreases at each iteration and is bounded below. It is conjectured that, following [5, 15, 22], the proposed algorithm converges to a coordinate-wise minimum, which is also a stationary point. Our literature search does not suggest a way to establish convergence to the global optimizer for the proposed approach (and under what conditions). We postpone rigorous research on convergence to future studies. In our numerical studies, convergence is achieved in a small to moderate number of iterations.

Tuning parameter selection: The tuning/regularization parameters have similar implications as in the literature and are selected using commonly adopted approaches. Specifically, we set $\lambda_2 = \sqrt{L}\lambda_1$ and impose comparable penalization to linear and nonlinear effects. λ_1 (and so λ_2) and λ_3 control the degree of shrinkage and are selected using V-fold cross validation ($V = 5$ in our numerical study). The regularization parameter ζ balances

between unbiasedness and concavity. Breheny and Huang [1], Zhang [29] and other studies suggest experimenting with a few values (including 1.8, 3, 4.5, 6, and 10) or fixing its value. In our numerical study, we examine this sequence and find that the results are not sensitive to the value of ζ and set $\zeta = 3$. In practice, to be prudent, ζ values other than 3 should also be examined. There are many publications on tuning parameter selection with splines. In our numerical study, we choose cubic splines and $\sim n^{1/5}$ equally spaced interior knots.

Implementation: To facilitate data analysis, we implement the above procedure in R. The code is available at https://github.com/shuanggema.

3.2. Consistency properties

Theoretical study of penalized robust regression under high-dimensional settings is limited [25]. Compared with the existing ones, this study is more complicated with the partially linear modeling and introduction of the network-based penalty. In addition, the proposed method involves simultaneously estimating a large number of high-dimensional models, which brings significantly more challenges than estimating a single high-dimensional model. Our theoretical study not only provides a strong basis for the proposed method but also sheds light on several existing methods and has independent value.

Recall that $X = (X_1,..., X_k, X_{k+1}) = (x_1,...,x_n)^\top$ is the $n \times (k + 1)$ design matrix. Define $x_{i,I}$ as the subvector of x_i indexed by $I \subseteq \{1,...,k + 1\}$. Let I^c and $|I|$ denote the complement and cardinality of set I, respectively. Then the model can be rewritten, for all $m \in \{1,..., p\}$ and $I \in \{1,...,n\}$, as

$$y_i^m = f_m(X_{mi}) + x_{i,m^C}^\top \alpha_{m^C}^m + \varepsilon_i^m, \quad (9)$$

where x_{i,m^C} denotes x_i with the mth element removed, and $\alpha_{m^C}^m$ is the regression coefficient vector associated with x_{i,m^C} for the mth GE.

Let $\left\{B_{m\ell}(x) = \sqrt{L}\left(S_\ell(x) - \sum_{i=1}^n S_\ell(X_{mi})/n\right) : \ell = 1, \ldots, L\right\}$ be the set of normalized basis functions, where $\{S_\ell(x)\}_{\ell=1}^L$ is the set of B-spline basis. With basis expansion, one has, for all $m \in \{1,..., p\}$ and $i \in \{1, ... ,n\}$,

$$y_i^m \approx B_m(X_{mi})^\top \gamma_m + x_{i,m^C}^\top \alpha_{m^C}^m + \varepsilon_i^m.$$

With a slight abuse of notation, we also denote $\alpha_m^m = \gamma_m$. Denote the true value of regression coefficients as $\alpha^* = (\alpha_t^{m*})$. Then γ_m^*, the true value of γ_m, is also labeled as α_t^{m*}. Let f_m^* be the true value of f_m. Denote $\mathcal{S} = \left\{(t,m) : \alpha_t^{m*} \neq 0\right\}$, which is the union of $\mathcal{F} = \left\{(m,m) : \alpha_m^{m*} \neq 0\right\}$ and $\mathcal{C} = \left\{(t,m) : \alpha_t^{m*} \neq 0, t \neq m\right\}$. Let $\mathcal{D} = \left\{m : f_m^* \neq 0\right\}$ be the set of

important nonparametric effects. For $m = \in \{1,\ldots, p\}$, define $\mathcal{S}_m = \{t:(t,m) \in \mathcal{S}\}$ and $\mathcal{C}_m = \{t:(t,m) \in \mathcal{C}\}$. Let $\|\cdot\|_q$ be the ℓ_q norm for vectors and $C^v([a, b])$ be the space of v-times continuously differentiable functions defined on $[a, b]$.

Let $\alpha_s = \{\alpha_t^m:(t,m) \in \mathcal{S}\}$, $\gamma_{\mathcal{D}} = \{\gamma_m:m \in \mathcal{D}\}$, and $\alpha_{\mathcal{C}} = \{\alpha_t^m:(t,m) \in \mathcal{C}\}$. First consider

$$\mathcal{O}_n(\alpha_S) = \frac{1}{n}\sum_{m=1}^{P}\sum_{i=1}^{n}\left|y_i^m - B_m(X_{mi})^\top \gamma_m \mathbf{1}(m \in \mathcal{D}) - x_{i,\mathcal{C}_m}^\top \alpha_{\mathcal{C}_m}^m\right| \quad (10)$$
$$+ \lambda_3 \sum_{m=1}^{p}\left\{\sum_{(j,t)\in\mathcal{A}_{1m}}|a_{jt}| \times |\alpha_j^m - \mathrm{sgn}(a_{jt})\alpha_t^m| + \sum_{(j,t)\in\mathcal{A}_{1m}}|a_{jt}| \times |\alpha_i^m|\right\},$$

Where $\mathcal{A}_{1m} = \{(j,t): j, t \in \mathcal{C}_m, j < t \le k; a_{jt} \neq 0\}$ and $\mathcal{A}_{2m} = \{(j,t): j \in \mathcal{C}_m, t \notin \mathcal{C}_m, j < t \le k; a_{jt} \neq 0\}$. $\mathcal{O}_n(\alpha_{\mathcal{S}})$ is the oracle counterpart of the proposed objective function. Define

$$U_{i,m} = \begin{cases} \left(B_m(X_{mi})^\top, x_{i,C_m}^\top\right)^\top & \text{if } m \in D, \\ x_{i,C_m} & \text{if } m \notin D. \end{cases}$$

The following conditions are needed to establish the asymptotic properties.

(C1) The cardinality of $\mathcal{D}$, denoted as d_0, is fixed. For $m \in \mathcal{D}$, $f_m^* \in C^v([a,b])$.

(C2) X_i has a compact support on $[a,b]$, for $i \in \{1,\ldots,k\}$. There exist positive constants $M1 < M2$ such that $M_1 \le \lambda_{\min}(n^{-1}\sum_{i=1}^n U_{i,m}U_{i,m}^\top) \le \lambda_{\max}(n^{-1}\sum_{i=1}^n U_{i,m}U_{i,m}^\top) \le M_2$ for all m, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalues, respectively.

(C3) Denote the density function of ε_i^m conditional on $U_{i,m}$ as $g_{i,m}(\cdot|U_{i,m})$. Uniformly over i and m, in a neighborhood of zero, $g_{i,m}(\cdot|U_{i,m})$ is bounded away from zero and infinity and has a bounded first order derivative.

These conditions are mild, and comparable ones have been assumed in the literature; see, e.g., [3, 13]. Conditions (C2) and (C3) also imply that, for all m, there exist positive constants $M_3 < M_4$ such that

$$M_3 \leq \lambda_{\min}\left\{n^{-1}\sum_{i=1}^{n} g_{i,m}\left(0|U_{i,m}\right)U_{i,m}U_{i,m}^{\top}\right\} \leq \lambda_{\max}\left\{n^{-1}\sum_{i=1}^{n} g_{i,m}\left(0|U_{i,m}\right)U_{i,m}U_{i,m}^{\top}\right\} \leq M_4 .$$

(11)

Consider the estimator $\tilde{\alpha}_S = \arg\min \mathcal{O}_n(\alpha_S)$. Let $s_m = |\mathscr{C}_m|$ and $J = \max_m \sum_{t \neq m}^{k} |a_{mt}|$. The following theorem establishes the consistency of $\tilde{\alpha}_{\mathcal{S}}$.

Theorem 1. *Assume that (C1)–(C3) hold and that* $\kappa_n \to \infty$ *as* $n \to \infty$. *In addition,* $L \to \infty$, $(L + s_m)^2\kappa^n/n \to 0$ $(m \in \mathscr{D})$, $s_m^2 k_n/n \to 0 (m \in \mathscr{D})$, *and* $\lambda_3 J \to 0$, *where* $C_1,\ldots,C_p$ *are large constants. With probability at least*

$$1 - \sum_{m \in \mathscr{D}} \exp\left\{-\left(L + s_m\right)k_n/8\right\} - \sum_{m \notin \mathscr{D}} \exp\left(-s_m k_n/8\right)$$

the estimator $\tilde{\alpha}_s = \left\{\tilde{\alpha}_t^m : (t,m) \in \mathcal{S}\right\}$ satisfies

For $m \in \mathscr{D}$, $\left|\tilde{f}_m(x) - f_m^*(x)\right| + \left\|\tilde{\alpha}_{\mathscr{C}_m}^m - \alpha_{\mathscr{C}_m}^{m*}\right\|_2 \leq \delta_m$, *where* $\delta_m = C_m\left\{\sqrt{(L + s_m)k_n/n} + L^{-\nu} + \lambda_3 J\right\}$ *and* $\tilde{f}_m(x) = B_m(x)^{\top}\tilde{\gamma}_m$ *for any* $x \in$ [a, b].

For $m \notin \mathscr{D}$, $\left\|\tilde{\alpha}_{\mathscr{C}_m}^m - \alpha_{\mathscr{C}_m}^{m*}\right\|_2 \leq \delta_m$, *where* $\delta_m = C_m\left(\sqrt{s_m k_n/n} + \lambda_3 J\right)$.The tail probability in Theorem 1 is exponentially small. In other words, the proposed method is able to accommodate high-dimensional data with $\ln p = o\left(k_n \min_{m \notin D} s_m\right)$. Consider the oracle estimator $\hat{\alpha}^o = \left\{\hat{\alpha}_S, \hat{\alpha}_{S^{\mathrm{C}}}\right\}$, where $\hat{\alpha}_S = \tilde{\alpha}_S$ and $\hat{\alpha}_{S^{\mathrm{C}}} = 0$. Let

$$b_l\left(\lambda_1,\lambda_3\right) = \max_{j,m}\left[-\lambda_1 - \lambda_3\left\{\sum_{t \in m^{\mathrm{C}} - \mathscr{C}_m} \left|a_{jt}\right| - \sum_{t \in \mathscr{C}_m} a_{jt}\operatorname{sgn}\left(\alpha_t^{m*}\right)\right\}\right],$$

$$b_u\left(\lambda_1,\lambda_3\right) = \min_{j,m}\left[\lambda_1 + \lambda_3\left\{\sum_{t \in m^{\mathrm{C}} - \mathscr{C}_m} \left|a_{jt}\right| - \sum_{t \in \mathscr{C}_m} a_{jt}\operatorname{sgn}\left(\alpha_t^{m*}\right)\right\}\right].$$

The following additional conditions are needed.

(C4) $b(\lambda_1,\lambda_3) = \min\{-b_l(\lambda_1,\lambda_3), b_u(\lambda_1,\lambda_3)\} > 0$, $nb(\lambda_1,\lambda_3) \to \infty$.

(C5) $(\max_m \delta_m)^{-1} \min_{(t,m)\in\mathscr{E}} |\alpha_t^{m*}| \to \infty$ and $(\max_{m\in\mathscr{D}} \delta_m)^{-1} \min_{m\in\mathscr{D}} \sum_{i=1}^n f_m^*(X_{mi})^2/n \to \infty$, where δm is defined in Theorem 1.

(C6) $\max_m \delta_m \le \lambda_1$ and $\max_{m\in\mathscr{D}} \delta_m \le \lambda_2$.

Condition (C4) requires that the smallest signal does not decay too fast. Comparable conditions have been assumed in Wang et al. [24] and others.

Theorem 2. *Assume that (C1)–(C6) and conditions in Theorem 1 hold. If* $\ln p = o\left(k_n \min_{m\notin\mathscr{D}} s_m\right)$, $\max_m \sqrt{L+s_m}\,\delta_m = o\{b(\lambda_1,\lambda_3)\}$, $(\max_m s_m \vee L)\ln n + \ln k + \ln p = o\{nb(\lambda_1,\lambda_3)\}$ *and* $\max_{m<\mathscr{D}} s_m \ln n + \ln k + \ln p = o(\lambda_2)$, *then with probability converging to 1,* $\hat{\alpha}^o$ *is a local minimizer of the penalized LAD objective function with penalty* (5).

This theorem establishes that the proposed estimator enjoys the same asymptotic consistency as the oracle estima-tor with probability approaching one. This property holds under ultrahigh dimensions without restrictive (for example, moment) conditions on the random errors. The proofs are presented in an Online Supplement.

4. Simulation

We set $(n, k, p) = (200, 200, 200)$. Note that although k and p may seem modest, the number of unknown effects $((k+1)\times p)$ is in fact very large. The analysis is even more challenging with the basis expansion for nonlinear effects. An important component of the proposed method is to accommodate the connections among CNVs. In practice, the number of highly correlated CNVs is not expected to be large. As to be shown in data analysis, this simulation setting mimics that of a pathway. GEs/CNVs in the same pathway are more likely to have related functions and correlated measurements and are sensible to be analyzed together, whereas different pathways are largely different and can be analyzed separately.

In TCGA and other data, the processed CNV data marginally have unimodal continuous distributions close to normal. Here we simulate the CNVs to have a multivariate normal distribution with marginal means zero and a block diagonal covariance structure. There are 40 blocks, with 5 CNVs per block. CNVs within the same block have correlation coefficient ρ, and those in different blocks are uncorrelated. Three levels of correlation are considered.

To generate the parametric parameters, we first simulate the 200×200 matrix $M = \mathrm{diag}(U_1 * I_1, \ldots, U_{40} * I_{40})$, where "$*$" denotes the element-wise product between the 5×5 matrices U_s and I_s with $s \in \{1,\ldots, 40\}$. Each entry of I_s is generated from a Bernoulli distribution with a success probability of 0.8. For $s \in \{2,\ldots, 10\}$, the entries in the odd and even columns of U_s are simulated from $\mathscr{U}(0.8, 1)$ and $\mathscr{U}(-1, -0.8)$, respectively. For $s \in \{12,\ldots, 20\}$, the entries in the odd and even columns of U_s are simulated from $\mathscr{U}(-1, -0.8)$ and $\mathscr{U}(0.8, 1)$, respectively. For $s \in \{1, 11\}$ and $s > 20$, all components of U_s are 0. The diagonal elements

of M are then set as zero. For each GE, the parametric parameters correspond to a row of M. Under this setup, trans-acting CNVs with nonzero effects belong to the same block as the cis-acting CNVs and are correlated. This is motivated by the presence of small functional groups, each of which consists of a small number of correlated CNVs and coregulated GEs [31].

For the effects of cis-acting CNVs, set $f_m(x) = 2\sin(x\pi/2)$ for $m \in \{1,\ldots, 5, 11,\ldots, 55, 61, \ldots, 100\}$ and = 0 otherwise. For blocks 1 and 11, GEs are regulated by cis-acting CNVs only. For blocks 2 and 12, GEs are regulated by trans-acting CNVs only. For the rest of the blocks with $s \leq 20$, GEs are regulated by both types of CNVs. For blocks with $s \in \{21,\ldots, 40\}$, there is no detectable regulation. The settings thus comprehensively cover all possible scenarios. The true parametric effects are generated randomly. By expectation, there are 288 of them. There are 90 true nonparametric effects. The intercepts are set as zero. Consider three random error distributions: N(0, 1) (Error 1), 0.85 N(0, 1)+0.15 Cauchy (Error 2), and 0.75 N(0, 1)+0.25 Cauchy (Error 3), which have different contamination levels.

Beyond the proposed method, we also consider seven alternatives (Table A.1, Online Supplement): (A_2) The partially linear modeling is adopted, and $P_1 + P_2$ is applied. There is no accommodation of the network structure. (A_3) All CNV effects are assumed to be linear, P_1 is used for estimation and selection, and the network structure is accommodated using P_3. (A_4) All CNV effects are assumed to be linear, and there is no accommodation of the network structure. Methods A_5–A_8 are parallel to A_1–A_4 but adopt the non-robust LS (least squares) loss. Although there are other alternatives, these seven have an analysis framework closest to that of the proposed method and can directly establish the merit of the proposed partially linear modeling, robust loss, and accommodation of the CNV network structure.

When evaluating the proposed and alternative methods, we mainly focus on selection. If CNVs that are relevant to GEs can be accurately identified, there are multiple ways of generating satisfactory estimation. Summary results are presented in Tables 1 and 2. The means and standard deviations of true and false positives are computed for the parametric and nonparametric effects separately based on 100 replicates. The main findings are as follows:

- **(a)** Simulation shows the advantage of partially linear modeling. For example with Error 1 and $\rho = 0.5$, TP1 and TP2 under A_1 are 287.6 and 90.0, respectively. In comparison, TP1 and TP2 under A_3 are 232.7 and 50.9, respectively.
- **(b)** When there is no contamination and correlation is weak to moderate, the non-robust methods can have satisfactory performance. However, with contamination, the robust methods outperform. For example with Error 2 and $\rho = 0.5$, TP1 and TP2 under A_1 are 286.8 and 89.5, respectively. TP1 and TP2 under A_5 are 274.5 and 87.9, respectively, which are also very satisfactory. However, the price is that A_5 has a large number of false positives with FP1 = 173.8.
- **(c)** When there exist moderate to strong CNV correlations, it pays off to accommodate the network structure. For example with Error 2 and $\rho = 0.9$, A_1 and A_2 have FP1 162.0 and 190.3, respectively. We note that with strong

correlations, the proposed method may have more false positives. This observation is reasonable. The network-based penalty shrinks the differences between regression coefficients. When an unimportant variable is correlated with an important one (within the same block in this simulation), this penalty tends to "pull" their coefficients together, also select the unimportant one, and increase false positives. Overall, across the whole spectrum of simulations, we observe superior performance of the proposed method.

In the second set of simulation, we examine whether the superiority of the proposed method over the alternatives depends on signal level. Specifically, under Error 2, we reduce the signal levels to 0.8 of those described above. Results are shown in Table A.2 (Online Supplement). A_1 outperforms A_2–A_4 in a similar way as observed above. Results on the non-robust methods, which have inferior performance, are omitted.

In the third set of simulation, we examine if performance of the proposed method depends on the number of signals. Specifically, the expected number of parametric effects is doubled, and the number of nonparametric effects remains at 90. Other settings are similar to those under the first set of simulation. Results of the robust methods are shown in Table A.3 (Online Supplement). We draw similar conclusions as those from Tables 1 and 2.

In the fourth set of simulation, we consider more realistic CNV distributions. Specifically, we use the real data analyzed in the next section. For each simulation replicate, we randomly sample 200 subjects from the TCGA data, each with 426 observed CNV measurements. The parametric coefficient matrix is generated in a similar way as under the first set of simulation. The same nonlinear f_m's and error distributions are adopted. Results are summarized in Table A.4 (Online Supplement). With more complex CNV distributions, all methods perform worse than in the first three sets. However, the patterns are similar, with the proposed method significantly outperforming the alternatives.

As in many other studies, the simulated CNV distributions in the first three sets may be much simpler than practically encountered. Luckily, this is "compensated" by the last set. The simulated settings have comprehensively covered different numbers and strengths of signals and correlations. Similar settings have been adopted in the literature [20]. In the literature, research on nonlinear cis-acting CNV effects is limited. It is sensible to start with simple forms, which have been extensively adopted in semiparametric modeling. As in any other study, the simulation set-tings have limitations. However, the superiority of the proposed method is clearly seen. We have also experimented with a few other settings and drawn similar conclusions (results omitted).

With the robust loss function and more complicated penalty, the proposed method has higher computational cost than some alternatives. However, simulation suggests that it is still computationally affordable. Specifically, with fixed tunings, the analysis of one replicate in the first set of simulation takes 5.1 (A_1), 3.4 (A_2), 4.7 (A_3), 2.0 (A_4), 3.9 (A_5), 1.8 (A_6), 3.1 (A_7), and 1.3 (A_8) minutes on a desktop with standard configurations. We have also observed that, when sample size and data dimensionality increase, computational cost increases moderately (details omitted). The current code is written in R. Computer time may be much reduced if the computational core is written in C and parallel computing is adopted.

5. Analysis of TCGA data

We analyze the TCGA data on cutaneous melanoma. In our analysis, the processed level 3 data are downloaded from the TCGA portal. Details on data collection and processing are available at TCGA website. Briefly, the GE data were collected using the Illumina Hiseq 2000 RNA-SEQ Ver 2 platforms. The processed data are the robust Z-scores, which have been lowess-normalized, log-transformed, and median-centered. The measurements indicate the under- or over-expression of genes in tumor with respect to normal tissues. The CNV data were collected using the Affymetrix SNP 6.0 platforms. Data went through segmentation analysis and were transformed into segment mean values, with amplified regions having positive values and deletions having negative values.

In principle, the proposed method can accommodate whole-genome data. However that would involve estimating a huge number of nonparametric effects and a giant coefficient matrix. In addition, as previously described, it is sensible to analyze different pathways separately. Specifically, we analyze the PCD (programmed cell death) pathway, which is well known for resistance to induction of apoptosis. The PCD pathway is related to the regulated cell suicide process where cells go through death to prevent themselves from proliferating or in response to certain signals to cells, such as stress or DNA damage. Genes within this pathway are identified using the annotation package in GSEA (http://www.broadinstitute.org/gsea). A total of 428 CNV and 426 GE measurements are available on 333 subjects. We remove the two CNVs that are not matched to GEs, resulting in 426 pairs of CNV and GE measurements. Under the proposed model, there are a total of 181,050 parametric parameters, 426 intercepts, and 426 nonparametric parameters.

Summary analysis results are presented in Table 3. The proposed method identifies 2260 parametric and 307 nonparametric effects. Each GE is found to be regulated by at least one CNV. Since a large number of effects are iden-tified, it is infeasible to present all results. For the identified parametric effects, we calculate the median as −0.025, and interquartile range as (−0.115, 0.099). For the identified nonlinear effects, we compute the normalized $\grave{}_1$ norms, which have median 1.029, and interquartile range (0.714, 1.668). In Figure 1, for four representative GEs, we show the estimated cis-acting CNV effects. For BCL2 and SNCA, we observe an overall decreasing and increasing trend, respectively. The “bumps” close to the boundaries need to be interpreted cautiously, as the numbers of observations are small in these regions. For IL2RA and TXNL1, the effects are mostly constants in the middle, but differences are observed for regions with high levels of deletions/amplifications. Such results suggest that the commonly adopted linear modeling may be insufficient. The deviation from linear can be caused by interactions, other regulating mech-anisms (microRNAs, methylation, etc.), and others. Mechanistic studies will be needed to fully validate and interpret the identified nonlinear trends.

To complement the above analysis, we implement a resampling-based method [7], use part of the data for iden-tification/estimation, make prediction for the rest, and compute the absolute prediction errors. As shown in Table 3, the proposed method has prediction performance slightly better than A_2 but much better than the other alternatives. It is noted that the non-robust methods have inferior prediction, which partly justifies the necessity of

robust analysis. Also using the resampling approach, we compute the probability of a specific effect being identified, which provides a way of assessing stability. For effects identified using full data, the mean probabilities (of being identified) are 59% (A_1), 45% (A_2), 46% (A_3), and 41% (A_4), respectively. The probabilities are lower for non-robust methods.

As a representative example, we take a closer look at gene BCL2, which is a confirmed biomarker for melanoma etiology and prognosis. In the process of apoptosis, the BCL2 protein plays an important role in inhibiting cell death and promoting cell survival. Elevated BCL2 gene expressions have been associated with the presence of ulceration and poorer survival. Using the proposed method, the cis-acting CNV is identified as having a nonlinear effect (Figure 1). The parametric effects identified using different methods are presented in Tables A.5 and A.6 (Online Supplement). Different methods generate significantly different findings. In Figure A.3 (Online Supplement), we show the network connection of CNVs identified using different methods. The proposed method identifies a tightly connected module composed of six CNVs: PPARD, RIPK1, SERPINB9, TNF, TXNDC5, and VEGFA. The corresponding genes are all located on chromosome 6p. The instability of chromosome 6 has been associated with melanoma progression. The proposed method also identifies two other connected CNVs: PMAIP1 and SERPINB2. The corresponding genes are located closely on the chromosome (18q21.33), with PMAP1 on18q21.32 and SERPINB2 on 18q21.3. Two isolated CNVs are also identified, with one located on chromosome 2 and the other on the X chromosome. The other robust methods also identify connected CNVs but not the six-CNV module. The non-robust methods do not identify connected CNVs.

In studies with a much smaller scale (with a few GEs and CNVs), the identified GE-CNV relationship can be functionally validated. The proposed analysis involves a large number of GEs and CNVs, and it may be unrealistic to validate all results. If there is an interest, subsets of the results can be examined in greater details. The improvement observed in simulation and improved prediction and stability provide support to the validity of analysis.

6. Discussion

In recent studies, multiple types of omics measurements have been collected, making it possible to study their regulation relationships. The regulation of GE by CNV is one of the first known regulating mechanisms and has attracted special attention. In this article, we have developed a new analysis approach for modeling the GE-CNV regulation. With the partially linear modeling, the proposed approach can more accurately describe the dominant cis-acting CNV effects. The proposed estimation also has notable advantages: the robust loss function accommodates long-tailed GE distributions and contamination; the simultaneous analysis ensures that all GEs are analyzed on the same ground; and the correlations among CNVs are effectively accommodated using a network structure and a shrinkage penalty. It is remarkable that with significant methodological advancements, the proposed method is still theoretically and computationally manageable. Under a variety of simulation settings, it significantly outperforms seven direct competitors. In the analysis of TCGA data, it leads to biologically plausible findings and improved prediction and stability.

The proposed approach is not limited to GE-CNV analysis. In omics studies, other regulation relationships (for example, proteins by GEs, GEs by methylation and microRNAs) are also of significant interest. In addition, data with both high-dimensional responses and high-dimensional covariates also arise in other fields. It is noted that with other data, there may not be a simple match between responses and covariates. More carefully examining the proposed approach (methodology, computation, and theory) suggests that there is actually no need for matching, and it can accommodate multiple nonlinear effects per response. Determining which covariate effects should be nonlinear is a "classic" statistical problem and has been examined in multiple studies. The proposed approach can be coupled with the techniques for determining nonlinear effects and have broader applications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Editor-in-Chief and reviewers for their careful review and insightful comments, which have led to a significant improvement of the article. This study was supported by the National Natural Science Foundation of China (11401561), National Bureau of Statistics of China (2016LD01), National Institutes of Health (CA204120, CA216017), Fundamental Research Funds for the Central Universities (20720171064, 20720181003), an Innovative Research Award from the Johnson Cancer Research Center at Kansas State University and a Kansas State University Faculty Enhancement Award.

References

[1]. Breheny P, Huang J, Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection, Ann. Appl. Statist 5 (2011) 232–253.

[2]. Buhlmann P, Van De Geer S, Statistics for High-dimensional Data: Methods, Theory and Applications, Springer, New York.

[3]. Fan J, Fan Y, Barut E, Adaptive robust variable selection, Ann. Statist 42 (2014) 324–351.

[4]. Fisher RA, Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, Biometrika 10 (1915) 507–521.

[5]. Friedman J, Hastie T, Hofling H, Tibshirani RJ, Pathwise coordinate optimization, Ann. Appl. Statist 1 (2007) 302–332.

[6]. Henrichsen CN, Chaignat E, Reymond A, Copy number variants, diseases and gene expression, Human Molecular Genetics 18 (2009) R1–R8. [PubMed: 19297395]

[7]. Huang J, Ma S, Variable selection in the accelerated failure time model via the bridge method, Lifetime Data Anal 16 (2010) 176–195. [PubMed: 20013308]

[8]. Huang J, Ma S, Li H, Zhang C, The sparse laplacian shrinkage estimator for high-dimensional regression, Ann. Statist 39 (2011) 2021– 2046.

[9]. Kim S, Sohn K-A, Xing EP, A multivariate regression approach to association analysis of a quantitative trait network, Bioinformatics 25 (2009) i204–i212. [PubMed: 19477989]

[10]. Knight K, Limiting distributions for l1 regression estimators under general conditions, Ann. Statist 26 (1998) 755–770.

[11]. Koenker R, Quantile Regression, Cambridge University Press, Cambridge, 2005.

[12]. Leday G, van der Vaart A, van Wieringen W, van de Wiel M, Modeling association between dna copy number and gene expression with constrained piecewise linear regression splines, Ann. Appl. Statist 7 (2013) 823–845.

[13]. Lian H, Liang H, Generalized additive partial linear models with high-dimensional covariates, Econometric Theory 29 (2013) 1136–1161.

[14]. Liu J, Huang J, Ma S, Incorporating network structure in integrative analysis of cancer prognosis data, Econometric Theory 37 (2013) 173–183.

[15]. Peng B, Wang L, An iterative coordinate descent algorithm for high–dimensional nonconvex penalized quantile regression, J. Comput. Graph. Stat 24 (2015) 676–694.

[16]. Peng J, Zhu J, Bergamaschi A, Han W, Noh DY, Pollack J, Wang P, Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer, Ann. Appl. Statist 4 (2010) 53–77.

[17]. Schafer M, Schwender H, Merk S, Haferlach C, Ickstadt K, Dugas M, Integrated analysis of copy number alterations and gene expression: A bivariate assessment of equally directed abnormalities, Bioinformatics 25 (2010) 3228–3235.

[18]. Schumaker L, Spline Functions: Basic Theory, Cambridge University Press, 2007.

[19]. Shi X, Liu J, Huang J, Zhou Y, Shia B, Ma S, Integrative analysis of high-throughput cancer studies with contrasted penalization, Genetic Epidemiology 38 (2014) 144–151. [PubMed: 24395534]

[20]. Shi X, Zhao Q, Huang J, Xie Y, Ma S, Deciphering the associations between gene expression and copy number alteration using a sparse double laplacian shrinkage approach, Bioinformatics 31 (2015) 3977–3983. [PubMed: 26342102]

[21]. Speed T, Statistical Analysis of Gene Expression Microarray Data, CRC Press, London, 2003.

[22]. Tseng P, Convergence of a block coordinate descent method for nondifferentiable minimization, J. Optim. Theory Appl 109 (2001) 475–494.

[23]. Wang H, Zhu Z, Zhou J, Quantile regression in partially linear varying coefficient models, Ann. Statist (2009) 3841–3866.

[24]. Wang L, Wu Y, Li R, Quantile regression for analyzing heterogeneity in ultra-high dimension, J. Amer. Statist. Assoc 107 (2001) 214–222.

[25]. Wu C, Ma S, A selective review of robust variable selection with applications in bioinformatics, Briefings in Bioinformatics 16 (2015) 873–883. [PubMed: 25479793]

[26]. Xiong L, Kuan P, Tian J, Keles S, Wang S, Multivariate boosting for integrative analysis of high-dimensional cancer genomic data, Cancer Inform 13 (Suppl 7) (2014) 123–131. [PubMed: 25520552]

[27]. Xue L, Qu A, Variable selection in high-dimensional varying-coefficient models with global optimality, J. Machine Learning Res 13 (2012) 1973–1998.

[28]. Zhang B, Horvath S, A general framework for weighted gene co-expression network analysis, Statistical Appl. Genetics Molecular Biology 4 (2005) 123–131. doi:10.2202/1544--6115.1128.

[29]. Zhang C, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist 38 (2010) 894–942.

[30]. Zhou Y, Wang P, Wang X, Zhu J, Song P, Sparse multivariate factor analysis regression models and its applications to integrative genomics analysis, Genet Epidemiol 41 (2017) 70–80. [PubMed: 27862229]

[31]. Zhu R, Zhao Q, Zhao H, Ma S, Integrating multidimensional omics data for cancer outcome, Biostatistics 41 (2016) 70–80.

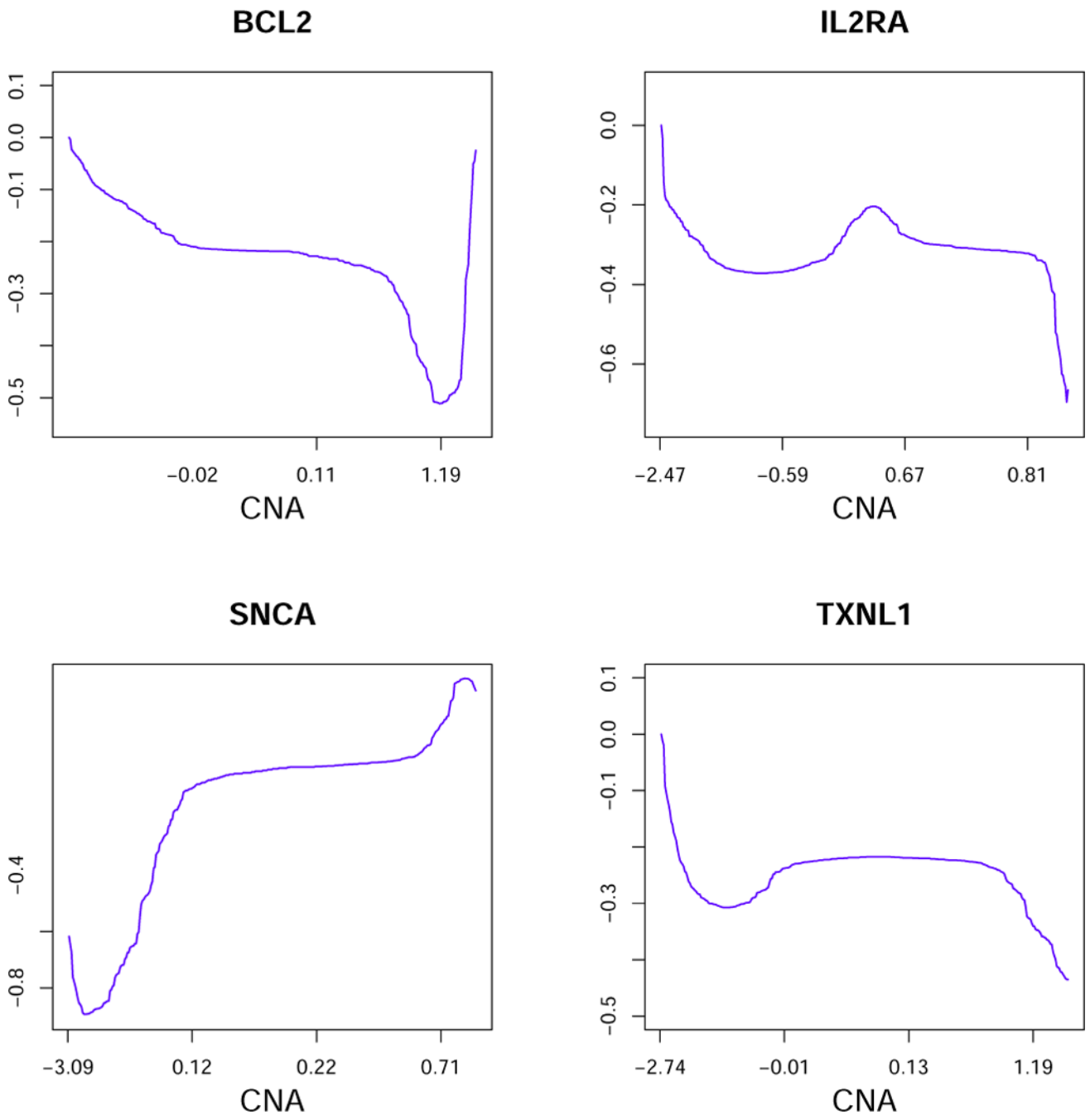

Figure 1:
Analysis of the TCGA data: estimated cis-acting CNV effects for four genes.

Table 1:

Simulation: summary based on 100 replicates. In each cell, mean (sd). TP1/FP1 and TP2/FP2: number of true/false positives for parametric and nonparametric effects.

	ρ		TP1	FP1	TP2	FP2
Error 1	0.1	A_1	283.1 (2.9)	0.0 (0.0)	90.0 (0.0)	0.1 (0.2)
		A_2	283.1 (2.7)	0.0 (0.0)	90.0 (0.0)	0.1 (0.2)
		A3	246.3 (9.6)	0.0 (0.0)	81.7 (2.6)	0.0 (0.0)
		A4	246.8 (9.5)	0.0 (0.0)	81.7 (2.6)	0.0 (0.0)
	0.5	A_1	287.6 (0.7)	2.5 (1.6)	90.0 (0.0)	0.0 (0.0)
		A2	287.6 (0.6)	2.4 (1.7)	90.0 (0.0)	0.0 (0.0)
		A3	232.7 (3.4)	1.5 (1.0)	50.9 (0.8)	2.3 (1.3)
		A_4	232.2 (3.7)	1.6 (0.9)	50.8 (0.8)	2.4 (1.3)
	0.9	A_1	285.0 (2.4)	147.6 (14.3)	90.0 (0.0)	0.1 (0.3)
		A2	286.2 (1.5)	179.9 (16.1)	90.0 (0.0)	0.1 (0.3)
		A_3	219.8 (3.1)	60.6 (9.7)	49.6 (0.7)	8.6 (0.6)
		A_4	222.1 (4.4)	64.6 (8.9)	49.4 (0.9)	8.5 (0.6)
Error 2	0.1	A_1	278.1 (4.6)	0.1(0.3)	89.7 (0.5)	0.2 (0.4)
		A_2	278.4 (4.4)	0.1 (0.3)	89.7 (0.5)	0.3 (0.5)
		A_3	227.9 (9.3)	0.1 (0.3)	78.9 (2.9)	0.0 (0.0)
		A_4	228.0 (8.8)	0.1 (0.3)	78.9 (3.0)	0.0 (0.0)
	0.5	A_1	286.8 (1.4)	3.1 (1.7)	89.5 (0.7)	0.2 (0.5)
		A_2	286.9 (1.5)	3.0 (1.5)	89.5 (0.8)	0.2 (0.5)
		A3	234.2 (3.6)	1.7 (1.1)	51.2 (1.0)	2.5 (1.3)
		A_4	233.2 (3.3)	1.6 (1.2)	51.2 (1.0)	2.6 (1.4)
	0.9	A_1	283.3 (2.8)	162.0 (12.0)	89.7 (0.5)	0.3 (0.6)
		A_2	284.8 (2.5)	190.3 (12.4)	89.7 (0.6)	0.3 (0.5)
		A_3	222.3 (4.1)	81.7 (8.6)	49.4 (0.7)	9.0 (0.5)
		A4	223.6 (5.0)	85.9 (12.5)	49.4 (0.9)	8.8 (0.6)
Error 3	0.1	A_1	279.2 (4.7)	0.0 (0.0)	89.4 (0.9)	0.2 (0.5)
		A_2	279.0 (4.7)	0.0 (0.0)	89.4 (0.9)	0.2 (0.5)
		A3	218.8 (8.9)	0.0 (0.0)	76.3 (2.9)	0.0 (0.0)
		A_4	218.7 (8.4)	0.1 (0.2)	76.5 (2.8)	0.0 (0.0)
	0.5	A_1	286.5 (1.4)	3.9 (1.3)	89.3 (0.9)	0.3 (0.6)
		A_2	286.7 (1.2)	4.2 (1.5)	89.4 (0.9)	0.3 (0.6)
		A_3	233.7 (3.4)	2.0 (1.5)	50.8 (0.9)	2.9 (1.3)
		A4	233.3 (3.5)	2.0 (1.3)	51.0 (0.9)	2.9 (1.2)
	0.9	A_1	282.4 (2.9)	162.9 (14.6)	89.6 (0.9)	0.4 (0.7)
		A2	283.8 (2.2)	202.1 (19.1)	89.6 (0.9)	0.5 (0.6)
		A_3	223.2 (3.9)	79.4 (13.2)	49.6 (0.5)	9.1 (0.6)
		A_4	221.6 (3.5)	74.1 (9.7)	49.6 (0.7)	9.0 (0.6)

Table 2:

Simulation: summary based on 100 replicates. In each cell, mean (sd). TP1/FP1 and TP2/FP2: number of true/false positives for parametric and nonparametric effects.

	ρ		TP1	FP1	TP2	FP2
Error 1	0.1	A_5	288 (0)	0.1 (0.2)	90.0 (0.0)	0.0 (0.0)
		A_6	288 (0)	0.1 (0.2)	90.0 (0.0)	0.0 (0.0)
		A7	288 (0)	1.6 (1.2)	90.0 (0.0)	0.0 (0.0)
		A8	288 (0)	1.6 (1.2)	90.0 (0.0)	0.0 (0.0)
	0.5	A_5	283.6 (2.1)	2.4 (1.5)	90.0 (0.0)	0.0 (0.0)
		A6	283.2 (1.9)	2.4 (1.5)	90.0 (0.0)	0.0 (0.0)
		A7	287.2 (0.8)	2.4 (1.6)	88.8 (1.1)	0.0 (0.0)
		A_8	287.2 (0.9)	2.4 (1.6)	88.8 (1.0)	0.0 (0.0)
	0.9	A_5	247.6 (3.8)	35.0 (4.6)	89.9 (0.3)	0.0 (0.0)
		A_6	171.6 (4.9)	16.9 (2.3)	89.9 (0.3)	0.0 (0.0)
		A_7	262.7 (6.9)	41.0 (11.6)	50.2 (1.3)	0.3 (0.5)
		A8	204.7 (9.4)	6.5 (2.6)	43.0 (2.5)	0.0 (0.0)
Error 2	0.1	A_5	287.0 (1.7)	196.5 (157.9)	89.9 (0.4)	0.4 (0.7)
		A_6	287.1 (1.3)	198.1 (158.0)	89.9 (0.3)	0.4 (0.7)
		A_7	287.3 (1.4)	201.5 (155.4)	89.5 (0.7)	0.3 (0.6)
		A_8	287.3 (1.4)	199.0 (155.4)	89.5 (0.7)	0.3 (0.6)
	0.5	A_5	274.5 (11.1)	171.7 (85.0)	87.9 (2.4)	0.6 (0.7)
		A_6	273.8 (11.5)	173.8 (86.9)	87.8 (2.6)	0.6 (0.7)
		A7	274.6 (11.0)	176.0 (88.4)	78.3 (9.5)	0.33 (0.6)
		A_8	273.5 (12.0)	174.9 (89.9)	78.4 (9.1)	0.37 (0.6)
	0.9	A_5	242.1 (17.2)	182.6 (104.4)	88.6 (5.4)	0.9 (1.1)
		A_6	158.8 (23.4)	159.6 (115.3)	80.3 (12.3)	0.7 (0.9)
		A_7	227.6 (31.8)	160.1 (83.6)	45.4 (4.9)	0.6 (0.7)
		A8	159.7 (39.8)	149.1 (116.0)	35.9 (7.7)	0.3 (0.6)
Error 3	0.1	A_5	286.5 (1.9)	262.7 (133.7)	89.5 (0.7)	1.1 (0.8)
		A_6	286.7 (1.7)	265.4 (141.7)	89.4 (0.9)	1.0 (0.8)
		A7	285.5 (1.9)	262.8 (134.0)	89.5 (0.7)	1.1 (0.8)
		A_8	286.7 (1.7)	263.8 (141.0)	89.4 (0.9)	1.0 (0.8)
	0.5	A_5	269.8 (11.2)	281.7 (126.5)	87.3 (3.0)	1.1 (1.0)
		A_6	269.0 (11.5)	288.6 (144.6)	87.4 (2.7)	1.1 (1.0)
		A_7	268.2 (9.9)	279.3 (147.5)	73.6 (2.6)	0.5 (0.7)
		A8	264.1 (11.8)	268.9 (135.0)	72.1 (8.1)	0.6 (0.7)
	0.9	A_5	225.7 (21.0)	279.8 (154.9)	87.1 (6.7)	1.7 (1.4)
		A6	145.1 (20.4)	253.6 (194.2)	74.8 (12.6)	1.4 (1.4)
		A_7	211.3 (33.3)	250.8 (142.9)	43.3 (4.9)	1.2 (0.9)
		A_8	139.8 (31.0)	233.6 (171.5)	32.4 (6.3)	1.0 (1.0)

Table 3:

Analysis of TCGA data. Numbers of identified nonzero parametric (P) and nonparametric (NP) effects, and sum of absolute prediction errors (PE).

	P	NP	PE
A_1	2260	307	282.02
A_2	2088	299	293.16
A3	2735		307.67
A4	2748		308.12
A_5	2491	70	549.99
A6	2648	76	538.14
A7	2784		438.62
A_8	2293		436.58