Published in final edited form as:

Genet Epidemiol. 2017 July ; 41(5): 437–454. doi:10.1002/gepi.22047.

Inferring gene regulatory relationships with a high-dimensional robust approach

Yangguang Zang[1,2], **Qing Zhao**[3], **Qingzhao Zhang**[4], **Yang Li**[5], **Sanguo Zhang**[1], and **Shuangge Ma**[2,4]

[1]School of Mathematical Sciences, University of Chinese Academy of Sciences

[2]Department of Biostatistics, Yale University

[3]Merck Research Lab, Rahway, NJ

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

[5]School of Statistics, Remin University of China

Abstract

Gene expression (GE) levels have important biological and clinical implications. They are regulated by copy number alterations (CNAs). Modeling the regulatory relationships between GEs and CNAs facilitates understanding disease biology and can also have value in translational medicine. The expression level of a gene can be regulated by its *cis*-acting as well as *trans*-acting CNAs, and the set of *trans*-acting CNAs is usually not known, which poses a high-dimensional selection and estimation problem. Most of the existing studies share a common limitation in that they cannot accommodate long-tailed distributions or contamination of GE data. In this study, we develop a high-dimensional robust regression approach to infer the regulatory relationships between GEs and CNAs. A high-dimensional regression model is used to accommodate the effects of both *cis*-acting and *trans*-acting CNAs. A DPD (density power divergence) loss function is used to accommodate long-tailed GE distributions and contamination. Penalization is adopted for regularized estimation and selection of relevant CNAs. The proposed approach is effectively realized using a coordinate descent algorithm. Simulation shows that it has competitive performance compared to the nonrobust benchmark and the robust LAD (least absolute deviation) approach. We analyze TCGA (The Cancer Genome Atlas) data on cutaneous melanoma and study GE-CNA regulations in the RAP (regulation of apoptosis) pathway, which further demonstrates satisfactory performance of the proposed approach.

Keywords

Gene regulatory relationship; Robustness; High-dimensional regression

1 Introduction

Gene expressions (GEs) have important biological implications, and the analysis of GE data has led to important findings with basic, translational, and clinical value for many complex diseases (Sparano et al., 2015; Deng et al., 2006). GE levels are regulated by copy number alterations (CNAs), possibly along with other mechanisms. Modeling the regulatory relationships between GEs and CNAs has important implications: it can lead to a better understanding of disease etiology, assist building disease outcome models with translational value (Marbach et al., 2016), and play an important role in drug discovery.

Identifying which CNAs regulate the levels of GEs and in what way has been studied but remains a challenging problem (Wang et al., 2011; Yuan et al., 2012). For the expression level of a specific gene, some studies focus on the "local" effect and analyze only the *cis*-acting CNA. Such studies are limited as GE levels are regulated by both *cis*-acting and *trans*-acting CNAs. Some studies analyze the *cis*-acting CNA as well as a small number of pre-selected *trans*-acting CNAs (Blackburn et al., 2015). However, they may not be broadly applicable, as for many genes, the sets of relevant *trans*-acting CNAs are not accurately known (Henrichsen et al., 2009). In recent studies, more effective approaches have been developed (Shi et al., 2015), which adopt high-dimensional regression techniques to accommodate the effects of a large number of CNAs and data-dependently identify the relevant ones.

A common limitation shared by most of the existing studies is that they adopt nonrobust estimation. For example, in regression analysis, the ordinary least squared loss function has been commonly adopted (Yuan et al., 2012). GE data may have long-tailed distributions or be contaminated. In Figure 1, we show examples of GE distributions from the data analyzed in this article. The long-tails and deviation from normality are clearly seen. The long tails (especially extremely high GE levels) may happen for multiple biological reasons. Contamination may happen because of technical reasons, as has been noted in published studies (Osborne and Overbay, 2004; Shieh and Hung, 2009). In addition, complex diseases may have multiple subtypes, which have distinct underlying biological processes. In data analysis, the small subtypes can be viewed as "contamination" for the largest subtype. In statistical analysis with low-dimensional data, it has been shown that with long-tailed distributions and contamination, nonrobust estimation, for example the ordinary least squares, can lead to severely biased estimation and wrong conclusions on the importance of effects. In the literature, one way of accommodating long-tailed distributions is transformation. However, it may not be suitable in the current context: it may not be possible to find a transformation that fits all GEs, and applying different transformations to different GEs leads to a lack of comparability and interpretability. With high-dimensional genetic data, our literature review suggests that robust approaches are still limited. This is especially true for the analysis of GE-CNA relationship, possibly because of the additional complexity as both sides of the relationship are high-dimensional.

In this article, we study the regulatory relationships between GEs and CNAs. This study has the following notable features, making it warranted beyond the existing studies. First, a high-dimensional regression model is adopted to accommodate the effects of both *cis*-acting CNAs as well as a large number of candidate *trans*-acting CNAs. Second, a DPD (density power divergence) loss function is adopted to accommodate long-tailed distributions and contamination. A robust loss can be more appropriate than the nonrobust ones for data shown in Figure 1 and those alike. Compared to some alternative robust approaches, the DPD has several notable advantages but has been much less investigated, especially under high-dimensional settings. It is thus of interest to develop it for the present problem. Third, penalization is adopted in estimation and can easily accommodate high data dimensionality and select relevant CNAs. Fourth, numerical studies, including simulation and data analysis, provide convincing evidences to the practical advantage of the proposed approach. Overall, this study provides a practically useful tool for inferring the GE-CNA regulatory relationships.

2 Methods

2.1 Data and model settings

Consider a dataset with n independent subject. For each subject, data are available for q GEs and p CNAs. For subject i (= 1, …, n), let $Y_i = (Y_{i1}, \cdots, Y_{iq})^\top$ denote the vector of GEs and $X_i = (X_{i1}, \cdots, X_{ip})^\top$ denote the vector of CNAs. Further denote $Y = (Y_1, \cdots, Y_n)^\top \in \mathbb{R}^{n\times q}$ and $X = (X_1, \cdots, X_n)^\top \in \mathbb{R}^{n\times p}$ as the data matrices for GEs and CNAs, respectively. With $Y_{,j} = (Y_{1j}, \cdots, Y_{nj})^\top$ and $X_{,j} = (X_{1j}, \cdots, X_{nj})^\top$, we can also rewrite as $X = (X_{,1}, \cdots, X_{,p})$ and $Y = (Y_{,1}, \cdots, Y_{,q})$.

For accommodating long-tailed distributions and contamination, we adopt a mixture model framework, which has been a popular choice in low-dimensional data analysis. The underlying structure is that the majority of the subjects are “normal” and satisfy a certain regression model/distribution. Denote this set of subjects as A_0 (note that this set is not known prior to analysis). In addition, a small subset of the subjects are “abnormal” with contaminated measurements or corresponding to different subtypes that satisfy a different model/distribution. Denote this set of subjects as A_1. In this study, *the analysis goal is to infer the regulatory relationships between GEs and CNAs for the majority of the subjects (A_0), while properly accommodating the “abnormal” ones (A_1).*

For a subject, say subject i, in A_0, consider the model

$$Y_{\mathrm{ij}}=\sum_{l=1}^{p} X_{\mathrm{il}}\beta_{\mathrm{ij}}+\varepsilon_{\mathrm{ij}},\quad j=1,\cdots,q. \qquad (1)$$

Assume that $\{\varepsilon_{ij} : j = 1, \ldots, q\}$ are independent and distributed as $N(0, \sigma_j^2)$. Denote $\beta_{,j} = (\beta_{1j}, \cdots, \beta_{pj})^\top$ and $\beta = (\beta_{,1}, \cdots, \beta_{,q}) \in \mathbb{R}^{p\times q}$. Inferring the regulatory relationships amounts to estimating the coefficient matrix β. Assume that the data have been properly centralized (so that the intercepts are zero) and standardized. Linear regression has been adopted for modeling the GE-CNA relationship in quite a few recent studies and shown to be effective. See for example Shi et al. (2015) and references therein. Although nonlinear CNA effects are possible, extensive nonlinear modeling is computationally prohibitive and may lead to unreliable estimation.

The proposed modeling framework has notable advantages. A significant one is that the regression model/distribution for subjects in A_1 does not need to be specified. This flexibility is much desired – as in practice one does not know the mechanism of abnormality or contamination – but not shared by many of the existing approaches. For example, the popular Gaussian mixture model techniques assume that all distribution components are normal. In addition, the proposed approach does not need to specify the percentage (or even existence) of abnormality or contamination. The high-dimensional model can accommodate a large number of candidate CNA effects and allow for data-dependently identifying the relevant ones. Overall, the proposed approach can be more data-driven and flexible.

2.2 The DPD-Lasso estimation

The DPD method is first proposed in Basu et al. (1998) for the robust estimation of a single distribution parameter. It is later extended to robust regression analysis with low-dimensional data (Durio and Isaia, 2011; Ghosh et al., 2013; Fujisawa and Eguchi, 2006). These studies have shown that the DPD approach has multiple statistical and numerical advantages over the nonrobust and robust alternatives. For the integrity of this article, below we briefly describe the general DPD strategy and refer to published studies for more details.

For two density functions g and f, the DPD measure $d_\alpha(g, f)$ is defined as

$$d_\alpha = \int \left\{ f^{1+\alpha} - \left(1+\frac{1}{\alpha}\right) f^\alpha g + \frac{1}{\alpha} g^{1+\alpha} \right\}, \alpha>0, \tag{2}$$

$$d_0 = \int g \ln \left(\frac{g}{f}\right). \tag{3}$$

Note that d_0 is the limit of d_α as $\alpha \rightarrow 0$ and also a version of the Kullback-Leibler divergence. The parameter α balances between efficiency and robustness, with a smaller α value corresponding to more efficient but less robust estimation (Basu et al., 1998).

For subjects in A_0, denote the true unknown density function of the jth GE as g_j. In estimation, our goal is to estimate (or "approximate") g_j with f_{θ_j}, where f_{θ_j} represents a family of density functions indexed by parameter θ_j. Under the DPD framework, θ_j is estimated by minimizing

$$\frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n} d_\alpha\left(g_j(\mathrm{Y_{ij}}), f_{\theta_j}(\mathrm{Y_{ij}})\right). \tag{4}$$

In practice, g_j is unknown and replaced with the empirical density function $\hat{g}_j(u) = \frac{1}{n}\sum_{i=1}^{n} \delta(u - \mathrm{Y_{ij}})$, where $\delta(x)$ is the Dirac delta function. Note that the third term in equation (2) does not depend on θ_j and can be ignored. Therefore, the DPD loss function is

$$\frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n}\left[\int f_{\theta_j}^{1+\alpha}(u_j)\mathrm{d}u_j - \left(1+\frac{1}{\alpha}\right) f_{\theta_j}^{\alpha}(\mathrm{Y_{ij}})\right] = \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n} \mathrm{V_{ij}}. \tag{5}$$

Note that, when $\alpha = 0$, the DPD loss becomes $\frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n}[-\ln(f_{\theta_j}(Y_{\mathrm{ij}}))]$, which leads to the maximum likelihood estimate.

Under the assumed high-dimensional linear regression model, for subject i which belongs to A_0, $Y_{\mathrm{ij}} \sim N(X_i^\top \beta_{\cdot j}, \sigma_j^2)$. With $\theta_j = (\beta_{\cdot j}^\top, \sigma_j^2)^\top$, we have

$$V_{\mathrm{ij}} = \frac{1}{(2\pi)^{\alpha/2}\sigma_j^\alpha\sqrt{1+\alpha}} - \frac{1+\alpha}{\alpha}\frac{1}{(2\pi)^{\alpha/2}\sigma_j^\alpha}e^{-\alpha(Y_{\mathrm{ij}}-X_i^\top\beta_{\cdot j})^2/2\sigma_j^2}. \quad (6)$$

For data analyzed in this article and those alike, the dimensionality of covariates (CNAs) can be high compared to sample size. With a limited sample size and potential long-tailed distributions/contamination, regularized estimation is desired. In addition, out of a large number of CNAs measured, only a few are expected to regulate the expression level of a specific gene. With such considerations, we propose to apply penalization, which has also been adopted in recent studies (Shi et al., 2015) and can be preferred over other regularization techniques because of better statistical and numerical properties.

Denote $\theta = (\theta_1^\top, \ldots, \theta_q^\top)^\top$. We propose the DPD-Lasso loss function as

$$L(\theta) = \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n}\left[\frac{1}{(2\pi)^{\alpha/2}\sigma_j^\alpha\sqrt{1+\alpha}} - \frac{1+\alpha}{\alpha}\frac{1}{(2\pi)^{\alpha/2}\sigma_j^\alpha}e^{-\alpha(Y_{\mathrm{ij}}-X_i^\top\beta_{\cdot j})^2/2\sigma_j^2}\right] + \sum_{j=1}^{q}\lambda\|\beta_{\cdot j}\|_1, \quad (7)$$

where $\|\beta_{\cdot j}\|_1 = \sum_{k=1}^{p}|\beta_{\mathrm{kj}}$, and $\lambda > 0$ is the data-dependent tuning parameter. The DPD-Lasso estimate is defined as the minimizer of $L(\theta)$. A nonzero component of this estimate represents a regulatory relationship between the corresponding CNA and GE. In (7), the Lasso penalty is adopted for its computational simplicity and satisfactory numerical performance and can be replaced with other penalties such as SCAD or MCP. Note that the proposed approach is different from analyzing each GE separately. Specifically, it applies the same λ to all GEs to ensure the same level of penalty and comparability, which is desired as no gene is "preferred" over the others. If each GE is analyzed separately, the tuning parameters are likely to differ across genes, resulting in a lack of comparability.

Although the DPD loss has been previously adopted in regression analysis, the existing studies are limited to low-dimensional settings. This study is the first to adopt it under high-dimensional settings. The analysis is further challenged by the high dimensionality of the response variable and application of penalized estimation. In addition, to the best of our knowledge, the DPD approach has never been applied in the context of GE-CNA analysis.

2.3 Computation

For optimizing the penalized loss function (7), we propose an iterative coordinate descent (CD) algorithm. Consider optimization with rest to β_{kj}, with the rest of the regression parameters fixed. Let Δ_{kj} represent the partial derivative with respect to β_{kj}. We have

$$\Delta_{\mathrm{kj}} L(\theta) = -\sum_{i=1}^{n}\left(\frac{1+\alpha}{(2\pi)^{\alpha/2}\sigma_j^{\alpha+2}} e^{-\frac{\alpha(Y_{\mathrm{ij}}-X_i^{\top}\beta_{,j})^2}{2\sigma_j^2}} (Y_{\mathrm{ij}}-X_i^{\top}\beta_{,j})X_{\mathrm{ik}}\right) + \lambda\frac{\beta_{\mathrm{kj}}}{|\beta_{\mathrm{kj}}|},$$

and the partial derivative with respect to σ_j^2 is

$$-\sum_{i=1}^{n}\frac{1}{(2\pi)^{\alpha/2}}\left[\frac{\alpha}{2\sigma_j^{\alpha+2}\sqrt{1+\alpha}} - \frac{1+\alpha}{2\sigma_j^{\alpha+2}} e^{-\frac{\alpha(Y_i-X_i^{\top}\beta_{,j})^2}{2\sigma_j^2}}\left\{1-\frac{(Y_{\mathrm{ij}}-X_i^{\top}\beta_{,j})^2}{\sigma_j^2}\right\}\right].$$

Set the partial derivatives equal to zero, and we get the estimating equations

$$-\sum_{i=1}^{n} e^{-\frac{\alpha(Y_{\mathrm{ij}}-X_i^{\top}\beta_{,j})^2}{2\sigma_j^2}} (Y_{\mathrm{ij}}-X_i^{\top}\beta_{,j})X_{\mathrm{ik}} + \tilde{\lambda}\frac{\beta_{\mathrm{kj}}}{|\beta_{\mathrm{kj}}|} = 0, \quad (8)$$

$$\sum_{i=1}^{n}\left[1-\frac{(Y_{\mathrm{ij}}-X_i^{\top}\beta_{,j})^2}{\sigma_j^2}\right] e^{-\frac{\alpha(Y_{\mathrm{ij}}-X_i^{\top}\beta_{,j})^2}{2\sigma_j^2}} = \frac{\alpha}{(1+\alpha)^{\frac{3}{2}}}; \quad (9)$$

where $\tilde{\lambda} = \frac{\lambda(2\pi)^{\alpha/2}\sigma_j^{\alpha+2}}{1+\alpha}$. We propose solving (8) and (9) iteratively to obtain the estimates of $\beta_{,j}$ and σ_j^2. More specifically, with σ_j^2 fixed, solving for a single β_{kj} poses a weighted Lasso-type problem. With $\beta_{,j}$ fixed, we estimate σ_j^2 using a bisection method.

Denote $\beta_{,j}^{(s)}$ and $\sigma_j^{2(s)}$ as the estimates of $\beta_{,j}$ and $\sigma_{,j}^2$ in the sth step. With fixed λ and α, the overall algorithm is described in Algorithm 1.

Algorithm 1

```
Initialize $s = 0$. For $j = 1, \cdots, q$, compute the initial estimate $\beta_{,j}^{(s)}$ (which is the ridge estimate in our numerical study).
With $\beta_{,j}$ fixed as $\beta_{,j}^{(s)}$, estimate $\sigma_j^{2(s)}$ by solving equation (9).
 repeat
   $s = s + 1$;
```

```
  for $j = 1, \cdots, q$ do
    Compute weights $w_{\mathrm{ij}}^{(s)} = \frac{e^{-\frac{\alpha(Y_{\mathrm{ij}} - X_i^\top \beta_{,j}^{(s-1)})^2}{2\sigma_j^2}}}{2(2\pi)^{\alpha/2}\sigma_j^{\alpha+2}/(1+\alpha)}$.
    Calculate $\beta_{,j}^{(s)}$, where the component $\beta_{\mathrm{kj}}^{(s)}$ can be obtained by solving
      $\beta_{kj}^{(s)} = \arg\min_{\beta_{kj}} \sum_{i=1}^{n} \left( w_{ij}^{(s)} (Y_{ij} - X_i^\top \beta_{,j})^2 + \lambda |\beta_{kj}| \right)$.
    Calculate $\sigma_j^{2(s)}$ by solving the following equation with the bisection method
      $\sum_{i=1}^{n} \left[ 1 - \frac{(Y_{ij} - X_i^\top \beta_{,j}^{(s)})^2}{\sigma_j^2} \right] e^{-\frac{\alpha(Y_{ij} - X_i^\top \beta_{,j}^{(s)})^2}{2\sigma_j^2}} = \frac{\alpha}{(1+\alpha)^{\frac{3}{2}}}$.
  end for
until the Frobenius norm of the difference between two consecutive $\beta$ estimates is less than a predefined threshold ($10^{-3}$ in our numerical study).
return the estimate of $\beta$ at convergence.
```

The proposed penalized loss function has a complex form. Our literature search does not lead to any simple technique that can be directly applied to establish the convergence property. For all of our simulated and real datasets, convergence is successfully achieved within 30 overall iterations (mostly within 10 iterations). We defer theoretical investigation on convergence to future study. The proposed approach involves α and λ. λ has the same implications as with other penalization methods. The definition of DPD and published studies under low-dimensional settings suggest that α balances between robustness and efficiency. To better comprehend its impact, in our simulation, we consider a sequence of α values as well as data-dependently selected α. To facilitate future data analysis, we have developed an R program which is available at https://github.com/shuanggema/mdpd.

2.4 *Ad hoc* identification of set A_1

After applying the approach described above and conducting parameter estimation for the "normal" subjects, one possible followup analysis is to discriminate subjects in A_0 from those in A_1. This analysis may potentially provide useful information. If A_1 is not empty, then it may suggest that there exist distinct biological processes that may define clinically meaningful subgroups and demand further investigation (Gosh, 2013). Comparing A_1 against A_0 may reveal the unique features of these subgroups. In addition, analysis can be re-done on A_0, which is composed of more homogeneous subjects.

Identifying A_1 can be more complicated than that in the existing studies, as the response variable (GEs) is high-dimensional. We propose the following calculation of p-value for subject i (= 1, …, n) to belong to A_0:

$$p_{\mathrm{ij}}=\frac{1}{n}\sum_{l=1}^{n}\mathbf{1}_{\{(Y_{\mathrm{lj}}-X_l^\top\beta_{,j})^2>(Y_{\mathrm{ij}}-X_i^\top\beta_{,j})^2\}},$$

$$p_i=-2\sum_{j=1}^{q}\ln(p_{\mathrm{ij}})\sim\chi^2_{2k}.$$

Note that the estimate of β needs to be used in calculating p_{ij}. Subject i is then classified as in A_1 if $p_i \le 0.05$. In this calculation, p_{ij}, the p-value for subject i and GE j, is calculated using a nonparametric method. The q p-values for subject i are then combined into one using the Fisher's method. Note that this analysis, although may be informative, is not necessary. In the literature, there are several methods that can detect outliers, and they can be potentially extended to the present analysis. We adopt the above *ad hoc* approach because of its simplicity and satisfactory performance in simulation.

3 Simulation

Simulation is conducted to evaluate performance of the proposed approach and compare against alternatives. Motivated by the TCGA (The Cancer Genome Atlas) data (analyzed in the next section) and those alike, we generate CNA measurements from a multivariate normal distribution with marginal means zero and variances one. Similar procedures have been adopted in Shi et al. (2015) and references therein. Use ρ_{jk} to denote the correlation coefficient between CNAs j and k. Consider the following correlation structures: (i) independent, where $\rho_{jk} = 0$ if $j \ne k$, (ii) AR (auto-regressive), where $\rho_{jk} = 0.4^{|j-k|}$, and (iii) banded, where $\rho_{jk} = 0.4$ if $|j - k| = 1$ and $\rho_{jk} = 0$ otherwise. The AR and banded correlations have been popular in high-dimensional simulations. The considered correlation levels are reasonable, as in for example the TCGA data, the (absolute values of) observed correlations are mostly below 0.4 (summary available from the authors). Set $n = 200$, $q = 100$, and $p = 100$, which mimics the analysis of data of a pathway as in the next section. Note that although p and q may not seem "dramatic", the number of unknown parameters is actually much larger than n. In addition, consider the following settings.

Simulation I

First consider subjects in A_0. For each GE, the following CNAs have nonzero effects: the *cis*-acting CNA, the first four *trans*-acting CNAs, and then fifteen randomly selected ones. Thus, a total of twenty CNAs are associated with each GE. Note that the number of nonzero CNA effects may be larger than that in practical data analysis, making the analysis more challenging. The nonzero regression coefficients are randomly generated from *Unif*(0.4, 1.2). The random errors have a standard normal distribution. The GE levels are computed from the linear regression models.

Denote $|A_1|$ as the size of A_1. Consider the following scenarios for subjects in A_1:

(S1) $|A_1|=0$. That is, there is no long-tailed distribution/contamination. This scenario favors nonrobust analysis.

(S2) $|A_1| = 0.15 \times n$. For each GE, the following CNAs have nonzero effects: the *cis*-acting CNA, *trans*-acting CNAs #6–#10, and then fifteen randomly selected ones. The rest of the settings are the same as for subjects in A_0.

(S3) The same as (S2), with the exception that $|A_1| = 0.3 \times n$.

(S4) $|A_1| = 0.15 \times n$. The random errors have a $N(0, 49)$ distribution. The rest of the settings are the same as for subjects in A_0.

(S5) The same as (S4), with the exception that $|A_1| = 0.3 \times n$.

(S6) The same as (S4), with the random errors having a Cauchy distribution.

(S7) The same as (S6), with the exception that $|A_1| = 0.3 \times n$.

Under (S2) and (S3), there exists a subset in which subjects have a different GE regression model. That is, the mean structures for subjects in A_0 and A_1 are different. Under (S4)–(S7), all subjects have the same GE regression model, with those in A_1 having larger variances. That is, all subjects have the same mean structure but different variance structures. This corresponds to the "classic" setting of contamination. To test the effectiveness in robustness, we intentionally set the percentages of contamination to be higher than those in many of the existing studies.

Simulation II

For subjects in A_0, the regression coefficient matrix β has a block-diagonal structure. This is motivated by the consideration that genes close to each other on the chromosome are often coordinated. A similar structure has been considered in Shi et al. (2015) and others. For subjects in A_1, the settings are mostly the same as under Simulation I. The difference is that, under (S2) and (S3), twenty randomly selected CNAs have nonzero effects, which can test performance of the proposed approach under different correlations of CNA signals.

Lower signal levels

We also test performance of the proposed approach under lower signal levels. Specifically, for both Simulation I and II and the AR correlation structure, we keep the other settings unchanged and generate the nonzero regression coefficients from $Unif[0.1, 0.5]$.

Different signal signs

Under the above simulation settings, all nonzero coefficients are positive. For both Simulation I and II and the AR correlation structure, we generate half of the nonzero regression coefficients from $Unif[0.1, 0.5]$ and the other half from $Unif[-0.5, -0.1]$. The other settings are kept the same.

Performance of the proposed approach is evaluated in multiple aspects. (a) We first examine performance in identifying important CNA effects. The tuning parameter λ directly affects identification. Performance of a tuning parameter selection method can differ for different analysis approaches (the proposed and alternatives described below). To have a fair comparison of different approaches, following the literature, we consider a sequence of λ values, evaluate identification performance at each value, and use the area under the ROC curve (AUC) as the overall identification accuracy measure. For each fixed λ value, we consider a set of α values, including 0.01, 0.1, 0.3, 0.5, 0.8, and 1, as well as the α value selected using V-fold cross validation (V=5 in our numerical study). (b) We next evaluate estimation performance. When α is fixed, λ is chosen using V-fold cross validation. We also consider that both λ and α are cross validation-selected. Estimation accuracy is evaluated using the SSE (sum of squared errors). (c) We also evaluate whether the proposed *ad hoc* approach can effectively separate subjects in A_0 and A_1. As this is an identification problem, we evaluate using the TPR (true positive rate) and FPR (false positive rate).

To better gauge performance of the proposed approach, we consider two alternatives, which have penalized objective functions as

$$\text{LS:} \qquad \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n}(Y_{\mathrm{ij}} - X_i^{\top}\beta_{;j})^2 + \sum_{j=1}^{q}\lambda\|\beta_{,j}\|_1.$$

$$\text{LAD:} \qquad \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n}|Y_{\mathrm{ij}} - X_i^{\top}\beta_{,j}| + \sum_{j=1}^{q}\lambda\|\beta_{;j}\|_1.$$

Here the LS approach adopts the nonrobust least squared loss function, which has been a popular choice in quite a few studies. See for example Shi et al. (2015) and references therein. The LAD approach adopts the robust least absolute deviation loss function, which is a special case of the popular quantile regression. To ensure comparability, the LS and LAD approaches are applied and evaluated in the same manner as the proposed approach. Summary statistics are computed based on 200 replicates. Results for Simulation I with independent correlation are presented in Table 1, and those for the other simulation settings are presented in Tables 3–11 (Appendix).

With a more complex loss function, inevitably, the proposed approach is computationally more expensive. However, simulation shows that it is still affordable. For one replicate, with a fixed α and a sequence of twenty λ values, the proposed approach takes about 850 seconds on a regular desktop, compared to about 120 seconds and 520 seconds for the LS and LAD approaches. Simulation suggests that it is preferred to have the α value small but not too small – similar observations have been made in the literature for low-dimensional problems. For our simulated data, comprehensively considered, α = 0.3 seems to have the most competitive performance. The α values selected using CV mostly fall in the range of (0.2, 0.4). In practical data analysis, to be prudent, other α values still should be examined. When A_1 is empty (scenario S1), the proposed approach has performance very similar or slightly inferior to the LS. For example, in Table 1 under (S1), all three approaches are very successful in identifying important CNA effects, while having SSEs 22 (LS), 31 (LAD), 25 (proposed with α = 0.3), and 23 (proposed with CV-selected α), respectively. It is noted that similar observations are made when the signal levels are lower (and hence the AUC values are also smaller). With long-tailed distributions/contamination, the proposed approach significantly outperforms the LS. Specifically, the LS approach has inferior identification. For example in Table 1 under (S7), it has AUC 0.79, compared to 0.99 of the proposed approach (both $\alpha = 0.3$ and CV-selected α). The most prominent problem of LS is its biased estimation. For example in Table 1 under (S7), its mean SSE is as large as 10^7. The LAD has identification performance comparable or slightly inferior to the proposed approach. However, the proposed approach outperforms LAD with more accurate estimation (smaller SSEs). For example in Table 1 under (S2), the proposed approach and LAD have SSEs 47 (both $\alpha = 0.3$ and CV-selected α) and 82, respectively. Under most of the simulation scenarios, the proposed approach can perfectly identify the subjects in A_1.

The observed superiority of the proposed approach may seem unfair, as it has an additional parameter α. However, we note that the three approaches have been compared in a relatively fair way. Beyond using the CV-selected α, setting $\alpha = 0.3$ also seems to provide competitive performance for all of our simulation settings. This fits the suggestion made in low-dimensional studies that α should be small but not too small.

4 Data Analysis

Many recent studies have collected data on both GE and CNA. Here we analyze TCGA data, which have a high quality and are publicly available. Specifically, we analyze data on cutaneous melanoma, which poses a serious public health concern, and extensive profiling studies have been conducted on it. As in many published studies, we analyze the processed level 3 data which were downloaded from cbioportal using the CGDS-R package. Detailed information on data processing is available in the literature (The Cancer Genome Atlas Network, 2015). Briefly, mRNA gene expressions were initially measured using the Illumina Hiseq RNAseq V2 platform. The downloaded data are the robust Z-scores which have been lowess-normalized, log-transformed, and median-centered and represent the gene expression status (up or down regulated) in tumor samples relative to normal tissues. CNA measurements were first obtained using the Affymetrix Genome-wide Human SNP array 6.0 platform. The loss and gain levels of copy number changes of tumors compared to normal tissues were identified using segmentation analysis and expressed in the log2 transformed form. Thus, what is analyzed is a relative CNA measure. Standard data processing is conducted following published studies. The analyzed dataset contains records on 208 subjects.

Here we analyze the GE-CNA regulatory relationships for genes in one pathway. With different pathways having different biological functions, across-pathway *trans*-acting CNA effects, although may exist, are expected to be small. In addition, with a limited sample size, regressing a GE level on all available CNA measurements may lead to unreliable results. Pathway information is obtained from Gene Ontology (GO) using the annotation package in GSEA (www.broadinstitute.org/gsea). The pathway of special interest is the regulation of apoptosis (RAP) pathway. A well-known hypothesis supported by genetic, functional, and biomedical studies is that melanoma cells are "born to survive" (Soengas et al., 2003). The aggressive behavior of melanoma cells stems from intrinsic resistance to apoptosis from their paternal melanocytes nourished by additional alterations acquired during tumor progression. It is of interest to study the GE-CNA regulation for this pathway, which may contribute to the understanding of the underlying biological mechanisms that are related to such survival mechanisms. A total of 333 GEs and 229 CNAs belong to this pathway.

In Figure 1, we show examples of GE distributions. The long tails and deviation from normality are clearly seen, suggesting that it is reasonable to conduct robust analysis. We apply the proposed as well as LS and LAD approaches, as in simulation. All tuning parameters are selected using cross validation. Different approaches lead to significantly different findings. Specifically, for the 333 × 229 regression coefficient matrix β, 794 (LS), 6,641 (LAD), and 5,338 (proposed) nonzero elements are identified. Compared to the LS, the robust LAD and proposed approaches identify many more regulations. The LS approach shares 436 and 443 common nonzero elements with the LAD and proposed approaches, respectively. The LAD and proposed approaches share 2,732 nonzero elements. Unlike in simulation, there is a lack of objective measure on identification accuracy. To get an indirect support, we conduct a five-fold cross validation-based prediction evaluation. The prediction MSEs are 1.47 (LS), 1.39 (LAD), and 1.18 (proposed), respectively.

To be comprehensive, we further conduct analysis and identify subjects that may belong to A_1. It is noted again that this analysis, although may be informative, is not necessary. With the proposed *ad hoc* approach, a total of 68 (LS), 80 (LAD), and 74 (proposed) subjects are identified as in A_1. Different approaches identify different sets of A_1 (details available from the authors). We also remove A_1 and re-analyze data. In this "after" analysis, for the regression coefficient matrix β, a total of 1,010 (LS), 5,215 (LAD), and 3,408 (proposed) nonzero elements are identified. The LS approach shares 512 and 573 common nonzero elements with the LAD and proposed approaches, respectively. The LAD and proposed approaches share 1,846 nonzero elements. The prediction MSEs are 0.92 (LS), 0.98 (LAD), and 0.86 (proposed), respectively. All approaches have improved prediction by removing the long-tailed/contaminated subjects. Here we note that the analysis results before and after removing A_1 are considerably different. In theory, the proposed approach conducts estimation and inference for A_0 no matter A_1 is present or not. However for this specific dataset, as a considerable number of subjects are identified as in A_1, in the analysis with all subjects, the results can be "pulled" to "balance" between the two sets of subjects. This problem may get more prominent with the high dimensionality and shrinkage estimation. In addition, a closer examination of the "before" analysis shows that some of the estimates are very small, which contribute to the difference between the "before" and "after" identification results but can be practically ignored.

We take a closer look at the subjects in A_0 and A_1. In Figure 2, we show the heatmaps of the GE correlation matrices for the two sets. The difference is clearly seen. We further apply the approach developed in Jennrich (1970), which is realized using the R package *psych*, and test the equivalence of the two correlation matrices. The resulting p-value is $< 10^{-6}$, suggesting a highly significant difference. The clear difference between A_0 and A_1 provides a strong support to the necessity of robust analysis as well as effectiveness of the proposed approach. In Figure 3, we plot the positions of the nonzero elements of β from analyzing A_0 and A_1 separately. The difference is again obvious.

Examining the analysis results for individual genes suggests that the identified CNA effects can be biologically meaningful. As a representative example, we examine the analysis results for gene PTEN, which is a known tumor suppressor and frequently inactivated in a variety of cancers including melanoma (Stahl et al., 2003). The identified CNA effects are presented in Table 2 for all three approaches. For each approach, we present the results for before (denoted using the subscript "*b*") and after removing A_1 (denoted using the subscript "*a*"). Note that the LS_b analysis fails to identify the *cis*-acting CNA effect, which is likely to be unreasonable, while the other five analyses do. This also suggests that the nonrobust analysis with all subjects may be inappropriate. Beyond the *cis*-acting CNA effect, the proposed approach also identifies a few meaningful *trans*-acting CNA effects. For example, it identifies regulatory effects of the apoptosis-related regulators such as Bcl-2 family members, BTK, Fas, RUNX3, and Sh3glb1. In addition, we also identify CNA effects from tumor-related genes such as Ras effector B-Raf, pro-apoptotic protein Bid, high penetrance susceptibility gene CDKN2A, and Notch signaling gene Notch2.

In summary, for this dataset, the proposed approach identifies GE-CNA regulations different from those using the alternatives. Its validity is supported by the smaller prediction MSEs, significant difference between A_0 and A_1, and important biological implications of many of the identified regulations. We have also examined other TCGA data/pathways and made similar observations (results omitted).

5 Discussion

Many studies have been conducted, identifying which *cis*- and *trans*-acting CNAs regulate gene expression levels and in what ways. This study examines the same scientific problem but advances from the existing studies by developing a novel new analysis approach. The most prominent advancement is the adoption of the DPD loss to accommodate long-tailed GE distributions and contamination, which have been well acknowledged in the literature but insufficiently investigated. Compared to some alternative robust techniques for example quantile regression, the DPD is less popular but may have several notable advantages, as have been noted in low-dimensional studies. Our study is the first to apply the DPD approach to the high-dimensional analysis of GE-CNA regulations. Robust methods are very limited in the context of GE-CNA regulation. In low-dimensional data analysis, no robust method dominates the others. It is thus of interest to develop new robust methods in addition to the existing ones. Other notable features of the proposed approach include adopting high-dimensional regression models (to accommodate a large number of candidate *trans*-acting CNAs) and penalized estimation and selection. The development of an effective computational algorithm and R code makes the proposed approach ready to be used in practice. Our simulation and data analysis suggest that the proposed approach can outperform the nonrobust approach and the robust LAD approach. In data analysis, we have implemented a two-step procedure (with the "before" and "after" analysis). We have also experimented applying it to simulated data and found that the identification results are almost identical and there is a small improvement in estimation (for example, for the simulation settings in Table 3, the "after" analysis has 2–7% improvement in terms of SSE). Overall, this study provides a practically useful tool for an important biological problem.

Multiple analyses can be conducted following the proposed estimation and identification. In this study, we consider an *ad hoc* approach and identify the long-tailed/contaminated subjects. Other “post-analysis” can also be conducted, as in published robust analyses. This study can be extended in multiple aspects, including for example coupling the DPD loss with other regularization techniques, establishment of statistical properties, and others. In low-dimensional studies, the statistical efficiency of estimates has also been examined. Under the present high-dimensional settings, we have focused on identification and estimation. It may also be of interest to study efficiency. In data analysis, we analyze one pathway, which may generate more reliable results than conducting a whole-genome analysis. For datasets with larger sample sizes, in principle, the proposed approach can be directly applied to whole-genome analysis. The proposed approach is developed for CNAs. GE levels may also be regulated by other mechanisms (microRNAs, methylation, etc.). Conceptually, the proposed approach can be directly applied to the model “GE~CNA + other mechanisms”.

Acknowledgments

We thank the editor and reviewers for their careful review and insightful comments, which have led to a significant improvement of this article. The study was partly supported by NIH/NCI (CA191383, CA204120), National Natural Science Foundation of China (71301162, 11401561), and National Bureau of Statistics of China (2016LD01).

References

Basu A, Harris IR, Hjort NL, Jones M. Robust and efficient estimation by minimising a density power divergence. Biometrika. 1998; 85(3):549–559.

Blackburn A, Almeida M, Dean A, Curran JE, Johnson MP, Moses EK, Abraham LJ, Carless MA, Dyer TD, Kumar S, Almasy L. Effects of copy number variable regions on local gene expression in white blood cells of Mexican americans. European Journal of Human Genetics. 2015; 23(9):1229–35. [PubMed: 25585699]

Deng MC, Eisen HJ, Mehra MR, Billingham M, Marboe CC, Berry G, Kobashigawa J, Johnson FL, Starling RC, Murali S, Pauly DF. Noninvasive discrimination of rejection in cardiac allograft recipients using gene expression profiling. American Journal of Transplantation. 2006; 6(1):150–160. [PubMed: 16433769]

Durio A, Isaia ED. The minimum density power divergence approach in building robust regression models. Informatica. 2011; 22(1):43–56.

Fujisawa H, Eguchi S. Robust estimation in the normal mixture model. Journal of Statistical Planning and Inference. 2006; 136(11):3989–4011.

Ghosh A, Basu A. Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. Electronic Journal of statistics. 2013; 7:2420–2456.

Gosh D. Genomic outlier detection in high-throughput data analysis. Statistical Methods for Microarray Data Analysis: Methods and Protocols. 2013:141–153.

Henrichsen CN, Vinckenbosch N, Zöllner S, Chaignat E, Pradervand S, Schütz F, Ruedi M, Kaessmann H, Reymond A. Segmental copy number variation shapes tissue transcriptomes. Nature Genetics. 2009; 41(4):424–429. [PubMed: 19270705]

Jennrich RI. An asymptotic χ^2 test for the equality of two correlation matrices. Journal of the American Statistical Association. 1970; 65(330):904–912.

Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nature Methods. 2016; 13(4):366–70. [PubMed: 26950747]

Osborne JW, Overbay A. The power of outliers (and why researchers should always check for them). Practical assessment, research & evaluation. 2004; 9(6):1–12.

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. 2015; 31(24):3977–3983. [PubMed: 26342102]

Shieh AD, Hung YS. Detecting outlier samples in microarray data. Statistical Applications in Genetics and Molecular Biology. 2009; 8(1):1–24.

Soengas MS, Lowe SW. Apoptosis and melanoma chemoresistance. Oncogene. 2003; 22(20):3138–3151. [PubMed: 12789290]

Sparano J, Gray R, Makower DF, Pritchard KI, Albain KS, Hayes DF, Geyer CE Jr, Dees EC, Perez EA, Olson JA Jr, Zujewski J. Prospective validation of a 21-gene expression assay in breast cancer. New England Journal of Medicine. 2015; 373(21):2005–2014. [PubMed: 26412349]

Stahl JM, Cheung M, Sharma A, Trivedi NR, Shanmugam S, Robertson GP. Loss of PTEN promotes tumor development in malignant melanoma. Cancer Research. 2003; 63(11):2881–2890. [PubMed: 12782594]

The Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. Cell. 2015; 161(7):1681–1696. [PubMed: 26091043]

Wang RT, Ahn S, Park CC, Khan AH, Lange K, Smith DJ. Effects of genome-wide copy number variation on expression in mammalian cells. BMC Genomics. 2015; 12:562.

Wu C, Ma S. A selective review of robust variable selection with applications in bioinformatics. Briefings in Bioinformatics. 2015; 16(5):873–883. [PubMed: 25479793]

Yuan Y, Curtis C, Caldas C, Markowetz F. A sparse regulatory network of copy-number driven gene expression reveals putative breast cancer oncogenes. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2012; 9(4):947–954.

Appendix

Table 3

Simulation I with AR correlation: mean (sd) based on 200 replicates (the values of AUC, TPR, and FPR are multiplied by 100).

				DPD (α =)						
Scenario		**LS**	**LAD**	**0.01**	**0.1**	**0.3**	**0.5**	**0.8**	**1**	**CV**
S1	AUC	LS	99(0)	100(0)	100(0)	100(0)	100(0)	99(0)	99(0)	100(0)
	SSE	19(1)	28(1)	20(1)	19(1)	22(1)	26(2)	44(3)	53(4)	20(1)
S2	AUC	88(1)	97(0)	98(1)	99(0)	99(0)	99(0)	98(0)	98(0)	99(0)
	SSE	115(16)	64(5)	118(16)	70(7)	39(4)	41(4)	62(6)	73(6)	38(4)
	TPR	100(1)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	4(2)	1(1)	5(2)	1(1)	0(0)	0(0)	0(0)	0(0)	0(0)
S3	AUC	84(1)	94(1)	94(1)	94(2)	96(1)	95(1)	95(1)	94(1)	96(1)
	SSE	253(23)	168(22)	252(26)	222(30)	125(17)	113(20)	136(21)	155(26)	120(18)
	TPR	64(5)	83(6)	63(7)	73(7)	91(6)	93(5)	93(4)	93(4)	92(5)
	FPR	9(2)	3(2)	9(2)	7(2)	2(1)	1(1)	1(2)	1(1)	2(1)
S4	AUC	92(1)	96(0)	96(1)	99(0)	99(0)	99(0)	99(0)	99(0)	99(0)
	SSE	154(6)	50(3)	138(8)	54(3)	33(2)	44(2)	59(4)	68(5)	33(2)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S5	AUC	87(1)	94(0)	92(1)	94(1)	98(0)	98(0)	98(0)	98(0)	98(0)

				DPD (α =)						
Scenario		**LS**	**LAD**	**0.01**	**0.1**	**0.3**	**0.5**	**0.8**	**1**	**CV**
	SSE	258(14)	185(16)	256(10)	177(10)	128(5)	63(4)	80(8)	92(7)	87(5)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S6	AUC	91(2)	99(0)	99(0)	100(0)	100(0)	99(0)	99(0)	99(0)	100(0)
	SSE	10^4(10^5)	72(11)	84(11)	44(9)	62(9)	72(11)	82(14)	93(15)	53(9)
	TPR	100(0)	100(1)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	8(2)	0(0)	1(1)	1(1)	1(1)	1(1)	0(1)	1(1)	1(1)
S7	AUC	83(2)	99(0)	96(1)	99(0)	100(0)	100(0)	99(0)	99(0)	100(0)
	SSE	10^8(10^8)	193(23)	196(31)	79(11)	99(15)	112(21)	126(20)	132(25)	80(13)
	TPR	96(3)	99(1)	99(1)	99(1)	99(1)	99(1)	99(1)	99(1)	99(1)
	FPR	8(2)	0(0)	1(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)

Table 4

Simulation I with banded correlation: mean (sd) based on 200 replicates (the values of AUC, TPR, and FPR are multiplied by 100).

				DPD (α =)						
Scenario		**LS**	**LAD**	**0.01**	**0.1**	**0.3**	**0.5**	**0.8**	**1**	**CV**
S1	AUC	100(0)	99(0)	100(0)	100(0)	100(0)	100(0)	99(0)	99(0)	100(0)
	SSE	21(1)	30(1)	21(1)	20(1)	23(2)	28(2)	46(3)	57(4)	22(2)
S2	AUC	93(1)	97(0)	98(1)	99(0)	99(0)	99(0)	98(0)	98(0)	99(0)
	SSE	121(15)	69(7)	113(18)	74(7)	43(4)	45(4)	63(6)	78(7)	43(4)
	TPR	100(1)	100(0)	100(1)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	5(2)	1(1)	4(2)	2(1)	0(0)	0(0)	0(1)	0(0)	0(0)
S3	AUC	88(1)	93(1)	94(1)	94(1)	95(1)	95(1)	95(1)	94(1)	95(1)
	SSE	263(27)	178(27)	251(23)	233(26)	139(23)	127(17)	138(20)	169(20)	131(20)
	TPR	66(7)	74(6)	68(6)	73(6)	88(8)	92(4)	93(5)	90(5)	90(7)
	FPR	10(23)	3(2)	9(2)	7(2)	2(1)	1(1)	1(1)	1(1)	2(1)
S4	AUC	91(1)	97(0)	96(1)	99(0)	99(0)	99(0)	99(0)	99(0)	99(0)
	SSE	160(10)	54(4)	147(9)	59(5)	36(3)	45(4)	61(4)	74(6)	36(3)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S5	AUC	86(1)	93(0)	91(1)	93(1)	98(0)	98(0)	98(0)	97(0)	98(0)
	SSE	274(16)	93(7)	266(13)	192(13)	64(8)	65(5)	88(10)	103(8)	63(7)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S6	AUC	90(2)	97(2)	99(0)	100(0)	100(0)	100(0)	99(0)	99(0)	100(0)
	SSE	10^4(10^5)	78(11)	93(15)	48(9)	60(7)	76(10)	84(10)	94(11)	53(7)
	TPR	100(0)	100(1)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	8(2)	1(1)	2(1)	0(0)	1(1)	1(1)	1(1)	1(1)	1(1)

				DPD (α =)						
Scenario		**LS**	**LAD**	**0.01**	**0.1**	**0.3**	**0.5**	**0.8**	**1**	**CV**
S7	AUC	81(2)	98(0)	96(1)	99(0)	99(0)	99(0)	99(0)	99(0)	99(0)
	SSE	10^8(10^8)	197(26)	208(31)	84(13)	108(20)	112(19)	131(20)	142(22)	97(16)
	TPR	95(3)	99(1)	99(2)	99(1)	99(2)	100(1)	99(2)	99(1)	99(1)
	FPR	9(2)	0(0)	2(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)

Table 5

Simulation II with independent correlation: mean (sd) based on 200 replicates (the values of AUC, TPR, and FPR are multiplied by 100).

				DPD (α =)						
Scenario		**LS**	**LAD**	**0.01**	**0.1**	**0.3**	**0.5**	**0.8**	**1**	**CV**
S1	AUC	100(0)	99(0)	100(0)	100(0)	100(0)	100(0)	100(0)	99(0)	100(0)
	SSE	21(1)	31(1)	21(1)	21(1)	23(1)	29(2)	45(2)	57(3)	23(1)
S2	AUC	92(0)	98(0)	98(0)	99(0)	99(0)	99(0)	98(0)	98(0)	99(0)
	SSE	110(5)	71(4)	106(5)	74(5)	46(1)	48(4)	67(4)	84(6)	46(1)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	1(1)	0(0)	1(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S3	AUC	89(1)	94(1)	95(1)	95(1)	95(1)	96(0)	94(1)	93(1)	96(1)
	SSE	232(10)	181(22)	232(11)	208(15)	145(15)	117(10)	149(15)	174(13)	128(12)
	TPR	98(2)	100(1)	98(2)	98(0)	100(0)	100(0)	100(1)	100(1)	100(0)
	FPR	4(2)	1(1)	4(1)	2(1)	1(1)	0(0)	1(1)	1(1)	1(1)
S4	AUC	92(1)	97(0)	95(1)	98(0)	99(0)	99(0)	99(1)	99(0)	98(0)
	SSE	165(8)	54(4)	157(10)	60(4)	36(2)	44(2)	58(3)	69(4)	36(2)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S5	AUC	86(1)	96(0)	90(1)	93(1)	97(0)	98(0)	98(0)	97(0)	97(0)
	SSE	291(14)	92(6)	281(11)	200(10)	63(3)	62(4)	84(6)	90(7)	61(3)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S6	AUC	88(2)	99(0)	98(0)	100(0)	100(0)	100(0)	99(0)	99(0)	100(0)
	SSE	10^5(10^5)	65(10)	91(11)	49(8)	56(7)	65(10)	78(7)	89(13)	52(7)
	TPR	100(1)	100(0)	100(0)	100(0)	100(0)	100(0)	100(1)	100(0)	100(0)
	FPR	2(1)	0(0)	1(1)	0(0)	0(0)	0(0)	0(0)	0(1)	0(0)
S7	AUC	80(2)	98(0)	95(1)	99(0)	100(0)	99(0)	99(0)	99(0)	100(0)
	SSE	10^6(10^7)	190(21)	210(25)	78(12)	90(12)	122(43)	112(20)	131(30)	83(11)
	TPR	97(2)	100(1)	99(1)	100(1)	100(1)	99(1)	99(1)	100(0)	100(0)
	FPR	2(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)

Table 6

Simulation II with AR correlation: mean (sd) based on 200 replicates (the values of AUC, TPR, and FPR are multiplied by 100).

				DPD (α =)						
Scenario		**LS**	**LAD**	**0.01**	**0.1**	**0.3**	**0.5**	**0.8**	**1**	**CV**
S1	AUC	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	SSE	14(1)	20(1)	14(1)	14(1)	16(1)	18(1)	25(1)	35(2)	15(1)
S2	AUC	93(1)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	SSE	88(6)	57(3)	86(6)	53(4)	30(2)	31(2)	37(2)	43(2)	30(2)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(1)	0(0)	1(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S3	AUC	90(1)	98(0)	97(1)	98(0)	98(0)	98(0)	98(0)	97(1)	98(0)
	SSE	197(9)	125(10)	197(12)	170(9)	103(6)	76(9)	78(10)	89(14)	83(6)
	TPR	96(0)	100(1)	96(3)	98(2)	100(1)	100(0)	100(0)	100(1)	100(0)
	FPR	3(1)	0(0)	3(1)	1(1)	0(0)	0(0)	0(0)	0(0)	0(0)
S4	AUC	93(1)	100(0)	98(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	SSE	112(6)	36(3)	104(7)	40(3)	24(2)	31(3)	40(4)	47(3)	24(2)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S5	AUC	90(1)	99(0)	95(1)	97(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	SSE	192(9)	59(3)	184(8)	128(6)	44(3)	43(4)	56(6)	61(5)	42(3)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S6	AUC	92(1)	98(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	SSE	$10^5(10^5)$	69(10)	66(10)	38(8)	52(11)	57(9)	65(8)	72(14)	41(3)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	2(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S7	AUC	84(2)	99(0)	98(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	SSE	$10^6(10^7)$	95(17)	161(33)	67(12)	80(20)	102(25)	118(35)	123(24)	75(13)
	TPR	97(2)	100(1)	99(1)	100(1)	100(1)	99(1)	99(1)	100(1)	100(1)
	FPR	1(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)

Table 7

Simulation II with banded correlation: mean (sd) based on 200 replicates (the values of AUC, TPR, and FPR are multiplied by 100).

				DPD (α =)						
Scenario		**LS**	**LAD**	**0.01**	**0.1**	**0.3**	**0.5**	**0.8**	**1**	**CV**
S1	AUC	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	SSE	17(1)	24(1)	17(1)	17(1)	19(1)	21(1)	33(3)	44(3)	18(1)
S2	AUC	92(1)	99(0)	99(0)	100(0)	100(0)	100(0)	100(0)	99(0)	100(0)

Scenario		LS	LAD	DPD (a =)						
				0.01	0.1	0.3	0.5	0.8	1	CV
	SSE	96(6)	53(4)	192(7)	60(5)	26(2)	36(2)	43(3)	54(4)	26(2)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	1(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S3	AUC	90(2)	97(0)	96(1)	97(1)	98(0)	98(0)	97(1)	97(1)	98(0)
	SSE	209(9)	139(13)	208(10)	185(12)	101(8)	90(9)	89(11)	102(10)	96(9)
	TPR	97(2)	99(1)	95(3)	97(2)	100(1)	100(1)	100(0)	100(0)	100(0)
	FPR	3(1)	0(0)	3(1)	2(1)	0(0)	0(0)	0(0)	0(0)	0(0)
S4	AUC	94(1)	100(0)	98(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	SSE	129(5)	43(3)	118(6)	47(3)	30(3)	36(3)	48(3)	55(4)	30(3)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S5	AUC	90(1)	99(0)	94(1)	96(1)	99(0)	99(0)	99(0)	99(0)	99(0)
	SSE	220(9)	72(3)	213(9)	151(8)	50(3)	50(4)	66(8)	72(7)	49(3)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S6	AUC	90(1)	100(0)	99(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	SSE	10^6(10^7)	66(11)	77(14)	45(8)	56(12)	72(12)	74(10)	83(9)	49(10)
	TPR	100(1)	100(0)	100(0)	100(0)	100(0)	100(0)	100(1)	100(0)	100(0)
	FPR	2(1)	0(0)	0(1)	0(0)	0(1)	0(0)	0(0)	0(0)	0(0)
S7	AUC	82(2)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	SSE	10^7(10^8)	156(18)	73(11)	71(14)	87(10)	93(18)	97(17)	127(20)	82(11)
	TPR	97(2)	100(0)	100(0)	100(1)	100(1)	100(1)	100(0)	100(1)	100(1)
	FPR	1(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)

Table 8

Simulation I with AR correlation and a lower signal level: mean (sd) based on 200 replicates (the values of AUC, TPR, and FPR are multiplied by 100).

Scenario		LS	LAD	DPD (a =)						
				0.01	0.1	0.3	0.5	0.8	1	CV
S1	AUC	94(0)	93(1)	94(0)	94(0)	93(1)	92(2)	88(1)	87(1)	94(1)
	SSE	34(1)	46(1)	34(1)	34(1)	39(1)	40(2)	86(4)	88(2)	37(2)
S2	AUC	89(1)	88(1)	89(1)	89(0)	89(0)	87(1)	85(1)	83(1)	89(0)
	SSE	78(4)	76(3)	77(4)	70(3)	62(2)	75(4)	112(7)	120(5)	62(2)
	TPR	99(1)	100(0)	99(1)	100(1)	100(0)	99(1)	99(1)	100(0)	100(0)
	FPR	6(2)	6(2)	6(3)	5(2)	4(1)	6(1)	7(3)	7(2)	4(1)
S3	AUC	83(1)	82(1)	83(1)	84(1)	85(1)	81(1)	78(1)	77(1)	85(1)
	SSE	130(7)	126(6)	127(4)	121(6)	114(7)	127(6)	165(10)	172(7)	113(7)
	TPR	80(3)	78(5)	77(4)	78(5)	83(6)	82(4)	80(5)	85(4)	83(6)
	FPR	11(3)	10(2)	12(3)	10(3)	7(2)	11(4)	10(1)	8(2)	7(2)

Scenario		LS	LAD	DPD (α =)						
				0.01	0.1	0.3	0.5	0.8	1	CV
S4	AUC	77(1)	87(0)	77(1)	86(1)	90(1)	88(1)	87(0)	85(0)	91(1)
	SSE	172(7)	79(3)	163(5)	85(3)	56(4)	84(6)	109(4)	117(5)	55(4)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S5	AUC	68(1)	79(1)	69(1)	73(1)	84(1)	85(1)	83(1)	82(1)	85(1)
	SSE	237(7)	120(5)	232(5)	193(4)	94(6)	110(9)	138(14)	153(7)	93(7)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S6	AUC	75(1)	89(1)	87(1)	93(0)	92(0)	91(0)	88(1)	86(1)	93(1)
	SSE	$10^4(10^4)$	77(5)	105(13)	58(6)	78(20)	98(17)	113(15)	125(15)	63(10)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	99(1)	100(0)
	FPR	7(2)	1(1)	2(1)	0(0)	0(0)	0(0)	1(1)	0(0)	0(0)
S7	AUC	65(2)	89(0)	79(1)	91(1)	91(1)	90(1)	88(1)	86(1)	92(1)
	SSE	$10^6(10^6)$	102(8)	199(1)	87(8)	104(12)	43(33)	162(14)	169(20)	84(9)
	TPR	96(2)	100(1)	99(1)	99(1)	99(0)	99(1)	99(1)	99(1)	99(1)
	FPR	7(2)	0(0)	1(0)	0(1)	0(0)	0(0)	0(0)	0(0)	0(0)

Table 9

Simulation II with AR correlation and a lower signal level: mean(sd) based on 200 replicates (the values of AUC, TPR, and FPR are multiplied by 100).

Scenario		LS	LAD	DPD (α =)						
				0.01	0.1	0.3	0.5	0.8	1	CV
S1	AUC	97(0)	97(0)	98(0)	98(0)	92(0)	96(0)	96(0)	96(0)	98(0)
	SSE	14(1)	19(1)	14(1)	14(1)	15(1)	17(1)	23(1)	30(1)	14(1)
S2	AUC	95(1)	95(1)	95(1)	95(1)	96(0)	95(0)	94(1)	93(1)	96(0)
	SSE	33(2)	29(1)	31(2)	28(2)	23(1)	24(1)	30(2)	37(3)	23(1)
	TPR	100(0)	100(1)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	2(1)	2(1)	3(1)	2(1)	1(1)	1(1)	2(1)	2(1)	1(1)
S3	AUC	90(1)	91(0)	90(1)	91(1)	91(1)	90(1)	90(1)	88(1)	91(1)
	SSE	60(4)	51(1)	58(2)	54(2)	45(2)	44(4)	51(3)	59(3)	45(2)
	TPR	96(2)	97(3)	95(3)	96(2)	98(1)	99(2)	98(1)	98(2)	99(1)
	FPR	2(2)	2(1)	3(2)	2(1)	1(1)	1(1)	1(1)	2(1)	1(1)
S4	AUC	84(1)	95(1)	85(1)	93(1)	96(1)	95(0)	94(0)	92(0)	96(1)
	SSE	78(2)	32(2)	71(3)	35(1)	24(2)	31(3)	40(3)	48(3)	22(2)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S5	AUC	76(1)	85(1)	77(1)	82(1)	92(1)	93(0)	92(1)	92(1)	92(1)
	SSE	112(4)	48(3)	107(5)	87(4)	43(4)	48(5)	58(9)	69(13)	44(5)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)

Scenario		LS	LAD	DPD (α =)						
				0.01	0.1	0.3	0.5	0.8	1	CV
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S6	AUC	83(1)	91(0)	93(1)	97(1)	97(0)	96(1)	95(1)	94(1)	97(0)
	SSE	77(3)	39(3)	49(7)	30(4)	43(5)	46(5)	64(13)	70(15)	33(4)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	1(1)	0(0)	0(0)	0(1)	0(0)	1(1)	0(0)	0(0)	0(0)
S7	AUC	71(1)	90(0)	86(0)	96(0)	96(1)	96(1)	95(1)	94(1)	96(1)
	SSE	$10^6(10^7)$	60(8)	112(24)	50(8)	65(12)	95(22)	95(14)	98(15)	52(9)
	TPR	98(2)	99(0)	100(1)	99(2)	100(1)	100(1)	100(1)	100(1)	100(1)
	FPR	0(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)

Table 10

Simulation I with AR correlation and half positive/negative signals: mean (sd) based on 200 replicates (the values of AUC, TPR, and FPR are multiplied by 100).

Scenario		LS	LAD	DPD (α =)						
				0.01	0.1	0.3	0.5	0.8	1	CV
S1	AUC	93(0)	92(0)	93(0)	93(0)	92(1)	90(1)	87(1)	86(1)	92(0)
	SSE	37(2)	50(2)	38(1)	38(1)	42(2)	56(3)	88(2)	90(2)	38(1)
S2	AUC	86(1)	86(1)	86(1)	86(1)	87(1)	85(1)	82(1)	81(1)	87(1)
	SSE	93(3)	90(4)	92(3)	82(3)	72(3)	86(2)	120(3)	128(5)	73(3)
	TPR	98(2)	99(1)	99(2)	100(0)	99(1)	99(1)	99(1)	100(0)	100(0)
	FPR	5(1)	4(1)	4(2)	4(2)	2(1)	4(1)	6(3)	7(3)	2(1)
S3	AUC	77(2)	77(2)	77(1)	80(1)	78(1)	77(1)	75(1)	74(1)	80(1)
	SSE	155(8)	155(10)	152(5)	149(5)	139(6)	148(8)	175(10)	182(12)	138(6)
	TPR	73(4)	76(5)	77(6)	77(5)	78(5)	80(5)	75(4)	81(4)	80(5)
	FPR	12(3)	10(3)	13(2)	10(2)	8(3)	8(3)	8(2)	8(2)	8(3)
S4	AUC	75(1)	84(1)	75(1)	85(0)	89(1)	87(1)	85(1)	85(1)	90(1)
	SSE	186(7)	84(4)	175(5)	92(3)	64(5)	85(4)	114(4)	120(4)	66(5)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S5	AUC	67(1)	79(1)	67(1)	72(1)	83(1)	83(1)	82(1)	81(1)	83(1)
	SSE	245(7)	131(5)	239(6)	204(3)	102(7)	115(10)	148(6)	157(8)	105(8)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S6	AUC	74(2)	87(0)	86(1)	91(1)	91(0)	89(0)	86(1)	85(1)	91(1)
	SSE	$10^8(10^8)$	180(5)	111(10)	64(5)	74(8)	95(9)	123(8)	135(14)	67(6)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	3(1)	0(0)	1(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S7	AUC	64(1)	87(0)	77(1)	90(1)	90(0)	88(1)	86(1)	84(1)	91(1)
	SSE	$10^9(10^9)$	111(7)	202(33)	92(6)	109(12)	142(13)	173(13)	183(17)	94(8)

Scenario		LS	LAD	DPD (α =)						
				0.01	0.1	0.3	0.5	0.8	1	CV
	TPR	98(2)	99(1)	99(1)	99(1)	100(1)	99(1)	98(1)	99(1)	99(1)
	FPR	4(2)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)

Table 11

Simulation II with AR correlation and half positive/negative signals: mean (sd) based on 200 replicates (the values of AUC, TPR, and FPR are multiplied by 100).

Scenario		LS	LAD	DPD (α =)						
				0.01	0.1	0.3	0.5	0.8	1	CV
S1	AUC	98(1)	97(0)	98(0)	98(0)	97(1)	97(0)	95(0)	95(1)	97(0)
	SSE	14(1)	19(5)	13(1)	14(2)	15(1)	17(1)	22(1)	30(1)	15(1)
S2	AUC	94(1)	95(1)	94(0)	95(1)	95(1)	94(1)	93(1)	92(1)	95(1)
	SSE	34(2)	32(2)	35(2)	30(1)	25(1)	25(1)	30(1)	36(3)	25(1)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	2(1)	2(1)	2(1)	2(1)	1(1)	1(1)	1(1)	2(1)	1(1)
S3	AUC	88(1)	89(1)	88(1)	89(1)	90(1)	89(1)	88(1)	87(1)	90(1)
	SSE	66(3)	57(3)	63(2)	59(2)	48(3)	46(2)	52(2)	8(5)	47(3)
	TPR	95(3)	97(2)	96(1)	97(2)	98(2)	99(1)	99(1)	98(2)	99(1)
	FPR	5(2)	3(2)	5(1)	4(2)	3(2)	1(1)	2(1)	3(2)	3(2)
S4	AUC	84(1)	92(1)	85(1)	93(1)	96(0)	95(0)	94(0)	94(1)	96(0)
	SSE	76(3)	33(2)	73(2)	35(2)	24(2)	32(3)	37(2)	46(4)	23(3)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S5	AUC	76(1)	88(1)	76(1)	81(1)	92(1)	93(1)	92(1)	92(1)	93(1)
	SSE	110(4)	48(2)	110(6)	86(3)	40(3)	43(3)	54(7)	60(10)	39(3)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S6	AUC	81(2)	93(0)	93(1)	97(1)	97(0)	96(0)	95(0)	95(1)	97(0)
	SSE	85(16)	39(5)	53(6)	30(4)	39(6)	47(7)	65(20)	60(9)	33(5)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	1(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S7	AUC	70(2)	90(1)	87(1)	96(1)	96(0)	96(0)	95(0)	94(1)	96(1)
	SSE	$10^6(10^7)$	62(7)	94(13)	47(8)	70(11)	95(27)	110(28)	98(14)	49(9)
	TPR	99(1)	100(1)	100(1)	100(0)	99(1)	99(1)	100(0)	100(0)	100(0)
	FPR	0(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)

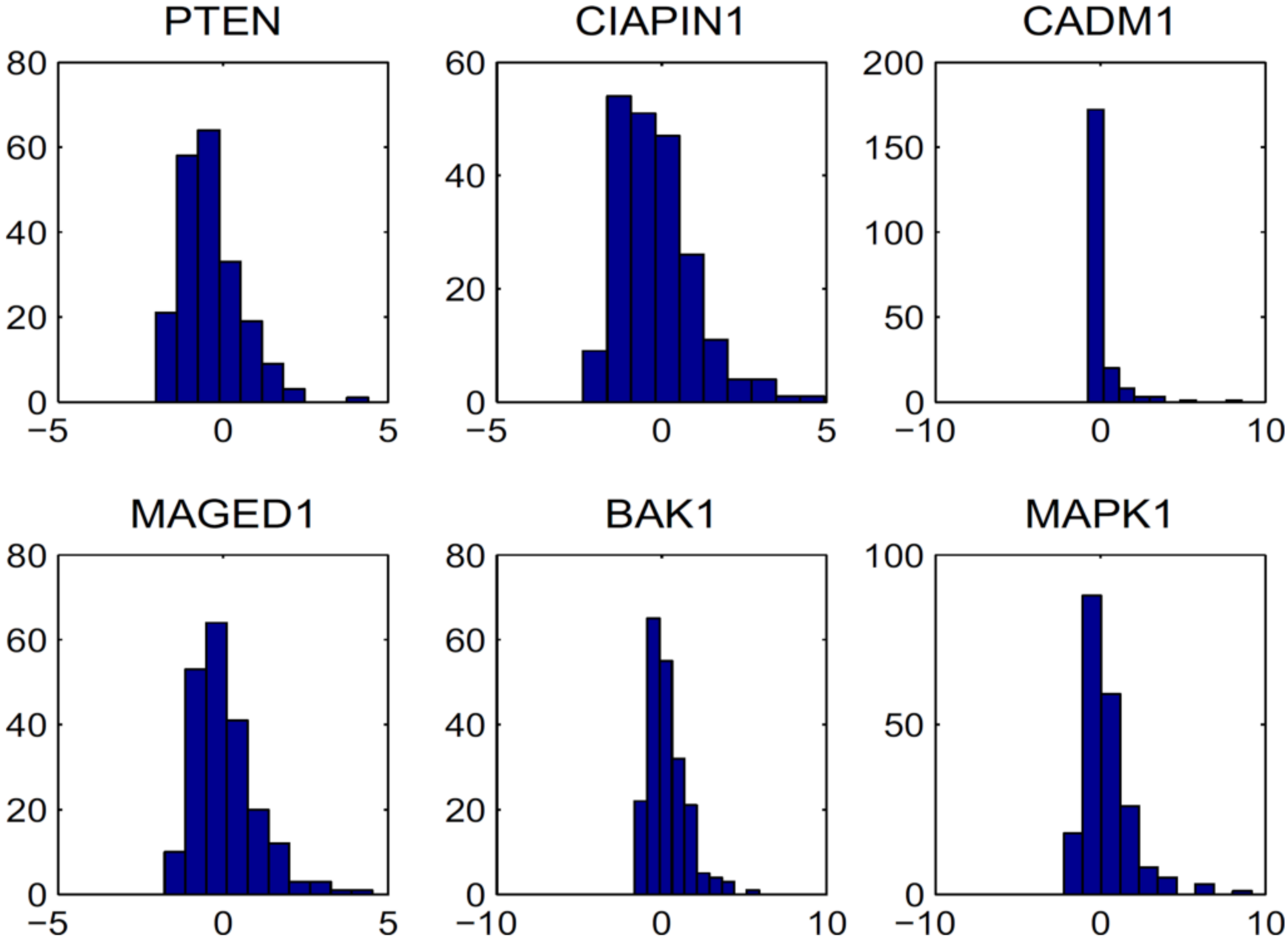

Figure 1.

Data analysis: histograms of GE distributions with long-tails (x-axis: GE levels).

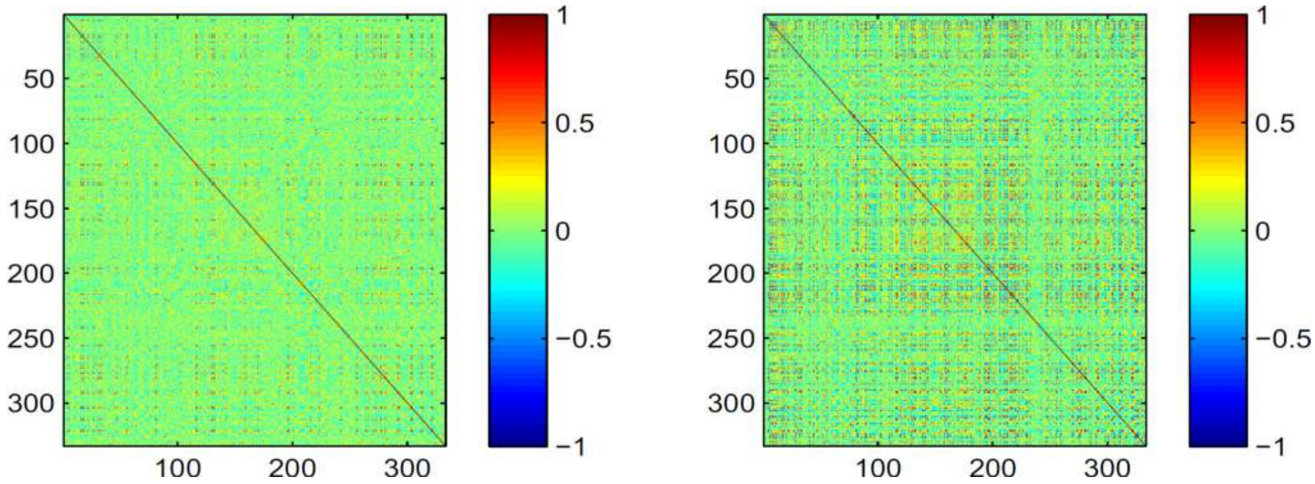

Figure 2.

Data analysis: heatmaps of the GE correlation matrices for A_0 (left) and A_1 (right). Both x-axis and y-axis are gene numbers.

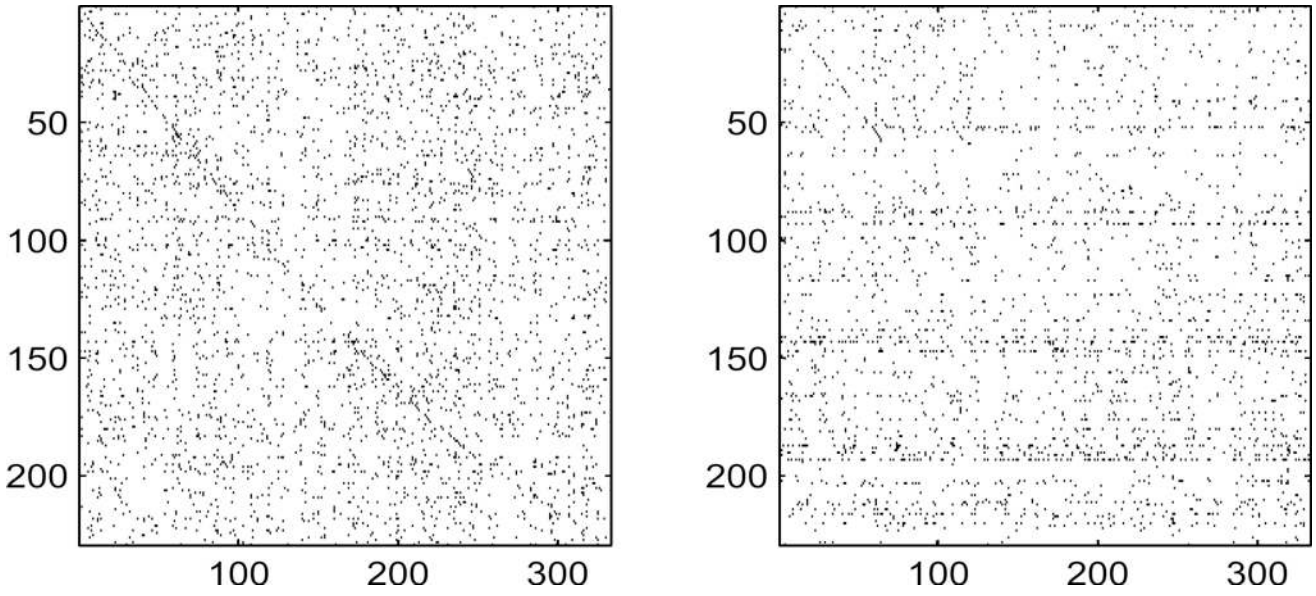

Figure 3.

Data analysis: positions of the nonzero components of β for A_0 (left) and A_1 (right). x-axis and y-axis are GE and CNA numbers, respectively.

Table 1

Simulation I with independent correlation: mean (sd) based on 200 replicates (the values of AUC, TPR, and FPR are multiplied by 100).

Scenario		LS	LAD	DPD (α =)						
				0.01	**0.1**	**0.3**	**0.5**	**0.8**	**1**	**CV**
S1	AUC	100(0)	99(0)	100(0)	100(0)	100(0)	100(0)	99(0)	99(0)	100(0)
	SSE	22(2)	31(2)	21(1)	22(1)	25(2)	30(2)	49(3)	59(4)	23(2)
S2	AUC	88(1)	98(0)	98(1)	98(0)	98(0)	98(0)	98(1)	97(1)	98(0)
	SSE	117(16)	82(6)	111(13)	78(9)	47(4)	52(4)	76(7)	90(12)	47(4)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	5(2)	2(1)	4(2)	2(1)	0(0)	1(1)	1(0)	1(1)	0(0)
S3	AUC	83(1)	94(1)	93(1)	94(1)	94(1)	94(1)	93(1)	91(1)	94(1)
	SSE	256(24)	211(21)	251(25)	217(26)	171(26)	144(14)	183(30)	212(24)	155(22)
	TPR	74(6)	87(5)	75(6)	78(7)	94(3)	94(4)	93(4)	92(4)	94(3)
	FPR	10(2)	6(2)	10(2)	7(2)	3(2)	3(2)	3(2)	4(2)	3(2)
S4	AUC	94(1)	96(0)	95(1)	98(0)	99(0)	99(0)	99(0)	98(0)	99(0)
	SSE	171(9)	54(3)	60(12)	60(4)	36(2)	47(3)	63(4)	74(5)	35(2)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	1(1)	0(0)	1(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S5	AUC	88(1)	97(1)	89(1)	93(1)	97(0)	97(0)	97(0)	97(1)	97(0)
	SSE	303(18)	96(8)	293(20)	138(12)	65(3)	70(7)	89(10)	102(9)	66(3)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
S6	AUC	89(2)	99(0)	98(0)	100(0)	100(0)	99(0)	99(0)	99(0)	100(0)
	SSE	$10^7(10^8)$	63(12)	95(11)	52(9)	60(11)	69(8)	77(11)	91(7)	54(10)
	TPR	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	FPR	2(9)	1(1)	2(1)	1(1)	1(1)	0(0)	2(1)	1(1)	1(1)
S7	AUC	79(2)	98(0)	95(1)	99(0)	99(0)	99(0)	99(0)	99(0)	99(0)
	SSE	$10^7(10^8)$	102(13)	201(22)	78(11)	91(12)	105(18)	122(24)	145(32)	85(11)
	TPR	97(2)	99(1)	98(2)	100(1)	99(1)	99(1)	100(0)	99(0)	100(1)
	FPR	8(2)	0(0)	3(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)

Table 2

Data analysis: CNAs identified as associated with the expression level of gene PTEN.

CNA	DPD$_b$	DPD$_a$	LAD$_b$	LAD$_a$	LS$_b$	LS$_a$	CNA	DPD$_b$	DPD$_a$	LAD$_b$	LAD$_a$	LS$_b$	LS$_a$
BTK	−0.02	−0.04	−0.01				FAS	0.01					0.04
KRT18	−0.01			−0.02			TXNDC5	−0.03		−0.02	−0.04		
BCL2L1	−0.02						IKBIP	−0.01					
PTEN	0.35	0.36	0.30	0.35		0.26	AIFM1	−0.08		−0.04			
PSEN1	0.03	0.02	0.03				TNFRSF10D	0.03	0.01				
IFT57	−0.05						BIK	−0.03		−0.01			
GDNF	0.04		0.08	0.05			PRUNE2	−0.12		−0.06			
PRKCA	0.04						BTG1	−0.05	−0.01	−0.04			
SH3GLB1	0.13		0.12				BCL2L10		−0.04				
NOTCH2		0.03					STK4		−0.02		−0.03		
GCLC		−0.02					RUNX3		0.01		0.01		
ABL1		−0.01					DCC		−0.04				
ERCC2		−0.02					FOXL2		−0.03				
BCL2L11			−0.03				CD74			−0.03			
SEMA4D			−0.02				XIAP			−0.04	−0.03		
CDKN2A			−0.02				HRK			−0.04	−0.03		
CCL2			−0.02	−0.02			RPS3A			0.02			
CTSB			0.06				MAPK1				−0.02		
TMX1				0.01			SFRP1				0.02		
BAG4				0.02			PROK2				−0.01		
HMGB1				−0.02			DYRK2				−0.02		
CLCF1				−0.05			IGF1R				0.01		
BNIP3					0.02		CDK1					0.05	
IKBKG						0.02							