Statistica Sinica **27** (2017), 71-88
doi:http://dx.doi.org/10.5705/ss.202015.0180

REGRESSION IN HETEROGENEOUS PROBLEMS

Hanwen Huang

University of Georgia

Abstract: We develop a new framework for modeling the impact of sub-cluster structure of data on regression. The proposed framework is specifically designed for handling situations where the sample is not homogeneous in the sense that the response variables in different regions of covariate space are generated through different mechanisms. In such situation, the sample can be viewed as a composition of multiple data sets each of which is homogeneous. The traditional linear and general nonlinear methods may not work very well because it is hard to find a model to fit multiple data sets simultaneously. The proposed method is flexible enough to ensure that the data generated from different regions can be modeled using different functions. The key step of our method incorporates the k-means clustering idea into the traditional regression framework so that the regression and clustering tasks can be performed simultaneously. The k-means clustering algorithm is extended to solve the optimization problem in our model that groups the samples with similar response-covariate relationship together. General conditions under which the estimation of the model parameters is consistent are investigated. By adding appropriate penalty terms, the proposed model can conduct variable selection to eliminate the uninformative variables. The conditions under which the proposed model can achieve asymptotic selection consistency are also studied. The effectiveness of the proposed method is demonstrated through simulations and real data analysis.

Key words and phrases: Asymptotic consistency, heterogeneous problem, k-means clustering, LASSO, regression.

1. Introduction

Our method is motivated by a collection of problems as illustrated by a simulated example in which $\mathbf{x} \in R^{20}$, $y \in R$ and y only depends on the first coordinate x_1. Figure 1 shows the dependence of y on x_1. Clearly $\mathbf{x}$ can be divided into three regions according to x_1, and the relationships between y and $\mathbf{x}$ are different in different regions. This situation emerges often in practice when the response variables depend on the covariates via a number of distinct mechanisms, with different mechanisms applying in different regions of the feature space. For example, breast cancer data reveal distinct genomic sub-types and different subtypes display different clinical outcomes (Parker et al. (2009)). Our method can

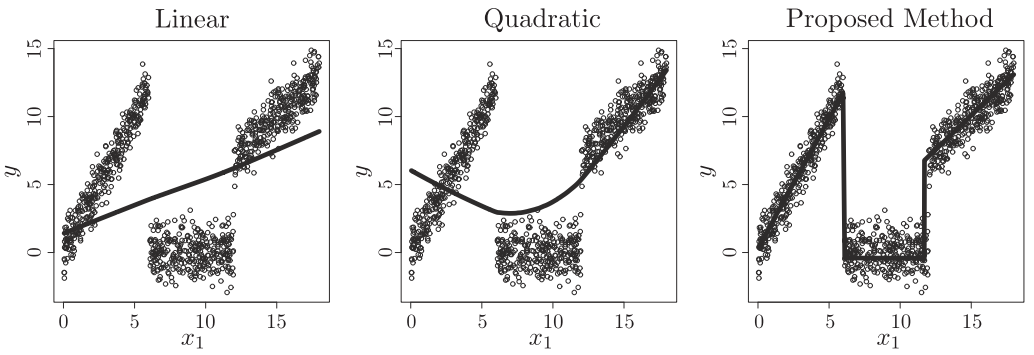

Figure 1. Scatter plots of y versus x_1 for a simulated example with three different regression curves shown using solid lines. Here the covariate $\mathbf{x} \in R^{20}$ and x_1 is the first coordinate. Note the proposed method gives a major improvement in performance.

be particularly applied to the analysis of the quantitative structure-activity relationship (QSAR) in which the structural characteristics of the polychlorinated biphenyl influence their potency and mechanism of action (Andersson (2000)). Section 5 applies our method to a QSAR example.

In such situations, the traditional linear and general nonlinear methods may not work well. A linear model was used to fit these data, see the left panel of Figure 1. The middle panel of Figure 1 shows a full quadratic fit. Our goal is to fit the data drawn from different regions using different functions, the right panel of Figure 1 results from fitting our model. It can not only improve the goodness of the fit but also find distinct sub-groups within the data set.

We incorporate k-means clustering into the traditional regression framework such that the regression and clustering tasks can be performed simultaneously. Assume that the feature space contains K regions and we conduct K linear regressions, one for each region. Our method is especially useful for high dimensional data with distinct sub-group structures. If there is only one covariate, as in Figure 1, it is easy to find the sub-structure through visualization and to apply traditional nonparametric techniques, such as piecewise splines, to improve the goodness of fit. In high dimensions, it is hard to identify the data structure. Traditional nonparametric methods are mainly designed for estimating the unknown function of one or a few variables. When y is an unknown function of many variables, our proposed method is flexible and appealing.

Aside from inhomogeneous data with potential sub-cluster structures that are the main motivation here, our method can also be applied when the dependence between the response and input variables is not linear. Our method can effectively partition the feature space into approximately linear regions and replace the nonlinear function with a linear approximation in each region (see

Simulations 3 and 4 in Section 4.1). In contrast to general nonlinear methods, our method has a much simpler function space and thus retains most of the good parsimony of linear methods. The advantages of our method over general nonlinear methods include less tendency for overfitting in high dimensional settings, and the capability of variable selection, since its coefficients have clear interpretation as in the linear case.

A new classification method, Bi-Dimensional Discrimination, has been developed recently (Huang, Liu and Marron (2012)) that shares the same spirit. It generalizes linear classification from one hyperplane to two or more. Our framework is more general and applies to both classification and regression.

The remainder of this article is organized as follows. Section 2 describes the proposed penalized composite regression model and introduces its computational algorithm. Some asymptotic properties of our method are provided in Section 3. We test the performance of our method on simulated data in Section 4, and on real data in Section 5. The article concludes with a discussion in Section 6. The technical conditions and details of proofs are relegated to the appendix and the online supplementary material.

2. Method

2.1. Formulation

Suppose we are given a training dataset consisting of n observations $(\mathbf{x}_i, y_i)$ for $i = 1, \ldots, n$. Here $\mathbf{x}_i \in R^d$ represents an input vector, and $y_i \in R$ denotes the corresponding response value. Assume that each $(\mathbf{x}_i, y_i)$ is an independent random vector distributed according to some unknown distribution function $P(\mathbf{x}, y)$. To estimate the relationship between y and $\mathbf{x}$, the commonly used LASSO penalized linear regression (Tibshirani (1996)) solves the optimization problem

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|) \right\}, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T \in R^d$ and $p_{\lambda_n}(\cdot)$ is a penalty function indexed by the tuning parameter λ_n. We propose a composite model that uses different linear expressions to fit the data drawn from different regions in feature space. This kind of idea has appeared in computer science works (Stanforth, Kolossov and Mirkin (2007); Manwani and Sastry (2012)), designed solely for clustering without considering statistical inference and feature selection. We extend it to sparse statistical inference and study its various asymptotic properties.

In the composite model, we divide the feature space into K domains and apply a linear function in each domain. Consider linear functions $\boldsymbol{\beta}_k$, $k = 1, \ldots, K$, and the problem

$$\min_{C(\cdot),\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{(y_i-\mathbf{x}_i^T\boldsymbol{\beta}_{C(i)})^2+\tau\|\mathbf{x}_i-\boldsymbol{\mu}_{C(i)}\|^2\right\}+\sum_{k=1}^{K}\sum_{j=1}^{d}p_{\lambda_n}(|\beta_{kj}|)\right], \tag{2.2}$$

where $\tau > 0$ is a fixed constant, $\|\cdot\|$ is the standard Euclidean norm, $\boldsymbol{\mu}_k$ is the centroid of the kth domain, and $C(\cdot) : \{1,\ldots,n\} \to \{1,\ldots,K\}$ is a cluster assignment function with $C(i) \in \{1,\ldots,K\}$ representing the cluster membership for the ith subject. The second term here enforces proximity of points in the same cluster that may have the same relationship with the response variable.

This approach can also be useful if the response variable y depends on $\mathbf{x}$ in a nonlinear fashion. It effectively partitions the feature space into approximately linear regions and replaces the nonlinear function with a linear approximation in each region. If $K = 1$, (2.2) reduces to (2.1). As K increases, more sub-cluster structures are discovered and the model approaches a general nonlinear model.

The first term in (2.2) measures the within-cluster residual sum of squares (RSS), the second measures the within-cluster distance from each observation to its corresponding cluster center in feature space. The tuning parameter τ decides relative weight of the two terms. One is not restricted to quadratic loss, other choices include L_1 loss, Huber's loss (Huber (1964)), Tukey's bisquare, and Hampel' ψ, among others.

2.2. Penalty function

The regularization term in (2.2) can be the LASSO penalty $p_{\lambda_n}(|\beta_{kj}|) = \lambda_n|\beta_{kj}|$ (Tibshirani (1996)) or the adaptive LASSO penalty $p_{\lambda_n}(|\beta_{kj}|) = \lambda_n w_{kj}|\beta_{kj}|$ (Zou (2006)), where λ_n is tuning parameter and w_{kj} $(k = 1,\ldots,K; j = 1,\ldots,d)$ are known weights. We use the adaptive LASSO penalty here and set $\hat{w}_{kj} = 1/|\tilde{\beta}_{kj}|$, where $\tilde{\beta}_{kj}$ is an estimate of β_{kj} when solving

$$\min_{C(\cdot),\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K}\left[\frac{1}{n}\sum_{i=1}\left\{(y_i-\mathbf{x}_i^T\boldsymbol{\beta}_{C(i)})^2+\tau\|\mathbf{x}_i-\boldsymbol{\mu}_{C(i)}\|^2\right\}\right]. \tag{2.3}$$

If we want to treat each direction in a grouped manner, we can apply the group LASSO penalty (Yuan and Lin (2006)) or the adaptive group LASSO penalty (Wang and Leng (2008)). We can also use other penalty functions such as the L_p-penalty with $0 < p < 1$ (Frank and Friedman (1993)), or the SCAD penalty (Fan and Li (2001)).

2.3. Computational algorithm

For (2.2), we adopt an iterative scheme as in solving the k-means clustering problem. We update $(\boldsymbol{\beta}_k, \boldsymbol{\mu}_k)$ and $C(i)$ separately at each iteration, holding the other one fixed. At the $t+1$-th iteration, when $(\boldsymbol{\beta}_k^{(t)}, \boldsymbol{\mu}_k^{(t)})$ are fixed, $C^{(t+1)}(i) =$

$\text{argmin}_k\{(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k^{(t)})^2 + \tau\|\mathbf{x}_i - \boldsymbol{\mu}_k^{(t)}\|^2\}$. When $C(i+1)$ is fixed, $\boldsymbol{\mu}_k$ is updated as the mean of the updated kth cluster and

$$\boldsymbol{\beta}_k^{(t+1)} = \text{argmin}_{\boldsymbol{\beta}_k}\left\{\frac{1}{n}\sum_{C^{(t+1)}(i)=k}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k)^2 + \lambda_n\sum_{j=1}^{d}\hat{w}_{kj}|\beta_{kj}|\right\}. \tag{2.4}$$

This can be solved using the coordinate descent algorithm as described in Zou (2006).

The details of the proposed algorithm are as follows.

Step 1. Obtain initial estimates of the cluster membership by applying the standard k-means clustering algorithm on the data $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$.

Step 2. Using the initial estimates of the cluster membership, obtain the cluster means and regression coefficients for each cluster by minimizing (2.4).

Step 3. Conditional on the given coefficients for each cluster, find the cluster assignment.

Step 4. Reiterate Steps 2 and 3 until all cluster memberships are unchanged.

As with k-means clustering, the optimization for (2.2) is not a convex problem, so results depend on the initial cluster membership generated by the standard k-means algorithm. To overcome the sensitivity to initialization, the standard k-means clustering is randomly done multiple times and the one which gives the smallest final objective value is selected.

2.4. Prediction

Our model can predict the response value for any given new input, but the prediction process is complicated due to the fact that there is more than one linear function involved. We use a classification algorithm to estimate the cluster label en route to the prediction, as follows

Step 1. Use the training data to build a classifier on the feature space. Many classification algorithm can be used at this stage, for example, the multiclass Support-Vector-Machine (Lee, Lin and Wahba (2004)) or the multiclass Distance-Weighted-Discrimination (Huang et al. (2013)).

Step 2. Predict to which cluster the test sample belongs based on the classifier created in Step 1.

Step 3. Compute the response value for the test sample using the linear function corresponding to the cluster label predicted from Step 2.

2.5. Related work

Other nonparametric approaches in the literature also have the potential to address the problem as shown by Figure 1. Piecewise polynomial and spline methods can fit the data where the sample is inhomogeneous, but they only work for low dimensional situations (Wahba (1983); Heckman (1986); Chen (1988); Speckman (1988); Cuzick (1992); Hastie and Loader (1993)). Classification and Regression Trees (CART) and more advanced tree methods make use of the assumption that nearby observations should have the same relationship with the response (Breiman et al. (1984)), but are less flexible for general inhomogeneous structures.

The change-point detection methods aim to identify changes at unknown times and to estimate the location of changes in stochastic processes (Page (1954); Yao (1993a,b); Ombao, von Sachs and Guo (2005); Aue et al. (2009); Muggeo (2008); Jeng, Cai and Li (2010); Fryzlewicz and Rao (2014); Frick, Munk and Sieling (2014), among many others). State of the art work on change-point detection, e.g. Fryzlewicz (2014), can deal with high dimensions, and the number of change points can be estimated automatically (Schroeder and Fryzlewicz (2013)).

Khalili and Chen (2007) proposed a finite mixture of regression (FMR) method to model data arising from heterogeneous populations. There, the conditional density function of y given $\mathbf{x}$ has the form

$$f(y; \mathbf{x}, \mathbf{\Psi}) = \sum_{k=1}^{K} \pi_k g(y; \mathbf{x}^T \boldsymbol{\beta}_k, \phi_k), \tag{2.5}$$

where $g(y; \mathbf{x}^T \boldsymbol{\beta}_k, \phi_k)$ is a parametric density function depending on linear predictor $\mathbf{x}^T \boldsymbol{\beta}_k$ and dispersion parameter ϕ_k. The model allows the response variable to follow a mixture distribution and the contributions of covariates to the response variable vary from one component to another, but the same π_k and $\boldsymbol{\beta}_k$ are used over the entire $\mathbf{x}$ region.

3. Asymptotic Properties

3.1. Asymptotic estimation consistency

Assuming the number of clusters K is pre-specified, we examine the asymptotic consistency of the parameters in our model as the number of the samples goes to infinity. Strong consistency of the standard k-means clustering was established in Pollard (1981), who specified a set of sufficient conditions for the empirical cluster centers to converge to the true cluster centers almost surely. Sun, Wang and Fang (2012) studied the consistency of a penalized k-means clustering model in which an adaptive penalized term is added to the standard k-means clustering formula. Here we employ similar techniques to study the consistency of the composite regression problem (2.2).

Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ be independent and identically distributed observations from an unknown joint distribution $P(\mathbf{x}, y)$, and let $P_n(\mathbf{x}, y)$ be the associated empirical measure. Let $A = \{(\boldsymbol{\beta}_1, \boldsymbol{\mu}_1), \ldots, (\boldsymbol{\beta}_K, \boldsymbol{\mu}_K)\}$ be a finite subset of $\mathbb{R}^{2d}$ containing K points, and take

$$g_A(\mathbf{x}, y) = \min_{(\boldsymbol{\beta}_k, \boldsymbol{\mu}_k) \in A} \{(y - \mathbf{x}^T \boldsymbol{\beta}_k)^2 + \tau \|\mathbf{x} - \boldsymbol{\mu}_k\|^2\}.$$

For any given probability measure $\Omega(\cdot, \cdot)$ on $\mathbb{R}^d \times R$, let

$$\Phi(A, \Omega) = \int g_A(\mathbf{x}, y) \Omega(d\mathbf{x}, dy). \tag{3.1}$$

Write $\mathcal{A}_K = \{A : A \text{ contains K points in } \mathbb{R}^{2d}\}$ and take

$$m_K(\Omega) = \inf_{A \in \mathcal{A}_K} \{\Phi(A, \Omega)\}. \tag{3.2}$$

Let $\bar{A}(K) = \{(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\mu}}_1), \ldots, (\bar{\boldsymbol{\beta}}_K, \bar{\boldsymbol{\mu}}_K)\}$ denote the subsets that satisfy $\Phi(\bar{A}(K), P) = m_K(P)$.

We choose the adaptive LASSO penalty $p_{\lambda_n}(|\beta_{kj}|) = \lambda_n |\beta_{kj}| / |\tilde{\beta}_{kj}|$, where $\tilde{\beta}_{kj}$ is the solution of (2.3). Regarding (2.2) as a function of regression coefficients, cluster centers and empirical measure P_n, the penalized composite regression finds a set $A = \{(\boldsymbol{\beta}_1, \boldsymbol{\mu}_1), \ldots, (\boldsymbol{\beta}_K, \boldsymbol{\mu}_K))$ to minimize

$$W(A, P_n) = \Phi(A, P_n) + \lambda_n \sum_{k=1}^{K} \sum_{j=1}^{d} \frac{|\beta_{kj}|}{|\tilde{\beta}_{kj}|} \tag{3.3}$$

over $A \in \mathcal{A}_K$. Once A is fixed, the cluster assignment of each subject is also fixed. Let $\hat{A}_n(K) = \{(\hat{\boldsymbol{\beta}}_1^{(n)}, \hat{\boldsymbol{\mu}}_1^{(n)}), \ldots, (\hat{\boldsymbol{\beta}}_K^{(n)}, \hat{\boldsymbol{\mu}}_K^{(n)})\}$ be the estimated coefficients solving (3.3).

Theorem 1. *If Conditions* 1–4 *in Appendix A hold, and* $n^{1/2}\lambda_n \to 0$, *then* $\hat{A}_n(K) \to \bar{A}(K)$ *almost surely.*

3.2. Asymptotic selection consistency

We study the asymptotic selection property of our model, investigating whether, under proper choice of λ_n, the uninformative variables in the regression can be eliminated with probability tending to one as n goes to infinity.

Denote the informative variable set with $\mathcal{B}$: all index pairs (k, j), $k = 1, \ldots, K$ and $j = 1, \ldots, d$, such that $\bar{\beta}_{kj} \neq 0$; and the informative variable set with $\mathcal{B}^c$: all index pairs (k, j) such that $\bar{\beta}_{kj} = 0$.

Theorem 2. *If Conditions* 1–6 *in Appendix A hold,* $n\lambda_n \to \infty$, *and* $n^{1/2}\lambda_n \to 0$, $P\left(\hat{\beta}_{kj}^{(n)} = 0\right) \to 1$ *for any* $(k, j) \in \mathcal{B}^c$.

The assumptions on λ_n here are the same as in Zou (2006), except for a scale $1/n$. The condition $n\lambda_n \to \infty$ is the minimum amount of penalization needed for variable selection, while $n^{1/2}\lambda_n \to 0$ ensures consistent estimation of large coefficients. The data dependent $\hat{w}_{kj} = 1/|\tilde{\beta}_{kj}|$ are the key ingredient: the sample size grows, the corresponding weights approach infinity for zero coefficient whereas they converge to a finite constant for nonzero coefficient.

In the proofs of consistency, we treat $\tau \geq 0$ as a fixed parameter. The parameter λ_n controls the variable selection, but changing τ does not change the consistency result. Empirically τ has a large impact on the results and it needs to be well tuned in a numerical study.

4. Simulation

We carried out some Monte Carlo simulations to study the performance of our method in comparison with other approaches. We took K fixed and treated τ and λ_n as the tuning parameters. We considered $K = 3$ and $K = 4$. Each simulation included a training set, a tuning set, and a test set with sample sizes $n_{\text{train}} = 200$, $n_{\text{tuning}} = 200$ and n_{test} =1,000, respectively. To analyze the performance of different methods, we repeated each example 100 times.

In each simulation, we first used the training and tuning sets to select the optimal tuning parameters τ and λ_n. To achieve this, we did a two-dimensional grid search. The range for $\log \lambda_n$ was from -5 to 5. For τ, we first rescaled the sum of square terms at (2.2) to make them comparable–we divided the first term by the estimated residual sum of squares from a single linear regression, then chose the range for the relative weight $\tau/(1+\tau)$ from 0 to 1. For each value, we applied our method to the training set to get the estimate of the cluster label for each data point, as well as the linear regression coefficients for each cluster, using the algorithm described in Section 2.3. Then based on the estimated cluster label, we constructed a classifier in the feature space $\mathbf{x} \in \mathbb{R}^d$ using the multiclass SVM method that allowed us to predict the cluster label for each tuning data point. The response value y for each tuning data point was computed based on the regression formula estimated for the corresponding cluster to which it belonged. The criterion we used for evaluating the performance of a model was based on the distance between the predicted y value and the corresponding true y value, $\sqrt{\sum_{i=1}^{n_{\text{tuning}}}(y_i - \hat{y}_i)^2/n_{\text{tuning}}}$, where $\hat{y}_i$ is the estimated value for the i-th tuning point. We calculated this distance for each τ and λ_n, and selected the τ and λ_n arrives the smallest distance as the optimal tuning parameters used in the prediction of the response value y for the test set.

For comparison, we include the results from the LASSO method, and from a regularized full quadratic model with L_1 penalty. It is of interest to compare our method with CART and change point detection methods. For CART, we used the

R package *tree*; for change point, we used the R package *segmented*. In *segmented*, we took both the "segmented" variables and the number of breakpoints from their true values. For example, in Simulation 1, we input the "segmented" variable as x_1, and the number of breakpoints as three.

4.1. Scenario I: three clusters $K = 3$

In the first scenario, we took the response value y to depend only on the first covariate, $y = f(x_1) + \epsilon$, where ϵ is standard normal. Here x_1 was drawn from a uniform [0,18] distribution and the remaining $d-1$ variables were sampled from a standard normal. We considered four simulation settings.

- Simulation 1 ($d = 20$, upper left panel of Figure 2):
$$f(x_1) = \begin{cases} 2x_1 & 0 \le x_1 \le 6, \\ 0 & 6 < x_1 \le 12, \\ -5 + x_1 & 12 < x_1 \le 18. \end{cases}$$
- Simulation 2 ($d = 100$, upper right panel of Figure 2):
$$f(x_1) = \begin{cases} x_1 & 0 < x_1 \le 6, \\ 12 - x_1 & 6 < x_1 \le 12, \\ -12 + x_1 & 12 < x_1 \le 18. \end{cases} \tag{4.1}$$
- Simulation 3 ($d = 20$, lower left panel of Figure 2):
$$f(x_1) = \frac{10 \exp(0.5x_1)}{1 + \exp(0.5x_1)}.$$
- Simulation 4 ($d = 20$, lower right panel of Figure 2):
$$f(x_1) = \frac{10 \exp(x_1)}{1 + \exp(x_1)}.$$

Table 1 lists the normalized test errors for each simulation setting using different methods. For Simulation 1, CART is best, and for the other three simulations, change point is best. Our proposed method, Composite, is close to best for all simulation settings. The performances from the the linear and quadratic methods are poor. The change point method does well in most of the situations because of the right choice for the "segmented" variable. The current version of *segmented* package cannot handle situations where all d variables are considered as the "segmented" variables, it crashes or takes forever to run. Fryzlewicz (2014) developed a sparsed-version change point method that can do variable selection but we did not find an existing R package.

Figure 2 shows the scatter plots and the predicted curves using the different methods for the simulated test example under different settings. Clearly, the

Table 1. Summary table of the normalized test error $\sqrt{\sum_{i=1}^{n_{\text{test}}}(y_i - \hat{y}_i)^2/n_{\text{test}}}$ over 100 replications based on different methods for Simulations 1−4. The mean and the standard error (in parenthesis) are reported.

	LASSO	Quadratic	CART	Change Point	Composite
Simulation 1	4.52(0.01)	4.19(0.01)	1.53(0.019)	2.73(0.11)	1.8 (0.05)
Simulation 2	2.82(0.02)	32.14(2.93)	1.36(0.01)	1.23(0.01)	1.30(0.01)
Simulation 3	1.34(0.004)	1.42(0.005)	1.17(0.004)	1.12(0.005)	1.14(0.008)
Simulation 4	1.88(0.005)	2.00(0.007)	1.15(0.002)	1.12(0.007)	1.14(0.007)

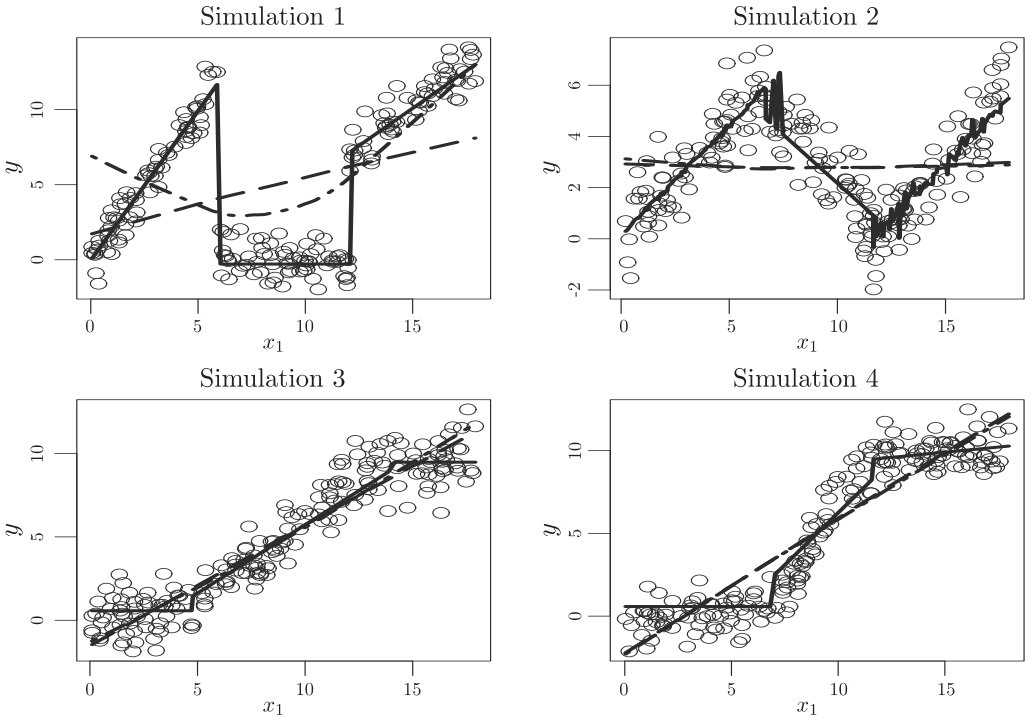

Figure 2. Scatter plots of y versus x_1 for four simulated test examples. Predicted curves using the Composite, Linear, and Quadratic methods are shown.

estimated curves from the Composite method are much closer to the true points than both linear and quadratic methods.

4.2. Scenario II: four clusters $K = 4$

In the second scenario, we took the response variable y to depend on the first two covariates, $y = f(x_1, x_2) + \epsilon$ with ϵ being standard normal. In Simulation 5, x_1 and x_2 were drawn from a uniform [0,12] distribution and the remaining $d-2$

Table 2. Summary table of the normalized test error over 100 replications based on different methods for Simulations 5-6. The mean and the standard deviation (in parenthesis) are reported.

	LASSO	Quadratic	CART	Change Point	Composite
Simulation 5	2.79(0.008)	1.32(0.004)	1.76(0.02)	1.06(0.003)	1.21(0.008)
Simulation 6	2.71(0.008)	2.64(0.007)	2.08(0.034)	2.54(0.008)	1.09(0.008)

variables were sampled from a standard normal. In Simulation 6, x_1 and x_2 were drawn from a uniform [-5,5] distribution and the remaining 18 variables were sampled from a standard normal. We had $f(x_1, x_2)$ change in the coordinate directions in Simulation 5, and in a rotated direction in Simulation 6.

- Simulation 5 ($d = 20$)

$$f(x_1,x_2) = \begin{cases} x_1 + x_2 & 0 \le x_1 \le 6 \quad \& \quad 0 \le x_2 \le 6, \\ 12 + x_1 - x_2 & 0 \le x_1 \le 6 \quad \& \quad 6 < x_2 \le 12, \\ 12 - x_1 + x_2 & 6 < x_1 \le 12 \quad \& \quad 0 \le x_2 \le 6, \\ 24 - x_1 - x_2 & 6 < x_1 \le 12 \quad \& \quad 6 < x_2 \le 12. \end{cases}$$

- Simulation 6 ($d = 20$)

$$f(x_1,x_2) = \begin{cases} 10 - x_1 - x_2 & x_1 + x_2 \ge 0, \\ 10 + x_1 + x_2 & x_1 + x_2 \le 0. \end{cases}$$

The performances of the different methods for Simulations 5 and 6 are summarized in Table 2. Our proposed Composite method delivered superior results against the other methods in terms of the test errors for Simulation 6. The poor performance of both CART and change point here is because they are designed to split in coordinate directions. In Simulation 5, change point is the best and our method is the second best, we take the "segmented" variables using the true variables x_1 and x_2, but in practice they are not known.

4.3. Variable selection

To study variable selection and interpretability, we show in Figure 3 the relative contributions of each variable considered in the Composite method for Simulation 2. Figure 3 indicates that our method correctly picks the first variable. Table 3 lists the estimated coefficients for the intercept and first variable of the three linear regressions seen in the Composite method. They are close to the true values. We also studied the variable selection property of our method for the other simulation settings in both scenarios. The results are similar and not reported here.

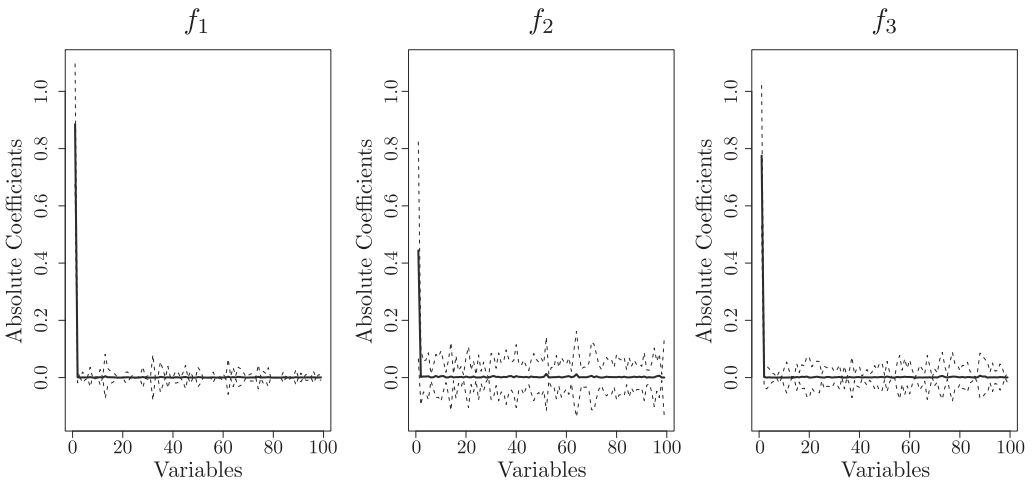

Figure 3. The averages, shown as solid lines with ± 2 standard deviations shown as dotted lines, over 100 replications of the absolute values of the regression coefficients for the regression functions in Simulation 2.

Table 3. Summary table of the estimated coefficients for the first variable over 100 replications based on the Composite method for Simulation 2.

	Intercept 1	Slope 1	Intercept 2	Slope 2	Intercept 3	Slope 3
True Value	0	0.89	12	-1	-12	1
Mean	0.48	0.88	7.54	-0.44	-8.56	0.78
Standard Error	(1.34)	(0.11)	(1.25)	(0.19)	(1.77)	(0.13)

5. A Real Data Example

We applied the proposed composite regression method to the analysis of quantitative structure-activity relationships (QSAR) used in the chemical and biological sciences. QSAR models study the relationship between chemical structures and biological activity, and use it to predict the activities of new chemicals. The predictors consist of physico-chemical properties or theoretical molecular descriptors of chemicals and the response-variable is a biological activity. The data consists of dihydrofolate reductase inhibitors and have been prepared by Klebe and Abraham (1999) and studied by Sutherland, OBrien and Weaver (2004). The response variable is called pIC_{50} values for rat liver enzyme ranging from 3.3 to 9.8. The predictors include the connection table of molecules, the physico-chemical properties of molecules in their bio-active conformation, as well as the traditional molecular descriptors, such the χ indices, counts of rotatable bonds, and molecular weight. After removing the missing data and highly correlated variables, the final data set used in the analysis included 273 subjects and 44 dimensions.

Table 4. Summary table of the test errors for the QSAR data over 100 random splitting based on different methods.

	LASSO	Quadratic	CART	FMR	Composit $K = 2$	Composit $K = 3$
Mean	0.629	0.741	0.691	0.644	0.609	0.646
S. E.	0.006	0.014	0.013	0.013	0.008	0.008

We studied the generalization properties of our method using cross-validation. The data set was randomly split into a training set (80%), a tuning set (10%), and a test set (10%). The tuning set is used for the selection of the tuning parameters τ and λ_n. The division of the data was randomly repeated 100 times and Table 4 reports the summary of the test errors over 100 replications for different methods. Here we considered three options for the composite model with $K = 1, 2, 3$, respectively. The linear model is equivalent to the composite model with $K = 1$. We did not apply the change point method here because the package cannot run for situations with more than one "segmented" variable. We included FMR method with $K = 2$ for comparison. From Table 4, the composite model with $K = 2$ gives the lowest test error followed by the LASSO and FMR methods. The worst two methods are Quadratic and CART. The results indicate that it is appropriate to divide the data into two groups and to apply separate linear regressions to each group.

We applied the composite model with $K = 2$ to the entire data set and chose the tuning parameters τ and λ_n to be those which gave the smallest BIC value $\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\hat{C}(i)})^2/(n\hat{\sigma}^2) + \hat{\nu}\log(n)$, where $\hat{\nu}$ is the estimated model degree of freedom, equal to the total number of nonzero coefficients. Our method automatically divided the data into two clusters including 140 and 133 samples. Here 17 active variables contributed to the regression for the first group: S_aaCH, S_sCl, N_aasC, IC, BIC, Rotlbonds, Hbond.acceptor, AlogP98, RadOfGyration, Jurs.PPSA.1, Jurs.PNSA.1, Jurs.RNCS, Shadow.XZ, Shadow.XYfrac, Shadow.XZfrac, Shadow.nu, and Shadow.Ylength, while 10 active variables contributed to the regression for the second group: Order, S_aasC, IC, BIC, Hbond.donor, SC.3_C, Jurs.PNSA.1, Jurs.PNSA.2, Jurs.FPSA.3, and Jurs.RPCS. We applied our method to several other QSAR data sets and obtained some similar results.

6. Conclusion

This article proposes a generalization of the linear regression with high flexibility and data adaptability. The proposed approach is able to automatically divide the data into multiple regions and apply different regression models in

different regions. The inclusion of the penalty terms allows us to perform variable selection simultaneously. It is shown that such a technique has an oracle property. Numerical studies show that our method outperforms some commonly used methods in certain situations where the data are heterogeneous.

In most real applications, the number of clusters of the samples K is not known. A major challenges in clustering analysis is to estimate the number of clusters. The main difficulty is the absence of an objective measure to compare the quality of various clusterings of the same dataset. A number of methods have been proposed for estimating the number of clusters, see Milligan and Cooper (1985); Gordon (1999); Sugar (1998); Sugar, Lenert and Olshen (1999); Tibshirani, Walther and Hastie (2001); Tibshirani and Walther (2005), among many others. It is difficult to find a method that works uniformly better in most situations.

Our composite model includes the response variables and thus enables us to directly assess the performance of different methods by evaluating how close the predicted response value is to the true value. In practice we can also treat K as tuning parameter in addition to to τ and λ_n. However, a three-dimensional search is computationally expensive. For this reason, we here assume the number of clusters K fixed, and treat different K as different models. Our method was designed for situations in which K is not too large (typically not more than 5). If K is large, the general nonlinear model is more appropriate.

The main computational burden for the current algorithm is from solving LASSO problem (2.4) in Step 2. We currently use coordinate descent algorithm that works fine for moderately high dimension. For a simulated Example 2 with $n = 200$, $d = 100$, and $K = 3$ in a computer with RAM 16GB and processor 3.5GHz, it took less than 2 minutes to run the analysis, including the two-dimension grid search for the tuning parameters. In order to apply to high-dimensional data, we need to improve the LASSO solver for (2.4) using fast algorithms such as recently developed iterative shrinkage and thresholding algorithm (Beck and Teboulle (2009)) or alternative direction method of multipliers algorithm (Tseng (1991)).

As pointed out by one reviewer, in many applications, different regression relationships cannot be well separated based on feature space at all. For example, subjects may react differently to the dosage increase of certain drug, resulting in several different regression lines that may or may not cross each other. To deal with such situations, we can extend the proposed method to more general mixture regression setups. One possible solution is to split feature space into M regions and extend the FMR method (2.5) to allow for different π_k and β_k in different regions. If $K = 1$, this general model reduces to our composite model; if $M = 1$, it reduces to the FMR model (2.5). It is quite challenging to study this general model for theoretical properties and numerical implementation.

Another direction would extend the current framework to the task of classification. This is quite straightforward and we only need to replace the L_2 loss function in (2.2) with such classification loss functions as hinge loss or logistic loss.

Supplementary Materials

The online supplement materials contain proofs of Theorems 1 and 2.

Acknowledgement

The author thanks the Editor, an associate editor, and two referees for many helpful comments and suggestions which led to a much improved presentation.

Appendix A: Regularity Conditions

Assume the data ($\mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$) are sampled from the distribution P and the marginal distribution of $\mathbf{x}$ is Q. Let P_n, Q_n be the corresponding empirical measures. Let $\mathcal{O} = \{\mathbf{u} \in \mathbb{R}^d \colon \|\mathbf{u}\|^2 = 1\}$ and $U = \{\mathbf{u}_1, \ldots, \mathbf{u}_k, \ldots, \mathbf{u}_K\}$ be a subset of $\mathcal{O}$ containing K points, $\mathbf{u}_k \in \mathcal{O}$. For fixed U, take $h_U(\mathbf{x}) = \min_{\mathbf{u}_k \in U}\{(\mathbf{x}^T\mathbf{u}_k)^2\}$. Let $\mathcal{U}_K = \{U \colon \text{U contains K points in } \mathcal{O}\}$, and for a probability measure Q, define

$$w_K(Q) = \inf_{U \in \mathcal{U}_K} \left\{ \int h_U(\mathbf{x}) Q(d\mathbf{x}) \right\}. \tag{A.1}$$

Let $\phi_k(\mathbf{x}, y) = (y - \mathbf{x}^T\boldsymbol{\beta}_k)^2 + \tau\|\mathbf{x} - \boldsymbol{\mu}_k\|^2$. For any given $A = \{(\boldsymbol{\beta}_1, \boldsymbol{\mu}_k), \ldots, (\boldsymbol{\beta}_K, \boldsymbol{\mu}_K)\}$, the space $\mathbb{R}^d \times R$ can be divided into K distinct regions B_k, $k = 1, \ldots, K$, such that if $(\mathbf{x}, y) \in B_k$, $\phi_k(\mathbf{x}, y) < \phi_j(\mathbf{x}, y)$ for any $j \neq k$. Then for the indicator function $I(\cdot)$, we have

$$I(B_k) = \prod_{j \neq k} I\left(\phi_k(\mathbf{x}, y) < \phi_j(\mathbf{x}, y)\right). \tag{A.2}$$

Take $G_{kj}(\mathbf{x}, y) = \phi_j(\mathbf{x}, y) - \phi_k(\mathbf{x}, y)$, and denote by $m_{kj}(\mathbf{x}, y)$ the Jacobian

$$m_{kj}(\mathbf{x}, y) = \det\left(\frac{\partial^2 G_{kj}(\mathbf{x}, y)}{\partial \mathbf{x} \partial y}\right).$$

Denote by $\sigma_{kj}(F(\mathbf{x}, y))$ the integral of function $F(\mathbf{x}, y)$ over the boundary surface between B_k and B_j.

Condition 1. The joint distribution $P(\mathbf{x}, y)$ satisfies

$$a = \int (\|\mathbf{x}\|^2 + y^2) P(d\mathbf{x}, dy) < \infty. \tag{A.3}$$

Condition 2. For each $k = 1, \ldots, K$, the solution to the populational distribution $\bar{A}(k)$ is unique lies in a compact region of $\mathbb{R}^{2d}$.

Condition 3. The marginal distribution $Q(\mathbf{x})$ satisfies $w_K(Q) > 0$.

Condition 4. For any $\epsilon > 0$, there exists a $\delta_0 > 0$ such that, with $\mathrm{I}(\cdot)$ an indicator function,

$$\sup_{\mathbf{u}\in\mathcal{O}} \int \|\mathbf{x}\|^2 \mathrm{I}(|\mathbf{x}^T\mathbf{u}| < \delta_0) Q(d\mathbf{x}) < \epsilon. \tag{A.4}$$

Condition 5. The joint distribution P has a continuous density $f(\mathbf{x}, y)$ with respect to $d+1$ dimensional Lebesgue measure.

Condition 6. For each k and j, there exists

$$\sigma_{kj}\left(\frac{\{(y-\mathbf{x}^T\boldsymbol{\beta}_k)\mathbf{x}+\tau(\mathbf{x}-\boldsymbol{\mu}_k)\}\{(y-\mathbf{x}^T\boldsymbol{\beta}_j)\mathbf{x}+\tau(\mathbf{x}-\boldsymbol{\mu}_j)\}^T}{|m_{kj}(\mathbf{x},y)|}f(\mathbf{x},y)\right).$$

Condition 1 explicitly depicts the boundedness of the second order moment of both the joint distribution $P(d\mathbf{x}, dy)$ and the marginal distribution $Q(d\mathbf{x})$. Condition 2 indicates that for each $k = 1, \ldots, K$, there is a unique set $\bar{A}(k)$ for which $\Phi(\bar{A}(k), P) = m_k(P)$. Similar to the argument given in Pollard (1981) for k-means clustering, the uniqueness on $\bar{A}(k)$ implies that $m_1(P) > m_2(P) > \cdots m_K(P)$. Condition 3 ensures that the distribution of $\mathbf{x}$ spreads over all directions. Condition 4 assures that the marginal distribution $Q(\mathbf{x})$ gives zero probability to every $d-1$ dimensional hyperplane in $\mathbb{R}^d$.

Conditions 5 and 6 are to insure that the loss function in (2.2) has a second order derivative over parameters $(\boldsymbol{\beta}_k, \boldsymbol{\mu}_k)$, and thus can be approximated locally by a quadratic form. Particularly, Condition 6 ensures that the integrals of functions $\partial\phi_k(\mathbf{x}, y)/\partial\boldsymbol{\beta}_k$ and $\partial\phi_k(\mathbf{x}, y)/\partial\boldsymbol{\mu}_k$ exist over the boundary surface.

References

Andersson, P. (2000). Physico-chemical characteristics and quantitative structureactivity relationships of pcbs. Dissertation: Texas A & M University.

Aue, A., Hormann, S., Horvath, L. and Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *Ann. Statist.* **37**, 4046-4087.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.* **2**, 183-202.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall, New York.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* **16**, 136-146.

Cuzick, J. (1992). Semiparametric additive regression. *J. Roy. Statist. Soc. Ser. B* **54**, 831-843.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and it oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Frank, E. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-135.

Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference. *J. Roy. Statist. Soc. Ser. B* **76**, 495-580.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42**, 2243-2281.

Fryzlewicz, P. and Rao, S. S. (2014). Multiple-change-point detection for autoregressive conditional heteroscedastic processes. *J. Roy. Statist. Soc. Ser. B* **76**, 903-924.

Gordon, A. (1999). *Classification*. 2nd edition. Chapman & Hall. CRC Press.

Hastie, T. and Loader, C. (1993). Local regression: Automatic kernel carpentry. *Statist. Sci.* **8**, 120-129.

Heckman, N. (1986). Spline smoothing in a partly linear model. *J. Roy. Statist. Soc. Ser. B* **48**, 244-248.

Huang, H., Liu, Y., Du, Y., Perou, C. M., Hayes, D. N., Todd, M. J. and Marron, J. S. (2013). Multiclass distance-weighted discrimination. *J. Comput. Graph. Statist.* **22**, 953-969.

Huang, H., Liu, Y. and Marron, J. S. (2012). Bi-directional discrimination with application to data visualization. *Biometrika* **99**, 851-864.

Huber, P. (1964). Robust estimation of location parameter. *Ann. Math. Statist.* **35**, 73-101.

Jeng, X. J., Cai, T. T. and Li, H. (2010). Optimal sparse segment identification with application in copy number variation analysis. *J. Amer. Statist. Assoc.* **105**, 1156-1166.

Khalili, A. and Chen, J. (2007). Variables selection in finite mixture of regression models. *J. Amer. Statist. Assoc.* **102**, 1025-1038.

Klebe, G. and Abraham, U. (1999). Comparative molecular similarity index analysis (comsia) to study hydrogen-bonding properties and to score combinatorial libraries. *J Computer-Aided Molecular Design* **13**, 1-10.

Lee, Y., Lin, Y. and Wahba, G. (2004). Multicategory support vector machines. *J. Amer. Statist. Assoc.* **99**, 67-81.

Manwani, N. and Sastry, P. S. (2012). K-plane regression. CoRR abs/1211.1513.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50**, 159-179.

Muggeo, V. M. R. (2008). segmented: An R package to fit regression models with broken-line relationships. *R News* **8**, 20-25.

Ombao, H., von Sachs, R. and Guo, W. (2005). Slex analysis of multivariate nonstationary time series. *J. Amer. Statist. Assoc.* **100**, 519-531.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika* **41**, 100-115.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z. and et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clinical Oncol.* **27**, 1160-1167.

Pollard, D. (1981). Strong consistency of k-means clustering. *Ann. Statist.* **9**, 135-140.

Schroeder, A. L. and Fryzlewicz, P. (2013). *Statist. and Its Interface* **6**, 449-461.

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50**, 413-436.

Stanforth, R., Kolossov, E. and Mirkin, B. (2007). Hybrid k-means: Combining regression-wise and centroid-based criteria for QSAR. Selected Contributions in Data Analysis and Classification: *Studies in Classification, Data Analysis, and Knowledge Organization*, 225-233.

Sugar, C. (1998). Techniques for clustering and classification with applications to medical problems. Ph.D. dissertation in Statistics, Stanford University.

Sugar, C., Lenert, L. and Olshen, R. (1999). An application of cluster analysis to health services research: empirically defined health states for depression from the SF-12. Technical report. Division of Biostatistics, Stanford.

Sun, W., Wang, J. and Fang, Y. (2012). Regularized k-means clustering of highdimensional data and its asymptotic consistency. *Electron. J. Statist.* **6**, 148-167.

Sutherland, J. J., OBrien, L. A. and Weaver, D. F. (2004). A comparison of methods for modeling quantitative structure-activity relationships. *J. Med. Chem.* **47**, 5541-5554.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Tibshirani, R. and Walther, G. (2005). Cluster validation by prediction strength. *J. Comput. Graph. Statist.* **14**, 511-528.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. Roy. Statist. Soc. Ser. B* **63**, 411-423.

Tseng, P. (1991). Applications of splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control Optim.* **29**, 119-138.

Wahba, G. (1983). *Cross Validated Spline Methods for the Estimation of Multivariate Functions from Data on Functionals*. Mathematical Sciences Research Institute.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Comput. Statist. Data Anal.* **52**, 5277-5286.

Yao, Q. (1993a). Asymptotically optimal detection of a change in a linear model. *Sequent. Anal.* **12**, 201-210.

Yao, Q. (1993b). Tests for change-points with epidemic alternatives. *Biometrika* **80**, 179-191.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49-67.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA 30602, USA.
E-mail: huanghw@uga.edu

(Received May 2015; accepted December 2015)