Statistica Sinica **26** (2016), 235-253
doi:http://dx.doi.org/10.5705/ss.2014.233

ADAPTIVE ESTIMATION WITH PARTIALLY OVERLAPPING MODELS

Sunyoung Shin[1], Jason Fine[2] and Yufeng Liu[2]

[1] *The University of Wisconsin, Madison and*
[2] *The University of North Carolina, Chapel Hill*

Abstract: In many problems, one has several models of interest that capture key parameters describing the distribution of the data. Partially overlapping models are taken as models in which at least one covariate effect is common to the models. A priori knowledge of such structure enables efficient estimation of all model parameters. However, in practice, this structure may be unknown. We propose adaptive composite M-estimation (ACME) for partially overlapping models using a composite loss function, which is a linear combination of loss functions defining the individual models. Penalization is applied to pairwise differences of parameters across models, resulting in data driven identification of the overlap structure. Further penalization is imposed on the individual parameters, enabling sparse estimation in the regression setting. The recovery of the overlap structure enables more efficient parameter estimation. An oracle result is established. Simulation studies illustrate the advantages of ACME over existing methods that fit individual models separately or make strong a priori assumption about the overlap structure.

Key words and phrases: Composite loss function, lasso, model selection, oracle property, overlapping, penalization, sparsity.

1. Introduction

Regression modeling aims to explain the association between a response variable and covariates in a dataset. A regression model targets a profile of the conditional distribution of the response given the predictors. It is of interest to consider several linear models to describe a more complete picture of the conditional distribution. We can simultaneously fit the models on the dataset and estimate the parameters. Such joint estimation borrows information across the models and is referred as to composite estimation.

Composite estimation may be based on combining loss functions as weighted averages of loss functions tailored to individual models. Given n independent identically distributed samples, $\boldsymbol{z}_1 = (\boldsymbol{x}_1, y_1), \ldots, \boldsymbol{z}_n = (\boldsymbol{x}_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, consider the K empirical convex loss functions:

$$\frac{1}{n}\sum_{i=1}^{n} L_k(\boldsymbol{z}_i, (\alpha_k, \boldsymbol{\beta}_k)) \equiv \frac{1}{n}\sum_{i=1}^{n} L_k(y_i, \alpha_k + \boldsymbol{x}_i^T\boldsymbol{\beta}_k),\ k = 1, \ldots, K, \tag{1.1}$$

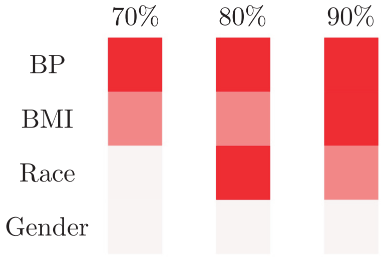

Figure 1. Diabetes patients' risk factors

where α_k's are the intercept terms across the models and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K \in \mathbb{R}^p$ are the parameter vectors for all models of interest. Our composite loss function is

$$L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)) \equiv \sum_{k=1}^{K} w_k L_k(y_i, \alpha_k + \boldsymbol{x}_i^T \boldsymbol{\beta}_k), \tag{1.2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T)^T \in \mathbb{R}^{K \times p}$, and $\boldsymbol{w} = (w_1, \ldots, w_K)^T$ is a positive weight vector. Minimizing (1.2) without further assumptions on parameter overlap is equivalent to minimizing the loss functions separately. As an example, in composite quantile regression (CQR), each L_k is a check function used to fit a model to a quantile Zou and Yuan (2008). Combining the check function for median regression (L_1) with the usual least squares loss function (L_2) is another.

Composite estimation is useful when the underlying parameter structures are partially overlapped. In partially overlapping models, some covariates have the same effect on the response across at least two models, while others do not. The CQR and L_1-L_2 loss functions may have overlapping parameters for different quantiles or median and expectation. Figure 1 shows a simple example of the partially overlapping models. Here possible risk factors to diabetes patients include blood pressure (BP), body mass index (BMI), race, and gender, and the response is the blood glucose level. Interest is in three patient groups with levels of blood glucose, 70%, 80%, and 90%. Each parameter vector corresponds to the check function for each quantile (β_1, β_2, and β_3). The BP has rows of the same color, which impart the same parameter values across all three quantiles. We call this arrangement overlapping structure. According to the definition of overlapping structure, the effects of BMI overlap across patients with 70% and 80% quantiles, and the effects of gender overlap across all three quantiles. The row of the gender appears white-shaded, which indicates that it is not a risk factor across all three levels.

A complete overlapping structure is one extreme of partially overlapping structures, where all parameters are common to all loss functions. For completely overlapping models, Bradic, Fan, and Wang (2011) and Zou and Yuan (2008) used composite loss functions with the goal of improving efficiency of the regression parameter estimators. They considered the composite loss function as an approximation to the unknown log-likelihood function of the error distribution Bradic, Fan, and Wang (2011) while ACME considers each loss component as a model targeting different profiles of the conditional distribution. The completely overlapping modeling in composite loss estimation can limit flexibility in statistical modelling. Consider a linear location-scale model whose several covariates affect the scale of response and error is centered to zero but not symmetric. Different loss functions estimate different parameters defined both by the mean and variance of the response. The parameters are the same for the covariates which have no effect on the variance function Carroll and Ruppert (1988). A parameter vector for L_2 is the same as the regression parameter vector of the model while a parameter vector for L_1 is the weighted sum of the regression parameter vector and the scale parameter vector.

We aim for efficient composite estimation under the partially overlapping structure, which can overcome the drawback of completely overlapping models and allow the flexibility of having different parameter values. To adapt such overlapping structure in the models, we incorporate penalization into (1.2). The penalty is applied to all absolute pairwise differences between coefficients corresponding to each covariate. In addition to this overlapping penalty, we also employ a penalty for sparse estimation. The objective function for our empirical composite loss function with double penalties is

$$\sum_{k=1}^{K}\sum_{i=1}^{n} w_k L_k(y_i, \alpha_k + \boldsymbol{x}_i^T\boldsymbol{\beta}_k) + n\sum_{k=1}^{K}\sum_{j=1}^{p} p_{\lambda_{1n}}(|\beta_{kj}|) + n\sum_{k<k'}\sum_{j=1}^{p} p_{\lambda_{2n}}(|\beta_{k'j} - \beta_{kj}|). \tag{1.3}$$

The penalty terms in (1.3) applied to each coefficient encourage sparsity by shrinking small coefficients toward zero. The penalty terms applied to the difference in the coefficients enable recovery of the overlapping structure by shrinking small differences toward zero. Penalization of the differences is used not for variable selection, but for selecting the overlapping structure across the multiple loss functions. The fused lasso Tibshirani et al. (2004) also has a sparse penalty term combined with a penalty term for pairwise differences to identify local consistency of coefficients in a single model.

In the sequel, we propose and study adaptive composite M-estimation (ACME) based on (1.3); it simultaneously shrinks toward the true overlapping model structure while estimating the shared coefficients in that structure. For the models from Figure 1, ACME automatically chooses risk factors strongly associated with high blood glucose levels and estimates their same effects across different levels. Our procedure yields estimators with improved efficiency by information combination across the models. It correctly selects both the true overlap structure and the true non-zero parameters with probability 1 in large samples. The parameter estimators hereby are oracle in the sense that they have the same distribution as the oracle estimator based on knowing the true model structure a prior.

The rest of the paper is organized as follows. In Section 2, we introduce notation for the distinct parameter vector across models, based on overlap in the $\boldsymbol{\beta}_k$'s, and define the oracle estimator. The large sample properties of the oracle estimator are established under partially overlapping models. Section 3 presents ACME for partially overlapping models and describes its implementation along with a discussion of its theoretical properties. Section 4 contains numerical results from an extensive simulation study, and Section 5 reanalyzes a well-known dataset on annual salaries of professional baseball players. Proofs and some numerical results are presented in a web-appendix.

2. Oracle M-estimator for Overlapping Models

2.1. Models and notations

We first consider the K separate models with their corresponding loss functions in (1.1). The risk function for the kth model is the expectation of the kth loss function, $R_k(\alpha_k, \boldsymbol{\beta}_k) = \mathbb{E}_{\boldsymbol{z}}[L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k)]$ for $\boldsymbol{\beta}_k \in \mathbb{R}^p,\ k = 1, \ldots, K$. The true parameter vector for the kth model is the minimizer of the corresponding risk function, $R_k(\alpha_k, \boldsymbol{\beta}_k)$, with $(\alpha_k^0, \boldsymbol{\beta}_k^{0T})^T = \operatorname*{argmin}_{(\alpha_k, \boldsymbol{\beta}_k^T)^T \in \Theta \subset \mathbb{R}^{p+1}} R_k(\alpha_k, \boldsymbol{\beta}_k)$. We estimate the parameter vector of each model by minimizing its corresponding loss function. We consider a stack of all parameter vectors across all models, and write the $K \cdot (p+1)$-dimensional true parameter vector as $(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})^T = (\alpha_1^0, \ldots, \alpha_K^0, \boldsymbol{\beta}_1^{0T}, \ldots, \boldsymbol{\beta}_K^{0T})^T$.

We describe the underlying parameter structure across the multiple models using set notation. Let $\mathcal{A}_k = \{j \in \{1, \ldots, p\} : \beta_{kj}^0 \neq 0\}$ be the index set of the non-zero parameters in the kth model and $\mathcal{A}_k^c = \{1, \ldots, p\} \backslash \mathcal{A}_k$ be its complement, with the underlying sparse structure for all models as $\mathcal{A}^0 \equiv \{\mathcal{A}_k\}_{k=1}^K$. With $\mathcal{A}^0$, we can decompose the parameters as the true zero parameters $\boldsymbol{\beta}_{\mathcal{A}_k^c} = [\beta_{kj}]_{j \in \mathcal{A}_k^c} \in \mathbb{R}^{|\mathcal{A}_k^c|}$, $k = 1, \ldots, K$, and the true non-zero and intercept parameters, $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}_{\mathcal{A}}^T)^T = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}_{\mathcal{A}_1}^T, \boldsymbol{\beta}_{\mathcal{A}_2}^T,\ \ldots,\ \boldsymbol{\beta}_{\mathcal{A}_K}^T)^T$, where $\boldsymbol{\beta}_{\mathcal{A}_k} = [\beta_{kj}]_{j \in \mathcal{A}_k}$. Let $\mathcal{O}_{kk'} = \{j \in \{1, \ldots, p\} : \beta_{kj}^0 = \beta_{k'j}^0 \neq 0\}$ be the index set of the same-valued non-zero parameters between $\boldsymbol{\beta}_k^0$ and $\boldsymbol{\beta}_{k'}^0$ for $k \neq k'$, with the underlying overlapping structure available as $\mathcal{G}^0 \equiv \{O_{kk'}\}_{k<k'}$. We can identify the underlying sparse and overlapping structure with the sparsity sets and the overlap sets.

For joint estimation, the composite loss function is taken as the linear combination of all loss functions with weights, as in (1.2), and the composite risk function as $R(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T) = \mathbb{E}\sum_{k=1}^{K} w_k L_k(\alpha_k, \boldsymbol{\beta}_k) = \sum_{k=1}^{K} w_k R_k(\alpha_k, \boldsymbol{\beta}_k)$. The minimizer of $R(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)$, $(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})^T = \operatorname{argmin}_{(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)^T \in \Theta \subset \mathbb{R}^{K\cdot(p+1)}} R(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)$, is the true parameter vector for all K models. The true non-zero and intercept parameter vector is the minimizer of the composite risk function restricted to the non-zero parameters with the overlapping constraint:

$$(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}_{\mathcal{A}}^{0T})^T = \operatorname*{argmin}_{(\boldsymbol{\alpha}^T,\boldsymbol{\beta}_{\mathcal{A}}^T)^T} \sum_{k=1}^{K} w_k \mathcal{R}_k(\alpha_k, \boldsymbol{\beta}_{\mathcal{A}_k}) \qquad (2.1)$$

$$\text{subject to } \beta_{A_k j} = \beta_{A_{k'} j}\ \forall j \in \mathcal{O}_{kk'},\ \forall k < k',$$

where $\mathcal{R}_k(\alpha_k, \boldsymbol{\beta}_{\mathcal{A}_k}) = \mathbb{E}_{\boldsymbol{z}} L_k(y, \alpha_k + \boldsymbol{x}^{kT}\boldsymbol{\beta}_{\mathcal{A}_k})$ and $\boldsymbol{x}_i^k = [\boldsymbol{x}_{ij}]_{j\in\mathcal{A}_k}$.

The oracle M-estimator of $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ for partially overlapping models is the unpenalized M-estimator obtained under the assumption that the sparsity and overlapping structure is known in advance, say $(\hat{\boldsymbol{\alpha}}^{oT}, \hat{\boldsymbol{\beta}}^{oT})^T$. It can be decomposed into its zero parameter and non-zero parameter parts: $\hat{\boldsymbol{\beta}}^o_{\mathcal{A}_k^c} = [\beta^o_{kj}]_{j\in\mathcal{A}_k^c} = \mathbf{0}_{|\mathcal{A}_k^c|} \in \mathbb{R}^{|\mathcal{A}_k^c|}$, $k = 1, \ldots, K$, and $(\hat{\boldsymbol{\alpha}}^{oT}, \hat{\boldsymbol{\beta}}^{oT}_{\mathcal{A}})^T = (\hat{\boldsymbol{\alpha}}^{oT}, \hat{\boldsymbol{\beta}}^{oT}_{\mathcal{A}_1}, \ldots, \hat{\boldsymbol{\beta}}^{oT}_{\mathcal{A}_K})^T \in \mathbb{R}^{K+\sum_{k=1}^{K}|\mathcal{A}_k|}$, where $\hat{\boldsymbol{\beta}}^o_{\mathcal{A}_k} = [\beta^o_{kj}]_{j\in\mathcal{A}_k}$. Since we know the sparsity pattern of the models, $\mathcal{A}_1^c, \ldots, \mathcal{A}_K^c$, we estimate the corresponding parameters as zeros. Analogous to the definition of the true parameters in (2.1), the oracle estimator to the non-zero parameters minimizes the empirical weighted multiple loss functions with the overlapping structure constraint: $(\hat{\boldsymbol{\alpha}}^{oT}, \hat{\boldsymbol{\beta}}^{oT}_{\mathcal{A}})^T = \operatorname{argmin}_{(\boldsymbol{\alpha}^T,\boldsymbol{\beta}_{\mathcal{A}}^T)^T} (1/n)\sum_{i=1}^{n}\sum_{k=1}^{K} w_k L_k(y_i, \alpha_k + \boldsymbol{x}_i^{kT}\boldsymbol{\beta}_{\mathcal{A}_k})$ subject to $\beta_{A_k j} = \beta_{A_{k'} j}$ $\forall j \in \mathcal{O}_{kk'}$, for any $k < k'$.

2.2. Distinct parametrization and distinct oracle M-estimator

The common parametrization in Section 2.1 includes the duplication of the same valued parameters from the overlapping structure. The left panel of Figure 2 shows an example of such redundant parametrization. We use two 4-dimensional parameter vectors, $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^4$, to describe the models. The first and second parameter pairs have the same values respectively ($\beta_{11} = \beta_{21}$ and $\beta_{12} = \beta_{22}$). We can use one parameter, θ_{11}, for β_{11} and β_{21}, and another parameter, θ_{21}, for β_{12} and β_{22} as in the right panel of Figure 2. Furthermore, this parametrization excludes the zero-valued parameters, β_{23} and β_{14}. We call such parametrization distinct parametrization or non-redundant parametrization. The underlying sparse and overlapping structure is imposed on the non-redundant parametrization. The parametrization is lower-dimensional formulation for the true parameter vector and the oracle M-estimator.

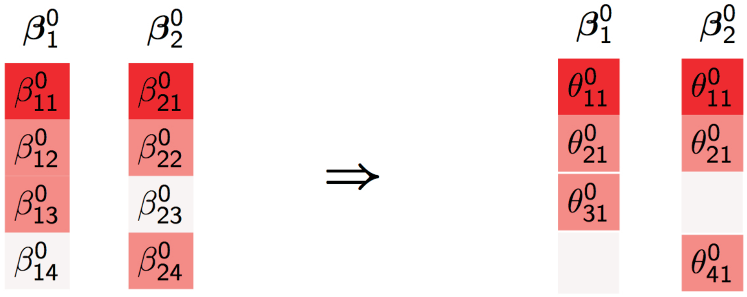

Figure 2. Illustration of distinct parametrization with $\beta^0_{14} = \beta^0_{23} = 0$.

To define our distinct oracle estimator, we borrow notation from Bondell and Reich (2007). Consider the union of the index sets of the non-zero parameters of all models, $\bigcup_{k=1}^K \mathcal{A}_k = \{j_1, \ldots, j_Q\}$; it corresponds to the index set of covariates with a non-zero true parameter in at least one model. Denote its cardinality as $Q = |\bigcup_{k=1}^K A_k| (\le p)$. Given a variable, x_{j_q}, $j_q \in \bigcup_{k=1}^K \mathcal{A}_k$, we consider the unique true non-zero parameter values among the elements of $\{\beta^0_{\mathcal{A}_k j_q} : \forall k \ s.t.\ j_q \in \mathcal{A}_k\}$. They are called the true distinct parameters to the variable, x_{j_q}.

Suppose we have the $G_q (\le K)$ true distinct parameters denoted as $\theta^0_{q1}, \ldots, \theta^0_{qG_q}$ for the variable, x_{j_q}. We denote the true distinct parameter vector across all covariates as $\boldsymbol{\theta}^0 = (\boldsymbol{\theta}^0_0, \boldsymbol{\theta}^0_1, \ldots, \boldsymbol{\theta}^0_Q)^T = (\theta^0_{01}, \ldots, \theta^0_{0K}, \theta^0_{11}, \ldots, \theta^0_{1G_1} \ldots, \theta^0_{Q1}, \ldots, \theta^0_{QG_Q})^T \in \mathbb{R}^{K+\sum_{q=1}^Q G_q}$, where $\boldsymbol{\theta}^0_0$ is the true intercept vector, $\boldsymbol{\alpha}^0$. This parameter vector is the non-redundant enumeration of the true parameters in terms of overlapping structure for all models along the predictors.

We define the distinct composite loss function with the non-redundant parametrization as $\mathcal{L}(\boldsymbol{z}_i, \boldsymbol{\theta}) = \sum_{k=1}^K w_k \mathcal{L}_k(y_i, \boldsymbol{\theta}_{0k} + \boldsymbol{x}_i^{kT} \boldsymbol{\beta}_{\mathcal{A}_k}(\boldsymbol{\theta}))$, where $[\boldsymbol{\beta}_{\mathcal{A}_k}(\boldsymbol{\theta})]_j$ is an element of $\boldsymbol{\theta}$ to $\boldsymbol{\beta}_{\mathcal{A}_k j}, j \in \mathcal{A}_k$. The distinct composite loss function is a random convex function on $\mathbb{R}^{K+\sum_{q=1}^Q G_q}$. The distinct composite risk function is the expectation of the distinct composite loss function with $\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{z}}[\mathcal{L}(\boldsymbol{z}, \boldsymbol{\theta})] = \sum_{k=1}^K w_k \mathcal{R}_k(\theta_{0k}, \boldsymbol{\beta}_{\mathcal{A}_k}(\boldsymbol{\theta}))$. The minimizer of the distinct composite risk function is the true distinct parameter vector.

The distinct oracle M-estimator of $\boldsymbol{\theta}$ is defined as the minimizer of the distinct loss function: $\hat{\boldsymbol{\theta}}^o = (\hat{\boldsymbol{\theta}}^0_0, \hat{\boldsymbol{\theta}}^0_1, \ldots, \hat{\boldsymbol{\theta}}^0_Q)^T = \text{argmin}_{\boldsymbol{\theta}} (1/n) \sum_{i=1}^n \mathcal{L}(\boldsymbol{z}_i, \boldsymbol{\theta}) \in \mathbb{R}^{K+\sum_{q=1}^Q G_q}$. We assume that the dimension of the distinct oracle M-estimator, $K + \sum_{q=1}^Q G_q$, is less than the sample size, n. The distinct oracle M-estimator can be viewed as the non-redundant enumeration of the oracle M-estimator, $(\hat{\boldsymbol{\alpha}}^{oT}, \hat{\boldsymbol{\beta}}^{oT}_{\mathcal{A}})^T$, in terms of overlaps. Specifically, every element of $\hat{\boldsymbol{\theta}}^o_q$ corresponds to one or some nonzero elements among $\hat{\beta}^o_{1j_q}, \ldots, \hat{\beta}^o_{Kj_q}$.

2.3. Asymptotic properties of distinct oracle M-estimator

We establish the asymptotic properties of the distinct oracle M-estiamtor. Some assumptions on the K separate loss functions are required.

A1. $(\alpha_k^0, \boldsymbol{\beta}_k^{0T})^T = \operatorname{argmin}_{(\alpha_k,\boldsymbol{\beta}_k^T)^T \in \Theta \subset \mathbb{R}^{p+1}} \mathbb{E} L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k),\ k = 1, \ldots, K$ are bounded and unique.

A2. $\mathbb{E} L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k) < \infty$ for each $(\alpha_k, \boldsymbol{\beta}_k^T) \in \mathbb{R}^{p+1},\ k = 1, \ldots, K$.

A3. (a) $L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k)$ is differentiable w.r.t. $(\alpha_k, \boldsymbol{\beta}_k^T)^T$ at $(\alpha_k^0, \boldsymbol{\beta}_k^0)$ for $\mathbb{P}_{\boldsymbol{z}}$-almost every $\boldsymbol{z} = (\boldsymbol{x}, y)$ with derivative $\nabla_{(\alpha_k,\boldsymbol{\beta}_k^T)^T} L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k^0)$ and $J_k(\alpha_k^0, \boldsymbol{\beta}_k^0) \equiv \mathbb{E}[\nabla_{(\alpha_k,\boldsymbol{\beta}_k^T)^T} L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k^0) \cdot \nabla_{(\alpha_k,\boldsymbol{\beta}_k^T)^T} L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k^0)^T] < \infty$.
(b) The risk function $R_k(\alpha_k, \boldsymbol{\beta}_k) = \mathbb{E}[L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k)]$ is twice differentiable w.r.t. $(\alpha_k, \boldsymbol{\beta}_k^T)^T$ at $(\alpha_k^0, \boldsymbol{\beta}_k^{0T})^T$ with a positive definite Hessian matrix, $H_k(\alpha_k^0, \boldsymbol{\beta}_k^0)$.

A4. The loss function, $L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k)$, is convex with respect to $(\alpha_k, \boldsymbol{\beta}_k^T)^T$ for $\mathbb{P}_{\boldsymbol{z}}$-almost every $\boldsymbol{z}$.

Similar conditions can be found for one model setting in Section 2.1 of Rocha, Wang, and Yu (2009). The assumption, A1, ensures that the parameter for the kth model, $(\alpha_k^0, \boldsymbol{\beta}_k^{0T})^T$, is well defined. The second assumption, A2, guarantees that the pointwise limit of the loss function is the risk function. From A3, we can consider local quadratic asymptotic approximations to the risk function around the parameter, approximating the loss function to the risk function at each point near the parameter. A4 is used to apply Convexity Lemma Pollard (1991) for the uniformity of approximation.

Lemma 1. *If $L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k),\ k = 1, \ldots, K$, satisfy* A1, …, A4*, then the composite loss function, $L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T))$ satisfies* A1, …, A4.

Lemma 2 is essential to proving consistency and asymptotic normality of the distinct oracle M-estimator and the $\sqrt{n}$-consistency, selection and overlapping consistency, and asymptotic normality of ACME.

Lemma 2. *If $L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k),\ k = 1, \ldots, K$, satisfy* A1–A4*, then*

(a) *there exists a $K \cdot (p+1)$ dimensional random vector $\boldsymbol{W} \sim N(0, J(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}))$ such that, for each $\boldsymbol{u} \in \mathbb{R}^{K\cdot(p+1)}$,*

$$\sum_{i=1}^{n} \Big[L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) + \frac{\boldsymbol{u}^T}{\sqrt{n}}) - L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})) \Big] \\ - \Big[\frac{1}{2} \boldsymbol{u}^T \cdot H(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) \cdot \boldsymbol{u} + \boldsymbol{W}^T \cdot \boldsymbol{u} \Big] \xrightarrow{p} 0.$$

(b) *for every compact set* $K \subset \mathbb{R}^{K\cdot(p+1)}$*,*

$$\sup_{\boldsymbol{u}\in K}\Bigg\| \sum_{i=1}^{n}[L(\boldsymbol{z}_i,(\boldsymbol{\alpha}^{0T},\boldsymbol{\beta}^{0T})+\frac{\boldsymbol{u}^T}{\sqrt{n}}) - L(\boldsymbol{z}_i,(\boldsymbol{\alpha}^{0T},\boldsymbol{\beta}^{0T}))] \\ -\Big[\frac{1}{2}\boldsymbol{u}^T\cdot H((\boldsymbol{\alpha}^{0T},\boldsymbol{\beta}^{0T}))\cdot\boldsymbol{u}+\boldsymbol{W}^T\cdot\boldsymbol{u}\Big]\Bigg\| \xrightarrow{p} 0.$$

Lemma 2 generalizes Lemma 2 of Rocha, Wang, and Yu (2009), which considers the setting of a single loss function. The distinct oracle M-estimator is a special type of M-estimators based on the distinct loss function. Lemma 3 shows consistency of the distinct oracle M-estimator.

Lemma 3. *If* A1−A4 *are satisfied for all* K *loss functions, then* $\hat{\boldsymbol{\theta}}^o$ *converges in probability to* $\boldsymbol{\theta}^0$ *as* $n\to\infty$*.*

Theorem 1. *If* A1−A4 *are satisfied for all* K *loss functions, then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^o-\boldsymbol{\theta}^0)\xrightarrow{d} N(0,\mathcal{H}(\boldsymbol{\theta}^0)^{-1}\mathcal{J}(\boldsymbol{\theta}^0)\mathcal{H}(\boldsymbol{\theta}^0)^{-1})),\ \text{as } n\to\infty,$$

where $[\mathcal{H}(\boldsymbol{\theta}^0)]_{ij}=\frac{\partial^2\mathcal{R}(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}$*, and* $\mathcal{J}(\boldsymbol{\theta}^0)=\mathbb{E}[\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{z},\boldsymbol{\theta}^0)\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{z},\boldsymbol{\theta}^0)^T]$*.*

The non-redundant oracle estimator across models asymptotically follows a normal distribution, similar to some oracle estimators based on a single model.

3. Adaptive Composite M-estimation for Overlapping Structure

We establish the theoretical properties of ACME when A1−A4 hold for all models. We develop the asymptotic theories based on the objective function in (1.3), which is denoted as $Q_n(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)$. In particular, we focus on the oracle properties of ACME for partially overlapping models. For $p_{\lambda_{1n}}(|t|)$ and $p_{\lambda_{2n}}(|t|)$ we consider folded concave penalties, one-step folded concave penalties, and weighted L_1 penalties Fan, Xue, and Zou (2014); Zou and Li (2008).

Lemma 4. *If* $\lambda_{1n}\to 0$*,* $\lambda_{2n}\to 0$ *for folded concave, one-step folded concave penalty functions, and* $\sqrt{n}\lambda_{1n}\to 0$*,* $\sqrt{n}\lambda_{2n}\to 0$ *for weighted* L_1 *penalty functions, there is a local minimizer of* $Q_n(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)$ *such that*

$$\sqrt{n}|(\hat{\boldsymbol{\alpha}}^T,\hat{\boldsymbol{\beta}}^T)^T-(\boldsymbol{\alpha}^{0T},\boldsymbol{\beta}^{0T})^T| = O_p(1).$$

If both $p_{\lambda_{1n}}(t)$ *and* $p_{\lambda_{2n}}(t)$ *are weighted* L_1 *penalty functions, then* $(\hat{\boldsymbol{\alpha}}^T,\hat{\boldsymbol{\beta}}^T)^T$ *is the unique global minimizer.*

Lemma 4 demonstrates the existence of a $\sqrt{n}$-consistent penalized M-estimator with a proper choice of λ_n. Theorem 2 implies that the ACME achieves selection consistency and overlapping consistency. The notion of overlapping consistency is analogous with that of selection consistency. For any index $j \in \mathcal{O}_{kk'}$ for any $k < k'$, both $\hat{\beta}_{kj}$ and $\hat{\beta}_{k'j}$ have the exactly same values with probability tending to 1.

Theorem 2. *Suppose that* $\lambda_{1n} \to 0$, $\lambda_{2n} \to 0$, $\sqrt{n}\lambda_{1n} \to \infty$, *and* $\sqrt{n}\lambda_{2n} \to \infty$ *for folded concave, one-step folded concave penalty functions. For weighted* L_1 *penalty functions, suppose* $\sqrt{n}\lambda_{1n} \to 0$, $\sqrt{n}\lambda_{2n} \to 0$, $n^{(s+1)/2}\lambda_{1n} \to \infty$, *and* $n^{(s+1)/2}\lambda_{2n} \to \infty$. *If there exists at least one* $j \in \mathcal{O}_{kk'}$ *for some* $k < k'$, *then* $P(\bigcap_{k=1}^{K} \bigcap_{j \in \mathcal{A}_k^c} \{\hat{\beta}_{kj} = 0\} \cap \bigcap_{k<k'} \bigcap_{j \in \mathcal{O}_{kk'}} \{\hat{\beta}_{kj} = \hat{\beta}_{k'j}\}) \to 1$ *as* $n \to \infty$.

Let $\hat{\mathcal{A}}_k = \{j \in \{1, \ldots, p\} : \hat{\beta}_{kj} \neq 0\}$ denote the non-zero coefficient index set corresponding to the kth loss function. Denote $\hat{\mathcal{G}}$ as the estimated grouping. The selection and overlapping consistency can be written as $P(\{\hat{\mathcal{A}}_k = \mathcal{A}_k, k = 1, \ldots, K\} \cap \{\hat{\mathcal{G}} = \mathcal{G}^0\}) \to 1$.

Let $\hat{\boldsymbol{\theta}}_{\mathcal{A}^0}(\mathcal{G}^0)$ denote our distinct ACME from (1.3) provided we know the true overlapping structure, $\mathcal{G}^0$, and the true sparse structure, $\mathcal{A}^0$. We study the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{\mathcal{A}^0}(\mathcal{G}^0)$ since our estimator recovers the true sparsity and overlapping structure with probability tending to one; its dimension is the dimension of the distinct oracle estimator.

Theorem 3. *If the assumptions in Theorem* 2 *are satisfied, then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}^0}(\mathcal{G}^0) - \boldsymbol{\theta}^0) \xrightarrow{d} N(0, \mathcal{H}(\boldsymbol{\theta}^0)^{-1}\mathcal{J}(\boldsymbol{\theta}^0)\mathcal{H}(\boldsymbol{\theta}^0)^{-1})).$$

Theorem 3 states that the distinct estimator has the same asymptotic distribution as the distinct oracle estimator in Theorem 1. The ACME across the multiple models follows a normal distribution in terms of non-zero non-redundant enumeration as the penalized estimators of a single model for the non-zero parameters follow a normal distribution Fan and Li (2001).

The asymptotic distribution of the distinct ACME in Theorem 3 leads to theoretically optimal weights to achieve the efficiency across the multiple models. The criterion for the choice of weights is to maximize the efficiency of the estimator Bradic, Fan, and Wang (2011). We can use the determinant of the asymptotic covariance matrix of the estimator or its trace as the criterion; its asymptotic covariance is a function of the unknown matrices of $\mathcal{J}(\boldsymbol{\theta}^0)$ and $\mathcal{H}(\boldsymbol{\theta}^0)$, and both depend on the weight vector, $\boldsymbol{w}$. Completely overlapping models also have an asymptotic normal distribution and their asymptotic covariance depends on the weight vector Bradic, Fan, and Wang (2011). In this setup, the asymptotic covariance matrix can be simplified as the multiplication of a scalar function and a function of predictors. Since the scalar function only takes the weight vector as its variable, the weight vector can be decoupled from the asymptotic covariance matrix. Bradic, Fan, and Wang (2011) chooses the weight vector by minimizing the scalar function. However, such decoupling cannot be obtained for partially overlapping models, due to the complex form of the asymptotic covariance.

To address the problem, we suggest a data dependent approach to select weights. We first obtain the separate penalized M-estimators as the initial separate estimators with $\hat{\boldsymbol{\beta}}_k^{(0)} = \text{argmin}_{(\alpha_k,\boldsymbol{\beta}_k^T)^T\in\Theta\subset\mathbb{R}^{p+1}} \sum_{i=1}^n L_k(y_i, \alpha_k + \boldsymbol{x}_i^T\boldsymbol{\beta}_k) + n\sum_{j=1}^p p_{\lambda_{1n}}(|\beta_{kj}|), k = 1, \ldots, K$. The preliminary M-estimators achieve sparse estimation, but do not attain overlapping estimation. Next we calculate data-driven weights, $\boldsymbol{w} = (w_1, \ldots, w_K)^T$ based on the preliminary estimators. We set w_k to be proportional to the reciprocal of the empirical loss function of the initial estimators with $w_k \propto [(1/n)\sum_{i=1}^n L_k(y_i, \alpha_k + \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_k^{(0)})]^{-1}$, $k = 1, \ldots, K$. We recommend this weight ratio for the same leverage of each loss function to the composite loss function. For computational efficiency, they are rescaled to have sum to one as $\sum_{k=1}^K w_k = 1$. We adopt this choice of weights in the numerical studies of Section 4, which yields excellent performance. We assume positive weights because the presence of a zero weight automatically removes the parameter vector of the corresponding model.

Next we solve the optimization problem, (1.3), with the plug-in weights. Zero-estimated parameters in the preliminary step can be estimated as non-zero in the ACME procedure. For implementation, we adopt one-step SCAD penalties and select a suitable algorithm with respect to the composite loss of interest. For example, we use a coordinate descent alrogithm for the L_1-L_2 composite loss. The optimization problem for ACME with the CQR can be recast as a linear programming problem with some slack variables Wu and Liu (2009). To obtain the optimal tuning parameters for λ_{1n} and λ_{2n}, we use five-fold cross validation. A two-dimensional grid search is performed for the selection of $(\lambda_{1n}, \lambda_{2n})$. A proper choice of the tuning parameters is required to simultaneously recover the sparsity and overlapping structure.

4. Simulation Studies

We performed simulation studies under a classical linear model and a linear location-scale model. Each dataset in Sections 4.1−4.2 was generated from these two models. We obtained ACME for both least absolute deviations (LAD) regression and least squares (LS) regression with a composite L_1-L_2 loss function. We compared it with separate LAD and LS estimators such as ordinary unpenalized LAD and LS estimators (Ordinary), adaptive Lasso penalized LAD and LS estimators (AdLasso), and one-step SCAD penalized LAD and LS estimators (SCAD). We also compared with penalized composite quasi-likelihood (PCQ) in

Bradic, Fan, and Wang (2011), which was developed for a classical linear model. PCQ assumes the completely overlapping structure across all loss functions. We employed one-step SCAD penalty for PCQ.

For comparison, we report the median of model errors (MME), the standard error of model errors (SE), the number of correctly classified non-zero estimators (TP), and the number of incorrectly classified zero estimators (FP). The model error of each estimator is defined as $ME(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T \mathbb{E}(\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$. We also evaluated the overlapping performance across the LAD and LS models. The overlapping structures are categorized into four types: truly grouped estimators, truly grouped non-zero estimators, truly grouped zero estimators, and truly ungrouped estimators, with the index set of the categories as TG, NG, ZG, and UG, respectively. We measured the performance of the overlapping recovery using overlapping ratios corresponding to these four categories. More details are provided in the web-appendix.

4.1. Classical linear regression model

We considered the classical linear model from Fan and Li (2001): $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}^0 + \epsilon_i$, where $\boldsymbol{\beta}^0 = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0)$. The covariate $\boldsymbol{x}_i$ was multivariate normal with zero mean and covariance, $\mathrm{Cov}(x_{ij_1}, x_{ij_2}) = 0.5^{|j_1 - j_2|}$, $1 \le j_1, j_2 \le 12$. We took the error term, $\epsilon_1, \ldots, \epsilon_n$, to follow a normal distribution ($N(0,3)$), a double exponential distribution (DE), and a t distribution with d.f. 4 ($t(4)$). We considered both LAD regression and LS regression. The true models were completely overlapped since the true parameter vector of the LS regression was the same as the true parameter vector of the LAD regression. For these models, both PCQ and ACME used the composite L_1-L_2 loss function. Our choice of weight for ACME was $(w_1, w_2) \propto (1/MAE(\hat{\alpha}_{lad}^{SCAD}, \hat{\boldsymbol{\beta}}_{lad}^{SCAD}), 1/MSE(\hat{\alpha}_{ls}^{SCAD}, \hat{\boldsymbol{\beta}}_{ls}^{SCAD}))$, with $MAE(\hat{\alpha}_{lad}^{SCAD}, \hat{\boldsymbol{\beta}}_{lad}^{SCAD})$ as the mean of absolute errors of the SCAD-LAD estimator, and $MSE(\hat{\alpha}_{ls}^{SCAD}, \hat{\boldsymbol{\beta}}_{ls}^{SCAD}))$ as the mean of squared errors of the SCAD-LS estimator. The results were obtained from 100 simulated datasets with $n = 100$ and $n = 500$.

From the first three columns of Tables 1−2, the performance of ACME is the best for both L_1 and L_2 under DE error with $n = 100, 500$ and under $t(4)$ with $n = 100$ in terms of MME. Under $N(0,3)$ with $n = 100, 500$, the MMEs of the PCQ are smaller than those of ACME, but ACME outperforms the others. In this setting, PCQ is generally comparable to ACME because PCQ achieves the oracle overlapping structure. All estimators successfully selected the significant variables, β_1^0, β_2^0, β_5^0, as evidenced by TP. ACME performed the best in terms of FP in most cases.

For the overlaps, we had TG$= \{1, 2, \ldots, 11, 12\}$, NG$=\{1, 2, 5\}$, ZG$=\{3, 4, 6, \ldots, 12\}$ and UG$= \emptyset$. In the first three rows of Table $S1$ in the web-appendix,

Table 1. Simulation results with model errors and numbers of correct nonzeros/incorrect zeros ($n = 100$).

		N(0,3)	DE	t(4)	LLS
	Estimation	MME	MME	MME	MME
		(TP, FP)	(TP, FP)	(TP, FP)	(TP, FP)
LAD	Oracle	0.1192	0.0484	0.0482	0.4853
		(3, 0)	(3, 0)	(3, 0)	(10, 0)
	Ordinary	0.5643	0.34	0.2493	0.9383
		(3, 9)	(3, 9)	(3, 9)	(10, 8)
	AdLasso	0.2713	0.1115	0.1008	0.7472
		(3, 2.52)	(3, 1.84)	(3, 2.44)	(9.97, 2.42)
	SCAD	0.2632	0.091	0.1014	0.6476
		(3, 2.48)	(3, 1.59)	(3, 2.17)	(9.96, 1.56)
	PCQ oracle	0.0738	0.067	0.0386	6.8094
		(3, 0)	(3, 0)	(3, 0)	(7, 0)
	PCQ	0.1395	0.1356	0.0981	14.3802
		(3, 1.97)	(3, 3.72)	(3, 3)	(9.59, 7.1)
	ACME oracle	0.0786	0.0642	0.0411	0.6278
		(3, 0)	(3, 0)	(3, 0)	(10, 0)
	ACME	0.1761	0.085	0.0694	0.6717
		(3, 1.62)	(3, 1.16)	(3, 1.4)	(9.78, 1.03)
LS	Oracle	0.0727	0.0881	0.0428	2.866
		(3, 0)	(3, 0)	(3, 0)	(7, 0)
	Ordinary	0.3892	0.3871	0.2794	10.0807
		(3, 9)	(3, 9)	(3, 9)	(7, 11)
	AdLasso	0.1569	0.1647	0.1054	6.0877
		(3, 1.79)	(3, 1.88)	(3, 1.83)	(6.88, 3.26)
	SCAD	0.1436	0.1719	0.1038	6.4209
		(3, 1.96)	(3, 2.11)	(3, 2.06)	(6.88, 4.82)
	PCQ oracle	0.0738	0.067	0.0386	1.6273
		(3, 0)	(3, 0)	(3, 0)	(7, 0)
	PCQ	0.1395	0.1356	0.0981	8.3698
		(3, 1.97)	(3, 3.72)	(3, 3)	(7, 9.69)
	ACME oracle	0.0786	0.0642	0.0411	1.48
		(3, 0)	(3, 0)	(3, 0)	(7, 0)
	ACME	0.1434	0.1238	0.0802	5.3363
		(3, 1.63)	(3, 1.41)	(3, 1.51)	(6.85, 2.38)

ACME has reasonable ratios of NG as well as ZG. Most ZGs are higher than NGs since the two penalty terms for overlapping and sparsity encourage increase in the ZG ratio. We can view the NG ratio as a more accurate measure on the performance of the overlapping penalization than the ZG ratio. The ZG ratio of ACME is almost 30% higher than that of all separate estimators under $n = 100$ and $n = 500$. ACME has almost two thirds the NG ratio, except for the nor-

Table 2. Simulation results with model errors and numbers of correct nonzeros/incorrect zeros ($n = 500$).

		N(0,3)	DE	t(4)	LLS
	Estimation	MME	MME	MME	MME
		(TP, FP)	(TP, FP)	(TP, FP)	(TP, FP)
LAD	Oracle	0.0255	0.0072	0.0072	0.0453
		(3, 0)	(3, 0)	(3, 0)	(10, 0)
	Ordinary	0.1074	0.0409	0.0403	0.0589
		(3, 9)	(3, 9)	(3, 9)	(10, 7.99)
	AdLasso	0.0453	0.0134	0.0148	0.0544
		(3, 1.69)	(3, 1.52)	(3, 1.79)	(10, 1.17)
	SCAD	0.0393	0.0126	0.0132	0.0489
		(3, 1.53)	(3, 1.42)	(3, 1.58)	(10, 0.85)
	PCQ oracle	0.014	0.0082	0.0074	5.6941
		(3, 0)	(3, 0)	(3, 0)	(7, 0)
	PCQ	0.0174	0.0224	0.0148	8.8911
		(3, 1.12)	(3, 3.38)	(3, 2.56)	(9.99, 7.85)
	ACME oracle	0.0156	0.0088	0.0071	0.059
		(3, 0)	(3, 0)	(3, 0)	(10, 0)
	ACME	0.0311	0.0108	0.01	0.0542
		(3, 0.82)	(3, 1.17)	(3, 1.14)	(10, 0.3)
LS	Oracle	0.0135	0.0133	0.0096	0.6803
		(3, 0)	(3, 0)	(3, 0)	(7, 0)
	Ordinary	0.0712	0.0671	0.0471	1.7359
		(3, 9)	(3, 9)	(3, 9)	(7, 11)
	AdLasso	0.0229	0.0238	0.0178	1.0036
		(3, 1.16)	(3, 1.27)	(3, 1.39)	(7, 2.31)
	SCAD	0.0191	0.024	0.012	1.1313
		(3, 1.22)	(3, 1.56)	(3, 1)	(7, 3.39)
	PCQ oracle	0.014	0.0082	0.0074	1.4777
		(3, 0)	(3, 0)	(3, 0)	(7, 0)
	PCQ	0.0174	0.0224	0.0148	1.5568
		(3, 1.12)	(3, 3.38)	(3, 2.56)	(7, 10.84)
	ACME oracle	0.0156	0.0088	0.0071	0.2633
		(3, 0)	(3, 0)	(3, 0)	(7, 0)
	ACME	0.0189	0.0206	0.0132	0.7471
		(3, 0.92)	(3, 1.32)	(3, 1.28)	(7, 1.01)

mal distribution with $n = 100$. Ordinary, AdLasso, and SCAD have zero NG ratios because the separate estimation does not involve any overlapping penalization. PCQ possesses complete overlapping because the dataset is assumed to be generated from a classical linear model. Hence, PCQ successfully recovers the overlapping structure.

4.2. Linear location-scale model

Under linear location-scale models, LS regression and LAD regression are partially overlapping models as some covariates affect the scale of the response. Our dataset was generated from a linear location-scale model: $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}^0 + \boldsymbol{x}_i^T\boldsymbol{\gamma}^0\epsilon_i$, where $\boldsymbol{\beta}^0 = (3,3,3,3,3,3,3,0,0,0,0,0,0,0,0,0,0,0)^T$ and $\boldsymbol{\gamma}^0 = (0,0,0,0,3,-3,3,-3,3,-3,0,0,0,\ 0,0,0,0,0)^T$. The covariate, $\boldsymbol{x}_i = (x_{i1},\ \ldots,\ x_{i18})^T$, was generated from a multivariate standard normal distribution, $N(\mathbf{0}, I_{18\times18})$. We took that the error term, ϵ_i, as a shifted gamma distribution, $\Gamma(0.25,\ 2)-0.5$. This distribution is skewed to the right and centered to mean 0. The true parameter vector of the LS regression model was $\boldsymbol{\beta}^0_{ls} = \boldsymbol{\beta}^0$ and the true parameter vector of LAD regression model was $\boldsymbol{\beta}^0_{lad} = (3,3,3,3,1.762,4.238,1.762,1.238,-1.238,1.238,0,0,0,0,0,0,0,0)^T$. As in Section 4.1, we used the composite L_1-L_2 loss function. We did 100 repetitions for $n = 100$ and $n = 500$.

From the last columns of Tables 1−2, the ACME has the second smallest MME for LAD regression, and the smallest MME for LS regression with $n = 100, 500$. The SCAD has the smallest MME for LAD and the SCAD has the second smallest MME for LS. The separate estimators and the ACME show much better performance for the LAD regression than the LS regression due to the skewed error distribution. From this point of view, it is desirable to have a trade-off between LAD and LS estimation performance in ACME. The ACME sacrifices LAD estimation performance about 5% with $n = 100$, and 10% with $n = 500$, while it gains in LS estimation performance almost 15% with $n = 100$, and 30% with $n = 500$. Overall, ACME has very competitive performance in terms of MME, sparsity and overlapping structure recovery. The performance of PCQ is poor as expected since the LAD and LS regression models are assumed to be completely overlapped.

The grouping performance results under this model are summarized at the bottom of Table $S1$ in the web-appendix. We have TG= $\{1,2,3,4,11,\ldots,18\}$, NG= $\{1,2,3,4\}$, ZG= $\{11,\ldots,18\}$ and UG= $\{5,6,\ldots,10\}$. ACME has much higher TG, NG, and ZG ratios than separate estimation. Both NG and ZG ratios increase as the sample size increases. ACME also has higher UG ratio, whose oracle target is zero, but the ratio drops to 0.005 from 0.2217 as the sample size is increased to $n = 500$ from $n = 100$. PCQ shows good performance for underlying grouped variables (TG, NG, ZG), while it groups the variables which are not truly overlapped (UG).

5. Baseball Data Analysis

We analyzed the major league baseball (MLB) players' annual salary dataset, obtained from `http://lib.stat.cmu.edu`. We were interested in the salary determinants of low-paid, median-paid, and highly-paid players respectively. We obtained ACME for three quantile regression models to the quantiles, 0.25, 0.5, 0.75. The dataset consists of the records and information on 263 North American MLB players in 1986 season and their salary in 1987 season. This dataset was previously studied by He, Ng, and Portnoy (1998) and Li, Liu, and Zhu (2007). They assumed that salary is a function of only the number of home runs in the previous year (HR) and the number of years in MLB (YEARS).

In addition to HR and YEARS, we considered covariates such as their performance in the previous years and their league, division, and position information. The response is the annual salary on opening day in 1987, in thousands of dollars. The first seven predictors were the number of hits (HIT), the number of runs (RUN), the number of runs batted in (RBI), the number of walks (WALK), the number of put outs (PUTOUT), the number of assists (ASSIST), and the number of errors (ERROR). We employed seven dummy variables for league and division, and position information: National East (NE), National West (NW), American East (AE), Infielder (IN), Outfielder (OUT), Catcher (CC), and Designated Hitter (DH). We treated American West (AW) and Utility Players (UP) as the base groups of the league and division, and position, respectively. We dropped the players' batting in 1986 (BAT) since BAT is highly correlated with such other variables as HIT, HR, RUN, RBI, and WALK. Especially, the correlation between the BAT and HIT is 0.9640. Most of the correlations among the performance records during career are almost 0.9, which indicates severe collinearity.

Our goal was to determine important covariates on the first, second, and third quantiles of the players' salaries. We used a CQR loss function for the analysis with the quantile vector, $\tau = (0.25, 0.5, 0.75)$, corresponding to the low-paid, median-paid, and highly-paid players. We performed separate quantile regression estimation methods, PCQ, and ACME. The separate regression methods included ordinary, adaptive Lasso and one-step SCAD penalized quantile regression estimation.

ACME provides interpretable results by grouping the similar effects across the different quantiles. The results are summarized in Table 3. ACME selects HIT, YEARS, PUTOUT, league and division, and positions across the three quantiles. The second quantile regression model is partially overlapped with the third quantile regression for the three covariates: HIT, YEARS, and PUTOUT; they are seen to have the same strength of impacts on the median-paid and highly-paid baseball players' salary; their effects are weaker in the case of low-paid players' salaries. HR was found to be significant only for the highly-paid players. The other coefficients, such as RUN and RBI, go to zero across all quantiles. WALK and ASSIST are non-zero in the preliminary estimator for the third quantile, but they go to zero in the ACME procedure.

Table 3. Regression coefficients for the baseball dataset.

	Ordinary (SE)	Sig.	AdLasso	SCAD	PCQ	ACME
(Intercept)	-245.5120 (73.4387)		3.5418	-219.1371	-515.5512	-222.0246
HIT	0.7907 (1.7183)		0	1.2864	2.9716	1.2815
HR	-5.3061 (4.9069)		0	0	2.0697	0
RUN	1.8274 (2.7044)		1.2953	0	0	0
RBI	2.4403 (2.6514)		0.2118	0	0	0
WALK	0.7804 (1.5287)		0	0	2.4083	0
YEARS	30.2551 (4.3717)	(**)	25.0385	31.0540	34.7556	31.2286
PUTOUT	-0.0890 (0.0978)		0	0.0015	0.1878	0.0118
ASSIST	-0.1639 (0.2459)		0	0	-0.0423	0
ERROR	-4.0178 (4.5298)		0	0	-5.4835	0
NE	-0.3179 (50.4264)		0	0	119.9565	0
NW	14.4817 (46.9023)		0	24.2768	49.2665	19.6613
AE	45.4914 (48.4438)		0	38.5924	94.8061	40.6199
IN	158.2192 (70.4252)	(**)	0	131.8874	146.3136	130.0462
OUT	103.3899 (71.0636)		0	163.0241	104.6079	160.9292
CC	192.0264 (75.5660)	(**)	0	144.9067	180.0394	147.7828
DH	-79.9613 (122.5664)		0	-10.3131	-37.9423	-11.7313
(Intercept)	-433.8376 (70.6211)		-377.3501	-350.5207	-389.9337	-345.8087
HIT	4.0231 (1.5517)	(**)	2.9242	2.9508	2.9716	2.9707
HR	6.6351 (6.2462)		2.5825	0	2.0697	0
RUN	-1.8305 (2.7047)		0	0	0	0
RBI	-1.4046 (2.5405)		0	0	0	0
WALK	2.0973 (1.3878)		1.7366	0	2.4083	0
YEARS	40.8095 (4.6872)	(**)	38.4487	42.1105	34.7556	42.5428
PUTOUT	0.2477 (0.1416)	(*)	0.2641	0.3109	0.1878	0.2662
ASSIST	-0.2267 (0.2770)		-0.0258	0	-0.0423	0
ERROR	-1.8804 (4.0841)		-0.5691	0	-5.4835	0
NE	108.5532 (52.4478)	(**)	93.8747	128.5615	119.9565	130.8570
NW	12.7587 (47.3871)		0	29.9324	49.2665	32.3740
AE	40.8497 (45.3921)		23.4914	81.1657	94.8061	73.9340
IN	190.6089 (78.3024)	(**)	89.7357	54.3756	146.3136	66.6862
OUT	136.6861 (62.7354)	(**)	95.4291	104.7711	104.6079	103.4506
CC	145.0478 (81.5529)	(*)	103.9739	80.8829	180.0394	90.0636
DH	-1.8963 (133.4392)		0	0	-37.9423	0
(Intercept)	-391.8350 (81.0963)		-361.7759	-399.4956	-245.9810	-374.7126
HIT	4.8975 (2.1460)	(**)	4.1554	3.4490	2.9716	2.9707
HR	13.3862 (7.9316)	(*)	12.4493	9.6505	2.0697	13.0354
RUN	-2.4222 (3.7428)		-1.4637	0	0	0
RBI	-1.9237 (3.7097)		-1.6779	0	0	0
WALK	3.2575 (1.9991)		3.5655	1.9914	2.4083	0
YEARS	39.3092 (6.4817)	(**)	41.4364	40.8961	34.7556	42.5428
PUTOUT	0.2982 (0.1529)	(*)	0.3053	0.2727	0.1878	0.2662
ASSIST	-0.6020 (0.3831)		-0.5430	-0.3295	-0.0423	0
ERROR	-1.7205 (6.3196)		-0.4648	0	-5.4835	0
NE	172.1072 (61.4199)	(**)	151.9045	156.2564	119.9565	183.7244
NW	46.0431 (60.8648)		33.0716	54.2641	49.2665	66.9276
AE	112.6242 (70.0346)		95.4571	101.6325	94.8061	82.9392
IN	224.1558 (100.9911)	(**)	164.4403	137.2256	146.3136	120.8592
OUT	62.4650 (87.4832)		42.4966	86.4714	104.6079	149.6180
CC	49.1510 (106.0594)		17.4022	63.3216	180.0394	91.7776
DH	-129.9760 (216.2998)		-174.4692	-69.4182	-37.9423	7.7997

Note. (**) indicates significant level 0.05 and (*) indicates significant level 0.1.

Table 4. Test errors of baseball dataset for three quantiles

	Ordinary	AdLasso	SCAD	PCQ	ACME
Q1	75.9326	75.9482	72.7219	81.9500	74.2914
Q2	106.4342	105.3914	106.0978	103.6183	105.4999
Q3	92.7157	92.3668	93.7098	93.7860	91.9224

The players' position was shown to be another important factor on the annual salary. Across all quantiles, the outfielders (OUT) are seen as the most-paid position. The catchers' (CC) and the infielders' (IN) salaries are the second and third highest, and the designated hitters (DH) and the utility players (UP) have the second-lowest and lowest salaries. Similar to position, we can analyze the league and division factor on the players' salaries. Table 3 also reports the standard errors of the ordinary coefficients and their significance. They were obtained from the Markov chain marginal bootstrap (MCMB) with 500 repetitions (Kocherginsky, He, and Mu, 2005). ACME selects all variables known to be significant by MCMB under the significance level of 0.1.

Table 4 shows the test errors for all estimation procedures from 10 repetitions. In each iteration, randomly selected 28 data points were assigned as a test set and the remaining 235 data points were assigned as a training set. ACME outperformed the ordinary quantile regression models at all quantiles. Compared with the other penalized estimators, ACME had better performance at two of the three quantiles. The performance of PCQ was substantially biased at the first quantile. Because PCQ assumes complete overlapping models, the first quantile regression modeling was dragged upward toward other two quantiles.

6. Concluding Remarks

We have proposed adaptive composite estimation for partially overlapping models, first introducing the notion of partially overlapping regression models on a given dataset. Overlapping structure has the same effect of a covariate on the response across multiple models. Partially overlapping models have at least one overlapping structure. We have also considered the sparse structure of the regression parameters for all models. ACME achieves both goals with a doubly penalized composite loss function. Its regular penalty function encourages sparse structure recovery while the other penalty function induces the overlapping structure recovery. The arguments of the second penalty function are all pairwise differences of the coefficients for each covariate across the models. We have shown its selection and overlapping consistency under the proper choice of the tuning parameters. We have also established the asymptotic normality of non-redundant ACME, given the true sparse and overlapping structure. In numerical studies, ACME has outperformed the separate penalized M-estimation and the composite

M-estimation under the complete overlapping structure assumption. Our study has focused on a moderate number of covariates and a moderate number of loss functions due to computational burden. Extension to high-dimensional covariates and models requires future research.

Supplementary Materials

Supplementary materials available at the *Statistica Sinica* journal website include additional results for simulated examples, and proofs of theorems.

Acknowledgement

The authors are indebted to Joint Editor Professor Naisyin Wang, an associate editor, and three referees, whose helpful comments and suggestions led to a much improved presentation. The authors are supported in part by NSF grant DMS-1407241, and NIH grants NIH/NCI R01 CA-149569 and NIH/NCI P01 CA142538.

References

Bondell, H. and Reich, B. (2007). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics* **64**, 115-123.

Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. Roy. Statist. Soc. Ser. B* **73**, 325-349.

Carroll, R. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman & Hall/CRC.

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

Fan, J., Xue, L. and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42**, 819-849.

He, X., Ng, P. and Portnoy, S. (1998). Bivariate quantile smoothing splines. *J. Roy. Statist. Soc. Ser. B* **60**, 537-550.

Kocherginsky, M., He, X. and Mu, Y. (2005). Practical confidence intervals for regression quantiles. *J. Comput. Graph. Statist.* **14**, 41-55.

Li, Y., Liu, Y. and Zhu, J. (2007). Quantile regression in reproducing kernel hilbert spaces. *J. Amer. Statist. Assoc.* **102**, 255-268.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econom. Theory*, **7**, 186-199.

Rocha, G., Wang, X. and Yu, B. (2009). Asymptotic distribution and sparsistency for $l1$-penalized parametric m-estimators with applications to linear svm and logistic regression. arXiv preprint arXiv:0908.1940v1.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2004). Sparsity and smoothness via the fused lasso. *J. Roy. Statist. Soc. Ser. B* **67**, 91-108.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statist. Sinica*, **19**, 801-817.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36**, 1509-1533.

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36**, 1108-1126.

Department of Statistics, Department of Biostatistics and Medical Informatics, The University of Wisconsin-Madison, Madison, WI 53706, U.S.A.

E-mail: shin@stat.wisc.edu

Department of Biostatistics, Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.

E-mail: jfine@bios.unc.edu

Department of Statistics and Operations Research, Department of Biostatistics, Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.

E-mail: yfliu@email.unc.edu

(Received June 2014; accepted January 2015)