

A Cluster Elastic Net for Multivariate Regression

Bradley S. Price

*College of Business and Economics
West Virginia University
Morgantown, WV 26505, USA*

BRAD.PRICE@MAIL.WVU.EDU

Ben Sherwood

*School of Business
University of Kansas
Lawrence, KS 66045, USA*

BEN.SHERWOOD@KU.EDU

Editor: Sara van de Geer

Abstract

We propose a method for simultaneously estimating regression coefficients and clustering response variables in a multivariate regression model, to increase prediction accuracy and give insights into the relationship between response variables. The estimates of the regression coefficients and clusters are found by using a penalized likelihood estimator, which includes a cluster fusion penalty, to shrink the difference in fitted values from responses in the same cluster, and an L_1 penalty for simultaneous variable selection and estimation. We propose a two-step algorithm, that iterates between k-means clustering and solving the penalized likelihood function assuming the clusters are known, which has desirable parallel computational properties obtained by using the cluster fusion penalty. If the response variable clusters are known *a priori* then the algorithm reduces to just solving the penalized likelihood problem. Theoretical results are presented for the penalized least squares case, including asymptotic results allowing for $p \gg n$. We extend our method to the setting where the responses are binomial variables. We propose a coordinate descent algorithm for the normal likelihood and a proximal gradient descent algorithm for the binomial likelihood, which can easily be extended to other generalized linear model (GLM) settings. Simulations and data examples from business operations and genomics are presented to show the merits of both the least squares and binomial methods.

Keywords: Multivariate Regression, Clustering, Fusion Penalty

1. Introduction

In this article we consider the pair $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$, with $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip}) \in \mathcal{R}^p$ and $\mathbf{y}_i = (y_{i1}, \dots, y_{ir})^T \in \mathcal{R}^r$. Define $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T \in \mathcal{R}^{n \times p}$ and $Y = (\mathbf{y}_1, \dots, \mathbf{y}_n)^T \in \mathcal{R}^{n \times r}$. We initially assume the linear model

$$\mathbf{y}_i = B^{*T} \mathbf{x}_i + \boldsymbol{\epsilon}_i, \quad (1)$$

where $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{ir})^T \in \mathcal{R}^r$ are realizations of an i.i.d. random variable with mean zero and covariance matrix Σ , $B^* = (\boldsymbol{\beta}_1^*, \dots, \boldsymbol{\beta}_r^*) \in \mathcal{R}^{p \times r}$ and $\boldsymbol{\beta}_k^* = (\beta_{1k}^*, \dots, \beta_{pk}^*)^T \in \mathcal{R}^p$. We will refer to the matrix of error as $E = (\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_n)^T \in \mathcal{R}^{n \times r}$. Under mild assumptions a

consistent estimator of β_k^* is the ordinary least squares (OLS) estimator of

$$\tilde{\beta}_k = \operatorname{argmin}_{\beta_k} \sum_{i=1}^n (y_{ik} - \mathbf{x}_i^T \beta_k)^2.$$

If ϵ_i are i.i.d. and $\epsilon_i \sim N(\mathbf{0}_r, \Sigma)$ the estimator $\tilde{\beta}_k$ is the MLE. This estimator does not use the other responses, ignoring potentially useful information.

Throughout this paper for a vector \mathbf{a} define $\|\mathbf{a}\|_q$ as the L_q norm and for a matrix A we define $\|A\|_q$ as the entrywise L_q norm. If there is *a priori* information that the fitted values of response k and m should be close then we could impose a penalty on the difference in the fitted values and consider the estimators

$$(\tilde{\beta}_k, \tilde{\beta}_m) = \operatorname{argmin}_{\beta_k, \beta_m} \sum_{i=1}^n \{(y_{ik} - \mathbf{x}_i^T \beta_k)^2 + (y_{im} - \mathbf{x}_i^T \beta_m)^2\} + \frac{\gamma}{n} \|X(\beta_k - \beta_m)\|_2^2, \quad (2)$$

where γ is a tuning parameter controlling the amount of agreement between the two fitted values vectors. We propose an objective function that generalizes (2) for multiple responses from multiple clusters that may not be known *a priori*. The proposed objective function also includes an L_1 penalty for simultaneous estimation and variable selection, which allows our method to be used to increase prediction accuracy, select relevant variables for each response, and detect groupings of response variables without assuming or estimating a covariance structure. In our theory, simulations, and applied examples we consider cases where $p \gg n$. We extend the proposed method to the generalized linear model framework, specifically focusing on multiple binary responses. This extension allows the method to be used in many different contexts, such as understanding co-morbidities related to patient information recorded in electronic medical records, or product level purchasing habits of customers based on information obtained from a loyalty program. We propose a coordinate descent algorithm for the least squares case and proximal coordinate descent algorithm for the binomial GLM case, which provides a general framework for extending the method to other GLM or M-estimator settings.

Our work has been influenced by previous work in estimating high dimensional models. When $\frac{1}{n} X'X = I_p$ the penalty function is equivalent to a ridge penalty (Hoerl and Kennard, 1970) on the difference of the coefficient vectors for the two responses. We add the L_1 penalty as proposed in Tibshirani (1996) to do simultaneous variable selection and estimation. Similar to the work of Zou and Hastie (2005) we combine the ridge and L_1 penalties. The proposed estimator simultaneously estimates clusters of the response and fuses the fitted values of the clustered responses. Previous work has been done on clustering covariates for high dimensional regression with a univariate response. This work is most similar to the work of Witten et al. (2014) who proposed the cluster elastic net (CEN) that simultaneously estimates clusters of covariates and fuses the effects of covariates within the same cluster. Our proposed method is also similar to Grace estimators proposed in Li and Li (2008) and Li and Li (2010), which use regularization based on external network information to minimize the difference of coefficients for related predictors and use a lasso penalty for sparsity. Huang et al. (2011) proposed the sparse Laplacian shrinkage method, which performs variable selection and promotes similarities among coefficients of correlated covariates. Zhao and Shojaie (2016) proposed the Grace Test, a testing framework for Grace estimators, that

allows for some uncertainty in the graph and showed that if the external graph is informative it increases the power of the Grace test. Bühlmann et al. (2013) proposed two different penalized methods for clustered covariates in high-dimensional regression: cluster representative lasso (CRL) and cluster group lasso (CGL). In CRL the covariates are clustered, dimension reduction is done by replacing the original covariates with the cluster centers and a lasso model is fit using the cluster centers as covariates. In CGL the group penalty of Yuan and Lin (2005) is applied using the previously found clusters as the groups. Zhou et al. (2017) demonstrated that averaging over models using different cluster centers for both responses and predictors can improve prediction accuracy of DNase I hypersensitivity using gene expression data. Kim et al. (2009) proposed graph-guided fused lasso (GGFL) to the specific problem of association analysis to quantitative trait networks. GGFL presents a fused lasso framework in multivariate regression that leverages correlated traits based on a network structure. Our work is related to the fused lasso literature as well, though we do not achieve exact fusion (Tibshirani et al., 2005; Rinaldo, 2009; Hoefling, 2010; Tibshirani, 2014). The proposed method differs from the works mentioned in this setting because it focuses on using correlation between the response variables to improve estimation, however all of the works mentioned were instrumental in helping us derive our final estimator.

The idea of using information from different responses to improve estimation in multivariate regression is not new and our work builds upon previous works in this area. Breiman and Friedman (1997) introduced the Curds and Whey method whose predictions are an optimal linear combination of least squares predictions. Rothman et al. (2010) proposed multivariate regression with covariance estimation (MRCE), which is a penalized likelihood approach to simultaneously estimate the regression coefficients and the inverse covariance matrix of the errors. MRCE leverages correlation in unexplained variation to improve estimation, while our proposed method leverages correlation in explained variation to improve estimation. Other estimators assume both the response and covariates are multivariate normal and exploit this structure to derive estimators (Lee and Liu, 2012; Molstad and Rothman, 2016). Rai et al. (2012) proposed a penalized likelihood method for multivariate regression that simultaneously estimates regression coefficients, the inverse covariance matrix of the errors, and the covariance matrix of the regression coefficients across responses using lasso type penalties. Peng et al. (2010) introduced regularized multivariate regression for identifying master predictors (remMap), which relies on *a priori* information about valuable predictors and imposes a group L_1 and L_2 norm, across responses, on all covariates not prespecified as being useful predictors. Kim and Xing (2012) proposed the tree guided group lasso, which uses an *a priori* hierarchical clustering of the responses to define overlapping group lasso penalties for the multivariate regression model. They propose a weighting method that ensures all coefficients are penalized equally, while using the hierarchical structure to impose a similar sparsity structure across highly correlated responses.

Another approach to improving efficiency is by doing dimension reduction on Y to find a smaller subspace that retains the material information needed for estimation of the regression coefficients (Cook et al., 2010; Cook and Zhang, 2015; Sun et al., 2015). Cook et al. (2010) introduced the envelope estimator for the multivariate linear model, which projects the maximum likelihood estimator onto the estimated subspace with the material information. Cook and Zhang (2015) provided envelope models for GLMs and weighted least squares. Sun et al. (2015) proposed a sparse regression model (SPReM) for estimating mod-

els where r is very large. SPReM projects the response variables into a lower-dimensional space while maintaining the structure needed for a specific hypothesis test. The key difference between our proposed method and these approaches is that we are interested in simultaneously estimating clustering of the response variables and fusing the fitted values from responses within the same cluster.

The proposed method simultaneously estimates clusters of the response and coefficients. Changes in cluster groups are discrete changes and as a result our objective function is discontinuous, similar to k-means clustering, thus making it difficult to derive an efficient algorithm that will find the optimal estimates for coefficients and groups. Witten et al. (2014) dealt with a similar difficulty for the CEN estimator, but noticed that if the groups are fixed then the problem is convex, while if the regression coefficients are fixed the problem becomes a k-means clustering problem. We modify the approach proposed in Witten et al. (2014) to our problem of grouping responses and extend the approach to the case of generalized linear models, specifically the binomial logistic model. In our theoretical results we assume the clustering groups are known, but the problem remains challenging as we are dealing with multiple responses, allow for $p \gg n$ and for p to increase with n .

In Section 2 we present our method for the multivariate linear regression model and provide theoretical results, including consistency of our estimator, to better understand the basic properties of the penalized likelihood solution. In Section 3 we provide details on the two-step iterative algorithm and show estimating the regression coefficients for the different clusters is an embarrassingly parallel problem, which is a property of our cluster fusion penalty that fuses within group fitted values. This avoids issues that would arise in fusing all possible combinations of regression coefficients, or having to specify a fusion set *a priori*. Examples of the issues that can arise can be found in Price et al. (2017), who discussed the importance of choosing the fusion set, and the original fused lasso paper which fused only consecutive coefficients (Tibshirani et al., 2005). In Section 4 we present the model for binomial responses along with an algorithm, demonstrating how the use of the cluster fusion penalty can exploit relationships of response variables beyond the traditional Gaussian problem. Simulations for both conditional Gaussian and binomial responses are presented in Section 5. The least squares version of our method is applied to model baby birth weight, placental weight and cotinine levels given maternal gene expression and demographic information. The binomial case is applied to model concession stand purchases using customer information as covariates. Both applied analysis are presented in Section 6. We conclude with a summary in Section 7.

2. Least Squares Model

2.1 Method

First, we consider estimating (1) when there are Q unknown clusters of the r responses. We further assume that $\sum_{i=1}^n y_{ik} = 0$ for all $k = 1, \dots, r$, $\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n x_{ij}^2 \leq n$ for all $j = 1 \dots, p$. The model requires rp parameters to be estimated for prediction, which is problematic when r or p are large. Let $D = (D_1, \dots, D_Q)$ be a partition of the set $\{1, \dots, r\}$. For a set A define $|A|$ as the cardinality of that set. We propose the multivariate cluster

elastic net (MCEN) estimator as

$$\begin{aligned}
 (\hat{B}, \hat{D}) = \arg \min_{B \in \mathcal{R}^{p \times r}, D_1, \dots, D_Q} & \frac{1}{2n} \sum_{i=1}^n \sum_{c=1}^r (y_{ic} - \mathbf{x}_i^T \boldsymbol{\beta}_c)^2 + \delta \|B\|_1 \\
 & + \frac{\gamma}{2n} \sum_{q=1}^Q \frac{1}{|D_q|} \sum_{l, m \in D_q} \|X(\boldsymbol{\beta}_l - \boldsymbol{\beta}_m)\|_2^2,
 \end{aligned} \tag{3}$$

where Q is the number of clusters and γ and δ are non-negative user specified tuning parameters. In addition Q , the total number of clusters, can be considered a tuning parameter. The cluster fusion penalty, associated with tuning parameter γ , is used to exploit similarities in the fitted values. The lasso penalty, with tuning parameter δ , is used to perform simultaneous estimation and variable selection. When $\gamma = 0$ or $Q = r$, the optimization in (3) reduces to r independent lasso penalized least squares problems with tuning parameter δ . If \hat{D} is known then the optimization in (3) can be split into Q independent optimizations that are similar to the optimizations presented in Li and Li (2008), Li and Li (2010), and Witten et al. (2014) and can be solved in parallel. We exploit this computational feature in our algorithm, which is a result of using the cluster fusion penalty.

The proposed method uses a combination of L_1 and L_2 penalties as proposed by Zou and Hastie (2005). Similar methods have been proposed for grouping the effects of predictors with a univariate response such as CEN (Witten et al., 2014) and Grace estimators (Li and Li, 2008, 2010; Zhao and Shojaie, 2016). Kim and Xing (2012) proposed a method that uses a predetermined hierarchical clustering of the responses that provides an L_1 penalty for all coefficients and a group L_2 penalty for responses that are grouped together. Chen et al. (2016) proposed a method using conjoint clustering to incorporate similarities in preferences between individuals in conjoint analysis. This method does not simultaneously estimate coefficients and groupings. It requires a two-step algorithm to estimate the number of clusters, and then estimates coefficients using regularization based on the estimated cluster. The proposed approach uses non-hierarchical clusters, allows for the clustering structure to be unknown before estimation of the coefficients and focuses more on imposing similar fitted values for grouped responses, compared to directly imposing a similar sparsity structure.

Selecting the triplet, (Q, γ, δ) , of tuning parameters can be done by K-fold cross validation minimizing the squared prediction error. Let \mathcal{F}_k be the set of indices in the k th fold, $k \in \{1, \dots, K\}$, and $\hat{\boldsymbol{\beta}}_c^{(-\mathcal{F}_k)}(Q, \gamma, \delta)$ be the estimated regression coefficient vector using Q , γ and δ for response c produced from the training set with \mathcal{F}_k removed. Then select the triplet, $(\hat{Q}, \hat{\gamma}, \hat{\delta})$, that minimizes

$$V(Q, \delta, \gamma) = \sum_{k=1}^K \sum_{c=1}^r \sum_{i \in \mathcal{F}_k} \left\{ y_{ic} - x_i^T \hat{\boldsymbol{\beta}}_c^{(-\mathcal{F}_k)}(Q, \gamma, \delta) \right\}^2. \tag{4}$$

2.2 Theoretical Results

For theoretical discussions we assume that D is known for some fixed value of Q . This is because for D unknown the objective function in (3) is discontinuous because of the discrete changes in groups, however if D is known (3) is a convex function. In this section we will

look at properties of the MCEN estimator for the special case of fixed n and p with $\delta = 0$. In addition, we present a consistency result that allows for $p \gg n$ when $\delta = o(1)$ and $\gamma = o(1)$.

Thus, the first two theorems refer to the following estimator

$$\begin{aligned} \bar{B} = \arg \min_{B \in \mathcal{R}^{p \times r}} & \frac{1}{2n} \sum_{i=1}^n \sum_{c=1}^r (y_{ic} - \mathbf{x}_i^T \beta_c)^2 + \delta \|B\|_1 \\ & + \frac{\gamma}{2n} \sum_{q=1}^Q \frac{1}{|D_q|} \sum_{l,m \in D_q} \|X(\beta_l - \beta_m)\|_2^2. \end{aligned} \quad (5)$$

The estimator \bar{B} does not simultaneously estimate the groups, it assumes they are known *a priori*, and thus is different than \hat{B} . There are instances where the grouping structure is known before data analysis and thus using \bar{B} would be preferable in practice. In addition \bar{B} is a key component to the algorithm discussed in Section 3. We begin by relating the estimator in (5) to ordinary least squares (OLS), for the special case of $\delta = 0$. Removing the L_1 penalty allows us to derive a closed form for the estimator.

Theorem 1 *Assume $n > p$, $\delta = 0$, and Q and γ are fixed values. Define $\dot{B} = (\dot{\beta}_1, \dots, \dot{\beta}_r)$ to be the OLS estimates for the r response variables and $\bar{B} = (\bar{\beta}_1, \dots, \bar{\beta}_r)$ be the solution to (5) with tuning parameter γ . Given $l \in D_q$ then $\bar{\beta}_l$ has the closed form solution of*

$$\bar{\beta}_l = \dot{\beta}_l + \frac{2\gamma}{(1+2\gamma)|D_q|} \sum_{c \in D_q, c \neq l} (\dot{\beta}_c - \dot{\beta}_l). \quad (6)$$

Theorem 1 provides some intuition about the MCEN estimator. As γ increases the MCEN estimator approaches a weighted average of the OLS coefficients within a cluster. In addition the results from Theorem 1 can be used to calculate the bias and variance of \bar{B} , which are needed for proving Theorem 2. The proof of Theorem 1 and the following Theorems can be found in the appendix.

Theorem 2 *Assume $E(\epsilon_{ic}^2) = 1$ for all $i \in \{1, \dots, n\}$ and $c \in \{1, \dots, r\}$ and $E(\epsilon_{ic}\epsilon_{ik}) = \rho$ for $c \neq k$, where $\rho \in (0, 1)$. Set $\delta = 0$, then for a fixed n and p where $n > p$ there exists a positive γ such that*

$$E\left(\|\bar{B} - B^*\|_2^2\right) < E\left(\|\dot{B} - B^*\|_2^2\right), \quad (7)$$

where B^* are the true regression coefficients, \dot{B} is as defined in Theorem 1 and \bar{B} is as defined in (5).

Similar to ridge regression Theorem 2 shows that for some positive γ the estimator from (5) has a smaller mean squared error than OLS. Note, we are not assuming that for $l, s \in D_m$ that $\beta_l^* = \beta_s^*$ and unless this condition holds the estimator \bar{B} is biased. Thus, there exists a value of γ for which there is a favorable bias-variance trade off.

Next we examine the asymptotic performance of the estimator with the L_1 penalty. At times it will be easier to refer to a vectorized version of a matrix and for any matrix $A \in \mathcal{R}^{a \times b}$, $\text{vec}(A) \in \mathcal{R}^{ab}$. Where $\text{vec}(A)$ is the vector formed by stacking the columns of A .

Define S as the set of active predictors. That is, S is a subset of $\{1, \dots, rp\}$ where $m \in S$ if $\text{vec}(B^*)_m \neq 0$. The subspace for the active predictors is

$$\mathcal{M}(S) \equiv \{\boldsymbol{\theta} \in \mathcal{R}^{pr} \mid \theta_j = 0 \text{ if } j \notin S\}.$$

The parameter space will be separated using projections of vectors into orthogonal complements. We define a projection of a vector \mathbf{u} into space $\mathcal{M}(S)$ as

$$\mathbf{u}_{\mathcal{M}(S)} \equiv \arg \min_{\mathbf{v} \in \mathcal{M}(S)} \|\mathbf{u} - \mathbf{v}\|_2.$$

The orthogonal complement of space $\mathcal{M}(S) \subseteq \mathcal{R}^p$ is

$$\mathcal{M}^\perp(S) \equiv \{\mathbf{v} \in \mathcal{R}^{pr} \mid \langle \mathbf{u}, \mathbf{v} \rangle = 0 \text{ for all } \mathbf{u} \in \mathcal{M}(S)\}.$$

The following set is central to our proof of consistency,

$$\mathcal{C} \equiv \{\boldsymbol{\theta} \in \mathcal{R}^{pr} \mid \|\boldsymbol{\theta}_{\mathcal{M}^\perp(S)}\|_1 \leq \|\boldsymbol{\theta}_{\mathcal{M}}\|_1\}.$$

For our proof of the consistency of \bar{B} we make the following six assumptions:

A1 Define \mathbf{X}_j to be the j th column vector of X , then $\mathbf{X}_j \in \mathcal{R}^p$ has the condition that $\frac{\|\mathbf{X}_j\|_2^2}{n} \leq 1$.

A2 Define $\boldsymbol{\epsilon}_c = (\epsilon_{1c}, \dots, \epsilon_{nc})^T \in \mathcal{R}^n$ as the error vector for response c . The error vector $\boldsymbol{\epsilon}_c$ has a mean of zero and sub-Gaussian tails for all $c \in \{1, \dots, r\}$. That is, there exists a constant σ_c such that for any $\mathbf{a} \in \mathcal{R}^n$, with $\|\mathbf{a}\|_2 = 1$,

$$P(|\langle \boldsymbol{\epsilon}_c, \mathbf{a} \rangle| > t) \leq 2\exp\left(-\frac{t^2}{2\sigma_c^2}\right).$$

Define $\sigma = \max_c \sigma_c$.

A3 Define $\tilde{X} = I_r \otimes X \in \mathcal{R}^{rn \times rp}$, where \otimes is the standard Kronecker product. There exists a positive constant κ such that

$$\kappa \|\boldsymbol{\theta}\|_2^2 \leq \min_{\boldsymbol{\theta} \in \mathcal{C}} n^{-1} \|\tilde{X}\boldsymbol{\theta}\|_2^2.$$

A4 There exists a positive constant \acute{b} such that $\max_{q=1, \dots, Q} \max_{(l,k) \in D_q} \|\boldsymbol{\beta}_l^* - \boldsymbol{\beta}_k\|_2 \leq \acute{b}$.

A5 Given $l, k \in D_q$, if $\beta_{lj}^* = 0$ then $\beta_{kj}^* = 0$, for all $j \in \{1, \dots, p\}$ and $q \in \{1, \dots, Q\}$.

A6 Define $\rho_{\max}(A)$ as the maximum eigenvalue of square matrix A and $X_{S_{D_q}}$ as the matrix of true predictors for cluster q , where the j th predictor is a true predictor if $\beta_{lj}^* \neq 0$ for any $l \in D_q$. There exists a positive constant ρ_{\max} such that

$$\max_{q=1, \dots, Q} \rho_{\max}\left(\frac{1}{n} X_{S_{D_q}}^T X_{S_{D_q}}\right) \leq \rho_{\max}.$$

Assumption A1 is a standard assumption for lasso-type penalties and can be achieved by appropriately scaling the covariates, which is commonly done in penalized regression. Assumption A2 is a generalization of the sub-Gaussian error assumption for penalized regression for a univariate response. Assumption A1 could be relaxed to allow for certain unbounded covariates, but then A2 would be replaced by assuming the errors are normally distributed (Candes and Tao, 2007; Meinshausen and Yu, 2009). Assumption A3 is a generalization of the common restricted eigenvalue assumption. Motivation for assumption A3 is discussed in great detail by Negahban et al. (2012) and a version for $r = 1$ has been used in several works analyzing asymptotic behaviors of the lasso estimator (Bickel et al., 2009; van de Geer and Bühlmann, 2009; Meinshausen and Yu, 2009). Assumptions A4 and A5 provide that the true coefficients are similar for responses in the same group. Assumption A5 provides that they have the same sparsity structure. While, assumption A4 ensures that the difference in the non-zero elements can be bounded by a finite constant, even if the number of predictors increases with n . Assumption A6 assumes the maximum eigenvalues of the sample covariance of the true predictors are bounded, a common assumption in high-dimensional work. Assumptions A4-A6 can be replaced by an assumption similar to assumption A2 from Witten et al. (2014) that if $b, c \in D_m$ then $\beta_b^* = \beta_c^*$, for all $m \in \{1, \dots, Q\}$, thus the bias of the MCEN estimator only comes from the L_1 penalty.

Using assumptions A3 and A5 we can provide a closed form definition of the asymptotic bias when $\delta = 0$. This relationship will be central to our proof of consistency of \bar{B} .

Corollary 3 *Let B^* be an s -sparse matrix, whose column vectors are all sparse and $E[X^T X/n] \in \mathcal{R}^{p \times p}$ to be a positive definite matrix. Assume Q and γ are fixed values. Define,*

$$\hat{B} = (\hat{\beta}_1, \dots, \hat{\beta}_r) = \arg \min_{\beta_1, \dots, \beta_r \in \mathcal{R}^p} E \left(\frac{1}{2n} \sum_{i=1}^n \sum_{c=1}^r (y_{ic} - \mathbf{x}_i^T \beta_c)^2 + \frac{\gamma}{2n} \sum_{q=1}^Q \frac{1}{|D_q|} \sum_{l, m \in D_q} \|X(\beta_l - \beta_m)\|_2^2 \right),$$

Assume $l \in D_q$ then $\hat{\beta}_l$ has closed form solution,

$$\hat{\beta}_l = \beta_l^* + \frac{2\gamma}{(1 + 2\gamma)|D_q|} \sum_{c \in D_q, c \neq l} (\beta_c^* - \beta_l^*).$$

Corollary 3 provides insight into what \bar{B} would converge to for a fixed γ . Knowing this exact relationship is used in our consistency proof because it allows us to understand the exact nature of the bias caused by the L_2 penalty and for γ going to zero at a given rate we can show that the bias is asymptotically negligible.

Theorem 4 *Let B^* be an s -sparse matrix, whose column vectors are all sparse and $E[X^T X/n] \in \mathcal{R}^{p \times p}$ to be a positive definite matrix. Given $\delta = 16\sigma \sqrt{\frac{\log(rp)}{n}}$, $\gamma \leq \frac{5}{4\rho_{\max} b} \sigma \sqrt{\frac{\log(rp)}{n}}$ and assumptions A1-A6 hold then there exist constants c_1 , c_2 , c_3 and c_4 such that*

$$\| \text{vec}(\bar{B} - B^*) \|_2 \leq \sigma \sqrt{\frac{s \log(rp)}{n}} \left(\frac{c_3}{\kappa} + \frac{c_4}{\rho_{\max}} \right), \quad (8)$$

with probability at least $1 - c_1 \exp(-c_2 n \delta^2)$.

The convergence rate derived is similar to rates found in lasso-type estimators with a univariate response, with $\log(rp)$ replacing $\log(p)$ to accommodate for the multiple responses (Bickel et al., 2009; Candes and Tao, 2007; Meinshausen and Yu, 2009; Negahban et al., 2012). Thus, under the conditions of Theorem 4 if $pr \rightarrow \infty$ then $\|\text{vec}(\bar{B} - B^*)\|_2 = O_p \left\{ \sqrt{\frac{s \log(rp)}{n}} \right\}$. Our results prove consistency of our estimator when the group structure is known. Zhao and Shojaie (2016) propose the Grace test for an estimator with a similar penalty for grouping predictors with a univariate response and establish asymptotic results that allow for inference even if there is some uncertainty to the grouping structure.

3. Algorithm

The optimization in (3) is discontinuous because of the estimation of cluster assignments. To simplify the optimization we propose an iterative algorithm that alternates between estimating the groups with the regression coefficients fixed, and estimating the regression coefficients with the groups fixed. If the clusters are known (5) then it is a convex optimization problem that can be solved by a coordinate descent algorithm. Let $R = \frac{1}{n} X^T X$, define \mathbf{R}_j as the j th column of R . The super script $(-h)$ denotes the h th element of the vector has been removed, and r_{jj} is j th diagonal element of R . Define $S(a, b) = \text{sign}(a) \max(0, |a| - b)$. To solve (5), we use a coordinate descent algorithm where each update is preformed by

$$\bar{\beta}_{jk} \leftarrow \frac{S \left[\frac{1}{n} y_k^T \mathbf{X}_j - \left\{ 1 + \frac{\gamma(|D_q|-1)}{|D_q|} \right\} \mathbf{R}_j^{(-j)T} \bar{\beta}_k^{(-j)} + \frac{\gamma}{|D_q|} \sum_{s \in D_q, s \neq k} \mathbf{R}_j^T \bar{\beta}_s, \delta/2 \right]}{r_{jj} \left(1 + \gamma \frac{|D_q|-1}{|D_q|} \right)}. \quad (9)$$

Thus, (5) is solved by iterating through $j \in \{1, \dots, p\}$ and $k \in \{1, \dots, r\}$ until the solution converges, similar to other coordinate descent solutions (Witten et al., 2014; Li and Li, 2010, 2008; Friedman et al., 2008). If B is known then the solution to (3) reduces to the well studied k-means clustering problem. Recognizing this, we propose a two-step iterative procedure to obtain a local minimum. To start the algorithm an initial estimate of D or B is needed. We propose initializing the regression coefficients for the different responses separately with the elastic net estimator of response c of

$$\hat{\beta}_c^1 = \arg \min_{\beta_c \in \mathcal{R}^p} \frac{1}{2n} \sum_{i=1}^n (y_{ic} - \mathbf{x}_i^T \beta_c)^2 + \delta \|\beta_c\|_1 + \gamma \|\beta_c\|_2^2, \quad (10)$$

where $\hat{B}^w = (\hat{\beta}_1^w, \dots, \hat{\beta}_r^w)$ represents the w th iterative estimate of B^* . Given a fixed (Q, γ, δ) we propose the following algorithm.

1. Begin with initial estimates, $\hat{\beta}_1^1, \dots, \hat{\beta}_r^1$.
2. For the w th step, where $w > 1$, repeat the steps below until the group estimates do not change:
 - (a) Hold \hat{B}^{w-1} fixed and minimize,

$$\left(\hat{D}_1^w, \dots, \hat{D}_Q^w \right) = \underset{D_1, \dots, D_Q}{\text{minimize}} \left\{ \sum_{q=1}^Q \frac{1}{|D_q|} \sum_{l, m \in D_q} \left\| X \left(\hat{\beta}_l^{w-1} - \hat{\beta}_m^{w-1} \right) \right\|_2^2 \right\}. \quad (11)$$

The above can be solved by performing K -means clustering on the r n -dimensional vectors $X\hat{\beta}_1^{w-1}, \dots, X\hat{\beta}_r^{w-1}$.

(b) Holding $\hat{D}_1^w, \dots, \hat{D}_Q^w$ fixed the w th estimate of B^* is

$$\begin{aligned} \hat{B}^w = \arg \min_{B \in \mathcal{R}^{p \times r}} & \frac{1}{2n} \sum_{i=1}^n \sum_{c=1}^r (y_{ic} - \mathbf{x}_i^T \beta_c)^2 + \delta \|B\|_1 \\ & + \frac{\gamma}{2n} \sum_{q=1}^Q \frac{1}{|\hat{D}_q^w|} \sum_{l,m \in \hat{D}_q^w} \|X(\beta_l - \beta_m)\|_2^2. \end{aligned} \tag{12}$$

Note that for the groups known, instead of estimated, \hat{B}^w is equivalent to \bar{B} . Thus (12) can be solved using the coordinate descent solution from (9) using \hat{B}^{w-1} as the initial estimates for the coordinate descent algorithm.

Convergence is reached once the groups at the w th and $(w - 1)$ th iteration are the same. The optimization in (5) is separable with respect to \hat{D} , and results in Q independent optimization problems that can be solved in parallel. The algorithm for (5) can be solved in solution path type form where we iterate across different values of δ in a similar fashion as proposed in the glmnet algorithm (Friedman et al., 2008). If all of the initial elastic net estimators are fully sparse, we set the solution to be a zero matrix and thus following Friedman et al. (2008), initialize the algorithm by beginning the sequence with δ_{\max} at

$$\delta_{\max} = 2 \max_{j,k} \left| \frac{\sum_{i=1}^n y_{ik} x_{ij}}{n} \right|.$$

Our two-step approach is closely related to the CEN algorithm proposed by Witten et al. (2014), who proposed a two-step algorithm where the two steps are solved by coordinate descent and k-means algorithms. The major difference in our proposal is that we cluster the responses rather than the predictors, and have the ability to solve the optimization in parallel due to the nature of our regularization in a multiple response setting.

4. Binomial Model

4.1 Method

Next we extend the multivariate cluster elastic net to generalized linear models. We focus specifically on the binomial response case, but our discussion here will scale to other exponential families. A fusion penalty has been proposed for merging groups from a multinomial response (Price et al., 2017), but our method differs as it aims to leverage association between multiple binomial responses. Kasap et al. (2016) proposed an ensemble method that combines association rule mining and binomial logistic regression via a multiple linear regression model. Our method differs from this by simultaneously estimating the clusters of the response variables and estimating the regression coefficients. An example is n customers, with p covariates, such as demographic and historic purchasing variables, and r indicators of product purchasing statuses for each customer. You could run r independent models, but this would not allow for modeling the relationship between the different products. Extending the multivariate cluster elastic net to multiple binomial responses would

allow us to group products by purchase probabilities to identify and use relationships between products. This could also be used to create a probabilistic model for diseases based on patient demographic and medical information.

For the linear model we ignore the intercept term as it can be removed by appropriately scaling Y and X . This is not possible in logistic regression, therefore the model needs an intercept term. We define $\mathbf{u}_i = (1, \mathbf{x}_i^T)^T \in \mathcal{R}^{p+1}$, $U = (\mathbf{u}_1^T, \dots, \mathbf{u}_n^T)^T \in \mathcal{R}^{n \times p+1}$, $\mathbf{U}_k \in \mathcal{R}^n$ as the k th column vector of U and $\hat{R} = U^T U$. The true coefficients for response k is defined as $\boldsymbol{\theta}_k^* \in \mathcal{R}^{p+1}$, $\Theta^* = (\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_r^*) \in \mathcal{R}^{p+1 \times r}$, $\Theta_{-1}^* \in \mathcal{R}^{p \times r}$ is the matrix with the first row, the row of intercept coefficients, of Θ^* removed and $\boldsymbol{\theta}_{(-1)k}^* \in \mathcal{R}^p$ is the k th column vector of Θ_{-1}^* . In this model y_{ik} is an independent draw from

$$\text{Bin}(1, \pi_{ik}^*), \quad (13)$$

where

$$\pi_{ik}^* = \frac{\exp(\mathbf{u}_i^T \boldsymbol{\theta}_k^*)}{1 + \exp(\mathbf{u}_i^T \boldsymbol{\theta}_k^*)}. \quad (14)$$

The penalized negative log-likelihood function is

$$\begin{aligned} & \sum_{k=1}^r \sum_{i=1}^n y_{ik} \mathbf{u}_i^T \boldsymbol{\theta}_k - \log \{1 + \exp(\mathbf{u}_i^T \boldsymbol{\theta}_k)\} \\ & + \frac{\gamma}{2n} \sum_{q=1}^q \frac{1}{|D_q|} \sum_{l,m \in D_q} \|U(\boldsymbol{\theta}_l - \boldsymbol{\theta}_m)\|_2^2 + \delta \|\Theta_{-1}\|_1. \end{aligned} \quad (15)$$

4.2 Algorithm

We propose solving (15) by approximating it with a penalized quadratic function similar to the glmnet algorithm proposed by Friedman et al. (2008). Define,

$$g(\pi_{ik}) = \log \left(\frac{\pi_{ik}}{1 - \pi_{ik}} \right) = \mathbf{u}_i^T \boldsymbol{\theta}_k. \quad (16)$$

To implement this approximation we define

$$z_{ik} = g(y_{ik}) = g(\pi_{ik}) + \frac{y_{ik} - \pi_{ik}}{\pi_{ik}(1 - \pi_{ik})}, \quad (17)$$

$$w_{ik} = \pi_{ik}(1 - \pi_{ik}), \quad (18)$$

$$-l_{Ak}(\boldsymbol{\theta}_k) = \sum_{i=1}^n w_{ik} (z_{ik} - \mathbf{u}_i^T \boldsymbol{\theta}_k)^2. \quad (19)$$

Note that z_{ik} is just the first order Taylor approximation of $g(y_{ik})$, and that w_{ik} is the conditional variance of z_{ik} given \mathbf{u}_i . Define $\mathbf{Z}_k = (z_{1k}, \dots, z_{nk})^T \in \mathcal{R}^n$ and $\mathbf{W} = (w_{1k}, \dots, w_{nk})^T \in \mathcal{R}^n$.

The MCEN estimator for the binomial model is

$$\begin{aligned} (\hat{\Theta}, \hat{D}) = & \arg \min_{\Theta \in \mathcal{R}^{p+1 \times r}, D_1, \dots, D_Q} \sum_{k=1}^r -l_{Ak}(\boldsymbol{\theta}_k) + \delta \|\Theta_{-1}\|_1 \\ & + \frac{\gamma}{2n} \sum_{q=1}^q \frac{1}{|D_q|} \sum_{l,m \in D_q} \|U(\boldsymbol{\theta}_l - \boldsymbol{\theta}_m)\|_2^2. \end{aligned} \quad (20)$$

If the groups are known *a priori* the solution is

$$\begin{aligned} \bar{\Theta} = \arg \min_{\Theta \in \mathcal{R}^{p+1 \times r}} & \sum_{k=1}^r -l_{Ak}(\boldsymbol{\theta}_k) + \delta \|\Theta_{(-1)}\|_1 \\ & + \frac{\gamma}{2n} \sum_{q=1}^q \frac{1}{|D_q|} \sum_{l,m \in D_q} \|U(\boldsymbol{\theta}_l - \boldsymbol{\theta}_m)\|_2^2. \end{aligned} \quad (21)$$

For same length vectors \mathbf{a} and \mathbf{b} let $\mathbf{a} \circ \mathbf{b}$ represent the component wise multiplication of the two vectors. To solve (21), we use a proximal coordinate descent algorithm where each update is performed by

$$\bar{\theta}_{jk} \leftarrow \frac{S\{(\mathbf{w}_k \circ \mathbf{z}_k)^T \mathbf{U}_j - M_{jk}, I(j \neq 0)\delta/2\}}{r_{jj}\gamma \frac{|D_q|-1}{n|D_q|} + \mathbf{U}_j^T(\mathbf{w}_k \circ \mathbf{U}_j)}, \quad (22)$$

where

$$\begin{aligned} M_{jk} &= \sum_{c=1, c \neq h}^p \mathbf{U}_j^T(\mathbf{w}_k \circ \mathbf{U}_c) \bar{\Theta}_{cj} + \frac{\gamma(|D_q|-1)}{n|D_q|} \tilde{\mathbf{R}}_j^{(-j)T} \bar{\boldsymbol{\theta}}_k^{(-j)} \\ &- \frac{\gamma}{n|D_q|} \sum_{s \in D_q, s \neq k} \tilde{\mathbf{R}}_j^T \bar{\boldsymbol{\theta}}_s. \end{aligned}$$

The final solution is found by iterating through $j \in \{1, \dots, p\}$ and $k \in \{1, \dots, r\}$ until convergence. Again this is a solution similar to the glmnet algorithm proposed by Friedman et al. (2008).

To solve (20), we propose an algorithm that is similar in nature to the penalized least squares solution proposed in Section 3. The main difference is that we solve (20) with D_1, \dots, D_Q fixed using an iteratively reweighed least squares (IRWLS) solution with a proximal coordinate descent algorithm. The initial estimator for each response is done separately with

$$\hat{\boldsymbol{\theta}}_k^1 = \arg \min_{\boldsymbol{\theta}_k \in \mathcal{R}^{p+1}} -l_{Ak}(\boldsymbol{\theta}_k) + \delta \|\boldsymbol{\theta}_{(-1)k}\|_1 + \gamma \|\boldsymbol{\theta}_{(-1)k}\|_2^2. \quad (23)$$

The following is our proposed algorithm for estimating (20).

1. Begin with initial estimates of $\hat{\Theta}^1 = (\hat{\boldsymbol{\theta}}_1^1, \dots, \hat{\boldsymbol{\theta}}_r^1) \in \mathcal{R}^{p+1 \times r}$.
2. For the w th step, where $w > 1$, repeat the steps below until the group estimates do not change:
 - (a) Hold $\hat{\Theta}^{w-1}$ fixed and minimize

$$\left(\hat{D}_1^w, \dots, \hat{D}_Q^w \right) = \underset{D_1, \dots, D_Q}{\text{minimize}} \left\{ \sum_{q=1}^Q \frac{1}{|D_q|} \sum_{l,m \in D_q} \left\| U(\boldsymbol{\theta}_l^{w-1} - \boldsymbol{\theta}_m^{w-1}) \right\|_2^2 \right\}. \quad (24)$$

The above can be solved by performing K -means clustering.

(2b) Holding $\hat{D}_1^w, \dots, \hat{D}_Q^w$ fixed the w th update for the coefficients is

$$\begin{aligned} \hat{\Theta}^w = \arg \min_{\Theta \in \mathcal{R}^{p+1 \times r}} & \sum_{k=1}^r -l_{Ak}(\boldsymbol{\theta}_k) + \delta \|\Theta_{-1}\|_1 \\ & + \frac{\gamma}{2n} \sum_{q=1}^q \frac{1}{|\hat{D}_q^w|} \sum_{l,m \in \hat{D}_q^w} \|U(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)\|_2^2. \end{aligned} \quad (25)$$

Where (25) can be solved using the proximal coordinate descent solution presented in (22), using $\hat{\Theta}^{w-1}$ as the initial estimates for the proximal coordinate descent algorithm.

The triplet (Q, γ, δ) can be selected using K-Fold cross validation maximizing the validation log-likelihood. Let \mathcal{F}_k be the set of indices in the k th fold ($k \in \{1, \dots, K\}$) and $\hat{\pi}_{ic}^{(-\mathcal{F}_k)}(Q, \gamma, \delta)$ be the estimated probability for observation i and response c produced from the model with \mathcal{F}_k removed using Q, γ and δ . Specifically we select the triplet that maximizes

$$V(Q, \delta, \gamma) = \sum_{v=1}^K \sum_{c=1}^r \sum_{i \in \mathcal{F}_v} \left[y_{ic} \log \left\{ \hat{\pi}_{ic}^{(-\mathcal{F}_k)} \right\} + (1 - y_{ic}) \log \left\{ 1 - \hat{\pi}_{ic}^{(-\mathcal{F}_k)} \right\} \right]. \quad (26)$$

The quadratic approximation defined by (19), is a standard technique used to estimate parameters in generalized linear models, making this framework and our algorithm scalable to other exponential family settings (Faraway, 2006). Tuning parameter selection would then be done by updating (26) with the appropriate likelihood.

5. Simulations

5.1 Gaussian Simulations

In this section we compare the performances of the MCEN estimator (3), the true MCEN (TMCEN) (5), with clustering structure known *a priori*, the separate elastic net (SEN) estimator (10), the joint elastic net (JEN) estimator

$$\hat{B}_{\text{JEN}} = \arg \min_B \frac{1}{2n} \sum_{k=1}^r \sum_{i=1}^n (y_{ik} - \mathbf{x}_i^T \boldsymbol{\beta}_k)^2 + \delta \sum_{j=1}^p \sqrt{\beta_{j1}^2 + \dots + \beta_{jr}^2} + \gamma \sum_{k=1}^r \sum_{j=1}^p \beta_{jk}^2, \quad (27)$$

and the tree-guided group lasso (TGL) (Kim and Xing, 2012). Define $\mathbf{B}_j \in \mathcal{R}^r$ as the j th row vector of matrix B . Given a tree T with vertices V , where each node $v \in V$ is associated with group G_v define $B_j^{G_v}$ as a vector of the j th predictors from responses in group G_v . The TGL estimator is

$$\hat{B}_{\text{TGL}} = \arg \min_B \frac{1}{2} \sum_{k=1}^r \sum_{i=1}^n (y_{ik} - \mathbf{x}_i^T \boldsymbol{\beta}_k)^2 + \delta \sum_{j=1}^p \sum_{v \in V} w_v \|B_j^{G_v}\|_2, \quad (28)$$

where w_v are weights that can vary with the nodes. See Kim and Xing (2012) for a detailed presentation of TGL, including how the weights, w_v , are derived.

The JEN and SEN models are fit using the `glmnet` package in R (Friedman et al., 2008). Tuning parameters for all methods are selected using 10-folds cross validation. For the MCEN and TMCEN methods cluster sizes of 2, 3 and 4 are considered. We include the TMCEN estimator for two reasons. First, in practice the TMCEN estimator could be used if the practitioner has a predetermined clustering of the responses. Second, the TMCEN is useful as a benchmark to compare with the MCEN estimator because if the grouping of responses is useful TMCEN provides the optimal grouping. In all of the simulations the sample size is 100 and the number of responses is 15. For the number of covariates we considered 12, 100 and 300. Next we define how the covariates are generated and then will present the generating process for the response variables.

Define $\tilde{\Sigma} \in \mathcal{R}^{12 \times 12}$ with entries $\tilde{\sigma}_{ii} = 1$ and $\tilde{\sigma}_{ij} = \rho$, for $i \neq j$. Let $0_{a,b} \in \mathcal{R}^{a \times b}$ be a matrix with all entries equal to zero. The covariates are generated by $\mathbf{x}_i \sim N(\mathbf{0}_p, \Sigma_x)$, where $\Sigma_x = \tilde{\Sigma}$ for $p = 12$ and otherwise

$$\Sigma_x = \begin{pmatrix} \tilde{\Sigma} & 0_{p-12, p-12} \\ 0_{p-12, p-12} & I_{p-12} \end{pmatrix},$$

with $\rho = .7$.

For a group of responses we define the grouped coefficients as $\mathbf{b}_q(\eta, \lambda) = (\boldsymbol{\eta}_q - \lambda, \boldsymbol{\eta}_q^*, \boldsymbol{\eta}_q + \lambda, \boldsymbol{\eta}_q + 2\lambda, \boldsymbol{\eta} + 3\lambda) \in \mathcal{R}^{q \times 5}$, where λ is a constant and $\boldsymbol{\eta}_q \in \mathcal{R}^q$ with each element equal to η . In the case of $p = 12$ the matrix of coefficients is

$$B_{\eta, \lambda}^* = \begin{pmatrix} \mathbf{b}_4(\eta, \lambda) & \mathbf{0}_{4,5} & \mathbf{0}_{4,5} \\ \mathbf{0}_{4,5} & \mathbf{b}_4(\eta, \lambda) & \mathbf{0}_{4,5} \\ \mathbf{0}_{4,5} & \mathbf{0}_{4,5} & \mathbf{b}_4(\eta, \lambda) \end{pmatrix},$$

otherwise

$$B_{\eta, \lambda}^* = \begin{pmatrix} \mathbf{b}_{10}(\eta, \lambda) & \mathbf{0}_{10,5} & \mathbf{0}_{10,5} \\ \mathbf{0}_{10,5} & \mathbf{b}_{10}(\eta, \lambda) & \mathbf{0}_{10,5} \\ \mathbf{0}_{10,5} & \mathbf{0}_{10,5} & \mathbf{b}_{10}(\eta, \lambda) \\ \mathbf{0}_{p-30,5} & \mathbf{0}_{p-30,5} & \mathbf{0}_{p-30,5} \end{pmatrix}.$$

Define $\Sigma_\epsilon \in \mathcal{R}^{15 \times 15}$ with $\sigma(\epsilon)_{ij}$ being the entry for the i th row and j th column of Σ_ϵ . The generating process for the response is

$$\mathbf{y}_i = B_{\eta, \lambda}^*{}^T \mathbf{x}_i + \boldsymbol{\epsilon}_i, \quad (29)$$

where $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}_{15}, \Sigma_\epsilon)$, $\sigma(\epsilon)_{ii} = 1$ and $\sigma(\epsilon)_{ij} = 0$, for i not equal to j . In all simulations we set the sample size to 100, perform 50 replications and with p set consider the following 9 different combinations for the true coefficient matrix,

$$(\eta, \lambda) \in \{0.25, 0.5, 0.75, 1\} \times \{0.02, 0.05, 0.10\}.$$

Models are fit using the training data with a sample size of 100. The tree for TGL is defined by performing complete-linkage hierarchical clustering on the responses in the training data. In addition we generate 1000 additional testing samples to assess the prediction accuracies of the models. Let y_{ij}^* represent the i th training sample for the j th response and

\hat{y}_{ij} represent a predicted value of that sample and response. The average squared prediction error (ASPE) is defined as

$$\frac{1}{15000} \sum_{i=1}^{1000} \sum_{j=1}^{15} (y_{ij}^* - \hat{y}_{ij})^2. \quad (30)$$

We also report the mean squared error (MSE) of the estimators where for an estimator B

$$\text{MSE}(B) = \sum_{j=1}^{15} \|\beta_j - \beta_j^*\|_2^2. \quad (31)$$

In addition we report the number of true variables selected (TV), out of a maximum of 60 for $p = 12$ and 150 otherwise, and the number of false variables selected (FV). Box plots of the statistics for $p = 300$ and the different combinations of η and λ are reported in Figures 1–4. These results show that TMCEN generally outperforms all methods in terms of ASPE and MSE. The one exception being when $\eta = 1$, particularly for larger values of λ , TGL is competitive with or outperforms TMCEN. For larger values of λ we expect more bias in the MCEN and TMCEN solutions and our simulation setting is favorable to TGL because the sparsity structure is the same for responses in the same cluster. With regards to ASPE and MSE, MCEN generally does better than TGL when $\eta = .5$ or $.75$. This suggests that the MCEN approach is advantageous with several smaller signals, but the signals need to be strong enough to correctly identify the clustering of the responses. The MCEN method also outperforms JEN and SEN in terms of ASPE and MSE, except in the case of $\eta = .25$ where JEN outperforms MCEN. In this case the signal is too small resulting in the MCEN method not finding the true clustering structure, and thus the grouping penalty will not be optimal. The MCEN and TMCEN methods tend to pick a larger model than SEN, but a smaller model than JEN. This results in the MCEN and TMCEN methods correctly choosing more true predictors than SEN and fewer false positive predictors than JEN for weaker signal cases. In terms of variable selection MCEN and TMCEN tend to do better than TGL in terms of both true and false variable selection. For the stronger signal cases the SEN approach does the best in terms of variable selection, tending to have the maximum number of true covariates selected, while a smaller number of false covariates selected. Similar conclusions can be derived for the plots of $p = 12$ and $p = 100$, which are available in the supplementary material.

5.2 Binomial Simulations

In this setting we have a binomial response variable and compare performance of the MCEN estimator (20) to SEN (23), for $r = 15$ and $p = 12, 100$ or 300 . The SEN models were fit using the `glmnet` package in R (Friedman et al., 2008). Similar to the previous section, the covariates are generated by $\mathbf{x}_i \sim N(\mathbf{0}_p, \Sigma_x)$, where Σ_x has the same structure provided in the Gaussian simulations with $\rho = .9$.

We use the same structure of B presented in Section 5.1, consider the same values of η and λ and again perform 50 replications with a sample size of 100. Tuning parameters for the models are estimated via 10-folds cross validation as explained in Section 4.2. For Q , the number of groups, we consider values of 2, 3, and 4. For the SEN method each response

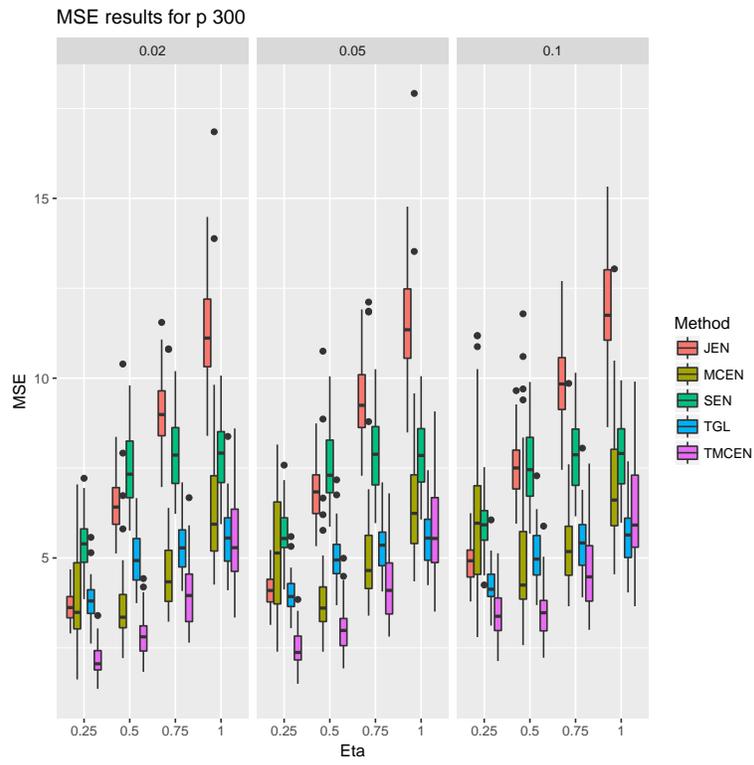


Figure 1: MSE results for the Gaussian simulations with p equal to 300. Different box plots correspond to different values of λ , while x-axis values are for different values of η .

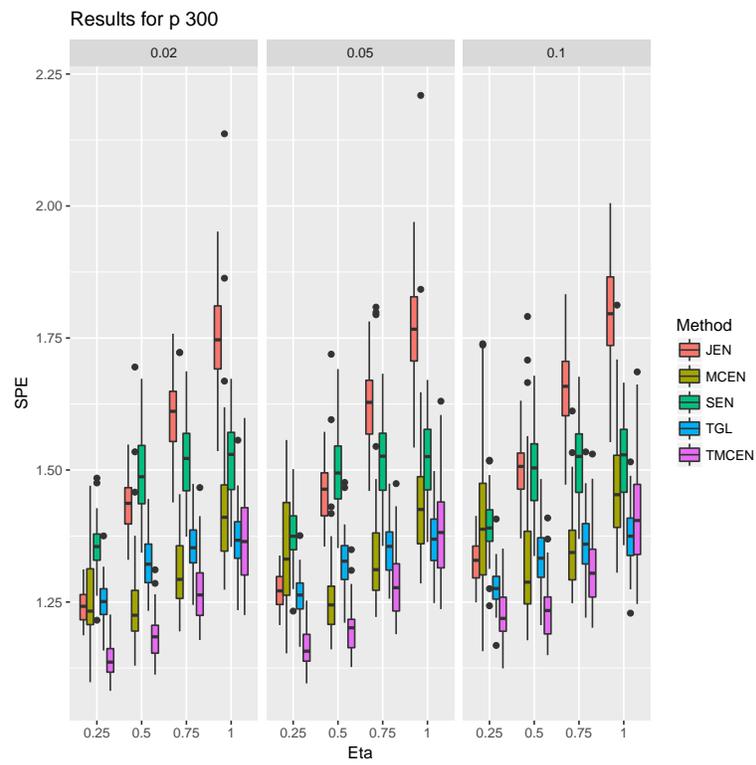


Figure 2: ASPE results for the Gaussian simulations with p equal to 300. Different box plots correspond to different values of λ , while x-axis values are for different values of η .

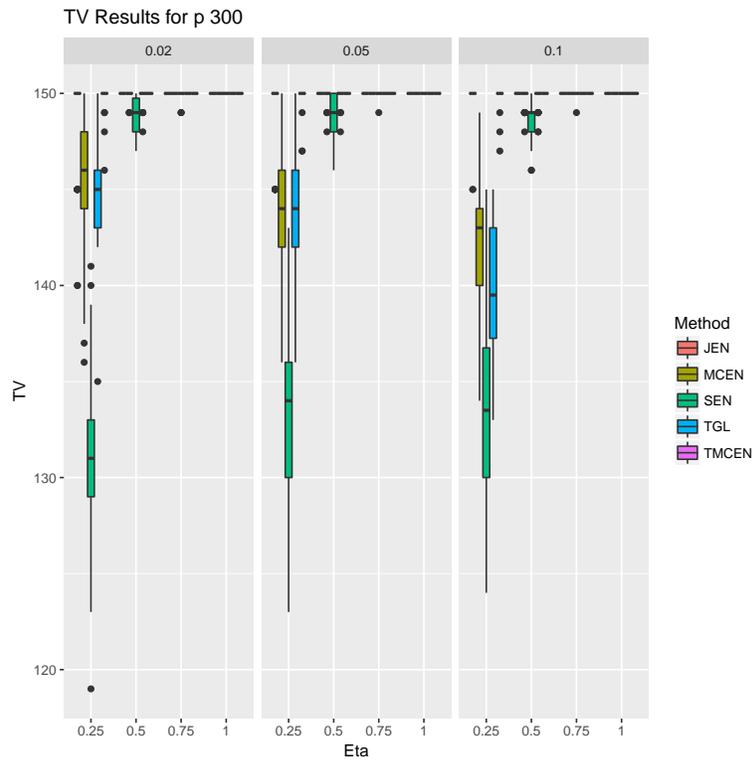


Figure 3: TV results for the Gaussian simulations with p equal to 300. Different box plots correspond to different values of λ , while x-axis values are for different values of η .

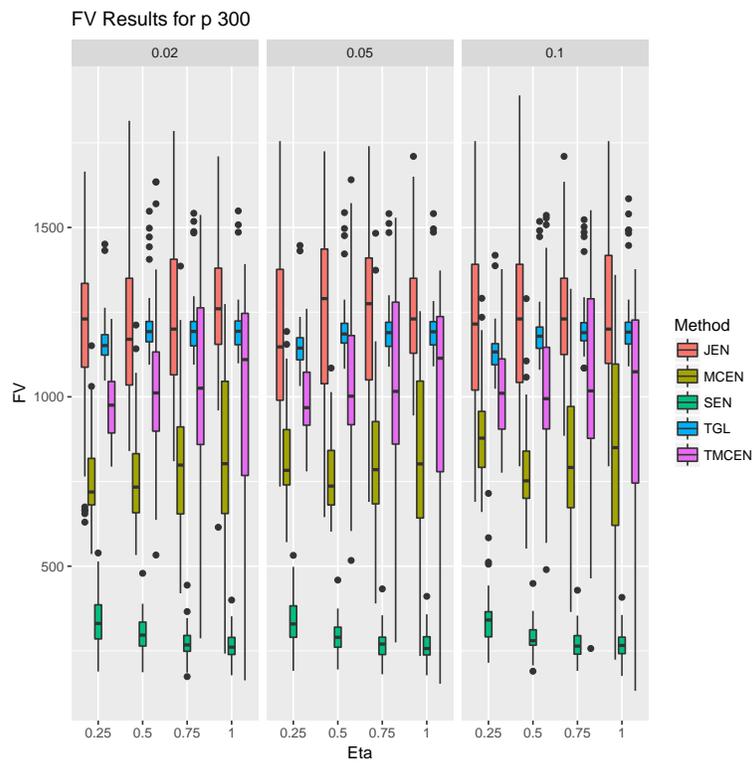


Figure 4: FV results for the Gaussian simulations with p equal to 300. Different box plots correspond to different values of λ , while x-axis values are for different values of η .

$c \in \{1, \dots, r\}$ will be associated with its own tuning parameters of γ_c and δ_c that will be selected by maximizing the equivalent of (26) for only one response.

Define $\beta_k^*(\eta, \lambda)$ as the k th column vector of $B_{\eta, \lambda}^*$. In all settings the k th response of the i th observation, y_{ik} , is an independent draw from $\text{Bin}(1, \pi_{ik}^*)$ where

$$\pi_{ik}^* = \frac{\exp\{\mathbf{x}'_i \beta_k^*(\eta, \lambda)\}}{1 + \exp\{\mathbf{x}'_i \beta_k^*(\eta, \lambda)\}}.$$

To evaluate the methods, 1000 validation observations are generated from the data generating model and the KL divergence is measured for each of the 50 replications. The KL divergence for a replication is defined as,

$$\frac{1}{1000} \sum_{i=1}^{1000} \sum_{k=1}^{15} \left\{ \log \left(\frac{\hat{\pi}_{ik}}{\pi_{ik}^*} \right) \hat{\pi}_{ik} + \log \left(\frac{1 - \hat{\pi}_{ik}}{1 - \pi_{ik}^*} \right) (1 - \hat{\pi}_{ik}) \right\},$$

where π_{ik}^* is the true probability and $\hat{\pi}_k(x_i, \delta, \gamma)$ is the estimated probability for response k for validation observation i .

Box plots are presented to compare the KL divergence of MCEN and SEN for the different settings in the case of $p = 300$. The results of simulation in cases where $p = 12$ and 100 are available in the supplementary material. Figure 5 presents the KL divergence results from the 50 replications for the different settings of η and λ . In terms of KL divergence MCEN outperforms SEN in all settings.

A comparison of MSE of coefficient estimates between methods is shown using box plots in Figure 6 and shows similar results to the cases of $p = 12$ and 100 available in the supplementary material. The results show that based on MSE binomial MCEN either outperforms or performs as well as binomial SEN. Figures 7 and 8 report the number of true positive and false positive predictors selected by each method for each combination of η and λ when $p = 300$, and MCEN outperforms SEN by generally selecting more true positive predictors, while the number of false predictors selected varies by the signal size. For smaller signals MCEN selects a smaller number of false predictors, but for larger signals MCEN tends to select more false predictors.

6. Data Example

6.1 Genomics Data

Votavova et al. (2011) collected gene expression profiles, demographic and birth information from 72 pregnant mothers. Using these data we modeled four response variables: placental weight, newborn weight, cotinine level from the mothers' peripheral blood sample and cotinine level from the umbilical cord blood sample. Smoking status, mother's age, mother's BMI, parity, gestational age and expression data for 24,526 gene probes from the mother's peripheral blood sample were used as covariates. Our analysis was limited to the 65 mothers with complete data. From a clinical perspective an accurate model for birth weight would be the primary interest as birth weight is associated with both short and long term negative health outcomes (Turan et al., 2012). Including placental weight as an additional response could potentially be helpful in the MCEN model because previous studies found placental and newborn weight are correlated (Molteni et al., 1978; Panti et al., 2012; Thame et al.,

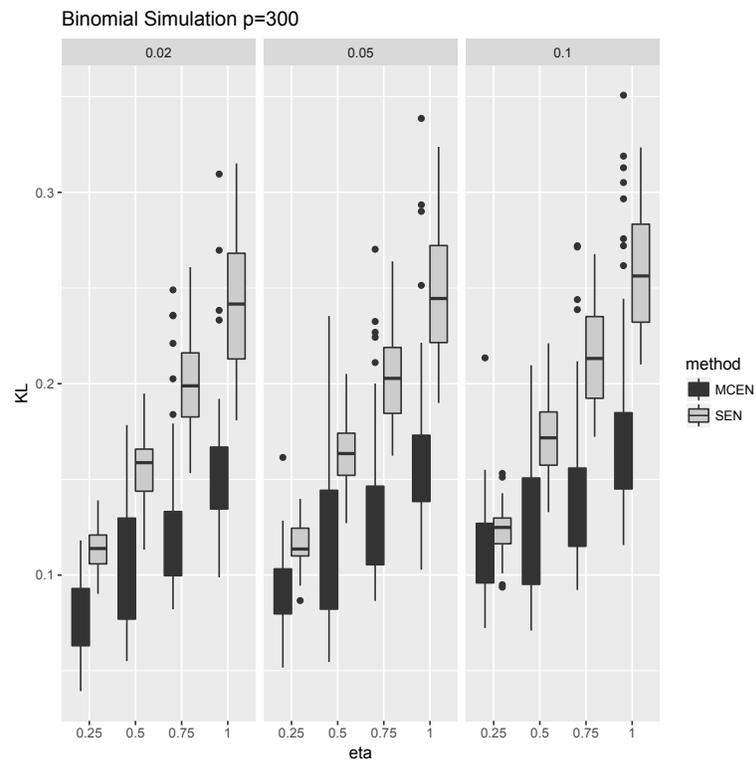


Figure 5: Simulation results comparing binomial SEN and binomial MCEN for $p=300$ at varying values of λ and η . Each box plot represents results for a different value of η , given at the top of the plot.

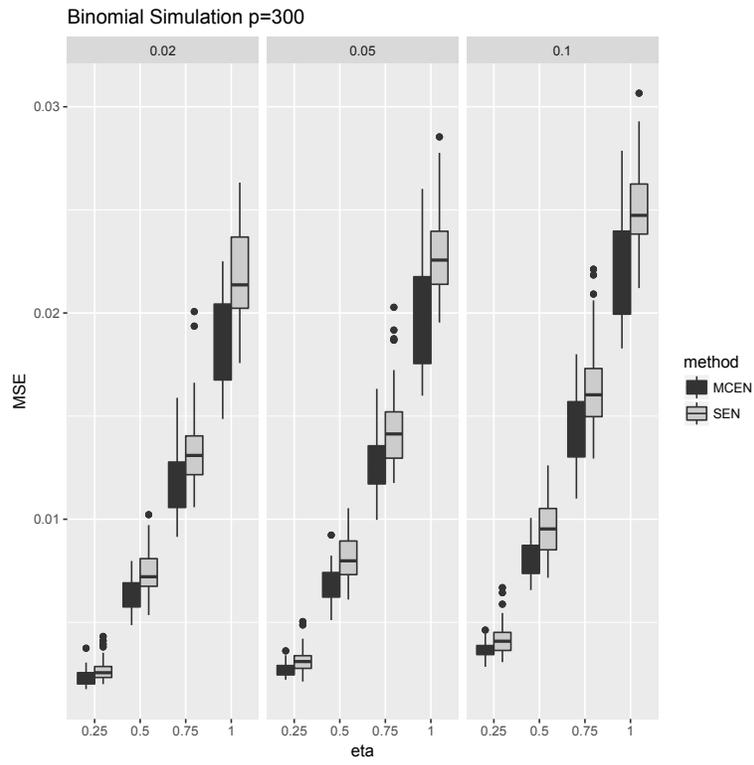


Figure 6: Simulation results comparing MSE of binomial SEN and binomial MCEN when $p=300$ at varying values of λ and η . Each box plot represents results for a different value of η , given at the top of the plot.

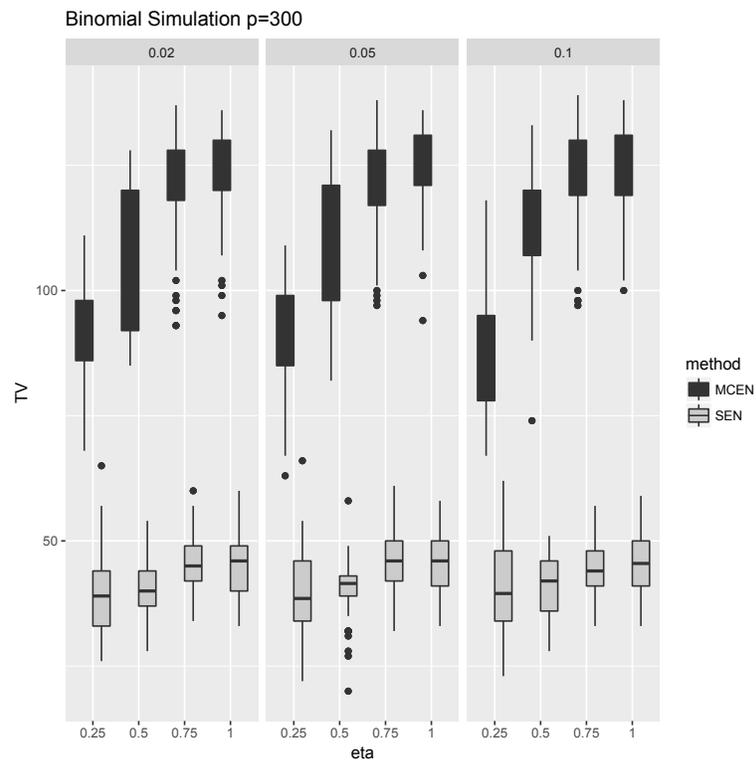


Figure 7: Simulation results comparing TV results by binomial SEN and binomial MCEN when $p=300$ at varying values of λ and η . Each box plot represents results for a different value of η , given at the top of the plot.

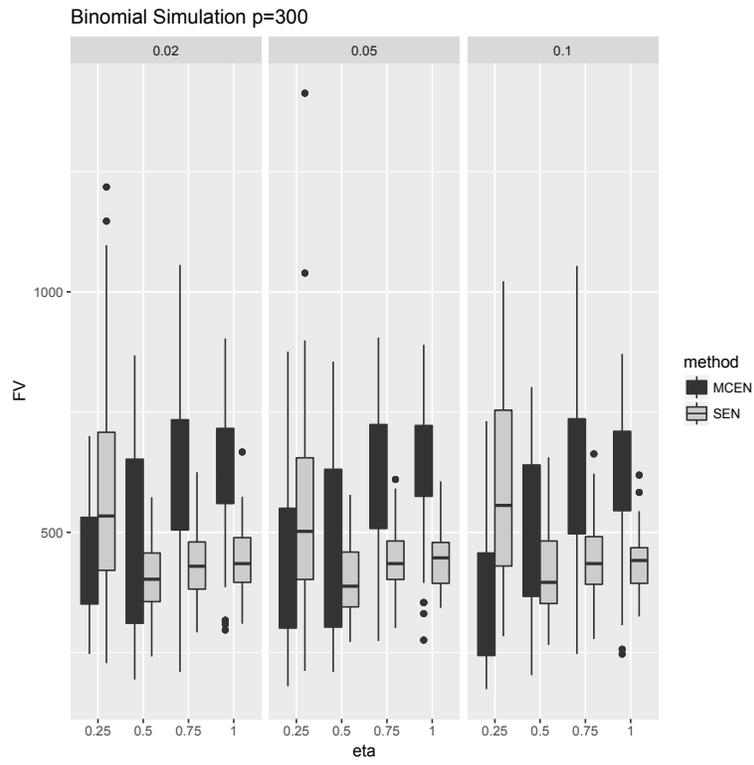


Figure 8: Simulation results comparing FV results by binomial SEN and binomial MCEN when $p=300$ at varying values of λ and η . Each box plot represents results for a different value of η , given at the top of the plot.

2004), but placental weight is hard to use as a predictor since it is observed at birth. The two measurements of cotinine levels are essentially measuring the same thing and are clearly related to smoking status. Thus we can test if these variables were correctly clustered and smoking status selected in the MCEN models.

The same methods used in Section 5.1 are used to fit the data, except we did not implement the TMCEN method as we did not assume to know the true clustering structure of the response variables. To evaluate the methods we randomly partitioned the data into 50 training and 15 testing samples. All four response variables are modeled on the log scale. In the training data all variables are centered and scaled to have mean zero and a standard deviation of one. We filter the gene expression data for each response by using the top 25 genes in terms of absolute value of correlation with a given response. For the joint modeling methods we use a union of the top 25 genes for each response. Models are fit using the training data, then predictions are evaluated on the testing samples. We compare the methods by looking at the ASPE, as defined in Section 5.1. For MCEN we consider clusters of size 1, 2 and 3. The process is repeated 100 times and the ASPE for all methods and responses are included in Figure 9. The MCEN method performs the best for modeling birth weight, the most clinically interesting variable, and is about the same as the other methods for modeling placenta weight. However, it does worse than the other three methods for modeling cotinine level. In all 100 random partitions the MCEN method correctly grouped the two cotinine responses together and selected smoking status as a predictor for those two responses.

6.2 Concession Data

We analyzed 2000 concession transactions from a major event venue. Each transaction is linked with the customer’s information from the venue’s loyalty program. These data are proprietary and cannot be made publicly available. Whether a customer purchases a specific item, 0 if they do and 1 if they do not, is the response and customer information from the loyalty program, such as seat identification and amount spent on previous concession sales, are treated as the covariates. The multiple response setting comes from there being multiple items available for sale at the concession stands. In total there are 34 predictor variables, stemming from purchase history from the venue, ticketing, and seating. The same customer may appear in the data more than once, but any correlation structure is ignored. We analyze two different sets of responses with the same covariates. The point-of-sale system records purchases in two different item set groupings; menu group (7 items) and food group (12 items). The different groups provide different insights into customer habits as the items form different groups.

Similar to the simulation section we compared SEN and MCEN, with tuning parameter selected as described in Section 5.2. For Q , the number of groups, we consider values of $\{1, 2, \dots, 7\}$. We divide 2000 transactions into training and validation sets. There is a time component to our data, which we ignore, but use to evaluate the predictive performance of our models. The first 1000 transactions are used to train our models, with 3-fold cross validation used to select the tuning parameters for both MCEN and SEN. The predictive performance of the models are then compared using the next 1000 transactions.

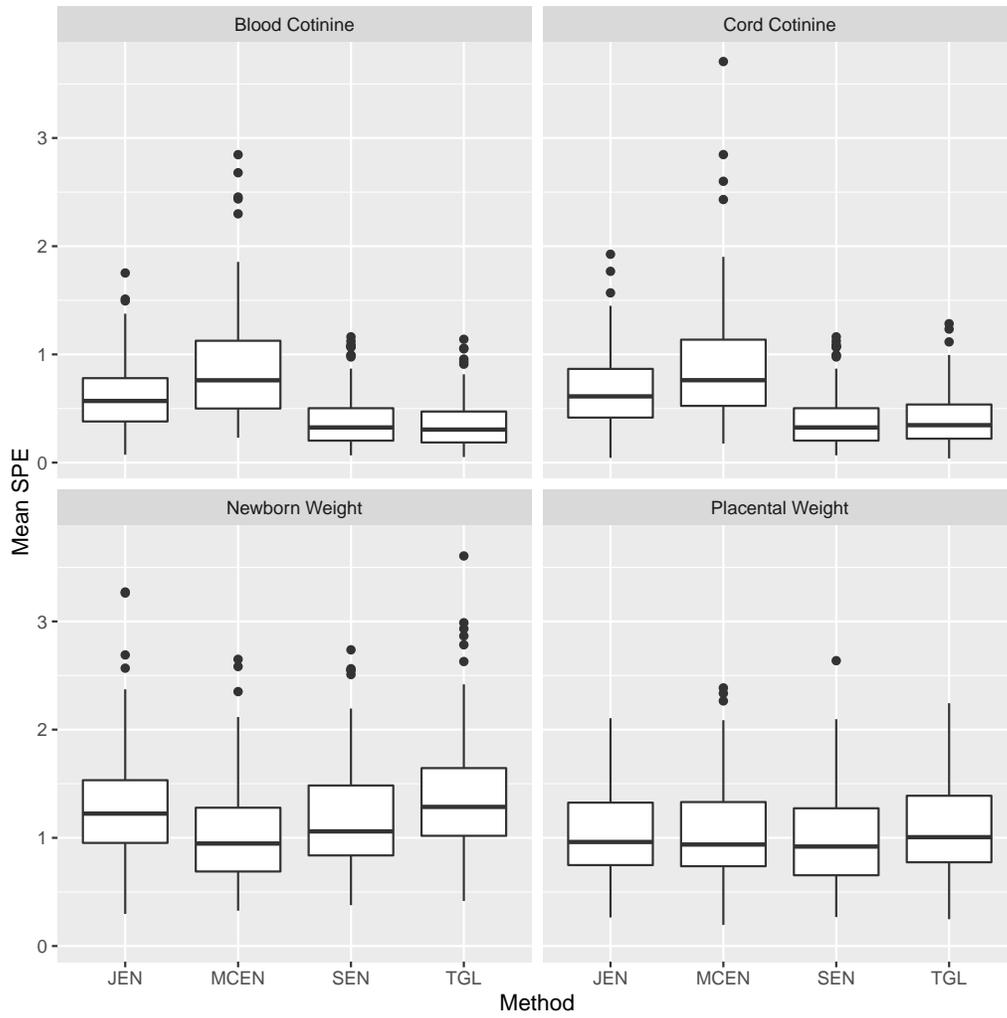


Figure 9: Mean SPE from 100 random partitions

For comparison of the methods we present the ROC curves as a metric for classification performance on the 1000 validation observations. Figure 10 presents the ROC curves and shows that in most situations the binomial logistic MCEN was competitive with SEN. In this analysis MCEN found 3 response clusters where the first cluster contained concession food, the second cluster contained both alcoholic and non-alcoholic beverages, and the third cluster contained all specialty item groups. For comparison we used k-means clustering on the predicted values of the independent elastic net, and selected the number of clusters based on the gap statistic. It selected 2 clusters. The first cluster had concession and both beverage types, while the second cluster contained all specialty items.

The resulting ROC curves for the food group analysis are presented in Figure 11. Five clusters were found by binomial logistic MCEN. The first cluster has popcorn, hamburger, french fries, bottled water, appetizers, and a chicken basket. These correspond to low selling non-alcoholic items. The second cluster consists of hot dogs, craft beer and misc sides, which represents a group of higher selling items. The last three clusters are singleton clusters consisting of non-alcoholic beverages, domestic beer, and liquor. These clusters represent high selling items with different demographics important in each. We also ran k-means clustering on the predicted values from the EN results, and found no distinct clustering using the gap statistic to select the number of clusters. Thus the MCEN method clusters all cold beverages together, while using k-means on fitted values from SEN does not find this clustering. The results of both analyses show that SEN outperforms MCEN using ROC curves. This could be due to the coarseness of MCEN framework, which assumes a similar sparsity structure for all responses. The grouping insights given from the resulting MCEN clusters provide a starting point for investigating each cluster individually with its own MCEN models. This procedure would allow for different levels of sparsity for different clusters. Flexibility such as this should be addressed in extensions of MCEN.

7. Discussion

We present a method for simultaneous estimation of regression coefficients and response clustering for a multivariate response model. The method is introduced for the case of continuous and binary responses. Future work could include extending the model to other GLM settings. Currently, our model imposes the same amount of sparsity on all response models, but this could be relaxed by allowing a sparsity tuning parameter for each individual response or each response group. The `mcen` R package that implements the methods outlined in this article is available on CRAN (Sherwood and Price, 2018).

Define $\ell(B)$ as a likelihood or convex objective function, $P(B, D_q)$ as a distance function between all elements where $\sum_{q=1}^Q P(B, D_q)$ is an optimization problem to separate the r p -dimensional coefficient vectors into Q clusters and $p_\delta(B)$ as a penalty function with tuning parameter δ . Then the MCEN method could be generalized to a larger class of estimators where

$$(\hat{B}, \hat{D}) = \arg \min_{B, D_1, \dots, D_Q} \ell(B) + \gamma \sum_{q=1}^Q P(B, D_q) + p_\delta(B). \quad (32)$$

One example would be to define $P(B, D_q)$ as an L_1 norm to penalize the difference between fitted values, similar to a fused lasso penalty (Tibshirani et al., 2005; Tibshirani, 2014). An

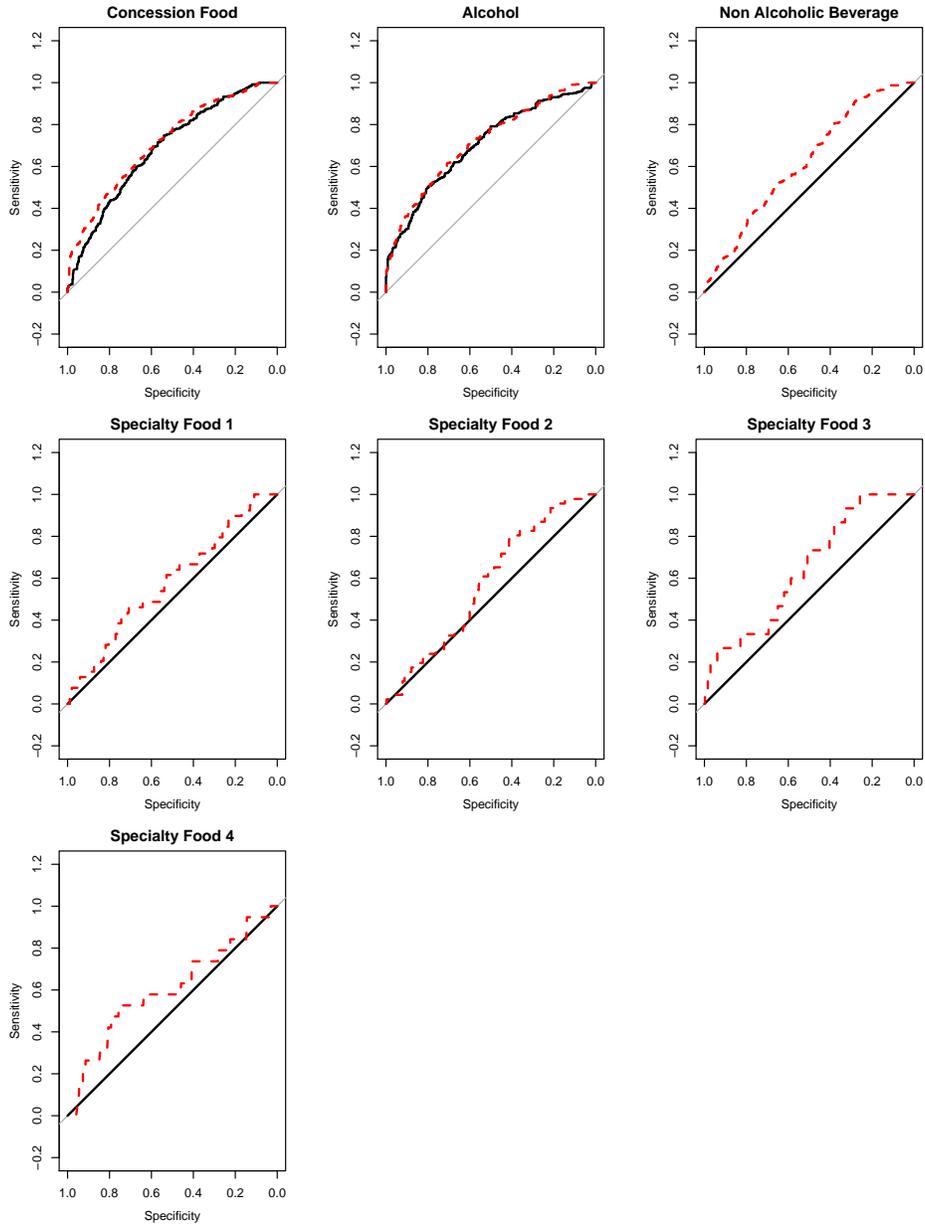


Figure 10: ROC curves for the 1000 validation observations for the menu group item responses. The black lines represent the ROC for MCEN and red for SEN.

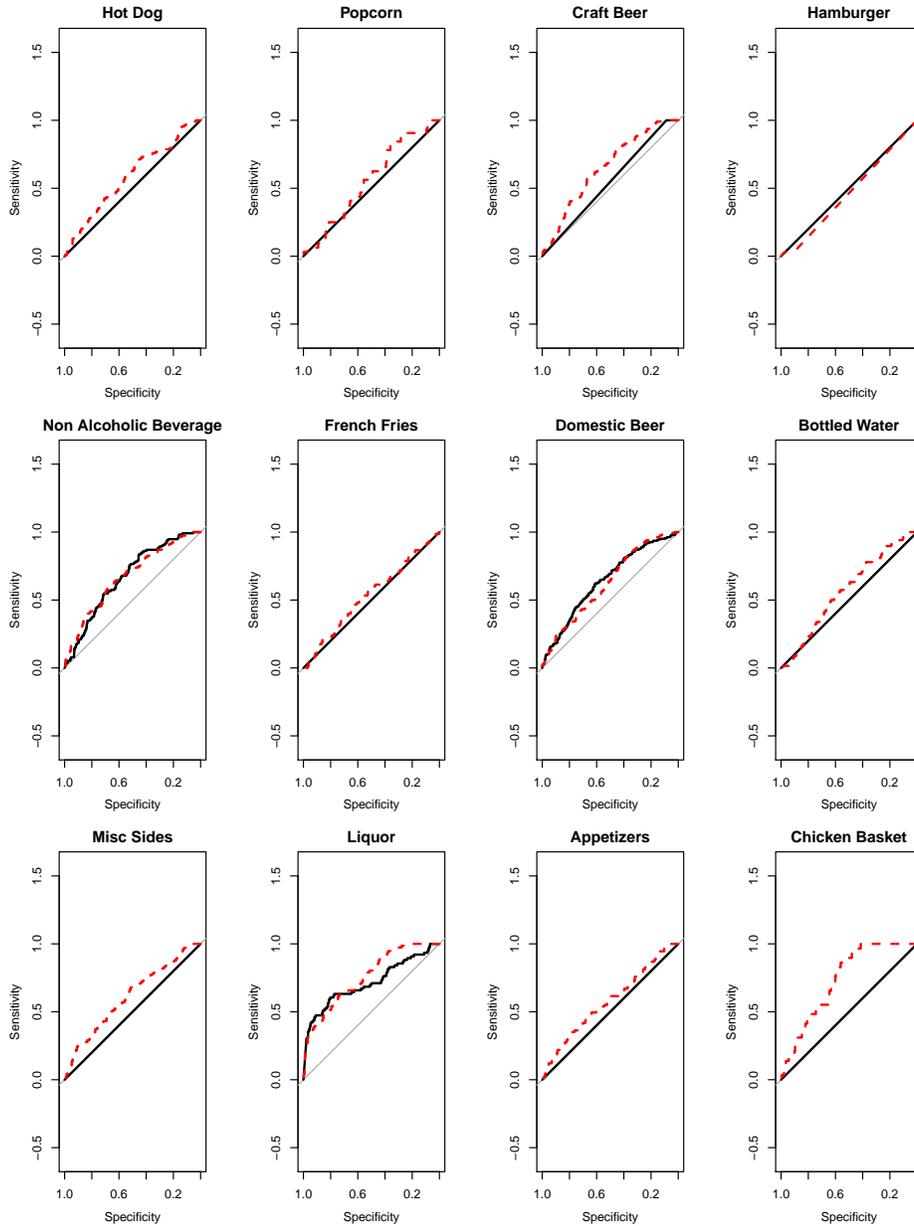


Figure 11: ROC curves for the 1000 validation observations for the food group item responses comparing EN and MCEN. The black lines represent the ROC for MCEN and red for SEN.

advantage of the estimator proposed in this paper is that by defining $P(B, D_q)$ as the L_2 norm squared, when the coefficients are fixed, the minimization problem is equivalent to a k-means problem. However, different definitions of $P(B, D_q)$ may not have well studied clustering algorithms to solve the optimization to define the groupings. One challenge of extending this work would be finding functions $P(B, D_q)$ that become well defined clustering problems when B is known or proposing new algorithms for solving $P(B, D_q)$. Otherwise the two-step algorithm proposed in this paper would not work.

The asymptotics in this paper are limited to consistency of the estimator when groups are known. Zhao and Shojaie (2016) presented an inference framework for a similar estimator that uses a fusion penalty and demonstrated that inference is still possible even if the structure of the graph that determines the fusion penalty is not correctly specified. Extending the results provided here to include inference would be of great use to practitioners and a good topic for future research.

Acknowledgments

We would like to thank Dr. Sara van de Geer and an anonymous reviewer for their insightful comments on early versions of this work. We would also like to thank Suzanna Emelio of the University of Kansas and Dr. Nancy McIntyre for their careful editing of our manuscript. This work was also partially sponsored by Big XII Faculty Fellowships from West Virginia University and the University of Kansas.

Appendix

A.1. Proof of Theorem 1

Proof Define

$$L(B) = \frac{1}{2n} \sum_{i=1}^n \sum_{c=1}^r (y_{ic} - \mathbf{x}_i^T \beta_c)^2 + \frac{\gamma}{2n} \sum_{q=1}^Q \frac{1}{|D_q|} \sum_{l, m \in D_q} \|X(\beta_l - \beta_m)\|_2^2.$$

For $l \in D_q$

$$\frac{\partial}{\partial \beta_l} L(B) = -\frac{1}{n} (X^T Y - X^T X \beta_l) + X^T X \frac{2\gamma}{n|D_q|} \sum_{c \in D_q, c \neq l} \beta_l - \beta_c.$$

Thus,

$$\bar{\beta}_l \left\{ 1 + \frac{2\gamma(|D_q| - 1)}{|D_q|} \right\} - \dot{\beta}_l - \frac{2\gamma}{|D_q|} \sum_{c \in D_q, c \neq l} \bar{\beta}_c = 0. \quad (33)$$

Therefore for $l, m \in D_q$

$$\begin{aligned} & \bar{\beta}_l \left\{ 1 + \frac{2\gamma(|D_q| - 1)}{|D_q|} \right\} - \dot{\beta}_l - \frac{2\gamma}{|D_q|} \sum_{c \in D_q, c \neq l} \bar{\beta}_c \\ & - \bar{\beta}_m \left\{ 1 + \frac{2\gamma(|D_q| - 1)}{|D_q|} \right\} - \dot{\beta}_m - \frac{2\gamma}{|D_q|} \sum_{c \in D_q, c \neq m} \bar{\beta}_c \\ & = (\bar{\beta}_l - \bar{\beta}_m) (1 + 2\gamma) - \dot{\beta}_l + \dot{\beta}_m = 0. \end{aligned}$$

Therefore for $l, m \in D_q$ and $l \neq m$

$$\bar{\beta}_m = \bar{\beta}_l + \frac{1}{1+2\gamma} (\dot{\beta}_m - \dot{\beta}_l). \quad (34)$$

Combining (33) and (34) gives

$$\begin{aligned} \bar{\beta}_l \left\{ 1 + \frac{2\gamma(|D_q| - 1)}{|D_q|} \right\} &= \dot{\beta}_l + \frac{2\gamma}{|D_q|} \sum_{c \in D_q, c \neq l} \bar{\beta}_l + \frac{1}{1+2\gamma} (\dot{\beta}_c - \dot{\beta}_l) \\ &= \dot{\beta}_l + \frac{2\gamma(|D_q| - 1)}{|D_q|} \bar{\beta}_l + \frac{2\gamma}{(1+2\gamma)|D_q|} \sum_{c \in D_q, c \neq l} (\dot{\beta}_c - \dot{\beta}_l), \end{aligned}$$

which completes the proof. ■

A.2. Proof of Theorem 2

Proof It is assumed that $E(\epsilon_{ic}^2) = 1$ and for $c \neq k$ that $E(\epsilon_{ic}\epsilon_{ik}) = \rho$. Thus, note that for any $v \in \{1, \dots, r\}$

$$\begin{aligned} \text{Var}(\bar{\beta}_v) &= \text{Var} \left\{ \frac{|D_q| + 2\gamma}{(1+2\gamma)|D_q|} \dot{\beta}_v + \frac{2\gamma}{(1+2\gamma)|D_q|} \sum_{s \in D_q, s \neq v} \dot{\beta}_s \right\} \\ &= (X^T X)^{-1} \left\{ \frac{|D_q|(|D_q| + 4\gamma + 4\gamma^2)}{(1+2\gamma)^2 |D_q|^2} + 4\rho\gamma(|D_q| - 1) \frac{|D_q| + 2\gamma|D_q| - 2\gamma}{(1+2\gamma)^2 |D_q|^2} \right\}. \end{aligned}$$

Define $\mathbf{b}_v = \sum_{s \in D_q, s \neq v} (\beta_s^* - \beta_v^*)$. The squared bias term is then

$$\begin{aligned} &E \left[\{E(\bar{\beta}_v) - \beta_v^*\}' \{E(\bar{\beta}_v) - \beta_v^*\} \right] \\ &= E \left[\left\{ \beta_v^* + \frac{2\gamma}{(1+2\gamma)|D_q|} \mathbf{b}_v - \beta_v^* \right\}' \left\{ \beta_v^* + \frac{2\gamma}{(1+2\gamma)|D_q|} \mathbf{b}_v - \beta_v^* \right\} \right] \\ &= \frac{4\gamma^2}{(1+2\gamma)^2 |D_q|^2} \|\mathbf{b}_v\|_2^2. \end{aligned}$$

Let $\omega = \text{Trace} \left\{ (X^T X)^{-1} \right\}$ then MSE of $\bar{\beta}_v$ will be smaller than MSE of $\dot{\beta}_v$ if

$$\begin{aligned} &\omega \left\{ \frac{|D_q|(|D_q| + 4\gamma + 4\gamma^2)}{(1+2\gamma)^2 |D_q|^2} + 4\rho\gamma(|D_q| - 1) \frac{|D_q| + 2\gamma|D_q| - 2\gamma}{(1+2\gamma)^2 |D_q|^2} \right\} \\ &+ \frac{4\gamma^2}{(1+2\gamma)^2 |D_q|^2} \|\mathbf{b}_v\|_2^2 \\ &< \omega, \end{aligned}$$

which is equivalent to

$$\gamma \|\mathbf{b}_v\|_2^2 < \omega \left\{ |D_q|(|D_q| - 1) + \gamma |D_q|(|D_q| - 1) \right. \\ \left. - \rho \{ (|D_q| - 1)|D_q| + 2\gamma(|D_q| - 1)^2 \} \right\}.$$

Note that, $\omega |D_q|(|D_q| - 1)(1 - \rho) > 0$ and thus if $\|\mathbf{b}_v\|_2^2 \leq \omega(|D_q| - 1) \{ |D_q| - 2\rho(|D_q| - 1) \}$ then the MSE of $\tilde{\beta}_v$ is smaller than the MSE of $\hat{\beta}_v$ for any $\gamma > 0$. Otherwise, the MSE of $\tilde{\beta}_v$ will be smaller for any $\gamma \in \left(0, \frac{\omega |D_q|(|D_q| - 1)(1 - \rho)}{\|\mathbf{b}_v\|_2^2 - \omega(|D_q| - 1) \{ |D_q| - 2\rho(|D_q| - 1) \}} \right)$. Thus for any $v \in \{1, \dots, r\}$ then any $\gamma > 0$ or any γ sufficiently small will result in $\tilde{\beta}_v$ having a smaller MSE than $\hat{\beta}_v$. The proof is complete because we can then find a γ sufficiently small that will result in $\tilde{\beta}_v$ having a smaller MSE than $\hat{\beta}_v$ for all $v \in \{1, \dots, r\}$. \blacksquare

A.3. Proof of Corollary 3

The proof of Corollary 3 is similar to the proof of Theorem 1 and only changes with respect to the expected loss rather than the observed loss.

A.4. Theorem 4

The proof of Theorem 4 will include some new definitions and an alternative formulation of (5). In our proof we use a vectorized version of many of the matrices. Let $\tilde{\mathbf{Y}} = \text{vec}(Y)$, $\tilde{\mathbf{B}} = \text{vec}(B)$, $\tilde{\mathbf{B}}' = \text{vec}(\hat{B})$ and $\tilde{\mathbf{E}} = \text{vec}(E)$. Define $\mathbf{A}_{m,s} \in \mathcal{R}^r$, where $(m, s) \in D_q$, with $\sqrt{\frac{1}{|D_q|}}$ in the m th element, $-\sqrt{\frac{1}{|D_q|}}$ in the s th element and 0 in all other elements, $A_{D_q} \in \mathcal{R}^{|D_q|(|D_q|-1) \times r}$ as the matrix with row vectors $\mathbf{A}_{m,s}$ where $(m, s) \in D_q$, and $A_D \equiv \left(A_{D_1}^T, \dots, A_{D_Q}^T \right)^T \in \mathcal{R}^{\sum_{q=1}^Q |D_q|(|D_q|-1) \times r}$.

Then the objective function from (5) can be restated as

$$\frac{1}{2n} \left[\tilde{\beta}^T \left\{ \tilde{X}^T \tilde{X} + \gamma (A_D \otimes X)^T (A_D \otimes X) \right\} \tilde{\beta} - 2\tilde{\mathbf{Y}}^T \tilde{X} \tilde{\beta} \right] + \delta \|\tilde{\beta}\|_1 \\ = \ell(\tilde{\beta}) + \delta g(\tilde{\beta}).$$

In addition define, $\tilde{\ell}(\Delta, \tilde{\beta}) \equiv \ell(\tilde{\beta} + \Delta) - \ell(\tilde{\beta}) - \langle \nabla \ell(\tilde{\beta}), \Delta \rangle$.

First, we will present some lemmas that are helpful in proving Theorem 4. A general outline of the proof for Theorem 4 is by using the triangle inequality we have $\|\text{vec}(\tilde{B} - B^*)\|_2 \leq \|\text{vec}(\tilde{B} - \hat{B})\|_2 + \|\tilde{\beta}' - \tilde{\beta}^*\|_2$. Completing the proof is done by establishing upper bounds for $\|\text{vec}(\tilde{B} - \hat{B})\|_2$ and $\|\tilde{\beta}' - \tilde{\beta}^*\|_2$. Much of the proof will require working with $\tilde{\beta}'$ and we introduce the following notation to easily relate $\tilde{\beta}'$ and $\tilde{\beta}^*$. For response l in group q define $\mathbf{H}_l = \frac{1}{\sqrt{|D_q|}} \sum_{c \in D_q, c \neq l} \mathbf{A}_{c,l}$ where $\mathbf{H}_l \in \mathcal{R}^r$ and $H = (\mathbf{H}_1, \dots, \mathbf{H}_r)^T \in \mathcal{R}^{r \times r}$. Then we have

$$\tilde{\beta}' = \left\{ \left(I_r + \frac{2\gamma}{2\gamma + 1} H \right) \otimes I_p \right\} \tilde{\beta}^*.$$

For response l is in group q define $\mathbf{U}_l = \frac{1}{|D_q|} \sum_{k \in D_q} (\beta_k^* - \beta_l^*)$ where $\mathbf{U}_l \in \mathcal{R}^p$ and $U = (\mathbf{U}_1, \dots, \mathbf{U}_r)$ with $U \in \mathcal{R}^{p \times r}$ and $\tilde{\mathbf{U}} = \text{vec}(U) \in \mathcal{R}^{pr}$, then

$$\left\| \text{vec}(\hat{B} - B^*) \right\|_2 = \frac{2\gamma}{1+2\gamma} \left\| (H \otimes I_p) \tilde{\beta}^* \right\|_2 = \frac{2\gamma}{1+2\gamma} \left\| \tilde{\mathbf{U}} \right\|_2.$$

Lemma 5 *Under assumption A3*

$$\tilde{\ell}(\Delta, \tilde{\beta}') \geq \kappa \|\Delta\|_2^2 \text{ for all } \Delta \in \mathcal{C}.$$

Proof From the definition of $\tilde{\ell}(\Delta, \tilde{\beta}')$, assumption A3 and that $\Delta \in \mathcal{C}$ it follows that

$$\begin{aligned} \tilde{\ell}(\Delta, \tilde{\beta}') &= \frac{1}{2n} \Delta^T \left\{ \tilde{X}^T \tilde{X} + \gamma (A_D \otimes X)^T (A_D \otimes X) \right\} \Delta \\ &\geq \frac{1}{2n} \Delta^T \tilde{X}^T \tilde{X} \Delta \\ &\geq \frac{\kappa}{2} \|\Delta\|_2^2. \end{aligned}$$

■

For any vector $\mathbf{a} = (a_1, \dots, a_{pr})^T \in \mathcal{R}^{pr}$ we define the $\|\mathbf{a}\|_\infty$ as the L_∞ norm of \mathbf{a} , that is $\|\mathbf{a}\|_\infty = \max_i |a_i|$.

Lemma 6 *For \bar{B} from (5), under assumptions A1-A4 with $\delta \geq 2 \left\| \nabla \ell(\tilde{\beta}') \right\|_\infty$ then there exists a positive constant c_3 such that*

$$\left\| \text{vec}(\bar{B} - \hat{B}) \right\|_2^2 \leq 9 \frac{\delta^2}{\kappa^2} s.$$

Proof Define the set $\acute{S} = \{j \in \{1, \dots, rp\}, \tilde{\beta}'_j \neq 0\}$. By assumption A5 and Corollary 3 $\acute{S} = S$, that is $\tilde{\beta}'_j = 0$ if and only if $\tilde{\beta}^*_j = 0$. Define $\psi(\mathcal{M}) \equiv \sup_{\mathbf{u} \in \mathcal{M} \setminus \{\mathbf{0}\}} \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_2}$. Note that $\psi\{\mathcal{M}(S)\} = \sqrt{s}$. Also, note that the dual norm of the L_1 norm is the L_∞ norm. Results follow from Theorem 1 of Negahban et al. (2012) and Lemma 5. ■

Lemma 7 *Under the conditions of Theorem 4 there exists positive c_1, c_2 and c_3 such that*

$$\left\| \text{vec}(\bar{B} - \hat{B}) \right\|_2 \leq \frac{48\sigma}{\kappa} \sqrt{\frac{s \log(rp)}{n}},$$

with probability at least $1 - c_1 \exp(-c_2 n \delta^2)$.

Proof If we can find positive constants c_1 and c_2 such that with probability at least $1 - c_1 \exp(-c_2 n \delta^2)$ that $\delta \geq 2 \left\| \nabla \ell(\tilde{\beta}') \right\|_\infty$ then proof will be complete by Lemma 6 and by

the condition that $\delta = 16\sigma\sqrt{\frac{\log(rp)}{n}}$. Note that

$$\begin{aligned}
 2 \left\| \nabla \ell(\tilde{\beta}') \right\|_{\infty} &= 2 \left\| \frac{1}{n} \left[\left\{ \tilde{X}^T \tilde{X} + \gamma(A_D \otimes X)^T (A_D \otimes X) \right\} \tilde{\beta}' - \tilde{X}^T \tilde{Y} \right] \right\|_{\infty} \\
 &= 2 \left\| \frac{1}{n} \left[\left\{ \tilde{X}^T \tilde{X} + \gamma(A_D \otimes X)^T (A_D \otimes X) \right\} \left\{ \left(I_r + \frac{2\gamma}{2\gamma+1} H \right) \otimes I_p \right\} \tilde{\beta}^* - \tilde{X}^T (\tilde{X} \tilde{\beta}^* + \tilde{\mathbf{E}}) \right] \right\|_{\infty} \\
 &\leq 2 \left\| \frac{2\gamma}{n(1+2\gamma)} \tilde{X}^T \tilde{X} \tilde{\mathbf{U}} \right\|_{\infty} + 2 \left\| \frac{\gamma}{n} (A_D \otimes X)^T (A_D \otimes X) \tilde{\beta}^* \right\|_{\infty} \\
 &\quad + 2 \left\| \frac{2\gamma^2}{n(1+2\gamma)} (A_D \otimes X)^T (A_D \otimes X) \tilde{\mathbf{U}} \right\|_{\infty} + 2 \left\| \frac{1}{n} \tilde{X}^T \tilde{\mathbf{E}} \right\|_{\infty}.
 \end{aligned}$$

Next, we will establish upper bounds for the first three terms. Define $I(l \in D_q)$ to be 1 if $l \in D_q$ and zero otherwise. Using the definition of $\tilde{\mathbf{U}}$ and assumptions A4-A6,

$$\begin{aligned}
 2 \left\| \frac{2\gamma}{n(1+2\gamma)} \tilde{X}^T \tilde{X} \tilde{\mathbf{U}} \right\|_{\infty} &= \frac{4\gamma}{1+2\gamma} \max_{l \in \{1, \dots, r\}} \left\| \frac{1}{n} X^T X \sum_{q=1}^Q I(l \in D_q) \sum_{k \in D_q} \frac{1}{|D_q|} (\beta_k^* - \beta_l^*) \right\|_{\infty} \\
 &\leq \frac{4\gamma}{1+2\gamma} \rho_{\max} \max_{l \in \{1, \dots, r\}} \left\| \sum_{q=1}^Q I(l \in D_q) \sum_{k \in D_q} \frac{1}{|D_q|} (\beta_k^* - \beta_l^*) \right\|_2 \\
 &\leq \frac{4\gamma}{1+2\gamma} \rho_{\max} \acute{b}.
 \end{aligned}$$

Using assumptions A4-A6,

$$\begin{aligned}
 2 \left\| \frac{\gamma}{n} (A_D \otimes X)^T (A_D \otimes X) \tilde{\beta}^* \right\|_{\infty} &= 2\gamma \max_{l \in \{1, \dots, r\}} \left\| \frac{1}{n} X^T X \sum_{k, l \in D_q, k \neq l} \frac{1}{|D_q|} (\beta_k^* - \beta_l^*) \right\|_{\infty} \\
 &\leq 2\gamma \rho_{\max} \max_{l \in \{1, \dots, r\}} \left\| \sum_{k, l \in D_q} \frac{1}{|D_q|} (\beta_k^* - \beta_l^*) \right\|_2 \\
 &\leq 2\gamma \rho_{\max} \acute{b}.
 \end{aligned}$$

Note that for $a \in D_q$ and $b \in D_q$ that

$$\begin{aligned}
 \mathbf{U}_a - \mathbf{U}_b &= \frac{1}{|D_q|} \left(\sum_{l \in D_q} \beta_l^* - \beta_a^* - \sum_{l \in D_q} \beta_l^* - \beta_b^* \right) \\
 &= \frac{1}{|D_q|} \sum_{l \in D_q} \beta_b^* - \beta_a^* = \beta_b^* - \beta_a^*.
 \end{aligned}$$

Therefore

$$\begin{aligned}
 2 \left\| \frac{2\gamma^2}{n(1+2\gamma)} (A_D \otimes X)^T (A_D \otimes X) \tilde{\mathbf{U}} \right\|_\infty &= \frac{4\gamma^2}{1+2\gamma} \max_{l \in \{1, \dots, r\}} \left\| \frac{1}{n} X^T X \sum_{q=1}^Q I(l \in D_q) \sum_{k \in D_q} \frac{1}{|D_q|} (\mathbf{U}_k - \mathbf{U}_l) \right\|_\infty \\
 &= \frac{4\gamma^2}{1+2\gamma} \max_{l \in \{1, \dots, r\}} \left\| \frac{1}{n} X^T X \sum_{q=1}^Q I(l \in D_q) \sum_{k \in D_q} \frac{1}{|D_q|} \beta_l^* - \beta_k^* \right\|_\infty \\
 &\leq \frac{4\gamma^2}{1+2\gamma} \rho_{\max} \max_{l \in \{1, \dots, r\}} \left\| \sum_{q=1}^Q I(l \in D_q) \sum_{k \in D_q} \frac{1}{|D_q|} \beta_l^* - \beta_k^* \right\|_2 \\
 &\leq \frac{4\gamma^2}{1+2\gamma} \rho_{\max} \acute{b} \\
 &\leq 2\gamma \rho_{\max} \acute{b}.
 \end{aligned}$$

Under assumptions A1 and A2 it follows that

$$P \left(\left\| \frac{1}{n} \tilde{X}^T \tilde{\mathbf{E}} \right\|_\infty > t \right) \leq 2 \exp \left\{ \frac{-nt^2}{2\sigma^2} + \log(rp) \right\}. \quad (35)$$

Thus,

$$\begin{aligned}
 P \left\{ \delta \geq 2 \left\| \nabla \ell(\tilde{\beta}') \right\|_\infty \right\} &\geq P \left\{ \delta \geq 2 \left\| \frac{1}{n} \tilde{X}^T \tilde{\mathbf{E}} \right\|_\infty + \rho_{\max} \acute{b} \left(\frac{4\gamma}{1+2\gamma} + 2\gamma + 2\gamma \right) \right\} \\
 &\geq P \left(\delta \geq 2 \left\| \frac{1}{n} \tilde{X}^T \tilde{\mathbf{E}} \right\|_\infty + 8\gamma \rho_{\max} \acute{b} \right) \\
 &\geq P \left(\frac{3}{16} \delta \geq \left\| \frac{1}{n} \tilde{X}^T \tilde{\mathbf{E}} \right\|_\infty \right) \\
 &\geq 1 - 2 \exp \left\{ \frac{-9n\delta^2}{16^2 2\sigma^2} + \log(rp) \right\} \\
 &= 1 - 2 \exp \left\{ -\frac{7}{2} \log(rp) \right\}.
 \end{aligned}$$

Set $c_1 = 2$ and $c_2 = \frac{7}{2}$ and the proof is complete. \blacksquare

Proof of Theorem 4

Proof Applying the triangle inequality we have

$$\left\| \text{vec}(\hat{\beta} - \beta^*) \right\|_2 \leq \left\| \text{vec}(\hat{\beta} - \tilde{\beta}') \right\|_2 + \left\| \tilde{\beta}' - \tilde{\beta}^* \right\|_2. \quad (36)$$

For the second term using the upper bound for γ stated in the conditions for Theorem 4 and assumptions A4 and A5 it follows that

$$\begin{aligned}
 \left\| \tilde{\beta}' - \tilde{\beta}^* \right\|_2 &= \frac{2\gamma}{1+2\gamma} \left\| \tilde{\mathbf{U}} \right\|_2 \\
 &\leq 2\gamma \sqrt{s\acute{b}} \leq \frac{5\sigma}{2\rho_{\max}} \sqrt{\frac{s \log(rp)}{n}}.
 \end{aligned}$$

Combining the above inequality with (36) and Lemma 7 it follows that there exists positive constants c_1 and c_2 such that

$$\|\text{vec}(\bar{B} - B^*)\|_2 \leq \frac{48\sigma}{\kappa} \sqrt{\frac{s \log(rp)}{n}} + \frac{5\sigma}{2\rho_{\max}} \sqrt{\frac{s \log(rp)}{n}},$$

with probability at least $1 - c_1 \exp(-c_2 n \delta^2)$. To complete the proof set $c_3 = 48$ and $c_4 = \frac{5}{2}$. ■

References

- Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.
- Leo Breiman and Jerome H. Friedman. Predicting multivariate responses in multiple linear regression. *J. R. Statist. Soc. B*, 59(1):3–54, 1997.
- Peter Bühlmann, Philipp Rütimann, Sara van de Greer, and Cun-Hui Zhang. Correlated variables in regression: Clustering and sparse estimation. *J. Statist. Plannng Inf*, 143(11):1835–1858, 2013.
- Emmanuel Candes and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n . *Ann. Statist.*, 35(6):2313–2351, 2007.
- Yupeng Chen, Raghuram Iyengar, and Garud Iyengar. Modeling multimodal continuous heterogeneity in conjoint analysis—a sparse learning approach. *Marketing Science*, 36(1):140–156, 2016.
- R. Dennis Cook and Xin Zhang. Foundations for envelope models and methods. *J. Am. Statist. Ass*, 110(510):599–611, 2015.
- R. Dennis Cook, Bing Li, and Francesca Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statistica Sinica*, 20:927–1010, 2010.
- Julian J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press, 2006.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2008.
- Holger Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.
- Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- Jian Huang, Shuangge, Hongzhe Li, and Cun-Hui Zhang. The sparse laplacian shrinkage estimator for high-dimensional regression. *Annals of Statistics*, 39(4), 2011.

- Ozge Yucler Kasap, Nevzat Ekmekci, and Utku Gorkem Ketenci. Combining logistic regression analysis and association rule mining via mlr algorithm. In *ICSEA 2016 The Eleventh International Conference on Software Engineering Advances*, pages 154–159. IARA, 2016.
- Seyoung Kim and Eric P. Xing. Tree-guided group lasso for multi-response regression with structured sparsity with an application to eqtl mapping. *The Annals of Applied Statistics*, 6(3):1095–1117, 2012.
- Seyoung Kim, Kyung-Ah Sohn, and Eric P. Xing. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25(12):i204–i212, 2009.
- Wonyul Lee and Yufeng Liu. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *Journal of Multivariate Analysis*, 111:241–255, 2012.
- Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.
- Caiyan Li and Hongzhe Li. Variable selection and regression analysis for graph-structure covariates with an application to geneomics. *The Annals of Applied Statistics*, 4(3):1498–1516, 2010.
- Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 2009.
- Aaron J. Molstad and Adam J. Rothman. Indirect multivariate response linear regression. *Biometrika*, 3(103):595–607, 2016.
- R. A. Molteni, S. J. Stys, and F. C. Battaglia. Relationship of fetal and placental weight in human beings: fetal/placental weight ratios at various gestational ages and birth weight distributions. *The Journal of Reproductive Medicine*, 21:327–334, 1978.
- Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m -estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.
- Abubakar A. Panti, Bissala A. Ekele, Emmanuel I. Nwobodo, and Ahmed Yakubu. The relationship between the weight of the placenta and birth weight of the neonate in a nigerian hospital. *Nigerian Medical Journal*, 53(2):80–84, 2012.
- Jie Peng, Ji Zhu, Anna Beramaschi, Wonshik Han, Dong-Young Noh, Jonathan R. Pollac, and Pei Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics*, 4(1):53–77, 2010.
- Bradley S. Price, Charles J. Geyer, and Adam J. Rothman. Automatic response category combination in multinomial logistic regression. <https://arxiv.org/abs/1705.03594>, May 2017.

- Piyush Rai, Abhishek Kumar, and Hal Daume. Simultaneously leveraging output and task structures for multiple-output regression. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3185–3193. Curran Associates, Inc., 2012. URL <http://papers.nips.cc/paper/4501-simultaneously-leveraging-output-and-task-structures-for-multiple-output-regression.pdf>.
- Alessandro Rinaldo. Properties and refinements of the fused lasso. *Ann. Statist.*, 37(5B): 2597–3097, 2009.
- Adam J. Rothman, Elizaveta Levina, and Ji Zhu. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962, 2010.
- Ben Sherwood and Bradley S. Price. *mcen: Multivariate Cluster Elastic Net*, 2018. R package version 1.0.
- Qiang Sun, Hongtu Zhu, Yufeng Liu, and Joseph G. Ibrahim. Sprem: Sparse projection regression model for high-dimensional linear regression. *J. Am. Statist. Ass.*, 110(509): 289–302, 2015.
- M. Thame, C. Osmond, F. Bennett, R. Wilks, and T. Forrester. Fetal growth is directly related to maternal anthropometry and placental volume. *European Journal of Clinical Nutrition*, 58:894–900, 2004.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58(1):267–288, 1996.
- Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B*, pages 91–108, 2005.
- Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.
- Nahid Turan, Mohamed F. Ghalwash, Sunita Katari, Christos Coutifaris, Zoran Obradovic, and Carmen Sapienza. Dna methylation differences at growth related genes correlate with birth weight: a molecular signature linked to developmental origins of adult disease? *BMC Medical Genomics*, 5(10):1–21, 2012.
- Sara van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.
- Hana Votavova, Michaela Dostalova Merkerova, Kamilia Fejglova, A. Vasikova, Zdenek Krejcik, A. Pastorkova, N. Tabashidze, J. Topinka, M. Veleminsky Jr., R.J. Sram, and R. Brdicka. Transcriptome alterations in maternal and fetal cells induced by tobacco smoke. *Placenta*, 32:763–770, 2011.
- Daniela M. Witten, Ali Shojaie, and Fan Zhang. The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics*, 56(1):112–122, 2014.

- Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68(1):49–67, 2005.
- Sen Zhao and Ali Shojaie. A significance test for graph constrained estimation. *Biometrics*, 72(2):484–493, 2016.
- Weiqliang Zhou, Ben Sherwood, Zhicheng Ji, Yingchao Xue, Fang Du, Jiawei Bai, Mingyao Ying, and Hongkai Ji. Genome-wide prediction of dnase i hypersensitivity using gene expression. *Nature Communications*, 8:1–17, 2017.
- Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67:301–320, 2005.