On the Use of Minimum Penalties in Statistical Learning

Ben Sherwood\textsuperscript{a} and Bradley S. Price\textsuperscript{b}

\textsuperscript{a}School of Business, University of Kansas, Lawrence, KS; \textsuperscript{b}Management Information Systems Department, West Virginia University, Morgantown, WV

\textbf{ABSTRACT}

Modern multivariate machine learning and statistical methodologies estimate parameters of interest while leveraging prior knowledge of the association between outcome variables. The methods that do allow for estimation of relationships do so typically through an error covariance matrix in multivariate regression which does not generalize to other types of models. In this article we proposed the MinPen framework to simultaneously estimate regression coefficients and detect relationships between outcome variables using common assumptions. The MinPen framework uses a novel penalty based on the minimum function to simultaneously detect and exploit relationships between responses. An iterative algorithm is proposed as a solution to the nonconvex optimization. Theoretical results such as high-dimensional convergence rates, model selection consistency, and a framework for post selection inference are provided. We extend the proposed MinPen framework to other exponential family loss functions, with a specific focus on multiple binomial responses. Tuning parameter selection is also addressed. Finally, simulations and two data examples are presented to show the finite sample properties of this framework. Supplemental material providing proofs, additional simulations, code, and datasets are available online.

\section{Introduction}

Multivariate, also known as multiple response or multitask, regression is a popular approach for modeling multiple responses (outcomes) given a common set of predictors (inputs). Joint modeling of responses, as opposed to modeling each response separately, is only interesting if the responses are related. Leveraging these relationships should improve estimation of the model, but in practice these relationships may not always be known. In this work we propose a framework to simultaneously estimate the regression coefficients and detect relationships between the response variables. The framework also allows for simultaneous variable selection.

Consider the sample of independent data, \(\{(x_i, y_i)\}_{i=1}^n\), where \(x_i \in \mathbb{R}^p\) and \(y_i \in \mathbb{R}^r\) with a model of

\[ y_i = B^* x_i + \epsilon_i, \tag{1} \]

where \(B^* = (\beta_1^*, \ldots, \beta_r^*) \in \mathbb{R}^{p \times r}\) and \(\epsilon_i \in \mathbb{R}^r\) are iid random normal vectors with mean zero and diagonal covariance matrix \(\Sigma\). A popular estimator for \(B^*\) is

\[ \hat{B} = \arg \min_{B \in \mathbb{R}^{p \times r}} \frac{1}{2n} \sum_{k=1}^p \sum_{i=1}^n (y_{ik} - x_i^T \beta_k)^2, \tag{2} \]

which has desirable qualities such as consistency and asymptotic normality under mild conditions, and is the maximum likelihood estimator. However, there is something unsatisfying about not using correlation of the responses in the estimation. To improve efficiency in estimation of coefficients, while accounting for correlation across the errors Rothman, Levina, and Zhu (2010) (MRCE), Witten and Tibshirani (2009) (SCOUT), and Lee and Liu (2012) have all proposed methods which simultaneously estimates the regression coefficients and the inverse covariance matrix of the errors. These methods all aim to improve estimation of regression coefficients by exploiting correlation across the responses that is not explained by the predictors.

We propose an alternative approach that focuses on relationships across the responses that can be explained, due to similarity in the regression coefficients. Specifically, we consider the \(k\)th and \(l\)th response related if \(\beta_k^*\) is similar to \(\beta_l^*\) or \(-\beta_l^*\). If relationships were known a priori, then for each response \(k\) one could define three disjoint sets: (a) \(P_k\), a set of responses positively related to response \(k\); (b) \(N_k\), a set of responses negatively related to response \(k\); and (c) \(Z_k\), a set of responses that are not related to \(k\). These relationships could be used to improve estimation using a penalized likelihood such as

\[ \arg \min_{B \in \mathbb{R}^{p \times r}} \frac{1}{2n} \sum_{k=1}^p \sum_{i=1}^n (y_{ik} - x_i^T \beta_k)^2 + \frac{\gamma}{2} \sum_{k=1}^r \sum_{l=1}^r (\beta_k - \beta_l)^2, \tag{3} \]

where \(\gamma\) is a tuning parameter. The penalty is a regularization term that encourages sparsity in the regression coefficients. This penalty has desirable properties such as consistency and asymptotic normality under mild conditions.

\textbf{ARTICLE HISTORY}

Received June 2022
Accepted April 2023

\textbf{KEYWORDS}

Graph constrained models; High-dimensional convergence; Nonconvex optimization; Post-Selection inference; Selection consistency
where, for a vector \( a \in \mathbb{R}^d \) define \( \| a \|_q = (\sum_{j=1}^{d} a_j^q)^{\frac{1}{q}} \) as the \( L_q \) norm and \( \gamma > 0 \) is a tuning parameter to promote similarities between groups of related coefficients. Kim and Xing (2009) proposed a similar penalized loss function, but used a lasso-fusion penalty instead of ridge fusion and did not consider \( Z_k \) sets. The major difference between their approach and the one in this article is they consider the graph sets \( P_k \) and \( N_k \) known either from a priori knowledge or found during a pre-processing step. They also used an \( L_1 \) penalty for feature selection.

The method we propose also uses an \( L_1 \) penalty for feature selection. The major difference from the approach of Kim and Xing (2009) is we generalize (3) to the setting where the sets \( P_k, N_k, \) and \( Z_k \) are not known a priori. The proposed method simultaneously estimates these unknown sets, while using the framework of (3) to include these detected relationships in the estimation. To simultaneously estimate the structure of the relationships of the responses we propose a novel penalty based on the minimum function. The penalized objective function is nonconvex but can be restated as solving a finite number of convex problems. However, this quickly becomes computationally intractable for even small values of \( r \). This approach can be easily generalizable to simultaneous modeling of binary response variables or other noncontinuous values by replacing the least squares loss function with the appropriate generalized linear model loss function.

In the least squares case we provide rates of convergence, model selection consistency and post selection inference results which hold for high-dimensional predictors, \( p \gg n \). The results hold for any grouping of the response variables and thus do not depend on the relationships between responses to be estimated correctly. To derive rates of convergence under standard conditions we generalize the results of Negahban et al. (2012), which do not directly apply because of the ridge fusion penalty used in this framework.

Our proposal builds on methods which simultaneously estimate clusters and univariate regression models (Witten, Shojaei, and Zhang 2014), multivariate regression models (Price and Sherwood 2018), and precision matrices (Price, Molstad, and Sherwood 2021). All of which propose iterative algorithms that alternate between estimating the clustering structure, using k-means, and the model of interest. These approaches only accommodate positive relationships and rely on k-means, which provides distinct sets of clusters and may be unreliable for high-dimensional problems. The proposed method allows for more complex structures in the response variables and provides a global minimum at each iteration.

Extensive work has been done on exploiting relationships in multivariate regression to improve efficiency. For a recent survey of the field see Price, Allenbrand, and Sherwood (2021) and citations within. Reduced rank regression is a related method that reduces the dimension of the problem by constraining the rank of the coefficient matrix \( B \) while minimizing the multivariate least squares objective function to find a set of latent variables that increase prediction accuracy (Anderson 1951; Velu and Reinsel 2013; Chen and Huang 2016). Other dimension reduction methods have been proposed which find a smaller subspace of \( Y \) that maintains the important information needed to estimate regression coefficients (Cook, Li, and Chiaromonte 2010; Cook and Zhang 2015; Sun et al. 2015). While dimension-reduction techniques can improve predictions, they can suffer from interpretability issues, while this proposal has interpretable coefficients. Various group lasso penalties have been used to promote structure across response or a combination of response and predictors (Kim and Xing 2012; Li, Nan, and Zhu 2015). Instead of the proposed approach of simultaneously estimating relationships and coefficients, many have used a two-stage clustering approach. In the first stage the clusters are estimated and in the second stage the structure is incorporated into the estimation of the coefficients. Examples include hierarchical clustering for tree guided lasso (Kim and Xing 2012), convex clustering in joint analysis (Chen, Liyengar, and Liyengar 2016), separating global and response specific features (Xu et al. 2015) and replacing high-dimensional responses or predictors with clusters (Bühlmann et al. 2013; Zhou et al. 2017).

Other methods in univariate regression, such as the GRACE method, use known relationships, or relationships that are estimated a priori, to assist in increasing accuracy and select relevant variables (Li and Li 2008, 2010). Zhao and Shojaei (2016) investigated statistical inference frameworks for univariate graph constrained models that can accommodate error in the estimation of the graph Laplacian. The minimum function has been used as a penalty to approximate 0–1 loss for likelihood based estimation (Shen, Pan, and Zhu 2012) and the simultaneous estimation of a regression model and an undirected graph of the predictors (Yang et al. 2012; Zhu, Shen, and Pan 2013).

The key difference between the proposed method and state of the art methods in the literature, is that our method simultaneously detects relationships between the responses, regardless of sign, and estimates regression coefficients without needing to estimate a covariance matrix of the errors. In Section 2 we introduce the MinPen framework with the least squares loss function along with theoretical results and the proposed algorithm. In Section 3 we extend the framework to multiple binomial responses with discussion of generalizing to other exponential family based loss functions. Sections 4 and 5 present simulations to investigate finite sample properties of the proposed methods. Finally we present examples of our methodology in applications in genomics and substance abuse overdoses in Section 6.

2. Least Squares Model

2.1. Method

We consider estimating \( B^* \) in (1) when there may be similarities in the coefficients of the different responses. Define,

\[
P(\beta_1; \beta_k) = \min (\| \beta_1 - \beta_k \|_2^2, \| \beta_1 + \beta_k \|_2^2, \| \beta_1 \|_2^2).
\]

(4)

If \( P(\beta_1^*, \beta_k^*) = \| \beta_1^* - \beta_k^* \|_2^2 \) this implies that \( \beta_1 \) and \( \beta_k \) are similar. If \( P(\beta_1^*, \beta_k^*) = \| \beta_1^* + \beta_k^* \|_2^2 \) this implies \( \beta_k^* \) and \( -\beta_k^* \) are similar, while \( P(\beta_1^*, \beta_k^*) = \| \beta_1^* \|_2^2 \) implies \( \beta_k^* \) is not sufficiently similar to \( \beta_k^* \) or \( -\beta_k^* \) to merit fusion toward either vector. Therefore, a penalty based on the minimum function can be used to simultaneously identify and leverage relationships between coefficients. Motivated by this, we propose the minimum penalty elastic net multivariate regression (MinPen) estimator as the solution to
where $\delta$ and $\gamma$ are nonnegative tuning parameters specified by the user. We refer to this as an elastic net estimator as it uses the combination of the lasso and ridge penalties as first introduced by Zou and Hastie (2005). Kim and Xing (2009) proposed a similar penalized loss function, but with a different fusion penalty. Instead of a ridge based penalty they penalized $||\beta_j - \beta_k||_1$ or $||\beta_j + \beta_k||_1$, but the practitioner knew a priori whether the plus or negative version was used for each pair of $(i,k)$, where we propose simultaneously estimating the coefficients and determining the best grouping penalty. The lasso penalty, associated with tuning parameter $\delta$, is used to simultaneously perform variable selection and estimate regression coefficients. The proposed ridge type penalty, associated with $\gamma$, is a non-convex penalty that simultaneously identifies and exploits relationships between coefficient vectors of different responses. Note that the use of the minimum penalty can generalize beyond the elastic net penalty described here. The proposed framework and proposed algorithm create the foundation for the use of the minimum as a penalty in multivariate regression. Figure 1 presents how the feasible set changes with $\gamma$ and different definitions of $P(\beta_j, \beta_k)$.

The penalty $P(\beta_j, \beta_k)$ is not symmetric due to the inclusion of $||\beta_j||_2^2$ in the penalty. If $||\beta_j||_2^2$ is replaced with some nonnegative constant, $a$, then the function is symmetric. However, this raises the problem of what value should $a$ be set to. If $a = 0$ then a relationship will be imposed upon all responses and shrinking all response coefficients toward, or away from, each other causes the model selection performance to suffer. Finding an optimal value of $a$ adds another tuning parameter to the problem and the optimal value of $a$ may not be the same for all responses. While using $||\beta_j||_2^2$ provides a data adaptive solution and provides a common form of shrinkage that helps model selection performance. Though, the penalty also may fail to identify cases where coefficient vectors are very similar but of varying magnitudes. Taking an insightful comment from one of the reviewers, consider the case where $\beta_k = b\beta_l$, where $b > 2$ then $P(\beta_l, \beta_k) = ||\beta_l||_2^2$. Replacing $||\beta_l||_2^2$ with $a$ would not solve this issue if $a \ll (b - 1)||\beta_l||_2^2$ and so this is not necessarily resolved by making the penalty function symmetric. To mitigate this issue we recommend scaling each response so it has a mean of zero and a standard deviation of one. In addition, it is important to note that the relationships being found are for coefficients that have similar values and magnitudes. Note that penalties besides the ridge penalty proposed in this manuscript can be used inside of the minimum function to detect various other types of relationships between responses. These functions will depend on the various applications and structures of interest by the researchers and practitioners.

Both our theory and algorithms will deal with a vectorized version of the solution to (5). For a matrix $A$, define vec$(A)$ as the vector formed by stacking the columns of $A$ on top of one another. Define $Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^{nxr}$, $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{nxp}$, $E = (e_1, \ldots, e_n)^T \in \mathbb{R}^{nxr}$, $\tilde{Y} = \text{vec}(Y) \in \mathbb{R}^{nr}$ and $\tilde{X} = \Sigma \otimes \mathbf{I}_n \in \mathbb{R}^{nr \times nr}$. Then the vectorized version of (1) is

$$\tilde{Y} = \tilde{A}\beta^* + \tilde{\epsilon}.$$  

Define the set

$$\mathcal{A} = \left\{ A \in \{-1,0,1\}^{(r-1)p \times r} \right\}$$

$$||A\beta||_2^2 = \sum_{l=1}^r \sum_{i=1, i \neq m}^r ||\beta_i - d_{lm}||_2^2$$

$$\text{where } d_{lm} \in \{-1,0,1\}.$$  

The estimator in (5) is equivalent to

$$(\hat{\beta}, \hat{A}) = \arg\min_{\beta \in \mathbb{R}^r} \min_{A \in \mathcal{A}} \left( \frac{1}{2n} ||Y - \tilde{X}\beta||_2^2 + \delta||\beta||_1 + \frac{\gamma}{2}||A\beta||_2^2 \right).$$

Due to the minimum function the objective function is non-convex, which typically makes theoretical study of the global minimizer challenging. However, the global solution to (7) is the pair $(\hat{\beta}, \hat{A})$ that minimizes across the $|\mathcal{A}|$ potential convex objective functions. So while the objective function is nonconvex, the solution $\hat{\beta}$ minimizes a convex function. If the matrix $A$ is known a priori the estimator of $\beta^*$ is

$$\hat{\beta}(A) = \arg\min_{\beta \in \mathbb{R}^r} \frac{1}{2n} \sum_{i=1}^r \sum_{k=1}^r (y_{i,k} - x_{i,k}^T \beta_k)^2$$

$$+ \delta \sum_{k=1}^r ||\beta_k||_1 + \frac{\gamma}{2} A^T A \beta_k.$$  

2.2. Theoretical Results

Mild or standard conditions were used to derive the rates of convergence. Before presenting the conditions, we define some notation that is used in our theorems and conditions. The subspace for the active predictors is defined as $\mathcal{M} = \{a \in \mathbb{R}^p | a_j = 0 \text{ if } \beta_j^* = 0\}$, with $s$ being the number of nonzero elements of $\beta^*$. The parameter space is separated using projections of vectors into orthogonal complements. Define $\mathcal{M}_0$ as a projection of a vector $u$ into space $\mathcal{M}_0$, $\mathcal{M}_0^\perp$ as the orthogonal complement of $\mathcal{M}_0$ and $\mathcal{C} = \{a \in \mathbb{R}^p | ||a_{\mathcal{M}_0^\perp}||_1 \leq 3 ||a_{\mathcal{M}_0}||_1\}$.

Rates of convergence were derived using the following conditions.

A1 Define $x_j \in \mathbb{R}^p$ to be the jth column vector of $X$. For all $j \in \{1, \ldots, p\}$ $||x_j||_2^2/n \leq 1$.

A2 There are positive constants $\kappa_1$ and $\kappa_2$ where $\kappa_1||\theta||_2^2 \leq \theta^T X^T \tilde{X} \theta \leq \kappa_2||\theta||_2^2$, for any $\theta \in \mathcal{C}$.

A3 Assume $e_1 \sim N_r(0, \Sigma)$, are iid and $\Sigma = \sigma^2 D_0$, where $D_0 \in \mathbb{R}^{r \times r}$ is a diagonal matrix where diagonal values are on the support $(0, 1]$.

A4 There is at least one nonzero entry in $\beta^*$.
Conditions A1 and A2 are commonly made for penalized estimators. See Negahban et al. (2012) and the citations within. The framework in Negahban et al. (2012) does not directly apply to (7), because they assume the penalty is a norm which is not the case for our penalty, even for deterministic $A$, because of the inclusion of the ridge penalty. However, under these conditions we can extend the approach of Negahban et al. (2012) and prove rates of convergence under milder conditions than some similar methods (Li and Li 2010; Price and Sherwood 2018). For simplicity of presentation we assume the errors follow a multivariate normal distribution because later results on model selection and post selection inference use this assumption, but consistency could be derived under the more general assumption of sub-Gaussian tails. Note under A3 the maximum variance of the error is $\sigma^2$. In our results the upper bound of the tuning parameter $\gamma$ depends on the inverse of $||\beta^*||_{\infty}$, which is well defined under A4.

**Theorem 1.** Assume Conditions A1–A4 hold and $\gamma < \delta/(16r||\beta^*||_{\infty})$ then for any $A \in A$

$$P \left[ ||\hat{\beta}(A) - \beta^*||_2^2 \leq \frac{9\delta^2}{4k^2} \right] \geq 1 - 2\exp \left[ -\frac{n\delta^2}{32\sigma^2} + \log(rp) \right].$$

Proof of Theorem 1, and all other theoretical results, are provided in the supplemental material. Theorem 1 can be used to derive convergence rates for $\hat{\beta}(A)$ or $\hat{\beta}$, conditional on $\hat{A} = A$, for a set sequence of $\delta$. 

---

**Figure 1.** Feasible plots for MinPen with $r = 2$ and $p = 1$. Feasible set of (5) with $\delta = 1$, but varying $\gamma$ and $P(\beta_1, \beta_2)$. In first row min penalty is set to $P(\beta_1, \beta_2) = \min((\beta_1 - \beta_2)^2, (\beta_1 + \beta_2)^2)$. Second row $P(\beta_1, \beta_2)$ is as defined in (4). In third row $P(\beta_1, \beta_2) = \min(|\beta_1 - \beta_2|, |\beta_1 + \beta_2|, |\beta_1|)$. 
Corollary 1. Assume Conditions A1–A4 hold, \( \gamma < \delta / (16r) \) \( \| \beta^* \|_\infty \) and \( \delta = \sqrt{\frac{8kd^2 \log(rp)}{n}} \) then
\[
P \left[ \| \hat{\beta}(A) - \beta^* \|_2^2 \leq 144 \frac{d^2 \log(rp)}{nk_2^2} \frac{s}{1} \right] \geq 1 - 2 \exp \left[ - \log(rp) \right].
\]

Corollary 1 provides that for a fixed sequence of \( A \) and \( rp \to \infty \) that the estimator \( \hat{\beta}(A) \) achieves the same rate of converge as the lasso estimator (Negahban et al. 2012) and elastic net (Hebiri and van de Geer 2011).

### 2.3. Post-Selection Inference

This section details how the approach of Lee et al. (2016) for post-selection confidence intervals for nonzero coefficients of lasso estimators can be extended to our elastic-net type penalty with multivariate responses. The supplemental material includes a model selection consistency theorem and proof. Interpretation of models depends on the variables in the model and we restrict our analysis to inference conditional on the selected model. A consequence of this is \( \beta^* \) will change depending on the model.

The goal of this work is to define an interval for \( \beta^* \). We consider an arbitrary vector of selection intervals being conditional on the selected model and restrict our analysis to inference conditional on the selected model. A set of models depends on the variables in the model and we restrict our analysis to inference conditional on the selected model. A consequence of this is \( \beta^* \) will change depending on the model.

Let \( Q \subseteq \{1, \ldots, pr\} \) be a subset of selected predictor-response combinations. For simplicity of notation we assume that each response has at least one predictor selected. Define \( \tilde{X}_Q \equiv \tilde{X}(:,Q) \) to be the matrix made from the columns of \( \tilde{X} \) that belong to \( Q \).

Define the \( \hat{\beta}^*_Q = \arg \min\{ \| \tilde{y} - \tilde{X}_Q \beta \|_2^2 | \beta \in R^{m |Q|} \} \),

\[
\hat{\beta}^*_Q = \left( \hat{X}_Q^T \hat{X}_Q \right)^{-1} \hat{X}_Q^T \tilde{y}.
\]

The index of nonzero coefficients of \( \hat{\beta} \) is a subset of \( \hat{Q} \) and typically these sets are the same.

Let \( \hat{Q} \) represent the selected model and let \( \hat{A} \equiv \hat{X}(:,Q) \) be the graph selected from (7). Consider a specific index \( j \in Q \), the goal of this work is to define an interval \( C_{\hat{A},Q}^j \) for a given level \( \alpha \), such that

\[
P(\hat{\beta}^*_{Q,j} \in \bar{C}_{\hat{A},Q}^j | \hat{Q} = Q, \hat{A} = A) \geq 1 - \alpha.
\]

See Lee et al. (2016) for a justification of coverage of post-selection intervals being conditional on the selected model and a review of other approaches to post-selection inference. To derive this interval we consider an arbitrary vector of \( \eta_Q \in R^{|Q|} \) examine the conditional distribution of

\[
\eta_Q^T \tilde{y}(\hat{Q} = Q, \hat{A} = A).
\]

Note that this definition does include all covariates that have nonzero coefficients, but could potentially include a nonzero coefficient as \( \hat{s}_j \in [-1, 1] \). For \( \beta \) and \( \hat{s} \) to solve (8), conditional on \( \hat{A} = A \), the following sufficient and necessary KKT conditions must be satisfied,

\[
\left( \frac{1}{n} \tilde{X}^T \tilde{X} + \gamma A^T \hat{A} \right) \hat{\beta}(A) - \frac{1}{n} \tilde{X}^T \tilde{y} + \delta_s = 0, \quad (13)
\]

and

\[
\hat{s}_j = \text{sign}(\hat{\beta}_j) \text{ if } \hat{\beta}_j \neq 0, \quad (14)
\]

\[
\hat{s}_j \in [-1, 1] \text{ if } \hat{\beta}_j = 0. \quad (15)
\]

For a given \( A \), \( \hat{\beta}(A) \) is a unique minimizer of (8) and confidence intervals are constructed using these KKT conditions.

For the following results we need the OLS solution to be well defined and thus need \( \hat{X}_Q^T \hat{X}_Q \) to be positive definite.

A5 The matrix \( \hat{X}_Q^T \hat{X}_Q \) is positive definite.

Define \( e_j \) as the vector with zeros everywhere, but with a one in the jth position. The following theorem presents how to construct confidence intervals conditional on \( \hat{Q} \) and \( \hat{A} \).

Theorem 2. Let \( F_{\mu, \sigma}^{[a,b]}(x) \) be the CDF of a random normal variable with mean \( \mu \), variance \( \sigma^2 \), truncated to the interval \([a, b]\). Let \( \eta \equiv \tilde{X}^T \tilde{X}^{-1} \hat{e}_j \) and define \( L \) and \( U \) such that

\[
F_{\mu L, \sigma L}^{[\eta L, \eta U]}(\eta_T \tilde{y}) = 1 - \frac{\alpha}{2} \quad \text{and} \quad F_{\mu U, \sigma U}^{[\eta L, \eta U]}(\eta_T \tilde{y}) = \frac{\alpha}{2}.
\]

Where \( \eta_L \) and \( \eta_U \) are constants that depend on \( \eta_Q \in R^{1|Q|} \), \( A, Q, \hat{\Sigma}, \tilde{X}_Q, \gamma, \delta \) and \( \tilde{y} \). If Conditions A3 and A5 hold, then

\[
P(\hat{\beta}^*_{Q,j} \in [L, U] | \hat{Q} = Q, \hat{\Sigma} = \Sigma, \hat{A} = A) = 1 - \alpha.
\]

Theorem 2 provides a way to construct confidence intervals for a least squares model that accounts for the fact that variable selection was first done by satisfying the KKT conditions of the estimator \( \hat{\beta}(A) \). The result uses the framework proposed in Lee et al. (2016), which provides a more detailed discussion of the issue of post-selection inference and the intuition behind the interval construction. Theorem 2 assumes that \( \Sigma \), where \( \Sigma = \Sigma \otimes I_m \), is known. Usually in practice it will need to be estimated. If \( p < n \), then the covariance matrix of the residuals from a saturated multivariate regression model could be used. In order to unify our assumptions across results, we assume a block diagonal structure, but Theorem 2 holds for any positive definite \( \Sigma \).

### 2.4. Algorithm

To obtain the estimator defined in (5) we propose using a formulation that is similar to the optimization defined in (3). Define the following sets

\[
P_i = \left\{ m: \| \beta_i - \beta_m \|_2^2 \leq \| \beta_i + \beta_m \|_2^2, \| \beta_i - \beta_m \|_2^2, m \in \{1, \ldots, r\} \setminus \{i\} \right\},
\]

\[
N_i = \left\{ m: \| \beta_i + \beta_m \|_2^2 < \| \beta_i - \beta_m \|_2^2, \| \beta_i + \beta_m \|_2^2, m \in \{1, \ldots, r\} \setminus \{i\} \right\},
\]

\[
Z_i = \left\{ m: \| \beta_i \|_2^2 < \| \beta_i - \beta_m \|_2^2, \| \beta_i \|_2^2 \leq \| \beta_i - \beta_m \|_2^2, m \in \{1, \ldots, r\} \setminus \{i\} \right\},
\]

\[
Q_i = \left\{ m: \| \beta_i - \beta_m \|_2^2 = \| \beta_i + \beta_m \|_2^2, \| \beta_i - \beta_m \|_2^2, m \in \{1, \ldots, r\} \setminus \{i\} \right\}.
\]

\[\text{if } \hat{s}_j \neq 0 \text{ and } \hat{\beta}_j \neq 0, \text{ then } \hat{s}_j = \text{sign}(\hat{\beta}_j) \text{ if } \hat{\beta}_j \neq 0.
\]

\[\text{and } \hat{s}_j \in [-1, 1] \text{ if } \hat{\beta}_j = 0.
\]
such that $P_l \cup N_l \cup Z_l = \{1, \ldots, r\} \setminus \{l\}$, $P_l \cap N_l = \emptyset$, $P_l \cap Z_l = \emptyset$, and $Z_l \cap N_l = \emptyset \ \forall l \in \{1, \ldots, r\}$. Solving (5) is equivalent to solving

$$
\arg \min_{B \in \mathbb{R}^{p \times r}, P_l, N_l, Z_l, \ldots, P_r, N_r, Z_r} \frac{1}{2n} \sum_{k=1}^{r} \sum_{i=1}^{n} (y_{ik} - x_i^T \beta_k)^2 + \delta \sum_{k=1}^{r} \| \beta_k \|_1
$$

$$+ \frac{\gamma}{2} \sum_{k=1}^{r} \left( \sum_{l \in P_k} \| \beta_l - \beta_k \|_2^2 + \sum_{m \in N_k} \| \beta_m + \beta_k \|_2^2 + \sum_{s \in Z_k} \| \beta_k \|_2^2 \right).$$

(16)

By using this formulation we are able to propose an iterative algorithm that iterates between estimating the sets $P_l, N_l$, and $Z_l$ for each $l \in \{1, \ldots, r\}$, with the regression coefficients fixed, and estimating the regression coefficients with the sets fixed. To initialize the algorithm estimates of coefficients, $B$, or sets, $P_l, N_l$, and $Z_l$ for each $l \in \{1, \ldots, r\}$, are needed. We define initial values for the coefficients of the $t$th response as

$$
\hat{\beta}_t^{(0)} = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y_{it} - x_i^T \beta)^2 + \delta \| \beta \|_1 + \frac{\gamma}{2} \| \beta \|_2^2.
$$

(17)

and define $\hat{B}^{(0)} = \left( \hat{\beta}_1^{(0)}, \ldots, \hat{\beta}_r^{(0)} \right)$ to be the $t$th iterative estimate of $B^*$. Note we will also use $\hat{B}^{(w)}$ to represent the vectorized version of $\hat{B}^{(w)}$, consistent with previous notation. Given a fixed $(\delta, \gamma)$ we propose the following algorithm as a solution to (16).

1. Initialize $\hat{\beta}_1^{(0)}, \ldots, \hat{\beta}_r^{(0)}$ as defined in (17).

2. For the $w$th iteration, where $w > 0$, fix $\hat{B}^{(w-1)}$ and obtain estimates for $\hat{\beta}_k^{(w)}, \hat{N}^{(w)}, \hat{Z}^{(w)}$ for all $k \in \{1, \ldots, r\}$ by solving the optimization

$$
\text{minimize} \quad \sum_{k=1}^{r} \left( \sum_{l \in P_k} \| \hat{\beta}_l^{(w-1)} - \hat{\beta}_k^{(w-1)} \|_2^2 
+ \sum_{m \in N_k} \| \hat{\beta}_m^{(w-1)} + \hat{\beta}_k^{(w-1)} \|_2^2 + \sum_{s \in Z_k} \| \hat{\beta}_k^{(w-1)} \|_2^2 \right).
$$

(18)

3. Holding $\hat{\beta}_k^{(w)}, \hat{N}^{(w)}, \hat{Z}^{(w)}$ for all $k \in \{1, \ldots, r\}$ fixed obtain the estimate $\hat{B}^{(w)}$ by solving the optimization

$$
\arg \min_{B \in \mathbb{R}^{p \times r}} \frac{1}{2n} \sum_{k=1}^{r} \sum_{i=1}^{n} (y_{ik} - x_i^T \beta_k)^2 + \delta \sum_{k=1}^{r} \| \beta_k \|_1
$$

$$+ \frac{\gamma}{2} \sum_{k=1}^{r} \left( \sum_{l \in \hat{P}_k^{(w)}} \| \beta_l - \beta_k \|_2^2 + \sum_{m \in \hat{N}_k^{(w)}} \| \beta_m \|_2^2 + \beta_k \|_2^2 + \sum_{s \in \hat{Z}_k^{(w)}} \| \beta_k \|_2^2 \right).
$$

(19)

4. Repeat the above two steps until the estimated sets $P_l, N_l$, and $Z_l$ do not change from iterate $w-1$ to iterate $w$ for all $l \in \{1, \ldots, r\}$.

The update steps, shown in (18) and (19), respectively, break the nonconvex problem into two problems that can be solved directly with well-studied solutions. The optimization in (18) can be solved directly in a single pass, where the set assignments to the $w+1$ iteration for $l \in \{1, \ldots, r\}$ are defined as

$$
\hat{P}_l^{(w+1)} = \left\{ m : \| \hat{\beta}_l^{(w)} - \hat{\beta}_m^{(w)} \|_2^2 \leq \| \hat{\beta}_l^{(w)} + \hat{\beta}_m^{(w)} \|_2^2, \right\}
$$

$$
\| \hat{\beta}_l^{(w)} - \hat{\beta}_m^{(w)} \|_2^2 < \| \hat{\beta}_l^{(w)} \|_2^2, m \in \{1, \ldots, r\} \setminus \{l\} \right\},
$$

$$
\hat{N}_l^{(w+1)} = \left\{ m : \| \hat{\beta}_l^{(w)} + \hat{\beta}_m^{(w)} \|_2^2 < \| \hat{\beta}_l^{(w)} - \hat{\beta}_m^{(w)} \|_2^2, \right\}
$$

$$
\| \hat{\beta}_l^{(w)} + \hat{\beta}_m^{(w)} \|_2^2 < \| \hat{\beta}_l^{(w)} \|_2^2, m \in \{1, \ldots, r\} \setminus \{l\} \right\},
$$

$$
\hat{Z}_l^{(w+1)} = \left\{ m : \| \hat{\beta}_l^{(w)} \|_2^2 \leq \| \hat{\beta}_l^{(w)} - \hat{\beta}_m^{(w)} \|_2^2, \right\}
$$

$$
\| \hat{\beta}_l^{(w)} \|_2^2 \leq \| \hat{\beta}_l^{(w)} + \hat{\beta}_m^{(w)} \|_2^2, m \in \{1, \ldots, r\} \setminus \{l\} \right\}.
$$

Note that this update requires calculating $r^2$ terms and performing $r(r-1)$ comparisons, but the sets are deterministic and no iterates are needed.

In this second step, $A$ is considered fixed and thus each estimator derived from (19) is consistent under the conditions of Corollary 1. The optimization in (19) can be solved using a gradient descent algorithm where each $\hat{\beta}_{mk}$ is solved iteratively with all other regression coefficients held fixed. Let $S(x, a) = \text{sign}(x)(|x| - a)_+$ be the soft-thresholding operator. To solve (19) we use a coordinate descent algorithm where the update for the $m$th predictor for the $k$th response, $\hat{\beta}_{mk}$, is

$$
\hat{\beta}_{mk} = \frac{S \left( n^{-1} \sum_{i=1}^{n} x_{im} y_{ik} - \sum_{j=1, j \neq m}^{p} (n^{-1} \sum_{i=1}^{n} x_{im} x_{ij}) \hat{\beta}_{jk} + \gamma H_{mk}, \delta/2 \right)}{n^{-1} \left( \sum_{i=1}^{n} x_{im}^2 + \gamma \left( (r-1) + \sum_{g=1}^{r} I \left[ k \in \hat{P}_g^{(w)} \right] \right) \right)}
$$

where

$$
H_{mk} = \sum_{l \in \hat{P}_k^{(w)}} \hat{\beta}_{ml} - \sum_{l \in \hat{N}_k^{(w)}} \hat{\beta}_{ml} + \sum_{g=1}^{r} \hat{\beta}_{mg} I \left[ k \in \hat{P}_g^{(w)} \right]
$$

$$- \sum_{y=1}^{r} \hat{\beta}_{my} I \left[ k \in \hat{N}_y^{(w)} \right],
$$

and $\hat{\beta}_{my}$ are the current iterates of the coordinate descent algorithm. The algorithm iterates through all $m \in \{1, \ldots, p\}$ and $k \in \{1, \ldots, r\}$ in a similar fashion to other coordinate descent algorithms and converges when the change in the estimates across iterations is small. We propose selecting $\delta$ and $\gamma$ using k-fold cross-validation minimizing the validation residual sum of squares. The proposed coordinate descent update is similar to GRACE, but differs due to structured use of the minimum penalty across multivariate responses (Li and Li 2008, 2010). We
recommend scaling the responses and predictors because the MinPen estimator is not equivariant.

While this algorithm is related to the iterative algorithms proposed by Witten, Shojaie, and Zhang (2014), Price and Sherwood (2018), and Price, Molstad, and Sherwood (2021), there are some differences. First, the penalties investigated by those authors are considered *cluster fusion penalties*. Second, the algorithms proposed in those papers are two stage procedure that require a solution to the well-studied k-means problem, which is not guaranteed to find optimal sets, and without regularization can be problematic in high-dimensional settings (Sun, Wang, and Fang 2012). The proposed algorithm provides an efficient algorithm that is guaranteed to find the optimal sets of the responses given the coefficients and can be easily adjusted to accommodate if other penalty functions, besides the proposed ridge fusion, are used within the minimum function.

An alternative approach would be an exhaustive search of all possible \( A \in \mathcal{A} \) which would provide a global minimizer. However, this becomes computational intractable for medium or large \( r \) as there are \( 3^r(r-1) \) combinations that must be searched before tuning parameter selection is considered. The iterative method we propose is not constrained in this way but is not guaranteed to achieve a global solution to the nonconvex optimization problem. However, the algorithm produces an estimator that is equivalent to \( \hat{\beta}(A) \), for some \( A \in \mathcal{A} \). Conditioning on the selected graph, Corollary 1 provides the rate of convergence for the algorithm estimator. The other theoretical results also hold as they are conditional on the selected graph. Furthermore we are able to provide certain computational guarantees that show the effectiveness of our approach. For simplicity define \( L_{\delta,\gamma}(A,B) \) as the objective function found in (7) evaluated at \( A \) and \( B \), respectively.

**Theorem 3.** For \( \delta, \gamma > 0 \) the proposed algorithm is guaranteed to converge to a local minimizer in a finite number of outer iterations.
Theorem 3 provides that the proposed algorithm will converge to a local minimizer. Due to the non-convex objective function this does not imply convergence to a global minimizer and the solution of the proposed algorithm will depend on the initial estimator. The previous theory demonstrated consistency, model selection consistency and how to do inference conditional on the selected graph. Suggesting that regardless of the initial estimator what initial estimator is used the derived estimator will have strong asymptotic theoretical guarantees. The initial estimator the solution of the proposed algorithm will depend on this does not imply convergence to a global minimizer. Due to the non-convex objective function this does not imply convergence to a local minimizer. The simulation results that demonstrate that the algorithm is fairly robust to the choice of the initial estimator and verify that the algorithm tends to converge in a small number of outer iterations for different values of $r$ and $p$. Multiple initial values of $A$ or $B$ could be used to seed the algorithm, similar to setting multiple starting points in the k-means algorithm, to achieve more precise results (Fränti and Sieranoja 2019).

3. Extension to Multiple Binomial Responses

This section details how to extend the proposed method to the setting where conditional on the predictors each response follows a binomial distribution. The method presented in this section can be generalized to multiple responses from the same exponential family with different mean functions. Define $\theta_k = (\alpha_k, \omega_k^T)$, where $\alpha_k$ is a response specific intercept, and $u_i = (1, x_i)$ be the $p + 1$ dimensional vector of covariates for the $i$th observation. Let $y_{ik}$ be a realization of the random variable $Y_{ik} \sim \text{Bin}(n_{ik}, \pi_{ik})$, where

$$\pi_{ik} = \frac{\exp(u_i^T \theta_k)}{1 + \exp(u_i^T \theta_k)}.$$  

We define the penalized likelihood for the minimum penalized model as

$$\sum_{k=1}^r \sum_{i=1}^n y_{ik} u_i^T \theta_k - n_{ik} \log \left[ 1 + \exp(u_i^T \theta_k) \right] + \delta \sum_{k=1}^r \|\omega_k\|_1$$

$$+ \frac{\gamma}{2} \sum_{l=1}^r \sum_{m=1, l \neq m} \min \left( \|\theta_l - \theta_m\|_2^2, \|\theta_l + \theta_m\|_2^2, \|\theta_l\|_2^2 \right).$$

The lasso penalty does not include the intercept, because we assume that an intercept is part of the model for each response. However, the intercept is part of the penalty in the minimum function because we wish to group responses based on the relationships between fitted values. To select $\delta$ and $\gamma$ we propose using k-fold cross-validation minimizing validation likelihood loss.

4. OLS Simulations

In this section we investigate the performance of the MinPen estimator defined by (5), the MinPen estimator with the correct matrix $A$ known a priori (TMMinPen), a correlation based version of MinPen (Cor), the multivariate cluster elastic net estimator (MCEN) (Price and Sherwood 2018), sparse multivariate regression with covariance estimation (MRCE) (Rothman, Levina, and Zhu 2010), sparse reduced rank regression (SRRR) (Chen and Huang 2012), the separate elastic net estimator (SEN) which fits the elastic net on each response using common tuning parameters minimize the predicted sum of squares error on a different set of 50 test observations. The selected tuning parameter is trained on $n = 100$ observations and then evaluated on a different set of 50 test observations. The selected tuning parameters minimize the predicted sum of squares error on the test observations. We establish the TMMinPen estimator as a baseline for the MinPen estimator when the relationships would be known by a practitioner prior to fitting the model.

The SEN and JEN estimators are fit using the glmnet package in R (Friedman, Hastie, and Tibshirani 2010). Tuning parameters are selected using a train-test procedure, where each model is trained on $n = 100$ observations and then evaluated on a different set of 50 test observations. The selected tuning parameters minimize the predicted sum of squares error on the test observations. We establish the TMMinPen estimator as a baseline for the MinPen estimator when the relationships would be known by a practitioner prior to fitting the model.

Let $\tilde{\Sigma}_X \in \mathbb{R}^{4 \times 4}$ with entries $\tilde{\sigma}_{ii} = 1$ and $\tilde{\sigma}_{ij} = 0.7$, for $i \neq j$. The covariates are generated by $x_i \sim N(0_p, \Sigma_x)$, where $\Sigma_x$ is a block diagonal matrix with $p/4$ blocks of $\tilde{\Sigma}_X$ on the diagonal and all other entries are set to 0. Additional simulations and results using this general framework are provided in the supplemental material.
Figure 4. Results of the simulation described in Section 5 for \( p = 300 \).

The responses are generated from (1) with \( \epsilon_i \sim N_{15}(0_{15}, \Sigma_{\epsilon,15}) \), where \( \Sigma_{\epsilon,15} \) has an AR(1) structure with \( \sigma_{ij} = .25(\cdot75)^{|i-j|} \). In all simulations we perform 100 replications where each replication consists of using training dataset with \( n = 100 \) and tuning parameters are selected using predicted squared error loss on 50 independently sampled observations. To evaluate the model we generate an independent validation set consisting of 1000 observations and calculate the average squared prediction error (SPE)

\[
\frac{1}{1000r} \sum_{k=1}^{r} \sum_{i=1}^{1000} (y_{ik} - \hat{y}_{ik})^2,
\]

where \( y_{ik} \) is the \( k \)th element of \( y_i \) and \( \hat{y}_{ik} \) is a prediction for the \( i \)th observation and \( k \)th response. We also use the mean square error (MSE) of the estimators of the regression coefficient matrix \( B \), which we defined as

\[
\frac{1}{15p} \sum_{k=1}^{15} \| \hat{\beta}_k - \beta_k^* \|_2^2.
\]

The number of true variables and false variables selected through true positive (TP) and false positive (FP) rates respectively over the 100 replications are reported in the supplemental material. Let \( a_{ij} \) be the entry of \( A \) at the \( i \)th row and \( j \)th column. Let \( A^TM \) be the graph used in TMinPen. To compare how well MinPen and Cor identify the true structure we report the percentage of disagreements with the estimated \( A \) and \( A^TM \) as

\[
\frac{1}{r^2(r-1)p^2} \sum_{i=1}^{r(r-1)p} \sum_{j=1}^{rp} I(\hat{a}_{ij} \neq a^TM_{ij}).
\]
To compare the method for a variety of signal-to-noise ratios and values of $p$, we investigate the settings $(\eta, p) \in [0.0625, 0.125, 0.25] \times \{40, 100, 300\}$ over 100 replications. Figures 2 reports MSE and SPE results, while Figure 2 presents the percentage of disagreement between $\hat{A}$ and $A^{TM}$. Both figures are for $p = 300$. The supplemental material includes similar figures for other values of $p$ and true positive and false positive results. Results presented in Figure 2 demonstrate that with respect to MSE and SPE, MinPen works about as well as TMinPen, so long as there is sufficient signal. While TMinPen with respect to MSE and SPE, MinPen works about as well as TMinPen. The supplemental material also presents results that tend to converge after a small number of outer iterations.

5. Binomial Simulations

Next we study the impact of the minimum penalty methodology in binomial logistic regression setting and compare it to SEN and MCEN which also have binomial logistic regression cases for the respective methods. We generate $\{y\}_{i=1}^n$ in the same way as proposed in Section 4. The responses are generated by

$$y_{ik} \sim \text{Bin}(1, \pi_{ik}^*)$$

where $\pi_{ik}^* = \frac{\exp(\beta_k^T x_{ik})}{1 + \exp(\beta_k^T x_{ik})}$.

We define,

$$B^* = \begin{pmatrix}
\Delta_{10}(\eta, \lambda) & 0_{10} & 0_{10} \\
0_{10} & \Delta_{10}(\eta, \lambda) & 0_{10} \\
0_{p-30} & 0_{p-30} & 0_{p-30}
\end{pmatrix},$$

where $\Delta_q(\eta, \lambda) = (-\eta_q - \lambda, -\eta_q + \lambda, -\eta_q - 2\lambda, -\eta_q + 3\lambda) \in \mathbb{R}^{q \times 5}$ and $\eta_q$ is a $q$-dimensional vector with each element set equal to $\eta$ and $\lambda$ is a constant. We investigate settings of $(p, \eta, \lambda) \in \{40, 100, 300\} \times \{0.5, 1.0\} \times \{0.02, 0.05, 0.10\}$. In this setting there are three distinct clusters of five responses. The value $\lambda$ controls how close the coefficients are to each other, for the case of $\lambda = 0$ they are identical except for sign differences. The true graph, $A^{TM}$, is defined so that there is no relationship defined between the three clusters and within each cluster the relationship between the responses is determined by the signs of $\eta$. For each of the 100 replicates tuning parameters are selected using a train-test procedure with 100 observations in the training set, and 100 observations in the test set. A validation of 1000 observations is used to evaluate methods using Kullback-Leibler (KL) divergence, that is

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

where $\hat{\pi}_{ik}$ is the resulting estimated probability. We also compare methods using MSE, TP, and FP.

Figure 4 presents the KL divergence and MSE results of the simulations in the case of $p = 300$, while the supplemental material includes results for $p = 40$ and $p = 100$ and TP and FP for all values of $p$. The results show that MinPen performs as well as TMinPen and outperforms the other competitors with respect to MSE and KL. Results in the supplemental material demonstrate that with regard to variable selection, MinPen is comparable, if not better, than the other methods with respect to TP and outperforms all approaches in terms of FP. In all situations MinPen and TMinPen perform similarly demonstrating that the MinPen approach works as well as if the optimal $A$ was known a priori. Figure 5 presents the proportion of disagreements between $\hat{A}$ and $A^{TM}$. The larger the signal the closer $\hat{A}$ is to $A^{TM}$. Results in the supplemental material demonstrate that estimation of the graph also improves if $p$ decreases. The supplemental results for $p \in \{40, 100\}$ are similar to the results provided here, all methods perform better for smaller values of $p$. Again, the MinPen approach performs similar to TMinPen and performs better than the alternatives, particularly with respect to MSE and KL-divergence.
6. Applied Examples

6.1. Genomics Data

In this section we compare the post model selection inference approach outlined in this article, after using the proposed least squares method, with the post model selection inference approach proposed by Lee et al. (2016), after using lasso for model selection. The models are fit to data analyzed by Votavova et al. (2011), who collected demographic, birth and gene expression data from 72 postpartum women and their newborns, but our analysis is limited to 64 of the women, 65 had complete data and one was dropped due to outliers in the demographic data. Four response variables are modeled: placenta weight, newborn weight and two measures of cotinine level, one from the mother’s peripheral blood and a second from the umbilical cord. The predictors are smoking status, mother’s age, mother’s BMI, parity, gestational age and 33 gene expression probes. The 33 probes were selected by taking the absolute value of the correlation for each response and the 24,526 probes measured in the study and using a union of the top ten for each response. All variables were centered and scaled to have mean zero and standard deviation one before fitting the MinPen or lasso models. Results are presented after transforming back to the original scale of the data.

The lasso models are fit separately for each response. For both methods tuning parameters are selected to minimize the mean squared prediction error from five folds cross-validation. For post selection inference, $\hat{\Sigma}$ is estimated using the variance of residuals from the full multivariate linear regression model for the diagonal entries and zero for the off-diagonal entries. The MinPen selected graph identifies that the coefficients of the two cotinine measurements should be similar, while the penalty enforces no similarity between the other potential relationships.

Figure 6 presents the post selection confidence intervals for the lasso and MinPen methods. The selected models are very similar with the biggest model difference being in the model of the weight for placenta, where MinPen selects four more variables including mother’s age which is the only demographic variable difference between the models. In addition, for this model the post selection results for MinPen find a negative relationship between smoking and placenta weight, but the lasso results are inconclusive. Wang et al. (2014) found smoking has a negative relationship with placenta weights suggesting that for this particular relationship the lasso approach suffers from a Type II error, while MinPen correctly identifies this relationship. However, when modeling cotinine level in the umbilical cord, the one setting where the variables selected are the same, the lasso confidence intervals are noticeably smaller. In terms of variable selection for the two cotinine level measurements, the lasso models have 3 disagreements, while MinPen has one disagreement. These responses are almost identical and ideally would have no disagreement. Indicating that MinPen can provide similar coefficient estimates and variable selection for similar responses. Using a Dell Latitude 740 BTX, the lasso models were fit in a combined time of 0.16 sec, while the MinPen took 14.17 sec. The lasso method is faster due to the simpler model being fit and the highly optimized code from years of research on efficient algorithms for lasso.

6.2. Connecticut Multiple Substance Use Data

We investigate the use of the minimum penalty to identify co-occurring drug use based on accidental drug related deaths in Connecticut from 2012 to 2018. The dataset consists of 5097 accidental overdoses in the state of Connecticut and corre-
Figure 7. Results of the validation set of 500 observations from 2018 with regard to four substances: Cocaine, Fentanyl Analogue, Methadone, and the ability to detect any opioid. The results show that MinPen performs as well if not better with regard to AUC in classification.

sponding toxicology, death certificate, and scene investigation information from the Chief Medical Examiner Office, publicly available at https://data.ct.gov. The predictors were centered and scaled to have mean zero and standard deviation 1, but since the responses are all binary values they were not scaled or centered. We identify 17 response variables which are 16 individual drugs that were investigated plus a response that identifies if any opiate is involved. The goal of this analysis is to identify how different drugs appear (or do not appear) in cases of overdose, with a hope of better understanding and being able to predict co-occurrence of drugs. By being able to predict co-occurrence of drugs, both clinical practitioners and law enforcement are better able to understand emerging trends in drug usage allowing them to quickly identify, intervene, and mitigate issues.

The covariates in this data are indicators of the city and state of the residence and death, along with using a bag of words approach to find common language used in the cause of death and description of injury provided by police and the medical examiner. As the goal is to better understand the relationships between drugs being used, information in the cause of death may not be relevant to all drugs found in the system. In total 111 covariates were used in this analysis. A validation set was created from 500 randomly selected observations from 2018. Tuning parameters were selected using a 90/10 train/test approach using total classification error rate. Once the tuning parameters were obtained the model was fit using the chosen tuning parameters on all data less the validation set. We fit MinPen, MCEN, and elastic net models using a common sparsity parameter for all 17 responses, and compare methods using ROC curves on the validation set. Figure 7 presents ROC curves for Cocaine, Fentanyl Analogue, Methadone, and the ability to detect any opioid. The supplemental material includes ROC curves for the remaining 13 responses. The results show the MinPen is competitive or out performs MCEN and the elastic net approach. Furthermore, MinPen provides relationships between responses. Only positive or no relationships were found between responses, that is no
negative relationships were found. **Figure 8** provides a graph of these relationships. Most notably heroin is not predictive or related to fentanyl, fentanyl analogue, benzodiazapine, or inform the presence of other opioids. The opioid variable was mostly related to prescription based drugs such as oxymorphone, hydrocodone, methadone, among others. Methadone was positively related to every other variable which is notable because it is used as a medical assisted treatment for opiate abuse. Further, the model finds benzodiazapine is not related to cocaine, which is of interest as benzodiazapine is used to treat cocaine toxicity. The results show MinPen provides better prediction accuracy than both MCEN and elastic net, providing evidence that use of the minimum penalty can be beneficial in the presence of complex and potentially unknown relationships.

7. Discussion

This article introduces a new penalty based on the minimum function to exploit similarity and differences between regression coefficients of responses from a multiple response setting while simultaneously estimating the relationships between the responses. A two stage iterative algorithm is proposed to solve the non-convex objective function. The structured matrix that describes the relationships, \( A \in \mathbb{R}^{p \times p} \), grows rapidly with \( p \) and \( r \). This large number of constraints introduces computational complexities. One solution is to use screening methods, such as sure independence screening proposed by Fan and Lv (2008), to reduce the complexity of the problem. A future extension would be to either develop computational methods that can handle these extremely large dimensional problems. An alternative approach would be to modify the objective function to restrict the potential comparisons considered, and thus reducing the dimension of the search space for \( A \).

The proposed algorithm suggests using cross-validation to select the tuning parameters \( \delta \) and \( \gamma \). Alternatively, one could use the theory as a guide and select \( \delta = \sqrt{\frac{6\sigma^2 \log(np)}{n}} \) and \( \gamma < \delta/(16\sigma^2||\beta^*||_\infty) \), but this would require estimating \( \sigma^2 \) and \( ||\beta^*||_\infty \). For \( r = 1 \) and \( p < n \) the estimated variance from a model with all \( p \) predictors provides a consistent estimator, while for \( p > n \) Lee et al. (2016) suggested using a conservative estimate of \( \sigma^2 \). For cases of \( r > 1 \) plugging in \( \hat{\sigma}^2 = \max_{\beta \in \mathbb{R}^p} \hat{\sigma}^2_k \) would be a natural estimator, but how valid this approach is unknown to us. Estimating \( ||\beta^*||_\infty \) is also challenging because usually penalized methods are used to determine \( ||\beta^*||_\infty \), so how to obtain an estimate of this without relying on a penalized method, or some other form of model selection, is an open research question. Finding optimal data adaptive parameters in this setting merits future exploration.

This article proposes a minimum penalty to simultaneously group responses and estimate coefficients. The proposed penalty could be modified. For instance, (4) could be replaced with

\[
P_a(\beta_1, \beta_k) = \min (||\beta_1 - \beta_k||_2^2, ||\beta_1 + \beta_k||_2^2, a).
\]

If we set \( a = 0 \), then, barring special consideration of an edge case of \( ||\beta_1 - \beta_k||_2^2 = ||\beta_1 + \beta_k||_2^2 \), the penalty would create a complete graph. For positive values of \( a \) the penalty would be symmetric and allow for disconnected graphs to be estimated. Alternatively the functions of \( || \cdot ||_2^2 \) could be replaced with \( || \cdot ||_1 \), or some other norm or penalty function. Determining how to optimally define \( a \) or the penalty function used within the minimum function is an interesting question worth further research. We expect the optimal choice of the functions or \( a \) will vary based on the field of study and specific goals of the practitioner.

**Supplementary Materials**

**Proofs and Additional Simulations:** An Appendix containing details for the theoretical results, additional simulations, and simulation results described in this article. (SuppMaterialSubmission.pdf)

**Code and Datasets** The R-Code and datasets along with documentation to reproduce simulations and examples described in this article. A readme file is supplied describing each file supplied. (submission_code_jcgs.zip)

**Funding**

The project described was supported by the National Institute Of General Medical Sciences, 2U54GM104942-02 and National Institute On Minority Health And Health Disparities of the National Institutes of Health under Award Number U01MD017419. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was also supported by the National Science Foundation Major Research Instrumentation Program Award 1726534.

**References**

Anderson, T. W. (1951), “Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions,” *The Annals of Mathematical Statistics*, 22, 327–351. [139]

Bühlmann, P., Rütimann, P., van de Greer, S., and Zhang, C.-H. (2013), “Correlated Variables in Regression: Clustering and Sparse Estimation,” *Journal of Statistical Planning and Inference*, 143, 1835–1858. [139]
