A Convex-Nonconvex Strategy for Grouped Variable Selection

Xiaoqian Liu

Department of Statistics

North California State University

Collaborators:
Aaron Molstad, University of Florida
Eric Chi, Rice University

November 10, 2021

Overview

1. Convex-Nonconvex Penalization
   - Motivation
   - Generalized Minimax Concave (GMC) penalty
2. Group GMC for Grouped Variable Selection
   - The group GMC estimator
   - Algorithms for the group GMC model
   - Error bound for the group GMC estimator
   - Simulations and a real data application
3. Discussion

Introduction

The task of recovering a sparse representation is often formulated as

$$\text{minimize } F(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\psi(\boldsymbol{\beta}), \tag{1}$$

- Statistics – sparse linear regression
  - $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector
  - $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the design matrix
  - $\boldsymbol{\beta}$ is the vector of coefficients
- Signal processing – signal recovery/denoising
  - $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of observations
  - $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a linear operator
  - $\boldsymbol{\beta}$ is the underlying signal vector
- $\psi : \mathbb{R}^p \mapsto \mathbb{R}$ is a penalty function promoting sparsity in $\boldsymbol{\beta}$.

Convex penalization

Commonly used convex penalties:

- $\psi(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$
  - Lasso (Tibshirani, 1996)
  - Basis Pursuit (Chen and Donoho, 1994)
- $\psi(\boldsymbol{\beta}) = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2$
  - Elastic Net (Zou and Hastie, 2005)

Characteristics of convex penalties:

+ no suboptimal local minimizers

− underestimate large magnitude components

Nonconvex penalization

Commonly used nonconvex penalties:

- the smoothly clipped absolute deviations (SCAD) penalty
  - (Fan and Li, 2001)
- the minimax concave penalty (MCP)
  - (Zhang et al., 2010)

Characteristics of nonconvex penalties:

+ more accurate estimation
− existence of suboptimal local minimizers

Introduction

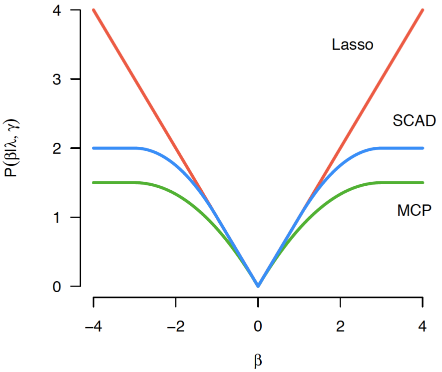


Figure: Visualization of Lasso, SCAD and MCP (Adopted from Patrick Breheny's lecture on BIOS 7240).

- Non-differentiability at the origin $\rightarrow$ sparsity

Introduction

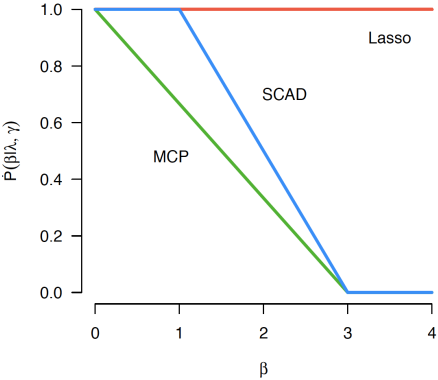


Figure: Visualization of derivatives of Lasso, SCAD and MCP (Adopted from Patrick Breheny's lecture on BIOS 7240)

- derivative $\rightarrow$ penalization rate (estimation bias)

The GMC penalization

A convex-nonconvex strategy:

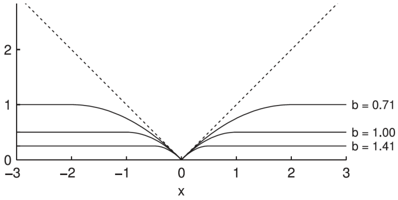
Design a nonconvex penalty but maintain the convexity of the problem.

The **GMC penalty** (Selesnick, 2017):

$$\psi_{\boldsymbol{B}}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 - \min_{\boldsymbol{v}\in\mathbb{R}^p}\{\|\boldsymbol{v}\|_1 + \frac{1}{2}\|\boldsymbol{B}(\boldsymbol{\beta}-\boldsymbol{v})\|_2^2\}, \qquad (2)$$

where $\boldsymbol{B} \in \mathbb{R}^{n\times p}$ is a matrix parameter for $\psi_{\boldsymbol{B}}$.

The GMC penalization



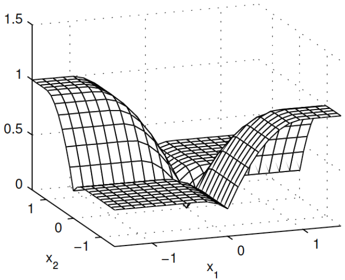


Figure: Visualization of the GMC penalty in the univariate case (left) and the multivariate case (right). Adopted from Selesnick (2017).

The GMC penalization

The optimization problem:

$$\text{minimize } F(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\psi_{\boldsymbol{B}}(\boldsymbol{\beta}), \tag{3}$$

maintains convex if

$$\boldsymbol{X}^\top \boldsymbol{X} \succeq \lambda \boldsymbol{B}^\top \boldsymbol{B}. \tag{4}$$

(4) is the **convexity-preserving condition** for the GMC model (3).

The GMC penalization

A key factor in GMC: the matrix parameter $\boldsymbol{B}$

Functions of $\boldsymbol{B}$:

- Preserves the convexity of the model
- Controls the degree of the convexity
- Affects the computation of the optimization problem
- Impacts the estimation/recovery performance

The GMC penalization

An open question for the GMC penalization:

how to set the matrix parameter $\boldsymbol{B}$?

An approach in (Selesnick, 2017):

$$\boldsymbol{B} = \sqrt{\theta/\lambda}\boldsymbol{X}, \quad \text{with } \theta \in (0, 1),$$

then $\lambda \boldsymbol{B}^\top \boldsymbol{B} = \theta \boldsymbol{X}^\top \boldsymbol{X}$, which satisfies condition (4).

Grouped variable selection

Consider the classical linear regression setting:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

- $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector
- $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the design matrix whose columns are p **covariate variables with natural group structures**
- $\boldsymbol{\epsilon}$ is a vector of noise variables with mean zero and variance σ^2

grouped variable selection and coefficient estimation

Grouped variable selection

Existing methods for grouped variables selection in linear regression:

- Convex penalization
  Group Lasso (Yuan and Lin, 2006) and its variants

$$\hat{\boldsymbol{\beta}}_{\text{grLasso}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{2n}\|\boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{X}_j\boldsymbol{\beta}_j\|_2^2 + \lambda \sum_{j=1}^{J} K_j\|\boldsymbol{\beta}_j\|_2 \tag{5}$$

  - $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_J^T)^T \in \mathbb{R}^p$ with $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}$ and $\sum_{j=1}^{J} p_j = p$
  - $\boldsymbol{X}_j$ is the submatrix of $\boldsymbol{X}$ whose columns correspond to the variables in the j-th group
  - K_js are used to adjust for the group sizes, e.g. $K_j = \sqrt{p_j}$

- Nonconvex penalization
  Group SCAD (Wang et al., 2007), Group MCP (Huang et al., 2012)

The group GMC estimator

We define the **group GMC penalty** as

$$\phi_{\boldsymbol{B}}(\boldsymbol{\beta}) = \sum_{j=1}^{J} K_j \|\boldsymbol{\beta}_j\|_2 - \min_{\boldsymbol{v} \in \mathbb{R}^p} \left\{ \sum_{j=1}^{J} K_j \|\boldsymbol{v}_j\|_2 + \frac{1}{2n} \|\boldsymbol{B}(\boldsymbol{\beta} - \boldsymbol{v})\|_2^2 \right\} \quad (6)$$

- $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_J^T)^T \in \mathbb{R}^p$
- $\boldsymbol{v} = (\boldsymbol{v}_1^T, ..., \boldsymbol{v}_J^T)^T \in \mathbb{R}^p$
- For each j, $\boldsymbol{\beta}_j, \boldsymbol{v}_j \in \mathbb{R}^{p_j}$ with $\sum_{j=1}^{J} p_j = p$

The group GMC estimator

The group GMC model:

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\beta\|_2^2+\lambda\phi_{\boldsymbol{B}}(\boldsymbol{\beta}), \tag{7}$$

- $\|\boldsymbol{y}-\boldsymbol{X}\beta\|_2^2 = \|\boldsymbol{y}-\sum_{j=1}^{J}\boldsymbol{X}_j\boldsymbol{\beta}_j\|_2^2$
- $\lambda \geq 0$ is the tuning parameter, which represents the degree of penalization
- $\boldsymbol{B}$ is a matrix parameter, which controls the concavity of the group GMC penalty

The group GMC estimator

The group GMC problem (7) is a convex optimization problem if

$$\boldsymbol{X}^\top \boldsymbol{X} \succeq \lambda \boldsymbol{B}^\top \boldsymbol{B} \tag{8}$$

- **convexity-preserving condition** for group GMC

The group GMC estimator

Set matrix $\boldsymbol{B}$ for the group GMC:

$$\lambda \boldsymbol{B}^\top \boldsymbol{B} = \theta \boldsymbol{X}^\top \boldsymbol{X}, \quad \theta \in [0, 1]. \tag{9}$$

- θ: the **convexity-preserving parameter** of the group GMC model
  - $\theta = 0$: group GMC $\rightarrow$ group Lasso
  - $\theta = 1$: a maximally nonconvex penalty

The group GMC estimator

Relationship between the group GMC and the group MCP (Huang et al., 2012):

Remark

The group GMC method is equivalent to the group MCP method when $\mathbf{B}^\mathsf{T}\mathbf{B}$ *is diagonal and the diagonal elements are suitably designed. This equivalence also holds for the GMC and MCP.*

The group GMC estimator

Properties of the solution path:

Theorem

Suppose $\boldsymbol{X}^\top \boldsymbol{X} \succ \lambda \boldsymbol{B}^\top \boldsymbol{B}$*, then the solution path* $\beta^\star(\lambda)$ *to the group GMC problem* (7) *exists, is unique, and is continuous in* λ*.*

- Problem (7) is well-posed
- Warm start when solving a sequence of problems over a grid of λ values

The group GMC estimator

Properties of the solution path:

Theorem

The group GMC problem (7) *has a unique solution* $\beta^{\star}(\lambda) = \mathbf{0}$ *for all* λ *greater than* $\lambda_0 = \max_j \left\{ \frac{\|\boldsymbol{X}_j^{\top} \boldsymbol{y}\|_2}{nK_j} \right\}$, *where* $\boldsymbol{X}_j$ *and* K_j *are as defined in* (5) *for* $j = 1, \cdots, J$.

- A precise range of λ, $[0, \lambda_0]$, to sample the full dynamic range of the coefficient estimation

Algorithms for the group GMC model

Recast the optimization problem (7) as a saddle-point problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \max_{\boldsymbol{v} \in \mathbb{R}^p} f(\beta) + \beta^\mathsf{T} \boldsymbol{Z} \boldsymbol{v} - g(\boldsymbol{v}), \tag{10}$$

where

$$f(\beta) = \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda \sum_{j=1}^{J} K_j \|\beta_j\|_2 - \frac{\lambda}{2n} \|\boldsymbol{B}\beta\|_2^2,$$

$$g(\boldsymbol{v}) = \frac{\lambda}{2n} \|\boldsymbol{B}\boldsymbol{v}\|_2^2 + \lambda \sum_{j=1}^{J} K_j \|\boldsymbol{v}_j\|_2,$$

$$\boldsymbol{Z} = \frac{\lambda}{n} \boldsymbol{B}^\mathsf{T} \boldsymbol{B}.$$

- Primal-Dual Hybrid Gradient (PDHG) method

Algorithms for the group GMC model

Algorithm 1 Basic PDHG steps for problem (10)

1: Set $\beta_0 \in \mathbb{R}^p, \boldsymbol{v}_0 \in \mathbb{R}^p, \sigma_k > 0, \tau_k > 0$
2: **for** $k = 1$ to K **do**
3: $\quad \hat{\boldsymbol{\beta}}_{k+1} = \boldsymbol{\beta}_k - \tau_k \boldsymbol{Z}^T \boldsymbol{v}_k$
4: $\quad \boldsymbol{\beta}_{k+1} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} f(\boldsymbol{\beta}) + \frac{1}{2\tau_k} \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{k+1}\|_2^2$
5: $\quad \hat{\boldsymbol{v}}_{k+1} = \boldsymbol{v}_k + \sigma_k \boldsymbol{Z}(2\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k)$
6: $\quad \boldsymbol{v}_{k+1} = \arg\min_{\boldsymbol{v} \in \mathbb{R}^p} g(\boldsymbol{v}) + \frac{1}{2\sigma_k} \|\boldsymbol{v} - \hat{\boldsymbol{v}}_{k+1}\|_2^2$
7: **end for**

Algorithms for the group GMC model

Updating $\boldsymbol{\beta}_{k+1}$ and $\boldsymbol{v}_{k+1}$ using FASTA:

$$\begin{aligned}\boldsymbol{\beta}_{k+1} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} f(\boldsymbol{\beta}) + \frac{1}{2\tau_k}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{k+1}\|_2^2 \\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 - \frac{\lambda}{2n}\|\boldsymbol{B}\boldsymbol{\beta}\|_2^2 + \frac{1}{2\tau_k}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{k+1}\|_2^2\right\} \\ &\quad + \lambda \sum_{j=1}^{J} K_j \|\boldsymbol{\beta}_j\|_2\end{aligned}$$

$$\begin{aligned}\boldsymbol{v}_{k+1} &= \underset{\boldsymbol{v}\in\mathbb{R}^p}{\operatorname{argmin}} g(\boldsymbol{v}) + \frac{1}{2\sigma_k}\|\boldsymbol{v} - \hat{\boldsymbol{v}}_{k+1}\|_2^2 \\ &= \underset{\boldsymbol{v}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{\frac{\lambda}{2n}\|\boldsymbol{B}\boldsymbol{v}\|_2^2 + \frac{1}{2\sigma_k}\|\boldsymbol{v} - \hat{\boldsymbol{v}}_{k+1}\|_2^2\right\} + \lambda \sum_{j=1}^{J} K_j \|\boldsymbol{v}_j\|_2\end{aligned}$$

Error bound for the group GMC estimator

Some definitions:

- $\mathbf{v}^\star = \underset{\mathbf{v}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{j=1}^{J} K_j \|\mathbf{v}_j\|_2 + \frac{1}{2n} \|\mathbf{B}(\beta^\star - \mathbf{v})\|_2^2 \right\}$
- $\mathcal{S} := \{j : \|\beta_j^\star\|_2 \neq 0, j \in [J]\}$ and $\mathcal{S}^c := [J] \setminus \mathcal{S}$
- $$\nu_j = \begin{cases} K_j + n^{-1}\|[\mathbf{B}^\top\mathbf{B}]_{j,\cdot}(\beta^\star - \mathbf{v}^\star)\|_2, & j \in \mathcal{S} \\ K_j - n^{-1}\|[\mathbf{B}^\top\mathbf{B}]_{j,\cdot}(\beta^\star - \mathbf{v}^\star)\|_2, & j \in \mathcal{S}^c \end{cases}$$
- $\bar{\nu} := \max_{j\in\mathcal{S}} \nu_j$ and $\underline{\nu} := \min_{k\in\mathcal{S}^c} \nu_k$

Error bound for the group GMC estimator

Conditions and assumptions:

- **X** satisfies a "block-normalization" condition:

$$\|\mathbf{X}_{\cdot,j}\| \leq \sqrt{n}, \quad j \in [J]$$

- **A1.** (Subgaussian errors). The data are generated from (13) where $\boldsymbol{\epsilon} \in \mathbb{R}^n$ has independent entries which are σ-subgaussian random variables for $0 < \sigma < \infty$. That is, $\mathbb{E}(\epsilon_i) = 0$ and for all $t \in \mathbb{R}$, $\mathbb{E}\{\exp(t\epsilon_i)\} \leq \exp(t^2\sigma^2/2)$ for each $i \in [n]$.
- **A2.** (Convexity) The matrix **B** is chosen so that $\mathbf{X}^\mathsf{T}\mathbf{X} \succeq \lambda \mathbf{B}^\mathsf{T}\mathbf{B}$.
- **A3.** (Sample size) The sample size n is sufficiently large so that $\nu_k > 0$ for all $k \in \mathcal{S}^c$.

Conditions and assumptions:

- **A4.** (Restricted eigenvalue condition) For a fixed $c > 1$, define

$$\mathbb{C}_n(\mathcal{S}, \nu, c) = \left\{ \mathbf{\Delta} \in \mathbb{R}^p : \mathbf{\Delta} \neq \mathbf{0}, \sum_{k \in \mathcal{S}^c} \left(\nu_k - \frac{\nu}{c} \right) \|\mathbf{\Delta}_k\|_2 \leq \sum_{j \in \mathcal{S}} \left(\nu_j + \frac{\nu}{c} \right) \|\mathbf{\Delta}_j\|_2 \right\}.$$

We assume there exists a constant $k > 0$ such that for all n and p,

$$0 < k \leq \kappa_{\mathbf{B}}(\mathcal{S}, c) := \inf_{\mathbf{\Delta} \in \mathbb{C}_n(\mathcal{S}, \nu, c)} \frac{\mathbf{\Delta}^{\mathsf{T}} (\mathbf{X}^{\mathsf{T}} \mathbf{X} - \lambda \mathbf{B}^{\mathsf{T}} \mathbf{B}) \mathbf{\Delta}}{2n \|\mathbf{\Delta}\|_2^2}.$$

Error bound for the group GMC estimator

Theorem

(Error bound for group GMC) Let $c > 1$ and $k_1 > 0$ be fixed constants. If assumptions **A1–A4** *hold and*

$$\lambda = \frac{2c\sigma}{\underline{\nu}} \left(\max_{j \in [J]} \sqrt{\frac{p_j}{n}} + \sqrt{\frac{k_1 \log(J)}{n}} \right),$$

then with probability at least $1 - 2\exp(-2k_1 \log(J))$,

$$\|\hat{\beta}(\lambda) - \beta^\star\|_2 \leq \frac{2c\sigma}{\kappa_{\mathbf{B}}(\mathcal{S}, c)} \left(\frac{\bar{\nu}}{\underline{\nu}} + \frac{1}{c} \right) \left\{ \left(\max_{j \in [J]} \sqrt{\frac{|\mathcal{S}| p_j}{n}} \right) + \sqrt{\frac{|\mathcal{S}| k_1 \log(J)}{n}} \right\},$$

where $\hat{\beta}(\lambda)$ is the group GMC estimator obtained from (7).

Error bound for the group GMC estimator

- Same asymptotic error rate as the group Lasso estimator
- Choose **B** such that $\kappa_{\mathbf{B}}(\mathcal{S}, c)$ is large and $\bar{\nu}/\underline{\nu}$ is small

Error bound for the group GMC estimator

Theorem

(Error bound for GMC) Let $c > 1$ *and* $k_2 \in (0,1)$ *be fixed constants. Let* $p_j = 1$ *for* $j \in [p]$ *so that* $\mathcal{S} = \{j : \beta^\star_j \neq 0, j \in [p]\}$*. If assumptions* **A1–A4** *hold and* $\lambda = (c\sigma/\underline{\nu})\sqrt{2\log(p/k_2)/n}$*, then with probability at least* $1 - 2k_2$

$$\|\hat{\beta}(\lambda) - \beta^\star\|_2 \leq \frac{c\sigma}{\kappa_{\mathbf{B}}(\mathcal{S}, c)} \left(\frac{\bar{\nu}}{\underline{\nu}} + \frac{1}{c}\right) \sqrt{\frac{2|\mathcal{S}|\log(p/k_2)}{n}},$$

where $\hat{\beta}(\lambda)$ *is the corresponding GMC estimator.*

Simulation Experiments

We explore some simulation experiments based on the simulations in Yuan and Lin (2006).

- Models:
  - an additive model including both categorical and continuous variables
  - an ANOVA model with all two-way interactions
- Factors of interest:
  - signal-to-noise ratio (SNR) of the model
  - correlation among groups
  - problem dimension
  - convexity-preserving parameter (for the group GMC)

Simulation experiments

Data generation of the additive model:

- Continuous covariates $X_1, \cdots, X_{20}$ are defined as $X_i = Z_i + cW$
  - Z_i and W are independently sampled from $N(0,1)$
  - c is a constant controlling the correlation between X_i and X_j
- $X_{11}, \cdots, X_{20}$ are trichotomized to $0, 1$ or 2
  - 0 if smaller than $\Phi^{-1}(\frac{1}{3})$
  - 1 if larger than $\Phi^{-1}(\frac{1}{3})$
  - 2 if in between

Data generation of the additive model:

The true regression model is

$$y = X_3^3 + X_3^2 + X_3 + \frac{1}{3}X_6^3 - X_6^2 + \frac{2}{3}X_6 + 2\mathbb{1}(X_{11} = 0) + \mathbb{1}(X_{11} = 1) + \epsilon$$

- $\mathbb{1}(\cdot)$ is the indicator function
- $\epsilon \sim N(0, \sigma^2)$
- 50 covariate variables from 20 groups

Simulation experiments

Performance in three aspects:

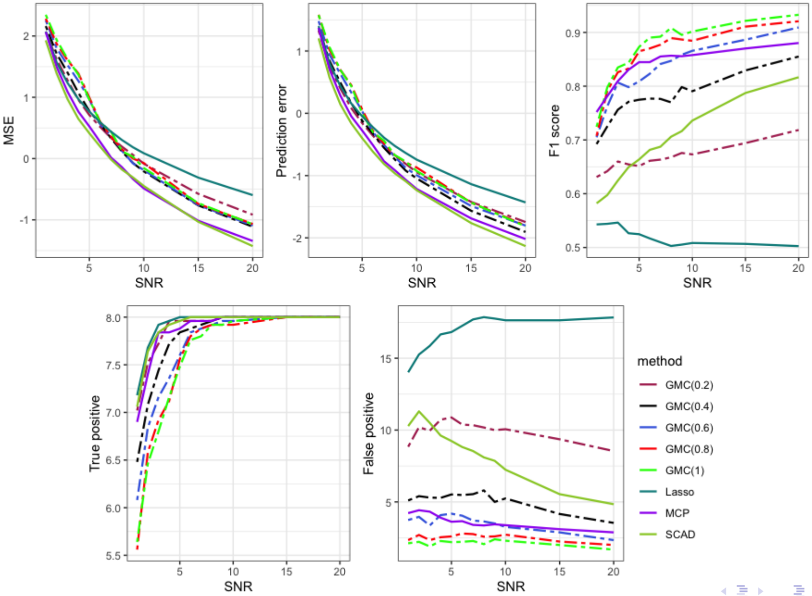
- Coefficient estimation
  - SE $= \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$
- Prediction performance
  - prediction error $= \frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$
- Support recovery
  - F1 score $= \dfrac{2TP}{2TP + FP + FN}$
  - true positive (TP) and false positive (FP)

		Estimation	
		$\hat{\beta}_j != 0$	$\hat{\beta}_j = 0$
Truth	$\beta_j != 0$	TP	FN
	$\beta_j = 0$	FP	TN

Simulation experiments

- **Case I: effect of the SNR**
  - uncorrelated groups ($c = 0$)
  - problem dimension $p = 50$
  - sample size $n = 100$
  - SNR $\in \{1, 2, \cdots, 9, 10, 15, 20\}$
  - $\theta \in \{0.2, 0.4, 0.6, 0.8, 1\}$

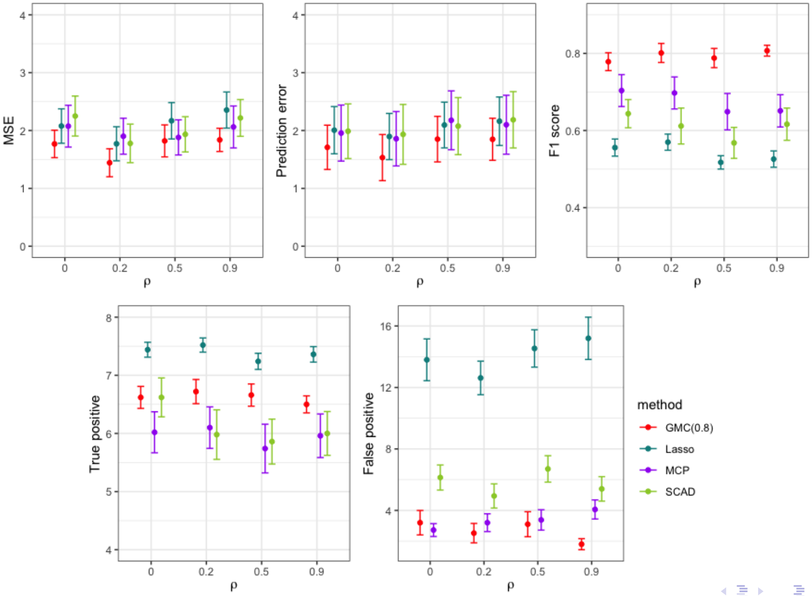
Simulation experiments



Simulation experiments

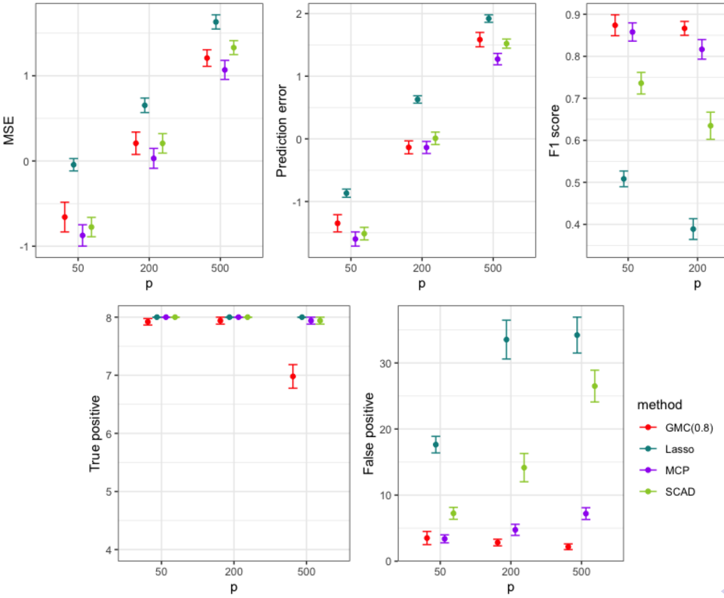
- **Case II: effect of the correlation among groups**
  - SNR $= 10$
  - problem dimension $p = 50$
  - sample size $n = 100$
  - $\theta = 0.8$
  - correlation $\rho = \frac{c^2}{1+c^2} \in \{0, 0.2, 0.5, 0.9\}$

Simulation experiments



- **Case III: effect of the problem dimension**
  - uncorrelated groups ($c = 0$)
  - SNR $= 10$
  - sample size $n = 100$
  - $\theta = 0.8$
  - $p \in \{50, 200, 500\}$

Simulation experiments



Real data application

The birth weight data set investigated in Yuan and Lin (2006):

- risk factors associated with low rank infant birth weight
- 189 observations of one response variable (infant birth weight)
- eight explanatory variables (continuous and categorical)

Real data application

Table 1. *Description of the birth weight data set*

Name	Type	Variable description
Birth weight	Continuous	Infant birth weight in kilograms
Mother's age	Continuous	Mother's age in years
Mother's weight	Continuous	Mother's weight in pounds at last menstrual period
Race	Categorical	Mother's race (white, black or other)
Smoking	Categorical	Smoking status during pregnancy (yes or no)
# Premature	Categorical	Previous premature labors (0, 1, or more)
Hypertension	Categorical	History of hypertension (yes or no)
Uterine irritability	Categorical	Presence of uterine irritability (yes or no)
# Phys. visits	Categorical	Number of physician visits during the first trimester (0, 1, 2, or more)

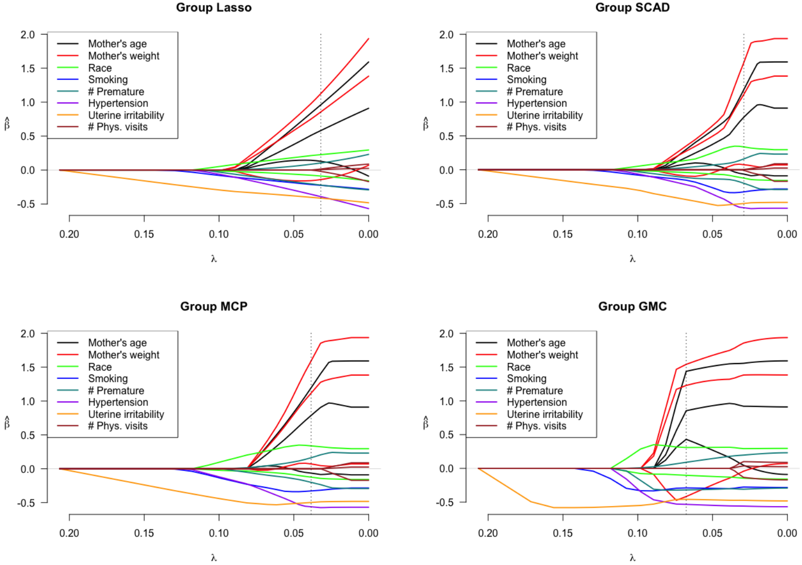
We have 16 covariate variables from 8 groups.

Real data application

Table 2. *Summarized results for the birth weight data*

	Prediction error	# nonzero groups	Excluded groups
Group Lasso	0.36	8	none
Group SCAD	0.35	8	none
Group MCP	0.35	7	# Phys. visits
Group GMC	0.35	7	# Phys. visits

Real data application



Discussion

Summary:

- A group GMC method for grouped variable selection and coefficient estimation in linear regression
- Convexity preserving condition, relation to existing methods, and properties of solution path
- Algorithms for computing the solution path
- Error bounds of the group GMC estimator, as well as the original GMC estimator
- Simulations and a real data application

Discussion

Future directions:

- Guidance on setting the matrix parameter $\boldsymbol{B}$
- Extension to generalized linear models
- Computation of the (group) GMC problem

Reference I

Chen, S. and Donoho, D. (1994). Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4).

Selesnick, I. (2017). Sparse regularization via convex analysis. *IEEE Transactions on Signal Processing*, 65(17):4481–4494.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Reference II

Wang, L., Chen, G., and Li, H. (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.