SCOP: A New Method for Model-Free Variable Selection

Angela Minster

Department of Statistics, Temple University

February 13, 2015

Outline

Notation

- Scalar: lower case letter (a)
- Vector: bold lower case letter ($\mathbf{a}$)
- Matrix by bold upper case letter ($\mathbf{A}$)
- Random variable: uppercase letter such as X or Y
- Random vector: lower case bold letter such as $\mathbf{x}$.
- Random vectors are assumed to be represented as column vectors, i.e. $\mathbf{x} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ where $\mathbf{x}$ is p-dimensional.
- $\mathbb{R}$ the set of all real numbers
- $\mathbb{R}^p$ the p-dimensional Euclidean space
- $\perp\!\!\!\perp$ indicates independence
- A set is indicated by italic capitals such as $\mathcal{A}$ or $\mathcal{F}$, with the exception of $\mathcal{S}$ which is reserved to indicate a linear subspace.
- The cardinality of a set $\mid \mathcal{F} \mid$

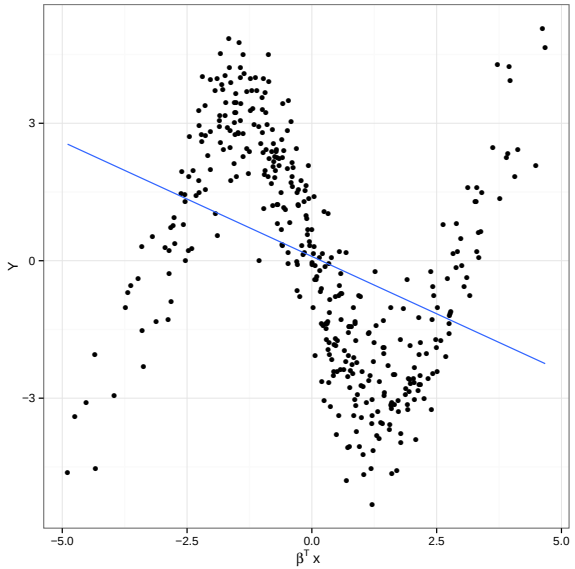
A note on "model-free"

In this presentation we mean:
$Y = f(X_1, X_2, \cdots, X_p) + \varepsilon$

Other models that could be considered model-free:
$Y = f(X_1, X_2, \cdots, X_p, \varepsilon)$
$Y = f(X_1) + f(X_2) + \cdots + f(X_p) + \varepsilon$

Motivation

Consider the model: $Y = 3\sin(1.05\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \pi) + \varepsilon,$ where $\varepsilon \sim \mathcal{N}(0, 1)$
$\boldsymbol{\beta} = (1, -1,\ 0,\ 0,\ 0,\ 0.8,\ 1.2,\ 0,\ 0,\ 0)$



$\boldsymbol{\beta}^* = (-0.3,\ 0.2,\ 0.0,\ 0.0,\ 0.2,\ -0.3,\ -0.4,\ 0.1,\ -0.1,\ 0.0)$

Outline

SDR: the definition

- $\mathbf{x} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$, univariate response $Y \in \mathbb{R}$
- Sufficient Dimension Reduction (SDR) maps $\mathbf{x}$ to $R(\mathbf{x}) \in \mathbb{R}^d$ where $d < p$ in such a way that $\mathrm{E}(Y|\mathbf{x})$ is a function only of $R(\mathbf{x})$.
- Typically $R(\mathbf{x})$ is linear combinations of $\mathbf{x}$, e.g. $R(\mathbf{x}) = \mathbf{A}^{\mathrm{T}}\mathbf{x}$ where $\mathbf{A}$ is a $p \times d$ matrix.
- Let $\boldsymbol{\alpha}$ be a basis for the column space of $\mathbf{A}$. The space spanned by $\boldsymbol{\alpha}$ is span$(\boldsymbol{\alpha})$ and is called a mean dimension reduction subspace (DRS).

- If $Y \perp\!\!\!\perp \mathrm{E}(Y \mid \mathbf{x}) \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}$ then $\mathrm{span}(\boldsymbol{\alpha})$ is a mean dimension reduction subspace for the regression of Y on $\mathbf{x}$.
- We can equivalently say:
  - $\mathrm{cov}\{(Y, E(Y|\mathbf{x})) \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}\}$
  - $\mathrm{E}(Y \mid \mathbf{x}) = \mathrm{E}(Y \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})$. (Since $\mathrm{E}(Y \mid \mathbf{x})$ is a function of $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}$, then given $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}$, $\mathrm{E}(Y \mid \mathbf{x})$ is a constant and hence the two statements are equivalent.)

- The intersection of all such mean dimension reduction subspaces is also a mean dimension reduction subspace and it is called the central mean space:

$$\bigcap \text{span}(\boldsymbol{\alpha}_i) = \mathcal{S}_{\text{E}(\mathbf{y}|\mathbf{x})}$$

Sufficient Dimension Reduction Methods

Many sufficient dimension reduction methods rely on the following conditions where $\mathbf{P}_{\Sigma} = \mathbf{\Sigma B}(\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma B})^{-1}\mathbf{B}^{\mathrm{T}}$ is the projection matrix onto the column space of $\mathbf{B}$ with the $\mathbf{\Sigma}$-inner product $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{\Sigma}} = \mathbf{a}^{\mathrm{T}}\mathbf{\Sigma b}$. $\mathbf{Q}_{\mathbf{\Sigma}} = \mathbf{I} - \mathbf{P}_{\mathbf{\Sigma}}$ where $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{x}$.

1. Linear Conditional Mean Assumption: $\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = \mathbf{P}_{\mathbf{\Sigma}}\mathbf{x}$
   - i.e. $\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$ is a linear function of $\mathbf{B}^{\mathrm{T}}\mathbf{x}$

Ordinary Least Squares (OLS) Regression

- OLS is actually a sufficient dimension reduction technique.
- $\beta_{\text{OLS}} = \boldsymbol{\Sigma}^{-1}\text{E}(\mathbf{x}Y) \subseteq \mathcal{S}_{\text{E}(\mathbf{y}|\mathbf{x})}$ under condition (1) (Li and Duan 1989)
- This holds only under models of the form:

$$Y = f(\boldsymbol{\theta}^{\text{T}}\mathbf{x}) + \varepsilon,$$

where f is some unknown link function, $\boldsymbol{\theta} \in \mathbb{R}^p$, and $\varepsilon \perp\!\!\!\perp \mathbf{x}$.

Outline

Setup

- Univariate response: $Y \in \mathbb{R}$ and p-dimensional predictor $\mathbf{x} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$
- Goal: select the predictors, chosen from $\Omega = \{1, 2, \cdots, p\}$, which are relevant to the regression mean $\mathrm{E}(Y \mid \mathbf{x})$ by estimating $\mathcal{A}$ such that:

$$\mathcal{A} = \{k \in \Omega \colon \mathrm{E}(Y \mid \mathbf{x}) \text{ functionally depends on } X_k\},$$

$$\mathcal{A}^c = \{k \in \Omega \colon \mathrm{E}(Y \mid \mathbf{x}) \text{ does not functionally depend on } X_k\}.$$

Connecition to SDR

- These definitions imply that $\mathrm{E}(Y \mid \mathbf{x}) \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}}$ and $\mathrm{E}(Y \mid \mathbf{x}) = \mathrm{E}(Y \mid \mathbf{x}_{\mathcal{A}})$.
- We can reframe the definitions from Sufficient Dimension Reduction to be:
  1. $Y \perp\!\!\!\perp E(Y|\mathbf{x})|\mathbf{x}_{\mathcal{A}}$;
  2. $\mathrm{cov}\{(Y, E(Y|\mathbf{x}))|\mathbf{x}_{\mathcal{A}}\} = 0$;
  3. $\mathrm{E}(Y|\mathbf{x})$ is a function of only $\mathbf{x}_{\mathcal{A}}$.

- Estimating $\mathcal{A}$ can be viewed as finding the smallest set $\mathcal{F}$ satisfying:

$$\mathrm{E}(Y \mid \mathbf{x}) = \mathrm{E}(Y \mid \mathbf{x}_{\mathcal{F}})$$

- Now these statements are also equivalent:
  1. $Y \perp\!\!\!\perp E(Y|\mathbf{x}_{\mathcal{F}}, X_j)|\mathbf{x}_{\mathcal{F}}$
  2. $\mathrm{cov}\{(Y, E(Y|\mathbf{x}_{\mathcal{F}}, X_j))|\mathbf{x}_{\mathcal{F}}\} = 0$
  3. $\mathrm{E}(Y|\mathbf{x}_{\mathcal{F}}, X_j)$ does not depend on X_j.
- In the model free setting we wish to estimate the set $\mathcal{A}$ without specific model assumptions about $\mathrm{E}(Y \mid \mathbf{x})$.

Some common methods for variable selection

- Classical stepwise regression
- Penalized regression
  - LASSO (Tibsirani, 1996), SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005) and many more.
  - Rely on a linear model and parameter sparsity.
- Methods using sufficient dimension reduction...

Variable Selection through SDR

SDR: light assumptions and model-free.

Variable Selection: potential for smaller covariance matricies.

Goal: Combine SDR and variable selection to get the "best of both worlds"

SDR vs Variable Selection

SDR is not explicitly a variable selection technique, but some techniques have been developed to use dimension reduction for variable selection:

- Hypothesis tests of e.d.r. direction vector entries (Cook, 2004 and Li et al., 2005)
- Regularized or sparse estimation using penalized procedures (Zhong et al., 2005; Li, 2007; Bondell and Li, 2009)
- SIR-based stepwise procedures (*most similar to our work*)
  - Correlation Pursuit (COP) by Zhong et al. (2012)
  - Sliced Inverse Regression with Interaction detection (SIRI) by Jiang and Liu (2014)

New Method

- Consider the following optimization problem:

$$\underset{\boldsymbol{\beta}}{\text{argmax}}\ \text{corr}(Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}),$$

- subject to $\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta} = 1.$

Proposition

Consider $\max \operatorname{corr}(Y, \boldsymbol{\beta}^T \mathbf{x})$ *over* $\boldsymbol{\beta} \in \mathbb{R}^p$ *subject to* $\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} = 1$. *Then the maximizer* $\boldsymbol{\theta}$ *will be proportional to* $\boldsymbol{\beta}_{\text{OLS}} = \boldsymbol{\Sigma}^{-1} \mathrm{E}(\mathbf{x}Y)$.
Furthermore, the maximum satisfies $\operatorname{corr}(Y, \boldsymbol{\theta}^T \mathbf{x}) = \sigma_Y^{-2} \operatorname{tr}(\mathbf{M})$ *with* $\mathbf{M} = \boldsymbol{\Sigma}^{-1/2} \mathrm{E}(\mathbf{x}Y) \mathrm{E}^T(\mathbf{x}Y) \boldsymbol{\Sigma}^{-1/2}$.

- Proposition 1 suggests an alternate use of β_{OLS} as the central mean space estimator, estimated through maximizing the correlation between Y and $\beta^{\text{T}}\mathbf{x}$.
- To introduce variable selection, consider the double optimization over both $\mathcal{F} \subseteq \Omega$ and $\beta_{\mathcal{F}} \in \mathbb{R}^{|\mathcal{F}|}$:

$$\max \text{corr}(Y, \beta_{\mathcal{F}}^{\text{T}}\mathbf{x}_{\mathcal{F}}) \text{ subject to } \beta_{\mathcal{F}}^{\text{T}}\boldsymbol{\Sigma}_{\mathcal{F}}\beta_{\mathcal{F}} = 1,$$

- In this double optimization we want to find the *largest* correlation with the *smallest* $\mathcal{F}$.
  - i.e. the smallest $\mathcal{F}$ such that $\text{corr}(Y, \boldsymbol{\theta}_{\mathcal{F}}^{\text{T}}\mathbf{x}_{\mathcal{F}}) = \text{corr}(Y, \boldsymbol{\theta}^{\text{T}}\mathbf{x})$

Proposition

For any index set $\mathcal{F}$ *such that* $\mathcal{A} \subseteq \mathcal{F} \subseteq \Omega$, *we have*

$$\text{corr}(Y, \boldsymbol{\theta}_{\mathcal{A}}^T \mathbf{x}_{\mathcal{A}}) = \text{corr}(Y, \boldsymbol{\theta}_{\mathcal{F}}^T \mathbf{x}_{\mathcal{F}}) = \text{corr}(Y, \boldsymbol{\theta}^T \mathbf{x}),$$

where
$\boldsymbol{\theta} = \text{argmax}\, \text{corr}(Y, \boldsymbol{\beta}^T \mathbf{x})$ *subject to* $\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} = 1$,
$\boldsymbol{\theta}_{\mathcal{A}} = arg\ max\ \text{corr}(Y, \boldsymbol{\beta}_{\mathcal{A}}^T \mathbf{x}_{\mathcal{A}})$ *subject to* $\boldsymbol{\beta}_{\mathcal{A}}^T \boldsymbol{\Sigma}_{\mathcal{A}} \boldsymbol{\beta}_{\mathcal{A}} = 1$, *and*
$\boldsymbol{\theta}_{\mathcal{F}} = arg\ max\ \text{corr}(Y, \boldsymbol{\beta}_{\mathcal{F}}^T \mathbf{x}_{\mathcal{F}})$ *subject to* $\boldsymbol{\beta}_{\mathcal{F}}^T \boldsymbol{\Sigma}_{\mathcal{F}} \boldsymbol{\beta}_{\mathcal{F}} = 1$ *with fixed* $\mathcal{F}$.

- The proposition shows that irrelevant predictors will not increase the correlation between Y and $\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}$.
- Thus solving the double optimization problem leads to $\mathcal{F}$ as an estimate of $\mathcal{A}$.

SCOP - Stepwise Correlation OPtimiztion

- Following similar methods such as COP and of course traditional linear regression, we propose a stepwise procedure.
- At each step: Given the current index set $\mathcal{F}$, should the candidate index $j \in \mathcal{F}^c$ be added to $\mathcal{F}$?
- Formally:

$$H_0: \; Y \perp\!\!\!\perp E(Y|\mathbf{x}_{\mathcal{F}}, X_j)|\mathbf{x}_{\mathcal{F}}$$

v.s.

$$H_a: \; Y \not\perp\!\!\!\perp E(Y|\mathbf{x}_{\mathcal{F}}, X_j)|\mathbf{x}_{\mathcal{F}}$$

- Consider the pivotal quantity

$$t_{\mathcal{F},j} = \text{corr}(Y, \boldsymbol{\theta}_{\mathcal{F}\cup j}^{\mathrm{T}} \, \mathbf{x}_{\mathcal{F}\cup j}) - \text{corr}(Y, \boldsymbol{\theta}_{\mathcal{F}}^{\mathrm{T}} \, \mathbf{x}_{\mathcal{F}})$$

where

$$\begin{aligned} \boldsymbol{\theta}_{\mathcal{F}\cup j} &= \text{argmax}\, \text{corr}(Y, \boldsymbol{\beta}_{\mathcal{F}\cup j}^{\mathrm{T}} \, \mathbf{x}_{\mathcal{F}\cup j}) \\ &\qquad \text{subject to} \quad \boldsymbol{\beta}_{\mathcal{F}\cup j}^{\mathrm{T}} \, \boldsymbol{\Sigma}_{\mathcal{F}\cup j} \, \boldsymbol{\beta}_{\mathcal{F}\cup j} = 1 \\ \boldsymbol{\theta}_{\mathcal{F}} &= \text{argmax}\, \text{corr}(Y, \boldsymbol{\beta}_{\mathcal{F}}^{\mathrm{T}} \, \mathbf{x}_{\mathcal{F}}) \\ &\qquad \text{subject to} \quad \boldsymbol{\beta}_{\mathcal{F}}^{\mathrm{T}} \quad \boldsymbol{\Sigma}_{\mathcal{F}} \quad \boldsymbol{\beta}_{\mathcal{F}} \quad = 1 \end{aligned}$$

Proposition

For any $\mathcal{F} \subseteq \Omega$ *and any* $j \in \mathcal{F}^c$, *suppose* $\mathrm{E}(X_j \mid \mathbf{x}_\mathcal{F})$ *is a linear function of* $\mathbf{x}_\mathcal{F}$ *and let* $X_{j|\mathcal{F}} = X_j - \mathrm{E}(X_j \mid \mathbf{x}_\mathcal{F})$, *then*

a) $t_{\mathcal{F},j} = \mathrm{E}^2\left[\left(\frac{X_{j|\mathcal{F}}}{\sigma_{j|\mathcal{F}}}\right)\left(\frac{Y}{\sigma_Y}\right)\right]$, *where* $\sigma^2_{j|\mathcal{F}} = \mathrm{var}\{X_j - \mathrm{E}(X_j \mid \mathbf{x}_\mathcal{F})\}$.

b) $t_{\mathcal{F},j} = 0$ *under* $H_0 : \mathrm{E}(Y \mid \mathbf{x}) \perp\!\!\!\perp X_j \mid \mathbf{x}_\mathcal{F}$.

i.e. the pivotal quality on the previous slide will actually be 0 when j is irrelevant, and not zero when j should be added to $\mathcal{F}$.

- We have not made any assumptions about the distribution of the pivotal quantity.
- How can we know if it is significantly different from zero?
- Solution: through a permutation test
- The following proposition suggests that we can nonparametrically approximate the distribution of the sample statistic $T^{\mathcal{F},j}$ under H_0 through random permutation.

Proposition

Under the null hypothesis that $Y \perp\!\!\!\perp E(Y|\mathbf{x}_{\mathcal{F}}, X_j)|\mathbf{x}_{\mathcal{F}}$, *if*
a) $X_{j|\mathcal{F}} \perp\!\!\!\perp \mathbf{x}_{\mathcal{F}}$,
b) $Y \perp\!\!\!\perp E(Y|\mathbf{x}_{\mathcal{F}}, X_j)|\mathbf{x}_{\mathcal{F}}$ *implies that* $Y \perp\!\!\!\perp X_j \mid \mathbf{x}_{\mathcal{F}}$.
Then $X_{j|\mathcal{F}} \perp\!\!\!\perp Y$.

Permutation Test

- Given i.i.d. observations $\left\{(\mathbf{x}^{(i)}, Y^{(i)}), i = 1, \ldots, n\right\}$, fix $\mathbf{x}$ and randomly permute Y.
- New sample after the kth permutation is $\left\{(\mathbf{x}^{(i)}, \tilde{Y}_k^{(i)}), i = 1, \ldots, n\right\}$.
- The test statistic based on the kth permuted sample is

$$\tilde{T}_k^{\mathcal{F},j} = \mathrm{E}_n^2 \left[\left(\frac{X_{j|\mathcal{F}}}{\hat{\sigma}_{j|\mathcal{F}}} \right) \left(\frac{\tilde{Y}_k}{\hat{\sigma}_{\tilde{Y}_k}} \right) \right]$$

- $\hat{\sigma}^2_{\tilde{Y}_k}$ and $\hat{\sigma}^2_Y$ are equal.

Permutation Test (cont'd)

Under the conditions in the previous propositions, we have independence between $X_{j|\mathcal{F}}$ and Y, thus

- $\mathrm{E}_n^2[X_{j|\mathcal{F}} Y]$ and $\mathrm{E}_n^2[X_{j|\mathcal{F}} \tilde{Y}_k]$ have the same distribution.
- $\tilde{T}_k^{\mathcal{F},j}$ has the same distribution as $T^{\mathcal{F},j}$ under H_0.
- After permuting K times the approximate p-value is:

$$p^{\mathcal{F},j} = \frac{1}{K} \sum_{k=1}^{K} I\left(T^{\mathcal{F},j} > \tilde{T}_k^{\mathcal{F},j}\right), \text{ where } I(\cdot) \text{ is the indicator function.}$$

Proposition

Suppose $\mathbf{x}$ *is normal and* $(\mathbf{x}, Y)$ *follows the model* $Y = f(\mathbf{x}) + \varepsilon$. *Then as* $n \to \infty$, $T^n_{j|\mathcal{F},\{b\}} \overset{D}{=} T^n_{j|\mathcal{F}}$ *under* $H_0 : Y \perp\!\!\!\perp X_j | \mathbf{x}_{\mathcal{F}}$.

Forward Stepwise Sample Algorithm

1. Set initial working set $\mathcal{F} = \varnothing$.
2. In the addition step, find index j_a such that $j_a = \underset{j \in \mathcal{F}^c}{\operatorname{argmin}}\, p^{\mathcal{F},j}$. If $p^{\mathcal{F},j_a} < p_a$, update $\mathcal{F}$ to be $\mathcal{F} \cup j_a$. If $p^{\mathcal{F},j_a} \geq p_a$, no predictor is added.
3. In the deletion step, find index j_d such that $j_d = \underset{j \in \mathcal{F}}{\operatorname{argmax}}\, p^{(\mathcal{F} \setminus j),j}$. If $p^{(\mathcal{F} \setminus j_d),j_d} > p_d$, update $\mathcal{F}$ to be $(\mathcal{F} \setminus j_d)$. If $p^{(\mathcal{F} \setminus j_d),j_d} \leq p_d$, no predictor is deleted.
4. Iterate between steps 2 and 3 until no predictors can be added or deleted.

Note that p_a and p_d are the predetermined p-value thresholds for the addition step and the deletion step respectively.

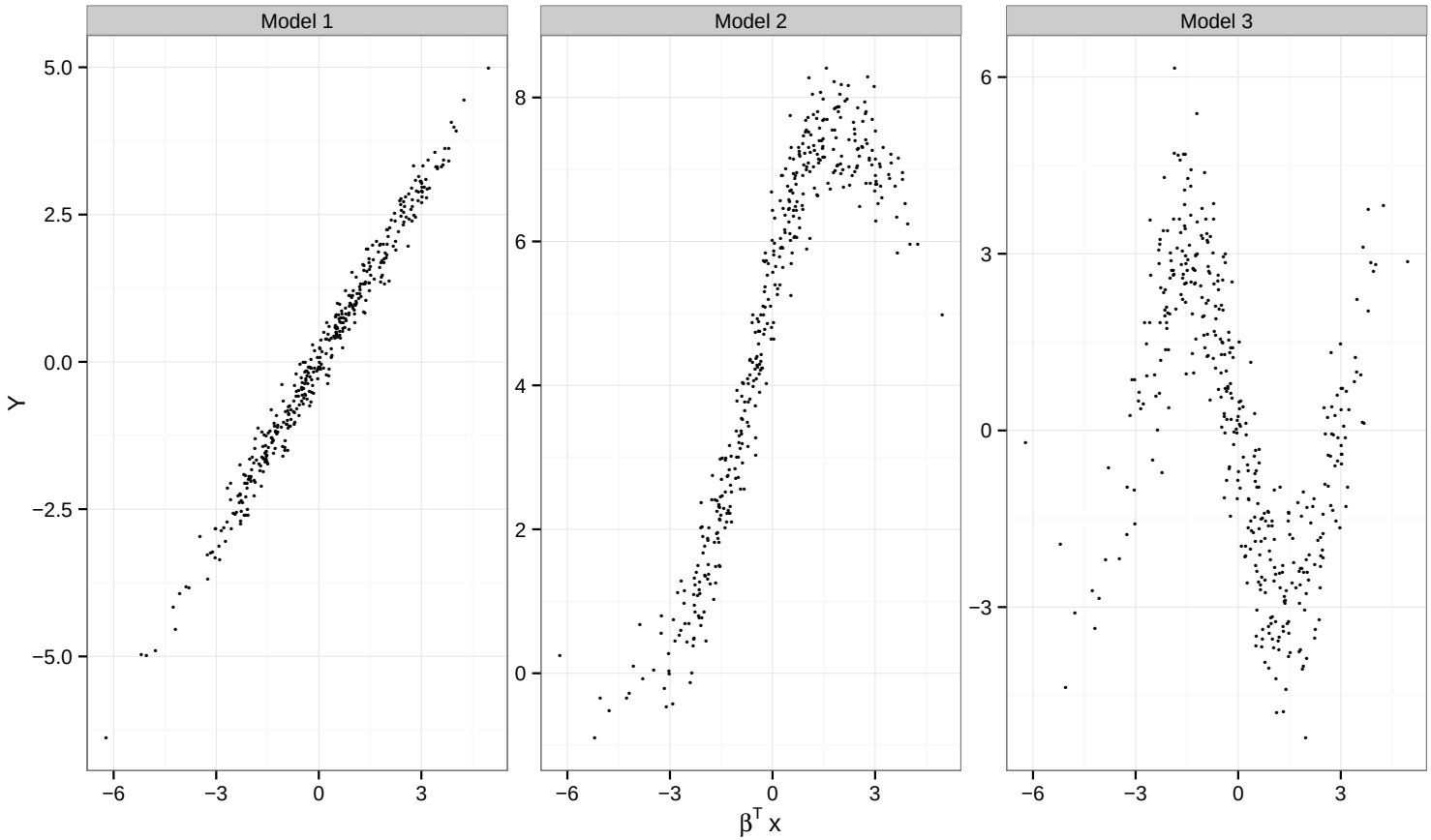
Simulation Results

- Simulated data are constructed under the following models with $n = 400, p = 10, \boldsymbol{\beta} = (1, -1,\ 0,\ 0,\ 0,\ 0.8,\ 1.2,\ 0,\ 0,\ 0)^{\mathrm{T}}$ and $\mathbf{x}$ is normally distributed with mean $\mathbf{0}$ variance $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$, where $\sigma_{ij} = \rho^{|i-j|}$ rho varies from 0.1 to 0.9 in increments of 0.1:

$$
\begin{aligned}
\text{Model 1: } Y &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0, 0.25) \\
\text{Model 2: } Y &= 2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + 2)^2/[(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})^2 + 9] + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0, 0.5) \\
\text{Model 3: } Y &= 3\sin(1.05\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \pi) + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0, 1) \\
\text{Model 4: } Y &= (0.6X_1 + 0.8X_2)^3 + 3\,\mathrm{sign}(X_3) + \exp(X_{10}) + \varepsilon, \\
&\quad \text{where } \varepsilon \sim \mathcal{N}(0, 1)
\end{aligned}
$$

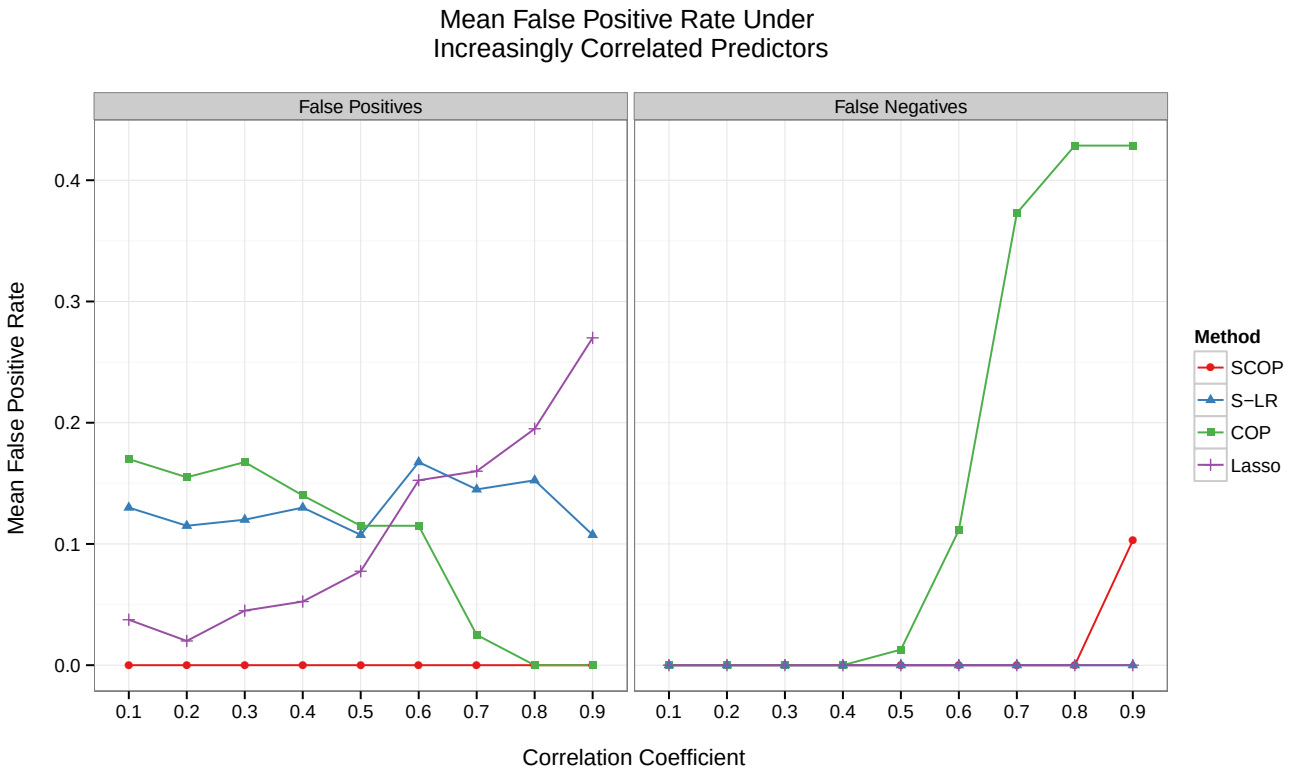
- We compare to: LASSO, Stepwise Linear Regression (SLR) and Correlation Pursuit (COP).

Graphs of Models 1 to 3



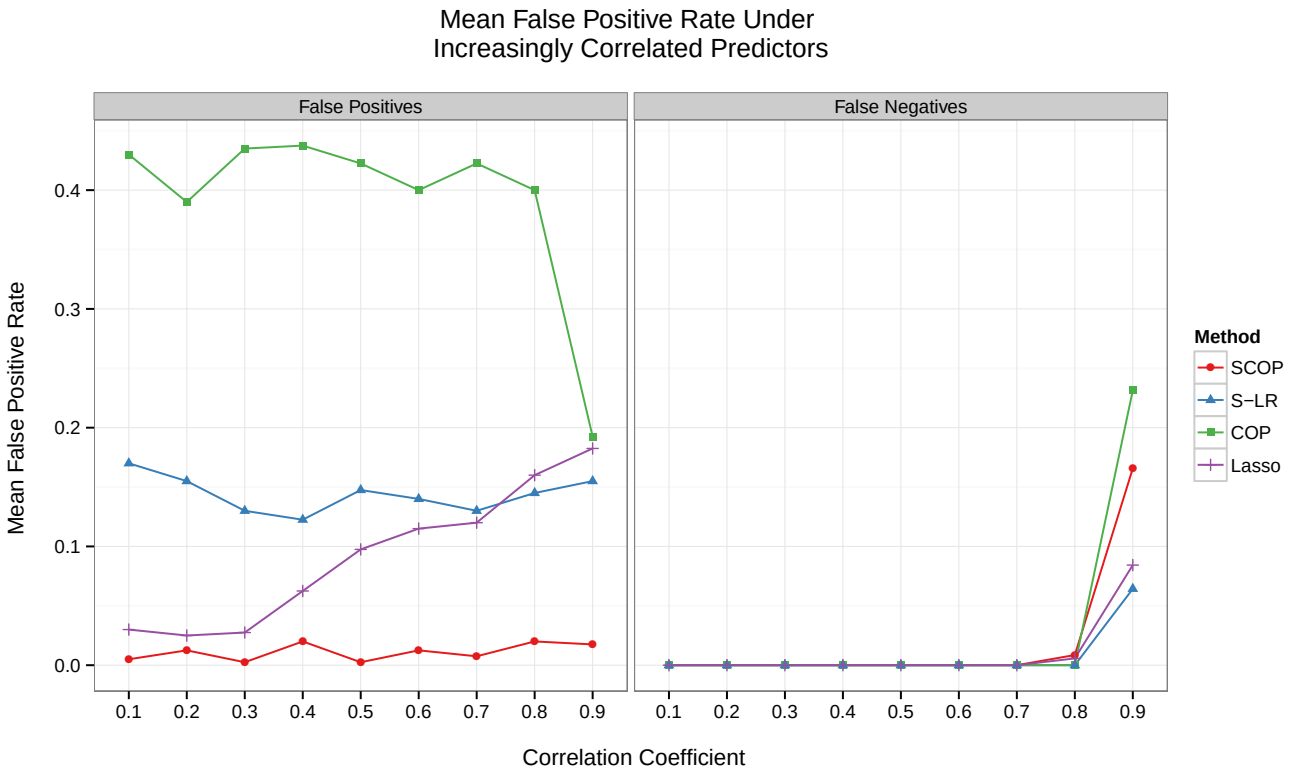
Model 1 Simulation Results

$$Y = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0, 0.25)$$



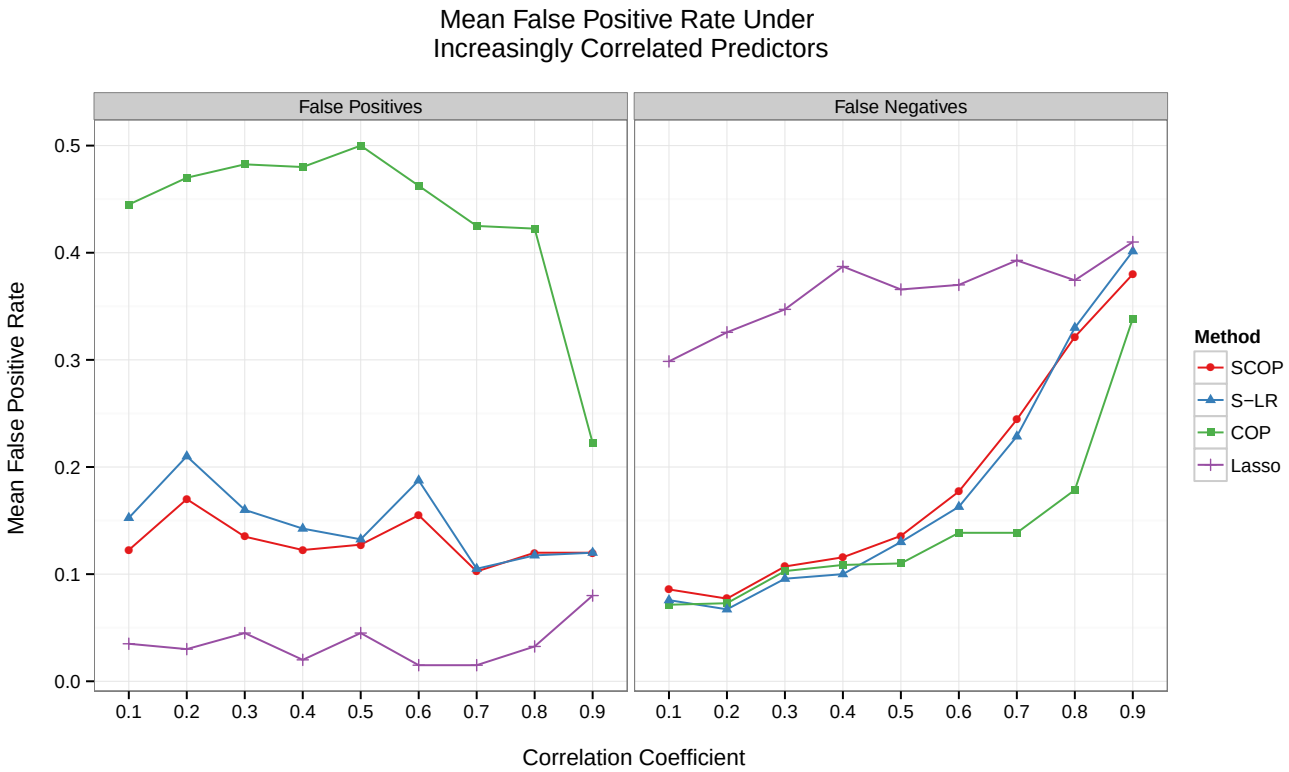
Model 2 Simulation Results

$$Y = 2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + 2)^2/[(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})^2 + 9] + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0, 0.5)$$

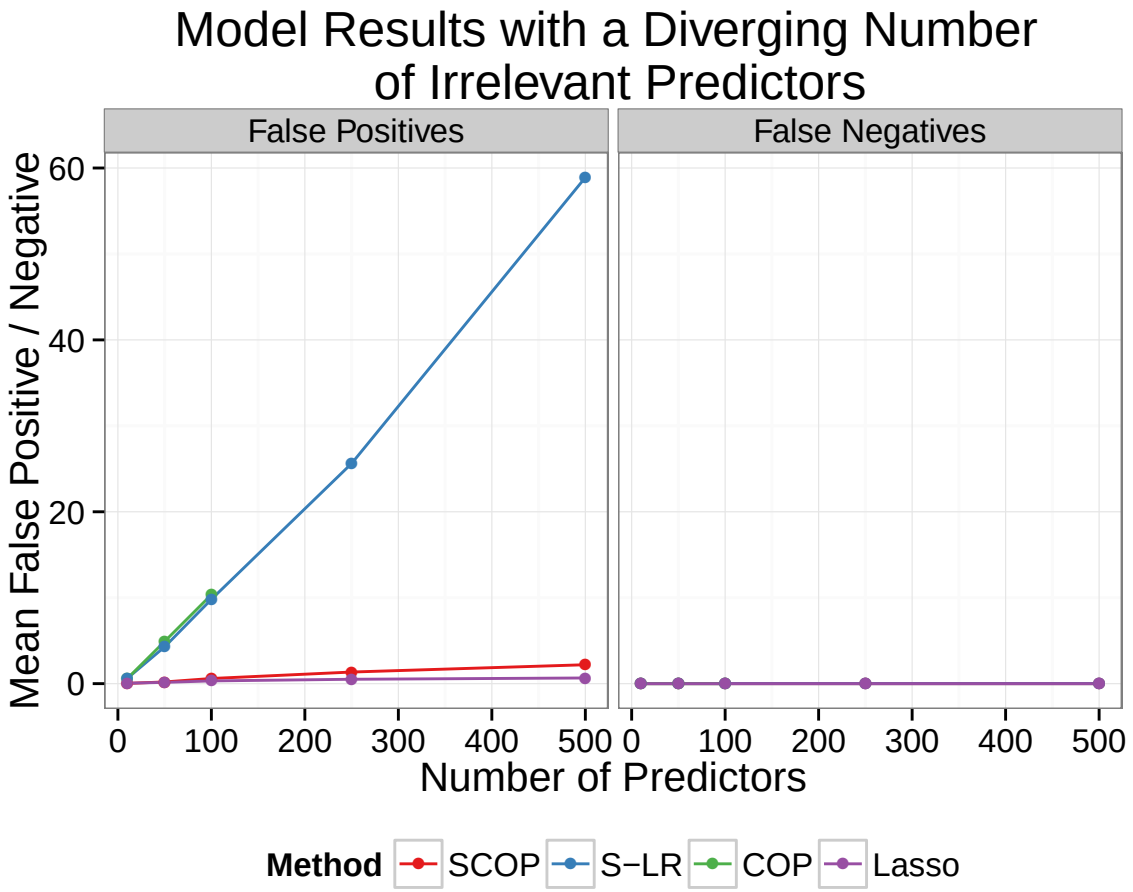


Model 3 Simulation Results

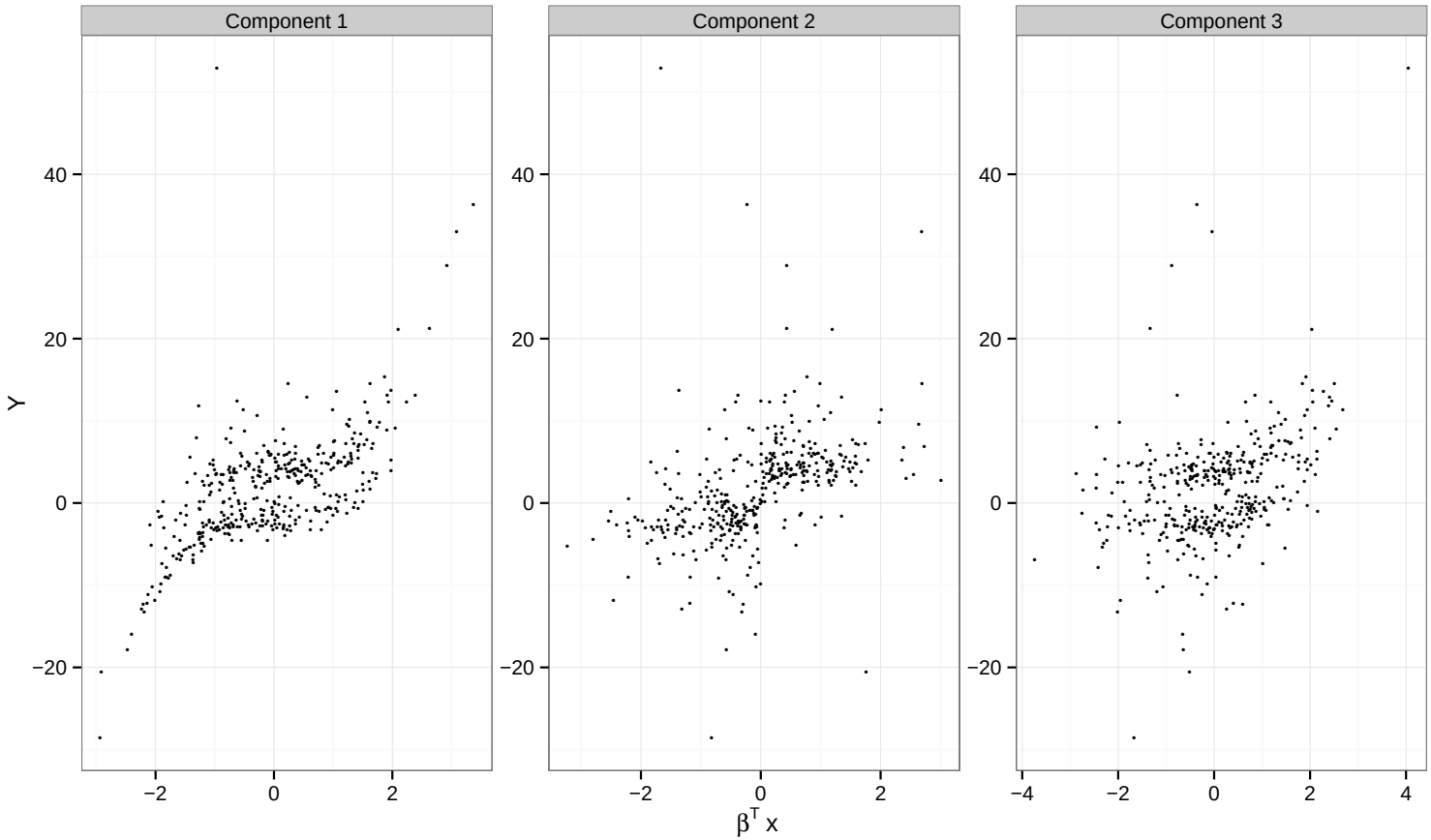
$$Y = 3\sin(1.05\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \pi) + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0, 1)$$



Simulation Results with Diverging p and $\rho = 0.3$

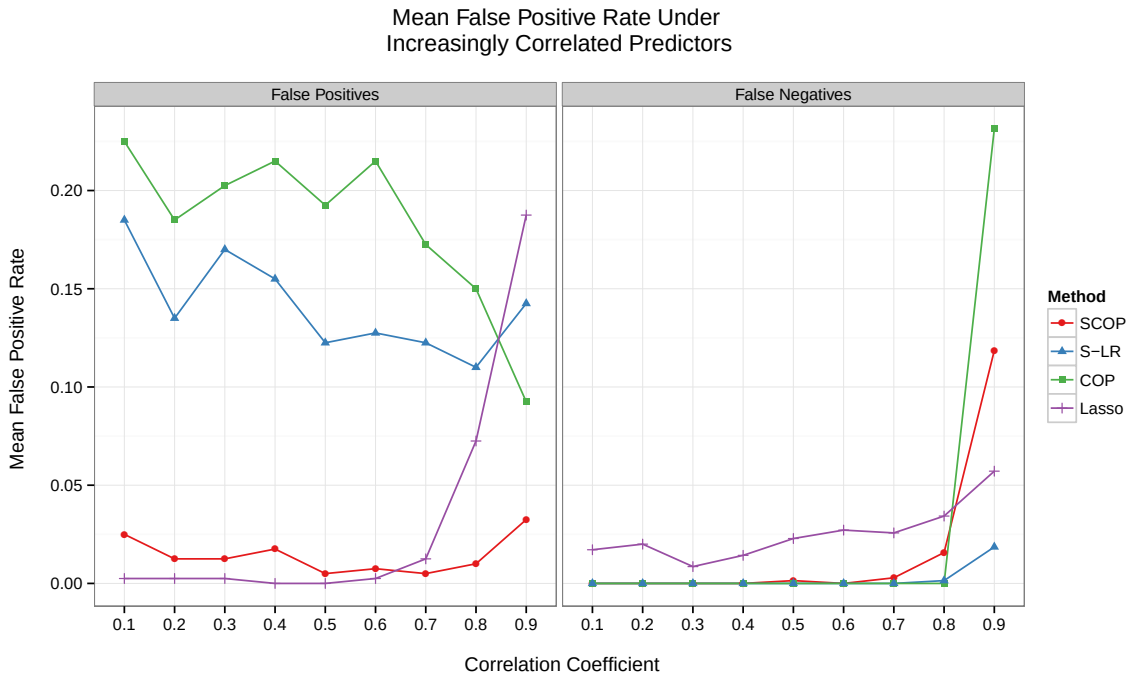


Plots of Simulated Model 4



Model 4 Simulation Results

$$Y = (0.6X_1 + 0.8X_2)^3 + 3\,\text{sign}(X_3) + \exp(X_{10}) + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0,1)$$



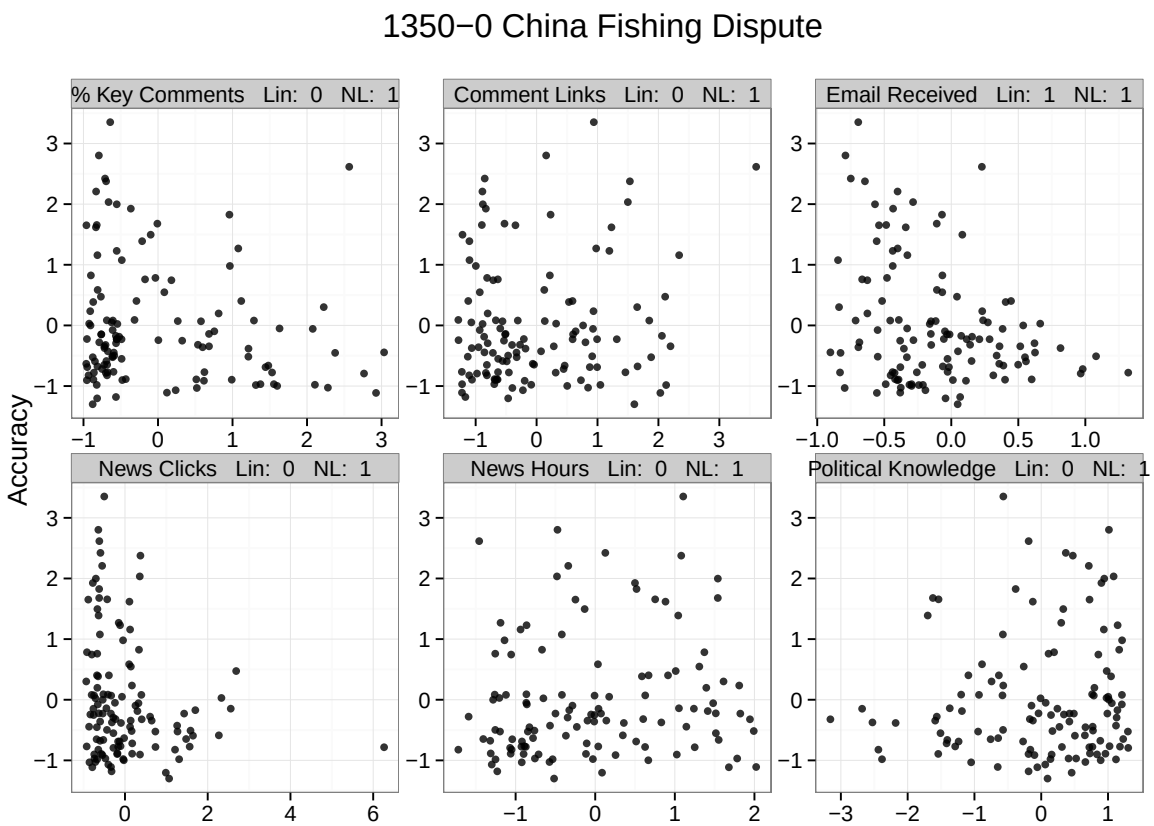
The Good Judgement Project Forecasting Tournament Data

- Goal: to accurately predict the outcome of various geopolitical events by aggregating crowd-sourced probability forecasts.
  - eg. Before 1 May 2014, will China confiscate the catch or equipment of any foreign fishing vessels in the South China Sea for failing to obtain prior permission to enter those waters?
- Tournament is set up with various experimental conditions, we'll look at data from forecasters working in teams.
- Accuracy is measured using squared error loss.

The Good Judgement Project Forecasting Tournament Data

- Question: which aspects of a forecaster (forecasting behavior, personality tests, etc.) most contribute to accuracy?
- Ultimate Goal: Categorize question types and understand which variables can be used to upweight certain forecasters when aggregating forecasts.
- We have 44 variables and between 62 and 299 observations (forecasters per question) and all variables are marginally standardized.
- We find that SCOP selects variables that stepwise linear regression misses.

Forecasting Data (cont'd)



Modeling

- Generalized Additive Model

	AIC	MSE	Test MSE*
SLR	331.5	0.89	1.25
SCOP	305.7	0.51	0.97

* Test MSE is the MSE after splitting the data into 80% training and 20% testing.

- Linear Model

	R^2	MSE
SLR	0.03	1.00
SCOP	0.22	0.91

THANK YOU
QUESTIONS?