On Variable Selection and Effective Estimations of Interactive and Quadratic Sensitivity Coefficients: A Collection of Regularized Regression Techniques

M&C 2015

Weixiong Zheng and Ryan G. McClarren

Texas A&M University
Section 1

1 Introduction
   • Background
   • Regularization

2 Model description
   • Problem settings

3 Tests
   • Variable selection
   • Coefficient Estimation

4 Summary and Future Work
Background for variable selection and sensitivity estimation

- For parametric uncertainties the curse of dimensionality is still a problem
  - This is especially true for pairwise interactions and second-order sensitivity coefficients
- In some problems in engineered systems, high order sensitivity coefficients and variable significance are important
- Two potential ways: perturbation theory and random sampling based estimation
  - High order perturbation theory could be hard to implement in multiphysics codes
  - Random sampling based estimation equipped with regression is simple to implement, but for second-order and interaction coefficients, multi-collinearity leads to ill-conditioned problems.
- Our focus: regularized regressions
  - Add small constraints to the regression could bring in numerical stability and well-posedness
  - Different constraints result in different estimation process and results
Regression problems

- The general regression problem is written as

\[ Y = X\beta + \varepsilon \] (1)

- \( Y \): data (outcomes), \( X \): input matrix, \( \beta \): regression coefficients, \( \varepsilon \): errors

\[
\begin{align*}
Y &= \begin{pmatrix}
Y_1 \\
Y_2 \\
s\vdots \\
Y_n
\end{pmatrix}, \quad 
X &= \begin{pmatrix}
1 & X_{11} & X_{12} & \cdots & X_{1p} \\
1 & X_{21} & X_{22} & \cdots & X_{2p} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & X_{n1} & X_{n2} & \cdots & X_{np}
\end{pmatrix}, \quad 
\beta &= \begin{pmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_p
\end{pmatrix}
\end{align*}
\]

(2)

- \( n \) is number of samples and \( p \) is the number of independent variables
- Regression aim: estimate the coefficients, \( \beta \), in Eq. (1).
The direct “solve” by ordinary least squares (OLS)

\[ \beta \approx (X^T X)^{-1} X^T Y. \]

Several common situations can make OLS ill-conditioned or ill-posed:
- \( n < p \): Number of samples is smaller than number of parameters
- \( X \) contains interdependencies, i.e., multi-collinearity, if high order terms are included
- In either case, \( X^T X \) is rank deficient and not invertible
- Alternative approaches like the pseudo-inverse can give unreasonable results as has been demonstrated in previous work.

A possible cure is regularization: change the regression problem to make the system well-posed and give it better properties.
Basic ideas

- Another way to think of OLS regression is as the minimizer of the $\ell_2$ norm of the error between the fit and the original data:

$$\beta = \arg\min_\beta \|Y - X\beta\|_2^2$$  \hspace{1cm} (3)

- Equivalent to a direct solve of the regression problem:

$$\beta_{\text{OLS}} = (X^T X)^{-1} X^T Y$$

- Ineffective and inaccurate for ill-conditioned problems

- Regularization: add additional information
  - Add a constraint term to the Lagrangian or cost function of the minimization problem
  - Different types of constraints have different effects
  - Certain regularizations can guarantee well-posedness.
Non-Bayesian Regularization Regression Approaches

In these methods we explicitly change the minimization problem.

- **Lasso regression** (OLS plus an $\ell_1$ penalty based on size of $\beta$’s):
  \[
  \beta = \arg\min_{\beta} \{ \| Y - X\beta \|_2^2 + \lambda_1 \| \beta \|_1 \} \tag{4}
  \]

- **Ridge regression** (OLS plus an $\ell_2$ penalty based on size of $\beta$’s):
  \[
  \beta = \arg\min_{\beta} \{ \| Y - X\beta \|_2^2 + \lambda_2 \| \beta \|_2^2 \} \tag{5}
  \]

- **Elastic net regression** (Combination of Lasso and Ridge):
  \[
  \beta = \arg\min_{\beta} \{ \| Y - X\beta \|_2^2 + \alpha \lambda_1 \| \beta \|_1 + (1 - \alpha) \lambda_2 \| \beta \|_2^2 \} \tag{6}
  \]

- **Dantzig selector** (Minimize $\ell_\infty$ error in fit with $\ell_1$ penalty on $\beta$’s):
  \[
  \beta = \arg\min_{\beta} \{ \| \beta^T(Y - X\beta) \|_\infty + \lambda_1 \| \beta \|_1 \} \tag{7}
  \]
Non-Bayesian L-2 norm constraint put too much strength on limiting parameters with higher magnitudes: over-penalization
The Bayesian version of regularized-regressions differs from non-Bayesian in the sense that hyperparameters, i.e. $\lambda$, are sampled in the Bayesian inference process.

In other words, the Bayesian methods take similar forms to the non-Bayesian problems, but estimate the parameters through a Bayesian framework.

Bayesian theory:

$$p(\beta|D) = \frac{p(D|\beta)p(\beta)}{\int d\beta \ p(D|\beta)p(\beta)}$$

Bayesian inference short introduction:

- Sample realizations of parameters from priors
- Calculate posteriors
- Modify the priors for the next iteration and repeat until reaching the maximum iteration
- Do statistics with the results from the iterations
Bayesian Regularization Regression Approaches

Bayesian lasso prior and posterior:

\[
p(\beta | \sigma^2, \lambda_1) = \prod_{j=1}^{p} \frac{\lambda_1}{2\sqrt{\sigma^2}} \exp \left\{ -\frac{\lambda_1 |\beta_j|}{\sqrt{\sigma^2}} \right\} \tag{9a}
\]

\[
p(\beta | \sigma^2, \lambda_1, Y, X) \propto \exp \left\{ -\frac{1}{2\sigma^2} \| Y - X\beta \|_2^2 - \frac{\lambda_1 \|\beta\|_1}{\sqrt{\sigma^2}} \right\} \tag{9b}
\]

Bayesian ridge prior and posterior:

\[
p(\beta | \sigma^2, \lambda_2) = \left( \frac{\lambda_2}{2\pi\sigma^2} \right)^{(n+1)/2} \exp \left\{ -\frac{\lambda_2 \|\beta\|_2^2}{2\sigma^2} \right\} \tag{10a}
\]

\[
p(\beta | \sigma^2, \lambda_2, Y, X) \propto \exp \left\{ -\frac{1}{2\sigma^2} \| Y - X\beta \|_2^2 - \frac{\lambda_2 \|\beta\|_2^2}{\sigma^2} \right\} \tag{10b}
\]
Automatic relevance determination (ARD) prior and posterior

\[ p(\beta | \sigma^2, \lambda_2) \propto \exp \left\{ - \sum_{j=1}^{p} \frac{\lambda_2}{2\sigma_j^2} |\beta_j|^2 \right\}, \]  \hspace{1cm} (11a)

\[ p(\beta | \sigma^2, \lambda_2, Y, X) \propto \exp \left\{ - \frac{1}{2\sigma^2} ||Y - X\beta||^2_2 - \sum_{j=1}^{p} \frac{\lambda_2}{2\sigma_j^2} |\beta_j|^2 \right\} \]  \hspace{1cm} (11b)

- ARD is very similar to Ridge regression except that it has a different \( \sigma_j \), controlling the variance, for each variable.
Section 2

1 Introduction
   - Background
   - Regularization

2 Model description
   - Problem settings

3 Tests
   - Variable selection
   - Coefficient Estimation

4 Summary and Future Work
Problem settings

Lattice of TRIGA fuels pin modeled with MCNP

- QoI: $k_{\text{eff}}$
There are 299 sensitivity coefficients taken into account in this problem:

- 23 input parameters:
  - 6 geometric parameters: e.g. \( r\)-fuel (fuel radius)
  - 17 material parameters: e.g. \( \rho\)–Zr (Zr rod mass density)
- 253 pairwise interactions (23 choose 2)
- 23 quadratic terms

The aim is to investigate the sensitivity of the criticality to the parameters, especially the second order terms. The model is:

\[
\frac{\delta k}{k} \approx \sum_{i=1}^{23} c_i \left( \frac{\delta x_i}{x_i} \right) + \sum_{i=1}^{22} \sum_{j=i+1}^{23} c_{ij} \left( \frac{\delta x_i}{x_i} \right) \left( \frac{\delta x_j}{x_j} \right) + \sum_{i=1}^{23} c_{ii} \left( \frac{\delta x_i}{x_i} \right)^2
\]

(12)

where \( c_i, c_{ij} \) and \( c_{ii}, i = 1, \cdots, 23, j \neq i \), are the first order, interactive and quadratic sensitivity coefficients, respectively.
We are going to compare reference sensitivity coefficients to the coefficients computed by various regularized regression techniques using many few code runs (cases).

The reference coefficients are computed using 1058 cases.
- We need 46 total simulations for the linear and quadratic parameters
- 1012 simulations are needed for the 253 interactions (4 simulations for each)

The goal of this research is to see if regularized regression techniques can give coefficient estimates close to the references using many fewer simulation runs than the 1058 cases.
Quasi-uniform multi-D sampling

- For the regression results we sample from the 23 parameters using by Latin Hypercube sampling.
- For any number of samples we fit the entire 299-sample sensitivity model for $k_{\text{eff}}$.
- A 12-sample example is shown below. 2D projections are uniform.
Section 3

1. Introduction
   - Background
   - Regularization

2. Model description
   - Problem settings

3. Tests
   - Variable selection
   - Coefficient Estimation

4. Summary and Future Work
One use of sensitivity analysis is to down-select from the large parametric uncertainty space to a smaller set of important parameters.

After this variable selection process, a more detailed study can be performed on the important variables.

In our case we would like to use a small number of samples (code-runs) to select the important variables.

Below we’ll discuss the selection of significant pairwise interaction and quadratic terms.
Variable Selection: Interaction Terms

- Coefficients with a magnitude above 10% of the highest magnitude (from corresponding method) will be selected as significant.
- Reference result has 15 significant pairwise interactions
<table>
<thead>
<tr>
<th>Sample size</th>
<th>True Positive</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>Lasso</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>150</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>200</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>250</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>299</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

- Least square regression (OLS): gives hundreds of false positives
- Regularization helps remove false positives, though no method gets all 15 true parameters using the small number of samples considered.
- Lasso: 5 right with 0 wrong
- Dantzig selector (DS): more true positives with 3 wrong picks (borderline picks)
- Ridge: only 3 true positives but 2 false negatives: over-penalization
Variable Selection: Interactions (cont’d)

<table>
<thead>
<tr>
<th>Sample size</th>
<th>True Positives</th>
<th>False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>Lasso</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>150</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>200</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>250</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>299</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

- Bayesian ridge and Bayesian lasso are comparable as both get 8 correct parameters at 200 samples.
- ARD seems makes most conservative picks: small false positives and small true positives.
Variable Selection: Quadratic

- Same 10% threshold from interaction case.
- Reference result has 3 significant variables

<table>
<thead>
<tr>
<th>Sample size</th>
<th>True Positive</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>Lasso</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>150</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>200</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>299</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

- OLS and Ridge not useful in this case.
- Lasso and elastic net converge to the correct answer.
- Dantzig selector does have a high number of false positives with 299 samples, but these could be borderline cases (near 10%)
Variable Selection: Quadratic (cont’d)

<table>
<thead>
<tr>
<th>Sample size</th>
<th>OLS</th>
<th>Lasso</th>
<th>BRidge</th>
<th>BLasso</th>
<th>ARD</th>
<th>OLS</th>
<th>Lasso</th>
<th>BRidge</th>
<th>BLasso</th>
<th>ARD</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>15</td>
<td>17</td>
<td>5</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>17</td>
<td>12</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>150</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>18</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>200</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>15</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>299</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>15</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

- BLasso, BRidge: similar with DS, borderline picks
- ARD: conservative
Now we ask a more difficult question of the methods: estimate the numeric value of the coefficients and compare with the reference result.

- Each parameter is assigned an ID
- IDs from 24 to 276: interactive coefficients
- IDs from 277 to 299: quadratic coefficients

The results that follow all use 299 samples, about 28% of those used in the reference calculation.
Blue dots are regression estimations, red lines are reference
Coefficient Estimation: Interactions (cont’d)

Parameter IDs
50 100 150 200 250

Coef. EN
-0.4
-0.3
-0.2
-0.1
0
0.1

Estimations
Reference

Parameter IDs
50 100 150 200 250

Coef. BRidge
-0.4
-0.3
-0.2
-0.1
0
0.1

Estimations
Reference

Parameter IDs
50 100 150 200 250

Coef. BLasso
-0.4
-0.3
-0.2
-0.1
0
0.1

Estimations
Reference

Parameter IDs
50 100 150 200 250

Coef. ARD
-0.4
-0.3
-0.2
-0.1
0
0.1

Estimations
Reference
Coefficient Estimation: Quadratic

Weixiong Zheng (TAMU)
Coefficient estimation: quadratic (cont’d)

Parameter IDs
280 285 290 295 300

Coef. EN
-0.8
-0.6
-0.4
-0.2
0

Estimations
Reference

Parameter IDs
280 285 290 295 300

Coef. BRidge
-0.8
-0.6
-0.4
-0.2
0

Estimations
Reference

Parameter IDs
280 285 290 295 300

Coef. BLasso
-0.8
-0.6
-0.4
-0.2
0

Estimations
Reference

Parameter IDs
280 285 290 295 300

Coef. ARD
-0.8
-0.6
-0.4
-0.2
0

Estimations
Reference
Section 4

1. Introduction
   - Background
   - Regularization

2. Model description
   - Problem settings

3. Tests
   - Variable selection
   - Coefficient Estimation

4. Summary and Future Work
Summaries

- Investigated seven types of regularization methods on second order variable selection and sensitivity coefficient estimations.
- On variable selection, we found Bayesian lasso, Bayesian ridge, Dantzig selector and elastic net are promising and comparable to lasso, a commonly used method in the statistics community.
- On coefficient estimation:
  - L-2 norm regularized methods: ARD and ridge are too conservative
    - Ridge has over-penalization
  - Lasso and EN present similar estimations that selects significant variables out but not with correct magnitudes
  - Dantzig selector, Bayesian lasso and Bayesian ridge present similar high accuracy on second order coefficient estimations
    - BRidge fixes the over-penalization
Future work

- Other regularizations are worth investigation: e.g. $\ell_{0.5}$ “norm”
- Apply the methods with nuclear data sensitivity research
  - Include impact of covariances
  - Even higher dimensional problems common.
On Variable Selection and Effective Estimations of Interactive and Quadratic Sensitivity Coefficients: A Collection of Regularized Regression Techniques

M&C 2015

Weixiong Zheng and Ryan G. McClarren

Texas A&M University