A Generalized Approach to Ridge Regression/Tikhanov

Katya Vasilaky

February 7, 2017

Overview of talk

1. Part I
   1.1 Review Standard Ridge
   1.2 Derivation of Generalized Ridge Results
2. Part II
   2.1 Test Generalized Ridge under simulated case
   2.2 Test Generalized Ridge with real data and cross validation
   2.3 Reconstruct an image from projections

Motivation of Inverse Problems

- Inverse problems: compute information about some “interior” properties using “exterior” measurements.
    - Inference: Covariates $->$ Coefficients $->$ Outcome
    - Tomography: Xray source $->$ Object $->$ X Ray Dampening
    - ML: Features $->$ Effect Size $->$ Classifier/Prediction

Why is Regularization Used?

- OLS is BLUE when the covariate matrix (A) is full rank
- But when A is ill-conditioned (covariates correlated), estimators will be sensitive to noise
- Regularization methods are used to dampen the effects of the sensitivity to noise
- I present a generalization to the frequently used Ridge Regression
  - Performs as well or better than Standard Ridge
  - Allows for a more flexible weighting of singular values than Standard Ridge
  - Useful for data where covariates are correlated: large consumer data sets, health data

Reviewing the Basics of Gauss Markov

Recall the Gauss Markov estimator for the least squares problem:

$$y = Ax + e$$

$$min_x||Ax - y||_2^2$$

A is nxp, with rank p,

$$\hat{x} = (A'A)^{-1}A'y$$

$$Var(\hat{x}) = \sigma^2(A'A)^{-1}$$

and the MSE$(\hat{x}) = \sigma^2 Tr(A'A)^{-1} = \sigma^2 \sum_i \frac{1}{\sigma_i^2}$,
where σ_i^2 is the *ith* eigenvalue of $A'A$

This will serve as a comparison later on.

When is Gauss Markov MSE Large?

- When $A'A$ is ill-conditioned (nearly not full rank), the solution to OLS is sensitive to noise ($y = \bar{y} + \epsilon$)
- This can occur when covariates are highly correlated, the number of covariates exceeds the number observations, or A is sparse
  - OLS is still BLUE
  - But the standard errors of $\hat{x}$, and MSE, will be large (σ_i's are small)
    $\text{Var}(\hat{x}) = \sigma^2(A'A)^{-1}$
    $\text{MSE}(\hat{x}) = \sigma^2 Tr(A'A)^{-1} = \sigma^2 \sum_i \frac{1}{\sigma_i^2}$
- And $\hat{x}$ can deviate very far from the true solution x

Example: Noise is magnified if the covariate matrix is ill-conditioned

$$A = U\Sigma V' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 10^{-6} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{1}$$

$$(A'A)^{-1}A'y = \begin{bmatrix} 1 & 0 \\ 0 & 10^{6} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 10^{6} \end{bmatrix} \text{Noiseless solution} \tag{2}$$

$$(A'A)^{-1}A'(y+e) = \begin{bmatrix} 1 & 0 \\ 0 & 10^{6} \end{bmatrix} \begin{bmatrix} 1 \\ 1.1 \end{bmatrix} = \begin{bmatrix} 1 \\ 10^{6} + 10^{5} \end{bmatrix} \text{Naive solution} \tag{3}$$

Regularization problems formulate a nearby problem, which has a more stable solution:

$$Min_x(||Ax - y||_2^2 + \lambda||x||_2^2), \lambda > 0$$

where we introduce the term $\lambda||x||_2^2$ perturbing the least-squares formulation.

- The estimated $\hat{x}$ is no longer unbiased, however, the MSE $=$ Variance $+ \mathrm{Bias}^2$, may be smaller than BLUE.

The regularization can be L1 (Lasso) or L2. The objective is to choose a lambda that brings x close to the noiseless solution.

The regularized least squares problem problem becomes:

$$Min_x||Ax - y||_2^2 + \lambda||x - x_0||_2^2$$

$$\hat{x} = (A'A + \lambda I_n)^{-1}A'y$$

$$\text{MSE} = \sigma^2 \sum_{n=1}^{\infty} \frac{\sigma_i^2}{(\sigma_i^2+\lambda)^2} + \lambda^2 \sum_{n=1}^{\infty} \frac{\alpha_i^2}{(\sigma_i^2+\lambda)^2}$$

(Hoerl and Kennard, 1970)[1]

[1] Note: where $\alpha = Vx$, $A'A = V\Sigma^2 V'$

Other regularization methods include:

- Lasso, with a L1 penalty weighted by λ
- Truncated SVD
- Elastic Net, a weighted sum of L1 and L2
- 'Generalized' Tikhanov

Introduction to Generalized Iterative Ridge

- Rather than computing one solution I will compute a sequence of solutions which approximates the noiseless solution $(A'A)^{-1}A'\bar{y}$ on its way to the naive solution $(A'A)^{-1}A'(\bar{y}+e)$.
- Since the operator $(A'A+\lambda I)^{-1} = UDiag(\frac{1}{\sigma_i^2+\lambda})V'$ has eigenvalues that are less than one, it is contracting.
- I show that repeated application of this operator will converge to the naive solution.

New normal equations with added constraint become:

$$(A'A+\lambda I)x_1 = A'y + \lambda x_0$$

Introduction to Generalized Iterative Ridge

$$(A'A + \lambda I)x_1 = A'y + \lambda x_0$$

The solution for x_1 is:

$$\hat{x}_\lambda^1 = (A'A + \lambda I)^{-1}A'y + \lambda(A'A + \lambda I)^{-1}x_0 \text{ (Regular Ridge, } x_0 = 0)$$

Then substituting $\hat{x}_\lambda^1$ into x_0, we obtain $\hat{x}_\lambda^2$, and if we substitute $\hat{x}_\lambda^{k-1}$ into $\hat{x}_\lambda^{k-2}$, we obtain:

$$\hat{x}_\lambda^k = \Sigma_{i=1}^k \lambda^{i-1}((A'A + \lambda I)^{-i}(A'y) + \lambda^k(A'A + \lambda I)^{-k}x_0)$$

where, $\Sigma_{i=1}^k \lambda^{i-1}((A'A + \lambda I)^{-i}(A'y)$, is a contracting opearator.

Generalized Iterative Solution, x_λ^k Now we sum the geometric series using single value decomposition (SVD) of A: $A = U\Sigma V^T$, where $\Sigma = \mathrm{Diag}[\sigma_1, .., \sigma_n, 0....0] \in \mathbb{R}^{mxn}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$, and where n is the rank of A. The iterative solution after k iterations, we have:

$$V \begin{bmatrix} \frac{1}{\sigma_1} - \frac{1}{\sigma_1}(\frac{\lambda}{\lambda+\sigma_1^2})^k & \cdots & & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & & \\ 0 & \cdots & \frac{1}{\sigma_n} - \frac{1}{\sigma_n}(\frac{\lambda}{\lambda+\sigma_n^2})^k & 0 & \cdots & 0 \end{bmatrix} U^T y$$

For practical ease we can set x_0=0. It does not affect the solution, as the last term attached to x_0 is pulled quickly towards 0.

The limits of Generalized Iterative Ridge (GIR)

The GIR solution tends towards the naive solution as
k $- > \infty$:
$(A'A)^{-1}A'y = V\sum^{-1} U'y$, or

$$V \begin{bmatrix} \frac{1}{\sigma_1} & & \dots & & 0 & \dots & 0 \\ \vdots & \ddots & & \vdots & & & \\ 0 & \dots & & \frac{1}{\sigma_n} & 0 & \dots & 0 \end{bmatrix} U^T y$$

The limits of Generalized Iterative Ridge (GIR)

When k =1, GIR collapses to Standard Ridge:

$$V \begin{bmatrix} \frac{\sigma_1}{\lambda+\sigma_1^2} & & \cdots & & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots & & & \\ 0 & \cdots & & \frac{\sigma_n}{\lambda+\sigma_n^2} & 0 & \cdots & 0 \end{bmatrix} U^T y$$

The diagonal entries are the filters for the singular values, and it is clear that the iterative solution's filter is a generalization of the two extremes.

Summary, Generalized Iterative Ridge

- Falls between Standard Ridge and OLS
- The filter is more general than Standard Ridge

Choosing k, and λ using the Residual Error

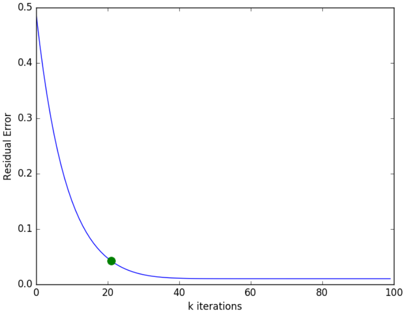
We can use this simple and nice expression for the residual error, $||A\hat{x}_k - y||$, as a function of k and λ, and choose it's lowest point or maximum curvature in convexity.

$$RE(\lambda, k) = || \begin{bmatrix} (\frac{\lambda}{\lambda+\sigma_1^2})^k & & \dots & & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots & & & \\ 0 & & \dots & (\frac{\lambda}{\lambda+\sigma_n^2})^k & 0 & \dots & 0 \end{bmatrix} U'y||$$

Choosing k, and λ using the Residual Error

We can see that the residual error, $||Ax_k - y||$, is convex and decreases as k increases, for a given lambda.

Residual error as a function of k, given lambda. Choose k at the elbow.



Intuition for choosing k at residual error's elbow

Residual error's convexity can be seen when we take the difference of $\hat{x}_\lambda^k$ and the noiseless OLS solution $\hat{x}$.

- First term, iteration error declines monotonically as k $- > \infty$
- Second term, noise term, increases monotonically with k (to the OLS residual).

(S1)

$$-V \begin{bmatrix} \frac{1}{\sigma_1}(\frac{\lambda}{\lambda+\sigma_1^2})^k & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & & \\ 0 & \dots & \frac{1}{\sigma_n}(\frac{\lambda}{\lambda+\sigma_n^2})^k & 0 & \dots & 0 \end{bmatrix} U^T \bar{y}$$

$$\text{(S2) } V \begin{bmatrix} \frac{1}{\sigma_1} - \frac{1}{\sigma_1}(\frac{\lambda}{\lambda+\sigma_1^2})^k & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & & \\ 0 & \dots & \frac{1}{\sigma_n} - \frac{1}{\sigma_n}(\frac{\lambda}{\lambda+\sigma_n^2})^k & 0 & \dots & 0 \end{bmatrix} U^T e$$

Summary, Choosing k and λ

- Grid search over λ
- Choose k at the elbow

Comparing the Filters of Standard Ridge and GIR

- A key contribution of the iterative solution is in the filters.
- The filters dampen the effects of small singular values.

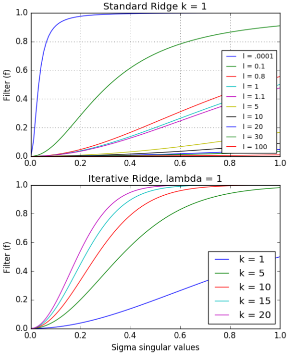
Blue $\frac{1}{\sigma_i}$

Standard Ridge $\frac{1}{\sigma_i}\boldsymbol{(1-(\frac{\lambda}{\lambda+\sigma_1^2}))=\frac{\sigma_i^2}{\lambda+\sigma_i^2}}$

Iterative Ridge$\frac{1}{\sigma_i}\boldsymbol{(1-(\frac{\lambda}{\lambda+\sigma_1^2})^k)}$

What is the extra advantage of the k iteration parameter in the filter?

Iterative Ridge introduces additional flexibility in the weighting of each covariate



- With Standard Ridge/Tikhanov, even medium valued sigma's are heavily penalized (for lambda $= 0.1$)
- Notice $\lambda = 0.0001$ Standard Ridge and Iterative Ridge $\lambda = 0.01, k = 20$ have similar shapes, however, a small λ increases the noise.

Comparing the MSEs for BLUE, Standard Ridge, and GIR

Last but not least, I have the generalized MSE for Iterative Ridge, exposes the variance and the bias:[2]

MSE OLS

$$\sigma^2 \sum_i \frac{1}{\sigma_i^2}$$

MSE Ridge

$$\sigma^2 \sum_{n=1}^{\infty} \frac{\sigma_i^2}{(\sigma_i^2+\lambda)^2} + \lambda^2 \sum_{n=1}^{\infty} \frac{\alpha_i^2}{(\sigma_i^2+\lambda)^2}$$

MSE Iterative Ridge

$$\sigma^2 \sum_{n=1}^{\infty} \frac{1}{\sigma_i^2}(1-(\frac{\lambda}{\lambda+\sigma_i^2})^k)^2 + \lambda^2 k \sum_{n=1}^{\infty} \alpha_i^2 (\frac{1}{\sigma_i^2+\lambda})^{2k}$$

[2]Note: where $\alpha = Vx$, $A'A = V\Sigma^2 V'$

Part II

How do we know we can do better with Generalized Ridge?

- Provide a small simulated example with known x, noisy y, and ill-conditioned A
- Provide a real data example where we test our out-of-sample predictions
- Provide an image example where we recover an image

Example 1: Simulated Data

Let's take the Golub matrix. An Ill conditioned matrix.

$$\begin{bmatrix} 1 & 3 & 11 & 0 & -11 & -15 \\ 18 & 55 & 209 & 15 & -198 & -277 \\ -23 & -33 & 144 & 532 & 259 & 82 \\ 9 & 55 & 405 & 437 & -100 & -285 \\ 3 & -4 & -111 & -180 & 39 & 219 \\ -13 & -9 & 202 & 346 & 401 & 253 \end{bmatrix}$$

The singular values are:

s = [9.545e+02, 7.240e+02, 1.767e+02, 7.301e+01, 3.536e-01, 3.169e-10]

So the matrix's condition is $\frac{9.545e+02}{3.169e-10}$, or 3 quadrillion.

Framework of Simulated Problem

Suppose we know the true x = [1, 1, 1, 1, 1, 1].

So we also know true $\bar{y} = A * x$

We add a small normally distributed error to $\bar{y}$, with s.d. 0.1

Now, we would like to recover known x, but with the presence of noise.

Recall, the noiseless solution, x = [1, 1, 1, 1, 1, 1], is:

$$V \begin{bmatrix} \frac{1}{\sigma_1} & & \dots & & 0 & \dots & 0 \\ \vdots & \ddots & & \vdots & & & \\ 0 & \dots & & \frac{1}{\sigma_n} & 0 & \dots & 0 \end{bmatrix} U^T \bar{y}$$

And the naive solution, $(A'A)^{-1}A'(\bar{y}+e)$:

$$V \begin{bmatrix} \frac{1}{\sigma_1} & & \dots & & 0 & \dots & 0 \\ \vdots & \ddots & & \vdots & & & \\ 0 & \dots & & \frac{1}{\sigma_n} & 0 & \dots & 0 \end{bmatrix} U^T (\bar{y}+e)$$

Solutions to Simulated Problem

$\hat{x}$ with OLS, Ridge, and Iterative Ridge:

OLS with noise
$[3.08e{+}08, -1.44e{+}08, 1.12e{+}07, 1.36e{+}06, -9.11e{+}04, -1.49e{+}04]$
$||\hat{x} - x|| = 340,863,251!$

Ridge, $\lambda = 0.1$
$[0.067, 0.24, 1.25, 0.89, 0.88, 1.054]$
$||\hat{x} - x|| = 0.92$

Iterative Ridge, $\lambda = 0.1$, k =5
$[0.08, 0.29, 1.23, 0.89, 0.89, 1.05]$
$||\hat{x} - x|| = 0.68$

λ was chosen through a grid search, and k via the elbow algorithm.

Example 2: Real Data

- Using Body Fat (Carnegie Mellon Data Statistics Library)
- Predict body fat based on 17 variables
- Covariates include: neck, chest, abdomen, hip, thigh, wrist circumference, and hence correlated
- The data are ill conditioned with $\sigma_{17} = .000027$
- We split data into training [202:17] and test [50:17]

Results

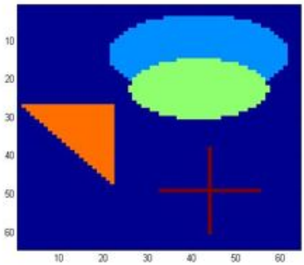
- BLUE unperturbed MSPE = 0.7740
- With added error, N(0,1)[3]. , MSPE = 2.47 a 302% increase.[4]
- Ridge, $\lambda = 100$, k = 1, MSPE = 2.16, about 15% improvement over BLUE.
- Iterative Ridge k = 2 MSP = 1.73, about 30% improvement over BLUE.
- Iterative Ridge $\lambda = 300$, k=5, MSP=1.72
- Iterative Ridge $\lambda = 500$ k=7, MSP= 1.71

[3] std(y) = 7.7509 so that e with std=1 is not excessive

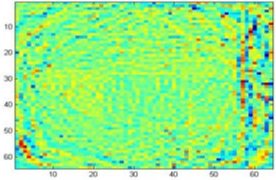
[4] Mean Squared Predicted Error is the norm of predicted and true y.

Example 3: Image Data

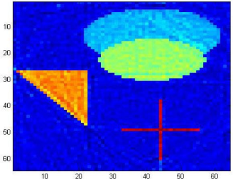
Example of an image that is 64 by 64 square pixels, where A is sparse:



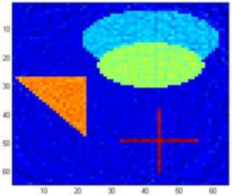
Blue



Ridge, $\lambda = 0.0034$[5]



Iterative Ridge, $\lambda = 1$, $k = 20$



Summary

- Developed a generalization of ridge regression
- The additional parameter k provides more flexibility in balancing bias and noise
- Iterative Ridge performs better when there are number of small singular values, A is sparse, and y is noisy