MAKING BIG DATA A LITTLE BIT SMALLER
Generalizing Regularizing

Katya Vasilaky

Cal Poly

BACKGROUND

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

PURPOSE AND RESULTS

- 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

BACKGROUND

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

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

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

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

RIDGE TIKHANOV

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)[2]

EXTENDED 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:

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

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

EXTENDED RIDGE

(S1)

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

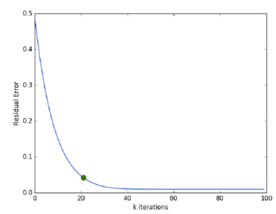
$$\text{(S2) } V\begin{bmatrix}\frac{1}{\sigma_1}-\frac{1}{\sigma_1}(\frac{\lambda}{\lambda+\sigma_1^2})^k & \dots & 0 & 0 & \dots & 0\\ \vdots & \ddots & \vdots & & & \\ 0 & \dots & \frac{1}{\sigma_n}-\frac{1}{\sigma_n}(\frac{\lambda}{\lambda+\sigma_n^2})^k & 0 & \dots & 0\end{bmatrix}U^T e$$

CHOOSING PARAMETERS

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.



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.

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

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 − > ∞
- Second term, noise term, increases monotonically with k (to the OLS residual).

COMPARING FILTERS

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 $\frac{1}{\sigma_i}$

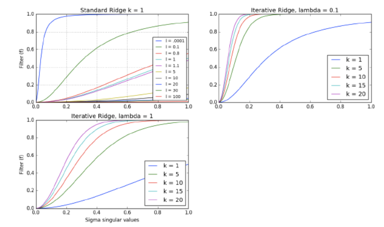
Standard Ridge $\frac{1}{\sigma_i}(1-(\frac{\lambda}{\lambda+\sigma_i^2})) = \frac{\sigma_i^2}{\lambda+\sigma_i^2}$

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

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

MAJOR BENEFITS

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.

Summary, Generalized Iterative Ridge

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

- 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

REFERENCES

References

[1] G. H. Golub, M. Heath, G. Wahba, Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter, Technometrics, Vol. 21, No. 2 (May., 1979).

[2] Arthur E. Hoerl and Robert W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, Vol. 12, No. 1 (Feb., 1970), pp. 55-67.

[3] H. W. Engl, M. Hanke and A. Neubauer, Regularization of Inverse Problems, Kluwer, Dordrecht 1996.

[4] M. Hanke and M. P. C. Hansen, Regularization methods for large-scale problems, Surveys Math. Indust. 3 (1993), pp. 253–315.

[5] G. Wahba, A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem, Ann. Statist. 13 (1985), pp. 1378–1402.

[6] Rafał Zdunek, Multigrid Regularized Image Reconstruction or Limited-Data Tomography, Computational Methods in Science and technology 13(1), 67-77 (2007)

[7] Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. Journal of the Royal Statistical Society. Series B (Methodological), pages 103–106.