STA315 Paper Review: **Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter**

Zoilo Wu, Pengzhou Zhu

12th April, 2025

Authors: Gene H. Golub, Michael Heath, Grace Wahba
Published in: *Technometrics*, Vol. 21, No. 2, May 1979

1. Motivation and Background

In regression modeling scenarios, when there are high-dimensional or large number of predictors, traditional ordinary least squares (OLS) methods become unstable. Since the design matrix X is ill conditioned as demonstrated in the paper, which can increase the variance and result in poor predictions. In addition, when the variables of X are linearly dependent, the result would also be unstable. Ridge regression is a regularization method that addresses these issues by adding a penalty term to shrink the regression coefficients, which would decrease the number of predictors that don't make big effect on the result. Ridge regression solves the modified optimization problem:

$$\hat{\beta}_\lambda = \arg\min_{\beta} \left\{ \|y - X\beta\|^2 + n\lambda\|\beta\|^2 \right\},$$

where $\lambda \geq 0$ is the regularization parameter that balances the bias-variance trade off. However, to use ridge regression accurately, it heavily depends on the appropriate choice of λ. If λ is too small, we overfit the model, because it would be too sensitive to noise. Ifλ is too large, we oversmooth the data which would potentially losing some important predictors. The goal in this paper is how to select the ridge parameter λ in an efficient and stable way.

2. Limitations of Existing Methods

Several methods were proposed before this paper for selecting λ, each with some pros and cons:

(a) PRESS (Prediction Residual Error Sum of Squares)

PRESS is based on leave-one-out cross-validation. For each data point, the model is trained without the selected point, and the prediction error on that point will be calculated and recorded. We choose the λ that can best minimize the sum of squared prediction errors. By definition and from the paper we selected, PRESS method minimizes:

$$P(\lambda) = \frac{1}{n} \sum_{k=1}^{n} \left(\left[X\beta^{(k)}(\lambda) \right]_k - y_k \right)^2$$

Pros of PRESS

- **Intuitively appealing:** PRESS is easy to be understood for beginners.
- **Does not require noise variance σ^2**

Cons of PRESS

- **Not rotation-invariant:** The performance of PRESS depends on the coordinate system which means in near-diagonal or ill-conditioned cases PRESS would behave badly and inconsistent for each replication.
- **Poor performance:** In the paper's simulations, PRESS performed very poorly, especially in low-noise settings. Prediction inefficiency values which is I_R were as high as 2.1×10^5, and estimation inefficiency I_D were as high as 3.84×10^3)

(b) Maximum Likelihood Estimation (MLE)

This method assumes a prior distribution $\beta \sim N(0, \alpha I)$, and is obtained from the model

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

Then the posterior distribution of y is

$$y \sim \mathcal{N}\left(0,\, a\left(XX^T + n\lambda I\right)\right)$$

where $\lambda = \frac{\sigma^2}{na}$. The MLE estimate for λ from the model is then the minimizer of $M(\lambda)$, given by

$$M(\lambda) = \frac{1}{n} \cdot \frac{y^T(I - A(\lambda))y}{[\det(I - A(\lambda))]^{1/n}}$$

Pros of MLE

- **Theoretical optimal if assumptions are correct:** If the prior $\beta \sim \mathcal{N}(0, aI)$ is correct and σ^2 is known, then MLE would provides a good unbiased estimate for λ.

Cons of MLE

- **Sensitive to prior distribution:** If the true regression coefficients β do not follow the prior $\mathcal{N}(0, aI)$, the resulting estimate of λ would be biased.
- **Bad performance in simulations:** In the paper the authors report that in their Monte Carlo experiments, the MLE method often resulted in extremely high estimation errors. For example, in the low-noise case ($\sigma^2 = 10^{-8}$), estimation inefficiency I_D values is 9120.
- **Requires σ^2:** This method depends on the value of σ^2 or needs to be able to estimating the noise variance σ^2, which may be difficult in practice, especially in small-sample case.

(c) Mallows' C_p / Range Risk (RR)

This method needs the value of σ^2 and minimizes:

$$\hat{T}(\lambda) = \frac{1}{n} \|(I - A(\lambda))y\|^2 - \frac{2\hat{\sigma}^2}{n} \operatorname{Tr}(I - A(\lambda)) + \hat{\sigma}^2$$

where $A(\lambda) = X(X^TX + n\lambda I)^{-1}X^T$ is the ridge smoothing matrix.

Pros of RR

- **Good performance (with correct σ^2):** In the authors' simulations (especially for $\sigma^2 = 10^{-4}$), RR performed very well, even better than GCV in some replicates.

Cons of RR

- **Requires knowledge of σ^2:** This is the most significant limitation of RR. In many real applications, σ^2 is unknown.
- **Sensitive to noise level:** If σ^2 is incorrectly estimated, the selected λ would be very biased.

3. The Generalized Cross-Validation Method

To address the issues with the existing methods from the previous section, the authors propose the Generalized Cross-Validation method.

Key Advantages

- No need to estimate σ^2.
- Rotation-invariant: independent of coordinate system.
- Effective when $n \approx p$ or even when $p > n$.

GCV Method Process

$$V(\lambda) = \frac{1}{n} \cdot \frac{\|(I - A(\lambda))y\|^2}{\left(\frac{1}{n}\operatorname{Tr}(I - A(\lambda))\right)^2}$$

where:

- $A(\lambda) = X(X^T X + n\lambda I)^{-1} X^T$ is the ridge hat matrix.
- $y \in R^n$ is the response vector.
- $A(\lambda)y$ is the fitted value.
- $\|(I - A(\lambda))y\|^2$ This measures the distance between the fitted value and the actual value which also implies how well the model fits the data. A small value means better fit to data
- $\operatorname{Tr}(I - A(\lambda))$ is the effective number of parameters left for estimating the error adjusts for model complexity. A larger trace means a more complex model.
- Compute the GCV score:

  $$V(\lambda) = \frac{\text{Adjusted Residual Sum of Squares}}{(\text{Effective Sample Size})^2}$$

  This penalizes models with good fit but high complexity.
- Select the optimal λ: Choose the value of λ that minimizes $V(\lambda)$.

4. Monte Carlo Results

To evaluate the performance of different ridge parameter selected by different methods, the author did a Monte Carlo simulation study. The author repeatedly simulated data under controlled conditions where the true regression parameters are known so that we can easily compare training error and test error.

Setup

- Model used:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

- $n = 21$, $p = 10$
- Design matrix X: Taken from a discretized Laplace transform problem described by Varah (1973). This matrix was intentionally ill-conditioned, with a condition number of approximately 1.54×10^5, to test robustness.
- The value for β was fixed such that $\|X\beta\|^2 = 370.84$.
- Four values of σ^2: $10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}$
- Replications: For each σ^2, the experiment was repeated 4 times (16 runs total).
- Errors: The noise vector ε was generated using pseudo-random samples from a normal distribution $\mathcal{N}(0, \sigma^2)$.
- Metrics: inefficiency in coefficient estimation (I_D), and in predicting the response (I_R)

For each method and simulation, the following two performance metrics were computed:

- Estimation Inefficiency (I_D):

$$I_D = \frac{\|\beta - \hat{\beta}_\lambda\|^2}{\min_\lambda \|\beta - \hat{\beta}_\lambda\|^2}$$

  This measures how far the method's coefficient estimate is from the true β.

- Prediction Inefficiency (I_R):

$$I_R = \frac{T(\lambda)}{\min_\lambda T(\lambda)}, \quad \text{where } T(\lambda) = E\|X\hat{\beta}_\lambda - X\beta\|^2$$

  This measures the method's prediction error.

Numerical Results

We selected the numerical resultes from the paper, and we use the range of the results of the four repeated experiments to review and analyze.
For example, if the value of I_D for GCV from the four experiments are 4.43, 1.65, 16.71, 1.02, we represent these valuse as 1.02 - 16.71.

Table 1: Observed Inefficiencies for $\sigma^2 = 10^{-8}$

Method	I_D (Estimation Error)	I_R (Prediction Error)
GCV	1.02–16.71	1.01–1.10
RR	1.22–8.69	1.00–1.03
MLE	145–9120	1.23–1.53
PRESS	631–3840	4.8×10^4–2.1×10^5
Min Solution	1.00–1.00	1.00–2.27
Min Data	1.00–5.97	1.00–1.00

Table 2: Observed Inefficiencies for $\sigma^2 = 10^{-6}$

Method	I_D (Estimation Error)	I_R (Prediction Error)
GCV	1.32–151	1.00–1.26
RR	1.18–703	1.00–1.10
MLE	149–199	1.19–1.45
PRESS	5.80–241	1.01–607
Min Solution	1.00–1.00	1.02–1.38
Min Data	1.28–41.29	1.00–1.00

Table 3: Observed Inefficiencies for $\sigma^2 = 10^{-4}$

Method	I_D (Estimation Error)	I_R (Prediction Error)
GCV	1.00–1.50	1.00–2.58
RR	1.00–1.18	1.03–2.27
MLE	1.56–12.16	1.07–3.43
PRESS	2.03–8.66	1.57–24.34
Min Solution	1.00–1.00	1.03–2.05
Min Data	1.16–3.26	1.00–1.00

Table 4: Observed Inefficiencies for $\sigma^2 = 10^{-2}$

Method	I_D (Estimation Error)	I_R (Prediction Error)
GCV	1.40–31.20	1.01–17.20
RR	1.38–10.80	1.02–10.60
MLE	2.00–28.80	1.00–16.80
PRESS	1.00–2.16	1.01–21.50
Min Solution	1.00–1.00	1.01–1.98
Min Data	1.00–2.66	1.00–1.00

Table 5: Comparison of Ridge Parameter Selection Methods

Method	Needs σ^2?	Rotation-Invariant	Works in High Dimensions	Theoretical Optimality	Computational Complexity
Existing Methods					
PRESS	No	No	No	No guarantee	Low
Mallows' C_p / Range Risk (RR)	Yes	Yes	No	Finite-sample optimal	Low
MLE	Yes	Yes	No	Prior-dependent	High
New Method					
GCV	No	Yes	Yes	Asymptotically optimal	Moderate

Observations:

- GCV performs well overall, with a relatively low prediction error and a low estimation error.
- RR method is also quite good and even better for prediction error, but it assumes that σ^2 is known.
- MLE fails badly in estimation when the prior is inappropriate.
- PRESS performs extremely especially when $\sigma^2 = 10^{-8}$ because X is often ill-conditioned.
- The results of MLE and PRESS have improved compared to when the $\sigma^2 = 10^{-8}$, but is still less stable than GCV.