Ridge Regression: Structure, Cross-Validation, and Sketching

Sifan Liu[1] Edgar Dobriban[2]

[1]Department of Statistics
Stanford University

[2]Department of Statistics
University of Pennsylvania

ICLR, 2020

Outline

Outline

Overview

- We study ridge regression (ℓ_2-regularized regression) in linear models with large number of parameter and datapoints.

Overview

- We study ridge regression (ℓ_2-regularized regression) in linear models with large number of parameter and datapoints.
- We characterize the structure of the estimator (bias+variance).

Overview

- We study ridge regression (ℓ_2-regularized regression) in linear models with large number of parameter and datapoints.
- We characterize the structure of the estimator (bias+variance).
- We evaluate the bias of cross-validation for choosing the optimal regularization parameter (& correct it).

Overview

- We study ridge regression (ℓ_2-regularized regression) in linear models with large number of parameter and datapoints.
- We characterize the structure of the estimator (bias+variance).
- We evaluate the bias of cross-validation for choosing the optimal regularization parameter (& correct it).
- We study the effectivness of sketching to speed up computation (surprisingly useful)

Ridge regression

- Ridge regression (ℓ_2-regularized regression) is a widely used method for estimation and prediction in high-dimensional data analysis.

Ridge regression

- Ridge regression (ℓ_2-regularized regression) is a widely used method for estimation and prediction in high-dimensional data analysis.
- Consider the data generating model

$$Y = X\beta + \varepsilon, \quad Y \in \mathbb{R}^n, \ X \in \mathbb{R}^{n \times p},$$

where $\varepsilon \in \mathbb{R}^n$ has $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$,

Ridge regression

- Ridge regression (ℓ_2-regularized regression) is a widely used method for estimation and prediction in high-dimensional data analysis.
- Consider the data generating model

$$Y = X\beta + \varepsilon, \quad Y \in \mathbb{R}^n,\ X \in \mathbb{R}^{n\times p},$$

where $\varepsilon \in \mathbb{R}^n$ has $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$,
- Ridge regression solves the optimization problem ($\lambda > 0$),

$$\hat{\beta} = \text{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

The solution has the closed form

$$\hat{\beta} = \left(X^\top X/n + \lambda I_p\right)^{-1} X^\top Y/n.$$

An asymptotic framework

- We assume $p/n \to \gamma \in (0, \infty)$ as $n \to \infty$.

An asymptotic framework

- We assume $p/n \to \gamma \in (0, \infty)$ as $n \to \infty$.
- The empirical spectral distribution (ESD) of a $p \times p$ symmetric matrix Σ is the CDF of its eigenvalues.
  The ESD of the $n \times p$ matrix X is the ESD of the empirical covariance matrix $\hat{\Sigma} = X^\top X/n$.

An asymptotic framework

- We assume $p/n \to \gamma \in (0, \infty)$ as $n \to \infty$.
- The empirical spectral distribution (ESD) of a $p \times p$ symmetric matrix Σ is the CDF of its eigenvalues. The ESD of the $n \times p$ matrix X is the ESD of the empirical covariance matrix $\hat{\Sigma} = X^\top X/n$.
- We say the sequence $\{\Sigma_p\}$ has a limiting spectral distribution (LSD) if the ESD of Σ_p converges weakly to a probability measure. We say $\{X_p\}$ has a limiting spectral distribution if the ESD of $\{\hat{\Sigma}_p\}$ converges weakly to a probability measure, with probability one.

Outline

Structure of ridge estimator I

- Suppose $X = U\Sigma^{1/2}$ has a LSD with compact support bounded away from the origin. Suppose the noise is Gaussian and $\limsup \|\beta\|_2 < \infty$.
- **Theorem.** The ridge estimator has the asymptotic equivalent expression (linear combinations are close)

$$\hat{\beta}(\lambda) \asymp A(\Sigma, \lambda) \cdot \beta + B(\Sigma, \lambda) \cdot \sigma \cdot p^{-1/2} Z.$$

Structure of ridge estimator II

- Here $Z \sim \mathcal{N}(0, I_p)$ is a random vector that is stochastically dependent only on the noise ε, and A, B are deterministic matrices defined by applying the scalar functions below to Σ:

  $$A(x,\lambda) = (c_p x + \lambda)^{-2}(c_p + c_p')x, \ \ B(x,\lambda) = (c_p x + \lambda)^{-1} c_p x.$$

  And $c_p := c(n, p, \Sigma, \lambda)$ is the unique positive solution of the fixed point equation

  $$1 - c_p = \frac{c_p}{n} \operatorname{tr}\left[\Sigma(c_p\Sigma + \lambda I)^{-1}\right].$$

Some comments:

- In particular, we have

$$(\widehat{\Sigma} + \lambda I_p)^{-1} \asymp (c_p \Sigma + \lambda I_p)^{-1}.$$

Some comments:

- In particular, we have

$$(\widehat{\Sigma} + \lambda I_p)^{-1} \asymp (c_p \Sigma + \lambda I_p)^{-1}.$$

- The quantity c_p can be viewed as a *resolvent bias factor*, which tells us by what factor Σ is multiplied when evaluating the resolvent $(\widehat{\Sigma} + \lambda I)^{-1}$, and comparing it to its naive counterpart $(\Sigma + \lambda I)^{-1}$.

Some comments:

- In particular, we have

$$(\widehat{\Sigma} + \lambda I_p)^{-1} \asymp (c_p \Sigma + \lambda I_p)^{-1}.$$

- The quantity c_p can be viewed as a *resolvent bias factor*, which tells us by what factor Σ is multiplied when evaluating the resolvent $(\widehat{\Sigma} + \lambda I)^{-1}$, and comparing it to its naive counterpart $(\Sigma + \lambda I)^{-1}$.
- The quantity c_p' is the derivative of c_p, when viewing it as a function of $z := -\lambda$.

Some comments:

- In particular, we have

$$(\widehat{\Sigma} + \lambda I_p)^{-1} \asymp (c_p \Sigma + \lambda I_p)^{-1}.$$

- The quantity c_p can be viewed as a *resolvent bias factor*, which tells us by what factor Σ is multiplied when evaluating the resolvent $(\widehat{\Sigma} + \lambda I)^{-1}$, and comparing it to its naive counterpart $(\Sigma + \lambda I)^{-1}$.
- The quantity c_p' is the derivative of c_p, when viewing it as a function of $z := -\lambda$.
- For uncorrelated features, $\Sigma = I_p$, A, B reduce to multiplication by scalars.

Estimation and prediction error in random-effect model I

- We consider the *random-effect model*, where β has iid entries with $\mathbb{E}[\beta_i] = 0$, $\mathrm{Var}[\beta_i] = \alpha^2/p$, $i = 1, \ldots, p$ and β is independent of X and ε.

Estimation and prediction error in random-effect model II

Theorem (MSE and training error of ridge)

$$\lim_{n\to\infty} \mathbb{E}\|\hat{\beta} - \beta\|_2^2 = \alpha^2\lambda^2\theta_2 + \gamma\sigma^2[\theta_1 - \lambda\theta_2],$$

$$\lim_{n\to\infty} \mathbb{E}\|Y - X\hat{\beta}\|_2^2 = \alpha^2\lambda^2[\theta_1 - \lambda\theta_2] + \sigma^2\left[1 - \gamma(1 + \lambda\theta_1 - \lambda^2\theta_2)\right].$$

where $\theta_i(\lambda) = \int \frac{1}{(x+\lambda)^i} dF_\gamma(x)$.

Bias-Variance tradeoff

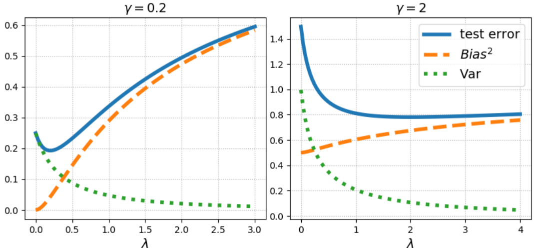


Figure: Ridge regression bias-variance tradeoff. Left: $\gamma = p/n = 0.2$; right: $\gamma = 2$. The data matrix X has iid Gaussian entries. The coefficient β has distribution $\beta \sim \mathcal{N}(0, I_p/p)$, while the noise $\varepsilon \sim \mathcal{N}(0, I_p)$.

- The theorem provides explicit formulas for the bias and variance.

Outline

1. Introduction
2. Representation of the ridge estimator
   - Estimation and prediction error
3. Cross-Validation
4. Sketching

Cross-Validation I

- For K-fold cross-validation, the ridge regression estimator has the form

$$\hat{\beta}_{-k}(\lambda) = \left(X_{-k}^\top X_{-k} + n_1 \lambda I_p\right)^{-1} X_{-k}^\top Y_{-k},$$

  where $n_1 = (K-1)n/K$. The data matrix X_{-k} has aspect ratio $\gamma_1 = \frac{K-1}{K}\gamma$.

Cross-Validation II

- In the random effects model with $\mathbb{E}\beta_i = 0$, $\mathrm{Var}\beta_i = \alpha^2/p$, the minimizer of $\mathbb{E}\widehat{CV}(\lambda)$ tends to $\lambda_k^* = \gamma_1\sigma^2/\alpha^2$.
- Suppose we have found $\hat{\lambda}_k^*$, the minimizer of $\widehat{CV}(\lambda)$ in cross-validation. We propose to use the debiased regularization parameter

$$\hat{\lambda}^* := \hat{\lambda}_k^* \frac{K-1}{K}$$

to refit on the entire dataset, i.e. find

$$\hat{\beta}(\hat{\lambda}^*) = (X^\top X + \hat{\lambda}^* nI)^{-1} X^\top Y.$$

- This bias-correction does not depend on any unknown parameters.

Outline

Sketching

- The computation cost of ridge regression, $O(np\min(n,p))$, can be large. Sketching is an approach to reducing the complexity by reducing the sample size and/or dimension, by random projection or sampling.

Sketching

- The computation cost of ridge regression, $O(np\min(n,p))$, can be large. Sketching is an approach to reducing the complexity by reducing the sample size and/or dimension, by random projection or sampling.
- There are two equivalent formulas for $\hat{\beta}$:

$$\hat{\beta} = \left(X^\top X/n + \lambda I_p\right)^{-1} X^\top Y/n = n^{-1}X^\top \left(XX^\top/n + \lambda I_n\right)^{-1} Y$$

Sketching

- The computation cost of ridge regression, $O(np\min(n,p))$, can be large. Sketching is an approach to reducing the complexity by reducing the sample size and/or dimension, by random projection or sampling.
- There are two equivalent formulas for $\hat\beta$:

$$\hat\beta = \left(X^\top X/n + \lambda I_p\right)^{-1} X^\top Y/n = n^{-1}X^\top \left(XX^\top/n + \lambda I_n\right)^{-1} Y$$

- *Primal sketching*: approximate $X^\top X$ by $X^\top L^\top LX$ for some $m \times n$ sketching matrix L $(m/n \to \xi \in (0,1))$.

Sketching

- The computation cost of ridge regression, $O(np\min(n,p))$, can be large. Sketching is an approach to reducing the complexity by reducing the sample size and/or dimension, by random projection or sampling.
- There are two equivalent formulas for $\hat{\beta}$:

$$\hat{\beta} = \left(X^\top X/n + \lambda I_p\right)^{-1} X^\top Y/n = n^{-1}X^\top \left(XX^\top/n + \lambda I_n\right)^{-1} Y$$

- *Primal sketching*: approximate $X^\top X$ by $X^\top L^\top L X$ for some $m \times n$ sketching matrix L ($m/n \to \xi \in (0,1)$).
- *Dual sketching*: approximate $XX^\top$ by $XRR^\top X^\top$ for some $p \times d$ sketching matrix R ($d/p \to \zeta \in (0,1)$).

Orthogonal sketching

For orthogonal sketching, L and R are partial orthogonal matrices.

Theorem (Orthogonal sketching)

The MSE of primal and dual orthogonal sketching has the limits

$$\alpha^2 \frac{\left[(\lambda+\xi-1)^2+\gamma(1-\xi)\right]\theta_2}{\xi^2} + \gamma\sigma^2 \frac{\xi\theta_1-(\lambda+\xi-1)\theta_2}{\xi^2},$$

$$\frac{\alpha^2}{\gamma}\left[\gamma-1+(\lambda-\gamma+\zeta)^2\bar{\theta}_2+(\gamma-\zeta)\bar{\theta}_1^2\right]+\sigma^2\left[\bar{\theta}_1-(\lambda+\zeta-\gamma)\bar{\theta}_2\right],$$

where

$$\theta_i = \int (x+\lambda)^{-i} dF_\gamma(x), \;\; \bar{\theta}_i = (1-\zeta)/\lambda^i + \zeta \int (x+\lambda)^{-i} dF_\zeta(x),$$

and F_ξ, F_ζ are the standard Marchenko-Pastur laws.

Gaussian sketching

For Gaussian sketching, L and R are random matrices with iid standard normal entries.

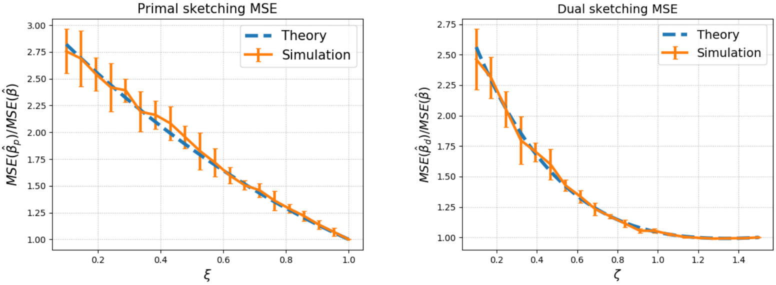
Theorem (Bias of Gaussian dual sketching)

The bias of dual Gaussian sketching equals

$$\mathrm{Bias}^2(\hat{\beta}_d) = \alpha^2 + \alpha^2/\gamma \cdot \left[m'(z) - 2m(z)\right] |_{z=0},$$

where $m^{-1}(z) = 1/[1 + z/\zeta] - [\gamma + 1 - \sqrt{(\gamma-1)^2 + 4\lambda z}]/(2z)$ *for complex* z *with positive imaginary part.*

Simulations



Acknowledgments

The authors thank Ken Clarkson for helpful discussions and for providing the reference [Chen et al.(2015)Chen, Liu, Lyu, King, and Zhang]. ED was partially supported by NSF BIGDATA grant IIS 1837992. SL was partially supported by a Tsinghua University Summer Research award. A version of our manuscript is available on arxiv at `https://arxiv.org/abs/1910.02373`.

References I

T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere.
The million song dataset.
In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

S. Chen, Y. Liu, M. R. Lyu, I. King, and S. Zhang.
Fast relative-error approximation algorithm for ridge regression.
In *UAI*, pages 201–210, 2015.

H. Wickham.
nycflights13: Flights that Departed NYC in 2013, 2018.
URL `https://CRAN.R-project.org/package=nycflights13`.
R package version 1.0.0.

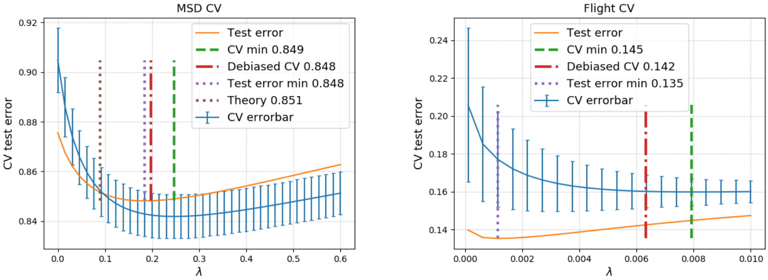


Figure: Left: Cross-validation on the Million Song Dataset [Bertin-Mahieux et al.(2011)Bertin-Mahieux, Ellis, Whitman, and Lamere]. For the error bar, we take $n = 1000$, $p = 90$, $K = 5$, and average over 90 different sub-datasets. For the test error, we train on 1000 training datapoints and fit on 9000 test datapoints. The debiased λ reduces the test error by 0.00024, and the minimal test error is 0.8480. Right: Cross-validation on the flights dataset [Wickham(2018)]. For the error bar, we take $n = 300$, $p = 21$, $K = 5$, and average over 180 different sub-datasets. For the test error, we train on 300 datapoints and fit on 27000 test datapoints. The debiased λ reduces the test error by 0.0022, and the minimal test error is 0.1353.

Comparing different ways of doing cross-validation

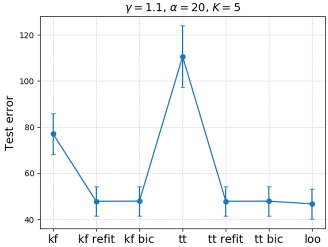


Figure: Comparing different ways of doing cross-validation. We take $n = 500$, $p = 550$, $\alpha = 20$, $\sigma = 1$, $K = 5$. As for train-test validation, we take 80% of samples to be training set and the rest 20% be test set. The error bars are the mean and standard deviation over 20 repetitions.

Bias and variance of orthogonal sketching

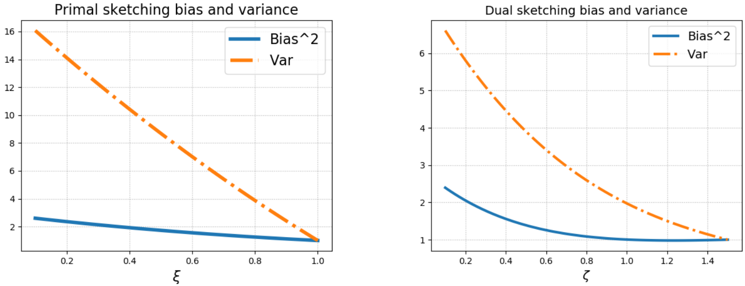


Figure: Bias and variance of primal (left) and dual (right) orthogonal sketching normalized by the bias and variance of ridge regression, respectively.

Bias-variance tradeoff at optimal regularization

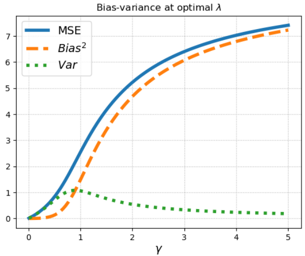


Figure: Bias-variance tradeoff at optimal $\lambda^* = \gamma\sigma^2/\alpha^2$, when $\alpha = 3, \sigma = 1$.

Simulation for dual Gaussian sketching

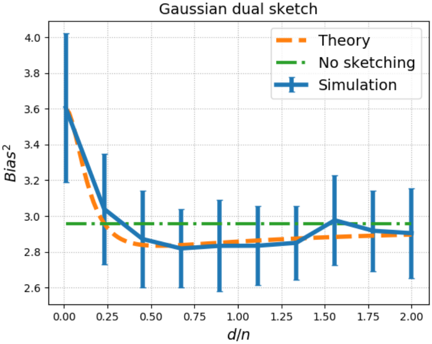


Figure: Gaussian dual sketch when there is no noise, $\gamma = 0.4$, $\alpha = 1$, $\lambda = 1$.