Ridge Regression

Prerequisite concepts

Linear regression and ordinary least squares (OLS), matrix algebra (transpose, inverse, trace), singular value decomposition, convex optimization, and the bias-variance tradeoff.

What you will learn

- Formulate ridge regression as penalized least squares and as a constrained optimization problem.
- Derive the closed-form estimator and interpret shrinkage through a spectral lens.
- Tune the regularization strength and compute ridge solutions efficiently in practice.
- Diagnose under- and over-regularization and interpret ridge coefficients.

Notation and conventions

Notation

- n is the number of observations and p is the number of features.
- $X \in \mathbb{R}^{n\times p}$ is the design matrix and $y \in \mathbb{R}^n$ is the response vector.
- $\beta \in \mathbb{R}^p$ is the coefficient vector and $\hat{\beta}_\lambda$ is the ridge estimator.
- $\lambda \geq 0$ is the ridge regularization parameter.
- I_p and I_n are the identity matrices of sizes p and n.
- $\varepsilon \in \mathbb{R}^n$ is the noise term with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$.
- The SVD of X is $X = USV^T$, where $U \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{p\times p}$ are orthonormal and $S = \mathrm{diag}(s_1, \ldots, s_p)$ with singular values $s_j \geq 0$.
- $A = (X^T X + n\lambda I_p)^{-1} X^T$ is the ridge linear operator mapping y to $\hat{\beta}_\lambda$.

Conventions

- Data shapes and symbol meanings: X has n rows and p columns, and y has length n.
- Centering and standardizing conventions: columns of X and y are centered; features are optionally standardized to unit variance before fitting.
- Intercept treatment: the intercept is handled separately and is not penalized; with centered data, the intercept is zero.
- Objective scaling conventions: we minimize $(1/2n)\|y - X\beta\|_2^2 + (\lambda/2)\|\beta\|_2^2$.
- Mapping to common library parameter names (if relevant): in scikit-learn, `alpha` multiplies $\|\beta\|_2^2$ in $\|y - X\beta\|_2^2 + \texttt{alpha}\|\beta\|_2^2$, so $\texttt{alpha} = n\lambda$ under our scaling.

1 Problem setup and motivation

Ordinary least squares can be unstable when predictors are highly correlated or when p is large relative to n, because X^TX becomes ill-conditioned and small perturbations in y lead to large changes in $\hat{\beta}$. Ridge regression addresses this by shrinking coefficients toward zero, which trades a bit of bias for a large reduction in variance and often improves out-of-sample prediction.

2 General idea

Ridge regression modifies least squares by adding an ℓ_2 penalty on the coefficient vector, producing a unique, well-conditioned solution even when X^TX is nearly singular. The penalty discourages large coefficients, stabilizes estimation along poorly identified directions, and yields a smooth path of solutions as the regularization strength varies.

3 Intuition

Lens 1: Geometric view

Least-squares fits correspond to ellipsoidal contours of the residual sum of squares in β-space. Ridge regression adds a spherical penalty, so the solution becomes the point where a residual contour first touches an ℓ_2 ball, which typically lies closer to the origin than the OLS solution. **Takeaway:** The ridge solution is the closest low-residual point that also stays inside an ℓ_2 ball, so coefficients are shrunk toward zero.

Lens 2: Spectral (SVD) view

Using the SVD from the notation section, decompose the fit along right-singular vectors. OLS divides by singular values, which amplifies noise when singular values are small. Ridge replaces division by s_j with division by $s_j^2 + n\lambda$, softly damping directions with small s_j and leaving well-identified directions nearly unchanged. **Takeaway:** Ridge regression shrinks most strongly along directions where the data are least informative.

Lens 3: Probabilistic or Bayesian view

Assume a Gaussian prior $\beta \sim \mathcal{N}(0, \tau^2 I_p)$ and a Gaussian noise model for y. The maximum a posteriori estimate balances the likelihood and prior, yielding the ridge objective with λ proportional to $\sigma^2/(n\tau^2)$. **Takeaway:** Ridge regression is equivalent to a Gaussian prior that expresses a belief in small coefficients.

4 Formal definition

Let y, X, β, and λ be as defined above, with centered data and an unpenalized intercept handled separately. The ridge estimator is defined by

$$\hat{\beta}_\lambda = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n}\|y - X\beta\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2. \tag{1}$$

Equation (1) states the penalized least-squares objective whose minimizer trades data fit against coefficient size.

$$(X^T X + n\lambda I_p)\hat{\beta}_\lambda = X^T y. \tag{2}$$

Equation (2) is the normal-equation form that characterizes the unique ridge solution when $\lambda > 0$.

5 Key results map

1. **Closed-form ridge estimator** $\hat{\beta}_\lambda = (X^T X + n\lambda I_p)^{-1} X^T y$. (Derived in Appendix A.1.)

2. **Constrained-form equivalence** For each $\lambda > 0$ there exists $t > 0$ such that $\hat{\beta}_\lambda$ solves $\min_{\|\beta\|_2^2 \le t}(1/2n)\|y - X\beta\|_2^2$. (Derived in Appendix A.2.)

3. **Spectral shrinkage** In the SVD basis, ridge scales components by $s_j^2/(s_j^2 + n\lambda)$ relative to OLS. (Derived in Appendix A.3.)

4. **Bias and variance** With A as defined in the notation section, $\mathbb{E}[\hat{\beta}_\lambda] = AX\beta$ and $\mathrm{Var}(\hat{\beta}_\lambda) = \sigma^2 AA^T$. (Derived in Appendix A.4.)

5. **Hat matrix and effective degrees of freedom** $\hat{y} = H_\lambda y$ with $H_\lambda = X(X^T X + n\lambda I_p)^{-1} X^T$ and $\mathrm{df}_\lambda = \mathrm{tr}(H_\lambda)$. (Derived in Appendix A.5.)

6 Estimation, tuning, and computation

Ridge regression is a convex quadratic problem with a unique solution for $\lambda > 0$. In practice, solve the normal equations with a Cholesky factorization when p is moderate, or use QR or SVD for better numerical stability when $X^T X$ is ill-conditioned. For very large or sparse problems, iterative solvers such as conjugate gradient or stochastic gradient methods can be used because the objective is smooth and strongly convex.

Choosing λ is typically done with a validation set or cross-validation; generalized cross-validation can approximate leave-one-out error using the hat matrix. Because the penalty depends on feature scaling, standardize predictors (or use penalty factors) before tuning so the selected λ has consistent meaning.

7 Diagnostics and interpretation

Interpret ridge coefficients as shrunken effects on the standardized scale; back-transform if you need coefficients in original units. Inspect ridge traces (coefficients versus λ) to see which predictors stabilize quickly and which remain unstable. Evaluate predictive error across λ and check residual patterns to confirm that the linear model remains reasonable. The effective degrees of freedom $\operatorname{tr}(H_\lambda)$ provides a measure of model complexity that decreases smoothly as λ grows.

8 Common confusions and failure modes

- **Symptom:** Coefficients change dramatically when features are rescaled. **Cause:** The ℓ_2 penalty is not scale invariant. **Fix:** Center and standardize features or use penalty factors.
- **Symptom:** The intercept is shrunk toward zero. **Cause:** The intercept was included in the penalty or the data were not centered. **Fix:** Fit an unpenalized intercept or center X and y.
- **Symptom:** The model underfits with overly small coefficients. **Cause:** λ is too large. **Fix:** Tune λ with cross-validation and inspect the validation curve.
- **Symptom:** The ridge solution matches OLS but is numerically unstable. **Cause:** λ is effectively zero relative to $X^T X$. **Fix:** Use a positive λ or switch to a robust OLS solver.

9 Connections and extensions

Ridge regression is equivalent to Tikhonov regularization in inverse problems and to a Gaussian-prior Bayesian linear model. It is closely related to principal component regression, but uses soft shrinkage instead of hard truncation. Elastic net blends ridge and lasso penalties, kernel ridge regression replaces X with a kernel matrix, and generalized ridge allows feature-specific penalties through a positive semidefinite penalty matrix.

10 Further reading

Foundational paper

- Hoerl and Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems.” [1]

Best notes or survey

- Vinod, “A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Squares.” [2]

Textbook

- Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning* (2nd ed.). [3]

Implementation docs

- scikit-learn documentation for `Ridge`. [4]

Bibliography

[1] A. E. Hoerl and R. W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, 12(1):55–67, 1970.

[2] H. D. Vinod. "A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Squares." *The Review of Economics and Statistics*, 60(1):121–131, February 1978.

[3] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed., Springer, 2009.

[4] scikit-learn developers. "Ridge" documentation, scikit-learn (stable). `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html`.

A Derivations

A.1 Closed-form ridge solution

We differentiate the penalized objective to obtain the normal equations.

$$f(\beta) = \frac{1}{2n}(y - X\beta)^T(y - X\beta) + \frac{\lambda}{2}\beta^T\beta. \tag{3}$$

This equation restates the ridge objective so we can compute its gradient explicitly.

$$\nabla f(\beta) = \frac{1}{n}(X^T X\beta - X^T y) + \lambda\beta. \tag{4}$$

This gradient expression is needed to find the stationary point of the convex objective.

$$\begin{aligned} \nabla f(\beta) = 0 &\Rightarrow \frac{1}{n}X^T X\beta + \lambda\beta = \frac{1}{n}X^T y \qquad (5)\\ &\Rightarrow (X^T X + n\lambda I_p)\beta = X^T y. \qquad (6) \end{aligned}$$

These equations show the normal equations that characterize the unique minimizer for $\lambda > 0$.

$$\hat{\beta}_\lambda = (X^T X + n\lambda I_p)^{-1} X^T y. \tag{7}$$

This final expression solves the normal equations to give the closed-form ridge estimator.

A.2 Constrained-form equivalence

We show that the penalized and constrained formulations yield the same solution for a suitable constraint level.

$$\min_\beta \frac{1}{2n}\|y - X\beta\|_2^2 \quad \text{subject to} \quad \|\beta\|_2^2 \le t. \tag{8}$$

This constrained problem makes the shrinkage explicit by limiting the ℓ_2 norm of β.

$$\mathcal{L}(\beta, \gamma) = \frac{1}{2n}\|y - X\beta\|_2^2 + \frac{\gamma}{2}(\|\beta\|_2^2 - t). \tag{9}$$

The Lagrangian introduces a multiplier $\gamma \geq 0$ for the norm constraint.

$$\nabla_\beta \mathcal{L}(\beta, \gamma) = \frac{1}{n}(X^T X\beta - X^T y) + \gamma\beta = 0. \tag{10}$$

The stationarity condition matches ridge normal equations with $\lambda = \gamma$.

$$(X^T X + n\gamma I_p)\beta = X^T y. \tag{11}$$

This equation shows that the constrained solution equals the ridge solution for $\lambda = \gamma$ when the constraint is active.

A.3 Spectral shrinkage in the SVD basis

Using the SVD from the notation section, the singular values $s_1, \ldots, s_p$ appear on the diagonal of S.

$$\hat{\beta}_\lambda = (X^T X + n\lambda I_p)^{-1} X^T y \tag{12}$$

$$= (VS^2V^T + n\lambda I_p)^{-1} VSU^T y. \tag{13}$$

This step substitutes the SVD into the closed-form solution to expose shrinkage along singular vectors.

$$(VS^2V^T + n\lambda I_p)^{-1} = V(S^2 + n\lambda I_p)^{-1}V^T. \tag{14}$$

This identity uses orthogonality of V to diagonalize the ridge system.

$$\hat{\beta}_\lambda = V(S^2 + n\lambda I_p)^{-1} SU^T y. \tag{15}$$

This expression gives ridge coefficients in the right-singular-vector basis.

Let $z = U^T y$. The jth coefficient in the V basis is

$$\hat{\theta}_{\lambda,j} = \frac{s_j}{s_j^2 + n\lambda} z_j. \tag{16}$$

This formula shows that ridge scales each component of z by a factor that depends on s_j.

If $\hat{\theta}_{\text{OLS},j} = z_j/s_j$, then

$$\hat{\theta}_{\lambda,j} = \frac{s_j^2}{s_j^2 + n\lambda} \hat{\theta}_{\text{OLS},j}. \tag{17}$$

This equation makes the shrinkage factor $s_j^2/(s_j^2 + n\lambda)$ explicit.

A.4 Bias and variance of the ridge estimator

Using the notation A, the ridge estimator can be written as $\hat{\beta}_\lambda = Ay$.

$$\mathbb{E}[\hat{\beta}_\lambda] = A\mathbb{E}[y] = AX\beta. \tag{18}$$

This equation computes the expectation under the linear model $y = X\beta + \varepsilon$.

$$\mathbb{E}[\hat{\beta}_\lambda] = (X^T X + n\lambda I_p)^{-1} X^T X\beta. \tag{19}$$

This form shows the shrinkage bias relative to the true β.

$$\mathrm{Var}(\hat{\beta}_\lambda) = A \, \mathrm{Var}(y) \, A^T = \sigma^2 A A^T. \tag{20}$$

This equation uses the noise variance assumption to propagate uncertainty through the linear estimator.

$$\mathrm{Var}(\hat{\beta}_\lambda) = \sigma^2 (X^T X + n\lambda I_p)^{-1} X^T X (X^T X + n\lambda I_p)^{-1}. \tag{21}$$

This expression gives the covariance matrix of the ridge estimator.

A.5 Hat matrix and effective degrees of freedom

The fitted values are linear in y.

$$\hat{y} = X\hat{\beta}_\lambda = X(X^T X + n\lambda I_p)^{-1} X^T y. \tag{22}$$

This equation defines the ridge hat matrix H_λ through $\hat{y} = H_\lambda y$.

$$H_\lambda = X(X^T X + n\lambda I_p)^{-1} X^T. \tag{23}$$

This definition is included to characterize ridge as a linear smoother.

$$\mathrm{df}_\lambda = \mathrm{tr}(H_\lambda). \tag{24}$$

This equation defines the effective degrees of freedom as the trace of the hat matrix.