

# Regularization Methods II: Parameter Tuning, Inference, and Extensions

## 1 Parameter Tuning

### 1.1 Review: The Role of the Regularization Parameter

Regularization methods such as LASSO, ridge regression, and elastic net introduce a penalty term controlled by a tuning parameter  $\lambda$ .

- $\lambda$  controls the strength of regularization.
- Larger  $\lambda$  leads to stronger shrinkage and simpler models.
- Smaller  $\lambda$  leads to weaker shrinkage and more complex models.

As  $\lambda$  increases:

- Model complexity decreases
- Bias increases
- Variance decreases

Thus,  $\lambda$  directly governs the **bias–variance tradeoff**.

### 1.2 How Is $\lambda$ Selected in Practice?

In practice,  $\lambda$  is almost always chosen in a data-driven way, most commonly using **cross-validation**.

Two commonly reported values are:

- $\lambda_{\min}$ : the value minimizing cross-validated prediction error
- $\lambda_{1se}$ : the largest  $\lambda$  within one standard error of the minimum

These choices correspond to different modeling goals:

- Prediction accuracy favors smaller  $\lambda$
- Interpretability and sparsity favor larger  $\lambda$

### 1.3 Choice of Metrics for Tuning $\lambda$

Cross-validation provides a general framework for choosing the regularization parameter  $\lambda$ , but it does not specify *what criterion* should be optimized. The choice of metric reflects the modeling goal.

### 1.3.1 Prediction-Oriented Metrics

When the primary goal is prediction,  $\lambda$  is typically chosen to minimize a measure of predictive error, such as:

- Mean squared error (MSE)
- Mean absolute error (MAE)
- Deviance or negative log-likelihood
- Classification error or AUC (for classification problems)

### 1.3.2 Metrics for Variable Selection

Variable selection poses a different challenge. Ideal metrics would assess:

- Recovery of the true active set
- False discovery and false omission rates
- Model sparsity and interpretability

However, these quantities depend on unknown ground truth and involve discrete decisions, making them difficult to optimize directly using cross-validation.

In practice, several heuristic strategies are commonly used:

- Choosing a more conservative value of  $\lambda$ , such as  $\lambda_{1se}$ , i.e., the largest value of  $\lambda$  whose cross-validated error is within one standard error of the minimum, which favors sparser models
- Using information criteria (e.g., AIC, BIC, or extended BIC) to penalize model complexity
- Evaluating stability of selected variables across resampled datasets

These approaches prioritize parsimony and robustness over minimal prediction error, but none provides a universally optimal solution.

## 2 Statistical Inference

### 2.1 Can We Trust Regularized Estimates for Inference?

**Short answer:**

- For prediction: yes
- For inference: not directly

Regularization methods deliberately introduce bias to reduce variance and improve predictive performance. While this bias is beneficial for prediction, it invalidates the assumptions underlying classical statistical inference. This motivates a careful examination of inference under regularization.

## 2.2 Why Regularization Produces Biased Estimates

In classical linear regression, the ordinary least squares (OLS) estimator satisfies the score equation

$$X^\top(y - X\hat{\beta}_{\text{OLS}}) = 0,$$

which implies that, under standard assumptions,

$$\mathbb{E}[\hat{\beta}_{\text{OLS}}] = \beta \quad \text{and} \quad \hat{\beta}_{\text{OLS}} \sim N(\beta, \Sigma).$$

A general regularized estimator can be written as

$$\hat{\beta} = \arg \min_{\beta} \{ \mathcal{L}(y, X\beta) + \lambda \mathcal{P}(\beta) \},$$

where  $\mathcal{P}(\beta)$  is a penalty function and  $\lambda > 0$  controls the strength of regularization.

From a statistical inference perspective, the crucial point is that regularization modifies the estimating equation. Instead of the classical score equation, the estimator satisfies

$$X^\top(y - X\hat{\beta}) = \lambda \partial \mathcal{P}(\hat{\beta}),$$

where  $\partial \mathcal{P}(\hat{\beta})$  denotes the (sub)gradient of the penalty.

Because the right-hand side does not vanish when  $\lambda > 0$ , the estimating equation is no longer centered at zero. Consequently,

$$\mathbb{E}[\hat{\beta}] \neq \beta,$$

and the estimator is biased by construction. This bias is intentional and reflects the tradeoff between variance reduction and shrinkage.

### 2.2.1 LASSO as an Illustrative Example [Reading]

The LASSO estimator is defined as

$$\hat{\beta} = \arg \min_{\beta} \{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \}.$$

The optimal solution satisfies the Karush–Kuhn–Tucker (KKT) conditions. In particular,

$$X^\top(y - X\hat{\beta}) = \lambda z,$$

where

$$z_j \in \begin{cases} \{\text{sign}(\hat{\beta}_j)\}, & \hat{\beta}_j \neq 0, \\ [-1, 1], & \hat{\beta}_j = 0. \end{cases}$$

These conditions reveal two fundamental properties of the LASSO estimator.

**Shrinkage and Bias.** For coefficients with  $\hat{\beta}_j \neq 0$ , the estimating equation includes the penalty term  $\lambda \text{sign}(\hat{\beta}_j)$ , which does not vanish even asymptotically when  $\lambda > 0$ . As a result,

$$\mathbb{E}[\hat{\beta}_j] \neq \beta_j,$$

and the estimator is biased by construction.

**Variable Selection.** For coefficients satisfying

$$|X_j^\top (y - X\hat{\beta})| \leq \lambda,$$

the KKT conditions force  $\hat{\beta}_j = 0$ . This thresholding behavior leads to exact sparsity and makes variable selection an intrinsic part of estimation.

### 2.2.2 Implications for Statistical Inference

Regularization modifies the estimating equations underlying classical linear regression. This has fundamental consequences for statistical inference.

Classical inference relies on several key assumptions:

1. Estimators are (asymptotically) unbiased,
2. The model is fixed and pre-specified,
3. Sampling distributions are approximately Gaussian.

Regularized estimators, particularly sparse methods such as LASSO, violate all three assumptions. First, the penalty term introduces shrinkage bias, so that

$$\mathbb{E}[\hat{\beta}] \neq \beta.$$

As a result, classical variance formulas are no longer valid, since regularized estimators do not admit a simple linear representation of the form

$$\hat{\beta} = \beta + (X^\top X)^{-1} X^\top \varepsilon.$$

Second, sparse regularization methods induce data-dependent variable selection. Whether a coefficient is estimated as exactly zero depends on the realized noise in the data and the choice of the tuning parameter  $\lambda$ . Consequently, the selected model itself is a random object rather than a fixed entity.

Third, the non-smooth nature of the  $\ell_1$  penalty leads to distorted and non-Gaussian sampling distributions, particularly near zero. Classical inferential procedures implicitly assume smoothness and fixed model structure, and therefore fail in this setting.

As a result, naive post-regularization standard errors and p-values are generally invalid.

## 2.3 Remedies: Correcting Bias and Accounting for Selection

The inferential challenges introduced by regularization motivate the development of methods that either correct the bias of regularized estimators or explicitly account for the data-dependent selection process. We first discuss approaches that modify the estimator itself, and then methods that address inference after variable selection.

## 2.4 Methods that fixing the estimator

### 2.4.1 Adaptive LASSO

Adaptive LASSO modifies the standard  $\ell_1$  penalty by assigning variable-specific weights:

$$\hat{\beta}^{\text{adapt}} = \arg \min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^p w_j |\beta_j| \right\}, \quad w_j = \frac{1}{|\tilde{\beta}_j|^\gamma},$$

where  $\tilde{\beta}$  is an initial estimator and  $\gamma > 0$ .

By penalizing small coefficients more heavily and large coefficients less, adaptive LASSO reduces shrinkage bias for strong signals and improves variable selection consistency. However, the resulting estimator remains biased and is not designed to provide valid standard errors or hypothesis tests.

### 2.4.2 Debiased (Desparsified) LASSO

Debiased LASSO explicitly corrects the shrinkage bias introduced by regularization. Starting from the LASSO estimator, a bias-correction term is added to recover asymptotic normality:

$$\hat{\beta}^{\text{debias}} = \hat{\beta}^{\text{lasso}} + \text{bias correction}.$$

Under suitable regularity conditions, each component of the debiased estimator satisfies

$$\sqrt{n} \left( \hat{\beta}_j^{\text{debias}} - \beta_j \right) \xrightarrow{d} N(0, \sigma_j^2),$$

which enables the construction of confidence intervals and hypothesis tests. This correction typically increases variance and relies on strong assumptions about sparsity and the design matrix.

## 2.5 Methods addressing the consequences of selection

### 2.5.1 Selective Inference (post-hoc inference)

Selective inference addresses a different problem: valid inference *conditional on variable selection*. Rather than modifying the estimator, selective inference conditions on the selection event itself and derives the sampling distribution of test statistics given that a variable was selected by a particular procedure (e.g., LASSO at a fixed  $\lambda$ ).

**Post-LASSO Inference as Selective Inference (by Tibshirani).** For a fixed tuning parameter  $\lambda$ , the event that LASSO selects a particular active set and sign pattern can be expressed as a set of linear inequalities in the response  $y$ , i.e.,  $Ay \leq b$ . Conditional on this selection event, the sampling distribution of test statistics is no longer Gaussian but follows a truncated normal distribution. By deriving p-values and confidence intervals from this conditional distribution, post-LASSO inference provides valid finite-sample inference that properly accounts for data-dependent selection. The resulting inference is conditional on the selection event and should be interpreted accordingly.

### 2.5.2 Stability selection (by Meinshausen & Bühlmann)

Stability selection is a resampling-based approach designed to assess the robustness of variable selection under regularization. Rather than performing formal statistical inference, it focuses on identifying variables that are consistently selected across repeated perturbations of the data.

#### How It Works (Algorithmically).

1. Subsample the data repeatedly.
2. Apply a variable selection method (e.g., LASSO) to each subsample.
3. Record how often each variable is selected.

4. Retain variables that are selected with high probability.

For each variable  $j$ , stability selection estimates the selection probability

$$\hat{\Pi}_j = \mathbb{P}(\text{variable } j \text{ is selected}),$$

which measures how robustly the variable is chosen across subsamples.

It outputs selection probabilities rather than p-values and emphasizes robustness over coefficient-level inference.

### 3 Bayesian Interpretation of Regularization Methods

Regularization is not merely an optimization technique. From a Bayesian perspective, it corresponds to *implicit prior beliefs* about the parameters.

#### 3.1 Big Idea: Penalized Likelihood as MAP Estimation

Consider a penalized likelihood estimator of the form

$$\hat{\beta} = \arg \min_{\beta} \{-\log p(y | X, \beta) + \lambda \Omega(\beta)\}.$$

This optimization problem is equivalent to maximum a posteriori (MAP) estimation in a Bayesian model:

$$\hat{\beta}_{\text{MAP}} = \arg \max_{\beta} \{\log p(y | X, \beta) + \log p(\beta)\}.$$

The identification is given by

$$\Omega(\beta) \longleftrightarrow -\log p(\beta),$$

so that the penalty function corresponds to the negative log-prior density.

**Penalty = negative log-prior.**

Thus, regularization imposes prior beliefs on the coefficients, even when framed as a frequentist optimization problem.

#### 3.2 Example: LASSO and the Laplace Prior

The LASSO estimator solves the penalized optimization problem

$$\hat{\beta}^{\text{lasso}} = \arg \min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}.$$

From a Bayesian perspective, this estimator corresponds to maximum a posteriori (MAP) estimation under a Laplace (double-exponential) prior. Specifically, assume the linear model

$$y | X, \beta \sim N(X\beta, \sigma^2 I),$$

together with independent Laplace priors on the coefficients:

$$\beta_j \sim \text{Laplace}(0, b), \quad p(\beta_j) \propto \exp\left(-\frac{|\beta_j|}{b}\right).$$

The log-prior density is proportional to

$$-\frac{1}{b} \sum_{j=1}^p |\beta_j|,$$

which corresponds exactly to the  $\ell_1$  penalty in the LASSO objective. Identifying terms yields

$$\lambda = \frac{1}{b}.$$

### 3.2.1 Interpretation

The Laplace prior has several distinctive features:

- A sharp peak at zero
- Heavy tails compared to a Gaussian
- Strong encouragement of exact sparsity

This explains why LASSO produces exact zeros in the estimated coefficients: the prior mass concentrates near zero, making sparse solutions a priori plausible.

### 3.3 Regularization Is Not Fully Bayesian

Regularization methods provides only a point estimation via MAP:

- No posterior distribution
- No uncertainty quantification

Fully Bayesian sparse models (e.g., spike-and-slab, horseshoe priors) explicitly model uncertainty and avoid hard selection.

## References

- Zou, H. (2006). *The adaptive lasso and its oracle properties*. Journal of the American Statistical Association, 101(476), 1418–1429.  
(Adaptive LASSO: reduced shrinkage bias and improved variable selection consistency.)
- Park, T. and Casella, G. (2008). *The Bayesian lasso*. Journal of the American Statistical Association, 103(482), 681–686.  
(Bayesian interpretation of LASSO via Laplace priors and MAP estimation.)
- Griffin, J. E. and Brown, P. J. (2010). *Inference with normal-gamma prior distributions in regression problems*. Bayesian Analysis, 5(1), 171–188.  
(Bayesian shrinkage priors linking regularization to adaptive sparsity.)
- van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). *On asymptotically optimal confidence regions and tests for high-dimensional models*. The Annals of Statistics, 42(3), 1166–1202.  
(Debiased LASSO: asymptotically valid confidence intervals under regularization.)

- Javanmard, A. and Montanari, A. (2014). *Confidence intervals and hypothesis testing for high-dimensional regression*. Journal of Machine Learning Research, 15, 2869–2909.  
(Debiased LASSO: inference by correcting shrinkage bias.)
- Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). *Exact post-selection inference, with application to the lasso*. The Annals of Statistics, 44(3), 907–927.  
(Selective inference: exact conditional inference after LASSO selection.)
- Tibshirani, R., Taylor, J., Lockhart, R., and Tibshirani, R. J. (2014). *Exact post-selection inference for sequential regression procedures*. Journal of the American Statistical Association, 109(508), 495–504.  
(Post-hoc inference framework for LASSO and related procedures.)
- Meinshausen, N. and Bühlmann, P. (2010). *Stability selection*. Journal of the Royal Statistical Society: Series B, 72(4), 417–473.  
(Stability selection: robustness-based variable selection under regularization.)

## Extended Reading

- Neyshabur, B., Tomioka, R., and Srebro, N. (2015). *In search of the real inductive bias*. Proceedings of the International Conference on Learning Representations (ICLR) Workshop.  
(Implicit regularization and inductive bias in modern machine learning.)
- Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). *Understanding deep learning requires rethinking generalization*. Proceedings of the International Conference on Learning Representations (ICLR).  
(Overparameterization and the role of implicit regularization in deep learning.)