Week 7: Model Selection and Information Criteria

Francisco Richter and Ernst Wit

Introduction to Data Science (MSc)

1 Introduction to Model Selection

Model selection represents one of the most fundamental challenges in statistical learning and data science, addressing the critical question of how to choose among competing models that vary in complexity and explanatory power. The theoretical foundation of model selection rests on the recognition that the goal of statistical modeling extends beyond merely fitting the observed data to encompass the broader objective of making accurate predictions on future, unseen observations.

The central tension in model selection emerges from the bias-variance tradeoff, a fundamental principle that governs the behavior of all statistical estimators. Simple models tend to exhibit high bias but low variance, potentially failing to capture important patterns in the data (underfitting). Complex models, conversely, tend to have low bias but high variance, potentially capturing noise rather than signal (overfitting). The optimal model achieves the best balance between these competing sources of error.

Information criteria provide a principled framework for navigating this tradeoff by explicitly penalizing model complexity while rewarding goodness of fit. These criteria emerge from deep connections to information theory, Bayesian inference, and asymptotic statistical theory, providing both theoretical justification and practical guidance for model selection decisions.

2 Mathematical Framework for Model Selection

2.1 The Prediction Error Decomposition

The theoretical foundation of model selection begins with a precise characterization of prediction error and its components. For a target variable Y and a prediction function $\hat{f}(x)$ estimated from training data, the prediction error at a new point x can be decomposed into fundamental components that illuminate the sources of model performance.

Theorem 2.1 (Bias-Variance-Noise Decomposition)**.** *For a target variable $Y = f(x) + \epsilon$ where $E[\epsilon] = 0$ and $Var(\epsilon) = \sigma^2$, and a prediction $\hat{f}(x)$ based on training data, the expected squared prediction error is:*

$$E[(Y - \hat{f}(x))^2] = Bias^2[\hat{f}(x)] + Var[\hat{f}(x)] + \sigma^2$$

where:

$$Bias[\hat{f}(x)] = E[\hat{f}(x)] - f(x) \tag{1}$$

$$Var[\hat{f}(x)] = E[(\hat{f}(x) - E[\hat{f}(x)])^2] \tag{2}$$

$$\sigma^2 = E[\epsilon^2] \text{ (irreducible error)} \tag{3}$$

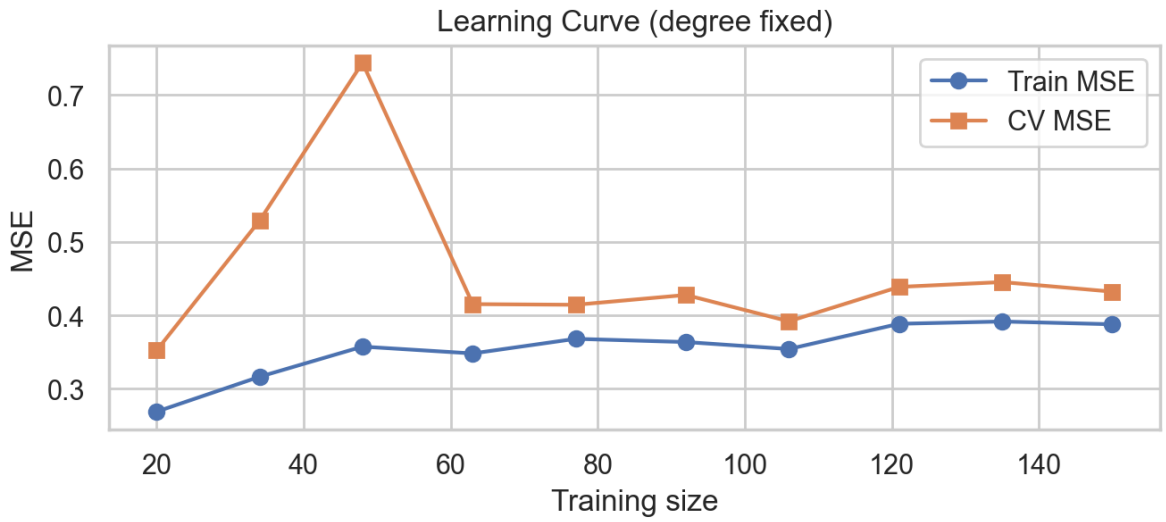

Figure 1: Learning curve for a fixed model degree: training and CV errors versus training set size.

This decomposition reveals that prediction error consists of three distinct components: bias arising from systematic deviations of the expected prediction from the true function, variance arising from sensitivity to particular training samples, and irreducible noise inherent in the data generating process. The bias and variance components are under the control of the model selection process, while the noise component represents a fundamental limit on achievable performance.

The bias-variance tradeoff manifests as an inverse relationship between these two components as model complexity varies. Simple models with few parameters tend to have high bias because they cannot capture complex patterns, but low variance because their predictions are relatively stable across different training samples. Complex models with many parameters tend to have low bias because they can approximate complex functions, but high variance because they are sensitive to particular features of the training data.

2.2 Generalization and the True Risk

The fundamental goal of model selection is to minimize the true risk, defined as the expected loss over the entire population from which the training data are drawn. For squared error loss, the true risk is:

$$R(f) = E_{X,Y}[(Y - f(X))^2]$$

Since the true risk cannot be computed directly (as it requires knowledge of the population distribution), model selection procedures must estimate this quantity from the available training data. The challenge lies in the fact that the training error systematically underestimates the true risk, particularly for complex models that can fit the training data closely.

The optimism of the training error, defined as the difference between the true risk and the training error, increases with model complexity. Information criteria address this challenge by adding penalty terms that estimate this optimism, providing approximately unbiased estimates of the true risk.

3 Information-Theoretic Foundations

3.1 Kullback-Leibler Divergence and Model Distance

The theoretical foundation of information criteria rests on the concept of Kullback-Leibler (KL) divergence, which provides a natural measure of the distance between probability distributions. In the context of model selection, KL divergence quantifies how well a candidate model approximates the true data generating process.

Definition 3.1 (Kullback-Leibler Divergence)**.** *For probability densities $f(x)$ and $g(x)$, the Kullback-Leibler divergence from g to f is:*

$$D_{KL}(f||g) = \int f(x) \log \frac{f(x)}{g(x)} dx = E_f \left[\log \frac{f(X)}{g(X)} \right]$$

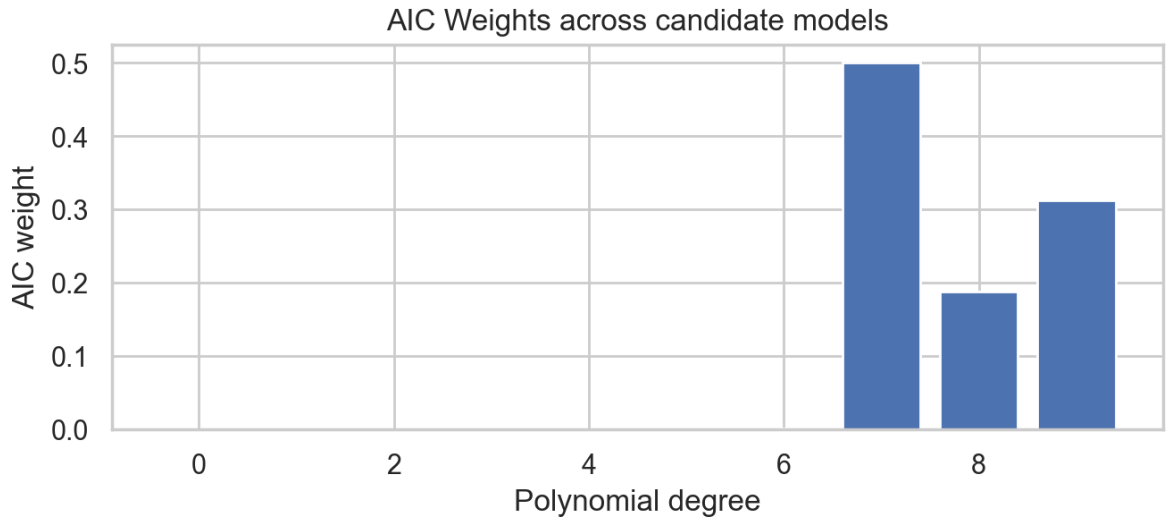

Figure 2: AIC weights assigned to each candidate polynomial degree; weights sum to 1.

The KL divergence is always non-negative and equals zero if and only if $f = g$ almost everywhere. It is not symmetric, reflecting the fact that the divergence from the true model to a candidate model differs from the divergence in the opposite direction. In model selection, we are interested in the divergence from the true model to candidate models.

The expected KL divergence between the true model f and a fitted model $g_{\hat{\theta}}$ can be decomposed as:

$$E[D_{KL}(f||g_{\hat{\theta}})] = E\left[\int f(x) \log f(x) dx \right] - E\left[\int f(x) \log g_{\hat{\theta}}(x) dx \right]$$

The first term is constant across models and can be ignored for model selection purposes. The second term, the expected log-likelihood of the fitted model evaluated at the true distribution, becomes the focus of information criteria.

3.2 Akaike Information Criterion: Derivation and Properties

The Akaike Information Criterion emerges from an asymptotic analysis of the expected KL divergence between the true model and a fitted parametric model. Akaike's fundamental insight was

that the maximum log-likelihood provides a biased estimate of the expected log-likelihood under the true model, with the bias depending on the number of parameters.

Theorem 3.1 (AIC Derivation)**.** *Under regularity conditions, for a parametric model with k parameters, the expected KL divergence between the true model and the fitted model is asymptotically:*

$$E[D_{KL}(f||g_{\hat{\theta}})] \approx -E[\log L(\hat{\theta})] + k + \text{constant}$$

where $L(\hat{\theta})$ is the maximized likelihood. This leads to the AIC:

$$AIC = -2\log L(\hat{\theta}) + 2k$$

The factor of 2 in the penalty term arises from the asymptotic analysis and reflects the expected bias in the log-likelihood as an estimator of the expected log-likelihood under the true model. The AIC provides an approximately unbiased estimator of the expected KL divergence, making it suitable for selecting models that minimize prediction error.

The asymptotic properties of AIC include efficiency in the sense that it selects the model that minimizes prediction error among the candidate models. However, AIC is not consistent, meaning that it may not select the true model even with infinite data if the true model is among the candidates. This reflects AIC's focus on prediction rather than model identification.

3.3 Bayesian Information Criterion: Consistency and Model Identification

The Bayesian Information Criterion takes a different approach to model selection, emerging from Bayesian model comparison and the principle of selecting the model with the highest posterior probability. Under certain conditions, BIC provides a consistent model selection criterion.

Theorem 3.2 (BIC Derivation and Consistency)**.** *Under regularity conditions, the BIC is defined as:*

$$BIC = -2\log L(\hat{\theta}) + k\log(n)$$

where n is the sample size. If the true model is among the candidates, BIC selects the true model with probability approaching 1 as $n \to \infty$.

The $\log(n)$ penalty in BIC grows with sample size, making it more conservative than AIC for large samples. This stronger penalty reflects BIC's focus on model identification rather than prediction. The consistency property means that BIC will eventually identify the correct model structure if it exists among the candidates, making it particularly valuable when the goal is scientific understanding rather than pure prediction.

The relationship between BIC and Bayesian model selection becomes clear when considering the marginal likelihood of each model. Under uniform priors on model parameters and equal prior probabilities for models, BIC approximates twice the negative log marginal likelihood, making model selection equivalent to choosing the model with the highest posterior probability.

4 Cross-Validation: Theory and Practice

4.1 Mathematical Foundation of Cross-Validation

Cross-validation provides a direct approach to estimating prediction error by using data splitting to simulate the prediction of new observations. The theoretical foundation rests on the principle that prediction error should be estimated using data that were not used for model fitting.

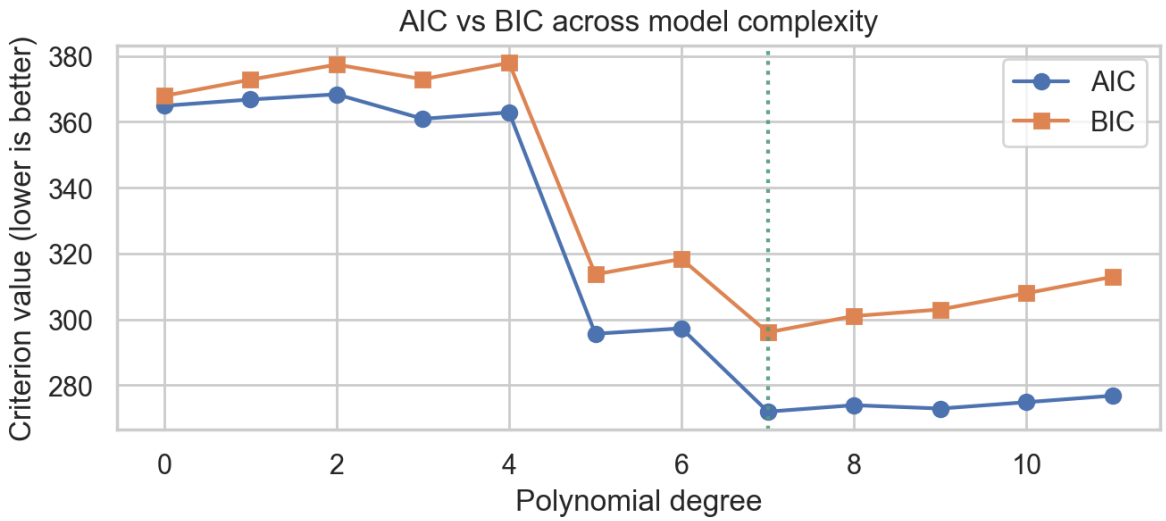

Figure 3: AIC and BIC across polynomial degrees. Vertical lines indicate the minimizing degrees for each criterion.

Definition 4.1 (K-Fold Cross-Validation)**.** *The data* $\{(x_i, y_i)\}_{i=1}^n$ *are randomly partitioned into* K *approximately equal-sized folds* $C_1, C_2, \ldots, C_K$*. For each fold* k*:*

1. *Fit the model using data from all folds except* C_k*:* $\hat{f}^{(-k)}$
2. *Compute prediction error on fold* k*:* $PE_k = \sum_{i \in C_k} L(y_i, \hat{f}^{(-k)}(x_i))$

The cross-validation estimate is: $CV = \frac{1}{K} \sum_{k=1}^K PE_k$

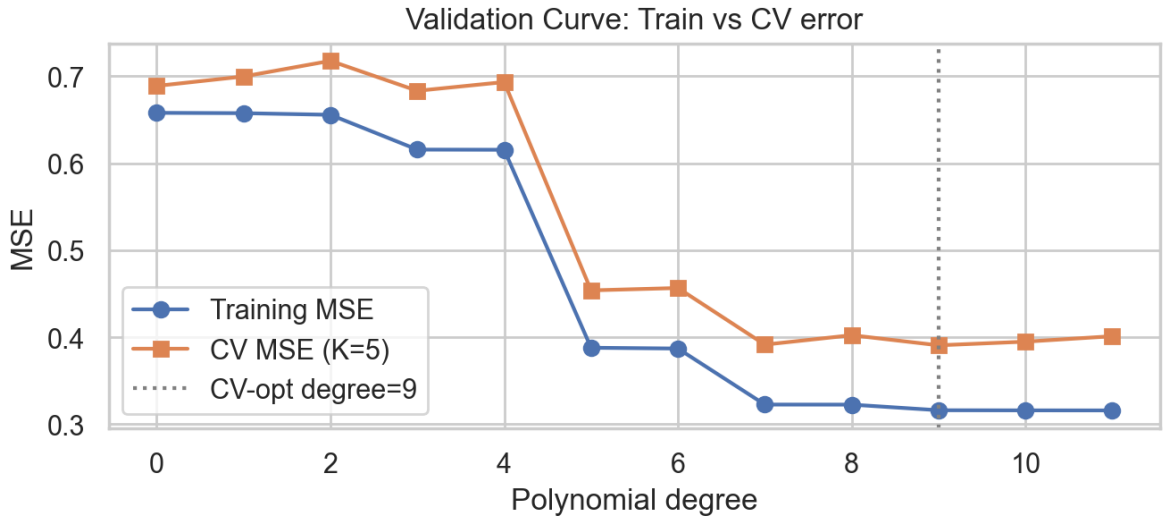

Figure 4: Validation curve across polynomial degrees: training MSE decreases monotonically; CV error shows a U-shape with an optimal degree.

The choice of K involves a bias-variance tradeoff. Larger values of K (approaching leave-one-out) provide nearly unbiased estimates of prediction error but have high variance because the

training sets are highly correlated. Smaller values of K introduce bias because the training sets are smaller than the full dataset, but have lower variance. In practice, $K = 5$ or $K = 10$ often provide good compromises.

4.2 Leave-One-Out Cross-Validation and Computational Shortcuts

Leave-one-out cross-validation (LOOCV) represents the extreme case where $K = n$, providing the most nearly unbiased estimate of prediction error. For linear models, LOOCV can be computed efficiently without actually fitting n separate models.

Theorem 4.1 (LOOCV Formula for Linear Models)**.** *For linear regression with design matrix* $\mathbf{X}$ *and hat matrix* $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$*, the LOOCV estimate is:*

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i}{1 - h_{ii}}\right)^2$$

where e_i *are the ordinary residuals and* h_{ii} *are the diagonal elements of the hat matrix.*

This formula reveals that LOOCV can be computed from a single model fit, making it computationally feasible even for large datasets. The denominator $(1-h_{ii})$ adjusts for the leverage of each observation, with high-leverage points receiving larger weights in the cross-validation estimate.

The relationship between LOOCV and information criteria provides additional theoretical insight. For linear models with normal errors, LOOCV is asymptotically equivalent to AIC, providing a connection between the information-theoretic and cross-validation approaches to model selection.

5 Regularization and Penalized Likelihood Methods

5.1 Ridge Regression: Bias-Variance Tradeoff

Regularization methods address the model selection problem by fitting a single model with a penalty term that controls complexity. Ridge regression adds a quadratic penalty to the least squares objective, shrinking coefficient estimates toward zero.

Definition 5.1 (Ridge Regression)**.** *Ridge regression minimizes the penalized sum of squares:*

$$RSS(\lambda) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

The solution is: $\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$

The regularization parameter λ controls the strength of the penalty, with larger values producing more shrinkage. The bias-variance tradeoff manifests as λ varies: small values of λ produce estimates with low bias but high variance, while large values produce estimates with high bias but low variance.

The ridge regression estimator can be expressed in terms of the singular value decomposition of the design matrix, revealing that ridge regression shrinks the coefficients corresponding to smaller singular values more heavily. This selective shrinkage is particularly beneficial when the design matrix is ill-conditioned or when there are many correlated predictors.

5.2 LASSO: Sparsity and Feature Selection

The Least Absolute Shrinkage and Selection Operator (LASSO) replaces the quadratic penalty of ridge regression with an L_1 penalty, inducing sparsity in the coefficient estimates.

Definition 5.2 (LASSO)**.** *LASSO minimizes the penalized sum of squares:*

$$RSS(\lambda) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

The L_1 penalty has the unique property of producing exactly zero coefficient estimates for sufficiently large λ, effectively performing automatic feature selection. This sparsity property makes LASSO particularly valuable in high-dimensional settings where interpretability is important.

The LASSO solution path, showing how coefficient estimates vary with λ, provides insights into the relative importance of different predictors. Variables that remain in the model for small values of λ are typically the most important, while variables that are quickly eliminated as λ increases are less relevant for prediction.

5.3 Elastic Net: Combining Ridge and LASSO

The elastic net combines the ridge and LASSO penalties, addressing some limitations of each method when used alone.

Definition 5.3 (Elastic Net)**.** *The elastic net minimizes:*

$$RSS(\lambda_1, \lambda_2) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2$$

The elastic net addresses the limitation of LASSO in selecting at most n variables when $p > n$, and its tendency to select only one variable from groups of highly correlated variables. The ridge component encourages the selection of groups of correlated variables, while the LASSO component maintains sparsity.

6 Advanced Model Selection Strategies

6.1 Forward Selection and Backward Elimination

Stepwise selection procedures provide computationally efficient approaches to model selection when the number of potential predictors is large. These methods build or reduce models sequentially, using statistical criteria to guide the selection process.

Forward selection begins with the null model and adds variables sequentially based on their contribution to model fit. At each step, the variable that most improves the selection criterion (such as AIC or BIC) is added to the model. The process continues until no remaining variable improves the criterion.

Backward elimination begins with the full model and removes variables sequentially. At each step, the variable whose removal least worsens the selection criterion is eliminated. The process continues until removing any remaining variable would significantly worsen the criterion.

Stepwise selection combines forward selection and backward elimination, allowing variables to be added or removed at each step. This flexibility can lead to better final models but increases computational complexity and the risk of overfitting to the particular dataset.

6.2 Best Subset Selection

Best subset selection considers all possible subsets of predictors and selects the one that optimizes the chosen criterion. While computationally intensive for large p, this approach guarantees finding the globally optimal subset for any given criterion.

For p predictors, there are 2^p possible subsets, making exhaustive search feasible only for moderate values of p (typically $p \leq 20$). For larger problems, approximate algorithms such as genetic algorithms or simulated annealing can be used to search the space of possible models.

The theoretical appeal of best subset selection lies in its guarantee of finding the optimal model according to the specified criterion. However, this optimality comes at the cost of increased computational complexity and potential overfitting, particularly when the number of candidate models is large relative to the sample size.

7 Model Averaging and Uncertainty Quantification

7.1 Bayesian Model Averaging

Rather than selecting a single "best" model, model averaging acknowledges uncertainty in model selection by combining predictions from multiple models. Bayesian model averaging provides a principled framework for weighting different models based on their posterior probabilities.

Definition 7.1 (AIC Weights)**.** *For models $M_1, \ldots, M_R$ with AIC values $AIC_1, \ldots, AIC_R$, the AIC weights are:*

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R} \exp(-\frac{1}{2}\Delta_r)}$$

where $\Delta_i = AIC_i - \min(AIC_1, \ldots, AIC_R)$.

These weights can be interpreted as approximate posterior model probabilities under certain conditions. Models with lower AIC values receive higher weights, but the exponential transformation ensures that models with similar AIC values receive similar weights, reflecting uncertainty in model selection.

Model averaging often provides better predictive performance than selecting a single model, particularly when several models have similar performance. The averaging process reduces the variance of predictions while potentially increasing bias, representing another manifestation of the bias-variance tradeoff.

7.2 Information-Theoretic Model Selection

The information-theoretic approach to model selection extends beyond AIC and BIC to encompass a broader framework based on information theory and minimum description length principles. The fundamental insight is that the best model is the one that provides the most compressed representation of the data.

The Minimum Description Length (MDL) principle formalizes this intuition by seeking the model that minimizes the total description length of the model and the data given the model. This approach provides a unified framework that encompasses both Bayesian and frequentist perspectives on model selection.

The connection between information criteria and coding theory reveals deep relationships between statistical inference and information transmission. Models that compress data efficiently correspond to models that capture the underlying structure of the data generating process, providing a fundamental justification for information-based model selection.

8 Applications in Modern Data Science

8.1 High-Dimensional Model Selection

Modern data science applications often involve high-dimensional settings where the number of predictors p is comparable to or larger than the sample size n. Traditional model selection methods may not be applicable in these settings, requiring specialized approaches that account for the curse of dimensionality.

Regularization methods such as LASSO become essential in high-dimensional settings, providing automatic feature selection while maintaining computational tractability. The theoretical properties of LASSO in high-dimensional settings have been extensively studied, with results showing that LASSO can achieve near-optimal performance under appropriate sparsity assumptions.

The sure independence screening (SIS) procedure provides a two-stage approach to high-dimensional model selection, first reducing the dimensionality through marginal screening and then applying traditional model selection methods to the reduced set of variables. This approach can handle ultra-high-dimensional settings where p grows exponentially with n.

8.2 Machine Learning Model Selection

In machine learning applications, model selection extends beyond choosing variables to include selecting among different algorithms, architectures, and hyperparameters. Cross-validation becomes the primary tool for model selection, as information criteria may not be applicable to non-parametric methods.

Grid search and random search provide systematic approaches to hyperparameter tuning, using cross-validation to evaluate different parameter combinations. Bayesian optimization offers a more sophisticated approach that uses probabilistic models to guide the search for optimal hyperparameters.

The concept of model selection extends to ensemble methods, where the goal is to select and combine multiple base learners to improve predictive performance. Techniques such as stacking and blending use cross-validation to learn optimal combination weights, representing a form of model averaging adapted to machine learning contexts.

8.3 Causal Model Selection

When the goal is causal inference rather than prediction, model selection criteria must be adapted to account for confounding and selection bias. Traditional prediction-focused criteria may select models that achieve good predictive performance but fail to identify causal relationships.

Directed acyclic graphs (DAGs) provide a framework for representing causal assumptions and guiding model selection for causal inference. The backdoor criterion and other causal identification strategies determine which variables must be included in the model to achieve valid causal inference, potentially overriding purely statistical selection criteria.

Instrumental variable methods and other causal inference techniques require specialized model selection approaches that account for the identifying assumptions underlying causal inference. These methods illustrate the importance of aligning model selection criteria with the ultimate goals of the analysis.

9 Summary

Model selection represents a fundamental challenge in statistical learning that requires balancing goodness of fit with model complexity to achieve optimal predictive performance. The theoretical foundations of model selection rest on information theory, Bayesian inference, and asymptotic statistical theory, providing principled approaches to this challenging problem.

Information criteria such as AIC and BIC provide computationally efficient approaches to model selection that are grounded in deep theoretical principles. AIC focuses on prediction and provides asymptotically efficient model selection, while BIC focuses on model identification and provides consistent model selection under appropriate conditions.

Cross-validation offers a direct approach to estimating prediction error that is widely applicable across different types of models and learning algorithms. The bias-variance tradeoff inherent in the choice of cross-validation parameters reflects the fundamental tensions underlying all model selection procedures.

Regularization methods provide an alternative approach that fits a single model with complexity controlled by penalty parameters. These methods are particularly valuable in high-dimensional settings and provide automatic feature selection capabilities that are essential in modern data science applications.

The extension of model selection principles to machine learning, causal inference, and other contemporary applications demonstrates the continued relevance of these fundamental statistical concepts. As data science continues to evolve, the principles underlying model selection remain essential for developing methods that achieve optimal performance while maintaining interpretability and scientific validity.

10 Exercises

1. **Information Criteria Derivation:** Derive the relationship between AIC and the expected Kullback-Leibler divergence. Show that AIC provides an approximately unbiased estimator of $-2E[\log L(\hat{\theta})]$ where the expectation is taken over future data from the true model.

2. **Bias-Variance Decomposition:** For ridge regression with penalty parameter λ, derive expressions for the bias and variance of the coefficient estimates. Show how the bias-variance tradeoff depends on λ and the eigenvalues of $\mathbf{X}^T\mathbf{X}$.

3. **Cross-Validation Theory:** Prove the LOOCV formula for linear regression. Explain why this formula allows LOOCV to be computed from a single model fit and discuss the computational advantages this provides.

4. **LASSO Solution Path:** Describe the algorithm for computing the LASSO solution path as λ varies. Explain why the LASSO produces sparse solutions and how this relates to the geometry of the L_1 penalty.

5. **Model Selection Consistency:** Prove that BIC is consistent for model selection under appropriate regularity conditions. Compare this with the properties of AIC and explain why AIC is not consistent despite being asymptotically efficient.

6. **High-Dimensional Model Selection:** In a setting where $p > n$, explain why traditional model selection methods may fail and describe how regularization methods address these challenges. Discuss the role of sparsity assumptions in high-dimensional inference.

7. **Model Averaging:** Derive the AIC weights formula and explain its interpretation as approximate posterior model probabilities. Compare model averaging with model selection and discuss when each approach is preferable.

8. **Causal Model Selection:** Explain how model selection for causal inference differs from model selection for prediction. Describe the role of directed acyclic graphs in guiding variable selection for causal inference and discuss potential conflicts between statistical and causal selection criteria.