Model Averaging

Statistics · Module 10 — Model Selection Learning

Node 10.6

Position in Knowledge Graph

System: Statistics. Structural Domain: Model Selection and Learning (Module 10). Node 10.6.

Model selection (Nodes 10.2–10.3) commits to a single model, discarding uncertainty about which model generated the data. Model averaging instead combines predictions across all candidates, weighting each by its posterior probability (Bayesian model averaging) or its predictive performance (stacking). This node unifies the Bayesian predictive framework (Node 6.5), Bayes factors (Node 6.6), and information criteria (Node 10.2) into a coherent prediction strategy that accounts for model uncertainty.

Formal Definition

Let $\mathcal{M}_1, \ldots, \mathcal{M}_K$ be candidate models with prior probabilities $\pi(\mathcal{M}_j)$, $\sum_j \pi(\mathcal{M}_j) = 1$.

Bayesian Model Averaging (BMA). The posterior model probability is

$$\pi(\mathcal{M}_j \mid x) = \frac{m_j(x)\,\pi(\mathcal{M}_j)}{\sum_{k=1}^{K} m_k(x)\,\pi(\mathcal{M}_k)},$$

where $m_j(x) = \int p(x \mid \theta^{(j)}, \mathcal{M}_j)\,\pi(\theta^{(j)} \mid \mathcal{M}_j)\,d\theta^{(j)}$ is the marginal likelihood.

The **BMA predictive distribution** is

$$p(\tilde{x} \mid x) = \sum_{j=1}^{K} \pi(\mathcal{M}_j \mid x)\,p(\tilde{x} \mid x, \mathcal{M}_j).$$

Intuition

Model selection bets everything on one model. If that model is wrong, predictions inherit systematic bias. Model averaging hedges: it spreads probability mass across multiple models in proportion to how well each fits the data (weighted by the prior). When model uncertainty is substantial—no single model clearly dominates—averaging delivers better calibrated predictions and more honest uncertainty quantification than selection.

The cost is interpretability: a weighted average of models is harder to explain than a single selected model.

Structural Role

Dependencies: Builds on information criteria (Node 10.2) for BIC-approximated weights, posterior predictive distributions (Node 6.5), and Bayes factors (Node 6.6) for model comparison. **Enables:** This is a terminal node in Module 10.

Model averaging operationalizes the Bayesian principle that inference should integrate over all sources of uncertainty, including uncertainty about model structure. It connects the computational tools of Nodes 10.2–10.3 to the Bayesian prediction framework, unifying model selection and prediction.

Mathematical Development

BMA Optimality Under Proper Scoring Rules

Theorem. Among all predictive distributions of the form $q(\tilde{x}) = \sum_j w_j \, p(\tilde{x} \mid x, \mathcal{M}_j)$ with $w_j \geq 0$, $\sum_j w_j = 1$, the BMA predictive distribution (with $w_j = \pi(\mathcal{M}_j \mid x)$) minimizes the posterior expected loss under any proper scoring rule.

Proof. A scoring rule $S(q, \tilde{x})$ is **proper** if $\mathbb{E}_{\tilde{x} \sim p}[S(q, \tilde{x})]$ is minimized at $q = p$. The posterior expected score is

$$\mathbb{E}\big[S(q, \tilde{X}) \mid x\big] = \sum_j \pi(\mathcal{M}_j \mid x) \, \mathbb{E}_{\tilde{X} \sim p_j}\big[S(q, \tilde{X})\big] \, .$$

For the **log score** $S(q, \tilde{x}) = -\log q(\tilde{x})$:

$$\mathbb{E}\big[-\log q(\tilde{X}) \mid x\big] = \sum_j \pi(\mathcal{M}_j \mid x) \, \mathbb{E}_{p_j}\big[-\log q(\tilde{X})\big] \, .$$

This is a mixture of KL divergences $D_{\mathrm{KL}}(p_j \| q)$ plus entropy terms. The unique minimizer over mixture distributions is the mixture with weights $w_j = \pi(\mathcal{M}_j \mid x)$, since for any other weights w',

$$\sum_j \pi_j D_{\mathrm{KL}}(p_j \| q_{w'}) \geq \sum_j \pi_j D_{\mathrm{KL}}(p_j \| q_{\pi}),$$

by the convexity of KL divergence and the definition of the posterior mixture. The result extends to all proper scoring rules by the characterization theorem for proper scores.

BIC Approximation to BMA Weights

Substituting the BIC approximation $\log m_j(x) \approx \ell_j - (k_j/2) \log n$ into the posterior model weights:

$$\pi(\mathcal{M}_j \mid x) \approx \frac{\exp(-\mathrm{BIC}_j/2) \, \pi(\mathcal{M}_j)}{\sum_k \exp(-\mathrm{BIC}_k/2) \, \pi(\mathcal{M}_k)}.$$

Under equal priors $\pi(\mathcal{M}_j) = 1/K$, the weights simplify to

$$w_j^{\mathrm{BIC}} = \frac{\exp(-\mathrm{BIC}_j/2)}{\sum_k \exp(-\mathrm{BIC}_k/2)}.$$

This provides a practical BMA implementation without computing marginal likelihoods directly.

Asymptotic Concentration

If the true model $\mathcal{M}^*$ is among the candidates, the BMA weights concentrate:

$$\pi(\mathcal{M}^* \mid x) \to 1 \quad \text{a.s. as } n \to \infty.$$

This follows from the consistency of the Bayes factor: $\log(m_j(x)/m^*(x)) \to -\infty$ for $j \neq *$ (Node 6.6). In finite samples, however, multiple models may have non-negligible weights, and BMA outperforms selection by exploiting this uncertainty.

Frequentist Model Averaging and Stacking

In the frequentist framework, model averaging combines estimators with data-adaptive weights:

$$\tilde{f}(x) = \sum_{j=1}^{K} w_j \, \hat{f}_j(x), \quad w_j \geq 0, \quad \sum_j w_j = 1.$$

Mallows Model Averaging (MMA). Select weights minimizing an estimate of prediction error:

$$\hat{w}^{\text{MMA}} = \arg \min_{w \geq 0, \sum w_j = 1} \left[\|Y - \sum_j w_j \hat{Y}_j\|^2 + 2\sigma^2 \sum_j w_j k_j \right],$$

where k_j is the degrees of freedom of model j. The penalty $2\sigma^2 \sum_j w_j k_j$ corrects for the optimism of in-sample fit, analogous to Mallows' C_p.

Stacking. Select weights minimizing leave-one-out cross-validated prediction error:

$$\hat{w}^{\text{stack}} = \arg \min_{w \geq 0, \sum w_j = 1} \sum_{i=1}^{n} \left(Y_i - \sum_j w_j \, \hat{f}_j^{(-i)}(X_i) \right)^2,$$

where $\hat{f}_j^{(-i)}$ is model j trained without observation i. This is a constrained least squares problem in K weights, solvable by quadratic programming.

Stacking is model-agnostic, requires no marginal likelihood computation, and does not assume that any candidate model is true. It is asymptotically optimal in the sense of minimizing the KL divergence to the true distribution among all convex combinations of the candidate models (Wolpert, 1992).

Model Averaging vs. Model Selection

Criterion	Selection	Averaging
Computational cost	Lower	Higher
Interpretability	Single model	Weighted combination
Risk when one model dominates	Equivalent	Slightly worse (dilution)
Risk under model uncertainty	Higher (wrong model bias)	Lower (hedging)
Uncertainty quantification	Ignores model uncertainty	Propagates model uncertainty

Asymptotically, if a true model exists, both converge to it. In finite samples with substantial model uncertainty, averaging typically has lower prediction MSE.

Worked Examples

Example 1: BMA with Two Nested Models

Model 1: $Y = \alpha + \varepsilon$ (intercept only), $k_1 = 2$, $\ell_1 = -52.3$. Model 2: $Y = \alpha + \beta X + \varepsilon$, $k_2 = 3$, $\ell_2 = -49.1$. $n = 40$, equal priors.

BIC: $\text{BIC}_1 = 104.6 + 2\log 40 = 104.6 + 7.38 = 111.98$; $\text{BIC}_2 = 98.2 + 3\log 40 = 98.2 + 11.07 = 109.27$.

Weights: $w_1 = \exp(-111.98/2)/(\exp(-111.98/2) + \exp(-109.27/2))$.

$\Delta\text{BIC} = 111.98 - 109.27 = 2.71$. So $w_1/w_2 = \exp(-2.71/2) = \exp(-1.355) = 0.258$. Thus $w_2 = 1/(1 + 0.258) = 0.795$, $w_1 = 0.205$.

BMA prediction at new x^*: $\hat{Y}^{\text{BMA}} = 0.205 \cdot \hat{\alpha}_1 + 0.795 \cdot (\hat{\alpha}_2 + \hat{\beta}_2 x^*)$. Model 2 dominates but Model 1 retains 20.5% weight, reflecting non-negligible uncertainty about the predictor's relevance.

Example 2: Stacking Three Regression Models

Models: (A) linear regression with 3 predictors, (B) polynomial regression degree 3, (C) ridge regression with $\lambda = 5$. $n = 60$.

Compute LOO predictions $\hat{f}_A^{(-i)}(X_i)$, $\hat{f}_B^{(-i)}(X_i)$, $\hat{f}_C^{(-i)}(X_i)$ for $i = 1, \ldots, 60$. Form the 60×3 matrix Z with $Z_{ij} = \hat{f}_j^{(-i)}(X_i)$.

Solve: $\min_{w \geq 0, \sum w_j = 1} \|Y - Zw\|^2$.

Suppose the solution is $w_A = 0.35$, $w_B = 0.10$, $w_C = 0.55$. Ridge regression (C) gets the most weight because it has the best leave-one-out predictions. The polynomial model (B) contributes little—its LOO error is high due to overfitting.

For a new observation x^*: $\hat{Y}^{\text{stack}} = 0.35\hat{f}_A(x^*) + 0.10\hat{f}_B(x^*) + 0.55\hat{f}_C(x^*)$.

Cross-Links

AI. Ensemble methods—random forests, boosting, mixture of experts—are frequentist model averaging procedures. Stacking is equivalent to the super-learner algorithm in targeted learning. In deep learning, model ensembles (averaging predictions from multiple trained networks) consistently improve prediction and calibration.

Economics. BMA is used in growth regressions (which variables determine economic growth?), where the set of plausible predictors far exceeds what any single model can support. Forecast combination (averaging multiple economic forecasts) typically outperforms any single forecaster, a robust empirical finding known as the "forecast combination puzzle."

Linear Algebra. The stacking weights solve a constrained least squares problem: $\min_{w \in \Delta^{K-1}} \|Y - Zw\|^2$, where Z is the matrix of leave-one-out predictions. The simplex constraint makes this a quadratic program solvable by active-set or interior-point methods.

Quantum Mechanics. In quantum state estimation, Bayesian model averaging over different state-space dimensions (ranks of the density matrix) produces predictive distributions that account for uncertainty in the state's rank, analogous to BMA over models of different complexity.

Structural Compression

Invariant

BMA predictions are the posterior-weighted average of predictions across models; each model's contribution is proportional to its marginal likelihood times its prior probability.

Mechanism

Marginal likelihoods (or BIC approximations) reweight prior model probabilities by the data's compatibility with each model. BMA is optimal under proper scoring rules because the posterior mixture minimizes the expected KL divergence to the true data-generating process across models. Stacking achieves the analogous frequentist optimality by minimizing cross-validated prediction error over convex combinations.

Cross-System Echo

BMA is the coherent Bayesian solution to model uncertainty. In AI, ensemble methods and mixture-of-experts architectures operationalize the same principle: multiple models combined outperform any single model when uncertainty is non-trivial. Stacking is the super-learner of targeted learning; forecast combination in economics is the applied instantiation.

Exercises

1. Derive the BMA predictive distribution and show it equals $p(\tilde{x} \mid x) = \sum_j \pi(\mathcal{M}_j \mid x)\, p(\tilde{x} \mid x, \mathcal{M}_j)$ starting from the joint posterior $p(\tilde{x}, j \mid x) = p(\tilde{x} \mid x, \mathcal{M}_j)\, \pi(\mathcal{M}_j \mid x)$ and marginalizing over j.
2. Prove that BMA is optimal under the log scoring rule: show that $\mathbb{E}[-\log q(\tilde{X}) \mid x]$ is minimized over mixture distributions $q = \sum_j w_j p_j$ when $w_j = \pi(\mathcal{M}_j \mid x)$.
3. For three models with BIC values $100.0, 102.5, 108.0$ and equal priors, compute the BMA weights. How does the result change if the prior on Model 1 is doubled?
4. Show that as $n \to \infty$, BMA with the true model among candidates concentrates all weight on the true model (i.e., BMA inherits the consistency of Bayes factors).
5. Formulate the stacking optimization as a quadratic program. Write out the objective and constraints for $K = 3$ models and show that the solution lies on the boundary of the simplex when one model strictly dominates the others in LOO performance.
6. Compare BMA and stacking on a concrete example: $K = 2$ models, one correctly specified and one misspecified, with $n = 50$. Show that BMA concentrates weight on the correct model while stacking may assign positive weight to the misspecified model if it improves out-of-sample prediction in the finite sample.
7. Argue, structurally, why model averaging must outperform model selection in expected prediction error when model uncertainty is substantial, and identify the regime where selection is preferable (one model clearly dominates, interpretability is paramount).

Statistics · Module 10 — Model Selection Learning · Node 10.6
Concepts: model-selection, prior-posterior, bayesian-updating
Prerequisites: 10.2, 6.5, 6.6

hbar.university — own.your.cognition