Model Selection

CEVE 543 - Fall 2025

Dr. James Doss-Gollin

2025-11-10

1 Motivation

Statistical modeling requires making many choices: which distribution family to use, which covariates (if any) to include for nonstationarity, which parameters to model as nonstationary, how to pool information across space, and more. There is no single "right" answer to these questions. How can we proceed in a principled way?

This challenge extends well beyond extreme value analysis to all of statistical modeling [1], [2]. In this lecture, we explore quantitative criteria for model selection and comparison.

> **Linear Regression as a Mental Model**
>
> Many of the quantitative approaches we discuss were originally developed with linear regression in mind, though they apply more generally. When interpreting these methods for extreme value analysis or other complex models, it's worth keeping in mind that some of their theoretical guarantees may not hold exactly.

> **Further Reading**
>
> For accessible discussions of model selection challenges, see D. J. Navarro [2] and G. Heinze, C. Wallisch, and D. Dunkler [3]. For more mathematical depth on Bayesian predictive methods, see J. Piironen and A. Vehtari [4]. R. McElreath [5] (Chapter 7) provides an excellent conceptual introduction to information criteria.

2 Quantitative Model Selection

> **Technical Content**
>
> This section introduces some mathematical formalism. You should focus on understanding the key concepts and trade-offs rather than memorizing equations.

2.1 The Challenge

We want to make probabilistic predictions about unobserved future data $\tilde{y}$. This is challenging because Earth systems are high-dimensional, multi-scale, nonlinear, and complex.

To approximate the true system, we define a model space $\mathcal{M}$ containing candidate models, then use data to select among them [4]. The key question becomes: how do we measure and compare the predictive performance of different models?

2.2 Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence measures how similar two probability distributions are. Let $\mathcal{X}$ denote the set of all possible outcomes (e.g., possible values of annual maximum precipitation), and let x denote a specific outcome. The KL divergence from distribution P to distribution Q is:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left[\frac{P(x)}{Q(x)}\right]$$

One interpretation is the amount of information lost when Q is used to approximate P. Another is the information gained by revising one's beliefs from Q to P. For continuous random variables, the sum becomes an integral.

Minimizing KL divergence from a candidate model to the true data-generating distribution is equivalent to maximizing expected predictive accuracy.

2.3 Measures of Predictive Accuracy

We measure predictive performance using the log predictive density $\log p(\tilde{y} \mid D, M)$, where D represents our data and M represents our model. Since future observations $\tilde{y}$ are unknown, we evaluate this in expectation:

$$\overline{u}(M) = \mathbb{E}[\log p(\tilde{y} \mid D, M)] = \int p_t(\tilde{y}) \log p(\tilde{y} \mid D, M)\, d\tilde{y}$$

where $p_t(\tilde{y})$ is the unknown true data-generating distribution.

Maximizing this expected log predictive density is equivalent to minimizing KL divergence from our candidate model to the true distribution.

In practice, we don't know the true distribution, so we approximate using the log pointwise predictive density (lppd):

$$\mathrm{lppd} = \sum_{i=1}^{N} \log\left[\frac{1}{S}\sum_{s=1}^{S} p(y_i \mid \theta^s)\right]$$

where we have approximated the posterior distribution with S samples from MCMC.

The lppd computed on observed data y overestimates the expected predictive density for future data. Information criteria correct for this optimism.

3 Model Comparison Approaches

3.1 Significance Testing

A common approach in practice is to use null hypothesis significance testing (NHST) to decide whether to include variables [1], [3]. For example, to decide whether to include a trend term with coefficient β:

1. Specify a null hypothesis: $\beta = 0$
2. Compute a test statistic and p-value
3. If $p < \alpha$ (commonly $\alpha = 0.05$), include the variable; otherwise exclude it

This approach is widely used but has well-documented problems [1], [6]. It assumes the existence of a true model ($\mathcal{M}$-closed) and is sensitive to arbitrary significance thresholds. Sequential testing (testing multiple variables one at a time) compounds these issues through the problem of multiple comparisons.

3.2 Information Criteria

Information criteria provide quantitative measures for comparing models by balancing fit quality against model complexity. All of these criteria can be computed from fitted models and used to compare alternatives.

3.2.a Akaike Information Criterion (AIC)

The AIC provides a simple correction for overfitting [5]. If a model with maximum likelihood estimate $\hat{\theta}_{\text{mle}}$ fits k parameters, then:

$$\text{AIC} = 2k - 2\ln\hat{\mathcal{L}}$$

where $\hat{\mathcal{L}}$ is the maximized likelihood.

The first term penalizes model complexity (more parameters increase AIC). The second term rewards fit to the data (higher likelihood decreases AIC). We select the model that minimizes AIC.

When to use: AIC is appropriate for maximum likelihood estimation and focuses on predictive performance in an $\mathcal{M}$-open setting. It aims to select the model that will best predict future data, even if none of the candidate models is "true."

Assumptions: The AIC assumes that parameters are asymptotically normally distributed and that residuals are independent given $\hat{\theta}$. These assumptions often hold approximately for linear regression but may be questionable for more complex models.

Key limitation: For complex models, determining the effective number of parameters k is not straightforward.

3.2.b Bayesian Information Criterion (BIC)

The BIC takes a different approach by approximating the marginal probability of the data $p(y)$ [7]:

$$\text{BIC} = k\ln(n) - 2\ln\hat{\mathcal{L}}$$

where n is the sample size and $\hat{\mathcal{L}} = \max_\theta p(y \mid \theta, M)$.

The BIC penalizes model complexity more heavily than AIC when $n > 7$, favoring simpler models.

When to use: BIC is appropriate when you believe the true model is among your candidates (an $\mathcal{M}$-closed perspective) and want to identify it. The model with the lowest BIC will asymptotically be the

"true" model if the true model is one of the choices. This makes BIC useful for hypothesis testing and model identification rather than pure prediction.

Key assumption: BIC assumes that one of the candidate models is the true data-generating process. Under this assumption, BIC provides consistent model selection as sample size increases.

3.2.c Watanabe-Akaike Information Criterion (WAIC)

WAIC is a more fully Bayesian approach that is asymptotically equivalent to Bayesian cross-validation [8]. Unlike AIC and BIC, which use point estimates, WAIC uses the full posterior distribution.

The WAIC is computed as:

$$\text{WAIC} = -2(\text{lppd} - p_{\text{WAIC}})$$

where lppd is the log pointwise predictive density (defined earlier) and p_{WAIC} is the effective number of parameters, computed as:

$$p_{\text{WAIC}} = \sum_{i=1}^{N} \text{Var}_{\text{post}}[\log p(y_i \mid \theta)]$$

The variance is computed over the posterior distribution of θ.

When to use: WAIC is the recommended default for Bayesian model comparison [8]. Use it when you have MCMC samples and want to focus on predictive performance. It properly accounts for posterior uncertainty and works in both $\mathcal{M}$-open and $\mathcal{M}$-closed settings.

Advantages: WAIC is fully Bayesian and works with posterior distributions rather than point estimates.

Limitations: Like other information criteria, WAIC assumes that future data will be generated from the same process as the observed data.

3.3 Model Averaging

Rather than selecting a single "best" model, we can treat model selection as a source of uncertainty [4]. If we have candidate models $\{M_\ell\}_{\ell=1}^{L}$, the posterior distribution over models is:

$$p(M \mid D) \propto p(D \mid M)p(M)$$

We can then make predictions by averaging over all models:

$$p(\tilde{y} \mid D) = \sum_{\ell=1}^{L} p(\tilde{y} \mid D, M_\ell)p(M_\ell \mid D)$$

This approach acknowledges model uncertainty explicitly rather than conditioning all inferences on a single selected model. Bayesian model averaging can improve predictive performance, especially when multiple models have similar support from the data [8]. More advanced approaches like stacking can further enhance predictive accuracy [9].

4 $\mathcal{M}$-Open vs $\mathcal{M}$-Closed Perspectives

A fundamental distinction in model selection is whether we assume the "true" model is in our candidate set [8].

$\mathcal{M}$-Closed: Assumes one of the candidate models is the true data-generating process. Under this assumption, methods like BIC have strong theoretical guarantees—as sample size increases, we will identify the true model. This perspective motivates selection approaches that choose a single "best" model.

$\mathcal{M}$-Open: Recognizes that all models are approximations, and the true data-generating process is not in our candidate set. This is the reality we face in practice, especially for complex Earth systems. Under this perspective, we should focus on predictive performance rather than identifying a "true" model [4].

The distinction has practical implications. In an $\mathcal{M}$-closed world, theoretical guarantees about convergence and optimality apply. In an $\mathcal{M}$-open world (where we actually live), these guarantees break down, and judgment becomes essential [2], [10].

This is why model selection requires more than just computing information criteria and selecting the minimum. We must consider the scientific context, check assumptions, examine multiple metrics, and acknowledge uncertainty. Statistical theory provides useful tools, but cannot replace domain knowledge and careful thinking about what we're trying to accomplish.

5 Key Takeaways

Model comparison and selection involves subjective judgments [2], [10]. There is no purely objective, automatic way to identify the "best" model [3]. Several important principles guide good practice:

1. **No single criterion is definitive.** Be skeptical of analyses that rely on a single metric (like "AIC selected this model") without additional justification [3].
2. **Transparency matters.** Make your assumptions and decision criteria explicit so others can evaluate and critique your choices [10], [11].
3. **Context and purpose guide selection.** Different applications may prioritize different aspects of model performance (e.g., tail behavior vs. central tendency, parsimony vs. flexibility) [2].
4. **Uncertainty about model form deserves attention.** When multiple models perform similarly, model averaging acknowledges this uncertainty rather than pretending we know the true model [4].

Statistical modeling requires judgment informed by domain knowledge, diagnostic checking, and awareness of the assumptions underlying different selection criteria [12]. Subjective does not mean arbitrary —we can and should use principled approaches while acknowledging the limits of formal procedures [10].

6 References

Bibliography

[1] G. Heinze and D. Dunkler, "Five Myths about Variable Selection," *Transplant International*, vol. 30, no. 1, pp. 6–10, 2017, doi: 10.1111/tri.12895.

[2] D. J. Navarro, "Between the Devil and the Deep Blue Sea: Tensions Between Scientific Judgement and Statistical Model Selection," *Computational Brain & Behavior*, Nov. 2018, doi: 10.1007/s42113-018-0019-z.

[3] G. Heinze, C. Wallisch, and D. Dunkler, "Variable Selection – a Review and Recommendations for the Practicing Statistician," *Biometrical Journal*, vol. 60, no. 3, pp. 431–449, 2018, doi: 10.1002/bimj.201700067.

[4] J. Piironen and A. Vehtari, "Comparison of Bayesian Predictive Methods for Model Selection," *Statistics and Computing*, vol. 27, no. 3, pp. 711–735, May 2017, doi: 10.1007/s11222-016-9649-y.

[5] R. McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, Second edition. in Texts in Statistical Science Series. Boca Raton ;: CRC Press, Taylor & Francis Group, 2020.

[6] A. Gelman and C. R. Shalizi, "Philosophy and the Practice of Bayesian Statistics," *British Journal of Mathematical and Statistical Psychology*, vol. 66, no. 1, pp. 8–38, 2013, doi: 10/f4k2h4.

[7] R. E. Kass and A. E. Raftery, "Bayes Factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, Jun. 1995, doi: 10.1080/01621459.1995.10476572.

[8] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[9] Y. Yao, A. Vehtari, D. Simpson, and A. Gelman, "Using Stacking to Average Bayesian Predictive Distributions," *Bayesian Analysis*, 2018, doi: 10.1214/17-ba1091.

[10] A. Gelman and C. Hennig, "Beyond Subjective and Objective in Statistics," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 180, no. 4, pp. 967–1033, 2017, doi: 10.1111/rssa.12276.

[11] J. Doss-Gollin and K. Keller, "A Subjective Bayesian Framework for Synthesizing Deep Uncertainties in Climate Risk Management," *Earth's Future*, vol. 11, no. 1, Jan. 2023, doi: 10.1029/2022EF003044.

[12] A. Gelman *et al.*, "Bayesian Workflow," Nov. 03, 2020. doi: 10.48550/arXiv.2011.01808.