

A large, abstract, light blue wavy graphic that resembles a stylized knot or a complex wave pattern, serving as a background for the title text.

AIC, BIC, Bayesian Evidence

Information criteria

- Process f is described with models g_1 and g_2 . If we do not know f , then we cannot calculate Kullback–Leibler divergence
- via AIC we can estimate how much more (or less) information is lost by g_1 than by g_2 . The estimate, though, is only valid asymptotically - if the number of data points is small, then some correction is often necessary (e.g. AICc)

Akaike information criterion

Selection Target

Akaike (1973, 1974, 1985, 1994) showed that the critical issue for getting an applied K-L model selection criterion was to estimate

$$\mathbf{E}_y \mathbf{E}_x [\log(g(x|\hat{\theta}(y)))],$$

where x and y are independent random samples from the same distribution and both statistical expectations are taken with respect to truth (f). This double expectation, both with respect to truth f , is the target of all model selection approaches, based on K-L information.

Akaike information criterion

The Key Result

Thus, an approximately unbiased estimator of

$$\mathbf{E}_y \mathbf{E}_x [\log(g(x|\hat{\theta}(y)))]$$

for large samples and “good” models is

$$\log(\mathcal{L}(\hat{\theta}|data)) - K.$$

This result is equivalent to

$$\log(\mathcal{L}(\hat{\theta}|data)) - K = \text{constant} - \hat{\mathbf{E}}_{\hat{\theta}}[I(f, \hat{g})],$$

where $\hat{g} = g(\cdot|\hat{\theta})$.

Akaike information criterion

$$AIC = 2K - 2\log(\mathcal{L}(\hat{\theta}|y))$$

- A mean for model selection
- AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model
- AIC rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters

Akaike information criterion

The Least Squares Case

If all the models in the set assume normally distributed errors with a constant variance, then AIC can be easily computed from least squares regression statistics as

$$\text{AIC} = n \log(\hat{\sigma}^2) + 2K,$$

where

$$\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n} \text{ (the MLE of } \sigma^2 \text{),}$$

and $\hat{\epsilon}_i$ are the estimated residuals for a particular candidate model. A common mistake with LS model fitting, when computing AIC, is to take the estimate of σ^2 from the computer output, instead of computing the ML estimate, above. **Also, for LS model fitting, K is the total number of estimated regression parameters, including the intercept and σ^2 .**

Hypothesis testing 1

- Input: a random sample from each of the two populations

- Two models:

$$- \mu_1, \sigma_1, \mu_2, \sigma_2$$

$$- \mu_1 = \mu_2, \sigma_1, \sigma_2$$

- The likelihood function:

$$\mathcal{L}(\mu_1, \sigma_1, \mu_2, \sigma_2) = \prod_{i=1}^{n_1} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2}\right) \cdot \prod_{i=n_1+1}^{n_1+n_2} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x_i - \mu_2)^2}{2\sigma_2^2}\right)$$

- $\exp((AIC_1 - AIC_2)/2) = 0.01 \rightarrow$ the two populations have different means
- Advantages: the sizes of the two samples can be different, no assumptions on deviations

Hypothesis testing 2

- Input: two packs of white and black balls → binomially distributed; first pack contains m balls, m_1 of them are white, n and n_1 are the same numbers for the second pack.

- Two models:

- p, q

- $p = q$

- The likelihood function:

$$\mathcal{L}(p, q) = \frac{m!}{m_1!(m - m_1)!} p^{m_1} (1 - p)^{m - m_1} \cdot \frac{n!}{n_1!(n - n_1)!} q^{n_1} (1 - q)^{n - n_1}$$

- $\exp((AIC_1 - AIC_2)/2) = 0.01 \rightarrow$ the two populations have different distributions

Statistical inference using AIC

- Point estimation can be done within the AIC paradigm: it is provided by maximum likelihood estimation.

$$\hat{\theta} \in \{\arg \max_{\theta \in \Theta} \mathcal{L}(\theta; x)\},$$

- Interval estimation can also be done within the AIC paradigm: it is provided by likelihood intervals.

$$\left\{ \theta : \frac{\mathcal{L}(\theta | x)}{\mathcal{L}(\hat{\theta} | x)} \geq \frac{p}{100} \right\}.$$

Bayesian information criterion

$$\text{BIC} = -2 \log(\mathcal{L}) + K \cdot \log(n).$$

- independent of the prior.
- can measure the efficiency of the parameterized model in terms of predicting the data.
- penalizes the complexity of the model where complexity refers to the number of parameters in the model.

BUT:

- n must be much larger than the number of parameters k in the model.
- cannot handle complex collections of models as in the variable selection problem in high-dimension

Bayesian information criterion

$$p(x \mid M) = \int p(x \mid \theta, M) \pi(\theta \mid M) d\theta$$

where $\pi(\theta \mid M)$ is the prior for θ under model M .

The log(likelihood), $\ln(p(x \mid \theta, M))$, is then expanded to a second order Taylor series :

$$\ln(p(x \mid \theta, M)) = \ln(\hat{L}) - 0.5(\theta - \hat{\theta})' n \mathcal{I}(\hat{\theta})(\theta - \hat{\theta}) + R(x, \theta),$$

$$p(x \mid M) \approx \hat{L} (2\pi/n)^{k/2} |\mathcal{I}(\hat{\theta})|^{-1/2} \pi(\hat{\theta})$$

As n increases, we can ignore $|\mathcal{I}(\hat{\theta})|$ and $\pi(\hat{\theta})$ as they are $O(1)$. Thus,

$$p(x \mid M) = \exp\{\ln \hat{L} - (k/2) \ln(n) + O(1)\} = \exp(-\text{BIC}/2 + O(1)),$$

$$p(M \mid x) \propto p(x \mid M) p(M) \approx \exp(-\text{BIC}/2) p(M)$$

AIC vs BIC

- If the "true model" is in the set of candidates, then BIC will select the "true model" with probability 1, as $n \rightarrow \infty$; in contrast, when selection is done via AIC, the probability can be less than 1.
- for finite n , BIC can have a risk of selecting a very bad model from the candidate set
- In regression, AIC is asymptotically optimal for selecting the model with the least mean squared error, under the assumption that the "true model" is not in the candidate set

Bayesian evidence

- Bayes' theorem

$$P(m|d, M) = P(m|M) \frac{P(d|m, M)}{P(d|M)}$$

- Evidence can be estimated

$$\begin{aligned} P(d|M, I) &= \int dm P(m, d|M, I) \\ &= \int dm P(m|M, I) P(d|m, M, I), \end{aligned}$$

- If there are two competing theories

$$\frac{P(M_1, d|I)}{P(M_2, d|I)} = \frac{P(M_1|I) P(d|M_1, I)}{P(M_2|I) P(d|M_2, I)}$$

$$\frac{P(d|I) P(M_1|d, I)}{P(d|I) P(M_2|d, I)} = \frac{P(M_1|I) P(d|M_1, I)}{P(M_2|I) P(d|M_2, I)}$$

$$\frac{P(M_1|d, I)}{P(M_2|d, I)} = \frac{P(M_1|I)}{P(M_2|I)} \frac{P(d|M_1, I)}{P(d|M_2, I)}$$

Bayesian evidence

the Bayes factor or odds ratio

$$\text{OR} = \frac{P(d|M_1, I)}{P(d|M_2, I)}$$

or, equivalently, the log odds ratio

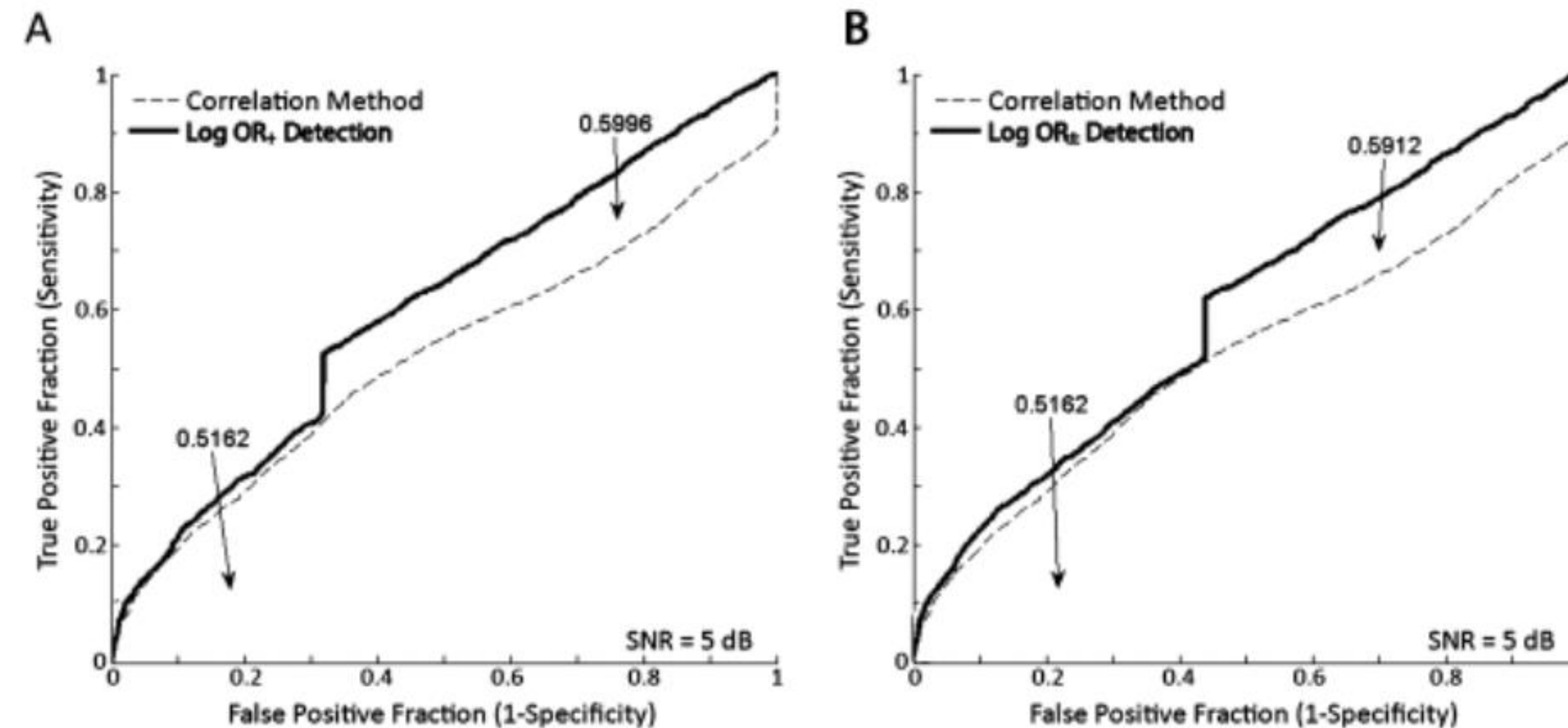
$$\log \text{OR} = \log P(d|M_1, I) - \log P(d|M_2, I).$$

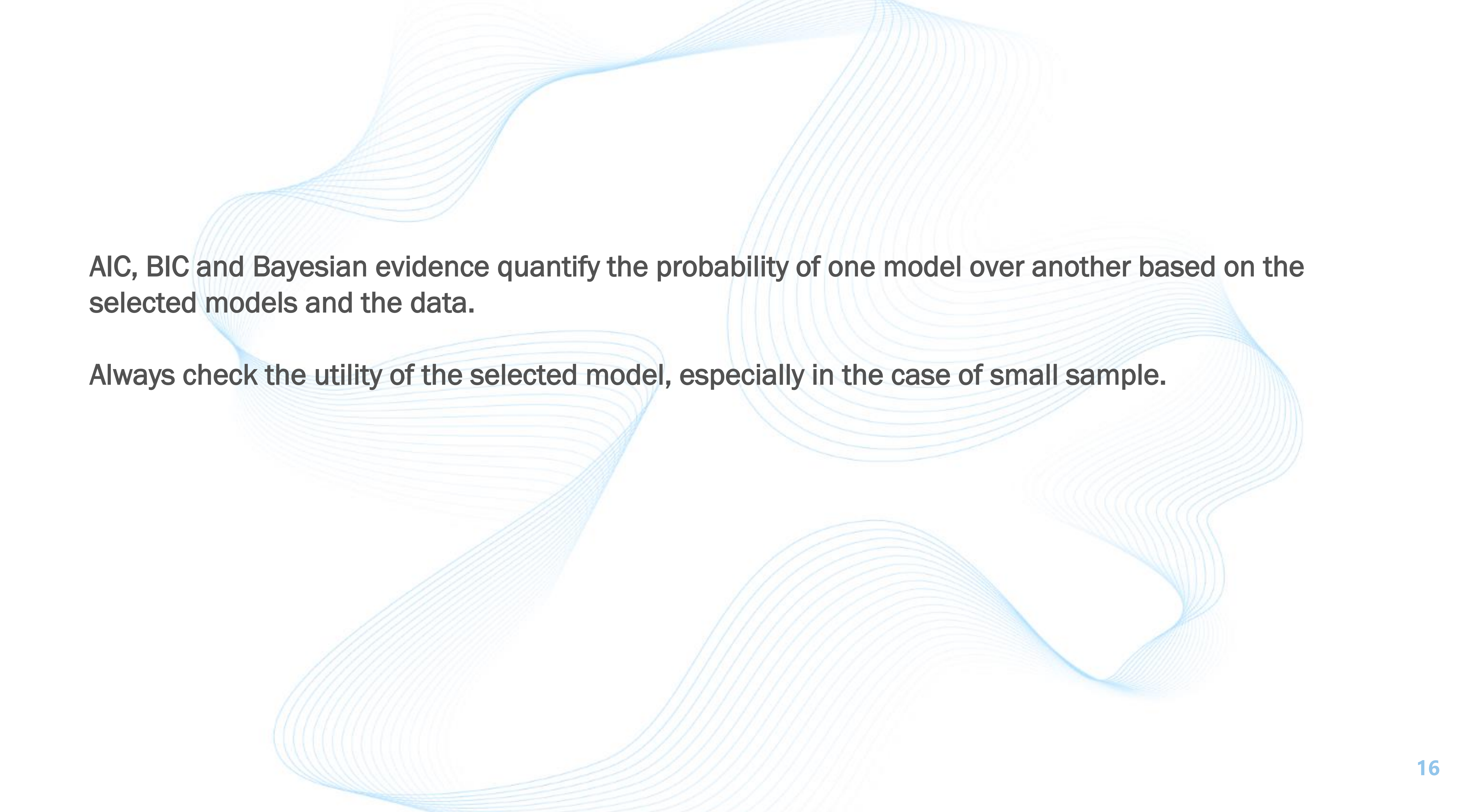
With this definition, we can write the ratio of posterior probabilities
the two different theories M_1 and M_2 in terms of the odds ratio

$$\frac{P(M_1|d, I)}{P(M_2|d, I)} = \frac{P(M_1|I)}{P(M_2|I)} \times \text{OR},$$

Bayesian evidence. Examples.

- Force field selection in biomolecular structure determination
- Exoplanet detection
- Light sensor characterization
- Signal detection



The background of the slide features a series of thin, light blue lines that flow and curve across the page, creating a sense of movement and depth. These lines are more densely packed in some areas, forming soft, cloud-like shapes, while in other areas they are more sparse.

AIC, BIC and Bayesian evidence quantify the probability of one model over another based on the selected models and the data.

Always check the utility of the selected model, especially in the case of small sample.