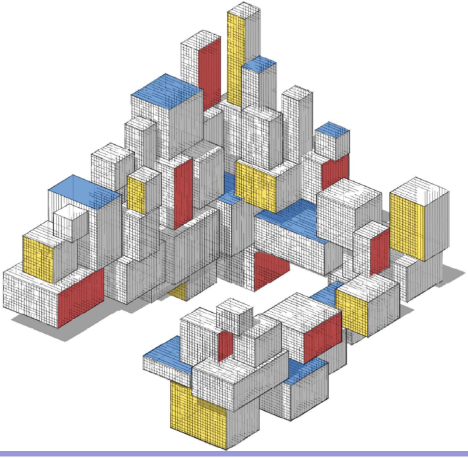
Bayesian Information Criterion



Purpose

In this lecture we discuss:

- Bayesian Unsupervised Learning
- Bayesian Information Criterion
- Bayesian Model Selection

Bayesian Unsupervised Learning

We return to the unsupervised learning setting, but consider this from a Bayesian perspective. Recall that the Kullback–Leibler risk for an approximating function g is

$$\ell(g) = \int f(\tau_n')\,[\ln f(\tau_n') - \ln g(\tau_n')]\,\mathrm{d}\tau_n',$$

where it is convenient to consider here τ' as test (rather than training) data.

Since $\int f(\tau_n') \ln f(\tau_n')\,\mathrm{d}\tau_n'$ plays no role in minimizing the risk, we consider instead the cross-entropy risk, defined as

$$\ell(g) = -\int f(\tau_n') \ln g(\tau_n')\,\mathrm{d}\tau_n'.$$

Cross-Entropy Risk

The smallest possible cross-entropy risk is

$$\ell_n^* = -\int f(\tau_n')\ln f(\tau_n')\,\mathrm{d}\tau_n'.$$

The expected generalization risk of the Bayesian learner can then be decomposed as

$$\mathbb{E}\,\ell(g_{\mathcal{T}_n}) = \ell_n^* + \underbrace{\int f(\tau_n')\ln\frac{f(\tau_n')}{\mathbb{E}\,g(\tau_n' \,|\, \mathcal{T}_n)}\,\mathrm{d}\tau_n'}_{\text{“bias” component}} + \underbrace{\mathbb{E}\int f(\tau_n')\ln\frac{\mathbb{E}\,g(\tau_n' \,|\, \mathcal{T}_n)}{g(\tau_n' \,|\, \mathcal{T}_n)}\,\mathrm{d}\tau_n'}_{\text{“variance” component}},$$

where $g_{\mathcal{T}_n}(\tau_n') = g(\tau_n' \,|\, \mathcal{T}_n) = \int g(\tau_n' \,|\, \boldsymbol{\theta})\, g(\boldsymbol{\theta} \,|\, \mathcal{T}_n)\,\mathrm{d}\boldsymbol{\theta}$ is the posterior predictive density after observing $\mathcal{T}_n$.

Generalization Risk for the Iid Case

For the case where the sets $\mathcal{T}_n$ and $\mathcal{T}_n'$ are comprised of $2n$ iid random variables with density f, the expected generalization risk simplifies (exercise!) to

$$\mathbb{E}\,\ell(g_{\mathcal{T}_n}) = \mathbb{E}\ln g(\mathcal{T}_n) - \mathbb{E}\ln g(\mathcal{T}_{2n}), \tag{1}$$

where $g(\tau_n)$ and $g(\tau_{2n})$ are the prior predictive densities of τ_n and τ_{2n}, respectively.

Let $\overline{\boldsymbol{\theta}}_n := \operatorname{argmax}_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}\,|\,\mathcal{T}_n)$ be the MAP estimator and let $\boldsymbol{\theta}^* := \operatorname{argmax}_{\boldsymbol{\theta}}\, \mathbb{E}\ln g(\boldsymbol{X}\,|\,\boldsymbol{\theta})$.

For large n, $\overline{\boldsymbol{\theta}}_n \approx \boldsymbol{\theta}^*$.

Approximation of the Expected Generalization Risk

Assuming that $\overline{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}^*$ (with probability one) and $\frac{1}{n}\mathbb{E}\ln g(\mathcal{T}_n \,|\, \overline{\boldsymbol{\theta}}_n) = \mathbb{E}\ln g(\boldsymbol{X} \,|\, \boldsymbol{\theta}^*) + \mathcal{O}(1/n)$, we can use the following large-sample approximation of the expected generalization risk.

Theorem: Approximating the Bayesian CE Risk

Let p be the dimension of $\boldsymbol{\theta}$. For $n \to \infty$, the expected cross-entropy generalization risk satisfies:

$$\mathbb{E}\ell(g_{\mathcal{T}_n}) \simeq -\mathbb{E}\ln g(\mathcal{T}_n) - \frac{p}{2}\ln n, \tag{2}$$

where

$$\mathbb{E}\ln g(\mathcal{T}_n) \simeq \mathbb{E}\ln g(\mathcal{T}_n \,|\, \overline{\boldsymbol{\theta}}_n) - \frac{p}{2}\ln n. \tag{3}$$

Proof

To show (3), we apply Laplace's theorem to approximate $\ln \int \mathrm{e}^{-n r_n(\boldsymbol{\theta})} g(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$, where

$$r_n(\boldsymbol{\theta}) := -\frac{1}{n} \ln g(\mathcal{T}_n \,|\, \boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \ln g(\boldsymbol{X}_i \,|\, \boldsymbol{\theta}) \xrightarrow{\text{a.s.}} -\mathbb{E} \ln g(\boldsymbol{X} \,|\, \boldsymbol{\theta}) =: r(\boldsymbol{\theta}).$$

This gives (with probability one)

$$\ln \int g(\mathcal{T}_n \,|\, \boldsymbol{\theta})\, g(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \simeq -n\, r(\boldsymbol{\theta}^*) - \frac{p}{2} \ln(n).$$

Taking expectations on both sides and using $n\, r(\boldsymbol{\theta}^*) = n\, \mathbb{E} r_n(\overline{\boldsymbol{\theta}}_n) + \mathcal{O}(1)$, we deduce (3).

To demonstrate (2), we derive the asymptotic approximation of $\mathbb{E} \ln g(\mathcal{T}_{2n})$ by repeating the argument for (3), but replacing n with $2n$, where necessary. Thus, we obtain:

$$\mathbb{E} \ln g(\mathcal{T}_{2n}) \simeq -2n r(\boldsymbol{\theta}^*) - \frac{p}{2} \ln(2n).$$

Then, (2) follows from the identity (1).

Model Evidence

The results of this theorem have two major implications for model selection and assessment.

The first result of the theorem suggests that $-\ln g(\mathcal{T}_n)$ can be used as an approximation to the expected generalization risk for large n and fixed p.

In this context, the prior predictive density $g(\mathcal{T}_n)$ is usually called the model evidence or marginal likelihood for the class $\mathcal{G}_p$.

Since the integral $\int g(\mathcal{T}_n \mid \boldsymbol{\theta})\, g(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$ is rarely available in closed form, the exact computation of the model evidence is typically not feasible and may require Monte Carlo estimation methods.

Bayesian Information Criterion

The second result of the theorem suggests that we can use the following large-sample approximation:

$$-2\mathbb{E}\ln g(\mathcal{T}_n) \simeq -2\ln g(\mathcal{T}_n \,|\, \overline{\boldsymbol{\theta}}_n) + p\ln(n). \tag{4}$$

The asymptotic approximation on the right-hand side of (4) is called the Bayesian information criterion (BIC). We prefer the class $\mathcal{G}_p$ with the smallest BIC. The BIC is typically used when the model evidence is difficult to compute and n is sufficiently larger than p. For a fixed p, and as n becomes larger and larger, the BIC becomes a more and more accurate estimator of $-2\mathbb{E}\ln g(\mathcal{T}_n)$. Note that the BIC approximation is valid even when the true density $f \notin \mathcal{G}_p$.

Bayesian Supervised Learning

Although the above Bayesian theory has been presented in an unsupervised learning setting, it can be readily extended to the supervised case. We only need to relabel the training set $\mathcal{T}_n$.

In particular, when (as is typical for regression models) the training responses $Y_1, \ldots, Y_n$ are considered as random variables but the corresponding feature vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are viewed as being fixed, then $\mathcal{T}_n$ is the collection of random responses $\{Y_1, \ldots, Y_n\}$.

Alternatively, we can simply identify $\mathcal{T}_n$ with the response vector $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^\top$. In particular, for a normal linear model:

$$\boldsymbol{Y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}_n).$$

We will adopt this notation in the next example.

Normal Linear Model

Consider the normal linear model, but now in a Bayesian framework, where the prior knowledge on $(\sigma^2, \boldsymbol{\beta})$ is specified by $g(\sigma^2) = 1/\sigma^2$ and $\boldsymbol{\beta} \,|\, \sigma^2 \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{D})$, and $\mathbf{D}$ is a (matrix) hyperparameter. Let $\boldsymbol{\Sigma} := (\mathbf{X}^\top \mathbf{X} + \mathbf{D}^{-1})^{-1}$. Then the posterior can be written as:

$$g(\boldsymbol{\beta}, \sigma^2 \,|\, \boldsymbol{y}) = \frac{\exp\left(-\frac{\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{n/2}} \times \frac{\exp\left(-\frac{\boldsymbol{\beta}^\top \mathbf{D}^{-1} \boldsymbol{\beta}}{2\sigma^2}\right)}{(2\pi\sigma^2)^{p/2}\, |\mathbf{D}|^{1/2}} \times \frac{1}{\sigma^2} \bigg/ g(\boldsymbol{y})$$

$$= \frac{(\sigma^2)^{-(n+p)/2-1}}{(2\pi)^{(n+p)/2}\, |\mathbf{D}|^{1/2}} \exp\left(-\frac{\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\beta} - \overline{\boldsymbol{\beta}})\|^2}{2\sigma^2} - \frac{(n+p+2)\,\overline{\sigma}^2}{2\sigma^2}\right) \bigg/ g(\boldsymbol{y}),$$

where $\overline{\boldsymbol{\beta}} := \boldsymbol{\Sigma}\mathbf{X}^\top \boldsymbol{y}$ and $\overline{\sigma}^2 := \boldsymbol{y}^\top (\mathbf{I} - \mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^\top)\boldsymbol{y}/(n+p+2)$ are the MAP estimates of $\boldsymbol{\beta}$ and σ^2.

Model Evidence

The model evidence, $g(\boldsymbol{y})$, for $\mathcal{G}_p$ is:

$$\begin{aligned} g(\boldsymbol{y}) &= \iint g(\boldsymbol{\beta}, \sigma^2, \boldsymbol{y})\, \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\sigma^2 \\ &= \frac{|\boldsymbol{\Sigma}|^{1/2}}{(2\pi)^{n/2}|\mathbf{D}|^{1/2}} \int_0^\infty \frac{\exp\left(-\frac{(n+p+2)\,\overline{\sigma}^2}{2\sigma^2}\right)}{(\sigma^2)^{n/2+1}}\, \mathrm{d}\sigma^2 \\ &= \frac{|\boldsymbol{\Sigma}|^{1/2}\Gamma(n/2)}{|\mathbf{D}|^{1/2}(\pi(n+p+2)\,\overline{\sigma}^2)^{n/2}}. \end{aligned}$$

BIC

Based on (2), we thus have

$$2\mathbb{E}\ell(g_{\mathcal{T}_n}) \simeq -2\ln g(\boldsymbol{y}) = n\ln\left[\pi(n+p+2)\,\overline{\sigma}^2\right] - 2\ln\Gamma(n/2) + \ln|\mathbf{D}| - \ln|\boldsymbol{\Sigma}|.$$

Moreover, the minus of the log-likelihood of $\boldsymbol{Y}$ can be written as

$$\begin{aligned} -\ln g(\boldsymbol{y}\,|\,\boldsymbol{\beta},\sigma^2) &= \frac{\|\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}\|^2}{2\sigma^2} + \frac{n}{2}\ln(2\pi\sigma^2) \\ &= \frac{\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}})\|^2}{2\sigma^2} + \frac{(n+p+2)\,\overline{\sigma}^2}{2\sigma^2} + \frac{n}{2}\ln(2\pi\sigma^2). \end{aligned}$$

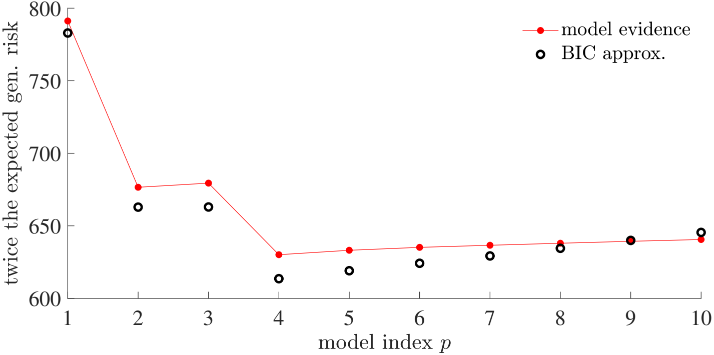
Therefore, the BIC approximation (4) is

$$-2\ln g(\boldsymbol{y}\,|\,\overline{\boldsymbol{\beta}},\overline{\sigma}^2) + (p+1)\ln(n) = n[\ln(2\pi\,\overline{\sigma}^2)+1] + (p+1)\ln(n) + (p+2).$$

BIC for the Polynomial Regression Example

The figure shows the model evidence and its BIC approximation for the polynomial regression example, where we used a hyperparameter $\mathbf{D} = 10^4 \times \mathbf{I}_p$ for the prior density of $\boldsymbol{\beta}$.

We can see that both approximations exhibit a pronounced minimum at $p = 4$, thus identifying the true polynomial regression model.



Bayesian Model Complexity

It is possible to give the model complexity parameter p a Bayesian treatment, in which we define a prior density on the set of all models under consideration.

For example, let $g(p),\ \ p = 1, \ldots, m$ be a prior density on m candidate models.

Treating the model complexity index p as an additional parameter to $\boldsymbol{\theta} \in \mathbb{R}^p$, and applying Bayes' formula, the posterior for $(\boldsymbol{\theta}, p)$ can be written as:

$$\begin{aligned} g(\boldsymbol{\theta}, p \,|\, \tau) &= g(\boldsymbol{\theta} \,|\, p, \tau) \times g(p \,|\, \tau) \\ &= \underbrace{\frac{g(\tau \,|\, \boldsymbol{\theta}, p)\, g(\boldsymbol{\theta} \,|\, p)}{g(\tau \,|\, p)}}_{\text{posterior of } \boldsymbol{\theta} \text{ given model } p} \times \underbrace{\frac{g(\tau \,|\, p)\, g(p)}{g(\tau)}}_{\text{posterior of model } p} . \end{aligned}$$

Model Selection

The model evidence for a fixed p is now interpreted as the prior predictive density of τ, conditional on the model p:

$$g(\tau \mid p) = \int g(\tau \mid \boldsymbol{\theta}, p)\, g(\boldsymbol{\theta} \mid p)\, \mathrm{d}\boldsymbol{\theta},$$

and the quantity $g(\tau) = \sum_{p=1}^{m} g(\tau \mid p)\, g(p)$ is interpreted as the marginal likelihood of all the m candidate models. Finally, a simple method for model selection is to pick the index $\widehat{p}$ with the largest posterior probability:

$$\widehat{p} = \underset{p}{\operatorname{argmax}}\, g(p \mid \tau) = \underset{p}{\operatorname{argmax}}\, g(\tau \mid p)\, g(p).$$

Polynomial Regression (cont.)

Let us revisit the polynomial regression example by giving the parameter $p = 1, \dots, m$, with $m = 10$, a Bayesian treatment. Recall that we used the notation $\tau = \boldsymbol{y}$ in that example. We assume that the prior $g(p) = 1/m$ is flat and uninformative so that the posterior is given by

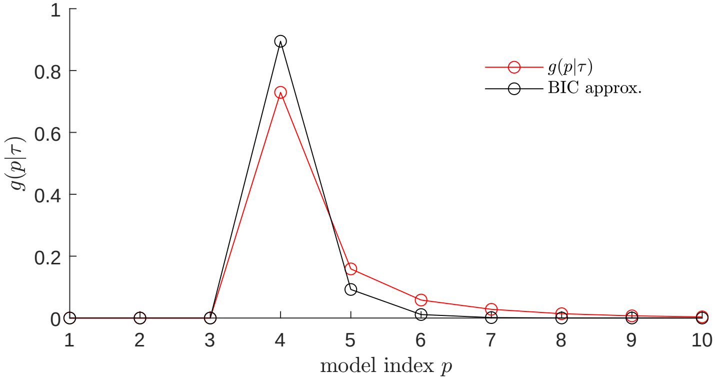
$$g(p \mid \boldsymbol{y}) \propto g(\boldsymbol{y} \mid p) = \frac{|\boldsymbol{\Sigma}|^{1/2}\,\Gamma(n/2)}{|\mathbf{D}|^{1/2}(\pi(n+p+2)\,\overline{\sigma}^2)^{n/2}},$$

where all quantities in $g(\boldsymbol{y} \mid p)$ are computed using the first p columns of $\mathbf{X}$.

The figure shows the resulting posterior density $g(p \mid \boldsymbol{y})$.
The figure also shows the posterior density $\widehat{g}(\boldsymbol{y} \mid p) \big/ \sum_{p=1}^{10} \widehat{g}(\boldsymbol{y} \mid p)$, where

$$\widehat{g}(\boldsymbol{y} \mid p) := \exp\left(-\frac{n[\ln(2\pi\,\overline{\sigma}^2) + 1] + (p+1)\ln(n) + (p+2)}{2}\right)$$

is derived from the BIC approximation. In both cases, there is a clear maximum at the true model $p = 4$.



Bayes Factor

Suppose that we wish to compare two models, say model $p = 1$ and model $p = 2$. Instead of computing the posterior $g(p \,|\, \tau)$ explicitly, we can compare the posterior odds ratio:

$$\frac{g(p = 1 \,|\, \tau)}{g(p = 2 \,|\, \tau)} = \frac{g(p = 1)}{g(p = 2)} \times \underbrace{\frac{g(\tau \,|\, p = 1)}{g(\tau \,|\, p = 2)}}_{\text{Bayes factor } B_{1 \,|\, 2}} .$$

This gives rise to the Bayes factor $B_{i \,|\, j}$, whose value signifies the strength of the evidence in favor of model i over model j. In particular $B_{i \,|\, j} > 1$ means that the evidence in favor for model i is larger.

Savage–Dickey Ratio

Suppose that we have two models.

- Model $p = 2$ has a likelihood $g(\tau \,|\, \mu, \nu, p = 2)$, depending on two parameters.
- Model $p = 1$ has the same functional form for the likelihood but now ν is fixed to some (known) ν_0; that is,

$$g(\tau \,|\, \mu, p = 1) = g(\tau \,|\, \mu, \nu = \nu_0, p = 2).$$

We also assume that the prior information on μ for model 1 is the same as that for model 2, conditioned on $\nu = \nu_0$. That is, we assume $g(\mu \,|\, p = 1) = g(\mu \,|\, \nu = \nu_0, p = 2)$. As model 2 contains model 1 as a special case, the latter is said to be nested inside model 2.

Savage–Dickey Density Ratio

We can formally write:

$$\begin{aligned} g(\tau \,|\, p=1) &= \int g(\tau \,|\, \mu, p=1)\, g(\mu \,|\, p=1)\, \mathrm{d}\mu \\ &= \int g(\tau \,|\, \mu, \nu=\nu_0, p=2)\, g(\mu \,|\, \nu=\nu_0, p=2)\, \mathrm{d}\mu \\ &= g(\tau \,|\, \nu=\nu_0, p=2) = \frac{g(\tau, \nu=\nu_0 \,|\, p=2)}{g(\nu=\nu_0 \,|\, p=2)}. \end{aligned}$$

Hence, the Bayes factor simplifies to

$$B_{1\,|\,2} = \frac{g(\tau \,|\, p=1)}{g(\tau \,|\, p=2)} = \frac{g(\tau, \nu=\nu_0 \,|\, p=2)}{g(\nu=\nu_0 \,|\, p=2)} \bigg/ g(\tau \,|\, p=2) = \frac{g(\nu=\nu_0 \,|\, \tau, p=2)}{g(\nu=\nu_0 \,|\, p=2)}.$$

In other words, $B_{1\,|\,2}$ is the ratio of the posterior density to the prior density of ν, evaluated at $\nu = \nu_0$ and both under the unrestricted model $p = 2$. This is called the Savage–Dickey density ratio.

Bayesian or Frequentist?

Whether to use a classical (frequentist) or Bayesian model is largely a question of convenience.

- Classical inference is useful because it comes with a huge repository of ready-to-use results, and requires no (subjective) prior information on the parameters.
- Bayesian models are useful because the whole theory is based on the elegant Bayes' formula, and uncertainty in the inference (e.g., confidence intervals) can be quantified much more naturally (e.g., credible intervals).

A usual practice is to "Bayesify" a classical model, simply by adding some prior information on the parameters.