Model Validation

The validation of a statistical model is very application dependent. Within the Bayesian framework we have the ability obtain posterior predictive distributions. These can be considered as a general tool for model checking.

Model Validation

The validation of a statistical model is very application dependent. Within the Bayesian framework we have the ability obtain posterior predictive distributions. These can be considered as a general tool for model checking.

Do inferences from the model make sense? This question relates both to the estimation of model parameters and the prediction of future observations. Substantive knowledge about the considered model is essential to provide sensible answers.

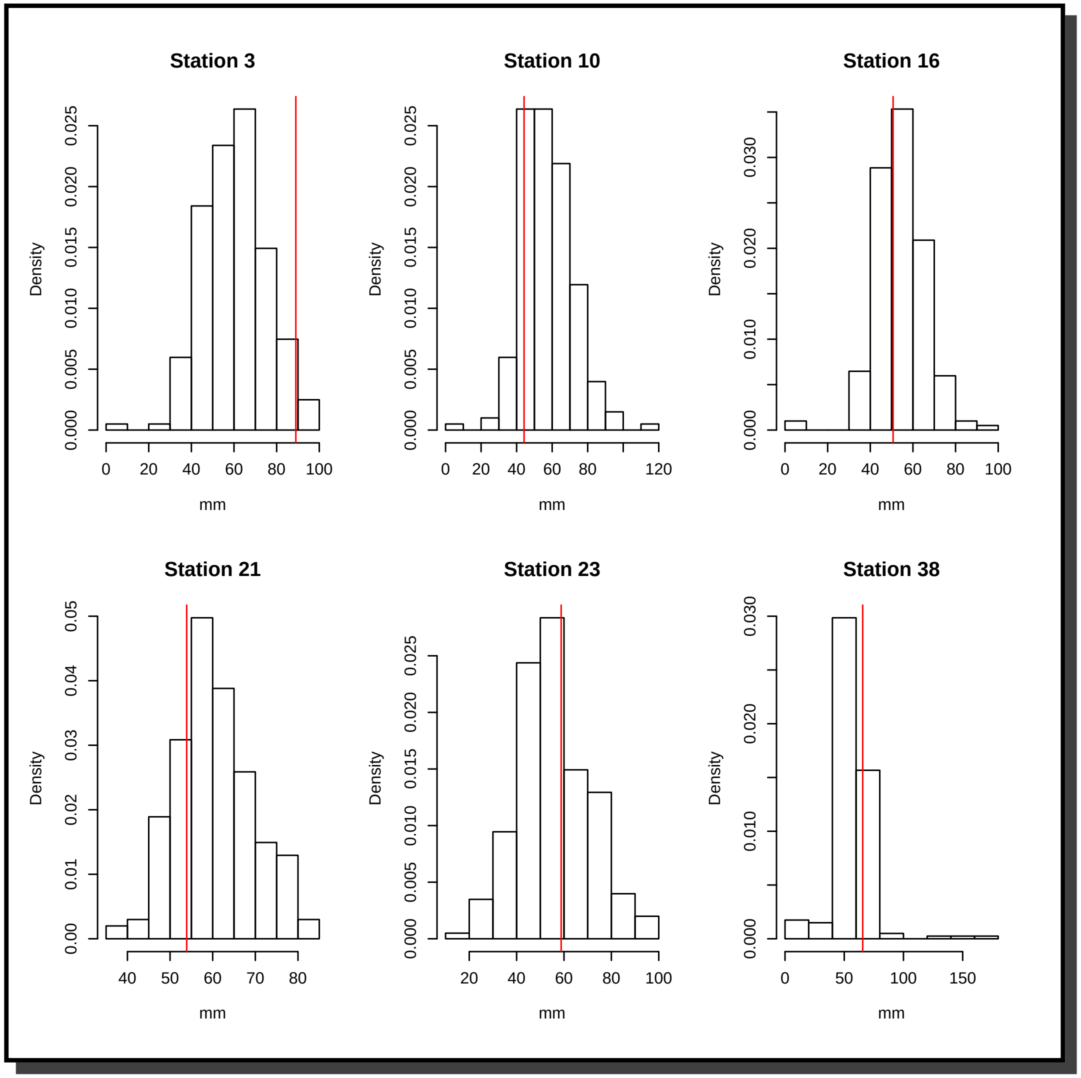
Model Validation

The validation of a statistical model is very application dependent. Within the Bayesian framework we have the ability obtain posterior predictive distributions. These can be considered as a general tool for model checking.

Do inferences from the model make sense? This question relates both to the estimation of model parameters and the prediction of future observations. Substantive knowledge about the considered model is essential to provide sensible answers.

Is the model consistent with the data? A model that fits well should be able to replicate the data. Based on this idea we compare the posterior predictive density $p(z|\boldsymbol{x})$ to the observations. More generally, we can consider the posterior predictive distribution of a function of z.

Example of Model Validation



Example: Predictive distribution of the rainfall over Nebraska in May 1989. The histograms correspond to the predictions obtained by leaving one station out at a time. The dotted red lines are the actual observations

Test Quantities

We can measure the discrepancy between model and data by defining a test quantity $T(X, \theta)$. This is usually a scalar summary involving parameters and data.

Test Quantities

We can measure the discrepancy between model and data by defining a test quantity $T(X, \theta)$. This is usually a scalar summary involving parameters and data.

Following the P-value tradition, we can consider the probability that model predictions have a behavior which is more extreme that the observed data.

$$Pr(T(z,\theta) \geq T(\boldsymbol{x},\theta)|\boldsymbol{x}) = \int_{\Theta}\int_{Z} \mathbb{1}_{T(z,\theta)\geq T(\boldsymbol{x},\theta)} p(z|\boldsymbol{x},\theta)p(\theta|\boldsymbol{x})dzd\theta$$

Large values of this probability will indicate that the model is not replicating the behavior of the data.

Test Quantities

We can measure the discrepancy between model and data by defining a test quantity $T(X, \theta)$. This is usually a scalar summary involving parameters and data.

Following the P-value tradition, we can consider the probability that model predictions have a behavior which is more extreme that the observed data.

$$Pr(T(z, \theta) \geq T(\boldsymbol{x}, \theta)|\boldsymbol{x}) = \int_{\Theta} \int_{Z} \mathbb{1}_{T(z,\theta) \geq T(\boldsymbol{x},\theta)} p(z|\boldsymbol{x}, \theta) p(\theta|\boldsymbol{x}) dz d\theta$$

Large values of this probability will indicate that the model is not replicating the behavior of the data.

In practice the calculation is usually performed using simulated samples from $p(z|\boldsymbol{x})$. This gives great flexibility to consider almost any kind of test quantity.

Consider the problem of testing the hypotheses $H_1, \ldots, H_k$ regarding the parameter space Θ. By placing prior probabilities $p(H_i), i = 1, \ldots, k$ we can compare them using the posterior probabilities

$$p(H_i|\boldsymbol{x}) \propto p(\boldsymbol{x}|H_i)p(H_i), \quad i = 1, \ldots, k$$

We focus on the two hypotheses problem

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1$$

We consider a decision theoretic setting with two actions a_0 and a_1, a_i denotes acceptance of H_i.

The loss function is

$$L(\theta_i, a_i) = \begin{cases} 0 & \text{if } \theta \in \Theta_i \\ k_i & \text{if } \theta \in \Theta_j, j \neq i \end{cases}$$

so that the posterior expected losses of a_0 and a_1 are $k_0 p(\Theta_1|\boldsymbol{x})$ and $k_1 p(\Theta_0|\boldsymbol{x})$ respectively. Thus H_0 is rejected if

$$\frac{k_0}{k_1} > \frac{p(\Theta_0|\boldsymbol{x})}{P(\Theta_1|\boldsymbol{x})} \quad \text{or} \quad P(\Theta_1|\boldsymbol{x}) > \frac{k_1}{k_0 + k_1}$$

if Θ_0 and Θ_1 are a partition of Θ. Clearly, if the loss is the same for both types of errors, then the optimal choice is that of the hypothesis with the highest posterior probability.

The posterior odds are given by

$$\frac{p(\Theta_0|\boldsymbol{x})}{p(\Theta_1|\boldsymbol{x})} = \frac{p(\boldsymbol{x}|H_0)}{p(\boldsymbol{x}|H_1)}\frac{p(H_0)}{p(H_1)}$$

The factor $B_{01} = p(\boldsymbol{x}|H_0)/p(\boldsymbol{x}|H_1)$ updates the prior odds to the posterior odds. This is known as the Bayes Factor in favor of H_0.

When H_0 and H_1 are both simple hypotheses, the Bayes factor corresponds to the likelihood ratio $B_{01} = p(\boldsymbol{x}|\theta_0)/p(\boldsymbol{x}|\theta_1)$. In general, if $g_i(\theta)$ is the prior distribution for θ under H_i, then

$$B_{01} = \frac{\int_{\Theta_0} p(\boldsymbol{x}|\theta)g_0(\theta)d\theta}{\int_{\Theta_1} p(\boldsymbol{x}|\theta)g_1(\theta)d\theta} = \frac{m_0(\boldsymbol{x})}{m_1(\boldsymbol{x})}$$

Example: Assume that $X \sim N(\theta, \sigma^2)$ where σ^2 is known and let $\pi(\theta) \propto 1$. Then, if

$$H_0 : \theta \leq \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0$$

then

$$p(\Theta_0|\boldsymbol{x}) = p(\theta \leq \theta_0|\boldsymbol{x}) = \Phi((\theta_0 - x)/\sigma)$$

this coincides with the P-value that is produced by the classical test which is defined as the probability of observing a sample more extreme than the actual data.

For many one-sided tests, P-values can be seen to be equivalent to the posterior probability of the null hypothesis.

Example: Assume that $X_1, \ldots, X_n$ is a random sample from $N(\theta, \sigma^2)$ where σ^2 is known and consider the hypotheses

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

Consider a prior that puts probability π on θ_0 and $(1-\pi)g_1(\theta)$ on $\theta \in \Theta_1$, where g_1 is a proper density. Then the posterior probability of Θ_0 is

$$p(\Theta_0|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta_0)\pi}{p(\boldsymbol{x}|\theta_0)\pi + m_1(\boldsymbol{x})(1-\pi)} = \left[1 + \frac{(1-\pi)}{\pi}\frac{m_1(\boldsymbol{x})}{f(\boldsymbol{x}|\theta_0)}\right]^{-1}$$

$$m_1(\boldsymbol{x}) = \int_{\theta \neq \theta_0} p(\boldsymbol{x}|\theta) g_1(\theta) d\theta$$

and the Bayes Factor is

$$B_{01} = \frac{p(\boldsymbol{x}|\theta_0)}{m_1(\boldsymbol{x})}$$

Example: In the previous example suppose that g_1 is $N(\theta|\theta_0, \sigma^2)$ and $\pi = 1/2$. Then

$$p(\Theta_0|\boldsymbol{x}) = \left[1 + \frac{\exp\{\frac{n}{n+1}\frac{z^2}{2}\}}{\sqrt{1+n}}\right]^{-1}$$

where $z = \sqrt{n}|\bar{x} - \theta_0|/\sigma$. The table shows the comparison between the P-value and the posterior probability of H_0 for some values of z.

$\|z\|$	P-value	Posterior probability			
		$n = 1$	$n = 10$	$n = 100$	$n = 1000$
1.96	0.05	0.35	0.37	0.60	0.80
2.576	0.01	0.21	0.14	0.27	0.53

The table shows that there is an enormous conflict between the P-values and the posterior probabilities of H_0. In fact a P-value of 0.05 suggest pretty strong evidence against the Null, but the posterior probability ranges from 35% to 80%.

The table shows that there is an enormous conflict between the P-values and the posterior probabilities of H_0. In fact a P-value of 0.05 suggest pretty strong evidence against the Null, but the posterior probability ranges from 35% to 80%.

The previous results depend on the particular choice of prior. A lower bound for $p(\Theta_0|\boldsymbol{x})$ can be obtained for g_1 being any distribution on $\theta \neq \theta_0$. Let $r(\boldsymbol{x}) = \sup_{\theta \neq \theta_0} f(\boldsymbol{x}|\theta)$, then

$$p(\Theta_0|\boldsymbol{x}) \geq \left[1 + \frac{r(\boldsymbol{x})}{f(\boldsymbol{x}|\theta_0)}\right]^{-1} = \left[1 + \exp\{z^2/2\}\right]^{-1}$$

The corresponding bound for the Bayes Factor is

$$B_{01} \geq \frac{f(\boldsymbol{x}|\theta)}{r(\boldsymbol{x})} = \exp\{-z^2/2\}$$

This bound is just the minimum likelihood ratio of H_0 to H_1.

The values for these lower bounds are

$\lvert z \rvert$	P-value	Bound on $P(\Theta_0 \lvert \boldsymbol{x})$	Bound on B_{01}
1.96	0.05	0.205	$1/6.83$
2.576	0.01	0.035	$1/27.60$

Implying that, for P-value of 5%, the likelihood ratio of H_0 to H_1 is, at least, $1/6.83$. So the data favors H_1 by a factor of, at most, 6.83.

Consider m models, $M_1, \ldots, M_m$ defined by the likelihood $p_i(\boldsymbol{x}|\theta_i)$ and prior $p_i(\theta_i)$. We compare M_i to M_j using the Bayes Factor

$$B_{ij}(\boldsymbol{x}) = \frac{m_i(\boldsymbol{x})}{m_j(\boldsymbol{x})}, \qquad p(M_k|\boldsymbol{x}) = \left(\sum_{i=1}^{m} \frac{p(M_i)}{p(M_k)} B_{jk}\right)^{-1},$$

$$m_k(\boldsymbol{x}) = \int_{\Theta_k} p_k(\boldsymbol{x}|\theta_k) p_k(\theta_k) d\theta_k$$

Improper Priors

Consider m models, $M_1, \ldots, M_m$ defined by the likelihood $p_i(\boldsymbol{x}|\theta_i)$ and prior $p_i(\theta_i)$. We compare M_i to M_j using the Bayes Factor

$$B_{ij}(\boldsymbol{x}) = \tfrac{m_i(\boldsymbol{x})}{m_j(\boldsymbol{x})}, \qquad p(M_k|\boldsymbol{x}) = \left(\textstyle\sum_{i=1}^{m} \tfrac{p(M_i)}{p(M_k)} B_{jk}\right)^{-1},$$
$$m_k(\boldsymbol{x}) = \textstyle\int_{\Theta_k} p_k(\boldsymbol{x}|\theta_k) p_k(\theta_k) d\theta_k$$

If the priors $p_k(\theta_k)$ are defined up to a proportionality constant c_k, then the Bayes Factors will depend on the ratio of such constants.

If the priors are proper densities, then we have

$$c_k^{-1} = \int_{\Theta_k} p_k(\theta_k) d\theta_k$$

and then the Bayes Factor is uniquely defined. For improper priors, the Bayes Factor depends on the ratio of two unknown constants.

Bayesian Information Criterion

A popular method for model comparison is the Bayesian Information Criterion (BIC) that is given as

$$BIC = -2 \log f(\boldsymbol{x}|\hat{\theta}) + p \log n$$

where n is the sample size, $\hat{\theta}$ is the MLE and p is the number of parameters in the model. Changing $\log n$ for 2 produces the AIC.

Bayesian Information Criterion

A popular method for model comparison is the Bayesian Information Criterion (BIC) that is given as

$$BIC = -2\log f(\boldsymbol{x}|\hat{\theta}) + p\log n$$

where n is the sample size, $\hat{\theta}$ is the MLE and p is the number of parameters in the model. Changing $\log n$ for 2 produces the AIC.

It can be shown that, for two given non-hierarchical models, the difference of their BICs approximates the log of the bayes factor. For hierarchical models the penalty factor $p\log n$ is not clearly specified.

Deviance Information Criterion

The Deviance Information Criterion (DIC) is a generalization of the AIC for hierarchical models. This is achieved by estimating the complexity or effective number of parameters in the model.

The deviance statistics is given by

$$D(\theta) = -2\log f(y|\theta) + 2\log h(y)$$

where $h(y)$ is a standardizing function. The DIC is defined as

$$DIC = \overline{D} + p_D = \overline{D} + (\overline{D} - D(\bar{\theta})) = 2\overline{D} - D(\bar{\theta})$$

where the overline denotes posterior expectation. Smaller values of DIC indicate better-fitting models.

Posterior Predictive Loss Criteria

Gelfand and Ghosh (1998) propose a criterion based on the predictive ability of a given model. The idea is to compare the posterior predictive distribution to the actual observed data to assess the suitability of the model.

Denote the observed data as $\boldsymbol{x}$ and z as replicates of the observed data. Let $\mu_l = E(z_l|\boldsymbol{x})$ and $\sigma_l^2 =$var$(z_l|\boldsymbol{x})$. Then

$$D = G + P \quad G = \underbrace{\sum_{l=1}^{n} (\mu_l - x_l)^2}_{\text{Goodness of Fit}} \quad \text{and} \;\; P = \underbrace{\sum_{l=1}^{n} \sigma_l^2}_{\text{Penalty}}$$

D seeks to reward goodness of fit penalizing complexity. So the smaller the D the better.