Session 4: Bayesian Model Assessment

4.1 Posterior Model Probabilities

4.2 Criterion-Based Methods

4.3 Conditional Predictive Ordinate

4.4 The Other Methods

4.5 Bayesian Model Diagnostics

Overview

In this fourth and final session, we will discuss Bayesian model comparison and Bayesian model diagnostics in survival analysis. The topics covered include Bayes factors and posterior model probabilities, the Bayesian Information Criterion (BIC), the Conditional Predictive Ordinate (CPO), and the L measure for Bayesian model comparison; and Bayesian latent residuals and prequential methods for Bayesian model diagnostics. Detailed examples using real data are presented, and issues involving the computational implementation are addressed.

4.1 Posterior Model Probabilities

♠ **General Notation**

- **Model space**: $\mathcal{M}$
- **Model index**: m, a specific model in $\mathcal{M}$.
- **Model parameter vector**: $\boldsymbol{\theta}^{(m)}$ associated with model m.
- **Posterior model probability of model** m:

$$p(m|D) = \frac{p(D|m)p(m)}{\sum_{m\in\mathcal{M}} p(D|m)p(m)},$$

where D denotes the data,

$$p(D|m) = \int L(\boldsymbol{\theta}^{(m)}|D)\pi(\boldsymbol{\theta}^{(m)})\, d\boldsymbol{\theta}^{(m)},$$

$L(\boldsymbol{\theta}^{(m)}|D)$ is the likelihood, and $p(m)$ denotes the prior probability of model m.

♠ Variable Selection in the Cox Model

• Difficulties

Bayesian variable selection is often difficult to carry out because of the challenge in

(i) specifying prior distributions for the regression parameters for all possible models in $\mathcal{M}$;

(ii) specifying a prior distribution on the model space; and

(iii) computations.

• **Specific Notation**

Let p denote the number of covariates for the full model and let $\mathcal{M}$ denote the model space. We enumerate the models in $\mathcal{M}$ by $m = 1, 2, \ldots, \mathcal{K}$, where $\mathcal{K}$ is the dimension of $\mathcal{M}$ and model $\mathcal{K}$ denotes the full model. Also, let $\boldsymbol{\beta}^{(\mathcal{K})} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ denote the regression coefficients for the full model including an intercept, and let $\boldsymbol{\beta}^{(m)}$ denote a $p_m \times 1$ vector of regression coefficients for model m with an intercept, and a specific choice of $p_m - 1$ covariates. We write $\boldsymbol{\beta}^{(\mathcal{K})} = (\boldsymbol{\beta}^{(m)'}, \boldsymbol{\beta}^{(-m)'})'$, where $\boldsymbol{\beta}^{(-m)}$ is $\boldsymbol{\beta}^{(\mathcal{K})}$ with $\boldsymbol{\beta}^{(m)}$ deleted.

• **Model**

We consider the semiparametric model described in Session 1, which is based on a discretized gamma process on the baseline hazard function with independent increments. Under model m, the likelihood can be written as

$$L(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}|D^{(m)}) = \prod_{j=1}^{J} \left\{ \exp\{-\delta_j(a_j + b_j)\} \times \prod_{k \in \mathcal{D}_j} \left[1 - \exp\{-\eta_k^{(m)} T_j\}\right] \right\},$$

where $\eta_k^{(m)} = \exp(\boldsymbol{x}_k^{(m)'} \boldsymbol{\beta}^{(m)})$, $x_k^{(m)}$ is a $p_m \times 1$ vector of covariates for the i^{th} individual under model m, $X^{(m)}$ denotes the $n \times p_m$ covariate matrix of rank p_m, $D^{(m)} = (n, \boldsymbol{y}, X^{(m)}, \boldsymbol{\nu})$ denotes the data under model m,

$$a_j = \sum_{l=j+1}^{J} \sum_{k \in \mathcal{D}_l} \eta_k^{(m)} (s_{l-1} - s_{j-1}), \quad b_j = \sum_{l=j}^{J} \sum_{k \in \mathcal{C}_l} \eta_k^{(m)} (s_l - s_{j-1}),$$

$T_j = (s_j - s_{j-1}) \sum_{l=1}^{j} \delta_l$, $\mathcal{D}_j$ be the set of subjects failing, $\mathcal{C}_j$ is the set of subjects that are censored, and $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_J)'$. Here, $\boldsymbol{\delta}$ can be viewed as a nuisance parameter, which does not depend on m.

• **Power Prior for Model Parameters**

We use the power prior

$$\pi(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}, a_0 | D_0^{(m)}) \propto L(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta} | D_0^{(m)})^{a_0} \pi_0(\boldsymbol{\beta}^{(m)} | c_0) \times \pi_0(\boldsymbol{\delta} | \boldsymbol{\theta}_0) \pi(a_0 | \alpha_0, \lambda_0),$$

where $D_0^{(m)} = (n_0, \boldsymbol{y}_0, X_0^{(m)}, \boldsymbol{\nu}_0)$ is the historical data under model m,

$$\pi_0(\boldsymbol{\delta} | \boldsymbol{\theta}_0) \propto \prod_{j=1}^{J} \delta_j^{f_{0j}-1} \exp\{-\delta_j g_{0j}\},$$

$$\pi(a_0 | \alpha_0, \lambda_0) \propto a_0^{\alpha_0 - 1} (1 - a_0)^{\lambda_0 - 1},$$

and $\boldsymbol{\theta}_0 = (f_{01}, g_{01}, \ldots, f_{0J}, g_{0J})'$ and (α_0, λ_0) are prespecified hyperparameters. For the purposes of prior elicitation, it is easier to work with $\mu_0 = \alpha_0/(\alpha_0 + \lambda_0)$ and $\sigma_0^2 = \mu_0(1 - \mu_0)(\alpha_0 + \lambda_0 + 1)^{-1}$.

An attractive feature of the power prior for $\boldsymbol{\beta}^{(m)}$ in variable selection problems is that it is semiautomatic in the sense that one only needs a one time input of $(D_0^{(m)}, c_0, \boldsymbol{\theta}_0, \alpha_0, \lambda_0)$ to generate the prior distributions for all $m \in \mathcal{M}$.

• Prior Distribution on the Model Space

Let the initial prior for the model space be denoted by $p_0(m)$. Given the historical data $D_0^{(m)}$, the prior probability of model m for the current study based on an update of y_0 via Bayes theorem is given by

$$p(m) \equiv p(m|D_0^{(m)}) = \frac{p(D_0^{(m)}|m)p_0(m)}{\sum_{m\in\mathcal{M}} p(D_0^{(m)}|m)p_0(m)},$$

where $p(D_0|m) = \int L(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}|D_0^{(m)})\pi_0(\boldsymbol{\beta}^{(m)}|d_0)\pi_0(\boldsymbol{\delta}|\boldsymbol{\kappa}_0)d\beta^{(m)}\ d\boldsymbol{\delta}$, $L(\boldsymbol{\delta}, \boldsymbol{\beta}^{(m)}|D_0^{(m)})$ is the likelihood function of the parameters based on $D_0^{(m)}$, $\pi_0(\boldsymbol{\beta}^{(m)}|d_0)$ is the initial prior for $\boldsymbol{\beta}^{(m)}$, and $\pi_0(\boldsymbol{\delta}|\kappa_0)$ is the initial prior for $\boldsymbol{\delta}$. Specifically, we take $\pi_0(\boldsymbol{\beta}^{(m)}|d_0)$ to be a $N_{p_m}(0, d_0 W_0^{(m)})$, where $W_0^{(m)}$ is the submatrix of the diagonal matrix $W_0^{(\mathcal{K})}$ corresponding to model m, and

$$\pi_0(\boldsymbol{\delta}|\boldsymbol{\kappa}_0) \propto \prod_{j=1}^{J} \delta_j^{f_{0j}^*-1}\ \exp\left\{-\delta_j g_{0j}^*\right\},$$

where $\boldsymbol{\kappa}_0 = (f_{01}^*, g_{01}^*, \ldots, f_{0J}^*, g_{0J}^*)'$.

Properties

- $p(m)$ corresponds to the usual Bayesian update of $p_0(m)$ using $D_0^{(m)}$ as the data.
- As $d_0 \to 0$, $p(m)$ reduces to $p_0(m)$. Therefore, as $d_0 \to 0$, the historical data $D_0^{(m)}$ have a minimal impact in determining $p(m)$.
- As $d_0 \to \infty$, $\pi_0(\boldsymbol{\beta}^{(m)}|d_0)$ plays a minimal role in determining $p(m)$, and in this case, the historical data play a larger role in determining $p(m)$.
- The parameter d_0 thus serves as a tuning parameter to control the impact of $D_0^{(m)}$ on the prior model probability $p(m)$.

A Note

When there is little information about the relative plausibility of the models at the initial stage, taking $p_0(m) = \frac{1}{\mathcal{K}}$, $m = 1, 2, \ldots, \mathcal{K}$, *a priori* is a reasonable "neutral" choice.

• **Computing Prior Model Probabilities**

Suppose that under the full model, we have a sample $\{(\boldsymbol{\beta}_{0,l}^{(\mathcal{K})}, \boldsymbol{\delta}_{0,l}),\ l = 1, 2, \ldots, L\}$ from

$$\pi_0(\boldsymbol{\beta}^{(\mathcal{K})}, \boldsymbol{\delta}|D_0^{(\mathcal{K})}) \propto \pi_0^*(\boldsymbol{\beta}^{(\mathcal{K})}, \boldsymbol{\delta}|D_0^{(\mathcal{K})}),$$

where

$$\pi_0^*(\boldsymbol{\beta}^{(\mathcal{K})}, \boldsymbol{\delta}|D_0^{(\mathcal{K})}) = L(\boldsymbol{\beta}^{(\mathcal{K})}, \boldsymbol{\delta}|D_0^{(\mathcal{K})})\pi_0(\boldsymbol{\beta}^{(\mathcal{K})}|d_0)\pi_0(\boldsymbol{\delta}|\boldsymbol{\kappa}_0).$$

Then, the prior probability of model m can be estimated by

$$\hat{p}(m) = \hat{p}(m|D_0^{(m)}) = \frac{\frac{1}{L}\sum_{l=1}^{L} \frac{\pi_0^*(\boldsymbol{\beta}_{0,l}^{(m)}, \boldsymbol{\delta}_{0,l}|D_0^{(m)}) w_0(\boldsymbol{\beta}_{0,l}^{(-m)}|\beta_{0,l}^{(m)})}{\pi_0^*(\boldsymbol{\beta}_{0,l}^{(\mathcal{K})}, \boldsymbol{\delta}_{0,l}|D_0^{(\mathcal{K})})} p_0(m)}{\frac{1}{L}\sum_{j=1}^{\mathcal{K}}\sum_{l=1}^{L} \frac{\pi_0^*(\boldsymbol{\beta}_{0,l}^{(j)}, \boldsymbol{\delta}_{0,l}|D_0^{(j)}) w_0(\boldsymbol{\beta}_{0,l}^{(-j)}|\boldsymbol{\beta}_{0,l}^{(j)})}{\pi_0^*(\boldsymbol{\beta}_{0,l}^{(\mathcal{K})}, \boldsymbol{\delta}_{0,l}|D_0^{(\mathcal{K})})} p_0(j)},$$

where

$$\pi_0^*(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}|D_0^{(\mathcal{K})}) = L(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}|D_0^{(m)})\pi_0(\boldsymbol{\beta}^{(m)}|d_0)\pi_0(\boldsymbol{\delta}|\boldsymbol{\kappa}_0),$$

and $w_0(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})$ is a *completely* known conditional density whose support is contained in, or equal to, the support of the conditional

density of $\boldsymbol{\beta}^{(-m)}$ given $\boldsymbol{\beta}^{(m)}$ with respect to the full model joint prior distribution.

Choice of w_0

We take

$$w_0(\boldsymbol{\beta}^{(-m)}|\boldsymbol{\beta}^{(m)}) = (2\pi)^{-(p-p_m)/2}|\tilde{\Sigma}_{11.2m}|^{-1/2}$$
$$\times \exp\Big\{ -\tfrac{1}{2}(\boldsymbol{\beta}^{(-m)} - \tilde{\boldsymbol{\mu}}_{11.2m})'\tilde{\Sigma}^{-1}_{11.2m}(\boldsymbol{\beta}^{(-m)} - \tilde{\boldsymbol{\mu}}_{11.2m})\Big\},$$

where

$$\tilde{\Sigma}_{11.2m} = \tilde{\Sigma}_{11m} - \tilde{\Sigma}_{12m}\tilde{\Sigma}^{-1}_{22m}\tilde{\Sigma}'_{12m},$$

$\tilde{\Sigma}_{11m}$ is the covariance matrix from the marginal distribution of $\boldsymbol{\beta}^{(-m)}$, $\tilde{\Sigma}_{12m}$ consists of the covariances between $\boldsymbol{\beta}^{(-m)}$ and $\boldsymbol{\beta}^{(m)}$, and $\tilde{\Sigma}_{22m}$ is the covariance matrix of the marginal distribution of $\boldsymbol{\beta}^{(m)}$ with respect to the joint normal distribution $N_p(\tilde{\boldsymbol{\beta}}_0, \tilde{\Sigma}_0)$ for $\boldsymbol{\beta}^{(\mathcal{K})}$. Also

$$\tilde{\boldsymbol{\mu}}_{11.2m} = \tilde{\boldsymbol{\mu}}^{(-m)} + \tilde{\Sigma}_{12m}\tilde{\Sigma}^{-1}_{22m}(\boldsymbol{\beta}^{(m)} - \tilde{\boldsymbol{\mu}}^{(m)}),$$

where $\tilde{\boldsymbol{\mu}}^{(-m)}$ is the mean of the normal marginal distribution of $\boldsymbol{\beta}^{(-m)}$ and $\tilde{\boldsymbol{\mu}}^{(m)}$ is the mean of the normal marginal distribution of $\boldsymbol{\beta}^{(m)}$.

Features

There are several nice features of the above Monte Carlo procedure:

(i) we need *only one random draw* from $\pi_0(\boldsymbol{\beta}^{(\mathcal{K})}, \boldsymbol{\delta}|D_0^{(\mathcal{K})})$, which greatly eases the computational burden;

(ii) it is more numerically stable since we calculate ratios of the densities; and

(iii) $\pi_0(\boldsymbol{\beta}^{(\mathcal{K})}, \boldsymbol{\delta}|D_0^{(\mathcal{K})})$ plays the role of a ratio importance sampling density (see Chen and Shao, 1997) which needs to be known only up to a normalizing constant since this common constant cancels out in the calculation.

• Computing Posterior Model Probabilities

Difficulty

Computing the posterior model probability $p(D^{(m)}|m)$ requires evaluating the ratio of two analytically intractable integrals, one from the prior distribution and another from the posterior distribution.

A Key Theoretical Result

It can be shown that the posterior probability of model m is given by

$$p(m|D^{(m)}) = \frac{\frac{\pi(\boldsymbol{\beta}^{(-m)}=0|D^{(\mathcal{K})})}{\pi(\boldsymbol{\beta}^{(-m)}=0|D_0^{(\mathcal{K})})}p(m)}{\sum_{j=1}^{\mathcal{K}} \frac{\pi(\boldsymbol{\beta}^{(-j)}=0|D^{(\mathcal{K})})}{\pi(\boldsymbol{\beta}^{(-j)}=0|D_0^{(\mathcal{K})})}p(j)},$$

where $\pi(\boldsymbol{\beta}^{(-m)}=0|D_0^{(\mathcal{K})})$ and $\pi(\boldsymbol{\beta}^{(-m)}=0|D^{(\mathcal{K})})$ are the marginal prior and posterior densities of $\boldsymbol{\beta}^{(-m)}$ evaluated at $\boldsymbol{\beta}^{(-m)}=0$ for $m=1,2,\ldots,\mathcal{K}$.

Monte Carlo Estimation

Due to the complexity of the prior and posterior distributions, the analytical forms of $\pi(\boldsymbol{\beta}^{(-m)}|D_0^{(\mathcal{K})})$ and $\pi(\boldsymbol{\beta}^{(-m)}|D^{(\mathcal{K})})$ are not available. However, we can adopt the importance-weighted marginal posterior density estimation (IWMDE) method of Chen (1994) to estimate these marginal prior and posterior densities.

The IWMDE method requires using only two respective Markov chain Monte Carlo (MCMC) samples from the prior and posterior distributions for the full model, making the computation of complicated posterior model probabilities feasible.

It directly follows from the IWMDE method that a simulation consistent estimator of $\pi(\boldsymbol{\beta}^{(-m)} = 0|D^{(\mathcal{K})})$ is given by

$$\hat{\pi}(\boldsymbol{\beta}^{(-m)} = 0|D^{(\mathcal{K})})$$

$$= \frac{1}{L}\sum_{l=1}^{L} w(\boldsymbol{\beta}_l^{(-m)}|\boldsymbol{\beta}_l^{(m)}) \frac{\pi(\boldsymbol{\beta}_l^{(m)}, \boldsymbol{\beta}^{(-m)} = 0, \boldsymbol{\delta}_l, a_{0,l}|D^{(\mathcal{K})})}{\pi(\boldsymbol{\beta}_l^{(\mathcal{K})}, \boldsymbol{\delta}_l, a_{0,l}|D^{(\mathcal{K})})},$$

where $w(\boldsymbol{\beta}^{(-m)}|\boldsymbol{\beta}^{(m)})$ is a completely known conditional density of $\boldsymbol{\beta}^{(-m)}$ given $\boldsymbol{\beta}^{(m)}$, whose support is contained in, or equal to, the support of the conditional density of $\boldsymbol{\beta}^{(-m)}$ given $\boldsymbol{\beta}^{(m)}$ with respect to the full model joint posterior distribution,
$\{(\boldsymbol{\beta}_l^{(\mathcal{K})}, \boldsymbol{\delta}_l, a_{0,l}), l = 1, 2, \ldots, L\}$ is a sample from the joint posterior distribution $\pi(\boldsymbol{\beta}^{(\mathcal{K})}, \boldsymbol{\delta}, a_0|D^{(\mathcal{K})})$. To construct a good $w(\boldsymbol{\beta}^{(-m)}|\boldsymbol{\beta}^{(m)})$, we can use a procedure similar to the one used to construct $w_0(\boldsymbol{\beta}^{(-m)}|\boldsymbol{\beta}^{(m)})$ for calculating the prior model probabilities.

• **Example: Multiple myeloma data**

Objectives:

Our main goals in this example are to illustrate the prior elicitation and variable selection techniques and to examine the sensitivity of the posterior probabilities to the choices of (μ_0, σ_0^2), c_0, and d_0.

The Data:

We have two similar studies in the multiple myeloma study E2479, Study 1 (*historical*), and Study 2 (*current*). Our analysis uses $p = 8$ covariates. These are blood urea nitrogen (x_1), hemoglobin (x_2), platelet count (x_3) (1 if normal, 0 if abnormal), age (x_4), white blood cell count (x_5), bone fractures (x_6), percentage of the plasma cells in bone marrow (x_7), and serum calcium (x_8). A total of $n = 339$ observations were available from Study 2, with 8 observations being right censored, while Study 1 consisted of $n_0 = 65$ observations of which 17 were right censored.

Initial Priors and Others:

- We take $W_0^{(\mathcal{K})}$ to be the diagonal elements of the inverse of the Fisher information matrix based on the Cox's partial likelihood where $\mathcal{K} = 2^8 = 256$ in this example.
- We use a uniform initial prior on the model space, that is, $p_0(m) = \frac{1}{\mathcal{K}}$ for $m = 1, 2, \ldots, \mathcal{K}$.
- We take $\boldsymbol{\theta}_0 = \boldsymbol{\kappa}_0$ and use $f_{0j} = s_j - s_{j-1}$ if $s_j - s_{j-1} \geq 1$ and $f_{0j} = 1.1$ if $s_j - s_{j-1} < 1$, and $g_{0j} = 0.001$. For the last interval, we take $g_{0j} = 10$ for $j = J$ since very little information in the data is available for this last interval. The above choices of f_{0j} and g_{0j} ensure the log-concavity of $\pi_0(\boldsymbol{\delta}|\boldsymbol{\theta}_0)$, as this is required in sampling $\boldsymbol{\delta}$ from its conditional prior and posterior distributions.
- We use $J = 28$, with the intervals chosen so that with the combined datasets from the historical and current data, at least one failure or censored observation falls in each interval.
- A stepwise variable selection procedure in SAS for the current study yields $(x_2, x_3, x_4, x_7, x_8)$ as the top model.

Choice of c_0

TABLE 4.1. The Posterior Model Probabilities for $(\mu_0, \sigma_0^2) = (0.5, 0.004)$, $d_0 = 3$ and Various Choices of c_0.

c_0	m	$p(m)$	$p(D\|m)$	$p(m\|D)$
3	(1234578)	0.015	0.436	0.769
10	(1234578)	0.015	0.310	0.679
30	(1234578)	0.015	0.275	0.657

Note that we use (1234578) to denote the model indexed by $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$.

Table 4.1 gives the model with the largest posterior probability using $(\mu_0, \sigma_0^2) = (0.5, 0.004)$, (i.e., $\alpha_0 = \lambda_0 = 30$) for several values of c_0. For each value of c_0 in Table 4.1, the model $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ obtains the largest posterior probability, and thus model choice is not sensitive to these values. In addition, for $d_0 = 3$ and for any $c_0 \geq 3$, the $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ model obtains the largest posterior probability. Although not shown in Table 4.1, values of $c_0 < 3$ do not yield $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ as the top model. Thus, model choice may become sensitive to the choice of c_0 when $c_0 < 3$.

Choice of d_0

TABLE 4.2. The Posterior Model Probabilities for $(\mu_0, \sigma_0^2) = (0.5, 0.004)$, $c_0 = 3$ and Various Choices of d_0.

d_0	m	$p(m)$	$p(D\|m)$	$p(m\|D)$
5	(1234578)	0.011	0.436	0.750
10	(1234578)	0.005	0.436	0.694
30	(1234578)	0.001	0.436	0.540

From Table 4.2, we see how the prior model probability is affected as d_0 is changed. In each case, the true model obtains the largest posterior probability. Under the settings of Table 4.2, the $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ model obtains the largest prior probability when $d_0 \geq 3$. With values of $d_0 < 3$, however, model choice may be sensitive to the choice of d_0. For example, when $d_0 = 0.0001$ and $c_0 = 10$, the top model is $(x_1, x_2, x_4, x_5, x_7, x_8)$ with posterior probability of 0.42 and the second-best model is $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ with posterior probability of 0.31. Finally, we mention that as both c_0 and d_0 become large, the $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ model obtains the largest posterior model probability.

Incorporation of Historical Data ((μ_0, σ_0^2))

TABLE 4.3. The Posterior Model Probabilities for $c_0 = 10$, $d_0 = 10$ and Various Choices of (μ_0, σ_0^2).

(μ_0, σ_0^2)	m	$p(m)$	$p(D\|m)$	$p(m\|D)$
(0.5, 0.008)	(1234578)	0.005	0.274	0.504
(0.5, 0.004)	(1234578)	0.005	0.310	0.558
(0.98, 0.0004)	(1234578)	0.005	0.321	0.572

Table 4.3 shows a sensitivity analysis with respect to (μ_0, σ_0^2). Under these settings, model choice is not sensitive to the choice of (μ_0, σ_0^2). We see that in each case, $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ obtains the largest posterior probability. In addition, there is a monotonic increase in the posterior model probability as more weight is given to the historical data.

4.2 Criterion-Based Methods

♠ An Introduction

- Many of the proposed Bayesian methods for model comparison usually rely on posterior model probabilities or Bayes factors, and it is well known that to use these methods, proper prior distributions are needed. It is also well known that posterior model probabilities are generally sensitive to the choices of prior parameters, and thus one cannot simply select vague proper priors to get around the elicitation issue.
- Criterion based methods do not require proper prior distributions in general, and thus have an advantage over posterior model probabilities in this sense.

♠ L Measure

The L measure criterion is constructed from the posterior predictive distribution of the data, and can be written as a sum of two components, one involving the means of the posterior predictive distribution and the other involving the variances.

Consider an experiment that yields the data $\boldsymbol{y} = (y_1, \ldots, y_n)'$. Denote the joint sampling density of the y_i's by $f(\boldsymbol{y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of indexing parameters.

Let $\boldsymbol{z} = (z_1, \ldots, z_n)'$ denote future values of a replicate experiment. That is, $\boldsymbol{z}$ is a future response vector with the same sampling density as $\boldsymbol{y}|\theta$

We note that $\boldsymbol{y}$ and $\boldsymbol{z}$ may represent a transformation of the original data. For example, in survival analysis, it is common to take the logarithms of the survival times, and thus y would represent the logs of the survival times.

• **Case 1: $\boldsymbol{y}$ is fully observed**

When $\boldsymbol{y}$ is fully observed, the L measure can be calculated as

$$L(y) = \sum_{i=1}^{n} \mathrm{Var}(z_i|y_i) + \nu \sum_{i=1}^{n} (\mu_i - y_i)^2,$$

where $0 \leq \nu < 1$, $\mu_i = E(z_i|\boldsymbol{y})$, and $[z_i|\boldsymbol{y}]$ denotes the posterior predictive distribution with a density proportion to

$$f(\boldsymbol{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y}).$$

Small values of the L measure imply a good model.

• **Case 2: $\boldsymbol{y}$ is censored**

Let $\boldsymbol{y} = (\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{cens}})$, where $\boldsymbol{y}_{\text{obs}}$ denotes the completely observed components of $\boldsymbol{y}$, and $\boldsymbol{y}_{\text{cens}}$ denotes the censored components. Here, we assume that $\boldsymbol{y}_{\text{cens}}$ is a random quantity and $\boldsymbol{a}_l < \boldsymbol{y}_{\text{cens}} < \boldsymbol{a}_r$, where $\boldsymbol{a}_l$ and $\boldsymbol{a}_r$ are known.

Let $D = (n, \boldsymbol{y}_{\text{obs}}, \boldsymbol{a}_l, \boldsymbol{a}_r)$ denote the observed data. Then the L measure is modified as

$$\begin{aligned} & L(\boldsymbol{y}_{\text{obs}}) \\ =& E_{\boldsymbol{y}_{\text{cens}}|D}[1\{\boldsymbol{a}_l < \boldsymbol{y}_{\text{cens}} < \boldsymbol{a}_r\} L(\boldsymbol{y})] \\ =& \int \int_{\boldsymbol{a}_l}^{\boldsymbol{a}_r} L(\boldsymbol{y}) f(\boldsymbol{y}_{\text{cens}}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|D) \, d\boldsymbol{y}_{\text{cens}} \, d\boldsymbol{\theta}, \end{aligned}$$

where $1\{\boldsymbol{a}_l < \boldsymbol{y}_{\text{cens}} < \boldsymbol{a}_r\}$ is a generic indicator function taking the value 1 if $\boldsymbol{a}_l < \boldsymbol{y}_{\text{cens}} < \boldsymbol{a}_r$ and 0 otherwise.

• **Choice of ν**

The choice of ν has much potential influence on the properties of the L measure, calibration distribution, and model choice in general.

For the linear model, Ibrahim, Chen, and Sinha (2001) theoretically show that certain values of ν yield highly desirable properties of the L measure and the calibration distribution compared to other values of ν.

Based on their theoretical exploration, $\nu = \frac{1}{2}$ is a desirable and justifiable choice for model selection.

• **Computation**

It can be shown that $L(\boldsymbol{y})$ can be expressed as a posterior expectation, so that

$$L(\boldsymbol{y}) = \sum_{i=1}^{n} \{E_{\boldsymbol{\theta}|D}(E[(z_i)^2|\boldsymbol{\theta}]) - \mu_i^2\} + \nu \sum_{i=1}^{n} (\mu_i - y_i)^2,$$

where $\mu_i = E_{\boldsymbol{\theta}|D}[E(z_i|\boldsymbol{\theta})]$, and the expectation $E_{\boldsymbol{\theta}|D}$ is taken with respect to the posterior distribution $\pi(\boldsymbol{\theta}|D)$.

Suppose that $\{\boldsymbol{\theta}_q,\ q = 1, 2, \ldots, Q\}$ is an MCMC sample from $\pi(\boldsymbol{\theta}|D)$ and $\{\boldsymbol{y}_{\text{cens},q},\ q = 1, 2, \ldots, Q\}$ is an MCMC sample from the truncated posterior predictive distribution

$$1\{\boldsymbol{a}_l < \boldsymbol{y}_{\text{cens}} < \boldsymbol{a}_r\} f(\boldsymbol{y}_{\text{cens}}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|D).$$

Then an Monte Carlo estimate of $L(\boldsymbol{y}_{\text{obs}})$ is given by

$$\begin{aligned}\hat{L}(\boldsymbol{y}_{\text{obs}}) = & \sum_{i=1}^{n} \left\{ \frac{1}{Q} \sum_{q=1}^{Q} \left(E\left[(z_i)^2 | \boldsymbol{\theta}_q \right] \right) - \hat{\mu}_i^2 \right\} \\ & + \nu \left\{ \sum_{\{i:\ y_i \text{ observed}\}} (\hat{\mu}_i - y_i)^2 \right. \\ & \left. + \frac{1}{Q} \sum_{q=1}^{Q} \left[\sum_{\{i:\ y_i \text{ censored}\}} (\hat{\mu}_i - y_{\text{cens},iq})^2 \right] \right\},\end{aligned}$$

where $\hat{\mu}_i = (1/Q) \sum_{q=1}^{Q} E(z_i | \boldsymbol{\theta}_q)$, and $y_{\text{cens},iq}$ is the i^{th} component of $\boldsymbol{y}_{\text{cens},q}$.

In the cases where $E[(z_i)^2|\boldsymbol{\theta}]$ and $E(z_i|\boldsymbol{\theta})$ are not analytically available, we need an MCMC sample $\{(\boldsymbol{z}_q, \boldsymbol{\theta}_q),\ q = 1, 2, \ldots, Q\}$ from the joint distribution $f(\boldsymbol{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|D)$. Then, in $\hat{L}(\boldsymbol{y}_{\text{obs}})$, we replace

$$\frac{1}{Q}\sum_{q=1}^{Q}(E[(z_i)^2|\boldsymbol{\theta}_q]) \ \text{ and } \ \frac{1}{Q}\sum_{q=1}^{Q}E(z_i|\boldsymbol{\theta}_q)$$

by

$$\frac{1}{Q}\sum_{q=1}^{Q}(z_{i,q})^2 \ \text{ and } \ \frac{1}{Q}\sum_{q=1}^{Q}z_{i,q},$$

where $z_{i,q}$ is the i^{th} component of $\boldsymbol{z}_q$.

♠ The Calibration Distribution

• Why Calibration?

Consider, for example, variable subset selection, in which a model with 5 predictors achieves the minimum criterion value, but a model with 3 predictors achieves a slightly larger criterion value. Which model should be chosen? On the basis of the criterion value alone, the model with 5 predictors wins, but it is less parsimonious than the 3 predictor model. This situation arises often in practice, and in these cases, it is desirable to have a calibration of the criterion to formally compare criterion values between the candidate models. Thus, one of the crucial steps in using criterion based methods for model assessment and model choice is to define a calibration for the criterion.

• **Calibration Distribution**

Let c denote the candidate model under consideration, and let t denote the true model. Further, let $L_c(\boldsymbol{y}_{obs})$ denote the L measure for the candidate model c, and let $L_t(\boldsymbol{y}_{obs})$ denote the L measure for the true model t. Now consider the difference in L measures,

$$D(\boldsymbol{y}_{obs}, \nu) \equiv L_c(\boldsymbol{y}_{obs}) - L_t(\boldsymbol{y}_{obs}).$$

To calibrate the L measure, we construct the marginal distribution of $D(\boldsymbol{y}_{\text{obs}}, \nu)$, computed with respect to the prior predictive distribution of $\boldsymbol{y}_{\text{obs}}$ under the true model t, denoted by

$$p_t(\boldsymbol{y}_{\text{obs}}) = \int f_t(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta})\pi_t(\boldsymbol{\theta})\ d\boldsymbol{\theta}.$$

Thus, the calibration distribution is defined as

$$p_{L_c} \equiv p(D(\boldsymbol{y}_{\text{obs}}, \nu)),$$

which is the marginal distribution of $D(\boldsymbol{y}_{\text{obs}}, \nu)$, computed with respect to $p_t(\boldsymbol{y}_{\text{obs}})$.

• **Summary of Calibration Distribution**

Once p_{L_c} is computed, several statistical summaries can be obtained from it to summarize the calibration. These include various HPD intervals and the mean of $D(\boldsymbol{y}^*_{\text{obs}}, \nu)$. The mean of the calibration distribution is denoted by

$$\mu_c(\nu) = E_t(D(\boldsymbol{y}^*_{\text{obs}}, \nu)),$$

where $E_t(\cdot)$ denotes the expectation with respect to the prior predictive distribution of the true model. This summary, $\mu_c(\nu)$, is attractive since it measures, on average, how close the centers are of the candidate and true models. If the candidate model is a good model, then $\mu_c(\nu)$ should be close to 0, whereas if the candidate model is far from the true model, then $\mu_c(\nu)$ should be far from 0. We note that $\mu_c(\nu)$ depends on the candidate model and therefore changes with every c. If $c = t$, then $\mu_c(\nu) = 0$ for all ν.

- **Comments**
  - For p_{L_c} to be well defined, we need a proper prior distribution for $\boldsymbol{\theta}$. This definition of the calibration distribution is appealing since it avoids the potential problem of a double use of the data as discussed by Bayarri and Berger (1999).
  - Since the true model t will not be known in practice, we use the criterion minimizing model $t_{\min}$ to compute

$$\hat{D}(\boldsymbol{y}_{\text{obs}}, \nu) = L_c(\boldsymbol{y}_{\text{obs}}) - L_{t_{\min}}(\boldsymbol{y}_{\text{obs}}),$$

    and

$$\hat{p}_{L_c} = p(\hat{D}(\boldsymbol{y}_{\text{obs}}, \nu)),$$

    where $\hat{p}_{L_c}$ is computed with respect to the prior predictive distribution of the criterion minimizing model.

• **Computation**

Computing the calibration distribution requires the following two steps:

(i) Generate a pseudo-observation $\tilde{\boldsymbol{y}}$ from the prior predictive distribution $f_{t_{\min}}(\boldsymbol{y}|\boldsymbol{\theta})\pi_{t_{\min}}(\boldsymbol{\theta})$; and

(ii) Set $\boldsymbol{y}_{\text{obs}} = \tilde{\boldsymbol{y}}$ and compute MC estimates of $L_c(\boldsymbol{y}_{\text{obs}})$ and $L_{t_{\min}}(\boldsymbol{y}_{\text{obs}})$.

We repeat (i) and (ii) Q times to obtain MCMC samples of $L_c(\boldsymbol{y}_{\text{obs}})$ and $L_{t_{\min}}(\boldsymbol{y}_{\text{obs}})$. Using these MCMC samples, we can compute the entire calibration distribution p_{L_c}, for example, by using the kernel method.

We note that step (ii) may be computationally intensive. However, the entire computational procedure is quite straightforward.

♠ Example: Breast Cancer Data

• The Data

The data given on page 7 are from Finkelstein and Wolf (1985), which consists of a data set of (case-2) interval censored data. In this data set, 46 early breast cancer patients receiving only radiotherapy (covariate value $x = 0$) and 48 patients receiving radio-chemotherapy ($x = 1$) were monitored for cosmetic changes through weekly clinic visits.

Sinha, Chen, and Ghosh (1999) consider a semiparametric Bayesian analysis of these data using three models based on a discretized version of the Cox model (Cox 1972). Specifically, the hazard, $\lambda(y|x)$, is taken to be a piecewise constant function with $\lambda(y|x) = \lambda_j \theta_j^x$ for $y \in I_j$, where $\theta_j = e^{\beta_j}$, $I_j = (a_{j-1}, a_j]$ for $j = 1, 2, \ldots, g$, $0 = a_0 < a_1 < \cdots < a_g = \infty$, and g is the total number of grid intervals.

• **Models**

We consider the following three models:

$\mathcal{M}_1$**:** (i) $\lambda_j \overset{\text{i.i.d.}}{\sim} \mathcal{G}(\eta_j, \gamma_j)$ for $j = 1, \ldots, g$; and

(ii) $\beta \sim N(\beta_0, w_0^2)$.

$\mathcal{M}_2$**:** (i) λ_j's have the same prior as in model $\mathcal{M}_1$; and

(ii) $\beta_{j+1}|\beta_1, \ldots, \beta_{g-1} \sim N(\beta_j, w_j^2)$ for $j = 0, \ldots, g-1$.

$\mathcal{M}_3$**:** (i) $\alpha_{j+1}|\alpha_1, \ldots, \alpha_j \sim N(\alpha_j, v_j^2)$, where $\alpha_j = \ln(\lambda_j)$; for $j = 0, 1, \ldots, g-1$; and

(ii) same as in $\mathcal{M}_2$.

• **Results**

Table 4.4 shows the results using $\nu = \frac{1}{2}$, and reveals that model $\mathcal{M}_1$ is the criterion minimizing model with an L measure value of 80.45. Models $\mathcal{M}_2$ and $\mathcal{M}_3$ have $\mu_c(\frac{1}{2})$ values of $\mu_2(\frac{1}{2}) = 5.36$ and $\mu_3(\frac{1}{2}) = 28.91$, respectively, and therefore, model $\mathcal{M}_2$ is much closer to the criterion minimizing model than model $\mathcal{M}_3$.

TABLE 4.4. L Measure and Calibration Summaries for Breast Cancer Data.

Model	L Measure	$\mu_c(\frac{1}{2})$	95% HPD
1*	80.45	–	–
2	87.24	5.36	(4.24, 6.34)
3	113.54	28.91	(27.23, 30.23)

* Criterion minimizing model.

This is also clearly displayed in Figure 4.1, which gives the calibration distributions for models $\mathcal{M}_2$ and $\mathcal{M}_3$. We see from Figure 4.1 that there is a wide separation between p_{L_2} and p_{L_3}, and p_{L_2} has smaller dispersion than p_{L_3}. The HPD intervals for models $\mathcal{M}_2$ and $\mathcal{M}_3$ do not contain 0. We conclude here that both models $\mathcal{M}_2$ and $\mathcal{M}_3$ are sufficiently different from one another as well as being sufficiently different from the criterion minimizing model.

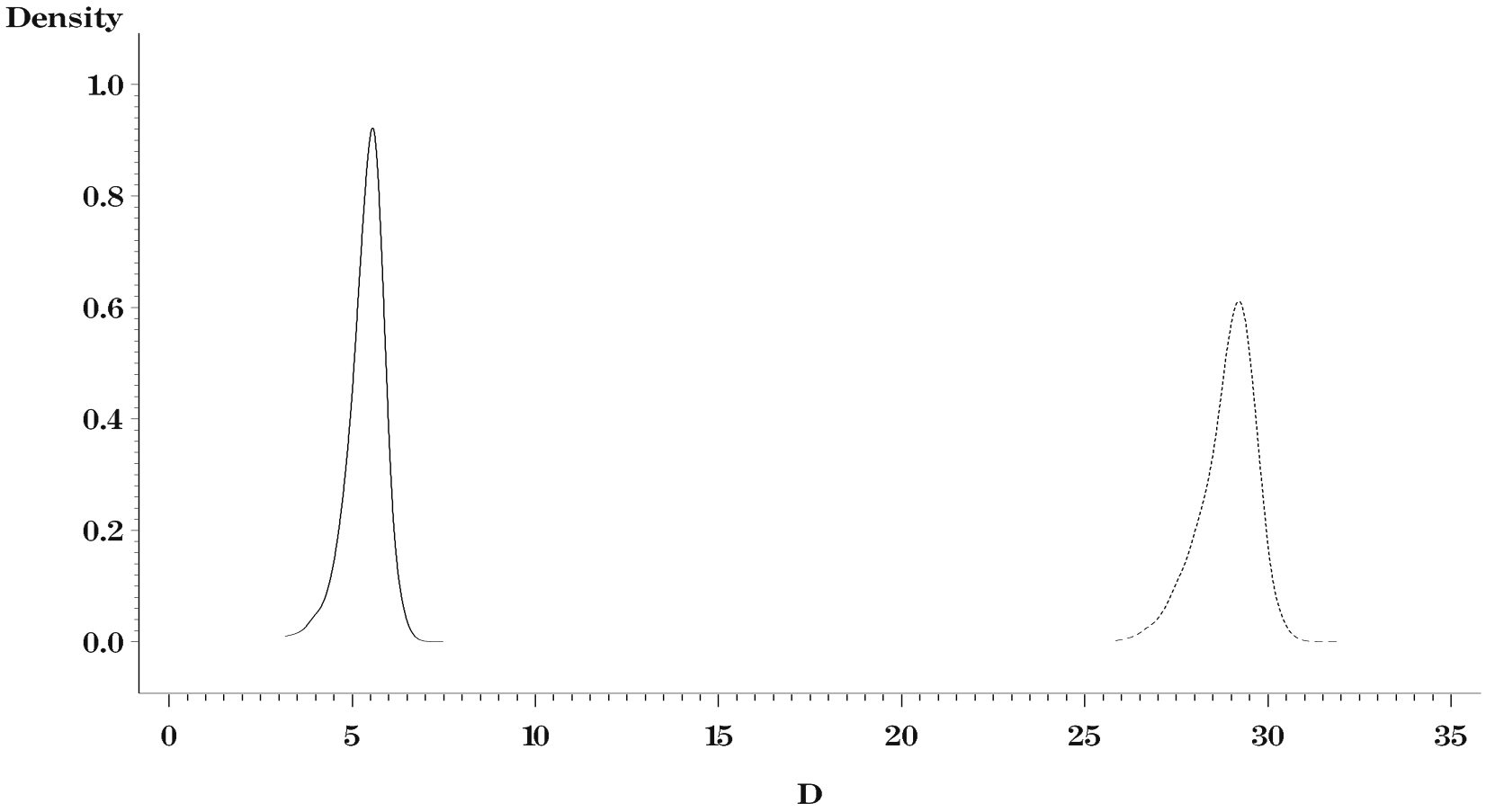

FIGURE 4.1. Calibration distributions for breast cancer data; solid curve: model $\mathcal{M}_2$, and dashed curve: model $\mathcal{M}_3$.

4.3 Conditional Predictive Ordinate

♠ CPO Statistic

The Conditional Predictive Ordinate (CPO) statistic is a very useful model assessment tool which has been widely used in the statistical literature under various contexts. For a detailed discussion of the CPO statistic and its applications to model assessment, see Geisser (1993), Gelfand, Dey, and Chang (1992), Dey, Chen, and Chang (1997), and Sinha and Dey (1997). For the i^{th} observation, the CPO statistic is defined as

$$\text{CPO}_i = f(y_i|D^{(-i)}) = \int f(y_i|\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{x}_i)\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D^{(-i)})\ d\boldsymbol{\beta}\ d\boldsymbol{\lambda},$$

where y_i denotes the response variable and $\boldsymbol{x}_i$ is the vector of covariates for case i, $D^{(-i)}$ denotes the data with the i^{th} case deleted, and $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D^{(-i)})$ is the posterior density of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ based on the data $D^{(-i)}$.

CPO_i is the marginal posterior predictive density of y_i given $D^{(-i)}$, and can be interpreted as the height of this marginal density at y_i. Thus, large values of CPO_i imply a better fit of the model.

♠ CPO Plot

For comparing two competing models, we examine the CPO_i's under both models. The observation with a larger CPO value under one model will support that model over the other. Therefore, a plot of CPO_i's under both models against observation number should reveal that the better model has the majority of its CPO_i's above those of the poorer fitting model. In comparing several competing models, the CPO_i values under all models can be plotted against the observation number in a single graph.

♠ Pseudomarginal likelihood (LPML)

An alternative to CPO plots is the summary statistic called the logarithm of the Pseudomarginal likelihood (LPML) defined as

$$\text{LPML} = \sum_{i=1}^{n} \log(\text{CPO}_i).$$

To compare LPML's from two different studies for a given model, we propose to use a modification of LPML, which is the average LPML, given by

$$\text{ALPML} = \frac{\text{LPML}}{n},$$

where n is the sample size. The statistic ALPML can be interpreted as the relative pseudomarginal likelihood.

♠ Advantages of the CPO Approch

- Compared to Bayes factor approach, CPO or LPML is always well defined as long the posterior predictive density is proper. Thus, LPML is well defined under improper priors, and in addition, it is very computationally stable. However, the Bayes factor is not well defined with improper priors, and is generally quite sensitive to vague proper priors.
- Compared to L measure, The L measure is a Bayesian criterion requiring finite second moments of the sampling distribution of y_i, whereas the CPO or LPML statistic does not require existence of any moments. Since the cure rate models have improper survival functions, no moments of the sampling distribution exist, and therefore the L measure is not well defined for these models. Thus, the LPML statistic is well motivated.

♠ Example: Melanoma data

• Data

We use the E1684 (historical) and E1690 (current) melanoma datasets to illustrate the CPO statistic.

• Models

We consider the piecewise exponential (PE) model and the semiparametric cure rate (SPCR) model. For the piecewise exponential model, we consider a fully parametric analysis (i.e., $J = 1$) and a semiparametric analysis using $J = 5$. For the semiparametric cure rate model, we use $J = 5$. We note that $J = 1$ corresponds to a fully parametric cure rate model.

• Incorporation of Historical Data

We consider several choices of a_0, including $a_0 = 0$ and $a_0 = 1$ with probability 1, $E(a_0|D) = 0.05$, $E(a_0|D) = 0.30$, and $E(a_0|D) = 0.60$.

- **Results**

TABLE 4.5. CPO Statistics for E1684 and E1690.

Study	Model	ALPML
E1684	PE ($J = 5$)	-1.3775
E1690	PE ($J = 5$)	-1.2232
E1684	SPCR ($J = 1$)	-1.3407
E1690	SPCR ($J = 1$)	-1.2172
E1684	SPCR ($J = 5$)	-1.3439
E1690	SPCR ($J = 5$)	-1.2184

Table 4.5 shows results of ALPML for the E1684 and E1690 studies separately, based on $a_0 = 0$ with probability 1. We see from Table 4.5 that the results for PE and SPCR are quite similar, yielding similar ALPML statistics. In addition, the PE model with $J = 5$ gives comparable results to the cure rate models. However, the exponential model (i.e., the PE model with $J = 1$) yields a smaller CPO statistic relative to the other models, indicating a poorer fit. These results suggest that the SPCR models appear to provide a more adequate fit to the E1690 data compared to the exponential model and are comparable to, but slightly better than, the PE model with $J = 5$.

TABLE 4.6. LPML Statistics for PE and SPCR Models.

Model	$E(a_0\|D)$	$J=1$	$J=5$	$J=10$
PE	0	−575.60	−522.30	−523.62
	0.05	−575.45	−522.05	−523.20
	0.20	−575.23	−521.67	−522.39
	0.30	−575.13	−521.59	−522.12
	0.60	−574.95	−521.61	−522.02
	1	−574.64	−522.24	−522.71
SPCR	0	−519.75	−520.24	−524.42
	0.05	−519.61	−519.89	−523.82
	0.20	−519.39	−519.43	−522.83
	0.30	−519.34	−519.31	−522.53
	0.60	−519.40	−519.67	−522.56
	1	−519.67	−520.16	−522.97

Table 4.6 is quite informative.

- First, we see that $J = 5$ is better than $J = 1$ or $J = 10$. However, for the SPCR model, $J = 1$ and $J = 5$ are fairly close.
- Second, for both $J = 1$ or $J = 5$, the cure rate model yields a better fit than the PE model.
- Third, the incorporation of the E1684 data into the analysis improves the model fit.
- Fourth, for all the cases, LPML is a concave function of $E(a_0|D)$ (see Figure 4.2). This is an interesting feature in LPML in that it demonstrates that there is an "optimal" weight for the historical data with respect to the statistic LPML, and thus this property is potentially very useful in selecting a model.

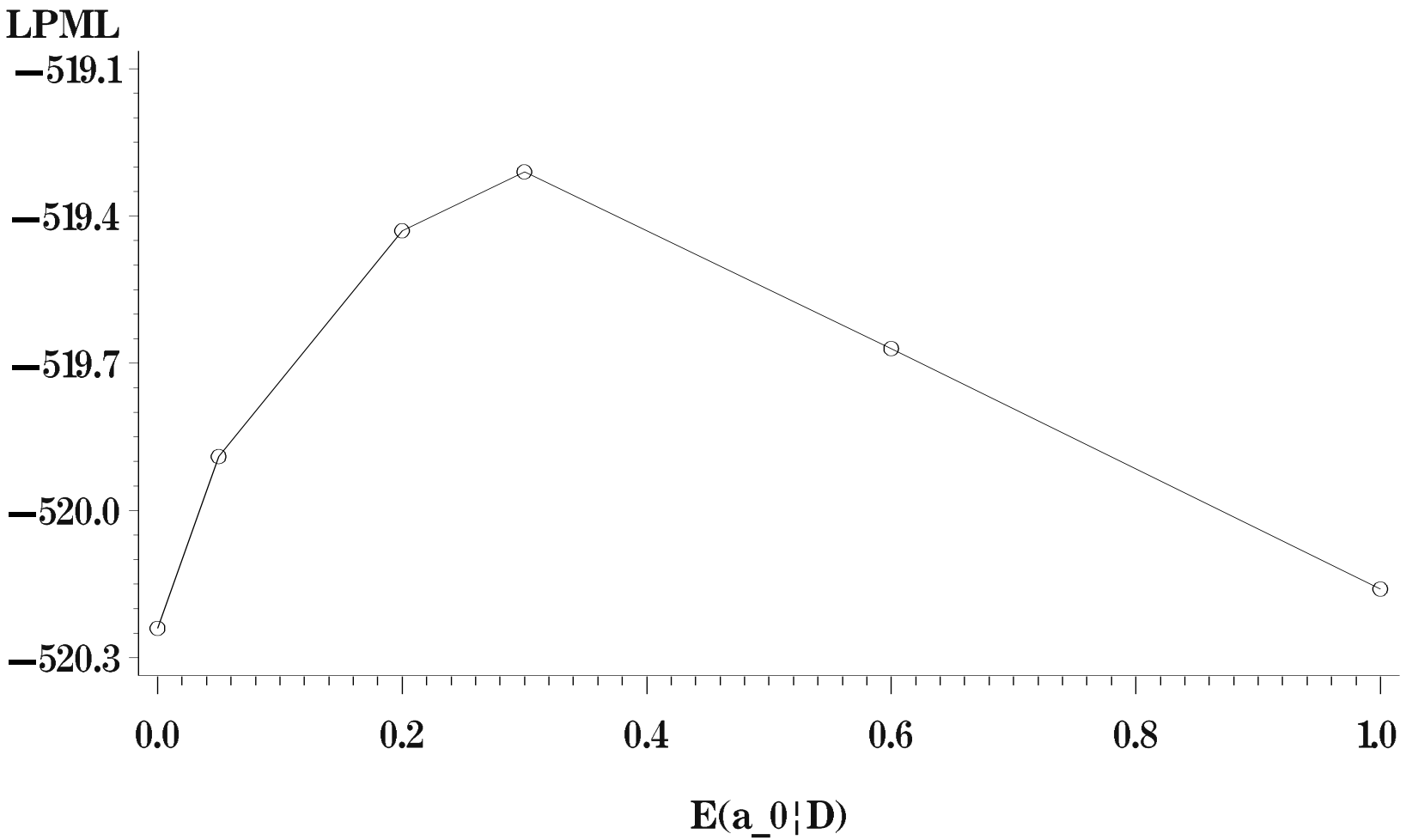

FIGURE 4.2. Plot of LPML's for SPCR with $J = 5$.

♠ **Example: Breast cancer data**

• **Computational Detail for Interval Censored Data**

For interval censored data, the CPO statistic for the i^{th} observation is defined as

$$\text{CPO}_i = P(Y_i \in (a_{l_i}, a_{r_i}]|x_i, D^{(-i)}),$$

where $D^{(-i)}$ denotes the interval censored data with the i^{th} patient removed. CPO_i is the posterior predictive probability of the observed data for the i^{th} patient given the modified data $D^{(-i)}$. Let $\boldsymbol{\theta}$ denote the vector of model parameters. CPO_i can be computed as

$$\text{CPO}_i = \left(\text{E} \left[\frac{1}{P(Y_i \in (a_{l_i}, a_{r_i}]|\boldsymbol{\theta}, x_i)} \right] \right)^{-1},$$

where the expectation is taken with respect to the joint posterior $\pi(\boldsymbol{\theta}|D)$.

Note that

$$P(Y_i \in (a_{l_i}, a_{r_i}]|\boldsymbol{\theta}, z_i) = \exp\left\{-\sum_{k=1}^{l_i} \lambda_k \theta_k^{x_i} \tilde{\Delta}_k\right\} - \exp\left\{-\sum_{k=1}^{r_i} \lambda_k \theta_k^{x_i} \tilde{\Delta}_k\right\},$$

where $\tilde{\Delta}_k = a_k - a_{k-1}$ and $\theta_k = \exp(\beta_k)$. Thus, a Monte Carlo estimate of CPO_i is given by

$$\begin{aligned}\widehat{\text{CPO}}_i =& \text{E}\left[\frac{1}{P(Y_i \in (a_{l_i}, a_{r_i}]|\boldsymbol{\theta}, x_i)}\right] \\ =& \frac{1}{L}\sum_{l=1}^{L}\left[\exp\left\{-\sum_{k=1}^{l_i} \lambda_{kl} \theta_{kl}^{x_i} \tilde{\Delta}_k\right\} - \exp\left\{-\sum_{k=1}^{r_i} \lambda_{kl} \theta_{kl}^{x_i} \tilde{\Delta}_k\right\}\right]^{-1},\end{aligned}$$

where $\{\theta_{kl},\ l = 1, 2, \cdots, L\}$ (L is large) is an MCMC sample from the posterior distribution $\pi(\boldsymbol{\theta}|D)$.

• **Results**

The values of the LPML's are -157.61 and -188.33 for $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively, and the CPO_i's are displayed in Figure 4.3. Based on the LPML statistics, it is clear that $\mathcal{M}_1$ is more preferable than $\mathcal{M}_2$. The plots of the pairwise log CPO ratios are consistent with the single summary measure LPML. In Figure 4.3, 84% of the log CPO ratios for $\mathcal{M}_1$ versus $\mathcal{M}_2$ are positive. Therefore, the data support $\mathcal{M}_1$ instead of $\mathcal{M}_2$, which is also consistent with the L measure criterion.

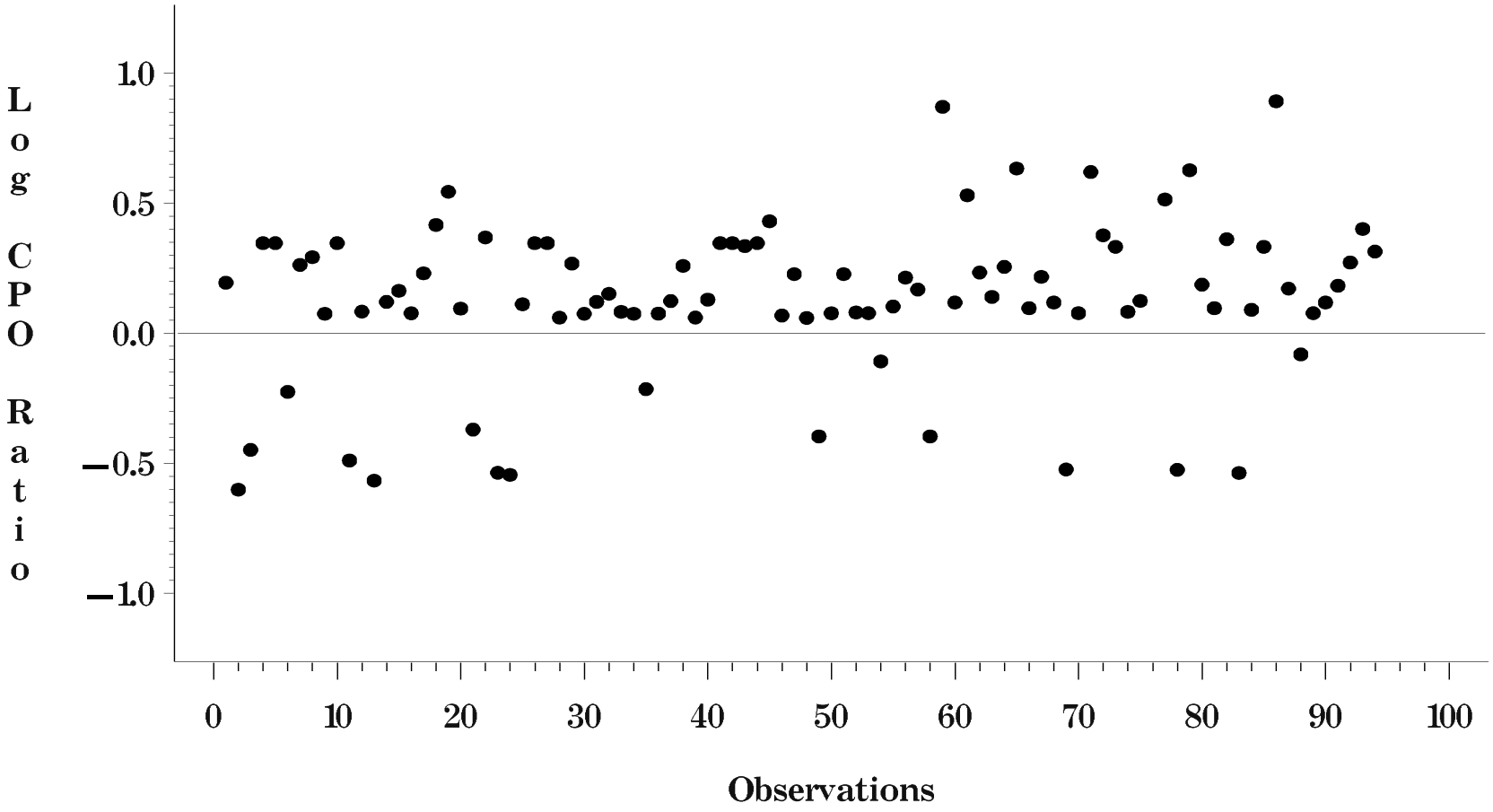

FIGURE 4.3. Plot of log CPO ratios for $\mathcal{M}_1$ versus $\mathcal{M}_2$ for breast cancer data.

4.4 The Other Methods

The other methods for model comparison and model assessment include *Bayesian Model Averaging* and *Bayesian Information Criterion.* Bayesian Model Averaging (BMA) is one of the popular approaches to model selection. In this approach, one bases inference on an average of all possible models in the model space $\mathcal{M}$, instead of a single "best" model. Suppose $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_\mathcal{K}\}$, and let Δ denote the quantity of interest such as a future observation, a set of regression coefficients, or the utility of a course of action. Then, the posterior distribution of Δ is given by

$$\pi(\Delta|D) = \sum_{k=1}^{\mathcal{K}} \pi(\Delta|D, \mathcal{M}_k) p(\mathcal{M}_k|D),$$

where D denotes the data, $\pi(\Delta|D, \mathcal{M}_k)$ is the posterior distribution of Δ under model $\mathcal{M}_k$, and $p(\mathcal{M}_k|D)$ is the posterior model probability.

The Bayesian Information Criterion (BIC) is defined as

$$\text{BIC} = -2\{\ell_k(\hat{\boldsymbol{\theta}}_k) - \ell_0(\hat{\boldsymbol{\theta}}_0)\} + (p_k - p_0)\log(n),$$

where $\ell_k(\hat{\boldsymbol{\theta}}_k)$ and $\ell_0(\hat{\boldsymbol{\theta}}_0)$ are the log maximized likelihoods under $\mathcal{M}_k$ and a reference model $\mathcal{M}_0$, whose parameter has dimension p_0, where n is the sample size.

Due to the time constraint, the details for these two methods will not be discussed further in this session. We refer the interesting audiences to the following papers for more discussions: Madigan and Raftery (1994), Madigan and York (1995), Raftery (1996), Volinsky, Madigan, Raftery, and Kronmal (1997), and Volinsky and Raftery (2000).

4.5 Bayesian Model Diagnostics

♠ Bayesian Latent Residuals

• The Frailty Model

Under the proportional hazards frailty model,

$$h(y|w_i, x_{ij}) = h_0(y) w_i \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}),$$

let $H_0(t)$ be the cumulative hazard function. Suppose we divide the time axis into J prespecified intervals $I_k = (s_{k-1}, s_k]$ for $k = 1, 2, \ldots, J$, and assume the baseline hazard to be constant within these intervals. Then it follows that for $t \in I_k = (s_{k-1}, s_k]$, $k = 1, 2, \ldots, J$, $H_0(t) = \sum_{l=1}^{k} \Delta_l \lambda_l$, where $\Delta_l = s_l - s_{l-1}$.

• **Latent Residuals**

Define

$$u_{ij} = u(y_{ij}|w_i, \boldsymbol{x}_{ij}) = H_0(y_{ij})\theta_{ij}w_i,$$

where $\theta_{ij} = \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta})$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m_i$. Recall $h_0(t) = \lambda_k$ if $t \in I_k = (s_{k-1}, s_k]$. Then the survival function is given by

$$\begin{aligned} & S_{ij}(t|w_i, \boldsymbol{x}_{ij}) \\ &= \exp\left\{-\theta_{ij}w_i \int_0^t h_0(u)\ du\right\} \\ &= \exp\{-\theta_{ij}w_i H_0(t)\}. \end{aligned}$$

It follows that the probability density function of the survival time T_{ij} is given as

$$f_{ij}(t|w_i, \boldsymbol{x}_{ij}) = h_0(t)\theta_{ij}w_i \exp\{-\theta_{ij}w_i H_0(t)\}.$$

Thus,

$$\begin{aligned} & P(u_{ij} \le u) \\ =& P(\theta_{ij} w_i H_0(t) \le u) = P(T_{ij} \le H_0^{-1}(u/(\theta_{ij} w_i))) \\ =& 1 - \exp\{-\theta_{ij} w_i (u/(\theta_{ij} w_i))\} = 1 - \exp(-u), \end{aligned}$$

which implies that given w_i, u_{ij} has a standard exponential distribution. Further, if

$$v_{ij}(t) = \exp(-u_{ij}(y_{ij})),$$

then v_{ij} given w_i has a standard uniform distribution. Also, conditional on w_i, the v_{ij}'s are independent.Then, the v_{ij}'s can be treated as standardized residuals. We call the v_{ij}'s Bayesian latent residuals since they are functions of the unobserved frailty random variable w_i.

• **Residual Plots**

Using the v_{ij}'s, Aslanidou, Dey, and Sinha (1998) propose two diagnostic plots, using the output from the MCMC samples. First, if the model is correct, then the v_{ij}'s have a uniform distribution. Thus, a box-plot of the Monte Carlo estimates of the v_{ij}'s can be used to check model adequacy for a given dataset. Alternatively, a Q-Q plot of v_{ij} versus a standard uniform distribution produces similar features. The estimates of the v_{ij}'s are obtained as

$$\hat{v}_{ij} = L^{-1} \sum_{l=1}^{L} v_{ij}^{l},$$

where

$$v_{ij}^{l} = \exp(-H_0(y_{ij})^{l} \theta_{ij}^{l} w_i^{l}),$$

$H_0(y_{ij})^{l}$, θ_{ij}^{l}, and w_i^{l} are the values of $H_0(y_{ij})$, θ_{ij}, and w_i computed at the l^{th} MCMC iteration for $l = 1, 2, \ldots, L$, and L is the total number of MCMC iterations.

• **Example: Kidney Infection Data**

Aslanidou, Dey, and Sinha (1998) reanalyze the kidney infection data given in Example 1.4 of the textbook. They assume that $w_i \sim \mathcal{G}(\kappa^{-1}, \kappa^{-1})$, so that $E(w_i) = 1$ and $\text{Var}(w_i) = \kappa$. The prior for the hyperparameter κ is $\mathcal{G}(6, 1)$ with $E(\kappa) = \text{Var}(\kappa) = 6$ to assure enough heterogeneity among the patients. Since $\eta = \kappa^{-1}$, the prior for η is taken to be an inverse gamma, $\mathcal{IG}(6, 1)$. For the λ_k's, $k = 1, 2, \ldots, J$, they consider the priors $\lambda_1 \sim \mathcal{G}(0.7, 0.7)$ and $\lambda_k|\lambda_{k-1} \sim \mathcal{G}(0.7, 0.7/\lambda_{k-1})$ for $k = 2, 3, \ldots, J$. They consider only one covariate, i.e., sex, in their analysis. For β, the regression coefficient corresponding to sex, they take a $N(-1.2, 100)$ prior. The prior mean is chosen near the estimate found by other analyses of this dataset. The variance of the prior is taken large enough to incorporate sufficient diffuseness. Finally, they divide the survival times of the patients into $J = 20$ equal intervals.

For the model checking analysis, they obtain the following results. The first infection of the 19th, 35th, and 36th patients did not have a nice fit in the model like the rest of the subjects for the first infection. This can be concluded from the box-plots of the quantities $v_{19,1}$, $v_{35,1}$, and $v_{36,1}$, which did not seem to follow a uniform distribution. A similar phenomenon is observed for the second infections of the 14th, 15th, and 22nd patients. For illustration, Figure 4.4 shows the box-plots of the first infection for the 16th patient, who fit nicely to the model, and the first infection of the 36th patient who did not have a nice fit.

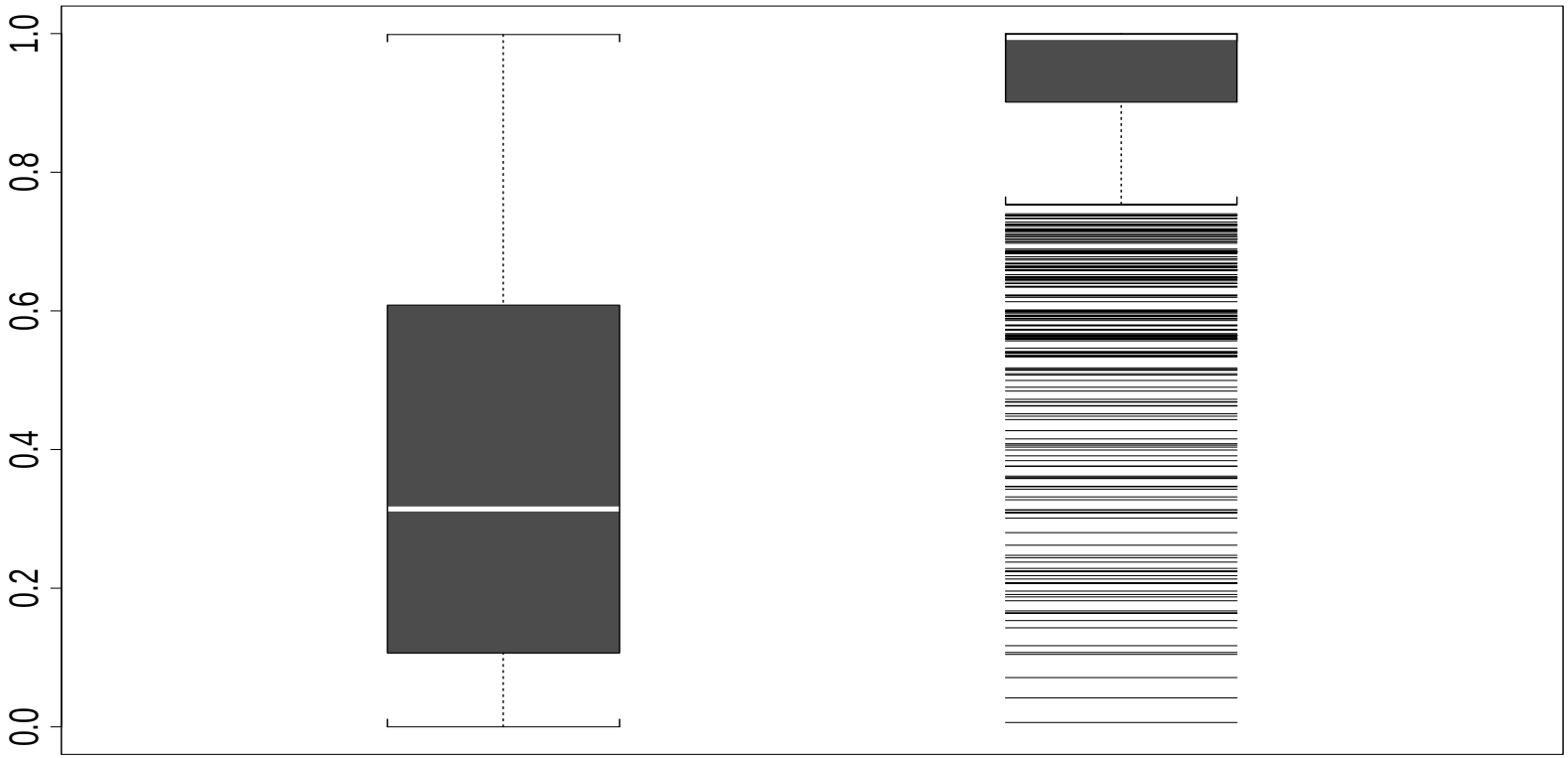

FIGURE 4.4. Box-plots of $v_{16,1}$ and $v_{36,1}$.

♠ Prequential Methods

• The Contributors of the Method

The prequential method discussed here is proposed by Arjas and Gasbarra (1997, Biometrika), who extend the prequential approach introduced by Dawid (1992) to continuous time marked point processes.

• Notation

– **Data**:

Suppose that the data are of the form $\{(Y_n, X_n)\}$, where $0 \equiv Y_0 < Y_1 < Y_2 < \ldots < Y_N$ are "the observed times of occurrence" and X_n is a description of the event which occurred at Y_n. For simplicity, the marks X_n are assumed to take values in a countable set E.

– **The Pre-t History**

$$\mathcal{H}_t = \{(Y_n, X_n) : \ Y_n \leq t\},$$

consisting of the events in the data which occurred before (and including) time t, and by $\mathcal{H}_{t-}$ the corresponding history when the inequality in the definition of $\mathcal{H}_t$ is strict.

– **The Internal Filtration of Process**

$$\mathcal{A}_t = \sigma\{\mathcal{H}_t\}$$

is the σ-field of $\mathcal{H}_t$.

– **The Canonical Path Space Ω**

It consists of the sequences $\omega = \{(y_n, x_n) : n \geq 0\}$ such that $x_n \in E$, where E is some measurable space, $0 \equiv y_0 \leq y_1 \leq y_2 \leq \ldots$, and $y_n < \infty$ implies $y_n < y_{n+1}$, on the one hand, and on the distribution of the "initial mark" X_0 and the $(\mathcal{A}_t)$-compensators of the counting processes

$$N_t(x) = \sum_{Y_n \leq t} 1_{\{Y_n \leq t, X_n = x\}}(t), \quad t \geq 0, x \in E,$$

on the other.

• Statistical Model of the Marked Point Process

Let θ denote an unknown parameter θ, which can be viewed as an initial unobserved mark X_0 of the marked point process at time $Y_0 = 0$. The parameter space Θ may be finite-dimensional real or abstract, but its role will always be the same: for any given value $\theta \in \Theta$, we assume that there is a corresponding probability P^θ defined on the path space

$$\Omega = \{\omega = (X_n)_{n=0}^{\infty} : \ X_i \in E\}.$$

In practice, as noted above, P^θ is specified most conveniently in terms of an initial distribution for X_0 and the corresponding conditional intensities, $h_t^\theta(x)$, say, $x \in E, t \geq 0$. In classical statistical inference, the family

$$\mathcal{M} = \{P^\theta : \ \theta \in \Theta\}$$

is called a statistical model of the marked point process.

• **Bayesian Inference**

In Bayesian inference, both the parameter θ and the process sample path $\mathcal{H}_\infty$ are viewed as random elements. The joint distribution of $(\theta, \mathcal{H}_\infty)$ is then determined by probabilities of the form

$$P(\theta \in A, \mathcal{H}_\infty \in B) = P_{(\theta, \mathcal{H}_\infty)}(A \times B) = \int_A P^\theta(B)\pi(d\theta),$$

where $\pi(\theta)$ is the prior distribution for θ.

• **Dynamical Prediction Problem**

The prediction at time t concerns the unobserved future sample path, say, $\mathcal{H}_{(t,\infty)}$, and the predictions are updated continuously on the basis of the observed $\mathcal{H}_t$. At time $t = 0$, the predictive distribution is simply the marginal of H_∞, obtained from $P(\theta \in A, \mathcal{H}_\infty \in B)$ by letting $A = \Theta$. Let P denote this probability and also let E denote the corresponding expectation. Then, the updating of P corresponds to viewing $\mathcal{H}_\infty$ as a pair $(\mathcal{H}_t, \mathcal{H}_{(t,\infty)})$ and conditioning the joint distribution of θ and $\mathcal{H}_\infty$ on $\mathcal{H}_t$, again integrating out the parameter θ. This continuous updating of predictions, which in practice is done by applying Bayes' formula on the corresponding posterior distributions for θ, is at the heart of the method proposed by Arjas and Gasbarra (1997) for model assessment.

• **A Useful Theoretical Result**

Let $\pi(d\theta|\mathcal{H}_{t-})$ denote the posterior distribution given the pre-t history $\mathcal{H}_{t-}$. It follows that the *predictive* $(P, \mathcal{A}_t)$-intensities $\hat{h}_t(x)$ is given by

$$\hat{h}_t(x) = E(h_t^\theta(x)|\mathcal{H}_{t-}).$$

Notice that $\hat{h}_t(x)$ needs to be evaluated sequentially, which could be difficult to compute, unless the analytic form for this posterior expectation is available.

For a particular mark $x \in E$, the corresponding sequence of precise times at which x occurs, say

$$\tau_0^x \equiv 0, \;\; \tau_{k+1}^x = \inf\{Y_n > \tau_k^x; X_n = x\}.$$

We are led to the following result, given in Arjas and Gasbarra (1997).

Result 4.1: *The spacings* $\widehat{H}_{\tau_{k+1}^x}(x) - \widehat{H}_{\tau_k^x}(x)$, $k = 0, 1, 2, ...,$ *of the* $(P, \mathcal{A}_t)$*-compensator* $\widehat{H}_t(x) = \int_0^t \hat{h}_s(x)\, ds$ *form a sequence of independent* $\mathcal{E}(1)$ *random variables.*

• **Total Time on Test Plot**

The plot process,

$$n \to S_n = \sum_{i \le n} \widehat{H} {\tau_i}^i,$$

under the Bayesian forecasting system, has independent $\mathcal{E}(1)$ increments. Asymptotically, the Kolmogorov law of the iterated logarithm gives a sharp result on the behavior of the random walk process $\{S_n - n\}$: with probability 1,

$$\limsup \frac{S_n - n}{\sqrt{2n \log\log n}} = +1$$

and

$$\liminf \frac{S_n - n}{\sqrt{2n \log\log n}} = -1.$$

Thus,

$$n \pm \sqrt{2n \log\log n}$$

constitutes two bounds for S_n. If $\{S_n\}$ infringes the boundaries and does not return, this can be used as evidence against the model.

• **Prequential P-value**

To quantify the evidence for or against the model, Arjas and Gasbarra (1997) consider some functionals of the whole sample path $\{S_n\}$. Since $S_n \sim \mathcal{G}(n,1)$, then

$$G_n = F_\gamma(S_n; n, 1) \sim \mathcal{U}(0,1),$$

where $F_\gamma(\cdot\,; n, 1)$ denotes the cdf of $\mathcal{G}(n,1)$. The test statistics are

$$G_{1n} = \max_{k \le n} G_k \;\text{ and }\; G_{2n} = \min_{k \le n} G_k.$$

By computing the reference distributions of G_{1N} and G_{2N} we can assign p-values to the whole sample, i.e.,

$$P(G_{1N} \geq g_{1N}) \text{ and } P(G_{2N} \leq g_{2N}),$$

where $g_{1N} = \max_{k \leq N} g_k$, $g_{2N} = \min_{k \leq N} g_k$ are the observed values of the statistics. Arjas and Gasbarra (1997) call these statistics *prequential P-values*. As usual, P-values close to 0 would be used as evidence against the model P. In principle, the distribution of G_{1n} could be computed recursively. Here approximate prequential P-values were determined by a simple Monte Carlo method, by generating independent identically distributed samples of the process $\{G_n\}$.

• Example: Point Process with Serially Correlated Spacings

Models

Let $0 < Y_1 < Y_2 < \ldots$ denote a simple point process and also let $\eta_n = Y_n - Y_{n-1}$ denote the spacings, with $Y_0 \equiv 0$. Arjas and Gasbarra (1997) consider the following three competing Bayesian models:

Model M_0. Suppose that (i) the model parameter θ is a real-valued random variable with prior distribution $F_\gamma(\cdot; \alpha, \beta)$ (α = shape parameter and β = scale parameter), and that (ii) conditionally on θ, the spacings $\{\eta_n; n \geq 1\}$ are independent and distributed as $\mathcal{E}(\theta)$. In other words, $\{Y_n\}$ is a doubly stochastic Poisson process, or Cox process, with conditional intensity given by $h_t^\theta \equiv \theta$.

Model M_1. Suppose that (i) θ is as in model M_0 above, but that (ii), conditionally on $(\theta, \eta_1, \eta_2, ..., \eta_{n-1})$, the spacings η_n are distributed according to the exponential distribution with parameter θ/η_{n-1}. The conditional intensity is now given by $h_t^\theta = \theta/\eta_{N_t-}$. According to model M_1, long (short) spacings are typically followed by long (short) spacings, and therefore the points Y_n tend to be clustered.

Model M_{01}. This model includes both M_0 and M_1 as special cases and it in principle permits consideration of model selection probabilities adapting to the data, would have the following mixture form. Let $q \in [0,1]$ be given and let ξ be a $\{0,1\}$-valued random variable with $P_{0,1}(\xi = 0) = 1 - q$, $P_{0,1}(\xi = 1) = q$. Then define the probability $P_{0,1}$ on $\Theta \times \Omega \times \{0,1\}$ by specifying the conditional probability $P_{0,1}(\cdot|\xi)$ through $P_{0,1}(\cdot|\xi) = \xi P_1(\cdot) + (1-\xi)P_0(\cdot)$.

True Model: M_1

A sample path segment of the process $\{Y_n\}$ consisting of 500 points was generated by a computer from model M_1, with $\theta = 0.6$. The hyperparameters were given the values $\alpha = 0.1$ and $\beta = 0.001$. This prior has a very large mean $\alpha/\beta = 100$, compared to the true value of 0.6, but it is also very flat, having variance $\alpha/\beta^2 = 10^5$.

Results:

The prequential P-values of the generated data under these three different models are shown in Table 4.7. Clearly, the values of these statistics provide the strong evidence that the models M_1 and M_{01} fit the data while the model M_0 does not.

TABLE 4.7. Prequential P-values

	M_1	M_0	$M_{0,1}$
$P(G_{1,500} \geq g_{1,500})$	0.594	$\simeq 0$	0.629
$P(G_{2,500} \leq g_{2,500})$	0.337	$\simeq 0$	0.253

The total-time-on-test plots displayed in Figure 4.5, where the S_n are plotted against n for $n = 1, 2, \ldots, 500$ along with the bounds $n \pm \sqrt{2n \log\log n}$ arising from the law of the iterated logarithm. Again one can see that, while the process $\{S_n\}$ has a nice random walk behavior under M_1 and $M_{0,1}$ (graphically coinciding in the figure), staying mostly in the region prescribed by the law of the iterated logarithm, under model M_0 it leaves this region abruptly.

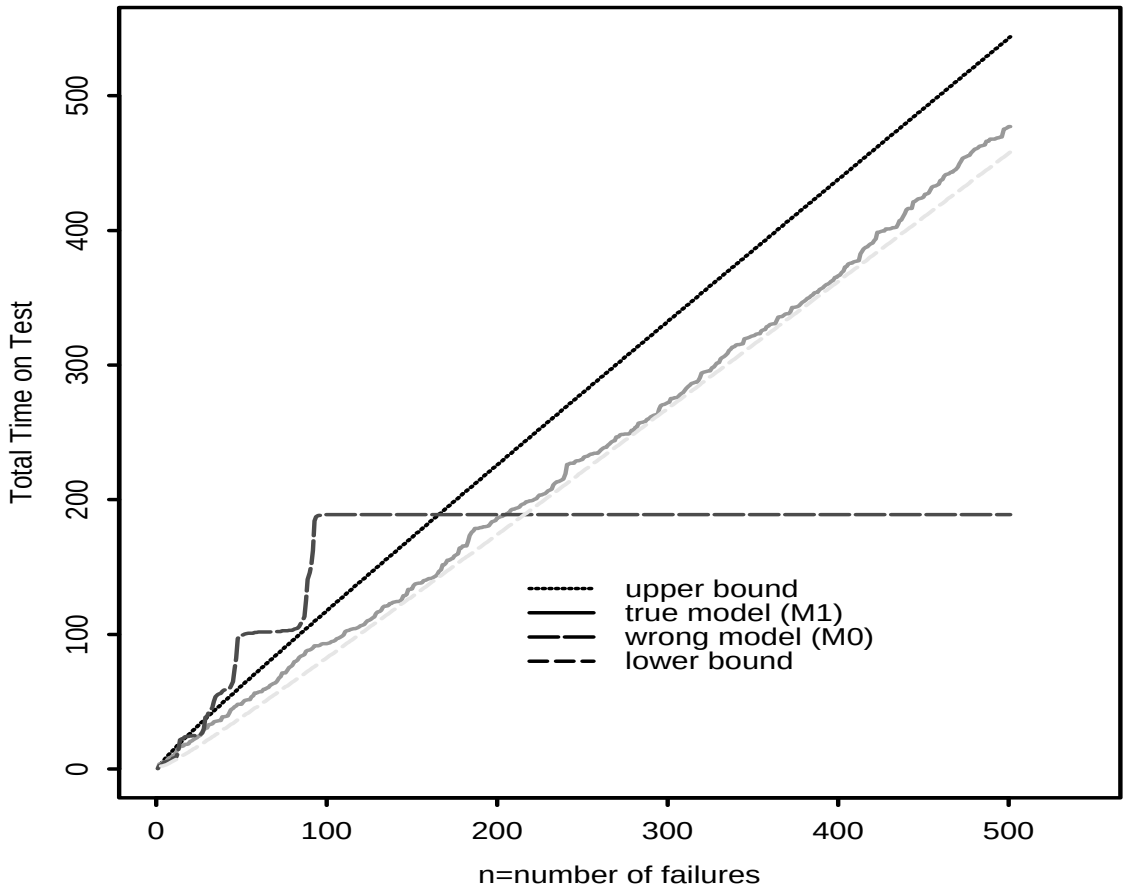

FIGURE 4.5. Total-time-on-test plot for S_n versus n, the number of failures.