12. Other topics

Objective

Introduce the basic ideas of robust Bayesian analysis and nonparametric Bayesian methods.

Recommended reading

- Berger, J. (1994) An overview of robust Bayesian analysis (with discussion). *Test*, **3**, 5–124.

- Ríos Insua, D. and Ruggeri, F. (eds.) (2000). *Robust Bayesian Analysis*. Springer Verlag.

Robustness

In any Bayesian analysis, especially when expert priors have been solicited, it is important to assess the sensitivity of the results to the election of the prior distribution $p(\boldsymbol{\theta})$.

Sensitivity to the loss function and likelihood function are also considered in Dey and Micheas (2000), Kadane et al (2000) and Shyamalkumar (2000).

An informal sensitivity analysis considers robustness to the use of various different prior distributions.

Example

Example 85

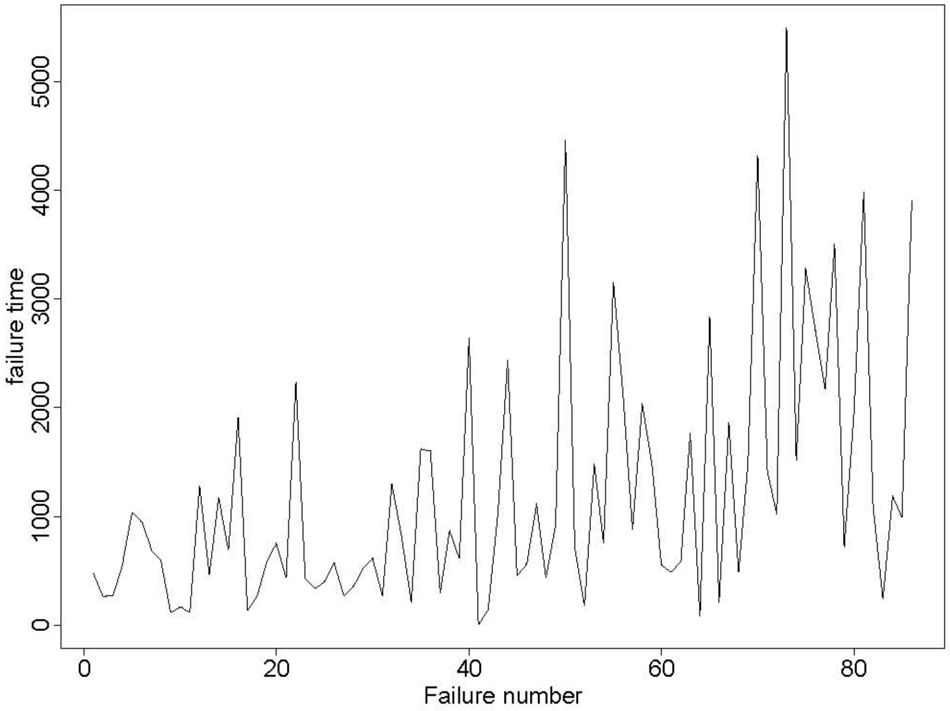
Wilson and Wiper (2000) analyzed the Jelinski Moranda (1972) model for software reliability.

Let $T_1, T_2, \ldots$ be the times between successive software failures. Then the Jelinski Moranda model assumes that

$$T_i | N, \phi \sim \mathcal{E}\left((N - i + 1)\phi\right)$$

so that initially, the program contains N faults, each of which has the same size or importance, ϕ, and then after each failure, the fault causing the failure is identified and removed from the program.

The first $m = 86$ inter failure times for a program were observed as in the following diagram.



The objective is to predict the next failure time and the number of bugs left in the program.

The likelihood function is

$$l(N,\phi|\mathbf{t}) \propto \frac{N!}{(N-m)!}\phi^m \exp\left(-\left[(N+1)m\bar{t}-\sum_{i=1}^{m} it_i\right]\phi\right)$$

and semi-conjugate priors are

$$\begin{array}{rcl} N & \sim & \mathcal{P}(\lambda) \quad \text{where we assume that } \lambda = 100 \\ \phi & \sim & \mathcal{G}(\alpha,\beta) \quad \text{where } \alpha = 1 \text{ and } \beta = .0001 \end{array}$$

Given these priors, it can be shown that $E[N|\text{data}] \approx 104$ (MLE $= 106$) and the posterior median of the distribution of the time to next failure is 2440×10^{-2} seconds (MLE $= 2177$).

Assume that we contaminate the prior distribution with a long tailed distribution. We shall consider the class of prior distributions

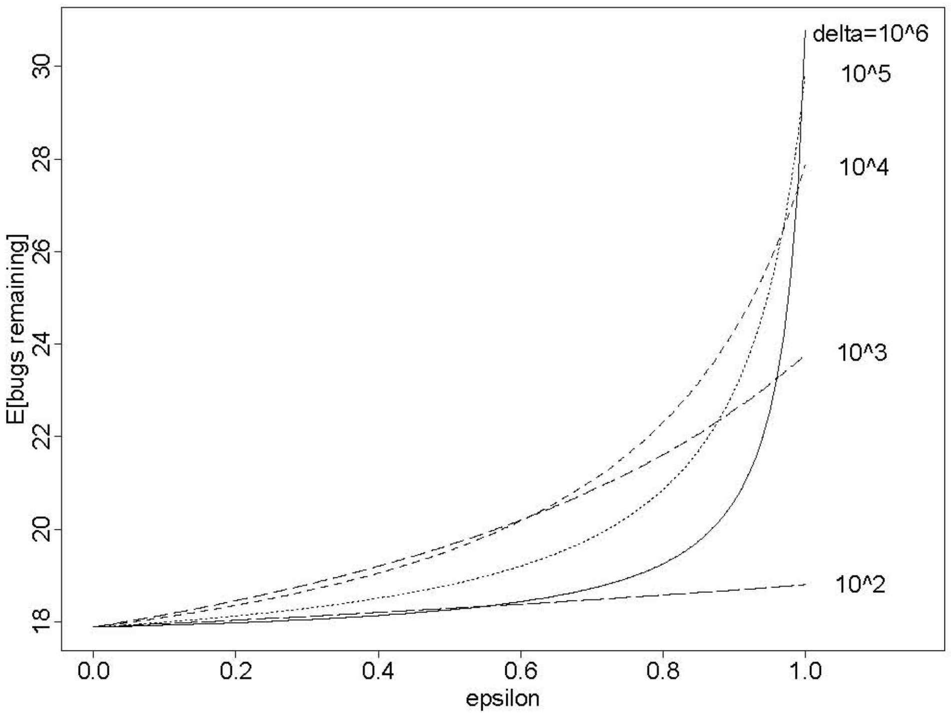
$$\Gamma = \{(1-\epsilon)P(N) + \epsilon Q(N)\}$$

where $P(N)$ represents the Poisson density and

$$Q(N) \propto \left((N-\lambda+\frac{1}{2})^2+\delta\right)^{-1}.$$

Q is a density with the same mode as the Poisson but with no mean.

The following diagram illustrates the effects of the contamination on the posterior mean of the number of remaining faults for different values of ϵ and δ.



For $\epsilon = 0.2$ the posterior median of the time to next failure varies between 2410 and 2440 for $100 \leq \delta \leq 1000000$ but when $\epsilon = 1$, the median is 1960 in the worst case, $\delta = 1000000$. We can conclude that the results are relatively insensitive to small changes in the prior.

Global sensitivity analysis

In the global approach, the prior is included within a wider class of distributions, Γ, and the sensitivity of some function of interest such as the posterior mean is examined. If there are large differences between the maximum and minimum estimates over the prior class then the inference is sensitive.

Possible classes are:

- ϵ-contamination classes

$$\Gamma = \{\pi : \ \pi(\theta) = (1-\epsilon)P(\theta) + \epsilon Q(\theta), \ g \in \mathcal{Q}\}$$

  where $\mathcal{Q}$ is a general class of contaminating distributions, e.g. unimodal distributions.

- Generalized moment classes: that is all distributions with a given set of specified moments or quantiles.

- Classes of density bands:

$$\Gamma = \{\pi : \; L(\theta) < \pi(\theta) < U(\theta)\}$$

for example, $L(\theta) = (1-\epsilon)f(\theta)$ and $U(\theta) = (1+\epsilon)f(\theta)$.

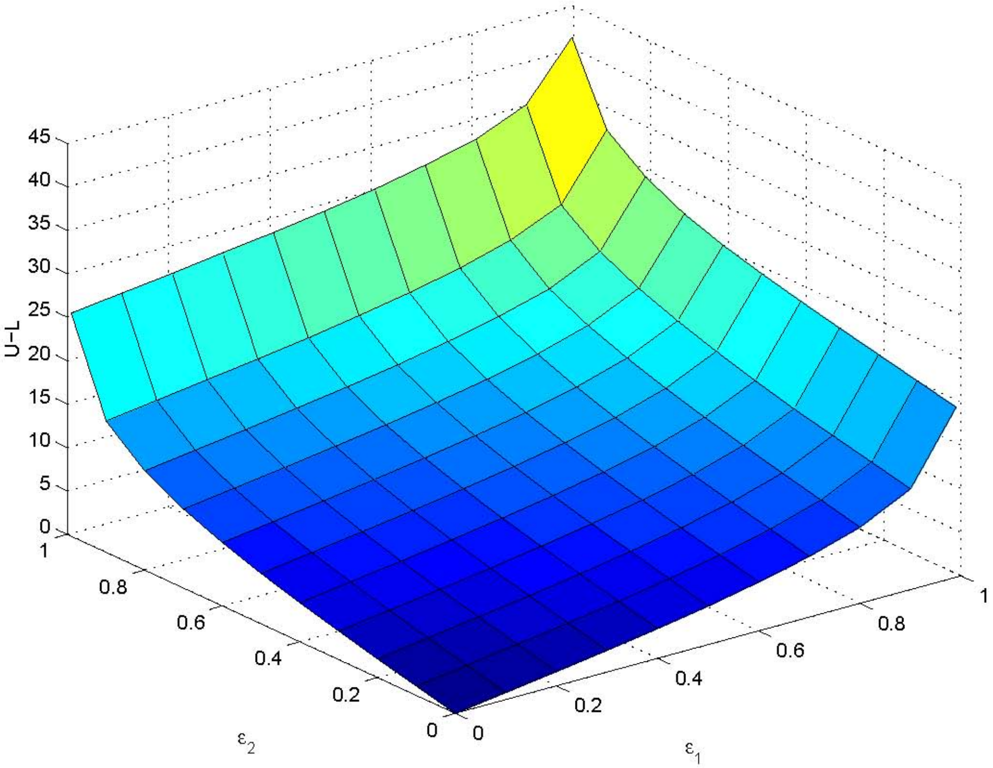
Example 86

Returning to the previous example, suppose that we wish to maintain the assumption of prior independence between N and ϕ. Then, we can define the class

$$\Gamma = \{\pi : \ \pi(N,\phi) = \pi_1(N)\pi_2(\phi)\}$$

where $(1-\epsilon_1)P(N) < \pi_1(N) < (1+\epsilon_1)P(N)$ and $(1-\epsilon_2)P(\phi) < \pi_2(\phi) < (1+\epsilon_2)P(\phi)$ and $P(N)$ and $P(\phi)$ are the Poisson and gamma priors we assumed earlier.

The diagram illustrates the differences between the upper and lower limits in the value of the posterior mean of N for $0 < \epsilon_1, \epsilon_2 < 1$.



The inference is more sensitive to contaminations of the prior of N than to contaminations in the prior for ϕ although it is still quite robust to small contaminations.

Problems with the global robustness approach

The main difficulty with this approach is that it is not always clear how to elect a reasonable contamination class. Most standard classes are too large and include unreasonable choices of prior.

A second problem is that the calculation of the minima and maxima of the quantity of interest is often very complex except for a very few, relatively simple contamination classes where analytic results are known.

Local robustness

An alternative approach is local robustness. In this case, influence measures based on, for example, the norm of the Frechet derivative of the posterior distribution relative to the prior are used to reflect the sensitivity to the the prior distribution. See e.g. Gustafson and Wassermann (1995) for more details.

Bayesian nonparametrics

Suppose that $X|f \sim f$ and that given a sample, $\mathbf{x}$, we wish to carry out inference about f.

In order to do this from a Bayesian standpoint, it is necessary to define a prior distribution over the (infinite dimensional) space of distributions. The simplest and most studied class of prior distributions is Dirichlet process priors.

The Dirichlet process

The definition of the Dirichlet process is a generalization of the Dirichlet distribution. It was first considered by Ferguson (1973).

Definition 30

Suppose that F is a random probability measure and that t F_0 is a (known) distribution function and α a scalar parameter. For any finite partition, $\{C_1, \ldots, C_r\}$, of the probability space, the Dirichlet process prior distribution for F, with parameters α and F_0 assigns the distribution

$$\{F(C_1), \ldots, F(C_r)\} \sim \mathcal{D}\left(\alpha F_0(C_1), \ldots, \alpha F_0(C_r)\right).$$

In this case, we write $F \sim \mathcal{DP}(\alpha, F_0)$.

It is straightforward to show that the Dirichlet process prior is conjugate.

Theorem 44

If $\{X\}_i$ is a sequence of exchangeable random variables with $X_i|F \sim F$ and $F \sim \mathcal{DP}(\alpha, F_0)$, then:

- the marginal distribution of X_i is F_0.
- the conditional distribution of F given a sample, $\mathbf{x} = (x_1, \ldots, x_n)$, is also a Dirichlet process such that

$$\{F(C_1), \ldots, F(C_r)\}|\mathbf{x} \sim \mathcal{D}\left(\alpha F_0(C_1) + \sum_{i=1}^{n} I_{C_1}(x_i), \ldots, \alpha F_0(C_r) + \sum_{i=1}^{n} I_{C_r}(x_i)\right)$$

that is $F|\mathbf{x} \sim \mathcal{DP}\left(\alpha + n, \frac{\alpha}{\alpha+n}F_0 + \frac{n}{\alpha+n}\hat{F}\right)$ where $\hat{F}$ is the empirical c.d.f.

Proof Firstly, the marginal c.d.f. of X is

$$\begin{aligned} P(X \leq x) &= \int P(X \leq x|F)p(F)\,dF \\ &= \int F(x)p(F)\,dF \\ &= \frac{\alpha F_0(x)}{\alpha} = F_0(x) \end{aligned}$$

Secondly, we have

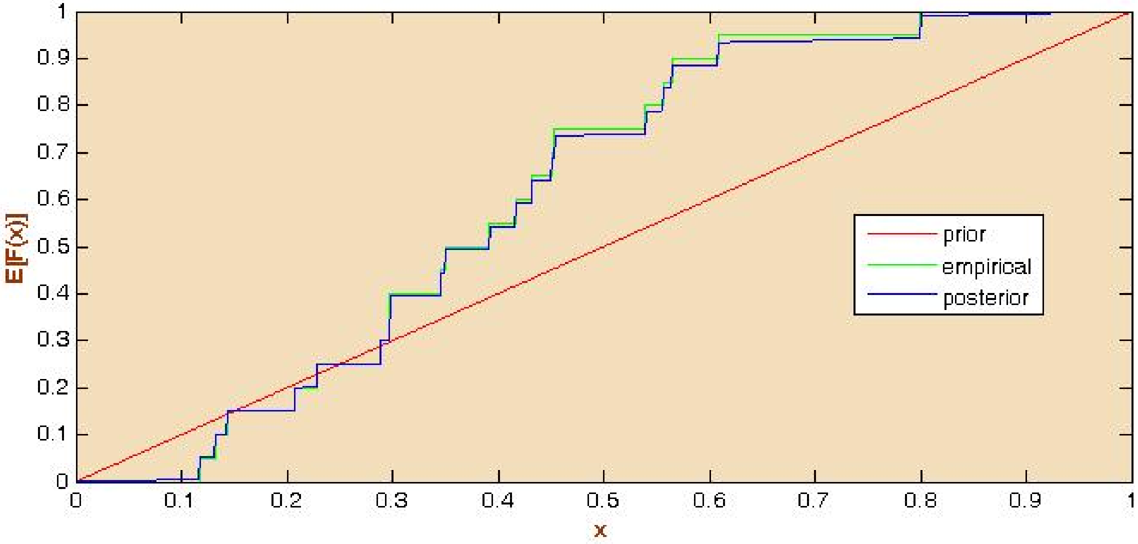
$$\begin{aligned} P(F(C_1),\ldots,F(C_r)|\mathbf{x}) &\propto P(F(C_1),\ldots,F(C_r))f(\mathbf{x}|F(C_1),\ldots,F(C_r)) \\ &\propto \prod_{j=1}^{r} F(C_j)^{\alpha F_0(C_i)+\sum_{i=1}^{n} I_{C_j}(x_i)} \end{aligned}$$

which is another Dirichlet distribution and proves the result. ■

When n increases, the predictive distribution function of X_{n+1} approaches the empirical c.d.f.

Example 87

20 data were generated from a beta distribution, $\mathcal{B}(4,6)$. A Dirichlet process prior with $\alpha = 1$ and $F_0(x) = x$, for $0 < x < 1$, i.e. a uniform distribution was assumed. The following diagram shows the predictive and empirical distribution functions.



Mixtures of Dirichlet processes

An important theoretical disadvantage of the Dirichlet process is that it can be shown that it assigns probability one to discrete probability measures, see Blackwell (1973). One way of getting around this is to use continuous mixtures of Dirichlet processes as developed in Antoniak (1974).

This leads to a hierarchical model

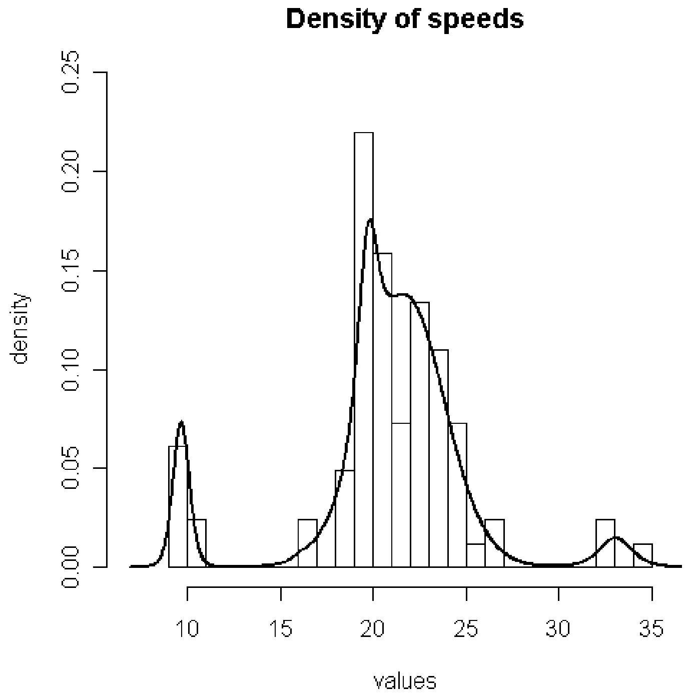
$$\begin{aligned} X_i|\theta_i &\sim f(x|\theta_i) \\ \theta_i|P &\sim P(\theta) \\ P &\sim \mathcal{DP}(\alpha, P_0) \end{aligned}$$

This model can be interpreted as a countably infinite, mixture model, i.e. as the limit of a finite mixture model, $f(x) = \sum_{i=1}^{k} w_i f(x|\theta_i)$, as the number of terms, k, goes to infinity.

One problem is how to choose the conditional density $f(x|\theta)$. Most applications have chosen to use a normal density as a flexible option. See e.g. McEachern and Muller (1998) or Neal (2000). In this case, inference is then carried out using MCMC techniques.

Example 88

The following diagram shows a fit of the well known galaxy data using the DP mixture model.



Nonparametric regression

A nonparametric regression model can be expressed as

$$y_i = f(x_i) + \epsilon_i$$

for $i = 1, \ldots, n$ where the function f is unknown. A number of classical estimation techniques are available for fitting such models, e.g. splines, neural networks or SVM's.

Bayesian penalized regression splines are implemented in Winbugs by e.g. Crainiceanu et al (2007) and a review of Bayesian neural nets is given by Lee (2004). The Bayesian equivalent of an SVM is the Gaussian process.

The Gaussian process

A Gaussian process defines a distribution over functions, f, where f is a function mapping some input space, $\mathcal{X}$ into $\mathbb{R}$. We shall call this distribution $P(f)$.

Let $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$ be an n dimensional vector of function points evaluated at $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Then $\mathbf{f}$ is a random variable.

Now, $P(f)$ is a Gaussian process if for any finite set, $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \subset \mathcal{X}$ then the marginal distribution $P(\mathbf{f})$ over that subset has a multivariate Gaussian distribution.

Gaussian processes are characterized by a mean value function $\mu(\mathbf{x})$ and a covariance function $c(\mathbf{x}, \mathbf{x}')$, so that

$$\begin{array}{rcl} P(f(\mathbf{x}), f(\mathbf{x}') & \sim & \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \quad \text{where} \\ \boldsymbol{\mu} & = & \begin{pmatrix} \mu(\mathbf{x}) \\ \mu(\mathbf{x}') \end{pmatrix} \\ \boldsymbol{\Sigma} & = & \begin{pmatrix} c(\mathbf{x}, \mathbf{x}) & c(\mathbf{x}, \mathbf{x}') \\ c(\mathbf{x}', \mathbf{x}) & c(\mathbf{x}', \mathbf{x}') \end{pmatrix} \end{array}$$

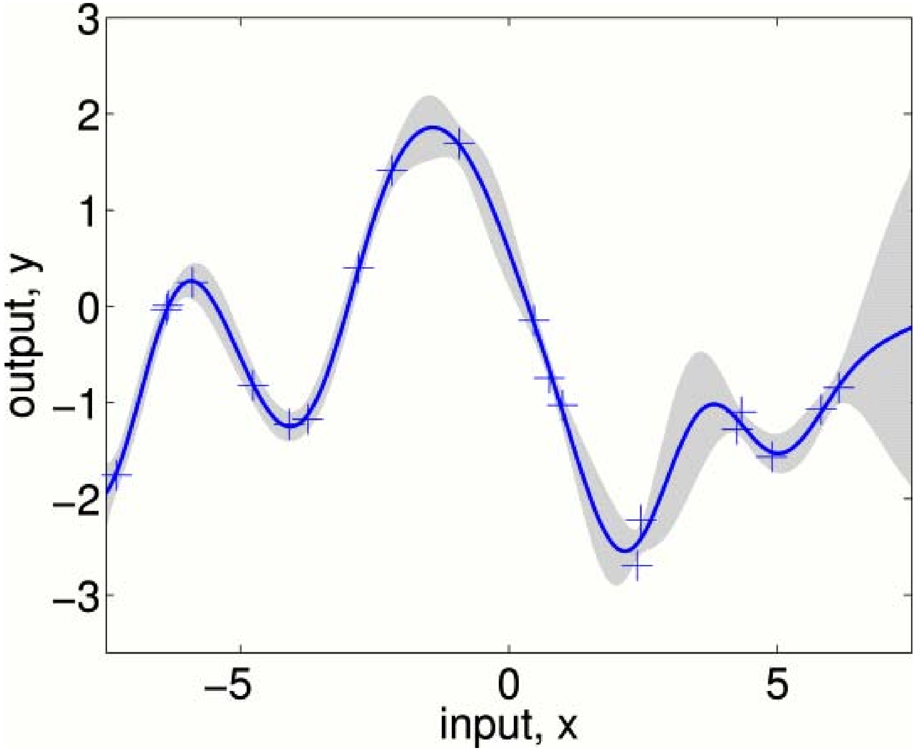
and similarly. Various forms for the covariance function have been considered, e.g.

$$c(x_i, x_j) = \nu_0 \exp\left(-\frac{|x_i - x_j|^\alpha}{\lambda}\right) + \nu_1 + \nu_2 \delta_{ij}$$

Some software for Gaussian process regression is available. See e.g.

`http://www.gaussianprocess.org/gpml/code/matlab/doc/regression.html`

Example


output, y
input, x