BAYESIAN ANALYSIS: FOREWORDS

Notation

1. "System" means the real thing and a "model" is an assumed mathematical form for the system.
2. The probability model class M contains the set of the all admissible models and the corresponding prior probability distribution over these models. The symbol M will be dropped whenever there is no confusion. One should always keep in mind that everything is conditioning on M.
3. X contains the uncertain model parameters in the assumed model class as well as other uncertain variables of interest; x is a possible value of X; $\hat{X}$ is a sample of X.
4. u is the known input to the system; uncertainties in the input should be modeled and parameterized by X; u after the conditioning symbol "|" will be dropped whenever there is no confusion.
5. Y is the uncertain response of the system and $\hat{Y}$ is the observed response from the system, i.e. the data.
6. $f(.)$ denotes probability density function; $F(.)$ denotes cumulative density function; $P(.)$ denotes probability.
7. All variables are vectors, unless particularly specified in the texts.
8. Cov(.) denotes the covariance matrix of uncertain variables; c.o.v. denotes the coefficient of variation.
9. T denotes known duration and Σ denotes covariance matrix.

Probability model class and Bayesian computation

A probability model class M specifies the prior PDF $f(x)$ and the prior predictive PDF $f(y|x)$.

Example 1:

$Y = X + E$, where $f(x) = N(2,1^2)$; $E \sim N(0,1^2)$ is the uncertain modeling error; $X \perp E$. It is clear that the prior predictive PDF $f(y|x) = N(x,1^2)$. Consider we get a data $\hat{Y} = 1$, the posterior PDF is then

$$f\left(x\mid\hat{Y}\right)=\frac{f\left(y=1\mid x\right)f\left(x\right)}{f\left(y=1\right)}\propto f\left(y=1\mid x\right)f\left(x\right)$$

$$\propto e^{-\frac{1}{2}(1-x)^2}e^{-\frac{1}{2}(x-2)^2}\propto e^{-\frac{1}{2}\left(x^2-4x+4+x^2-2x+1\right)}\propto e^{-\frac{1}{2\left(\sqrt{\frac{1}{2}}\right)^2}\left(x-\frac{3}{2}\right)^2}$$

$$=N\left(\frac{3}{2},\left(\sqrt{\frac{1}{2}}\right)^2\right)$$

Example 2:

$$A_k=g_k(X,u)\qquad k=1,..,T\qquad X\sim N(2,1^2)$$

$$Y_k=A_k+E_k\qquad E_k\sim i.i.d.N(0,1^2)\qquad E_k\perp X$$

Consider we get data $\hat{Y}=\{\hat{Y}_1,...,\hat{Y}_T\}$,

$$f\left(x\mid\hat{Y}\right)=\frac{\prod_{k=1}^{T}f\left(y=\hat{Y}_k\mid x\right)\cdot f\left(x\right)}{f\left(\hat{Y}\right)}$$

$$\propto\prod_{k=1}^{T}f\left(y=\hat{Y}_k\mid x\right)\cdot f\left(x\right)\propto\prod_{k=1}^{T}e^{-\frac{1}{2}\left(\hat{Y}_k-g_k(x,u)\right)^2}\cdot e^{-\frac{1}{2}(x-2)^2}$$

Example 3:

$$X_k=g_{k-1}(X_{k-1},u_{k-1})+W_{k-1}\qquad k=1,..,T\qquad X_0\sim N(2,1^2)\qquad W_k\sim N(0,2^2)$$
$$Y_k=h_k(X_k,u_k)+V_k\qquad V_k\sim i.i.d.N(0,1^2)\qquad V_k\perp W_k\perp X$$

Consider we get data $\hat{Y}=\{\hat{Y}_1,...,\hat{Y}_T\}$,

$$f\left(x\mid\hat{Y}\right)=f\left(x_0,...,x_T\mid\hat{Y}\right)=\frac{\prod_{k=1}^{T}f\left(y=\hat{Y}_k\mid x_k\right)\cdot\prod_{k=1}^{T}f\left(x_k\mid x_{k-1}\right)\cdot f\left(x_0\right)}{f\left(\hat{Y}\right)}$$

$$\propto\prod_{k=1}^{T}f\left(y=\hat{Y}_k\mid x_k\right)\cdot\prod_{k=1}^{T}f\left(x_k\mid x_{k-1}\right)\cdot f\left(x_0\right)$$

$$\propto\prod_{k=1}^{T}e^{-\frac{1}{2}\left(\hat{Y}_k-h_k(x_k,u_k)\right)^2}\cdot\prod_{k=1}^{T}e^{-\frac{1}{8}\left(x_k-g_{k-1}(x_{k-1},u_{k-1})\right)^2}\cdot e^{-\frac{1}{2}(x-2)^2}$$

Example 4:

$$X_k = g_{k-1}(X_{k-1}, u_{k-1}, \Theta) + W_{k-1} \qquad X_0 \sim N(\mu_0, \Sigma_0) \qquad W_k \sim N(0, \Sigma_W(\Theta))$$
$$Y_k = h_k(X_k, u_k, \Theta) + V_k \qquad V_k \sim i.i.d.N(0, \Sigma_V(\Theta)) \qquad \Theta \sim N(\mu, \Sigma)$$
$$V_k \perp W_k \perp X \perp \Theta \qquad k = 1,..,T$$

Consider we get data $\hat{Y} = \{\hat{Y}_1, ..., \hat{Y}_T\}$,

$$f\left(x \mid \hat{Y}\right) = f\left(x_0, ..., x_T, \theta \mid \hat{Y}\right) = \frac{\prod_{k=1}^{T} f\left(y = \hat{Y}_k \mid x_k, \theta\right) \cdot \prod_{k=1}^{T} f\left(x_k \mid x_{k-1}, \theta\right) \cdot f\left(x_0\right) \cdot f(\theta)}{f\left(\hat{Y}\right)}$$

$$\propto \prod_{k=1}^{T} f\left(y = \hat{Y}_k \mid x_k, \theta\right) \cdot \prod_{k=1}^{T} f\left(x_k \mid x_{k-1}, \theta\right) \cdot f\left(x_0\right) \cdot f(\theta)$$

$$\propto \prod_{k=1}^{T} e^{-\frac{1}{2}\left(\hat{Y}_k - h_k(x_k, u_k, \theta)\right)^T \Sigma_V(\theta)^{-1}\left(\hat{Y}_k - h_k(x_k, u_k, \theta)\right)}$$
$$\cdot \prod_{k=1}^{T} e^{-\frac{1}{2}\left(x_k - g_{k-1}(x_{k-1}, u_{k-1}, \theta)\right)^T \Sigma_W(\theta)^{-1}\left(x_k - g_{k-1}(x_{k-1}, u_{k-1}, \theta)\right)}$$
$$\cdot e^{-\frac{1}{2}(x_0 - \mu_0)^T \Sigma_0^{-1}(x_0 - \mu_0)} \cdot e^{-\frac{1}{2}(\theta - \mu)^T \Sigma^{-1}(\theta - \mu)}$$

Remarks:

1. We always know how to evaluate the prior PDF $f\left(x\right)$ and the predictive prior PDF $f\left(y \mid x\right)$ (or the likelihood when the data is given) because they are chosen by us when we setup the probability model class M. However, we usually don't know how to evaluate the posterior PDF $f\left(x \mid \hat{Y}\right)$ because there is unknown normalizing constant $f\left(\hat{Y}\right)$.

2. The posterior PDF $f\left(x \mid \hat{Y}\right)$ is usually not the end story for Bayesian analysis. Instead, we are usually interested in estimating some function g(X) conditioning on the data, i.e. estimate

   $E\left(g(X) \mid \hat{Y}\right) \equiv \int g(x) f(x \mid \hat{Y}) dx$

   One way of estimating this quantity is to draw N samples from $f\left(x \mid \hat{Y}\right)$ and according to the Law of Large Number:

$$E\left(g(X)\,|\,\hat{Y}\right) \equiv \int g(x) f(x\,|\,\hat{Y}) dx \approx \frac{1}{N}\sum_{i=1}^{N} g(\hat{X}_i)$$

Unfortunately, we usually cannot directly draw sample (i.e. using Monte Carlo) from $f\left(x\,|\,\hat{Y}\right)$. Most of the time, we need stochastic simulation techniques.

3. When the uncertain variable X does not change with time (Examples 1 and 2), we will solve the Bayesian problems using the block mode; when X changes with time (Example 3), we'll solve them using the sequential mode. When part of X changes with time and the other part does not, we'll solve it by combining the block mode and sequential mode.

   Block mode: Monte Carlo, rejection sampling, importance sampling, Markov chain Monte Carlo (Metropolis-Hastings, Gibbs sampler, hybrid Monte Carlo), transitional Markov chain Monte Carlo

   Sequential mode: Kalman filter, RTS smoother, backward sampler, particle filter, particle smoother

- **Prior PDF selection:**

1. Non-informative prior and Jeffrey's prior

   When we have no prior information about X, it is desirable to choose a prior PDF that does not provide any prior information. Such a prior PDF is called non-informative prior. Intuitively, the uniform PDF is non-informative; many people choose it for mathematical simplicity. But it can be seen that if X is uniformly distributed, exp(X) is not. This contradicts with the intuition that if the prior PDF of X is non-informative, that of exp(X) should also be non-informative.

   A better choice of non-informative prior is the Jeffrey's prior, which is proportional to

$$f(x) \propto \sqrt{\left|E_Y\left(-\nabla_x^2 \log\left[f\left(Y\,|\,x\right)\right]\right)\right|}$$

2. Conjugate prior

   In some rare cases, the posterior PDF will be of the same type of the prior PDF. When this happens, we say that the prior PDF is conjugate to the corresponding setting.

   Example 1:

   $Y = aX + E$, where E is Gaussian with known mean and covariance matrix; a is a known matrix. It can be shown that if X is also Gaussian with known mean and covariance matrix plus (X, E) are jointly Gaussian, the posterior PDF

$f\left(x|\hat{Y}\right)$ is also Gaussian. In fact, Gaussian prior PDF $f\left(x\right)$ is conjugate to linear models with Gaussian uncertainties.

Example 2:

$Y=E$, where $E\sim N\left(0,X\cdot I\right)$ and X is the uncertain variance parameter. It can be shown that if X is inverse Gamma, the posterior PDF $f\left(x|\hat{Y}\right)$ is also inverse Gamma. In fact, inverse Gamma prior PDF $f\left(x\right)$ is conjugate to problems with uncertain variance.

The choice of conjugate priors is usually motivated by its computational simplicity because the posterior PDF can be easily sampled.

3. Maximum entropy prior

   Under moment information about X, we can choose the least-informative prior (maximizing the differential entropy) subject to the moment constraints. This prior PDF can be found by solving the following optimization problem:

   $$\max_{f(\cdot)} -\int \log f\left(x\right)\cdot f\left(x\right)dx$$

   subject to $\int f(x)dx=1$, $\int g_1(x)f\left(x\right)dx=\mu_1$, $\int g_2(x)f\left(x\right)dx=\mu_2$, …

   $$\Rightarrow f^*\left(x\right)=e^{-(\lambda_0+\lambda_1 g_1(x)+\lambda_2 g_2(x)+\ldots)}$$