1 Priors Continued

1.1 Nuisance Parameters

Assume that unknown parameter is (θ_1, θ_2) but we are interested only in θ_1. The parameter θ_2 is called nuisance parameter. How do we handle nuisance parameters?

Suppose $\pi(\theta_1, \theta_2)$ is the joint prior for (θ_1, θ_2). The posterior is

$$\pi(\theta_1, \theta_2|x) \propto f(x|\theta_1, \theta_2) \cdot \pi(\theta_1, \theta_2).$$

The marginal posterior of interest is obtained by averaging over the nuisance parameter,

$$\pi(\theta_1|x) = \int \pi(\theta_1, \theta_2|x) d\theta_2,$$

or

$$\pi(\theta_1|x) = \int \pi(\theta_1|\theta_2, x)\pi(\theta_2|x) d\theta_2.$$

Example 1: Let $X = (X_1, \ldots, X_n)$ be a sample from normal $\mathcal{N}(\mu, \sigma)$ distribution.

Assume $\pi(\mu, \sigma^2) = \frac{1}{\sigma^2}$.

The posterior distribution is (by slight abuse of notation)

$$\begin{aligned}\pi(\mu, \sigma^2|X) &\propto \frac{1}{\sigma^{-n-2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2} \\ &= \frac{1}{\sigma^{-n-2}} e^{-\frac{1}{2\sigma^2}[(n-1)s^2+n(\bar{X}-\mu)^2]},\end{aligned}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

• Show that $\pi(\sigma^2|X)$ is $\mathcal{IG}amma(\frac{n-1}{2}, \frac{(n-1)s^2}{2})$. This distribution is sometimes referred as scaled inverse χ^2, $inv-\chi^2(n-1, s^2)$, see Handout 0.

The joint posterior can be represented as $\pi(\mu, \sigma^2|X) = \pi(\mu|\sigma^2, X) \cdot \pi(\sigma^2|X)$,

$$\begin{aligned}\pi(\mu, \sigma^2|X) &\propto \frac{1}{\sigma^{n+2}} e^{-\frac{(n-1)s^2}{2\sigma^2}} \cdot \frac{\sqrt{\pi\sigma^2/n}}{\sqrt{\pi\sigma^2/n}} e^{-n/(2\sigma^2)(\mu-\bar{X})^2} \\ &\propto \frac{1}{\sigma^{n+1}} e^{-\frac{(n-1)s^2}{2\sigma^2}} \cdot \mathcal{N}(\bar{X}, \sigma^2/n) \\ &\propto \frac{1}{(\sigma^2)^{\frac{n-1}{2}+1}} e^{-\frac{(n-1)s^2}{2\sigma^2}} \cdot \mathcal{N}(\bar{X}, \sigma^2/n) \\ &\propto \mathcal{IG}amma(\frac{n-1}{2}, \frac{(n-1)s^2}{2}) \cdot \mathcal{N}(\bar{X}, \sigma^2/n)\end{aligned}$$

Exercise 1: Derive prior and posterior predictive distributions for the model in the Example 1 above.

Exercise 2: Consider the Normal Inverse Gamma prior. This prior is conjugate. Find the parameters of the posterior.

Exercise 3: What if the prior is

$$\pi(\theta, \sigma^2) = \frac{1}{\sigma}? \tag{1}$$

This prior is noninformative independence prior since it is a product of non-informative, translation invariant priors for θ and σ^2. This prior, although not Jeffreys' for the problem, was ultimately recommended by Jeffreys in his book from (1961). For the mathematically thirsty, the prior in (1) is the right invariant Haar measure for the problem, see Berger (1985), Chapter 6.

Recall that the posterior for the parameter of interest was obtained by averaging over the nuisance parameter,

$$\pi(\theta_1|x) = \int \pi(\theta_1|\theta_2, x)\pi(\theta_2|x)d\theta_2,$$

Alternatively, one can first find the marginal likelihood, if this is convenient, and then use the prior on the parameter of interest to arrive to the same marginal posterior. Suppose $\pi(\theta_1, \theta_2) = \pi(\theta_1)\pi(\theta_2|\theta_1)$. The marginal likelihood is

$$f(x|\theta_1) = \int_{\Theta_2} f(x|\theta_1, \theta_2) \cdot \pi(\theta_1|\theta_2)d\theta_2.$$

Now,

$$\pi(\theta_1|x) \propto f(x|\theta_1)\pi(\theta_1).$$

Example 2: The following example is a simplification of a model from Vidakovic and Ruggeri (2001). The observations denoted by d are observed wavelet coefficients. The model for d is normal, with the mean θ being the parameter of interest and variance σ^2 being a nuisance parameter. If $[d|\theta, \sigma^2] \sim N(\theta, \sigma^2)$ and the prior on σ^2 is independent on θ, $[\sigma^2] \sim \mathcal{E}(\mu), \ \ \mu > 0$, with density $f(\sigma^2|\mu) = \mu e^{-\mu\sigma^2}$, the resulting marginal likelihood is

$$[d|\theta] \sim \mathcal{DE}\left(\theta, \frac{1}{\sqrt{2\mu}}\right), \quad \text{with density } f(d|\theta) = \frac{1}{2}\sqrt{2\mu}e^{-\sqrt{2\mu}|d-\theta|}.$$

Now, if the prior on θ is

$$[\theta] \sim \mathcal{DE}(0, \tau), \ \ \tau \text{ known},$$

then the (prior) predictive distribution of d is

$$[d] \sim m(d) = \frac{\tau e^{-|d|/\tau} - \frac{1}{\sqrt{2\mu}}e^{-\sqrt{2\mu}|d|}}{2\tau^2 - 1/\mu}.$$

Find the posterior distribution for $[\theta|d]$.

1.2 MaxEnt Priors

For a discrete probability distribution $p_1, \ldots, p_n$, $\sum_i p_i = 1$, the Shannon entropy is defined as

$$\mathcal{E}(p) = -\sum_i p_i \log(p_i).$$

Assume that the following restriction (better said *Information*) on the prior π is available:

$$E^{\pi}[g_k(\theta)] = \sum_i g_k(\theta_i) p(\theta_i) = \mu_k, \; k = 1, 2, \ldots, m,$$

where g_k are known functions. The maxent prior is given by

$$\pi^*(\theta_i) = \frac{\exp\{\sum_k \lambda_k g_k(\theta_i)\}}{\sum_i \exp\{\sum_k \lambda_k g_k(\theta_i)\}}.$$

The multipliers λ_i are obtained by solving the optimization problem.

Example 3: Assume $\Theta = \{0, 1, 2, \ldots\}$. Suppose that $E^{\pi}\theta = 5$. Here $g_1(\theta) = \theta$ and $\mu_1 = 5$. Thus,

$$\pi^*(\theta) = \frac{e^{\lambda_1 \theta}}{\sum_{\theta=0}^{\infty} e^{\lambda_1 \theta}} = (1 - e^{\lambda_1})(e^{\lambda_1})^{\theta}.$$

This is $\mathcal{G}eom(e^{\lambda_1})$ density and solving $\frac{1-e^{\lambda_1}}{e^{\lambda_1}} = 5$ gives $e^{\lambda_1} = 1/6$. Hence, the maxent prior is $\mathcal{G}eom(1/6)$

If the problem requires a continuous prior, the maxent approach becomes complicated. First, what should be the definition of entropy for a continuous distribution π? Jaynes (1968) argues that the entropy should be defined via Kullback-Leibler divergence between π and some invariant noninformative prior for the problem, π_0,

$$\mathcal{E}(\pi) = -E^{\pi_0}\left[\log \frac{\pi(\theta)}{\pi_0(\theta)}\right] = -\int \log \frac{\pi(\theta)}{\pi_0(\theta)} \pi_0(\theta) d\theta.$$

The maxent prior, under constraints as in the discrete case, is given by

$$\pi^*(\theta) = \frac{\exp\{\sum_k \lambda_k g_k(\theta_i)\}\pi_0(\theta)}{\int \exp\{\sum_k \lambda_k g_k(\theta_i)\}\pi_0(\theta) d\theta}.$$

The reference π_0 is thus instrumental in defining the maxent prior.

Exercise 4: If $E^{\pi}\theta = \mu$ and π_0 is a flat prior (Lebesgue measure), show that the maxent solution is $\pi^*(\theta) \propto e^{\lambda\theta}$, which cannot be normalized to a proper density. If in addition $Var(\theta) = \sigma^2$, then the maxent solution is normal $\mathcal{N}(\mu, \sigma^2)$ distribution.

1.3 Multivariate Priors

As a fundamental multivariate model we consider Multivariate Normal (MVN) likelihood and overview the MVN/MVN Bayes model. This model is not only important as an educational example that mimics normal/normal case, but also as a useful modeling tool.

Assume that $X_1, \ldots, X_n$ and their location θ are p-dimensional, and distributed as $\mathcal{MVN}(\theta, \Sigma)$. The covariance matrix Σ is assumed known. The likelihood of θ based on $X_1, \ldots, X_n$ is

$$f(x|\theta) = |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)'\Sigma^{-1}(x_i - \theta)\right)$$

which after some matrix algebra becomes

$$f(x|\theta) = |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\text{tr}(\Sigma^{-1}S)\right), \tag{2}$$

and $S = \sum_{i=1}^{n}(x_i - \theta)(x_i - \theta)'$ is the sum-of-squares matrix.

If the prior on θ is also multivariate normal $\mathcal{MVN}(\mu, \Pi)$, with vector μ and covariance matrix Π known, the posterior is

$$\pi(\theta|x) \propto \exp\left(-\frac{1}{2}\left[\sum_{i=1}^{n}(x_i - \theta)'\Sigma^{-1}(x_i - \theta) + (\theta - \mu)'\Pi^{-1}(\theta - \mu)\right]\right),$$

which, after standard matrix algebra becomes

$$\pi(\theta|x) \propto \exp\left(-\frac{1}{2}(\theta - \mu_1)'\Pi_1^{-1}(\theta - \mu_1)\right),$$

i.e., $\mathcal{MVN}(\mu_1, \Pi_1)$. The posterior parameters are matrix reformulations of the univariate case,

$$\begin{aligned} \mu_1 &= (\Pi^{-1} + n\Sigma^{-1})^{-1} \cdot (\Pi^{-1}\mu + n\Sigma^{-1}\bar{x}), \\ \Pi_1 &= (\Pi^{-1} + n\Sigma^{-1})^{-1}. \end{aligned}$$

As an illustration, we discuss the estimation of the multivariate mean for the student test scores. The subjects are: mechanics, vectors, algebra, analysis, and statistics, and records for $n = 88$ students are available (Data set and description in Mardia, Kent and Bibby, 1979).

The likelihood is $\mathcal{MVN}_p(\theta, \Sigma)$, with $\Sigma = 256I + 256J$, where I is the identity matrix and J is the matrix of ones. Part of the matlab code that calculates the Bayes estimator for θ is

```
Sigma = 256*eye(p)+256*ones(p,p);
mu = 50*ones(p,1);
Pi = 4* eye(p)+ 4* ones(p,p);
mu_1 = inv (inv(Pi) + n * inv(Sigma)) * ...
     ( inv(Pi) * mu + n * inv(Sigma) * (mean(scores))' );
Pi1 = inv (inv(Pi) + n * inv(Sigma));
```

and the complete m-file producing the figure and containing the data set is `bayes6_1.m`.

References

[1] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer Verlag.

[2] Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate Analysis.* London: Academic Press.

[3] Robert, C. (2001). *Bayesian Choice*, Second Edition, Springer Verlag.

[4] Vidakovic, B. and Ruggeri, F. (2001). BAMS Method: Theory and Simulations. *Sankhyā, Series B,* **63**,*2* (Special Issue on Wavelets), 234–249.

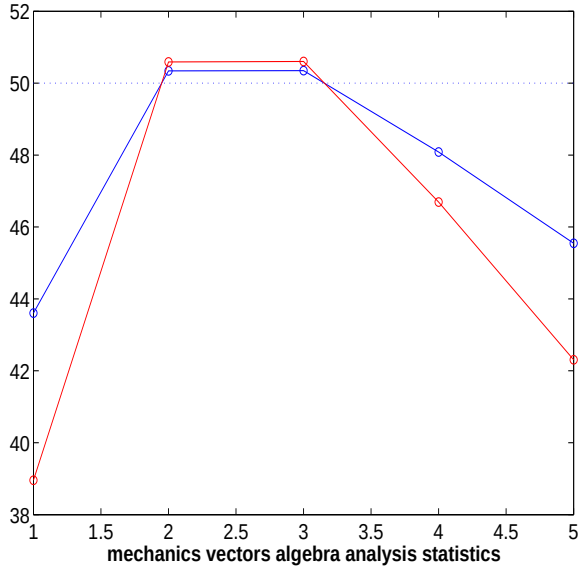

Figure 1: Multivariate Bayes estimator for the student scores from Mardia, Kent and Bibby, 1979. The subjects are: mechanics, vectors, algebra, analysis, and statistics.