Stat260: Bayesian Modeling and Inference **Lecture Date: February 10th, 2010**

Jeffreys priors

Lecturer: Michael I. Jordan *Scribe: Timothy Hunter*

1 Priors for the multivariate Gaussian

Consider a multivariate Gaussian variable X of size p. Its probability density function can be parametrized by a mean vector $\mu \in \mathbb{R}^p$ and a covariance matrix $\Sigma \in \mathcal{S}_p^1$:

$$p(X|\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}} \frac{1}{\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)\right) \tag{1}$$

We will consider three cases of conjugate priors: the case when the covariance is fixed, the case when the mean is fixed and the general case.

1.1 The case of fixed variance

The conjugate prior is a multivariate Gaussian of mean μ_0 and covariance matrix Σ_0. The derivations are the same as in the univariate case.

1.2 The case of fixed mean

The conjugate prior is the *inverse Wishart distribution*. One can see this using the trace trick:

$$(X-\mu)^T \Sigma^{-1}(X-\mu) = \mathrm{Tr}\left(\Sigma^{-1}(X-\mu)(X-\mu)^T\right) \tag{2}$$

This is an inner product in the matrix space between the matrices Σ^{-1} and $(X-\mu)(X-\mu)^T$. This particular inner product corresponds to summing all the pairwise products of elements of each matrix. We will derive here the Wishart distribution, and you can derive the inverse Wishart by a change of variable. The computations are similar for the inverse Wishart distribution.

The distribution over the precision matrix is the Wishart distribution with one degree of freedom is:

$$p(M|V) = \frac{|M|^{p/2} e^{-\frac{1}{2}\mathrm{Tr}(MV^{-1})}}{2^{p/2}|V|^{1/2}\Gamma_p\left(\frac{1}{2}\right)} \tag{3}$$

where Γ_p is the generalized (multivariate) Gamma function, and M, V are positive definite matrices. When we forget the normalization constants, we recognize the product of the determinant (up to some exponent) with the exponential of the trace, which is the familiar form of the multivariate Gaussian:

$$p(M|V) \propto |M|^{p/2} e^{-\frac{1}{2}\mathrm{Tr}(MV^{-1})} \tag{4}$$

This is the multivariate analog of a Gamma distribution. In order to work with the covariance matrix and get the inverse Wishart distribution, one has to apply the change of variable $\Sigma = P^{-1}$. This shape of the inverse Wishart looks very close to that of the inverse gamma:

$$p(\Sigma|V) \propto |\Sigma|^{-(p+1)/2} e^{-\frac{1}{2}\mathrm{Tr}(\Sigma^{-1}V^{-1})} \tag{5}$$

If we want to get a predictive distribution, we integrate the inverse Wishart against the multivariate Gaussian which gives gives the multivariate Student distribution:

$$T \propto \frac{1}{\left(1 + (X-\mu)^T \Sigma^{-1} (X-\mu)\right)^{p/2}} \tag{6}$$

with a complicated with a heavy tail.

1.3 The general case

The computations are the same as before with an inverse Wishart for the covariance and a scaled Gaussian (scaled by the Wishart).

1.4 Sampling from the Wishart distribution: the Bartlett decomposition

If one needs to sample from the Wishart, there is an nice way to sample it called the Bartlett decomposition. Consider the Cholesky decomposition of the parameter:

$$V = LL^T \tag{7}$$

Samples of Σ are obtained by sampling

$$\Sigma = LZZ^TL^T \tag{8}$$

where

$$Z = \begin{pmatrix} \sqrt{c_1} & & & & \\ z_{21} & \sqrt{c_2} & & & \\ z_{31} & & \sqrt{c_3} & & \\ \vdots & & & \ddots & \\ z_{p1} & z_{p2} & & \cdots & \sqrt{c_p} \end{pmatrix} \tag{9}$$

in which the diagonal coefficients are from the χ^2 distribution with p degrees of freedom and the z_{ij} are from the univariate Gaussian distribution $\mathcal{N}(0,1)$.

There is a similar way to sample from the multivariate Gaussian distribution. Consider the multivariate Gaussian with identity matrix $X \sim \mathcal{N}(0, I_p)$. This is easy to sample from: each coefficient can be sampled independently by a univariate Gaussian. We use the Cholesky (or the square root) decomposition of the covariance matrix

$$\Sigma = LL^T \tag{10}$$

We then define a new random variable $W = LX$, with 0 mean and covariance $\mathrm{Var}(W) = \mathbb{E}[WW^T] - 0 = LL^T = \Sigma$. Therefore W as defined this way can be described a 0-mean Gaussian with covariance Σ. Getting a different mean is simply a matter of translation W.

This concludes our introduction to conjugate priors. Conjugate priors are a matter of convenience, easy to implement and as such widely used in software implementations. They have some nice properties, in particular they are optimal asymptotically. They are often used in applications, when one lacks prior knowledge. Using conjugate priors, only needs to assess the prior parameters.

2 Jeffreys priors

Though conjugate priors are computationally nice, objective Bayesians instead prefer priors which do not strongly influence the posterior distribution. Such a prior is called an *uninformative prior*.

This is a hard problem, and a number of things we might try are not appropriate. The historical approach, followed by Laplace and Bayes, was to assign flat priors. This prior seems reasonably uninformative. We do not know where the actual value lies in the parameter space, so we might as well consider all values equiprobable. This prior however is not invariant. Consider for example a binomial distribution $X \sim \text{Binom}(n, \theta)$ in which we want to put a prior on θ. We know that θ lies between 0 and 1. The flat prior on θ is the uniform distribution: $\pi(\theta) = 1$. Since θ lies between 0 and 1, we can use a new parametrization using the log-odds ratio: $\rho = \log \frac{\theta}{1-\theta}$. This is a perfectly valid parametrization, and a natural one if we want to map θ to the full scale of the reals. Under this parametrization the prior distribution $\pi(\rho)$ is not flat anymore. This example shows a prior that is uninformative in one parametrization, but becomes informative through a change of variables.

This becomes more problematic in higher dimensions: the uniform prior in large dimension does not integrate anymore. In addition, the flat prior becomes very informative: it tells that most of the probability mass lies at $+\infty$, far from the origin. If instead one considers a high-dimensional Gaussian distribution $X \sim \mathcal{N}(0, 1)$, most of the mass is concentrated in a (high dimensional) unit sphere centered at the origin.

Faced with these issues, we see that flat priors and uninformative priors raise mathematical and philosophical problems. These examples show that finding prior distributions that have a minimal impact as possible on the data raises deep practical issues.

We first consider some special cases in one dimension, then consider the general case.

2.1 Examples

2.1.1 The example of an uninformative location prior

Consider the case where we have a location parameter: a probability distribution over a variable X of density $f(X - \theta)$ where θ is a *location parameter* that we endow with a prior. A candidate for a prior would be $\pi(\theta) \propto 1$. If θ lies in an interval, we can consider the uniform distribution as a prior estimate. If θ can take any value in $\mathbb{R}$, the flat prior is not a probability density because it does not integrate. Such a prior is called an *improper prior*. It expresses our state of ignorance (hence the flat prior) and can be defined as the limit of a proper prior.

2.1.2 The example of an uninformative scaling prior

Consider a density factor θ:

$$f_\theta(x) = \frac{1}{\theta} f\left(\frac{x}{\theta}\right) \tag{11}$$