Gaussian Conjugate Prior Cheat Sheet

Tom SF Haines

1 Purpose

This document contains notes on how to handle the multivariate Gaussian[1] in a Bayesian setting. It focuses on the conjugate prior, its Bayesian update given evidence and how to collapse (integrate out) drawing from the resulting posterior. Sampling is also covered for completeness. All content has been initially gleaned from the book *Bayesian Data Analysis* by Gelman, Carlin, Stern and Rubin, with extra input from many other sources, plus a certain amount of working out on my part, mostly for consistency verification purposes. Main motivation is that even the referenced book does not give all the useful equations, and I wanted all the equations in one clean document with consistent notation to minimise the chances of a mistake. I also prefer to work with the precision matrix rather than the covariance matrix.

2 Introducing...

The multivariate Gaussian distribution can be given in terms of its density as

$$P(x|\mu,\Lambda) = \frac{\exp(-0.5(x-\mu)^T\Lambda(x-\mu))}{(2\pi)^{d/2}|\Lambda|^{-0.5}} \tag{1}$$

where x and μ are length d vectors and Λ is the $d \times d$ precision matrix. $|\cdot|$ indicates the determinant. Λ, the precision matrix, is the inverse of the covariance matrix that is usually used, $\Lambda = \Sigma^{-1}$. Both the precision and covariance matrices are symmetric and positive definite (Inverse operation maintains these two properties.). Denominator of the fraction is the normalising constant. The Gaussian is generally denoted $x \sim \mathcal{N}(\mu, \Sigma)$,

[1] Otherwise known as the normal distribution, which is really silly as there is nothing inherently normal about it. People just like to pretend they can apply it to just about anything without consequence.

a convention held throughout this document despite the use of precision rather than covariance.

3 Conjugate prior

The conjugate prior of the multivariate Gaussian is comprised of the multiplication of two distributions, one for each parameter, with a relationship to be implied later. Over the mean, μ, is another multivariate Gaussian; over the precision, Λ, is the Wishart distribution.

For the purpose of understanding the Wishart distribution a draw can be represented as[2]

$$\Lambda \sim \mathcal{W}(V, n) = \sum_{i \in [1,n]} x_i x_i^T, \qquad x_i \sim \mathcal{N}(\mathbf{0}, V) \tag{2}$$

where $\mathcal{N}(\mathbf{0}, V)$ is a draw from the Gaussian with a mean of zero and a covariance of V. This is quite simply the scatter matrix of n draws from a Gaussian. The actual distribution, which is only valid when $n \geq d$, d being the number of dimensions, is given by its density as

$$P(\Lambda|V, n) = \frac{|\Lambda|^{(n-d-1)/2} \exp(-0.5\text{trace}(\Lambda V^{-1}))}{2^{nd/2}|V|^{n/2}\Gamma_d(n/2)} \tag{3}$$

where $\Gamma_d(\cdot)$ is the generalised multivariate Gamma function, which is defined in terms of the normal Gamma function as

$$\Gamma_d(n/2) = \pi^{d(d-1)/4}\Pi_{i \in [1,d]}\Gamma((n+1-i)/2) \tag{4}$$

Note that this definition of the Wishart allows n to be any real value, rather than just a natural number, which can be useful for a weak prior. Naming convention is to refer to n as the *degrees of freedom* and V as the scale matrix.

Using the Wishart distribution we may define a draw of the parameters necessary for a Gaussian, $\mathcal{N}(\mu, \Lambda^{-1})$, as

$$\Lambda \sim \mathcal{W}(\Lambda_0, n_0) \tag{5}$$

$$\mu|\Lambda \sim \mathcal{N}(\mu_0, (k_0\Lambda)^{-1}) \tag{6}$$

The four given parameters - $n_0, k_0 \in \mathbb{R}$, $\mu_0 \in \mathbb{R}^d$ and $\Lambda_0 \in \mathbb{R}^{d \times d}$ - parametrise the conjugate prior over a multivariate Gaussian distribution.

[2]It can help to contrast this with the definition of the Gamma distribution, for which this is one possible multivariate generalisation.

4 Bayesian Update

Given a prior and new evidence naturally updating to get a posterior is desired. The previously given choice of parameters make this easy:

$$n_m = n_0 + m \tag{7}$$

$$k_m = k_0 + m \tag{8}$$

$$\mu_m = \frac{k_0\mu_0 + m\bar{x}}{k_0 + m} \tag{9}$$

$$\Lambda_m = \left(\Lambda_0^{-1} + S + \frac{k_0 m}{k_0 + m}(\bar{x} - \mu_0)(\bar{x} - \mu_0)^T\right)^{-1} \tag{10}$$

where

$$S = \sum_{i \in [1,m]} (x_i - \bar{x})(x_i - \bar{x})^T \tag{11}$$

which is the scatter matrix of the evidence. As should be quite obvious from the context the x_i are the m samples that constituent the evidence and $\bar{x}$ is their mean. Given the nature of the update of Λ_m it can be computationally advantageous to store its inverse instead, at least whilst performing lots of Bayesian updates. It would also not be that unusual to merge n_m and k_m given their identical update, just storing separate initial values.

5 Integrating out

Given a posterior calculated using the above one would traditionally draw a Gaussian from it, which is in turn used to determine the probability of a specific sample, x, being drawn. Alternatively one can integrate out the intermediate Gaussian, which is highly advantageous if the evidence only has a few samples such that the Gaussian is not well defined. This occurs with a Dirichlet process mixture model for instance - when Gibbs sampling you have to work out the probability of a sample being drawn from a Gaussian drawn directly from the prior, without any extra evidence. There are two variables to integrate out - μ and Λ, and they can be done in sequence.

To remove the mean, μ, it has to be summed out; for the moment we can ignore the probability distribution on Λ as it has no relationship to μ

$$P(x|k_0, \mu_0, \Lambda) = \int P(x|\mu, \Lambda)P(\mu|\mu_0, k_0, \Lambda)d\mu \tag{12}$$

where the two rhs probabilities are Gaussian, and using a slight abuse of notation are given by

$$P(x|\mu,\Lambda) = x \sim \mathcal{N}(\mu,\Lambda^{-1}) \tag{13}$$

$$P(\mu|k_0,\mu_0,\Lambda) = \mu \sim \mathcal{N}(\mu_0,(k_0\Lambda)^{-1}) \tag{14}$$

This can be interpreted as the convolution of one Gaussian by another

$$f(\mu) = \mu \sim \mathcal{N}(\mu_0,(k_0\Lambda)^{-1}) \tag{15}$$

$$g(x-\mu) = x-\mu \sim \mathcal{N}(0,\Lambda^{-1}) \tag{16}$$

$$P(x|k_0,\mu_0,\Lambda) = \int f(\mu)g(x-\mu)d\mu \tag{17}$$

for which the result is well known to be yet another Gaussian

$$P(x|k_0,\mu_0,\Lambda) = x \sim \mathcal{N}\left(\mu_0,\left(\frac{\Lambda}{1+1/k_0}\right)^{-1}\right) \tag{18}$$

The next, and final, step is to integrate out Λ, for which the equation is

$$P(x|n_0,k_0,\mu_0,\Lambda_0) = \int P(x|k_0,\mu_0,\Lambda)P(\Lambda|\Lambda_0,n_0)d\Lambda \tag{19}$$

where $P(x|k_0,\mu_0,\Lambda)$ is given by equation 18 and

$$P(\Lambda|\Lambda_0,n_0) = \Lambda \sim \mathcal{W}(\Lambda_0,n_0) \tag{20}$$

which is the Wishart distribution. The answer is defined in terms of the multivariate Student-t[3] distribution, which has the density

$$P(\theta|v,\mu,\Sigma) = \frac{\Gamma((v+d)/2)}{\Gamma(v/2)(v\pi)^{d/2}|\Sigma|^{1/2}}\left(1+\frac{1}{v}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\right)^{-(v+d)/2} \tag{21}$$

where d is the number of dimensions, $v \in \mathbb{R}, v > 0$ the degrees of freedom, $\mu \in \mathbb{R}^d$ the location and $\Sigma \in \mathbb{R}^{d\times d}$ the scale matrix, which is symmetric and positive definite. Note that this equation involves extremely large values that cancel out - all implementations should use logarithms to avoid numerical overflow, and directly calculate the log of the gamma function. For notation

[3]Unlike the univariate case there is more than one definition of the Student-t distribution in the multivariate case. The definition given is the most common however.

$\theta \sim \mathcal{T}(v, \mu, \Sigma)$ is used, at least within this document[4]. Skipping directly to the answer to equation 19 you get

$$P(x|n_m, k_m, \mu_m, \Lambda_m) = x \sim \mathcal{T}\left(n_m - d + 1, \mu_m, \left(\frac{k_m(n_m - d + 1)}{(k_m + 1)}\Lambda_m\right)^{-1}\right) \tag{22}$$

which gives the probability of a new sample given the evidence, with the intermediate draw integrated out. Note that it is easy to avoid inverting Λ_m using the rule that $|\Sigma| = |\Sigma^{-1}|^{-1}$.

6 Sampling

Whilst the previous section discussed how to avoid sampling in one useful situation by integrating the parameters of the Gaussian out it is inevitable to actually want to draw a sample - this can be divided into two steps - drawing from a Wishart distribution and then drawing from a multivariate Gaussian distribution, as in equations 5 and 6.

6.1 Sampling the Wishart distribution

The Bartlett decomposition of $\Lambda \sim \mathcal{W}(V, n)$ is $\Lambda = LAA^TL^T$, where L is the Cholesky decomposition of V and A is formed of 3 parts - a diagonal, an upper triangle and a lower triangle. The upper triangle is zeroed out, the diagonal consists of the square root of draws from $\sim \mathcal{X}^2(n - d + 1)$, the Chi-squared distribution, and the lower triangle consists of draws from a Gaussian, $\sim \mathcal{N}(0, 1)$, i.e.

$$\begin{bmatrix} \sqrt{a_1} & 0 & 0 & \cdots & 0 \\ b_{21} & \sqrt{a_2} & 0 & \cdots & 0 \\ b_{31} & b_{32} & \sqrt{a_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{d1} & b_{d2} & b_{d3} & \cdots & \sqrt{a_d} \end{bmatrix} \tag{23}$$

$$\forall i \in [1, d]; a_i \sim \mathcal{X}^2(n - d + 1) \tag{24}$$

$$\forall i \in [1, d], j \in [1, d], i > j; b_{ij} \sim \mathcal{N}(0, 1) \tag{25}$$

[4] A lower-case t is more traditional, but is inconsistent with the other distributions; also v is often given as a subscript rather than a parameter, which is just stupid.

For completeness the Chi-squared distributions density is given by

$$x \sim \mathcal{X}^2(k) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)} \qquad , x \in \mathbb{R}, x \geq 0, k \in \mathbb{N}, k > 0 \tag{26}$$

It is limited by being defined only for k as a positive integer, however, it is a special case of the Gamma distribution, which allows a continuous generalisation, specifically

$$x \sim \mathcal{X}^2(k) = x \sim \Gamma(k/2, 1/2) \tag{27}$$

where the Gamma distributions density is given by[5]

$$x \sim \Gamma(\alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)} \tag{28}$$

6.2 Sampling the Gaussian distribution

This is easily Googleable knowledge, but is included for completeness. Given $x \sim \mathcal{N}(\mu, \Sigma)$ then $x = \mu + Az$, where A is the Cholesky decomposition of Σ and z is a vector of draws from $\mathcal{N}(0, 1)$.

7 1D equations

All of the previous equations obviously work directly with 1D data without modification, but for convenience the important equations from above are now provided in their 1D form, with the relevant simplifications. The 1D Gaussian can be given by

$$x \sim \mathcal{N}(\mu, \sigma^2) = P(x|\mu, \sigma^2) = \frac{\exp(-0.5\sigma^{-2}(x-\mu)^2)}{\sqrt{2\pi\sigma^2}} \tag{29}$$

where standard deviation, σ, has been used, such that $[\sigma^{-2}] = \Lambda$. All quantities are obviously now scalars rather than vectors/matrices.

The Wishart distribution simplifies to become the Gamma distribution, as given in equation 28

$$P(\sigma^{-2}|V, n) = \frac{\sigma^{-2(n/2-1)}\exp(-0.5\sigma^{-2}V^{-1})}{(2V)^{n/2}\Gamma(n/2)} \tag{30}$$

[5]There are two definitions used in the literature - the version given and a version where the inverse of β is used instead - it is often not clear which version is in use.

$$P(\sigma^{-2}|V,n) = \sigma^{-2} \sim \Gamma(\frac{n}{2}, \frac{1}{2V}) \tag{31}$$

Drawing from the prior therefore consists of

$$\sigma^{-2} \sim \Gamma(\frac{n_0}{2}, \frac{\sigma_0^2}{2}) \tag{32}$$

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/k_0) \tag{33}$$

and has the four parameters n_0, k_0, μ_0 and σ_0^2.

Performing a Bayesian update is mostly notationally identical with some slight adjustments due to the use of variance rather than inverse variance, and consists of

$$n_m = n_0 + m \tag{34}$$

$$k_m = k_0 + m \tag{35}$$

$$\mu_m = \frac{k_0\mu_0 + m\bar{x}}{k_0 + m} \tag{36}$$

$$\sigma_m^2 = \sigma_0^2 + \sum_{i\in[1,m]} (x_i - \bar{x})^2 + \frac{k_0 m}{k_0 + m}(\bar{x} - \mu_0)^2 \tag{37}$$

Integrating out the draw from the distribution again results in a student-t distribution, but this time the well known univariate case, which is given by

$$\theta \sim t(v, \mu, \sigma^2) = P(\theta|v, \mu, \sigma^2) = \frac{\Gamma((v+1)/2)}{\Gamma(v/2)\sigma\sqrt{v\pi}} \left(1 + \frac{(\theta - \mu)^2}{v\sigma^2})\right)^{-(v+1)/2} \tag{38}$$

The probability of x given the posterior parameters is then

$$P(x|n_m, k_m, \mu_m, \sigma_m^2) = x \sim t\left(n_m, \mu_m, \frac{(k_m + 1)}{k_m n_m}\sigma_m^2\right) \tag{39}$$

Sampling from the posterior parameters is obvious and involves only basic functions, hence it is omitted.