Exponential Families:
Gaussian, Gaussian-Gamma, Gaussian-Wishart, Multinomial

Yee Whye Teh
Gatsby Computational Neuroscience Unit,
University College London,
17 Queen Square, London WC1N 3AR, United Kingdom
ywteh@gatsby.ucl.ac.uk

August 13, 2007

1 Gaussian-Wishart

This section derives the formulas for a Gaussian distribution along with a Gaussian-Wishart conjugate prior.

Data:

n	The number of data vectors.
$\boldsymbol{x}$	The data vectors $x_1, \ldots, x_n$.
X	$= \sum_{i=1}^{n} x_i$.
C	$= \sum_{i=1}^{n} x_i x_i^\top$.

Parameters:

μ	Mean of data.
R	Precision of data.

Hyperparameters:

d	Dimensionality of data.
r	Relative precision of μ versus data. The precision of μ is rR.
ν	Degrees of freedom of precision of R.
m	Mean of μ is m.
S	Mean of R is νS^{-1}.

Prior:

$$\text{Wishart:} \qquad p(R) = 2^{-\nu d/2} \pi^{-d(d-1)/4} |S|^{\nu/2} \prod_{i=1}^{d} \Gamma\left(\frac{\nu+1-i}{2}\right)^{-1} |R|^{(\nu-d-1)/2} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[RS\right]\right) \tag{1}$$

$$\text{Gaussian:} \qquad p(\mu \,|\, R) = (2\pi)^{-d/2} |rR|^{1/2} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[rR\left((\mu-m)(\mu-m)^\top\right)\right]\right) \tag{2}$$

$$p(\mu, R) = \frac{1}{Z(d,r,\nu,S)} |R|^{(\nu-d)/2} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[R\left(r(\mu-m)(\mu-m)^\top + S\right)\right]\right) \tag{3}$$

$$= \frac{1}{Z(d,r,\nu,S)} |R|^{(\nu-d)/2} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[R\left(r\mu\mu^\top - 2\mu(rm)^\top + rmm^\top + S\right)\right]\right) \tag{4}$$

$$\text{where} \quad Z(d,r,\nu,S) = 2^{\frac{(\nu+1)d}{2}} \pi^{d(d+1)/4} r^{-d/2} |S|^{-\nu/2} \prod_{i=1}^{d} \Gamma\left(\frac{\nu+1-i}{2}\right) \tag{5}$$

Likelihood of data:

$$p(\boldsymbol{x} \,|\, \mu, R) = (2\pi)^{-nd/2} |R|^{n/2} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[R\left(\sum_{i=1}^{n}(\mu-x_i)(\mu-x_i)^\top\right)\right]\right) \tag{6}$$

$$= (2\pi)^{-nd/2} |R|^{n/2} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[R\left(n\mu\mu^\top - 2X\mu^\top + C\right)\right]\right) \tag{7}$$

Joint likelihood:

$$p(\boldsymbol{x}, \mu, R) = \frac{(2\pi)^{-nd/2}}{Z(d,r,\nu,S)} |R|^{(\nu+n-d)/2} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[R\left((r+n)\mu\mu^\top - 2\mu(rm+X)^\top + rmm^\top + C + S\right)\right]\right) \tag{8}$$

$$= \quad \cdots \quad \left[R\left((r+n)\left(\mu - \tfrac{rm+X}{r+n}\right)\left(\mu - \tfrac{rm+X}{r+n}\right)^\top - \tfrac{(rm+X)(rm+X)^\top}{r+n} + rmm^\top + C + S\right)\right] \tag{9}$$

Posterior hyperparameters:

$$\begin{aligned} r' &= r + n \\ \nu' &= \nu + n \\ m' &= \frac{rm + X}{r + n} \\ S' &= S + C + rmm^\top - r'm'm'^\top \end{aligned} \tag{10}$$

Marginal probability:

$$p(\boldsymbol{x}) = (2\pi)^{-nd/2} \frac{Z(d, r', \nu', S')}{Z(d, r, \nu, S)} = \pi^{-nd/2} \frac{r^{d/2} |S|^{\frac{\nu}{2}}}{r'^{d/2} |S'|^{\frac{\nu'}{2}}} \prod_{i=1}^{d} \frac{\Gamma\left(\frac{\nu'+1-i}{2}\right)}{\Gamma\left(\frac{\nu+1-i}{2}\right)} \tag{11}$$

Predictive probability:

$$p(y|\boldsymbol{x}) = \frac{p(y, \boldsymbol{x})}{p(\boldsymbol{x})} \tag{12}$$

Posterior:

$$p(\mu, R \,|\, \boldsymbol{x}) = \frac{1}{Z(d, r', \nu', S')} |R|^{(\nu'-d)/2} \exp\left(-\frac{1}{2} \operatorname{Tr}\left[R\left(r'(\mu - m')(\mu - m')^\top + S'\right)\right]\right) \tag{13}$$

Computations in mixture models:

We are interested in mixture models, in which each component is a Gaussian and the prior for parameters is the Gaussian-Wishart distribution. All mixture components share the same hyperparameters, but each component has its own set of parameters. In the posterior of the mixture model, the distribution for which data vectors belong to which components and the distribution over parameters are coupled and cannot be solved exactly. We have two choices: variational Bayes, or Markov chain Monte Carlo (MCMC) sampling.

We develop our technique for MCMC sampling here. In particular we Gibbs sample indicator variables which determine the components to which data vectors belong to, while integrating out the parameters. The only computations needed are marginal likelihood computations, and posterior computations for the parameters (which amounts to computing the posterior hyperparameters r', ν', m' and S').

First consider computing the marginal likelihood. Note that to compute conditional likelihoods $p(\boldsymbol{x} \,|\, \boldsymbol{y})$ we just evaluate $p(\boldsymbol{x}, \boldsymbol{y})/p(\boldsymbol{y})$. There are two complications, computing $|S|$ and $|S'|$ efficiently, and computing the ratio of Gamma terms in (11).

We deal with computing the determinants $|S|$ and $|S'|$ by representing S and S' using their Cholesky decompositions. In particular, updates to S and S' will be carried out by directly updating their Cholesky decompositions. Given the Cholesky decomposition the determinant is just the product of the diagonal terms. We will come to updates to the Cholesky decomposition later.

To efficiently compute the ratio of Gamma terms, we note that ν is the same for every component, and the ratio of Gamma terms can only taken on at most $N + 1$ values, one for each $n = 0, 1, \ldots, N$, where N is the total number of data vectors. By expanding the Gamma terms, we obtain:

$$\prod_{i=1}^{d} \frac{\Gamma\left(\frac{\nu'+1-i}{2}\right)}{\Gamma\left(\frac{\nu+1-i}{2}\right)} = \begin{cases} \left(\prod_{j=1}^{\frac{n}{2}} \left(\prod_{i=1}^{d} \frac{\nu-1-i}{2} + j\right)\right) & \text{if } n \text{ is even.} \\ \left(\prod_{j=1}^{\frac{n-1}{2}} \left(\prod_{i=1}^{d} \frac{\nu-i}{2} + j\right)\right) \left(\prod_{i=1}^{d} \frac{\Gamma\left(\frac{\nu+2-i}{2}\right)}{\Gamma\left(\frac{\nu+1-i}{2}\right)}\right) & \text{if } n \text{ is odd.} \end{cases} \tag{14}$$

We see that we can compute all (14) with $O(dN)$ simple arithmetic operations, and only need to evaluate the Gamma function $2d$ times. We compute these once and store the results to be used in the future. In fact, we do not need to precompute (14) for all n. We can compute and store them as we need.

Now consider updates to the posterior hyperparameters. In our setting we only need to worry about when a data vector is added to or removed from a component. Without loss of generality we assume r, ν, m and S are the

posterior hyperparameters for the component before adding/removing this new data vector into the component. The only complication is in updates to S', since we are representing S and S' using their Cholesky decompositions, say $S = M^\top M, S' = M'^\top M'$ where M, M' are upper triangular matrices. Note that S' differs from S only via three symmetric rank one matrices. Hence we can compute M' from M using three rank one Cholesky update, which takes $O(d^2)$ operations each.