Latent Gaussian Bayesian Parameter Learning (v2.0)

EJ LOUW
15600998

1 The Marginal over X

$$P(K) = \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} e^{-tr(V^{-1}K)/2}$$

$$P(\mu|K) = \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])}$$

$$P(X|\mu, K) = \frac{|K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([X-\mu]^T K[X-\mu])}$$

Note: It was initially (and stupidly) assumed that the $P(\mu|K)$ and $P(K)$ can simply be multiplied to give a Gaussian-Wishart distribution. This is obviously not the case, as the distributions are not independent and the product will therefore not be normalised. The correction factor was derived (accidentally) during the calculation of the X marginal (which should of course be normalised, but wasn't). The correctly normalised Gaussian-Wishart density is given below (to be confirmed)

$$\begin{aligned}\mathcal{NW}(\mu, K) &= \left(\frac{\Gamma_d(\frac{v}{2})}{\Gamma_d(\frac{v-d+1}{2})}\right) \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} e^{-tr(V^{-1}K)/2} \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])} \\ &= \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d(\frac{v-d+1}{2})} e^{-tr(V^{-1}K)/2} \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])}\end{aligned}$$

where K is the precision matrix of the gaussian distribution over X. Note that, in contrast to previous use of the gaussian expression, the K matrix and μ vector in the above expression are not constants but a random vector and matrix respectively. The above distribution is therefore no longer gaussian, but rather a conditional distribution over gaussian distributions, we're any observation of the mean and precision will collapse the distribution into a gaussian distribution. Luckily, it seems that we will only every have to marginalise out X or K and μ during inference.

Random Note: Uninformative GW Prior:
K. Murphy(pp 132): *One can show (Minka 2000f) that the (improper) uninformative prior has the form... In practice, it is often better to use a weakly informative data-dependent prior. A common choice (see e.g., (Chipman et al. 2001, p81), (Fraley and Raftery 2007, p6)) is to use* $S_0 = diag(S_{\bar{x}})/N$, *and* $v_0 = D + 2$, *to ensure* $E[\Sigma] = S_0$, *and to set* $\mu 0 = x$ *and* κ_0 *to some smallnumber, such as 0.01.*

$$P(X, \mu, K) = P(X|\mu, K)P(\mu|K)P(K)$$
$$P(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(X|\mu, K)P(\mu|K)P(K)d\mu dK$$

$$\begin{aligned} P(\mu, K) &= \int_{-\infty}^{\infty} P(X|\mu, K)P(\mu|K)P(K)dX \\ &= \left(\int_{-\infty}^{\infty} P(X|\mu, K)dX\right) P(\mu|K)P(K) \\ &= P(\mu|K)P(K) \end{aligned}$$

$$\begin{aligned} P(X) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(X|\mu, K)P(\mu|K)P(K)d\mu dK \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} \frac{|K|^{1/2}}{(2\pi)^{d/2}} e^{F(X,\mu,K)} d\mu dK \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\lambda_0^{d/2}|K|^{(v-d+1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^d} e^{-0.5F(X,\mu,K)} d\mu dK \end{aligned}$$

$$
\begin{aligned}
F(X,\mu,K) =& tr(V^{-1}K) \\
&+ ([\mu-\mu_0]^T\lambda_0 K[\mu-\mu_0]) \\
&+ ([X-\mu]^T K[X-\mu]) \\
=& tr(V^{-1}K) \\
&+ \mu^T\lambda_0 K\mu + \mu_0^T\lambda_0 K\mu_0 - 2\mu^T\lambda_0 K\mu_0 \\
&+ X^T KX + \mu^T K\mu - 2X^T K\mu \\
=& tr(V^{-1}K) \\
&+ \mu^T\lambda_0 K\mu + \mu^T K\mu - 2\mu^T KX - 2\mu^T\lambda_0 K\mu_0 \\
&+ X^T KX + \mu_0^T\lambda_0 K\mu_0 \\
=& tr(V^{-1}K) \\
&+ \mu^T(\lambda_0+1)K\mu - 2\mu^T K(X+\lambda_0\mu_0) \\
&+ X^T KX + \mu_0^T\lambda_0 K\mu_0 \\
=& tr(V^{-1}K) \\
&(\mu-\omega)^T(\lambda_0+1)K(\mu-\omega) - \omega^T(\lambda_0+1)K\omega \\
&+ X^T KX + \mu_0^T\lambda_0 K\mu_0 + \\
=& tr(V^{-1}K) \\
&(\mu-\omega)^T(\lambda_0+1)K(\mu-\omega) \\
&+ X^T KX + \mu_0^T\lambda_0 K\mu_0 - \omega^T(\lambda_0+1)K\omega \\
=& (\mu-\omega)^T(\lambda_0+1)K(\mu-\omega) \\
&+ tr(V^{-1}K) + tr(X^T KX) + tr(\mu_0^T\lambda_0 K\mu_0) - tr(\omega^T(\lambda_0+1)K\omega) \\
=& (\mu-\omega)^T(\lambda_0+1)K(\mu-\omega) \\
&+ tr(V^{-1}K + XX^T K + \mu_0\mu_0^T\lambda_0 K - \omega\omega^T(\lambda_0+1)K) \\
=& (\mu-\omega)^T(\lambda_0+1)K(\mu-\omega) + H(X,K)
\end{aligned}
$$

with

$$
\omega = \frac{X+\lambda_0\mu_0}{1+\lambda_0}
$$

$$
\begin{aligned}
H(X,K) =& tr(V^{-1}K + XX^T K + \mu_0\mu_0^T\lambda_0 K - \omega\omega^T(\lambda_0+1)K) \\
=& tr((V^{-1} + XX^T + \lambda_0\mu_0\mu_0^T - (\lambda_0+1)\omega\omega^T)K)
\end{aligned}
$$

So, we can rewrite the integral as:

$$
\begin{aligned}
P(X) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\lambda_0^{d/2}|K|^{(v-d+1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^d} e^{-0.5F(X,\mu,K)} d\mu dK \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\lambda_0^{d/2}|K|^{(v-d+1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^d} e^{-0.5((\mu-\omega)^T(\lambda_0+1)K(\mu-\omega)+H(X,K))} d\mu dK \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\lambda_0^{d/2}|K|^{(v-d+1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^d} e^{-0.5(\mu-\omega)^T(\lambda_0+1)K(\mu-\omega)} e^{-0.5H(X,K))} d\mu dK \\
&= \int_{-\infty}^{\infty} \frac{\lambda_0^{d/2}|K|^{(v-d+1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^d} \left(\int_{-\infty}^{\infty} e^{-0.5(\mu-\omega)^T(\lambda_0+1)K(\mu-\omega)} d\mu\right) e^{-0.5H(X,K))} dK \\
&= \int_{-\infty}^{\infty} \frac{\lambda_0^{d/2}|K|^{(v-d+1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^d} \left(\frac{(2\pi)^{d/2}}{(\lambda_0+1)^{d/2}|K|^{1/2}}\right) e^{-0.5H(X,K))} dK \\
&= \int_{-\infty}^{\infty} \frac{\lambda_0^{d/2}|K|^{(v-d)/2}}{(\lambda_0+1)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^{d/2}} e^{-0.5H(X,K))} dK \\
&= \int_{-\infty}^{\infty} \frac{\lambda_0^{d/2}|K|^{(v-d)/2}}{(\lambda_0+1)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^{d/2}} e^{-0.5tr((V^{-1}+XX^T+\mu_0\mu_0^T\lambda_0-\omega\omega^T(\lambda_0+1))K)} dK \\
&= \int_{-\infty}^{\infty} \frac{\lambda_0^{d/2}|K|^{(v-d)/2}}{(\lambda_0+1)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^{d/2}} e^{-0.5tr(U^{-1}K)} dK \\
&= \frac{\lambda_0^{d/2}}{(\lambda_0+1)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^{d/2}} \int_{-\infty}^{\infty} |K|^{(v-d)/2} e^{-0.5tr(U^{-1}K)} dK \\
&= \frac{\lambda_0^{d/2}}{(\lambda_0+1)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^{d/2}} \int_{-\infty}^{\infty} |K|^{((v+1)-d-1)/2} e^{-0.5tr(U^{-1}K)} dK \\
&= \frac{\lambda_0^{d/2}}{(\lambda_0+1)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^{d/2}} \int_{-\infty}^{\infty} |K|^{(v'-d-1)/2} e^{-0.5tr(U^{-1}K)} dK
\end{aligned}
$$

with

$$U^{-1} = V^{-1} + XX^T + \mu_0\mu_0^T\lambda_0 - \omega\omega^T(\lambda_0 + 1)$$
$$v' = v + 1$$

We can now solve the above integral by recognising it as the integral of a Wishart distribution. The definition of the Wishart distribution is given again below for convenience.

$$P(K) = \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)}e^{-tr(V^{-1}K)/2}$$

$$\begin{aligned}
P(X) =& \frac{\lambda_0^{d/2}}{(\lambda_0+1)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^{d/2}}\int_{-\infty}^{\infty}|K|^{(v'-d-1)/2}e^{-tr(U^{-1}K)/2}dK \\
=& \frac{\lambda_0^{d/2}}{(\lambda_0+1)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^{d/2}}\left(2^{v'd/2}|U|^{v'/2}\Gamma_d(\frac{v'}{2})\right) \\
=& \frac{2^{v'd/2}\lambda_0^{d/2}\Gamma_d(\frac{v'}{2})}{(\lambda_0+1)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^{d/2}}\left(|U|^{v'/2}\right) \\
=& \frac{2^{(v+1)d/2}\lambda_0^{d/2}\Gamma_d(\frac{(v+1)}{2})}{2^{vd/2}(\lambda_0+1)^{d/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^{d/2}}\left(|(V^{-1}+XX^T+\mu_0\mu_0^T\lambda_0-\omega\omega^T(\lambda_0+1))^{-1}|^{(v+1)/2}\right) \\
=& \frac{2^{d/2}\lambda_0^{d/2}\Gamma_d(\frac{(v+1)}{2})}{(\lambda_0+1)^{d/2}\Gamma_d\left(\frac{v}{2}\right)(2\pi)^{d/2}}|V|^{-v/2}\left(|V^{-1}+XX^T+\mu_0\mu_0^T\lambda_0-\omega\omega^T(\lambda_0+1)|\right)^{-(v+1)/2} \\
=& \frac{\lambda_0^{d/2}\Gamma_d(\frac{(v+1)}{2})}{(\lambda_0+1)^{d/2}\Gamma_d\left(\frac{v}{2}\right)\pi^{d/2}}|V|^{-v/2}\left(|V^{-1}+XX^T+\mu_0\mu_0^T\lambda_0-\omega\omega^T(\lambda_0+1)|\right)^{-(v+1)/2}
\end{aligned}$$

But, remember:

$$\omega = \frac{X + \lambda_0 \mu_0}{1 + \lambda_0}$$

so

$$\begin{aligned}
P(X) =& \frac{1}{Z}|V|^{-v/2}\left(|V^{-1} + XX^T + \mu_0\mu_0^T\lambda_0 - \omega\omega^T(\lambda_0+1)|\right)^{-(v+1)/2} \\
=& \frac{1}{Z}|V|^{-v/2}\left(|V^{-1} + XX^T + \mu_0\mu_0^T\lambda_0 - \frac{1}{(1+\lambda_0)^2}(X+\lambda_0\mu_0)(X+\lambda_0\mu_0)^T(\lambda_0+1)|\right)^{-(v+1)/2} \\
=& \frac{1}{Z}|V|^{-v/2}\left(|V^{-1} + XX^T + \mu_0\mu_0^T\lambda_0 - \frac{1}{1+\lambda_0}(X+\lambda_0\mu_0)(X+\lambda_0\mu_0)^T|\right)^{-(v+1)/2} \\
=& \frac{1}{Z}|V|^{-v/2}\left(|V^{-1} + XX^T + \mu_0\mu_0^T\lambda_0 - \frac{1}{1+\lambda_0}(XX^T + 2\lambda_0\mu_0X^T + \lambda_0^2\mu_0\mu_0^T)|\right)^{-(v+1)/2} \\
=& \frac{1}{Z}|V|^{-v/2}\left(|V^{-1} + (\frac{\lambda_0}{1+\lambda_0})XX^T + \mu_0\mu_0^T\lambda_0 - \frac{1}{1+\lambda_0}(2\lambda_0\mu_0X^T + \lambda_0^2\mu_0\mu_0^T)|\right)^{-(v+1)/2} \\
=& \frac{1}{Z}|V|^{-v/2}\left(|V^{-1} + \left(\frac{\lambda_0}{1+\lambda_0}\right)XX^T + \left(\frac{\lambda_0^2+\lambda_0}{1+\lambda_0}\right)\mu_0\mu_0^T - \left(\frac{2\lambda_0}{1+\lambda_0}\right)\mu_0X^T - \left(\frac{\lambda_0^2}{1+\lambda_0}\right)\mu_0\mu_0^T)|\right)^{-(v+1)/2} \\
=& \frac{1}{Z}|V|^{-v/2}\left(|V^{-1} + \left(\frac{\lambda_0}{1+\lambda_0}\right)XX^T + \left(\frac{\lambda_0}{1+\lambda_0}\right)\mu_0\mu_0^T - \left(\frac{2\lambda_0}{1+\lambda_0}\right)\mu_0X^T|\right)^{-(v+1)/2} \\
=& \frac{1}{Z}|V|^{-v/2}\left(|V^{-1} + \left(\frac{\lambda_0}{1+\lambda_0}\right)\left(XX^T + \mu_0\mu_0^T - 2\mu_0X^T\right)|\right)^{-(v+1)/2} \\
=& \frac{1}{Z}|V|^{-v/2}\left(|V^{-1} + \left(\frac{\lambda_0}{1+\lambda_0}\right)(X-\mu_0)(X-\mu_0)^T|\right)^{-(v+1)/2}
\end{aligned}$$

with

$$\frac{1}{Z} = \frac{\lambda_0^{d/2}\Gamma_d(\frac{(v+1)}{2})}{(\lambda_0+1)^{d/2}\Gamma_d\left(\frac{v}{2}\right)\pi^{d/2}}$$

Now, using the matrix determinant lemma:

$$|A + uv^T| = (1 + v^tA^{-1}u)|A|$$

we can rewrite the above as follows

$$
\begin{aligned}
P(X) =& \frac{1}{Z}|V|^{-v/2}\left(|V^{-1} + \left(\frac{\lambda_0}{1+\lambda_0}\right)(X-\mu_0)(X-\mu_0)^T|\right)^{-(v+1)/2} \\
=& \frac{1}{Z}|V|^{-v/2}\left(\left(1 + \left(\frac{\lambda_0}{1+\lambda_0}\right)(X-\mu_0)^T V(X-\mu_0)\right)|V^{-1}|\right)^{-(v+1)/2} \\
=& \frac{1}{Z}|V|^{-v/2}|V|^{(v+1)/2}\left(1 + (X-\mu_0)^T\left(\frac{\lambda_0 V}{1+\lambda_0}\right)(X-\mu_0)\right)^{-(v+1)/2} \\
=& \frac{1}{Z}|V|^{1/2}\left(1 + (X-\mu_0)^T\left(\frac{\lambda_0 V}{1+\lambda_0}\right)(X-\mu_0)\right)^{-(v+1)/2} \\
=& \frac{\lambda_0^{d/2}\Gamma_d(\frac{(v+1)}{2})}{(\lambda_0+1)^{d/2}\Gamma_d\left(\frac{v}{2}\right)\pi^{d/2}}|V|^{1/2}\left(1 + (X-\mu_0)^T\left(\frac{\lambda_0 V}{1+\lambda_0}\right)(X-\mu_0)\right)^{-(v+1)/2} \\
=& \frac{\Gamma_d(\frac{(v+1)}{2})}{\Gamma_d\left(\frac{v}{2}\right)\pi^{d/2}}\left|\left(\frac{\lambda_0}{1+\lambda_0}\right)V\right|^{1/2}\left(1 + (X-\mu_0)^T\left(\frac{\lambda_0 V}{1+\lambda_0}\right)(X-\mu_0)\right)^{-(v+1)/2} \\
=& \frac{\Gamma_d(\frac{((v-d+1)+d)}{2})\left|\frac{(v-d+1)\lambda_0 V}{1+\lambda_0}\right|^{1/2}}{\Gamma_d\left(\frac{v}{2}\right)(v-d+1)^{d/2}\pi^{d/2}}\left(1 + \frac{1}{(v-d+1)}(X-\mu_0)^T\left(\frac{(v-d+1)\lambda_0 V}{1+\lambda_0}\right)(X-\mu_0)\right)^{-(v-d+1)+d)/2}
\end{aligned}
$$

We can now recognise the above expression as a multivariate t-distribution and, by looking at the definition of the distribution (see below), we can determine the parameters of the above multivariate t-distribution in terms of the parameters of our mean and precision priors.

$$
t_{v_X}(X;\mu_X,\Sigma_X = K_X^{-1}) = \frac{\Gamma(\frac{v_X+d}{2})|K_X|^{1/2}}{\Gamma(\frac{v_X}{2})v_X^{d/2}\pi^{d/2}}\left(1 + \frac{1}{v_X}(X-\mu_X)^T K_X (X-\mu_X)\right)^{-(v_X+d)/2}
$$

The parameters are therefore as follows:

$$
\begin{aligned}
v_X =& v - d + 1 \\
\mu_X =& \mu_0 \\
K_X =& \left(\frac{(v-d+1)\lambda_0}{1+\lambda_0}\right)V
\end{aligned}
$$

But the distribution is not properly normalised: $\Gamma_d(\frac{v}{2})$ in the denominator should be $\Gamma_d(\frac{v-d+1}{2})$
Suspicion: It is not normalised because the Gaussian-Wishart that we started with was unormalised. If this is true then the Gaussian-Wishart definition above should be multiplied by

$$
\left(\frac{\Gamma_d(\frac{v-d+1}{2})}{\Gamma_d(\frac{v}{2})}\right)^{-1}
$$

(Extra) The variance of the multivariate student-t above is as follows

$$\begin{aligned}\Sigma =& \frac{v_X}{v_X - 2} K_X^{-1} \\ =& \frac{v - d + 1}{v - d - 1} \left(\frac{1 + \lambda_0}{(v - d + 1)\lambda_0} \right) V^{-1} \\ =& \left(\frac{1 + \lambda_0}{(v - d - 1)\lambda_0} \right) V^{-1}\end{aligned}$$

2 The Marginal over μ

$$P(K) = \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)}e^{-tr(V^{-1}K)/2}$$

$$P(\mu|K) = \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}}e^{-0.5([\mu-\mu_0]^T\lambda_0 K[\mu-\mu_0])}$$

$$P(X|\mu, K) = \frac{|K|^{1/2}}{(2\pi)^{d/2}}e^{-0.5([X-\mu]^T K[X-\mu])}$$

$$
\begin{aligned}
P(\mu) =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(X|\mu,K)P(\mu|K)P(K)dXdK \\
=& \int_{-\infty}^{\infty} P(\mu|K)P(K)\int_{-\infty}^{\infty} P(X|\mu,K)dXdK \\
=& \int_{-\infty}^{\infty} P(\mu|K)P(K)dK \\
=& \int_{-\infty}^{\infty} \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T\lambda_0 K[\mu-\mu_0])} \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} e^{-tr(V^{-1}K)/2}dK \\
=& \frac{\lambda_0^{1/2}}{\pi^{d/2}} \int_{-\infty}^{\infty} \frac{|K|^{1/2}}{2^{d/2}} e^{-0.5(\lambda_0[\mu-\mu_0][\mu-\mu_0]^T K)} \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} e^{-tr(V^{-1}K)/2}dK \\
=& \frac{\lambda_0^{1/2}}{\pi^{d/2}} \int_{-\infty}^{\infty} \frac{|K|^{1/2}}{2^{d/2}} e^{-0.5tr(\lambda_0[\mu-\mu_0][\mu-\mu_0]^T K)} \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} e^{-tr(V^{-1}K)/2}dK \\
=& \frac{\lambda_0^{1/2}}{\pi^{d/2}} \int_{-\infty}^{\infty} \frac{|K|^{((v+1)-d-1)/2}}{2^{(v+1)d/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} e^{-tr([V^{-1}+V_\mu^{-1}]K)/2}dK \\
=& \frac{\lambda_0^{1/2}}{\pi^{d/2}2^{(v+1)d/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} \int_{-\infty}^{\infty} |K|^{((v+1)-d-1)/2} e^{-tr([V^{-1}+V_\mu^{-1}]K)/2}dK \\
=& \frac{\lambda_0^{1/2}}{\pi^{d/2}2^{(v+1)d/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} \left(2^{v_1 d/2}|V_1|^{v_1/2}\Gamma_d\left(\frac{v+1}{2}\right)\right) \\
=& \frac{\lambda_0^{1/2}\Gamma_d\left(\frac{v+1}{2}\right)}{\pi^{d/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} \left(|V_1|\right)^{(v+1)/2} \\
=& \frac{\lambda_0^{1/2}\Gamma_d\left(\frac{v+1}{2}\right)}{\pi^{d/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} \left(|(V^{-1}+V_u^{-1})^{-1}|\right)^{(v+1)/2} \\
=& \frac{\lambda_0^{1/2}\Gamma_d\left(\frac{v+1}{2}\right)}{\pi^{d/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} \left(|V^{-1}+\lambda_0[\mu-\mu_0][\mu-\mu_0]^T|\right)^{-(v+1)/2} \\
=& \frac{\lambda_0^{1/2}\Gamma_d\left(\frac{(v+1-d)+d}{2}\right)}{\pi^{d/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} \left(1+\lambda_0[\mu-\mu_0]^T V[\mu-\mu_0]|\right)^{-((v+1-d)+d)/2}
\end{aligned}
$$

With

$$
\begin{aligned}
V_u^{-1} &= \lambda_0[\mu-\mu_0][\mu-\mu_0]^T \\
V_1^{-1} &= V^{-1}+V_u^{-1} \\
V_1 &= (V^{-1}+V_u^{-1})^{-1} \\
v_1 &= v+1-d
\end{aligned}
$$

$$|A + uv^T| = (1 + v^t A^{-1} u)|A|$$

Note that the $\Gamma_d\left(\frac{v}{2}\right)$ term in the denominator should be $\Gamma_d\left(\frac{v+1-d}{2}\right)$. As discussed in the X marginal section, this is probably due to the unnormalised density used in this derivation. The correction term in the definition is therefore

$$\frac{\Gamma_d\left(\frac{v}{2}\right)}{\Gamma_d\left(\frac{v+1-d}{2}\right)}$$

This is exactly the same correction factor that we found in the X marginal calculations.

3 Updating The Gaussian-Wishart Parameters With a Gaussian Message

$$P(K) = \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} e^{-tr(V^{-1}K)/2}$$

$$P(\mu|K) = \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])}$$

$$P(X|\mu, K) = \frac{|K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([X-\mu]^T K[X-\mu])}$$

$$P(X, \mu, K) = P(X|\mu, K)P(\mu|K)P(K)$$

Lets first look at a simplified version where we instead receive a single gaussian update message at the parameter cluster. The update message that we receive is

$$P_m(X) = \frac{|K_m|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([X-\mu_m]^T K_m[X-\mu_m])}$$

So we can calculate the parameter message as follows

$$\begin{aligned} P_m(\mu, K) &= \int_{-\infty}^{\infty} P(X|\mu, K)P_m(X)P(\mu|K)P(K)dX \\ &= \left(\int_{-\infty}^{\infty} P(X|\mu, K)P_m(X)dX\right) P(\mu|K)P(K) \end{aligned}$$

$$
\begin{aligned}
P(X|\mu,K)P_m(X) &= \frac{|K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([X-\mu]^T K[X-\mu])} \frac{|K_m|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([X-\mu_m]^T K_m[X-\mu_m])} \\
&= \frac{|K|^{1/2}}{(2\pi)^{d/2}} \frac{|K_m|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([X-\mu]^T K[X-\mu])} e^{-0.5([X-\mu_m]^T K_m[X-\mu_m])} \\
&= \frac{|K|^{1/2}}{(2\pi)^{d/2}} \frac{|K_m|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([X-\mu_p]^T K_p[X-\mu_p])} e^{-0.5[-\mu_p^T K_p \mu_p + \mu^T K \mu + \mu_m^T K_m \mu_m]}
\end{aligned}
$$

with

$$
\begin{aligned}
&[X-\mu]^T K[X-\mu] + [X-\mu_m]^T K_m[X-\mu_m] \\
&= X^T K X - 2X^T K\mu + \mu^T K\mu + X^T K_m X - 2X^T K_m \mu_m + \mu_m^T K_m \mu_m \\
&= X^T (K+K_m) X - 2X^T K\mu + \mu^T K\mu - 2X^T K_m \mu_m + \mu_m^T K_m \mu_m \\
&= X^T (K+K_m) X - 2X^T h + \mu^T h - 2X^T h_m + \mu_m^T h_m \\
&= X^T (K+K_m) X - 2X^T (h+h_m) + \mu^T h + \mu_m^T h_m \\
&= X^T K_p X - 2X^T K_p \mu_p + \mu^T h + \mu_m^T h_m \\
&= [X-\mu_p]^T K_p [X-\mu_p] - \mu_p^T K_p \mu_p + \mu^T h + \mu_m^T h_m
\end{aligned}
$$

(tested tst546.m)

$$
\begin{aligned}
K_p &= K_m + K \\
h_p &= K_m \mu_m + K\mu \\
g_p &= \mu^T h + \mu_m^T h_m
\end{aligned}
$$

$$
\begin{aligned}
&\int_{-\infty}^{\infty} P(X|\mu,K)P_m(X)dX \\
&\qquad = \frac{|K|^{1/2}}{(2\pi)^{d/2}} \frac{|K_m|^{1/2}}{(2\pi)^{d/2}} \left(\int_{-\infty}^{\infty} e^{-0.5([X-\mu_p]^T K_p[X-\mu_p])} dX \right) e^{-0.5(-\mu_p^T K_p \mu_p + \mu^T K \mu + \mu_m^T K_m \mu_m)} \\
&\qquad = \left(\frac{|K_m|^{1/2}}{(2\pi)^{d/2}} \right) \frac{|K|^{1/2}}{(2\pi)^{d/2}} \frac{(2\pi)^{d/2}}{|K_p|^{1/2}} e^{-0.5(\mu-\mu_m)^T (K_m - K_m K_p^{-1} K_m)(\mu-\mu_m)}
\end{aligned}
$$

Where the above uses (see Random Derivations 1.1 for the proof)

$$\mu^T K\mu - \mu_p^T K_p \mu_p + \mu_m^T K_m \mu_m = (\mu - \mu_m)^T (K_m - K_m K_p^{-1} K_m)(\mu - \mu_m)$$

Now,using the Woodbury formula

$$(E - BD^{-1}C) = \left[E^{-1}B(D - CE^{-1}B)^{-1}CE^{-1} + E^{-1}\right]^{-1}$$

$$\begin{aligned} E =& B = C = K_m \\ D =& K_p \end{aligned}$$

$$\begin{aligned} (K_m - K_m K_p^{-1} K_m) &= \left[K_m^{-1} K_m (K_p - K_m K_m^{-1} K_m)^{-1} K_m K_m^{-1} + K_m^{-1}\right]^{-1} \\ &= \left[(K_p - K_m)^{-1} + K_m^{-1}\right]^{-1} \\ &= \left[K^{-1} + K_m^{-1}\right]^{-1} \end{aligned}$$

$$\begin{aligned} \int_{-\infty}^{\infty} P(X|\mu, K) P_m(X) dX &= \left(\frac{|K_m|^{1/2}}{(2\pi)^{d/2}}\right) \frac{|K|^{1/2}}{(2\pi)^{d/2}} \frac{(2\pi)^{d/2}}{|K_p|^{1/2}} e^{-0.5(\mu-\mu_m)^T (K^{-1}+K_m^{-1})^{-1}(\mu-\mu_m)} \\ &= \left(\frac{|K_m|^{1/2}}{(2\pi)^{d/2}}\right) \frac{|K|^{1/2}}{(2\pi)^{d/2}} \frac{(2\pi)^{d/2}}{|K_p|^{1/2}} e^{-0.5(\mu-\mu_m)^T K_D (\mu-\mu_m)} \\ &= \left(\frac{1}{(2\pi)^{d/2}}\right) \frac{|K_m|^{1/2}|K|^{1/2}}{|K_p|^{1/2}} e^{-0.5(\mu-\mu_m)^T K_D (\mu-\mu_m)} \\ &= \frac{|K_D|^{1/2}}{(2\pi)^{d/2}} e^{-0.5(\mu-\mu_m)^T K_D (\mu-\mu_m)} \end{aligned}$$

With

$$\begin{aligned} K_D &= (K^{-1} + K_m^{-1})^{-1} \\ K_p &= K + K_m \end{aligned}$$

4 a Gaussian-Wishart Approximation to the true (soft evidence updated) Posterior

ASSUMPTION (Check this - See Assumptions Checks Section at end of document): the K 'coordinate' of the mode of a gaussian wishart is the same as for the wishart marginal.

(IM?)POSSIBLE IMPROVEMENT: Can we really not solve $\int \mathcal{F} dK$ or $\int K\mathcal{F} dK$? Or find $d\mathcal{F}/dK = 0$? Or get the fixed K Hessian analytically? If we could, it would make the above much easier .

Testing:
Model selection comparison? - same data?
noisy covariance larger than sample covariance?

Maybe we can't approximate the posterior analytically, but we can do two things that might help us get a reasonable approximation. We can quickly find a posterior that should be close to the true posterior with the exact data update using the mean as the data. If the covariance of the message is relatively small this should be closer to the data updated posterior (DUP) should be close to the true posterior, but if the covariance is large the true posterior will probably be closer to the prior. Perhaps we can compare the mode of the prior to the precision of the message to decide if we should use the DUP or the prior parameters to initialise an optimisation search for the true posterior parameters. Secondly, we can evaluate the exact posterior at any given point. Could we not somehow use a series of such evaluations to find the Gaussian-Wishart that bests approximates the true posterior? But we will probably need to find the normalising constant of the true posterior function first?

(here again for convenience)

$$P(K) = \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)} e^{-tr(V^{-1}K)/2}$$

$$P(\mu|K) = \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])}$$

$$P(X|\mu, K) = \frac{|K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([X-\mu]^T K[X-\mu])}$$

Note that $P(\mu|K)P(K)$ does not result in a normalised Gaussian-Wishart. The distribution below is in the properly normalised form.

$$\begin{aligned}\mathcal{NW}(\mu, K) &= \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d(\frac{v-d+1}{2})} e^{-tr(V^{-1}K)/2} \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T\lambda_0 K[\mu-\mu_0])} \\ &= \frac{\lambda_0^{d/2}|K|^{(v-d)/2}}{(2\pi)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d(\frac{v-d+1}{2})} e^{-0.5([\mu-\mu_0]^T\lambda_0 K[\mu-\mu_0])} e^{-0.5(tr(V^{-1}K))}\end{aligned}$$

We can also write the above distribution in a full canonical form and add an additional weight (or log-weight) parameter to allow the distribution to be unnormalised.

$$\begin{aligned}\mathcal{NW}(\mu, K) &= \frac{w\lambda_0^{d/2}|K|^{(v-d)/2}}{(2\pi)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d(\frac{v-d+1}{2})} e^{-0.5([\mu-\mu_0]^T\lambda_0 K[\mu-\mu_0])} e^{-0.5(tr(V^{-1}K))} \\ &= |K|^{(v-d)/2} e^{log(w)} e^{\log(\lambda_0^{d/2}) - \log((2\pi)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d(\frac{v-d+1}{2}))} e^{-0.5([\mu-\mu_0]^T\lambda_0 K[\mu-\mu_0])} e^{-0.5(tr(V^{-1}K))} \\ &= |K|^{(v-d)/2} e^{\log(w)+g} e^{-0.5([\mu-\mu_0]^T\lambda_0 K[\mu-\mu_0])} e^{-0.5(tr(V^{-1}K))} \\ &= \frac{w\lambda_0^{d/2}|K|^{(v-d)/2}}{(2\pi)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d(\frac{v-d+1}{2})} e^{-0.5(\mu^T\lambda_0 K\mu - 2\mu^T\lambda_0 K\mu_0 + \mu_0^T\lambda_0 K\mu_0)} e^{-0.5(tr(V^{-1}K))} \\ &= \text{...will we ever do NW multiplication - we wont in perfact data case,} \\ &\quad \text{nor in soft data approximate update - so is this needed?}\end{aligned}$$

Where

$$g = \log(\lambda_0^{d/2}) - \log\left((2\pi)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d(\frac{v-d+1}{2})\right)$$

See `GaussianWishart_K_mode_suspicion_test.py`

The mode of a Wishart distribution is

$$argmax_K\{\mathcal{W}(K|v, V)\} = (v-d-1)V$$

But note that the value of K for the Gaussian-Wishart distribution is not the same, for a Gaussian-Wishart distribution $K_{argmax} = (v-d)V$. The difference between these two does of course become smaller as v gets larger (this can make debugging more difficult). Perhaps it is important to list the 3 different cases here for comparison

- Wishart Mode: $argmax_K\{\mathcal{W}(K|v, V)\} = (v-d-1)V$
- Gaussian-Wishart (Joint) Mode (K): $argmax_K\{\mathcal{NW}(\mu, K|v, V)\} = (v-d)V$
  (+1 because of $|K|^{1/2}$ term from Gaussian)
- Gaussian-Wishart (K - marginal) Mode (K): $argmax_K\{\int_{\mathbb{R}} \mathcal{NW}(\mu, K|v, V)d\mu\} = (v-d-1)V$
  (Gaussian removed completely, so -1 returns)

The potential at the mode (if it is normalised) is

$$\begin{aligned}\mathcal{W}_{max}(K|v,V) &= \frac{|(v-d-1)V|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)}e^{-tr(V^{-1}(v-d-1)V)/2} \\ &= \frac{(v-d-1)^{(v-d-1)d/2}|V|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)}e^{-tr((v-d-1)I)/2} \\ &= \frac{(v-d-1)^{(v-d-1)d/2}|V|^{(d-1)/2}}{2^{vd/2}\Gamma_d\left(\frac{v}{2}\right)}e^{-((v-d-1)d)/2}\end{aligned}$$

The mode of the gaussian is

$$argmax\{N(\mu|K,\lambda_0,\mu_0)\} = \mu_0$$

The potential at the mode (if it is normalised) is

$$\begin{aligned}\mathcal{N}_{max}(\mu|K,\lambda_0,\mu_0) &= \frac{|\lambda_0 K_{argmax}|^{1/2}}{(2\pi)^{d/2}} \\ &= \frac{|\lambda_0(v-d-1)V|^{1/2}}{(2\pi)^{d/2}}\end{aligned}$$

The mean of a Wishart distribution is at $K = vV$

and the potential at the mean is

$$\begin{aligned}W(K=vV) &= \frac{|vV|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)}e^{-tr(V^{-1}vV)/2} \\ &= \frac{v^{(v-d-1)d/2}|V|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d\left(\frac{v}{2}\right)}e^{-vd/2} \\ &= \frac{v^{(v-d-1)d/2}|V|^{(-d-1)/2}}{2^{vd/2}\Gamma_d\left(\frac{v}{2}\right)}e^{-vd/2}\end{aligned}$$

The mode of a Gaussian-Wishart distribution is

$$argmax\{\mathcal{NW}_(\mu,K|\lambda_0,\mu_0,v,V)\} = \{\mu=\mu_q=\mu_0, \quad K=K_q=(v-d)V\}$$

And the potential at the mode (if it is normalised) is

$$
\begin{aligned}
\mathcal{F}(\mu, K) = \mathcal{NW}(\mu, K) &= \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d(\frac{v-d+1}{2})} e^{-tr(V^{-1}K)/2} \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])} \\
&= \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d(\frac{v-d+1}{2})} e^{-tr(V^{-1}K)/2} \frac{\lambda_0^{d/2}|K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu_0-\mu_0]^T \lambda_0 K[\mu_0-\mu_0])} \\
&= \frac{\lambda_0^{d/2}|K|^{(v-d)/2}}{(2\pi)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d(\frac{v-d+1}{2})} e^{-tr(V^{-1}K)/2} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])} \\
\mathcal{F}_{max} &= \mathcal{NW}(\mu = \mu_0, K = (v-d)V) \\
&= \frac{\lambda_0^{d/2}|V(v-d)|^{(v-d)/2}}{(2\pi)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d(\frac{v-d+1}{2})} e^{-tr(V^{-1}K)/2} \\
&= \frac{\lambda_0^{d/2}(v-d)^{(v-d)d/2}|V|^{(v-d)/2}}{(2\pi)^{d/2}2^{vd/2}|V|^{v/2}\Gamma_d(\frac{v-d+1}{2})} e^{-tr(V^{-1}K)/2} \\
&= \frac{\lambda_0^{d/2}(v-d)^{(v-d)d/2}|V|^{-d/2}}{(2\pi)^{d/2}2^{vd/2}\Gamma_d(\frac{v-d+1}{2})} e^{-tr(V^{-1}K)/2} \\
&= \frac{\lambda_0^{d/2}(v-d)^{(v-d)d/2}|V|^{-d/2}}{(2\pi)^{d/2}2^{vd/2}\Gamma_d(\frac{v-d+1}{2})} e^{-tr(V^{-1}(v-d)V)/2} \\
&= \frac{\lambda_0^{d/2}(v-d)^{(v-d)d/2}}{(2\pi)^{d/2}2^{vd/2}|V|^{d/2}\Gamma_d(\frac{v-d+1}{2})e^{(v-d)d/2}}
\end{aligned}
$$

The above expression is however only for normalised distributions. The posterior won't be (after the soft evidence 'message' is 'absorbed'). We can account for this by adding a weight term w. Here we will also start viewing $\mathcal{F}$ as the updated posterior and therefore start using the 1 subscripts for the parameters to indicate that they are the parameters of the updated posterior.

$$
\begin{aligned}
\mathcal{F}_{max} &= \frac{w\lambda_1^{d/2}(v_1-d)^{(v_1-d)d/2}}{(2\pi)^{d/2}2^{v_1 d/2}|V_1|^{d/2}\Gamma_d(\frac{v_1-d+1}{2})e^{(v_1-d)d/2}} \\
&= \frac{w_1\lambda_0^{d/2}(v_1-d)^{(v_1-d)d/2}}{(2\pi)^{d/2}2^{v_1 d/2}|K_q(v_1-d)^{-1}|^{d/2}\Gamma_d(\frac{v_1-d+1}{2})e^{(v_1-d)d/2}} \\
&= \frac{w_1\lambda_1^{d/2}(v_1-d)^{(v_1-d)d/2}}{(2\pi)^{d/2}2^{v_1 d/2}(v_1-d)^{-d^2/2}|K_q|^{d/2}\Gamma_d(\frac{v_1-d+1}{2})e^{(v_1-d)d/2}} \\
&= \frac{w_1\lambda_1^{d/2}(v_1-d)^{(v_1 d/2-d^2/2)}}{(2\pi)^{d/2}2^{v_1 d/2}(v_1-d)^{-d^2/2}|K_q|^{d/2}\Gamma_d(\frac{v_1-d+1}{2})e^{(v_1-d)d/2}} \\
&= \frac{w_1\lambda_1^{d/2}(v_1-d)^{v_1 d/2}}{(2\pi)^{d/2}2^{v_1 d/2}|K_q|^{d/2}\Gamma_d(\frac{v_1-d+1}{2})e^{(v_1-d)d/2}} \\
w_1 &= \mathcal{F}_{max}\left(\frac{(2\pi)^{d/2}2^{v_1 d/2}|K_q|^{d/2}\Gamma_d(\frac{v_1-d+1}{2})e^{(v_1-d)d/2}}{\lambda_1^{d/2}(v_1-d)^{v_1 d/2}}\right)
\end{aligned}
$$

(old version lacking 1 subscripts commented

Where we have set $V_1 = K_q(v_1-d)^{-1}$

We can now write a more general expression for an unnormalised the Gaussian-Wishart with the weight term defined above. It is important to note that we have derived the expression for the weight specifically in terms of the function maximum and corresponding matrix value of K (K_q). We can find both these values through optimisation methods and this will be crucial when we approximate the true posterior as a Gaussian-Wishart.

$$\begin{aligned}
\mathcal{F}(\mu,K) = \mathcal{NW}(\mu,K) &= w\frac{|K|^{(v_1-d-1)/2}}{2^{v_1 d/2}|V_1|^{v_1/2}\Gamma_d(\frac{v_1-d+1}{2})}e^{-tr(V_1^{-1}K)/2}\frac{|\lambda_1 K|^{1/2}}{(2\pi)^{d/2}}e^{-0.5([\mu-\mu_1]^T\lambda_1 K[\mu-\mu_1])} \\
&= w\frac{|K|^{(v_1-d-1)/2}}{2^{v_1 d/2}|V_1|^{v_1/2}\Gamma_d(\frac{v_1-d+1}{2})}\frac{|\lambda_1 K|^{1/2}}{(2\pi)^{d/2}}e^{-0.5([\mu-\mu_1]^T\lambda_1 K[\mu-\mu_1])}e^{-tr(V_1^{-1}K)/2} \\
&= w\frac{|K|^{(v_1-d-1)/2}}{2^{v_1 d/2}|V_1|^{v_1/2}\Gamma_d(\frac{v_1-d+1}{2})}\frac{|\lambda_1 K|^{1/2}}{(2\pi)^{d/2}}G(\mu,K) \\
&= \mathcal{F}_{max}\left(\frac{(2\pi)^{d/2}2^{v_1 d/2}|K_q|^{d/2}\Gamma_d(\frac{v_1-d+1}{2})e^{(v_1-d)d/2}}{\lambda_1^{d/2}(v_1-d)^{v_1 d/2}}\right)\frac{|K|^{(v_1-d-1)/2}}{2^{v_1 d/2}|V_1|^{v_1/2}\Gamma_d(\frac{v_1-d+1}{2})}\frac{|\lambda_1 K|^{1/2}}{(2\pi)^{d/2}}G(\mu,K) \\
&= \mathcal{F}_{max}\left(\frac{(2\pi)^{d/2}2^{v_1 d/2}|K_q|^{d/2}\Gamma_d(\frac{v_1-d+1}{2})e^{(v_1-d)d/2}}{\lambda_1^{d/2}(v_1-d)^{v_1 d/2}}\right)\frac{|K|^{(v_1-d-1)/2}}{2^{v_1 d/2}|V_1|^{v_1/2}\Gamma_d(\frac{v_1-d+1}{2})}\frac{\lambda_1^{d/2}|K|^{1/2}}{(2\pi)^{d/2}}G(\mu,K) \\
&= \mathcal{F}_{max}\left(\frac{|K_q|^{d/2}e^{(v_1-d)d/2}}{(v_1-d)^{v_1 d/2}}\right)\frac{|K|^{(v_1-d)/2}}{|K_q(v_1-d)^{-1}|^{v_1/2}}G(\mu,K) \\
&= \mathcal{F}_{max}\left(\frac{|K_q|^{d/2}e^{(v_1-d)d/2}}{(v_1-d)^{v_1 d/2}}\right)\frac{|K|^{(v_1-d)/2}}{(v_1-d)^{-v_1 d/2}|K_q|^{v_1/2}}G(\mu,K) \\
&= \mathcal{F}_{max}\left(\frac{|K_q|^{(d-v_1)/2}e^{(v_1-d)d/2}}{(v_1-d)^{v_1 d/2}}\right)\frac{|K|^{(v_1-d)/2}}{(v_1-d)^{-v_1 d/2}}G(\mu,K) \\
&= \mathcal{F}_{max}|K_q|^{(d-v_1)/2}e^{(v_1-d)d/2}|K|^{(v_1-d)/2}G(\mu,K) \\
&= \mathcal{F}_{max}|K_q|^{(d-v_1)/2}e^{(v_1-d)d/2}|K|^{(v_1-d)/2}e^{-0.5([\mu-\mu_1]^T\lambda_1 K[\mu-\mu_1])}e^{-tr(V_1^{-1}K)/2}
\end{aligned}$$

(old version lacking 1 subscripts commented

Now that we have simplified the above expression, we need to evaluate the function at another point (not argmax) in order to calculate v. We will choose this point to be

$$\{\mu_c, K_c\} = \{\mu_q, \quad (1+\delta)K_q\}$$

$$\begin{aligned}
\mathcal{F}(\mu,K) &= \mathcal{F}_{max}|K_q|^{(d-v_1)/2}e^{(v_1-d)d/2}|K|^{(v_1-d)/2}e^{-0.5([\mu-\mu_1]^T\lambda_0 K[\mu-\mu_1])}e^{-tr(V_1^{-1}K)/2} \\
\mathcal{F}(\mu_c,K_c) &= \mathcal{F}_{max}|K_q|^{(d-v_1)/2}e^{(v_1-d)d/2}|(1+\delta)K_q|^{(v_1-d)/2}e^{-tr(K_q^{-1}(v_1-d)(1+\delta)K_q)/2} \\
&= \mathcal{F}_{max}|K_q|^{(d-v_1)/2}e^{(v_1-d)d/2}|(1+\delta)K_q|^{(v_1-d)/2}e^{-(v_1-d)(1+\delta)d/2} \\
&= \mathcal{F}_{max}e^{(v_1-d)d/2}(1+\delta)^{(v_1-d)d/2}e^{-(v_1-d)(1+\delta)d/2} \\
&= \mathcal{F}_{max}e^{(v_1-d)d/2}(1+\delta)^{(v_1-d)d/2}e^{-(v_1-d)d/2}e^{-(v_1-d)\delta d/2} \\
&= \mathcal{F}_{max}(1+\delta)^{(v_1-d)d/2}e^{-(v_1-d)\delta d/2}
\end{aligned}$$

(old version lacking 1 subscripts commented

Using

$$V_1 = K_q(v_1 - d)^{-1}$$
$$V_1^{-1} = K_q^{-1}(v_1 - d)$$

Using this function, we can derive an expression for v_1

$$\begin{aligned}
\mathcal{F}(\mu_c, K_c) &= \mathcal{F}_{max}(1+\delta)^{(v_1-d)d/2} e^{-(v_1-d)\delta d/2} \\
2log(\mathcal{F}(\mu_c, K_c)) &= 2log(\mathcal{F}_{max}) + (v_1 d - d^2)log(1+\delta) - v_1\delta d + \delta d^2 \\
2log(\mathcal{F}(\mu_c, K_c)) &= 2log(\mathcal{F}_{max}) + v_1(d)log(1+\delta) - d^2 log(1+\delta) - v_1\delta d + \delta d^2 \\
2log(\mathcal{F}(\mu_c, K_c)) &= 2log(\mathcal{F}_{max}) - d^2 log(1+\delta) + \delta d^2 + v_1(log(1+\delta) - \delta)d \\
v_1(log(1+\delta) - \delta)d &= 2log(\mathcal{F}(\mu_c, K_c)) - 2log(\mathcal{F}_{max}) + d^2 log(1+\delta) - \delta d^2 \\
v_1 &= \frac{2log(\mathcal{F}(\mu_c, K_c)) - 2log(\mathcal{F}_{max}) + d^2 log(1+\delta) - \delta d^2}{(log(1+\delta) - \delta)d}
\end{aligned}$$

(old version lacking 1 subscripts commented

Now we can also get the weight, since we have calculated the rest of the parameters. As usual though, we will prefer to rather use the log-weight in order to prevent numerical overflow.

$$\begin{aligned}
log(w_1) &= log(\mathcal{F}_{max}) + \log\left((2\pi)^{d/2} 2^{v_1 d/2} |K_q|^{d/2} \Gamma_d(\frac{v_1 - d + 1}{2}) e^{(v_1-d)d/2}\right) - \log\left(\lambda_1^{d/2}(v_1 - d)^{v_1 d/2}\right) \\
&= log(\mathcal{F}_{max}) + \log\left((2\pi)^{d/2} 2^{v_1 d/2} |K_q|^{d/2} \Gamma_d(\frac{v_1 - d + 1}{2}) e^{(v_1-d)d/2}\right) - \log\left(\lambda_1^{d/2}(v_1 - d)^{v_1 d/2}\right)
\end{aligned}$$

(old version lacking 1 subscripts commented out here)

The following algorithm serves as a summary of the above calculations. Note that, even though it seems as though all the new parameters are written in terms of each other, they depend on the previous parameters implicitly. This dependence is instantiated through the $\mathcal{F}_{max}$, K_c and K_q parameters, which are dependent on the true posterior, which is dependent on the prior and therefore the prior parameters. The prior (old) parameters are also used directly in the algorithm to calculate a starting point for the optimisation.

Algorithm 1: Approximate Soft-Evidence Update for Gaussian-Wishart Priors

Input: Prior Parameters (λ_0, μ_0, v_0, V_0),
Soft Evidence Message Parameters(K_m, μ_m),
(True Posterior Function $\mathcal{F}(\mu, K)$ - from above paramters)

Output: Approximate Gaussian-Wishart Posterior parameters (λ_1, μ_1, v_1, V_1)

$\{\mu_s, K_s\} = \mu_0, (v_0 - d - 1)V_0)\}$ (initial guess)

$\{\mu_1, (v_1 - d - 1)V_1)\} = argmax\{\mathcal{F}(\mu, K)\}$, initialise:$\{\mu_s, K_s\}$, constraint: K:PSD

(Also get $\mathcal{F}_{max}$)

So:

$\mu_1 = \mu_q$

$(v_1 - d - 1)V_1) = K_q$

$H_{\mu q} = Hessian(\quad -\log(\mathcal{F}(K = K_q, \mu))\quad)$ at $\mu = \mu_q$

$\lambda_1 = H_{\mu q}/K_q$

$\mu_c = \mu_1$ $K_c = (1 + \delta)K_q$ $A = 2log(\mathcal{F}(\mu_c, K_c)) - 2log(\mathcal{F}_{max}) + d^2 log(1 + \delta) - \delta d^2$

$v_1 = A/(log(1 + \delta)d - \delta d)$

$V_1 = K_q/(v_1 - d - 1)$

$\log(w_1) = log(\mathcal{F}_{max}) + \log\left((2\pi)^{d/2} 2^{v_1 d/2} |K_q|^{d/2} \Gamma_d(\frac{v_1 - d + 1}{2}) e^{(v_1 - d)d/2}\right) - \log\left(\lambda_1^{d/2}(v_1 - d)^{v_1 d/2}\right)$

Problems with this Approach

Although the above algorithm seems work very well at estimating the parameters of Gaussian-Wishart distributions, it looks as though the true posterior is in some case and in subtle ways quite different to a Gaussian-Wishart. Specifically, it looks as though one of the biggest differences between a real Gaussian-Wishart and the true posterior (after a soft evidence update) is that (if we try and view the true posterior as a Gaussian-Wishart) the λ parameter is not constant with respect to K. This can be seen in the graph below.

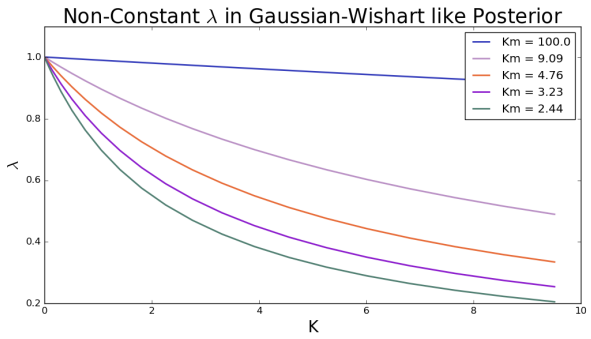

Figure 1: λ as a function of K for the 1 dimensional case of the true posterior (after soft evidence update) viewed as a Gaussian-Wishart like distribution

As the variance of the $\mathcal{N}(X)$ message gets larger, the lambda parameter does however approach a

constant (horizontal straight line). This is in line with our previous observation that the soft evidence update distribution approaches that of the perfect evidence (data update) update as K_m goes to infinity.

Furthermore, it is not clear how this characteristic of the true posterior will generalise to higher dimensions.

it is also worth noting that as K_m gets larger and the approximation gets better, the update also gets closer to the perfect evidence case, so using the data update becomes more valid and the use for the above algorithm diminishes. So the above algorithm, unfortunately, gives good approximations in cases where we don't need it and bad approximations in the cases that we do.

It would be convenient if lambda got less dependent on K as more and more messages were received. Then we could have possibly put of the approximation until lambda was fairly constant. Unfortunately, this does not seems to be the case. The graph below shows how $\lambda(K)$ changes as more $\mathcal{N}(X)$ messages (n is the number of messages) are received.

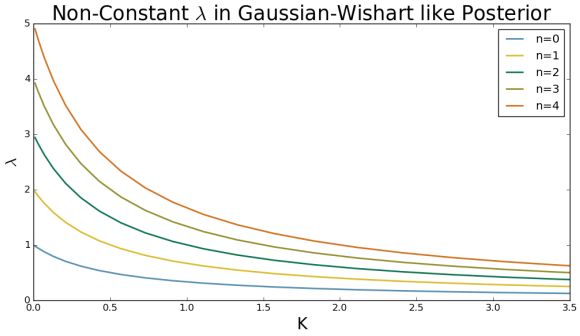

Figure 2: λ as a function of K for the 1 dimensional case of the true posterior (after soft evidence update) viewed as a Gaussian-Wishart like distribution. n is the number of soft evidence messages that have been received

Solution? Can we not solve this inaccuracy by letting λ be a function of K? How would this affect the distribution. Would it even help if we could do exact inference (see paragraph below).

Problems with the Exact Distribution

It seems as though even the exact soft evidence updates are not able to correctly infer the precision of the data. It looks as though, because of the large variance, the variance over the mean stays large and as a result of this the mode of the precision is often wrong and typically converges to a precision that is lower than that of the data and even lower than that of the noisy data? does this make any sense? In contrast, if the sample and noise covariances are small, the exact posterior often converges to the correct sample variance and mean. Furthermore (this might be related to the above mentioned problem, or not) it seems as though there are some fundamental problems with the inverse-Wishart (and also Wishart?) distribution as a prior for parameter inference. Other distributions (such as he scaled inverse-Wishart) have been proposed, that apparently have more desirable properties, but are more complex and therefore probably also more difficult to work with.

5 Assumption Checks

Assumption: the K 'coordinate' of the mode of a gaussian wishart is the same as for the wishart marginal.

$$
\begin{aligned}
W &= \frac{V^{-\frac{v}{2}}}{2^{vd/2}\Gamma_d(\frac{v}{2})} K^{-\frac{d}{2}+\frac{v}{2}-\frac{1}{2}} e^{-\frac{K}{2V}} \\
GW &= \frac{K^{0.5} K^{-\frac{d}{2}+\frac{v}{2}-\frac{1}{2}}}{2\pi^{d/2} \cdot 2^{vd/2}\Gamma_d(\frac{v}{2})} V^{-\frac{v}{2}} e^{-\frac{K}{2V}} e^{K(-0.5\mu+0.5\mu_0)(\mu-\mu_0)} \\
\frac{d\mathcal{W}(\mu,K)}{dK} &= \frac{K^{-\frac{d}{2}+\frac{v}{2}-\frac{3}{2}}}{2 \cdot 2^{vd/2}\Gamma_d(\frac{v}{2})} V^{-\frac{v}{2}-1} \left(-K - Vd + Vv - V\right) e^{-\frac{K}{2V}} \\
\frac{d\mathcal{NW}(\mu=\mu_0,K)}{dK} &= \frac{K^{-\frac{d}{2}+\frac{v}{2}-1.0} V^{-\frac{v}{2}-1}}{2 \cdot 2\pi^{d/2} \cdot 2^{vd/2}\Gamma_d(\frac{v}{2})} \left(-K^{1.0} - Vd + Vv\right) e^{-\frac{K}{2V}} \\
\frac{d\mathcal{NW}(\mu,K)}{dK} &= \frac{K^{-\frac{d}{2}+\frac{v}{2}-1.0}}{2\pi^{d/2} \cdot 2^{vd/2}\Gamma_d(\frac{v}{2})} V^{-\frac{v}{2}-1} \left(-0.5K^{1.0}V\mu^2 + 1.0K^{1.0}V\mu\mu_0 - 0.5K^{1.0}V\mu_0^2 - 0.5K^{1.0} - 0.5Vd + 0.5Vv\right) e^{K(-0.5} \\
\frac{d\mathcal{W}(\mu,K)}{dK} = 0: &\qquad K = \left[V\left(-d+v-1\right), \quad -V\left(d-v+1\right)\right] \\
\frac{d\mathcal{NW}(\mu,K)}{dK} = 0: &\qquad K = \left[\frac{V\left(-d+v\right)}{V\mu^2 - 2.0V\mu\mu_0 + V\mu_0^2 + 1.0}\right] \\
\frac{d\mathcal{NW}(\mu=\mu_0,K)}{dK} = 0: &\qquad K = \left[0^{\frac{1}{v}}, \quad V\left(-d+v\right)\right]
\end{aligned}
$$

[1] This is only for the scalar case. Still not sure whats going on here

see `Gaussian_Latent_Variable_Learning_Assumptions_Checks.py`

[1] These calculations were done with sympy

6 A simpler Case - Learning a Gaussian from Direct Observations

$$P(K) = \frac{|K|^{(v_0-d-1)/2}}{2^{v_0 d/2}|V_0|^{v_0/2}\Gamma_d\left(\frac{v_0}{2}\right)} e^{-tr(V_0^{-1}K)/2}$$

$$P(\mu|K) = \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])}$$

$$P(X|\mu, K) = \frac{|K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([X-\mu]^T K[X-\mu])}$$

The product $P(\mu|K)P(K)$ of the above distributions will not be normalised. The joint prior below should therefore rather be used in this derivation.

$$\mathcal{NW}(\mu, K) = (w_0)\frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])} \frac{|K|^{(v_0-d-1)/2}}{2^{v_0 d/2}|V_0|^{v_0/2}\Gamma_d(\frac{v_0-d+1}{2})} e^{-tr(V_0^{-1}K)/2}$$

We do however still include the weight term w_0 for generality. This will allow us to derive an update equation for the weight as well.The change to the previous derivation therefore simply corresponds to the additional multiplicative term

$$\frac{w_0\Gamma_d(\frac{v_0}{2})}{\Gamma_d(\frac{v_0-d+1}{2})}$$

We can write the correct marginal as follows

$$P(\mu, K) = w_0 P(\mu|K)P'(K) = (w_0)\left(\frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])}\right)\left(\frac{|K|^{(v_0-d-1)/2}}{2^{v_0 d/2}|V_0|^{v_0/2}\Gamma_d(\frac{v_0-d+1}{2})} e^{-tr(V_0^{-1}K)/2}\right)$$

$$\begin{aligned} P(X, \mu, K) &= P(X|\mu, K)P(\mu|K)P'(K) \\ P(\mu, K, X = x_i) &= P(X = x_i|\mu, K)P(\mu|K)P'(K) \end{aligned}$$

$$\begin{aligned} P(\mu, K|X = x_i) &= \frac{P(X = x_i|\mu, K)P(\mu|K)P'(K)}{P(X = x_i)} \\ P(\mu, K|X = x_i) &= \frac{P(X = x_i|\mu, K)P(\mu|K)P'(K)}{\int_{\mathbb{M}_{d,d}}\int_{\mathbb{R}^d} P(X = x_i|\mu, K)P(\mu|K)P'(K)d\mu dK} \\ P(\mu, K|X = x_i) &\propto P(X = x_i|\mu, K)P(\mu|K)P'(K) \end{aligned}$$

Where $\mathbb{M}_{d,d}$ is the space spanned by all d dimensional symmetric positive semi definite and $\mathbb{R}^d$ the space spanned by all d dimensional vectors.

$$
\begin{aligned}
P(X = x_i|\mu, K)P(\mu|K) &= \frac{|K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([x_i-\mu]^T K[x_i-\mu])} \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])} \\
&= \frac{|K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-x_i]^T K[\mu-x_i])} \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])} \\
&= \frac{|K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-x_i]^T K[\mu-x_i])} \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])} \\
&= \frac{|\lambda_0 K|^{1/2}|K|^{1/2}}{(2\pi)^{d/2}(2\pi)^{d/2}} e^{-0.5([\mu-x_i]^T K[\mu-x_i])} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])} \\
&= \frac{|\lambda_0 K|^{1/2}|K|^{1/2}}{(2\pi)^{d/2}(2\pi)^{d/2}} e^{-0.5([\mu-x_i]^T K[\mu-x_i])} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])} \\
&= \frac{|\lambda_0 K|^{1/2}|K|^{1/2}}{(2\pi)^{d/2}(2\pi)^{d/2}} e^{-0.5(H(\mu,K,\mu_0,\lambda_0,x_i))}
\end{aligned}
$$

$$
\begin{aligned}
H(\mu, K, \mu_0, \lambda_0, x_i) &= [\mu - x_i]^T K[\mu - x_i] + [\mu - \mu_0]^T \lambda_0 K[\mu - \mu_0] \\
&= \mu^T K\mu - 2\mu^T K x_i + x_i^T K x_i + \mu^T \lambda_0 K\mu - 2\mu^T \lambda_0 K\mu_0 + \mu_0^T \lambda_0 K\mu_0 \\
&= \mu^T K\mu - 2\mu^T K x_i - 2\mu^T \lambda_0 K\mu_0 + \mu^T \lambda_0 K\mu + x_i^T K x_i + \mu_0^T \lambda_0 K\mu_0 \\
&= \mu^T K\mu - 2\mu^T K x_i - 2\mu^T \lambda_0 K\mu_0 + \mu^T \lambda_0 K\mu + \mathbf{tr}(x_i x_i^T K + \lambda_0 \mu_0 \mu_0^T K) \\
&= \mu^T K\mu - 2\mu^T K(x_i + \lambda_0\mu_0) + \mu^T \lambda_0 K\mu + \mathbf{tr}(x_i x_i^T K + \lambda_0 \mu_0 \mu_0^T K) \\
&= \mu^T (1+\lambda_0) K\mu - 2\mu^T K(x_i + \lambda_0\mu_0) + \mathbf{tr}(x_i x_i^T K + \lambda_0 \mu_0 \mu_0^T K) \\
&= (1+\lambda_0)\left(\mu^T K\mu - 2\mu^T K\left(\frac{x_i + \lambda_0\mu_0}{1+\lambda_0}\right)\right) + \mathbf{tr}(x_i x_i^T K + \lambda_0 \mu_0 \mu_0^T K) \\
&= (1+\lambda_0)\left(\mu^T K\mu - 2\mu^T K\left(\frac{x_i + \lambda_0\mu_0}{1+\lambda_0}\right)\right) + \mathbf{tr}(x_i x_i^T K + \lambda_0 \mu_0 \mu_0^T K) \\
&= (1+\lambda_0)\left(\mu^T K\mu - 2\mu^T K\mu_1\right) + \mathbf{tr}(x_i x_i^T K + \lambda_0 \mu_0 \mu_0^T K) \\
&= [\mu - \mu_1]^T (1+\lambda_0) K[\mu - \mu_1] - (1+\lambda_0)\mu_1^T K\mu_1 + \mathbf{tr}(x_i x_i^T K + \lambda_0 \mu_0 \mu_0^T K) \\
&= [\mu - \mu_1]^T (1+\lambda_0) K[\mu - \mu_1] + \mathbf{tr}(x_i x_i^T K + \lambda_0 \mu_0 \mu_0^T K - (1+\lambda_0)\mu_1 \mu_1^T K) \\
&= [\mu - \mu_1]^T (1+\lambda_0) K[\mu - \mu_1] + \mathbf{tr}(x_i x_i^T K + \lambda_0 \mu_0 \mu_0^T K - (1+\lambda_0)\mu_1 \mu_1^T K) \\
&= [\mu - \mu_1]^T \lambda_1 K[\mu - \mu_1] + \mathbf{tr}(V_u^{-1} K)
\end{aligned}
$$

With

$$
\begin{aligned}
\lambda_1 &= \lambda_0 + 1 \\
\mu_1 &= \frac{x_i + \lambda_0\mu_0}{1+\lambda_0} \\
V_u^{-1} &= x_i x_i^T + \lambda_0 \mu_0 \mu_0^T - \lambda_1 \mu_1 \mu_1^T
\end{aligned}
$$

So

$$
\begin{aligned}
P(X = x_i|\mu, K)P(\mu|K) &= \frac{|\lambda_0 K|^{1/2}|K|^{1/2}}{(2\pi)^{d/2}(2\pi)^{d/2}} e^{-0.5\left([\mu-\mu_1]^T \lambda_1 K[\mu-\mu_1] + tr(V_u^{-1}K)\right)} \\
&= \frac{|\lambda_0 K|^{1/2}|K|^{1/2}}{(2\pi)^{d/2}(2\pi)^{d/2}} e^{-0.5[\mu-\mu_1]^T \lambda_1 K[\mu-\mu_1]} e^{-0.5tr(V_u^{-1}K)}
\end{aligned}
$$

And

$$
\begin{aligned}
&P(X=x_i,\mu,K) = P(X=x_i|\mu,K)P(\mu|K)P'(K) \\
&= (w_0)\frac{|\lambda_0 K|^{1/2}|K|^{1/2}}{(2\pi)^{d/2}(2\pi)^{d/2}}e^{-0.5[\mu-\mu_1]^T\lambda_1 K[\mu-\mu_1]}e^{-0.5tr(V_u^{-1}K)}\frac{|K|^{(v_0-d-1)/2}}{2^{v_0 d/2}|V|^{v_0/2}\Gamma_d(\frac{v_0-d+1}{2})}e^{-tr(V^{-1}K)/2} \\
&= (w_0)\frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}(2\pi)^{d/2}}e^{-0.5[\mu-\mu_1]^T\lambda_1 K[\mu-\mu_1]}\frac{|K|^{((v_0+1)-d-1)/2}}{2^{(v_0+1)d/2}|V|^{v_0/2}\Gamma_d(\frac{v_0-d+1}{2})}e^{-tr((V^{-1}+V_u^{-1})K)/2} \\
&= (w_0)\frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}(2\pi)^{d/2}}e^{-0.5[\mu-\mu_1]^T\lambda_1 K[\mu-\mu_1]}\frac{|K|^{(v_1-d-1)/2}}{2^{v_1 d/2}|V|^{v_0/2}\Gamma_d(\frac{v_0-d+1}{2})}e^{-tr(V_1^{-1}K)/2} \\
&= (w_0)\frac{\lambda_0^{d/2}|\lambda_1 K|^{1/2}}{\lambda_1^{d/2}(2\pi)^{d/2}(2\pi)^{d/2}}e^{-0.5[\mu-\mu_1]^T\lambda_1 K[\mu-\mu_1]}\frac{\Gamma_d(\frac{v_1-d+1}{2})|V_1|^{v_1/2}|K|^{(v_1-d-1)/2}}{\Gamma_d(\frac{v_0-d+1}{2})|V_0|^{v_0/2}2^{v_1 d/2}|V_1|^{v_1/2}\Gamma_d(\frac{v_1-d+1}{2})}e^{-tr(V_1^{-1}K)/2} \\
&= (w_0)\left(\frac{\lambda_0^{d/2}\Gamma_d(\frac{v_1-d+1}{2})|V_1|^{v_1/2}}{\lambda_1^{d/2}(2\pi)^{d/2}\Gamma_d(\frac{(v_0-d+1)}{2})|V_0|^{v_0/2}}\right)\frac{|\lambda_1 K|^{1/2}}{(2\pi)^{d/2}}e^{-0.5[\mu-\mu_1]^T\lambda_1 K[\mu-\mu_1]}\frac{|K|^{(v_1-d-1)/2}}{2^{v_1 d/2}|V_1|^{v_1/2}\Gamma_d(\frac{v_1-d+1}{2})}e^{-tr(V_1^{-1}K)/2} \\
&= (w_1)\frac{|\lambda_1 K|^{1/2}}{(2\pi)^{d/2}}e^{-0.5[\mu-\mu_1]^T\lambda_1 K[\mu-\mu_1]}\frac{|K|^{(v_1-d-1)/2}}{2^{v_1 d/2}|V_1|^{v_1/2}\Gamma_d(\frac{v_1-d+1}{2})}e^{-tr(V_1^{-1}K)/2}
\end{aligned}
$$

Comparing the above expression to the standard expression for a Gaussian-Wishart distribution:

$$\mathcal{NW}(\mu,K) = (w)\frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}}e^{-0.5([\mu-\mu_0]^T\lambda_0 K[\mu-\mu_0])}\frac{|K|^{(v_0-d-1)/2}}{2^{v_0 d/2}|V_0|^{v_0/2}\Gamma_d(\frac{v_0-d+1}{2})}e^{-tr(V_0^{-1}K)/2}$$

We see that is has the same form, so it is also a Gaussian-Wishart distribution and with following parmaters (i.t.o the previous parameters)[2]

$$
\begin{aligned}
w_1 =& (w_0)\left(\frac{\lambda_0^{d/2}\Gamma_d(\frac{v_1-d+1}{2})|V_1|^{v_1/2}}{\lambda_1^{d/2}(2\pi)^{d/2}\Gamma_d(\frac{(v_0-d+1)}{2})|V_0|^{v_0/2}}\right) \\
\lambda_1 =& \lambda_0 + 1 \\
\mu_1 =& \frac{x_i + \lambda_0\mu_0}{1+\lambda_0} \\
v_1 =& v+1 \\
V_1^{-1} =& V^{-1} + V_u^{-1} \\
=& V^{-1} + x_i x_i^T + \lambda_0\mu_0\mu_0^T - \lambda_1\mu_1\mu_1^T
\end{aligned}
$$

Below we compare the results of this case, where X is observed to the case where X is latent.

Observed Case:

$$P(X=x_i|\mu,K) = \frac{|K|^{1/2}}{(2\pi)^{d/2}}e^{-0.5([x_i-\mu]^T K[x_i-\mu])}$$

[2]See *K. Murphy - Machine Learning: A Probabilistic Perspective* (pp.134) for a similar result of a derivation using an Inverse-Wishart Prior on the covariance.

Latent Case:

$$\int_{\mathbb{R}^d} P(X|\mu,K)P_m(X)dX = \qquad = \frac{|(K^{-1}+K_m^{-1})^{-1}|^{1/2}}{(2\pi)^{d/2}} e^{-0.5(\mu-\mu_m)^T(K^{-1}+K_m^{-1})^{-1}(\mu-\mu_m)}$$

If we let the covariance of the message in the above expression go to zero (corresponding to infinite precision), it reduces to the following

$$\int_{\mathbb{R}^d} P(X|\mu,K)P_m(X)dX = \frac{|K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5(\mu-\mu_m)^T K(\mu-\mu_m)}$$

We can see that the above term is identical to the term for the observed case. This is as one would intuitively think it should be and is therefore a indication that the latent case makes sense.

7 Message Passing in a Cluster Graph

In this section we will discuss the message passing procedure that will be needed in order to learn the parameters of a latent gaussian variable from observations of related (a variable that has some dependence path to the latent variable) variables.

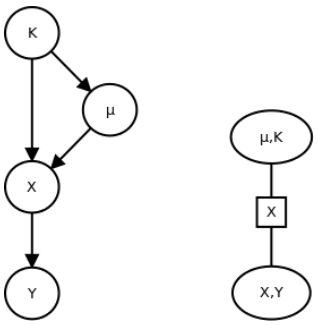

Figure 3: Basic Bayes Network and Corresponding Cluster Graph for Latent Gaussian Parameter Learning

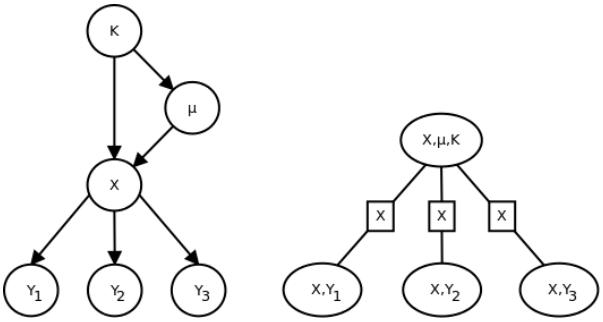

Figure 4: Bayes Network and Corresponding Cluster Graph for Latent Gaussian Parameter Learning from multiple observations

Looking at the cluster graph above, and assuming a linear gaussian dependence between X and Y (for now), we can see that the message passing schedule will be very simple. In the linear gaussian case, we can compute and send the messages from the $\{X, Y\}$ clusters to the $\{X, \mu, K\}$ cluster without first receiving a message about X at each of the $\{X, Y\}$ clusters [3]. We can do this because all the information is coming from the observed Y variables. Message passing in the above graph will therefore consist of each $\{X, Y\}$ cluster sending a Gaussian message to the $\{X, \mu, K\}$ cluster. Once the message passing is complete, the $\{\mu, K\}$ marginal, is the posterior over the parameters of X. This posterior will therefore be of the following form.

[3] In fact, we can probably do the same in the non-linear case, if we can work with the inverse transform Y from to X

$$
\begin{aligned}
P'(\mu, K) &= \int_{\mathbb{R}} \prod_{i}^{N} \delta_i(X) P(X|\mu, K) P(\mu, K) dX \\
&= P(\mu, K) \int_{\mathbb{R}} \prod_{i}^{N} \delta_i(X) P(X|\mu, K) dX
\end{aligned}
$$

With

$$
\begin{aligned}
P(X|\mu, K) &= \frac{|K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([X-\mu]^T K[X-\mu])} \\
P(\mu, K) &= \frac{|K|^{(v-d-1)/2}}{2^{vd/2}|V|^{v/2}\Gamma_d(\frac{v-d+1}{2})} e^{-tr(V^{-1}K)/2} \frac{|\lambda_0 K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([\mu-\mu_0]^T \lambda_0 K[\mu-\mu_0])} \\
\delta_i(X) &= \mathcal{N}(X|\mu_i, K_i)
\end{aligned}
$$

Question: Correct Model?

The above latent model does however not seem to result in sensible inference. The variance of the precision stays large, even with many updates. The below model corresponds to the inference done in the first testing simulation (to test the posterior approximation). This model yields much more accurate inference. There does however not seem to be a good explanation for why the first model should be wrong? The main differences between the two models is that the first accumulates all the messages over the data distribution X into a single product and the prior is never updated (unless the marginal is manually taken of course, but this is not part of the natural message passing). The second model needs to update the prior after every data distribution message is received. This also seems as though this is what is happening during inference for the observed evidence case.

Question: Inferring Correct Precision with Noise Grater than Data Variance

It seems that we should still be able to infer the precision of data (X) even if the variance of the noise is greater than that of the data (if we know what the noise variance is). The models tested so far does not seem to be able to do this. Is there something missing in the models? Or is this not possible?

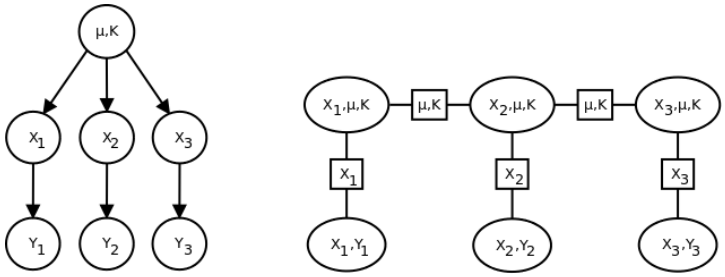

Figure 5: Bayes Network and Corresponding Cluster Graph for Latent Gaussian Parameter Learning from multiple observations

Below is an example of a model with X observed, for comparison.

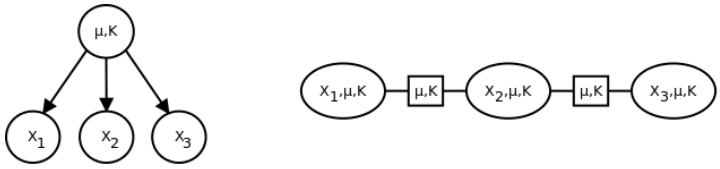

Figure 6: Bayes Network and Corresponding Cluster Graph for Gaussian Parameter Learning from multiple observations

Question: Inferring Correct Precision with Noise Greater than Data Variance

Edit: It seems as though this (see heading) is indeed possible, but only if the sample precision is larger than one? So if the sample variance is larger than one, the variance estimate seems to converge to that of the total (sample plus the noise) variance? This seems very strange. Could it have something to do with the prior or inherent properties of the Wishart distribution?

It seems that we should still be able to infer the precision of data (X) even if the variance of the noise is greater than that of the data (if we know what the noise variance is). The models tested so far does not seem to be able to do this. Is there something missing in the models? Or is this not possible?

Observed X - A special Case

If $Y = X$ (linear "transform" with no added noise), the $\{X, \mu, K\}$ cluster will Gaussian messages with zero variance. These functions are effectively dirac delta functions centered at the mean of the Gaussian. In this case the

Suppose that

$$K_1 = K_2 = K = \lim_{k \to \infty} kI$$

$$
\begin{aligned}
&(X-\mu_1)^T K_1(X-\mu_1) + (X-\mu_2)^T K_1(X-\mu_2) \\
=&X^T K_1 X - 2X^T K_1\mu_1 + \mu_1^T K_1\mu_1 \\
&+ X^T K_2 X - 2X^T K_2\mu_2 + \mu_2^T K_2\mu_2 \\
=&X^T(K_1+K_2)X - 2X^T(K_1\mu_1 + K_2\mu_2) + \mu_1^T K_1\mu_1 + \mu_2^T K_2\mu_2 \\
=&X^T(2K)X - 2X^T(K(\mu_1+\mu_2)) + \mu_1^T K\mu_1 + \mu_2^T K\mu_2 \\
=&X^T(K')X - 2X^T h' + \mu_1^T K\mu_1 + \mu_2^T K\mu_2 \\
=&X^T(K')X - 2X^T K'((K')^{-1}h') + \mu_1^T K\mu_1 + \mu_2^T K\mu_2 \\
=&X^T(K')X - 2X^T K'\mu' + \mu_1^T K\mu_1 + \mu_2^T K\mu_2 \\
=&(X-\mu')^T(K')(X-\mu') - (\mu')^T K'\mu' + \mu_1^T K\mu_1 + \mu_2^T K\mu_2 \\
=&(X-\mu')^T(K')(X-\mu') - (\mu')^T K'\mu' + \mu_1^T K\mu_1 + \mu_2^T K\mu_2 \\
=&(X-\mu')^T(K')(X-\mu') - 0.5(\mu_1+\mu_2)^T K(\mu_1+\mu_2) + \mu_1^T K\mu_1 + \mu_2^T K\mu_2 \\
=&(X-\mu')^T(K')(X-\mu') \\
&- 0.5\mu_1^T K\mu_1 - 0.5\mu_1^T K\mu_2 \\
&- 0.5\mu_2^T K\mu_1 - 0.5\mu_2^T K\mu_2 \\
&+ \mu_1^T K\mu_1 + \mu_2^T K\mu_2 \\
=&(X-\mu')^T(K')(X-\mu') \\
&- 0.5\mu_1^T K\mu_1 - \mu_1^T K\mu_2 - 0.5\mu_2^T K\mu_2 \\
&+ \mu_1^T K\mu_1 + \mu_2^T K\mu_2 \\
=&(X-\mu')^T(K')(X-\mu') \\
&+ 0.5\mu_1^T K\mu_1 - \mu_1^T K\mu_2 + 0.5\mu_2^T K\mu_2 \\
=&(X-\mu')^T(K')(X-\mu') \\
&+ 0.5(\mu_1^T K\mu_1 - 2\mu_1^T K\mu_2 + \mu_2^T K\mu_2) \\
=&(X-\mu')^T(K')(X-\mu') \\
&+ 2(\mu_1-\mu_2)^T K(\mu_1-\mu_2)
\end{aligned}
$$

Where the following was used

$$
\begin{aligned}
h' &= K(\mu_1+\mu_2) = 0.5K'(\mu_1+\mu_2) \\
\mu' &= (K')^{-1}h' \\
&= (K')^{-1}0.5K'(\mu_1+\mu_2) \\
&= 0.5(\mu_1+\mu_2)
\end{aligned}
$$

Using the above result, and setting $K = k\mathbf{I}$ and letting

$$\lim_{k\to\infty}$$

the gaussian product is:

$$
\begin{aligned}
&\frac{|K|^{1/2}|K|^{1/2}}{(2\pi)^d} e^{-0.5(X-\mu')^T(K')(X-\mu')} e^{-(\mu_1-\mu_2)^T K(\mu_1-\mu_2)} \\
=&\frac{|k\mathbf{I}|}{(2\pi)^d} e^{-0.5(X-\mu')^T(2k\mathbf{I})(X-\mu')} e^{-(\mu_1-\mu_2)^T k\mathbf{I}(\mu_1-\mu_2)} \\
=&\frac{k^d}{(2\pi)^d} e^{-0.5(X-\mu')^T(2k\mathbf{I})(X-\mu')} e^{-(\mu_1-\mu_2)^T(\mu_1-\mu_2)k} \\
=&\frac{k^d}{(2\pi)^d} e^{-0.5(X-\mu')^T(2k\mathbf{I})(X-\mu')} e^{-\sum_i^d \mu_i^2 k} \\
=&\frac{k^d}{(2\pi)^2} e^{-0.5(X-\mu')^T(2k\mathbf{I})(X-\mu')} e^{-S_\mu k} \\
=&\lim_{k\to\infty} \frac{k^d}{e^{S_\mu k}(2\pi)^d} e^{-0.5(X-\mu')^T(2k\mathbf{I})(X-\mu')}
\end{aligned}
$$

The above function is a dirac delta function, whose height goes to zero?

(This paragraph is not directly related to the above, where two dirac deltas were multiplied together - perhaps I should include another derivation with a finite precision gaussian times a dirac delta and then integrating over the scope of the dirac delta) If a dirac delta function is multiplied with a distribution and the product is integrated over the scope of the dirac delta, the resulting function is only non-zero at the points where the dirac delta is non-zero. Multiplying with a dirac delta (a Gaussian distribution, whose precision tends to infinity) is therefore equivalent to the evidence observation[4] operation, and is therefore a special case of normal message passing, where we perform the same operations, but with Gaussian's (or other types of distributions) with finite precision.

[4]the ObserveAndReduce functions also reduce the dimensionality of the disribution after the evidence is observed by removing the dimensions of the observed variables, as these variables can have only one value, are therefore deterministic and the joint distribution therefore no longer has a distribution in these dimensions. This is equivalent to the additional operation of integrating over the scope of the dirac delta, when viewing the operation(s) in the alternative way.

Question: Possible Logical Contradiction

If Observing evidence in a certain factor is equivalent to receiving a infinite precision Gaussian (IPG) message at that factor, then it makes sense that observing multiple data points is the same as receiving multiple IPG messages at that factor. If it is also true that the product of an IPG is again an infinite precision Gaussian with a mean that is the average of the individual means, then it seems as though an entire data set can be summarised by a single IPG? Or stated differently, it seems as though a specific, single IPG update can have the same effect on the distribution as the cumulative updates from an arbitrary number of IPG's. This does not seem reasonable, even if the data is Gaussian distributed (as it should be if the variable we are observing is a Gaussian random variable) and the data set is fully summarised by only it's mean and covariance, it seems we are still losing information by apparently throwing away the covariance information in the single equivalent IPG case?

A possible solution (sort of, not really)

If the precisions were not completely infinite, but just very large, the 'sample' variance would be encoded in the product of all the distributions - the closer the 'samples' are together the smaller the precision and vice versa. The covariance 'sample' information would also be encoded in the covariance terms of the product distribution. It seems therefore as though we are loosing information because of the infinite precision. If we could atleast keep track of the relative rates at which the precision terms of the product distribution go to infinity, perhaps we can salvage this information. This 'solution' does not however help us with sample distributions that are non-Gaussian? The other problem is of course that we also loose the closed-form property of inference if we do not take the classical approach.

The Stupid Way

This is the way that it is done in the simulation:

$$
\begin{aligned}
P_1(\mu, K) &= P(\mu, K) \int_{\mathbb{R}} \delta_1(X) P(X|\mu, K) dX \\
P_2(\mu, K) &= P_1(\mu, K) \int_{\mathbb{R}} \delta_2(X) P(X|\mu, K) dX \\
&= P(\mu, K) \int_{\mathbb{R}} \delta_1(X) P(X|\mu, K) dX \int_{\mathbb{R}} \delta_2(X) P(X|\mu, K) dX
\end{aligned}
$$

The last term above should be

$$
P_2(\mu, K) = P(\mu, K) \int_{\mathbb{R}} \delta_1(X) \delta_2(X) P(X|\mu, K) dX
$$

8 Expectation Maximisation Approach

8.1 Expectation Maximisation Brief Recap

See https://www2.ee.washington.edu/techsite/papers/documents/UWEETR-2010-0002.pdf

In EM, we want to find the parameters that maximise the data likelihood as shown in the expression below:

$$\theta_{MLE} = argmax_{\theta \in \Theta} p(y|\theta)$$

or equivalently

$$\theta_{MLE} = argmax_{\theta \in \Theta} log(p(y|\theta))$$

In some problems, however, it is difficult to perform this maximisation directly and EM offers an alternative solution. A brief recap of the EM algorithm is given below.

With:

- Z being the hidden variables
- $Y = \mathbf{y}$ the observed variables
- θ being the unknown constant parameters

Expectation Maximisation Algorithm:

1. Initial guess for $\theta = \theta_{m=0}$
2. Get $P(Z|Y = \mathbf{y}, \theta_m)$
3. We need to maximise $P(Z = \mathbf{z}, Y = \mathbf{y}|\theta) \propto P(Z = \mathbf{z}|Y = \mathbf{y}, \theta)$, but we do not know $\mathbf{z}$. Instead we will maximize the expected $P(\mathbf{z}|Y = \mathbf{y}, \theta)$, which we can write as follows
   Alternative:

$$\theta_{MLE} = argmax_{\theta \in \Theta} \int_Z P(Z|Y = \mathbf{y}, \theta = \theta_n) ln(P(Y = \mathbf{y}, Z|\theta)$$
$$\theta_{MLE} = argmax_{\theta \in \Theta} \left(\mathbb{E}_{\mathbf{z}|Y=\mathbf{y}, \theta=\theta_n} \{ ln(P(Y = \mathbf{y}, \mathbf{z}|\theta) \} \right)$$

   see https://www.cs.utah.edu/~piyush/teaching/EM_algorithm.pdf

Note, although EM is an established algorithm and should allow us to estimate the most likely parameters of the distribution, it will not allow us to maintain and update distributions over these parameters. In the large data limit, the true posterior over parameters might become more and more concentrated and the maximum likelihood estimate provided by EM will become more accurate.

$$
\begin{aligned}
P(Z|\mu, K) &= \frac{|K|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([Z-\mu]^T K[Z-\mu])} \\
P(Y|Z) &= \frac{|K_1|^{1/2}}{(2\pi)^{d/2}} e^{-0.5([Z-\mu_1]^T K_1[Z-\mu_1])} \\
P(Y, Z, \mu, K) &= P(Y|Z)P(Z|\mu, K)P(\mu, K) \\
P(Y, Z|\mu, K) &= P(Y|Z)P(Z|\mu, K)P(\mu, K)/P(\mu, K)) \\
&= P(Y|Z)P(Z|\mu, K) \\
P(Y|\mu, K) &= \int_Z P(Y|Z)P(Z|\mu, K)
\end{aligned}
$$

Where Z is the unobserved Gaussian random variable (the same as X in the rest of this document).

Maximum likelihood:

$$
\theta_{MLE} = argmax_{\theta \in \Theta} \int_Z P(Z|Y = \mathbf{y}, \theta = \theta_n) ln(P(Y = \mathbf{y}, Z|\theta)
$$

with:

$$
\theta = \{\mu, K\}
$$

$$
\begin{aligned}
P(Z|Y = \mathbf{y}, \theta = \theta_n) &= \prod_{i=0}^{m} P(Z|Y = \mathbf{y}_m, \mu_n, K_n) \\
ln(P(Y = \mathbf{y}, Z|\theta) &= ln\left(\prod_{i=0}^{m} P(Y = \mathbf{y}_m, Z|\mu, K)\right) \\
&= \sum_{i=0}^{m} ln\left(P(Y = \mathbf{y}_m, Z|\mu, K)\right)
\end{aligned}
$$

$$
\begin{aligned}
P(Y = \mathbf{y}_m, Z|\mu, K) &= P(Y|Z)P(Z|\mu, K) \\
&= \left(\frac{|K_{YZ}|^{1/2}}{(2\pi)^{d_X/2}} e^{-0.5([Z-\mu_{YZ}]^T K_{YZ}[Z-\mu_{YZ}])}\right)\left(\frac{|K|^{1/2}}{(2\pi)^{d_Z/2}} e^{-0.5([Z-\mu]^T K[Z-\mu])}\right) \\
ln\left(P(Y = \mathbf{y}_m, Z|\mu, K)\right) &= ln\left(\frac{|K_{YZ}|^{1/2}}{(2\pi)^{d_X/2}}\right) - 0.5(Z-\mu_{YZ}]^T K_{YZ}[Z-\mu_{YZ}]) \\
&+ ln\left(\frac{|K|^{1/2}}{(2\pi)^{d_Z/2}}\right) - 0.5([Z-\mu]^T K[Z-\mu]) \\
&= 0.5ln\left(\frac{|K_{YZ}||K|}{(2\pi)^{d_X+d_Z}}\right) - 0.5(Z-\mu_{YZ}]^T K_{YZ}[Z-\mu_{YZ}]) - 0.5([Z-\mu]^T K[Z-\mu])
\end{aligned}
$$

PROBLEM: We cant integrate the product of the above with a Gaussian

9 Random Derivations

Random Derivation 1.1

$$
\begin{aligned}
& -\mu_p^T K_p \mu_p + \mu^T K \mu + \mu_m^T K_m \mu_m \\
=& -\mu^T K_p \mu - \mu^T K_m K_p^{-1} K_m \mu + 2\mu^T K_m K_p^{-1} K_m \mu_m - \mu_m^T K_m K_p^{-1} K_m \mu_m - 2\mu^T K_m \mu_m + \mu^T 2 K_m \mu \\
& + \mu^T K \mu + \mu_m^T K_m \mu_m \\
=& -\mu^T (K_p - K) \mu - \mu^T K_m K_p^{-1} K_m \mu + 2\mu^T K_m K_p^{-1} K_m \mu_m - \mu_m^T K_m K_p^{-1} K_m \mu_m \\
& - 2\mu^T K_m \mu_m + \mu^T 2 K_m \mu + \mu_m^T K_m \mu_m \\
=& -\mu^T K_m K_p^{-1} K_m \mu + 2\mu^T K_m K_p^{-1} K_m \mu_m - \mu_m^T K_m K_p^{-1} K_m \mu_m \\
& - 2\mu^T K_m \mu_m + \mu^T K_m \mu + \mu_m^T K_m \mu_m \\
K_C =& K_m K_p^{-1} K_m \\
=& -\mu^T K_C \mu + 2\mu^T K_C \mu_m - \mu_m^T K_C \mu_m \\
& - 2\mu^T K_m \mu_m + \mu^T K_m \mu + \mu_m^T K_m \mu_m \\
=& -(\mu - \mu_m)^T K_C (\mu - \mu_m) - 2\mu^T K_m \mu_m + \mu^T K_m \mu + \mu_m^T K_m \mu_m \\
=& -(\mu - \mu_m)^T K_C (\mu - \mu_m) + (\mu - \mu_m)^T K_m (\mu - \mu_m) \\
=& (\mu - \mu_m)^T (K_m - K_C)(\mu - \mu_m) \\
=& (\mu - \mu_m)^T (K_m - K_m K_p^{-1} K_m)(\mu - \mu_m)
\end{aligned}
$$