Appendix A

Conjugate Exponential family examples

The following two tables present information for a variety of exponential family distributions, and include entropies, KL divergences, and commonly required moments. Where used, tilde symbols (e.g. $\tilde{\boldsymbol{\theta}}$), denote the parameters of a different distribution of the same form. Therefore $\mathrm{KL}(\tilde{\boldsymbol{\theta}}||\boldsymbol{\theta})$ is shorthand for the KL divergence between the distribution with parameter $\tilde{\boldsymbol{\theta}}$ and the distribution with parameter $\boldsymbol{\theta}$ (averaging with respect to the first distribution that is specified). The remainder of the notation should be self-explanatory.

Distribution	Notation & Parameters	Density function	Moments, entropy, KL-divergence, etc.
Exponential Family	$\boldsymbol{\theta} \sim ExpFam(\eta, \boldsymbol{\nu})$ number η and value $\boldsymbol{\nu}$ of pseudo-observations	$p(\boldsymbol{\theta} \mid \eta, \boldsymbol{\nu}) = \frac{1}{Z_{\eta\boldsymbol{\nu}}} g(\boldsymbol{\theta})^{\eta} e^{\boldsymbol{\phi}(\boldsymbol{\theta})^{\top}\boldsymbol{\nu}}$	$H_{\boldsymbol{\theta}} = \ln Z_{\eta\boldsymbol{\nu}} - \eta\langle \ln g(\boldsymbol{\theta})\rangle - \boldsymbol{\nu}^{\top}\langle \boldsymbol{\phi}(\boldsymbol{\theta})\rangle$
Uniform	$\theta \sim U(a, b)$ boundaries a,b with $b > a$	$p(\theta \mid a, b) = \frac{1}{b-a}, \theta \in [a, b]$	$H_{\theta} = \ln(b-a)$ $\langle\theta\rangle = \frac{a+b}{2}, \langle\theta^2\rangle - \langle\theta\rangle^2 = \frac{(b-a)^2}{12}$
Laplace	$\theta \sim Laplace(\mu, \lambda)$ μ mean λ decay scale	$p(\theta \mid \mu, \lambda) = \frac{1}{2\lambda} e^{-\frac{\lvert\theta-\mu\rvert}{\lambda}}$ $\lambda > 0$	$H_{\theta} = 1 + \ln(2\lambda)$
Multivariate normal (Gaussian)	$\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \Sigma)$ $\boldsymbol{\mu}$ mean vector Σ covariance	$p(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \Sigma) = (2\pi)^{-d/2} \lvert\Sigma\rvert^{-1/2} e^{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})(\boldsymbol{\theta}-\boldsymbol{\mu})^{\top}\right]}$	$H_{\boldsymbol{\theta}} = \frac{d}{2}(\ln 2\pi e) + \frac{1}{2}\ln\lvert\Sigma\rvert$ $\mathrm{KL}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma} \mid\mid \boldsymbol{\mu}, \Sigma) = -\frac{1}{2}\left(\ln\left\lvert\tilde{\Sigma}\Sigma^{-1}\right\rvert + \mathrm{tr}\left[I - \left[\tilde{\Sigma} + (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\top}\right]\Sigma^{-1}\right]\ln e\right)$ $\langle\boldsymbol{\theta}\rangle = \boldsymbol{\mu}$ $\langle\boldsymbol{\theta}\boldsymbol{\theta}^{\top}\rangle = \Sigma$ $K_{\boldsymbol{\theta}} = \frac{\langle\boldsymbol{\theta}^4\rangle}{\langle\boldsymbol{\theta}^2\rangle^2} - 3 = 0$ (relative kurtosis)
Gamma	$\tau \sim Ga(\alpha, \beta)$ shape $\alpha > 0$ inv. scale $\beta > 0$	$p(\tau \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau}$	$H_{\tau} = \ln\Gamma(\alpha) - \ln\beta + (1-\alpha)\psi(\alpha) + \alpha$ $\langle\tau^n\rangle = \frac{\Gamma(\alpha+n)}{\beta^n\Gamma(\alpha)}$ $\langle(\ln\tau)^n\rangle = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\frac{\partial^n}{\partial\alpha^n}\left(\frac{\Gamma(\alpha)}{\beta^{\alpha}}\right)$ $\langle\tau\rangle = \alpha/\beta$ $\langle\tau^2\rangle - \langle\tau\rangle^2 = \alpha/\beta^2$ $\langle\ln\tau\rangle = \psi(\alpha) - \ln\beta$ $\mathrm{KL}(\tilde{\alpha}, \tilde{\beta} \mid\mid \alpha, \beta) = \tilde{\alpha}\ln\tilde{\beta} - \alpha\ln\beta - \ln\frac{\Gamma(\tilde{\alpha})}{\Gamma(\alpha)} + (\tilde{\alpha}-\alpha)(\psi(\tilde{\alpha}) - \ln\tilde{\beta}) - \tilde{\alpha}(1 - \frac{\beta}{\tilde{\beta}})$

Distribution	Notation & Parameters	Density function	Moments, entropy, KL-divergence, etc.
Wishart	$W \sim Wishart_\nu(S)$ deg. of freedom ν precision matrix S	$p(W \mid \nu, S) = \frac{1}{Z_{\nu S}} \lvert W\rvert^{(\nu-k-1)/2} e^{-\frac{1}{2}\mathrm{tr}\left[S^{-1}W\right]}$ $Z_{\nu S} = 2^{\nu k/2}\pi^{k(k-1)/4} \lvert S\rvert^{\nu/2} \prod_{i=1}^{k} \Gamma\left(\frac{\nu+1-i}{2}\right)$	$H_W = \ln Z_{\nu S} - \frac{\nu-k-1}{2}\langle \ln \lvert W\rvert \rangle + \frac{1}{2}\nu k$ $\langle W \rangle = \nu S$ $\langle \ln \lvert W\rvert \rangle = \sum_{i=1}^{k} \psi\left(\frac{\nu+1-i}{2}\right) + k \ln 2 + \ln \lvert S\rvert$ $\mathrm{KL}(\tilde{\nu}, \tilde{S} \lvert\rvert \nu, S) = \ln \frac{Z_{\nu S}}{Z_{\tilde{\nu}\tilde{S}}}$ $+ \frac{\tilde{\nu}-\nu}{2}\langle \ln \lvert W\rvert \rangle_{\tilde{Q}} + \frac{1}{2}\tilde{\nu}\,\mathrm{tr}\left[S^{-1}\tilde{S} - \mathrm{I}\right]$
Inverse-Wishart	$W \sim Inv\text{–}Wishart_\nu(S^{-1})$ deg. of freedom ν covariance matrix S	$p(W \mid \nu, S^{-1}) = \frac{1}{Z} \lvert W\rvert^{-(\nu+k+1)/2} e^{-\frac{1}{2}\mathrm{tr}\left[SW^{-1}\right]}$ $Z = 2^{\nu k/2}\pi^{k(k-1)/4} \prod_{i=1}^{k} \Gamma\left(\frac{\nu+1-i}{2}\right) \times \lvert S\rvert^{-\nu/2}$	$\langle W \rangle = (\nu - k - 1)^{-1} S$
Student-t (1)	$\theta \sim t_\nu(\mu, \sigma^2)$ deg. of freedom $\nu > 0$ mean μ; scale $\sigma > 0$	$p(\theta \mid \nu, \mu, \sigma^2) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma} \left(1 + \frac{1}{\nu}\left(\frac{\theta-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2}$	$\langle \theta \rangle = \mu$, for $\nu > 1$ $\langle \theta^2 \rangle - \langle \theta \rangle^2 = \frac{\nu}{\nu-2}\sigma^2$, for $\nu > 2$
Student-t (2)	$\theta \sim t(\mu, \alpha, \beta)$ shape $\alpha > 0$; mean μ scale^2 $\beta > 0$	$p(\theta \mid \mu, \alpha, \beta) = \frac{\Gamma(\alpha+1/2)}{\Gamma(\alpha)\sqrt{2\pi\beta}} \left(1 + \frac{(\theta-\mu)^2}{2\beta}\right)^{-(\alpha+1/2)}$	$H_\theta = \left[\psi(\alpha + \frac{1}{2}) - \psi(\alpha)\right](\alpha + \frac{1}{2})$ $+ \ln \sqrt{2\beta} B(\frac{1}{2}, \alpha)$ $K_\theta = \frac{3}{\alpha-2}$ (relative to Gaussian) equiv. $\alpha \to \frac{\nu}{2}; \beta \to \frac{\nu}{2}\sigma^2$
Multivariate Student-t	$\boldsymbol{\theta} \sim t_\nu(\boldsymbol{\mu}, \Sigma)$ deg. of freedom $\nu > 0$ mean $\boldsymbol{\mu}$; scale^2 matrix Σ	$p(\boldsymbol{\theta} \mid \nu, \boldsymbol{\mu}, \Sigma) = \frac{1}{Z}\left(1 + \frac{1}{\nu}\mathrm{tr}\left[\Sigma^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})^\top\right]\right)^{-(\nu+d)/2}$ $Z = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)(\nu\pi)^{d/2}} \lvert\Sigma\rvert^{-1/2}$	$\langle \boldsymbol{\theta} \rangle = \boldsymbol{\mu}$, for $\nu > 1$ $\langle \boldsymbol{\theta}\boldsymbol{\theta}^\top \rangle - \langle \boldsymbol{\theta} \rangle \langle \boldsymbol{\theta} \rangle^\top = \frac{\nu}{\nu-2}\Sigma$, for $\nu > 2$
Beta	$\theta \sim Beta(\alpha, \beta)$ prior sample sizes $\alpha > 0, \beta > 0$	$p(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$ $\theta \in [0, 1]$	See Dirichlet with $k = 2$
Dirichlet	$\boldsymbol{\pi} \sim Dir(\boldsymbol{\alpha})$ prior sample sizes $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_k\}$ $\alpha_j > 0; \alpha_0 = \sum_{j=1}^{k} \alpha_j$	$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} \pi_1^{\alpha_1-1} \cdots \pi_k^{\alpha_k-1}$ $\pi_1, \ldots, \pi_k \geq 0; \sum_{j=1}^{k} \pi_j = 1$	$\langle \boldsymbol{\pi} \rangle = \boldsymbol{\alpha}/\alpha_0$ $\langle \boldsymbol{\pi}\boldsymbol{\pi}^\top \rangle - \langle \boldsymbol{\pi} \rangle \langle \boldsymbol{\pi} \rangle^\top = \frac{\alpha_0 \mathrm{diag}(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top}{\alpha_0^2(\alpha_0+1)}$ $\langle \ln \pi_j \rangle = \psi(\alpha_j) - \psi(\alpha_0)$ $\mathrm{KL}(\tilde{\boldsymbol{\alpha}} \lvert\rvert \boldsymbol{\alpha}) = \ln \frac{\Gamma(\tilde{\alpha}_0)}{\Gamma(\alpha_0)} - \sum_{j=1}^{k}\Big[\ln \frac{\Gamma(\tilde{\alpha}_j)}{\Gamma(\alpha_j)}$ $- (\tilde{\alpha}_j - \alpha_j)\left(\psi(\tilde{\alpha}_j) - \psi(\tilde{\alpha}_0)\right)\Big]$

Appendix B

Useful results from matrix theory

B.1 Schur complements and inverting partitioned matrices

In chapter 5 on Linear Dynamical Systems, we needed to obtain the cross-covariance of states across two time steps from the precision matrix, calculated from combining the forward and backward passes over the sequences. This precision is based on the joint distribution of the states, yet we are interested only in the cross-covariance between states. If A is of 2×2 block form, we can use Schur complements to obtain the following results for the partitioned inverse of A, and its determinant in terms of its blocks' constituents.

The partitioned inverse is given by

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} F_{11}^{-1} & -A_{11}^{-1}A_{12}F_{22}^{-1} \\ -F_{22}^{-1}A_{21}A_{11}^{-1} & F_{22}^{-1} \end{pmatrix} \tag{B.1}$$

$$= \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}F_{22}^{-1}A_{21}A_{11}^{-1} & -F_{11}^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}F_{11}^{-1} & A_{22}^{-1} + A_{22}^{-1}A_{21}F_{11}^{-1}A_{12}A_{22}^{-1} \end{pmatrix} \tag{B.2}$$

and the determinant by

$$\begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = |A_{22}| \cdot |F_{11}| = |A_{11}| \cdot |F_{22}| \,, \tag{B.3}$$

where

$$F_{11} = A_{11} - A_{12}A_{22}^{-1}A_{21} \tag{B.4}$$

$$F_{22} = A_{22} - A_{21}A_{11}^{-1}A_{12} \,. \tag{B.5}$$

Notice that inverses of A_{12} or A_{21} do not appear in these results. There are other Schur complements that are defined in terms of the inverses of these 'off-diagonal' terms, but they are not needed for our purposes, and indeed if the states involved have different dimensionalities or are independent, then these off-diagonal quantities are not invertible.

B.2 The matrix inversion lemma

Here we present a sketch proof of the matrix inversion lemma, included for reference only. In the derivation that follows, it becomes quite clear that there is no obvious way of carrying the sort of expectations encountered in chapter 5 through the matrix inversion process (see comments following equation (5.105)).

The matrix inversion result is most useful when A is a large diagonal matrix and B has few columns (equivalently D has few rows).

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}\,. \tag{B.6}$$

To derive this lemma we use the Taylor series expansion of the matrix inverse

$$(A+M)^{-1} = A^{-1}(I+MA^{-1})^{-1} = A^{-1}\sum_{i=0}^{\infty}(-1)^i(MA^{-1})^i\,, \tag{B.7}$$

where the series is only well-defined when the spectral radius of MA^{-1} is less than unity. We can easily check that this series is indeed the inverse by directly multiplying by $(A+M)$, yielding the identity,

$$\begin{aligned}(A+M)A^{-1}\sum_{i=0}^{\infty}(-1)^i(MA^{-1})^i = {} & AA^{-1}\left[I - MA^{-1} + (MA^{-1})^2 - (MA^{-1})^3 + \ldots\right] \\ & + MA^{-1}\left[\quad I \quad - \quad MA^{-1} \quad + (MA^{-1})^2 - \ldots\right] \qquad \text{(B.8)} \\ = {} & I\,. \qquad \text{(B.9)}\end{aligned}$$

In the series expansion we find an embedded expansion, which forms the inverse matrix term on the right hand side, as follows

$$
\begin{aligned}
(A+BCD)^{-1} &= A^{-1}(I+BCDA^{-1})^{-1} && \text{(B.10)}\\
&= A^{-1}\sum_{i=0}^{\infty}(-1)^i(BCDA^{-1})^i && \text{(B.11)}\\
&= A^{-1}\left(I+\sum_{i=1}^{\infty}(-1)^i(BCDA^{-1})^i\right) && \text{(B.12)}\\
&= A^{-1}\left(I-BC\left[\sum_{i=0}^{\infty}(-1)^i(DA^{-1}BC)^i\right]DA^{-1}\right) && \text{(B.13)}\\
&= A^{-1}\left(I-BC(I+DA^{-1}BC)^{-1}DA^{-1}\right) && \text{(B.14)}\\
&= A^{-1}-A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}\,. && \text{(B.15)}
\end{aligned}
$$

In the above equations, we assume that the spectral radii of $BCDA^{-1}$ (B.11) and $DA^{-1}BC$ (B.13) are less than one for the Taylor series to be convergent. Aside from these constraints, we can post-hoc check the result simply by showing that multiplication of the expression by its proposed inverse does in fact yield the identity.

Appendix C

Miscellaneous results

C.1 Computing the digamma function

The digamma function is defined as

$$\psi(x) = \frac{d}{dx} \ln \Gamma(x) \, , \tag{C.1}$$

where $\Gamma(x)$ is the Gamma function given by

$$\Gamma(x) = \int_0^\infty d\tau \; \tau^{x-1} e^{-\tau} \, . \tag{C.2}$$

In the implementations of the models discussed in this thesis, the following expansion is used to compute the $\psi(x)$ for large positive arguments

$$\psi(x) \simeq \ln x - \frac{1}{2x} - \frac{1}{12x^2} + \frac{1}{120x^4} - \frac{1}{252x^6} + \frac{1}{240x^8} + \dots \, . \tag{C.3}$$

If we have small arguments, then we would expect this expansion to be inaccurate if we only used a finite number of terms. However, we can make use of a recursion of the digamma function to ensure that we always pass this expansion large arguments. The Gamma function has the well known recursion:

$$x! = \Gamma(x+1) = x\Gamma(x) = x(x-1)! \, , \tag{C.4}$$

from which the recursion for the digamma function readily follows:

$$\psi(x+1) = \frac{1}{x} + \psi(x) \, . \tag{C.5}$$

In our experiments we used an expansion (C.3) containing terms as far as $\mathcal{O}(1/x^{14})$, and used the recursion to evaluate this only for arguments of $\psi(x)$ greater than 6. This is more than enough precision.

C.2 Multivariate gamma hyperparameter optimisation

In hierarchical models such as the VB LDS model of chapter 5, there is often a gamma hyperprior over the noise precisions on each dimension of the data. On taking derivatives of the lower bound with respect to the shape a and inverse scale b of this hyperprior distribution, we obtain fixed point equations of this form:

$$\psi(a) = \ln b + \frac{1}{p}\sum_{s=1}^{p}\overline{\ln \rho_s}\,, \qquad \frac{1}{b} = \frac{1}{pa}\sum_{s=1}^{p}\overline{\rho_s} \tag{C.6}$$

where the notation $\overline{\ln \rho_s}$ and $\overline{\rho_s}$ is used to denote the expectations of quantities under the variational posterior distribution (see section 5.3.6 for details). We can rewrite this as:

$$\psi(a) = \ln b + c\,, \qquad \frac{1}{b} = \frac{d}{a}\,, \tag{C.7}$$

where

$$c = \frac{1}{p}\sum_{s=1}^{p}\overline{\ln \rho_s}\,, \qquad \text{and} \quad d = \frac{1}{p}\sum_{s=1}^{p}\overline{\rho_s}\,. \tag{C.8}$$

Equation (C.7) is the generic fixed point equation commonly arrived at when finding the variational parameters a and b which minimise the KL divergence on a gamma distribution.

The fixed point for a is found at the solution of

$$\psi(a) = \ln a - \ln d + c\,, \tag{C.9}$$

which can be arrived at using the Newton-Raphson iterations:

$$a_{\text{new}} \leftarrow a\left[1 - \frac{\psi(a) - \ln a + \ln d - c}{a\psi'(a) - 1}\right]\,, \tag{C.10}$$

where $\psi'(x)$ is the first derivative of the digamma function. Unfortunately, this update cannot ensure that a remains positive for the next iteration (the gamma distribution is only defined for $a > 0$) because the gradient information is taken locally.

There are two immediate ways to solve this. First if a should become negative during the Newton-Raphson iterations, reset it to a minimum value. This is a fairly crude solution. Alter-

natively, we can solve a different fixed point equation for a' where $a = \exp(a')$, resulting in the multiplicative updates:

$$a_{\text{new}} \leftarrow a \exp\left[-\frac{\psi(a) - \ln a + \ln d - c}{a\psi'(a) - 1}\right] . \tag{C.11}$$

This update has the same fixed point but exhibits different (well-behaved) dynamics to reach it. Note that equation C.10 is simply the first two terms in the Taylor series of the exponential function in the above equation.

Once the fixed point a^* is reached, the corresponding b^* is found simply from

$$b^* = \frac{a^*}{d} . \tag{C.12}$$

C.3 Marginal KL divergence of gamma-Gaussian variables

This note is intended to aid the reader in computing the lower bound appearing in equation (5.147) for variational Bayesian state-space models. Terms such as the KL divergence between two Gaussian or two gamma distributions are straightforward to compute and are given in appendix A. However there are more complicated terms involving expectations of KL divergences for joint Gaussian and gamma variables, for which we give results here.

Suppose we have two variables of interest, $\mathbf{a}$ and $\mathbf{b}$, that are jointly Gaussian distributed. To be more precise let the two variables be linearly dependent on each other in this sense:

$$q(\mathbf{a}, \mathbf{b}) = q(\mathbf{b})q(\mathbf{a} \,|\, \mathbf{b}) = \mathrm{N}(\mathbf{b} \,|\, \boldsymbol{\mu}_b, \Sigma_b) \cdot \mathrm{N}(\mathbf{a} \,|\, \boldsymbol{\mu}_a, \Sigma_a) \tag{C.13}$$

$$\text{where} \quad \boldsymbol{\mu}_a = \mathbf{y} - G\mathbf{b} . \tag{C.14}$$

Let us also introduce a prior distribution $p(\mathbf{a} \,|\, \mathbf{b})$ in this way:

$$p(\mathbf{a} \,|\, \mathbf{b}) = \mathrm{N}(\mathbf{a} \,|\, \tilde{\boldsymbol{\mu}}_a, \tilde{\Sigma}_a) \tag{C.15}$$

where neither parameter $\tilde{\boldsymbol{\mu}}_a$ nor $\tilde{\Sigma}_a$ are functions of b.

The first result is the KL divergence between two Gaussian distributions (given in appendix A)

$$\mathrm{KL}\left[q(\mathbf{a} \,|\, \mathbf{b}) \,\|\, p(\mathbf{a} \,|\, \mathbf{b})\right] = \int d\mathbf{a}\; q(\mathbf{a} \,|\, \mathbf{b}) \ln \frac{q(\mathbf{a} \,|\, \mathbf{b})}{p(\mathbf{a} \,|\, \mathbf{b})} \tag{C.16}$$

$$= -\frac{1}{2} \ln \left|\tilde{\Sigma}_a^{-1}\Sigma_a\right| + \frac{1}{2} \mathrm{tr}\, \tilde{\Sigma}_a^{-1} \left[\Sigma_a - \tilde{\Sigma}_a + (\boldsymbol{\mu}_a - \tilde{\boldsymbol{\mu}}_a)(\boldsymbol{\mu}_a - \tilde{\boldsymbol{\mu}}_a)^\top\right] . \tag{C.17}$$

Note that this divergence is written w.r.t. the $q(\mathbf{a} \,|\, \mathbf{b})$ distribution. The dependence on $\mathbf{b}$ is not important here, but will be required later. The important part to note is that it obviously depends on each Gaussian's covariance, but also on the Mahalanobis distance between the means as measured w.r.t. the non-averaging distribution.

Consider now the KL divergence between the full joint posterior and full joint prior:

$$\mathrm{KL}\left[q(\mathbf{a}, \mathbf{b}) \,\|\, p(\mathbf{a}, \mathbf{b})\right] = \int d\mathbf{a}\, d\mathbf{b}\; q(\mathbf{a}, \mathbf{b}) \ln \frac{q(\mathbf{a}, \mathbf{b})}{p(\mathbf{a}, \mathbf{b})} \tag{C.18}$$

$$= \int d\mathbf{b}\; q(\mathbf{b}) \int d\mathbf{a}\; q(\mathbf{a} \,|\, \mathbf{b}) \ln \frac{q(\mathbf{a} \,|\, \mathbf{b})}{p(\mathbf{a} \,|\, \mathbf{b})} + \int d\mathbf{b}\; q(\mathbf{b}) \ln \frac{q(\mathbf{b})}{p(\mathbf{b})} \;. \tag{C.19}$$

The last term in this is equation is simply the KL divergence between two Gaussians, which is straightforward, but the first term is the *expected* KL divergence between the conditional distributions, where the expectation is taken w.r.t. the marginal distribution $q(\mathbf{b})$. After some simple manipulation, this first term is given by

$$\left\langle \mathrm{KL}\left[q(\mathbf{a} \,|\, \mathbf{b}) \,\|\, p(\mathbf{a} \,|\, \mathbf{b})\right] \right\rangle_{q(\mathbf{b})} = \int d\mathbf{b}\; q(\mathbf{b}) \int d\mathbf{a}\; q(\mathbf{a} \,|\, \mathbf{b}) \ln \frac{q(\mathbf{a} \,|\, \mathbf{b})}{p(\mathbf{a} \,|\, \mathbf{b})} \tag{C.20}$$

$$= -\frac{1}{2} \ln \left| \tilde{\Sigma}_a^{-1} \Sigma_a \right| + \frac{1}{2} \mathrm{tr}\; \tilde{\Sigma}_a^{-1} \Big[\Sigma_a - \tilde{\Sigma_a} + G \Sigma_b G^\top + (\mathbf{y} - G\boldsymbol{\mu}_b - \tilde{\boldsymbol{\mu}}_a)(\mathbf{y} - G\boldsymbol{\mu}_b - \tilde{\boldsymbol{\mu}}_a)^\top \Big] \;. \tag{C.21}$$

Let us now suppose that the covariance terms for the prior $\tilde{\Sigma}$ and posterior Σ_a have the same multiplicative dependence on another variable ρ^{-1}. This is the case in the variational state-space model of chapter 5 where, for example, the uncertainty in the entries for the output matrix C should be related to the setting of the output noise $\boldsymbol{\rho}$ (see equation (5.44) for example). In equation (C.17) it is clear that if both covariances are dependent on the same ρ^{-1}, then the KL divergence will not be a function of ρ^{-1} *provided* that the means of both distributions are the same. If they are different however, then there is a residual dependence on ρ^{-1} due to the $\tilde{\Sigma}_a^{-1}$ term from the non-averaging distribution $p(\mathbf{a} \,|\, \mathbf{b})$. This is important as there will usually be distributions over this ρ variable of the form

$$q(\rho) = \mathrm{Ga}(\rho \,|\, e_\rho, f_\rho) \tag{C.22}$$

with e and f shape and precision parameters of a gamma distribution. The most complicated term to compute is the penultimate term in (5.147), which is

$$\Big\langle \left\langle \mathrm{KL}\left[q(\mathbf{a} \,|\, \mathbf{b}, \rho) \,\|\, p(\mathbf{a} \,|\, \mathbf{b}, \rho)\right] \right\rangle_{q(\mathbf{b})} \Big\rangle_{q(\rho)} = \int d\rho\; q(\rho) \int d\mathbf{b}\; q(\mathbf{b} \,|\, \rho) \int d\mathbf{a}\; q(\mathbf{a} \,|\, \mathbf{b}, \rho) \ln \frac{q(\mathbf{a} \,|\, \mathbf{b}, \rho)}{p(\mathbf{a} \,|\, \mathbf{b}, \rho)} \;. \tag{C.23}$$

In the variational Bayesian state-space model, the prior and posterior for the parameters of the output matrix C (and D for that matter) are defined in terms of the same noise precision variable

ρ. This means that all terms but the last one in equation (C.21) are not functions of ρ and pass through the expectation in (C.23) untouched. The final term has a dependence on ρ, but on taking expectations w.r.t. $q(\rho)$ this simply yields a multiplicative factor of $\langle\rho\rangle_{q(\rho)}$. It is straightforward to extend this to the case of data with several dimensions, in which case the lower bound is a sum over all p dimensions of similar quantities.