Extra Material for *Fast Variational Bayesian Linear State-Space Model*

Jaakko Luttinen

July 4, 2013

Abstract

This document contains gradients for the rotation speed-ups presented in the paper *Fast Variational Bayesian Linear State-Space Model* (Luttinen, ECML 2013).

1 Useful matrix formulas

$$
\begin{aligned}
d\mathbf{R}^{-1} &= -\mathbf{R}^{-1}(d\mathbf{R})\mathbf{R}^{-1} \\
d\log|\mathbf{R}| &= \mathrm{tr}(\mathbf{R}^{-1}d\mathbf{R}) \\
d\,\mathrm{tr}(\mathbf{AR}) &= \mathrm{tr}(\mathbf{A}d\mathbf{R}) \\
d\,\mathrm{tr}(\mathbf{ARBR}^{\mathrm{T}}) &= \mathrm{tr}\left((\mathbf{ARB}+\mathbf{A}^{\mathrm{T}}\mathbf{RB}^{\mathrm{T}})d\mathbf{R}\right)
\end{aligned}
$$

2 Rotations

2.1 Simple Gaussian

Prior:

$$
p(\mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n|\mathbf{0}, \mathbf{\Lambda}^{-1}\right)
$$

Transformed posterior:

$$
\begin{aligned}
q(\mathbf{X}) &= \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n|\mathbf{R}\boldsymbol{\mu}_n, \mathbf{R}\mathbf{\Sigma}_n\mathbf{R}^{\mathrm{T}}\right) \\
\langle\mathbf{X}\rangle_R &= \mathbf{R}\langle\mathbf{X}\rangle \\
\langle\mathbf{X}\mathbf{X}^{\mathrm{T}}\rangle_R &= \mathbf{R}\langle\mathbf{X}\mathbf{X}^{\mathrm{T}}\rangle\mathbf{R}^{\mathrm{T}}
\end{aligned}
$$

Lower bound terms:

$$\langle \log q(\mathbf{X}) \rangle_R = -N \log |\mathbf{R}| + \text{const}$$
$$\langle \log p(\mathbf{X}) \rangle_R = -\frac{1}{2} \operatorname{tr} \left(\langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_R \mathbf{\Lambda} \right) + \text{const}$$

Gradient terms:

$$\frac{\partial}{\partial \mathbf{R}} \log |\mathbf{R}| = \mathbf{R}^{-\mathrm{T}}$$
$$\frac{\partial}{\partial \mathbf{R}} \operatorname{tr} \left(\langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_R \mathbf{\Lambda} \right) = 2\mathbf{\Lambda}\mathbf{R}\langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle$$

2.2 Gaussian with ARD prior

Prior:

$$p(\mathbf{X}|\boldsymbol{\alpha}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n | \mathbf{0}, \operatorname{diag}(\boldsymbol{\alpha})^{-1}\right)$$
$$p(\boldsymbol{\alpha}) = \prod_{d=1}^{D} \mathcal{G}(\alpha_d | a_0, b_0)$$

Transformed posteriors:

$$q(\mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n | \mathbf{R}\boldsymbol{\mu}_n, \mathbf{R}\mathbf{\Sigma}_n\mathbf{R}^{\mathrm{T}}\right)$$
$$\langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_R = \mathbf{R} \sum_{n=1}^{N} (\boldsymbol{\mu}_n \boldsymbol{\mu}_n^{\mathrm{T}} + \mathbf{\Sigma}_n)\mathbf{R}^{\mathrm{T}}$$

$$q(\boldsymbol{\alpha}) = \prod_{d=1}^{D} \mathcal{G}(\alpha_d | a, b_d)$$
$$a = a_0 + \frac{N}{2}$$
$$b_d = b_0 + \frac{1}{2} \left[\langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_R \right]_{dd}$$
$$\langle \alpha_d \rangle_R = \frac{a}{b_d}$$
$$\langle \log \alpha_d \rangle_R = \psi(a) - \log(b_d)$$

Lower bound terms:

$$\langle \log q(\mathbf{X}) \rangle_R = -N \log |\mathbf{R}| + \text{const}$$

$$\langle \log p(\mathbf{X}|\boldsymbol{\alpha}) \rangle_R = -\frac{1}{2} \operatorname{tr}\left(\langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_R \langle \operatorname{diag}(\boldsymbol{\alpha}) \rangle_R \right) + \frac{N}{2} \sum_{d=1}^{D} \langle \log \alpha_d \rangle_R + \text{const}$$

$$\langle \log q(\boldsymbol{\alpha}) \rangle_R = \sum_{d=1}^{D} \langle \log \alpha_d \rangle_R + \text{const}$$

$$\begin{aligned} \langle \log p(\boldsymbol{\alpha}) \rangle_R &= \sum_{d=1}^{D} \langle \mathcal{G}(\alpha_d | a_0, b_0) \rangle_R \\ &= (a_0 - 1) \sum_{d=1}^{D} \langle \log \alpha_d \rangle_R - b_0 \sum_{d=1}^{D} \langle \alpha_d \rangle_d + \text{const} \end{aligned}$$

Gradient terms:

$$\frac{\partial}{\partial \mathbf{R}} \log |\mathbf{R}| = \mathbf{R}^{-\mathrm{T}}$$

$$\frac{\partial}{\partial \mathbf{R}} \sum_{d=1}^{D} \langle \alpha_d \rangle_R = -\begin{bmatrix} \frac{\langle \alpha_1 \rangle_R}{b_1} & & \\ & \ddots & \\ & & \frac{\langle \alpha_D \rangle_R}{b_D} \end{bmatrix} \mathbf{R} \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle$$

$$\frac{\partial}{\partial \mathbf{R}} \sum_{d=1}^{D} \langle \log \alpha_d \rangle_R = -\begin{bmatrix} \frac{1}{b_1} & & \\ & \ddots & \\ & & \frac{1}{b_D} \end{bmatrix} \mathbf{R} \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle$$

$$\begin{aligned} \frac{\partial}{\partial \mathbf{R}} \operatorname{tr}\left(\langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_R \langle \operatorname{diag}(\boldsymbol{\alpha}) \rangle_R \right) &= 2 \begin{bmatrix} \langle \alpha_1 \rangle_R & & \\ & \ddots & \\ & & \langle \alpha_D \rangle_R \end{bmatrix} \mathbf{R} \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle \\ &\quad - \begin{bmatrix} \frac{\langle \alpha_1 \rangle_R}{b_1} [\mathbf{R}]_{1:} \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle [\mathbf{R}]_{1:}^{\mathrm{T}} & & \\ & \ddots & \\ & & \frac{\langle \alpha_D \rangle_R}{b_D} [\mathbf{R}]_{D:} \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle [\mathbf{R}]_{D:}^{\mathrm{T}} \end{bmatrix} \mathbf{R} \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle \end{aligned}$$

2.3 Gaussian Markov chain

Prior:

$$\begin{aligned} p(\mathbf{x}_0) &= \mathcal{N}(\mathbf{x}_0 | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \\ p(\mathbf{x}_n | \mathbf{x}_{n-1}, \mathbf{A}) &= \mathcal{N}(\mathbf{x}_n | \mathbf{A}\mathbf{x}_{n-1}, \mathbf{I}) \\ &\quad \text{some gaussian model for } \mathbf{A} \end{aligned}$$

Transformed posteriors:

$$q_R(\mathbf{X}) = \mathcal{N}\left([\mathbf{X}]_{:} \,\middle|\, (\mathbf{I}\otimes\mathbf{R})\boldsymbol{\mu}, (\mathbf{I}\otimes\mathbf{R})\boldsymbol{\Sigma}(\mathbf{I}\otimes\mathbf{R})^{\mathrm{T}}\right)$$
$$q_R(\mathbf{A}) = \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{A}_{d:} \,\middle|\, \mathbf{R}_{d:}\mathbf{A}\mathbf{R}^{-1}, \left(\sum_{k=1}^{D}\mathbf{R}_{kd}\right)^2 \mathbf{R}^{-\mathrm{T}}\boldsymbol{\Sigma}_A^{(d)}\mathbf{R}^{-1}\right)$$

Lowerbound terms:

$$\langle \log q(\mathbf{X})\rangle_R = -(N+1)\log|\mathbf{R}| + \text{const}$$
$$\langle \log p(\mathbf{X}|\mathbf{A})\rangle_R = \mathrm{tr}\left(-\frac{1}{2}\boldsymbol{\Lambda}\langle \mathbf{x}_0\mathbf{x}_0^{\mathrm{T}}\rangle_R + \boldsymbol{\Lambda}\boldsymbol{\mu}\langle \mathbf{x}_0\rangle_R^{\mathrm{T}} + \sum_{n=1}^{N}\Big[-\frac{1}{2}\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle_R - \frac{1}{2}\langle \mathbf{A}^{\mathrm{T}}\mathbf{A}\rangle_R\langle \mathbf{x}_{n-1}\mathbf{x}_{n-1}^{\mathrm{T}}\rangle_R + \langle \mathbf{A}\rangle_R\langle \mathbf{x}_{n-1}\mathbf{x}_n^{\mathrm{T}}\rangle_R\Big]\right)$$

Relevant moments:

$$\mathbf{c}^{\mathrm{T}}\langle \mathbf{x}_n\rangle_R = \mathbf{c}^{\mathrm{T}}\mathbf{R}\langle \mathbf{x}_n\rangle$$
$$\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle_R = \mathbf{R}\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle\mathbf{R}^{\mathrm{T}}$$
$$\mathrm{tr}\left(\langle \mathbf{A}\rangle_R\langle \mathbf{x}_{n-1}\mathbf{x}_n^{\mathrm{T}}\rangle_R\right) = \mathrm{tr}\left(\mathbf{R}^{\mathrm{T}}\mathbf{R}\langle \mathbf{A}\rangle\langle \mathbf{x}_{n-1}\mathbf{x}_n^{\mathrm{T}}\rangle\right)$$
$$\mathrm{tr}\left(\langle \mathbf{A}^{\mathrm{T}}\mathbf{A}\rangle_R\langle \mathbf{x}_{n-1}\mathbf{x}_{n-1}^{\mathrm{T}}\rangle_R\right) = \mathrm{tr}\left(\langle \mathbf{A}\rangle^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{R}\langle \mathbf{A}\rangle\langle \mathbf{x}_{n-1}\mathbf{x}_{n-1}^{\mathrm{T}}\rangle\right) + \sum_{d=1}^{D}\left(\sum_{k=1}^{D}\mathbf{R}_{kd}\right)^2 \mathrm{tr}\left(\boldsymbol{\Sigma}_A^{(d)}\langle \mathbf{x}_{n-1}\mathbf{x}_{n-1}^{\mathrm{T}}\rangle\right)$$

Gradient terms:

$$\frac{\partial}{\partial\mathbf{R}}\log|\mathbf{R}| = \mathbf{R}^{-\mathrm{T}}$$
$$\frac{\partial}{\partial\mathbf{R}}\mathbf{c}^{\mathrm{T}}\langle \mathbf{x}_n\rangle_R = \mathbf{c}\langle \mathbf{x}_n\rangle^{\mathrm{T}}$$
$$\frac{\partial}{\partial\mathbf{R}}\mathrm{tr}\left(\mathbf{C}\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle_R\right) = 2\mathbf{C}\mathbf{R}\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle$$
$$\frac{\partial}{\partial\mathbf{R}}\mathrm{tr}\left(\langle \mathbf{A}\rangle_R\langle \mathbf{x}_{n-1}\mathbf{x}_n^{\mathrm{T}}\rangle_R\right) = \mathbf{R}\left(\langle \mathbf{x}_{n-1}\mathbf{x}_n^{\mathrm{T}}\rangle^{\mathrm{T}}\langle \mathbf{A}\rangle^{\mathrm{T}} + \langle \mathbf{A}\rangle\langle \mathbf{x}_{n-1}\mathbf{x}_n^{\mathrm{T}}\rangle\right)$$
$$\frac{\partial}{\partial\mathbf{R}}\mathrm{tr}\left(\langle \mathbf{A}^{\mathrm{T}}\mathbf{A}\rangle_R\langle \mathbf{x}_{n-1}\mathbf{x}_{n-1}^{\mathrm{T}}\rangle_R\right) = 2\mathbf{R}\langle \mathbf{A}\rangle\langle \mathbf{x}_{n-1}\mathbf{x}_{n-1}^{\mathrm{T}}\rangle\langle \mathbf{A}\rangle^{\mathrm{T}} + 2\begin{bmatrix}1 & \cdots & 1\\ \vdots & \ddots & \vdots\\ 1 & \cdots & 1\end{bmatrix}\mathbf{R}\begin{bmatrix}\mathrm{tr}\left(\boldsymbol{\Sigma}_A^{(1)}\langle \mathbf{x}_{n-1}\mathbf{x}_{n-1}^{\mathrm{T}}\rangle\right) & & \\ & \ddots & \\ & & \mathrm{tr}\left(\boldsymbol{\Sigma}_A^{(D)}\langle \mathbf{x}_{n-1}\mathbf{x}_{n-1}^{\mathrm{T}}\rangle\right)\end{bmatrix}$$

2.4 Left-right rotation of Gaussian with ARD prior

Prior:

$$p(\mathbf{X}|\boldsymbol{\alpha}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n|\mathbf{0}, \mathrm{diag}(\boldsymbol{\alpha})^{-1}\right)$$
$$p(\boldsymbol{\alpha}) = \prod_{d=1}^{D} \mathcal{G}(\alpha_d|a_0, b_0)$$

Transformed posteriors, inspired by (but not exactly) $\mathbf{RXQ}^{\mathrm{T}}$:

$$q_*(\mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n|\sum_{m=1}^{N} q_{nm}\mathbf{R}\boldsymbol{\mu}_n, \left(\sum_{m=1}^{N} q_{mn}\right)^2 \mathbf{R}\boldsymbol{\Sigma}_n\mathbf{R}^{\mathrm{T}}\right)$$
$$q_*(\boldsymbol{\alpha}) = \prod_{d=1}^{D} \mathcal{G}(\alpha_d|a, b_d)$$
$$a = a_0 + \frac{N}{2}$$
$$b_d = b_0 + \frac{1}{2}\left[\langle\mathbf{X}\mathbf{X}^{\mathrm{T}}\rangle_*\right]_{dd}$$

Lower bound terms:

$$\langle\log q(\mathbf{X})\rangle_* = -N\log|\mathbf{R}| - D\sum_{n=1}^{N}\log\left|\sum_{m=1}^{N} q_{mn}\right| + \mathrm{const}$$
$$\langle\log p(\mathbf{X}|\boldsymbol{\alpha})\rangle_* = -\frac{1}{2}\,\mathrm{tr}\left(\langle\mathbf{X}\mathbf{X}^{\mathrm{T}}\rangle_*\langle\mathrm{diag}(\boldsymbol{\alpha})\rangle_*\right) + \frac{N}{2}\sum_{d=1}^{D}\langle\log\alpha_d\rangle_* + \mathrm{const}$$
$$\langle\log q(\boldsymbol{\alpha})\rangle_* = \sum_{d=1}^{D}\langle\log\alpha_d\rangle_* + \mathrm{const}$$
$$\langle\log p(\boldsymbol{\alpha})\rangle_* = (a_0 - 1)\sum_{d=1}^{D}\langle\log\alpha_d\rangle_* - b_0\sum_{d=1}^{D}\langle\alpha_d\rangle_* + \mathrm{const}$$

Relevant moments:

$$\langle\alpha_d\rangle_* = \frac{a}{b_d}$$
$$\langle\log\alpha_d\rangle_* = \psi(a) - \log(b_d)$$
$$\langle\mathbf{X}\mathbf{X}^{\mathrm{T}}\rangle_Q = \langle\mathbf{X}\rangle\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\langle\mathbf{X}\rangle^{\mathrm{T}} + \sum_{n=1}^{N}\left(\sum_{m=1}^{N} q_{mn}\right)^2 \boldsymbol{\Sigma}_n$$
$$\langle\mathbf{X}\mathbf{X}^{\mathrm{T}}\rangle_* = \mathbf{R}\langle\mathbf{X}\mathbf{X}^{\mathrm{T}}\rangle_Q\mathbf{R}^{\mathrm{T}}$$

Gradient terms with respect to $\mathbf{R}$:

$$\frac{\partial}{\partial \mathbf{R}} \log |\mathbf{R}| = \mathbf{R}^{-\mathrm{T}}$$

$$\frac{\partial}{\partial \mathbf{R}} \sum_{d=1}^{D} \langle \alpha_d \rangle_* = -\begin{bmatrix} \frac{\langle \alpha_1 \rangle_*}{b_1} & & \\ & \ddots & \\ & & \frac{\langle \alpha_D \rangle_*}{b_D} \end{bmatrix} \mathbf{R} \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_Q$$

$$\frac{\partial}{\partial \mathbf{R}} \sum_{d=1}^{D} \langle \log \alpha_d \rangle_* = -\begin{bmatrix} \frac{1}{b_1} & & \\ & \ddots & \\ & & \frac{1}{b_D} \end{bmatrix} \mathbf{R} \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_Q$$

$$\frac{\partial}{\partial \mathbf{R}} \operatorname{tr}\left(\langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_* \langle \operatorname{diag}(\boldsymbol{\alpha}) \rangle_* \right) = 2 \begin{bmatrix} \langle \alpha_1 \rangle_* & & \\ & \ddots & \\ & & \langle \alpha_D \rangle_* \end{bmatrix} \mathbf{R} \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_Q - \begin{bmatrix} \frac{\langle \alpha_1 \rangle_*}{b_1} [\mathbf{R}]_{1:} \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle [\mathbf{R}]_{1:}^{\mathrm{T}} & & \\ & \ddots & \\ & & \frac{\langle \alpha_D \rangle_*}{b_D} [\mathbf{R}]_{D:} \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle [\mathbf{R}]_{D:}^{\mathrm{T}} \end{bmatrix} \mathbf{R} \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_Q$$

Let us define a helpful function:

$$\begin{aligned} \boldsymbol{\Psi}(\mathbf{v}) &= \frac{\partial}{\partial \mathbf{Q}} \frac{1}{2} \operatorname{tr}\left(\operatorname{diag}(\mathbf{v}) \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_* \right) \\ &= \mathbf{Q} \langle \mathbf{X} \rangle^{\mathrm{T}} \mathbf{R}^{\mathrm{T}} \operatorname{diag}(\mathbf{v}) \mathbf{R} \langle \mathbf{X} \rangle \\ &\quad + \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \mathbf{Q} \begin{bmatrix} \operatorname{tr}\left(\mathbf{R}^{\mathrm{T}} \operatorname{diag}(\mathbf{v}) \mathbf{R} \boldsymbol{\Sigma}_1 \right) & & \\ & \ddots & \\ & & \operatorname{tr}\left(\mathbf{R}^{\mathrm{T}} \operatorname{diag}(\mathbf{v}) \mathbf{R} \boldsymbol{\Sigma}_N \right) \end{bmatrix} \end{aligned}$$

Gradient terms with respect to $\mathbf{Q}$:

$$\frac{\partial}{\partial \mathbf{Q}} \sum_{n=1}^{N} \log \left| \sum_{m=1}^{N} q_{mn} \right| = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} \left(\sum_{m=1}^{N} q_{m1} \right)^{-1} & \cdots & \left(\sum_{m=1}^{N} q_{mN} \right)^{-1} \end{bmatrix}$$

$$\frac{\partial}{\partial \mathbf{Q}} \sum_{d=1}^{D} \langle \alpha_d \rangle_* = -\boldsymbol{\Psi}\left(\frac{\langle \boldsymbol{\alpha} \rangle_*}{\mathbf{b}} \right)$$

$$\frac{\partial}{\partial \mathbf{Q}} \sum_{d=1}^{D} \langle \log \alpha_d \rangle_* = -\boldsymbol{\Psi}\left(\frac{1}{\mathbf{b}} \right)$$

$$\frac{\partial}{\partial \mathbf{Q}} \operatorname{tr}\left(\langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_* \langle \operatorname{diag}(\boldsymbol{\alpha}) \rangle_* \right) = 2\boldsymbol{\Psi}\left(\langle \boldsymbol{\alpha} \rangle_* \right) - \boldsymbol{\Psi}\left(\operatorname{diag} \langle \mathbf{X}\mathbf{X}^{\mathrm{T}} \rangle_* \circ \frac{\langle \boldsymbol{\alpha} \rangle_*}{\mathbf{b}} \right)$$