Derivatives of Gaussian KL-Divergence for some parameterizations of the posterior covariance for variational Gaussian-process inference

M. Rule

March 25, 2020

These notes provide the derivatives of the KL-divergence $D_{KL}[Q(\mathbf{z})\|P(\mathbf{z})]$ between two multivariate Gaussian distributions $Q(\mathbf{z})$ and $P(\mathbf{z})$ with respect to a few parameterizations θ of the covariance matrix $\Sigma(\theta)$ of Q. This is useful for variational Gaussian process inference, where clever parameterizations of the posterior covariance are required to make the problem tractable. Tables for differentiating matrix-valued functions can be found in The Matrix Cookbook.

Consider two multivariate Gaussian distributions $Q(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_q, \Sigma(\theta))$ and $P(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0 = \boldsymbol{\Lambda}^{-1})$ with dimension L. The KL divergence $D_{\text{KL}}[Q(\mathbf{z})\|P(\mathbf{z})]$ has the closed form

$$
\begin{aligned}
\mathcal{D} &:= D_{\text{KL}}[Q(\mathbf{z})\| \Pr(\mathbf{z})] \\
&= \tfrac{1}{2}\Big\{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_q)^\top \boldsymbol{\Lambda} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_q) \\
&\quad + \operatorname{tr}(\boldsymbol{\Lambda}\Sigma) - \ln|\Sigma| - \ln|\boldsymbol{\Lambda}|\Big\} + \text{constant}.
\end{aligned} \tag{1}
$$

In variational Bayesian inference, we minimize $\mathcal{D}$ while maximizing the expected log-probability of some observations with respect to $Q(\mathbf{z})$. Closed-form derivatives of $\mathcal{D}$ in terms of the parameters of Q are useful for manually optimizing code for larger problems. The derivatives of $\mathcal{D}$ in terms of $\boldsymbol{\mu}_q$ are straightforward: $\partial_{\boldsymbol{\mu}_q}\mathcal{D} = \boldsymbol{\Lambda}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_z)$ and $\text{H}_{\boldsymbol{\mu}_q}\mathcal{D} = \boldsymbol{\Lambda}$. In these notes, we explore derivatives of $\mathcal{D}$ with respect to a few different parameterizations ("θ") of $\Sigma(\theta)$.

We evaluate the following parameterizations for Σ: 1. Optimizing the full Σ directly 2. $\Sigma \approx \mathbf{X}\mathbf{X}^\top$ 3. $\Sigma \approx \mathbf{A}^\top \operatorname{diag}[\mathbf{v}]\mathbf{A}$ 4. $\Sigma \approx [\boldsymbol{\Lambda} + \operatorname{diag}[\mathbf{p}]]^{-1}$ 5. $\mathbf{F}^\top\mathbf{Q}\mathbf{Q}^\top\mathbf{F}$, where $\mathbf{Q} \in \mathbb{R}^{K\times K}$, $K<L$ and $\mathbf{F} \in \mathbb{R}^{K\times L}$, $\mathbf{F}\mathbf{F}^\top = \mathbf{I}$.

0.1 Σ

We first obtain gradients of $\mathcal{D}$ in Σ (assuming Σ is full-rank). These can be used to derive gradients in θ for some parameterizations $\Sigma(\theta)$ using the chain rule. The gradient of $\mathcal{D}$ in Σ can be obtained using identities (57) and (100) in The Matrix Cookbook:

$$
\begin{aligned}
\partial_\Sigma \mathcal{D} &= \partial_\Sigma \{\operatorname{tr}(\boldsymbol{\Lambda}\Sigma) - \ln|\Sigma|\} \\
&= \tfrac{1}{2}\left(\boldsymbol{\Lambda} - \Sigma^{-1}\right).
\end{aligned} \tag{2}
$$

The Hessian in Σ is a fourth-order tensor. It's simpler to express the Hessian in terms of a Hessian-vector product, which can be used with Krylov subspace solvers to efficiently compute the update in Newton's method. Considering an $L\times L$ matrix $\mathbf{M}$, the Hessian-vector product is given by

$$[\mathbf{H}_\Sigma \mathcal{D}]\,\mathbf{M} = \partial_\Sigma \langle \partial_\Sigma \mathcal{D}, \mathbf{M}\rangle = \partial_\Sigma \operatorname{tr}\left[(\partial_\Sigma \mathcal{D})^\top \mathbf{M}\right], \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar (Frobenius) product. This is given by identity (124) in the Matrix Cookbook:

$$\partial_\Sigma \operatorname{tr}\left[\frac{1}{2}\left(\boldsymbol{\Lambda} - \Sigma^{-1}\right)^\top \mathbf{M}\right] = -\frac{1}{2}\partial_\Sigma \operatorname{tr}\left[\Sigma^{-1}\mathbf{M}\right] = \frac{1}{2}\Sigma^{-1}\mathbf{M}^\top\Sigma^{-1}. \tag{4}$$

0.2 $\Sigma \approx \mathbf{X}\mathbf{X}^\top$

We consider an approximate posterior covariance of the form

$$\Sigma \approx \mathbf{X}\mathbf{X}^\top, \quad \mathbf{X} \in \mathbb{R}^{L\times K} \tag{5}$$

where $\mathbf{X}$ is a rank-$K < L$ matrix with L rows and K columns.

Since $\mathbf{X}$ is not full rank, the log-determinant $\ln|\Sigma| = \ln|\mathbf{X}\mathbf{X}^\top|$ in (1) diverges, due to the zero eigenvalues in the null space of $\mathbf{X}$. However, since this null-space is not being optimized, it does not affect our gradient. It is sufficient to replace the log-determinant with that of the reduced-rank representation, $\ln|\mathbf{X}^\top\mathbf{X}|$. Identity (55) in The Matrix Cookbook provides the derivative of this, $\partial_\mathbf{X} \ln|\mathbf{X}^\top\mathbf{X}| = 2\mathbf{X}^{+\top}$, where $(\cdot)^+$ is the pseudoinverse. Combined with identity (112), this gives the following gradient of $\mathcal{D}(\mathbf{X})$:

$$\partial_\mathbf{X}\mathcal{D} = \partial_\mathbf{X}\tfrac{1}{2}\left\{\operatorname{tr}\left[\boldsymbol{\Lambda}\mathbf{X}\mathbf{X}^\top\right] - \ln|\mathbf{X}^\top\mathbf{X}|\right\} = \boldsymbol{\Lambda}\mathbf{X} - \mathbf{X}^{+\top}. \tag{6}$$

The Hessian-vector product requires the derivative of $\partial_\mathbf{X} \operatorname{tr}[\mathbf{X}^+\mathbf{M}]$:

$$\partial_\mathbf{X}\langle \partial_\Sigma \mathcal{D}, \mathbf{M}\rangle = \partial_\mathbf{X} \operatorname{tr}\left[\left(\boldsymbol{\Lambda}\mathbf{X} - \mathbf{X}^{+\top}\right)^\top \mathbf{M}\right] = \partial_\mathbf{X} \operatorname{tr}[\boldsymbol{\Lambda}\mathbf{X}\mathbf{M}] - \partial_\mathbf{X} \operatorname{tr}\left[\mathbf{X}^+\mathbf{M}\right]. \tag{7}$$

Goulob and Pereya (1972) Eq. 4.12 gives the derivative of a fixed-rank pseudoinverse:

$$\partial\mathbf{X}^+ = -\mathbf{X}^+(\partial\mathbf{X})\mathbf{X}^+ + \mathbf{X}^+\mathbf{X}^{+\top}(\partial\mathbf{X})^\top(1 - \mathbf{X}\mathbf{X}^+) + (1 - \mathbf{X}^+\mathbf{X})(\partial\mathbf{X})^\top\mathbf{X}^{+\top}\mathbf{X}^+ \tag{8}$$

Since $\mathbf{X}$ is $N \times K$ with rank K, $\mathbf{X}^+\mathbf{X}$ is full-rank. Therefore $\mathbf{X}^+\mathbf{X} = \mathbf{I}_k$ and the final term in (**??**) vanishes. The derivative of the pseudoinverse can now be written as:

$$\partial\mathbf{X}^+ = -\mathbf{X}^+(\partial\mathbf{X})\mathbf{X}^+ + \mathbf{X}^+\mathbf{X}^{+\top}(\partial\mathbf{X})^\top(\mathbf{I}_n - \mathbf{X}\mathbf{X}^+) \tag{9}$$

Since the derivative of a trace of a matrix-valued function is just the (transpose) of the scalar derivative,

$$\begin{aligned}\partial_\mathbf{X} \operatorname{tr}\left[\mathbf{X}^+\mathbf{M}\right] &= \left\{-\mathbf{X}^+\mathbf{M}\mathbf{X}^+ + \mathbf{X}^+\mathbf{X}^{+\top}\mathbf{M}^\top(\mathbf{I}_n - \mathbf{X}\mathbf{X}^+)\right\}^\top \\ &= -\mathbf{X}^{+\top}\mathbf{M}^\top\mathbf{X}^{+\top} + (\mathbf{I} - \mathbf{X}^{+\top}\mathbf{X}^\top)\mathbf{M}\mathbf{X}^+\mathbf{X}^{+\top}.\end{aligned} \tag{10}$$

Overall, we obtain the following Hessian-vector product:

$$\partial_\mathbf{X}\langle \partial_\Sigma \mathcal{D}, \mathbf{M}\rangle = \boldsymbol{\Lambda}\mathbf{M}^\top + \mathbf{X}^{+\top}\mathbf{M}^\top\mathbf{X}^{+\top} - (\mathbf{I} - \mathbf{X}^{+\top}\mathbf{X}^\top)\mathbf{M}\mathbf{X}^+\mathbf{X}^{+\top} \tag{11}$$

0.2.1 Σ=$\mathbf{X}\mathbf{X}^\top$ when X is full-rank

Equations (6) and (11) are also valid if $\mathbf{X}$ is a rank-L triangular (Choleskey) factorization of Σ. In this case the pseudoinverse can be replaced by the full inverse, and various terms simplify:

$$
\begin{aligned}
\partial_{\mathrm{X}} \mathcal{D} &= \mathbf{\Lambda X} - \mathbf{X}^{-\top} \\
\partial_{\mathrm{X}} \langle \partial_{\mathrm{x}} \mathcal{D}, \mathbf{M} \rangle &= \mathbf{\Lambda M}^\top + \mathbf{X}^{-\top} \mathbf{M}^\top \mathbf{X}^{-\top}
\end{aligned}
\tag{12}
$$

0.3 $\Sigma = \mathbf{A}^\top \operatorname{diag}[\mathbf{v}]\mathbf{A}$

Let $\Sigma = \mathbf{A}^\top \operatorname{diag}[\mathbf{v}]\mathbf{A}$, where $\mathbf{A}$ is fixed and $\mathbf{v} \in \mathbb{R}^L$ are free parameters. Define $\operatorname{diag}[\cdot]$ as an operator that constructs a diagonal matrix from a vector, or extracts the main diagonal from a matrix if its argument is a matrix. The gradient of $\mathcal{D}$ in $\mathbf{v}$ is:

$$
\begin{aligned}
\partial_{\mathrm{X}} \mathcal{D} &= \partial_{\mathrm{X}} \tfrac{1}{2} \left\{ \operatorname{tr}\left[\mathbf{\Lambda A}^\top \operatorname{diag}[\mathbf{v}]\mathbf{A} \right] - \ln |\mathbf{A}^\top \operatorname{diag}[\mathbf{v}]\mathbf{A}| \right\} \\
&= \tfrac{1}{2} \left\{ \operatorname{diag}[\mathbf{A\Lambda A}^\top] - \tfrac{1}{\mathbf{v}} \right\}
\end{aligned}
\tag{13}
$$

The hessian in $\mathbf{v}$ is a matrix in this case:

$$
\mathrm{H}_{\mathbf{v}}\, \mathcal{D} = \tfrac{1}{2} \operatorname{diag}\left[\tfrac{1}{\mathbf{v}^2} \right]. \tag{14}
$$

This parameterization is useful for spatiotemporal inference problems, where the matrix $\mathbf{A}$ represents a fixed convolution which can be evaluated using the Fast Fourier Transform (FFT).

0.4 Inverse-diagonal approximation

Let $\Sigma^{-1} = \mathbf{\Lambda} + \operatorname{diag}[\mathbf{p}]$. To obtain the gradient in $\mathbf{p}$, combine the derivatives $\partial_\Sigma \mathcal{D}$ (Eq. (2)) and $\partial_{\mathbf{p}} \Sigma$ using the chain rule. If $\{(\Sigma)$ is a function of Σ, and $\Sigma(\theta_i)$ is a function of a parameter θ_i, then the chain rule is (The Matrix Cookbook; Eq. 136):

$$
\partial_{\theta_i} \{ = \left\langle \partial_\Sigma \{, \partial_{\theta_i} \Sigma \right\rangle = \sum_{kl} (\partial_{\Sigma_{kl}} \{)(\partial_{\theta_i} \Sigma_{kl}) \tag{15}
$$

From (2) we have $\partial_\Sigma \mathcal{D} = \frac{1}{2}\left(\mathbf{\Lambda} - \Sigma^{-1}\right)$; Since $\Sigma^{-1} = \mathbf{\Lambda} + \operatorname{diag}[\mathbf{p}]$, this simplifies to:

$$
\begin{aligned}
\partial_\Sigma \mathcal{D} &= \tfrac{1}{2} \left(\mathbf{\Lambda} - \Sigma^{-1} \right) \\
&= \tfrac{1}{2} \left(\mathbf{\Lambda} - \mathbf{\Lambda} - \operatorname{diag}[\mathbf{p}] \right) \\
&= -\tfrac{1}{2} \operatorname{diag}[\mathbf{p}]
\end{aligned}
\tag{16}
$$

We also need $\partial_{\mathrm{p}_i} \Sigma$. Let $\mathrm{Y} = \Sigma^{-1}$. The derivative $\partial \mathrm{Y}^{-1}$ is given as identity (59) in The Matrix Cookbook as $\partial \mathrm{Y}^{-1} = -\mathrm{Y}^{-1}(\partial \mathrm{Y})\mathrm{Y}^{-1}$. Using this, we can obtain $\partial_{\mathrm{p}_i} \Sigma$:

$$
\begin{aligned}
\partial_{\mathrm{p}_i} \Sigma &= \partial_{\mathrm{p}_i} \mathrm{Y}^{-1} = -\mathrm{Y}^{-1} \left(\partial_{\mathrm{p}_i} \mathrm{Y} \right) \mathrm{Y}^{-1} = -\Sigma \left(\partial_{\mathrm{p}_i} \Sigma^{-1} \right) \Sigma \\
&= -\Sigma \partial_{\mathrm{p}_i} \left[\mathbf{\Lambda} + \operatorname{diag}[\mathbf{p}_i] \right] \Sigma = -\Sigma \mathbf{J}_{ii} \Sigma \\
&= -\boldsymbol{\sigma}_i \boldsymbol{\sigma}_i^\top
\end{aligned}
\tag{17}
$$

where $\boldsymbol{\sigma}_i$ is the i^{th} row of Σ and $\mathbf{J}_{ii}$ is a matrix which is zero everywhere, except for at index (i, i), where it is 1.

Applying (15) to (16) and (17) for a particular element $\mathbf{p}_i$ gives:

$$\begin{aligned}
\partial_{\mathbf{p}_i}\mathcal{D} &= \sum_{kl}[\partial_{\Sigma_{kl}}\mathcal{D}][\partial_{\mathbf{p}_i}\Sigma_{kl}] = \sum_{kl}\left\{-\tfrac{1}{2}\operatorname{diag}[\mathbf{p}]\right\}_{kl}\left\{-\boldsymbol{\sigma}_i\boldsymbol{\sigma}_i^\top\right\}_{kl} \\
&= \tfrac{1}{2}\sum_{kl}\delta_{k=l}\mathbf{p}_k\boldsymbol{\sigma}_{ik}\boldsymbol{\sigma}_{il} = \tfrac{1}{2}\sum_{k}\mathbf{p}_k\boldsymbol{\sigma}_{ik}\boldsymbol{\sigma}_{ik} = \tfrac{1}{2}\sum_{k}\mathbf{p}_k\boldsymbol{\sigma}_{ik}^2 \\
&= \tfrac{1}{2}\mathbf{p}\boldsymbol{\sigma}_i^{\circ 2}
\end{aligned} \tag{18}$$

where $(\cdot)^{\circ 2}$ denotes the element-wise square of a vector or matrix. In matrix notation, this is:

$$\partial_{\mathbf{p}}\mathcal{D} = \tfrac{1}{2}\mathbf{p}\Sigma^{\circ 2} = \tfrac{1}{2}\operatorname{diag}[\Sigma\operatorname{diag}[\mathbf{p}]\,\Sigma], \tag{19}$$

The Hessian-vector product is cumbersome, since each term in the expression $\Sigma\,(\operatorname{diag}[\mathbf{p}])\,\Sigma$ depends on $\mathbf{p}$. In the case of the log-linear Poisson GLM, the gradient (**??**) simplifies further and optimization becomes tractable. We will explore this further in later notes.

This parameterization resembles the closed-form covariance update for a linear, Gaussian model, where $1/\mathbf{p}$ is a vector of measurement noise variances. It is also a useful parameterization for variational Bayesian solutions for non-conjugate Generalized Linear Models (GLMs), where $\mathbf{p}$ becomes a free parameter to be estimated.

0.5 $\Sigma = \mathbf{F}^\top\mathbf{Q}\mathbf{Q}^\top\mathbf{F}$

Let $\Sigma = \mathbf{F}^\top\mathbf{Q}\mathbf{Q}^\top\mathbf{F}$, where $\mathbf{Q} \in \mathbb{R}^{K\times K}$; $K<L$ is the free parameter and $\mathbf{F} \in \mathbb{R}^{K\times L}$ is a fixed transformation. If $\mathbf{Q}$ is a lower-triangular matrix, then this approximation involves optimizing $K(K+1)/2$ parameters.

Since the trace is invariant under cyclic permutation, $\operatorname{tr}[\boldsymbol{\Lambda}\mathbf{F}^\top\mathbf{Q}\mathbf{Q}^\top\mathbf{F}] = \operatorname{tr}[\mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top\mathbf{Q}\mathbf{Q}^\top]$. The derivatives have the same form as (12) with $\tilde{\boldsymbol{\Lambda}} = \mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top$:

$$\begin{aligned}
\partial_{\mathbf{Q}}\mathcal{D} &= \tilde{\boldsymbol{\Lambda}}\mathbf{Q} - \mathbf{Q}^{-\top} \\
&= \mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top\mathbf{Q} - \mathbf{Q}^{-\top} \\
\partial_{\mathbf{Q}}\left\langle\partial_{\mathbf{Q}}\mathcal{D}, \mathbf{M}\right\rangle &= \tilde{\boldsymbol{\Lambda}}\mathbf{M}^\top + \mathbf{Q}^{-\top}\mathbf{M}^\top\mathbf{Q}^{-\top} \\
&= \mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top\mathbf{M}^\top + \mathbf{Q}^{-\top}\mathbf{M}^\top\mathbf{Q}^{-\top}
\end{aligned} \tag{20}$$

This form is convenient for spatiotemporal inference problems that are sparse in frequency space. In this application, $\mathbf{F}$ corresponds a (unitary) Fourier transform with all by K of the resulting frequency components discarded. The product of $\mathbf{F}$ with a vector $\mathbf{v}$ can be computed in $\mathcal{O}[L\log(L)]$ time using the Fast Fourier Transform (FFT). Alternatively, if $K \le \mathcal{O}(\log(L))$, it is faster to simply multiply $\mathbf{F}\mathbf{v}$ directly. Furthermore, if $\mathbf{F}$ is semi-orthogonal ($\mathbf{F}\mathbf{F}^\top = \mathbf{I}$), then calculation of $\mathbf{F}^\top\mathbf{Q}$ can be re-used (for example $\operatorname{diag}[\Sigma] = [(\mathbf{F}^\top\mathbf{Q})^{\circ 2}]^\top\mathbf{1}$).

0.6 Conclusion

These notes provide the gradients and Hessian-vector products for four simplified parameterizations of the posterior covariance matrix for variational Gaussian process inference. If combined with the gradients and Hessian-vector products for the expected log-likelihood, these expressions can be used with Krylov-subspace solvers to compute the Newton-Raphson update to optimize Σ.

We evaluated the following parameterizations for Σ: 1. Σ:

$$\begin{aligned} \partial &= \tfrac{1}{2}\left(\Lambda - \Sigma^{-1}\right) \\ \partial\langle\partial, \mathbf{M}\rangle &= \frac{1}{2}\Sigma^{-1}\mathbf{M}^\top\Sigma^{-1} \end{aligned} \tag{21}$$

2. $\Sigma \approx \mathbf{X}\mathbf{X}^\top$:

$$\begin{aligned} \partial &= \Lambda\mathbf{X} - \mathbf{X}^{+\top}. \\ \partial\langle\partial, \mathbf{M}\rangle &= \Lambda\mathbf{M}^\top + \mathbf{X}^{+\top}\mathbf{M}^\top\mathbf{X}^{+\top} - (\mathbf{I} - \mathbf{X}^{+\top}\mathbf{X}^\top)\mathbf{M}\mathbf{X}^{+}\mathbf{X}^{+\top} \end{aligned} \tag{22}$$

3. $\Sigma \approx \mathbf{A}^\top \operatorname{diag}[\mathbf{v}]\mathbf{A}$:

$$\begin{aligned} \partial &= \tfrac{1}{2}\left\{\operatorname{diag}[\mathbf{A}\Lambda\mathbf{A}^\top] - \tfrac{1}{\mathbf{v}}\right\} \\ \partial\langle\partial, \mathbf{u}\rangle &= \tfrac{1}{2}\left[\tfrac{1}{\mathbf{v}^2}\right]^\top \mathbf{u} \end{aligned} \tag{23}$$

4. $\Sigma \approx [\Lambda + \operatorname{diag}[\mathbf{p}]]^{-1}$:

$$\partial = \tfrac{1}{2}\mathbf{p}\Sigma^{\circ 2} = \tfrac{1}{2}\operatorname{diag}\left[\Sigma \operatorname{diag}\left[\mathbf{p}\right]\Sigma\right], \tag{24}$$

5. $\mathbf{F}^\top\mathbf{Q}\mathbf{Q}^\top\mathbf{F}$:

$$\begin{aligned} \partial &= \mathbf{F}\Lambda\mathbf{F}^\top\mathbf{Q} - \mathbf{Q}^{-\top} \\ \partial\langle\partial, \mathbf{M}\rangle &= \mathbf{F}\Lambda\mathbf{F}^\top\mathbf{M}^\top + \mathbf{Q}^{-\top}\mathbf{M}^\top\mathbf{Q}^{-\top} \end{aligned} \tag{25}$$

In future notes, we will consider the full derivatives required for variational latent Gaussian-process inference for the Poisson and probit generalized linear models.