Supplemental material: Effiicent inference in matrix-variate Gaussian models with iid observation noise

Oliver Stegle[1]
Max Planck Institutes
Tübingen, Germany
stegle@tuebingen.mpg.de

Christoph Lippert[1]
Max Planck Institutes
Tübingen, Germany
clippert@tuebingen.mpg.de

Joris Mooij
Institute for Computing and Information
Radboud University Nijmegen
Nijmegen, The Netherlands
j.mooij@cs.ru.nl

Neil Lawrence
Department of Computer Science
University of Sheffield
Sheffield, UK
N.Lawrence@sheffield.ac.uk

Karsten Borgwardt
Max Planck Institutes & Eberhard Karls Universität
Tübingen, Germany
karsten.borgwardt@tuebingen.mpg.de

1 Preliminaries

- $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{U}^{\mathrm{T}}$ is the eigenvalue decomposition of the symmetric M-by-M matrix $\mathbf{A}$, where $\mathbf{U}$ is an M-by-M orthonormal matrix, of the M eigenvectors of $\mathbf{A}$ and $\mathbf{S}$ is an M-by-M diagonal matrix, holding the corresponding eigenvalues of $\mathbf{A}$ as diagonal entries.
- $\mathbf{A} \odot \mathbf{B}$ is the pointwise or Hadamard product of $\mathbf{A}$ and $\mathbf{B}$.
- $\mathbf{A} \otimes \mathbf{B}$ is the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$.
- $\mathbf{Y} \in \mathbb{R}^{N \times D}$ is the matrix of all samples, having N rows and D columns.
- $\mathrm{vec}(\mathbf{A}) = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{LM} \end{pmatrix}$ is the vec-operator, that concatenates the columns of a matrix $\mathbf{A}$ in a column vector.
- $\mathrm{vec}^{-1}_{L \times M} \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{LM} \end{pmatrix} = (\mathbf{A})$ is the inverse vec-operator, that arranges the entries of a vector in an $L \times M$ matrix $\mathbf{A}$.

2 Covariance estimation in Kronecker Gaussian processes

Here, we give details of an efficient implementation of the tensor Gaussian process model derived in the main paper.

[1]These authors contributed equally to this work.

The basic model

We start with a general form a Gaussian process model where the covariance has a Kronecker structure:

$$\ln p(\mathbf{Y} \mid \mathbf{R}, \mathbf{C}, \sigma^2) = \mathcal{N}\left(\mathrm{vec}(\mathbf{Y}) \,\middle|\, \mathbf{0}, \mathbf{C}(\mathbf{\Theta_C}) \otimes \mathbf{R}(\mathbf{\Theta_R}) + \sigma^2 \mathbf{I}\right). \tag{1}$$

Here, $\mathbf{Y}$ is the data matrix with N rows (samples) and D columns (features). We defined $\mathbf{R}(\mathbf{\Theta_R})$ as the row "row covariance" of the data matrix and $\mathbf{C}(\mathbf{\Theta_C})$ corresponds to the "column covariance".

For notational convenience we will drop the dependence of the covariance matrices on additional hyperparameters $\mathbf{\Theta_R}$ and $\mathbf{\Theta_C}$, respectively. Furthermore, we will make the simplifying assumption that $\mathbf{C}$ is kept constant, i.e. has no parameters that need to be adapted during learning. Importantly, this is no restriction for general solutions as all calculations can be performed with respect to other covariance as well, for example iteratively optimizing hyperparameters of $\mathbf{R}$ and $\mathbf{C}$ in turn.

To implement parameter optimization of the covariance parmeters of the model in Equation (1), we require efficient evaluation of the marginal likelihood and the gradients with respect to hyperparameters.

2.1 Efficient evaluation of the marginal likelihood

The term we want to evaluate is the log-marginal-likelihood, given by the log of the multivariate Normal density

$$\begin{aligned}&\mathcal{L}(\mathbf{\Theta_R}, \mathbf{\Theta_C}, \sigma^2) = \\ &-\tfrac{N \cdot D}{2} \ln 2\pi - \tfrac{1}{2} \underbrace{\ln |\mathbf{C} \otimes \mathbf{R} + \sigma^2 \mathbf{I}_{N \cdot D}|}_{\text{log-det}} - \tfrac{1}{2} \underbrace{\mathrm{vec}(\mathbf{Y})^{\mathrm{T}} (\mathbf{C} \otimes \mathbf{R} + \sigma^2 \mathbf{I}_{N \cdot D})^{-1} \mathrm{vec}(\mathbf{Y})}_{\text{squared form}}.\end{aligned} \tag{2}$$

In the following, we make heavy use of the eigenvalue decomposition of $\mathbf{C} \otimes \mathbf{R} = \mathbf{U}\mathbf{S}\mathbf{U}^{\mathrm{T}}$, where $\mathbf{U}$ is an $N \cdot D$-by-$N \cdot D$ orthonormal matrix, holding the eigenvectors of $\mathbf{C}$ and $\mathbf{S}$ is an N-by-NG diagonal matrix, holding the corrseponding eingenvalues on the diagonal. This decomposition can be efficiently obtained from the composition of the individual kronecker terms (after some reordering), i.e. $\mathbf{C} \otimes \mathbf{R} = (\mathbf{U_C} \otimes \mathbf{U_R})(\mathbf{S_C} \otimes \mathbf{S_R})(\mathbf{U_C^{\mathrm{T}}} \otimes \mathbf{U_R^{\mathrm{T}}})$.

We derive efficient solutions for the logarithm of the determinant of $\mathbf{C} \otimes \mathbf{R}$ and the squared form separately.

2.1.1 Efficient evaluation of the log-det

Assuming that we have the eigenvalue decompositiosn for $\mathbf{R}$ and $\mathbf{C}$, the logarithm of the determinant can be written as

$$\begin{aligned}\ln |\mathbf{C} \otimes \mathbf{R} + \sigma^2 \mathbf{I}_{N \cdot D}| \quad &= \ln |(\mathbf{U_C} \otimes \mathbf{U_R})(\mathbf{S_C} \otimes \mathbf{S_R})(\mathbf{U_C^{\mathrm{T}}} \otimes \mathbf{U_R^{\mathrm{T}}}) + \sigma^2 \mathbf{I}_{N \cdot D}| \\ &= \ln |(\mathbf{U_C} \otimes \mathbf{U_R})| + \ln |(\mathbf{S_C} \otimes \mathbf{S_R} + \sigma^2 \mathbf{I}_{N \cdot D})| + \ln |\mathbf{U_C^{\mathrm{T}}} \otimes \mathbf{U_R^{\mathrm{T}}}| \\ &= \ln |(\mathbf{S_C} \otimes \mathbf{S_R} + \sigma^2 \mathbf{I}_{N \cdot D})| \\ &= \sum_{r=1}^{N} \sum_{c=1}^{D} - \ln(\mathbf{S_R}[r, r] \cdot \mathbf{S_C}[c, c] + \sigma^2).\end{aligned} \tag{3}$$

This term can be evaluated in $O(N \cdot D)$.

2.1.2 Efficient evaluation of the squared form

Also the squared form in the log marginal likelihood can be evaluated efficiently as follows:

$$\begin{aligned}
&\mathrm{vec}(\mathbf{Y})^{\mathrm{T}}(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I})^{-1}\mathrm{vec}(\mathbf{Y}) = \mathrm{vec}(\mathbf{Y})^{\mathrm{T}}\mathbf{U}(\mathbf{S}+\sigma^2\mathbf{I})^{-1}\mathbf{U}^{\mathrm{T}}\mathrm{vec}(\mathbf{Y})\\
&\text{Using orthogonality of } \mathbf{U}=(\mathbf{U}^T)^{-1}\text{:}\\
&=\mathrm{vec}(\mathbf{Y})^{\mathrm{T}}(\mathbf{U_C}\otimes\mathbf{U_R})(\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}(\mathbf{U_C^{\mathrm{T}}}\otimes\mathbf{U_R^{\mathrm{T}}})\mathrm{vec}(\mathbf{Y})\\
&\text{using the vectorization identities of kronecker structures}\\
&=(\mathrm{vec}(\mathbf{U_R^{\mathrm{T}}}\mathbf{Y}\mathbf{U_C}))^{\mathrm{T}}(\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\mathrm{vec}(\mathbf{U_R^{\mathrm{T}}}\mathbf{Y}\mathbf{U_C}) \qquad (4)
\end{aligned}$$

We only need to compute the diagonal of $(\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}$, which already has been computed before. As this term involves the rotation of the data matrix $\mathbf{Y}$, it can be evaluated in $O(N^2D+ND^2)$.

2.2 Efficient evaluation of the gradients w.r.t. to covariance hyperparameters

Here, the aim is to evaluate the gradient of Equation (1) w.r.t. to a particular column covariance parameter $\theta_{\mathbf{C}}\in\mathbf{\Theta_C}$, row covariance parameter $\theta_{\mathbf{R}}\in\mathbf{\Theta_R}$ and the noise parameter σ^2.

2.3 Derivatiaves w.r.t. noise variance σ^2

2.3.1 Derivatiaves of the log-det term w.r.t. noise variance σ^2

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\ln|\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I}| = \mathrm{Tr}\left[(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I})^{-1}\frac{\mathrm{d}}{\mathrm{d}\sigma^2}(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I})\right]\\
&=\mathrm{Tr}\left[(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I})^{-1}\frac{\mathrm{d}}{\mathrm{d}\sigma^2}(\sigma^2\mathbf{I})\right]\\
&=\mathrm{Tr}\left[\mathbf{U}(\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{I}\right]
\end{aligned}$$

Using that the sum of the eigenvalues of a matrix equals its trace, the derivative becomes

$$\sum_{r=1}^{N}\sum_{c=1}^{D}\frac{1}{\mathbf{S_R}[r,r]\cdot\mathbf{S_C}[c,c]+\sigma^2}. \qquad (5)$$

2.3.2 Squared form derivatives w.r.t. noise variance σ^2

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\mathrm{vec}(\mathbf{Y})^{\mathrm{T}}\left(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}\mathrm{vec}(\mathbf{Y}) = \mathrm{vec}(\mathbf{Y})^{\mathrm{T}}\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\left(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}\mathrm{vec}(\mathbf{Y}).\\
&\text{Using } \frac{\mathrm{d}}{\mathrm{d}\sigma^2}\mathbf{A}^{-1} = -\mathbf{A}^{-1}\left[\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\mathbf{A}\right]\mathbf{A}^{-1}\text{, this becomes}\\
&-\mathrm{vec}(\mathbf{Y})^{\mathrm{T}}\left(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}\left[\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\left(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I}\right)\right]\left(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}\mathrm{vec}(\mathbf{Y})\\
&=-\mathrm{vec}(\mathbf{Y})^{\mathrm{T}}(\mathbf{U_C}\otimes\mathbf{U_R})(\mathbf{S}+\sigma^2\mathbf{I})^{-1}(\mathbf{U_C^{\mathrm{T}}}\otimes\mathbf{U_R^{\mathrm{T}}})(\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\sigma^2\mathbf{I})(\mathbf{U_C}\otimes\mathbf{U_R})(\mathbf{S}+\sigma^2\mathbf{I})^{-1}(\mathbf{U_C^{\mathrm{T}}}\otimes\mathbf{U_R^{\mathrm{T}}})\mathrm{vec}(\mathbf{Y})\\
&=-\mathrm{vec}(\mathbf{U_R^{\mathrm{T}}}\mathbf{Y}\mathbf{U_C})^{\mathrm{T}}(\mathbf{S}+\sigma^2\mathbf{I})^{-1}(\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\sigma^2\mathbf{I})(\mathbf{S}+\sigma^2\mathbf{I})^{-1}\mathrm{vec}(\mathbf{U_R^{\mathrm{T}}}\mathbf{Y}\mathbf{U_C}) \qquad (6)
\end{aligned}$$

2.4 Derivatives w.r.t a row covariance parameter $\theta_{\mathbf{R}}$

2.4.1 Derivatives of the determinant w.r.t. a row covariance parameter $\theta_{\mathbf{R}}$

$$\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}} \ln|\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I}| = \mathrm{Tr}\left[(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I})^{-1}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I})\right]$$
$$= \mathrm{Tr}\left[(\mathbf{U_C}\otimes\mathbf{U_R})(\mathbf{S}+\sigma^2)^{-1}(\mathbf{U_C^T}\otimes\mathbf{U_R^T})(\mathbf{C}\otimes\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R})\right].$$

Using the identity $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = \mathbf{AC}\otimes\mathbf{BD}$ this equals

$$\mathrm{Tr}\left[(\mathbf{U_C}\otimes\mathbf{U_R})(\mathbf{S}+\sigma^2\mathbf{I})^{-1}(\mathbf{U_C^T C}\otimes\mathbf{U_R^T}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R})\right]$$

Using $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$, this becomes

$$\mathrm{Tr}\left[(\mathbf{S}+\sigma^2\mathbf{I})^{-1}(\mathbf{U_C^T C}\otimes\mathbf{U_R^T}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R})(\mathbf{U_C}\otimes\mathbf{U_R})\right]$$
$$= \mathrm{Tr}\left[(\mathbf{S}+\sigma^2\mathbf{I})^{-1}(\mathbf{U_C^T C U_C}\otimes\mathbf{U_R^T}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}\mathbf{U_R})\right]$$
$$= \mathrm{Tr}\left[(\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\Big(\mathbf{S_C}\otimes(\mathbf{U_R^T}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}\mathbf{U_R})\Big)\right]$$
$$= \mathrm{diag}\left((\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\right)^{\mathrm{T}}\Big(\mathrm{diag}(\mathbf{S_C})\otimes\mathrm{diag}(\mathbf{U_R^T}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}\mathbf{U_R})\Big) \tag{7}$$

As this derivation only involves the trace, we just need the diagonal of the Kronecker product, which only involves the diagonal of $(\mathbf{U_R^T}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}\mathbf{U_R})$.

For the special case, where $\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}$ has only one non-zero row:

$$\mathrm{diag}\left((\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\right)^{\mathrm{T}}\Big(\mathrm{diag}(\mathbf{S_C})\otimes\mathrm{diag}(\mathbf{U_R^T}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}\mathbf{U_R})\Big)$$
$$= \mathrm{diag}\left((\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\right)^{\mathrm{T}}\Big(\mathrm{diag}(\mathbf{S_C})\otimes\mathrm{diag}([\mathbf{U_R}]_{i:}^{\mathrm{T}}[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{i:}\mathbf{U_R})\Big)$$
$$= \mathrm{diag}\left((\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\right)^{\mathrm{T}}\left(\mathrm{diag}(\mathbf{S_C})\otimes\left([\mathbf{U_R}]_{i:}\odot\left([\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{i:}\mathbf{U_R}\right)\right)^{\mathrm{T}}\right) \tag{8}$$

For the special case, where $\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}$ has only one non-zero column:

$$\mathrm{diag}\left((\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\right)^{\mathrm{T}}\Big(\mathrm{diag}(\mathbf{S_C})\otimes\mathrm{diag}(\mathbf{U_R^T}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}\mathbf{U_R})\Big)$$
$$= \mathrm{diag}\left((\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\right)^{\mathrm{T}}\left(\mathrm{diag}(\mathbf{S_C})\otimes\left(\left(\mathbf{U_R^T}[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{:i}\right)\odot[\mathbf{U_R}]_{i:}^{\mathrm{T}}\right)\right) \tag{9}$$

So for the case, where $\mathbf{K_R} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$, the derivative w.r.t. an entry of $\mathbf{X}$ is:

$$\mathrm{diag}\left((\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\right)^{\mathrm{T}}\left(\mathrm{diag}(\mathbf{S_C})\otimes\left([\mathbf{U_R}]_{i:}\odot\left([\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{i:}\mathbf{U_R}\right)\right)^{\mathrm{T}}\right)+$$
$$\mathrm{diag}\left((\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\right)^{\mathrm{T}}\left(\mathrm{diag}(\mathbf{S_C})\otimes\left(\left(\mathbf{U_R^T}[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{:i}\right)\odot[\mathbf{U_R}]_{i:}^{\mathrm{T}}\right)\right)$$
$$= 2\mathrm{diag}\left((\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\right)^{\mathrm{T}}\left(\mathrm{diag}(\mathbf{S_C})\otimes\left(\left(\mathbf{U_R^T}[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{:i}\right)\odot[\mathbf{U_R}]_{i:}^{\mathrm{T}}\right)\right)$$
$$= 2\mathrm{vec}\left(\left(\left(\mathbf{U_R^T}[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{:i}\right)^{\mathrm{T}}\odot[\mathbf{U_R}]_{i:}\right)\mathrm{vec}^{-1}_{N\times D}(\mathrm{diag}\left((\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\right))\,\mathrm{diag}(\mathbf{S_C})\right), \tag{10}$$

where we rearranged the transposes and used $\mathbf{A}\odot\mathbf{B} = \mathbf{B}\odot\mathbf{A}$.

2.4.2 Derivatiaves of the squared form w.r.t. a row covariance parameter $\theta_{\mathbf{R}}$

$$\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathrm{vec}(\mathbf{Y})^{\mathrm{T}}\left(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}\mathrm{vec}(\mathbf{Y}) = \mathrm{vec}(\mathbf{Y})^{\mathrm{T}}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\left(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}\mathrm{vec}(\mathbf{Y})$$

using $\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{A}^{-1} = -\mathbf{A}^{-1}\left[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{A}\right]\mathbf{A}^{-1}$

$$= -\mathrm{vec}(\mathbf{Y})^{\mathrm{T}}\left(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}\left[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\left(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I}\right)\right]\left(\mathbf{C}\otimes\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}\mathrm{vec}(\mathbf{Y})$$

$$= -\mathrm{vec}(\mathbf{Y})^{\mathrm{T}}(\mathbf{U_C}\otimes\mathbf{U_R})(\mathbf{S}+\sigma^2\mathbf{I})^{-1}(\mathbf{U_C^{\mathrm{T}}}\otimes\mathbf{U_R^{\mathrm{T}}})(\mathbf{C}\otimes\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R})(\mathbf{U_C}\otimes\mathbf{U_R})(\mathbf{S}+\sigma^2\mathbf{I})^{-1}(\mathbf{U_C^{\mathrm{T}}}\otimes\mathbf{U_R^{\mathrm{T}}})\mathrm{vec}(\mathbf{Y})$$

$$= -\mathrm{vec}(\mathbf{U_R^{\mathrm{T}}}\mathbf{Y}\mathbf{U_C})^{\mathrm{T}}(\mathbf{S}+\sigma^2\mathbf{I})^{-1}(\mathbf{U_C^{\mathrm{T}}}\otimes\mathbf{U_R^{\mathrm{T}}})(\mathbf{C}\otimes\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R})(\mathbf{U_C}\otimes\mathbf{U_R})(\mathbf{S}+\sigma^2\mathbf{I})^{-1}\mathrm{vec}(\mathbf{U_R^{\mathrm{T}}}\mathbf{Y}\mathbf{U_C})$$

$$= -(\mathrm{vec}(\mathbf{U_R^{\mathrm{T}}}\mathbf{Y}\mathbf{U_C}))^{\mathrm{T}}(\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}(\mathbf{U_C^{\mathrm{T}}}\mathbf{C}\mathbf{U_C}\otimes\mathbf{U_R^{\mathrm{T}}}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}\mathbf{U_R})(\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\mathrm{vec}(\mathbf{U_R^{\mathrm{T}}}\mathbf{Y}\mathbf{U_C})$$

$$= -\mathrm{vec}(\mathbf{U_R^{\mathrm{T}}}\mathbf{Y}\mathbf{U_C})^{\mathrm{T}}(\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\big(\mathbf{S_C}\otimes(\mathbf{U_R^{\mathrm{T}}}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}\mathbf{U_R})\big)(\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\mathrm{vec}(\mathbf{U_R^{\mathrm{T}}}\mathbf{Y}\mathbf{U_C})$$

$$= -\mathrm{vec}(\tilde{\mathbf{Y}})^{\mathrm{T}}\big(\mathbf{S_C}\otimes(\mathbf{U_R^{\mathrm{T}}}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}\mathbf{U_R})\big)\mathrm{vec}(\tilde{\mathbf{Y}})$$

$$= -\mathrm{vec}(\tilde{\mathbf{Y}})^{\mathrm{T}}\mathrm{vec}\big((\mathbf{U_R^{\mathrm{T}}}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}\mathbf{U_R})\tilde{\mathbf{Y}}\mathbf{S_C}\big), \tag{11}$$

where $\mathrm{vec}(\tilde{\mathbf{Y}}) = (\mathbf{S_C}\otimes\mathbf{S_R}+\sigma^2\mathbf{I})^{-1}\mathrm{vec}(\mathbf{U_R^{\mathrm{T}}}\mathbf{Y}\mathbf{U_C})$.

For the special case, where the derivative of $\mathbf{R}$ is a row matrix:

$$\begin{aligned}
&-\mathrm{vec}(\tilde{\mathbf{Y}})^{\mathrm{T}}\mathrm{vec}\big((\mathbf{U_R^{\mathrm{T}}}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}\mathbf{U_R})\tilde{\mathbf{Y}}\mathbf{S_C}\big)\\
&= -\mathrm{vec}\left(\tilde{\mathbf{Y}}\right)^{\mathrm{T}}\mathrm{vec}\left(([\mathbf{U_R}]_{i:}^{\mathrm{T}}[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{i:}\mathbf{U_R})\tilde{\mathbf{Y}}\mathbf{S_C}\right)\\
&= -\mathrm{Tr}\left(\left(\tilde{\mathbf{Y}}^{\mathrm{T}}[\mathbf{U_R}]_{i:}^{\mathrm{T}}\right)\left([\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{i:}\mathbf{U_R}\tilde{\mathbf{Y}}\mathbf{S_C}\right)\right)\\
&= -[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{i:}\mathbf{U_R}\tilde{\mathbf{Y}}\mathbf{S_C}\tilde{\mathbf{Y}}^{\mathrm{T}}[\mathbf{U_R}]_{i:}^{\mathrm{T}},
\end{aligned} \tag{12}$$

where we use the fact that the trace of an outer product between vectors equals to the inner product between the vectors.

For the special case, where the derivative of $\mathbf{R}$ is a column matrix:

$$\begin{aligned}
&= -\mathrm{vec}(\tilde{\mathbf{Y}})^{\mathrm{T}}\mathrm{vec}\big((\mathbf{U_R^{\mathrm{T}}}\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}\mathbf{U_R})\tilde{\mathbf{Y}}\mathbf{S_C}\big)\\
&= -\mathrm{vec}(\tilde{\mathbf{Y}})^{\mathrm{T}}\mathrm{vec}\big((\mathbf{U_R^{\mathrm{T}}}[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{:i}[\mathbf{U_R}]_{i:})\tilde{\mathbf{Y}}\mathbf{S_C}\big)\\
&= -\mathrm{Tr}\left(\tilde{\mathbf{Y}}^{\mathrm{T}}(\mathbf{U_R^{\mathrm{T}}}[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{:i}[\mathbf{U_R}]_{i:})\tilde{\mathbf{Y}}\mathbf{S_C}\right)\\
&= -[\mathbf{U_R}]_{i:}\tilde{\mathbf{Y}}\mathbf{S_C}\tilde{\mathbf{Y}}^{\mathrm{T}}\mathbf{U_R^{\mathrm{T}}}[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{:i}
\end{aligned} \tag{13}$$

So for the derivative w.r.t. entries in hidden factors $\mathbf{X}$ in a linear kernel $\mathbf{K_R} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ we have to compute the following term:

$$\begin{aligned}
&-[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{i:}\mathbf{U_R}\tilde{\mathbf{Y}}\mathbf{S_C}\tilde{\mathbf{Y}}^{\mathrm{T}}[\mathbf{U_R}]_{i:}^{\mathrm{T}} - [\mathbf{U_R}]_{i:}\tilde{\mathbf{Y}}\mathbf{S_C}\tilde{\mathbf{Y}}^{\mathrm{T}}\mathbf{U_R^{\mathrm{T}}}[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{:i}\\
&-2[\mathbf{U_R}]_{i:}\tilde{\mathbf{Y}}\mathbf{S_C}\tilde{\mathbf{Y}}^{\mathrm{T}}\mathbf{U_R^{\mathrm{T}}}[\frac{\mathrm{d}}{\mathrm{d}\theta_{\mathbf{R}}}\mathbf{R}]_{:i},
\end{aligned} \tag{14}$$

where we observe that the second term is the transpose of the first one.