The Bigraphical Lasso: Supplementary material

Alfredo Kalaitzis A.KALAITZIS@UCL.AC.UK
Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

John Lafferty LAFFERTY@GALTON.UCHICAGO.EDU
Department of Statistics, University of Chicago, 5734 S. University Ave., Chicago, IL 60637, USA

Shuheng Zhou SHUHENGZ@UMICH.EDU
Dept. of Statistics, University of Michigan, 439 West Hall, 1085 S. University Ave., Ann Arbor, MI 48109, USA

Neil D. Lawrence N.LAWRENCE@SHEFFIELD.AC.UK
Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK

1. Useful identities

All matrix derivatives are based on the following differential forms; for proofs see (Magnus & Neudecker, 1988):

$$\partial(\mathbf{X} \otimes \mathbf{Y}) = (\partial \mathbf{X}) \otimes \mathbf{Y} + \mathbf{X} \otimes (\partial \mathbf{Y}) \tag{1}$$

$$\partial \mathbf{X}^{-1} = -\mathbf{X}^{-1}(\partial \mathbf{X})\mathbf{X}^{-1} \tag{2}$$

$$\partial \ln|\mathbf{X}| = \mathrm{tr}\left(\mathbf{X}^{-1}\partial \mathbf{X}\right) \ . \tag{3}$$

Moreover, if the $\mathbf{X}$ in $\frac{\partial f}{\partial \mathbf{X}}$ is symmetric then

$$\frac{\partial f}{\partial \mathbf{X}} = \left[\frac{\partial f}{\partial \mathbf{X}}\right] + \left[\frac{\partial f}{\partial \mathbf{X}}\right]^\top - \mathbf{I} \circ \left[\frac{\partial f}{\partial \mathbf{X}}\right] \ . \tag{4}$$

2. Derivatives for BiGLasso

We denote $\mathbf{J}^{ij}$ as the single-entry matrix with $J_{ij} = 1$ and zeros elsewhere; $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

Gradient wrt $\mathbf{\Psi}_n$ Taking the gradient of (8) with respect to Ψ_{ij} and using identity (3), we get:

$$\begin{aligned}
&\frac{\partial}{\partial \Psi_{ij}} \ln|\mathbf{\Psi}_n \oplus \mathbf{\Theta}_p| \\
&= \mathrm{tr}\left\{(\mathbf{\Psi}_n \oplus \mathbf{\Theta}_p)^{-1} \frac{\partial(\mathbf{\Psi}_n \oplus \mathbf{\Theta}_p)}{\partial \Psi_{ij}}\right\} \\
&= \mathrm{tr}\left\{\mathbf{W}\left(\frac{\partial \mathbf{\Psi}_n}{\partial \Psi_{ij}} \otimes \mathbf{I}_p\right)\right\} \ , \text{ by } (1) \\
&= \mathrm{tr}\left\{\mathbf{W}\left((\mathbf{J}^{ij} + \mathbf{J}^{ji} - \mathbf{J}^{ij}\mathbf{J}^{ij}) \otimes \mathbf{I}_p\right)\right\} \ , \text{ by } (4) \\
&= \mathrm{tr}\left\{\mathbf{W}\begin{bmatrix} \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \mathbf{I}_p^{(\mathbf{i},\mathbf{j})} & \vdots \\ \mathbf{0} & \dots & \mathbf{0} \end{bmatrix}\right\} + \mathrm{tr}\left\{\mathbf{W}\left(\mathbf{J}^{ji} \otimes \mathbf{I}_p\right)\right\} \\
&\quad - \mathrm{tr}\left\{\mathbf{W}\left(\mathbf{J}^{ij}\mathbf{J}^{ij} \otimes \mathbf{I}_p\right)\right\} \\
&= 2 \ \mathrm{tr}\left\{\mathbf{W}_{(\mathbf{i},\mathbf{j})}\right\} - \delta_{ij}\mathrm{tr}\left\{\mathbf{W}_{(\mathbf{i},\mathbf{j})}\right\} \ ,
\end{aligned}$$

where $\mathbf{W} \triangleq (\mathbf{\Psi}_n \oplus \mathbf{\Theta}_p)^{-1}$; $\mathbf{I}_p^{(\mathbf{i},\mathbf{j})}$ is at the (i,j)-th block of size $p \times p$, that is, $(\mathbf{i},\mathbf{j}) = [(pi-p+1):pi \ , \ (pj-p+1):pj]$. Thereby,

$$\frac{\partial}{\partial \mathbf{\Psi}_n} \ln|\mathbf{\Psi}_n \oplus \mathbf{\Theta}_p| = 2 \ \mathrm{tr}_p\left(\mathbf{W}\right) - \mathrm{tr}_p\left(\mathbf{W}\right) \circ \mathbf{I} \ . \tag{5}$$

Also, using (4) gives

$$\frac{\partial \ p \, \mathrm{tr}\left(\mathbf{\Psi}_n \mathbf{T}\right)}{\partial \mathbf{\Psi}_n} = 2p \, \mathbf{T} - \mathbf{T} \circ \mathbf{I} \ . \tag{6}$$

3. Product of Gaussians

The product of two Gaussian distributions yields an *unnormalized* Gaussian:

$$\begin{aligned} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)\ \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B) &\propto \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_C, \boldsymbol{\Sigma}_C), \\ \text{where} \quad \boldsymbol{\mu}_C &= \boldsymbol{\Sigma}_C \left(\boldsymbol{\Sigma}_A^{-1}\boldsymbol{\mu}_A + \boldsymbol{\Sigma}_B^{-1}\boldsymbol{\mu}_B\right)^{-1} \\ \boldsymbol{\Sigma}_C &= \left(\boldsymbol{\Sigma}_A^{-1} + \boldsymbol{\Sigma}_B^{-1}\right)^{-1} . \end{aligned} \tag{7}$$

Note that the precision matrix of the unnormalized Gaussian is simply the *sum* of the individual precision matrices and the mean is the *convex sum* of the means, weighted by the individual precision matrices (Rasmussen & Williams, 2006, section A.2).

References

Magnus, J. R. and Neudecker, H. *Matrix differential calculus with applications in statistics and econometrics*. Wiley, 1988.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian processes for machine learning*. MIT Press, Cambridge, MA, Cambridge, MA, 2006. ISBN 0-262-18253-X.