Supplementary material for Paper "Matrix-normal models for fMRI analysis"

Anonymous Author(s)

1 Appendix A : Matrix-normal intersubject functional connectivity and simultaneous modeling

Here we derive matrix-normal variants of two additional models from the literature, intersubject functional connectivity (Simony et al., 2016), and simultaneous modeling (Turner et al., 2013, 2014, 2016).

1.1 Matrix-normal intersubject functional connectivity

The goal of the ISFC method is to estimate a "shared stimulus-induced covariance matrix" in fMRI data as a way to measure functional connectivity between brain regions while abstracting over subject-specific connectivity patterns and extracting only the patterns that are consistent across subjects. The intuition behind the method is simple: it computes pairwise correlations between each subject's patterns and averages them. To prove that the method is indeed free of subject-specific bias, Simony and colleagues frame their model in terms of a gaussian generative model. Here is this generative model, rewritten in the matrix-normal formalism:

$$\mathbf{A} \mid \mathbf{C} \sim \mathcal{MN}(0, \mathbf{C}, \mathbf{I}) \tag{1}$$

$$\mathbf{D}_i \mid \sigma^2_{\mathbf{D}} \sim \mathcal{MN}(0, \sigma^2_{\mathbf{D}}\mathbf{I}, \mathbf{I}) \tag{2}$$

$$\mathbf{S} \sim \mathcal{MN}(0, \mathbf{I}, \mathbf{I}) \tag{3}$$

$$\mathbf{E} \mid \mathbf{Q} \sim \mathcal{MN}(0, \mathbf{Q}, \mathbf{I}) \tag{4}$$

$$\mathbf{X}_i = (\mathbf{A} + \mathbf{D}_i)\mathbf{S} + \mathbf{E}_i \tag{5}$$

The "shared stimulus-induced covariance matrix" the method is intended to estimate is $\mathbf{C}$, the row covariance of the projection matrix into latent space. The somewhat redundant formulation is needed to motivate the closed-form estimator used in the original method. However, the formulation required for the closed-form estimator places severe restrictions on the projection matrix $\mathbf{S}$, both in terms of its rank (which must be full) and distribution (which is independent standard normal). We instead simplify the model

and integrate out the projection. Let $\mathbf{W}_i = \mathbf{A} + \mathbf{D}_i$, and rewrite:

$$\mathbf{A} \mid \mathbf{C} \sim \mathcal{MN}(0, \mathbf{C}, \mathbf{I}) \tag{6}$$

$$\mathbf{W}_i \mid \mathbf{A}, \sigma^2_{\mathbf{D}} \sim \mathcal{MN}(\mathbf{A}, \sigma^2_{\mathbf{D}}\mathbf{I}, \mathbf{I}) \tag{7}$$

$$\mathbf{S} \sim \mathcal{MN}(0, \mathbf{I}, \mathbf{I}) \tag{8}$$

$$\mathbf{X}_i \mid \mathbf{W}, \mathbf{S}, \mathbf{Q} \sim \mathcal{MN}(\mathbf{W}_i\mathbf{S}, \mathbf{Q}, \mathbf{I}) \tag{9}$$

Then marginalize A:

$$\mathbf{W}_i \mid C, \sigma^2_{\mathbf{D}} \sim \mathcal{MN}(0, \mathbf{C} + \sigma^2_{\mathbf{D}}\mathbf{I}, \mathbf{I}) \tag{10}$$

$$\mathbf{S} \sim \mathcal{MN}(0, \mathbf{I}, \mathbf{I}) \tag{11}$$

$$\mathbf{X}_i \mid \mathbf{W}, \mathbf{S}, \mathbf{Q} \sim \mathcal{MN}(\mathbf{W}_i\mathbf{S}, \mathbf{Q}, \mathbf{I}) \tag{12}$$

The resultant model is remarkably similar to MN-SRM: ISFC models the row (spatial) noise covariance as full-rank whereas MN-SRM models it as diagonal. MN-SRM models the shared response covariance as full-rank but ISFC models it as diagonal. Finally, and most importantly, MN-SRM models the projection into latent space as orthonormal whereas ISFC is specifically interested in its covariance (which MN-SRM can in fact estimate).

1.2 Matrix-normal simultaneous modeling

The simultaneous modeling framework (Turner et al., 2015) is organized around attempts to estimate the joint covariance of the vector $\{\psi_1, \psi_2, \ldots, \psi_p, \phi_i, \phi_2, \ldots, \phi_k\}$, which is a combined vector of cognitive model parameters ψ and features extracted from fMRI signal ϕ. As it is a broad framework, a number of specific instances have been provided, with specific cognitive models including accumulator models and signal detection theory models, and feature extraction mechanisms including ICA, PCA, and other methods.

There are a number of challenges with the current formulation of simultaneous modeling that we address: first, while the formulation in terms of correlations between brain and behavior allows for intuitive interpretation, it makes it challenging to regularize the model, or place priors on brain-behavior relationships, except for the special case of complete independence. Second, by performing the feature extraction in an unsupervised way, there is no guarantee that the features extracted will be relevant to the behavior or cognitive model; on the other hand, applying the framework to whole-brain data is not generally tractable, as it involves estimating a sizable covariance matrix by MCMC.

We show how matrix-normal simultaneous modeling can address all of these challenges. Since SM is a framework rather than one specific model, and no public implementation is available, we focus on a toy example to illustrate our contribution. We choose factor analysis as our factor model, leave the cognitive model unspecified for the derivation, which is applicable to any cognitive model, and any *linear* factor model.

First, we can use properties of partitioned Gaussians to write the conditional distribution of $\phi \mid \psi$, which is a simple linear regression:

$$\phi_i \mid \mu_\phi, \boldsymbol{\ell}, \Sigma_{\phi|\psi}, \Psi \sim \mathcal{N}(\mu_\phi + \Psi\boldsymbol{\ell}, \boldsymbol{\Sigma}_{\phi|\psi}) \tag{13}$$

As we show in the supplement, the intercepts and slopes here map directly to the full covariance of the simultaneous modeling framework. Now we add the cognitive model and factorization, and stack into matrix-variate form. We also add an additional design matrix for observed stimulus features $\mathbf{X}$ and its coefficient matrix β:

$$\mathbf{H} \mid \mathbf{\Psi} \sim \text{Cog.}(\mathbf{\Psi}, \mathbf{S}) \tag{14}$$

$$\mathbf{\Phi} \mid \boldsymbol{\beta}, \boldsymbol{\ell}, \mathbf{\Sigma}_{\phi|\psi}, \mathbf{S}, \mathbf{\Psi} \sim \mathcal{MN}(\mathbf{\Psi}\boldsymbol{\ell} + \mathbf{X}\beta, \Sigma_{\Phi t}, \mathbf{\Sigma}_{\Phi s}) \tag{15}$$

$$\mathbf{Y}^\top \mid \mathbf{\Phi}, \mathbf{W}, \mathbf{\Sigma}_s, \mathbf{\Sigma}_t \sim \mathcal{MN}(\mathbf{W}\mathbf{\Phi}, \mathbf{\Sigma}_s, \mathbf{\Sigma}_t) \tag{16}$$

This analysis combines the matrix-regression model for Φ and matrix-factor model for $\mathbf{Y}$. In this case since we only need the latent factors Φ to map to the cognitive parameters Ψ, we can marginalize over the factor mapping $\mathbf{W}$. For decoding cognitive parameters Ψ, we do not need the regression mapping either, so we marginalize over the coefficients, giving us a direct model from brain behavior via latent cognitive parameters and a neural factor space:

$$\mathbf{W} \sim \mathcal{MN}(0, \mathbf{\Sigma}_s, \mathbf{I}) \tag{17}$$

$$\boldsymbol{\beta} \sim \mathcal{MN}(0, \mathbf{\Sigma}_{\Phi s}, \mathbf{U}) \tag{18}$$

$$\boldsymbol{\ell} \sim \mathcal{MN}(0, \mathbf{\Sigma}_{\Phi s}, \mathbf{V}) \tag{19}$$

$$\mathbf{H} \mid \mathbf{\Psi} \sim \text{Cog.}(\mathbf{\Psi}, \mathbf{X}) \tag{20}$$

$$\mathbf{\Phi} \mid \mathbf{\Sigma}_{\Phi s}, \mathbf{\Sigma}_{\Phi t}, \mathbf{X}, \mathbf{\Psi} \sim \mathcal{MN}(0, \mathbf{\Sigma}_{\Phi s}, \mathbf{\Sigma}_{\Phi t} + \mathbf{X}^\top\mathbf{U}\mathbf{X} + \mathbf{\Psi}^\top\mathbf{V}\mathbf{\Psi}) \tag{21}$$

$$\mathbf{Y}^\top \mid \mathbf{\Phi}, \mathbf{\Sigma}_s, \mathbf{\Sigma}_t \sim \mathcal{MN}(0, \mathbf{\Sigma}_s, \mathbf{\Sigma}_t + \mathbf{\Phi}^\intercal\mathbf{\Phi}) \tag{22}$$

Given this marginalization, both the latent neural factors and the latent cognitive parameters appear in the model only as their inner products, and are perfectly nonidentifiable. Therefore, an equivalent model is a direct regression from voxels to cognitive parameters, marginalized over the mapping. This will be true for any linear factor model under marginalization:

$$\mathbf{H} \mid \mathbf{\Psi}, \mathbf{X} \sim \text{Cog.}(\mathbf{\Psi}, \mathbf{X}) \tag{23}$$

$$\mathbf{Y} \mid \mathbf{\Sigma}_{\Phi s}, \mathbf{\Sigma}_{\Phi t}, \mathbf{S}, \mathbf{\Psi}, \mathbf{U}, \mathbf{V} \sim \mathcal{MN}(0, \mathbf{\Sigma}_s, \mathbf{\Sigma}_t + \mathbf{X}\mathbf{U}^\top\mathbf{X} + \mathbf{\Psi}^\top\mathbf{V}\mathbf{\Psi}) \tag{24}$$

In this view we have arrived again at an RSA-type intuition, namely that while it may very challenging to know the true projection from $\mathbf{Y}$ to Ψ, mapping them on second-order statistics in time space can prove to be useful, especially as the dimensionality of $\mathbf{Y}$ (and hence the marginalized-over mapping) grows.

With the mapping marginalized, we can still perform prediction from the model by maximizing the likelihood of the cognitive parameters corresponding to new data given parameters estimated previously:

$$\mathbf{Y}_{new} \mid \Psi \sim \mathcal{MN}(\mathbf{M}, \hat{\Sigma}_s, \mathbf{C}) \tag{25}$$

$$\mathbf{M} = \mathbf{Y}_{old}(\hat{\Sigma}_t + \hat{\Psi}^\top\hat{\Psi})^{-1}(\hat{\Psi}^\top\Psi) \tag{26}$$

$$\mathbf{C} = \hat{\Sigma}_t + \Psi^\top\Psi - (\Psi^\top\hat{\Psi})(\hat{\Sigma}_t + \hat{\Psi}^\top\hat{\Psi})^{-1}(\hat{\Psi}^\top\Psi) \tag{27}$$

This maximization rotates the inner-products of the train and test sets into the same orientation. If the train and test sets have different numbers of TRs, we need to replace the temporal noise covariance matrix with a kernel function, but otherwise the derivation proceeds identically.

The resultant matrix-normal model mitigates the issues we identified previously: first, the only thing that scales with the number of voxels is the noise model rather than the mapping itself, allowing analysis to proceed using voxels directly assuming the noise model is efficient enough; second, it is targeted in that it automatically identifies the voxels most related to the cognitive model parameters; third, it is implicitly regularized via the priors on β and ℓ. As with all MNMs, it can also simultaneously handle both spatial and temporal noise in the fMRI signal.

2 Appendix B : Derivation of matrix normal identities

Consider the following three distributions:

$$\mathbf{X}_{ij} \sim \mathcal{MN}(\mathbf{A}_{ij}, \Sigma_{\mathbf{X}i}, \Sigma_{\mathbf{X}j}) \tag{28}$$

$$\mathbf{Y}_{jk} \sim \mathcal{MN}(\mathbf{B}_{jk}, \Sigma_{\mathbf{Y}j}, \Sigma_{\mathbf{Y}k}) \tag{29}$$

$$\mathbf{Z}_{ik} \mid \mathbf{X}_{ij}, \mathbf{Y}_{jk} \sim \mathcal{MN}(\mathbf{X}_{ij}\mathbf{Y}_{jk} + \mathbf{C}_{ik}, \Sigma_{\mathbf{Z}_i}, \Sigma_{\mathbf{Z}_k}) \tag{30}$$

We use lowercase subscripts to denote sizes, to make dimension constraints clearer. We first use the relationship between the matrix-normal and multivariate normal distribution to rewrite the densities in vectorized form. Next, we rewrite the vectorized product in the mean into kronecker form:

$$\mathrm{vec}(\mathbf{Z}_{ik}) \mid \mathbf{X}_{ij}, \mathbf{Y}_{jk} \sim \mathcal{N}(\mathrm{vec}(\mathbf{X}_{ij}\mathbf{Y}_{jk} + \mathbf{C}_{ik}), \Sigma_{\mathbf{Z}_k} \otimes \Sigma_{\mathbf{Z}_i}) \tag{31}$$

$$\mathrm{vec}(\mathbf{Z}_{ik}) \mid \mathbf{X}_{ij}, \mathbf{Y}_{jk} \sim \mathcal{N}((\mathbf{I}_k \otimes \mathbf{X}_{ij})\mathrm{vec}(\mathbf{Y}_{jk}) + \mathrm{vec}(\mathbf{C}_{ik}), \Sigma_{\mathbf{Z}_k} \otimes \Sigma_{\mathbf{Z}_i}) \tag{32}$$

We recognize the resultant distribution as following into the form $y \sim \mathcal{N}(Mx+b, \Sigma)$. Now, the standard gaussian marginalization identity (e.g. Bishop et al. 2006) can be applied:

$$\mathrm{vec}(\mathbf{Z}_{ik}) \mid \mathbf{X}_{ij} \sim \mathcal{N}((\mathbf{I}_k \otimes \mathbf{X}_{ij})\mathrm{vec}(\mathbf{B}_{jk}) + \mathrm{vec}(\mathbf{C}_{ik}), \Sigma_{\mathbf{Z}_k} \otimes \Sigma_{\mathbf{Z}_i} + (\mathbf{I}_k \otimes \mathbf{X}_{ij})(\Sigma_{\mathbf{Y}_k} \otimes \Sigma_{\mathbf{Y}_j})(\mathbf{I}_k \otimes \mathbf{X}_{ij})^\top) \tag{33}$$

We collect terms using the mixed-product property of kronecker products:

$$\mathrm{vec}(\mathbf{Z}_{ik}) \mid \mathbf{X}_{ij} \sim \mathcal{N}(\mathrm{vec}(\mathbf{X}_{ij}\mathbf{B}_{jk}) + \mathrm{vec}(\mathbf{C}_{ik}), \Sigma_{\mathbf{Z}_k} \otimes \Sigma_{\mathbf{Z}_i} + \Sigma_{\mathbf{Y}_k} \otimes \mathbf{X}_{ij}\Sigma_{\mathbf{Y}_j}{\mathbf{X}_{ij}}^\top) \tag{34}$$

Now, we can see that the marginal density is a matrix-variate normal only if $\Sigma_{\mathbf{Z}_k} = \Sigma_{\mathbf{Y}_k}$ – that is, the variable we're marginalizing over has the same covariance in the

dimension we are *not* marginalizing over as the marginal density. Otherwise the density is well-defined but not matrix-normal. If we let $\Sigma_k := \Sigma_{\mathbf{Z}_k} = \Sigma_{\mathbf{Y}_k}$, then we can factor out that term and rewrite the marginal density as a matrix normal:

$$\text{vec}(\mathbf{Z}_{ik}) \mid \mathbf{X}_{ij} \sim \mathcal{N}(\text{vec}(\mathbf{X}\mathbf{B}_{jk}) + \text{vec}(\mathbf{C}_{ik}), \Sigma_k \otimes \Sigma_{\mathbf{Z}_i} + \Sigma_k \otimes \mathbf{X}\Sigma_{\mathbf{Y}_j}\mathbf{X}^\top) \tag{35}$$

$$\text{vec}(\mathbf{Z}_{ik}) \mid \mathbf{X}_{ij} \sim \mathcal{N}(\text{vec}(\mathbf{X}\mathbf{B}_{jk}) + \text{vec}(\mathbf{C}_{ik}), \Sigma_k \otimes (\Sigma_{\mathbf{Z}_i} + \mathbf{X}\Sigma_{\mathbf{Y}_j}\mathbf{X}^\top)) \tag{36}$$

$$\mathbf{Z}_{ik} \mid \mathbf{X}_{ij} \sim \mathcal{MN}(\mathbf{X}\mathbf{B}_{jk} + \mathbf{C}_{ik}, \Sigma_{\mathbf{Z}_i} + \mathbf{X}\Sigma_{\mathbf{Y}_j}\mathbf{X}^\top, \Sigma_k) \tag{37}$$

Unlike the multivariate normal case, we can apply the same identity over either $\mathbf{X}$ or $\mathbf{Y}$, since if $\mathbf{X} \sim \mathcal{MN}(M, U, V)$ then $\mathbf{X}^\top \sim \mathcal{MN}(M^\top, V, U)$. We write it directly below:

$$\mathbf{Z}^\top{}_{ik} \mid \mathbf{X}_{ij}, \mathbf{Y}_{jk} \sim \mathcal{MN}(\mathbf{Y}_{jk}{}^\top \mathbf{X}_{ij}{}^\top + \mathbf{C}^\top{}_{ik}, \Sigma_{\mathbf{Z}_k}, \Sigma_{\mathbf{Z}_i}) \tag{38}$$

$$\text{let } \Sigma_i := \Sigma_{\mathbf{Z}_i} = \Sigma_{\mathbf{X}_i} \tag{39}$$

$$\cdots \tag{40}$$

$$\mathbf{Z}^\top{}_{ik} \mid \mathbf{Y}_{jk} \sim \mathcal{MN}(\mathbf{A}_{jk}{}^\top \mathbf{X}_{ij}{}^\top + \mathbf{C}^\top{}_{ik}, \Sigma_{\mathbf{Z}_k} + \mathbf{Y}^\top\Sigma_{\mathbf{Y}_j}\mathbf{Y}, \Sigma_{\mathbf{Z}_i}) \tag{41}$$

$$\mathbf{Z}_{ik} \mid \mathbf{Y}_{jk} \sim \mathcal{MN}(\mathbf{X}_{ij}\mathbf{A}_{jk} + \mathbf{C}_{ik}, \Sigma_{\mathbf{Z}_i}, \Sigma_{\mathbf{Z}_k} + \mathbf{Y}^\top\Sigma_{\mathbf{Y}_j}\mathbf{Y}) \tag{42}$$

Next, we do the same for the partitioned gaussian identity. First two vectorized matrix-normals that form our partition:

$$\mathbf{X}_{ij} \sim \mathcal{MN}(\mathbf{A}_{ij}, \Sigma_i, \Sigma_j) \to \text{vec}[\mathbf{X}_{ij}] \sim \mathcal{N}(\text{vec}[\mathbf{A}_{ij}], \Sigma_j \otimes \Sigma_i) \tag{43}$$

$$\mathbf{Y}_{ik} \sim \mathcal{MN}(\mathbf{B}_{ik}, \Sigma_i, \Sigma_k) \to \text{vec}[\mathbf{Y}_{ik}] \sim \mathcal{N}(\text{vec}[\mathbf{B}_{ik}], \Sigma_k \otimes \Sigma_i) \tag{44}$$

$$\begin{bmatrix}\text{vec}[\mathbf{X}_{ij}] \\ \text{vec}[\mathbf{Y}_{ik}]\end{bmatrix} \sim \mathcal{N}\left(\text{vec}\begin{bmatrix}\mathbf{A}_{ij} \\ \mathbf{B}_{ik}\end{bmatrix}, \begin{bmatrix}\Sigma_j \otimes \Sigma_i & \Sigma_{jk} \otimes \Sigma_i \\ \Sigma_{kj} \otimes \Sigma_i & \Sigma_k \otimes \Sigma_i\end{bmatrix}\right) \tag{45}$$

We apply the standard partitioned Gaussian identity and simplify using the properties of the vec operator and the mixed product property of kronecker products:

$$\text{vec}[\mathbf{X}_{ij}] \mid \text{vec}[\mathbf{Y}_{ik}] \sim \mathcal{N}(\text{vec}[\mathbf{A}_{ij}] + (\Sigma_{jk} \otimes \Sigma_i)(\Sigma_k^{-1} \otimes \Sigma_i^{-1})(\text{vec}[\mathbf{Y}_{ik}] - \text{vec}[\mathbf{B}_{ik}]), \tag{46}$$

$$\Sigma_j \otimes \Sigma_i - (\Sigma_{jk} \otimes \Sigma_i)(\Sigma_k^{-1} \otimes \Sigma_i^{-1})(\Sigma_{kj} \otimes \Sigma_i)) \tag{47}$$

$$= \mathcal{N}(\text{vec}[\mathbf{A}_{ij}] + (\Sigma_{jk}\Sigma_k^{-1} \otimes \Sigma_i\Sigma_i^{-1})(\text{vec}[\mathbf{Y}_{ik}] - \text{vec}[\mathbf{B}_{ik}]), \tag{48}$$

$$\Sigma_j \otimes \Sigma_i - (\Sigma_{jk}\Sigma_k^{-1}\Sigma_{kj} \otimes \Sigma_i\Sigma_i^{-1}\Sigma_i)) \tag{49}$$

$$= \mathcal{N}(\text{vec}[\mathbf{A}_{ij}] + (\Sigma_{jk}\Sigma_k^{-1} \otimes \mathbf{I})(\text{vec}[\mathbf{Y}_{ik}] - \text{vec}[\mathbf{B}_{ik}]), \tag{50}$$

$$\Sigma_j \otimes \Sigma_i - (\Sigma_{jk}\Sigma_k^{-1}\Sigma_{kj} \otimes \Sigma_i) \tag{51}$$

$$= \mathcal{N}(\text{vec}[\mathbf{A}_{ij}] + \text{vec}[\mathbf{Y}_{ik} - \mathbf{B}_{ik}\Sigma_k^{-1}\Sigma_{kj}], (\Sigma_j - \Sigma_{jk}\Sigma_k^{-1}\Sigma_{kj}) \otimes \Sigma_i) \tag{52}$$

Next, we recognize that this multivariate gaussian is equivalent to the following matrix variate gaussian:

$$\mathbf{X}_{ij} \mid \mathbf{Y}_{ik} \sim \mathcal{MN}(\mathbf{A}_{ij} + (\mathbf{Y}_{ik} - \mathbf{B}_{ik})\Sigma_k^{-1}\Sigma_{kj}, \Sigma_i, \Sigma_j - \Sigma_{jk}\Sigma_k^{-1}\Sigma_{kj}) \tag{53}$$

The conditional in the other direction can be written by working through the same algebra:

$$\mathbf{Y}_{ik} \mid \mathbf{X}_{ij} \sim \mathcal{MN}(\mathbf{B}_{ik} + (\mathbf{X}_{ij} - \mathbf{A}_{ij})\Sigma_j^{-1}\Sigma_{jk}, \Sigma_i, \Sigma_k - \Sigma_{kj}\Sigma_j^{-1}\Sigma_{jk}) \tag{54}$$

Finally, vertical rather than horizontal concatenation (yielding a partitioned row rather than column covariance) can be written by recognizing the behavior of the matrix normal under transposition:

$$\mathbf{X}^\top{}_{ji} \mid \mathbf{Y}^\top{}_{ki} \sim \mathcal{MN}(\mathbf{A}^\top{}_{ji} + \Sigma_{jk}\Sigma_k^{-1}(\mathbf{Y}^\top{}_{ki} - \mathbf{B}^\top{}_{ki}), \Sigma_j - \Sigma_{jk}\Sigma_k^{-1}\Sigma_{kj}, \Sigma_i) \tag{55}$$

$$\mathbf{Y}^\top{}_{ki} \mid \mathbf{X}^\top{}_{ji} \sim \mathcal{MN}(\mathbf{B}^\top{}_{ki} + \Sigma_{kj}\Sigma_j^{-1}(\mathbf{X}^\top{}_{ji} - \mathbf{A}^\top{}_{ji}), \Sigma_k - \Sigma_{kj}\Sigma_j^{-1}\Sigma_{jk}, \Sigma_i) \tag{56}$$

3 Appendix C : Expectation Conditional Maximization (ECM) derivation for Matrix-Normal Shared Response Model

The Q function, marginalized $\mathbf{W}$

$$\mathbf{X} \sim \mathcal{MN}(\mathbf{WS} + \mathbf{b1}^\top, \rho \otimes \Sigma_v, \Sigma_t) \tag{57}$$
$$\mathbf{S} \sim \mathcal{MN}(0, \mathbf{I}, \Sigma_t) \tag{58}$$
$$\mathbf{W} \sim \mathcal{MN}(0, \rho \otimes \Sigma_v, \mathbf{I}) \tag{59}$$

$$\begin{aligned}
\mathcal{L} := \mathbb{E}_{p(\mathbf{W}|\mathbf{X},\theta')} \log p(\mathbf{X}, \mathbf{W} \mid \theta) =& \frac{1}{2}\mathbb{E}\Big[nv \log|\Sigma_t^{-1}| + tv \log|\rho^{-1}| + tn \log|\Sigma_v^{-1}| \\
&- \mathrm{Tr}\left[\Sigma_t^{-1}(\mathbf{X} - \mathbf{WS} - \mathbf{b1}^\top)^\top(\rho \otimes \Sigma_v)^{-1}(\mathbf{X} - \mathbf{WS} - \mathbf{b1}^\top)\right] \\
&+ kv \log|\rho^{-1}| + kn \log|\Sigma_v^{-1}| - \mathrm{Tr}\left[\Sigma_w^{-1}\mathbf{W}^\top(\rho \otimes \Sigma_v)^{-1}\mathbf{W}\right] \\
&+ k \log|\Sigma_t^{-1}| - \mathrm{Tr}[\Sigma_t^{-1}\mathbf{S}^\top\mathbf{S}]\Big] + \mathrm{const.}_\theta \qquad (60) \\
=& \frac{1}{2}\left[(nv + k) \log|\Sigma_t^{-1}| + v(k + t) \log|\rho^{-1}| + n(k + t) \log|\Sigma_v^{-1}|\right. \\
&- \mathrm{Tr}\left[\Sigma_t^{-1}(\mathbf{X} - \mathbf{b1}^\top - \mathbf{W'S})^\top(\rho \otimes \Sigma_v)^{-1}(\mathbf{X} - \mathbf{b1}^\top - \mathbf{W'S})\right] \\
&- \mathrm{Tr}\left[\Sigma_w^{-1}\mathbf{W'}^\top(\rho \otimes \Sigma_v)^{-1}\mathbf{W'}\right] - \mathrm{Tr}[\rho^{-1}\rho'] \,\mathrm{Tr}[\Sigma_v^{-1}\Sigma_v'] \,\mathrm{Tr}[\Sigma_w'\mathbf{S}^\top\Sigma_t^{-1}\mathbf{S}] \\
&\left.- \mathrm{Tr}[\rho^{-1}\rho'] \,\mathrm{Tr}[\Sigma_v^{-1}\Sigma_v'] \,\mathrm{Tr}[\Sigma_w^{-1}\Sigma_w'] - \mathrm{Tr}[\Sigma_t^{-1}\mathbf{S}^\top\mathbf{S}]\right] + \mathrm{const.}_\theta \\
& \qquad (61)
\end{aligned}$$

The sufficient statistics are:

$$\mathbf{W} \mid \mathbf{X}, \theta \sim \mathcal{MN}\left(\mathbf{W'}, \rho_w' \otimes \Sigma_{vw}', \Sigma_w'\right) \tag{62}$$
$$\Sigma_w' := \mathbf{I} - S(\Sigma_t + \mathbf{S}^\top\mathbf{S})^{-1}\mathbf{S}^\top = (\mathbf{I} + \mathbf{S}\Sigma_t^{-1}\mathbf{S}^\top)^{-1} \tag{63}$$
$$\rho_w' := \rho_w \tag{64}$$
$$\Sigma_{vw}' := \Sigma_{vw} \tag{65}$$
$$\mathbf{W'} = (\mathbf{X} - \mathbf{b1}^\top)\mathbf{S}^\top(\Sigma_t + \mathbf{S}^\top\mathbf{S})^{-1} = (\mathbf{X} - \mathbf{b1}^\top)\Sigma_t^{-1}\mathbf{S}^\top\Sigma_w' \tag{66}$$

3.1 Gradients for $\mathbf{S}$

$$\begin{aligned}
\mathrm{d}_{\mathbf{S}}\mathcal{L} =& \frac{1}{2}\mathrm{d}\Big[-\operatorname{Tr}\Big[\Sigma_t^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})^\top(\rho\otimes\Sigma_v)^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})\Big] && (67)\\
& -\operatorname{Tr}[\rho^{-1}\rho'_w]\operatorname{Tr}[\Sigma_v^{-1}\Sigma'_{vw}]\operatorname{Tr}[\Sigma'_w\mathbf{S}\Sigma_t^{-1}\mathbf{S}]-\operatorname{Tr}[\Sigma_t^{-1}\mathbf{S}^\top\mathbf{S}]\Big] && (68)\\
=& \frac{1}{2}\Big[-2\operatorname{Tr}\Big[\Sigma_t^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})^\top(\rho\otimes\Sigma_v)^{-1}(\mathbf{W}'\mathrm{d}\mathbf{S})\Big] && (69)\\
& -2\operatorname{Tr}[\rho^{-1}\rho'_w]\operatorname{Tr}[\Sigma_v^{-1}\Sigma'_{vw}]\operatorname{Tr}[\Sigma'_w\mathrm{d}\mathbf{S}\Sigma_t^{-1}\mathbf{S}^\top]^{-1}\Sigma'_w]-2\operatorname{Tr}[\Sigma_t^{-1}\mathbf{S}^\top\mathrm{d}\mathbf{S}]\Big] && (70)\\
=& -\operatorname{Tr}\Big[\Sigma_t^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})^\top(\rho\otimes\Sigma_v)^{-1}\mathbf{W}'\mathrm{d}\mathbf{S}\Big] && (71)\\
& -\operatorname{Tr}[\rho^{-1}\rho'_w]\operatorname{Tr}[\Sigma_v^{-1}\Sigma'_{vw}]\operatorname{Tr}[\Sigma_t^{-1}\mathbf{S}^\top\Sigma'_w\mathrm{d}\mathbf{S}]-\operatorname{Tr}[\Sigma_t^{-1}\mathbf{S}^\top\mathrm{d}\mathbf{S}] && (72)\\
& && (73)\\
\frac{\partial\mathcal{L}}{\partial\mathbf{S}} =& \mathbf{W}'^\top(\rho^{-1}\otimes\Sigma_v^{-1})(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})\Sigma_t^{-1}-\operatorname{Tr}[\rho^{-1}\rho'_w]\operatorname{Tr}[\Sigma_v^{-1}\Sigma'_{vw}]\Sigma'_w\mathbf{S}\Sigma_t^{-1}-\mathbf{S}\Sigma_t^{-1} && (74)\\
=& \mathbf{W}'^\top(\rho^{-1}\otimes\Sigma_v^{-1})(\mathbf{X}-\mathbf{b}\mathbf{1}^\top)-(\mathbf{W}'^\top(\rho^{-1}\otimes\Sigma_v^{-1})\mathbf{W}'\mathbf{S}-\operatorname{Tr}[\rho^{-1}\rho'_w]\operatorname{Tr}[\Sigma_v^{-1}\Sigma'_{vw}]\Sigma'_w\mathbf{S}-\mathbf{S} && (75)\\
=& \mathbf{W}'^\top(\rho^{-1}\otimes\Sigma_v^{-1})(\mathbf{X}-\mathbf{b}\mathbf{1}^\top)-(\mathbf{W}'^\top(\rho^{-1}\otimes\Sigma_v^{-1})\mathbf{W}'+\operatorname{Tr}[\rho^{-1}\rho'_w]\operatorname{Tr}[\Sigma_v^{-1}\Sigma'_{vw}]\Sigma'_w+1)\mathbf{S} && (76)\\
& && (77)\\
\widehat{\mathbf{S}} =& (\mathbf{W}'^\top(\rho^{-1}\otimes\Sigma_v^{-1})\mathbf{W}'+\operatorname{Tr}[\rho^{-1}\rho'_w]\operatorname{Tr}[\Sigma_v^{-1}\Sigma'_{vw}]\Sigma'_w+1)^{-1}\mathbf{W}'^\top(\rho^{-1}\otimes\Sigma_v^{-1})(\mathbf{X}-\mathbf{b}\mathbf{1}^\top) && (78)
\end{aligned}$$

3.2 Gradients for $\mathbf{b}$

$$\begin{aligned}
\mathrm{d}_{\mathbf{b}}\mathcal{L} =& -\frac{1}{2}\operatorname{Tr}\Big[\Sigma_t^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})^\top(\rho\otimes\Sigma_v)^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})\Big] && (79)\\
=& -\operatorname{Tr}\Big[\mathbf{1}^\top\Sigma_t^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})^\top(\rho\otimes\Sigma_v)^{-1}\mathrm{d}\mathbf{b}\Big] && (80)\\
\frac{\partial\mathcal{L}}{\partial\mathbf{b}} =& (\rho\otimes\Sigma_v)^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})\Sigma_t^{-1}\mathbf{1} && (81)\\
& && (82)\\
0 =& (\rho\otimes\Sigma_v)^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})\Sigma_t^{-1}\mathbf{1} && (83)\\
& \mathbf{b}\mathbf{1}^\top\Sigma_t^{-1}\mathbf{1}=(\mathbf{X}-\mathbf{W}'\mathbf{S})\Sigma_t^{-1}\mathbf{1} && (84)\\
& \hat{\mathbf{b}}=\frac{(\mathbf{X}-\mathbf{W}'\mathbf{S})\Sigma_t^{-1}\mathbf{1}}{\sum\Sigma_t^{-1}} && (85)
\end{aligned}$$

3.3 Gradients for Σ_t

$$
\begin{aligned}
\mathrm{d}_{\Sigma_t^{-1}}\mathcal{L} =& \frac{1}{2}\mathrm{d}\left[(nv+k)\log|\Sigma_t^{-1}| + v(k+t)\log|\rho^{-1}| + n(k+t)\log|\Sigma_v^{-1}|\right. && (86)\\
&- \mathrm{Tr}\left[\Sigma_t^{-1}(\mathbf{X}-\mathbf{b1}^\top-\mathbf{W'S})^\top(\rho\otimes\Sigma_v)^{-1}(\mathbf{X}-\mathbf{b1}^\top-\mathbf{W'S})\right] && (87)\\
&- \mathrm{Tr}\left[\Sigma_w^{-1}\mathbf{W'}^\top(\rho\otimes\Sigma_v)^{-1}\mathbf{W'}\right] - \mathrm{Tr}[\rho^{-1}\rho']\,\mathrm{Tr}[\Sigma_v^{-1}\Sigma_v']\,\mathrm{Tr}[\Sigma_w'\mathbf{S}\Sigma_t^{-1}\mathbf{S}^\top] && (88)\\
&\left.- \mathrm{Tr}[\rho^{-1}\rho']\,\mathrm{Tr}[\Sigma_v^{-1}\Sigma_v']\,\mathrm{Tr}[\Sigma_w^{-1}\Sigma_w'] - \mathrm{Tr}[\Sigma_t^{-1}\mathbf{S}^\top\mathbf{S}]\right] && (89)\\
=& \frac{1}{2}\left[(nv+k)\,\mathrm{Tr}[\Sigma_t\mathrm{d}\Sigma_t^{-1}] - \mathrm{Tr}\left[\mathrm{d}\Sigma_t^{-1}(\mathbf{X}-\mathbf{b1}^\top-\mathbf{W'S})^\top(\rho\otimes\Sigma_v)^{-1}(\mathbf{X}-\mathbf{b1}^\top-\mathbf{W'S})\right]\right. && (90)\\
&\left.- \mathrm{Tr}[\rho^{-1}\rho']\,\mathrm{Tr}[\Sigma_v^{-1}\Sigma_v']\,\mathrm{Tr}[\Sigma_w'\mathbf{S}\mathrm{d}\Sigma_t^{-1}\mathbf{S}^\top] - \mathrm{Tr}[\mathrm{d}\Sigma_t^{-1}\mathbf{S}^\top\mathbf{S}]\right] && (91)\\
& && (92)\\
\frac{\partial\mathcal{L}}{\partial\Sigma_t^{-1}} =& \frac{1}{2}\left[(nv+k)\Sigma_t - (\mathbf{X}-\mathbf{b1}^\top-\mathbf{W'S})^\top(\rho\otimes\Sigma_v)^{-1}(\mathbf{X}-\mathbf{b1}^\top-\mathbf{W'S})\right. && (93)\\
&\left.- \mathrm{Tr}[\rho^{-1}\rho']\,\mathrm{Tr}[\Sigma_v^{-1}\Sigma_v']\mathbf{S}^\top\Sigma_w'\mathbf{S} - \mathbf{S}^\top\mathbf{S}\right] && (94)\\
& && (95)\\
\widehat{\Sigma_t^{-1}} =& \left(\frac{1}{nv+k}(\mathbf{X}-\mathbf{b1}^\top-\mathbf{W'S})^\top(\rho\otimes\Sigma_v)^{-1}(\mathbf{X}-\mathbf{b1}^\top-\mathbf{W'S})\right. && (96)\\
&\left.- \mathrm{Tr}[\rho^{-1}\rho']\,\mathrm{Tr}[\Sigma_v^{-1}\Sigma_v']\mathbf{S}^\top\Sigma_w'\mathbf{S} - \mathbf{S}^\top\mathbf{S}\right)^{-1} && (97)
\end{aligned}
$$

3.4 Gradients for Σ_v

Here again we assume ρ is diagonal, in which case:

$$
\begin{aligned}
\mathrm{d}_{\Sigma_v^{-1}}\mathcal{L} =& \mathrm{d}\frac{1}{2}\left[(nv+k)\log|\Sigma_t^{-1}| + v(k+t)\log|\rho^{-1}| + n(k+t)\log|\Sigma_v^{-1}|\right. && (98)\\
&- \mathrm{Tr}\left[\Sigma_t^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top - \mathbf{W}'\mathbf{S})^\top(\rho\otimes\Sigma_v)^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})\right] && (99)\\
&- \mathrm{Tr}\left[\Sigma_w^{-1}\mathbf{W}'^\top(\rho\otimes\Sigma_v)^{-1}\mathbf{W}'\right] - \mathrm{Tr}[\rho^{-1}\rho']\,\mathrm{Tr}[\Sigma_v^{-1}\Sigma_v']\,\mathrm{Tr}[\Sigma_w'\mathbf{S}^\top\Sigma_t^{-1}\mathbf{S}] && (100)\\
&\left.- \mathrm{Tr}[\rho^{-1}\rho']\,\mathrm{Tr}[\Sigma_v^{-1}\Sigma_v']\,\mathrm{Tr}[\Sigma_w^{-1}\Sigma_w'] - \mathrm{Tr}[\Sigma_t^{-1}\mathbf{S}^\top\mathbf{S}]\right] && (101)\\
&= \frac{1}{2}\left[n(k+t)\Sigma_v\mathrm{d}\Sigma_v^{-1}\right. && (102)\\
&- \sum_j \tau_j\,\mathrm{Tr}\left[\Sigma_t^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})^\top\mathrm{d}\Sigma_v^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})\right] && (103)\\
&\left.- \sum_j \tau_j\,\mathrm{Tr}\left[\Sigma_w^{-1}\mathbf{W}'^\top\mathrm{d}\Sigma_v^{-1}\mathbf{W}'\right] - \mathrm{Tr}[\rho^{-1}\rho']\,\mathrm{Tr}[\Sigma_v'\mathrm{d}\Sigma_v^{-1}]\,\mathrm{Tr}[\Sigma_w'(\mathbf{I}+\mathbf{S}^\top\Sigma_t^{-1}\mathbf{S})]\right. && (104)\\
& && (105)\\
\frac{\partial\mathcal{L}}{\partial\Sigma_v^{-1}} &= \frac{1}{2}\left[n(k+t)\Sigma_v\right. && (106)\\
&- \sum_j \tau_j(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})\Sigma_t^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})^\top && (107)\\
&\left.- \sum_j \tau_j\mathbf{W}'\Sigma_w^{-1}\mathbf{W}'^\top - \mathrm{Tr}[\rho^{-1}\rho']\,\mathrm{Tr}[\Sigma_w'(\mathbf{I}+\mathbf{S}\Sigma_t^{-1}\mathbf{S}^\top)]\Sigma_v'\right] && (108)\\
& && (109)\\
\widehat{\Sigma_v^{-1}} &= \left(\frac{1}{n(k+t)}\sum_j \tau_j(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})\Sigma_t^{-1}(\mathbf{X}-\mathbf{b}\mathbf{1}^\top-\mathbf{W}'\mathbf{S})^\top\right. && (110)\\
&\left.+ \sum_j \tau_j\mathbf{W}'\Sigma_w^{-1}\mathbf{W}'^\top - \mathrm{Tr}[\rho^{-1}\rho']\,\mathrm{Tr}[\Sigma_w'(\mathbf{I}+\mathbf{S}\Sigma_t^{-1}\mathbf{S}^\top)]\Sigma_v'\right)^{-1} && (111)
\end{aligned}
$$

3.5 Constrained covariances

For template constraints (e.g. diagonal, blocked, banded), we can elementwise-multiply the gradient by a template matrix, and construct the constrained update.

Algorithm 1 Solve $x = (L_0 \otimes L_1 \otimes \cdots \otimes L_{n-1})\backslash y$

1: **Input:** vector y, matrices $L_0, L_1, \cdots L_{n-1}$
2: **Output:** vector x
3: **if** n == 1 **then**
4: return matrix_triangular_solve(L_0, y)
5: **else**
6: $x = y$
7: na = dim(L_0)
8: nb = dim$(L_1)\times$ dim$(L_2)\times \cdots$ dim(L_{n-1})
9: **for** $i = 0$ **to** $na - 1$ **do**
10: $t = x[i * nb : (i + 1) * nb]/L_0[i, i]$
11: $x[i * nb : (i + 1) * nb] = (L_1 \otimes \cdots \otimes L_{n-1})\backslash t$
12: **for** $j = i + 1$ **to** $na - 1$ **do**
13: $x[j * nb : (j + 1) * nb] -= L_0[j, i] * t$
14: **end for**
15: **end for**
16: return x
17: **end if**

4 Appendix D : Algorithm for solving kronecker factored matrices

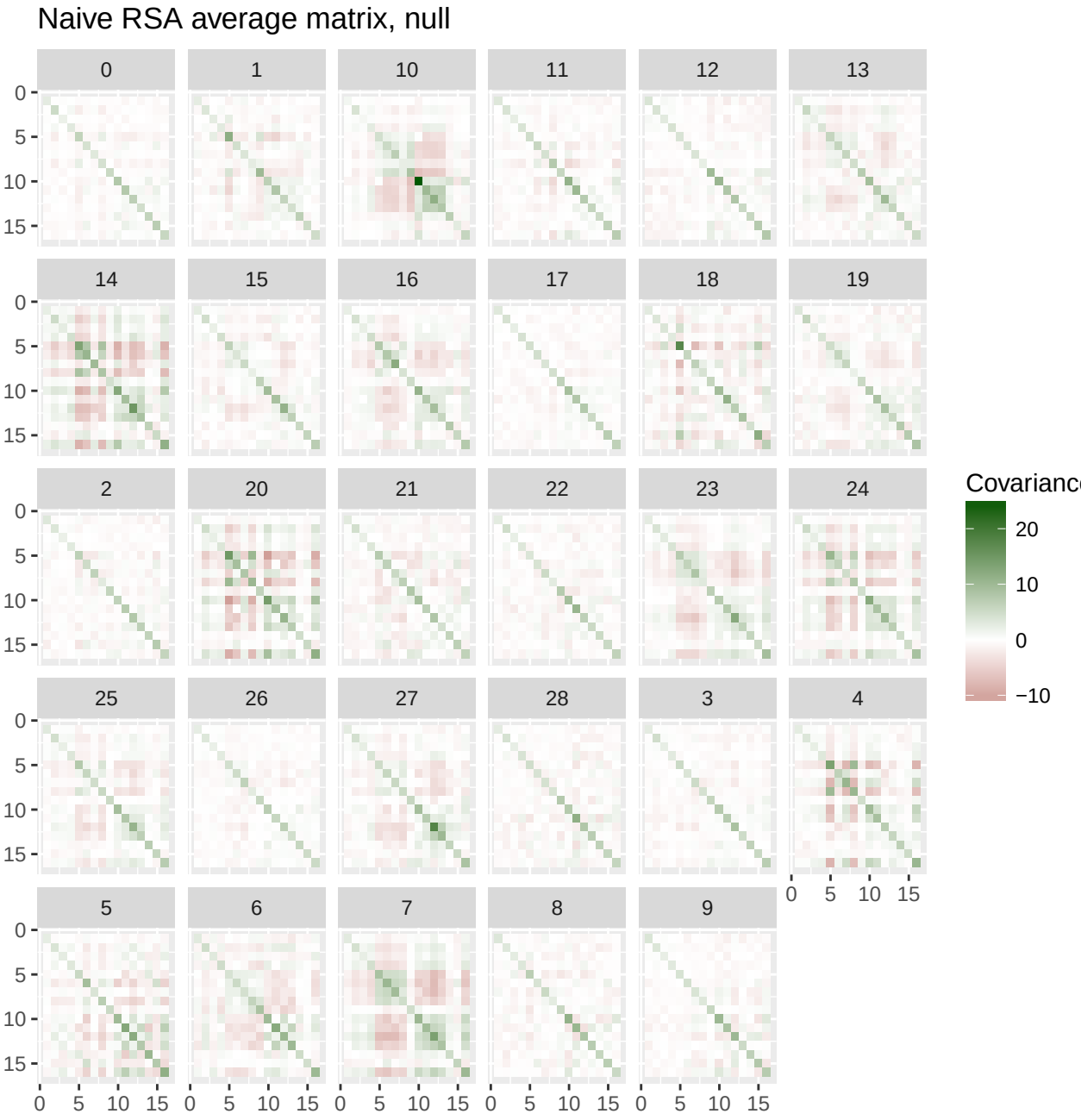
In algorithm 1, we show how to efficiently solve for a lower triangular matrix that is the kronecker product of smaller lower triangular matrices.

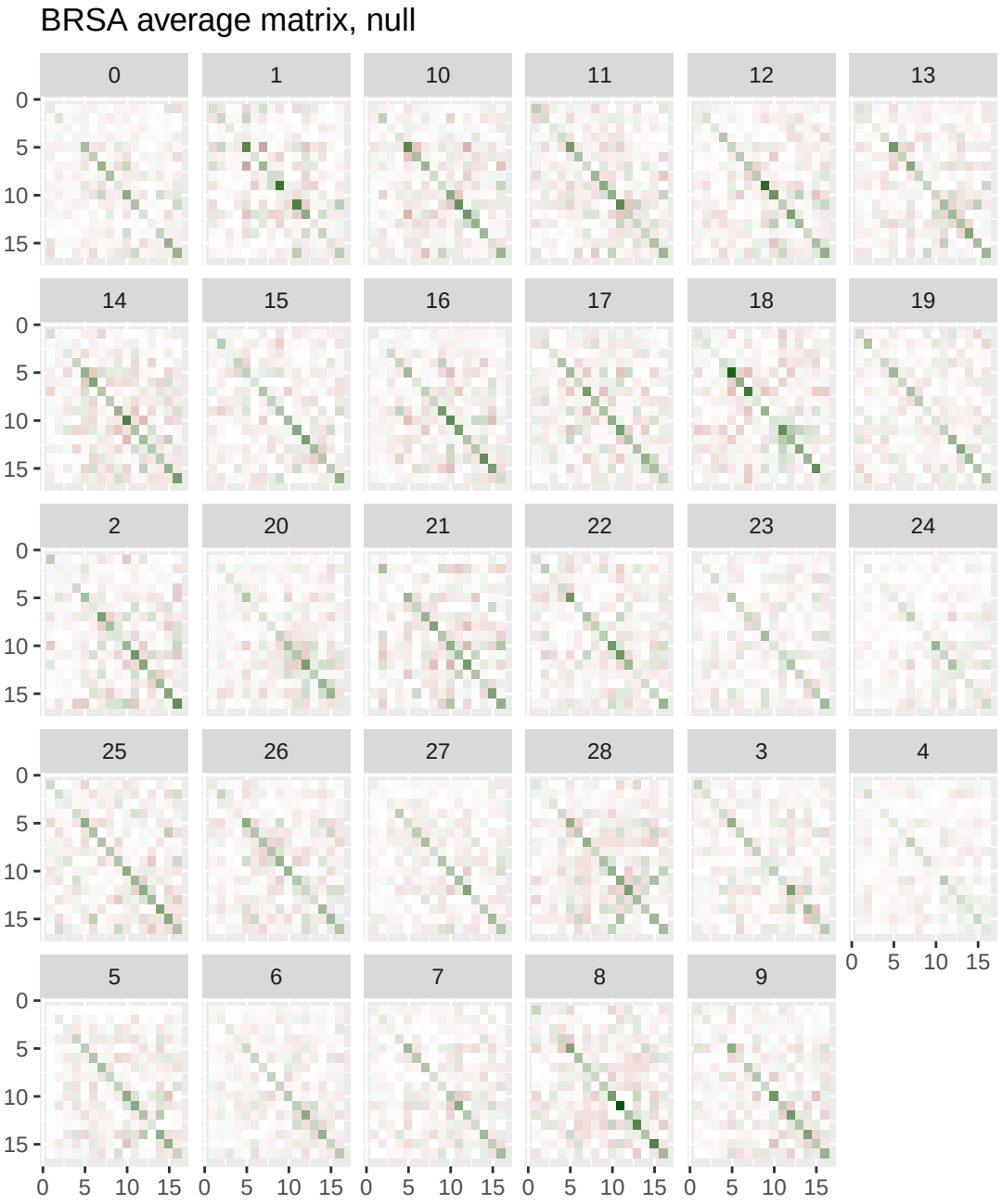
Since the cholesky of a kronecker product is the kronecker product of its cholesky factors, we avoid computing the cholesky factorization of a large matrix and instead only cholesky factorize the individual factors. Algorithm 1 is recursive: line 11 calls the same function but with one less kronecker factor. The masked variant of the algorithm is similar except for lines 4, 11 and 13. Lines 4 and 11 now perform matrix solves with a mask. Line 13 multiplies $L_0[j, i]$ not with t but with $t' = (L_1 \otimes \cdots \otimes L_{n-1}) \cdot x[i * nb : (i + 1) * nb]$. t and t' are identical when no rows and columns are masked, but differ when some of them are masked. Solving $\Sigma^{-1}\mathbf{X}$ now involves the following steps - (1) Cholesky factorize the kronecker factor matrices. (2) Use algorithm 1 to solve $Z = (L_0 \otimes L_1 \otimes \cdots \otimes L_{n-1})\backslash X$. (3) Apply the corresponding upper triangular variant to solve $(L_0 \otimes L_1 \otimes \cdots \otimes L_{n-1})^T\backslash Z$.

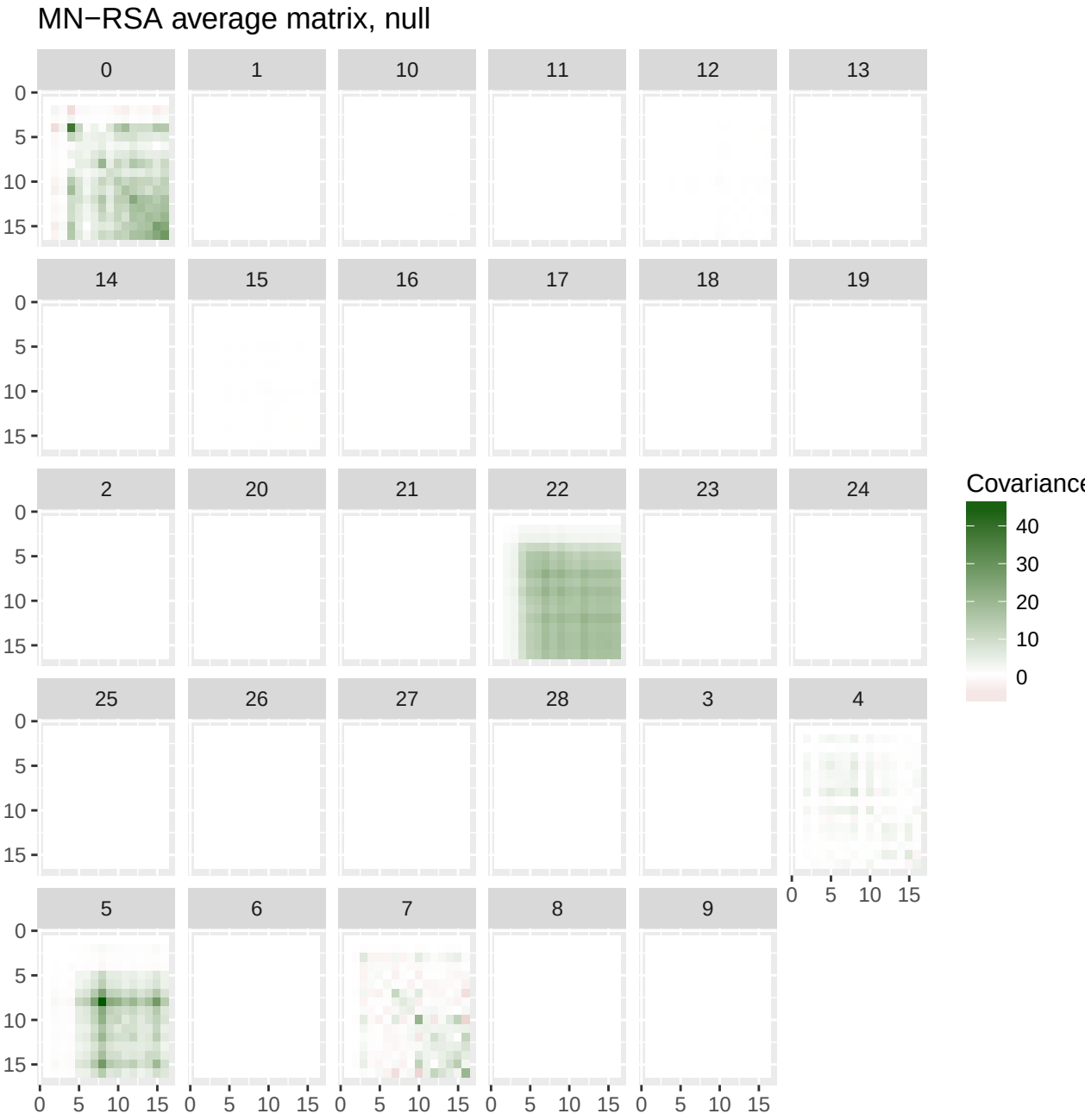
We can calculate log-determinant for kronecker products as follows. After cholesky factorization, $\log|\Sigma| = 2 \cdot \sum_i((\log|L_i|)(\prod_{j,j\neq i} \dim(L_j)))$. $\log|L_i|$ is easy to calculate for a triangular matrix L_i. For masked kronecker product, the latter product term in the previous expression is replaced by counting the number of valid rows/columns corresponding to that element in the mask.

5 Appendix E : Additional null hypothesis RSA results

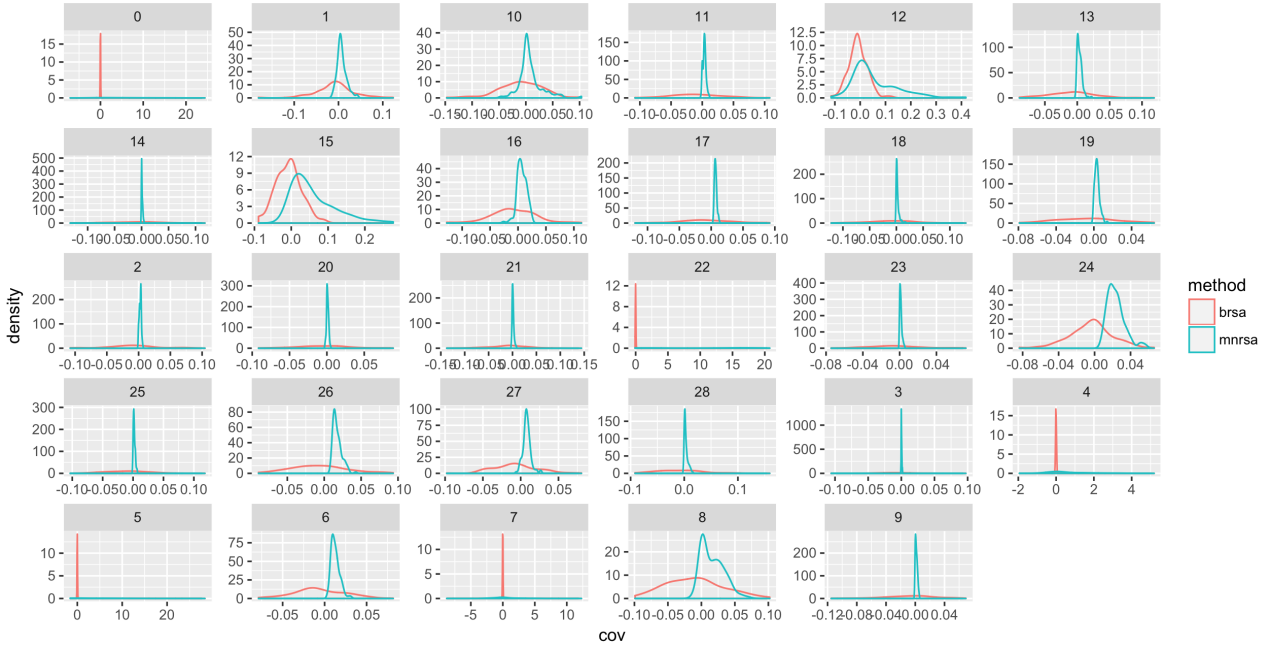
First, we show RSA matrices under the null hypothesis for all subjects and methods:




BRSA average matrix, null
0
1
10
11
12
13
14
15
16
17
18
19
2
20
21
22
23
24
25
26
27
28
3
4
5
6
7
8
9
0
5
10
15
Covariance
0.4
0.2
0.0

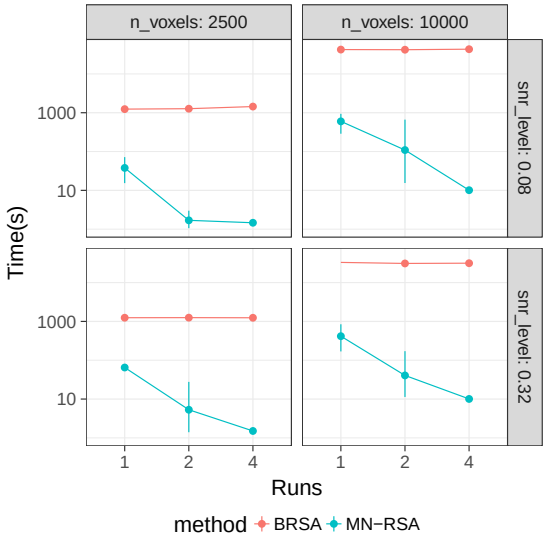


Notice that only for MN-RSA most of the covariances are noticeably degenerate. This is not a scaling effect on the figure driven by the color bar, but an effect on the underlying data, as we can see in the distribution of values in the covariance matrix for BRSA and MN-RSA:



6 Appendix F : timing figures for BRSA and MN-RSA

Experiment details mentioned in main text. Note time on the log scale.



Bibliography

E. Simony, C. J. Honey, J. Chen, O. Lositsky, Y. Yeshurun, A. Wiesel, and U. Hasson. Dynamic reconfiguration of the default mode network during narrative comprehension.

Nature Communications, 7(May 2015):12141, jul 2016.

B. M. Turner, B. U. Forstmann, E.-J. Wagenmakers, S. D. Brown, P. B. Sederberg, and M. Steyvers. A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, 72:193–206, may 2013.

B. M. Turner, P. B. Sederberg, and J. L. McClelland. Bayesian analysis of simulation-based models. *Journal of Mathematical Psychology*, 2014.

B. M. Turner, L. van Maanen, and B. U. Forstmann. Informing cognitive abstractions through neuroimaging: The neural drift diffusion model. *Psychological Review*, 122 (2):312–336, 2015.

B. M. Turner, C. A. Rodriguez, T. M. Norcia, S. M. McClure, and M. Steyvers. Why more is better: Simultaneous modeling of EEG, fMRI, and behavioral data. *NeuroImage*, 128:96–115, mar 2016.