Varimax gradient

Yves Bernaerts

1 Introduction

In neuroscience contexts of state space modelling, applying a varimax objective to loading matrices $\mathbf{C} \in \mathbb{R}^{N \times L}$ (N neurons, L latents) means to find a rotation matrix $\mathbf{R} \in \mathbb{R}^{L \times L}$ so as to maximise the variance of squared elements in each column of $\overline{\mathbf{C}} = \mathbf{CR}$:

$$\mathcal{V} = \sum_l \left(\frac{1}{N} \sum_n \bar{c}_{nl}^4 - \frac{1}{N^2} \sum_n \bar{c}_{nl}^2 \right). \tag{1}$$

2 Gradient derivation

We would like to maximize the objective in 1 w.r.t. rotation matrix $\mathbf{R}$ with the constraint that $\mathbf{R}$ is an orthogonal matrix, that is $\mathbf{RR}^\top = \mathbf{I}$.

2.1 Unprojected gradient ascent

First, we derive a gradient without constraints applied to $\mathbf{R}$. Let us first derive the change of $\mathcal{V}$ w.r.t. scalar matrix elements r_{ij}. For that, we need the chain rule:

$$\frac{\partial \mathcal{V}}{\partial r_{ij}} = \sum_{n'l'} \frac{\partial \mathcal{V}}{\partial \bar{c}_{n'l'}} \frac{\partial \bar{c}_{n'l'}}{\partial r_{ij}}.$$

By inspection of 1, we can quickly derive that

$$\frac{\partial \mathcal{V}}{\partial \bar{c}_{n'l'}} = \frac{4}{N} \bar{c}_{nplp}^3 - \frac{2}{N^2} \left(\sum_{n'} \bar{c}_{n'l'} \right) 2\bar{c}_{n'l'}$$

$$= \frac{4}{N} \left(\bar{c}_{n'l'}^3 - s_{l'} \bar{c}_{n'l'} \right),$$

where we defined $s_l := \frac{1}{N} \sum_{n'} \bar{c}_{n'l'}$.
The second factor reads:

$$\frac{\partial \bar{c}_{n'l'}}{\partial r_{ij}} = \frac{\partial \left(c_{n'1} r_{1l'} + \cdots c_{n'L} r_{Ll'} \right)}{\partial r_{ij}}$$

$$= \delta_{l'j} c_{n'i},$$

where
$$\delta_{l'j} = \begin{cases} 1, & \text{if } l' = j, \\ 0, & \text{if } l' \neq j. \end{cases}$$

is the Kronecker delta function.

Plugging both factor expressions back in 1, we find that

$$\frac{\partial \mathcal{V}}{\partial r_{ij}} = \frac{4}{N} \sum_{n'l'} \left(\bar{c}_{n'l'}^3 \delta_{l'j} c_{n'i} - s_{l'} \bar{c}_{n'l'} \delta_{l'j} c_{n'i} \right)$$

$$= \frac{4}{N} \sum_{n'} \left(\bar{c}_{n'j}^3 c_{n'i} - s_j \bar{c}_{n'j} c_{n'j} \right).$$

Thankfully, this can be written compactly in matrix notation:

$$\frac{\partial \mathcal{V}}{\partial \mathbf{R}} = \frac{4}{N} \left(\mathbf{C}^\top \overline{\mathbf{C}}^{\odot 3} - \mathbf{C}^\top \overline{\mathbf{C}} \odot \mathbf{S} \right), \tag{2}$$

where $\overline{\mathbf{C}}^{\odot 3} = \overline{\mathbf{C}} \odot \overline{\mathbf{C}} \odot \overline{\mathbf{C}}$, and $\odot$ denotes the element-wise matrix product (Hadamar product). Moreover,

$$\mathbf{S} = \begin{pmatrix} s_1 & s_2 & \cdots & s_L \\ s_1 & s_2 & \cdots & s_L \\ \vdots & & & \\ s_1 & s_2 & \cdots & s_L \end{pmatrix} \in \mathbb{R}^{N \times L}.$$

2.2 Projected gradient ascent

If we consider the manifold of orthogonal matrices, and would like to take a step from our current estimate to an estimate within that manifold which increases the varimax objective value 1, we need to project the gradient derived in 2 to the tangent space of that manifold. Essentially, we want a small gradient projection step $\varepsilon \mathbf{P}$ added to $\mathbf{R}$ to satisfy:

$$\mathbf{I} = (\mathbf{R} + \varepsilon \mathbf{P})(\mathbf{R} + \varepsilon \mathbf{P})^\top$$

$$= \mathbf{R}\mathbf{R}^\top + \varepsilon \mathbf{R}\mathbf{P}^\top + \varepsilon \mathbf{P}\mathbf{R}^\top + \mathcal{O}(\varepsilon^2)$$

$$= \mathbf{I} + \varepsilon (\mathbf{R}\mathbf{P}^\top + \mathbf{P}\mathbf{R}^\top) + \mathcal{O}(\varepsilon^2).$$

Up to first order, it is clear that we want $\mathbf{R}\mathbf{P}^\top + \mathbf{P}\mathbf{R}^\top = 0$. If we define $\mathbf{\Omega} := \mathbf{R}^\top \mathbf{P}$, we can see that it needs to satisfy $\mathbf{\Omega} = -\mathbf{\Omega}^\top$; that is $\mathbf{\Omega}$ needs to be skew-symmetric.

Now that we know what $\mathbf{P}$ needs to satisfy, we also want $\mathbf{P}$ to be as close as possible to $\frac{\partial \mathcal{V}}{\partial \mathbf{R}}$, the gradient we derived in the previous section. Let us first

rewrite this notion:

$$||\frac{\partial \mathcal{V}}{\partial \mathbf{R}} - \mathbf{P}||_F^2 = ||\frac{\partial \mathcal{V}}{\partial \mathbf{R}} - \mathbf{R}\mathbf{\Omega}||_F^2$$

$$= Tr\left((\frac{\partial \mathcal{V}}{\partial \mathbf{R}} - \mathbf{R}\mathbf{\Omega})^\top (\frac{\partial \mathcal{V}}{\partial \mathbf{R}} - \mathbf{R}\mathbf{\Omega})\right)$$

$$= Tr\left((\frac{\partial \mathcal{V}}{\partial \mathbf{R}} - \mathbf{R}\mathbf{\Omega})^\top \mathbf{R}\mathbf{R}^\top (\frac{\partial \mathcal{V}}{\partial \mathbf{R}} - \mathbf{R}\mathbf{\Omega})\right)$$

$$= Tr\left((\mathbf{R}^\top(\frac{\partial \mathcal{V}}{\partial \mathbf{R}} - \mathbf{R}\mathbf{\Omega}))^\top (\mathbf{R}^\top(\frac{\partial \mathcal{V}}{\partial \mathbf{R}} - \mathbf{R}\mathbf{\Omega}))^\top\right)$$

$$= ||\mathbf{R}^\top \frac{\partial \mathcal{V}}{\partial \mathbf{R}} - \mathbf{\Omega}||_F^2$$

$$= ||\mathbf{M} - \mathbf{\Omega}||_F^2,$$

where we defined $\mathbf{M} := \mathbf{R}^\top \frac{\partial \mathcal{V}}{\partial \mathbf{R}}$, and which we will write out as the sum of a symmetric part and an skew-symmetric part: $\mathbf{M} = \frac{1}{2}(\mathbf{M}+\mathbf{M}^\top)+\frac{1}{2}(\mathbf{M}-\mathbf{M}^\top) = \mathbf{S} + \mathbf{K}$.
Now,

$$\min_{\mathbf{\Omega}} ||\mathbf{K} + \mathbf{S} - \mathbf{\Omega}||_F^2 = \min_{\mathbf{\Omega}} \left(||\mathbf{S}||_F^2 + ||\mathbf{K} - \mathbf{\Omega}||_F^2 - 2 < \mathbf{K} - \mathbf{\Omega}, \mathbf{S} >\right)$$

$$= \min_{\mathbf{\Omega}} ||\mathbf{K} - \mathbf{\Omega}||_F^2,$$

as $||\mathbf{S}||_F^2$ does not depend on $\mathbf{\Omega}$ and the dot product between skew-symmetric and symmetric matrices is zero.
We effectively derived that we need to take the skew-symmetric part of $\mathbf{M}$, here denoted by $\mathbf{K}$, for our original projected gradient $\mathbf{P}$ to be as close as possible to $\frac{\partial \mathcal{V}}{\partial \mathbf{R}}$. We therefore have:

$$\mathbf{P} = \mathbf{R}\mathbf{\Omega} \tag{3}$$

$$= \mathbf{R}\mathbf{K} \tag{4}$$

$$= \mathbf{R}\frac{1}{2}(\mathbf{M} - \mathbf{M}^\top) \tag{5}$$

$$= \mathbf{R}\frac{1}{2}(\mathbf{R}^\top \frac{\partial \mathcal{V}}{\partial \mathbf{R}} - \frac{\partial \mathcal{V}}{\partial \mathbf{R}}^\top \mathbf{R}) \tag{6}$$

$$= \frac{1}{2}(\frac{\partial \mathcal{V}}{\partial \mathbf{R}} - \mathbf{R}\frac{\partial \mathcal{V}}{\partial \mathbf{R}}^\top \mathbf{R}). \tag{7}$$

We hence can algebraically efficiently compute first $\frac{\partial \mathcal{V}}{\partial \mathbf{R}}$ and then construct the projected gradient with $\mathbf{P} = \frac{1}{2}(\frac{\partial \mathcal{V}}{\partial \mathbf{R}} - \mathbf{R}\frac{\partial \mathcal{V}}{\partial \mathbf{R}}^\top \mathbf{R})$.
This however does not guarantee that the updated matrix $\mathbf{R} \leftarrow \mathbf{R} + \eta\mathbf{P}$ (η is some learning rate scalar) is also an orthogonal matrix. It indeed only guarantees it up to first order. If we want to make it absolutely certain that our next $\mathbf{R}$ is orthogonal we can safely take $\mathbf{U}\mathbf{V}^\top$ after applying *svd*-decomposition on $\mathbf{R} + \eta\mathbf{P}$. If we do not want to optimize over reflection matrices (who are orthogonal too), but optimize over rotation matrices only, we can furthermore ensure the determinant of the projection is always $+1$.

2.3 Summary

Given loading matrix $\mathbf{C}$ and rotation matrix $\mathbf{R}$, in order to maximize objective 1, we can take gradient ascent steps as:

Box 1: Projected gradient ascent

$$\frac{\partial \mathcal{V}}{\partial \mathbf{R}} = \frac{4}{N} \left(\mathbf{C}^\top \overline{\mathbf{C}}^{\odot 3} - \mathbf{C}^\top \overline{\mathbf{C}} \odot \mathbf{S} \right)$$

$$\mathbf{P} = \frac{1}{2}(\frac{\partial \mathcal{V}}{\partial \mathbf{R}} - \mathbf{R}\frac{\partial \mathcal{V}}{\partial \mathbf{R}}^\top \mathbf{R})$$

$$\mathbf{R} = \mathbf{U}\mathbf{V}^\top \leftarrow \mathcal{SVD}(\mathbf{R} + \eta\mathbf{P}), \text{ s.t. } \det(\mathbf{R}) = +1$$