Extracting coactivated features from multiple datasets
—Supplementary material—

Michael U. Gutmann and Aapo Hyvärinen

Dept. of Computer Science and HIIT
Dept. of Mathematics and Statistics
University of Helsinki
michael.gutmann@helsinki.fi   aapo.hyvarinen@helsinki.fi

S.1 Preliminaries

We give here a more mathematical development of the material in Subsection 2.1 and derive some preliminary equations that will be used in the following sections.

Let $\mathbf{s}_k = (s_k^1, \ldots, s_k^n)$ denote the vector which contains the k-th source from all the n data sets. With the independence assumptions from Subsection 2.1, we have that the joint density of all the sources $\mathbf{s} = (s_1^1, \ldots, s_d^1, \ldots, s_1^n, \ldots, s_d^n)$ factorizes into d factors,

$$p_{\mathbf{s}}(s_1^1, \ldots, s_d^n) = \prod_{k=1}^{d} p_{\mathbf{s}_k}(\mathbf{s}_k) = \prod_{k=1}^{d} p_{\mathbf{s}_k}(s_k^1, \ldots, s_k^n) \tag{i}$$

With the ICA model for $\mathbf{z}^i$ in Eq.(1) and the orthogonality of the $\mathbf{Q}^i$, we obtain

$$s_k^i = \mathbf{q}_k^{i\,T} \mathbf{z}^i. \tag{ii}$$

Using this relation and the orthogonality of the $\mathbf{Q}^i$, the joint density $p_{\mathbf{z}}$ of the random variables $\mathbf{z}^1, \ldots, \mathbf{z}^n$ becomes

$$p_{\mathbf{z}}(\mathbf{z}^1, \ldots, \mathbf{z}^n) = p_{\mathbf{s}}(\mathbf{q}_1^{1\,T}\mathbf{z}^1, \ldots, \mathbf{q}_d^{n\,T}\mathbf{z}^n) \tag{iii}$$

$$= \prod_{k=1}^{d} p_{\mathbf{s}_k}(\mathbf{q}_k^{1\,T}\mathbf{z}^1, \ldots, \mathbf{q}_k^{n\,T}\mathbf{z}^n) \tag{iv}$$

The joint density $p_{\mathbf{z}}$ is used in the log-likelihood ℓ,

$$\ell(\mathbf{q}_1^1, \ldots, \mathbf{q}_d^n) = \sum_{t=1}^{T} \sum_{k=1}^{d} \log p_{\mathbf{s}_k}(\mathbf{q}_k^{1\,T}\mathbf{z}^1(t), \ldots, \mathbf{q}_k^{n\,T}\mathbf{z}^n(t)). \tag{v}$$

The log-likelihood can be evaluated when the joint density $p_{\mathbf{s}_k}$ is known.

Inverting the linear transform in Eq.(2) gives

$$\tilde{\mathbf{s}}_k = \frac{1}{\sigma_k}\mathbf{s}_k, \tag{vi}$$

where $\tilde{\mathbf{s}}_k = (\tilde{s}_k^1, \dots, \tilde{s}_k^n)$. The determinant of this linear transformation is $1/\sigma_k^n$. Integrating out the variable σ_k with density p_{σ_k} leads to an expression for the density of $\mathbf{s}_k$,

$$p_{\mathbf{s}_k}(\mathbf{s}_k) = \int \frac{p_{\sigma_k}(\sigma_k)}{\sigma_k^n} p_{\tilde{\mathbf{s}}_k}\left(\frac{\mathbf{s}_k}{\sigma_k}\right) \mathrm{d}\sigma_k. \tag{vii}$$

Equivalently, we can specify a prior p_{ω_k} for $\omega_k = \sigma_k^2$. The density $p_{\mathbf{s}_k}$ is then

$$p_{\mathbf{s}_k}(\mathbf{s}_k) = \int \frac{p_{\omega_k}(\omega_k)}{\omega_k^{\frac{n}{2}}} p_{\tilde{\mathbf{s}}_k}\left(\frac{\mathbf{s}_k}{\sqrt{\omega_k}}\right) \mathrm{d}\omega_k. \tag{viii}$$

So far, we have assumed that the $s_k^1, \dots, s_k^n$ are dependent through a common variance variable, but we have not yet completely specified their joint distribution. Specifying the priors for $\tilde{\mathbf{s}}_k$ and σ_k (or ω_k) completes the model in the paper, and allows for the evaluation of the log-likelihood in Eq.(v), Two choices for the priors are discussed in Subsection 2.2 and 2.3.

S.2 Derivation of Eq.(3) in Subsection 2.2

The variables $\tilde{\mathbf{s}}_k$ are assumed jointly Gaussian with density $p_{\tilde{\mathbf{s}}_k}$,

$$p_{\tilde{\mathbf{s}}_k}(\tilde{\mathbf{s}}_k) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{\Sigma}_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\tilde{\mathbf{s}}_k^T \mathbf{\Sigma}_k^{-1} \tilde{\mathbf{s}}_k\right), \tag{ix}$$

where $\mathbf{\Sigma}_k$ is the covariance matrix. The variance variable $\omega_k = \sigma_k^2$ is assumed to follow the inverse Gamma distribution $\mathcal{G}^{-1}(\alpha_k, \beta_k)$ with parameters α_k, β_k,

$$p_{\omega_k}(\omega_k; \alpha_k, \beta_k) = \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} \omega_k^{-\alpha_k - 1} \exp\left(-\frac{\beta_k}{\omega_k}\right), \tag{x}$$

where $\Gamma(\alpha_k)$ is the gamma function,

$$\Gamma(\alpha_k) = \int_0^\infty u^{\alpha_k - 1} \exp(-u) \mathrm{d}u. \tag{xi}$$

The density $p_{\mathbf{s}_k}$ is with Eq. (viii)

$$p_{\mathbf{s}_k}(\mathbf{s}_k) = \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{\Sigma}_k|^{\frac{1}{2}}} \int_0^\infty \omega_k^{-\alpha_k - 1 - \frac{n}{2}} \exp\left(-\left(\beta_k + \frac{1}{2}\mathbf{s}_k^T \mathbf{\Sigma}_k^{-1} \mathbf{s}_k\right) \frac{1}{\omega_k}\right) \mathrm{d}\omega_k. \tag{xii}$$

Making the change of variables

$$\omega_k = \left(\beta_k + \frac{1}{2}\mathbf{s}_k^T \mathbf{\Sigma}_k^{-1} \mathbf{s}_k\right) \frac{1}{u} \tag{xiii}$$

we obtain

$$\begin{aligned} p_{\mathbf{s}_k}(\mathbf{s}_k) &= \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{\Sigma}_k|^{\frac{1}{2}}} \int_0^\infty \left(\beta_k + \frac{1}{2}\mathbf{s}_k^T \mathbf{\Sigma}_k^{-1} \mathbf{s}_k\right)^{-\alpha_k - \frac{n}{2}} u^{\alpha_k + \frac{n}{2} - 1} \exp(-u) \mathrm{d}u \\ &= \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{\Sigma}_k|^{\frac{1}{2}}} \left(\beta_k + \frac{1}{2}\mathbf{s}_k^T \mathbf{\Sigma}_k^{-1} \mathbf{s}_k\right)^{-\alpha_k - \frac{n}{2}} \Gamma\left(\alpha_k + \frac{n}{2}\right) \end{aligned} \tag{xiv}$$

which can be reorganized to give

$$p_{\mathbf{s}_k}(\mathbf{s}_k) = \frac{\Gamma\left(\alpha_k + \frac{n}{2}\right)}{(2\pi\beta_k)^{\frac{n}{2}}\Gamma(\alpha_k)}|\mathbf{\Sigma}_k|^{-\frac{1}{2}}\left(\frac{1}{1+\frac{1}{2\beta_k}\mathbf{s}_k^T\mathbf{\Sigma}_k^{-1}\mathbf{s}_k}\right)^{\alpha_k+\frac{n}{2}}. \tag{xv}$$

This equation can further be simplified by taking into account that the variance of the s_k^i is, by the ICA model in Eq.(1), one. Fixing the variances of the s_k^i eliminates the parameter β_k: The variance $\mathbb{V}(\mathbf{s}_k)$ of $\mathbf{s}_k$ can be computed as

$$\begin{aligned}
\mathbb{V}(\mathbf{s}_k) &= \int_0^\infty p_{\omega_k}(\omega_k)\mathbb{V}(\mathbf{s}_k|\omega_k)\mathrm{d}\omega_k &\text{(xvi)}\\
&= \mathbb{V}(\tilde{\mathbf{s}}_k)\int_0^\infty \omega_k p_{\omega_k}(\omega_k)\mathrm{d}\omega_k &\text{(xvii)}\\
&= \mathbf{\Sigma}_k\int_0^\infty \omega_k\frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)}\omega_k^{-\alpha_k-1}\exp\left(-\frac{\beta_k}{\omega_k}\right)\mathrm{d}\omega_k &\text{(xviii)}\\
&= \mathbf{\Sigma}_k\frac{\beta_k}{\alpha_k-1}. &\text{(xix)}
\end{aligned}$$

For the second equality, we have used that $\mathbb{V}(\mathbf{s}_k|\omega_k) = \omega_k\mathbb{V}(\tilde{s}_k)$, for the third equality, we have used the definitions of $\mathbf{\Sigma}_k$ and p_{ω_k}, and the last equality follows after a change of variables. Hence,

$$\mathbf{\Sigma}_k^{-1} = \frac{\beta_k}{\alpha_k-1}\mathbb{V}(\mathbf{s}_k)^{-1} \tag{xx}$$

and

$$|\mathbf{\Sigma}_k|^{-1} = |\mathbf{\Sigma}_k^{-1}| = \left(\frac{\beta_k}{\alpha_k-1}\right)^n|\mathbb{V}(\mathbf{s}_k)|^{-1}, \tag{xxi}$$

which gives

$$p_{\mathbf{s}_k}(\mathbf{s}_k) = \frac{\Gamma\left(\alpha_k + \frac{n}{2}\right)}{(2\pi(\alpha_k-1))^{\frac{n}{2}}\Gamma(\alpha_k)}|\mathbb{V}(\mathbf{s}_k)|^{-\frac{1}{2}}\left(\frac{1}{1+\frac{1}{2(\alpha_k-1)}\mathbf{s}_k^T\mathbb{V}(\mathbf{s}_k)^{-1}\mathbf{s}_k}\right)^{\alpha_k+\frac{n}{2}}. \tag{xxii}$$

Note that this expression does not depend on β_k any more. In the paper, we consider the case $n=2$. Since the s_k^i have variance one, $\mathbb{V}(\mathbf{s}_k)$ is equal to

$$\mathbb{V}(\mathbf{s}_k) = \begin{pmatrix} 1 & \rho_k \\ \rho_k & 1 \end{pmatrix}, \tag{xxiii}$$

where $\rho_k \in (-1\ 1)$ is the correlation coefficient between s_k^1 and s_k^2. Denoting the inverse of $\mathbb{V}(\mathbf{s}_k)$ by $\mathbf{\Lambda}_k$, we have

$$\mathbf{\Lambda}_k = \mathbb{V}(\mathbf{s}_k)^{-1} = \frac{1}{1-\rho_k^2}\begin{pmatrix} 1 & -\rho_k \\ -\rho_k & 1 \end{pmatrix}, \tag{xxiv}$$

which is Eq.(4) in the paper. For $n=2$, the expression for $p_{\mathbf{s}_k}$ is

$$p_{\mathbf{s}_k}(\mathbf{s}_k) = \frac{\Gamma\left(\alpha_k+1\right)}{(2\pi(\alpha_k-1))\Gamma(\alpha_k)}|\mathbf{\Lambda}_k|^{\frac{1}{2}}\left(\frac{1}{1+\frac{1}{2(\alpha_k-1)}\mathbf{s}_k^T\mathbf{\Lambda}_k\mathbf{s}_k}\right)^{\alpha_k+1}, \tag{xxv}$$

where we have used $\mathbf{\Lambda}_k = \mathbb{V}(\mathbf{s}_k)^{-1}$. The more standard parameter for student's t distributions is $\nu_k = 2\alpha_k$, which gives Eq.(3) in the paper:

$$p_{\mathbf{s}_k}(\mathbf{s}_k) = \frac{\Gamma\left(\frac{\nu_k+2}{2}\right)}{(\pi(\nu_k-2))\Gamma\left(\frac{\nu_k}{2}\right)}|\mathbf{\Lambda}_k|^{\frac{1}{2}}\left(\frac{1}{1+\frac{1}{(\nu_k-2)}\mathbf{s}_k^T\mathbf{\Lambda}_k\mathbf{s}_k}\right)^{\frac{\nu_k+2}{2}}, \tag{xxvi}$$

S.3 Detailed calculations for Subsection 2.2, Eq.(6) to Eq.(7)

Eq. (6) in the paper is

$$\ell(\mathbf{q}_1^1,\ldots,\mathbf{q}_d^2) = \text{const} - \sum_{t=1}^{T}\sum_{k=1}^{d}\frac{\nu_k+2}{2}\log\left(1+\frac{1}{\nu_k-2}\mathbf{y}_k(t)^T\mathbf{\Lambda}_k\mathbf{y}_k(t)\right) \tag{xxvii}$$

and Eq. (7) is

$$\ell(\mathbf{q}_1^1,\mathbf{q}_1^2,\ldots,\mathbf{q}_d^1,\mathbf{q}_d^2) \approx \text{const} + T\sum_{k=1}^{d}\frac{1}{1-\rho_k^2}\left(\rho_k\mathbf{q}_k^{1\,T}\widehat{\mathbf{\Sigma}}_{12}\mathbf{q}_k^2\right). \tag{xxviii}$$

We show here the detailed steps for going from Eq. (6) to Eq.(7), using that ν_k is large.

First, we compute $1/(\nu_k-2)\mathbf{y}_k^T\mathbf{\Lambda}_k\mathbf{y}_k$, where we drop for a moment the counter t for the samples. Using the definition of $\mathbf{y}_k$,

$$\mathbf{y}_k = (\mathbf{q}_k^{1\,T}\mathbf{z}^1,\ \mathbf{q}_k^{2\,T}\mathbf{z}^2)^T \tag{xxix}$$

and the definition of $\mathbf{\Lambda}_k$,

$$\mathbf{\Lambda}_k = \frac{1}{1-\rho_k^2}\begin{pmatrix}1 & -\rho_k\\ -\rho_k & 1\end{pmatrix}, \tag{xxx}$$

we obtain

$$\frac{\mathbf{y}_k^T\mathbf{\Lambda}_k\mathbf{y}_k}{\nu_k-2} = \frac{1}{\nu_k-2}\frac{1}{1-\rho_k^2}\left[\left(\mathbf{q}_k^{1\,T}\mathbf{z}^1\right)^2+\left(\mathbf{q}_k^{2\,T}\mathbf{z}^2\right)^2-2\rho_k\mathbf{q}_k^{1\,T}\mathbf{z}^1\mathbf{z}_2^T\mathbf{q}_k^2\right], \tag{xxxi}$$

For large ν_k the term $1/(\nu_k-2)\mathbf{y}_k^T\mathbf{\Lambda}_k\mathbf{y}_k$ is small. Hence,

$$\log\left(1+\frac{1}{\nu_k-2}\mathbf{y}_k^T\mathbf{\Lambda}_k\mathbf{y}_k\right) = \frac{1}{\nu_k-2}\mathbf{y}_k^T\mathbf{\Lambda}_k\mathbf{y}_k + O\left(\frac{1}{\nu_k^2}\right), \tag{xxxii}$$

where we have used the first-order Taylor expansion of $\log(1+x)$ around $x=0$. Dropping terms of order $1/\nu_k^2$ and smaller, we have for Eq. (6)

$$\begin{aligned}\ell(\mathbf{q}_1^1,\mathbf{q}_1^2,\ldots,\mathbf{q}_d^1,\mathbf{q}_d^2) \approx \text{const} - \sum_{t=1}^{T}\sum_{k=1}^{d}\frac{\nu_k+2}{2\nu_k-4}\frac{1}{1-\rho_k^2}\Big[\left(\mathbf{q}_k^{1\,T}\mathbf{z}^1(t)\right)^2 +\\ \left(\mathbf{q}_k^{2\,T}\mathbf{z}^2(t)\right)^2 - 2\rho_k\mathbf{q}_k^{1\,T}\mathbf{z}^1(t)\mathbf{z}_2(t)^T\mathbf{q}_k^2\Big].\end{aligned} \tag{xxxiii}$$

Since ν_k is assumed large,

$$\frac{\nu_k + 2}{2\nu_k - 4} \approx \frac{1}{2} \tag{xxxiv}$$

and thus

$$\ell(\mathbf{q}_1^1, \mathbf{q}_1^2, \ldots, \mathbf{q}_d^1, \mathbf{q}_d^2) \approx \text{const} - \sum_{t=1}^{T} \sum_{k=1}^{d} \frac{1}{1-\rho_k^2} \frac{1}{2} \Big[\left(\mathbf{q}_k^{1\,T} \mathbf{z}^1(t) \right)^2 + \left(\mathbf{q}_k^{2\,T} \mathbf{z}^2(t) \right)^2 - 2\rho_k \mathbf{q}_k^{1\,T} \mathbf{z}^1(t) \mathbf{z}_2(t)^T \mathbf{q}_k^2 \Big]. \tag{xxxv}$$

The sum over the samples is

$$\sum_{t=1}^{T} \left[\left(\mathbf{q}_k^{1\,T} \mathbf{z}^1(t) \right)^2 + \left(\mathbf{q}_k^{2\,T} \mathbf{z}^2(t) \right)^2 - 2\rho_k \mathbf{q}_k^{1\,T} \mathbf{z}^1(t) \mathbf{z}_2(t)^T \mathbf{q}_k^2 \right],$$

which equals

$$T \left[\mathbf{q}_k^{1\,T} \widehat{\boldsymbol{\Sigma}}_{11} \mathbf{q}_k^1 + \mathbf{q}_k^{2\,T} \widehat{\boldsymbol{\Sigma}}_{22} \mathbf{q}_k^2 - 2\rho_k \mathbf{q}_k^{1\,T} \widehat{\boldsymbol{\Sigma}}_{12} \mathbf{q}_k^2 \right],$$

where

$$\widehat{\boldsymbol{\Sigma}}_{ii} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{z}^i(t) \mathbf{z}^i(t)^T, \qquad \widehat{\boldsymbol{\Sigma}}_{12} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{z}^1(t) \mathbf{z}^2(t)^T \tag{xxxvi}$$

are the sample covariance and cross-correlation matrix. Now, either assuming that T is large, or that the data has been preprocessed such that it is white, we have that $\widehat{\boldsymbol{\Sigma}}_{ii}$ is the identity matrix. Since $\mathbf{q}_k^i$ are the columns of an orthonormal matrix, we obtain

$$\mathbf{q}_k^{i\,T} \widehat{\boldsymbol{\Sigma}}_{ii} \mathbf{q}_k^i = 1 \quad (i = 1, 2) \quad \forall k \tag{xxxvii}$$

Plugging these relations into Eq. (xxxv), we have

$$\ell(\mathbf{q}_1^1, \mathbf{q}_1^2, \ldots, \mathbf{q}_d^1, \mathbf{q}_d^2) \approx \text{const} - T \sum_{k=1}^{d} \frac{1}{1-\rho_k^2} \frac{1}{2} \left[2 - 2\rho_k \mathbf{q}_k^{1\,T} \widehat{\boldsymbol{\Sigma}}_{12} \mathbf{q}_k^2 \right], \tag{xxxviii}$$

from where Eq. (7) follows.

S.4 Detailed calculations for Eq.(8) in Subsection 2.3

Eq. (8) in Subsection 2.3 is

$$\ell(\mathbf{q}_1^1, \ldots, \mathbf{q}_d^n) = \sum_{t=1}^{T} \sum_{k=1}^{d} G_k \left(\sum_{i=1}^{n} (\mathbf{q}_k^{i\,T} \mathbf{z}^i(t))^2 \right),$$

where $\mathbf{z}^i(t)$ is the t-th data point in data set $i = 1, \ldots, n$, and G_k is a nonlinearity which depends on the distribution of the variance variable σ_k.

We show here how this equation follows from Eq.(v) and (vii) when the $\tilde{s}_k^i$ follow a standard normal distribution. We have

$$p_{\tilde{\mathbf{s}}_k}(\tilde{\mathbf{s}}_k) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\tilde{s}_k^i)^2\right), \tag{xxxix}$$

from where we obtain

$$p_{\tilde{\mathbf{s}}_k}\left(\frac{\mathbf{s}_k}{\sigma_k}\right) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma_k^2}\sum_{i=1}^{n}(s_k^i)^2\right). \tag{xl}$$

This density depends only via the sum $\sum_{i=1}^{n}(s_k^i)^2$ on $\mathbf{s}_k$ so that with Eq. (vii)

$$\log p_{\mathbf{s}_k}(\mathbf{s}_k) = \log p_{\mathbf{s}_k}(s_k^1, \ldots, s_k^n) = G_k\left(\sum_{i=1}^{n}(s_k^i)^2\right), \tag{xli}$$

where the function G_k is

$$G_k(u) = \log \int \frac{p_{\sigma_k}(\sigma_k)}{(2\pi\sigma_k^2)^{\frac{n}{2}}} \exp\left(-\frac{u^2}{2\sigma_k^2}\right) \mathrm{d}\sigma_k. \tag{xlii}$$

This function is defined via an integral and depends on the prior distribution p_{σ_k}. The integral will not be analytically computable for many choices of p_{σ_k}. It is, however, a one-dimensional integral which can efficiently be evaluated with numerical methods.

Using Eq.(xli) in the log-likelihood given in Eq.(v), we obtain

$$\ell(\mathbf{q}_1^1, \ldots, \mathbf{q}_d^n) = \sum_{t=1}^{T}\sum_{k=1}^{d} G_k\left(\sum_{i=1}^{n}(\mathbf{q}_k^{i\,T}\mathbf{z}^i(t))^2\right), \tag{xliii}$$

which is Eq.(8) in the paper.