The Independent Factor Analysis (IFA) model assumes a standard generative factor model for the observations $\boldsymbol{y}$ with $K$ factors,

$$\boldsymbol{y}_n \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{s}_n, \boldsymbol{\Lambda}),$$

but uses an expressive factorial mixture of Gaussians for the sources with $C$ components per factor $k$,

$$s_{nk} \sim \sum_{c=1}^{C} w_{kc} \mathcal{N}(\mu_{kc}, \sigma_{kc}^2),$$

By multiplying the univariate mixtures together, we can equivalently state that the multivariate source vector $\boldsymbol{s}$ is drawn from a mixture of Gaussians with diagonal covariance and $Q = C^K$ components,

$$\boldsymbol{s}_n \sim \sum_{q=1}^{Q} \bar{w}_q \mathcal{N}(\bar{\boldsymbol{\mu}}_q, \bar{\boldsymbol{\Sigma}}_q). \tag{1}$$

As with any mixture, we can express the mixture density using auxiliary indicators, where $z_{nq} = 1$ if observation $n$ was drawn from component $q$, and 0 otherwise.

$$\boldsymbol{s}_n | \boldsymbol{z}_n \sim \prod_{q=1}^{Q} \mathcal{N}(\bar{\boldsymbol{\mu}}_q, \bar{\boldsymbol{\Sigma}}_q)^{z_{nq}}, \quad \boldsymbol{z}_n \sim \mathrm{Mult}(\bar{\boldsymbol{w}}). \tag{2}$$

with $N$ observations, the full log-joint has terms

$$\begin{aligned}
\ln p(\boldsymbol{y}|\boldsymbol{s}) &= \sum_{n=1}^{N} \left( -\frac{1}{2}(\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n)^\top \boldsymbol{\Lambda}^{-1}(\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n) - \frac{1}{2}\ln|\boldsymbol{\Lambda}| - \frac{D}{2}\ln 2\pi) \right) \\
\ln p(\boldsymbol{s}|\boldsymbol{z}) &= \sum_{n=1}^{N}\sum_{q=1}^{Q} z_{nq} \left( -\frac{1}{2}(\boldsymbol{s}_n - \bar{\boldsymbol{\mu}}_q)^\top \bar{\boldsymbol{\Sigma}}_q^{-1}(\boldsymbol{s}_n - \bar{\boldsymbol{\mu}}_q) - \frac{1}{2}\ln|\bar{\boldsymbol{\Sigma}}_q| - \frac{D}{2}\ln 2\pi) \right) \\
\ln p(\boldsymbol{z}) &= \sum_{n=1}^{N}\sum_{q=1}^{Q} z_{nq} \ln \bar{w}_q
\end{aligned}$$

To compute gradients in $\ln p(\boldsymbol{y})$, we need to be able to integrate $\boldsymbol{s}$ and $\boldsymbol{z}$ over $p(\boldsymbol{s}, \boldsymbol{z}|\boldsymbol{y}) = p(\boldsymbol{s}|\boldsymbol{z}, \boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{y})$. First we have a standard Gaussian posterior,

$$p(\boldsymbol{s}_n | z_{nq} = 1, \boldsymbol{y}_n) = \mathcal{N}(\boldsymbol{V}_q \boldsymbol{A}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{y}_n, \boldsymbol{V}_q), \quad \boldsymbol{V}_q = (\boldsymbol{A}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{A} + \bar{\boldsymbol{\Sigma}}_q^{-1})^{-1} \tag{3}$$

and then using a standard responsibility argument, we can find the posterior of the assignment indicators $\boldsymbol{z}$ as

$$p(\boldsymbol{z}_{nq} = 1 | \boldsymbol{y}_n) \propto \bar{w}_q p(\boldsymbol{y}_n | z_{nq} = 1) = \bar{w}_q \mathcal{N}(\boldsymbol{y}_n | \boldsymbol{A}\bar{\boldsymbol{\mu}}_q, \boldsymbol{\Lambda} + \boldsymbol{A}\bar{\boldsymbol{\Sigma}}_q \boldsymbol{A}^\top). \tag{4}$$

## 0.1 Mixture weights

Since $\sum_{c=1}^{C} w_{kc} = 1$, we have to add Lagrangians, and due to the positivity constraint we reparameterize as $v_{kc} = \ln w_{kc}$. The gradient is then

$$\nabla_{v_{kc}} \left( \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z}) + \sum_{\ell=1}^{K} \rho_\ell (1 - \sum_{d=1}^{C} e^{v_{\ell d}}) \right) = \sum_{n=1}^{N} \sum_{q=1}^{Q} z_{nq} \xi_{kc}^{q} - \rho_k e^{v_{kc}} \tag{5}$$

Taking the expectation and solving we find,

$$w_{kc} \propto \sum_{n=1}^{N} \sum_{q=1}^{Q} \xi_{kc}^{q} \, \mathbb{E}[z_{nq}] \tag{6}$$

If we add a Dirichlet prior $\boldsymbol{w}_k \sim \mathrm{Dir}(\boldsymbol{\alpha}_k)$, then it contributes with the terms

$$\nabla_{v_{kc}} \ln p(\boldsymbol{w}_k) = \alpha_{kc} - 1 \tag{7}$$

which changes the analytical solution to

$$w_{kc} \propto (\alpha_{kc} - 1) + \sum_{n=1}^{N} \sum_{q=1}^{Q} \xi_{kc}^{q} \, \mathbb{E}[z_{nq}] \tag{8}$$

## 0.2 Mixture Variance

Taking the gradient in $\bar{\boldsymbol{\Sigma}}_q$, we get

$$\nabla_{\bar{\boldsymbol{\Sigma}}_q} \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z}) = -\frac{1}{2} \sum_{n=1}^{N} z_{nq} \left( \bar{\boldsymbol{\Sigma}}_q^{-1} (\boldsymbol{s}_n - \bar{\boldsymbol{\mu}}_q)(\boldsymbol{s}_n - \bar{\boldsymbol{\mu}}_q)^\top \bar{\boldsymbol{\Sigma}}_q^{-1} - \bar{\boldsymbol{\Sigma}}_q^{-1} \right) \tag{9}$$

Let the diagonal elements be denoted by $\bar{\sigma}_{qk}^2 = [\bar{\boldsymbol{\Sigma}}_q]_{kk}$. We take $\xi_{kc}^q$ to be an indicator of whether $\sigma_{kc}^2$ is a factor of $\bar{\sigma}_{qk}^2$. Then we can use the chain rule to get

$$\frac{\partial}{\partial \sigma_{kc}^2} \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z}) = -\frac{1}{2} \sum_{q=1}^{Q} \xi_{kc}^{q} \left( \sum_{n=1}^{N} z_{nq} \left( \frac{s_{nk} - \bar{\mu}_{qk}}{\bar{\sigma}_{qk}^2} \right)^2 - \frac{\sum_{n=1}^{N} z_{nq}}{\bar{\sigma}_{qk}^2} \right) \tag{10}$$

We can get the gradient in $\ln p(\boldsymbol{y})$ by taking the expectation of the above quantity with respect to the posteriors we calculated before, and we can then solve it analytically,

$$\sigma_{kc}^2 = \frac{\sum_{q=1}^{Q} \sum_{n=1}^{N} \xi_{kc}^{q} \, \mathbb{E}[z_{nq}] \, \mathbb{E}\left[(s_{nk} - \bar{\mu}_{qk})^2\right]}{\sum_{q=1}^{Q} \sum_{n=1}^{N} \xi_{kc}^{q} \, \mathbb{E}[z_{nq}]} \tag{11}$$

If $\sigma_{kc}^2$ is endowed with an inverse Gamma prior with shape $a_{kc}$ and scale $b_{kc}$,

$$\ln p(\sigma_{kc}^2) = -(1 + a_{kc}) \ln \lambda_{kc}^2 - \frac{b_{kc}}{\lambda_{kc}^2} \tag{12}$$

then it contributes the terms,

$$\nabla_{\sigma^2} \ln p(\sigma_{kc}^2) = -(1 + a_{kc})\frac{1}{\sigma_{kc}^2} + \frac{b_{kc}}{\sigma_{kc}^2} \tag{13}$$

and adding those terms to the gradient we can solve again and find

$$\sigma_{kc}^2 = \frac{2(1 + a_{kc}) + \sum_{q=1}^{Q}\sum_{n=1}^{N} \xi_{kc}^q \,\mathbb{E}[z_{nq}]\,\mathbb{E}\big[(s_{nk} - \bar{\mu}_{qk})^2\big]}{2b_{kc} + \sum_{q=1}^{Q}\sum_{n=1}^{N} \xi_{kc}^q \,\mathbb{E}[z_{nq}]} \tag{14}$$

## 0.3 Noise Variance

$$\nabla_{\mathbf{\Lambda}} \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z}) = \frac{1}{2}\sum_{n=1}^{N}\big(\mathbf{\Lambda}^{-1}(\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n)(\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n)^\top \mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1}\big) \tag{15}$$

Taking the diagonal elements $\lambda_k^2$

$$\frac{\partial}{\partial \lambda_k^2} \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z}) = \frac{1}{2}\left(\sum_{n=1}^{N}\left(\frac{y_{nk} - \boldsymbol{e}_k^\top \boldsymbol{A}\boldsymbol{s}_n}{\lambda_k^2}\right)^2 - N\frac{1}{\lambda_k^2}\right) \tag{16}$$

Taking the expectation and solving, we get

$$\lambda_k^2 = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\big[(y_{nk} - \boldsymbol{e}_k^\top \boldsymbol{A}\boldsymbol{s}_n)^2\big] = \frac{1}{N}\sum_{n=1}^{N}(y_{nk}^2 + \boldsymbol{e}_k^\top \boldsymbol{A}\,\mathbb{E}\big[\boldsymbol{s}_n\boldsymbol{s}_n^\top\big]\boldsymbol{A}^\top \boldsymbol{e}_k - 2y_{nk}\boldsymbol{e}_k^\top \boldsymbol{A}\,\mathbb{E}[\boldsymbol{s}_n]) \tag{17}$$

If $\lambda_k^2$ is again endowed with an inverse Gamma prior then it contributes the terms,

$$\nabla_{\lambda^2} \ln p(\lambda_k^2) = -(1 + \alpha_k)\frac{1}{\lambda_k^2} + \frac{\beta_k}{(\lambda_k^2)^2} \tag{18}$$

and adding those terms to the gradient we can solve again and find

$$\lambda_k^2 = \frac{1}{N + 2 + 2\alpha_k}\sum_{n=1}^{N}\big(2\beta_k + \mathbb{E}\big[(y_{nk} - \boldsymbol{e}_k^\top \boldsymbol{A}\boldsymbol{s}_n)^2\big]\big)$$

If we have a single $\lambda_0^2$ controlling the noise level (scaled unit diagonal covariance), then the gradient simplifies

$$\frac{\partial}{\partial \lambda_0^2} \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z}) = \frac{1}{(\lambda_0^2)^2}\left(\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n)^\top(\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n)\right) - \frac{ND}{2}\frac{1}{\lambda_0^2} \tag{19}$$

and taking the expectation and isolating yields,

$$\lambda_0^2 = \frac{\sum_{n=1}^{N}(\boldsymbol{y}_n^\top \boldsymbol{y}_n + \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{A}\,\mathbb{E}\big[\boldsymbol{s}_n\boldsymbol{s}_n^\top\big]) - 2\boldsymbol{y}_n^\top \boldsymbol{A}\,\mathbb{E}[\boldsymbol{s}_n])}{ND} \tag{20}$$

or if $\mathbf{\Lambda}$ is a scaled unit matrix and

$$\lambda_0^2 = \frac{2\beta_0 + \sum_{n=1}^{N}(\boldsymbol{y}_n^\top \boldsymbol{y}_n + \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{A}\,\mathbb{E}\big[\boldsymbol{s}_n\boldsymbol{s}_n^\top\big]) - 2\boldsymbol{y}_n^\top \boldsymbol{A}\,\mathbb{E}[\boldsymbol{s}_n])}{ND + 2 + 2\alpha_0} \tag{21}$$

## 0.4 Factor Loadings

$$\nabla_{\boldsymbol{A}} \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z}) = \boldsymbol{\Lambda}^{-1} \sum_{n=1}^{N} (\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n)\boldsymbol{s}_n^\top \tag{22}$$

we can then again take the expectation,

$$\mathbb{E}[\nabla_{\boldsymbol{A}} \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z})] = \boldsymbol{\Lambda}^{-1} \left( \sum_{n=1}^{N} \boldsymbol{y}_n \, \mathbb{E}[\boldsymbol{s}_n]^\top - \boldsymbol{A} \left( \sum_{n=1}^{N} \mathbb{E}\big[\boldsymbol{s}_n \boldsymbol{s}_n^\top\big] \right) \right) \tag{23}$$

and set the gradient to zero to find the optimal update

$$\boldsymbol{A} = \left( \sum_{n=1}^{N} \boldsymbol{y}_n \, \mathbb{E}[\boldsymbol{s}_n]^\top \right) \left( \sum_{n=1}^{N} \mathbb{E}\big[\boldsymbol{s}_n \boldsymbol{s}_n^\top\big] \right)^{-1}$$

If we add a Gaussian prior $[\boldsymbol{A}]_{ij} \sim \mathcal{N}(0, \sigma_{\boldsymbol{A}}^2)$ it contributes the term,

$$\mathbb{E}[\nabla_{\boldsymbol{A}} \ln p(\boldsymbol{A})] = -\frac{1}{\sigma_{\boldsymbol{A}}^2} \boldsymbol{A}$$

and we can solve again to find the MAP update,

$$\boldsymbol{A} = \left( \sum_{n=1}^{N} \boldsymbol{y}_n \, \mathbb{E}[\boldsymbol{s}_n]^\top \right) \left( \sum_{n=1}^{N} \mathbb{E}\big[\boldsymbol{s}_n \boldsymbol{s}_n^\top\big] + \frac{\lambda^2}{\sigma_{\boldsymbol{A}}^2} \right)^{-1}$$