The Projected mixture of Gaussians (proj-MoG) 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 a mixture of Gaussians with Q clusters for the sources,

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

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.

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

with N observations, the full log-joint has terms

$$\ln p(\boldsymbol{y}|\boldsymbol{s}) = \sum_{n=1}^{N} \left(-\frac{1}{2}(\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n)^\top \boldsymbol{\Lambda}^{-1} (\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n) - \frac{1}{2}\ln|\boldsymbol{\Lambda}| - \frac{D}{2}\ln 2\pi) \right)$$

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

$$\ln p(\boldsymbol{z}) = \sum_{n=1}^{N}\sum_{q=1}^{Q} z_{nq} \ln \bar{w}_q$$

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,

$$p(\boldsymbol{s}_n | z_{nq} = 1, \boldsymbol{y}_n) = \mathcal{N}(\boldsymbol{V}_q \left(\boldsymbol{A}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{y}_n + \bar{\boldsymbol{\Sigma}}_q^{-1} \bar{\boldsymbol{\mu}}_q \right), \boldsymbol{V}_q), \quad \boldsymbol{V}_q = (\boldsymbol{A}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{A} + \bar{\boldsymbol{\Sigma}}_q^{-1})^{-1} \tag{2}$$

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

$$p(\boldsymbol{z}_{nq} = 1 | \boldsymbol{y}_n) \propto \bar{w}_q p(\boldsymbol{y}_n | z_{nq} = 1) = \bar{w}_q \mathcal{N}(\boldsymbol{y}_n | \boldsymbol{A}\bar{\boldsymbol{\mu}}_q, \boldsymbol{\Lambda} + \boldsymbol{A}\bar{\boldsymbol{\Sigma}}_q \boldsymbol{A}^\top). \tag{3}$$

0.1 Mixture weights

Since $\sum_{q=1}^{Q} \bar{w}_q = 1$, we have to add Lagrangians, and due to the positivity constraint we reparameterize as $v_q = \ln \bar{w}_q$. The gradient is then

$$\nabla_{v_q} \left(\ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z}) + \sum_{\ell=1}^{K} \rho_\ell (1 - \sum_{d=1}^{C} e^{v_{\ell d}}) \right) = \sum_{n=1}^{N}\sum_{q=1}^{Q} z_{nq} - \rho_k e^{v_q} \tag{4}$$

Taking the expectation and solving we find,

$$\bar{w}_q \propto \sum_{n=1}^{N} \mathbb{E}[z_{nq}] \tag{5}$$

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

$$\nabla_{v_q} \ln p(\boldsymbol{w}) = \alpha_q - 1 \tag{6}$$

which changes the analytical solution to

$$\bar{w}_q \propto (\alpha_q - 1) + \sum_{n=1}^{N} \mathbb{E}[z_{nq}] \tag{7}$$

0.2 Mixture Mean

$$\nabla_{\bar{\boldsymbol{\mu}}_q} \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z}) = -\frac{1}{2} \sum_{n=1}^{N} z_{nq} \left(\bar{\boldsymbol{\Sigma}}_q^{-1} (\boldsymbol{s}_n - \bar{\boldsymbol{\mu}}_q) \right) \tag{8}$$

taking the expectation, setting to zero, and isolating,

$$\bar{\boldsymbol{\mu}}_q = \frac{1}{\sum_{n=1}^{N} \mathbb{E}[z_{nq}]} \sum_{n=1}^{N} \mathbb{E}[z_{nq}] \, \mathbb{E}[\boldsymbol{s}_n | z_{nq} = 1] \tag{9}$$

or we can introduce a prior in the form of a normal,

$$\nabla_{\bar{\boldsymbol{\mu}}_q} \ln \mathcal{N}(\bar{\boldsymbol{\mu}}_q | \bar{\boldsymbol{\mu}}_0, \tau \boldsymbol{I}) = -\frac{1}{2\tau} \nabla_{\bar{\boldsymbol{\mu}}_q} \left(\bar{\boldsymbol{\mu}}_q - \bar{\boldsymbol{\mu}}_0 \right)^\top \left(\bar{\boldsymbol{\mu}}_q - \bar{\boldsymbol{\mu}}_0 \right) = -\frac{1}{\tau} (\bar{\boldsymbol{\mu}}_q - \bar{\boldsymbol{\mu}}_0) \tag{10}$$

which we can add in before solving,

$$\bar{\boldsymbol{\mu}}_q = \left(\frac{1}{\tau} \bar{\boldsymbol{\Sigma}}_q + \left(\sum_{n=1}^{N} \mathbb{E}[z_{nq}] \right) \boldsymbol{I} \right)^{-1} \left(\frac{1}{\tau} \bar{\boldsymbol{\Sigma}}_q \bar{\boldsymbol{\mu}}_0 + \sum_{n=1}^{N} \mathbb{E}[z_{nq}] \, \mathbb{E}[\boldsymbol{s}_n | z_{nq} = 1] \right) \tag{11}$$

If we use the conjugate prior $\mathcal{N}(\bar{\boldsymbol{\mu}}_q | \bar{\boldsymbol{\mu}}_0, \tau \bar{\boldsymbol{\Sigma}}_q)$ instead, this simplifies to an expression that does not involve matrix inverses,

$$\bar{\boldsymbol{\mu}}_q = \frac{1}{\frac{1}{\tau} + \sum_{n=1}^{N} \mathbb{E}[z_{nq}]} \left(\frac{1}{\tau} \bar{\boldsymbol{\mu}}_0 + \sum_{n=1}^{N} \mathbb{E}[z_{nq}] \, \mathbb{E}[\boldsymbol{s}_n | z_{nq} = 1] \right) \tag{12}$$

0.3 Mixture Variance

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

$$\nabla_{\bar{\boldsymbol{\Sigma}}_q} \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z}) = \frac{1}{2} \sum_{n=1}^{N} z_{nq} \left(\bar{\boldsymbol{\Sigma}}_q^{-1} (\boldsymbol{s}_n - \bar{\boldsymbol{\mu}}_q)(\boldsymbol{s}_n - \bar{\boldsymbol{\mu}}_q)^\top \bar{\boldsymbol{\Sigma}}_q^{-1} - \bar{\boldsymbol{\Sigma}}_q^{-1} \right) \tag{13}$$

Taking the expectation and setting to 0, we can then isolate the covariance matrix,

$$\bar{\mathbf{\Sigma}}_q = \frac{1}{\sum_{n=1}^{N} \mathbb{E}[z_{nq}]} \left(\sum_{n=1}^{N} \mathbb{E}[z_{nq}] \, \mathbb{E}\big[(\boldsymbol{s}_n - \bar{\boldsymbol{\mu}}_q)(\boldsymbol{s}_n - \bar{\boldsymbol{\mu}}_q)^\top | z_{nq} = 1\big] \right) \tag{14}$$

If we impose a conjugate inverse Wishart prior with log-density,

$$\ln \mathcal{W}^{-1}(\bar{\mathbf{\Sigma}}|\mathbf{\Psi}, \nu) = -\frac{\nu + p + 1}{2} \ln |\bar{\mathbf{\Sigma}}| - \frac{1}{2} \operatorname{tr}(\mathbf{\Psi} \bar{\mathbf{\Sigma}}^{-1}) + \text{const.} \tag{15}$$

we can add the derivative given by,

$$\nabla_{\bar{\mathbf{\Sigma}}} \ln p(\bar{\mathbf{\Sigma}}) = -\frac{\nu + p + 1}{2} \bar{\mathbf{\Sigma}}^{-1} + \frac{1}{2} \bar{\mathbf{\Sigma}}^{-1} \mathbf{\Psi} \bar{\mathbf{\Sigma}}^{-1} \tag{16}$$

to compute the following MAP update as well,

$$\bar{\mathbf{\Sigma}}_q = \frac{1}{\sum_{n=1}^{N} \mathbb{E}[z_{nq}] + \nu + p + 1} \left(\Psi + \sum_{n=1}^{N} \mathbb{E}[z_{nq}] \, \mathbb{E}\big[(\boldsymbol{s}_n - \bar{\boldsymbol{\mu}}_q)(\boldsymbol{s}_n - \bar{\boldsymbol{\mu}}_q)^\top | z_{nq} = 1\big] \right). \tag{17}$$

0.4 Noise Variance

$$\nabla_{\mathbf{\Lambda}} \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z}) = \frac{1}{2} \sum_{n=1}^{N} \left(\mathbf{\Lambda}^{-1} (\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n)(\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n)^\top \mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1} \right) \tag{18}$$

Taking the diagonal elements λ_k^2

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

Taking the expectation and solving, we get

$$\lambda_k^2 = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\big[(y_{nk} - \boldsymbol{e}_k^\top \boldsymbol{A}\boldsymbol{s}_n)^2\big] = \frac{1}{N} \sum_{n=1}^{N} (y_{nk}^2 + \boldsymbol{e}_k^\top \boldsymbol{A} \, \mathbb{E}\big[\boldsymbol{s}_n \boldsymbol{s}_n^\top\big] \boldsymbol{A}^\top \boldsymbol{e}_k - 2 y_{nk} \boldsymbol{e}_k^\top \boldsymbol{A} \, \mathbb{E}[\boldsymbol{s}_n]) \tag{20}$$

If λ_k^2 is endowed with an inverse Gamma prior then it contributes the terms,

$$\nabla_{\lambda^2} \ln p(\lambda_k^2) = -(1 + \alpha_k) \frac{1}{\lambda_k^2} + \frac{\beta_k}{(\lambda_k^2)^2} \tag{21}$$

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

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

If we have a single λ_0^2 controlling the noise level (scaled unit diagonal covariance), then the gradient simplifies

$$\frac{\partial}{\partial \lambda_0^2} \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z}) = \frac{1}{(\lambda_0^2)^2} \left(\frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n)^\top (\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n) \right) - \frac{ND}{2} \frac{1}{\lambda_0^2} \tag{22}$$

and taking the expectation and isolating yields,

$$\lambda_0^2 = \frac{\sum_{n=1}^{N} (\boldsymbol{y}_n^\top \boldsymbol{y}_n + \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{A}\, \mathbb{E}\big[\boldsymbol{s}_n \boldsymbol{s}_n^\top\big]) - 2\boldsymbol{y}_n^\top \boldsymbol{A}\, \mathbb{E}[\boldsymbol{s}_n])}{ND} \tag{23}$$

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

$$\lambda_0^2 = \frac{2\beta_0 + \sum_{n=1}^{N} (\boldsymbol{y}_n^\top \boldsymbol{y}_n + \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{A}\, \mathbb{E}\big[\boldsymbol{s}_n \boldsymbol{s}_n^\top\big]) - 2\boldsymbol{y}_n^\top \boldsymbol{A}\, \mathbb{E}[\boldsymbol{s}_n])}{ND + 2 + 2\alpha_0} \tag{24}$$

0.5 Factor Loadings

$$\nabla_{\boldsymbol{A}} \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z}) = \boldsymbol{\Lambda}^{-1} \sum_{n=1}^{N} (\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n)\boldsymbol{s}_n^\top \tag{25}$$

we can then again take the expectation,

$$\mathbb{E}[\nabla_{\boldsymbol{A}} \ln p(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{z})] = \boldsymbol{\Lambda}^{-1} \left(\sum_{n=1}^{N} \boldsymbol{y}_n\, \mathbb{E}[\boldsymbol{s}_n]^\top - \boldsymbol{A} \left(\sum_{n=1}^{N} \mathbb{E}\big[\boldsymbol{s}_n \boldsymbol{s}_n^\top\big] \right) \right) \tag{26}$$

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,

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