Gaussian mixtures, factor models, independent factor analysis, and mixtures of factor analyzers are all different restrictive types of Gaussian mixtures with density

$$p(\boldsymbol{y}_n) = \sum_{q=1}^{Q} \bar{w}_q \, \mathcal{N}(\boldsymbol{y}_n | \bar{\boldsymbol{\mu}}_q, \bar{\boldsymbol{\Sigma}}_q) \tag{1}$$

The mean and second moment of a Gaussian mixture is given by

$$\mathbb{E}[\boldsymbol{y}_n] = \sum_{q=1}^{Q} \bar{w}_q \bar{\boldsymbol{\mu}}_q \tag{2}$$

$$\mathbb{E}\big[\boldsymbol{y}_n \boldsymbol{y}_n^\top\big] = \sum_{q=1}^{Q} \bar{w}_q (\bar{\boldsymbol{\Sigma}}_q + \bar{\boldsymbol{\mu}}_q \bar{\boldsymbol{\mu}}_q^\top) \tag{3}$$

The covariance then follows from the standard definition,

$$\mathrm{Cov}(\boldsymbol{y}_n) = \mathbb{E}\big[\boldsymbol{y}_n \boldsymbol{y}_n^\top\big] - \mathbb{E}[\boldsymbol{y}_n] \, \mathbb{E}[\boldsymbol{y}_n]^\top . \tag{4}$$

In the above we implicitly conditioned on a number of hyperparameters. If we assume that they are all independent and that $\mathbb{E}[\bar{w}_q] = 1/Q$ and $\mathbb{E}[\bar{\boldsymbol{\mu}}_q] = \mathbf{0}$ and $\mathbb{E}\big[\bar{\boldsymbol{\mu}}_q \bar{\boldsymbol{\mu}}_q^\top\big] = \boldsymbol{\Lambda}_{\bar{\mu}}$, then

$$\mathbb{E}[\boldsymbol{y}_n] = \mathbf{0} \tag{5}$$

$$\mathbb{E}\big[\boldsymbol{y}_n \boldsymbol{y}_n^\top\big] = \mathrm{Cov}(\boldsymbol{y}_n) = \boldsymbol{\Lambda}_{\bar{\mu}} + \frac{1}{Q} \sum_{q=1}^{Q} \mathbb{E}\big[\bar{\boldsymbol{\Sigma}}_q\big] \tag{6}$$

If $\bar{\boldsymbol{\mu}}_q = \mathbf{0}$, then we can comfortably calculate the covariance to be

$$\mathrm{Cov}(\boldsymbol{y}_n) = \sum_{q=1}^{Q} \bar{w}_q \bar{\boldsymbol{\Sigma}}_q \tag{7}$$

or

$$\mathrm{Cov}(\boldsymbol{y}_n) = \frac{1}{Q} \sum_{q=1}^{Q} \mathbb{E}\big[\bar{\boldsymbol{\Sigma}}_q\big] \tag{8}$$

in the fully marginalized case. This case includes factor analyzers of various sorts. These models typically also assume that the covariance of each component has low-rank structure

$$\bar{\boldsymbol{\Sigma}}_q = \boldsymbol{A} \boldsymbol{\Sigma}_q \boldsymbol{A}^\top + \boldsymbol{\Lambda} \tag{9}$$

which then leads to a covariance matrix of the form

$$\mathrm{Cov}(\boldsymbol{y}_n) = \boldsymbol{A} \left(\sum_{q=1}^{Q} \bar{w}_q \boldsymbol{\Sigma}_q \right) \boldsymbol{A}^\top + \boldsymbol{\Lambda} . \tag{10}$$

In this case it can also be interesting to calculate the trace of the covariance matrix, which is simply

$$\mathrm{tr}(\mathrm{Cov}(\boldsymbol{y}_n)) = \mathrm{tr}\left(\boldsymbol{A}^\top \boldsymbol{A}\left(\sum_{q=1}^{Q} \bar{w}_q \boldsymbol{\Sigma}_q\right)\right) + \mathrm{tr}(\boldsymbol{\Lambda}). \tag{11}$$

The total covariance is also amenable to marginalization over the hyperparameters yielding a,

$$\mathrm{tr}(\mathrm{Cov}(\boldsymbol{y}_n)) = \mathrm{tr}\left(\mathbb{E}\big[\boldsymbol{A}^\top \boldsymbol{A}\big]\left(\frac{1}{Q}\sum_{q=1}^{Q} \mathbb{E}[\boldsymbol{\Sigma}_q]\right)\right) + \mathrm{tr}(\mathbb{E}[\boldsymbol{\Lambda}]). \tag{12}$$

If we assume that $\boldsymbol{A}$ is a standard Gaussian ensemble with $[\boldsymbol{A}]_{ij} \sim \mathcal{N}(0,1)$ then $\boldsymbol{A}^\top \boldsymbol{A}$ follows a Wishart distribution with mean $d\boldsymbol{I}$ where d is the dimensionality of $\boldsymbol{y}_n$. The total covariance then simplifies to

$$\mathrm{tr}(\mathrm{Cov}(\boldsymbol{y}_n)) = \frac{d}{Q}\sum_{q=1}^{Q} \mathrm{tr}(\mathbb{E}[\boldsymbol{\Sigma}_q]) + \mathrm{tr}(\mathbb{E}[\boldsymbol{\Lambda}]). \tag{13}$$

0.0.1 Centered Independent Factor Analysis

For a centered independent factor analysis model, $\bar{\boldsymbol{\Sigma}}_q$ is diagonal and has inverse Gamma distributions on its diagonal elements. Take the distribution to have mean m_{σ^2}. While the original model has unconstrained $\boldsymbol{\Lambda}$, we assume it to be a scaled unit matrix with the scaling factor also following an inverse Gamma distribution with mean m_{λ^2}. The resulting total covariance is

$$\mathrm{tr}(\mathrm{Cov}(\boldsymbol{y}_n)) = dKm_{\sigma^2} + dm_{\lambda^2}. \tag{14}$$

0.0.2 Projected Mixture

We define projected mixtures to be mixtures on a low-rank space that are then projected up into the observation space using a factor matrix $\boldsymbol{A}$ as with a factor model. Again we assume that,

$$\bar{\boldsymbol{\Sigma}}_q = \boldsymbol{A}\boldsymbol{\Sigma}_q\boldsymbol{A}^\top + \boldsymbol{\Lambda} \tag{15}$$

but now we let the mean be non-zero and we do not assume $\boldsymbol{\Sigma}_q$ to be diagonal. We can still follow the derivation that culminated with equation 13, only applying it to the non-zero mean covariance from equation 6, leaving

$$\mathrm{tr}(\mathrm{Cov}(\boldsymbol{y}_n)) = \mathrm{tr}(\boldsymbol{\Lambda}_{\bar{\mu}}) + \frac{d}{Q}\sum_{q=1}^{Q} \mathrm{tr}(\mathbb{E}[\boldsymbol{\Sigma}_q]) + \mathrm{tr}(\mathbb{E}[\boldsymbol{\Lambda}]). \tag{16}$$

We make the same assumptions about noise as for the centered IFA. We take the mean to be low-rank $\bar{\boldsymbol{\mu}}_q = \boldsymbol{A}\bar{\boldsymbol{\mu}}'_q$ where $\bar{\boldsymbol{\mu}}'_q\ \mathcal{N}(\boldsymbol{0}, \alpha_0\boldsymbol{I})$ so that

$$\mathrm{tr}(\mathrm{Cov}(\bar{\boldsymbol{\mu}}_q)) = \alpha_0\,\mathrm{tr}(\mathbb{E}\big[\boldsymbol{A}^\top \boldsymbol{A}\big]) = dK\alpha_0 \tag{17}$$

but take $\boldsymbol{\Lambda}_{\bar{\mu}} = \alpha_0 \boldsymbol{I}$ and let $\bar{\boldsymbol{\Sigma}}_q$ follow an inverse Wishart which has mean

$$\mathbb{E}\big[\bar{\boldsymbol{\Sigma}}_q\big] = \frac{\alpha}{\nu - K - 1}\boldsymbol{I} \tag{18}$$

for scale matrix $\alpha\boldsymbol{I}$ and degrees of freedom ν. This yields the total covariance,

$$\mathrm{tr}(\mathrm{Cov}(\boldsymbol{y}_n)) = d\alpha_0 + \frac{dK\alpha}{\nu - K - 1} + dm_{\lambda^2} \tag{19}$$

If we employ Wishart priors $\mathrm{Wis}(\alpha\boldsymbol{I}, \nu)$ instead of the conjugate inverse Wishart the expression simplifies to,

$$\mathrm{tr}(\mathrm{Cov}(\boldsymbol{y}_n)) = d\alpha_0 + dK\nu\alpha + dm_{\lambda^2} \tag{20}$$

0.1 Tuning the Inverse Gamma

For an inverse Gamma distribution $\mathrm{InvGamma}(\alpha, \beta)$ the mean and variance does not exist unless $\alpha > 2$, so we consider a random variable $X \sim \mathrm{InvGamma}(2 + \alpha, \beta)$. X then has mean and variance equal to

$$\mathbb{E}[X] = \frac{\beta}{\alpha + 1} \tag{21}$$

$$\mathrm{Var}(X) = \frac{\beta^2}{\alpha(\alpha+1)^2} \tag{22}$$

If we set $\mathbb{E}[X] = m$ and $\mathrm{Var}(X) = v$, then we can isolate the alpha and beta parameters in terms of the mean and variance as,

$$\beta = m\left(1 + \frac{m^2}{v}\right), \quad \alpha = \frac{m^2}{v} \tag{23}$$

Employing Markov's inequality, we can bound the tail probability using both the mean and the variance as

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}, \quad \mathbb{P}[X \geq a] \leq \frac{\mathrm{Var}(X)}{(a - \mathbb{E}[X])^2}. \tag{24}$$

Assume that the mean is set equal to m. Then if we want less than t mass in the tails, we can select the variance to be

$$\mathrm{Var}(X) = t(a - m)^2 \tag{25}$$

For a noise variable, we are likely to want a low mean variance like $m = 10^{-1}$, and we likely want to contain (at least) $1 - t = 0.95$ of the probability mass to $[0, 1]$ requiring $a = 1$. This results in a proposed variance of

$$\mathrm{Var}(X) = \frac{1}{20}(1 - 10^{-1})^2 = 0.0405 \tag{26}$$

this roughly holds if $\beta = 0.1247$ and $\alpha = 0.2469$ (or $\alpha = 2.2469$ in the original parameterization).

0.2 Tuning the Projected Mixture

For the projected mixture, we need to fix four effective parameters. We make the assumption that,

$$K\alpha_0 = \frac{K\alpha}{\nu - K - 1} \tag{27}$$

which corresponds to matching the total covariance contribution of the component means and the observations. Assuming the total covariance is d (corresponding to an identity covariance matrix) then we have,

$$1 - m_{\lambda^2} = \frac{2K\alpha}{\nu - K - 1}. \tag{28}$$

Using this, we can isolate α,

$$\alpha = (\nu - K - 1)\frac{1 - m_{\lambda^2}}{2K}. \tag{29}$$

For the inverse Wishart the diagonal elements $\bar{\mathbf{\Sigma}}_{ii}$ have variance,

$$\mathrm{Var}\big(\bar{\mathbf{\Sigma}}_{ii}\big) = \frac{2\alpha^2}{(\nu - K - 1)^2(\nu - K - 3)}. \tag{30}$$

If we fix the variance, we can insert the above value for α and isolate ν as

$$\nu = K + 3 + 2\frac{(\frac{1-m_{\lambda^2}}{2K})^2}{\mathrm{Var}\big(\bar{\mathbf{\Sigma}}_{ii}\big)} \tag{31}$$

Adding it back in we get the following formula for α,

$$\alpha = \left(1 + \frac{\left(\frac{1-m_{\lambda^2}}{2K}\right)^2}{\mathrm{Var}\big(\bar{\mathbf{\Sigma}}_{ii}\big)}\right)\frac{1 - m_{\lambda^2}}{K} \tag{32}$$

With Wishart priors, we start with

$$\alpha = \frac{1 - m_{\lambda^2}}{2K\nu}, \tag{33}$$

and using the variance expression,

$$\mathrm{Var}\big(\bar{\mathbf{\Sigma}}_{ii}\big) = 2\nu\alpha^2, \tag{34}$$

we can isolate ν,

$$\nu = \frac{(1 - m_{\lambda^2})^2}{2K^2\,\mathrm{Var}\big(\bar{\mathbf{\Sigma}}_{ii}\big)} \tag{35}$$

and then find α,

$$\alpha = \frac{K\,\mathrm{Var}\big(\bar{\mathbf{\Sigma}}_{ii}\big)}{1 - m_{\lambda^2}}. \tag{36}$$

0.3 Low-rank Gaussian density

The Gaussian density function is given by,

$$\frac{1}{\sqrt{(2\pi)^d|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{37}$$

If $\mathbf{\Sigma}$ is low-rank,

$$\mathbf{\Sigma} = \boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^\top + \boldsymbol{D} \tag{38}$$

then we can employ tricks to calculate the inverse more efficiently. The Woodbury matrix identity yields,

$$(\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^\top + \boldsymbol{D})^{-1} = \boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{A}(\boldsymbol{S}^{-1} + \boldsymbol{A}^\top\boldsymbol{D}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{D}^{-1}, \tag{39}$$

which can be simplified further if the cholesky factorization $\boldsymbol{L}\boldsymbol{L}^\top = \boldsymbol{S}$ is known, as we can pull the factors outside,

$$(\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^\top + \boldsymbol{D})^{-1} = \boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{L}(\boldsymbol{I} + \boldsymbol{L}^\top\boldsymbol{A}^\top\boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{L})^{-1}\boldsymbol{L}^\top\boldsymbol{A}^\top\boldsymbol{D}^{-1}. \tag{40}$$

If we define $\boldsymbol{B} = \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{L}$ then this can be written fairly succinctly as,

$$(\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^\top + \boldsymbol{D})^{-1} = \boldsymbol{D}^{-1} - \boldsymbol{B}(\boldsymbol{I} + \boldsymbol{B}^\top\boldsymbol{D}\boldsymbol{B})^{-1}\boldsymbol{B}^\top. \tag{41}$$

which only requires inversion of diagonal matrices and matrices of the same shape as $\boldsymbol{S}$, which can be smaller than $\mathbf{\Sigma}$ by design.

We can similarly apply the matrix determinant lemma to calculate the determinant as

$$|\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^\top + \boldsymbol{D}| = |\boldsymbol{D}||\boldsymbol{I} + \boldsymbol{L}^\top\boldsymbol{A}^\top\boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{L}| = |\boldsymbol{D}||\boldsymbol{I} + \boldsymbol{B}^\top\boldsymbol{D}\boldsymbol{B}| \tag{42}$$