

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

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

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

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

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

The covariance then follows from the standard definition,

$$\text{Cov}(\mathbf{y}_n) = \mathbb{E}[\mathbf{y}_n \mathbf{y}_n^\top] - \mathbb{E}[\mathbf{y}_n] \mathbb{E}[\mathbf{y}_n]^\top. \quad (4)$$

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 $\mathbb{E}[\bar{\boldsymbol{\mu}}_q \bar{\boldsymbol{\mu}}_q^\top] = \mathbf{\Lambda}_{\bar{\boldsymbol{\mu}}}$, then

$$\mathbb{E}[\mathbf{y}_n] = \mathbf{0} \quad (5)$$

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

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

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

or

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

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

$$\bar{\boldsymbol{\Sigma}}_q = \mathbf{A} \boldsymbol{\Sigma}_q \mathbf{A}^\top + \mathbf{\Lambda} \quad (9)$$

which then leads to a covariance matrix of the form

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

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

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

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

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

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

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

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

$$\text{tr}(\text{Cov}(\mathbf{y}_n)) = dK m_{\sigma^2} + d m_{\lambda^2}. \quad (14)$$

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 \mathbf{A} as with a factor model. Again we assume that,

$$\bar{\boldsymbol{\Sigma}}_q = \mathbf{A} \boldsymbol{\Sigma}_q \mathbf{A}^\top + \boldsymbol{\Lambda} \quad (15)$$

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

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

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

$$\text{tr}(\text{Cov}(\bar{\boldsymbol{\mu}}_q)) = \alpha_0 \text{tr}(\mathbb{E}[\mathbf{A}^\top \mathbf{A}]) = dK \alpha_0 \quad (17)$$

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

$$\mathbb{E}[\bar{\mathbf{\Sigma}}_q] = \frac{\alpha}{\nu - K - 1} \mathbf{I} \quad (18)$$

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

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

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

$$\text{tr}(\text{Cov}(\mathbf{y}_n)) = d\alpha_0 + dK\nu\alpha + dm_{\lambda^2} \quad (20)$$

0.1 Tuning the Inverse Gamma

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

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

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

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

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

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

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

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

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

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

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

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,

$$K\alpha_0 = \frac{K\alpha}{\nu - K - 1} \quad (27)$$

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,

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

Using this, we can isolate α ,

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

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

$$\text{Var}(\bar{\Sigma}_{ii}) = \frac{2\alpha^2}{(\nu - K - 1)^2(\nu - K - 3)}. \quad (30)$$

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

$$\nu = K + 3 + 2 \frac{\left(\frac{1 - m_{\lambda^2}}{2K}\right)^2}{\text{Var}(\bar{\Sigma}_{ii})} \quad (31)$$

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

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

With Wishart priors, we start with

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

and using the variance expression,

$$\text{Var}(\bar{\Sigma}_{ii}) = 2\nu\alpha^2, \quad (34)$$

we can isolate ν ,

$$\nu = \frac{(1 - m_{\lambda^2})^2}{2K^2 \text{Var}(\bar{\Sigma}_{ii})} \quad (35)$$

and then find α ,

$$\alpha = \frac{K \text{Var}(\bar{\Sigma}_{ii})}{1 - m_{\lambda^2}}. \quad (36)$$

0.2.1 Alternative Calibration of the Inverse Wishart

Alternatively we can use that if $\bar{\Sigma} \sim \mathcal{W}^{-1}(\alpha \mathbf{I}, \nu)$ then,

$$\bar{\Sigma}_{ii} \sim \Gamma^{-1}\left(\frac{\alpha}{2}, \frac{\nu - K - 1}{2}\right) \quad (37)$$

which means we can calibrate the Wishart as an Inverse Gamma. We should then set

$$\alpha_0 = \frac{\alpha}{\nu - K - 1} \quad (38)$$

to balance

0.3 Low-rank Gaussian density

The Gaussian density function is given by,

$$\frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \quad (39)$$

If Σ is low-rank,

$$\Sigma = \mathbf{A} \mathbf{S} \mathbf{A}^\top + \mathbf{D} \quad (40)$$

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

$$(\mathbf{A} \mathbf{S} \mathbf{A}^\top + \mathbf{D})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1} \mathbf{A} (\mathbf{S}^{-1} + \mathbf{A}^\top \mathbf{D}^{-1} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{D}^{-1}, \quad (41)$$

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

$$(\mathbf{A} \mathbf{S} \mathbf{A}^\top + \mathbf{D})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1} \mathbf{A} \mathbf{L} (\mathbf{I} + \mathbf{L}^\top \mathbf{A}^\top \mathbf{D}^{-1} \mathbf{A} \mathbf{L})^{-1} \mathbf{L}^\top \mathbf{A}^\top \mathbf{D}^{-1}. \quad (42)$$

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

$$(\mathbf{A} \mathbf{S} \mathbf{A}^\top + \mathbf{D})^{-1} = \mathbf{D}^{-1} - \mathbf{B} (\mathbf{I} + \mathbf{B}^\top \mathbf{D} \mathbf{B})^{-1} \mathbf{B}^\top. \quad (43)$$

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

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

$$|\mathbf{A} \mathbf{S} \mathbf{A}^\top + \mathbf{D}| = |\mathbf{D}| |\mathbf{I} + \mathbf{L}^\top \mathbf{A}^\top \mathbf{D}^{-1} \mathbf{A} \mathbf{L}| = |\mathbf{D}| |\mathbf{I} + \mathbf{B}^\top \mathbf{D} \mathbf{B}| \quad (44)$$