Density estimation in arbitrary dimension using a Dirichlet process mixture of multivariate normal with normal-inverse-Wishart base distribution

Alice Doucet Beaupré

September 20, 2022

1 Model

Our non-parametric generative model performs density estimation in arbitrary dimension using "mixtures of Dirichlet processes" (Escobar & West 1994, MacEarchern 1994). This is a bit of a misnomer given that there is only a single Dirichlet process which controls the number of components in a mixture of multivariate normal distributions.

One can think of this approach as as alternative to kernel density estimation (KDE).It differs from KDE in that it is a clustering algorithm where sample points are assigned to mixture components and each mixture component has its own covariance matrix. This latter property can be thought of as a more flexible generalization of the bandwidth in KDE.

One salient feature of the model is that, as mentionned above, the number of mixture components is arbitrary and itself a distribution inferred from the data and so is the covariance matrix of each component.

The model uses a non-parametric Chinese Restaurant Process (CRP) prior (Aldous 1983) over partitions of the sample points, i.e. over both the number of mixture components and over the assignments of sample points to those components. For the base distribution of the CRP we use the normal-inverse-Wishart (niW) distribution given that it is the conjugate prior to our data likelihood the multivariate normal distribution (mvn) with unknown mean and covariance. Conjugacy is desirable because it greatly simplifies the accompanying algorithm because we are able to integrate exactly over the parameters—the mean and covariance—of each mixture component.

Diving straight into it, the generative model is given by

$$\begin{aligned} \pi \mid \alpha &\sim \mathrm{CRP}(\alpha), \\ \mu_\omega, \Sigma_\omega \mid \pi, \mu_0, \lambda_0, \Psi_0, \nu_0 &\sim \text{normal-inverse-Wishart}(\mu_0, \lambda_0, \Psi_0, \nu_0), \quad \omega \in \pi, \\ x_j \mid \omega, \mu_\omega, \Sigma_\omega &\sim \text{multivariate-normal}(\mu_\omega, \Sigma_\omega), \quad j \in \omega. \end{aligned} \tag{1}$$

Here π is a partition over the set of sample points $\{x_j\}_{j=1}^N$. ω are parts, or clusters, of the partition π such that $\bigcup_{\omega\in\pi}\omega = \{x_j\}_{j=1}^N$ and $\omega \cap \omega' = \varnothing$ for ω and ω' distincts. Means μ_0^ω and sample points x_j are vectors in d-dimensions. The precision matrix Ψ_0 and covariances Σ_ω are d-dimensional positive-definite matrices.

1.1 The normal-inverse-Wishart distribution

The probability distribution function of the niW is given by

$$\mathrm{niW}(\mu, \Sigma \mid \mu_0, \lambda_0, \Psi_0, \nu_0) = \frac{1}{Z^0_{\mathrm{niW}}} |\Sigma|^{-\frac{\nu_0+d+2}{2}} e^{-\frac{1}{2}\operatorname{Tr}\Psi_0\Sigma^{-1} - \frac{\lambda_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)}, \tag{2}$$

where

$$Z^0_{\mathrm{niW}} = \frac{(2\pi)^{d/2} 2^{\nu_0 d/2} \Gamma_d(\frac{\nu_0}{2})}{\lambda_0^{d/2} |\Psi_0|^{\nu_0/2}},$$

and $\Gamma_d(x)$ is the multivariate gamma function

$$\Gamma_d(x) = \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma(a + (1-i)/2).$$

The niW distribution is the conjugate prior of the multivariate normal distribution. This means that the product between the niW prior and a mvn data likelihood gives us back a niW distribution with updated parameters, namely

$$\mathrm{niW}(\mu, \Sigma \mid \mu_0, \lambda_0, \Psi_0, \nu_0) \prod_{j\in\omega} \mathrm{mvn}(x_j \mid \mu, \Sigma) \propto \mathrm{niW}(\mu, \Sigma \mid \mu_0^\omega, \lambda_0^\omega, \Psi_0^\omega, \nu_0^\omega), \tag{3}$$

where hyperparameters superscripted ω depend on x_j's in ω and the original hyperparameters. Let the sample mean of a cluster $\bar{x} = \frac{1}{|\omega|}\sum_{j\in\omega} x_j$. Updated hyperparameters are given by

$$\begin{aligned}
\mu_0^\omega &= \frac{\lambda_0\mu_0 + |\omega|\bar{x}}{\lambda_0 + |\omega|}, \\
\lambda_0^\omega &= \lambda_0 + |\omega|, \\
\Psi_0^\omega &= \Psi_0 + \sum_{j\in\omega}(x_j - \bar{x})(x_j - \bar{x})^T + \frac{\lambda_0|\omega|}{\lambda_0 + |\omega|}(\bar{x} - \mu_0)(\bar{x} - \mu_0)^T, \\
\nu_0^\omega &= \nu_0 + |\omega|.
\end{aligned} \tag{4}$$

1.1.1 Derivation of updated hyperparameters

Let

$$z = (2\pi)^{d/2}.$$

From the definitions of the niW Eq. 5 and the mvn, lets assume that Eq. 3 can indeed be rewritten

$$\begin{aligned}
&\mathrm{niW}(\mu, \Sigma \mid \mu_0, \lambda_0, \Psi_0, \nu_0) \prod_{j \in \omega} \mathrm{mvn}(x_j \mid \mu, \Sigma) \\
&= \frac{1}{Z^0_{\mathrm{niW}}} |\Sigma|^{-\frac{\nu_0+d+2}{2}} e^{-\frac{1}{2}\operatorname{Tr}\Psi_0\Sigma^{-1} - \frac{\lambda_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)} \frac{1}{z^{|\omega|}} |\Sigma|^{-\frac{|\omega|}{2}} e^{-\frac{1}{2}\sum_{j\in\omega}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)}, \\
&= \frac{1}{Z^0_{\mathrm{niW}}} \frac{1}{z^{|\omega|}} \frac{Z^\omega_{\mathrm{niW}}}{Z^\omega_{\mathrm{niW}}} |\Sigma|^{-\frac{\nu^\omega_0+d+2}{2}} e^{-\frac{1}{2}\operatorname{Tr}\Psi^\omega_0\Sigma^{-1} - \frac{\lambda^\omega_0}{2}(\mu-\mu^\omega_0)^T\Sigma^{-1}(\mu-\mu^\omega_0)}, \\
&= \frac{Z^\omega_{\mathrm{niW}}}{Z^0_{\mathrm{niW}} z^{|\omega|}} \mathrm{niW}(\mu, \Sigma \mid \mu^\omega_0, \lambda^\omega_0, \Psi^\omega_0, \nu^\omega_0).
\end{aligned} \tag{5}$$

Combining powers of the determinant $|\Sigma|$ we find the updated

$$\nu^\omega_0 = \nu_0 + |\omega|.$$

Expanding the second and third lines above and gathering the powers of μ, which we highlight in parentheses, we obtain by inspection four equalities, namely

$$\begin{aligned}
(\lambda_0 + \sum_{j\in\omega} 1)\left(\mu^T\Sigma^{-1}\mu\right) &= \lambda^\omega_0\left(\mu^T\Sigma^{-1}\mu\right), \\
\left(\mu^T\Sigma^{-1}\right)(\lambda_0\mu_0 + \sum_{j\in\omega} x_j) &= \left(\mu^T\Sigma^{-1}\right)\lambda^\omega_0\mu^\omega_0, \\
(\lambda_0\mu_0 + \sum_{j\in\omega} x_j)\left(\Sigma^{-1}\mu\right) &= \lambda^\omega_0\mu^\omega_0\left(\Sigma^{-1}\mu\right), \\
\operatorname{Tr}\Psi_0\Sigma^{-1} + \lambda_0\mu_0^T\Sigma^{-1}\mu_0 + \sum_j x_j^T\Sigma^{-1}x_j &= \operatorname{Tr}\Psi^\omega_0\Sigma^{-1} + \lambda^\omega_0(\mu^\omega_0)^T\Sigma^{-1}\mu^\omega_0.
\end{aligned} \tag{6}$$

From the first equality we can read off

$$\lambda^\omega_0 = \lambda_0 + |\omega|.$$

Letting the sample mean $\bar{x} = \frac{1}{|\omega|}\sum_{j\in\omega} x_j$, the second and third equalities equivalently imply that

$$\mu^\omega_0 = \frac{\lambda_0\mu_0 + |\omega|\bar{x}}{\lambda_0 + |\omega|}.$$

Finally the third equality for Ψ^ω_0 requires a bit more algebra. First, using the cyclicity of the trace and the trace identity $\operatorname{Tr}(ba^T) = a^Tb$ for a and b two column vectors we can write, after dropping the trace operator, that

$$\Psi^\omega_0 = \Psi_0 + \lambda_0\mu_0\mu_0^T + \sum_{j\in\omega} x_jx_j^T - \lambda^\omega_0\mu^\omega_0(\mu^\omega_0)^T, \tag{7}$$

which we can rewrite

$$\begin{aligned}\Psi_0^\omega &= \Psi_0 + \sum_{j\in\omega}(x_j-\bar{x})(x_j-\bar{x})^T + |\omega|\bar{x}\bar{x}^T + \lambda_0\mu_0\mu_0^T - \lambda_0^\omega\mu_0^\omega(\mu_0^\omega)^T, \\ &= \Psi_0 + \sum_{j\in\omega}(x_j-\bar{x})(x_j-\bar{x})^T + \frac{\lambda_0|\omega|}{\lambda_0+|\omega|}\left(\mu_0\mu_0^T + \bar{x}\bar{x}^T - \bar{x}\mu_0^T - \mu_0\bar{x}^T\right), \\ &= \Psi_0 + \sum_{j\in\omega}(x_j-\bar{x})(x_j-\bar{x})^T + \frac{\lambda_0|\omega|}{\lambda_0+|\omega|}(\bar{x}-\mu_0)(\bar{x}-\mu_0)^T.\end{aligned} \tag{8}$$

This completes the derivation of Eq. 4. Notice in passing that as the number $|\omega|$ of samples $x_i \in \omega$ increases, all updated hyperparameters become overwhelmed by the data. Indeed the updated mean $\mu_0^\omega \to \bar{x}$, the udpated scale parameter $\lambda_0^\omega \to |\omega|$, the updated precision matrix $\Psi_0^\omega \to \sum_{j\in\omega}(x_j-\bar{x})(x_j-\bar{x})^T$, and the updated number of degrees of freedom $\nu_0^\omega \to |\omega|$, all functions of the data only.

1.1.2 Hyperpriors

The model specification Eq. 1 contains five hyperparameters: α, μ_0, λ_0, Ψ_0, and ν_0. As it stands, these hyperparameters have to be fixed to some values in order to completely specify the model. We could use an empirical Bayes approach but we don't like the idea of "double-dipping" in the data. There are some suggestions in the literature for fixed choices of the hyperparameters of the niW but none of them are completely satisfactory. For α, a vague logarithmic prior is sometimes used but this also seems arbitrary and solely base on the fact that $\alpha > 0$. We would like instead to close the hierarchy by introducing hyperpriors for hyperparameters and in doing so make the model parameter-free. We choose instead to use the (improper) independence Jeffreys prior, i.e. the product of the five one-parameter Jeffreys prior

$$P_{hyper}(\alpha,\mu_0,\lambda_0,\Psi_0,\nu_0) = J_\alpha(\alpha)J_{\mu_0}(\mu_0)J_{\lambda_0}(\lambda_0)J_{\Psi_0}(\Psi_0)J_{\nu_0}(\nu_0),$$

where $J_{\vec{\theta}}(\vec{\theta}) \propto \sqrt{|\mathcal{I}(\vec{\theta})|}$ and the Fisher information matrix $\mathcal{I}(\vec{\theta}) = -E[\partial_{\theta_i}\partial_{\theta_j}\log P(\vec{\theta})]$. This is a well behaved alternative to the full multi-parameters Jeffreys prior which unfortunately we found to be, of course, improper, but more damingly divergent in λ_0 and ν_0 and to depend on the ordering of the parameters.

For α, we recall the probability distribution of the CRP

$$P(\pi) = \frac{\alpha^K}{\alpha^{(N)}}\prod_{\lambda\in\pi}(|\lambda|-1)!,$$

where $K = |\pi|$ and $\alpha^{(N)} = \Gamma(\alpha+N)/\Gamma(\alpha)$ is the Pochhammer symbol. The 1×1 Fisher

information matrix is given by

$$\begin{aligned} -E[\partial_\alpha^2 \log P(\pi)] &= E\left[\frac{K}{\alpha^2} + \partial_\alpha^2 \log\Gamma(\alpha+N) - \partial_\alpha^2 \log\Gamma(\alpha)\right], \\ &= E\left[\frac{K}{\alpha^2} + \psi'(\alpha+N) - \psi'(\alpha)\right], \\ &= \frac{E[K]}{\alpha^2} + \psi'(\alpha+N) - \psi'(\alpha), \end{aligned} \tag{9}$$

where $\psi(x)$ is the digamma function and $\psi'(x)$ its derivative. The expectation value of the number of clusters K in a partition π distributed according the the CRP is given by a sum of Bernouilli variables, each parametrized by the probability that the m'th customer sits at a new table, namely

$$\begin{aligned} E[K] &= \sum_{m=1}^{N} E[b_m], \\ \text{where} \quad b_m \;&\sim\; \text{Bernoulli}\left[\frac{\alpha}{m-1+\alpha}\right], \\ \Rightarrow E[K] &= \sum_{m=1}^{N} \frac{\alpha}{m-1+\alpha}, \\ &= \alpha(\psi(\alpha+N) - \psi(\alpha)). \end{aligned} \tag{10}$$

We have therefore for α the Jeffreys prior

$$J_\alpha(\alpha) \propto \sqrt{\frac{\psi(\alpha+N)-\psi(\alpha)}{\alpha} + \psi'(\alpha+N) - \psi'(\alpha)}. \tag{11}$$

This prior differs from the often used logarithmic prior $1/\alpha$ in that $J_\alpha(\alpha) \sim \alpha^{-1/2}$ for $\alpha \to 0$ and $J_\alpha(\alpha) \sim \alpha^{-3/2}$ for $\alpha \to \infty$. In other words it gives more weight than the logarithmic prior at small α and less weight at large α, which is to say the logarithmic prior comparatively overestimates the number of clusters.

For μ_0, the $d \times d$ Fisher information matrix

$$\begin{aligned} -E\left[\partial_{\mu_{0i}}\partial_{\mu_{0j}} \log \text{niW}(\mu, \Sigma \mid \mu_0, \lambda_0, \Psi_0, \nu_0)\right] &= \frac{\lambda_0}{2} E[\partial_{\mu_i}\partial_{\mu_j}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)], \\ &= \lambda_0 E[\Sigma^{-1}], \\ &= \lambda_0\nu_0\Psi_0^{-1}. \end{aligned} \tag{12}$$

The last equality was obtained from the correspondance between the inverse-Wishart and Wishart distributions. Now notice that it does not depend on μ_0 and therefore

$$J_{\mu_0}(\mu_0) \propto 1$$

the uniform prior over $\mathbb{R}^d$.

For λ_0, the 1×1 Fisher information matrix is straightforward. Notice that the only term in the logarithm of the niW that does not disappear under the second derivative w.r.t. λ_0 is in the normalization constant. Indeed

$$\begin{aligned} -E\left[\partial^2_{\lambda_0} \log \mathrm{niW}(\mu, \Sigma \mid \mu_0, \lambda_0, \Psi_0, \nu_0)\right] &= -\frac{d}{2} E[\partial^2_{\lambda_0} \log \lambda_0], \\ &= \frac{d}{2}\frac{1}{\lambda_0^2}, \end{aligned} \tag{13}$$

and therefore

$$J_{\lambda_0}(\lambda_0) \propto \frac{1}{\lambda_0}$$

the logarithmic prior.

For Ψ_0, which we will write Ψ for now, the Fisher information matrix is a bit more interesting. Indeed

$$\begin{aligned} -E\left[\partial_{\Psi_{ij}}\partial_{\Psi_{kl}} \log \mathrm{niW}(\mu, \Sigma \mid \mu_0, \lambda_0, \Psi, \nu_0)\right] &= -\frac{\nu_0}{2} E[\partial_{\Psi_{ij}}\partial_{\Psi_{kl}} \log |\Psi|], \\ &= -\frac{\nu_0}{2}\partial_{\Psi_{ij}}\frac{1}{|\Psi|}\partial_{\Psi_{kl}}|\Psi|, \\ &= -\frac{\nu_0}{2}\partial_{\Psi_{ij}}\frac{1}{|\Psi|}|\Psi|\Psi^{-1}_{kl}, \\ &= -\frac{\nu_0}{2}\partial_{\Psi_{ij}}\Psi^{-1}_{kl}. \end{aligned} \tag{14}$$

We can show by taking the derivative of $A^{-1}A = \mathbb{1}$ that $\partial_A A^{-1} = -A^{-T} \otimes A^{-1}$ and therefore

$$-E\left[\partial^2_{\Psi_0} \log \mathrm{niW}(\mu, \Sigma \mid \mu_0, \lambda_0, \Psi_0, \nu_0)\right] = \frac{\nu_0}{2}\Psi_0^{-T} \otimes \Psi_0^{-1}.$$

It follows that the Jeffreys prior

$$\begin{aligned} J_{\Psi_0}(\Psi_0) &\propto \sqrt{\left|\Psi_0^{-T} \otimes \Psi_0^{-1}\right|}, \\ &= \frac{1}{|\Psi_0|^d}. \end{aligned} \tag{15}$$

Finally for ν_0, the 1×1 Fisher information matrix

$$-E\left[\partial^2_{\nu_0} \log \mathrm{niW}(\mu, \Sigma \mid \mu_0, \lambda_0, \Psi_0, \nu_0)\right] = \partial^2_{\nu_0} \log \Gamma_d(\nu_0/2)$$

where the multivariate gamma function

$$\Gamma_d(x) = \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\left(x + \frac{1-i}{2}\right).$$

Thus

$$\partial^2_{\nu_0} \log \Gamma_d(\nu_0/2) = \frac{1}{4} \sum_{i=1}^{d} \psi' \left(\frac{\nu_0}{2} + \frac{1-i}{2} \right),$$

and the Jeffreys prior

$$J_{\nu_0}(\nu_0) \propto \sqrt{\sum_{i=1}^{d} \psi' \left(\frac{\nu_0}{2} + \frac{1-i}{2} \right)}.$$

The behavior of this prior is such that $J_{\nu_0}(\nu_0) \sim (\nu_0 - d + 1)^{-1}$ for $\nu_0 \to d-1$ and $J_{\nu_0}(\nu_0) \sim \nu_0^{-1/2}$ for $\nu_0 \to \infty$.

Combining all single-parameter hyperpriors we found above, the full independence Jeffreys prior for the model

$$\begin{aligned} P_{hyper}(\alpha, \mu_0, \lambda_0, \Psi_0, \nu_0) \propto & \frac{1}{\lambda_0} \frac{1}{|\Psi_0|^d} \\ & \times \sqrt{\frac{\psi(\alpha+N) - \psi(\alpha)}{\alpha} + \psi'(\alpha+N) - \psi'(\alpha)} \\ & \times \sqrt{\sum_{i=1}^{d} \psi' \left(\frac{\nu_0}{2} + \frac{1-i}{2} \right)}. \end{aligned} \tag{16}$$

This hyperprior closes the hierarchy of hyperparameters and makes the model completely parameter-free.

1.1.3 Inadequacy of the multi-parameters Jeffreys prior

To give a sense of how badly behaved the full multi-parameters Jeffreys prior is, let us first look at the Fisher information matrix with parameter ordering $\vec{\theta} = \{\Psi_0, \nu_0, \lambda_0, \mu_0\}$ and ignoring α. The only difference is the appearance of off-diagonal blocks in $\partial_{\nu_0} \partial_{\Psi_{0ij}}$ and it is given by

$$\mathcal{I}(\vec{\theta}) = \begin{pmatrix} \frac{\nu_0}{2} \Psi_0^{-T} \otimes \Psi_0^{-1} & -\frac{1}{2} \operatorname{vec} \Psi_0^{-1} & & \\ -\frac{1}{2} \left(\operatorname{vec} \Psi_0^{-1} \right)^T & \frac{1}{4} \sum_{i=1}^{d} \psi'(\frac{\nu_0}{2} + \frac{1-i}{2}) & & \\ & & \frac{d}{2} \frac{1}{\lambda_0^2} & \\ & & & \lambda_0 \nu_0 \Psi_0^{-1} \end{pmatrix}, \tag{17}$$

where $\operatorname{vec}(M)$ is the vectorization operator. Using the determinant identity for block matrices

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |A||D - CA^{-1}B|$$

together with $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, we get

$$|\mathcal{I}(\vec{\theta})| = \frac{d}{2}\frac{1}{\lambda_0^2}\left|\lambda_0\nu_0\Psi_0^{-1}\right|\left|\frac{\nu_0}{2}\Psi_0^{-T} \otimes \Psi_0^{-1}\right|\left|D - \frac{1}{2\nu_0}(\operatorname{vec}\Psi_0^{-1})^T(\Psi_0 \otimes \Psi_0)\operatorname{vec}\Psi_0^{-1}\right|. \quad (18)$$

where $D = \frac{1}{4}\sum_{i=1}^{d}\psi'(\nu/2 + (1-i)/2)$. Now using the symmetry of Ψ_0 and identities $(A \otimes B)v = \operatorname{vec}(B(\operatorname{vec}^{-1} v)A^T)$ and $(\operatorname{vec} A)^T \operatorname{vec} B = \operatorname{Tr} A^T B$, we rewrite

$$\begin{aligned}(\operatorname{vec}\Psi_0^{-1})^T(\Psi_0 \otimes \Psi_0)\operatorname{vec}\Psi_0^{-1} &= (\operatorname{vec}\Psi_0^{-1})^T \operatorname{vec}(\Psi_0\Psi_0^{-1}\Psi_0),\\ &= (\operatorname{vec}\Psi_0^{-1})^T \operatorname{vec}\Psi_0,\\ &= \operatorname{Tr}\Psi_0^{-1}\Psi_0,\\ &= d.\end{aligned} \quad (19)$$

Therefore

$$\begin{aligned}|\mathcal{I}(\vec{\theta})| &= \frac{d}{2}\frac{1}{\lambda_0^2}\frac{(\lambda_0\nu_0)^d}{|\Psi_0|}\left(\frac{\nu_0}{2}\right)^{d^2}\frac{1}{|\Psi_0|^{2d}}\left(D - \frac{d}{2\nu_0}\right),\\ &= \frac{d}{2^{d^2+1}}\frac{\lambda_0^{d-2}\nu_0^{d(d+1)}}{|\Psi_0|^{2d+1}}\left(D - \frac{d}{2\nu_0}\right).\end{aligned} \quad (20)$$

and thus the multi-parameters Jeffreys prior

$$J_{multi}(\vec{\theta}) \propto \sqrt{|\mathcal{I}(\vec{\theta})|} \propto \lambda_0^{\frac{d}{2}-1}\nu_0^{\frac{d(d+1)}{2}}\frac{1}{|\Psi_0|^{d+\frac{1}{2}}}\sqrt{D - \frac{d}{2\nu_0}}.$$

Notice how terms in λ_0 for $d > 2$ and especially in ν_0 diverge rather dramatically, making the prior not only improper, but ill-defined for those two parameters. The divergence stems from the fact that we cannot immediately get rid of the factors ν_0 and $\lambda_0\nu_0$ in the blocks of Ψ_0 and μ_0 after taking the determinant and claiming the proportionality between the Jeffreys prior and the square-root of the determinant of the Fisher information matrix. There is no apparent remedy to this problem and we therefore stick with the independence Jeffreys prior.

2 Algorithm

The generative model is sampled using MCMC. Our MCMC algorithm uses three kinds of sampling strategies. For sampling the CRP we use both traditional Gibbs sampling (Neal 2000, algorithm 3) and the more ambitious split-merge moves (Jain & Neal 2004). Hyperparameters of the model for both the CRP and the niW base distributions have their own hyperpriors which we sample using a traditional Metropolis-Hastings approach.

We must begin with the expression for full generative model (omitting the hyperpriors). Letting $K = |\pi|$ and $N = |\{x_j\}| = \sum_{\omega\in\pi} |\omega|$,

$$\begin{aligned}
&P_{gen}(\pi, \{\mu_\omega\}, \{\Sigma_\omega\}) \\
&= \left[\frac{\alpha^K}{\alpha^{(N)}} \prod_{\omega\in\pi} (|\omega| - 1)!\right] \left[\prod_{\omega\in\pi} \text{niW}(\mu_\omega, \Sigma_\omega \mid \mu_0, \lambda_0, \Psi_0, \nu_0) \prod_{x_j\in\omega} \text{mvn}(x_j \mid \mu_\omega, \Sigma_\omega)\right], \\
&= \left[\frac{\alpha^K}{\alpha^{(N)}} \prod_{\omega\in\pi} (|\omega| - 1)!\right] \prod_{\omega\in\pi} \frac{Z^\omega_{\text{niW}}}{Z^0_{\text{niW}} z^{|\omega|}} \text{niW}(\mu_\omega, \Sigma_\omega \mid \mu^\omega_0, \lambda^\omega_0, \Psi^\omega_0, \nu^\omega_0).
\end{aligned} \tag{21}$$

Notice how the conjugacy allows us to easily marginalize over μ_ω and Σ_ω. Indeed

$$\begin{aligned}
P_{gen}(\pi) &= \left[\prod_{\omega\in\pi} \int d\mu_\omega d\Sigma_\omega\right] P(\pi, \{x_j\}, \{\mu^\omega_0\}, \{\Sigma_\omega\}), \\
&= \left[\frac{\alpha^K}{\alpha^{(N)}} \prod_{\omega\in\pi} (|\omega| - 1)!\right] \prod_{\omega\in\pi} \frac{Z^\omega_{\text{niW}}}{Z^0_{\text{niW}} z^{|\omega|}} \int d\mu_\omega d\Sigma_\omega \, \text{niW}(\mu_\omega, \Sigma_\omega \mid \mu^\omega_0, \lambda^\omega_0, \Psi^\omega_0, \nu^\omega_0), \\
&= \frac{\alpha^K}{\alpha^{(N)}} \prod_{\omega\in\pi} (|\omega| - 1)! \frac{Z^\omega_{\text{niW}}}{Z^0_{niW} z^{|\omega|}}.
\end{aligned} \tag{22}$$

This is the central expression from which we can derive the Gibbs moves as described in algorithm 3 of Neal 2000, which we do next.

2.1 Gibbs sampling of partitions

Let π_{-i} be the partition where we remove the sample point x_i from whichever cluster in π it was in. We now want to reassign x_i to either an existing cluster in π_{-i} or to its own new cluster. This reassignment gives rise to a new partition π'. There are two possibilities: either we reassign x_i to an existing cluster c in π_{-i}, namely $\pi' = \pi_{-i} - c + c \cup \{x_i\}$, or to its own cluster, namely $\pi' = \pi_{-i} + \{x_i\}$. Let's start with the second case where x_i is assigned its own cluster. Cancelling common factors in the numerator and denominator,

$$\begin{aligned}
P_{gen}(\pi' \mid \pi_{-i}) &= \frac{P_{gen}(\pi')}{P_{gen}(\pi_{-i})}, \qquad\qquad \pi' = \pi_{-i} + \{x_i\}, \\
&= \frac{\alpha^K 0!}{\alpha^{(N)}} \frac{\alpha^{(N-1)}}{\alpha^{K-1}} \frac{Z^{\{x_i\}}_{\text{niW}}}{Z^0_{niW} z}, \\
&= \frac{\alpha}{\alpha + N - 1} \frac{Z^{\{x_i\}}_{\text{niW}}}{Z^0_{niW} z}.
\end{aligned} \tag{23}$$

Now for the first case, the probability that x_i gets assigned to a cluster $c \in \pi_{-i}$, namely that $\pi' = \pi_{-i} - c + c \cup \{x_i\}$, is given by

$$\begin{aligned} P_{gen}(\pi' \mid \pi_{-i}) &= \frac{P_{gen}(\pi')}{P_{gen}(\pi_{-i})}, \\ &= \frac{\alpha^K}{\alpha^{(N)}} \frac{\alpha^{(N-1)}}{\alpha^K} \frac{|c|!}{(|c|-1)!} \frac{Z_{\text{niW}}^{c\cup\{x_i\}}}{Z_{\text{niW}}^0 z^{|c|+1}} \frac{Z_{\text{niW}}^0 z^{|c|}}{Z_{\text{niW}}^c}, \\ &= \frac{|c|}{\alpha + N - 1} \frac{Z_{\text{niW}}^{c\cup\{x_i\}}}{Z_{\text{niW}}^c z}. \end{aligned} \tag{24}$$

The algorithm for the Gibbs sampling is therefore straightforward. Pick a sample x_i at random and remove it from its cluster. Then draw at random either one of the remaining existing cluster $c \in \pi_{-i}$, each of which is assigned the probability Eq. 24, or a new empty cluster with probability Eq. 23. Add the sample x_i to this cluster. Repeat for all samples either at random with replacement or without replacement. We elect to sample without replacement because it guarantees that every sample is redrawn after one sweep. If one instead draws samples at random without replacement, the probability p that a given samples is picked at least once after x of sweeps, i.e. Nx draws, satisfies $1 - p = 1 - (1 - 1/N)^{Nx}$, and therefore $x = 1/N \log(1-p)/\log(1-1/N)$. For large N, the number or sweeps $x \to -\log(1-p)$. To give an idea of what this means, the number of sweeps with replacement necessary to be 95% sure that every sample has been picked up at least once is roughly 3. To be 99% sure one would need around 5 sweeps with replacement. Using sweeps with replacement therefore lead to an algorithm with slower mixing time than using sweeps without replacement. This is somewhat obvious.

2.2 Restricted Gibbs split-merge moves

Because Gibbs sample proceeds one sample at a time, it often gets stuck around false minima because samples gets reassigned to the same clusters they were in with high probability. This lead to very slow mixing time and poor exploration of the set of possible partitions. To counteract this issue one can use more ambitious moves whereby an "ambiguous" cluster can split into two, or two "nearby" clusters can merge into one. These moves are of the Metropolis-Hastings type because split-merge moves are proposed and accepted according to some acceptance probability rather than accepted with probability one as in Gibbs sampling. Indeed while it is easy to enumerate the full set of possible moves for a single sample—there are as many possible move as there are clusters, which is of order $\alpha \log N$—enumerating all possible ways to split a cluster is very often combinatorially explosive. We note though that there are only two possibilities when performing a merge move: either merge the two clusters under consideration or leave them as is. Therefore in this case Metropolis-Hastings is equivalent to Gibbs sampling.

2.3 Metropolis-Hastings move over hyperparameters

For hyperparameters μ_0, λ_0, ν_0, and α we proceed in the usual manner.

2.3.1 Moves for μ_0

For μ_0 we proceed one component μ_{0i} at a time. Since $\mu_0 \in \mathbb{R}^d$, we can use a Gaussian or uniform kernel with some width ϵ and, since these kernels are symmetric, the Hastings ratio is 1. The acceptance probability is therefore

$$P_{acc}(\mu_{0i} \to \mu'_{0i}) = \min\left[1, \frac{g(\mu_{0i} \mid \mu'_{0i})}{g(\mu'_{0i} \mid \mu_{0i})} \frac{P_{gen}(\mu'_{0i})}{P_{gen}(\mu_{0i})} \frac{J_{\mu_0}(\mu'_{0i})}{J_{\mu_0}(\mu_{0i})}\right] = \min\left[1, \frac{P_{gen}(\mu'_{0i})}{P_{gen}(\mu_{0i})}\right]$$

where $\mu'_{0i} = \mu_{0i} + \varepsilon$ with $\varepsilon \sim$ AnySymmetricDist.

2.3.2 Moves for λ_0

For λ_0 we use symmetric moves in logarithmic space $\log \lambda_0$ because $\lambda_0 > 0$. To transform those moves into uniform moves over the positive real line, we can find the Hastings ratio using the law of conservation of probability, namely

$$\begin{aligned}
\frac{g(\lambda_0 \mid \lambda'_0) d\lambda_0}{g(\lambda'_0 \mid \lambda_0) d\lambda'_0} &= \frac{g(\log \lambda_0 \mid \log \lambda'_0) d \log \lambda_0}{g(\log \lambda'_0 \mid \log \lambda_0) d \log \lambda'_0}, \\
\Rightarrow \quad \frac{g(\lambda_0 \mid \lambda'_0)}{g(\lambda'_0 \mid \lambda_0)} &= \frac{g(\log \lambda_0 \mid \log \lambda'_0)}{g(\log \lambda'_0 \mid \log \lambda_0)} \frac{\left|\frac{d \log \lambda_0}{d\lambda_0}\right|}{\left|\frac{d \log \lambda'_0}{d\lambda'_0}\right|}, \\
&= \frac{\lambda'_0}{\lambda_0}.
\end{aligned} \tag{25}$$

Interestingly,

$$\frac{J_{\lambda_0}(\lambda'_0)}{J_{\lambda_0}(\lambda_0)} = \frac{1}{\lambda'_0} \frac{\lambda_0}{1}$$

and therefore the ratio of Jeffreys priors cancels the Hastings ratio. This is to be expected since the hyperprior J_{λ_0} is logarithmic and we could have skipped both previous steps. We are left with

$$P_{acc}(\lambda_0 \to \lambda'_0) = \min\left[1, \frac{g(\lambda_0 \mid \lambda'_0)}{g(\lambda'_0 \mid \lambda_0)} \frac{P_{gen}(\lambda'_0)}{P_{gen}(\lambda_0)} \frac{J_{\lambda_0}(\lambda'_0)}{J_{\lambda_0}(\lambda_0)}\right] = \min\left[1, \frac{P_{gen}(\lambda'_0)}{P_{gen}(\lambda_0)}\right]$$

where $\log \lambda'_0 = \log \lambda_0 + \varepsilon$ with $\varepsilon \sim$ AnySymmetricDist.

2.3.3 Moves for ν_0

For ν_0 we once more use symmetric moves in logarithmic space, but since $\nu_0 > d-1$, the moves are actually over the shifted logarithmic space $\log(\nu_0 - d + 1)$. Since $d\log(\nu_0 - d + 1)/d(\nu_0 - d + 1) = 1/(\nu_0 - d + 1)$ we as in the case for λ_0 that

$$\frac{g(\nu_0 \mid \nu_0')}{g(\nu_0' \mid \nu_0)} = \frac{\nu_0' - d + 1}{\nu_0 - d + 1},$$

and therefore

$$P_{acc}(\nu_0 \to \nu_0') = \min\left[1, \frac{\nu_0' - d + 1}{\nu_0 - d + 1}\frac{P_{gen}(\nu_0')}{P_{gen}(\nu_0)}\frac{J_{\nu_0}(\nu_0')}{J_{\nu_0}(\nu_0)}\right],$$

where $\log(\nu_0' - d + 1) = \log(\nu_0 - d + 1) + \varepsilon$ with $\varepsilon \sim$ AnySymmetricDist.

2.3.4 Moves for Ψ_0

For Ψ_0 the Hastings ratio is substantially more complicated. Since Ψ_0 is positive definite we can first use the Cholesky decomposition $\Psi_0 = LL^T$ where L is a lower triangular matrix with entries $L_{ij} \in \mathbb{R}$ for all $i \geqslant j$ and 0 otherwise. We know it is easy to make symmetric moves $L_{ij} \to L_{ij} + \varepsilon$ with $\varepsilon \sim$ AnySymmetricDist. While making symmetric moves in the space of lower triangular matrices is therefore straightforward, how can we these into uniform moves in the space of positive definite matrices Ψ_0? We know that the general case

$$\frac{g(\Psi_0 \mid \Psi_0')d\Psi_0}{g(\Psi_0' \mid \Psi_0)d\Psi_0'} = \frac{g(L \mid L')dL}{g(L' \mid L)dL'}.$$

and thus

$$\frac{g(\Psi_0 \mid \Psi_0')}{g(\Psi_0' \mid \Psi_0)} = \frac{\left|\frac{d\Psi_0'}{dL'}\right|}{\left|\frac{d\Psi_0}{dL}\right|}.$$

We therefore need, writing once more Ψ instead of Ψ_0 for the moment, the determinant of the Jacobian

$$\left|\left[\frac{\partial \Psi_{ij}}{\partial L_{mn}}\right]_{ij,mn}\right| = \left|\left[\frac{\partial \sum_k L_{ik}L_{jk}}{\partial L_{mn}}\right]_{ij,mn}\right| = \left|\left[L_{in}\delta_{jm} + L_{jn}\delta_{im}\right]_{ij,mn}\right|, \quad i \geqslant j, m \geqslant n.$$

For $d = 1, 2, 3$ we have

$$
\begin{aligned}
d = 1 \Rightarrow & \quad \left|\frac{\partial \Psi_{ij}}{\partial L_{mn}}\right| = 2|L_{11}|, \\
d = 2 \Rightarrow & \quad \left|\frac{\partial \Psi_{ij}}{\partial L_{mn}}\right| = \begin{vmatrix} 2L_{11} & 0 & 0 \\ L_{21} & L_{11} & 0 \\ 0 & 2L_{21} & 2L_{22} \end{vmatrix} = 2^2 |L_{11}|^2 |L_{22}| \\
d = 3 \Rightarrow & \quad \left|\frac{\partial \Psi_{ij}}{\partial L_{mn}}\right| = \begin{vmatrix} 2L_{11} & 0 & 0 & 0 & 0 & 0 \\ L_{21} & L_{11} & 0 & 0 & 0 & 0 \\ L_{31} & 0 & L_{11} & 0 & 0 & 0 \\ 0 & 2L_{21} & 0 & 2L_{22} & 0 & 0 \\ 0 & L_{31} & L_{21} & L_{32} & L_{22} & 0 \\ 0 & 0 & 2L_{31} & 0 & 2L_{32} & 2L_{33} \end{vmatrix} = 2^3 |L_{11}|^3 |L_{22}|^2 |L_{33}|
\end{aligned}
\tag{26}
$$

and we claim without proof that in dimension d

$$\left|\frac{\partial \Psi_{ij}}{\partial L_{mn}}\right| = 2^d |L_{11}|^d |L_{22}|^{d-1} \dots |L_{dd}|.$$

The Hastings ratio we will form using this expression for the determinant of the Jacobian allows us to uniformly explore the space of positive definite matrices using easy symmetric moves in L. The acceptance probability, remembering Eq. 15,

$$P_{acc}(\Psi_0 \to \Psi_0') = \min\left[1, \frac{|L_{11}'|^d |L_{22}'|^{d-1} \dots |L_{dd}'|}{|L_{11}|^d |L_{22}|^{d-1} \dots |L_{dd}|} \frac{P_{gen}(\Psi_0')}{P_{gen}(\Psi_0)} \frac{|\Psi_0|^d}{|\Psi_0'|^d}\right].$$

where $L_{ij}' = L_{ij} + \varepsilon$ and $\varepsilon \sim$ AnySymmetricDist.

2.3.5 Moves for α

Since α also takes value on the positive real line $\alpha > 0$, we use once again symmetric moves in logarithmic space but because the hyperprior J_α is not logarithmic we use the Hastings ratio

$$\frac{g(\alpha \mid \alpha')}{g(\alpha' \mid \alpha)} = \frac{\alpha'}{\alpha}.$$

The acceptance probability

$$P_{acc}(\alpha \to \alpha') = \min\left[1, \frac{\alpha'}{\alpha} \frac{P_{gen}(\alpha')}{P_{gen}(\alpha)} \frac{J_\alpha(\alpha')}{J_\alpha(\alpha)}\right] = \frac{\alpha'^{K+1}}{\alpha^{K+1}} \frac{\Gamma(\alpha + N)}{\Gamma(\alpha)} \frac{\Gamma(\alpha')}{\Gamma(\alpha' + N)} \frac{J_\alpha(\alpha')}{J_\alpha(\alpha)}$$

where $\log \alpha' = \log \alpha + \varepsilon$ and $\varepsilon \sim$ AnySymmetricDist.