EURECOM
Sophia Antipolis

amadeus

Machine Learning for Nonparametric Unsupervised Fraud Detection

Dirichlet Process Mixture Model - Derivation

RÉMI DOMINGUES

TÉLÉCOM PARISTECH

EURECOM - SOPHIA ANTIPOLIS

Contents

1 Derivation 1
1.1 Kullback-Leibler divergence 1
1.2 One exponential family to rule them all 2
1.3 Approximating the underlying distribution 5
1.4 Coordinate ascent algorithm 6
1.5 Lower bound 10
1.6 Predictive density 12
1.7 Incremental training 12

Appendix A Derivation of exponential-family distributions **14**

Appendix B Derivation of conjugate priors in exponential family **20**

References **29**

Introduction

The Dirichlet Process Mixture Model algorithm presented here aggregates the variational inference method presented by Bishop in [1], the use of a Beta prior on the Dirichlet process responsible for the mixing proportions in [2] and the use of a Gamma prior on the concentration parameter of the Dirichlet process proposed by [3].

The current variational inference algorithm approximates the posterior distribution of the dataset by a mixture of multivariate Gaussians, inferring the mixing proportions from a stick-breaking process which concentration is inferred from a Gamma distribution.

1 Derivation

Our goal is to approximate the model evidence $P(\boldsymbol{x})$ and the posterior distribution $P(\boldsymbol{W}|\boldsymbol{x})$ by a variational distribution $q(\boldsymbol{W})$ using a method called mean field approximation, where $\boldsymbol{W}$ is a set of latent variables learnt by the algorithm.

1.1 Kullback-Leibler divergence

This is achieved using the reversed Kullback-Leibler defined in equation 1 where $\boldsymbol{\theta}$ is a set of hyperparameters used by the prior distribution.

$$D_{KL}(q||p) = \int q(\boldsymbol{W}) \ln \frac{q(\boldsymbol{W})}{p(\boldsymbol{W}|\boldsymbol{x}, \boldsymbol{\theta})} d\boldsymbol{W} \tag{1}$$

The KL divergence is equal to 0 when $q(\boldsymbol{W}) = p(\boldsymbol{W}|\boldsymbol{x})$. We thus want to minimize this divergence to obtain $q(\boldsymbol{W})$ as close as possible to the true posterior distribution.

$$\begin{aligned}
D_{KL}(q||p) &= -\int q(\boldsymbol{W}) \ln \frac{p(\boldsymbol{W}|\boldsymbol{x}, \boldsymbol{\theta})}{q(\boldsymbol{W})} d\boldsymbol{W} \\
D_{KL}(q||p) &= -\int q(\boldsymbol{W}) \ln \frac{p(\boldsymbol{W}, \boldsymbol{x}|\boldsymbol{\theta})}{q(\boldsymbol{W})} d\boldsymbol{W} + \ln p(\boldsymbol{x}|\boldsymbol{\theta}) \\
\ln p(\boldsymbol{x}|\boldsymbol{\theta}) &= \int q(\boldsymbol{W}) \ln \frac{p(\boldsymbol{W}, \boldsymbol{x}|\boldsymbol{\theta})}{q(\boldsymbol{W})} d\boldsymbol{W} + D_{KL}(q||p) \\
\ln p(\boldsymbol{x}|\boldsymbol{\theta}) &= \boldsymbol{\mathcal{L}}(q, \boldsymbol{\theta}) + D_{KL}(q||p)
\end{aligned}$$

Maximizing the lower bound $\boldsymbol{\mathcal{L}}$ defined in equation 2 is equivalent to minimizing $D_{KL}(q||p)$.

$$\boldsymbol{\mathcal{L}} = \int q(\boldsymbol{W}) \ln \frac{p(\boldsymbol{W}, \boldsymbol{x}|\boldsymbol{\theta})}{q(\boldsymbol{W})} d\boldsymbol{W} \tag{2}$$

Optimizing equation 2 is also achieved by maximizing the log marginal likelihood defined in equation 3 where $\mathbb{E}_q$ is the expectation with respect to the distribution q.

$$\begin{aligned} \ln p(\boldsymbol{x}|\boldsymbol{\theta}) &\geq \boldsymbol{\mathcal{L}}(q, \boldsymbol{\theta}) \\ &\geq \int q(\boldsymbol{W}) \ln \frac{p(\boldsymbol{W}, \boldsymbol{x}|\boldsymbol{\theta})}{q(\boldsymbol{W})} d\boldsymbol{W} \\ &\geq \mathbb{E}_q[\ln p(\boldsymbol{W}, \boldsymbol{x}|\boldsymbol{\theta})] - \mathbb{E}_q[\ln q(\boldsymbol{W})] \end{aligned} \tag{3}$$

The algorithm described thereafter provide a deterministic way to optimize the lower bound. The result obtained by equation 3 should thus increase at each iteration. This equation can thus be used to check implementation errors in the algorithm.

1.2 One exponential family to rule them all

The current algorithm approximates the data using a mixture of exponential-family distributions. To perform this approximation,the parameters of these likelihoods will averaged or sampled from their base distribution, a.k.a the posterior.

In order for the model to represent accurately a wide range of inputs, the derivation of the algorithm has been performed in exponential family. This representation allows the algorithm to handle numerous probability distributions with little changes.

The mapping of several probability distributions with their exponential family representation is given in appendix A. Most distributions are interesting choices for likelihoods, for which the exponential family representation of their base distribution is given in appendix B.

Exponential-family likelihoods and conjugate priors

Table 0.1 gives possible choices of representations based on the format of a given feature or set of features. Note that any bounded continuous data can be scaled if the bounds are known to fit between $[0, 1]$ or $[0, +\infty[$. The conjugate prior of the Dirichlet was introduced in [4].

Given the exponential-family likelihood of a mixture model containing an infinite number of components (equation 4 where $[z = i]$ is the Iverson bracket), the base distribution is computed in equation 5 where $\boldsymbol{\lambda_1}$ has the same dimension as $\boldsymbol{\eta}_i^*$ and λ_2 is a scalar. $\boldsymbol{\eta}_i^*$ and $\boldsymbol{\lambda}$ respectively contain the natural parameter(s) of the likelihood and the natural parameters of the base distribution. The base distribution has thus one parameter more than the likelihood.

$$p(\boldsymbol{x_n}|z_n, \boldsymbol{\eta}^*) = \prod_{i=1}^{\infty} \Big(h_l(\boldsymbol{x_n}) \exp(\boldsymbol{\eta_i^{*T}} T(\boldsymbol{x_n}) - a_l(\boldsymbol{\eta_i^*})) \Big)^{[z_n=i]} \tag{4}$$

$$p(\boldsymbol{\eta_i^*}|\boldsymbol{\lambda}) = h_b(\boldsymbol{\eta_i^*}) \exp(\boldsymbol{\lambda_1}^T \boldsymbol{\eta_i^*} + \lambda_2(-a_l(\boldsymbol{\eta_i^*})) - a_b(\boldsymbol{\lambda})) \tag{5}$$

Here, we distinguish the base measure h and the log-partition a of the likelihood and the base distribution with a subscript. Since the parameters alone allow this distinction, we did not include it in the remaining of this study.

Data description	Domain	Multivariate	Likelihood	Conjugate prior
Float $\in [0, 1]$	$[0, 1]$	No	Beta	$\propto \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^{\lambda_0} x_0^{\alpha} y_0^{\beta}$
Float $\in [0, 1]$	$[0, 1]$	Yes	Dirichlet	$\propto \frac{1}{B(\alpha)^{\eta}} e^{-\sum_{t=1}^{d} v_t \alpha_t}$
Integer $\in [0, +\infty[$	$\mathbb{N}$	No	Poisson	Gamma
Integer $\in [0, +\infty[$	$\mathbb{N}$	Yes	Multivariate Poisson	?
Float $\in [0, +\infty[$	$\mathbb{R}^+$	No	Gamma	$\propto \frac{p^{\alpha-1} e^{-\beta q}}{\Gamma(\alpha)^r \beta^{-\alpha s}}$
Float $\in [0, +\infty[$	$\mathbb{R}^+$	Yes	Multivariate Gamma	?
Float $\in]-\infty, +\infty[$	$\mathbb{R}$	No	Normal	Normal-Gamma
Float $\in]-\infty, +\infty[$	$\mathbb{R}$	Yes	Multivariate Normal	Normal-Wishart
Boolean	$\{True, False\}$	No	Binomial	Beta
Boolean	$\{True, False\}$	Yes	Multivariate Binomial	?
String, Boolean	Any number of distinct values	No	Multinomial	Dirichlet

Table 0.1 – Likelihood and conjugate prior according to data format

Data transformations and constraints

In Table 0.1, rows highlighted in light gray describe cases for which the conjugate prior still has to be investigated while dark gray rows describe cases where the analytical form of the normalization factor for the conjugate prior is not known. Due to this proportional form, we cannot compute $\mathbb{E}_q[\boldsymbol{\eta}^*]$ and $\mathbb{E}_q[-a(\boldsymbol{\eta}^*)]$ where $\boldsymbol{\eta}^*$ represents the natural parameter(s) of the likelihood. This computation requires indeed the derivative of the unknown log-partition (normalization factor) of the posterior which is unknown since the posterior has the same form as the conjugate prior.

To solve this constraint, we here apply a **transformation on the data**[1] so that univariate and multivariate data for which the domain is $[0;1]$ or $[0;+\infty[$ now become defined in $]-\infty;+\infty[$ and can thus be approximated by a mixture of multivariate normal instead of using Beta or Gamma univariate distributions. Let $\phi_p(x)$ be the cumulative distribution function (CDF) of a probability distribution p and $F_p^{-1}(x)$ be the inverse cumulative distribution (quantile function) of this distribution,

For $x \in [0,1]$, $F_N^{-1}(x) \in]-\infty,+\infty[$. Similarly if $x \in [0,+\infty[$, $\phi_\Gamma(F_N^{-1}(x)) \in]-\infty,+\infty[$. Note that the inverse mapping can be applied to the transformed data and results in the original data without any loss of information.

The CDF and inverse CDF of $N(0,1)$ are given in Figures 1 and 2, while figures 3 and 4 show the CDF and inverse CDF of $\Gamma(1,2)$ (shape and scale parameters).

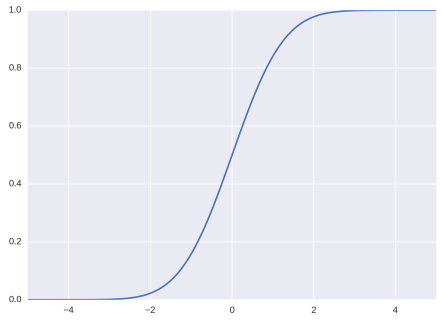

Figure 1 – $\phi_N(x), \mu = 0, \sigma = 1$

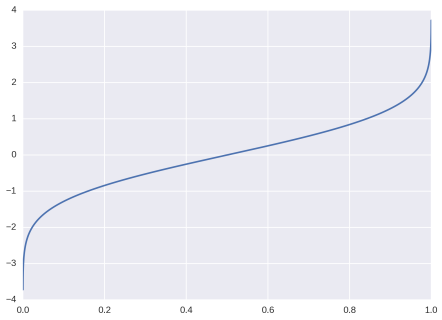

Figure 2 – $F_N^{-1}(x), \mu = 0, \sigma = 1$

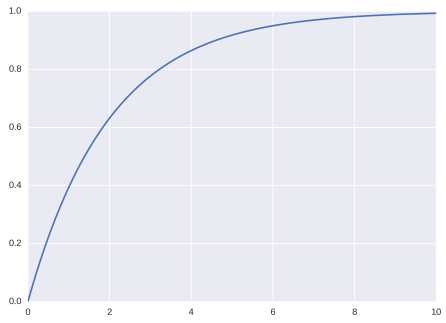

Figure 3 – $\phi_\Gamma(x), k = 1, \theta = 2$

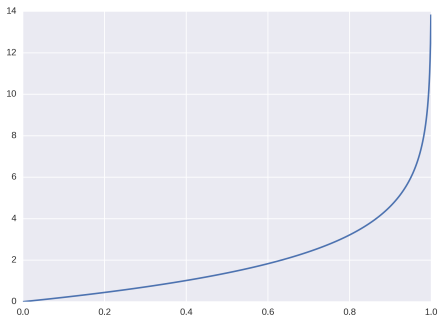

Figure 4 – $F_\Gamma^{-1}(x), k = 1, \theta = 2$

[1] $F_N^{-1}(Beta(\alpha,\beta),\mu_0,\sigma_0) \neq N(\mu,\sigma)$ except for $F_N^{-1}(Beta(1,1),\mu_0,\sigma_0) = N(\mu_0,\sigma_0)$. So we can't really give a prior on the data, nor get the inverse parameter mapping for the posterior. Hence prior parameters will have to be given for normal distributions instead of Beta or Gamma.

It must eventually be noted that multivariate distributions are able to efficiently express the correlations between features, while a loss of information will occur when using a product of distributions to represent a set of features. However, the Poisson distribution is very well suited to represent natural numbers. This is why using a product of mixtures of Poisson distributions must be compared with using a mixture of multivariate normal distributions when dealing with multivariate features for which the domain is $\mathbb{N}$.

1.3 Approximating the underlying distribution

Variational inference allows us to approximate likelihood and posterior distributions from a Dirichlet Process mixture prior. We now make the assumption that the data can be described by a product of probability distributions:

$$q(\boldsymbol{W}) = \prod_{i=1}^{M} q_i(\boldsymbol{W}_i) \tag{6}$$

We here choose the following approximation of the true posterior, setting $\boldsymbol{W} = \{\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}, w\}$:

$$q(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}, w) = q_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{v}) \cdot q_\tau(\boldsymbol{\eta}^*) \cdot q_r(\boldsymbol{z}) \cdot q_{g_1,g_2}(w) \tag{7}$$

Where $q_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{v})$ is a beta distribution, q_τ is an exponential-family distributions and $q_r(\boldsymbol{z})$ is a multinomial on the cluster assignment variable $\boldsymbol{z}$. Note that the product of exponential-family distributions is an exponential-family distribution, which allows q_τ to include Normal-Wishart, Gamma and Dirichlet posterior distributions.

The mixing proportions $\boldsymbol{\pi}$ computed in equation 8 are obtained from from a Dirichlet process, hence the use of a Beta distribution (equation 13) to sample the cluster weights v_i from a stick-breaking process ($v_i \sim Beta(1, w)$).

$$\pi_i(\boldsymbol{v}) = v_i \prod_{j=1}^{i-1}(1 - v_j) \tag{8}$$

w can be given as prior parameter, though it has a significant effect over the weight of each component and thus the number of components actually used by the fitted approximate posterior. This is why our model integrates over w, which become a latent variable.

A truncation parameter K on the number of clusters is used, which implies that $\pi_K(\boldsymbol{v}) = 0$ for $k > K$, thus $q(v_T = 1) = 1$. The current algorithm will be later extended by learning the truncation level K by variational inference, hence allowing an infinite number of clusters depending on the data complexity.

Notice that the first parameter of the Beta distribution is fixed to 1. We could have allowed a hyperparameter α_0 instead taking arbitrary values, then $q^*_{\alpha,\beta}(v)$ (eq.

20) would still be a Beta distribution of parameters $\alpha_k = \alpha_0 + N_k$ while β_k would be unchanged (eq. 23). However, we could no longer integrate out w as $q^*_{g_1,g_2}(w)$ (eq. 28) would no longer be a Gamma distribution.

We now write in equation 9 the joint probability of the random variables, with $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, s_0, r_0\}$.

$$p(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}, w|\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\eta}^*)p(\boldsymbol{z}|\boldsymbol{v})p(\boldsymbol{\eta}^*|\boldsymbol{\lambda})p(\boldsymbol{v}|w)p(w|s_0, r_0) \tag{9}$$

Defining hereafter the distributions, with N the size of the dataset:

$$p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\eta}^*) = \prod_{n=1}^{N}\prod_{k=1}^{K}\Big(h(\boldsymbol{x_n})\exp(\boldsymbol{\eta_k^{*T}}T(\boldsymbol{x_n}) - a(\boldsymbol{\eta_k^*}))\Big)^{z_{nk}} \tag{10}$$

$$\begin{aligned} p(\boldsymbol{z}|\boldsymbol{v}) &= \prod_{n=1}^{N}\prod_{k=1}^{K} Mult(\pi_k(\boldsymbol{v})) \\ &= \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k(\boldsymbol{v})^{z_{nk}} \\ &= \prod_{n=1}^{N}\prod_{k=1}^{K} (v_k \prod_{j=1}^{k-1}(1 - v_j))^{z_{nk}} \end{aligned} \tag{11}$$

$$p(\boldsymbol{\eta}^*|\boldsymbol{\lambda}) = \prod_{k=1}^{K} h(\boldsymbol{\eta_k^*})\exp(\boldsymbol{\lambda_1}^T\boldsymbol{\eta_k^*} + \lambda_2(-a(\boldsymbol{\eta_k^*})) - a(\boldsymbol{\lambda})) \tag{12}$$

$$p(\boldsymbol{v}|w) = \prod_{k=1}^{K} Beta(1, w) \tag{13}$$

$$p(w|s_0, r_0) = \Gamma(s_0, r_0) \tag{14}$$

Where s_0 and r_0 are respectively the shape and rate parameters of the Gamma prior on w.

1.4 Coordinate ascent algorithm

Since equation 7 is an approximation of equation 9, we now perform the derivation of each term of 7. Below, the star in $q_r*(\boldsymbol{z})$ denotes the expectation of this factor under all latent variables except $\boldsymbol{z}$. Those computation start from the joint probabilities defined in equation 9.

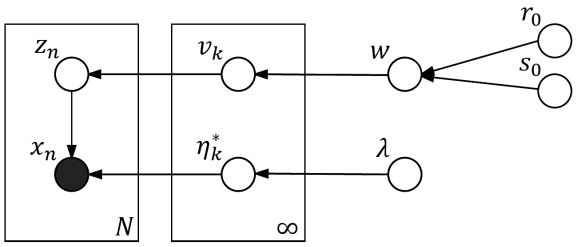

Figure 5 – Graphical model representing the Dirichlet Process Mixture Model according to the plate notation

$$\begin{aligned}
\ln q_r^*(\boldsymbol{z}) &= \mathbb{E}_{v,\boldsymbol{\eta}^*,w}[\ln p(\boldsymbol{x}, v, \boldsymbol{\eta}^*, \boldsymbol{z}, w)] + const \\
&= \mathbb{E}_{\boldsymbol{\eta}^*}[\ln p(\boldsymbol{x}|\boldsymbol{\eta}^*, \boldsymbol{z})] + \mathbb{E}_v[\ln p(\boldsymbol{z}|v)] + const \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\Big(\ln h(\boldsymbol{x_n}) + \mathbb{E}_q[\boldsymbol{\eta_k^*}]^T T(\boldsymbol{x_n}) + \mathbb{E}_q[-a(\boldsymbol{\eta_k^*})] \\
&+ \mathbb{E}[\ln v_k] + \sum_{i=1}^{k-1}\mathbb{E}[\ln(1-v_i)]\Big) + const
\end{aligned} \tag{15}$$

Taking the exponential of both sides, we get

$$q_r^*(\boldsymbol{z}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$$

With

$$\begin{aligned}
\ln \rho_{nk} &= \ln h(\boldsymbol{x_n}) + \mathbb{E}_q[\boldsymbol{\eta_k^*}]^T T(\boldsymbol{x_n}) + \mathbb{E}_q[-a(\boldsymbol{\eta_k^*})] \\
&+ \mathbb{E}[\ln v_k] + \sum_{i=1}^{k-1}\mathbb{E}[\ln(1-v_i)]
\end{aligned} \tag{16}$$

Where $h(\boldsymbol{x_n})$ and $T(\boldsymbol{x_n})$ are respectively the base measure and sufficient statistics of the likelihood distribution. Remember that $\forall n \ \sum_{k=1}^{K} z_{nk} = 1$ and $z_{nk} \in \{0, 1\}$. We can get rid of the proportionality by performing the following normalization:

$$q_r^*(\boldsymbol{z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}} \tag{17}$$

$$r_{nk} = \frac{\rho_{nk}}{\sum_{i=1}^{K}\rho_{ni}} \tag{18}$$

Thus

$$\mathbb{E}[z_{nk}] = r_{nk} \tag{19}$$

The current q distribution makes an approximation by setting an upper bound K on the number of clusters resulting in the truncation of the stick breaking process represented by the following beta distribution. This implies $q(v_K = 1) = 1$.

$$\begin{aligned}
\ln q^*_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{v}) &= \mathbb{E}_{\boldsymbol{\eta}^*,\boldsymbol{z},w}[\ln p(\boldsymbol{x},\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z},w)] + const \\
&= \mathbb{E}_{\boldsymbol{z}}[\ln p(\boldsymbol{z}|\boldsymbol{v})] + \mathbb{E}_w[\ln p(\boldsymbol{v}|w)] + const \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\left(\ln v_k + \sum_{i=1}^{k-1}\ln(1-v_i)\right) + \sum_{k=1}^{K-1}\Big((1-1)\ln v_k \\
&+ (\mathbb{E}[w]-1)\ln(1-v_k) - (\ln\Gamma(1) + \ln\Gamma(\mathbb{E}[w]) - \ln\Gamma(1+\mathbb{E}[w]))\Big) + const \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\ln v_k + \sum_{i=1}^{k-1}\ln(1-v_i)\right) \\
&+ \sum_{k=1}^{K-1}\Big((\mathbb{E}[w]-1)\ln(1-v_k) - \ln B(1,\mathbb{E}[w])\Big) + const \\
&= \sum_{k=1}^{K-1}\left(\mathbb{E}[w] + \sum_{n=1}^{N}\sum_{i=k+1}^{K} r_{ni} - 1\right)\ln(1-v_k) + \sum_{n=1}^{N} r_{nk}\ln v_k \\
&- \ln B(1,\mathbb{E}[w]) + const
\end{aligned} \tag{20}$$

Where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Taking the exponential of both sides, we recognize $q^*_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{v})$ as a Beta distribution.

$$q^*_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{v}) = \prod_{k=1}^{K-1} Beta(\alpha_k,\beta_k) \tag{21}$$

With

$$\alpha_k = 1 + N_k \tag{22}$$

$$\beta_k = \mathbb{E}[w] + \sum_{n=1}^{N}\sum_{i=k+1}^{K} r_{ni} \tag{23}$$

Where $N_k = \sum_{n=1}^{N} r_{nk}$

The next term of q is:

$$
\begin{aligned}
\ln q^*_{\boldsymbol{\tau}}(\boldsymbol{\eta}^*) &= \mathbb{E}_{\boldsymbol{v},\boldsymbol{z},w}[\ln p(\boldsymbol{x},\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z},w)] + const \\
&= \mathbb{E}_{\boldsymbol{z}}[\ln p(\boldsymbol{x}|\boldsymbol{\eta}^*,\boldsymbol{z})] + \ln p(\boldsymbol{\eta}^*|\boldsymbol{\lambda}) + const \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}[z_{nk}] \left(\ln h(\boldsymbol{x_n}) + \boldsymbol{\eta_k^{*T}} T(\boldsymbol{x_n}) - a(\boldsymbol{\eta_k^*})\right) \\
&+ \sum_{k=1}^{K} \left(\ln h(\boldsymbol{\eta_k^*}) + \boldsymbol{\lambda_1}^T \boldsymbol{\eta_k^*} - \lambda_2 a(\boldsymbol{\eta_k^*}) - a(\boldsymbol{\lambda})\right) + const \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \Bigg(\ln\left(h(\boldsymbol{x_n})^{r_{nk}} h(\boldsymbol{\eta_k^*})\right) + \left(r_{nk} T(\boldsymbol{x_n}) + \boldsymbol{\lambda_1}\right)^T \boldsymbol{\eta_k^*} - \left(\lambda_2 + r_{nk}\right) a(\boldsymbol{\eta_k^*}) \\
&- a(\boldsymbol{\lambda})\Bigg) + const
\end{aligned}
\tag{24}
$$

The exponential of this term is an exponential-family distribution taking the following parameters:

$$
q^*_{\boldsymbol{\tau}}(\boldsymbol{\eta}^*) = \prod_{k=1}^{K} h(\boldsymbol{\eta_k^*}) \exp(\boldsymbol{\tau_{k1}}^T \boldsymbol{\eta_k^*} + \tau_{k2}(-a(\boldsymbol{\eta_k^*})) - a(\boldsymbol{\tau_k})) \tag{25}
$$

$$
\boldsymbol{\tau_{k1}} = \boldsymbol{\lambda_1} + \sum_{n=1}^{N} r_{nk} T(\boldsymbol{x_n}) \tag{26}
$$

$$
\tau_{k2} = \lambda_2 + \sum_{n=1}^{N} r_{nk} \tag{27}
$$

Eventually, the last term of q is:

$$
\begin{aligned}
\ln q^*_{g_1,g_2}(w) &= \mathbb{E}_{\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}}[\ln p(\boldsymbol{x},\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z},w)] + const \\
&= \mathbb{E}_{\boldsymbol{v}}[\ln p(\boldsymbol{v}|w)] + \ln p(w|s_0,r_0) + const \\
&= \sum_{k=1}^{K-1} \left((w-1)\mathbb{E}_q[\ln(1-v_k)] - \ln\Gamma(w) + \ln\Gamma(w+1)\right) \\
&- \ln\Gamma(s_0) + s_0 \ln r_0 + (s_0-1)\ln w - r_0 w + const \\
&= (w-1)\sum_{k=1}^{K-1} \mathbb{E}_q[\ln(1-v_k)] + (K-1)\ln\frac{w\Gamma(w)}{\Gamma(w)} \\
&- \ln\Gamma(s_0) + s_0 \ln r_0 + (s_0-1)\ln w - r_0 w + const \\
&= (s_0 - 2 + K)\ln w - \left(r_0 - \sum_{k=1}^{K-1} \mathbb{E}_q[\ln(1-v_k)]\right) w \\
&- \mathbb{E}_q[\ln(1-v_k)] - \ln\Gamma(s_0) + s_0 \ln r_0 + const
\end{aligned}
\tag{28}
$$

$q^*_{g_1,g_2}(w)$ is thus a Γ distribution with shape g_1 and rate g_2.

$$g_1 = s_0 + K - 1 \tag{29}$$

$$g_2 = r_0 - \sum_{k=1}^{K-1} \mathbb{E}_q[\ln(1 - v_k)] \tag{30}$$

The expectations required to compute equation 19 are defined below, with ψ the derivative of the Γ function:

$$\mathbb{E}[\ln v_k] = \psi(\alpha_k) - \psi(\alpha_k + \beta_k) \tag{31}$$

$$\mathbb{E}[\ln(1 - v_k)] = \psi(\beta_k) - \psi(\alpha_k + \beta_k) \tag{32}$$

$$\mathbb{E}[w] = \frac{g_1}{g_2} \tag{33}$$

$$\mathbb{E}[\ln w] = \psi(g_1) - \ln g_2 \tag{34}$$

Note that $\mathbb{E}[\ln(1-v_K)]$ must be set to 0. $\mathbb{E}[\boldsymbol{\eta}^*_{\boldsymbol{k}}]$ and $\mathbb{E}[-a(\boldsymbol{\eta}^*_{\boldsymbol{k}})]$ depend on the analytical form of the posteriors and are detailed in Appendix B for most distributions.

This algorithm iterates between an expectation and maximization steps until a convergence is reached. The expectation step is composed of equations 31 to 34 and 19 which requires the expectation of the sufficient statistic terms depending on the underlying exponential-family distribution (See Appendix B). The maximization step contains equations 22, 23, 26, 27, 29 and 30. Equation 35 is used to monitor convergence, i.e. the iterations should stop when this bound does not increase more than a given threshold ϵ.

After convergence, the predictive density of a new data point is computed in equation 47.

1.5 Lower bound

We now continue the derivation of equation 3 by inserting the joint probability from equation 9 and the q distribution from equation 7.

$$\begin{aligned}
\ln p(\boldsymbol{x}|\boldsymbol{\theta}) &\geq \mathbb{E}_q[\ln p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\eta}^*, \boldsymbol{v}|\boldsymbol{\theta})] - \mathbb{E}_q[\ln q(\boldsymbol{z}, \boldsymbol{\eta}^*, \boldsymbol{v})] \\
&\geq \mathbb{E}_q[\ln p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\eta}^*)] + \mathbb{E}_q[\ln p(\boldsymbol{z}|\boldsymbol{v})] + \mathbb{E}_q[\ln p(\boldsymbol{\eta}^*|\boldsymbol{\lambda})] \\
&+ \mathbb{E}_q[\ln p(\boldsymbol{v}|w)] + \mathbb{E}_q[\ln p(w|s_0, r_0)] - \mathbb{E}_q[\ln q_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{v})] \\
&- \mathbb{E}_q[\ln q_{\boldsymbol{\tau}}(\boldsymbol{\eta}^*)] - \mathbb{E}_q[\ln q_r(\boldsymbol{z})] - \mathbb{E}_q[\ln q_{g_1,g_2}(w)]
\end{aligned} \tag{35}$$

Each term can be computed as follow, by taking the corresponding expectation of the logarithm under certain variables:

$$\mathbb{E}_q[\ln p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\eta}^*)] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\ln h(\boldsymbol{x_n}) + \mathbb{E}[\boldsymbol{\eta_k^*}]^T T(\boldsymbol{x_n}) + \mathbb{E}[-a(\boldsymbol{\eta_k^*})]\right) \tag{36}$$

$$\begin{aligned}\mathbb{E}_q[\ln p(\boldsymbol{z}|\boldsymbol{v})] &= \sum_{n=1}^{N}\sum_{k=1}^{\infty} r_{nk}\left(\mathbb{E}_q[\ln v_k] + \sum_{i=1}^{k-1}\mathbb{E}_q[\ln(1-v_i)]\right)\\ &= \sum_{n=1}^{N}\sum_{k=1}^{\infty}\left(\left(\sum_{i=k+1}^{\infty} r_{ni}\right)\mathbb{E}_q[\ln(1-v_k)] + r_{nk}\mathbb{E}_q[\ln v_k]\right)\\ &= \sum_{n=1}^{N}\sum_{k=1}^{K}\left(q(z_n>k)\mathbb{E}_q[\ln(1-v_k)] + q(z_n=k)\mathbb{E}_q[\ln v_k]\right)\end{aligned} \tag{37}$$

We truncated the summation at K, so $\mathbb{E}[\ln(1-v_K)] = 0$ and $q(z_n = k) = 0$ for $k > K$ where

$$q(z_n > k) = \sum_{i=k+1}^{K} r_{ni} \tag{38}$$

$$q(z_n = k) = r_{nk} \tag{39}$$

$$\mathbb{E}_q[\ln p(\boldsymbol{\eta}^*|\boldsymbol{\lambda})] = \sum_{k=1}^{K}\left(\ln h(\boldsymbol{\eta_k^*}) + \boldsymbol{\lambda_1}^T\mathbb{E}[\boldsymbol{\eta_k^*}] + \lambda_2\mathbb{E}[-a(\boldsymbol{\eta_k^*})] - a(\boldsymbol{\lambda})\right) \tag{40}$$

$$\mathbb{E}_q[\ln p(\boldsymbol{v}|w)] = \sum_{k=1}^{K}\left((\mathbb{E}[w]-1)\mathbb{E}[\ln(1-v_k)] - \ln\Gamma(\mathbb{E}[w]) + \ln\Gamma(\mathbb{E}[w]+1)\right) \tag{41}$$

$$\mathbb{E}_q[\ln p(w|s_0,r_0)] = s_0\ln r_0 - \ln\Gamma(s_0) + (s_0-1)\mathbb{E}[\ln w] - r_0\mathbb{E}[w] \tag{42}$$

$$\begin{aligned}\mathbb{E}_q[\ln q_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{v})] = \sum_{k=1}^{K}\Big((\alpha_k-1)\mathbb{E}[\ln(v_k)] + (\beta_k-1)\mathbb{E}[\ln(1-v_k)]\\ - \ln\Gamma(\alpha_k) - \ln\Gamma(\beta_k) + \ln\Gamma(\alpha_k+\beta_k)\Big)\end{aligned} \tag{43}$$

$$\mathbb{E}_q[\ln q_{\boldsymbol{\tau}}(\boldsymbol{\eta}^*)] = \sum_{k=1}^{K}\left(\ln h(\boldsymbol{\eta_k^*}) + \boldsymbol{\tau_{k1}}^T\mathbb{E}[\boldsymbol{\eta_k^*}] + \tau_{k2}\mathbb{E}[-a(\boldsymbol{\eta_k^*})] - a(\boldsymbol{\tau_k})\right) \tag{44}$$

$$\mathbb{E}_q[\ln q_r(\boldsymbol{z})] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ln r_{nk} \tag{45}$$

$$\mathbb{E}_q[\ln q_{g_1,g_2}(w)] = g_1\ln g_2 - \ln\Gamma(g_1) + (g_1-1)\mathbb{E}[\ln w] - g_2\mathbb{E}[w] \tag{46}$$

1.6 Predictive density

For a new data point $\boldsymbol{x}_{N+1}$, the predictive density is:

$$\begin{aligned} p(\boldsymbol{x}_{N+1}|\boldsymbol{x},\boldsymbol{\theta}) &= \int \sum_{k=1}^{\infty} \pi_k(\boldsymbol{v}) p(\boldsymbol{x}_{N+1}|\boldsymbol{\eta}_{\boldsymbol{k}}^*) dp(\boldsymbol{v},\boldsymbol{\eta}^*|\boldsymbol{x},\boldsymbol{\theta}) \\ &\approx \sum_{k=1}^{K} \mathbb{E}_q[\pi_k(\boldsymbol{v})] \mathbb{E}_q[p(\boldsymbol{x}_{N+1}|\boldsymbol{\eta}_{\boldsymbol{k}}^*)] \end{aligned} \tag{47}$$

Since $\pi_k(\boldsymbol{v})$ is given in equation 8 and v_i follows a Beta distribution, we obtain

$$\mathbb{E}_q[\pi_k(\boldsymbol{v})] = \frac{\alpha_k}{\alpha_k + \beta_k} \prod_{i=1}^{k-1} \left(1 - \frac{\alpha_i}{\alpha_i + \beta_i}\right) \tag{48}$$

This density is approximated by using a Monte Carlo estimate, i.e. we draw 1000 samples of $\boldsymbol{\eta}_k^*$ from the approximation of the posterior $q_{\boldsymbol{\tau}}^*(\boldsymbol{\eta}^*)$, each allowing us to compute the corresponding likelihood $p(\boldsymbol{x}_{N+1}|\boldsymbol{\eta}_k^*)$. The estimated likelihood for a given component is obtained by averaging the likelihood under each sample.

For a Normal likelihood and a Normal-Wishart approximation of the posterior, this density is a mixture of Student's t-distributions[1] given in equation 49. $\boldsymbol{\mu}_k$, λ_k, $\boldsymbol{V}_k$ and υ_k are the parameters of the Normal-Wishart distribution obtained from the inverse parameter mapping of $\boldsymbol{\tau}_k$.

$$p(\boldsymbol{x}_{N+1}|\boldsymbol{x},\boldsymbol{\theta}) = \sum_{k=1}^{K} \left(\frac{\alpha_k}{\alpha_k + \beta_k} \prod_{i=1}^{k-1} \left(1 - \frac{\alpha_i}{\alpha_i + \beta_i}\right) \mathrm{St}\left(\boldsymbol{x}_{N+1}|\boldsymbol{\mu}_k, \boldsymbol{L}_k, \upsilon_k + 1 - d\right) \right) \tag{49}$$

With

$$\boldsymbol{L}_k = \frac{(\upsilon + 1 - d)\lambda_k}{1 + \lambda_k} \boldsymbol{V}_k \tag{50}$$

We also report approximations of the log likelihood of a multivariate normal with $\Sigma^{-1} = L_k$ (Bishop. Note that $L_k \approx \frac{\sum^m W(V_k, \upsilon_k)}{m}$) in equation 51. The first approximation is must faster co compute, although we second one should be more accurate. Both remain rough approximations of the ground truth, which is the Student's distribution with L_k and a dedicated υ.

$$\ln N(x|m_k, \left(\frac{\sum^m W(V_k, \upsilon_k)}{m}\right)^{-1}) \approx \frac{\sum^m \ln N(x|m_k, W(V_k, \upsilon_k)^{-1})}{m} \tag{51}$$

1.7 Incremental training

In a production environment, the ideal model would be a never-ending learning one. Instead of adapting the whole algorithm in order to handle streaming data to

perform an incremental training, the simplest way to keep this model up to date with the current data flow while taking into account the data previously seen is to make the model evolve by performing periodic batch training where today's prior takes the value of yesterday's posterior.

Conclusion and future work

The following tasks have been achieved:

- Mean field variational inference for mixture of exponential-family distributions
- Handle several data types and complete derivation for most distributions required by the previous genericity
  - Float and integers of various ranges with data transformations and mixtures of multivariate normal and Poisson distributions
  - Strings with a mixture of categorical distributions
- Use a Dirichlet process to compute the the mixing proportions with a Beta prior
- Put a Gamma prior on the the scaling parameter w used by the Dirichlet process

Future work includes:

- Learn the truncation level K by variational inference
- Handle lists of actions by including HMM in the variational inference
- Distribute the algorithm

Additional steps while extending the algorithm are:

- Benchmark the algorithm on Amadeus' datasets
- Compare the algorithm to others, such as GMM, Collapsed Gibbs sampling and truncated Gibbs sampling

Appendix A

Derivation of exponential-family distributions

Exponential-family distributions are probability distributions which can be written into a specific form described in equation A.1, where $h(x)$ is a known function, $\eta(\theta)$ is the natural parameter, $T(x)$ the sufficient statistics and $A(\theta)$ the normalization factor of the exponential-family distribution.

$$h(x)\exp(\eta(\theta)\cdot T(x) - A(\theta)) \tag{A.1}$$

Table A.1 describes the mappings between the original distribution and its exponential-family representation. The inverse parameter mapping allows the computation of the distribution parameters from the natural parameters.

Distribution	**Parameter(s)** θ	**Natural parameters** η	**Inverse parameter mapping**	**Base measure** $h(x)$	**Sufficient statistic** $T(x)$	**Log-partition** $A(\theta)$
Binomial (known number of trials n)	p	$\ln \frac{p}{1-p}$	$\frac{1}{1+e^{-\eta}}$	$\binom{n}{x}$	x	$-n \ln(1-p)$
Multinomial (known number of trials n)	$p1, \cdots, p_k$ with $\sum_{i=1}^{k} p_i = 1$	$\begin{pmatrix} \ln p_1 \\ \vdots \\ \ln p_k \end{pmatrix}$	$\begin{pmatrix} e^{\eta_1} \\ \vdots \\ e^{\eta_k} \end{pmatrix}$ with $\sum_{i=1}^{k} e^{\eta_i} = 1$	$\frac{n!}{\prod_{i=1}^{k} x_i!}$	$\begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}$	0
Beta	α, β	$\begin{pmatrix} \alpha - 1 \\ \beta - 1 \end{pmatrix}$	$\begin{pmatrix} \eta_1 + 1 \\ \eta_2 + 1 \end{pmatrix}$	1	$\begin{pmatrix} \ln x \\ \ln(1-x) \end{pmatrix}$	$\ln \Gamma(\alpha) + \ln \Gamma(\beta) - \ln \Gamma(\alpha + \beta)$
Dirichlet	$\alpha_1, \cdots, \alpha_k$	$\begin{pmatrix} \alpha_1 - 1 \\ \vdots \\ \alpha_k - 1 \end{pmatrix}$	$\begin{pmatrix} \eta_1 + 1 \\ \vdots \\ \eta_k + 1 \end{pmatrix}$	1	$\begin{pmatrix} \ln x_1 \\ \vdots \\ \ln x_k \end{pmatrix}$	$\sum_{i=1}^{k} \ln \Gamma(\alpha_i) - \ln \Gamma\left(\sum_{j=1}^{k} \alpha_j\right)$
Gamma	α, β	$\begin{pmatrix} \alpha - 1 \\ -\beta \end{pmatrix}$	$\begin{pmatrix} \eta_1 + 1 \\ -\eta_2 \end{pmatrix}$	1	$\begin{pmatrix} \ln x \\ x \end{pmatrix}$	$\ln \Gamma(\alpha) - \alpha \ln \beta$
Poisson	λ	$\ln \lambda$	e^{η}	$\frac{1}{x!}$	x	λ
Multivariate normal (k dimensions)	$\boldsymbol{\mu}, \boldsymbol{\Sigma}$	$\begin{pmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{pmatrix}$	$\begin{pmatrix} -\frac{1}{2}\boldsymbol{\eta_2}^{-1}\boldsymbol{\eta_1} \\ -\frac{1}{2}\boldsymbol{\eta_2}^{-1} \end{pmatrix}$	$(2\pi)^{-\frac{d}{2}}$	$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{x}\boldsymbol{x}^T \end{pmatrix}$	$\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\ln\lvert\boldsymbol{\Sigma}\rvert$
Wishart (k dimensions)	$\boldsymbol{V}, n$	$\begin{pmatrix} -\frac{1}{2}\boldsymbol{V}^{-1} \\ \frac{n-d-1}{2} \end{pmatrix}$	$\begin{pmatrix} -\frac{1}{2}\boldsymbol{\eta_1}^{-1} \\ 2\eta_2 + d + 1 \end{pmatrix}$	1	$\begin{pmatrix} \boldsymbol{x} \\ \ln\lvert\boldsymbol{x}\rvert \end{pmatrix}$	$\frac{n}{2}(d \ln 2 + \ln\lvert\boldsymbol{V}\rvert) + \ln \Gamma_d(\frac{n}{2})$
Normal-Wishart (k dimensions)	$\boldsymbol{\mu}_0, \lambda, \boldsymbol{V}, n$	$\begin{pmatrix} \frac{n-d}{2} \\ -\frac{1}{2}(\boldsymbol{\mu}_0\boldsymbol{\mu}_0^T\lambda + \boldsymbol{V}^{-1}) \\ \boldsymbol{\mu}_0\lambda \\ -\frac{1}{2}\lambda \end{pmatrix}$	$\begin{pmatrix} -\frac{\boldsymbol{\eta_3}}{2\eta_4} \\ -2\eta_4 \\ \left(-2\boldsymbol{\eta_2} + \frac{\boldsymbol{\eta_3}\boldsymbol{\eta_3}^T}{2\eta_4}\right)^{-1} \\ 2\eta_1 + d \end{pmatrix}$	$(2\pi)^{-\frac{d}{2}}$	$\begin{pmatrix} \ln\lvert\boldsymbol{\Lambda}\rvert \\ \boldsymbol{\Lambda} \\ \boldsymbol{x}^T\boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}\boldsymbol{x}\boldsymbol{x}^T \end{pmatrix}$	$-\frac{d}{2}\ln\lambda + \frac{nd}{2}\ln 2 + \frac{n}{2}\ln\lvert\boldsymbol{V}\rvert + \ln \Gamma_d(\frac{n}{2})$
Conjugate prior of Gamma $f(\alpha, \beta \mid p, q, r, s) \propto \frac{p^{\alpha-1}e^{-\beta q}}{\Gamma(\alpha)^r \beta^{-\alpha s}}$	p, q, r, s	$\begin{pmatrix} r \\ s \\ \ln p \\ -q \end{pmatrix}$	$\begin{pmatrix} e^{\eta_3} \\ -\eta_4 \\ \eta_1 \\ \eta_2 \end{pmatrix}$	1	$\begin{pmatrix} \ln \Gamma(\alpha) \\ \alpha \ln \beta \\ \alpha \\ \beta \end{pmatrix}$	$\ln p$
Conjugate prior of Beta $\pi(\alpha, \beta \mid \lambda, x_0, y_0) \propto \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^{\lambda_0} x_0^{\alpha} y_0^{\beta}$	λ_0, x_0, y_0	$\begin{pmatrix} \lambda_0 \\ \ln x_0 \\ \ln y_0 \end{pmatrix}$	$\begin{pmatrix} \eta_1 \\ e^{\eta_2} \\ e^{\eta_3} \end{pmatrix}$	1	$\begin{pmatrix} \ln\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) \\ \alpha \\ \beta \end{pmatrix}$	0

Table A.1 – Exponential-family representation of several probability distributions

Multivariate normal

We here took advantage of following property $\text{tr}(a^T \cdot b) = vec(a) \cdot vec(b)$ with a and b vectors. We thus assume a vectorization of the matrices at the second-to-last step of the Normal and Normal-Wishart derivations. This allows us to use the trace property: $\text{tr}(a^T Bc) = \text{tr}(ca^T B)$.

Let d be the dimension of the space,

$$\begin{aligned}
N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \\
&= \frac{1}{\sqrt{2\pi}^d} \exp(-\frac{1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})) \\
&= \frac{1}{\sqrt{2\pi}^d} \exp(\text{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\boldsymbol{x}^T\right) - \frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\boldsymbol{x}^T\right) - \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\ln|\boldsymbol{\Sigma}|) \\
&= \frac{1}{\sqrt{2\pi}^d} \exp\left(\begin{pmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{pmatrix}^T \cdot \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{x}\boldsymbol{x}^T \end{pmatrix} - (\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\ln|\boldsymbol{\Sigma}|)\right) \\
&= h(\boldsymbol{x}) \exp(\eta(\boldsymbol{\theta}) \cdot T(\boldsymbol{x}) - A(\boldsymbol{\theta}))
\end{aligned} \tag{A.2}$$

This gives us the following parameter and inverse parameter mappings:

$$\begin{cases} \boldsymbol{\eta_1} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ \boldsymbol{\eta_2} = -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{cases} \qquad \begin{cases} \boldsymbol{\mu} = -\frac{1}{2}\boldsymbol{\eta_2}^{-1}\boldsymbol{\eta_1} \\ \boldsymbol{\Sigma} = -\frac{1}{2}\boldsymbol{\eta_2}^{-1} \end{cases}$$

Wishart

$$\begin{aligned}
W(\boldsymbol{x}|\boldsymbol{V}, n) &= \frac{|\boldsymbol{x}|^{\frac{n-d-1}{2}} e^{\frac{-\text{tr}(\boldsymbol{V}^{-1}\boldsymbol{x})}{2}}}{2^{\frac{nd}{2}}|\boldsymbol{V}|^{\frac{n}{2}}\Gamma_d(\frac{n}{2})} \\
&= \exp(-\frac{1}{2}\text{tr}(\boldsymbol{V}^{-1}\boldsymbol{x}) + \frac{n-d-1}{2}\ln|\boldsymbol{x}| - \frac{nd}{2}\ln 2 - \frac{n}{2}\ln|\boldsymbol{V}| - \ln\Gamma_d(\frac{n}{2})) \\
&= \exp\left(\begin{pmatrix} -\frac{1}{2}\boldsymbol{V}^{-1} \\ \frac{n-d-1}{2} \end{pmatrix}^T \cdot \begin{pmatrix} \boldsymbol{x} \\ \ln|\boldsymbol{x}| \end{pmatrix} - (\frac{n}{2}(d\ln 2 + \ln|\boldsymbol{V}|) + \ln\Gamma_d(\frac{n}{2}))\right) \\
&= h(\boldsymbol{x}) \exp(\eta(\boldsymbol{\theta}) \cdot T(\boldsymbol{x}) - A(\boldsymbol{\theta}))
\end{aligned} \tag{A.3}$$

Where Γ_d is the multivariate gamma function and the parameter mappings are

$$\begin{cases} \boldsymbol{\eta_1} = -\frac{1}{2}\boldsymbol{V}^{-1} \\ \boldsymbol{\eta_2} = \frac{n-d-1}{2} \end{cases} \qquad \begin{cases} \boldsymbol{V} = -\frac{1}{2}\boldsymbol{\eta_1}^{-1} \\ n = 2\eta_2 + d + 1 \end{cases}$$

Normal-Wishart

The following derivation also assumes a preliminary vectorization of the matrices as explained above.

$$\boldsymbol{\Lambda}|\boldsymbol{V}, n \sim W(\boldsymbol{\Lambda}|\boldsymbol{V}, n) \tag{A.4}$$

$$\boldsymbol{\mu}|\boldsymbol{\mu}_0, \lambda, \boldsymbol{\Lambda} \sim N(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\lambda\boldsymbol{\Lambda})^{-1}) \tag{A.5}$$

$$(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \sim NW(\boldsymbol{\mu}_0, \lambda, \boldsymbol{V}, n) \tag{A.6}$$

$$\begin{aligned}
NW(\boldsymbol{x}, \boldsymbol{\Lambda}|\boldsymbol{\mu}_0, \lambda, \boldsymbol{V}, n) &= \frac{|\boldsymbol{\Lambda}|^{\frac{n-d-1}{2}} e^{\frac{-\operatorname{tr}(\boldsymbol{V}^{-1}\boldsymbol{\Lambda})}{2}}}{2^{\frac{nd}{2}}|\boldsymbol{V}|^{\frac{n}{2}}\Gamma_d(\frac{n}{2})}(2\pi)^{-\frac{d}{2}}|(\lambda\boldsymbol{\Lambda})^{-1}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T\lambda\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu}_0)} \\
&= \frac{1}{\sqrt{2\pi}^d}\exp(\operatorname{tr}\left(\lambda\boldsymbol{\Lambda}\boldsymbol{\mu}_0\boldsymbol{x}^T\right) - \frac{1}{2}\operatorname{tr}\left(\lambda\boldsymbol{\Lambda}\boldsymbol{x}\boldsymbol{x}^T\right) - \frac{1}{2}\boldsymbol{\mu}_0^T\lambda\boldsymbol{\Lambda}\boldsymbol{\mu}_0 \\
&- \frac{1}{2}\ln|(\lambda\boldsymbol{\Lambda})^{-1}| - \frac{1}{2}\operatorname{tr}(\boldsymbol{V}^{-1}\boldsymbol{\Lambda}) + \frac{n-d-1}{2}\ln|\boldsymbol{\Lambda}| - \frac{nd}{2}\ln 2 \\
&- \frac{n}{2}\ln|\boldsymbol{V}| - \ln\Gamma_d(\frac{n}{2})) \\
&= \frac{1}{\sqrt{2\pi}^d}\exp(\frac{n-d-1}{2}\ln|\boldsymbol{\Lambda}| + \frac{d}{2}\ln\lambda + \frac{1}{2}\ln|\boldsymbol{\Lambda}| - \frac{1}{2}\boldsymbol{\mu}_0^T\lambda\boldsymbol{\Lambda}\boldsymbol{\mu}_0 \\
&- \frac{1}{2}\operatorname{tr}(\boldsymbol{V}^{-1}\boldsymbol{\Lambda}) + \operatorname{tr}\left(\lambda\boldsymbol{\Lambda}\boldsymbol{\mu}_0\boldsymbol{x}^T\right) - \frac{1}{2}\operatorname{tr}\left(\lambda\boldsymbol{\Lambda}\boldsymbol{x}\boldsymbol{x}^T\right) - \frac{nd}{2}\ln 2 \\
&- \frac{n}{2}\ln|\boldsymbol{V}| - \ln\Gamma_d(\frac{n}{2})) \\
&= \frac{1}{\sqrt{2\pi}^d}\exp\left(\begin{pmatrix} \frac{n-d}{2} \\ -\frac{1}{2}(\boldsymbol{\mu}_0\boldsymbol{\mu}_0^T\lambda + \boldsymbol{V}^{-1}) \\ \boldsymbol{\mu}_0\lambda \\ -\frac{1}{2}\lambda \end{pmatrix}^T \cdot \begin{pmatrix} \ln|\boldsymbol{\Lambda}| \\ \boldsymbol{\Lambda} \\ \boldsymbol{x}^T\boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}\boldsymbol{x}\boldsymbol{x}^T \end{pmatrix} - (-\frac{d}{2}\ln\lambda \right. \\
&\left. + \frac{nd}{2}\ln 2 + \frac{n}{2}\ln|\boldsymbol{V}| + \ln\Gamma_d(\frac{n}{2}))\right) \\
&= h(\boldsymbol{x})\exp(\eta(\boldsymbol{\theta})\cdot T(\boldsymbol{x}) - A(\boldsymbol{\theta}))
\end{aligned} \tag{A.7}$$

Which results in the following parameter mappings

$$\begin{cases} \eta_1 = \frac{n-d}{2} \\ \boldsymbol{\eta_2} = -\frac{1}{2}(\boldsymbol{\mu_0}\boldsymbol{\mu_0}^T\lambda + \boldsymbol{V}^{-1}) \\ \boldsymbol{\eta_3} = \boldsymbol{\mu_0}\lambda \\ \eta_4 = -\frac{1}{2}\lambda \end{cases} \qquad \begin{cases} \boldsymbol{\mu_0} = -\frac{\boldsymbol{\eta_3}}{2\eta_4} \\ \lambda = -2\eta_4 \\ \boldsymbol{V} = \left(-2\boldsymbol{\eta_2} + \frac{\boldsymbol{\eta_3}\boldsymbol{\eta_3}^T}{2\eta_4}\right)^{-1} \\ n = 2\eta_1 + d \end{cases}$$

Conjugate prior of the Beta distribution

The hyperparameters of this prior are λ_0, x_0 and y_0. However, its normalization factor does not have a closed form which limits the possible uses of this distribution.

$$\begin{aligned} \pi(\alpha, \beta|\lambda_0, x_0, y_0) &\propto \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^{\lambda_0} x_0^\alpha y_0^\beta \\ &\propto \exp\left(\lambda_0 \ln\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) + \alpha \ln x_0 + \beta \ln y_0\right) \\ &\propto \exp\left(\begin{pmatrix} \lambda_0 \\ \ln x_0 \\ \ln y_0 \end{pmatrix}^T \cdot \begin{pmatrix} \ln\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) \\ \alpha \\ \beta \end{pmatrix}\right) \\ &= h(\boldsymbol{x}) \exp(\eta(\boldsymbol{\theta}) \cdot T(\boldsymbol{x}) - A(\boldsymbol{\theta})) \end{aligned} \tag{A.8}$$

Where the parameter mappings are

$$\begin{cases} \eta_1 = \lambda_0 \\ \eta_2 = \ln x_0 \\ \eta_3 = \ln y_0 \end{cases} \qquad \begin{cases} \lambda_0 = \eta_1 \\ x_0 = e_2^\eta \\ y_0 = e_3^\eta \end{cases}$$

Conjugate prior of the Gamma distribution

The hyperparameters of this prior are p, q, r and s.

$$
\begin{aligned}
f(\alpha,\beta|p,q,r,s) &\propto \frac{p^{\alpha-1}e^{-\beta q}}{\Gamma(\alpha)^r\beta^{-\alpha s}} \\
&\propto p^{\alpha-1}e^{-\beta q}\Gamma(\alpha)^{-r}\beta^{\alpha s} \\
&\propto \exp\left((\alpha-1)\ln p-\beta q-r\ln\Gamma(\alpha)+\alpha s\ln\beta\right) \\
&\propto \exp\left(\begin{pmatrix} r \\ s \\ \ln p \\ -q \end{pmatrix}^T \cdot \begin{pmatrix} \ln\Gamma(\alpha) \\ \alpha\ln\beta \\ \alpha \\ \beta \end{pmatrix} - \ln p\right) \\
&= h(\boldsymbol{x})\exp(\eta(\boldsymbol{\theta})\cdot T(\boldsymbol{x})-A(\boldsymbol{\theta}))
\end{aligned}
\tag{A.9}
$$

Where the parameter mappings are

$$
\begin{cases} \eta_1 = r \\ \eta_2 = s \\ \eta_3 = \ln p \\ \eta_4 = -q \end{cases}
\qquad\qquad
\begin{cases} p = e^{\eta_3} \\ q = -\eta_4 \\ r = \eta_1 \\ s = \eta_2 \end{cases}
$$

Appendix B

Derivation of conjugate priors in exponential family

Given an exponential-family likelihood expressed in equation B.1, its base distribution is given by equation B.2, where λ_1 has the same dimension as η^* and λ_2 is a scalar. The base distribution has thus one parameter more than the likelihood.

$$p(x|\eta^*) = h_l(x)\exp(\eta^{*T}T(x) - a_l(\eta^*)) \tag{B.1}$$

$$p(\eta^*|\boldsymbol{\lambda}) = h_b(\eta^*)\exp(\lambda_1^T\eta^* + \lambda_2(-a_l(\eta^*)) - a_b(\boldsymbol{\lambda})) \tag{B.2}$$

In the case of a conjugate prior, the posterior is thus

$$p(\eta^*|\boldsymbol{\tau}) = h_b(\eta^*)\exp(\tau_1^T\eta^* + \tau_2(-a_l(\eta^*)) - a_b(\boldsymbol{\tau})) \tag{B.3}$$

We can then compute the expectation of each term of the sufficient statistics:

$$\mathbb{E}[\eta^*] = \frac{\partial a_b(\tau_1, \cdots)}{\partial \tau_1} \tag{B.4}$$

$$\mathbb{E}[-a_l(\eta^*)] = \frac{\partial a_b(\cdots, \tau_2)}{\partial \tau_2} \tag{B.5}$$

More generally, given a likelihood and a prior, the posterior is

$$
\begin{aligned}
p(\eta^*|x_{1:n},\lambda) &\propto p(\eta^*|\lambda)\prod_{i=1}^{n} p(x_i|\eta^*) \\
&= h_b(\eta^*)\exp(\lambda_1^T\eta^* + \lambda_2(-a_l(\eta^*)) - a_b(\lambda)) \\
&\cdot (\prod_{i=1}^{n} h_l(x))\exp(\eta^{*T}\sum_{i=1}^{n} T(x_i) - n a_l(\eta^*)) \\
&\propto h_b(\eta^*)\exp\left(\left(\lambda_1 + \sum_{i=1}^{n} T(x_i)\right)^T \eta^* + (\lambda_2 + n)(-a_l(\eta^*))\right)
\end{aligned}
\tag{B.6}
$$

The parameters of the posterior are thus

$$
\begin{cases}
\tau_1 = \lambda_1 + \sum_{i=1}^{n} T(x_i) \\
\tau_2 = \lambda_2 + n
\end{cases}
$$

Beta - Conjugate prior of Binomial likelihood

$$
\begin{aligned}
p(\eta^*|\boldsymbol{\lambda}) &= h(\eta^*)\exp(\lambda_1^T\eta^* + \lambda_2(-a(\eta^*)) - a(\boldsymbol{\lambda})) \\
&= \exp\left(\lambda_1 \ln\frac{p}{1-p} + \lambda_2 n\ln(1-p) - a(\boldsymbol{\lambda})\right) \\
&= \left(\frac{p}{1-p}\right)_1^{\lambda}(1-p)^{n\lambda_2}e^{-a(\boldsymbol{\lambda})} \\
&= p^{(\lambda_1+1)-1}(1-p)^{(n\lambda_2-\lambda_1+1)-1}e^{-a(\boldsymbol{\lambda})} \\
&= \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}
\end{aligned}
\tag{B.7}
$$

We recognize a Beta distribution with parameters

$$
\begin{cases}
\lambda_1 = \alpha - 1 \\
\lambda_2 = \frac{\beta+\alpha-2}{n}
\end{cases}
\qquad\qquad
\begin{cases}
\alpha = \lambda_1 + 1 \\
\beta = n\lambda_2 - \lambda_1 + 1
\end{cases}
$$

The expectation of terms of the sufficient statistics for the posterior are given thereafter, where ψ is the digamma function.

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{\eta^*}] &= \frac{\partial a(\tau_1,\cdots)}{\partial\tau_1} \\
&= \frac{\partial}{\partial\tau_1}\left(\ln\Gamma(\tau_1+1) + \ln\Gamma(\beta) - \ln\Gamma(\tau_1+\beta+1)\right) \\
&= \psi(\tau_1+1) - \psi(\tau_1+\beta+1) \\
&= \psi(\alpha) - \psi(\alpha+\beta)
\end{aligned}
\tag{B.8}
$$

$$
\begin{aligned}
\mathbb{E}[-a(\boldsymbol{\eta}^*)] &= \frac{\partial a(\cdots, \tau_2)}{\partial \tau_2} \\
&= \frac{\partial}{\partial \tau_2} \left(\ln \Gamma(\alpha) + \ln \Gamma(n\tau_2 - \tau_1 + 1) - \ln \Gamma(\alpha + n\tau_2 - \tau_1 + 1) \right) \\
&= n\psi(n\tau_2 - \tau_1 + 1) - n\psi(\alpha + n\tau_2 - \tau_1 + 1) \\
&= n\psi(\beta) - n\psi(\alpha + \beta)
\end{aligned}
\tag{B.9}
$$

Dirichlet - Conjugate prior of Multinomial likelihood

$$
\begin{aligned}
p(\boldsymbol{\eta}^*|\boldsymbol{\lambda}) &= h(\boldsymbol{\eta}^*) \exp(\boldsymbol{\lambda_1}^T \boldsymbol{\eta}^* + \lambda_2(-a(\boldsymbol{\eta}^*)) - a(\boldsymbol{\lambda})) \\
&= \exp\left(\boldsymbol{\lambda_1}^T \begin{pmatrix} \ln p_1 \\ \vdots \\ \ln p_m \end{pmatrix} - a(\boldsymbol{\lambda}) \right) \\
&= \prod_{i=1}^{m} p_i^{(\lambda_{1i}+1)-1} e^{-a(\boldsymbol{\lambda})} \\
&= \frac{1}{B(\alpha)} \prod_{i=1}^{m} p_i^{\alpha_i - 1}
\end{aligned}
\tag{B.10}
$$

We recognize a Dirichlet distribution with parameters

$$
\begin{cases} \boldsymbol{\lambda_1} = \begin{pmatrix} \lambda_{11} \\ \vdots \\ \lambda_{1m} \end{pmatrix} = \begin{pmatrix} \alpha_1 - 1 \\ \vdots \\ \alpha_m - 1 \end{pmatrix} \\ \lambda_2 = 0 \end{cases} \qquad \begin{cases} \boldsymbol{\alpha} = \begin{pmatrix} \lambda_{11} + 1 \\ \vdots \\ \lambda_{1m} + 1 \end{pmatrix} \end{cases}
$$

The expectations for the posterior are

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{\eta}^*] &= \frac{\partial a(\boldsymbol{\tau_1}, \cdots)}{\partial \boldsymbol{\tau_1}} \\
&= \frac{\partial}{\partial \boldsymbol{\tau_1}} \left(\sum_{i=1}^{m} \ln \Gamma(\tau_{1i} + 1) - \ln \Gamma \left(\sum_{i=1}^{m} (\tau_{1i} + 1) \right) \right) \\
&= \begin{pmatrix} \psi(\tau_{11} + 1) - \psi\left(\sum_{i=1}^{m} \tau_{1i} + 1\right) \\ \vdots \\ \psi(\tau_{1m} + 1) - \psi\left(\sum_{i=1}^{m} \tau_{1i} + 1\right) \end{pmatrix} \\
&= \begin{pmatrix} \psi(\alpha_1) - \psi\left(\sum_{i=1}^{m} \alpha_i\right) \\ \vdots \\ \psi(\alpha_m) - \psi\left(\sum_{i=1}^{m} \alpha_i\right) \end{pmatrix}
\end{aligned} \tag{B.11}
$$

$$
\mathbb{E}[-a(\boldsymbol{\eta}^*)] = 0 \tag{B.12}
$$

Gamma - Conjugate prior of Poisson likelihood

$$
\begin{aligned}
p(\eta^*|\lambda) &= h(\eta^*) \exp(\lambda_1 \eta^* + \lambda_2(-a(\eta^*)) - a(\lambda)) \\
&= \exp(\lambda_1 \ln \lambda_0 - \lambda_2 \lambda_0 - \ln \Gamma(\alpha) + \alpha \ln(\beta)) \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_0^{\lambda_1} e^{-\lambda_2 \lambda_0} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_0^{\alpha - 1} e^{-\beta \lambda_0}
\end{aligned} \tag{B.13}
$$

We recognize a Gamma where λ_0 is the parameter of the Poisson distribution.

$$
\begin{cases} \lambda_1 = \alpha - 1 \\ \lambda_2 = \beta \end{cases} \qquad \begin{cases} \alpha = \lambda_1 + 1 \\ \beta = \lambda_2 \end{cases}
$$

The expectations are

$$
\begin{aligned}
\mathbb{E}[\eta^*] &= \frac{\partial a(\tau_1, \cdots)}{\partial \tau_1} \\
&= \frac{\partial}{\partial \tau_1} (\ln \Gamma(\tau_1 + 1) - (\tau_1 + 1) \ln \beta) \\
&= \psi(\tau_1 + 1) - \ln \beta \\
&= \psi(\alpha) - \ln \beta
\end{aligned} \tag{B.14}
$$

$$
\begin{aligned}
\mathbb{E}[-a(\eta^*)] &= \frac{\partial a(\cdots, \tau_2)}{\partial \tau_2} \\
&= \frac{\partial}{\partial \tau_2}(\ln \Gamma(\alpha) - \alpha \ln(\tau_2)) \\
&= \frac{\alpha}{\tau_2} \\
&= \frac{\alpha}{\beta}
\end{aligned}
\tag{B.15}
$$

Normal-Wishart - Conjugate prior of Normal likelihood

$$
\begin{aligned}
p(\boldsymbol{\eta}^*|\boldsymbol{\lambda}) &= h(\boldsymbol{\eta}^*) \exp(\boldsymbol{\lambda_1}^T \boldsymbol{\eta}^* + \lambda_2(-a(\boldsymbol{\eta}^*)) - a(\boldsymbol{\lambda})) \\
&= (2\pi)^{-\frac{d}{2}} \exp\left(\boldsymbol{\lambda_1}^T \begin{pmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{pmatrix} + \lambda_2 \left(-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\ln|\boldsymbol{\Sigma}| \right) - a(\boldsymbol{\lambda}) \right) \\
&= (2\pi)^{-\frac{d}{2}} \exp\left(\boldsymbol{\lambda_{11}}^T\boldsymbol{\Lambda}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\lambda_{12}}^T\boldsymbol{\Lambda} - \frac{1}{2}\lambda_2\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu} - \frac{1}{2}\lambda_2 \ln|\boldsymbol{\Lambda}^{-1}| - a(\boldsymbol{\lambda}) \right) \\
&= (2\pi)^{-\frac{d}{2}} \exp\Big((\boldsymbol{\mu}_0\lambda_0)^T\boldsymbol{\Lambda}\boldsymbol{\mu} - \frac{1}{2}(\boldsymbol{\mu}_0\boldsymbol{\mu}_0^T\lambda_0 + \boldsymbol{V}^{-1})\boldsymbol{\Lambda} - \frac{1}{2}\lambda_0\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu} \\
&+ \frac{1}{2}((n-d-1)+1)\ln|\boldsymbol{\Lambda}| - a(\boldsymbol{\lambda}) \Big) \\
&= (2\pi)^{-\frac{d}{2}} \exp\Big(\operatorname{tr}(\lambda_0\boldsymbol{\Lambda}\boldsymbol{\mu}\boldsymbol{\mu}_0^T) - \frac{1}{2}\boldsymbol{\mu}_0^T\lambda_0\boldsymbol{\Lambda}\boldsymbol{\mu}_0 - \frac{1}{2}\operatorname{tr}(\lambda_0\boldsymbol{\Lambda}\boldsymbol{\mu}\boldsymbol{\mu}^T) - \frac{1}{2}\operatorname{tr}(\boldsymbol{V}^{-1}\boldsymbol{\Lambda}) \\
&+ \frac{n-d-1}{2}\ln|\boldsymbol{\Lambda}| + \frac{1}{2}\ln|\boldsymbol{\Lambda}| - \left(-\frac{d}{2}\ln\lambda_0 + \frac{nd}{2}\ln 2 + \frac{n}{2}\ln|\boldsymbol{V}| + \ln\Gamma_d(\frac{n}{2}) \right) \Big) \\
&= \frac{|\boldsymbol{\Lambda}|^{\frac{n-d-1}{2}} e^{\frac{-\operatorname{tr}(\boldsymbol{V}^{-1}\boldsymbol{\Lambda})}{2}}}{2^{\frac{nd}{2}}|\boldsymbol{V}|^{\frac{n}{2}}\Gamma_d(\frac{n}{2})} (2\pi)^{-\frac{d}{2}} |\lambda_0\boldsymbol{\Lambda}|^{\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\lambda_0\boldsymbol{\Lambda}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)}
\end{aligned}
\tag{B.16}
$$

We recognize a $NW(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{V}, n)$ distribution with the following parameters. The previous derivation used transformations such as $|\lambda A| = \lambda^d|A|$ or $|A^{-1}| = |A|^{-1}$ and assume the vectorization of the matrices.

$$
\begin{cases}
\boldsymbol{\lambda_{11}} = \boldsymbol{\mu_0}\lambda_0 \\
\boldsymbol{\lambda_{12}} = (\boldsymbol{\mu_0}\boldsymbol{\mu_0}^T\lambda_0 + \boldsymbol{V}^{-1})^T \\
\lambda_2 = \lambda_0 \\
\lambda_2 = n - d
\end{cases}
\qquad
\begin{cases}
\boldsymbol{\mu_0} = \frac{\boldsymbol{\lambda_{11}}}{\lambda_2} \\
\lambda_0 = \lambda_2 \\
\boldsymbol{V} = \left(\boldsymbol{\lambda_{12}} - \frac{\boldsymbol{\lambda_{11}}\boldsymbol{\lambda_{11}}^T}{\lambda_2} \right)^{-T} \\
n = \lambda_2 + d
\end{cases}
$$

Which implies the constraint $\lambda_0 = n - d$. We used the notation $\boldsymbol{B}^{-T} = (\boldsymbol{B}^{-1})^T$.

The expectations for the posterior are given below, where $\mathbb{E}[\boldsymbol{\eta}^*]$ contains a vector and a matrix and $\mathbb{E}[-a(\boldsymbol{\eta}^*)]$ is a scalar. $\boldsymbol{\tau}$ is the natural parameter of the posterior, corresponding to $\boldsymbol{\lambda}$ for the prior. We also used the previous inverse parameter mapping.

$$\mathbb{E}[\boldsymbol{\eta}^*] = \begin{pmatrix} \frac{\partial a(\boldsymbol{\tau_{11}}, \cdots)}{\partial \boldsymbol{\tau_{11}}} \\ \frac{\partial a(\cdots, \boldsymbol{\tau_{12}}, \cdots)}{\partial \boldsymbol{\tau_{12}}} \end{pmatrix} \tag{B.17}$$

$$\begin{aligned} \frac{\partial a(\boldsymbol{\tau_{11}}, \cdots)}{\partial \boldsymbol{\tau_{11}}} &= \frac{\partial}{\partial \boldsymbol{\tau_{11}}} \Bigg(-\frac{d}{2}\ln\tau_2 + \frac{(\tau_2+d)d}{2}\ln 2 + \frac{\tau_2+d}{2}\ln\left|\left(\boldsymbol{\tau_{12}} - \frac{\boldsymbol{\tau_{11}}\boldsymbol{\tau_{11}}^T}{\tau_2}\right)^{-T}\right| \\ &\quad + \ln\Gamma_d\left(\frac{\tau_2+d}{2}\right)\Bigg) \\ &= \frac{\partial}{\partial \boldsymbol{\tau_{11}}}\left(-\frac{\tau_2+d}{2}\ln\left|\boldsymbol{\tau_{12}} - \frac{\boldsymbol{\tau_{11}}\boldsymbol{\tau_{11}}^T}{\tau_2}\right|\right) \\ &= \frac{\partial}{\partial \boldsymbol{\tau_{11}}}\left(-\frac{\tau_2+d}{2}\ln\left(\left(1 - \frac{\boldsymbol{\tau_{11}}^T\boldsymbol{\tau_{12}}^{-1}\boldsymbol{\tau_{11}}}{\tau_2}\right)|\boldsymbol{\tau_{12}}|\right)\right) \\ &= \frac{\partial}{\partial \boldsymbol{\tau_{11}}}\left(-\frac{\tau_2+d}{2}\ln(\tau_2 - \boldsymbol{\tau_{11}}^T\boldsymbol{\tau_{12}}^{-1}\boldsymbol{\tau_{11}})\right) \\ &= \frac{(\tau_2+d)(\boldsymbol{\tau_{12}}^{-1} + \boldsymbol{\tau_{12}}^{-T})\boldsymbol{\tau_{11}}}{2\tau_2 - 2\boldsymbol{\tau_{11}}^T\boldsymbol{\tau_{12}}^{-1}\boldsymbol{\tau_{11}}} \end{aligned} \tag{B.18}$$

Where we used $|\boldsymbol{B} - \boldsymbol{x}\boldsymbol{x}^T| = (1 - \boldsymbol{x}^T\boldsymbol{B}^{-1}\boldsymbol{x})|\boldsymbol{B}|$ and $\frac{\partial \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{B} + \boldsymbol{B}^T)\boldsymbol{x}$.

$$\begin{aligned} \frac{\partial a(\cdots, \boldsymbol{\tau_{12}}, \cdots)}{\partial \boldsymbol{\tau_{12}}} &= \frac{\partial}{\partial \boldsymbol{\tau_{12}}} \Bigg(-\frac{d}{2}\ln\tau_2 + \frac{(\tau_2+d)d}{2}\ln 2 + \frac{\tau_2+d}{2}\ln\left|\left(\boldsymbol{\tau_{12}} - \frac{\boldsymbol{\tau_{11}}\boldsymbol{\tau_{11}}^T}{\tau_2}\right)^{-T}\right| \\ &\quad + \ln\Gamma_d\left(\frac{\tau_2+d}{2}\right)\Bigg) \\ &= \frac{\partial}{\partial \boldsymbol{\tau_{12}}}\left(-\frac{\tau_2+d}{2}\ln\left|\boldsymbol{\tau_{12}} - \frac{\boldsymbol{\tau_{11}}\boldsymbol{\tau_{11}}^T}{\tau_2}\right|\right) \\ &= \frac{\partial}{\partial \boldsymbol{\tau_{12}}}\left(-\frac{\tau_2+d}{2}\left(\ln\left(\tau_2 - \boldsymbol{\tau_{11}}^T\boldsymbol{\tau_{12}}^{-1}\boldsymbol{\tau_{11}}\right) + \ln|\boldsymbol{\tau_{12}}|\right)\right) \\ &= -\frac{\tau_2+d}{2}\left(\frac{\boldsymbol{\tau_{12}}^{-T}\boldsymbol{\tau_{11}}\boldsymbol{\tau_{11}}^T\boldsymbol{\tau_{12}}^{-T}}{\tau_2 - \boldsymbol{\tau_{11}}^T\boldsymbol{\tau_{12}}^{-1}\boldsymbol{\tau_{11}}} + \boldsymbol{\tau_{12}}^{-T}\right) \end{aligned} \tag{B.19}$$

Using $\frac{\partial a^T \boldsymbol{X}^{-1} b}{\partial \boldsymbol{X}} = -\boldsymbol{X}^{-T} a b^T \boldsymbol{X}^{-T}$ and $\frac{\partial |\boldsymbol{X}|}{\partial \boldsymbol{X}} = |\boldsymbol{X}| \boldsymbol{X}^{-T}$.

$$
\begin{aligned}
\mathbb{E}[-a(\boldsymbol{\eta}^*)] &= \frac{\partial a(\cdots, \tau_2)}{\partial \tau_2} \\
&= \frac{\partial}{\partial \tau_2} \Bigg(-\frac{d}{2} \ln \tau_2 + \frac{(\tau_2 + d)d}{2} \ln 2 + \frac{\tau_2 + d}{2} \ln \left| \left(\boldsymbol{\tau_{12}} - \frac{\boldsymbol{\tau_{11}} \boldsymbol{\tau_{11}}^T}{\tau_2} \right)^{-T} \right| \\
&+ \ln \Gamma_d \left(\frac{\tau_2 + d}{2} \right) \Bigg) \\
&= -\frac{d}{2\tau_2} + \frac{d}{2} \ln 2 + \frac{\partial}{\partial \tau_2} \Bigg(-\frac{\tau_2 + d}{2} \left(\ln \left(\frac{\tau_2 - \boldsymbol{\tau_{11}}^T \boldsymbol{\tau_{12}}^{-1} \boldsymbol{\tau_{11}}}{\tau_2} \right) + \ln |\boldsymbol{\tau_{12}}| \right) \\
&+ \left(\frac{d(d-1)}{4} \ln \pi + \sum_{i=1}^{d} \ln \Gamma \left(\frac{\tau_2 + d}{2} + \frac{1-i}{2} \right) \right) \Bigg) \\
&= -\frac{d}{2\tau_2} + \frac{d}{2} \ln 2 + \frac{1}{2} \sum_{i=1}^{d} \psi \left(\frac{\tau_2 + d + 1 - i}{2} \right) + \frac{\partial}{\partial \tau_2} \Bigg(-\frac{\tau_2}{2} \ln \left(\tau_2 - \boldsymbol{\tau_{11}}^T \boldsymbol{\tau_{12}}^{-1} \boldsymbol{\tau_{11}} \right) \\
&+ \frac{\tau_2}{2} \ln \tau_2 - \frac{\tau_2}{2} \ln |\boldsymbol{\tau_{12}}| - \frac{d}{2} \ln \left(\tau_2 - \boldsymbol{\tau_{11}}^T \boldsymbol{\tau_{12}}^{-1} \boldsymbol{\tau_{11}} \right) + \frac{d}{2} \ln \tau_2 \Bigg) \\
&= \frac{d}{2} \ln 2 + \frac{1}{2} \sum_{i=1}^{d} \psi \left(\frac{\tau_2 + d + 1 - i}{2} \right) - \frac{1}{2} \ln (\tau_2 - \boldsymbol{\tau_{11}}^T \boldsymbol{\tau_{12}}^{-1} \boldsymbol{\tau_{11}}) \\
&- \frac{\tau_2}{2\tau_2 - 2\boldsymbol{\tau_{11}}^T \boldsymbol{\tau_{12}}^{-1} \boldsymbol{\tau_{11}}} - \frac{1}{2} \ln |\boldsymbol{\tau_{12}}| + \frac{1}{2} \ln \tau_2 + \frac{1}{2} - \frac{d}{2\tau_2 - 2\boldsymbol{\tau_{11}}^T \boldsymbol{\tau_{12}}^{-1} \boldsymbol{\tau_{11}}} \\
&= \frac{1}{2} \Bigg(1 - \frac{d + \tau_2}{\tau_2 - \boldsymbol{\tau_{11}}^T \boldsymbol{\tau_{12}}^{-1} \boldsymbol{\tau_{11}}} + d \ln 2 - \ln |\boldsymbol{\tau_{12}}| + \ln \tau_2 - \ln (\tau_2 - \boldsymbol{\tau_{11}}^T \boldsymbol{\tau_{12}}^{-1} \boldsymbol{\tau_{11}}) \\
&+ \sum_{i=1}^{d} \psi \left(\frac{\tau_2 + d + 1 - i}{2} \right) \Bigg)
\end{aligned}
\tag{B.20}
$$

Conjugate prior of Gamma likelihood

$$
\begin{aligned}
p(\boldsymbol{\eta}^*|\boldsymbol{\lambda}) &= h(\boldsymbol{\eta}^*)\exp(\boldsymbol{\lambda_1}^T\boldsymbol{\eta}^* + \lambda_2(-a(\boldsymbol{\eta}^*)) - a(\boldsymbol{\lambda})) \\
&= \exp\left(\boldsymbol{\lambda_1}^T\begin{pmatrix}\alpha-1\\-\beta\end{pmatrix} + \lambda_2(-\ln\Gamma(\alpha)+\alpha\ln\Gamma(\beta)) - a(\boldsymbol{\lambda})\right) \\
&= \exp\left(\lambda_{11}(\alpha-1) - \lambda_{12}\beta - \lambda_2\ln\Gamma(\alpha) + \lambda_2\alpha\ln\beta - a(\boldsymbol{\lambda})\right) \\
&= \frac{(\ln\lambda_{11})^{\alpha-1}e^{-\lambda_{12}\beta}}{\Gamma(\alpha)^{\lambda_2}\beta^{\alpha\lambda_2}}e^{-a(\boldsymbol{\lambda})} \\
&\propto \frac{p^{\alpha-1}e^{-\beta q}}{\Gamma(\alpha)^r\beta^{-\alpha s}}
\end{aligned}
\tag{B.21}
$$

We recognize the corresponding conjugate prior with the following parameters, where $p,q,r,s>0$ and $f(\alpha,\beta|p,q,r,s)\propto\frac{p^{\alpha-1}e^{-\beta q}}{\Gamma(\alpha)^r\beta^{-\alpha s}}$ if $\alpha,\beta>0$, 0 otherwise.

$$
\begin{cases}\lambda_{11}=e^p\\ \lambda_{12}=q\\ \lambda_2=r\\ \lambda_2=-s\end{cases}
\qquad
\begin{cases}p=\ln\lambda_{11}\\ q=\lambda_{12}\\ r=\lambda_2\\ s=-\lambda_2\end{cases}
$$

Which implies the constraint $r=-s$.

The expectation of the sufficient statistic terms cannot be computed for the corresponding posterior since we don't have the analytical form of the normalization factor.

Conjugate prior of Beta likelihood

$$
\begin{aligned}
p(\boldsymbol{\eta}^*|\boldsymbol{\lambda}) &= h(\boldsymbol{\eta}^*)\exp(\boldsymbol{\lambda_1}^T\boldsymbol{\eta}^* + \lambda_2(-a(\boldsymbol{\eta}^*)) - a(\boldsymbol{\lambda})) \\
&= \exp\left(\boldsymbol{\lambda_1}^T\begin{pmatrix}\alpha-1\\\beta-1\end{pmatrix} + \lambda_2(-\ln\Gamma(\alpha)-\ln\Gamma(\beta)+\ln\Gamma(\alpha+\beta)) - a(\boldsymbol{\lambda})\right) \\
&= \exp\left(\lambda_{11}(\alpha-1) + \lambda_{12}(\beta-1) + \lambda_2\ln\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} - a(\boldsymbol{\lambda})\right) \\
&= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^{\lambda_2}(\ln\lambda_{11})^{\alpha-1}(\ln\lambda_{12})^{\beta-1}e^{-a(\boldsymbol{\lambda})} \\
&= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^{\lambda_2}(\ln\lambda_{11})^{\alpha}(\ln\lambda_{12})^{\beta}e^{-a(\boldsymbol{\lambda})-\lambda_{11}-\lambda_{12}} \\
&\propto \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^{\lambda_0}x_0^{\alpha}y_0^{\beta}
\end{aligned}
\tag{B.22}
$$

We recognize the corresponding conjugate prior $\pi(\alpha, \beta|\lambda_0, x_0, y_0)$ with the following parameters

$$\begin{cases} \lambda_{11} = e^{x_0} \\ \lambda_{12} = e^{y_0} \\ \lambda_2 = \lambda_0 \end{cases} \qquad \begin{cases} \lambda_0 = \lambda_2 \\ x_0 = \ln \lambda_{11} \\ y_0 = \ln \lambda_{12} \end{cases}$$

As previously, the expectations for the posterior cannot be computed due to the missing analytical form of the normalization factor.

References

[1] Christopher M Bishop. Pattern recognition. *Machine Learning*, 128:461–486, 2006.

[2] David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–144, 2006.

[3] Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

[4] Stamatios Lefkimmiatis, Petros Maragos, and George Papandreou. Bayesian inference on multiscale models for poisson intensity estimation: Applications to photon-limited image denoising. *IEEE Transactions on Image Processing*, 18(8):1724–1741, 2009.