



HAL
open science

Predictive distributions of random variables following a multivariate Gaussian distribution with Normal Inverse Wishart prior

Etienne Goffinet, Mustapha Lebbah, Hanane Azzag, Loïc Giraldi

► **To cite this version:**

Etienne Goffinet, Mustapha Lebbah, Hanane Azzag, Loïc Giraldi. Predictive distributions of random variables following a multivariate Gaussian distribution with Normal Inverse Wishart prior. LIPN. 2021. hal-03962723

HAL Id: hal-03962723

<https://hal.science/hal-03962723v1>

Submitted on 30 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

PREDICTIVE DISTRIBUTIONS OF RANDOM VARIABLES FOLLOWING A MULTIVARIATE GAUSSIAN DISTRIBUTION WITH NORMAL INVERSE WISHART PRIOR

Etienne Goffinet, Mustapha Lebbah, Hanane Azzag

Sorbonne Paris Nord University
99 Av. Jean Baptiste Clément,
93430 Villetaneuse
etienne.goffinet[at]gmail.com

Loïc Giraldi

CEA Cadarache, DES, IRESNE, DEC
13108 Saint-Paul-Lez-Durance, France

ABSTRACT

The Gaussian distribution is a popular choice of density for the modeling of continuous observations, often combined with a specific prior in the context of Bayesian inference. When this prior is conjugate (e.g., the normal-gamma inverse distribution in the univariate case), it can be proved that the posterior and predictive distributions have analytically tractable expressions. While these derivations are readily available in the literature for the univariate case, the multivariate case derivations are significantly harder to find. This short technical note proposes a detailed derivation of these closed-forms in the multivariate case with a conjugate Normal Inverse Wishart prior.

Keywords Bayesian Inference · Predictive Distributions · Multivariate Gaussian · Normal Inverse Wishart

The Gaussian distribution is a popular choice of density for the modeling of continuous observations, often combined with a specific prior in the context of Bayesian inference. When this prior is conjugate (e.g., the normal-gamma inverse distribution in the univariate case), it can be proved that the posterior and predictive distributions have analytically tractable expressions. The most cited source [1] (e.g., cited in [2, 3]) proves the equivalence in the univariate case.

However, the multivariate case derivations are significantly harder to find. In the following we propose a detailed derivation of the multivariate distribution, based on the exponential family expressions of the Gaussian and Normal Inverse Wishart distributions.

Lemma 1. *The probability density function of the multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ can be written as*

$$p(y \mid \eta) = h(y) \exp(\eta^T T(y) - \mathbb{1}^T g(\eta)),$$

Proof. First, observe that $y^T \Sigma^{-1} y = \text{Tr}(\Sigma^{-1} y y^T) = \text{vec}(\Sigma^{-1})^T \text{vec}(y y^T)$. As a consequence, using the usual definition of the density function yields

$$\begin{aligned} p(y \mid \mu, \Sigma) &= (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y - \mu)^T \Sigma^{-1} (y - \mu)\right) \\ &= (2\pi)^{-d/2} \exp\left(\left[\begin{array}{c} \Sigma^{-1} \mu \\ -\frac{1}{2} \text{vec}(\Sigma^{-1}) \end{array}\right]^T \left[\begin{array}{c} y \\ \text{vec}(y y^T) \end{array}\right] - \frac{1}{2} \mu^T \Sigma^{-1} \mu - \frac{1}{2} \log |\Sigma|\right) \\ &= h(y) \exp(\eta^T T(y) - \mathbb{1}^T g(\eta)) \end{aligned}$$

where

$$h(y) = (2\pi)^{-\frac{d}{2}}, \eta = \left[\begin{array}{c} \Sigma^{-1} \mu \\ -\frac{1}{2} \text{vec}(\Sigma^{-1}) \end{array}\right], g(\eta) = \left[\begin{array}{c} \frac{1}{2} \mu^T \Sigma^{-1} \mu \\ \frac{1}{2} \log |\Sigma| \end{array}\right], T(y) = \left[\begin{array}{c} y \\ \text{vec}(y y^T) \end{array}\right],$$

and vec is the vectorization operator. □

Lemma 2. The probability density function of the Normal Inverse Wishart distribution $NIW(\mu, \Sigma \mid \xi)$, with $\xi = (\mu_0, \kappa_0, \Sigma_0, \nu_0)$ can be written as

$$p(\mu, \Sigma \mid \Psi, \nu) = f(\Psi, \nu) \exp(\eta^T \Psi - \nu^T g(\eta)).$$

Proof.

$$\begin{aligned} p(\mu, \Sigma \mid \xi) &= \frac{\kappa_0^{d/2} |\Psi_0|^{\frac{\nu_0}{2}} |\Sigma|^{-\frac{\nu_0+d+2}{2}}}{2^{\frac{\nu_0 d}{2}} \pi^{d/2} \Gamma_d(\frac{\nu_0}{2})} \exp\left(-\frac{\kappa_0}{2} (\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0) - \frac{1}{2} \text{Tr}(\Psi_0 \Sigma^{-1})\right) \\ &= \frac{\kappa_0^{d/2} |\Psi_0|^{\frac{\nu_0}{2}}}{2^{\frac{\nu_0 d}{2}} \pi^{d/2} \Gamma_d(\frac{\nu_0}{2})} \exp\left(\eta^T \left[\begin{array}{c} \kappa_0 \mu_0 \\ \text{vec}(\kappa_0 \mu_0 \mu_0^T + \Psi_0) \end{array} \right] - \left[\begin{array}{c} \kappa_0 \\ \nu_0 + d + 2 \end{array} \right]^T g(\eta)\right) \\ &= f(\Psi, \nu) \exp(\eta^T \Psi - \nu^T g(\eta)) \end{aligned}$$

with $f(\Psi, \nu) = \frac{\kappa_0^{d/2} |\Psi_0|^{\frac{\nu_0}{2}}}{2^{\frac{\nu_0 d}{2}} \pi^{d/2} \Gamma_d(\frac{\nu_0}{2})}$, $\Psi = \left[\begin{array}{c} \kappa_0 \mu_0 \\ \text{vec}(\kappa_0 \mu_0 \mu_0^T + \Psi_0) \end{array} \right]$ and $\nu = \left[\begin{array}{c} \kappa_0 \\ \nu_0 + d + 2 \end{array} \right]$. □

Theorem 1. The prior predictive distribution of a random variable y drawn according to

$$y \sim \mathcal{N}(\mu, \Sigma), \quad (\mu, \Sigma) \sim NIW(\xi), \quad \xi = (\mu_0, \kappa_0, \Sigma_0, \nu_0),$$

is the multivariate t -distribution

$$y \mid \xi \sim t_{\nu_0-d+1} \left(y \mid \mu_0, \frac{(\kappa_0 + 1) \Psi_0}{\kappa_0(\nu_0 - d + 1)} \right).$$

Proof. According to Lemmas 1 and 2, we find that

$$\begin{aligned} p(y \mid \Psi, \nu) &= \int_{\mu, \Sigma} p(y \mid \mu, \Sigma) p(\mu, \Sigma \mid \Psi, \nu) d(\mu, \Sigma) \\ &= \int_{\eta} h(y) \exp(\eta^T T(y) - \mathbf{1}^T g(\eta)) f(\Psi, \xi) \exp(\eta^T \Psi - \nu^T g(\eta)) d\eta \\ &= \int_{\eta} h(y) \exp(\eta^T T(y) - \mathbf{1}^T g(\eta)) f(\Psi, \xi) \exp(\eta^T \Psi - \nu^T g(\eta)) d\eta \\ &= h(y) f(\Psi, \nu) \int_{\eta} \exp(\eta^T (\Psi + T(y)) - (\nu + \mathbf{1})^T g(\eta)) d\eta. \end{aligned}$$

The integral part is obtained by considering the Normal Inverse Wishart density with hyper-parameter $(\Psi + T(y), \nu + \mathbf{1})$:

$$\begin{aligned} p(\mu, \Sigma \mid \Psi + T(y), \nu + \mathbf{1}) &= f(\Psi + T(y), \nu + \mathbf{1}) \exp(\eta^T (\Psi + T(y)) - (\nu + \mathbf{1})^T g(\eta)) \\ \Rightarrow \int f(\Psi + T(y), \nu + \mathbf{1}) \exp(\eta^T (\Psi + T(y)) - (\nu + \mathbf{1})^T g(\eta)) d\eta &= 1 \\ \Rightarrow \int \exp(\eta^T (\Psi + T(y)) - (\nu + \mathbf{1})^T g(\eta)) d\eta &= \frac{1}{f(\Psi + T(y), \nu + \mathbf{1})} \end{aligned} \quad (1)$$

This last expression yields the following equality

$$p(y \mid \Psi, \nu) = h(y) \frac{f(\Psi, \nu)}{f(\Psi + T(y), \nu + \mathbf{1})},$$

where $\Psi + T(y) = \left[\begin{array}{c} \kappa_1 \mu_1 \\ \text{vec}(\kappa_1 \mu_1 \mu_1^T + \Psi_1) \end{array} \right] = \left[\begin{array}{c} \kappa_0 \mu_0 \\ \text{vec}(\kappa_0 \mu_0 \mu_0^T + \Psi_0) \end{array} \right] + \left[\begin{array}{c} y \\ \text{vec}(y y^T) \end{array} \right]$, and $\nu + \mathbf{1} = \left[\begin{array}{c} \kappa_1 \\ \nu_1 + d + 2 \end{array} \right] = \left[\begin{array}{c} \kappa_0 \\ \nu_0 + d + 2 \end{array} \right] + \left[\begin{array}{c} 1 \\ 1 \end{array} \right]$. Introducing these parameters into the last expression of $p(y \mid \xi)$ gives the result of the lemma.

$$p(y \mid \Psi, \nu) = \pi^{-d/2} \cdot \frac{\kappa_0^{d/2}}{\kappa_1^{d/2}} \cdot \frac{\Gamma_d(\nu_1/2)}{\Gamma_d(\nu_0/2)} \cdot \frac{|\Psi_0|^{\nu_0/2}}{|\Psi_1|^{\nu_1/2}}, \quad (2)$$

On one hand, $\Gamma_d(\nu_1/2)/\Gamma_d(\nu_0/2)$ simplifies using the definition of the multivariate Gamma function $\Gamma_d(a) = \pi^{d(d-1)/4} \prod_{j=1}^d \Gamma(a + (1-j)/2)$ yielding

$$\frac{\Gamma_d(\frac{\nu_1}{2})}{\Gamma_d(\frac{\nu_0}{2})} = \frac{\Gamma_d(\frac{\nu_0+1}{2})}{\Gamma_d(\frac{\nu_0}{2})} = \prod_{j=1}^d \frac{\Gamma(\frac{\nu_0+2-j}{2})}{\Gamma(\frac{\nu_0+1-j}{2})} = \frac{\Gamma(\frac{\nu_0+1}{2})}{\Gamma(\frac{\nu_0-d+1}{2})}.$$

On the other hand, $|\Psi_0|^{\nu_0/2}/|\Psi_1|^{\nu_1/2}$ reduces because

$$\begin{aligned} \Psi_1 &= \Psi_0 + yy^T + \kappa_0\mu_0\mu_0^T - \kappa_1\mu_1\mu_1^T \\ &= \Psi_0 + yy^T + \kappa_0\mu_0\mu_0^T - (\kappa_0 + 1) \left(\frac{\kappa_0\mu_0 + y}{\kappa_0 + 1} \right) \left(\frac{\kappa_0\mu_0 + y}{\kappa_0 + 1} \right)^T \\ &= \Psi_0 + \frac{\kappa_0 + 1}{\kappa_0 + 1} yy^T + \frac{\kappa_0(\kappa_0 + 1)}{\kappa_0 + 1} \mu_0\mu_0^T - \left(\frac{1}{\kappa_0 + 1} \right) (\kappa_0^2\mu_0\mu_0^T + \kappa_0\mu_0y^T + \kappa_0y\mu_0^T + yy^T) \\ &= \Psi_0 + \frac{\kappa_0}{\kappa_0 + 1} (y - \mu_0)(y - \mu_0)^T \\ &= \Psi_0 \left(1 + \frac{\kappa_0}{\kappa_0 + 1} (y - \mu_0)^T \Psi_0^{-1} (y - \mu_0) \right) \end{aligned}$$

and, using the matrix determinant lemma (i.e. $|A + uv^T| = (1 + v^T A^{-1}u)|A|$) gives

$$|\Psi_1| = \left(1 + \frac{\kappa_0}{\kappa_0 + 1} (y - \mu_0^T) \Psi_0^{-1} (y - \mu_0^T) \right) |\Psi_0|.$$

Substituting these expressions into Eq. (2) gives the density

$$p(y | \Psi, \nu) = \frac{\Gamma(\frac{\nu_0+1}{2})}{\left(\frac{\kappa_0+1}{\kappa_0}\right)^{d/2} \pi^{d/2} \Gamma(\frac{\nu_0-d+1}{2}) |\Psi_0|^{1/2}} \left(1 + \frac{\kappa_0}{\kappa_0 + 1} (y - \mu_0)^T \Psi_0^{-1} (y - \mu_0) \right)^{-(\nu_0+1)/2},$$

which is the density of a multivariate t -distribution $t_{\nu_0-d+1} \left(y | \mu_0, \frac{(\kappa_0+1)\Psi_0}{\kappa_0(\nu_0-d+1)} \right)$. \square

1 Posterior Predictive Distribution

With the same notation, and denoting $\mathbf{x} = (x_i)_n$ a set of observations, the posterior distribution follows a NIW distribution [4, 1]:

$$p(\mu, \Sigma | \mathbf{x}, \mu_0, \kappa_0, \Sigma_0, \nu_0) = NIW(\mu, \Sigma | \mu_n, \kappa_n, \Sigma_n, \nu_n),$$

with the updated hyper-parameter values obtained by:

$$\begin{aligned} \mu_n &= \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0 + n}, \\ \kappa_n &= \kappa_0 + n, \nu_n = \nu_0 + n, \\ \Psi_n &= \Psi_0 + \sum_i x_i x_i^T + \kappa_0\mu_0\mu_0^T - \kappa_n\mu_n\mu_n^T. \end{aligned}$$

With $\Psi' = \left[\text{vec}(\kappa_n\mu_n\mu_n^T + \Psi_n) \right]$ and $\nu' = \left[\nu_n + d + 2 \right]$, the posterior predictive distribution can be written as:

$$\begin{aligned} p(y | \Psi', \nu') &= \int_{\mu, \Sigma} p(y | \mu, \Sigma) p(\mu, \Sigma | \psi', \nu') d(\mu, \Sigma) \\ &= \int_{\eta} h(y) \exp(\eta^T T(y) - \mathbf{1}^T g(\eta)) f(\Psi', \nu') \exp(\eta^T \Psi' - \nu'^T g(\eta)) d\eta \\ &= h(y) f(\Psi', \nu') \int_{\eta} \exp[\eta^T (\Psi' + T(y)) - (\nu' + \mathbf{1})^T g(\eta)] d\eta. \end{aligned}$$

After integral simplification (c.f. Eq. 1), the posterior can be written as:

$$p(y | \mathbf{x}, \Psi, \nu) = h(y) \frac{f(\Psi', \nu')}{f(\Psi' + T(y), \nu' + \mathbb{1})}. \quad (3)$$

These expressions are similar to the ones of the prior distribution, with updated hyper-parameter values obtained from:

$$\begin{aligned} \Psi' + T(y) &= \begin{bmatrix} \kappa_1 \mu_1 \\ \text{vec}(\kappa_1 \mu_1 \mu_1^T + \Psi_1) \end{bmatrix} = \begin{bmatrix} \kappa_n \mu_n \\ \text{vec}(\kappa_n \mu_n \mu_n^T + \Psi_n) \end{bmatrix} + \begin{bmatrix} y \\ \text{vec}(y y^T) \end{bmatrix}, \text{ and} \\ \nu' + \mathbb{1} &= \begin{bmatrix} \kappa_1 \\ \nu_1 + d + 2 \end{bmatrix} = \begin{bmatrix} \kappa_n \\ \nu_n + d + 2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \end{aligned}$$

As in the prior predictive case, injecting the updated hyper-parameters values into Eq. 3 gives the following expression:

$$p(y | \mathbf{x}, \Psi, \nu) = \pi^{-d/2} \cdot \frac{\kappa_n^{d/2}}{\kappa_1^{d/2}} \cdot \frac{\Gamma_d(\nu_1/2)}{\Gamma_d(\nu_n/2)} \cdot \frac{|\Psi_n|^{\nu_n/2}}{|\Psi_1|^{\nu_1/2}},$$

which is the density of the multivariate t -distribution $t_{\nu_n-d+1} \left(y | \mu_n, \frac{(\kappa_n+1)\Psi_n}{\kappa_n(\nu_n-d+1)} \right)$

2 Joint Predictive Distribution

The previous sections presented the derivations of the predictive distributions of one multivariate observation y . The following section details the derivations of the predictive distribution of a set of multivariate observations $\mathbf{y} = (y_i)_m$, i.e., the joint predictive distributions. These distributions also have closed-forms when the prior is conjugate. In the joint prior predictive case, this expression is given by:

$$\begin{aligned} p(\mathbf{y} | \Psi, \nu) &= \int_{\mu, \Sigma} p(\mathbf{y}, \mu, \Sigma | \Psi, \nu) d(\mu, \Sigma) \\ &= \int_{\mu, \Sigma} p(\mathbf{y} | \mu, \Sigma) p(\mu, \Sigma | \psi, \nu) d(\mu, \Sigma) \\ &= \int_{\mu, \Sigma} \prod_i p(y_i | \mu, \Sigma) p(\mu, \Sigma | \psi, \nu) d(\mu, \Sigma) \\ &= \int_{\eta} \prod_i [h(y_i) \exp(\eta^T T(y_i) - \mathbb{1}g(\eta))] f(\Psi, \xi) \exp(\eta^T \Psi - \nu^T g(\eta)) d\eta \\ &= \prod_i h(y_i) f(\Psi, \xi) \int_{\eta} \exp[\eta^T (\Psi + T(\mathbf{y})) - (\nu + m\mathbb{1})^T g(\eta)] d\eta, \end{aligned}$$

with $T(\mathbf{y}) = \sum_i T(y_i)$. After simplifying the integral (c.f. Eq. 1), the posterior can be written as:

$$p(\mathbf{y} | \Psi, \nu) = \prod_i h(y_i) \frac{f(\Psi, \nu)}{f(\Psi + T(\mathbf{y}), \nu + m\mathbb{1})}, \quad (4)$$

with $\Psi + T(\mathbf{y}) = \begin{bmatrix} \kappa_m \mu_m \\ \text{vec}(\kappa_m \mu_m \mu_m^T + \Psi_m) \end{bmatrix} = \begin{bmatrix} \kappa_0 \mu_0 \\ \text{vec}(\kappa_0 \mu_0 \mu_0^T + \Psi_0) \end{bmatrix} + \begin{bmatrix} \sum_i y_i \\ \text{vec}(\sum_i y_i y_i^T) \end{bmatrix}$, and $\nu + m\mathbb{1} = \begin{bmatrix} \kappa_m \\ \nu_m + d + 2 \end{bmatrix} = \begin{bmatrix} \kappa_0 \\ \nu_0 + d + 2 \end{bmatrix} + \begin{bmatrix} m \\ m \end{bmatrix}$, which yields the same hyper-parameters update expressions than in the posterior predictive case detailed in the previous section Sect. 1.

Injecting the values of h , f and the prior hyper-parameters in Eq. 4 gives the closed-form of the joint prior predictive distribution:

$$p(\mathbf{y} | \Psi, \nu) = \pi^{-\frac{md}{2}} \frac{\kappa_0^{d/2}}{\kappa_m^{d/2}} \cdot \frac{\Gamma_d(\nu_m/2)}{\Gamma_d(\nu_0/2)} \cdot \frac{|\Psi_0|^{\nu_0/2}}{|\Psi_m|^{\nu_m/2}}.$$

The same derivations yields the following joint posterior predictive distribution expressions:

$$p(\mathbf{y} | \mathbf{x}, \Psi, \nu) = \prod_i h(y_i) \frac{f(\Psi', \nu')}{f(\Psi' + T(\mathbf{y}), \nu' + m\mathbb{1})},$$

Distribution	Expression	Closed-form
prior predictive	$p(y \Psi, \nu)$	$\pi^{-d/2} \cdot \frac{\kappa_0^{d/2}}{\kappa_1^{d/2}} \cdot \frac{\Gamma_d(\nu_1/2)}{\Gamma_d(\nu_0/2)} \cdot \frac{ \Psi_0 ^{\nu_0/2}}{ \Psi_1 ^{\nu_1/2}}$
posterior predictive	$p(y \mathbf{x}, \Psi, \nu)$	$\pi^{-d/2} \cdot \frac{\kappa_n^{d/2}}{\kappa_1^{d/2}} \cdot \frac{\Gamma_d(\nu_1/2)}{\Gamma_d(\nu_n/2)} \cdot \frac{ \Psi_n ^{\nu_n/2}}{ \Psi_1 ^{\nu_1/2}}$
joint prior predictive	$p(\mathbf{y} \Psi, \nu)$	$\pi^{\frac{-md}{2}} \cdot \frac{\kappa_0^{d/2}}{\kappa_m^{d/2}} \cdot \frac{\Gamma_d(\nu_m/2)}{\Gamma_d(\nu_0/2)} \cdot \frac{ \Psi_0 ^{\nu_0/2}}{ \Psi_m ^{\nu_m/2}}$
joint posterior predictive	$p(\mathbf{y} \mathbf{x}, \Psi, \nu)$	$\pi^{\frac{-md}{2}} \cdot \frac{\kappa_n^{d/2}}{\kappa_m^{d/2}} \cdot \frac{\Gamma_d(\nu_m/2)}{\Gamma_d(\nu_n/2)} \cdot \frac{ \Psi_n ^{\nu_n/2}}{ \Psi_m ^{\nu_m/2}}$

Table 1: Summary of predictive distributions closed forms.

that reduces to the following closed-form:

$$\pi^{\frac{-md}{2}} \cdot \frac{\kappa_n^{d/2}}{\kappa_m^{d/2}} \cdot \frac{\Gamma_d(\nu_m/2)}{\Gamma_d(\nu_n/2)} \cdot \frac{|\Psi_n|^{\nu_n/2}}{|\Psi_m|^{\nu_m/2}}.$$

The prior and posterior predictive distributions closed forms are summarized in Tab. 1.

References

- [1] Kevin P Murphy. Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2 σ):16, 2007.
- [2] William R Vega-Brown, Marek Doniec, and Nicholas G Roy. Nonparametric bayesian inference on multivariate exponential families. *Advances in Neural Information Processing Systems*, 27:2546–2554, 2014.
- [3] Rajarshi Das. Collapsed gibbs sampler for dirichlet process gaussian mixture models (dpgmm). *Technical report, Carnegie Mellon University, United States*, 2014.
- [4] Eric Benhamou, David Saltiel, Beatrice Guez, and Nicolas Paris. Bcma-es ii: revisiting bayesian cma-es. 2019.