# Estimation de la variance généralisée

Thu Pham-Gia

**ESTIMATION OF THE GENERALIZED VARIANCE IN MULTIVARIATE ANALYSIS**

by *Thu PHAM-GIA, (phamgit@umoncton.ca)*

*Département de math. et de stat., Université de Moncton, Canada E1A 3E9.*

**ABSTRACT**: Expressing the density of the determinant of a sample covariance matrix in terms of Meijer's G-function , we provide a confidence interval for the determinant if the population covariance matrix. The Bayesian approach is used to provide a set of bounds for the posterior generalized variance of the mean vector and other cases. Numerical applications of the results are also presented.

**RESUME:** Exprimant la densité du déterminant de la matrice de covariance d'échantillon sous forme de fonction de Meijer, nous dérivons un intervalle de confiance de la variance généralisée de la population. Ceci permet, à travers l'approche bayésienne, d'obtenir des bornes pour la variance généralisée postérieure du vecteur moyen. Des applications numériques sont aussi présentées.

**MOTS-CLES** : Statistique mathématique, Méthodes bayésiennes.

## Main results

We consider $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and a random sample $\{\mathbf{X}_1, ..., \mathbf{X}_N\}$ of $\mathbf{X}$. We define

$$\mathbf{A} = \sum_{\alpha=1}^{N} \left(\mathbf{X}_\alpha - \overline{\mathbf{X}}\right)\left(\mathbf{X}_\alpha - \overline{\mathbf{X}}\right)^T, \tag{1}$$

with $\overline{\mathbf{X}} = \sum_{\alpha=1}^{N} \mathbf{X}_\alpha / N$, the sample sum of squares and products matrix. The sample covariance matrix is: $\mathbf{S} = \mathbf{A}/n$. For $n > p$ we have $\mathbf{A} \sim W_p(n, \boldsymbol{\Sigma})$, where $W_p(n, \boldsymbol{\Sigma})$ is the central Wishart distribution in $R^p$, with $n$ degrees of freedom and covariance matrix $\boldsymbol{\Sigma}$, which is a multivariate generalization of $\chi_n^2$. Hence, $\mathbf{S} \sim W_p(n, \boldsymbol{\Sigma}/n)$. Also, it can be proven that :

$$|\mathbf{A}| / |\boldsymbol{\Sigma}| \cong \chi_n^2 ... \chi_{n-p+1}^2, \tag{2}$$

where the $\chi_j^2$ variables , with $j$ degrees of freedom , $n - p + 1 \le j \le n$ , are independent.

This is also the distribution of $n^p \dfrac{|\mathbf{S}|}{|\boldsymbol{\Sigma}|}$, with $n = N - 1$.

On the other hand, Meijer's function $\mathbf{G}(x)$ **is** defined as follows:

$\mathbf{G}^{\,m\,r}_{p\,q}\left[\,x\,\middle|\begin{matrix}a_1,\ldots,a_p\\ b_1,\ldots,b_p\end{matrix}\right]=\dfrac{1}{2\pi i}\displaystyle\int_L \frac{\prod_{j=1}^{m}\Gamma\left(b_j-s\right)\prod_{j=1}^{r}\Gamma\left(1-a_j+s\right)}{\prod_{j=m+1}^{q}\Gamma\left(1-b_j+s\right)\prod_{j=r+1}^{p}\Gamma\left(a_j-s\right)}x^s ds$, i.e. it is the integral along the complex contour $L$ of a ratio of products of gamma functions. G has a relationship with the generalized hypergeometric function: ${}_pF_{q-1}(.)$. Many common densities defined on (0, $\infty$), can be expressed as a G-function. For example, the gamma variable in two parameters, defined on (0, $\infty$), with density $f(x)=x^{\alpha-1}\exp(-x/\beta)/[\beta^{\alpha}\Gamma(\alpha)],\ 0\le x<\infty,\ \alpha,\beta>0$ is a **G**-function random variable since: $f(x)=\dfrac{1}{\beta\ \Gamma(\alpha)}\mathbf{G}^{\,1\ 0}_{0\ 1}\left[\dfrac{x}{\beta}\,\middle|\,(\alpha-1)\right],\ x>0$.

**THEOREM :** The ratio $Y=\dfrac{|\mathbf{A}|}{|\mathbf{\Sigma}|}=n^p\dfrac{|\mathbf{S}|}{|\mathbf{\Sigma}|}$, related to a random sample of size N = *n+1,* has its density given by:

$$h(y)=\frac{1}{2^p}\left(\prod_{j=0}^{p-1}\frac{1}{\Gamma(\frac{n-j}{2})}\right)\mathbf{G}^{\,p\ 0}_{0\ p}\left[\frac{y}{2^p}\,\middle|\,\frac{(n-2)}{2},\frac{(n-3)}{2},\ldots,\frac{(n-(p+1))}{2}\right],\ y\ge 0 \qquad (3)$$

Let $h_{\alpha/2}$ and $h_{1-(\alpha/2)}$ be the lower and upper $\alpha/2$ percentiles of this density, that can be numerically determined with precision, using MAPLE. We then have:

$h_{\alpha}\le\dfrac{n^p|S|}{|\Sigma|}\le h_{1-\alpha/2}\ \rightarrow\ \dfrac{n^p|\hat{S}|}{h_{1-\alpha}/2}\le|\Sigma|\le\dfrac{n^p|\hat{S}|}{h_{\alpha}}$, which provides the exact (1-$\alpha$)100% confidence interval for the population generalized variance.

a) Numerical example: Let us generate nine observations from a normal population (on Fisher's iris data), with covariance matrix $\mathbf{\Sigma}$ taken from one of our previous publications. In $R^4$ this distribution $N_4(\boldsymbol{\mu},\mathbf{\Sigma})$, has known parameters, $\boldsymbol{\mu}^T=(6.35,6.20,5.55.5.23)$ and

$\mathbf{\Sigma}=\begin{pmatrix}1.80 & 0.140 & 0.023 & 0.010\\ 0.140 & 0.900 & 0.070 & 0.030\\ 0.023 & 0.070 & 0.700 & 0.030\\ 0.010 & 0.030 & 0.030 & 0.600\end{pmatrix}$. Hence $|\mathbf{\Sigma}|=0.6645$.

The nine observations generated lead to the estimated sample **S**, with mean vector $\hat{\boldsymbol{\mu}}^T=\{6.485,6.479,5.102,5.014\}$ and the estimated sample covariance matrix :

$$\widehat{\mathbf{S}} = \begin{pmatrix} 2.465 & 0.694 & -0.162 & 0.017 \\ 0.994 & 0.578 & -0.097 & -0.004 \\ -0.162 & -0.097 & 0.857 & 0.506 \\ 0.017 & -0.004 & 0.506 & 0.521 \end{pmatrix}, \text{ with determinant value: } |\hat{\mathbf{S}}| = 0.113.$$

a) With *n* = *8*, the density of $8^4 \times |\mathbf{S}| / |\mathbf{\Sigma}|$, where $\mathbf{S}$ is a random sample covariance matrix from the above distribution is given by (3). The exact 95% confidence interval for $n^p |\mathbf{S}| / |\mathbf{\Sigma}|$, is found numerically to be (65.35, 7934) , as given by Fig.1. Hence, by (3) , the corresponding 95% confidence interval for $|\mathbf{\Sigma}|$, based on the observed sample is: (0.0583, 7.0826), since $8^4 \times |\hat{\mathbf{S}}| = 462.84$ , and contains the value of $|\mathbf{\Sigma}|$.

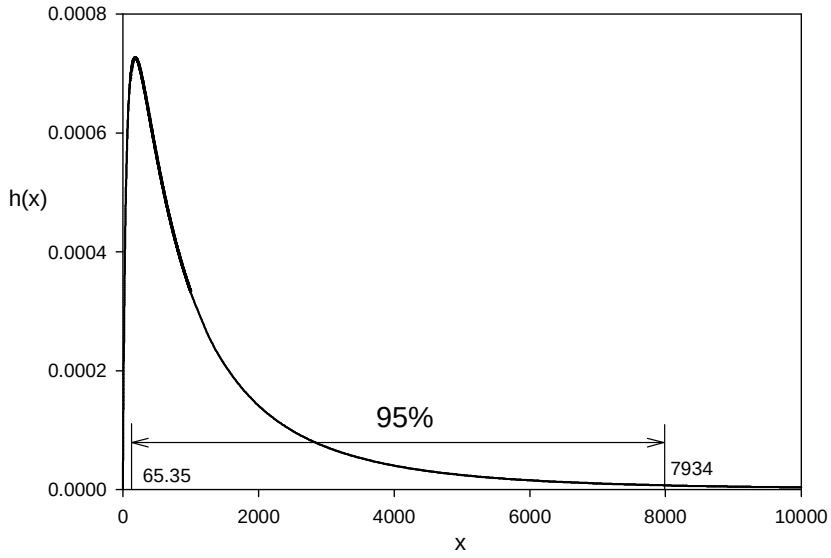


Fig. 1

Density of $n^p |\mathbf{S}| / |\mathbf{\Sigma}|, n = 8, p = 4$

i) Normal approximation : Anderson (1984, p.266) , based on the result that, asymptotically, $\sqrt{n}\left(\frac{|\mathbf{S}|}{|\mathbf{\Sigma}|} - 1\right) \sim N(0,\ 2p)$. This would lead to the approximate confidence interval :

$$|\mathbf{\Sigma}| \in \left(\frac{|\hat{\mathbf{S}}|}{1 + z_{\alpha/2}\sqrt{\frac{2p}{n}}},\ \frac{|\hat{\mathbf{S}}|}{1 - z_{\alpha/2}\sqrt{\frac{2p}{n}}}\right), \qquad (4)$$

where $P(Z \le z_{\alpha/2}) = \alpha/2$, and $\mathbf{S}$ is defined in the introduction.

ii) Other researchers: T. Kubokawa and Srivastava (2003) provided a detailed survey of the different approaches proposed to date to estimate $\boldsymbol{\Sigma}$ and its determinant.

Iliopoulos and Kourouklis (1998) also used the product of chi-square variables , but not G-functions.

Hao and Krishnamoorthy (2001) gave some results based on chi-square approximation and missing data approaches.

Using the very rich properties of the G-function, we can derive:

a) Moments of $Y$ : $E(|Y|^k) = 2^{pk}\prod_{j=1}^{p}\frac{\Gamma\left(\frac{n-(j-1)}{2}+k\right)}{\Gamma\left(\frac{n-(j-1)}{2}\right)}, k \geq 1$, from which the moments of $\mathbf{S}$, in function of $\boldsymbol{\Sigma}$, can be obtained.

b) Bayesian approach: Since the sample covariance matrix S is present in several results in Multivariate Bayesian analysis, we can derive the following result:

1) Let $\boldsymbol{\mu}$ be unknown in $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with a normal prior $N(\boldsymbol{\mu}_0, \boldsymbol{\tau}_0)$, where $\boldsymbol{\tau}_0$ being the matrix of precision. If $\boldsymbol{\Sigma}$ is known, we know that , with $\bar{\mathbf{x}}$ being the mean of the sample $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$, the posterior distribution of $\boldsymbol{\mu}$ is $N_p(\boldsymbol{\mu}^*, \boldsymbol{\tau}_0 + N\mathbf{R})$, where $\mathbf{R} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\mu}^* = \left(\boldsymbol{\tau}_0\boldsymbol{\mu}_0 + N\mathbf{R}\bar{\mathbf{x}}\right)(\boldsymbol{\tau}_0 + N\mathbf{R})^{-1}$. With the 95% confidence bounds for $\boldsymbol{\Sigma}$ above , we can derive the bounds for the difference between prior and posterior precisions (for the mean) is: $\left\{\frac{N}{l_u}, \frac{N}{l_l}\right\}$. Other similar results concern the cases:

2) Let $\boldsymbol{\Sigma}$ be unknown in $\mathbf{X} \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$, with a Wishart prior , $\boldsymbol{\Sigma} \sim Wi_p(\alpha, \boldsymbol{\tau})$, where $\boldsymbol{\tau}$ being the matrix of precision. We know that , with $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$ being the observed sample, the posterior distribution of $\boldsymbol{\Sigma}$ is Wishart, i.e.

$$\boldsymbol{\Sigma}\left|(\mathbf{x}_1, ..., \mathbf{x}_N) \sim Wi_p(\alpha^* = \alpha + n, \boldsymbol{\tau}^* = (\boldsymbol{\Sigma}^*)^{-1} = \boldsymbol{\tau} + \sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu}_0)(\mathbf{x}_i - \boldsymbol{\mu}_0)^T\right).$$

3) Both mean and precision matrix are unknown in $\mathbf{X} \sim N_p(\mathbf{M}, \mathbf{R})$. The prior of $(\mathbf{M}, \mathbf{R}) \sim NWi_p(\alpha, \boldsymbol{\tau}, \boldsymbol{\mu}, \nu)$ is the Normal-Wishart distribution where :

$$\mathbf{R} \sim Wi_p(\alpha, \boldsymbol{\tau}), \alpha \geq p, \ \mathbf{M}\left|\mathbf{R} = \mathbf{r} \sim N_p(\boldsymbol{\mu}, \nu\mathbf{r}), \ \nu\right.$$

The posterior distribution of $(\boldsymbol{\mu}, \mathbf{R}) \sim NWi_p(\alpha^*, \boldsymbol{\tau}^*, \boldsymbol{\mu}^*, \nu^*)$ , wher e

$$\alpha^* = \alpha + N,\ \nu^* = \nu + N,\ \boldsymbol{\mu}^* = \frac{\nu\boldsymbol{\mu} + N\bar{\mathbf{x}}}{\nu + N},$$

and $\boldsymbol{\tau}^* = \boldsymbol{\tau} + \mathbf{A} + \frac{\nu N}{\nu + N}(\boldsymbol{\mu} - \bar{\mathbf{x}})(\boldsymbol{\mu} - \bar{\mathbf{x}})^T$ .

Some bounds for $\boldsymbol{\tau}^* - \boldsymbol{\tau}$ can then be given.

**REFERENCES**

[1] Anderson, T.W., *Introduction to Multivariate Statistical Analysis*, 2nd Ed., John Wiley, New York, 1984.

[2] Iliopoulos, G. and Kourouklis, S. (1998), On Improved Interval Estimation for the Generalized Variance, *Journ. of Stat. Plann. and Inf*., 66, 305-420.

[3] Hao, J. and Krishnamoorthy, K.. (2001), Inferences on a Normal Covariance Matrix and Generalized Variance withMonotone Missing Data, *J. of Mult. Anal*.,78, 62-82.

[4] Kubokawa, T. and Srivastava, M.S. (2003), Estimating the Covariance Matrix: a New Approach, Discussion Paper CIRJE-F-52, the University of Toronto.

[5] Pham-Gia, T. and Turkkan, N., Density of the sample generalized Variance and applications, *Stat. Papers*, on line on the journal's site, 2009.