SUPPLEMENT TO "ADDRESSING NON-NORMALITY IN MULTIVARIATE ANALYSIS USING THE T-DISTRIBUTION"

FELIPE OSORIO, MANUEL GALEA, CLAUDIO HENRÍQUEZ,
AND REINALDO ARELLANO-VALLE

Abstract. In this supplement, we present some basic properties of the multivariate t-distribution based on the parameterization introduced by Sutradhar (1993) and Fiorentini et al. (2003). Then, we provide a detailed description of the maximum likelihood estimation procedure considering an EM algorithm.

Supplement A. Some properties of the multivariate t-distribution

Definition 1. The random variable U is said to have a gamma distribution, written as $U \sim \mathsf{Gamma}(a, b)$, if its probability density function is

$$f(u) = \frac{b^a u^{a-1} e^{-bu}}{\Gamma(a)}, \quad u, a, b > 0.$$

We have the following moments:

$$\mathrm{E}(U) = a/b, \qquad \mathrm{var}(U) = a/b^2, \qquad \mathrm{E}(\log(U)) = \psi(a) - \log(b),$$

where $\psi(u) = \mathrm{d}\log(\Gamma(u))/du$ is digamma function. The following result presents a summary of the properties of the t distribution.

Property 2. *Let $\boldsymbol{X} \sim \mathsf{T}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \eta)$ be a random vector with density function defined in Equation (1) from manuscript. Then, the random vector $\boldsymbol{X}$ has the following properties:*

(i) *The random vector $\boldsymbol{X}$ can be written using the following representation:*

$$\boldsymbol{X}|v \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/v) \quad \textit{and} \quad v \sim \mathsf{Gamma}\Big(\frac{1}{2\eta}, \frac{1}{2c(\eta)}\Big),$$

with $c(\eta) = \eta/(1 - 2\eta)$.

(ii) *For $v \sim \mathsf{Gamma}\Big(\frac{1}{2\eta}, \frac{1}{2c(\eta)}\Big)$, we have*

$$\mathrm{E}(u^{-r}) = \Big(\frac{1}{2c(\eta)}\Big)^r \frac{\Gamma\Big(\frac{1}{2\eta} - r\Big)}{\Gamma\Big(\frac{1}{2\eta}\Big)}, \qquad r < \frac{1}{\eta}.$$

(iii) $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ *and* $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. *Moreover, by using conditional expectations, we obtain*

$$\begin{aligned} \mathrm{E}(\boldsymbol{X}) &= \mathrm{E}(\mathrm{E}(\boldsymbol{X}|v)) = \boldsymbol{\mu}, \\ \mathrm{Cov}(\boldsymbol{X}) &= \mathrm{E}(\mathrm{Cov}(\boldsymbol{X}|v)) + \mathrm{Cov}(\mathrm{E}(\boldsymbol{X}|v)) = \boldsymbol{\Sigma}. \end{aligned}$$

The authors were partially supported by FONDECYT grants 1140580 and 1150325, respectively.

(iv) *Let* $\delta^2 = (\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$. *Then, the random variable*

$$F = \Big(\frac{1}{1-2\eta}\Big)\frac{\delta^2}{p} \sim F(p, 1/\eta),$$

with $\mathrm{E}(\delta^2) = p$.

(v) *The multivariate kurtosis coefficient (Mardia, 1970) is given by*

$$\beta_{2,p} = \mathrm{E}[\{(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\}^2] = p(p+2)(\kappa+1),$$

where $\kappa = 2\eta/(1-4\eta)$ *represents the excess of the kurtosis.*

We use $\boldsymbol{Z} = \boldsymbol{X}-\boldsymbol{\mu} \overset{\mathrm{d}}{=} v^{-1/2}\boldsymbol{Y}$, *where* $\boldsymbol{Y} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. *Then, we have that*

$$\begin{aligned}\beta_{2,p} &= \mathrm{E}(\boldsymbol{Z}^\top \mathrm{E}^{-1}(\boldsymbol{Z}\boldsymbol{Z}^\top)\boldsymbol{Z})^2 = \mathrm{E}(v^{-1}\boldsymbol{Y}^\top \mathrm{E}^{-1}(u^{-1}\boldsymbol{Y}\boldsymbol{Y}^\top)\boldsymbol{Y})^2 \\ &= \frac{\mathrm{E}(v^{-2})}{\mathrm{E}^2(v^{-1})}\mathrm{E}(\boldsymbol{Y}^\top \mathrm{E}^{-1}(\boldsymbol{Y}\boldsymbol{Y}^\top)\boldsymbol{Y})^2 = \Big(\frac{1-2\eta}{1-4\eta}\Big)p(p+2), \qquad \eta < 1/4.\end{aligned}$$

In addition, because $(1-2\eta)/(1-4\eta) > 1$ *for* $0 \le \eta < 1/4$, *we have that the kurtosis of the multivariate* t *distribution is always higher than that of the normal distribution (Yamaguchi, 2003).*

(vi) *Using results in Abramowitz and Stegun (1970), [p. 257], it follows that*

$$\lim_{\eta\to 0^+} K_p(\eta) = 1/(2\pi)^{p/2} \quad \text{and} \quad \lim_{\eta\to 0^+}\big(1 + c(\eta)\delta^2\big)^{-1/2\eta} = \exp(-\delta^2/2),$$

and then, when $\eta \to 0^+$, *the multivariate normal distribution is obtained, whose density is given by*

$$f_N(\boldsymbol{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp(-\delta^2/2), \quad \boldsymbol{x} \in \mathbb{R}^p.$$

The following lemma, adapted from Sutradhar (1993) and Bolfarine and Galea (1996), allows us to compute the expected information matrix (see also Lange et al., 1989).

Lemma 3. *If* $\boldsymbol{Z} \sim \mathsf{T}_p(\boldsymbol{0}, \boldsymbol{\Sigma}, \eta)$ *and if* $q = 1 + c(\eta)\delta^2$ *with* $\delta^2 = \boldsymbol{Z}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}$, *then*

(i) $\mathrm{E}(q^{-1}) = (1+\eta p)^{-1}$,
(ii) $\mathrm{E}(q^{-1}\boldsymbol{Z}) = \boldsymbol{0}$,
(iii) $\mathrm{E}(q^{-1}\boldsymbol{Z}\boldsymbol{Z}^\top) = \{(1-2\eta)/(1+\eta p)\}\boldsymbol{\Sigma}$,
(iv) $\mathrm{E}(q^{-1}\delta^2) = p(1-2\eta)/(1+p\eta)$,
(v) $\mathrm{E}(q^{-2}) = (1+2\eta)/(1+\eta p)(1+(p+2)\eta)$,
(vi) $\mathrm{E}(q^{-2}\boldsymbol{Z}) = \boldsymbol{0}$,
(vii) $\mathrm{E}(q^{-2}\boldsymbol{Z}\boldsymbol{Z}^\top) = \{(1-2\eta)/(1+\eta p)(1+(p+2)\eta)\}\boldsymbol{\Sigma}$,
(viii) $\mathrm{E}(q^{-2}\delta^2) = pc^{-1}(\eta)\eta/(1+\eta p)(1+(p+2)\eta)$,
(ix) $\mathrm{E}(q^{-2}\delta^2\boldsymbol{Z}) = \boldsymbol{0}$,
(x) $\mathrm{E}(q^{-2}\delta^2\boldsymbol{Z}\boldsymbol{Z}^\top) = \{(1-2\eta)^2(p+2)/(1+\eta p)(1+(p+2)\eta)\}\boldsymbol{\Sigma}$,
(xi) $\mathrm{E}(q^{-2}\delta^4) = p(p+2)(1-2\eta)^2/(1+\eta p)(1+(p+2)\eta)$,
(xii) $\mathrm{E}(\log q) = \psi\Big(\frac{1+\eta p}{2\eta}\Big) - \psi\Big(\frac{1}{2\eta}\Big)$,
(xiii) $\mathrm{E}\{c(\eta)\delta^2 q^{-1}\} = \frac{p\eta}{1+p\eta}$.

Next, an extension of Theorem 4.1 (i) given in Magnus and Neudecker (1979) is presented.

Lemma 4. *Assume that* $\boldsymbol{Y} \sim \mathsf{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$*; then,*

$$\mathrm{E}(\boldsymbol{Y}\boldsymbol{Y}^\top \otimes \boldsymbol{Y}\boldsymbol{Y}^\top) = 2(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\boldsymbol{N}_p + (\operatorname{vec}\boldsymbol{\Sigma})(\operatorname{vec}\boldsymbol{\Sigma})^\top,$$

where $\boldsymbol{N}_p = \frac{1}{2}(\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)$*, with* $\boldsymbol{K}_p$ *as the commutation matrix of order* p*.*

Lemmas 3 and 4 allow us to establish the following result:

Lemma 5. *Let* $\boldsymbol{Z} \sim \mathsf{T}_p(\mathbf{0}, \boldsymbol{\Sigma}, \eta)$ *and* $q = 1 + c(\eta)\delta^2$*. Then,*

$$\mathrm{E}(q^{-2}\boldsymbol{Z}\boldsymbol{Z}^\top \otimes \boldsymbol{Z}\boldsymbol{Z}^\top) = \frac{(1-2\eta)^2}{(1+\eta p)(1+(p+2)\eta)}\Big\{2(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\boldsymbol{N}_p + (\operatorname{vec}\boldsymbol{\Sigma})(\operatorname{vec}\boldsymbol{\Sigma})^\top\Big\}.$$

because $1 + c(\eta)\delta^2 \overset{\mathrm{d}}{=} 1 + p\eta F(p, 1/\eta)$*.*

Supplement B. ML estimation using the EM algorithm

The EM algorithm (Dempster et al., 1977) corresponds to an iterative procedure that applies a data augmentation scheme by introducing latent variables or missing data, indicated by $\boldsymbol{X}_{\mathsf{mis}}$, to enable the easy manipulation of the log-likelihood function of complete data $\boldsymbol{X}_{\mathsf{com}} = (\boldsymbol{X}_{\mathsf{obs}}^\top, \boldsymbol{X}_{\mathsf{mis}}^\top)^\top$, denoted by $\mathcal{L}_c(\boldsymbol{\theta})$. The EM algorithm then maximizes the log-likelihood function of observed data, $\mathcal{L}(\boldsymbol{\theta})$, iteratively by alternating between the following two steps:

E-step: For a current estimation $\boldsymbol{\theta}^{(k)}$, compute the conditional expectation

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \mathrm{E}[\mathcal{L}_c(\boldsymbol{\theta})|\boldsymbol{x}_{\mathsf{obs}}, \boldsymbol{\theta}^{(k)}],$$

M-step: Update $\boldsymbol{\theta}^{(k+1)}$ by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ in relation to $\boldsymbol{\theta}$.

It can be shown (see Wu, 1983) that under mild general conditions, the EM algorithm increases the observed data log-likelihood function after each iteration and that the sequence $\{\boldsymbol{\theta}^{(k)}\}$ converges to a stationary point of $\mathcal{L}(\boldsymbol{\theta})$.

To obtain the maximum likelihood estimate in our context, we augmented the observed data, $\boldsymbol{X}_{\mathsf{obs}} = \{\boldsymbol{X}_1^\top, \dots, \boldsymbol{X}_n^\top\}$ by incorporating latent variables to obtain $\boldsymbol{X}_{\mathsf{com}} = \{(\boldsymbol{X}_1^\top, v_1), \dots, (\boldsymbol{X}_n^\top, v_n)\}$. Thus, we consider the following hierarchical model:

$$\boldsymbol{X}_i|v_i \overset{\mathsf{ind}}{\sim} \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/v_i), \qquad \text{and} \qquad v_i \overset{\mathsf{ind}}{\sim} \mathsf{Gamma}\Big(\frac{1}{2\eta}, \frac{1}{2c(\eta)}\Big), \tag{S.1}$$

for $i = 1, \dots, n$. Based on the hierarchical structure of the model for the augmented data given in (S.1), clearly, the conditional distribution required to evaluate the expectation step of the EM algorithm takes the form

$$v_i|\boldsymbol{x}_i \overset{\mathsf{ind}}{\sim} \mathsf{Gamma}\Big(\frac{1/\eta + p}{2}, \frac{1/c(\eta) + \delta_i^2}{2}\Big), \qquad i = 1, \dots, n.$$

Thus, the conditional expectation of the complete-data log-likelihood function can be expressed as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = Q_1(\boldsymbol{\mu}, \boldsymbol{\phi}|\boldsymbol{\theta}^{(k)}) + Q_2(\eta|\boldsymbol{\theta}^{(k)}),$$

where

$$Q_1(\boldsymbol{\mu},\boldsymbol{\phi}|\boldsymbol{\theta}^{(k)}) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^n v_i^{(k)}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}), \tag{S.2}$$

$$Q_2(\eta|\boldsymbol{\theta}^{(k)}) = n\Big\{\frac{1}{2\eta}\log\Big(\frac{1}{2c(\eta)}\Big) - \log\Gamma\Big(\frac{1}{2\eta}\Big) + \frac{1}{2c(\eta)}\Big[\psi\Big(\frac{1/\eta^{(k)}+p}{2}\Big) - \log\Big(\frac{1/\eta^{(k)}+p}{2}\Big) + \frac{1}{n}\sum_{i=1}^n(\log v_i^{(k)} - v_i^{(k)})\Big]\Big\}. \tag{S.3}$$

In this case, the weights $v_i^{(k)}$ defined by the EM algorithm correspond to the following conditional expectations:

$$v_i^{(k)} = \mathrm{E}(u_i|\boldsymbol{x}_i,\boldsymbol{\theta}^{(k)}) = \frac{1/\eta^{(k)}+p}{1/c(\eta^{(k)}) + \delta_i^2(\boldsymbol{\tau}^{(k)})}.$$

To update $\boldsymbol{\mu}^{(k+1)}$ and $\boldsymbol{\phi}^{(k+1)}$, we maximize $Q_1(\boldsymbol{\mu},\boldsymbol{\phi}|\boldsymbol{\theta}^{(k)})$ given in (S.2) in relation to $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$ to obtain

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{\sum_{j=1}^n v_j^{(k)}}\sum_{i=1}^n v_i^{(k)}\boldsymbol{x}_i, \tag{S.4}$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n}\sum_{i=1}^n v_i^{(k)}(\boldsymbol{x}_i-\boldsymbol{\mu}^{(k+1)})(\boldsymbol{x}_i-\boldsymbol{\mu}^{(k+1)})^\top. \tag{S.5}$$

In addition, we can independently update the shape parameter of the t-distribution by maximizing $Q_2(\eta|\boldsymbol{\theta}^{(k)})$ defined in Equation (S.3) using a unidimensional optimization procedure. The parameter estimation approach proposed in this work has been implemented in an R package named `MVT`, which is available at CRAN.

References

Abramowitz, M., Stegun, I.A. (1970). *Handbook of Mathematical Functions*. Dover, New York.

Bolfarine, H., Galea, M. (1996). On structural comparative calibration under a t-model. *Computational Statistics* **11**, 63-85.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1-38.

Fiorentini, G., Sentana, E., Calzolari, G. (2003). Maximum likelihood estimation and inference in multivariate conditionally heteroscedastic dynamic regression models with Student t innovations. *Journal of Business & Economic Statistics* **21**, 532-546.

Lange, K., Little, R.J.A., Taylor, J.M.G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84**, 881-896.

Magnus, J.R., Neudecker, H. (1979). The commutation matrix: Some properties and applications. *The Annals of Statistics* **7**, 381-394.

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519-530.

Sutradhar, B.C. (1993). Score test for the covariance matrix of elliptical t-distribution. *Journal of Multivariate Analysis* **46**, 1-12.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11**, 95-103.

Yamaguchi, K. (2003). Robust model and the EM algorithm. In: Watanabe, M., Yamaguchi, K. (eds.) *The EM Algorithm and Related Statistical Models*, pp. 37-64. Marcel Dekker, New York.

Departamento de Matemática, Universidad Técnica Federico Santa María, Chile
Current address: Avenida España 1680, Valparaíso, Chile
Orcid ID: https://orcid.org/0000-0002-4675-5201
Email address: felipe.osorios@usm.cl

Departamento de Estadística, Pontificia Universidad Católica de Chile
Current address: Avenida Vicuña Mackena 4860, Santiago, Chile
Orcid ID: https://orcid.org/0000-0001-9819-5843 (M. Galea)
Orcid ID: https://orcid.org/0000-0002-5121-9702 (R. Arellano-Valle)
Email address: mgalea@mat.uc.cl, cnhenriquez@mat.uc.cl, reivalle@mat.uc.cl