REVSTAT – Statistical Journal
Volume 22, Number 1, January 2024, Supplementary Material, 1–6

# Supplementary Material for "Analysis of Antibody Data Using Skew-Normal and Skew-t Mixture Models"

Authors: Tiago Dias Domingues ✉
– CEAUL, Faculdade de Ciências Universidade de Lisboa, Portugal
tmdomingues@fc.ul.pt

Helena Mouriño
– CMAFcIO, Faculdade de Ciências, Universidade de Lisboa, Portugal
mhnunes@fc.ul.pt

Nuno Sepúlveda
– Faculty of Mathematics and Information Science, Warsaw University of Technology, Poland
– CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal
N.Sepulveda@mini.pw.edu.pl

Received: December 2020 Revised: March 2022 Accepted: March 2022

## A. APPENDIX – Theoretical details of the Skew-Normal and Skew-t distributions as members of the SMSN family

A random variable $Z_k$ belongs to the SMSN family with location parameter $\mu_k$, scale parameter $\sigma_k^2$ and skewness parameter $\alpha_k$ (denoted as $Z_k \sim \mathcal{SMSN}(\mu_k, \sigma_k^2, \alpha_k, H)$) if it can be written in the following way:

$$Z_k = \mu_k + \frac{W_k}{\sqrt{U_k}}, \tag{A.1}$$

where $\mu_k$ is the location parameter; $U_k$ is a random variable with distribution function $H_k(\cdot, \boldsymbol{v}_k)$ and pdf $h_k(\cdot, \boldsymbol{v}_k)$; $\boldsymbol{v}_k$ is either a scalar or a vector of parameters indexing the distribution of $U_k$; and $W_k \sim \mathcal{SN}(0, \sigma_k^2, \alpha_k)$ which is assumed to be independent of $U_k$ [1, 2].

✉ Corresponding author.

Based on expression (A.1), it is worth noting that the conditional distribution $Z_k|U_k = u$ takes the form

$$\begin{aligned} F_{Z_k|U_k=u}(z) &= P(Z_k \leq z) = P(\mu_k + \frac{1}{\sqrt{u}}W_k \leq z) \\ &= P(W_k \leq \sqrt{u}(z-\mu_k)) = F_{W_k}(\sqrt{u}(z-\mu_k)),\ z \in \mathbb{R}. \end{aligned} \tag{A.2}$$

Thus,

$$\begin{aligned} f_{Z_k|U_k=u}(z) &= \frac{d}{dz}F_{W_k}(z) = \sqrt{u} \times f_{W_k}(\sqrt{u}(z-\mu_k)) \\ &= \sqrt{u}\frac{2}{\sigma_k}\phi\left(\frac{\sqrt{u}(z-\mu_k)}{\sigma_k}\right)\Phi\left(\frac{\alpha_k\sqrt{u}(z-\mu_k)}{\sigma_k}\right),\ z \in \mathbb{R}, \end{aligned} \tag{A.3}$$

where $\phi(.)$ represents the pdf of the standard Normal distribution. Which is equivalent to, $f_{Z_k|U_k=u}(z) = 2\phi\left(z; \mu_k, \frac{\sigma_k^2}{u}\right)\Phi\left(\frac{\alpha_k(z-\mu_k)}{\sigma_k/\sqrt{u}}\right)$, $z \in \mathbb{R}$, where $\phi(\cdot; \mu_k, \frac{\sigma_k^2}{u})$ denotes the pdf of the $\mathcal{N}(\mu_k, \frac{\sigma_k^2}{u})$. Hence, $Z_k|U_k = u \sim \mathcal{SN}(\mu_k, \frac{\sigma_k^2}{u}, \alpha_k)$.

The marginal probability density distribution of $Z_k$ is given by

$$f_{Z_k}(z) = \int_0^{+\infty} 2\phi\left(z; \mu_k, \frac{\sigma_k^2}{u}\right)\Phi\left(\frac{\alpha_k(z-\mu_k)}{\sigma_k/\sqrt{u}}\right)dH(u; \boldsymbol{v}),\ z \in \mathbb{R}.$$

The name of this class of distributions relies on the fact that the density function of $Z_k$ (A.1) involves an infinite mixture of Skew-Normal distributions.

To model different patterns arising from serological data, we rely on 4 particular cases of the SMSN family. The first one is the case of the Skew-Normal distribution itself. This happens when $U_k$ is not a random variable but rather the scalar $u = 1$. Then, variable $Z_k$ in expression (A.1) simplifies to $Z_k = \mu_k + W_k$. Hence,

$$F_{Z_k}(z) = P(W_k \leq z - \mu_k) = F_{W_k}(z - \mu_k), z \in \mathbb{R}, \tag{A.4}$$

$$f_{Z_k}(z) = f_{W_k}(z - \mu_k) = 2\phi(z - \mu_k; 0, \sigma_k^2)\Phi\left(\alpha_k\left(\frac{z-\mu_k}{\sigma_k}\right)\right). \tag{A.5}$$

Therefore, $Z_k \sim \mathcal{SN}(\mu_k, \sigma_k^2, \alpha_k)$.

The second case is a simplification of the previous one when $\alpha_k = 0$. In this case, the Skew-Normal distribution reduces to the usual (symmetric) Normal distribution. In fact, when $\alpha_k = 0$ we get

$$f_{Z_k}(z) = 2\phi(z - \mu_k; 0, \sigma_k^2)\Phi(0) = \phi(z - \mu_k; 0, \sigma_k^2) = \phi(z; \mu_k, \sigma_k^2), z \in \mathbb{R},$$

where $\phi(\cdot; \mu_k, \sigma_k^2)$ represents the pdf of the $\mathcal{N}(\mu_k, \sigma_k^2)$ distribution.

The third and fourth cases are the skew Student's t-distribution and its symmetric counterpart, hereafter referred to as Skew-t and Student's t-distributions for short, respectively. These distributions can be obtained as follows.

Let $U_k$ be a Gamma distribution with shape and rate parameters $\frac{v}{2}$ and $\frac{v}{2}$, respectively, that is, $U_k \sim Gamma(\frac{v}{2}, \frac{v}{2})$. The formulation is such that the mean of $U_k$ is equal to one.

Note that $Z_k = \mu_k + \frac{W_k}{\sqrt{U_k}}$, where $W_k \sim \mathcal{SN}(0, \sigma_k^2, \alpha_k), U_k \sim Gamma(\frac{v}{2}, \frac{v}{2})$ are independent random variables, is equivalent to $Z_k = \mu_k + \frac{W_k}{\sqrt{\frac{R_k}{v}}}$ where $R_k$ is a $\chi^2$ distribution with $v$ degrees of freedom.

The conditional cumulative distribution function and the corresponding pdf of $Z_k|U_k = u$ are given by the expressions (A.2) and (A.3), respectively. According to expression (A.4), the marginal probability density distribution of $Z_k$ takes the form

$$
\begin{aligned}
(\text{A.6}) \quad f_{Z_k}(z) &= \int_0^{+\infty} f_{Z_k|U_k=u}(z) f_{U_k}(u) du \\
&= \int_0^{+\infty} 2\sqrt{u}\phi(\sqrt{u}(z-\mu_k); 0, \sigma_k^2)\Phi\left(\frac{\alpha_k\sqrt{u}(z-\mu_k)}{\sigma_k}\right)\frac{(\frac{v_k}{2})^{\frac{v_k}{2}} u^{\frac{v_k}{2}-1} e^{-\frac{v_k}{2}-u}}{\Gamma(\frac{v_k}{2})} du \\
&= \frac{2v_k^{\frac{v_k}{2}}}{\sigma_k\sqrt{\pi}2^{\frac{v_k+1}{2}}\Gamma(\frac{v_k}{2})}\int_0^{+\infty}\Phi(\sqrt{u}A)u^{\frac{1}{2}(v_k-1)}e^{-\frac{1}{2}u(d+v_k)}du,
\end{aligned}
$$

with $A = \frac{\alpha_k(z-\mu_k)}{\sigma_k}, d = \left(\frac{z-\mu_k}{\sigma_k}\right)^2$.

Integrating expression (A.6) by substitution of the variable $s = \frac{1}{2}u(d+v_k)$, we obtain

(A.7)

$$
\begin{aligned}
f_{Z_k}(z) &= \frac{2}{\sigma_k\sqrt{\pi v_k}\,\Gamma(\frac{v_k}{2})}\left(1+\frac{d}{v_k}\right)^{-\frac{1}{2}(v_k+1)}\int_0^{+\infty}\Phi\left(A\sqrt{\frac{2s}{d+v_k}}\right)s^{\frac{1}{2}(v_k-1)}e^{-s}ds \\
&= \frac{2\,\Gamma(\frac{v_k+1}{2})}{\sigma_k\sqrt{\pi v_k}\,\Gamma(\frac{v_k}{2})}\left(1+\frac{d}{v_k}\right)^{-\frac{1}{2}(v_k+1)}\int_0^{+\infty}\Phi\left(A\sqrt{\frac{2}{d+v_k}}\sqrt{s}\right)\frac{1}{\Gamma(\frac{v_k+1}{2})}s^{\frac{1}{2}(v_k-1)}e^{-s}ds \\
&= \frac{2\,\Gamma(\frac{v_k+1}{2})}{\sigma_k\sqrt{\pi v_k}\,\Gamma(\frac{v_k}{2})}\left(1+\frac{d}{v_k}\right)^{-\frac{1}{2}(v_k+1)}\times \\
&\quad\times\int_0^{+\infty}P\left(Z\le A\sqrt{\frac{2}{d+v_k}}\sqrt{s}|S=s\right)\frac{1}{\Gamma(\frac{v_k+1}{2})}s^{\frac{1}{2}(v_k-1)}e^{-s}ds.
\end{aligned}
$$

It is important to notice the following Lemma [3].

**Lemma:** *Suppose that* $Z \sim \mathcal{N}(0,1)$, $Y \sim Gamma(m,1), R \sim t_{2m}$, $m > 0$. *It can be proved that*

$$
E\left(\Phi(c\sqrt{Y})\right) = \int_0^{+\infty} P(Z \le c\sqrt{y}|Y=y) f_Y(y) dy = P(R \le c\sqrt{m}), c \in \mathbb{R}.
$$

Applying this Lemma to expression (A.7) leads to

$$
\begin{aligned}
f_{Z_k}(z) &= \frac{2\,\Gamma(\frac{v_k+1}{2})}{\sigma_k\sqrt{\pi v_k}\,\Gamma(\frac{v_k}{2})}\left(1+\frac{d}{v_k}\right)^{-\frac{1}{2}(v_k+1)} E\left(\Phi\left(A\sqrt{\frac{2}{d+v_k}}\sqrt{s}\right)\right) \\
&= 2\;t(z;\mu_k,\sigma_k,v_k+1)\;E\left(\Phi\left(A\sqrt{\frac{2}{d+v_k}}\sqrt{s}\right)\right) \\
&= 2\;t(z;\mu_k,\sigma_k,v_k+1)\;P\left(T\leq A\sqrt{\frac{v_k+1}{d+v_k}};v_k+1\right) \\
&= 2\;t(z;\mu_k,\sigma_k,v_k+1)\;T\left(A\sqrt{\frac{v_k+1}{d+v_k}};v_k+1\right),
\end{aligned}
\tag{A.8}
$$

where $t(\cdot;\mu_k,\sigma_k,v_k+1)$ denotes the probability density function of a Generalized Student-t distribution with location parameter $\mu_k$, scale parameter $\sigma_k$ and $v_k+1$ degrees of freedom; $T(\cdot;v_k+1)$ represents the cumulative distribution function of a standard Student-t distribution with $v_k+1$ degrees of freedom.

In short, if $Z_k \sim ST(\mu_k,\sigma_k^2,\alpha_k,v_k)$, then its pdf is given by

$$
f_{Z_k}(z) = 2\;t(z;\mu_k,\sigma_k,v_k+1)\;T\left(A\sqrt{\frac{v_k+1}{d+v_k}};v_k+1\right).
$$

## B. APPENDIX – Computational aspects of the proposed models

For the analysis of our data, we used a maximum of 10,000 iterations for the EM algorithm. To increase the chance of obtaining the correct ML estimates, we ran the algorithm with 100 different initial values for model parameters. The tolerance for the error between two consecutive iterations was set $10^{-5}$.

We did not experience any computational cost while analysing our data. This can be explained by two main reasons. Firstly, we only had a small number of antibodies under analysis. Secondly, in most of the cases, there was a clear separation of the seropositive and seronegative populations. As such, the convergence of the EM algorithm was obtained relatively quickly with an average of less than 500 iterations (see examples in Table 1). The only exception was the VZV data set where the estimation the Skew-t mixture models required more than 1500 iterations on average.

In general, we envision some computational cost when analysing the data sets with hundreds or even thousands of screened antibodies, as analysed by Loebel *et al.* [4], Blomberg *et al.* [5], van den Hoogen *et al.* [6] or Proeitti *et al.* [7]. On the one hand, larger data sets are likely to contain different data structures where seropositive and seronegative populations might overlap with different degrees. In this regard, a high overlap between these populations is expected to increase the number of iterations until the convergence of the EM algorithm.

Different data structures might also imply the necessity of estimating models with greater number of components, thus, increasing the number of estimated models. Another computational cost comes from the fact that current EM algorithm implemented in `mixsmsn` requires careful initialization in order to obtain the correct parameter estimates. In our example of application, we overcame this problem by running the EM algorithm with different initial values for the parameter estimates. However, such estimation strategy is time consuming when there is a large number of antibodies under analysis. For the case of Gaussian mixture models, there are modifications available of the EM algorithm that do not require careful initialization [8, 9]. These modified algorithms have also the advantage of estimating the number of the mixture components simultaneously. For the case of SMSN mixture models, there are no such modified EM algorithms in terms of their computational efficieny. There is then a research opportunity to improve current EM algorithm for a wide application of SMSN finite models in the context of high-throughput serology analysis.

**Table 1**: Average number of iterations of the EM algorithm for estimating some SMSN finite models to herpesviruses serological data, where $g$ represents the number of components in the mixture model, $p$ represents the number of parameters of the mixture model.

| **Virus** | **SMSN** | $g$ | $p$ | Average iterations |
|---|---|---|---|---|
| CMV | Skew-t | 1 | 4 | 80.6 |
| | | 2 | 8 | 193.5 |
| | | 3 | 12 | 126.5 |
| EBV | Skew-t | 1 | 4 | 143.4 |
| | | 2 | 8 | 128.2 |
| | | 3 | 12 | 383.6 |
| HSV-1 | Skew-Normal | 1 | 3 | 38.8 |
| | | 2 | 7 | 213.2 |
| | | 3 | 11 | 173.2 |
| HSV-2 | Normal | 1 | 2 | 2.0 |
| | | 2 | 5 | 11.5 |
| | | 3 | 8 | 127.5 |
| HHV-6 | Skew-Normal | 1 | 3 | 60.0 |
| | | 2 | 7 | 119.9 |
| | | 3 | 11 | 379.5 |
| VZV | Skew-t | 1 | 4 | 279.3 |
| | | 2 | 8 | 1675.8 |
| | | 3 | 12 | 2416.4 |

## C. APPENDIX – Exploratory data analysis

**Table 2**: Empirical skewness and excess kurtosis coefficients for hypothetical seronegative and seropositive populations using the cutoff poins suggested by the manufacturers of the commercial kits.

| | Seronegative | | Seropositive | |
|---|---|---|---|---|
| **Virus** | **Skewness** | **Excess Kurtosis** | **Skewness** | **Excess Kurtosis** |
| CMV | -0.695 | 1.062 | -0.889 | 0.198 |
| EBV | -2.599 | 9.399 | -0.326 | -0.517 |
| HHV-6 | -1.095 | 2.411 | 0.231 | -0.129 |
| HSV-1 | 0.011 | -1.304 | -1.021 | -0.352 |
| HSV-2 | 0.604 | -0.181 | 0.139 | -0.639 |
| VZV | -1.298 | 0.444 | -1.231 | 1.087 |

## REFERENCES


[1] Basso, R.M.; Lachos, V.H.; Cabral, C.R.B. and Gosh, P. (2010). Robust mixture modelling based on scale mixtures of skew-normal distributions, *Computational Statistics and Data Analysis*, **54**, 2926–2941.

[2] Lachos Dávila, V.H.; Zeller, C.B. and Cabral, C.R.B. (2018). *Finite Mixture Of Skewed Distributions*, Springer.

[3] Azzalini, A. and Capitanio, A. (2014). *The Skew-Normal and Related Families*, Cambridge University Press, Cambridge.

[4] Loebel, M.; Eckey, M.; Sotzny, F.; Hahn, E.; Bauer, S.; Grabowski, P.; Zerweck, J.; Holenya, P.; Hanitsch, L.G.; Wittke, K.; Borchmann, P.; Rüffer, J.U.; Hiepe, F.; Ruprecht, K.; Behrends, U.; Meindl, C.; Volk, H.D.; Reimer, U. and Scheibenbogen, C. (2017). Serological profiling of the EBV immune response in Chronic Fatigue Syndrome using a peptide microarray, *PloS One*, **12**(6), e0179124.

[5] Blomberg, J.; Rizwan, M.; Böhlin-Wiener, A.; Elfaitouri, A.; Julin, P.; Zachrisson, O.; Rosén, A. and Gottfries, C.G. (2019). Antibodies to Human Herpesviruses in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Patients, *Frontiers in Immunology*, **10**, 1946.

[6] van den Hoogen, L.L.; Présumé, J.; Romilus, I.; Mondélus, G.; Elismé, T.; Sepúlveda, N.; Stresman, G.; Druetz, T.; Ashton, R.A.; Joseph, V.; Eisele, T.P.; Hamre, K.; Chang, M.A.; Lemoine, J.F.; Tetteh, K.; Boncy, J.; Existe, A.; Drakeley, C. and Rogier, E. (2020). Quality control of multiplex antibody detection in samples from large-scale surveys: the example of malaria in Haiti, *Scientific Reports*, **10**(1), 1135.

[7] Proietti, C.; Krause, L.; Trieu, A.; Dodoo, D.; Gyan, B.; Koram, K.A.; Rogers, W.O.; Richie, T.L.; Crompton, P.D.; Felgner, P.L. and Doolan, D.L. (2020). Immune Signature Against Plasmodium falciparum Antigens Predicts Clinical Immunity in Distinct Malaria Endemic Communities, *Molecular & Cellular Proteomics*, **19**, 101–113.

[8] Figueiredo, M.A.T. and Jain, A.K. (2002). Unsupervised learning of finite mixture models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 381–396.

[9] Yang, M.; Lai, C. and Lin, C. (2012). A robust EM clustering algorithm for Gaussian mixture models, *Pattern Recognition*, **45**, 3950–3961.