REVSTAT – Statistical Journal
Volume 22, Number 1, January 2024, 111–132
https://doi.org/10.57805/revstat.v22i1.455

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

Abstract:

- Gaussian mixture models, which assume a Normal distribution for each component, are popular in antibody (or serological) data analysis to help determining antibody-positive and antibody-negative individuals. In this work, we advocate using finite mixture models based on Skew-Normal and Skew-t distributions for serological data analysis. These flexible mixing distributions have the advantage of describing right and left asymmetry often observed in the distributions of known antibody-negative and antibody-positive individuals, respectively. We illustrate the application of these alternative mixture models in a data set on the role of human herpesviruses in the Myalgic Encephalomyelitis/Chronic Fatigue Syndrome.

Keywords:

- *finite mixture models; Skew-Normal; Skew-t; seropositivity.*

AMS Subject Classification:

- 62H30, 62P10.

✉ Corresponding author.

1. INTRODUCTION

Antibodies are proteins produced by B cells upon recognition of an antigen derived from an infectious agent. In general, they contribute to microbial clearance and, if maintained in the body over time, they translate into a quicker and more efficient immune response upon repeated exposure to the same infection. In turn, autoantibodies bind to antigens from the body and they are usually present in autoimmunity diseases, such as multiple sclerosis and rheumatoid arthritis.

In routine laboratories, antibodies (or autoantibodies) against a specific antigen are quantified by the enzymatic-linked immunosorbent assays (ELISA) using serum samples. The readout of these assays is a light intensity, also known as optical density, which is converted into a concentration or a titre using a calibration curve of known antibody concentrations. In practice, these assays are easily standardized, widely available, and ideal for high-throughput analysis of antibodies against a single antigen [1]. Such advantages make them suitable for large-scale serological surveys where one aims to estimate the prevalence of exposure to a given pathogen in the population [1, 2, 3].

With the development of high-throughput technologies, antibody quantification is shifting from the ELISA to microarray, luminex, or cytometry bead assays, where many antibodies can be evaluated in the same serum sample. However, these technologies are still being optimized before their wide use.

Antibody (or serological) statistical analysis of antibody (or serological) data often assumes the existence of multiple latent populations each one representing a distinct level of exposure to a given antigen. This basic assumption calls for the use of finite mixture models. In general, these models can be more or less complex, depending on the number of mixing distributions used to describe the data [4]. In routine serological applications, one assumes a model with only two latent populations: seronegative and seropositive individuals or, equivalently, antibody-negative and antibody-positive individuals [5, 6, 7]. Models comprising more than two serological populations are also used in practice [8, 9, 10, 11, 12], but their interpretation is not straightforward [13].

A common choice for the mixing distribution is the Lognormal distribution in the original scale of the measurements or, equivalently, the Normal distribution after applying the logarithmic transformation to the data [6, 8]. Gamma and Weibull are other choices among textbook probability distributions [7, 11].

Less-trivial mixture models can be also used in the analysis. For example, a mixture of two truncated Normal distributions was used to describe data where observations could fall below the lower limit of detection or above the upper limit of detection of the assay [9]. Another alternative model was the mixture of a Normal distribution and a combination of half-Normal distributions for the seronegative and seropositive populations, respectively [5]. The rationale behind this model is that antibody levels decrease over time and, therefore, the seropositive populations should have left-skewed distributions [8]. Similarly, seronegative populations should have right-skewed distribution due to the detection of non-specific antibodies at lower concentrations of the target antibodies. Notwithstanding the suitability of these alternative models to tackle specific characteristics of serological data, none of the above models shows sufficient flexibility in terms of skewness and flatness of each mixing distribution that could be used serological data analysis and its automation in the context of high-throughput data.

We then propose using finite mixture models based on Skew-Normal and Skew-t distributions scale in routine serological data analysis. These alternative families of distributions are highly flexible due to three parameters that control the location, the scale, and the skewness of the resulting distribution. In the case of the Skew-t distribution, further flexibility can be achieved by an additional parameter that controls the weight of tails. These distributions also have the advantage of including the Normal distribution, the Generalized Student's t-distribution, and its skewed version as special cases [14]. As an example of application, we use these models to analyse a data set of 6 antibody responses to herpesviruses in the context of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) [15].

2. DATA UNDER ANALYSIS

ME/CFS is a multifactorial disease whose patients experience persistent fatigue that cannot be alleviated by rest, or they suffer from post-exertional malaise upon minimal physical and mental activity [16]. The cause of the disease remains unknown, but it is often linked to infections by herpesviruses.

The data set under analysis is part of the United Kingdom ME/CFS biobank, and it was published in a recent study with the aim of investigating the immunological component of the disease [15]. In the data set, there is a total of 406 individuals divided into three main groups: healthy controls (HC, $n = 107$; 26.4%), patients with ME/CFS ($n = 250$; 61.8%), and patients with multiple sclerosis (MS, $n = 49$; 12.1%). The group of patients with ME/CFS was subdivided into 196 patients with mild or moderate symptoms (ME-M) and 54 severely affected patients who are home- or even bed-bound (ME-S).

The data set comprises six serological antibody concentrations measured by commercial ELISA kits and related to the following common herpesviruses: human cytomegalovirus, CMV; Epstein-Barr virus, EBV; human herpesvirus-6, HHV-6; types 1 and 2 herpes simplex viruses, HSV-1 and HSV-2, respectively; and varicella-zoster virus, VZV. Note that the tested antibodies against EBV were specific to the viral-capsid antigen.

The concentration of the antibodies was expressed in arbitrary units per ml (U/ml). According to the kit manufacturers, individuals with antibody concentration ≤ 8 U/ml or ≥ 12 U/ml should be classified as seronegative or seropositive, respectively, for all antibodies except for the one against HHV-6. For antibodies against HHV-6, seronegative and seropositivity should be defined as ≤ 10.5 U/ml or ≥ 12.5 U/ml, respectively. Samples with concentrations between the above limits were considered equivocal.

3. STATISTICAL ANALYSIS OF SEROLOGICAL DATA

3.1. Finite mixture models

Let $G_1, ..., G_g$ be the partition from a superpopulation G (sample space) and $\pi_1, ..., \pi_g$ the probabilities of sampling an individual belonging to each latent population (with the usual

restriction of $\sum_{k=1}^{g} \pi_k = 1$ and $0 \leq \pi_k \leq 1$). A random variable Z is a finite mixture of independent random variables $Z_1, Z_2, ..., Z_g$ if the probability density function (pdf) of Z is given by

$$f(z) = \sum_{k=1}^{g} \pi_k \, f_{Z_k}(z; \boldsymbol{\theta}_k)\,, \tag{3.1}$$

where $f_{Z_k}(z; \boldsymbol{\theta}_k)$ is the mixing probability density function (pdf) of Z_k associated with the k-th latent population and parameterized by the vector $\boldsymbol{\theta}_k = \{\theta_1, ..., \theta_g\}$.

A common choice for the mixing distribution in the serological analysis is the Normal distribution which is symmetric around the mean, and it is a mesokurtic distribution (with kurtosis of 3 irrespective of the mean and standard deviation). Alternatively, the Generalized Student's t can be used as the mixing distribution because it has heavier tails than the Normal distribution. However, data from malaria seroepidemiological studies show long tails and marked right asymmetry in each latent population even after applying a logarithmic transformation [7]. In such cases, one aims to incorporate asymmetry and heavy tails in the finite mixture modelling. This is the purpose of using the Skew-Normal and Skew-t as mixing distributions [17, 18]. These alternative distributions are members of the so-called scale mixtures of Skew-Normal (SMSN) distributions [14]. This class of probability distributions is defined as follows.

Let Z_k be a random variable following a SMSN distribution with μ_k, σ_k^2, and α_k as the location, scale, and skewness parameters, respectively, and $H_k(\cdot\,; \boldsymbol{v}_k)$ as the mixing distribution parameterized by θ_k. Then, it can be written as

$$Z_k = \mu_k + \frac{W_k}{\sqrt{U_k}}\,, \tag{3.2}$$

where U_k is a random variable with distribution function $H_k(\cdot\,; \boldsymbol{v}_k)$ and $W_k \sim \mathcal{SN}(0, \sigma_k^2, \alpha_k)$, and W_k and U_k are two independent random variables [14]. See Appendix A in the Supplementary Material for additional theoretical discussion about this class of distributions.

3.1.1. Skew-Normal as a mixing distribution

Let W_k be a random variable with a Skew-Normal distribution with location parameter μ_k, scale parameter σ_k^2 and skewness parameter α_k (denoted as $W_k \sim \mathcal{SN}(\mu_k, \sigma_k^2, \alpha_k)$). The corresponding pdf is given by

$$\begin{aligned} f_{W_k}(w) &= 2\,\frac{1}{\sqrt{2\pi}\sigma_k}\, e^{-\frac{(w-\mu_k)^2}{2\sigma_k^2}} \times \int_{-\infty}^{\alpha_k \frac{(w-\mu_k)}{\sigma_k}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, dx \\ &= 2\,\phi\left(\frac{w-\mu_k}{\sigma_k}\right) \Phi\left(\frac{\alpha_k(w-\mu_k)}{\sigma_k}\right), \qquad w, \mu_k, \alpha_k \in \mathbb{R}\,, \quad \sigma_k \in \mathbb{R}^+, \end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denotes the pdf and the cumulative distribution function (cdf) of the standard Normal distribution, respectively [14, 19, 20].

When $\alpha_k = 0$, the above formula recreates the pdf of the Normal distribution. In this case, the Fisher information matrix of the Skew-Normal is singular, thus, influencing the asymptotic properties of the maximum likelihood estimators in the vicinity of zero. A detailed discussion about this topic can be found elsewhere [21, 22, 23].

When $\alpha_k = \infty$, the limiting distribution is the half-Normal distribution [21]. In this case, the location parameter μ_k determines the support of the distribution. This property makes the Skew-Normal distribution particularly useful to model data with a lower or an upper bound.

Note that the Skew-Normal distribution can be obtained from (3.2) when $H(\cdot\,;\theta_k)$ is a degenerate mixing distribution. Alternatively, the Skew-Normal distribution is a special case of the skew Normal-Normal [24] and the skew Student-t-Normal distribution [25]. These two flexible distributions are members of the so-called skew scale mixtures of Normal distributions [25]. This class of probability distributions differs from the class of SMSN in terms of the respective stochastic representation and dependence between skewness and kurtosis coefficients; see Ferreira *et al.* [25] for more details. In theory, distributions from this class can be seen as alternative candidates to the SMSN ones for the choice of mixing distributions. However, in practice, there are no estimation algorithms available for the context of finite mixture models.

3.1.2. Skew-t as a mixing distribution

Let Z_k be a random variable that follows a Skew-t distribution with location parameter μ_k, scale parameter σ_k^2, skewness parameter α_k, and v_k degrees of freedom. Then, its pdf is given by

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

where $d = (z-\mu_k)/\sigma_k$, $A = \alpha_k(z-\mu_k)/\sigma_k$, $t(\cdot\,;v)$ and $T(\cdot\,;v)$ are the pdf and cdf of the standard Student's t distribution with v degrees of freedom, respectively [14].

When $\alpha_k = 0$, the above distribution converts to the Generalized Student's t-distribution with location parameter μ_k, scale parameter σ_k and v_k degrees of freedom. When $v_k = 1$, one obtained the Skew-Cauchy distribution. Finally, when the degrees of freedom v_k tend to infinity, one obtains the Skew-Normal as the limiting distribution [14, 19, 20].

Note that the Skew-t distribution can be derived from (3.2) when U_k is a Gamma distribution with parameters $\alpha = v_k/2$ and $\beta = v_k/2$ [14]. As an additional note, Theodossiou [26] introduced the skew generalized t distribution with five parameters: location, scale, skewness, and two shape parameters. It can be derived from a ratio between a generalized gamma distribution and an appropriate transformation of a skew exponential power distribution, but it cannot be expressed as an SMSN distribution. As such, this alternative distribution has different skewness and kurtosis when compared to the above Skew-t distribution. See Arslan and Genç [27] and the references therein for more information.

3.2. Estimation of Skew-Normal and Skew-t mixture models

Let $X_1, ..., X_n$ be a random sample that represents the measured antibody levels in n individuals. In general, it is difficult to determine the maximum likelihood (ML) estimates of a finite mixture model by direct maximization of the log-likelihood function. To overcome

this problem, one can use the Expectation-Maximization (EM) algorithm given that the latent serological status of each individual is unknown and, thus, serological data are incomplete in that sense.

An EM-type algorithm for estimating SMSN mixture models is fully described elsewhere [14]. Briefly, the E-step is the same as in Gaussian mixture models, which has been largely studied in the literature [14, 17, 28]. Replacing the classical M-step with a sequence of conditional maximization steps (CM-steps), one obtains closed form expressions for the parameter estimates and the Fisher's information matrix [25]. To ensure convergence to the global maximum of the likelihood function, one should initiate the algorithm with different values for the parameter estimates. The final parameter estimates should be the ones that provide the highest value of the log-likelihood among all the different runs of the algorithm. Note that, for Gaussian mixture models, there are modifications of the classical EM algorithm that do not require the use of initial conditions and jointly determine the optimal number of the mixture components [29, 30]. These characteristics of the proposed algorithms reduces the computational time of analyses including a large number of screened antibodies. However, similar modifications remain to be done for the context of SMSN mixture models.

To obtain confidence intervals (CIs) for the model parameters, one can simply use the Wald's CIs. In the case of skewness parameters α_k's, the respective CIs are given by

$$\hat{\alpha}_k \pm \Phi^{-1}_{(\gamma+1)/2}\, se(\hat{\alpha}_k)\,,$$

where $\hat{\alpha}_k$ is the ML estimate of α_k, γ is the confidence level, and $\Phi^{-1}_{(\gamma+1)/2}$ is the probit function evaluated at $(\gamma+1)/2$. However, according to Zeller *et al.* [40], Wald's intervals for these parameters tend to inflate the underlying uncertainty in the case of a single Skew-Normal distribution. Such inflation can be derived from a poor quadratic approximation of the profile likelihood (PL) taken as a function of α [41]; see Pawitan for a more general discussion [42]. In addition, the PL is expected to show an inflexion point at $\alpha = 0$, which affects the asymptotic normal approximation for the distribution of the respective ML estimator [21]. Similar argument is expected to hold when estimating the same parameter of a single Skew-t distribution. In these cases, the PL can be used to determine a more accurate CI for α:

$$2\left\{l(\hat{\alpha}) - l(\alpha)\right\} < \chi^2_{\gamma,1}\,,$$

where $\hat{\alpha}$ is the ML estimate of α, $l(\alpha)$ is the PL taken as a function of α, and $\chi^2_{\gamma,1}$ is the γ quantile of the χ^2 distribution with one degree of freedom. See Zeller *et al.* [40] and Montenegro *et al.* [41] for the application of this CI to non-serological data. In the context of SMSN finite models, the PL approach is not a viable solution due to the presence of different subpopulations with their own skewness parameter.

3.3. Model selection

Model selection aims to determine the best mixture model for the data in terms of the number of the constituent components, g, and the respective mixing distributions. With this purpose, one can use information criteria based on penalized forms of the log-likelihood function: the Akaike's Information Criterion (AIC) [31], the Integrated Complete Likelihood (ICL) [32], the Bayesian Information Criterion (BIC) [33] and its modified versions [34, 35].

However, AIC tends to overestimate g in Gaussian mixture models even when n is very large [36]. This overestimation can be explained by a weak penalization of AIC to complex models with spurious mixing components that can arise from unbounded likelihood functions or from the presence of multiple local maximizers of the log-likelihood function [37]. In the case of serological applications, the overestimation of g compromises interpretability of a mixture model with more than 2 components [13]. In contrast, ICL tends to underestimate g and it is more adequate when the mixture components are well separated [32]. Finally, In this regard, BIC offers a higher penalization of models with a higher components when compared to AIC. However, the regularity conditions for using BIC do not necessarily hold in analysing finite mixture models [33, 35]. However, simulation studies suggested a satisfactory performance of this criterion (or its modified versions) in determining the true number of Gaussian mixture components [29, 35]. Therefore, at this stage, BIC seems the recommended measure when comparing different mixture models. Simulation studies should be conducted in the future to confirm this recommendation.

To complement the analysis based on information criteria, one can also carry out the likelihood ratio test (LRT) for determining the optimal number of mixture components, g [4]. However, the regularity conditions for the asymptotic χ^2 approximation of the test statistic are not met in finite mixture models, because the null hypothesis is specified in the boundary of the parameter space [4]. To overcome this problem, one can use a parametric Bootstrap approach to estimate the p-value of this non-standard LRT [38, 39], as described below.

Consider the test for confronting $H_0 : g = g_0$ versus $H_1 : g = g_1$ where $g_0 < g_1$. Let $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_1$ be the parameter vectors of the mixture models under H_0 and H_1, respectively; $\boldsymbol{x} = (x_1, ..., x_n)$ the observed data and $T(\boldsymbol{x}; \boldsymbol{\psi}_0, \boldsymbol{\psi}_1)$ the test statistic of LRT. The bootstrap approach is given by the following algorithm [39]:

1. Use the EM algorithm to estimate the $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_1$ estimates under the H_0 and H_1 hypotheses, respectively. Calculate $T(\boldsymbol{x}; \hat{\boldsymbol{\psi}}_0, \hat{\boldsymbol{\psi}}_1)$;
2. Simulate $N = 10,000$ independent samples $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$ using the mixture model under H_0 and parameterized by $\hat{\boldsymbol{\psi}}_0$;
3. For each bootstrap sample i, calculate $T(\boldsymbol{x}_i^*; \hat{\boldsymbol{\psi}}_{0_i}, \hat{\boldsymbol{\psi}}_{1_i})$, where $\hat{\boldsymbol{\psi}}_{0_i}$ and $\hat{\boldsymbol{\psi}}_{1_i}$ are the estimated parameter vectors for the bootstrap sample i under the H_0 and H_1 hypotheses, respectively;
4. Estimate the p-value as $\frac{1}{N} \sum_{i=1}^{N} I\left\{T(\boldsymbol{x}_i^*; \hat{\boldsymbol{\psi}}_{0_i}, \hat{\boldsymbol{\psi}}_{1_i}) > T(\boldsymbol{x}; \hat{\boldsymbol{\psi}}_0, \hat{\boldsymbol{\psi}}_1)\right\}$, where $I\{\cdot\}$ is the indicator function.

Finally, the estimated models should be assessed in terms of their goodness of fit. For a matter of simplicity, one can simply used the Pearson's χ^2 test [43, 44]. To apply this test, one can divide the data into bins defined by the respective 5%-quantiles or deciles. Alternatively, one can use the Kolmogorov-Smirnov, Anderson-Darling, and Walton's test among others [45].

3.4. Estimation of seroprevalence

After determining the best finite mixture model for the data, the next step of the analysis is usually to estimate the seroprevalence, that is, the prevalence of antibody-positive individuals in the population (or, the probability of an individual being antibody-positive).

Seropositivity is traditionally defined by a cutoff, denoted by c, in the respective antibody distribution above which individuals would be considered seropositive. In the context of finite mixture models, cutoff determination requires the interpretation of each latent population in terms of seronegativity and seropositivity. To do that, one typically assumes the seronegative population as the one with lowest average while the remaining components are interpreted as different levels of seropositivity upon recurrent infections. In this scenario, the seropositivity of i-th individual can be seen as resulting from a Bernoulli random variable $Y_i \sim \text{Ber}(p)$ where $p = P[X_i \geq c]$ and X_i $(i = 1, ..., n)$ represents the random variable representing the underlying antibody concentration. The probability p is also called seroprevalence and it embodies the probability of exposed individuals to a given antigen in the population. According to the maximum likelihood method, seroprevalence can be estimated as the proportion of seropositive individuals in the sample. Therefore, different estimates for the seroprevalence can be obtained according to the methods used to determine the cutoff.

In this work, we consider the following three different methods for determining the seropositivity cutoff:

- **Method 1:** It is based on the 99.9%-quantile associated with the estimated seronegative population. This method is the most popular in sero-epidemiology [13, 46]. It is often called as the 3σ rule, because the 99.9%-quantile is given by the mean plus 3 times the standard deviation of a normally distributed seronegative population;

- **Method 2:** It relies on the minimum of the density mixture functions. In the case of two latent populations, the cutoff corresponds to the absolute minimum, and in the case of three or more latent populations the cutoff corresponds to the lowest relative minimum. This point can be calculated using the Dekker's algorithm [47]. It should be noted that the minimum of the mixing function is not expected to coincide with the point of intersection of the probability densities of each individual subpopulation;

- **Method 3:** It imposes a threshold in the the so-called conditional classification curves [13]. Under the assumption that all components but the first one refer to seropositive individuals, the conditional classification curve of seropositive individuals given the antibody level x is defined as

$$p_{+|x} = \frac{\sum_{k=2}^{g} \pi_k \, f_k(x; \boldsymbol{\theta}_k)}{\sum_{k=1}^{g} \pi_k \, f_k(x; \boldsymbol{\theta}_k)} \, .$$

  In turn, the classification curve of seronegative individuals is given by

$$p_{-|x} = 1 - p_{+|x} \, .$$

  After calculating these curves, one can impose a minimum value for the classification of each individual. In this case, two cut-off values arise in the antibody distribution, one for the seronegative individuals and another for seropositive individuals. Mathematically, the classification rule is given as follows

$$C_i = \begin{cases} \text{seronegative} \, , & \text{if} \;\; x_i \leq c_- \, , \\ \text{equivocal} \, , & \text{if} \;\; c_- < x_i < c_+ \, , \\ \text{seropositive} \, , & \text{if} \;\; x_i \geq c_+ \, , \end{cases}$$

  where c_- and c_+ are the cutoff values in the antibody distribution that ensure a minimum classification probability, say 90%. To calculate these cutoff values in practice, one can use the bisection method providing an initial interval where they might be located [13].

Note that the cutoff values based on the above methods are dependent on the data under analysis and, therefore, they should be seen as random realizations of the respective estimator distributions. In other words, they have some uncertainty associated with them due to random sampling. However, this uncertainty is typically neglected in serological data analysis. This topic will be discussed elsewhere in the near future.

3.5. R packages

We used the package `mixsmsn` to fit different SMSN mixture models [48]. In the EM algorithm, the tolerance value for the norm of the difference between parameter estimates from two consecutive iterations was 10^{-5} with a maximum of 10,000 iterations. For each model and antibody under analysis, the EM algorithm was started with 100 random initial guesses for the parameter estimates. The reported estimates were the ones that led to the maximum of the likelihood function among all the runs of the algorithm. For fitting the Generalized Student's t-distribution, we considered the R package `extraDistr` [49], namely, the functions `dlst` and `plst` to calculate its pdf and cdf, respectively. The estimation of the Skew-Normal and Skew-t distributions was done in the package `sn` [50]. See Appendix B in the Supplementary Material for a detailed discussion about the computational costs of the proposed methodology.

4. RESULTS

Serological data refer to positive quantities bounded by an upper limit of detection. In theory, the Skew-Normal or the Skew-t distributions can describe bounded data by setting the respective skewness parameter close to infinity. However, this situation reduces model flexibility by forcing the analysis to be done with SMSN mixture models composed of highly asymmetric mixing distributions. Besides that, it is possible to obtain a good fit of the Gaussian mixture models to serological data after a data transformation [7]. To avoid reducing model flexibility while checking the appropriateness of Gaussian normal models, we applied the logarithmic transformation to the data. For an intuitive interpretation of the resulting data, we used the base 10 logarithmic transformation.

4.1. Exploratory data analysis

In this preliminary data analysis, we aimed to demonstrate the necessity of using alternative mixture models beyond the ones based on the Normal distribution. For this purpose, we partitioned each data set according to the cutoff values suggested by the manufacturers of the commercial kits (see Section 2). We assumed that antibody values below and above these values reflected somehow the distributions of the seronegative and seropositive populations, respectively. We then calculated the empirical skewness and excess kurtosis coefficients in each subset of data (Supplementary Table 2). Note that negative and positive estimates of the excess kurtosis indicated distributions with lighter or heavier tails than the Normal distribution, respectively.

As expected, the putative seropositive populations tended to have a skewness close to zero (HHV-6 and HSV-2) or a negative skewness (CMV, EBV, HSV-1, and VZV) of the respective antibody distribution. Similar evidence could be taken by a visual inspection of the histograms of the data (Figure 1A and B). The empirical estimates of the excess kurtosis were in most cases negative, which suggested distributions with lighter tails than the Normal distribution. However, these negative estimates might have simply resulted from dividing the data into two parts, and such a division limits the "size" of the tails associated with each serological population.

With respect to the putative seronegative populations, the skewness estimates were close to zero in the case of CMV, HSV-1, and HSV-2. For the remaining cases (EBV, HHV-6, and VZV), the skewness estimates were unexpectedly negative. The estimates of the excess kurtosis suggested similar weights of the tails for HSV-2 and VZV. For the remaining, the tails seemed to be lighter or heavier than the Normal distribution.

Finally, there was no evidence based on skewness and excess kurtosis alone for an antibody distribution in which both the seronegative and seropositive populations were similar to the Normal distribution. This suggested the necessity of considering finite mixture models based on families of probability distributions, such as the Skew-Normal or Skew-t, in which skewness and the weight of tails can be modelled appropriately.

4.2. Serological data analysis using Skew-Normal and Skew-t mixture models

To avoid selecting mixture models with difficult biological interpretation due to a high number of components g, we restricted our analysis to models with $g = 1$ (data exclusively composed of a single population, seronegative or seropositive), $g = 2$ (presence of both seronegative and seropositive populations), and $g = 3$. When fitting the Skew-t mixture models, the package `mixsmsn` only allowed to estimate models with the same degree of freedom for all the mixing distributions (*i.e.*, $v_1 = \cdots = v_g = v$).

Before fitting different SMSN mixture models, we first conducted a preliminary analysis based on Gaussian mixture models. In this analysis, we applied an alternative EM algorithm in which there was no need for setting initial values for the parameter estimates while simultaneously determining the optimal number of the components, $\hat{g}$ [30]. The criterion for determining $\hat{g}$ was the maximization of the likelihood function penalized by entropy. For the antibodies against EBV, HSV-2 and VZV viruses, the best Gaussian mixture models were composed of two serological populations. These populations could be interpreted as putative seropositive and seronegative populations. For the remaining antibodies, the best models suggested the presence of three serological populations in the respective data. In this case, the biological interpretation of the respective serological populations is not straightforward, as discussed elsewhere [13].

When compared to our preliminary analysis, the best SMSN mixture models according to BIC tended to require a lesser number of components. In particular, antibodies could be divided into three major classes:

(**i**) antibodies against HHV-6 and VZV in which data suggested the presence of a single serological population (Table 1 and Figure 1A);

(**ii**) antibodies against CMV, EBV, and HSV-2 for which there was evidence for two serological populations (Table 2 and Figure 1B);

(**iii**) antibodies against HSV-2 in which the optimal mixture model is composed of three serological populations (Table 2 and Figure 1B).

Table 1: Analysis of antibody data with evidence for a single serological population, where g represents the number of serological populations, p is the respective number of model parameters, $\mathcal{L}_{\max}$ is the value of the maximized log-likelihood function, p_{gof} is the maximum p-value for the goodness-of-fit test when dividing data into deciles or 5%-quantiles, and p_{boot} is the Bootstrap p-value for testing H_0: $g = 1$ versus H_1: $g = 2$. Best models according to BIC and the goodness-of-fit tests are written in bold.

Virus	**SMSN**	g	p	$\mathcal{L}_{\max}$	**AIC**	**BIC**	p_{gof}	p_{boot}
HHV-6	Normal	1	2	−129.46	263.00	270.94	0.064	0.064
		2	5	−116.97	244.13	263.97	0.169	
		3	8	−110.43	241.51	268.91	0.462	
	Skew-Normal	**1**	**3**	**−121.35**	**248.80**	**260.71**	**0.140**	**0.027**
		2	7	−117.35	249.03	276.75	0.084	
		3	11	−109.40	241.22	284.87	0.152	
	Student's t	1	3	−124.38	254.86	266.77	0.157	0.042
		2	6	−117.14	246.55	270.32	0.122	
		3	9	−105.36	229.06	264.78	0.254	
	Skew-t	1	4	−118.81	245.78	261.65	0.148	0.409
		2	8	−116.83	253.54	281.71	0.076	
		3	12	−104.00	234.73	282.36	0.001	
VZV	Normal	1	2	−108.76	221.58	229.53	< 0.001	0.000
		2	5	−7.28	24.72	44.60	0.159	
		3	8	−1.70	19.95	51.45	0.153	
	Skew-Normal	1	3	−23.94	53.99	65.90	< 0.001	0.180
		2	7	−0.11	14.69	42.27	0.406	
		3	11	0.10	16.87	65.87	0.068	
	Student's t	1	3	−61.90	129.88	141.80	< 0.001	0.000
		2	6	−7.41	26.99	50.86	0.082	
		3	9	−1.68	21.98	57.42	0.113	
	Skew-t	**1**	**4**	**−7.89**	**24.29**	**39.81**	**0.076**	**0.375**
		2	8	−0.05	16.76	48.16	0.211	
		3	12	5.47	25.31	62.14	0.134	

Data of antibodies against HHV-6 and VZV were best described by the Skew-Normal and the Skew-t distributions, respectively. The estimated distributions showed left asymmetry (Figure 1A) with the respective skewness parameter estimated at −1.87 and −5.14 for HHV-6 and VZV datasets, respectively. Accordingly, the Wald's and the PL 95%s CIs provided negative values for this parameter in the case of the HHV6 data: (−2.44; −1.02) and (−2.57; −1.25), respectively. In this case, the likelihood ratio based on the PL can be roughly approximated by a quadratic function, and, therefore, these two CIs did not substantially differ from each other (Figure 2A). According to the theoretical findings of Chiogna [21], this function showed an inflexion point at $\alpha = 0$. At the level of 5%, there was evidence for a single Skew-Normal against a mixture of two Skew-Normal distributions ($p_{\text{boot}} = 0.027$).

In the case of VZV antibody data, the Wald's and the PL 95% CIs also agreed in terms of a negative skew: $(-6.94; -2.14)$ and $(-8.00; -3.32)$, respectively. However, the likelihood ratio based on the profile likelihood was far from a quadratic function and, therefore, the Wald's CI is not expected to produce reliable results for these data. Finally, there was strong evidence for a single Skew-t distribution compared to a mixture of two Skew-t distributions ($p_{\text{boot}} = 0.375$).

Figure 1: Carry forward and percentage change indices.
Both indices tend to approximate in the months with less prices.

In terms of the respective serological interpretation, a single population for antibodies against HHV-6 and VZV is consistent with a seropositive population, given that HHV-6 and VZV are usually acquired during childhood, and more than 95% of the adult populations show the presence of antibodies against these viruses [51]. In addition, the core values of these distributions are higher than the cutoff for seropositivity suggested by the lab protocol. Finally, a left skewness is also predicted for a hypothetical seropositive population because the antibodies should decay over time in the absence of repeated infections [8].

Table 2: Analysis of antibody data with evidence for more than one serological population. See Table 1 for further details.

Virus	**SMSN**	g	p	$\mathcal{L}_{\max}$	**AIC**	**BIC**	p_{gof}
CMV	Normal	1	2	−409.11	822.29	830.24	<0.001
		2	5	−245.75	501.66	521.54	0.016
		3	8	−233.70	483.64	515.45	0.018
	Skew-Normal	1	3	−357.61	721.30	733.23	<0.001
		2	7	−233.82	482.66	509.69	0.038
		3	11	−226.64	489.78	519.35	0.146
	Student's t	1	3	−410.14	826.36	838.29	<0.001
		2	6	−238.54	489.27	513.12	0.038
		3	9	−231.23	480.81	516.59	0.046
	Skew-t	1	4	−357.71	723.55	739.45	<0.001
		2	**8**	**−231.55**	**479.34**	**511.45**	**0.072**
		3	12	−226.93	478.22	525.93	0.324
EBV	Normal	1	2	−342.30	688.67	696.62	<0.001
		2	5	−152.66	315.48	335.36	<0.001
		3	8	−129.30	274.84	306.65	0.173
	Skew-Normal	1	3	−226.42	458.93	470.86	<0.001
		2	7	−130.57	275.34	303.17	0.084
		3	11	−128.02	278.51	322.10	0.054
	Student's t	1	3	−240.21	486.50	498.43	<0.001
		2	6	−151.61	315.39	339.26	<0.001
		3	9	−129.41	277.09	312.88	0.117
	Skew-t	1	4	−173.14	354.40	370.31	<0.001
		2	**8**	**−125.63**	**267.65**	**299.32**	**0.248**
		3	12	−126.29	280.61	324.66	0.087
HSV-1	Normal	1	2	−442.27	888.61	896.56	<0.001
		2	5	−291.59	593.34	613.22	<0.001
		3	8	−264.94	546.14	577.94	0.003
	Skew-Normal	1	3	−394.55	806.62	807.11	<0.001
		2	7	−260.74	538.10	563.52	0.003
		3	**11**	**−252.32**	**527.39**	**570.70**	**0.104**
	Student's t	1	3	−443.73	893.55	905.48	<0.001
		2	7	−291.73	595.65	619.51	<0.001
		3	9	−264.98	548.23	584.02	0.002
	Skew-t	1	4	−395.43	812.55	814.88	<0.001
		2	8	−260.88	541.64	569.82	0.001
		3	12	−251.86	528.84	575.79	<0.001
HSV-2	**Normal**	1	2	−427.29	858.63	866.59	<0.001
		2	**5**	**−277.62**	**565.39**	**585.27**	**0.516**
		3	8	−269.24	565.92	586.54	0.007
	Skew-Normal	1	3	−337.36	684.60	692.74	<0.001
		2	7	−264.32	544.79	570.68	0.013
		3	11	−257.19	550.71	580.45	0.003
	Student's t	1	3	−428.40	862.88	874.81	<0.001
		2	6	−277.84	567.85	591.71	0.688
		3	9	−269.60	557.52	593.26	0.004
	Skew-t	1	4	−337.79	687.68	699.60	<0.001
		2	8	−264.52	547.40	577.10	0.007
		3	12	−257.38	562.77	586.83	0.001

Note that most of the SMSN mixture models could also provide a good fitting of the data of these two antibodies. This is the case of the mixture of two or three Normal distributions ($p_{\text{gof}} = 0.169$ and 0.462 for antibodies against HHV-6 and $p_{\text{gof}} = 0.159$ and 0.153), which are typically used in serological data analysis. Therefore, although not being the best models for HHV-6 and VZV-related antibodies, these models could have been used for subsequent serological analyses.

For the remaining antibodies, the respective data analysis was not straightforward because the model with lowest BIC estimate could not fit the data well according to the Pearson's goodness-of-fit test at 5% significance level (Table 2). This occurred for the mixtures of two Skew-Normal distributions for the antibodies against CMV (BIC = 509.69 and $p_{\text{gof}} = 0.038$), HSV-1 (BIC = 563.52 and $p_{\text{gof}} = 0.003$), and HSV-2 (BIC = 570.68 and $p_{\text{gof}} = 0.013$). For these antibodies, the best models were considered to be a mixture of two Skew-t distributions (BIC = 511.45 and $p_{\text{gof}} = 0.072$), a mixture of three Skew-Normal distributions (BIC = 570.70 and $p_{\text{gof}} = 0.104$), and a mixture of two Normal distributions (BIC = 585.27 and $p_{\text{gof}} = 0.516$), respectively, because they were the first models ranked by BIC with a good fit for the data (Figure 1B). Interestingly, for the HSV-2-related antibody data, when the mixture of two Normal distributions was compared to the mixture of two Skew-Normal distribution by a likelihood ratio test, the first model was strongly rejected ($p < 0.0001$), which suggested the asymmetry of at least one of the components. This inconsistency between this test and the selected model can be explained by the unavailability of fitting a mixture of a Normal distribution and a Skew-Normal distribution in the package `smsn`. For the antibody against EBV, the best model was a mixture of two Skew-t distributions, which also had a good fit for the data (BIC = 299.32 and $p_{\text{gof}} = 0.248$; Figure 1B).

With respect to the biological interpretation of each component, there was evidence of putative seronegative and seropositive populations for antibodies against CMV, EBV, and HSV-2 (Figure 1B). This interpretation was supported by the observation that the cutoff value suggested by the commercial kits lies between these hypothetical serological populations. In the case of antibodies against HSV-1, the respective interpretation was not so obvious, because

(**i**) the best mixture model was composed of three components and

(**ii**) the cutoff suggested by the commercial kits lies in the middle of the intermediate distribution, which shows right asymmetry.

In theory, the distribution of a putative seronegative population is expected to have right asymmetry [8] and, if so, this intermediate component should be interpreted accordingly. However, one cannot rule out that there are two seronegative populations resulting from distinct background signals in the absence of antibodies. Without additional information about the serological data, this intermediate component was considered to represent a putative seronegative population.

Finally, we performed a similar model selection using AIC instead. Again, we selected the best models with the lowest AIC estimates and with a good fit to the data ($p_{\text{gof}} > 0.05$) at the same time. In contrast with BIC results, this alternative model selection could not provide evidence for a single serological population in the data of HHV-6 and VZV (Table 1).

In these two cases of HHV-6, the best models were mixtures of three Generalized Student-t distributions (AIC = 229.06 and $p_{\text{gof}} = 0.254$) and of two Skew-Normal distributions (AIC = 14.69 and $p_{\text{gof}} = 0.406$), respectively. For antibody against CMV, the best model was a mixture of three Skew-t distributions (AIC = 478.2 and $p_{\text{gof}} = 0.32$), which reflected an increase in the number of components compared to model selection using BIC (Table 2). For the remaining antibodies, it was selected the same model (Table 2). In summary, AIC tended to select models with an increased number of components required to explain the data of each antibody.

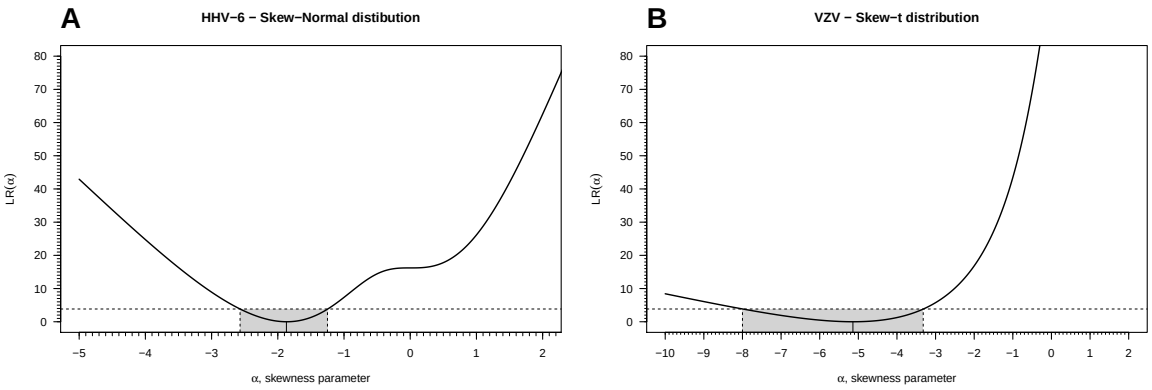

Figure 2: Likelihood ratio (LR) based on the profile likelihood as a function of the skew parameter α, when fitting the Skew-Normal and Skew-t distributions to HHV-6 (A) and VZV (B) data, respectively. The horizontal dashed lines represent the 95% quantile of a χ^2 distribution with one degree of freedom. The grey rectangles represent the 95% CI for α according to this method.

4.3. Estimation of cutoff for seropositivity

After fitting the mixture models to the data, the following step of the analysis was to estimate a cutoff value for seropositivity and the subsequent seroprevalence in the different study groups (Table 3).

For CMV and HSV-2 antibody data, the cutoff values did not vary substantially from one method to another. Interesting, the cutoff values estimated by method 1 (the 3σ rule) almost perfectly matched with the ones suggested by the commercial kits (12.6 U/ml and 12.0 U/ml for CMV and HSV-2 respectively versus 12.0). This good matching between estimates could be explained by a good approximation of the Normal distribution for the seronegative population (Figure 1B) and, therefore, we could infer that the cutoff value suggested by the commercial kits was derived from the 3σ rule; this information was absent from the original study [15]. Since the seronegative and seropositive populations were separated well in these antibody distributions, the seroprevalence estimates across the different study groups were almost invariant with respect to the cutoff value used.

With respect to the EBV antibody data, the hypothetical seronegative population is asymmetric to the right ($\alpha_1 = 1.74$; 95% CI = $(-1.30; 4.80)$; bootstrap 95% CI = $(0.04; 7.90)$; Figure 1B) with heavy tails ($v = 4.52$; 95% CI = $(0.79; 8.26)$; bootstrap 95% CI = $(3.00; 14.88)$).

Consequently, the cutoff value of 249.5 U/ml derived from method 1 was quite different from the one suggested by the commercial kit. However, this cutoff value was considered non-informative because it was well located within the seropositive population and implied seroprevalence estimates close to zero for the different study groups. In contrast, the cutoff values from the remaining methods were in the same order of magnitude of the one suggested by the commercial kits. Therefore, the subsequent seroprevalence estimates of each study group did not differ substantially among these methods. Again, the consistency of the resulting seroprevalence estimates was due to the fact that the seronegative and seropositive populations were well separated in these data.

Table 3: Seroprevalence (%) by cutoff method for seropositivity and by study group. c_- and c_+ are on the linear scale (U/ml). Seroprevalence was calculated based on c_+. The method denoted by "M" refers to the cutoff suggested by the protocol of the commercial kit. The confidence intervals (CI) refer to the Clopper-Pearson exact confidence interval for a proportion.

Virus	Method	c_-	c_+	**Seroprevalence (95% CI)**				
				Global	**HC**	**ME-M**	**ME-S**	**MS**
CMV	M	8.0	12.0	33.5 (28.9–38.4)	37.4 (28.2–47.3)	28.6 (22.4–35.4)	33.3 (21.1–47.5)	36.7 (23.4–51.7)
	1	—	12.6	33.5 (28.9–38.4)	37.4 (28.2–47.3)	28.6 (22.4–35.4)	33.3 (21.1–47.5)	36.7 (23.4–51.7)
	2	—	13.5	33.2 (28.6–38.1)	37.4 (28.2–47.3)	28.6 (22.4–35.4)	31.5 (19.5–45.6)	36.7 (23.4–51.7)
	3	9.4	14.1	32.9 (28.4–37.9)	37.4 (28.2–47.3)	28.1 (21.9–34.9)	31.5 (19.5–45.6)	36.7 (23.4–51.7)
EBV	M	8.0	12.0	87.3 (83.6–90.4)	87.9 (80.1–93.4)	86.2 (80.6–90.7)	81.5 (68.6–90.7)	75.5 (61.1–86.7)
	1	—	249.5	2.0 (0.09–3.9)	1.9 (0.02–6.6)	1.5 (0.03–4.4)	0.0 (0.0–6.6)	6.1 (1.3–16.9)
	2	—	11.5	87.3 (83.6–90.4)	87.9 (80.1–93.4)	86.2 (80.6–90.7)	81.5 (68.6–90.7)	75.5 (61.1–86.7)
	3	5.6	20.4	85.5 (81.7–88.9)	87.9 (80.1–93.4)	82.7 (76.6–87.7)	81.5 (68.6–90.7)	75.5 (61.1–75.5)
HSV-1	M	8.0	12.0	45.2 (40.2–50.2)	42.1 (32.6–51.9)	41.8 (34.8–49.1)	51.9 (37.8–65.6)	46.9 (32.5–61.7)
	1	—	271.0	0.0 (0.0–0.1)	0.0 (0.0–3.4)	0.0 (0.0–1.2)	0.0 (0.0–6.6)	0.0 (0.0–7.3)
	2	—	46.9	34.5 (29.8–39.4)	28.0 (19.8–37.5)	34.7 (28.1–41.8)	38.9 (25.9–53.1)	34.7 (21.7–49.6)
	3	42.7	83.2	30.7 (26.2–35.5)	24.3 (16.5–33.5)	32.1 (25.7–39.2)	33.3 (21.1–47.5)	28.6 (16.6–43.3)
HSV-2	M	8.0	12.0	38.1 (33.3–43.1)	33.6 (24.8–43.4)	38.8 (31.9–45.9)	40.7 (27.6–54.9)	32.7 (19.9–47.5)
	1	—	12.0	38.1 (33.3–43.1)	33.6 (24.8–43.4)	38.8 (31.9–45.9)	40.7 (27.6–54.9)	32.7 (19.9–47.5)
	2	—	10.7	38.8 (33.9–43.8)	33.6 (24.8–43.4)	39.3 (32.4–46.5)	40.7 (27.6–54.9)	36.7 (23.4–51.7)
	3	7.1	12.6	37.8 (33.0–42.8)	33.6 (24.8–43.4)	38.8 (31.9–45.9)	40.7 (27.6–54.9)	30.6 (18.3–45.4)

The largest differences in the cutoff values for seropositivity were observed for the HSV-1 antibody data. Coincidentally, this was the data set where the best mixture model was composed of three components. As discussed earlier in this paper, the intermediate component was considered a second hypothetical seronegative population, which resulted in a shift in the calculation of seropositivity towards higher values. As such, the cutoff seropositive based on the commercial kit led to the highest seroprevalence estimates for all study groups with a global estimate of 45.2% (95% CI = (40.2%; 50.2%). As an extreme case, the 3σ rule produced a too-high cutoff value again due to the right asymmetry of both seronegative populations. Such unrealistic cutoff value led to a zero seroprevalence estimates and rendered the respective analysis useless.

Finally, although not being the main objective of this study, the comparison of the four study groups suggested that, given a method for determining seropositivity and antibody under analysis, the seroprevalence of patients with ME/CFS did not appear to differ significantly from the one of healthy controls and patients with multiple sclerosis alike.

5. CONCLUSIONS

This study aimed to review the Skew-Normal and Skew-t mixture models and recommend their routine use in serological data analysis. Such recommendation sets its foundation in the high flexibility of these models in describing different data patterns, as illustrated with the data analysis of antibodies against 6 herpesviruses. In particular, high modelling flexibility is desirable given that right and left asymmetry can emerge from seronegative and seropositive populations, respectively. In this regard, most popular distributions used in Statistics are not able to exhibit either left or right asymmetry depending on the parameters specified. A less-known family of distributions that shows such stochastic property is the Generalized Tukey's λ distribution [54, 55]. This distribution offers a great variety of shapes owing to four parameters controlling the location, the scale, the skewness, and the flatness of the resulting distribution. However, the Generalized Tukey's λ distribution is only defined in terms of its quantile function and, hence, its estimation is cumbersome. This distribution has already been proposed for mixture modelling, but there are only theoretical and computational developments for the case of two components [52, 53]. This limits the application of these alternative models in data sets where there is evidence for more than two serological populations, such as the case of the antibodies against HSV-1 here analyzed or against the influenza virus reported elsewhere [11]. Therefore, Skew-Normal and Skew-t mixture models would appear the most general and flexible approach for analysing serological data.

For data analysis, we recommend using the package `mixsmsn` for estimating the finite mixture models [48]. Notwithstanding this recommendation, the package only estimates SMSN mixture models where all mixing distributions belong to the same family of SMSN probability distributions. Hence, it can only fit 4 different models per number of components. In theory, there are $4^2 = 16$ possible two-component mixture models resulting from the combination of Normal, Skew-Normal, Generalized Student's t, and Skew-t distributions as mixing distributions. Note that these possible models are nested in each other by imposing parametric restrictions to the most general mixture model based on the Skew-t distribution. For three-component mixture models, the number of possible models increases to $4^3 = 64$.

Therefore, the package `mixsmsn` excludes a vast number of possible models, which ultimately affects the detection of the most parsimonious model for the data; this model could be a combination of probability distributions from different families. The same limitation could also explain some inferential inconsistencies in the example of application. For instance, a single Skew-Normal distribution was considered the best model for the antibodies against HHV-6. However, the hypothesis of a single Skew-Normal distribution against a mixture of two Skew-Normal distributions could be rejected by bootstrap at the 5% significance level. A possible explanation for this statistical inconsistency is that the best model for these data could be a mixture of a Normal distribution for the seronegative population and a Skew-Normal distribution for the seropositive population. Therefore, there is a research opportunity to extend the package allowing each mixing component to be described by different families of SMSN distributions.

Another limitation of using `mixsmsn` package is that, for mathematical tractability, the mixtures of generalized Student t and Skew-t distributions were assumed to have the same degrees of freedom in all the mixing distributions. In theory, this assumption could be relaxed so this parameter could vary from one component of the mixture to another. This modelling option was available in the package `EMMIXuskew` for the mixture of Skew-t distributions [56]. However, this package is currently discontinued. In practice, we expect some degree of numerical instability when estimating different degrees of freedom in data where the serological populations overlap substantially with each other. In this regard, future research could be conducted to determine the stochastic and sampling conditions in which different degrees of freedom could infer from different components.

The problem of determining the optimal cutoff value for seropositivity has been intensively investigated, discussed, and revisited over the years [46, 57, 58, 59]. In this regard, the most popular cutoffs for seropositivity are simply defined by the mean plus a given number of times the standard deviation of the hypothetical seronegative population without checking the Normality assumption of the hypothetical seronegative population. The resulting cutoffs are associated with high-order quantiles of the Normal distribution, such as 97.7% or 99.9% for the 2σ and 3σ rules, respectively. In practice, these cutoffs imply a high specificity but show an arbitrary sensitivity for the respective serological classification. When the hypothetical seronegative population shows a right-skewed distribution, similar cutoffs can be obtained by calculating the same high quantiles of the estimated SMSN, as done here. The reverse argument can be made when analysing antibodies where seropositivity could be considered the default serological state of an individual, such as the case of antibodies against HHV-6 and VZV here analyzed or vaccine-related antibodies in populations where vaccination is mandatory. Similar cutoffs can be determined for these antibodies by the mean minus a given number of times the standard deviation of the hypothetical seropositive population assumed to be normally distributed. For a left-skewed seropositive population, the cutoff values for seropositivity are now calculated using the low order quantiles (*e.g.*, 2.3% and 0.1%-quantiles for the 2σ and 3σ rules, respectively). Inversely, these cutoffs generate a high sensitivity but an arbitrary specificity for the respective serological classification. It is worth noting that it is up to the analyst to decide on what she/he wants to control, whether specificity, sensitivity, or both with respect to the resulting serological classification. A similar decision problem occurs in analyses based on the Receiver Operating Characteristic curve. Given the multiplicity of criteria for estimating this cutoff and its uncertainty, several authors advocate a free-cutoff approach for serological analysis [6, 60]. However, a detailed discussion about the advantages and disadvantages of free-cutoff approaches was out of the scope of this study.

In summary, the mixture models based on Skew-Normal and Skew-t distributions show promise to become a routine tool for serological data analysis. They have the advantage of including the Gaussian mixture models as special cases. However, given the statistical complexity of these models and some inferential problems highlighted throughout the paper, their application should be done in a closer collaboration between biomedical researchers who generate the serological data and biostatisticians who have in principle the knowledge and skills to fit and compared these mode properly.

ACKNOWLEDGMENTS

The authors would like to thank two anonymous reviewers for their constructive comments and suggestions. The authors would also like to thank Eliana Lacerda, Luis Nacul, and Jackie-Cliff from the London School of Hygiene & Tropical Medicine (LSHTM) for sharing the data, and João Malato for helping to proof-read the manuscript. This work was partially funded by Fundação para a Ciência e a Tecnologia, Portugal (ref: UIDB/00006/2020 and UIDB/04561/2020) and resulted from a short-term scientific mission of TDS to the LSHTM, which was funded by the EUROMENE Cost Action (CA15111) from the European Union.

REFERENCES

[1] Wang, S.S.; Schiffman, M.; Shields, T.S.; Herrero, R.; Hildesheim, A.; Bratti, M.C.; Sherman, M.E.; Rodriguez, A.C.; Castle, P.E.; Morales, J.; Alfaro, M.; Wright, T.; Chen, S.; Clayman, B.; Burk, R.D. and Viscidi, R.P. (2003). Seroprevalence of human papillomavirus-16, -18, -31, and -45 in a population-based cohort of 10000 women in Costa Rica, *British Journal of Cancer*, **89**(7), 1248–1254.

[2] Cook, J.; Kleinschmidt, I.; Schwabe, C.; Nseng, G.; Bousema, T.; Corran, P.H.; Riley, E.M. and Drakeley, C.J. (2011). Serological markers suggest heterogeneity of effectiveness of malaria control interventions on Bioko Island, Equatorial Guinea, *Plos One*, **6**(9), e25137.

[3] Hsiang, M.S.; Hwang, J.; Kunene, S.; Drakeley, C.; Kandula, D.; Novotny, J.; Parizo, J.; Jensen, T.; Tong, M.; Kemere, J.; Dlamini, S.; Moonen, B.; Angov, E.; Dutta, S.; Ockenhouse, C.; Dorsey, G. and Greenhouse, B. (2012). Surveillance for malaria elimination in Swaziland: a national cross-sectional study using pooled PCR and serology, *PloS One*, **7**(1), e29550.

[4] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*, John Wiley & Sons, New York.

[5] Gay, N.J. (1996). Analysis of serological surveys using mixture models: application to a survey of parvovirus B19, *Statistics in Medicine*, **15**, 1567–1573.

[6] Chis Ster, I. (2012). Inference for serological surveys investigating past exposures to infections resulting in long-lasting immunity – an approach using finite mixture models with concomitant information, *Journal of Applied Statistics*, **39**(11), 2523–2542.

[7] Rogier, E.; Wiegand, R.; Moss, D.; Priest, J.; Angov, E.; Dutta, S.; Journel, I.; Jean, S.E.; Mace, K.; Chang, M.; Lemoine, J.F.; Udhayakumar, V. and Barnwell, J.W. (2015). Multiple comparisons analysis of serological data from an area of low Plasmodium falciparum transmission, *Malaria Journal*, **14**, 436.

[8] Parker, R.A.; Erdman, D.D. and Anderson, L.J. (1990). Use of mixture models in determining laboratory criterion for identification of seropositive individuals: application to parvovirus B19 serology, *Journal of Virological Methods*, **27**(2), 135–144.

[9] Baughman, A.L.; Bisgard, K.M.; Lynn, F. and Meade, B.D. (2006). Mixture model analysis for establishing a diagnostic cut-off point for pertussis antibody levels, *Statistics in Medicine*, **25**, 2994–3010.

[10] Rota, M.C.; Massari, M.; Gabutti, G.; Guido, M.; De Donno, A. and Ciofi degli Atti, M.L. (2008). Measles serological survey in the Italian population: interpretation of results using mixture model, *Vaccine*, **26**(34), 4403–4409.

[11] Nhat, N.; Todd, S.; de Bruin, E.; Thao, T.; Vy, N.; Quan, T.M.; Vinh, D.N.; van Beek, J.; Anh, P.H.; Lam, H.M.; Hung, E.; Lien, N.; Hong, T.; Farrar, J.; Simmons, C.P.; Chau, N.; Koopmans, M. and Boni, M.F. (2017). Structure of general-population antibody titer distributions to influenza A virus, *Scientific Reports*, **7**(1), 6060.

[12] Moreira da Silva, J.; Prata, S.; Domingues, T.D.; Leal, R.O.; Nunes, T.; Tavares, L.; Almeida, V.; Sepúlveda, N. and Gil, S. (2020). Detection and modeling of anti-Leptospira IgG prevalence in cats from Lisbon area and its correlation to retroviral infections, lifestyle, clinical and hematologic changes, *Veterinary and Animal Science*, **10**, 100144.

[13] Sepúlveda, N.; Stresman, G.; White, M.T. and Drakeley, C.J. (2015). Current mathematical models for analyzing anti-malarial antibody data with an eye to malaria elimination and eradication, *Journal of Immunology Research*, **2015**, 738030.

[14] 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.

[15] Cliff, J.M.; King, E.C.; Lee, J.S.; Sepúlveda, N.; Wolf, A.S.; Kingdon, C.; Bowman, E.; Dockrell, H.M.; Nacul, L.; Lacerda, E. and Riley, E.M. (2019). Cellular immune function in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS), *Frontiers in Immunology*, **10**, 796.

[16] Cortes Rivera, M.; Mastronardi, C.; Silva-Aldana, C.T.; Arcos-Burgos, M. and Lidbury, B.A. (2019). Myalgic encephalomyelitis/chronic fatigue syndrome: a comprehensive review, *Diagnostics*, **9**, 91.

[17] Lin, T.I.; Lee, J.C. and Yen, S.Y. (2007). Finite mixture modelling using the skew-normal distribution, *Statistica Sinica*, **17**, 909–927.

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

[19] Azzalini, A. (1985). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, **12**, 171–178.

[20] Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution, *Journal of the Royal Statistical Society B*, **65**, 367–389.

[21] Chiogna, M. (2005). A note on the asymptotic distribution of the maximum likelihood estimator for the scalar skew-normal distribution, *Statistical Methods & Applications*, **14**, 331–341.

[22] Hallin, M. and Ley, C. (2012). Skew-symmetric distributions and Fisher information – a tale of two densities, *Bernoulli*, **18**, 747–763.

[23] Hallin, M. and Ley, C. (2014). Skew-symmetric distributions and Fisher information: the double sin of the skew-normal, *Bernoulli*, **20**, 1432–1453.

[24] Gómez, H.W.; Venegas, O. and Bolfarine, H. (2007). Skew-symmetric distributions generated by the distribution function of the normal distribution, *Environmetrics*, **18**, 395–407.

[25] Ferreira, C.S.; Bolfarine, H. and Lachos, V.H. (2011). Skew scale mixtures of normal distributions: properties and estimation, *Statistical Methodology*, **8**, 154–171.

[26] Theodossiou, P. (1998). Financial data and the skewed generalized t distribution, *Management Science*, **44**, 1650–1661.

[27] Arslan, O. and Genç, A.I. (2009). The skew generalized t distribution as the scale mixture of a skew exponential power distribution and its applications in robust estimation, *Statistics*, **43**, 481–498.

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

[29] 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.

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

[31] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

[32] Biernacki, C.; Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 719–725.

[33] Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis, *The Computer Journal*, **41**, 578–588.

[34] Zhao, J.; Jin, L. and Shi, L. (2015). Mixture model selection via hierarchical BIC, *Computational Statistics & Data Analysis*, **88**, 139–153.

[35] Mehrjou, A.; Hosseini, R. and Araabi, B.N. (2016). Improved Bayesian information criterion for mixture model selection, *Pattern Recognition Letters*, **69**, 22–27.

[36] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions, *Psychometrika*, **52**, 345–370.

[37] Kim, Daeyoung and Seo, Byungtae (2014). Assessment of the number of components in Gaussian mixture models in the presence of multiple local maximizers, *Journal of Multivariate Analysis*, **125**, 100–120.

[38] Feng, Z.D. and McCullogh, C.E. (1996). Using bootstrap likelihood ratios in finite mixture models, *Journal of the Royal Statistical Society B*, **58**, 609–617.

[39] Yu, Y. and Harvill, J.L. (2019). Bootstrap likelihood ratio test for Weibull mixture models fitted to grouped data, *Communications in Statistics – Theory and Methods*, **48**(18), 4550–4568.

[40] Zeller, C.B.; Labra, F.V.; Lachos, V.H. and Balakrishnan, N. (2010). Influence analyses of skew-normal/independent linear mixed models, *Computational Statistics & Data Analysis*, **54**, 1266–1280.

[41] Montenegro, L.C.; Lachos, V.H. and Bolfarine, H. (2010). Inference for a skew extension of the Grubbs model, *Statistical Papers*, **51**, 701–715.

[42] Pawitan, Y. (2000). A reminder of the fallibility of the Wald statistic: likelihood explanation, The American Statistician, **54**, 54–56.

[43] Bunge, J. and Barger, K. (2008). Parametric models for estimating the number of classes, *Biometrical Journal*, **50**, 971–982.

[44] Ultsch, A.; Thrun, M.; Hansen-Goos, O. and Lötsch, J. (2015). Identification of molecular fingerprints in human heat pain thresholds by use of an interactive mixture model R toolbox (AdaptGauss), *International Journal of Molecular Sciences*, **16**, 25897–25911.

[45] Wichitchan, S.; Yao, W. and Yang, G. (2019). Hypothesis testing for finite mixture models, *Computational Statistics & Data Analysis*, **132**, 180–189.

[46] Saraswati, K.; Phanichkrivalkosil, M.; Day, N. and Blacksell, S.D. (2019). The validity of diagnostic cut-offs for commercial and in-house scrub typhus IgM and IgG ELISAs: a review of the evidence, *PLoS Neglected Tropical Diseases*, **13**(2), e0007158.

[47] Brent, R.P. (1973). *Algorithms For Minimization Without Derivatives*, Prentice-Hall, Englewood Cliffs, New Jersey, 73–76.

[48] Prates, M.O.; Lachos, V.H. and Cabral, C. (2013). Fitting finite mixture of scale mixture of skew-normal distributions, *Journal of Statistical Software*, **54**, 1–20.

[49] Wolodzko, T. (2020). Additional Univariate and Multivariate Distributions, *R CRAN*, `https://cran.r-project.org/web/packages/extraDistr/index.html`

[50] Azzalini, A. (2020). The Skew-Normal and Related Distributions Such as the Skew-t, *R CRAN*, `https://cran.r-project.org/web/packages/sn/sn.pdf`

[51] Braun, D.K.; Dominguez, G. and Pellett, P.E. (1997). Human herpesvirus 6, *Clinical Microbiology Reviews*, **10**(3), 521–567.

[52] Su, S. (2007). Fitting single and mixture of generalized lambda distributions to data via discretized and maximum likelihood methods: GLDEX in R, *Journal of Statistical Software*, **21**(9), 1–17.

[53] Su, S. (2011). Maximum log likelihood estimation using EM algorithm and partition maximum log likelihood estimation for mixtures of generalized lambda distributions, *Journal of Modern Applied Statistical Methods*, **10**, 17.

[54] Ramberg, J. and Schmeiser, B. (1974). An approximate method for generating asymmetric random variables, *Communications of the Association for Computing Machinery*, **17**, 78–82.

[55] Freimer, M.; Mudholkar, G.; Kollia, G. and Lin, C. (1988). A study of the generalised Tukey lambda family, *Communications in Statistics – Theory and Methods*, **17**, 3547–3567.

[56] McLachlan, G. and Lee, S. (2013). EMMIXuskew: an R package for fitting mixtures of multivariate skew t distributions via the EM algorithm, *Journal of Statistical Software*, **55**(12), 1–22.

[57] Ridge, S.E. and Vizard, A.L. (1993). Determination of the optimal cutoff value for a serological assay: an example using the Johne's Absorbed EIA, *Journal of Clinical Microbiology*, **31**(5), 1256–1261.

[58] Kafatos, G.; Andrews, N.J.; McConway, K.J.; Maple, P.A.; Brown, K. and Farrington, C.P. (2016). Is it appropriate to use fixed assay cut-offs for estimating seroprevalence?, *Epidemiology and infection*, **144**(4), 887–895.

[59] Migchelsen, S.J.; Martin, D.L.; Southisombath, K.; Turyaguma, P.; Heggen, A.; Rubangakene, P.P.; Joof, H.; Makalo, P.; Cooley, G.; Gwyn, S.; Solomon, A.W.; Holland, M.J.; Courtright, P.; Willis, R.; Alexander, N.D.; Mabey, D.C. and Roberts, C.H. (2017). Defining seropositivity thresholds for use in trachoma elimination studies, *PLoS Neglected Tropical Diseases*, **11**(1), e0005230.

[60] Bouman, J.A.; Riou, J.; Bonhoeffer, S. and Regoes, R.R. (2021). Estimating the cumulative incidence of SARS-CoV-2 with imperfect serological tests: exploiting cutoff-free approaches, *PLoS Computational Biology*, **17**, e1008728.