Analysis of antibody data using Finite Mixture Models based on Scale Mixtures of Skew-Normal distributions

Tiago Dias Domingues

CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal

Helena Mouriño

CMAFcIO, Faculdade de Ciências, Universidade de Lisboa, Portugal

Nuno Sepúlveda

CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal

London School of Hygiene and Tropical Medicine, UK

March 2021

Abstract

Finite mixture models have been widely used in antibody (or serological) data analysis in order to help classifying individuals into either antibody-positive or antibody-negative. The most popular models are the so-called Gaussian mixture models which assume a Normal distribution for each component of a mixture. In this work, we propose the use of finite mixture models based on a flexible class of scale mixtures of Skew-Normal distributions for serological data analysis. These distributions are sufficiently flexible to describe right and left asymmetry often observed in the distributions associated with hypothetical antibody-negative and antibody-positive individuals, respectively. We illustrate the advantage of these alternative mixture models with a data set of 406 individuals in which antibodies against six different human herpesviruses were measured in the context of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome.

Key-Words: Finite mixture models; scale mixtures of Skew-Normal distributions; seropositivity; chronic fatigue syndrome.

1 INTRODUCTION

Antibodies are key immunological proteins produced by B cells upon molecular 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

comprise the basis of the so-called immunological memory, which translates into a quicker and more efficient immune response in the case of repeated exposure to the same infectious agent. In turn, autoantibodies are antibodies recognizing components from the body and they are usually present in autoimmunity diseases, such as multiple sclerosis or rheumatoid arthritis. In the laboratory, antibodies (or autoantibodies) against a specific antigen are usually quantified by the enzymatic-linked immunosorbent assays (ELISA) using serum samples; see ref. [1, 2, 3] for some recent studies using these assays. The respective read-out is a light intensity, also known as optical density, which can be 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 ELISA particularly suitable for large-scale sero-epidemiology surveys where one aims to estimate the prevalence of exposure to a given pathogen in the population [1, 4, 5]. With the recent development of high-throughput technologies, antibody quantification is currently shifting from the traditional ELISA to more advance assays, such as microarray [6, 7], luminex [8, 9], or cytometry bead assays [10], where a large number of different antibodies can be evaluated in the same serum sample. However, some of these promising technologies still require some degree of optimization before their widely applicability in biomedical research [11, 12].

Statistical analysis of antibody (or serological) data is usually carried out under the assumption that the antibody distribution consists of different latent populations, each one representing a distinct antibody state or different degrees of exposure to a given antigen. This assumption gives rise to a serological data analysis based on finite mixture models [13]. These models can be more or less complex depending on the number of components and mixing distributions used to describe the data. Due to its conceptual simplicity and easy of interpretation, the most popular finite mixture model in routine serological applications invokes the existence of two components related to hypothetical seronegative and seropositive individuals or, equivalently, antibody-negative and antibody-positive individuals [14, 15, 16]. Models comprising more than two components have also been found appropriate to describe data from some studies [17, 18, 19, 20, 21], but they might bring some ambiguity when interpreting which components are associated with antibody positivity [22]. In turn, the most popular choice for the mixing distributions is the Lognormal distribution in the original scale of the measurements or, equivalently, the Normal distribution after logarithmic transformation of the data [15, 17]. Gamma and Weibull are other choices for the mixing distributions among textbook probability distributions [16, 20]. Alternatively, less trivial mixture models can be used in the analysis. This is the case of a mixture between two truncated Normal distributions describing the situation where observations might fall below the lower limit of detection or above the upper limit of detection of the assay [18]. Another interesting model is a mixture between a Normal and a combination of half-Normal distributions for the hypothetical seronegative and seropositive

populations, respectively [14]. The rationale behind this proposal is that the distribution of the seropositive population should be left skewed, because antibody levels tend to decrease over time [17]. Notwithstanding their suitability to tackle specific characteristics of serological data, none of the above models would appear to provide sufficiently flexibility in terms of skewness and flatness of each mixing distribution that could be used as the basis of data analysis automation in high-throughput serological studies.

In this scenario, we propose the scale mixture of Skew-Normal distributions (SMSN) as a flexible mixing distribution for serological data analysis. The flexibility of this family is attributed to four parameters that control the location, the scale, the skewness and the flatness of the resulting distribution. In addition, SMSN includes the Normal distribution, the Generalized Student's *t*-distribution, and its skewed version as special cases [23]. We illustrate the advantage of using these models by analysing a data set related to antibodies against 6 different common herpesviruses in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) [24].

2 DATA UNDER ANALYSIS

ME/CFS is complex disease whose patients experience a long-lasting fatigue that cannot be alleviated by rest or suffer from post-exertional malaise upon minimal physical and mental activity [25, 26]. The aetiology of the disease remains unknown, but it is often linked with common viral infections, including common herpesviruses [27].

To accelerate current knowledge on ME/CFS, it was created a large disease-specific biobank in the United Kingdom [28, 29]. The data set under analysis is part of this biobank and it was published in a recent study with the aim of investigating the immunological component of the disease [24]. In the data set, there is a total of 406 individuals, all adults, 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 further divided into a subgroup of 196 patients with mild or moderate symptoms (ME-M) and another subgroup of 54 severely affected patients who are home- or even bed-bound (ME-S). A detailed description about the recruitment of study participants, inclusion/exclusion criteria, and ethics can be found in the original reference [24].

The data set comprises six serological variables corresponding to the antibody concentration against 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.

In each serum sample, the concentration of each vira-specific antibody was expressed in arbitrary units per ml (U/ml) according the corresponding optical

density determined by commercial ELISA kits. According to the ELISA's kit manufacturers, samples with antibodies concentration ≤ 8 U/ml should be classified as seronegative and those with concentration ≥ 12 U/ml should be classified as seropositive for all antibodies with the exception of HHV-6. Samples with IgG concentration between 8 and 12 U/ml should be classified as equivocal. 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 these two limits were considered equivocal.

2.1 Ethical approval

All participants provided written informed consent for data collection (questionnaire, clinical measurement and laboratory tests), and for allowing their samples to be available to any research receiving ethical approval. Participants received an extensive information sheet and consent form in which there was an option for participation withdraw from the study at any time. Ethical approval was granted by the London School of Hygiene & Tropical Medicine (LSHTM) Ethics Committee (Ref. 6123) and the National Research Ethics Service (NRES) London-Bloomsbury Research Ethics Committee (REC ref. 11/10/1760, IRAS ID: 77765).

3 STATISTICAL ANALYSIS OF SEROLOGICAL DATA

3.1 Finite mixture models based on SMSN distributions

When analysing serological data related to the antibody responses against a specific antigen, it is usually assumed the existence of two or more latent, unobserved populations, which might represent different levels of exposure to that antigen. For simplicity, individuals that were never exposed to a given antigen are considered as seronegatives whilst individuals exposed to it are considered seropositives. In this scenario, the respective data from a specific antibody are typically described by a mixture of two or more probability distributions.

Let $G_1, \dots, G_g$ be the partition from a superpopulation G (sample space) and $\pi_1, \dots, \pi_g$ 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 $Z_1, Z_2, \dots, Z_g$ if the probability density function (pdf) of Z is given by

$$f(z) = \sum_{k=1}^g \pi_k f_{Z_k}(z; \theta_k), \quad (1)$$

where $f_{Z_k}(z; \theta_k)$ is the mixing probability density function of Z_k associated with the k -th latent population and parameterized by the vector $\theta_k = \{\theta_1, \dots, \theta_{l_k}\}$.

The most popular choice for the mixing distribution in serological analysis is the Normal distribution which is symmetric around the mean and it is a mesokurtic distribution (with a kurtosis of 3 irrespective of the mean and standard deviation of the distribution). However, serological data from populations on the brink of malaria elimination show long tails and marked right asymmetry [16] in each latent population even after applying log-transformation. In such cases, one can use instead the Generalized Student t as the mixing distribution, because it has heavier tails than the Normal distribution. However, this distribution remains in the realm of the symmetric distributions. To incorporate asymmetry in the modelling, one can alternatively use the less-known Skew-Normal as the mixing distribution [30].

A random variable W_k has a Skew-Normal distribution with location parameter μ_k , scale parameter σ_k^2 and skewness parameter α_k (denoted as $W_k \sim \mathcal{SN}(\mu_k, \sigma_k^2, \alpha_k)$) if its pdf can be written as

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

where $\phi(\cdot; \mu_k, \sigma_k^2)$ denotes the pdf of the Normal distribution with mean μ_k and variance σ_k^2 ; $\Phi(\cdot)$ denotes the cumulative distribution function of the standard Normal distribution [23, 31, 32]. The mean and variance of the Skew-Normal distribution are respectively given by,

$$E(W_k) = \mu_k + \sigma_k \sqrt{\frac{2}{\pi}} \frac{\alpha_k}{\sqrt{1 + \alpha_k^2}}, \quad V(W_k) = \left(1 - \left(\frac{2}{\pi} \frac{\alpha_k}{\sqrt{1 + \alpha_k^2}}\right)^2\right) \sigma_k^2.$$

Additionally, the Skew-Normal distribution can be used to construct a more general class of flexible distributions, the scale mixtures of Skew-Normal (SMSN) distributions.

The random variable Z_k in expression (1) belongs to the SMSN family with location parameter μ_k , scale parameter σ_k^2 and skewness parameter α_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:

$$Z_k = \mu_k + \frac{W_k}{\sqrt{U_k}}, \quad (2)$$

where μ_k is the location parameter; U_k is a random variable with distribution function $H_k(\cdot, \mathbf{v}_k)$ and pdf $h_k(\cdot, \mathbf{v}_k)$; $\mathbf{v}_k$ 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 [23, 33].

Based on expression (2), it is worth noting that the conditional distribution $Z_k|U_k = u$ takes the form

$$\begin{aligned} F_{Z_k|U_k=u}(z) &= P(Z_k \leq z) = P\left(\mu_k + \frac{1}{\sqrt{u}} W_k \leq z\right) \\ &= P(W_k \leq \sqrt{u}(z - \mu_k)) = F_{W_k}(\sqrt{u}(z - \mu_k)), \quad z \in \mathbb{R}. \end{aligned} \quad (3)$$

Thus,

$$\begin{aligned} f_{Z_k|U_k=u}(z) &= \frac{d}{dz} F_{W_k}(z) = \sqrt{u} \times f_{W_k}(\sqrt{u}(z - \mu_k)) \\ &= \sqrt{u} \frac{2}{\sigma_k} \phi\left(\frac{\sqrt{u}(z - \mu_k)}{\sigma_k}\right) \Phi\left(\frac{\alpha_k \sqrt{u}(z - \mu_k)}{\sigma_k}\right), \quad z \in \mathbb{R}, \end{aligned} \quad (4)$$

where $\phi(\cdot)$ 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

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

The name of this class of distributions relies on the fact that the density function of Z_k (2) 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 (2) simplifies to $Z_k = \mu_k + W_k$. Hence,

$$F_{Z_k}(z) = P(W_k \leq z - \mu_k) = F_{W_k}(z - \mu_k), \quad z \in \mathbb{R}, \quad (5)$$

$$f_{Z_k}(z) = f_{W_k}(z - \mu_k) = 2\phi(z - \mu_k; 0, \sigma_k^2) \Phi\left(\alpha_k \left(\frac{z - \mu_k}{\sigma_k}\right)\right). \quad (6)$$

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

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

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 Generalized 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, $U_k \sim \text{Gamma}(\frac{v}{2}, \frac{v}{2})$. 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 $W_k \sim \mathcal{SN}(0, \sigma_k^2, \alpha_k)$, $U_k \sim \text{Gamma}(\frac{v}{2}, \frac{v}{2})$ are independent random variables, is equivalent to $Z_k = \mu_k + \frac{W_k}{\sqrt{\frac{R_k}{v}}}$ where R_k is a

χ^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 (3) and (4), respectively. According to expression (5), the marginal probability density distribution of Z_k takes the form

$$\begin{aligned} f_{Z_k}(z) &= \int_0^{+\infty} f_{Z_k|U_k=u}(z) f_{U_k}(u) du \\ &= \int_0^{+\infty} 2\sqrt{u} \phi(\sqrt{u}(z - \mu_k); 0, \sigma_k^2) \Phi\left(\frac{\alpha_k \sqrt{u}(z - \mu_k)}{\sigma_k}\right) \frac{\left(\frac{v_k}{2}\right)^{\frac{v_k}{2}} u^{\frac{v_k}{2}-1} e^{-\frac{v_k}{2}u}}{\Gamma\left(\frac{v_k}{2}\right)} du \\ &= \frac{2v_k^{\frac{v_k}{2}}}{\sigma_k \sqrt{\pi} 2^{\frac{v_k+1}{2}} \Gamma\left(\frac{v_k}{2}\right)} \int_0^{+\infty} \Phi(\sqrt{u}A) u^{\frac{1}{2}(v_k-1)} e^{-\frac{1}{2}u(d+v_k)} du, \end{aligned} \quad (7)$$

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

Integrating expression (7) by substitution of the variable $s = \frac{1}{2}u(d+v_k)$, we obtain

$$\begin{aligned} f_{Z_k}(z) &= \frac{2}{\sigma_k \sqrt{\pi} v_k \Gamma\left(\frac{v_k}{2}\right)} \left(1 + \frac{d}{v_k}\right)^{-\frac{1}{2}(v_k+1)} \int_0^{+\infty} \Phi\left(A \sqrt{\frac{2s}{d+v_k}}\right) s^{\frac{1}{2}(v_k-1)} e^{-s} ds \\ &= \frac{2 \Gamma\left(\frac{v_k+1}{2}\right)}{\sigma_k \sqrt{\pi} v_k \Gamma\left(\frac{v_k}{2}\right)} \left(1 + \frac{d}{v_k}\right)^{-\frac{1}{2}(v_k+1)} \int_0^{+\infty} \Phi\left(A \sqrt{\frac{2}{d+v_k}} \sqrt{s}\right) \frac{1}{\Gamma\left(\frac{v_k+1}{2}\right)} s^{\frac{1}{2}(v_k-1)} e^{-s} ds \\ &= \frac{2 \Gamma\left(\frac{v_k+1}{2}\right)}{\sigma_k \sqrt{\pi} v_k \Gamma\left(\frac{v_k}{2}\right)} \left(1 + \frac{d}{v_k}\right)^{-\frac{1}{2}(v_k+1)} \times \\ &\quad \times \int_0^{+\infty} P\left(Z \leq A \sqrt{\frac{2}{d+v_k}} \sqrt{s} | S = s\right) \frac{1}{\Gamma\left(\frac{v_k+1}{2}\right)} s^{\frac{1}{2}(v_k-1)} e^{-s} ds. \end{aligned} \quad (8)$$

It is important to notice the following Lemma [34].

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

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

Applying this Lemma to expression (8) leads to

$$\begin{aligned}
 f_{Z_k}(z) &= \frac{2 \Gamma(\frac{v_k+1}{2})}{\sigma_k \sqrt{\pi v_k} \Gamma(\frac{v_k}{2})} \left(1 + \frac{d}{v_k}\right)^{-\frac{1}{2}(v_k+1)} E\left(\Phi\left(A\sqrt{\frac{2}{d+v_k}}\sqrt{s}\right)\right) \\
 &= 2 t(z; \mu_k, \sigma_k, v_k + 1) E\left(\Phi\left(A\sqrt{\frac{2}{d+v_k}}\sqrt{s}\right)\right) \\
 &= 2 t(z; \mu_k, \sigma_k, v_k + 1) P\left(T \leq A\sqrt{\frac{v_k+1}{d+v_k}}; v_k + 1\right) \\
 &= 2 t(z; \mu_k, \sigma_k, v_k + 1) T\left(A\sqrt{\frac{v_k+1}{d+v_k}}; v_k + 1\right),
 \end{aligned} \tag{9}$$

where $t(\cdot; \mu_k, \sigma_k, v_k + 1)$ denotes the probability density function of a Generalized Student-t distribution with location parameter μ_k , scale parameter σ_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

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

It should be noted that when the skewness parameter is equal to zero, i.e., $\alpha_k = 0$, the quantity $A = \frac{\alpha_k(z-\mu_k)}{\sigma_k} = 0$, and the above expression takes the form

$$f_{Z_k}(z) = 2 t(z; \mu_k, \sigma_k, v_k + 1) P(T \leq 0; v_k + 1) = t(z; \mu_k, \sigma_k, v_k + 1).$$

which corresponds to the probability density function of a Generalized Student-t distribution with location parameter μ_k , scale parameter σ_k and $v_k + 1$ degrees of freedom.

As the degrees of freedom tends to infinity, the Skew-t distribution converges to the Skew-Normal distribution [23, 31, 32].

The mean and variance of the Skew-t distribution are respectively given by,

$$E(Z_k) = \mu_k + \sigma_k b_{v_k} \delta_k, \text{ if } v_k > 1, \quad V(Z_k) = \sigma_k^2 \left[\frac{v_k}{v_k - 2} - \left(b_{v_k} \delta_k \right)^2 \right] \text{ if } v_k > 2,$$

$$\text{where } b_{v_k} = \frac{\sqrt{v_k} \Gamma(\frac{1}{2}(v_k-1))}{\sqrt{\pi} \Gamma(\frac{1}{2}v_k)} \text{ and } \delta_k = \frac{\alpha_k}{\sqrt{1+\alpha_k^2}}.$$

3.2 Estimation and model selection

Suppose that we have a random sample $X_1, \dots, X_n$ representing the antibody levels of n individuals. In general, it is very difficult to determine the maximum

likelihood (ML) estimates of the parameters of any given finite mixture model by direct maximization of the corresponding log-likelihood functions. One way to surpass this problem is to consider the Expectation-Maximization (EM) algorithm given that the latent serological status of each individual is unknown and, thus, we are in the presence of a problem of incomplete data.

A full derivation of an EM-type algorithm for fitting mixtures of SMSN can be found elsewhere [23]. In brief, the E-step is the same as in the traditional mixtures of Normal distributions, which has been largely studied in the literature [23, 30, 33]. Replacing the classical M-step with a sequence of conditional maximization steps (CM-steps), one obtains closed form expressions that simplify the implementation of the algorithm. Also, the observed information matrix can be derived analytically [35].

There are several methods to determine the optimal number of components that constitute the mixture, g [13, 36, 37]. A simple way to do it is to use penalized forms of the log-likelihood function: the information criteria. They rely on the idea that an increase in the number of components in the mixture leads to a better fit of the data, thus, increasing the maximized likelihood function. Invoking the parsimony principle to determine the best model for the data, enhancement in model fitting is penalized by an increase in the number of parameters included in the model. In this framework, two of the most popular measures are Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) [38]. In general the best model is the one that provides the lowest estimate of AIC or BIC value among all models tested. About the use of one or the other criterion, the BIC criterion demonstrates consistency in determining the number of components of the mixture models [38, 39]. In addition, BIC tends to select simpler models (ideally with less number of components) than AIC, which simplifies data interpretation. Hence, this information criterion is preferred for model selection.

Another way to assess the number of components in a mixture model is to carry out a hypothesis testing, namely the Likelihood Ratio Test (LRT). However, the regularity conditions for the validity of classical asymptotic approximation of the test statistic are not met in the context of finite mixture models, because the null hypothesis associated with this hypothesis is specified in the boundary of the parameter space rather than its interior [13]. In some cases, the true parameter is in a non-identifiable subset of the parameter space [40]. As a consequence, there is no guarantee that, under the null hypothesis, the likelihood ratio statistic asymptotically follows a χ^2 distribution with the degrees of freedom given by the difference between the number of parameters under the alternative and the null hypothesis [13]. To surpass this problem, a bootstrap approach can be carried out to estimate the p-value of this non-standard LRT [40, 41].

Let us consider the test specified by $H_0 : g = g_0$ versus $H_1 : g = g_1$ where $g_0 < g_1$. Let us also denote ψ_0 and ψ_1 the vector of parameters of the mixture models under H_0 and H_1 hypotheses, respectively; $\mathbf{x} = (x_1, \dots, x_n)$ the observed data and $T(\mathbf{x}; \psi_0, \psi_1)$ the test statistic of LRT. The bootstrap approach is given by the following algorithm [41]:

1. Use the EM algorithm to estimate the ψ_0 and ψ_1 estimates under the H_0 and H_1 hypotheses, respectively. Calculate $T(\mathbf{x}; \hat{\psi}_0, \hat{\psi}_1)$;
2. Simulate $N = 10,000$ independent samples $\mathbf{x}_1^*, \dots, \mathbf{x}_n^*$ using the mixture model under H_0 and parameterized by $\hat{\psi}_0$;
3. For each bootstrap sample i , calculate $T(\mathbf{x}_i^*; \hat{\psi}_{0_i}, \hat{\psi}_{1_i})$, where $\hat{\psi}_{0_i}$ and $\hat{\psi}_{1_i}$ 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 $\frac{1}{N} \sum_{i=1}^N I\{T(\mathbf{x}_i^*; \hat{\psi}_{0_i}, \hat{\psi}_{1_i}) > T(\mathbf{x}; \hat{\psi}_0, \hat{\psi}_1)\}$, where $I\{\cdot\}$ is the indicator function.

Another important statistical test in the context of mixtures based on SMSN is to address the significance of the asymmetry parameters of the mixing distributions composing the mixture model based on Skew-Normal or Skew-t distributions. To attain this goal, a LRT can also be carried out. Suppose that we have a mixture model with all g components given by either Skew-Normal or Skew-t distributions. In this test, the hypotheses under testing are the following:

$$H_0 : \alpha_k = 0, \forall k=1, \dots, g \text{ versus } H_1 : \exists k=1, \dots, g, \alpha_k \neq 0.$$

The test statistic is given by $\Lambda = -2 \ln \frac{\hat{L}_0}{\hat{L}_1}$, where $\hat{L}_0$ and $\hat{L}_1$ correspond to the likelihood functions of the mixture model evaluated at the maximum likelihood estimates under H_0 and H_1 , respectively. In contrast with the previous test related to the number of components, the usual asymptotic approximation for the distribution of the LRT statistic under H_0 holds, that is, a χ^2 distribution with g degrees of freedom.

Finally, the quality of the fit of the estimated models should be assessed. For a matter of simplicity, the Pearson's χ^2 test for goodness-of-fit can be used. To apply this test, the data under analysis can be divided into bins according to the sampled 5%-quantiles or deciles (*i.e.*, 10%-quantiles).

3.3 Estimation of seropositivity

After determining the best finite mixture model for the data, the next step 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 usually 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 value 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 ($i = 1, \dots, n$) 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 [22, 42]. 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 [43]. 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 [22]. 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

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

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 [22].

3.4 R packages

We used the package `mixsmsn` to fit different mixture models based on SMSN [44]. In particular, we used the function `smsn.mix` to estimate the model parameter via the EM algorithm and the function `rmix` to generate random samples from a given mixture model in the bootstrap method. For fitting the Student's t-distribution, we considered the R package `extraDistr` [45], namely, the function `dlst` to calculate their density and the function `plst` to define the cumulative distribution function. The fitting of the Skew-Normal distributions was performed with the package `sn` [46]. The functions `dsn` and `psn` were used to calculate the probability density function and the cumulative distribution function of the Skew-Normal distribution, respectively. In the case of the Skew-t distribution, the functions `dst` and `pst` were used to calculate the probability density function and the cumulative distribution function, respectively.

4 RESULTS

4.1 Analysis of serological data by finite models based on SMSN

The statistical analysis was performed after applying the base 10 logarithmic transformation to the data. The number of components g in the mixture models was allowed to vary from 1 (single distribution) to 3 components. When fitting the mixtures of Skew-t distributions, the package `mixsmsn` only allowed to fit models with a common degree of freedom for all mixing distributions (*i.e.*, $v_1 = \dots = v_g = v$).

Our results suggested that the 6 antibodies under evaluation could be divided into two major classes: (i) the first one including antibodies against HHV-6 and VZV where there was evidence for a single serological population (Table 1) and (ii) another one including the antibodies against the remaining four herpesviruses where there was evidence for the existence of more than one serological population in the respective data (Table 2).

According to BIC, the best models for the antibodies against HHV-6 and VZV were Skew-Normal and Skew-t distributions, respectively. The estimated distributions were both left skew (Figure 1A; $\alpha_{HHV6} = -1.82$ with 95% CI = (-2.44; -1.02) and $\alpha_{VZV} = -4.54$ with 95% CI = (-6.94; -2.14). They would appear to have a good fit to the data at the 5% significance level ($p_{gof} = 0.140$ and 0.076, respectively). In the case of antibodies against VZV, further evidence was obtained for a single population when testing one Skew-t distribution against

Virus	SMSN	g	p	$\mathcal{L}_{\max}$	BIC	p_{gof}	p_{boot}
HHV-6	Normal	1	2	-129.46	270.94	0.064	0.064
		2	5	-116.97	263.97	0.169	
		3	8	-110.43	268.91	0.462	
	Skew-Normal	1	3	-121.35	260.71	0.140	0.027
		2	7	-117.35	276.75	0.084	
		3	11	-109.40	284.87	0.152	
	Student's t	1	3	-124.38	266.77	0.157	0.042
		2	6	-117.14	270.32	0.122	
		3	9	-105.36	264.78	0.254	
	Skew-t	1	4	-118.81	261.65	0.148	0.409
		2	8	-116.83	281.71	0.076	
		3	12	-104.00	586.83	0.001	
VZV	Normal	1	2	-108.76	229.53	< 0.001	0.000
		2	5	-7.28	44.60	0.159	
		3	8	-1.70	51.45	0.153	
	Skew-Normal	1	3	-23.94	65.90	< 0.001	0.180
		2	7	-0.11	42.27	0.406	
		3	11	0.10	65.87	0.068	
	Student's t	1	3	-61.90	141.80	< 0.001	0.000
		2	6	-7.41	50.86	0.082	
		3	9	-1.68	57.42	0.113	
	Skew-t	1	4	-7.89	39.81	0.076	0.375
		2	8	-0.05	48.16	0.211	
		3	12	5.47	62.14	0.134	

Table 1: Analysis of antibody data with evidence for a single serological population, where g represents the number of components in the mixture models, p is the number of parameters of the model, $\mathcal{L}_{\max}$ 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 $H_1 : g = 2$.

a mixture of two Skew-t distributions, respectively ($p_{boot} = 0.375$). However, when testing one Skew-Normal distribution against a mixture of two Skew-Normal distributions for the antibodies against HHV-6, the respective result was in the borderline of the 5% statistical significance ($p_{boot} = 0.027$).

In terms of serological classification, the evidence for a single population would appear to represent a putative seropositive population. This interpretation is consistent with the prior knowledge that HHV-6 and VZV are usually acquired during childhood and more than 95% of the adult populations typically shows evidence of antibody positivity against these viruses [47]. 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 [17].

It is worth noting that most of the mixture models under comparison could also fit data of these two antibodies well. This the case of the mixture of two or three Normal distributions ($p_{gof} = 0.169$ and 0.462 for antibodies against HHV-6 and $p_{gof} = 0.159$ 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.

With respect to the antibodies against the remaining herpesviruses, the respective data analysis was not so 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 is the case of the mixtures of two Skew-Normal distributions for the antibodies against CMV (BIC=509.69 and $p_{gof} = 0.038$), HSV-1 (BIC=563.52 and $p_{gof} = 0.003$), and HSV-2 (BIC=570.68 and $p_{gof} = 0.013$). For these antibodies, the best models were considered to be a mixture of two Skew t distributions (BIC=511.45 and $p_{gof} = 0.072$), a mixture of three Skew-Normal distributions (BIC=570.70 and $p_{gof} = 0.104$), and a mixture of two Normal distributions (BIC=585.27 and $p_{gof} = 0.516$), 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 2 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 EBV-related antibody data, the best model according to BIC was a mixture of two Skew-t distributions, which also had a good fit for the data (BIC=299.32 and $p_{gof} = 0.248$; Figure 1B).

In terms of 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 tends to show right asymmetry [17] and, if so, this intermediate component should be interpreted accordingly. However, this interpretation opens the door for the presence of two sets of seronegative populations resulting from distinct background signals in absence of antibodies. In absence of additional information about the serological data, this intermediate component was considered to represent a putative seronegative population.

4.2 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

Virus	SMSN	g	p	$\mathcal{L}_{\max}$	BIC	p_{gof}
CMV	Normal	1	2	-409.11	830.24	< 0.001
		2	5	-245.75	521.54	0.016
		3	8	-233.70	515.45	0.018
	Skew-Normal	1	3	-357.61	733.23	< 0.001
		2	7	-233.82	509.69	0.038
		3	11	-226.64	519.35	0.146
	Student's t	1	3	-410.14	838.29	< 0.001
		2	6	-238.54	513.12	0.038
		3	9	-231.23	516.59	0.046
	Skew-t	1	4	-357.71	739.45	< 0.001
		2	8	-231.55	511.45	0.072
		3	12	-226.93	525.93	0.324
EBV	Normal	1	2	-342.30	696.62	< 0.001
		2	5	-152.66	335.36	< 0.001
		3	8	-129.30	306.65	0.173
	Skew-Normal	1	3	-226.42	470.86	< 0.001
		2	7	-130.57	303.17	0.084
		3	11	-128.02	322.10	0.054
	Student's t	1	3	-240.21	498.43	< 0.001
		2	6	-151.61	339.26	< 0.001
		3	9	-129.41	312.88	0.117
	Skew-t	1	4	-173.14	370.31	< 0.001
		2	8	-125.63	299.32	0.248
		3	12	-126.29	324.66	0.087
HSV-1	Normal	1	2	-442.27	896.56	< 0.001
		2	5	-291.59	613.22	< 0.001
		3	8	-264.94	577.94	0.003
	Skew-Normal	1	3	-394.55	807.11	< 0.001
		2	7	-260.74	563.52	0.003
		3	11	-252.32	570.70	0.104
	Student's t	1	3	-443.73	905.48	< 0.001
		2	7	-291.73	619.51	< 0.001
		3	9	-264.98	584.02	0.002
	Skew-t	1	4	-395.43	814.88	< 0.001
		2	8	-260.88	569.82	0.001
		3	12	-251.86	575.79	< 0.001
HSV-2	Normal	1	2	-427.29	866.59	< 0.001
		2	5	-277.62	585.27	0.516
		3	8	-269.24	586.54	0.007
	Skew-Normal	1	3	-337.36	692.74	< 0.001
		2	7	-264.32	570.68	0.013
		3	11	-257.19	580.45	0.003
	Student's t	1	3	-428.40	874.81	< 0.001
		2	6	-277.84	591.71	0.688
		3	9	-269.60	593.26	0.004
	Skew-t	1	4	-337.79	699.60	< 0.001
		2	8	-264.52	577.10	0.007
		3	12	-257.38	586.83	0.001

Table 2: Analysis of antibody data with evidence for more than one serological population, where g represents the number of components in the mixture models, p is the number of parameters of the model, $\mathcal{L}_{\max}$ is the value of the maximized log-likelihood function, and p_{gof} is the maximum p-value for the goodness-of-fit test when dividing data in deciles or 5%-quantiles.

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

Figure 1: Best models for the data under analysis. **A.** Antibody distributions with evidence for a single serological population (HHV-6 and VZV). **B.** Antibody distributions with evidence for more than one serological population (CMV, EBV, HSV1, and HSV2). Antibody concentration in x axis is given in $\log_{10}$ units.

the original study [24]. Since the seronegative and seropositive populations were separated well in these antibody distributions, the estimates of seroprevalence across the different study groups were almost invariant with respect to 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); Figure 1B) with heavy tails ($v = 4.52$; 95% CI=(0.79;8.26)). As a consequence, the cut-off 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.

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 again a too-high cutoff value due to the right asymmetry of both seronegative populations. Such unrealistic cutoff value led 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 finite mixture models based on SMSN and to recommend their routine use in serological data analysis. Such recommendation sets its foundation in the high flexibility of these models in describing different patterns of randomness, as illustrated with the analysis of antibodies against 6 different herpesviruses. In particular, a high modelling flexibility is necessary given that right and left asymmetry could emerge from hypothetical 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 remarkable stochastic property is the so-called the Generalized Tukey's λ distribution [50, 51]. 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 available for the two-component case [48, 49]. This limits the applicability of these

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)

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.

alternative models, namely, in data 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 [20]. Therefore, finite mixture models based on SMSN would appear the most general and flexible approach so far for analysing serological data.

For data analysis, we recommend the use of the package `mixsmsn` for estimating the finite mixture models [44]. Notwithstanding this recommendation, the package only allows to estimate finite mixture models where all mixing distributions belong to the same class of SMSN probability distributions. Hence, it is only possible to 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 result from 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 true best model for the data. This computational limitation might be the reason for some inconsistencies that can be found in the example of application. For instance, a single Skew-Normal distributions 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 contradicting evidence 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.

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 [52]. However, this package is currently discontinued. In practice, we expect some degree of numerical instability when trying to estimate different degrees of freedom for mixtures in which different components overlap with each other substantially. In this regard, future research could be conducted in order to determine under which conditions different degrees of freedom could infer for the different components.

The problem of determining the optimal cutoff value for seropositivity has been intensively investigated, discussed, and revisited over the years [42, 53, 54, 55]. 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 skew distribution, similar cutoffs can be obtained by calculating same high quantiles of the estimated SMSN, as done here. The reverse argument can be made when analysing antibodies where seropositivity is expected to be 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. For these antibodies, similar cutoffs can be determined by the mean minus a given number of times the standard deviation of the hypothetical seropositive population assumed to be normally distributed. For a hypothetical seropositive population with a left skew distribution, the cutoff values for seropositive 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, as expected, several authors advocate

a free-cutoff approach for serological analysis [15, 56]. However, a detailed discussion about the advantages and disadvantages of free-cutoff approaches was considered to be out of the scope of this study.

In terms of the results concerning the example of application, there is no evidence for a different level of exposure of the patients with ME/CFS to these herpesviruses when compared to healthy controls and patients with multiple sclerosis. This finding seems independent of the method for determining the seropositivity and it is in line with the findings reported in the original study [24] and with another serological investigation of these herpesviruses when comparing patients with healthy controls only [9]. A possible explanation for this "negative" finding might be explained by the choice of antibodies against highly immunogenic antigens used in this serological study. It is then possible that there is a specific set of viral-derived antigens associated with ME/CFS, as suggested by a comprehensive study about the role of antibodies against EBV in this disease [7]. Finally, a more detailed analysis of these data is currently carried out in order to understand whether the lack of association between ME/CFS and these antibodies could be explained by putative confounding effects of age and gender on the underlying antibody distributions. This detailed analysis will be reported elsewhere.

In summary, the finite mixture models based on SMSN show a good potential to become a routine tool for serological data analysis. They have the advantage of including the popular Gaussian mixture models as special cases. However, given the statistical complexity of these models, we recommend 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 them properly.

6 Acknowledgments

The authors would 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 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). Seropreva-

- lence 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] VILA NOVA, B.; CUNHA, E.; SEPÚLVEDA, N.; OLIVEIRA, M.; SÃO BRAZ, B.; TAVARES, L.; ALMEIDA, V. and GIL, S. (2018). Evaluation of the humoral immune response induced by vaccination for canine distemper and parvovirus: a pilot study, *BMC Veterinary Research*, **16**, 14, 348.
- [3] SHIKOVA, E.; RESHKOVA, V.; KUMANOVA, ?.; RALEVA, S.; ALEXANDROVA, D.; CAPO, N.; MUROVSKA, M. and EUROPEAN NETWORK ON ME/CFS (EUROMENE) (2020). Cytomegalovirus, Epstein-Barr virus, and human herpesvirus-6 infections in patients with myalgic ?ncephalomyelitis/chronic fatigue syndrome, *Journal of Medical Virology*, **92**, 12, 3682?-3688.
- [4] 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.
- [5] 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.
- [6] HELB, D. A.; TETTEH, K. K.; FELGNER, P. L.; SKINNER, J.; HUBBARD, A.; ARINAITWE, E.; MAYANJA-KIZZA, H.; SSEWANYANA, I.; KAMYA, M. R.; BEESON, J. G.; TAPPERO, J.; SMITH, D. L.; CROMPTON, P. D., ROSENTHAL, P. J., DORSEY, G., DRAKELEY, C. J., and GREENHOUSE, B. (2015). Novel serologic biomarkers provide accurate estimates of recent Plasmodium falciparum exposure for individuals and communities, *Proceedings of the National Academy of Sciences of the United States of America*, **112**, **32**, E4438?E4447.
- [7] LOEBEL, M.; ECKEY, M.; SOTZNY, F.; HAHN, E.; BAUER, S.; GRABOWSKI, P.; ZERWECK, J.; HOLENYA, P.; HANITSCH, L. G.; WITKE, K.; BORCHMANN, P.; RÜFFER, J. U.; HIEPE, F.; RUPRECHT, K.; BEHREND, 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.
- [8] LAMMIE, P. J.; MOSS, D. M.; BROOK GOODHEW, E.; HAMLIN, K.; KROLEWIECKI, A.; WEST, S. K. and PRIEST, J. W. (2012). Development of a new platform for neglected tropical disease surveillance, *International Journal for Parasitology*, **42**, 9, 797?-800.

- [9] 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.
- [10] SOWA, M., HIEMANN, R., SCHIERACK, P., REINHOLD, D., CONRAD, K., and ROGGENBUCK, D. (2017). Next-Generation Autoantibody Testing by Combination of Screening and Confirmation-the CytoBead® Technology, *Clinical Reviews in Allergy & Immunology*, **53**,1, 87?-104.
- [11] 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.
- [12] WU, L., HALL, T., SSEWANYANA, I., OULTON, T., PATTERSON, C., VASILEVA, H., SINGH, S., AFFARA, M., MWESIGWA, J., CORREA, S., BAH, M., D’ALESSANDRO, U., SEPÚLVEDA, N., DRAKELEY, C., and TETTEH, K. (2020). Optimisation and standardisation of a multiplex immunoassay of diverse Plasmodium falciparum antigens to assess changes in malaria transmission using sero-epidemiology, *Wellcome Open Research*, **4**, 26.
- [13] MCLACHLAN, G. and PEEL, D (2000). *Finite Mixture Models*. John Wiley & Sons, New York.
- [14] GAY, N.J. (1996). Analysis of serological surveys using mixture models: application to a survey of parvovirus B19, *Statistics in Medicine*, **15**, 1567–1573.
- [15] 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.
- [16] 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.
- [17] 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.

- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] FUKUDA, K.; STRAUS, S. E.; HICKIE, I.; SHARPE, M. C.; DOBBINS, J. G., and KOMAROFF, A. (1994). The chronic fatigue syndrome: a comprehensive approach to its definition and study. International Chronic Fatigue Syndrome Study Group, *Annals of Internal Medicine*, **121**, 12, 953–959.
- [26] CARRUTHERS, B. M.; JAIN, A. K.; DE MEIRLEIR, K. L.; PETERSON, D. L.; KLIMAS, N. G.; LERNER, A. M.; BESTED, A. C.; FLOR-HENRY, P.; JOSHI, P.; PETER POWLES, A. C., SHERKEY, J. A. and VAN DE SANDE, M. I. (2003). Myalgic Encephalomyelitis/Chronic Fatigue Syndrome, *Journal of Chronic Fatigue Syndrome*, **11**, 1, 7–115.

- [27] RASA, S.; NORA-KRUKLE, Z.; HENNING, N.; ELIASSEN, E.; SHIKOVA, E.; HARRER, T.; SCHEIBENBOGEN, C.; MUROVSKA, M.; PRUSTY, B. K. and EUROPEAN NETWORK ON ME/CFS (EUROMENE) (2018). Chronic viral infections in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS), *Journal of Translational Medicine*, **16**, 1, 268.
- [28] LACERDA, E. M.; BOWMAN, E. W.; CLIFF, J. M.; KINGDON, C. C.; KING, E. C.; LEE, J.S.; CLARK, T.G.; DOCKRELL, H.M.; RILEY, E.M.; CURRAN, H. and NACUL, L. (2017). The UK ME/CFS Biobank for biomedical research on Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) and Multiple Sclerosis, *Open Journal of Bioresources*, **4**, 4.
- [29] LACERDA, E. M.; MUDIE, K.; KINGDON, C.C.; BUTTERWORTH, J.D.; O'BOYLE, S. and NACUL, L. (2018). The UK ME/CFS Biobank: A Disease-Specific Biobank for Advancing Clinical Research Into Myalgic Encephalomyelitis/Chronic Fatigue Syndrome, *Frontiers in Neurology*, **9**, 1026.
- [30] LIN, T.I. and LEE, J.C. and YEN, S.Y. (2007). Finite mixture modelling using the Skew-Normal distribution, *Statistica Sinica*, **17**, 909–927.
- [31] AZZALINI, A. (1985). A Class of distributions which includes the normal Ones, *Scandinavian Journal of Statistics*, **12**, 171–178.
- [32] AZZALINI, A. and CAPITANIO, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution, *J.R.Statist.Soc.B*, **65**, 367–389.
- [33] LACHOS DÁVILA, V. H. and ZELLER, C. B. and CABRAL, C. R. B. (2018). *Finite mixture of skewed distributions*, Springer.
- [34] AZZALINI, A. (2014). *The skew-normal and related families*, Cambridge University Press.
- [35] FERREIRA, C.S.; BOLFARINE H.; and LACHOS V.H. (2011). Skew scale mixtures of normal distributions: Properties and estimation, *Statistical Methodology*, **8**, 154–171.
- [36] OLIVEIRA-BROCHADO, A. and MARTINS, F. V. (2005). Assessing the number of components in mixture models: a review, *Universidade do Porto, Faculdade de Economia do Porto*, **194**.
- [37] LUKO?IEN?, O. and VERMUNT, J. K. (2009). Determining the number of components in mixture models for hierarchical data, *In Advances in data analysis, data handling and business intelligence*, 241–249. Springer, Berlin, Heidelberg.

- [38] XIE, C. H.; CHANG, J. Y. and LIU, Y. J. (2013). Estimating the number of components in Gaussian mixture models adaptively for medical image, *Optik*, **124**(23), 6216-6221.
- [39] MEHRJOU, A.; HOSSEINI, R. and ARAABI, B. N. (2016). Improved Bayesian information criterion for mixture model selection, *Pattern Recognition Letters*, **69**, 22–27.
- [40] FENG, Z.D and McCULLOGH, C.E. (1996). Using Bootstrap Likelihood Ratios in Finite Mixture Models, *Journal of the Royal Statistical Society*, **58**, 3, 609–617.
- [41] 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.
- [42] 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.
- [43] BRENT, R.P. (1973). *Algorithms for Minimization Without Derivatives*, Prentice-Hall, Englewood Cliffs, New Jersey, 73–76.
- [44] 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.
- [45] WOŁODZKO, T. (2020). Additional Univariate and Multivariate Distributions, *R CRAN*, <https://cran.r-project.org/web/packages/extraDistr/index.html>.
- [46] 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>.
- [47] BRAUN, D. K.; DOMINGUEZ, G. and PELLETT, P. E. (1997). Human herpesvirus 6, *Clinical Microbiology Reviews*, **10**, 3, 521-?567.
- [48] 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.
- [49] 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**, 2, 17.
- [50] RAMBERG, J. and SCHMEISER, B. (1974). An Approximate Method for Generating Asymmetric Random Variables, *Communications of the Association for Computing Machinery*, **17**, 78?-82.

- [51] 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.
- [52] 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.
- [53] 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.
- [54] 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.
- [55] MIGCHELSEN, S. J.; MARTIN, D. L.; SOUTHISOMBATH, K.; TURKAGUMA, 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.
- [56] BOUMAN, J. A.; BONHOEFFER, S. and REGOES, R. R. (2020). Estimating seroprevalence with imperfect serological tests: exploiting cutoff-free approaches. *bioRxiv*, doi: <https://doi.org/10.1101/2020.04.29.068999>.