Strong consistency of pseudo-likelihood parameter estimator for univariate Gaussian mixture models

Jüri Lember^{a*}, jyril@ut.ee Raul Kangro^a, raul.kangro@ut.ee Kristi Kuljus^a, kristi.kuljus@ut.ee

^aInstitute of Mathematics and Statistics, University of Tartu; Narva mnt 18, 51009, Tartu, Estonia * Corresponding author

Abstract

We consider a new method for estimating the parameters of univariate Gaussian mixture models. The method relies on a nonparametric density estimator \hat{f}_n (typically a kernel estimator). For every set of Gaussian mixture components, \hat{f}_n is used to find the best set of mixture weights. That set is obtained by minimizing the L_2 distance between \hat{f}_n and the Gaussian mixture density with the given component parameters. The densities together with the obtained weights are then plugged in to the likelihood function, resulting in the so-called pseudo-likelihood function. The final parameter estimators are the parameter values that maximize the pseudo-likelihood function together with the corresponding weights. The advantages of the pseudo-likelihood over the full likelihood are: 1) its arguments are the means and variances only, mixture weights are also functions of the means and variances; 2) unlike the likelihood function, it is always bounded above. Thus, the maximizer of the pseudo-likelihood function – referred to as the pseudo-likelihood estimator – always exists. In this article, we prove that the pseudo-likelihood estimator is strongly consistent.

Keywords: distance-based estimation, Gaussian mixture distributions, kernel density, maximum likelihood estimation, pseudo-likelihood estimator, strong consistency

1 Introduction

1.1 Pseudo-likelihood estimator

We consider the problem of parameter estimation in a univariate Gaussian mixture model with k components. In [5], a new method called the *pseudo-likelihood approach* was proposed for estimating all parameters (means, variances, weights). In the present article, we establish the strong consistency of the pseudo-likelihood estimator. The pseudo-likelihood approach relies on a nonparametric density estimator \hat{f}_n (typically a kernel estimator), which is used to estimate the mixture weights. More precisely, letting $\theta_i = (\mu_i, \sigma_i)$ denote the mean and variance of the i-th mixture component, and $g(\theta_i, \cdot)$ the corresponding Gaussian density, the weights are obtained by minimizing the L_2 distance between \hat{f}_n and

the mixture density with fixed parameters $\theta = (\theta_1, \dots, \theta_k)$:

$$v^{n}(\theta) := \arg \inf_{w \in S_{k}} \|\hat{f}_{n}(\cdot) - \sum_{i=1}^{k} w_{i} g(\theta_{i}, \cdot)\|.$$
 (1)

Here, S_k denotes the (k-1)-dimensional simplex, defined as

$$S_k := \{(w_1, \dots, w_k) : w_i \ge 0, \sum_i w_i = 1\},\$$

and $\|\cdot\|$ denotes the L_2 norm. The second step of the pseudo-likelihood approach is to plug the obtained weights $v^n(\theta)$, together with the parameters θ , into the likelihood function to get the pseudo-likelihood function

$$L_n(\theta) := \prod_{t=1}^n \left(\sum_{i=1}^k v_i^n(\theta) g(\theta_i, y_t) \right),$$

where y_1, \ldots, y_n is the observed sample. In [5], it was proved that for distinct y_1, \ldots, y_n , $L_n(\theta)$ is bounded even when the variances are not bounded away from 0 (Theorem 2.2 in [5]). This implies that the maximizer $\hat{\theta}_n$ of $L_n(\theta)$ exists almost surely. The estimator $\hat{\theta}_n$ will be referred to as the maximum pseudo-likelihood estimator. The main goal of the present article is to show that, under an i.i.d. sample, the maximum pseudo-likelihood estimator is strongly consistent: $\hat{\theta}_n \stackrel{a.s.}{\to} \theta^*$ and $v^n(\hat{\theta}_n) \stackrel{a.s.}{\to} w^*$, where θ^* and w^* denote parameters and weights of the true distribution. In the sequel, the term "parameters" refers to the means and variances, excluding the weights – although, strictly speaking, the weights are also parameters of the mixture density. The consistency result is stated as Theorem 2.1.

Estimating the mixture weights using the L_2 distance has already been (implicitly) exploited in the so-called DUDE method for signal denoising [18, 3]. The setting and objective in [18, 3] differ somewhat from ours. In terms of the present paper, their case corresponds to that with known component densities g_i and the weights w are estimated as $\hat{w} = A^{-1}\hat{u}$, where $\hat{u}_i = \langle \hat{f}, g_i \rangle$ and $A = (a_{ij})$ is the Gram matrix with entries $a_{ij} = \langle g_i, g_j \rangle$. When $\hat{f} \neq f$, the estimate \hat{w} may lie outside the simplex. Therefore, we use the direct estimate (1) even when we do not have a closed form of $v^n(\theta)$ any more. In our setting, the emission densities are also unknown, so we optimize a different objective function – the pseudo-likelihood – to estimate simultaneously both the densities and the weights.

A key feature of our estimation procedure is that the L_2 distance is used solely for estimating the mixture weights and not for the entire mixture density. There exists a large body of literature on distance-based estimation of mixture distributions, where the entire mixture model is estimated by minimizing a distance between a nonparametric (typically kernel-based) density estimate $\hat{f}_n(\cdot)$ and the model density $\sum_i w_i g(\theta_i, \cdot)$; that is, the minimization is performed over both the weights and the component parameters. Commonly used distances include the L_2 and L_1 norms, as well as the Hellinger distance. Another option is to minimize the distance between empirical and theoretical distribution

functions using the Wolfowitz, Cramér-von Mises or Kolmogorov distance, see, e.g., [2] and the references therein. These estimators are typically consistent, consistency follows from the continuity of the metric projection and the (relative) compactness of parameter space. In contrast, our estimator combines distance-based estimation with a likelihood-based approach, resulting in an objective function of a different nature. As a consequence, standard tools based on metric projection are not sufficient to establish consistency.

The current article is a direct follow-up of [5], where the pseudo-likelihood method was introduced. The simulations in [5] demonstrate good behavior of the maximum pseudo-likelihood method – it typically outperforms the L_2 -based estimators and in some of the studied examples even beats the local maximizer of the likelihood function obtained with the EM algorithm. In particular, our method performs well when the number of mixture components k is relatively large. This is understandable, since in the pseudo-likelihood approach the mixture weights are no longer treated as independent parameters, which reduces the number of parameters to be estimated. The larger the value of k, the greater the reduction. For a further discussion of the relationship with other estimation methods, an overview of the relevant literature, and simulations, we refer the reader to [5].

The article is organized as follows. In Section 2, we introduce necessary notation and preliminaries, define the maximum pseudo-likelihood estimator, and state the main consistency theorem (Theorem 2.1). In Section 2.3, we give a brief overview of the proof and the guidelines for reading it. Sections 3-6 are devoted to the proof of Theorem 2.1.

1.2 Motivation for studying the pseudo-likelihood approach

The need for a pseudo-likelihood arises from the well-known fact that the likelihood of Gaussian mixtures is unbounded. Several approaches have been proposed to address this issue, including restricting the parameter space, using sieves, penalized maximum likelihood estimation, Bayesian methods, profile likelihood, and others; see, e.g., [8, 13, 14] and the references therein. In [1], consistency of the maximum likelihood estimator (MLE) is proved under the assumption that the variances of the mixture components are equal, a restriction that ensures the boundedness of the likelihood function. In [14], consistency of the MLE is proved under the condition that all standard deviations are bounded below by $\exp[-n^d]$. In [13], consistency is proved under the assumption that the ratio between the minimum and the maximum variance is bounded below by a sequence b_n with $b_n \to 0$, or, more generally, when this ratio is suitably penalized. In this paper, we remove any restriction or penalties on the variances by replacing the likelihood function by the pseudolikelihood function. The pseudo-likelihood function differs from the likelihood function only through the weights $v^n(\theta)$. In a sense, this represents the minimal modification of the likelihood function required to ensure boundedness without imposing any restrictions on the parameters. The price of this modification is that existing consistency proofs and results are not directly applicable.

There are obviously many other ways to modify the likelihood function to obtain an objective function with different properties. Since $v^n(\theta)$ is based on the kernel estimate \hat{f}_n , let us mention the so-called double-smoothed likelihood introduced in [9, 10]. Recall

that the standard MLE minimizes the Kullback-Leibler divergence between the empirical measure and the assigned model. In the double-smoothed likelihood, both arguments of the Kullback-Leibler divergence – the empirical measure and the assigned distribution – are smoothed using the same kernel. The resulting function is bounded and, as shown in [11], under rather general assumptions the maximizer of the double-smoothed likelihood (DS-MLE) is consistent. The proof is relatively straightforward and closely follows the classical proof of MLE consistency. To reduce the number of parameters, the weights in the double-smoothed likelihood function can be replaced with $v^n(\theta)$ (as in the pseudolikelihood function), and we conjecture that consistency still holds by standard arguments. An advantage of DS-MLE is that the kernel bandwidth can be fixed, i.e., chosen independently of the sample size n. In contrast, in our approach, the bandwidth must decrease sufficiently slowly to ensure the convergence $f_n \to f$. On the other hand, a drawback of DS-MLE is its computational complexity. In [10], the authors propose using Monte-Carlo estimation, which is computationally demanding. In contrast, our pseudo-likelihood function can be easily computed and optimized using standard optimization tools, see [5]. From a theoretical perspective, we believe that keeping the pseudo-likelihood as close as possible to the likelihood helps preserve the desirable properties of the standard MLE. Simulations in [5] suggest that this may indeed be the case.

1.3 Generalization beyond the i.i.d. case

Throughout the paper, we assume that Y_1, Y_2, \ldots are i.i.d. observations from a Gaussian mixture distribution. When inspecting the consistency proof, it becomes evident that the assumption of independent observations is used to apply the (uniform) strong law of large numbers, to ensure almost sure weak convergence of empirical measures and to guarantee almost sure convergence $||f_n||_{\infty} \to ||f||_{\infty}$, where f denotes the true density. However, all these convergence results also hold in more general settings, suggesting that the consistency theorem may extend beyond the i.i.d. mixture case to more general latent variable models. In particular, the following model is of interest. Let X_1, X_2, \ldots be a stationary ergodic process taking values in $\{1,\ldots,k\}$ and let the observations Y_1,Y_2,\ldots be as follows: 1) given X_1, X_2, \ldots , the observations are (conditionally) independent; 2) given $X_t = i$, the observation Y_t has a Gaussian distribution with parameter θ_i . Such models are commonly used in many applications, where the latent X-process represents an underlying signal, and the observed Y-process models the signal corrupted by Gaussian noise. A classical example of such a model is a hidden Markov model, where the X-process is a Markov chain. In such a model, the weights w_i^* are the probabilities $P(X_t = i)$ and the parameters θ_i are typically called the emission parameters. When estimating the emission parameters, the order of observations does not matter; thus, one can still use the pseudo-likelihood $L_n(\theta)$ as defined above even when the model is not an i.i.d. mixture model any more. Consistency in this context refers to the convergence of emission parameters as well as the marginal distribution of X_t . Due to the ergodicity, we conjecture that the consistency holds and the pseudo-likelihood method is justified for more general models than just i.i.d. mixtures. For maximum likelihood estimation, [7] proved that the maximum likelihood estimator of the parameters of a finite mixture distribution obtained under the assumption of independence (that is, ignoring the actual dependence structure) is consistent and asymptotically normally distributed when the regime process is an ergodic Markov chain. For the maximum spacing estimator, consistency of the estimator for the marginal parameters in hidden Markov models was established in [6].

2 Consistency of pseudo-likelihood estimator

2.1 Setting

Let $\Theta_o = (\mathbb{R} \times (0, \infty))^k$ be the parameter space. For every $\theta = (\theta_1, \dots, \theta_k) \in \Theta_o$, let $g(\theta_i, \cdot)$ stand for Gaussian density with parameter $\theta_i = (\mu_i, \sigma_i)$. It may happen that some components coincide, so let $s(\theta) \leq k$ be the number of different components. We shall identify all vectors θ with the same set of distinct components as a single parameter, thus Θ_o should be considered as the set of equivalence classes. For example, all permutations of θ are equivalent. When $s(\theta) = k$, then an equivalence class consists of only permutations, otherwise the class is larger. As a representative of an equivalence class, we consider θ under a natural ordering of the parameters: $\mu_1 \leq \ldots \leq \mu_k$, and, in cases where some of the means are equal, the ordering is determined by the corresponding variances. When we discuss uniqueness or equality of parameter vectors, the natural ordering is assumed.

We shall assume that the true parameter θ^* is such that all components are different, i.e., $s(\theta^*) = k$. We denote the true density by f, i.e., $f(\cdot) = \sum_{i=1}^k w_i^* g(\theta_i^*, \cdot)$, where $w_i^* > 0$, $i = 1, \ldots, k$. The requirement that $w_i^* > 0$ for every $i = 1, \ldots, k$ ensures that there exists no other parameter $\theta \neq \theta^*$ and weights w such that $\sum_i w_i g(\theta_i, \cdot) = f(\cdot)$. This follows from the identifiability of Gaussian mixtures (see, e.g., [15, 4]). The condition is also necessary for uniqueness, since when some weights of w^* equal to zero, then there exists $\theta \neq \theta^*$ and weights w such that $\sum_i w_i g(\theta_i, \cdot) = f(\cdot)$ – the true parameter would not be unique. For example, when $w_1^* = 0$, then θ could be taken as $(\theta_2^*, \theta_2^*, \theta_3^*, \ldots, \theta_k^*) \neq \theta^*$ and w could be taken as $(w_2^*/2, w_2^*/2, w_3^*, \ldots, w_k^*)$.

Since any θ is an equivalence class, the convergence $\theta^n \to \theta$ is also a convergence between equivalence classes. For that we consider every class as a set and define the convergence between the parameters (classes) as the convergence between the sets in Hausdorff's sense. In our notation, when $\theta = (\theta_1, \dots, \theta_k)$ and $\theta' = (\theta'_1, \dots, \theta'_k)$ are two vectors in Θ_o , then the Hausdorff distance $h(\theta, \theta')$ is defined as

$$h(\theta, \theta') := \max\{ \max_{i} \min_{j} \|\theta_i - \theta'_j\|, \max_{i} \min_{j} \|\theta'_i - \theta_j\| \}.$$

Clearly, $h(\theta, \theta') = 0$ if and only if θ and θ' are in the same class. Note that the convergence $\theta^n \to \theta$ is equivalent to the condition that every subsequence $\theta^{n'}$ has a further subsequence $\theta^{n''}$, which converges component-wise to a representative of the equivalence class of θ . Therefore, we work mostly with point-wise convergent parameter sequences in this paper.

It is important to understand that we can not replace our notion of convergence of sequences of parameter vectors with point-wise convergence of naturally ordered representatives: clearly we want to say that the sequence $((\frac{(-1)^n}{n}, 1), (\frac{(-1)^{n+1}}{n}, 2))$ converges to ((0, 1), (0, 2)), but the sequence of naturally ordered representatives does not converge point-wise and converging subsequences converge to different representatives of the equivalence class of the limiting vector.

Recall that S_k stands for the (k-1)-dimensional simplex. For r < k, let S_r denote the (r-1)-dimensional simplex. Recall the definition of weights $v^n(\theta)$ in (1), where $\theta \in \Theta_o$ and \hat{f}_n is any density function in L_2 . Throughout the article, for any $w \in S_k$ and for any k-dimensional vector of Gaussian densities $g = (g_1, \ldots, g_k)$, we denote $wg := \sum_{i=1}^k w_i g_i$. The following lemma guarantees that $v^n(\theta)$ always exists and that the corresponding density $v^n g$ is unique.

Lemma 2.1 Let $f, g_1, \ldots, g_k \in L_2$. Then there always exists at least one $v \in S_k$ such that $||f - vg|| = \inf_{w \in S_k} ||f - wg||$. If v_1 and v_2 are two such vectors, then $v_1g = v_2g$.

Proof. We show that a minimizer exists. The existence of v follows from the compactness of the simplex S_k and the continuity of $w \mapsto ||f - wg||$. If v_1 and v_2 are two different minimizers such that $v_1g \neq v_2g$, the strict convexity of L_2 norm would imply that for any $\lambda \in (0,1)$,

$$||f - (\lambda v_1 + (1 - \lambda)v_2)g|| = ||\lambda f - \lambda v_1 g + (1 - \lambda)f - (1 - \lambda)v_2 g|| < \lambda ||f - v_1 g|| + (1 - \lambda)||f - v_2 g||.$$

That contradicts the definition of v_1 and v_2 .

Lemma 2.1 ensures that for any θ , the solution of $\inf_{w \in S_k} \|f(\cdot) - wg(\theta, \cdot)\|$, let it be $v(\theta)$, always exists, but when $s(\theta) < k$, then it is not necessarily unique. However, given θ , the density $\sum_{i=1}^k v_i(\theta)g(\theta_i, \cdot)$ is always unique. Therefore, if $s(\theta) = k$, then there are no other solutions v' satisfying $\sum_{i=1}^k v_i(\theta)g(\theta_i, \cdot) = \sum_{i=1}^k v_i'g(\theta_i, \cdot)$. This follows from the identifiability of Gaussian mixtures – when $g(\theta_i, \cdot) \neq g(\theta_j, \cdot)$ for all $i \neq j$, then $w_1g = w_2g$ would imply $w_1 = w_2$ (recall that we have fixed the ordering). Therefore, our assumption $s(\theta^*) = k$ guarantees the uniqueness of w^* . The identifiability also implies that when v is any minimizer of $\inf_{w \in S_k} \|f - wg\|$, then $\sigma_o = \max\{\sigma_i : v_i > 0\}$ is unique, i.e., independent of the choice of a particular minimizer. Let us remark that we are not aware of the closed form representation of v^n except for the special case k = 2 (see [5], (6)).

2.2 Consistency theorem

Let $Y_1, Y_2, ...$ be a sequence of i.i.d. random variables with true density $f(\cdot) = w^*g(\theta^*, \cdot)$, where $\theta^* \in \Theta_o$ and $w^* = (w_1^*, ..., w_k^*)$ are the corresponding strictly positive weights. Given a nonparametric density estimator \hat{f}_n and $g(\theta) := (g_1(\theta), ..., g_k(\theta))$, denote

$$v^{n}(\theta) = \arg\inf_{w \in S_{k}} \|\hat{f}_{n} - wg(\theta)\|, \quad v(\theta) = \arg\inf_{w \in S_{k}} \|f - wg(\theta)\|.$$
 (2)

For every $\theta \in \Theta_o$, define the log-pseudo-likelihood function $\ell_n(\theta)$ as follows:

$$\ell_n(y,\theta) := \ln \left(v^n(\theta) g(\theta, y) \right), \quad \ell(y,\theta) := \ln \left(v(\theta) g(\theta, y) \right),$$
$$\ell_n(\theta) := \frac{1}{n} \sum_{t=1}^n \ell_n(Y_t, \theta), \quad \ell(\theta) := E\ell(Y_1, \theta).$$

Sometimes, to stress the dependence of $\ell_n(\theta)$ on $Y_1(\omega), \ldots, Y_n(\omega)$, we use the notation $\ell_n^{\omega}(\theta)$. To keep the technique simpler, throughout the paper we ignore the cases where ℓ_n is unbounded (it happens with probability zero). Recall that $v^n(\theta)g(\theta,\cdot)$ and $v(\theta)g(\theta,\cdot)$ are unique even when $v^n(\theta)$ or $v(\theta)$ are not. By our assumptions on f (that is, $s(\theta^*) = k$ and $w_i^* > 0$ for every i), for any $w \in S_k$ and for any $\theta \in \Theta_o$ such that $\theta \neq \theta^*$, it holds that $f(\cdot) \neq wg(\theta,\cdot)$, and thus by Gibb's inequality

$$\int f(y)\ln(wg(\theta,y))dy < \int f(y)\ln f(y)dy = \int f(y)\ln \left(w(\theta^*)g(\theta^*,y)\right)dy = \ell(\theta^*).$$

Hence, for any $\theta \neq \theta^*$, it holds that $\ell(\theta) < \ell(\theta^*)$. We are interested in consistency of the pseudo-likelihood estimator $\hat{\theta}^n$, where $\hat{\theta}^n$ is for $\epsilon_n \searrow 0$ defined so that

$$\ell_n(\hat{\theta}^n) \ge \sup_{\theta \in \Theta^o} \ell_n(\theta) - \epsilon_n. \tag{3}$$

Let $\hat{\theta}^n = ((\mu_{1,n}, \sigma_{1,n}), \dots, (\mu_{k,n}, \sigma_{k,n}))$. Sometimes, to stress the dependence on ω , we denote $\hat{\theta}^n_{\omega}$. The main result of the article is the following consistency theorem. In the following, the convergence between functions means convergence in the L_2 norm if not stated otherwise.

Theorem 2.1 Assume that $\hat{f}_n \stackrel{a.s.}{\to} f$ (in L_2) and $\exists C < \infty$ so that $P(\|\hat{f}_n\|_{\infty} < C$ eventually) = 1. Then the following convergences hold:

$$\hat{\theta}^n \overset{a.s.}{\to} \theta^*, \quad v^n(\hat{\theta}^n) \overset{a.s.}{\to} w^*, \quad v^n(\hat{\theta}^n)g(\hat{\theta}^n, \cdot) \overset{a.s.}{\to} f(\cdot).$$

In the theorem, we do not assume that f_n is a kernel estimator, although in practice it is the most natural choice. Since we deal with the estimation of normal mixtures, it is natural to take \hat{f}_n as the Gaussian kernel estimator. When the bandwidth of Gaussian kernel estimator tends to zero sufficiently slowly, then $\|\hat{f}_n - f\|_{\infty} \stackrel{a.s.}{\to} 0$ [12, 16], so that the assumptions on \hat{f}_n are fulfilled.

2.3 About the proof

In one way or another, all consistency proofs rely on (relative) compactness of the parameter space. Perhaps the most direct and well-known example of this is the famous Wald consistency proof (see, e.g., [1, 17]). Another common use of compactness is to establish the uniform convergence $\sup_{\theta} |\ell_n(\theta) - \ell(\theta)| \to 0$ almost surely, and then to use the fact that, on a compact space, uniform convergence implies the convergence of the maximizers (i.e., M-estimators). This is how we prove the consistency in the current article. Although this approach is standard and widely used, applying it in the present setting involves several technical difficulties.

Unbounded means and vanishing or unbounded variances. First, the compactification of the parameter space Θ_o includes zero and infinite variances, as well as infinite means. In the case of Gaussian distributions, zero variances are particularly problematic. Even in the case of i.i.d. Gaussian random variables, one needs to apply the so-called Kiefer-Wolfowitz trick to handle vanishing variances when using the Wald consistency proof (see, e.g., (5.15) in [17]).

To handle unbounded means and vanishing or unbounded variances, we start by showing that at least one component of $\hat{\theta}_n$ is such that its variance is bounded away from zero and above and its mean is bounded as well. In particular, we show that there exist constants $0 < u < U < \infty$ and $N < \infty$ (depending only on the true density) such that, for all sufficiently large n, there exists – with probability one – a component i(n) for which $|\mu^n_{i(n)}| < N$ and $u \le \sigma^n_{i(n)} \le U$. This property is stated as Proposition 3.1 and proved in Section 3. The proof uses some ideas from the proof of Lemma 3.1 in [1]. Proposition 3.1 ensures that for every convergent subsequence $\hat{\theta}^{n'} \to \theta$, the limit θ also contains a component whose parameters are bounded as described above. It also allows us to reduce the parameter space so that the parameters of at least one component are bounded as described above; this set is denoted by $\Theta_o(u, U, N)$.

Uniform convergence of the criterion function. The next step is to show that the uniform convergence of the criterion function holds over $\Theta_o(u,U,N)$, see (25). Proving the uniform convergence is the main technical challenge of this article. First, observe that even for a fixed parameter θ , we cannot directly apply the strong law of large numbers (SLLN) to deduce the convergence $\ell_n(\theta) \stackrel{a.s.}{\to} \ell(\theta)$. This is because the weights $v^n(\theta)$ depend on \hat{f}_n , and thus also on ω . Thus, the standard SLLN does not apply in our case, and we must generalize it to accommodate the pseudo-likelihood setting as well. This generalization is formalized in Lemma 4.1, which makes use of the Skorohod representation theorem. Lemma 4.1 yields pointwise convergence $\ell_n(\theta) \stackrel{a.s.}{\to} \ell(\theta)$ for fixed θ , but we also need the convergence of $\ell(\theta_n)$ for sequences $\theta_n \to \theta$, where the limit θ may involve zero or infinite variances and/or infinite means.

All possible limits beyond Θ_o require special treatment and, to some extent, novel techniques. These issues are addressed in Section 5. The main result of that section is Proposition 5.1, which, together with Lemma 4.1, leads to the uniform convergence result via Proposition 6.1. Once uniform convergence is established, the final consistency proof becomes standard, it is presented as the concluding argument of Section 6.

To recapitulate, from a broad perspective, our proof follows a standard path. However, almost every step along the way requires specific and largely novel techniques. The main difficulties arise from the fact that the estimates for the weights and parameters of the components are obtained by combining two different criterion functions (L_2 distance and likelihood). At the same time, we believe that this property – applying two different criterion functions to obtain parameter estimates – is one of the reasons behind the strong empirical performance of our estimator.

3 Proof that $\hat{\theta}^n$ belongs to $\Theta_o(u, U, N)$

Let $0 < u < U < \infty$ and $0 < N < \infty$ be fixed. Define

$$\Delta(u, U, N) := \mathbb{R} \times \mathbb{R}^+ \setminus [-N, N] \times [u, U], \quad \Theta_o(u, U, N) := \Theta_o \setminus (\Delta(u, U, N))^k. \tag{4}$$

Thus, $\theta \in \Theta_o(u, U, N)$ if and only if there exists i such that $\sigma_i \in (u, U)$ and $|\mu_i| \leq N$. Let $\Theta(u, U, N)$ be the closure of $\Theta_o(u, U, N)$.

The following lemma was proved in [5].

Lemma 3.1 Let $f \in L_2$ and let g_1, \ldots, g_k be Gaussian densities. Let $v = (v_1, \ldots, v_k)$ be any minimizer of (2) for given θ . Denote $\sigma_o = \max\{\sigma_i : v_i > 0\}$. Then

$$\frac{v_i}{\sigma_i} \le a + \frac{b}{\sigma_o}, \quad a = 2\sqrt{\pi} ||f||_{\infty}, \quad b = 2\sqrt{2}. \tag{5}$$

Throughout this section, we assume that there exists $C < \infty$ such that $P(\|\hat{f}_n\|_{\infty} \leq C \text{ eventually}) = 1$. In particular, this holds when $\|\hat{f}_n\|_{\infty} \stackrel{a.s.}{\to} \|f\|_{\infty}$. The latter holds when $\|\hat{f}_n - f\|_{\infty} \stackrel{a.s.}{\to} 0$, let Ω_o denote the corresponding set. When such a C exists, then by Lemma 3.1, we can assume without loss of generality the existence of universal constants a > 0 and b > 0 (depending on $\|f\|_{\infty}$), such that for every $\theta \in \Theta_o$ and for every $\omega \in \Omega_o$,

$$\frac{v_i^n(\theta)}{\sqrt{2\pi}\sigma_i} \le \left(a + \frac{b}{\sigma_0^n}\right), \quad \sigma_0^n := \max\{\sigma_i : v_i^n(\theta) > 0\},\tag{6}$$

provided $n > n_o(\omega)$.

For every u > 0, define the functions U(u) and N(u) as follows:

$$\frac{1}{\sqrt{2\pi}aU(u)} = e^{-\frac{1}{u}}, \quad \exp\left(\frac{-N^2(u)}{8U^2(u)}\right) = e^{-\frac{1}{u}}.$$
 (7)

Observe that both functions are decreasing in u and $\lim_{u\to 0} N(u) = \lim_{u\to 0} U(u) = \infty$. Let Y be a random variable with true distribution, consider

$$r_1(u) := 1 - P(|Y| > N(u)/2) - k||f||_{\infty} 2\sqrt{2u}.$$

Then $r_1(u) \nearrow 1$ in the process $u \searrow 0$. Thus, there exists u_o such that $r_1(u) \ge 3/4$, whenever $u \le u_o$. Define

$$r_2(u) := \sup_{z \in (0,u)} \left[\ln \left(a + \frac{b}{z} \right) + \ln k - \frac{1}{2z} \right].$$

Since $\ln\left(a+\frac{b}{z}\right)+\ln k-\frac{1}{2z}\to -\infty$ as $z\to 0$, there exists u_W for every $-W>-\infty$ such that $r_2(u)\leq -W$, whenever $u\leq u_W$. Take $-W<-\ln\left(\operatorname{Var}Y\right)/2-2$ and fix the constants u,U,N as follows:

$$0 < u < \min\{u_W, u_o, U(u)\}, \quad U := U(u), \quad N := N(u). \tag{8}$$

Observe that the choice of u, U, N depends solely on the true density f. For every $0 < u < U < \infty$ and N, define

$$\bar{\Theta}(u, U, N) := \{ \theta \in \Theta_o : \text{there exists a partition } \{ J_1, J_2, J_3 \} \text{ of } \{ 1, \dots, k \}$$
such that $\max_{i \in J_1} \sigma_i \leq u, \min_{i \in J_2} \sigma_i \geq U;$

$$\sigma_i \in (u, U), i \in J_3; \min_{i \in J_3} |\mu_i| > N \}.$$

Note that some of the sets in partition $\{J_1, J_2, J_3\}$ can be empty.

Proposition 3.1 With u, U and N defined as in (8), the following holds:

$$P(\limsup_{n} \{\hat{\theta}^{n} \in \bar{\Theta}(u, U, N)\}) = P(\hat{\theta}^{n} \in \bar{\Theta}(u, U, N) \text{ i.o.}) = 0, \tag{9}$$

thus

$$P(\hat{\theta}^n \in \Theta_o(u, U, N) \text{ eventually}) = 1.$$

Proof. Let u, U, N be defined as in (8). Fix $\omega \in \Omega_o$ and $n_o(\omega)$ so that (6) holds. Consider $\theta \in \overline{\Theta}(u, U, N)$ and let $\{J_1, J_2, J_3\}$ be the corresponding partition (depending on θ). Take any $v^n(\theta) = (v_1^n(\theta), \dots, v_k^n(\theta))$ minimizing (1), and denote $c_i^n(\theta) = \frac{v_i^n(\theta)}{\sqrt{2\pi}\sigma_i}$ and $c_0^n(\theta) = a + \frac{b}{\sigma_0^n}$. Then

$$\ln\left(v^n(\theta)g(\theta,y)\right) \le \ln c_0^n(\theta) + \ln k + \ln \max_i \left(\frac{c_i^n(\theta)}{c_0^n(\theta)} \exp\left(\frac{-(y-\mu_i)^2}{2\sigma_i^2}\right)\right).$$

Define

$$A_{\theta} = \{ y : |y| \le N/2; |y - \mu_i|^2 \ge 2\sigma_i, i \in J_1 \},$$

then due to (6) and by the choice of N and U,

$$\frac{c_i^n(\theta)}{c_0^n(\theta)} \exp\left(\frac{-(y-\mu_i)^2}{2\sigma_i^2}\right) \leq \begin{cases}
1, & y \notin A_{\theta}, \\
e^{-\frac{1}{\sigma_i}}, & y \in A_{\theta}, i \in J_1, \\
\frac{1}{\sqrt{2\pi}aU} = e^{-\frac{1}{u}}, & y \in A_{\theta}, i \in J_2, \\
\exp(\frac{-N^2}{8U^2}) = e^{-\frac{1}{u}}, & y \in A_{\theta}, i \in J_3.
\end{cases}$$

Since for every $i \in J_1$ such that $v_i^n > 0$,

$$e^{-\frac{1}{\sigma_i}} \le e^{-\frac{1}{\min(\sigma_0^n, u)}}.$$

and in the case $\sigma_0^n < u$ the sets J_2 and J_3 are empty or all weights are 0 for the components from those sets, we get

$$\ln\left(v^n(\theta)g(\theta,y)\right) \le \ln\left(a + \frac{b}{\min(u,\sigma_0^n)}\right) + \ln k - \frac{1}{\min(u,\sigma_0^n)}I_{\{y\in A_\theta\}}(y). \tag{10}$$

To recapitulate: we have shown that for any $\omega \in \Omega_o$ and for any $\theta \in \bar{\Theta}(u, U, N)$, the upper bound (10) holds, provided $n > n_o(\omega)$. Due to our choice of $u, P(A_\theta) \ge 3/4$, because

$$P(A_{\theta}) \ge 1 - P(|Y| > N/2) - |J_1| ||f||_{\infty} 2\sqrt{2u}$$

$$\geq 1 - P(|Y| > N/2) - k||f||_{\infty} 2\sqrt{2u} = r_1(u) \geq 3/4.$$

Since A_{θ} consists of at most k+1 intervals, the Glivenko-Cantelli theorem gives that the following inequality holds almost surely (let the corresponding set be Ω_{GC}):

$$\inf_{\theta \in \bar{\Theta}(u,U,N)} P_n(A_{\theta}) \ge \frac{1}{2} \text{ eventually.}$$

It follows by (10) that when $\omega \in \Omega_{GC} \cap \Omega_o$, then

$$\sup_{\theta \in \bar{\Theta}(u,U,N)} \ell_n(\theta) \le \sup_{z \in (0,u)} \left[\ln \left(a + \frac{b}{z} \right) + \ln k - \frac{1}{2z} \right] = r_2(u) \le -W \text{ eventually.}$$
 (11)

On the other hand, by taking $\theta_0^n = ((\mu_n, S_n), \dots, (\mu_n, S_n))$ (all components are equal), where $\mu_n = \frac{1}{n} \sum_{t=1}^n Y_t$ is the sample mean and $S_n^2 = \frac{1}{n} \sum_{t=1}^n (Y_t - \mu_n)^2$ is the sample variance, we obtain

$$\ell_n(\theta_0^n) = -\ln(\sqrt{2\pi}) - \ln(S_n) - \frac{1}{2} \ge -\ln(S_n) - 2. \tag{12}$$

By SLLN, $\ln(S_n) \stackrel{a.s.}{\to} \frac{1}{2} \ln (\operatorname{Var}(Y))$, thus

$$P\left(\ell_n(\hat{\theta}^n) \ge -\frac{1}{2}\ln\left(\operatorname{Var}(Y)\right) - 2 \text{ eventually}\right) = 1.$$
 (13)

Let Ω_V be the corresponding set. Recall that $-W < -\frac{1}{2}\ln\left(\operatorname{Var}(Y)\right) - 2$. Let $\omega \in \Omega_{GC} \cap \Omega_V \cap \Omega_o$. If the corresponding $\hat{\theta}^n_{\omega}$ is such that along a subsequence, $\hat{\theta}^{n'}_{\omega} \in \bar{\Theta}(u, U, N)$, then by (11), $\limsup_n \ell_n^{\omega}(\hat{\theta}^n_{\omega}) \leq -W$ – a contradiction. Hence,

$$\lim\sup_{n} \{\hat{\theta}^{n} \in \bar{\Theta}(u, U, N)\} \subset \Omega_{V}^{c} \cup \Omega_{GC}^{c} \cup \Omega_{o}^{c},$$

and $P(\hat{\theta}^n \in \Theta_o(u, U, N) \text{ eventually}) = 1 \text{ follows.}$

4 Modification of SLLN

The following lemma generalizes SLLN. Note that for $h_n \equiv h$, (14) reduces to the standard SLLN.

Lemma 4.1 Let P be a probability measure, and let h_n and h be functions such that for P-a.e. $y, y_n \to y$ implies $h_n(y_n) \to h(y)$. Let Y_1, Y_2, \ldots be a sequence of i.i.d. observations with distribution P, and let H be a continuous function such that $EH(Y_1) = \int H(y)P(dy) < \infty$. If $|h_n(y)| \leq H(y)$ for every n and $y \in \mathbb{R}$, then

$$\frac{1}{n} \sum_{t=1}^{n} h_n(Y_t) \stackrel{a.s.}{\to} Eh(Y_1). \tag{14}$$

When $h \equiv -\infty$, and $h_n(y) \leq H(y)$ for every n and $y \in \mathbb{R}$, then

$$\frac{1}{n} \sum_{t=1}^{n} h_n(Y_t) \stackrel{a.s.}{\to} -\infty. \tag{15}$$

Moreover, the set, where (14) and (15) hold is

$${P_n \Rightarrow P} \cap \left\{ \int H(y)P_n(dy) \to \int H(y)P(dy) \right\},$$

where P_n is the empirical measure corresponding to Y_1, \ldots, Y_n .

Proof. Let $P_n \Rightarrow P$ and $\int H(y)P_n(dy) \to \int H(y)P(dy)$. Let now Z_n and Z be random variables such that $Z_n \sim P_n$, $Z \sim P$ and $Z_n \stackrel{a.s.}{\to} Z$. By the Skorohod representation theorem, such random variables exist. Then $h_n(Z_n) \stackrel{a.s.}{\to} h(Z)$, $0 \le h_n(Z_n) + H(Z_n) \to H(Z) + h(Z)$, and $EH(Z_n) \to EH(Z)$, so by Fatou

$$E(h(Z) + H(Z)) \le \liminf_{n} (Eh_n(Z_n) + EH(Z_n)),$$

thus

$$Eh(Z) \le \liminf_n Eh_n(Z_n).$$

By Fatou, again,

$$E(H(Z) - h(Z)) \le \liminf_{n} E(H(Z_n) - h_n(Z_n)) = EH(Z) - \limsup_{n} Eh_n(Z_n), \quad (16)$$

so that $Eh_n(Z_n) \to Eh(Z)$. This establishes (14).

When $h \equiv \infty$, then by (16),

$$\infty \le \liminf_n E(H(Z_n) - h_n(Z_n)) = EH(Z) - \limsup_n Eh_n(Z_n),$$

so that $\limsup_n Eh_n(Z_n) \leq -\infty$. Since $P_n \Rightarrow P$ almost surely, and by SLLN, $\int H(y)P_n(dy) \to \int H(y)P(dy)$ almost surely, (14) and (15) hold with probability 1.

Corollary 4.1 Suppose that $\{h_n^{\omega}\}$ is a random sequence of measurable functions. Let Ω_o be the set such that, for every $\omega \in \Omega_o$, the following convergences hold: $P_n^{\omega} \Rightarrow P$; for P-a.e. y the convergence $y_n \to y$ implies $h_n^{\omega}(y_n) \to h(y)$; $|h_n^{\omega}(y)| \leq H^{\omega}(y)$, where H^{ω} is continuous, and $\int H^{\omega} dP_n^{\omega} \to \int H^{\omega} dP$. Then $\forall \omega \in \Omega_o$,

$$\frac{1}{n} \sum_{t=1}^{n} h_n^{\omega}(Y_t) \to Eh(Y_1). \tag{17}$$

5 Approximation of f_n for a given set of normal components

In this section, we shall consider k sequences of normal densities $g_i^n := g(\mu_{i,n}, \sigma_{i,n}; \cdot)$ such that for every $i \in \{1, \ldots, k\}$, the following limits exist:

$$\sigma_i = \lim_n \sigma_{i,n} \in [0, \infty], \quad \mu_i = \lim_n \mu_{i,n} \in [-\infty, \infty].$$

We also assume that for every i and j, the limit $\lim_n (\mu_{i,n} - \mu_{j,n}) \in [-\infty, \infty]$ exists. This assumption is automatically fulfilled when $|\mu_i| < \infty$ and $|\mu_j| < \infty$. It is important to realize that for any sequence of k normal densities, one can choose a subsequence such that all these limits exist.

Let $f_n \to f$ be any convergent sequence. We consider an approximation of f_n with $v^n g^n$, where the weights v^n are defined as

$$v^{n} := \arg \inf_{w \in S_{k}} \|f_{n} - wg^{n}\|.$$
 (18)

Recall that v^n is not necessarily unique, but $v^n g^n$ is. In the main proposition of this section, we will study the convergence of $v^n g^n$.

Define the following partition of the set of component indexes $\{1, \ldots, k\}$:

$$I_0 := \{i : \sigma_i \in (0, \infty), |\mu_i| < \infty\}, \quad r_0 := |I_0|;$$

$$I_1 := \{i : \sigma_i \in (0, \infty), |\mu_i| = \infty\}, \quad r_1 := |I_1|;$$

$$I_2 := \{i : \sigma_i = \infty\}, \quad r_2 := |I_2|;$$

$$I_3 := \{i : \sigma_i = 0\}, \quad r_3 := |I_3|.$$

Thus, I_0 consists of indexes such that variances converge to nondegenerate limits and means converge too; I_1 consists of indexes, where variances converge to nondegenerate limits, but the means diverge; I_2 consists of indexes, where variances diverge; and I_3 is the set of indexes, where variances tend to 0. For any $i \in I_0$, we denote $g_i := g(\mu_i, \sigma_i; \cdot)$, thus $g_i^n \to g_i$ for every $i \in I_0$. Since

$$||g_i^n||^2 = \frac{1}{2\sqrt{\pi}} \cdot \frac{1}{\sigma_{i,n}},$$

we see that

$$g_i^n \to 0, \quad \forall i \in I_2.$$
 (19)

Furthermore, note that for $i \in I_1$, the sequence of norms $||g_i^n||$ is bounded and $g_i^n(y) \to 0$ for every y, thus g_i^n converges weakly to 0 in L_2 (denoted by $g_i^n \to 0$). We shall denote

$$k_{ij}^n := \langle g_i^n, g_j^n \rangle = \frac{1}{\sqrt{2\pi(\sigma_{i,n}^2 + \sigma_{j,n}^2)}} \exp\left[-\frac{(\mu_{i,n} - \mu_{j,n})^2}{2(\sigma_{i,n}^2 + \sigma_{j,n}^2)}\right].$$

When $i, j \in I_0$, then $k_{ij}^n \to k_{ij} := \langle g_i, g_j \rangle$; when $i \in I_0$ and $j \in I_1$, then $k_{ij}^n \to 0$. Due to our additional assumption about the existence of the limit $\lim_n (\mu_{i,n} - \mu_{j,n})$, clearly $k_{ij}^n \to k_{ij}$ also when $i, j \in I_1$. Thus, the Gram matrix $K^n = (k_{ij}^n)_{i,j \in I_1}$ converges entry-wise to $K = (k_{ij})_{i,j \in I_1}$.

Throughout this section, we shall assume that $r_0 > 0$, that is, $\exists i \in \{1, ..., k\}$ such that $g_i^n \to g_i$. Without loss of generality, we denote this index by 1, so we shall assume that $1 \in I_0$.

Let us introduce some necessary notation. For any integer $r \geq 1$, let S_r be the (r-1)-dimensional simplex. Recall that for every vector $w = (w_1, \ldots, w_k) \in S_k$ and densities $g = (g_1, \ldots, g_k)$, we shall denote $wg := \sum_i w_i g_i$. For any subset $I \subset \{1, \ldots, k\}$ and for any vector $v \in S_k$, we denote the restrictions of vg as follows:

$$v_I g := \sum_{i \in I} v_i g_i, \quad v_I := (v_i)_{i \in I}.$$

Recall that $g_i^n \to g_i$ for every $i \in I_0$. Let

$$t: S_{r_0+r_1+r_2} \to \mathbb{R}^+, \quad t(w) := \|f - w_{I_0}g\|^2 + \sum_{i,j \in I_1} w_i w_j k_{ij},$$

$$u := \arg \min_{w \in S_{r_0+r_1+r_2}} t(w),$$

$$t^* := t(u).$$

Note that when $r_2 > 0$, we have $u_i = 0$ for $i \in I_1$, therefore u_{I_0} can also be found by

$$u_{I_0} = \arg \inf_{w_i \ge 0, \sum_{i \in I_0} w_i \le 1} ||f - w_{I_0}g||.$$

The next lemma proves that the approximation $u_{I_0}g$ of the true density f is unique.

Lemma 5.1 The function $u_{I_0}g$ is unique.

Proof. Note that $S_{r_0+r_1+r_2}$ is a convex set and that $K = (k_{ij})_{i,j \in I_1}$ is a symmetric and positive semidefinite matrix. Assume that a and b are two $(r_0 + r_1 + r_2)$ -dimensional vectors. Then, using the notation $a \cdot b$ for the scalar product between two vectors, we have the following equality:

$$\frac{t(a) + t(b)}{2} - t\left(\frac{a+b}{2}\right) = \frac{\|f - a_{I_0}g\|^2 + \|f - b_{I_0}g\|^2 - 2\|f - \frac{a_{I_0} + b_{I_0}}{2}g\|^2}{2} + \frac{2Ka_{I_1} \cdot a_{I_1} + 2Kb_{I_1} \cdot b_{I_1} - K(a_{I_1} + b_{I_1}) \cdot (a_{I_1} + b_{I_1})}{4}$$

As for any $x, y \in L_2$ we have

$$||x + y||^2 + ||x - y||^2 = 2(||x||^2 + ||y||^2),$$

we get for $x = f - a_{I_0}g$ and $y = f - b_{I_0}g$ the equality

$$\frac{\|f - a_{I_0}g\|^2 + \|f - b_{I_0}g\|^2 - 2\|f - \frac{a_{I_0} + b_{I_0}}{2}g\|^2}{2} = \frac{1}{4}\|a_{I_0}g - b_{I_0}g\|^2.$$

Similarly, for any symmetric matrix M and any vectors v and w, we have

$$M(v+w) \cdot (v+w) + M(v-w) \cdot (v-w) = 2(Mv \cdot v + Mw \cdot w),$$

therefore,

$$\frac{2Ka_{I_1} \cdot a_{I_1} + 2Kb_{I_1} \cdot b_{I_1} - K(a_{I_1} + b_{I_1}) \cdot (a_{I_1} + b_{I_1})}{4} = \frac{1}{4}K(a_{I_1} - b_{I_1}) \cdot (a_{I_1} - b_{I_1}).$$

Thus, we have shown that

$$\frac{t(a)+t(b)}{2}-t\left(\frac{a+b}{2}\right)=\frac{1}{4}\left(\|a_{I_0}g-b_{I_0}g\|^2+K(a_{I_1}-b_{I_1})\cdot(a_{I_1}-b_{I_1})\right).$$

Therefore, if a and b are two different minimum points of $t(\cdot)$ in a convex region, it follows from properties of K that the second term is non-negative, and thus necessarily $a_{I_0}g = b_{I_0}g$.

Recall an important bound from Lemma 3.1: for every $i \in \{1, ..., k\}$,

$$\frac{v_i^n}{\sigma_{i,n}} \le a_n + \frac{b}{\sigma_{0,n}}, \quad a_n = 2\sqrt{\pi} \|f_n\|_{\infty}, \quad b = 2\sqrt{2}, \quad \sigma_{0,n} := \max\{\sigma_{i,n} : v_i^n > 0\}. \tag{20}$$

In what follows, we shall assume that $\sup_n ||f_n||_{\infty} < \infty$ and therefore, the constant a_n in (20) can be chosen independently of n, so we shall use $a < \infty$ instead of a_n .

The following auxiliary lemma will be needed in the proof of the main proposition of this section.

Lemma 5.2 Let $a^n \in S_k$ be an arbitrary sequence of weights such that the sequence $||f_n - a^n g^n||$ is bounded above. Then $\sum_{i \in I_2} a_i^n \to 0$.

Proof. By assumption, $||f_n - a^n g^n||$ is bounded above. The reverse triangular inequality $||f_n - a^n g^n|| \ge ||a^n g^n|| - ||f_n||$ implies the boundedness of $||a^n g^n||$. Now, because of nonnegativity of all terms, for every n and i,

$$||a^n g^n|| \ge a_i^n ||g_i^n||.$$

Since $\forall i \in I_3$, $||g_i^n|| \to \infty$, it follows that $a_i^n \to 0$.

Corollary 5.1 For any choice of v^n , $\sum_{i \in I_3} v_i^n \to 0$.

Proof. By assumption, $g_1^n \to g_1$. Then $||f_n - v^n g^n|| \le ||f_n - g_1^n|| \to ||f - g_1||$, so that $||f_n - v^n g^n||$ is bounded above. Thus, the assumptions of Lemma 5.2 with $a^n = v^n$ are satisfied.

5.1 Convergence of the function $v^n g^n$

Let v^n be any vector of weights minimizing (18). The main result of the present section is the following proposition.

Proposition 5.1 Let $f_n \to f$ and $\sup_n ||f_n||_{\infty} < \infty$. Let $y_n \to y$ be a convergent sequence such that $y \notin \{\mu_i : i \in I_3\}$. Then $v^n g^n(y_n) \to u_{I_0} g(y)$ and

$$v_{I_0}^n g^n \to u_{I_0} g. \tag{21}$$

Proof. Denote

$$z_i^n = v_{I_i}^n g^n, \quad j \in \{0, 1, 2, 3\}.$$

Clearly, f_n, z_0^n and z_1^n are bounded in L_2 . Using the notation $\langle \cdot, \cdot \rangle$ for the inner product in L_2 , we can write

$$||f_n - v^n g^n||^2 = ||f_n - z_0^n||^2 + ||z_1^n||^2 + ||z_2^n + z_3^n||^2 - 2\langle f_n - z_0^n, z_1^n \rangle - 2\langle f_n - z_0^n, z_1^n, z_2^n + z_3^n \rangle.$$

Since $g_i^n \to 0$ and $|v_i| \le 1 \ \forall i \in I_2$, we have $z_2^n \to 0$. Next, we will show that $z_3^n \to 0$, then it is clear that the third and fifth terms above converge to 0 when $n \to \infty$.

According to Corollary 5.1, $v_i^n \to 0$ for every $i \in I_3$. Denote $J := I_0 \cup I_1 \cup I_2$ and $\sigma_{0,n} := \max_i \{ \sigma_{i,n} : v_i^n > 0 \}$. By Corollary 5.1, for every n big enough, there exists $i(n) \in J$ such that $\sigma_{0,n} = \sigma_{i(n),n}$. In other words, $\sigma_{0,n} \ge \min_{i \in J} \sigma_{i,n}$. This, in turn, implies $\liminf_n \sigma_{0,n} \ge \lim_n \min_{i \in J} \sigma_{i,n} = \min_{i \in J} \sigma_i > 0$. Since $\sup_n \|f_n\|_{\infty} < \infty$, by (20) there exist constants a and b such that $v_i^n/\sigma_{i,n} \le a + b/\sigma_{0,n}$. For every i,

$$||g_i^n||^2 = \frac{1}{2\sqrt{\pi}} \cdot \frac{1}{\sigma_{i,n}},\tag{22}$$

so that when $\sigma_{i,n} \to 0$, then

$$|v_i^n||g_i^n|| = (2\sqrt{\pi})^{-\frac{1}{2}} \cdot \frac{v_i^n}{\sigma_{i,n}} \sqrt{\sigma_{i,n}} \to 0.$$

Thus $||z_3^n|| \leq \sum_{i \in J^c} v_i^n ||g_i^n|| \to 0$, implying $z_3^n \to 0$.

Now we are ready to study the existence and properties of the limit of $v^n g^n$. By the compactness of S_k , there exists a converging subsequence $v^{n'} \to w'$. Recall that $g_i^n \to g_i$ for every $i \in I_0$, thus $f_{n'} - z_0^{n'} \to f - w'_{I_0} g$. Since $g_i^n \to 0 \ \forall i \in I_1$, we have $z_1^n \to 0$, and therefore, $\langle f_{n'} - z_0^{n'}, z_1^{n'} \rangle \to 0$. Finally, $||z_1^{n'}||^2 = K^{n'} v_{I_1}^{n'} \cdot v_{I_1}^{n'} \to K w'_{I_1} \cdot w'_{I_1}$. Hence,

$$||f_{n'} - v^{n'}g^{n'}||^2 \to t(w'_J).$$

On the other hand, if we define

$$\tilde{u}_i = \begin{cases} u_i, & i \in J, \\ 0, & i \in I_3, \end{cases}$$

then, by similar argument, we have

$$||f_n - \tilde{u}g^n||^2 \to t(u) = t^*.$$

According to the definition of v^n , the inequality $||f_n - \tilde{u}g^n||^2 \ge ||f_n - v^n g^n||^2$ holds for every n, therefore $t(u) \ge t(w'_J)$. According to the definition of u, this implies that $t(u) = t(w'_J)$, and because of the uniqueness of $u_{I_0}g$, we have $w'_{I_0}g = u_{I_0}g$. Thus, for this subsequence, the convergence (21) holds. Since from every subsequence of the original sequence we can extract a subsequence which converges to the same limit, we have established (21).

Since for every $i \in I_0$, $\mu_{i,n} \to \mu_i$ and $\sigma_{i,n} \to \sigma_i > 0$, for every convergent sequence $y_n \to y$, also $v_{I_0}^n g^n(y_n) \to u_{I_0} g(y)$. Since $y \in \mathbb{R}$, but for every $i \in I_1$, $|\mu_{i,n}| \to \infty$ and $\sigma_{i,n} \to \sigma_i \in (0,\infty)$, it follows that $g_i^n(y_n) \to 0$, thus $v_{I_1}^n g^n(y_n) \to 0$. For every $i \in I_2$, $g_i^n(y_n) \to 0$. Finally, if $y \notin \{\mu_i : i \in I_3\}$, then for every $i \in I_3$, $\exp[-(y_n - \mu_{i,n})^2/\sigma_{i,n}^2] \to 0$. Since $v_n/\sigma_{i,n} \le a + b/\sigma_{0,n}$, and $\sigma_{0,n}$ is bounded away from 0, we obtain $v_{I_3}^n g^n(y_n) \to 0$. Thus, $v^n g^n(y_n) \to u_{I_0} g(y)$.

Corollary 5.2 Let $f_n \to f$ and $\sup_n ||f_n||_{\infty} < \infty$. When $g_i^n \to g_i$ for every $i \in \{1, \ldots, k\}$, then $v^n g^n \to vg$, where $v = \arg\min_{w \in S_k} ||f - wg||$.

Proof. When $g_i^n \to g_i$ for every i, then $I_0 = \{1, \dots, k\}$, $t(w) = ||f - wg||^2$ and u = v. The convergence $v^n g^n \to vg$ follows from (21).

6 Uniform convergence of the criterion function

Extending the pseudo-likelihood function. Recall the log-pseudo-likelihood function $\ell_n(\theta)$. We enlarge Θ_o , allowing some of the variances to be zero or infinite, and some of the means to be infinite. Thus, we define $\Theta := ([-\infty, \infty] \times [0, \infty])^k$. We now extend the pseudo-likelihood to Θ . For every $\theta \in \Theta$, let

$$I_0(\theta) := \{i : \sigma_i \in (0, \infty), |\mu_i| < \infty\}, \quad I_1(\theta) := \{i : \sigma_i \in (0, \infty), |\mu_i| = \infty\},$$

 $I_2(\theta) := \{i : \sigma_i = \infty\}, \quad I_3(\theta) := \{i : \sigma_i = 0\}.$

When $I_1(\theta) \neq \emptyset$, we need the symmetric, nonnegatively definite matrix $K = (k_{i,j})_{i,j \in I_1(\theta)}$ defined in Section 5, where $k_{i,i} = \frac{1}{2\sigma_i\sqrt{\pi}}$. For the elements of Θ with $|I_0(\theta)| = r_0 > 0$, we extend the function $\ell(\theta)$ as follows $(g_i(\cdot) := g(\theta_i, \cdot), i \in I_0)$. Recall that for any $w \in S_{r_0+r_1+r_2}$, $t(w) := ||f - w_{I_0}g||^2 + \sum_{i,j \in I_1} w_i w_j k_{ij}$, and $u := \arg\min_{w \in S_{r_0+r_1+r_2}} t(w)$. Let

$$\ell(\theta, K) := E \ln \left(u_{I_0}(\theta) g(\theta, Y_1) \right).$$

By Lemma 5.1, the definition of $\ell(\theta, K)$ is correct. Observe that when $I_1(\theta) = \emptyset$, then $\ell(\theta, K)$ is independent of K, and when $\sum_{i \in I_0} u_i = 0$, then $\ell(\theta, K) = -\infty$. When $\theta \in \Theta_o$, then $I_0(\theta) = \{1, \ldots, k\}$, $t(w) = \|f - wg\|^2$, $I_2(\theta) = \emptyset$ and $u(\theta) = v(\theta)$. Hence, $\ell(\theta, K)$ extends $\ell(\theta)$. When $\sum_{i \in I_0} u_i \neq 0$, we define $|u| = \sum_{i \in I_0} u_i$ and $\tilde{f} = u_{I_0} g/|u|$. So \tilde{f} is a proper probability density function and if it is different from f, then by Gibb's inequality,

 $E \ln(\tilde{f}(Y_1)) < \ell(\theta^*)$. Now, whenever $\theta \in \Theta$ is such that $\theta \neq \theta^*$ (in the sense of equivalence classes), then for every K it holds that

$$\ell(\theta, K) = E \ln(\tilde{f}(Y_1)) + \ln|u| < E \ln(\tilde{f}(Y_1)) < \ell(\theta^*).$$
(23)

In the following proposition, $\theta^n \to \theta$ denotes the componentwise convergence $\mu_{i,n} \to \mu_i$ and $\sigma_{i,n} \to \sigma_i$, where we allow some limits μ_i to be infinite and σ_i to be 0 or ∞ . Hence, $\theta \in \Theta$. Moreover, we assume that $K_n = (\langle g_i^n, g_j^n \rangle)_{i,j \in I_1(\theta)}$ converges entrywise to the matrix K, denoted by $K_n \to K$. This convergence is required for Proposition 5.1.

Proposition 6.1 Assume that $\sup_n \|\hat{f}_n\|_{\infty} < \infty$ almost surely. Let $\theta^n \in \Theta_o$ and $\theta^n \to \theta \in \Theta$, $|I_0(\theta)| = r_0 > 0$ and $K_n \to K$. Then $\ell_n(\theta^n) \stackrel{a.s.}{\to} \ell(\theta, K)$ and $\ell(\theta^n) \to \ell(\theta, K)$. Moreover, the set of ω 's where the almost sure convergence holds is independent of the sequence $\{\theta^n\}$.

Proof. Recall that P_n is the empirical measure based on Y_1, \ldots, Y_n , and P is the probability measure with density f. Let

$$\Omega_o := \{\omega : \hat{f}_n \to f\} \cap \{P_n \Rightarrow P\} \cap \{\sup_n \|\hat{f}_n\|_{\infty} < \infty\} \cap \{\int y^2 P_n(dy) \to \int y^2 P(dy)\}.$$

Take $\omega \in \Omega_o$ and denote $v^n = v^n(\theta^n)$ (recall (2)), let P_n^{ω} be the corresponding empirical measures. Observe that v^n depends on ω . Take $f_n = \hat{f}_n$, then $f_n \to f$.

The upper bound of $\ln(v^n g^n(y))$. Consider a sequence $y_n \to y$, where $y \notin \{\mu_i : i \in I_3(\theta)\}$. By Proposition 5.1, $v^n g^n(y_n) \to u_{I_0} g(y)$. Apply Corollary 4.1 with

$$h_n^{\omega}(\cdot) := \ln \left(v^n g^n(\cdot) \right), \quad h(\cdot) := \ln \left(u_{I_0} g(\cdot) \right).$$

It may happen that $\sum_{i\in I_0}u_i=0$, in which case, by (21), we have $v_{I_0}^ng^n\to 0$. If this is the case, then (15) of Lemma 4.1 establishes that $\ell_n(\theta^n)\to -\infty$ (the function H is constant, see the upper bound in (24) below). We now consider the case where $\|u_{I_0}g\|>0$. Let us show that there exists a continuous function H such that $|h_n^\omega(y)|\leq H(y)$ and $\int HdP_n^\omega\to \int HdP$. By (21), it holds that $\|v_{I_0}^ng^n\|\to \|u_{I_0}g\|>0$, which implies that $\lim\inf_n\sum_{i\in I_0}v_i^n>0$. Thus, there exists $\alpha>0$ (depending on θ and K, but independent of the choice of v^n) such that $\sum_{i\in I_0}v_i^n>\alpha$ eventually. This means that for every n large enough, there exists $j\in I_0$ such that $v_j^n\geq \alpha/k$, and therefore $v_j^n/\sigma_{j,n}\geq \frac{\alpha}{k\max_{i\in I_0}\sigma_{i,n}}$. Recall from (20) that $\sigma_{0,n}=\max\{\sigma_{i,n}:v_i^n>0\}$ and for n large enough, $v_i^n/\sigma_{i,n}\leq a+b/\sigma_{0,n}$ (because for large n, $\|\hat{f}_n\|_\infty<\|f\|_\infty+1$). Thus, we have

$$k\left(a + \frac{b}{\sigma_{0,n}}\right) \ge v^n g^n(y) \ge v_{I_0}^n g^n(y) \ge \frac{1}{\sqrt{2\pi}} \frac{\alpha}{k \max_{i \in I_0} \sigma_{i,n}} \exp\left[-\frac{\max_{i \in I_0} (y - \mu_{i,n})^2}{\min_{i \in I_0} \sigma_{i,n}^2}\right]. \tag{24}$$

Since $\sum_{i \in I_0} v_i^n > \alpha$ eventually, it holds that

$$\lim\inf_n \sigma_{0,n} \geq \lim_n \min_{i \in I_0} \sigma_{i,n} = \min_{i \in I_0} \sigma_i > 0.$$

Therefore, $\ln(v^n g^n(y))$ is bounded above by a constant: $\ln(k(a+b/\sigma_{0,n})) \leq N_1$, where N_1 depends on $\min_{i \in I_0} \sigma_i$ and hence on θ .

The lower bound of $\ln(v^n g^n(y))$. Observe that $\max_{i \in I_0} \sigma_{i,n} \to \max_{i \in I_0} \sigma_i < \infty$ and $\min_{i \in I_0} \sigma_{i,n} \to \min_{i \in I_0} \sigma_i < \infty$, thus there exist constants $N_2 < \infty$ and $M_2 < \infty$ depending on θ and K such that

$$-\ln\left(v^{n}g^{n}(y)\right) \leq N_{2} + \frac{\max_{j \in I_{0}}(y - \mu_{i,n})^{2}}{\min_{i \in I_{0}}\sigma_{i,n}^{2}} \leq N_{2} + \frac{\sum_{i \in I_{0}}(y - \mu_{i,n})^{2}}{\min_{i \in I_{0}}\sigma_{i,n}^{2}}$$
$$\leq N_{2} + M_{2}\sum_{i \in I_{0}}(y - \mu_{i,n})^{2}.$$

Since for every $i \in I_0$ we have $\mu_{i,n} \to \mu_i \in \mathbb{R}$, there exist constants A and B depending on μ_i such that

$$\sum_{i \in I_0} (y^2 - 2y\mu_{i,n} + \mu_{i,n}^2) \le ky^2 + A|y| + B.$$

Hence, by taking $H(y) := N_2 + M_2(ky^2 + A|y| + B) + N_1$, we can see that $\int HP_n^{\omega} \to \int HdP$. Since the assumptions of Corollary 4.1 are fulfilled, $\ell_n^{\omega}(\theta^n) \to \ell(\theta, K)$ follows. Since $P(\Omega_0) = 1$, we obtain $\ell_n(\theta^n) \stackrel{a.s.}{\to} \ell(\theta, K)$.

For the proof of $\ell(\theta^n) \to \ell(\theta)$, take $f_n := f$ and use Proposition 5.1 to deduce that $\ln(c^n g^n(y_n)) \to h(y)$, where $c^n = v(\theta^n)$ and $h(y) = \ln(u_{I_0}g(y))$. Observe that c^n is independent of ω . The convergence

$$\ell(\theta^n) = E \ln(c^n g^n(Y_1)) \to Eh(Y_1) = \ell(\theta)$$

now follows either by the dominated convergence theorem or by the argument above when taking $P_n = P$.

Proposition 6.1 implies the almost sure uniform convergence of the log-pseudo-likelihood over $\Theta_o(u, U, N)$.

Corollary 6.1 Let the assumptions of Proposition 6.1 hold. Then

$$P\left(\sup_{\theta \in \Theta_n(u,U,N)} |\ell_n(\theta) - \ell(\theta)| \to 0\right) = 1.$$
 (25)

Proof. Let Ω_o be the set with probability measure 1, where the convergences $\theta^n \to \theta$ and $K_n \to K$ entail $\ell_n(\theta^n) \to \ell(\theta, K)$, provided that $I_0(\theta) \neq \emptyset$. Fix $\omega \in \Omega_o$. When $\sup_{\theta \in \Theta_o(u,U,N)} |\ell_n^{\omega}(\theta) - \ell(\theta)| \to 0$ fails, there exists a sequence $\theta_n \in \Theta_o(u,U,N)$ and some $\epsilon_o > 0$ such that $|\ell_n^{\omega}(\theta^n) - \ell(\theta^n)| > \epsilon_o$ for every n. It is easy to see that there exists a subsequence $\theta^{n'}$ such that $\theta^{n'} \to \theta \in \Theta(u,U,N)$ and $K_{n'} \to K$. Since $\omega \in \Omega_o$, Proposition 6.1 establishes $\ell_{n'}^{\omega}(\theta^{n'}) \to \ell(\theta,K)$ and $\ell(\theta^{n'}) \to \ell(\theta,K)$ – a contradiction.

We can now prove the main theorem of the article. The uniform convergence in (25) implies the consistency of $\hat{\theta}^n$, provided that $\hat{\theta}^n$ eventually belongs to $\Theta_o(u, U, N)$.

The proof of Theorem 2.1. a) We start by proving that $\hat{\theta}^n \stackrel{a.s.}{\to} \theta^*$. By Proposition 3.1, there exist constants u, U, N (depending solely on f) such that $P(\hat{\theta}^n \in \Theta_o(u, U, N))$ eventually f(u, U, N) eventually f(u, U, N) eventually f(u, U, N). Let f(u, U, N) eventually, and the uniform convergence (25) holds. Since all these events hold with probability one, clearly f(u, U, N) eventually. On this set, the following relationships hold:

$$\ell_n(\hat{\theta}^n) \ge \ell_n(\theta^*) - \epsilon_n \to \ell(\theta^*) \quad \Rightarrow \quad \liminf_n \ell_n(\hat{\theta}^n) \ge \ell(\theta^*).$$
 (26)

Because of the uniform convergence (25), $\limsup_n \ell_n(\hat{\theta}^n) = \limsup_n \ell(\hat{\theta}^n) \leq \ell(\theta^*)$, which together with (26) implies $\ell_n(\hat{\theta}^n) \to \ell(\theta^*)$. From every subsequence of $\hat{\theta}^n$, one can find a further subsequence n' such that $K_{n'} \to K$ and $\theta^{n'} \to \theta \in \Theta_o(u, U, N)$. By Proposition 6.1, $\ell_{n'}(\hat{\theta}^{n'}) \to \ell(\theta, K)$. On the other hand, $\ell_{n'}(\hat{\theta}^{n'}) \to \ell(\theta^*)$, thus $\ell(\theta, K) = \ell(\theta^*)$. By (23), $\theta = \theta^*$. This implies $\hat{\theta}^n \to \theta^*$.

b) Since the convergence $\hat{\theta}^n \to \theta^*$ entails the convergence $g^n := g(\hat{\theta}^n, \cdot) \to g(\theta^*, \cdot) =: g$, and $w^* = \arg\min_{w \in S_k} \|f - wg\|$, the convergence $v^n(\hat{\theta}^n)g^n \xrightarrow{a.s.} w^*g$ follows from Corollary 5.2. The convergence of weights follows from the uniqueness of Gaussian densities and our assumptions on f (recall that $s(\theta^*) = k$ and $w_i^* > 0$ for every i), which imply that any convergent subsequence of $v^n(\hat{\theta}^n)$ must have limit w^* . This concludes the proof.

Acknowledgements

This work was funded by the Estonian Research Council grant PRG865.

References

- [1] J. Chen (2017). Consistency of the MLE under mixture models. *Statistical Science*, 32(1), 47–63.
- [2] A. Cutler, O. I. Cordero-Brana (1996). Minimum Hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, 91(436), 1716–1723.
- [3] A. Dembo, T. Weissman, T. (2005). Universal denoising for the finite-input general-output channel, *IEEE Transactions on Information Theory*, 51(4), 1507–1517.
- [4] S. Frühwirth-Schnatter (2006). Finite mixture and Markov switching models. Springer, New York.
- [5] R. Kangro, K. Kuljus, J. Lember (2025). Pseudo-likelihood approach for parameter estimation in univariate normal mixture models. *Statistical Papers*, 66(22).
- [6] K. Kuljus, B. Ranneby (2025). Maximum spacing estimation for hidden Markov models. Statistical Inference for Stochastic Processes, 28(7).

- [7] G. Lindgren (1978). Markov regime models for mixed distributions and switching regressions. Scandinavian Journal of Statistics, 5(2), 81–91.
- [8] M. Ranalli, B. G. Lindsay, D. R. Hunter (2020). A classical invariance approach to the normal mixture problem. *Statistica Sinica*, 30(3), 1235–1254.
- [9] B. Seo (2017). The doubly smoothed maximum likelihood estimation for location-shifted semiparametric mixtures, *Computational Statistics and Data Analysis*, 108(1), 27–39.
- [10] B. Seo, B. G. Lindsay (2010). A computational strategy for doubly smoothed MLE exemplified in the normal mixture model, *Computational Statistics and Data Analysis*, 54(8), 1930–1941.
- [11] B. Seo, B. G. Lindsay (2013). A universally consistent modification of maximum likelihood, *Statistica Sinica*, 23(2), 467–487.
- [12] B. W. Silverman (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives, *The Annals of Statistics*, 6(1), 177–184.
- [13] K. Tanaka (2009). Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when penalty is imposed on the ratios of the scale parameters, *Scandinavian Journal of Statistics*, 36(1), 171–184.
- [14] K. Tanaka, A. Takemura (2006). Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when the scale parameters are exponentially small, *Bernoulli*, 12(6), 1003–1017.
- [15] H. Teicher (1963). Identifiability of finite mixtures. The Annals of Mathematical Statistics, 34(4), 1265–1269.
- [16] A. Z. Zambom, R. Dias (201). A review of kernel density estimation with applications to econometrics. *International Econometric Review*, 5(1), 20–42.
- [17] A. W. van der Vaart (2000). Asymptotic statistics. Cambridge University Press.
- [18] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, M. J. Weinberger (2005). Universal discrete denoising: known channel, *IEEE Transactions on Information Theory*, 51(1), 5-28.