A Recursive Algorithm for Mixture of Densities Estimation

Alessio Sancetta*

July 9, 2013

Abstract

Recursive algorithms for the estimation of mixtures of densities have attracted a lot of attention in the last ten years. Here an algorithm for recursive estimation is studied. It complements existing approaches in the literature, as it is based on conditions that are usually very weak. For example, the parameter space over which the mixture is taken does not need to be necessarily bounded. The essence of the procedure is to combine density estimation via empirical characteristic function together with an iterative Hilbert space approximation algorithm. The conditions for consistency of the estimator are verified for three important statistical problems. A simulation study is also included.

Key Words: Boosting, Copula, Elliptic Distribution, Empirical Characteristic Function, Hilbert Space, Location Model, Recursive Algorithm.

1 Introduction

Modelling by mixtures is an old problem dating back to Pearson (1894). Typical approaches are based on EM algorithms (implying a discrete mixture model, e.g., Wu, 1983, for asymptotic results), greedy and recursive estimation techniques (e.g., Li and Barron, 2000, Klemela, 2007) and nonparametric Bayesian techniques (e.g., Ghosh and Ramamoorthi, 2003). Greedy and recursive estimation algorithms tend to be much faster and more computationally feasible than other approaches (e.g., see remarks in Tokdar et al., 2009). For this reason, study of fast algorithms somehow in the same vein of boosting should be welcomed and their statistical properties should be understood. This paper proposes a greedy algorithm for estimating mixtures of densities and shows that the resulting density estimator is strongly consistent. Moreover, it is also shown that the estimated characteristic function converges in probability to the true one, under a modified L_2 norm, at the optimal parametric rate.

The consistency of mixture of densities algorithms often relies - either explicitly or implicitly - on a bounded support for the mixing distribution (e.g., Wu, 1983, Li and

*E-mail: <asancetta@gmail.com>, URL: <http://sites.google.com/site/wwwsancetta/> (corresponding author). Address for correspondence: Department of Economics, Royal Holloway, Egham Hill, Egham TW20 0EX, UK. Acknowledgments: I thank Ryan Martin for sharing some code that proved useful in the simulations, and the referees for valuable comments that led to improvements both in content and presentation.

Barron, 2000, Ghosal and van der Vaart, 2001, Rakhlin, 2005, Tokdar et al., 2009); Lijoi et al. (2005) is an exception in this respect in the Bayesian context. The restriction on the support of the mixing law is usually needed to avoid divergence to infinity between elements in the same class.

To retain the simplicity of recursive algorithms and also allow for a possibly unbounded support for the mixing law, this paper combines density estimation via empirical characteristic function (Feuerverger and McDunnough, 1981a, 1981b, Carrasco and Florens, 2002, and Yu, 2004, for a review) together with an iterative Hilbert space approximation algorithm (Li and Barron, 2000, Barron, 1993, Jones, 1992). The recursive aspect of the algorithm used here is also similar to the L_2 algorithm studied in Klemela (2007). Using the L_2 distance between densities, Klemela (2007) controls the estimation error via an entropy integral that might be difficult to bound unless the parameter space is bounded. Here, the stagewise optimization is carried out in the frequency domain. This allows to use implicitly weaker conditions. The consistency bound in this paper - though in the frequency domain - is shown to hold with the same convergence rate as Klemela (2007, Corollary 1), but almost surely rather than in L_2. Moreover, it is shown that in probability, one can achieve the parametric rate of convergence between characteristic functions.

It is well known that the set of densities with bounded support is dense in L_1 (e.g., Devroye and Gyorfi, 2002, Lemma 5.1). Hence, weakening the conditions to allow for an unbounded support for the mixing distribution might be considered of little practical importance. However, unless the procedure is consistent for the support of the mixing distribution going to infinity, it is impossible to justify such approximation asymptotically. Hence, approximating a distribution with unbounded support with one with bounded support may have implications on the convergence rates, leading to possibly suboptimal results in practice.

Estimation via characteristic function is not new. It has often been used for finite dimensional parameter estimation (Feuerverger and McDunnough, 1981a, 1981b, Carrasco and Florens, 2002, and Yu, 2004, for a review). The main motivation, within the finite dimensional parameter estimation literature, is tractability of characteristic function in cases when the likelihood might be intractable. Here, the motivation is that the characteristic function is usually a well behaved object, being bounded and allowing to easily deal with important models found in the literature, including scale and location mixture models. Some examples can be found in Section 3.3.

Section 2 describes the algorithm and the consistency results. Section 3 provides a discussion of the conditions used and the results. The conditions are verified for elliptic densities with the consistency property (Kano, 1994) and elliptic copulae that are nearly tail dependent (Manner and Segers 2011, Hua and Joe, 2011) as well as to the standard Gaussian location mixture model. The first two applications require the support of the mixing law to be unbounded, hence, cannot be dealt with by most of the existing procedures. Remarks on the quantities used to derive the estimator, such as number of iterations can also be found in Section 3. Finally, some simulation results are reported in order to highlight both pros and cons of the present estimator. In these simulations, the estimator is compared to Klemela's stagewise optimization, the Newton's recursive estimator, kernel density and the scale location mixture estimator by the EM algorithm. Proofs are deferred to Section 4.

1.1 Statement of the Problem

Let $X := (X_j)_{j>0}$ be a sequence of independent identically distributed observations with values in $\mathbb{R}^K$ (or a subset of it) and with density function

$$\int_\Theta \kappa_\theta(x)\, dP(\theta), \tag{1}$$

where $\{\kappa_\theta(x) : \theta \in \Theta\}$ is a class of densities and Θ is a Euclidean set. The mixing law P is unknown. Interest lies in estimating (1). To this end, a method based on the empirical characteristic function will be used.

Let $f_\theta(t)$ be the characteristic function corresponding to the density $\kappa_\theta(x)$. The empirical characteristic function of X, for a sample size n, is

$$f_n(t) = \frac{1}{n}\sum_{j=1}^{n} \exp\{i\langle t, X_j\rangle\},$$

where $\langle \bullet, \bullet \rangle$ is the inner product between two vectors. By the properties of the empirical characteristic function,

$$\begin{aligned} \mathbb{E}f_n(t) &= \int_\Theta f_\theta(t)\, dP(\theta), \\ &=: f_P(t) \end{aligned}$$

with the definition on the r.h.s. used throughout to represent the true characteristic function of the data. Also, write f_G when using G as mixing law rather than the true P. The above functions of t will be treated as elements in a liner vector space, so that, more generally, $\langle f_n, f_\theta\rangle = \int_{\mathbb{R}^K} f_n(t)\,\overline{f_\theta(t)}dt$, and similarly for related quantities. Here and elsewhere, $\overline{f}$ denotes the complex conjugate of f. For a linear operator W, let $\langle f, g\rangle_W = \langle f, Wg\rangle$, so that $|f-g|_W := \langle f-g, f-g\rangle_W^{1/2}$ is a metric when W is positive definite.

We shall minimize the following objective function

$$|f_n - f_G|_W^2 = \langle f_n - f_G, f_n - f_G\rangle_W \tag{2}$$

with respect to some mixing law G, where W is a positive definite operator.

2 An Algorithm to Approximate Mixture of Densities

The optimization in (2) is solved by the greedy algorithm in Table 1.

TABLE 1

Set

$J > 0$

$j = 1;$

$F_0 := 0$

While $j < J$

$$\hat{\theta}(j) = \arg\inf_{\theta\in\Theta} \left| f_n - \frac{j-1}{j} F_{j-1} - \frac{1}{j} f_\theta \right|_W^2$$

$$F_j = \frac{j-1}{j} F_{j-1} + \frac{1}{j} f_{\hat{\theta}(j)}$$

$$j = j + 1,$$

End

The algorithm is based on the ideas of Barron (1993), and - mutatis mutandis - the weighting scheme can be deduced from the proof of Theorem 1 in Li and Barron (2000). Suppose the following:

Condition 1 *For each $\epsilon > 0$, there exists a set Θ_ϵ with cardinality $N(\epsilon) < \infty$ such that*

$$\inf_{\theta(2),\theta(4)\in\Theta_\epsilon} \sup_{\theta(1),\theta(3)\in\Theta} \left\langle f_{\theta(1)} - f_{\theta(2)}, f_{\theta(3)} - f_{\theta(4)} \right\rangle_W \leq \epsilon.$$

Condition 2 *The operator W is strictly positive definite and such that its kernel (also denoted by W throughout the paper) satisfies $\int\int |W(s,t)| \leq 1$.*

Then, the estimation algorithm used here satisfies the following approximation bound:

Theorem 1 *Let $F_J = \sum_{j=1}^J f_{\hat{\theta}(j)}/J$, where $\left\{\hat{\theta}(j) : j \leq J\right\}$ is the output of the algorithm in Table 1. Then, under Conditions 1 and 2, for $J \geq 2$,*

$$|f_n - F_J|_W^2 \leq \inf_{G\in\mathcal{G}} |f_n - f_G|_W^2 + \frac{4\ln J}{J},$$

where $\mathcal{G}$ is the set of all laws with support in Θ.

Theorem 1 together with control of the estimation error leads to the following consistency result:

Theorem 2 *Let $F_J = \sum_{j=1}^J f_{\hat{\theta}(j)}/J$ be as in Theorem 1. Under Conditions 1-2, when $N(\epsilon) = O\left(\epsilon^{-V}\right)$ for some finite V,*

$$|f_P - F_J|_W^2 = O_{a.s.}\left(\sqrt{\frac{\ln n}{n}} + \frac{\ln J}{J}\right),$$

If $J/\ln J \geq cn$, for some (any) $c > 0$,

$$|f_P - F_J|_W^2 = O_p\left(\frac{1}{n}\right).$$

Moreover,

$$\frac{1}{J}\sum_{j=1}^J \kappa_{\hat{\theta}(j)}(x) \to \int_\Theta \kappa_\theta(x)\, dP(\theta)$$

a.s., almost everywhere in x.

3 Discussion

3.1 Remarks on Conditions

Condition 1 is used to control the complexity of the model. It says that there is an ϵ-net for the model $(f_\theta)_{\theta\in\Theta}$ under the norm induced by $\langle\bullet,\bullet\rangle_W$, and that the cardinality is equal to $N(\epsilon)$. Then, Theorem 2 restricts the growth of $N(\epsilon)$ (as $\epsilon \to 0$) to be exponential with exponent V. This exponent is proportional to the V-C dimension of $(f_\theta)_{\theta\in\Theta}$ (under $\langle\bullet,\bullet\rangle_W$) (e.g., van der Vaart and Wellner, 2000). The V-C dimension is also used in the bounds of Klemela (2005) for the L_2 stagewise algorithm. Under the L_2 norm, the Gaussian scale/location mixture model has V-C dimension that does not seem simple/possible to bound by a finite number, unless the parameter space for both location and scale are assumed to be bounded (e.g., Ghosal and van der Vaart, 2001).

Condition 2 is not innocuous, as it makes the estimator possibly inefficient. It is known that the choice of kernel W that leads to estimation as efficient as maximum likelihood is not in L_1 (e.g., Feuerverger and McDunnough, 1981a, 1981b, Carrasco and Florens, 2002). Section 3.5 provides details and insights to better understand the role of W in the estimation and the implications of minimizing the norm $|\bullet|_W$.

3.2 Remarks on Results

Mutatis mutandis, Klemela (2007) uses a similar updating approach via minimization of the L_2 distance between densities and where the weights of each component are independent of the sample, e.g., equal weights, as here. Using a likelihood based approach, Li and Barron (2000) allow each component to be added to the model to have a weight different than j^{-1}. At the cost of increased computational complexity, one can conjecture that the results presented here are also valid when we also optimize with respect to the mixing weights.

Theorem 2 gives a non-optimal bound which holds a.s. as well as a parametric rate that holds in probability. Rakhlin et al. (2005) show that a parametric convergence rate holds for the Li and Barron (2000) estimator, under the Kullback-Leibler divergence - a stronger criterion than the one used here- but at the cost of putting extra restrictions on the estimated model. Klemela (2007) derives a convergence rate which is essentially $O\left(\sqrt{n^{-1}\ln n}\right)$ in many practical situations (Klemela, 2007, Corollary 3). The same rate holds here, almost surely, but for $|\bullet|_W^2$ in the frequency domain. This norm is weaker than the L_2 distance between densities, but the restrictions on the model tend to be weaker. In fact, implicitly, Klemela's results may impose restrictions on the model - e.g., Θ bounded - in order to derive a bound for an entropy integral under the L_2 norm.

The convergence of the characteristic functions under $|\bullet|_W$ leads to pointwise density convergence. Under suitable conditions, the pointwise convergence can be turned into L_1 convergence by dominated convergence and to almost uniform convergence by Ergov's Theorem.

As shown by Rakhlin et al. (2005), the upper bound for the estimation error does not depend on the number of components in the mixtures, but only on the structure of the set Θ. Intuitively, the extremes of the estimation error in a mixture is reached at the edges of the simplex. Hence, as long as we choose J of at least the same order of magnitude as n then the error in Theorem 1 can be kept as small as possible in order of magnitude.

Adding extra terms may not improve the estimation, but does not not lead to a diverging estimation error. Mutatis mutandis, the simulations in Klemela (2007) highlight this feature in a finite sample and simulations carried out by the author, but not reported here also confirm this feature. Hence, the estimator is resilient to overfitting, but may require many discrete support points. This contrasts with the usual recommendations in the literature of having fewest support points (e.g., Priebe, 1994). Nevertheless, in finite sample, it is still desirable to find ways to choose the number of iterations in an optimal way. Section 3.6 provides further remarks in this respect.

The equal weighting given to the components has a simple intuitive explanation. Rather than trying to cover the set Θ with many little balls, we instead cover the range of P with equal intervals of size J^{-1}. This is the way the Lebesgue integral is constructed as opposed to the Riemann integral. This "Lebesgue integral" approach using equal weights leads to considerable simplification and faster estimation.

By Theorem 2, for large n, $F_J \simeq \mathbb{E} f_n$ so that the objective function in the estimation is an asymptotically degenerate V-statistic of order 2 (see the proof of Lemma 3 in Section 4.2). If $\kappa_\theta(x)$ is smooth for any $\theta \in \Theta$, one may conjecture that the asymptotic distribution of $\sum_{j=1}^{J} \kappa_{\hat{\theta}(j)}(x)/J$ turns out to be a weighted sum of chi-square random variables with weights depending on unknown quantities (e.g., Serfling, 1980). Unfortunately, direct control of $\sum_{j=1}^{J} \kappa_{\hat{\theta}(j)}(x)/J$ is not simple as the estimator is derived recursively. Of course, even if possible, the asymptotic distribution would depend on unknown quantities and the bootstrap for U and V statistics would be needed to construct confidence intervals (Arcones and Giné, 1992).

3.3 Applications

The following applications use the results in the previous section to show how practical problems can be solved by the current algorithm. Common assumptions used to derive consistency of the estimator are not necessarily satisfied by two of the problems discussed next, as the mixing law does not have bounded support.

Conditions for the validity of the results in Section 2 are stated for each application. For the first two applications, the domain of the kernel W is $\mathbb{R}^K \times \mathbb{R}^K$ (for some finite K) and Condition 2 needs to be strengthened to

$$\int_{\mathbb{R}^K} \int_{\mathbb{R}^K} |W(s,t)| \left(\sum_{k,l} |s_k s_l| \right) \left(\sum_{k,l} |t_k t_l| \right) dsdt \leq 1, \tag{3}$$

where s_k is the k^{th} element in s. The third and final example does not require any additional condition.

3.3.1 Elliptic Densities with the Consistency Property

Elliptic densities such that their marginals have all the same generator are said to satisfy the consistency property. Examples are the Gaussian and student's t density. The power exponential distribution is elliptic but does not satisfy such property (Kano, 1994). From Theorem 1 in Kano (1994), deduce that an elliptic density satisfying the consistency

property has the following representation

$$\int_{\Theta} \phi\left(x|\theta\Sigma\right) P\left(d\theta\right) \tag{4}$$

where $\phi\left(x|\theta\Sigma\right)$ is the centered K dimensional Gaussian density with covariance matrix $\theta\Sigma$, θ being a positive real number and P a law with support on the positive real line.

Verification of the conditions in Theorem 2 gives the following.

Corollary 1 *Consider the model in (4). Suppose Σ is positive definite and has bounded entries, and W satisfies 3. Then, Theorem 2 holds.*

By continuity w.r.t. Σ, the results hold when Σ is replaced by a root-n consistent estimator. For example, one may replace Σ with the empirical covariance function (assuming $K/n = O\left(n^{-1}\right)$).

Many of the existing methods cannot be used to estimate this model directly. Restricting P to have bounded support would rule out common examples like (4) being a t-density. For this example, eq. 5 in the recursive method in Li and Barron (2000) is not satisfied, as well as Condition A5 in Tokdar et al. (2009). As an illustration, we can compare the result for the Gaussian scale model to the results in Klemela (2007). For simplicity, consider the one dimensional case, so that $\Sigma = 1$. To apply the results in that paper we needed to find an ϵ-net in L_2, with finite cardinality for each $\epsilon > 0$, for the class of functions $\left(\phi\left(x|\theta\right)\right)_{\theta \geq 0}$, where $\phi\left(x|\theta\right)$ is the univariate Gaussian density with variance θ. This means that we need to verify the following: for each $\epsilon > 0$, there is a set Θ_{ϵ} with finite cardinality $N\left(\epsilon\right)$ such that

$$\inf_{\theta' \in \Theta_{\epsilon}} \sup_{\theta \in [0,\infty)} \int_{\mathbb{R}} \left|\phi\left(x|\theta\right) - \phi\left(x|\theta'\right)\right|^2 dx \leq \epsilon.$$

However, for $\theta \to 0$ or $\theta \to \infty$, it does not seem possible to bound the display by arbitrary $\epsilon > 0$. For example, Ghosal and van der Vaart (2001) derive entropy rates for scale location mixtures, but require the scale parameter θ to be in a compact interval bounded away from zero and infinity.

3.3.2 Nearly Tail Dependent Elliptic Copulae

Manner and Segers (2011) and Hua and Joe (2011) discuss the family of elliptic copulae that just fail to exhibit tail dependence. These authors use the Ledford and Tawn (1996) coefficient of tail dependence to continuously interpolate tail dependence with tail independence. In the case of a K dimensional copula C, the framework of Ledford and Tawn (1996) is, as $u \to 0^+$,

$$C\left(u1_K\right) = u^{\tau} l\left(u\right)\left(1 + o\left(u\right)\right),$$

where u is a scalar $\tau \geq 1$ and $l\left(u\right)$ is a slowly varying function at $u \to 0^+$ (e.g., Bingham et al., 1989). Note that in Ledford and Tawn (1996), $\eta = \tau^{-1}$ is used instead in their eq. 5.3. Lower tail dependence occurs when $\tau = 1$ and $\lim_{u \to 0+} l\left(u\right) > 0$. Upper tail dependence can be defined similarly (e.g., Hua and Joe, 2011). The intermediate case with $\tau = 1$ and $\lim_{u \to 0+} l\left(u\right) = 0$ shall be called near tail dependence: the η Ledford and Tawn coefficient is still equal to 1. Let C_{ρ} be a K dimensional Gaussian copula with

correlation matrix Σ with off diagonal entries all equal to $\rho \in [0,1)$. For any law P with support $[0,1)$,

$$\int_0^1 C_\rho\left(u1_K\right) dP\left(\rho\right) = ul\left(u\right)\left(1+o\left(u\right)\right)$$

as $u \to 0^+$. To see this, following Hua and Joe (2011) and applying similar arguments as in Manner and Segers (2011, Proposition 4), letting 1_K be the K dimensional vector of ones,

$$\int_0^1 C_\rho\left(u1_K\right) dP\left(\rho\right) \geq C_{\bar{\rho}}\left(u1_K\right) P\left(\rho \in [\bar{\rho},1)\right) = u^{\tau(\bar{\rho})} P\left(\rho \in [\bar{\rho},1)\right)$$

where $\tau\left(\rho\right) = \left\langle 1_K, \Sigma^{-1} 1_K \right\rangle = K/\left[1+\left(K-1\right)\rho\right]$ (Hua and Joe, 2011). Hence, for any $\epsilon > 0$ we can find $\bar{\rho} < 1$ such that

$$u^{\tau(\bar{\rho})}\left(1-F\left(\bar{\rho}\right)\right)/u^{1+\epsilon} \to \infty.$$

This implies that $\int_0^1 C_\rho\left(u1_K\right) dP\left(\rho\right)$ is of larger order than $u^{1+\epsilon}$ for any $\epsilon > 0$, but also of smaller order than u when $\rho < 1$.

The scaling matrix needs to be positive definite with diagonal equal to 1. An obvious representation to impose these constraints is

$$\Sigma_{kl} = \frac{\delta_{kl} + \sum_{r=1}^L B_{kr} B_{lr} \theta_r}{\sqrt{1+\sum_{r=1}^L B_{kr}^2 \theta_r}\sqrt{1+\sum_{r=1}^L B_{lr}^2 \theta_r}} \tag{5}$$

where δ is Kronecker's delta, and we can interpret B_{kr} as factor loadings in $\mathbb{R}$ and θ_r as the r^{th} factor variance, so that Θ only needs to be restricted to the positive orthant for the restrictions on $\Sigma\left(\theta\right)$ to hold. For $L = K$, the above is clearly dense in the space of positive semi-definite matrices. The matrix becomes singular as soon as $\theta_r \to \infty$ for any r if $B_{kr} B_{lr} \neq 0$ for some $k \neq l$. Sancetta and Satchell (2007) studied the above correlation structure when one puts a distribution on θ and derived several implications for portfolio diversification failure.

Using the above representation,

$$C\left(u|B,P\right) := \int C\left(u|B,\theta\right) dP\left(\theta\right) \tag{6}$$

where $C\left(u|B,\theta\right)$ is a Gaussian copula with scaling matrix Σ parametrized as in (5) and P is a law with support in $[0,\infty)^L$ (the L dimensional positive orthant). Near tail dependence is satisfied as long as P has support $[0,\infty)^L$. Once again, most common procedures cannot cope with the case of unbounded support. On the other hand, the following holds.

Corollary 2 *Consider the model in (6) with Σ as in (5). Suppose B has bounded entries and W satisfies (3). Then, Theorem 2 holds.*

3.3.3 Location Model

The above examples justify the procedure for having a scale model with unbounded scale parameter set. However, a very common model used to approximate densities is

the mixture of Gaussian location models (e.g., Lindsay, 1995, Ghosal and van der Vaart, 2001, and references therein). Here, it is shown that this important class of approximating models does satisfy Condition 1. Consider the following mixture

$$\int_{\Theta} \phi(x-\theta) P(d\theta) \tag{7}$$

where $\phi(x)$ is the standard Gaussian density and $\Theta \subseteq \mathbb{R}$. This model has good approximating properties even when Θ is a strict compact subset of $\mathbb{R}$. However, even if Θ were unbounded, Condition 1 would be satisfied and Theorem 2 be true under minimal conditions.

Corollary 3 *Consider the model in (7). Suppose Condition 2 holds. Then, Theorem 2 holds.*

3.4 Estimation in the Presence of Nuisance Parameters

The empirical characteristic function approach was first proposed in the context of parameter estimation. Hence, it is no surprise that it can be still be used when the kernel $\kappa_{\theta,\gamma}(x)$ depends on a nuisance parameter $\gamma \in \Gamma$ for some compact parameter space Γ. For notational simplicity, assume $\Gamma \subset \mathbb{R}$. The characteristic function of the kernel is now denoted by $f_{\theta,\gamma}(x)$. The objective function becomes

$$\left| f_n - \int_{\Theta} f_{\theta,\gamma} dG(\theta) \right|_W^2$$

and the first order condition w.r.t. γ, under regularity conditions, is

$$\left\langle f_n - \int_{\Theta} f_{\theta,\gamma} dG(\theta), \int_{\Theta} \frac{df_{\theta,\gamma}}{d\gamma} dG(\theta) \right\rangle_W = 0.$$

Hence, if a guess solution for the optimal γ is given, G can be estimated using the guess solution. Having estimated G, we can then plug G in the objective function and re-estimate γ and so on, until convergence of γ, up to a tolerance level. This procedure is costly and alternatives should be welcome. For results in this direction, using an approximate Bayesian framework, see Martin and Tokdar (2011).

3.5 Isometric Relations and Choice of W

As usual, there is a relation between convergence in the frequency domain and in the original space of functions. For simplicity, in the sequel, suppose that g and g' are two univariate densities with support in the reals and with characteristic function f and f', respectively. By Parseval's Theorem,

$$\int |g(x) - g'(x)|^2 dx = \frac{1}{2\pi} \int |f(t) - f'(t)|^2 dt,$$

which is $\langle f - f', f - f' \rangle_W$ when $W(s,t) = (2\pi)^{-1} \delta(s-t)$, where $\delta(t)$ is the Dirac delta function. This choice of W does not satisfy Condition 2, as $\int\int \delta(s-t) dsdt = \infty$. In

general, by definition of the characteristic function and Fubini's Theorem,

$$\begin{aligned}
|f-f'|_W^2 &= \int\int W(s,t)\,(f(s)-f'(s))\,(f(t)-f'(t))\,dsdt \\
&= \int\int W(s,t)\int e^{ixs}\,(g(x)-g'(x))\,dx \\
&\quad \times \int e^{-iyt}\,(g(y)-g'(y))\,dydsdt \\
&= \int\int \left(\int\int e^{ixs}e^{-iyt}W(s,t)\,dsdt\right) \\
&\quad \times (g(x)-g'(x))\,(g(y)-g'(y))\,dxdy. \qquad (8)
\end{aligned}$$

Now note that by Mercer's Theorem, the kernel of a positive definite operator can be written as

$$W(s,t)=\sum_{k=1}^{\infty}\rho_k\varphi_k(s)\,\varphi_k(t),$$

where $\rho_k \geq \rho_{k+1} > 0$ are the eigenvalues of the kernel and the $\varphi_k(t)$'s are the orthonormal eigenfunctions. It follows that

$$M(x,y) := \int\int W(s,t)\,e^{ixs}e^{-iyt}dsdt$$

is also positive definite and real if W is real. This last remark together with (8) shows that

$$|f-f'|_W^2 = |g-g'|_M^2$$

and convergence to zero implies convergence of the density under $|\bullet|_M$, which in turn implies convergence almost everywhere, and also L_1 convergence by the dominated convergence theorem if a function dominating the estimated density exists. (The latter may actually impose restrictions that are equivalent to a mixing law with bounded support.) The space of densities with finite norm $|\bullet|_M$ forms an Hilbert space under the inner product $\langle\bullet,\bullet\rangle_M$, but the topology induced by $|\bullet|_M$ is weaker than the one induced by the usual L_2distance.

The choice of kernel W affects the weight that is given to different frequencies of the original density. A kernel that has representation with eigenvalues ρ_k's decaying fast will give less weight to high frequency components of $g-g'$, i.e. implicitly smoothing out irregularities, and noise if the elements are estimators. However, as long as W has support on $\mathbb{R}\times\mathbb{R}$, convergence under $|\bullet|_W^2$ does imply convergence of all the frequencies.

In the numerical results presented below,

$$W(s,t)=\delta(s-t)\exp\left(-\left|\frac{s+t}{2}\right|^2\right) \qquad (9)$$

so that, by direct calculation using the functional form of the characteristic function of a Gaussian kernel,

$$M(x,y)\asymp\exp\left\{-\left(\frac{x-y}{2}\right)^2\right\}.$$

Convergence using this Gaussian kernel does not require the densities to be in L_2.

3.6 Number of Greedy Iterations J and Computational Cost

The total number of greedy steps J is a function of the sample size n. This makes sure that the approximation error as given in Theorem 1 is of the same order of magnitude as the estimation error.

The actual value of J is only specified in order of magnitude. For example, $J/\ln J = cn$ and fine tuning c can lead to improved performance. Choice of c could be based on crossvalidation. For example, define an estimation sample of size n_E and validation sample of size n_V ($n = n_E + n_V$, and the two samples are non-overlapping). Use the estimation sample to estimate the mixture for different values of c, say $F_{J(c)}$. Use the validation to compute the empirical characteristic function, say f_{n_V}. Then choose $J = \hat{c}n$, where

$$\hat{c} := \arg\min_{c} \left| f_{n_E} - F_{J(c)} \right|_W^2 .$$

One may conjecture that Theorem 2 implies that for a finite mixture with only a few support points the estimated $\hat{\theta}(j)$'s may cluster around the true support points. Experiments conducted on the Gaussian scale mixture, but not reported here, showed that this is not necessarily the case. At present, the author is unable to furnish a definite answer to this. However, it is most likely due to the fact that the objective function may have multiple local minima leading to estimated support points that are not necessary the same as the true ones. Mutatis mutandis, this is the same issue encountered in practice when maximizing the likelihood of Gaussian mixtures.

Because of the optimization step of θ over Θ, the current algorithm is not a proper one. The computational cost depends on the optimization procedure used. The present approach is similar to the one in Klemela (2007), the main difference is the objective function that is minimized at each step. Hence, the same remarks for the stagewise algorithm in Klemela (2007) directly apply to the present problem. Klemela (2007, Remark 3) looks at the computational cost of the brute force optimization, where θ is chosen to minimize the objective function over a finite set, say $\{\theta(1), \theta(2), ..., \theta(N)\}$. In this case, the computational cost is $O(J \times N \times EvalCost)$, where $EvalCost$ is the cost required to evaluate the objective function. It is worth noting that N grows exponentially with the dimension of Θ, and so the computational cost. In practice optimization of a smooth function is not carried out discretizing the parameter set, but by standard gradient based methods reducing considerably the computational cost, but increasing the possibility of finding only local minima.

3.7 Simulation

To highlight the small sample behaviour of the estimator, a set of simulations are carried out. Consider the model

$$\int_{\Theta} \phi(x|\theta)\, dP(\theta),$$

where $\phi(x|\theta)$ is the univariate centered Gaussian density with variance θ. Two cases are contemplated. In the first, P is the law of an inverse chi-square random variable divided by v, where v are the degrees of freedom. Hence, the mixture is just a Student t-density with v degrees of freedom. In the second case, following Friedman (2001), amongst others, simulations are also carried out from a random model. This is done to reduce the dependence of the results on the Monte Carlo. In particular, in each simulation, P is a

realization from a Dirichlet process with parameter αQ, where Q is the law corresponding to an exponential density with mean 1. Realizations of P are derived by stick breaking construction, so that any distribution can be approximated weakly by P constructed in such a way (e.g., Sethuraman, 1994, for details).

The sample size is $n = 10, 50, 100, 500$ and the number of simulations are $N = 250$. The estimator is computed replacing integrals with Monte Carlo integration, using importance sampling and restricting $W(s,t)$ to the diagonal $t = s$, i.e.

$$\frac{1}{1000}\sum_{s\in\mathcal{N}}\left|f_n(s) - \int_\Theta f_\theta(s)\, dG_n(\theta)\right|^2$$

where $\mathcal{N}$ is a set of 1000 iid mean zero normally distributed random variables with variance $1/2$, and G_n is the estimated mixing law. This corresponds to the Monte Carlo importance sampling version of (2) with $W(s,t)$ as in (9). For simplicity, the number of iterations was not made data dependent, but just set to $J = 200$.

The estimator is then evaluated based on the L_1 distance,

$$\int_{\mathbb{R}}\left|\int_0^\infty \phi(x|\theta)\, dP(\theta) - \int_0^\infty \phi(x|\theta)\, dG_n(\theta)\right| dx.$$

The integration over $\mathbb{R}$ is replaced by its Riemann approximation over a large enough interval. Results using Monte Carlo integration via importance sampling were similar but in some cases a bit unstable and required some care. To reduce the level of discretion in the evaluation process, these are not reported.

For comparison purposes, the following estimators are also computed: Klemela (2007) stagewise estimator, Newton (2002) recursive estimator of the mixing law, which is then used to compute the mixture, the Gaussian kernel density estimator and the Gaussian location and scale mixture estimated using the EM algorithm.

In particular, for the Newton's algorithm a uniform density with support $[0, 50]$ is used as first guess. The size of support seemed to have non-negligible impact on the estimator. The bandwidth for the Gaussian kernel is chosen as the one that minimizes the ex post L_1 distance over a fixed set of bandwidths. The EM algorithm is estimated using the Matlab function <gmdistribution> and BIC for selecting the number of mixing components. In implementing these competing models, some subjective judgment was used in the selection of additional parameters. This may affect the estimation performance. The goal of the simulations is to verify if the present estimator can be competitive at estimating a mixture of densities. It is not to demonstrate superiority of a method against another, as this depends on many factors, including Monte Carlo design, choice of tuning parameters etc.

Some of the methods considered in the simulations were actually devised with the specific goal of estimating the mixing law consistently and not with the focus on estimation of the corresponding mixture of densities (e.g., the Newton's method). Conversely, the algorithm discussed in this paper focuses on estimation of the mixture of densities and not the mixing law for which it can provide poor estimates (e.g., Section 3.6). Hence, due to different goals and scopes of the methods, one should be cautious in drawing comparisons and conclusions.

3.7.1 Student-t Density

For the Student density, data were simulated for the following choices of degrees of freedom $v = 4, 8, 16$. For economy of space, Figure 1 and 2 only report the boxplot for the L_1 distance for $v = 4$ and $n = 10, 500$, as other results are consistent with the reported ones.

Figure 1. Boxplots: Student $v = 4$ df, $n = 10$

Figure 2. Boxplots: Student $v = 4$ df, $n = 500$

3.7.2 Random Model

In the stick breaking construction, the following choices for the dispersion parameter where chosen $\alpha = 1, 4, 8$. For economy of space, Figure 3 and 4 only report the boxplot for the L_1 distance for $\alpha = 4$ and $n = 10,500$. In this simulation, the dominating measure in the mixing distribution used in the Newton's algorithm is still uniform, i.e. continuous. However, the true one is a Dirichlet mixing law, which is a.s. discrete. This miss-specification negatively affects the resulting estimated mixture (Martin and Ghosh, 2008, and Martin, 2013, for discussions and a remedy).

Figure 3. Boxplots: Dirichlet $\alpha = 4$, $n = 10$

Figure 4. Boxplots: Dirichlet $\alpha = 4$, $n = 500$

4 Proofs

4.1 Proof of Theorem 1

The proof makes use of the following lemma that shows that the true mixing law P can be approximated by an atomic law under a suitable topology.

Lemma 1 *Suppose P is a law with support in Θ. If Condition 1 and 2 hold, then, there is a purely atomic law G such that*

$$\left| \int_{\Theta} f_{\theta} d\left(P\left(\theta\right) - G\left(\theta\right)\right) \right|_{W}^{2} \leq \epsilon.$$

Proof. Let $(A_s)_{s\in\mathbb{N}}$ be a cover for Θ. Let $G\left(\theta\right) = \sum_{s\in\mathbb{N}} \left(\int_{A_s} dP\left(\theta\right) \right) \delta_{\theta(s)}\left(\theta\right)$, where $\delta_{\theta(s)}$ is the point mass at $\theta\left(s\right) \in A_s$. Then,

$$\int_{\Theta} f_{\theta} d\left(P\left(\theta\right) - G\left(\theta\right)\right) = \sum_{s\in\mathbb{N}} \int_{A_s} \left(f_{\theta} - f_{\theta(s)}\right) dP\left(\theta\right)$$

and substituting in the objective function

$$\left\langle \int_{\Theta} f_{\theta} d\left(P\left(\theta\right) - G\left(\theta\right)\right), \int_{\Theta} f_{\theta} d\left(P\left(\theta\right) - G\left(\theta\right)\right) \right\rangle_{W}$$

$$= \sum_{r\in\mathbb{N}}\sum_{s\in\mathbb{N}}\int_{A_r}\int_{A_s}\left\langle f_\theta - f_{\theta(r)}, f_{\theta'} - f_{\theta(s)}\right\rangle_W dP(\theta)\, dP(\theta')$$

[by linearity of the inner product]

$$\leq \max_{r,s\in\mathbb{N}}\sup_{\theta\in A_r,\theta'\in A_s}\left\langle f_\theta - f_{\theta(r)}, f_{\theta'} - f_{\theta(s)}\right\rangle_W \sum_{r\in\mathbb{N}}\sum_{s\in\mathbb{N}}\int_{A_r}\int_{A_s} dP(\theta)\, dP(\theta'),$$

and the result follows by integration together with Condition 1 choosing $\theta(s) \in A_s$ appropriately for all s. ■

Proof. [Theorem 1] The proof is a modification of the proof of Theorem 5 in Barron (1993), eq. 51-55, in particular. By Condition 1,

$$\inf_{\theta'\in\Theta_\epsilon}\sup_{\theta\in\Theta}|f_\theta - f_{\theta'}|_W^2 \leq \epsilon, \tag{10}$$

and applying Lemma 1, there are weights λ_j's (positive and summing to one), and parameter values $\theta'(j)$'s, $j = 1, ..., N$, such that

$$\left|\sum_{j=1}^N \lambda_j f_{\theta'(j)} - \mathbb{E}f_n\right|_W^2 \leq \epsilon. \tag{11}$$

Defining

$$\tilde{f} := \sum_{j=1}^N \lambda_j f_{\theta'(j)}, \tag{12}$$

after some algebra,

$$a_j := |f_n - F_j|_W^2 - \left|f_n - \sum_{j=1}^N \lambda_j f_{\theta'(j)}\right|_W^2 = -2\left\langle f_n, F_j - \tilde{f}\right\rangle_W + |F_j|_W^2 - \left|\tilde{f}\right|_W^2.$$

By definition of F_j,

$$\begin{aligned} a_j &= \inf_{\theta\in\Theta}\left\{-2\left\langle f_n, \left(1-\frac{1}{j}\right)F_{j-1} + \frac{1}{j}f_\theta - \tilde{f}\right\rangle_W + \left|\left(1-\frac{1}{j}\right)F_{j-1} + \frac{1}{j}f_\theta\right|_W^2 - \left|\tilde{f}\right|_W^2\right\} \\ &= \left(1-\frac{1}{j}\right)\left[-2\left\langle f_n, F_{j-1} - \tilde{f}\right\rangle_W + |F_{j-1}|_W^2 - \left|\tilde{f}\right|_W^2\right] - \frac{1}{j}\left[\left|\tilde{f}\right|_W^2 + \left(1-\frac{1}{j}\right)|F_{j-1}|_W^2\right] \\ &\quad + \frac{1}{j}\inf_{\theta\in\Theta}\left\{-2\left\langle f_n, f_\theta - \tilde{f}\right\rangle_W + 2\left(1-\frac{1}{j}\right)\langle F_{j-1}, f_\theta\rangle_W + \frac{1}{j}|f_\theta|_W^2\right\}. \end{aligned}$$

By the same argument in Jones (1992) and Barron (1993), for λ_j's as in (12),

$$\begin{aligned} &\inf_{\theta\in\Theta}\left\{-2\left\langle f_n, f_\theta - \tilde{f}\right\rangle_W + 2\left(1-\frac{1}{j}\right)\langle F_{j-1}, f_\theta\rangle_W + \frac{1}{j}|f_\theta|_W^2\right\} \\ \leq\ &\sum_{j=1}^N \lambda_j\left\{-2\left\langle f_n, f_{\theta'(j)} - \tilde{f}\right\rangle_W + 2\left(1-\frac{1}{j}\right)\langle F_{j-1}, f_{\theta'(j)}\rangle_W + \frac{1}{j}\left|f_{\theta'(j)}\right|_W^2\right\} \\ =\ &2\left(1-\frac{1}{j}\right)\left\langle F_{j-1}, \tilde{f}\right\rangle_W + \frac{1}{j}\sum_{j=1}^N \lambda_j\left|f_{\theta'(j)}\right|_W^2, \end{aligned}$$

by definition of $\tilde{f}$ in (12). Noting the definition of a_j, and substituting the last display in a_j,

$$\begin{aligned} a_j &\leq \left(1-\frac{1}{j}\right) a_{j-1} - \frac{1}{j}\left[\left|\tilde{f}\right|_W^2 + \left(1-\frac{1}{j}\right)|F_{j-1}|_W^2\right] \\ &\quad +2\frac{1}{j}\left(1-\frac{1}{j}\right)\left\langle F_{j-1}, \tilde{f}\right\rangle_W + \frac{1}{j^2}\sum_{j=1}^{N}\lambda_j \left|f_{\theta'(j)}\right|_W^2 \\ &\leq \left(1-\frac{1}{j}\right) a_{j-1} - \frac{1}{j}\left(1-\frac{1}{j}\right)\left|F_{j-1}-\tilde{f}\right|_W^2 + \frac{1}{j^2}\sum_{j=1}^{N}\lambda_j \left|f_{\theta'(j)}\right|_W^2 \\ &\leq \left(1-\frac{1}{j}\right) a_{j-1} + \frac{1}{j^2}\sup_{\theta\in\Theta}|f_\theta|_W^2, \end{aligned}$$

where the second inequality follows by noting that $-\left|\tilde{f}\right|_W^2 \leq -\left(1-j^{-1}\right)\left|\tilde{f}\right|_W^2$ and then completing the square. By Condition 2, for any characteristic function f, $|f|_W^2 \leq 1$. Then,

$$\begin{aligned} a_1: &= \inf_{\theta\in\Theta}|f_n - f_\theta|_W^2 - \left|f_n - \sum_{j=1}^{N}\lambda_j f_{\theta'(j)}\right|_W^2 \\ &\leq \inf_{\theta\in\Theta}|f_n - f_\theta|_W^2 \\ &\leq |f_n|_W^2 \\ &\leq 1. \end{aligned}$$

By the above display, and the fact that $\sup_{\theta\in\Theta}|f_\theta|_W^2 \leq 1$, the recursion becomes

$$\begin{aligned} a_1 &\leq 1 \\ a_j &\leq \left(\frac{j-1}{j}\right) a_{j-1} + \frac{1}{j^2}, \end{aligned}$$

for $j > 1$, This is then bounded by Lemma 2. ■

Lemma 2 *Suppose $(a_j)_{j\geq 1}$ is a sequence of non-negative numbers such that $a_1 \leq A$ and*

$$a_j = \left(1-\frac{1}{j}\right) a_{j-1} + \frac{A}{j^2}.$$

Then, for $J \geq 2$,

$$a_J \leq \frac{4A\ln J}{J}.$$

Proof. To find the order of magnitude of the recursion, iterate to find

$$a_J \leq \prod_{j=2}^{J}\left(\frac{j-1}{j}\right) a_1 + A\sum_{j=2}^{J}\frac{1}{j^2}\prod_{s=1}^{J-j}\left(\frac{J-s}{J+1-s}\right)$$

where the empty product $\prod_{s=1}^{0}$ is set to one. The products are seen to telescope as follows

$$\prod_{j=2}^{J}\left(\frac{j-1}{j}\right)=\frac{1}{J},$$

$$\prod_{s=1}^{J-j}\left(\frac{J-s}{J+1-s}\right)=\frac{j}{J},$$

so that

$$a_J \leq \left(\frac{1}{J}\right)a_1 + A\sum_{j=2}^{J}\frac{1}{j^2}\frac{j}{J} \leq \left(\frac{1}{J}\right)a_1 + \frac{A}{J}\left(1+\ln J\right)$$

and the result follows if $a_1 \leq A$, noting that $2+\ln J < 4\ln J$ when $J \geq 2$. ■

4.2 Proof of Theorem 2

The proof shall make use of some technical lemmata. The reader can directly go to the proof of Theorem 2 and refer to them, as needed.

The following allows to control the first two moments of the objective function using V and U statistics.

Lemma 3 *Let f be a characteristic function on $\mathbb{R}^K$ and suppose Condition 2 holds. If $\mathbb{E}f_n = f$,*

$$\mathbb{E}\left\langle\left(f_n - f\right),\left(f_n - f\right)\right\rangle_W = O\left(n^{-1}\right),$$

$$Var\left(\left\langle\left(f_n - f\right),\left(f_n - f\right)\right\rangle_W\right) = O\left(n^{-2}\right),$$

if $\mathbb{E}f_n \neq f$,

$$Var\left(\left\langle\left(f_n - f\right),\left(f_n - f\right)\right\rangle_W\right) = O\left(n^{-1}\right).$$

Moreover, for any characteristic function f,

$$\mathbb{E}\left\langle f_n - f, f_n - f\right\rangle_W \quad = \quad \left\langle \mathbb{E}f_n - f, \mathbb{E}f_n - f\right\rangle_W + O\left(n^{-1}\right).$$

Proof. Write

$$\begin{aligned}
& \left\langle\left(f_n - f\right),\left(f_n - f\right)\right\rangle_W \\
= & \int_{\mathbb{R}^K}\int_{\mathbb{R}^K}\left(\frac{1}{n}\sum_{j=1}^{n}\exp\left\{i\left\langle s, X_j\right\rangle\right\} - f\left(s\right)\right)\left(\frac{1}{n}\sum_{j=1}^{n}\exp\left\{-i\left\langle t, X_j\right\rangle\right\} - \overline{f\left(t\right)}\right)W\left(s,t\right)dsdt \\
= & \frac{1}{n^2}\sum_{j_1,j_2=1}^{n} g\left(X_{j_1}, X_{j_2}\right) = \frac{1}{n^2}\sum_{j_1\neq j_2} g\left(X_{j_1}, X_{j_2}\right) + \frac{1}{n^2}\sum_{j=1}^{n} g\left(X_j, X_j\right) \\
=: & \ I_1 + I_2,
\end{aligned}$$

where

$$g\left(X_{j_1}, X_{j_2}\right) := \int_{\mathbb{R}^K}\int_{\mathbb{R}^K}\left(\exp\left\{i\left\langle s, X_{j_1}\right\rangle\right\} - f\left(s\right)\right)\left(\exp\left\{-i\left\langle t, X_{j_2}\right\rangle\right\} - \overline{f\left(t\right)}\right)W\left(s,t\right)dsdt.$$

The above is a decomposition of a V-statistic in terms of two uncorrelated U-statistics, of order 2 and 1 respectively. If $\mathbb{E}f_n = f$, $g(X_{j_1}, X_{j_2})$ is a degenerate kernel, i.e. $\mathbb{E}g(x, X_j) = 0$. Then, $\mathbb{E}I_1 = 0$, while $\mathbb{E}I_2 \lesssim n^{-1}$. Moreover, by the variance of U-statistics for degenerate kernels (e.g., Serfling, 1980), $Var(I_1) \lesssim n^{-2}$, implying

$$\begin{aligned} Var\left(\langle (f_n - f), (f_n - f)\rangle_W\right) &= Var(I_1) + Var(I_2) \\ &\lesssim n^{-2}, \end{aligned}$$

as the two U-statistics are uncorrelated by degeneracy of the kernel, and because $Var(I_2) \lesssim n^{-3}$, as I_2 is the sum of n uncorrelated terms divided by n^{-2}. On the other hand, if $\mathbb{E}f_n \neq f$, the U-statistic is not degenerate and we have the second result for the variance. The last result follows using the fact that $g(X_{j_1}, X_{j_2})$ is a degenerate kernel ($j_1 \neq j_2$), so after some algebra,

$$\begin{aligned} & \mathbb{E}\langle f_n - f, f_n - f\rangle_W \\ & - \langle \mathbb{E}f_n - f, \mathbb{E}f_n - f\rangle_W \\ = & \frac{1}{n^2}\sum_{j=1}^{n}\int_{\mathbb{R}^K}\int_{\mathbb{R}^K} Cov\left(\exp\{i\langle s, X_j\rangle\} - f(s), \exp\{-i\langle t, X_j\rangle\} - \overline{f(t)}\right) W(s,t)\, ds dt \\ \lesssim & \frac{1}{n}. \end{aligned}$$

■

Control of the estimation error (a.s.) requires a strong uniform law of large numbers with rates of convergence.

Lemma 4 *Let $\mathcal{G}$ be the set of all laws with support in Θ. If Condition 1 and 2 hold with $N(\epsilon) = O\left(\epsilon^{-V}\right)$ for some finite V, then,*

$$\sup_{G\in\mathcal{G}}\left|(1-\mathbb{E})\left|f_n - \int_\Theta f_\theta dG(\theta)\right|_W^2\right| = O_{a.s.}\left(\sqrt{\frac{\ln n}{n}}\right).$$

Proof. Define

$$g(X_1^n|G) := (1-\mathbb{E})\left\langle f_n - \int_\Theta f_\theta dG(\theta), f_n - \int_\Theta f_\theta dG(\theta)\right\rangle_W,$$

and

$$g(X_1^n|\theta,\theta') := (1-\mathbb{E})\langle f_n - f_\theta, f_n - f_{\theta'}\rangle_W,$$

where $X_1^n := (X_1, ..., X_n)$. Note that

$$g(X_1^n|G) = \int_\Theta\int_\Theta g(X_1^n|\theta,\theta')\, dG(\theta)\, dG(\theta')$$

Let

$$X_1^n(j) = \left(X_1, ..., X_{j-1}, X_j', X_{j+1}, ..., X_n\right)$$

be just as X_1^n but with the j^{th} observation replaced by an independent copy X_j' of X_j. Suppose that there is a finite absolute constant C such that

$$\mathbb{E}\left[\sum_{j=1}^{n}\left|g(X_1^n|G) - g(X_1^n(j)|G)\right|^2 |X_1^n\right] \leq \frac{C}{n}. \tag{13}$$

Then, Corollary 3 in Boucheron et al. (2003) gives

$$\Pr\left(\left|g\left(X_1^n|G\right)\right| \geq x\right) \leq 2\exp\left\{-\frac{nx^2}{4C}\right\}, \tag{14}$$

implying sub-Gaussian tails. To verify (13), let

$$f_{nj}\left(t\right) := \frac{1}{n}\left[\exp\left\{i\exp\left\{i\left\langle t, X_j'\right\rangle\right\}\right\} + \sum_{s\neq j}\exp\left\{i\exp\left\{i\left\langle t, X_s\right\rangle\right\}\right\}\right], \tag{15}$$

i.e. f_n computed using $X_1^n\left(j\right)$ rather than X_1^n. Then, by definition of $g\left(\bullet|G\right)$

$$\begin{aligned}
&\sum_{j=1}^{n}\left|g\left(X_1^n|G\right) - g\left(X_1^n\left(j\right)|G\right)\right|^2 \\
= &\sum_{j=1}^{n}\left[\left(1-\mathbb{E}\right)\left(\left|f_n - \int_\Theta f_\theta dG\left(\theta\right)\right|_W^2 - \left|f_{nj} - \int_\Theta f_\theta dG\left(\theta\right)\right|_W^2\right)\right]^2 \\
\lesssim &\sum_{j=1}^{n}\left[\left|f_n - \int_\Theta f_\theta dG\left(\theta\right)\right|_W^2 - \left|f_{nj} - \int_\Theta f_\theta dG\left(\theta\right)\right|_W^2\right]^2 \\
= &\sum_{j=1}^{n}\left[\left(\left\langle f_n - \int_\Theta f_\theta dG\left(\theta\right), f_n - \int_\Theta f_\theta dG\left(\theta\right)\right\rangle_W - \left\langle f_{nj} - \int_\Theta f_\theta dG\left(\theta\right), f_n - \int_\Theta f_\theta dG\left(\theta\right)\right\rangle_W\right)\right. \\
&\left. + \left(\left\langle f_{nj} - \int_\Theta f_\theta dG\left(\theta\right), f_n - \int_\Theta f_\theta dG\left(\theta\right)\right\rangle_W - \left\langle f_{nj} - \int_\Theta f_\theta dG\left(\theta\right), f_{nj} - \int_\Theta f_\theta dG\left(\theta\right)\right\rangle_W\right)\right]^2 \\
=: &\sum_{j=1}^{n}\left[I_{1j} + I_{2j}\right]^2.
\end{aligned}$$

By linearity of the inner product, given that characteristic functions are bounded by 1, and that, using (15),

$$\begin{aligned}
\left|f_n\left(t\right) - f_{nj}\left(t\right)\right| &= \frac{1}{n}\left|\exp\left\{i\left\langle t, X_j'\right\rangle\right\} - \exp\left\{i\left\langle t, X_j\right\rangle\right\}\right| \\
&\leq \frac{2}{n},
\end{aligned}$$

one finds

$$\begin{aligned}
I_{1j} &= \left\langle f_n - f_{nj}, f_n - \int_\Theta f_\theta dG\left(\theta\right)\right\rangle_W \\
&\lesssim \frac{1}{n}.
\end{aligned}$$

The same bound holds for I_{2j}. Hence,

$$\begin{aligned}
\sum_{j=1}^{n}\left[I_{1j} + I_{2j}\right]^2 &\lesssim \sum_{j=1}^{n}\frac{1}{n^2} \\
&= \frac{1}{n},
\end{aligned}$$

implying that (13), hence (14) holds.

As remarked in Rakhlin et al. (2005, p. 225),

$$\max_{G\in\mathcal{G}} |g\left(X_1^n|G\right)| = \max_{\theta,\theta'\in\Theta} |g\left(X_1^n|\theta,\theta'\right)| \tag{16}$$

because the maximum is achieved at one of the edges of the convex hull. By Condition 1, there is a set Θ_ϵ of $N = N(\epsilon)$ points in Θ such that for any two points $\theta(1), \theta(3) \in \Theta$ one can find two point $\theta(2), \theta(4) \in \Theta_\epsilon$, satisfying $\left|f_{\theta(1)} - f_{\theta(2)}\right|_W \leq \epsilon^{1/2}$, and similarly for $\theta(3)$ and $\theta(4)$. By direct algebra, the Cauchy–Schwarz inequality, and the aforementioned remark,

$$\begin{aligned}
& \left\langle f_n - f_{\theta(1)}, f_n - f_{\theta(3)}\right\rangle_W \\
= \; & \left\langle f_n - f_{\theta(1)}, f_n - f_{\theta(3)}\right\rangle_W - \left\langle f_n - f_{\theta(2)}, f_n - f_{\theta(3)}\right\rangle_W \\
& + \left\langle f_n - f_{\theta(2)}, f_n - f_{\theta(3)}\right\rangle_W - \left\langle f_n - f_{\theta(2)}, f_n - f_{\theta(4)}\right\rangle_W \\
& + \left\langle f_n - f_{\theta(2)}, f_n - f_{\theta(4)}\right\rangle_W \\
\leq \; & \left\langle f_n - f_{\theta(2)}, f_n - f_{\theta(4)}\right\rangle_W + \left|f_{\theta(2)} - f_{\theta(1)}\right|_W \left|f_n - f_{\theta(3)}\right|_W \\
& + \left|f_{\theta(4)} - f_{\theta(3)}\right|_W \left|f_n - f_{\theta(2)}\right|_W \\
\leq \; & \left\langle f_n - f_{\theta(2)}, f_n - f_{\theta(4)}\right\rangle_W + 4\epsilon^{1/2}.
\end{aligned}$$

Hence, (16) and the previous display give

$$\sup_{G\in\mathcal{G}} |g\left(X_1^n|G\right)|^{1/2} \leq \max_{\theta,\theta'\in\Theta_\epsilon} |g\left(X_1^n|\theta,\theta'\right)| + 4\epsilon^{1/2}. \tag{17}$$

By the union bound, and (16), it also follows that

$$\Pr\left(\max_{\theta,\theta'\in\Theta_\epsilon} |g\left(X_1^n|\theta,\theta'\right)| \geq x\right) \leq 2N \exp\left\{-\frac{nx^2}{4C}\right\}.$$

By the conditions of the lemma, there exists a finite absolute constant a, depending on V, such that $\ln N(\epsilon) \leq a \ln n$ when $\epsilon = n^{-1}$. Hence, for $x \gtrsim \left[3a\left(4C\ln n\right)/n\right]^{1/2}$, the above display is summable in n, implying a.s. convergence by the Borel-Cantelli Lemma. Substituting in (17), when $\epsilon = n^{-1}$ gives the result. ■

The following local uniform control is needed to derive sharp rates of convergence, in probability, for the estimator.

Lemma 5 *For any $\delta > 0$ under Condition 2,*

$$\begin{aligned}
& \mathbb{E} \sup_{|f_G - f_{G'}|_W \leq \delta} \left| |f_n - f_G|_W^2 - |f_n - f_{G'}|_W^2 - \left(|\mathbb{E}f_n - f_G|_W^2 - |\mathbb{E}f_n - f_G|_W^2\right)\right| \\
\lesssim \; & \frac{\delta}{\sqrt{n}}.
\end{aligned}$$

Proof. By algebraic manipulation,

$$\begin{aligned}
& |f_n - f_G|_W^2 - |f_n - f_{G'}|_W^2 \\
= \; & \langle f_{G'} - f_G, f_n - f_G\rangle_W + \langle f_n - f_{G'}, f_{G'} - f_G\rangle_W \\
= \; & 2\langle f_{G'} - f_G, f_n - f_G\rangle_W - \langle f_{G'} - f_G, f_{G'} - f_G\rangle_W,
\end{aligned}$$

where the first equality follows from the definition of $|\bullet|_W^2$ and then adding and subtracting $\langle f_n - f_{G'}, f_n - f_G\rangle_W$. Similar calculations give

$$|\mathbb{E}f_n - f_G|_W^2 - |\mathbb{E}f_n - f_{G'}|_W^2 = 2\langle f_{G'} - f_G, \mathbb{E}f_n - f_G\rangle_W - \langle f_{G'} - f_G, f_{G'} - f_G\rangle_W .$$

The above displays imply

$$\begin{aligned}
& \mathbb{E}\sup_{|f_G - f_{G'}|_W \le \delta} \left| |f_n - f_G|_W^2 - |f_n - f_{G'}|_W^2 - \left(|\mathbb{E}f_n - f_G|_W^2 - |\mathbb{E}f_n - f_G|_W^2 \right) \right| \\
= \; & 2\mathbb{E}\sup_{|f_G - f_{G'}|_W \le \delta} |\langle f_{G'} - f_G, f_n - \mathbb{E}f_n\rangle_W| \\
\le \; & 2\sup_{|f_G - f_{G'}|_W \le \delta} |f_{G'} - f_G|_W \, \mathbb{E}\left|(1-\mathbb{E})\, f_n\right|_W \\
= \; & 2\delta \left(\mathbb{E}\left|(1-\mathbb{E})\, f_n\right|_W^2 \right)^{1/2} \\
\lesssim \; & \delta n^{-1/2},
\end{aligned}$$

by Lemma 3. ■

Equipped with the above technical tools, Theorem 2 can be proved.

Proof. [Theorem 2] Define $G_n(\bullet) = \sum_{j=1}^J \delta_{\hat\theta(j)}(\bullet)/J$ where $\delta_{\hat\theta(j)}(\bullet)$ is the point mass at $\hat\theta(j)$, where $\hat\theta(j)$ is as defined in the algorithm in Table 1. Then,

$$\frac{1}{J}\sum_{j=1}^J f_{\hat\theta(j)} = \int_\Theta f_\theta dG_n(\theta).$$

Noting that $\mathbb{E}f_n = f_P$ because the empirical characteristic function is unbiased, decompose

$$\left| \mathbb{E}f_n - \int_\Theta f_\theta dG_n(\theta) \right|_W^2 = \sum_{s=1}^5 T_s,$$

where

$$\begin{aligned}
T_1 \; := \; & \left| \mathbb{E}f_n - \int_\Theta f_\theta dG_n(\theta) \right|_W^2 \\
& - \mathbb{E}\left| f_n - \int_\Theta f_\theta dG_n(\theta) \right|_W^2
\end{aligned}$$

$$T_2 := (\mathbb{E}-1)\left| f_n - \int_\Theta f_\theta dG_n(\theta) \right|_W^2$$

$$\begin{aligned}
T_3 \; := \; & \left| f_n - \int_\Theta f_\theta dG_n(\theta) \right|_W^2 \\
& - \left| f_n - \int_\Theta f_\theta dP(\theta) \right|_W^2
\end{aligned}$$

$$T_4 := (1-\mathbb{E})\left| f_n - \int_\Theta f_\theta dP(\theta) \right|_W^2$$

$$
\begin{aligned}
T_5 &:= \mathbb{E}\left|f_n - \int_\Theta f_\theta dP(\theta)\right|_W^2 \\
&\quad - \left|\mathbb{E}f_n - \int_\Theta f_\theta dP(\theta)\right|_W^2
\end{aligned}
$$

$$
T_6 := \left|\mathbb{E}f_n - \int_\Theta f_\theta dP(\theta)\right|_W^2.
$$

Now, $T_1 = O\left(n^{-1}\right)$, by Lemma 3;

$$
\begin{aligned}
T_2 &\leq \sup_{G\in\mathcal{G}}\left|(1-\mathbb{E})\left\langle f_n - \int_\Theta f_\theta dG(\theta), f_n - \int_\Theta f_\theta dG(\theta)\right\rangle_W\right| \\
&= O_{a.s.}\left(\sqrt{\frac{\ln n}{n}}\right)
\end{aligned}
$$

by Lemma 4;

$T_3 = O(\ln J/J)$ by Theorem 1;

$T_4 = O_{a.s.}\left(\sqrt{\ln n/n}\right)$ by Lemma 4;

$T_5 = O\left(n^{-1}\right)$ by Lemma 3;

$T_6 = 0$, as $\mathbb{E}f_n = \int_\Theta f_\theta dP(\theta)$.

To show the rate in the convergence in probability, it is enough to verify the conditions of Theorem 3.4.1 in van der Vaart and Wellner (2000). In the present context, Problem 3.4.5 in van der Vaart and Wellner (2000) says that for any G such that

$$
\left|\int_\Theta f_\theta d\left(G_n(\theta) - G(\theta)\right)\right|_W \geq 4\left|\int_\Theta f_\theta d\left(P(\theta) - G(\theta)\right)\right|_W, \tag{18}
$$

then,

$$
\mathbb{E}\left|f_n - \int_\Theta f_\theta dG_n(\theta)\right|_W^2 - \mathbb{E}\left|f_n - \int_\Theta f_\theta dG(\theta)\right|_W^2 \geq \frac{1}{4}\left|\int_\Theta f_\theta d\left(G_n(\theta) - G(\theta)\right)\right|_W.
$$

Suppose, for the moment that (18) holds. Then, the above display satisfies the first condition in Theorem 3.4.1 in van der Vaart and Wellner (2000). The second condition in the same theorem is satisfied using Lemma 5. The third condition in that theorem requires to find a diverging sequence r_n such that, in this specific context,

$$
\left|\int_\Theta f_\theta d\left(P(\theta) - G(\theta)\right)\right|_W \lesssim r_n^{-1}
$$

and $r_n \leq n^{1/2}$. The sequence also needs to satisfy

$$
\left|f_n - \int_\Theta f_\theta dG_n(\theta)\right|_W^2 \leq \inf_{G\in\mathcal{G}}\left|f_n - \int_\Theta f_\theta dG(\theta)\right|_W^2 + O_p\left(r_n^{-2}\right),
$$

where $\mathcal{G}$ is as in Theorem 1. Hence, by Theorem 1, when $J/\ln J \gtrsim n$, one can choose $r_n = n^{1/2}$. The above two display also imply that (18) is true. Hence, Theorem 3.4.1

in van der Vaart and Wellner (2000) states that r_n^{-1} is the resulting convergence rate in probability.

Finally, the last part of the theorem follows from the fact that convergence under $|\bullet|_W$ with W satisfying Condition 2 implies convergence of the density function for almost all x's. ■

4.3 Proof of Corollaries

Corollaries 1, 2 and 3 are a consequence of Theorem 2 and the following three lemmata used to verify Condition 1 for each application.

Lemma 6 *Consider Corollary 1 and suppose its conditions hold. There is a cover* $(A_r)_{r\leq N}$ *for* Θ*, with* $\ln N = O(\ln(1/\epsilon))$*, such that*

$$\left\langle f_{\theta(1)} - f_{\theta(2)}, f_{\theta(3)} - f_{\theta(4)} \right\rangle_W \leq \epsilon$$

for $\theta(1), \theta(1) \in A_r$ *,* $\theta(3), \theta(4) \in A_s$ *for any* $r, s \leq N$*.*

Proof. Let $B(\epsilon) := \left\{t \in \mathbb{R}^K : |t_k| \leq \epsilon, k = 1, 2, ..., K\right\}$. Since $f_\theta(t) = \exp(-\theta \langle t, \Sigma t\rangle/2) \leq 1$, by (3),

$$\begin{aligned} & \left\langle \left(f_{\theta(1)} - f_{\theta(2)}\right) B(\epsilon), \left(f_{\theta(3)} - f_{\theta(4)}\right) B(\epsilon) \right\rangle_W \\ & + \left\langle \left(f_{\theta(1)} - f_{\theta(2)}\right) B(\epsilon), \left(f_{\theta(3)} - f_{\theta(4)}\right) \right\rangle_W \\ & + \left\langle \left(f_{\theta(1)} - f_{\theta(2)}\right), \left(f_{\theta(3)} - f_{\theta(4)}\right) B(\epsilon) \right\rangle_W \\ \lesssim \quad & \epsilon \end{aligned}$$

so that we only need to find a cover for

$$\left\langle \left(f_{\theta(1)} - f_{\theta(2)}\right) B^c(\epsilon), \left(f_{\theta(3)} - f_{\theta(4)}\right) B^c(\epsilon) \right\rangle_W$$

where $B^c(\epsilon)$ is the complement of $B(\epsilon)$. Since Σ is positive definite, there is a constant $\tau > 0$ such that $\langle 1_K, \Sigma 1_K\rangle \geq \tau$. For fixed $\epsilon > 0$, define $\bar{\theta} = \bar{\theta}(\epsilon)$ such that $\exp\left(-\bar{\theta}\tau\epsilon/2\right) \leq \epsilon$, e.g., $\tau\bar{\theta} = -\epsilon^{-1}\ln\epsilon$. Note that on $B(\epsilon)$, $\langle t, \Sigma t\rangle \geq \epsilon \langle 1_K, \Sigma 1_K\rangle \geq \epsilon\tau$. This implies that for $\theta(1), \theta(2) \geq \bar{\theta}$, and unrestricted $\theta(3), \theta(4)$, by linearity of the inner product and the integrability condition on W,

$$\left\langle \left(f_{\theta(1)} - f_{\theta(2)}\right) B^c(\epsilon), \left(f_{\theta(3)} - f_{\theta(4)}\right) B^c(\epsilon) \right\rangle_W \lesssim \epsilon.$$

Hence, it is sufficient to find a cover for $\theta < \bar{\theta}$. By the mean value theorem, for θ^* in the convex hull of $\{\theta(1), \theta(2)\}$,

$$f_{\theta(1)}(t) - f_{\theta(2)}(t) = -\frac{1}{2}\exp\left(-\theta^*\frac{\langle t, \Sigma t\rangle}{2}\right)\langle t, \Sigma t\rangle\left(\theta(1) - \theta(2)\right). \tag{19}$$

Hence, by linearity, letting θ^{**} be in the convex hull of $\{\theta(3), \theta(4)\}$, using (19),

$$\begin{aligned} & \left\langle \left(f_{\theta(1)} - f_{\theta(2)}\right) B^c(\epsilon), \left(f_{\theta(3)} - f_{\theta(4)}\right) B^c(\epsilon) \right\rangle_W \\ = \quad & \int_{B^c(\epsilon)} \int_{B^c(\epsilon)} \left(f_{\theta(1)}(s) - f_{\theta(1)}(s)\right)\left(f_{\theta(3)}(t) - f_{\theta(4)}(t)\right) W(s,t)\, dsdt \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{4}\int_{B^c(\epsilon)}\int_{B^c(\epsilon)} W(s,t)\exp\left(-\theta^*\frac{\langle s,\Sigma s\rangle}{2}\right)\exp\left(-\theta^*\frac{\langle t,\Sigma t\rangle}{2}\right)\langle s,\Sigma s\rangle\langle t,\Sigma t\rangle\, dsdt \\
&\quad \times |\theta(1)-\theta(2)|\,|\theta(3)-\theta(4)| \\
&\lesssim |\theta(1)-\theta(2)|\,|\theta(3)-\theta(4)|,
\end{aligned}
$$

by the condition of the Theorem using the fact that the entries in Σ are bounded. Hence, we can deduce that there is a cover of cardinality $N = O\left(\bar{\theta}/\epsilon^{1/2}\right) = O\left(\ln(1/\epsilon)/\epsilon^{3/2}\right)$ so that $\ln N = O(\ln(1/\epsilon))$ and the Lemma is proved. ■

Lemma 7 *Consider Corollary 2 and suppose its conditions hold. Suppose $\Sigma_{kl}(\theta)$ is as in (5). There is a cover $(A_r)_{r\leq N}$ for Θ, with $N = O\left(\epsilon^{-L}\right)$, depending on B only, such that*

$$\left\langle f_{\theta(1)} - f_{\theta(2)}, f_{\theta(3)} - f_{\theta(4)}\right\rangle_W \leq \epsilon$$

for $\theta(1), \theta(1) \in A_r$, $\theta(3), \theta(4) \in A_s$ for any $r, s \leq N$.

Proof. By change of variables, the Gaussian copula $C(u|B,\theta)$ corresponds to a Gaussian density $\phi(x|\Sigma)$ which has mean zero and covariance matrix $\Sigma := \Sigma(\theta)$ as in (5). The change of variables does not change the mixing law P. With abuse of notation, let

$$f_\Sigma(t) := \exp\left(-\frac{\langle t,\Sigma t\rangle}{2}\right)$$

be the characteristic function of $\phi(x|\Sigma)$. Then, by the mean value theorem,

$$f_{\Sigma(1)}(t) - f_{\Sigma(2)}(t) = -\frac{1}{2}\exp\left(-\frac{\langle t,\Sigma^* t\rangle}{2}\right)\sum_{k,l} t_k t_l \Sigma^*_{kl}\left(\Sigma_{kl}(1) - \Sigma_{kl}(2)\right), \tag{20}$$

for covariance matrices $\Sigma(1)$ and $\Sigma(2)$ and Σ^* in the convex hull of $\{\Sigma(1), \Sigma(2)\}$.

Since Σ is a correlation matrix, all the entries are in $[-1,1]$. Hence, an ϵ-cover for $[-1,1]^K$ is an ϵ-cover for the set of covariance matrices. Such a cover, say A_r, $r = 1,2,...,N$, is of cardinality $N = (2\epsilon)^{-K}$. Also, (20) and (3) imply that

$$\sup_{\Sigma(1),\Sigma(2)\in A_r,\Sigma(3),\Sigma(4)\in A_s}\left\langle f_{\Sigma(1)} - f_{\Sigma(2)}, f_{\Sigma(3)} - f_{\Sigma(4)}\right\rangle_W \lesssim \epsilon^2. \tag{21}$$

Pick up a unique point in each A_r, $r = 1,2,...,N$, (the cover of $[-1,1]^K$) and identify a (possibly non-unique) point in Θ, to generate a finite set with cardinality $N = (2\epsilon)^{-K}$, say Θ_ϵ. Then, from (21)

$$\inf_{\theta(2),\theta(4)\in\Theta_\epsilon}\sup_{\theta(1),\theta(3)\in\Theta}\left\langle f_{\theta(1)} - f_{\theta(2)}, f_{\theta(3)} - f_{\theta(4)}\right\rangle_W \lesssim \epsilon^2 \leq \epsilon,$$

as required. ■

Lemma 8 *Consider Corollary 3 and suppose its conditions hold. There is a cover $(A_r)_{r\leq N}$ for Θ, with $N = O\left(\epsilon^{-L}\right)$, such that*

$$\left\langle f_{\theta(1)} - f_{\theta(2)}, f_{\theta(3)} - f_{\theta(4)}\right\rangle_W \leq \epsilon$$

for $\theta(1), \theta(1) \in A_r$, $\theta(3), \theta(4) \in A_s$ for any $r, s \leq N$.

Proof. The characteristic function of $\phi(x-\theta)$ is $f_\theta(t) = \exp\left\{i\theta t - \left(t^2/2\right)\right\}$. By Euler's formula,

$$\exp\{i\theta(1)t\} - \exp\{i\theta(2)t\} = \cos(\theta(1)t) - \cos(\theta(2)t) + i\left[\sin((\theta(1)t)) - \sin((\theta(2)t))\right].$$

Since

$$\begin{aligned} Real\left(f_{\theta(1)}(t) - f_{\theta(2)}(t)\right) &\leq |\cos(\theta(1)t) - \cos(\theta(2)t)| \exp\left\{-t^2/2\right\} \\ &\leq t\exp\left\{-t^2/2\right\}|\theta(1) - \theta(2)| \\ &\lesssim |\theta(1) - \theta(2)|, \end{aligned}$$

and similarly for the complex part (i.e. the complex sine),

$$\begin{aligned} \left\langle\left(f_{\theta(1)} - f_{\theta(2)}\right), \left(f_{\theta(3)} - f_{\theta(4)}\right)\right\rangle_W &\lesssim \int_{\mathbb{R}}\int_{\mathbb{R}} W(s,t)\, dsdt\, |\theta(1) - \theta(2)|\, |\theta(3) - \theta(4)| \\ &\lesssim |\theta(1) - \theta(2)|\, |\theta(3) - \theta(4)|. \end{aligned}$$

Since $\Theta = \mathbb{R}$ is not compact, it is necessary to trim the sides of Θ and show that the supremum over the trimmed version is the same as over Θ. To this end, define $B(\epsilon) := \{t \in \mathbb{R} : |t| \leq \epsilon/8\}$. Then,

$$\begin{aligned} \int_{B(\epsilon)} \left|\exp\left\{i\theta t - \left(t^2/2\right)\right\}\right| dt &\leq \int_{B(\epsilon)} dt \\ &\leq \epsilon/4 \end{aligned}$$

implying

$$\sup_{\theta(1),\theta(2),\theta(3),\theta(4)} \left\langle\left(f_{\theta(1)} - f_{\theta(2)}\right), \left(f_{\theta(3)} - f_{\theta(4)}\right) B(\epsilon)\right\rangle_W \leq \epsilon. \tag{22}$$

Now, note that for any $\theta(1) \in \Theta$ and $t \in \mathbb{R}$, there is a $\theta(2) \in [0, 2\pi/t]$ such that $\exp\{i\theta(1)t\} - \exp\{i\theta(2)t\} = 0$. By the aforementioned remark and (22) let $t = \epsilon/8$ so that $[0, 2\pi/t]$ becomes $\left[0, 16\pi\epsilon^{-1}\right]$. Define $\Theta_\epsilon := \{\theta'(1), \theta'(2), ..., \theta'(N)\}$ where $\theta'(j) = j\epsilon$ and $N = 16\pi\epsilon^{-1}$. Then,

$$\begin{aligned} &\inf_{\theta(2),\theta(4)\in\Theta_\epsilon} \sup_{\theta(1),\theta(3)\in\Theta} \left\langle\left(f_{\theta(1)} - f_{\theta(2)}\right) B^c(\epsilon), \left(f_{\theta(3)} - f_{\theta(4)}\right) B^c(\epsilon)\right\rangle_W \\ \leq\ &\epsilon. \end{aligned}$$

Redefining $\epsilon = \epsilon/3$ shows that Condition 1 holds. ■

Bio Sketch

Alessio Sancetta is Professor in the Department of Economics at Royal Holloway University of London, UK. He received a PhD from the University of Cambridge.

References

[1] Arcones, M.A. and E. Giné (1992) On the Bootstrap of U and V Statistics. Annals of Statistics 20, 655-674.

[2] Barron A.R. (1993) Universal Approximation Bounds for Superpositions of a Sigmoidal Function. IEEE Transactions on Information Theory 39, 930-944.

[3] Bingham, N.H., C.M. Goldie and J.L. Teugels (1989) Regular Variation. Cambridge: Cambridge University Press.

[4] Boucheron, S., G. Lugosi and P. Massart (2003) Concentration Inequalities Using the Entropy Method. Annals of Probability 31, 1583-1614.

[5] Carrasco, M. and J.-P. Florens (2002) Efficient GMM Estimation Using the Empirical Characteristic Function. IDEI Working Paper 140.

[6] Devroye, L. and L. Gyorfi (2002) Distribution and Density Estimation. In L. Gyorfi (ed.), Principles of Nonparametric Learning. CISM Courses and Lectures No. 434, pp. 211-270. Vienna: Springer.

[7] Feuerverger, A. and P. McDunnough (1981a) On the efficiency of empirical characteristic function procedures. Journal of the Royal Statistical Society, Series B, 43, 20-27.

[8] Feuerverger, A. and P. McDunnough (1981b) On Some Fourier Methods for Inference. Journal of the American Statistical Association 76, 379-387.

[9] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics 29, 1189-1232.

[10] Ghosh, J.K. and R.V. Ramamoorthi (2003) Bayesian Nonparametrics. Springer.

[11] Ghosal, S., J.K. Ghosh and R.V. Ramamoorthi (1999) Posterior Consistency of Dirichlet Mixtures in Density Estimation. Annals of Statistics 27, 143-158.

[12] Ghosal, S. and A.W. van der Vaart (2001) Entropies and Rates of Convergence for Maximum Likelihood and Bayes Estimation for Mixtures of Normal Densities. Annals of Statistics 29, 1233-1263.

[13] Hua L. and H. Joe (2011) Tail Order and Intermediate Tail Dependence of Multivariate Copulas. Journal of Multivariate Analysis 102, 1454-1471.

[14] Jones, L. (1992) A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training. Annals of Statistic 20, 608-613.

[15] Kano, Y. (1994) Consistency Property of Elliptical Probability Density Functions. Journal of Multivariate Analysis 51, 139-147.

[16] Klemelä, J. (2007) Density Estimation with Stagewise Optimization of the Empirical Risk. Machine Learning 67, 169-195.

[17] Ledford, A. and J. Tawn (1996) Statistics for Near Independence in Multivariate Extreme Values. Biometrika 83, 169-187.

[18] Li, J.Q. and A.R. Barron (2000) Mixture Density Estimation. In S.A. Solla, T.K. Leen and K-R. Mueller (Eds.), Advances in Neural Information Processing Systems 12, 279-285. Cambridge, MA: MIT Press.

[19] Lijoi, A., I. Pruenster and S.G. Walker (2005) On Consistency of Nonparametric Normal Mixtures for Bayesian Density Estimation. Journal of the American Statistical Association 100, 1292-1296.

[20] Lindsay, B. (1995) Mixture Models: Theory, Geometry and Applications. Hayward, CA: IMS.

[21] Manner, H. and J. Segers (2011) Tails of Correlation Mixtures of Elliptical Copulas. Insurance: Mathematics and Economics 48, 153–160.

[22] Martin, R. (2013) An Approximate Bayesian Marginal Likelihood Approach for Estimating Finite Mixtures. Communications in Statistics: Simulation and Computation 42, 1533-1548.

[23] Martin, R. and J.K. Ghosh (2008) Stochastic Approximation and Newton's Estimate of a Mixing Distribution. Statistical Science 23, 365-382.

[24] Martin, R. and S.T. Tokdar (2011) Semiparametric Inference in Mixture Models with Predictive Recursion Marginal Likelihood. Biometrika 98, 567-582.

[25] Newton, M.A. (2002) On a Nonparametric Recursive Estimator of the Mixing Distribution. Sankhyā Ser. A 64, 306–322.

[26] Pearson, K. (1894) Contributions to the Mathematical Theory of Evolution. Philosophical Transactions of the Royal Society A, 185, 71-110.

[27] Rakhlin, A., D. Panchenko and S. Mukherjee (2005) Risk Bounds for Mixture Density Estimation. ESAIM: Probability and Statistics 9, 220-229.

[28] Sancetta, A. and S.E. Satchel (2007) Changing Correlation and Equity Portfolio Diversification Failure for Linear Factor Models during Market Declines. Applied Mathematical Finance 14, 227-242.

[29] Serfling, R.J. (1980) Approximation Theorems of Mathematical Statistics. New York: John Wiley & Sons.

[30] Sethuraman, J. (1994) A Constructive Definition of Dirichlet Priors. Statistica Sinica 4, 639-650.

[31] Tokdar, S.T., R. Martin and J.K. Ghosh (2009) Consistency of a Recursive Estimate of Mixing Distributions. Annals of Statistics 37, 250-2522.

[32] Wu, C.F.J (1983) On the Convergence Properties of the EM Algorithm. Annals of Statistics 11, 95-103.

[33] Yu, J. (2004) Empirical Characteristic Function Estimation and Its Applications. Econometric Reviews 23, 93-123.