Battey, Heather; Linton, Oliver

Working Paper

Nonparametric estimation of multivariate elliptic densities via finite mixture sieves

cemmap working paper, No. CWP41/13

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Battey, Heather; Linton, Oliver (2013) : Nonparametric estimation of multivariate elliptic densities via finite mixture sieves, cemmap working paper, No. CWP41/13, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2013.4113

This Version is available at:
https://hdl.handle.net/10419/97407

cemmap
centre for microdata methods and practice

Nonparametric estimation of multivariate elliptic densities via finite mixture sieves

Heather Battey
Oliver Linton

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP41/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Nonparametric estimation of multivariate elliptic densities via finite mixture sieves

Heather Battey
University of Bristol

Oliver Linton
University of Cambridge

August 21, 2013

Abstract

This paper considers the class of p-dimensional elliptic distributions ($p \geq 1$) satisfying the consistency property (Kano, 1994) and within this general framework presents a two-stage semiparametric estimator for the Lebesgue density based on Gaussian mixture sieves. Under the on-line *Exponentiated Gradient* (EG) algorithm of Helmbold et al. (1997) and without restricting the mixing measure to have compact support, the estimator produces estimates converging uniformly in probability to the true elliptic density at a rate that is independent of the dimension of the problem, hence circumventing the familiar curse of dimensionality inherent to many semiparametric estimators. The rate performance of our estimator depends on the tail behaviour of the underlying mixing density (and hence that of the data) rather than smoothness properties. In fact, our method achieves a rate of at least $O_p(n^{-1/4})$, provided only some positive moment exists. When further moments exists, the rate improves reaching $O_p(n^{-3/8})$ as the tails of the true density converge to those of a normal. Unlike the elliptic density estimator of Liebscher (2005), our sieve estimator always yields an estimate that is a valid density, and is also attractive from a practical perspective as it accepts data as a stream, thus significantly reducing computational and storage requirements. Monte Carlo experimentation indicates encouraging finite sample performance over a range of elliptic densities. The estimator is also implemented in a binary classification task using the well-known Wisconsin breast cancer dataset.

1 Introduction

Owing to generality considerations and breadth of application, density estimation is one of the most actively studied challenges in statistics. Although nonparametric density estimation was advanced dramatically by the introduction of the kernel density estimator (Fix and

Hodges, 1951), the performance of this estimator deteriorates rapidly for a fixed sample size as the number of dimensions grows large. Moreover, this performance depends heavily on the choice of b bandwidth parameters, where b grows quadratically with the dimension of the problem. This provides motivation for estimating nonparametrically within a restricted class of p-dimensional Lebesgue densities ($p \geq 1$): one that embeds many naturally arising distributions, allowing us to maintain a large degree of flexibility, whilst circumventing these problems that arise in high dimensions.

Much recent research has focussed on shape constrained density estimation. For instance, Cule et al. (2010) consider maximum likelihood estimation of a p-dimensional density that satisfies a log-concavity constraint, i.e. densities with tails decaying at least exponentially fast. This estimator involves no choice of smoothing parameter, and is able to estimate a range of symmetric and asymmetric densities consistently. Moreover, the estimator is shown to exhibit a certain degree of robustness to mispecification (Cule and Samworth, 2010).

This paper is concerned with nonparametric estimation within the class of elliptic densities in $\mathbb{R}^p$. This problem has been addressed in the literature before (Stute and Werner, 1991; Liebscher, 2005; Sancetta, 2009), but our work provides new contributions, which are highlighted below. Densities from the elliptic class are characterised by the property that their contours of equal density have the same elliptical shape as the Gaussian. Indeed, many of the convenient analytic properties possessed by the multivariate normal distribution (see e.g. Muirhead, 1982, Chapter 1) stem from the quadratic form in its characteristic function, which is actually a feature of the elliptic class more generally (Fang et al., 1990; Cambanis et al., 1981). Such features are, in part, responsible for the popularity of the elliptical symmetry assumption in applied work (see e.g. Kariya and Eaton (1977); Marsh (2007) for usage in the invariant testing literature, Owen and Rabinovitch (1983); Berk (1997) for usage in portfolio theory, and Chmielewski (1981) for a review of elliptical symmetry with applications).

More specifically, we consider a large subclass of elliptic distributions whose densities can be expressed as scale mixtures of normal densities, hence restrict attention only to distributions whose tails are heavier than those of a normal (Beale and Mallows, 1959). Members of this subclass are said to satisfy the *consistency property* (Kano, 1994) and are characterised by having all their d dimensional marginals $d < p$ from the same type of elliptic class as the p dimensional joint distribution. Unfortunately, the subclass excludes some well known members of the elliptic class such as the logistic, Pearson types II and VII, Kotz-type and Bessel distributions. It does however include (inter alia) the multivariate symmetric stable and Cauchy distributions, which arise as limit laws of normalised sums of i.i.d. random variables with fewer than 2 finite absolute moments, leading to their popularity as (multivariate) mutation distributions in evolutionary algorithms (see e.g. Rudolph, 1997; Arnold and Beyer, 2003), as well as the multivariate t, popular in finance. A key issue in some applications is heavy tails or non existence of moments and so in practice more

emphasis has been given to leptokurtic members of the elliptical class, which is aligned with our approach.

We propose a two-stage estimation procedure based on mixture likelihoods for the density of an elliptically distributed random vector with the consistency property. A major feature of this estimator is that it accepts data as a stream, which leads to a significant reduction in computational and storage requirements. In the elliptic framework without the consistency property, Stute and Werner (1991) proposed a density estimator that also circumvents the curse of dimensionality. Stute and Werner (1991) document the difficulty in estimating the density of an elliptic random variable due to the so called *volcano effect* that presents itself in a neighbourhood of the mean. Liebscher (2005) proposed a different estimator of the density that benefits from improved properties; he showed that his estimator achieved an optimal (one-dimensional) rate away from the mean for standard smoothness classes. As noted, his estimator does not completely overcome the problems arising near the mean. Furthermore, his main result requires the existence of at least four moments for the random variables of interest, which rules out many distributions of practical interest. Another problem is that the procedure relies on higher order kernels that must be chosen to satisfy a series of conditions, with the ultimate consequence that the resulting estimate can be negative and highly oscillatory at certain points of the support, an effect that is particularly prominent in small sample sizes (Marron and Wand, 1992). Our estimator, by contrast, always yields a valid density. Moreover, the implementation relies on delicate asymptotic analysis which requires knowledge of unknown quantities. Since the procedure does not support cross validation, it is hard in practice to implement the estimator described in Liebscher (2005). Further discussion of these problems appears in section 4.2. Although the construction of our estimator also relies on one unknown quantity, this may be computed, either by cross validation or by direct estimation. In fact, Monte Carlo evidence also suggests that the estimator is not unduly affected by an incorrect choice of this quantity. A key difference in the orientation of our approach is that we allow for heavy tailed data and our estimation procedure explicitly uses information or assumptions about tail behaviour rather than smoothness properties.

2 The model and its properties

Let $X_1, \ldots, X_n$ be i.i.d. random vectors in $\mathbb{R}^p$ having a density $f \in \mathcal{F}$. In this paper, we are only concerned with the case in which $\mathcal{F}$ is the set of elliptic densities with the consistency property (see below); it is of interest to study the implications for the estimator of f when the assumption of elliptical symmetry is violated, but this task is left for future work. A p dimensional random vector X is said to have an elliptic distribution with mean μ, scaling matrix Ω (positive definite), and Lebesgue measurable generator $g_p : \mathbb{R}^+ \mapsto \mathbb{R}^+$ (written

$X \sim El(\mu, \Omega, g_p))$ if its density function at x has the form

$$f(x) = c_p |\Omega|^{-1/2} g_p \left((x-\mu)^T \Omega^{-1} (x-\mu) \right).$$

The parameters μ, Ω and g_p uniquely determine the elliptic density up to a scaling factor: $El(\mu, \Omega, g_p) = El(\mu, c\Omega, g_{p,c})$ where $g_{p,c}(q) = g_p(q/c)$, which means we can always consider a Ω with diagonal elements all equal to one. This Ω is just the matrix of linear correlation coefficients in the case that the elements of X have finite variances, however, to subsume the more general cases, we will refer to Ω as the *orbital eccentricity matrix*. Provided Ω is full rank, X necessarily has the following stochastic representation (Cambanis et al., 1981, Theorem 1)

$$X \stackrel{d}{=} \mu + RAU^{(p)}; \qquad A = \Omega^{1/2}, \tag{2.1}$$

where $\stackrel{d}{=}$ means equality in distribution, $U^{(p)}$ is a random vector uniformly distributed on the unit sphere in $\mathbb{R}^p$, i.e. on $\{u \in \mathbb{R}^p : u^T u = 1\}$, A is the square root of Ω and R is a scalar random variable on $\mathbb{R}^+$, distributed independently of $U^{(p)}$. By the full rank condition, we may define $Z := A^{-1}(X - \mu) \stackrel{d}{=} RU^{(p)}$, which has a spherical distribution, hence is distributionally invariant under the orthogonal group (Muirhead, 1982, Definition 1.5.1). The density of Z is thus uniquely determined by the density of $Z^T Z \stackrel{d}{=} R^2$ according to $f(z) = c_p g_p(z^T z) \equiv c_p g_p(r^2)$, where this density exists if and only if R has density at r, $h_p(r)$, related to $g_p(r^2)$ as (Fang et al., 1990)

$$h_p(r) = \frac{2\pi^{p/2}}{\Gamma(p/2)} r^{p-1} g_p(r^2).$$

The subscript p on the generator indicates that, in general, g_p depends on the dimension p. When variables can be integrated out without changing the form of g, then the density generator is said to possess the *consistency property* (Kano, 1994); see Condition 3 below for a formal statement.

Condition 1. *Ω is full rank.*

Condition 2. *R possesses a density with respect to Lebesgue measure.*

Condition 3. *The density generator, $g(\cdot)$, possesses the consistency property, i.e.*

$$\int_{-\infty}^{\infty} g_p \left(\sum_{j=1}^{p} z_j^2 \right) dz_p = g_{p-1} \left(\sum_{j=1}^{p-1} z_j^2 \right) \tag{2.2}$$

for any $p \in \mathbb{N}$ and almost all $z \in \mathbb{R}^p$

To provide an example of when Condition 3 holds and when it does not, if the p dimensional joint distribution is, say, Gaussian, then all the $p-1$ dimensional marginals are also

Gaussian, yet if the p dimension joint distribution is power exponential with parameter α, then it is not true that the $p-1$ dimensional marginals are power exponential with parameter α for any $p \in \mathbb{N}$; further details on this latter example are provided in Kano (1994). Theorem 1 of Kano (1994) establishes equivalence between the consistency property and Z being a scale mixture of normal random vectors, which motivates our estimation procedure described below.

Assuming μ is known, we may centre as $Y := X - \mu$ and, by Condition 1 and the property $El(\mu, \Omega, g_p) = El(\mu, c\Omega, g_{p,c})$, consider,

$$Y \stackrel{d}{=} \Omega^{1/2} R U^{(p)}.$$

Then by the consistency property (Condition 3) and by Theorem 1 (iii) of Kano (1994), there exists a random variable $Q > 0$, unrelated to p such that for any $p \in \mathbb{N}$, $R \stackrel{d}{=} \sqrt{\chi_p^2/Q}$ with χ_p^2 a chi-square random variable with p degrees of freedom and Q, χ_p^2 and $U^{(p)}$ mutually independently distributed. It follows that $Y \stackrel{d}{=} N_p/\sqrt{Q}$ where N_p is a p dimensional normal random vector with mean zero and correlation matrix Ω and Q is unrelated to p and independent of N_p. By Condition 2, Y possesses a Lebesgue density, hence the probability density function $f(y)$ of Y at y may be written

$$f(y) = \int_0^\infty \phi(y|\Omega/q)\mathbb{P}(d(1/q)), \tag{2.3}$$

where $\phi(y|\Omega)$ is the normal kernel, given by $(2\pi)^{-p/2}|\Omega|^{-1/2}\exp\{-\frac{1}{2}(y^T\Omega^{-1}y)\}$ and $\mathbb{P}$ is the unknown law of the inverse mixing random variable $1/Q$ on $\mathcal{Q} = (0, \infty)$. The problem is now one of estimating Ω and the unknown law of $1/Q$, whose tails are assumed to satisfy Condition 4.

Condition 4. $\mathbb{P}(1/Q > x) \leq L(x)x^{-\alpha}; \quad \alpha > 0$, *where the slowly varying function* $L(x)$ *satisfies* $\lim_{x\to\infty} L(tx)/L(x) = 1$ *for all constant* $t > 0$.

Remark. Condition 4 allows us to estimate densities of random variables with heavy tails, with the convergence rate of our estimator depending on the parameter α. For instance, consider a random variable T_2 distributed as a student t with 2 degrees of freedom, then $Q \sim \chi_2^2/2$, where χ_2^2 is a chi-square random variable with 2 degrees of freedom. Since

$$\Pr\left(1/Q > x\right) = \Pr\left(2\chi_2^{-2} > x\right) = 2(1 - e^{-1/2x}), \tag{2.4}$$

T_2 satisfies Condition 4 with $\alpha = 1$.

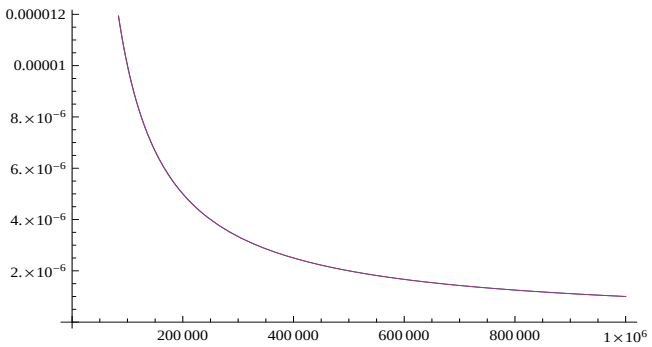

Figure 2.1: **Plot of** $\Pr(1/Q > x)$ **against** x **for** $1/Q \sim 2/\chi^2_2$**, (blue), together with** x^{-1} **against** x **(red) (lines lie on top of each other).**

Likewise, for a student t random variable with 1.5 degrees of freedom, the tail behaviour of $1/Q$ satisfies

$$\Pr\left(1/Q > x\right) = \Pr\left(1.5\chi^{-2}_{1.5} > x\right) = 1.22407(1.22542 - \Gamma(0.75, 0.5/x)) \approx x^{-0.75}, \qquad (2.5)$$

where $\Gamma(a, b)$ is the incomplete gamma function $\Gamma(a, b) = \int_b^\infty t^{a-1}e^{-t}dt$.

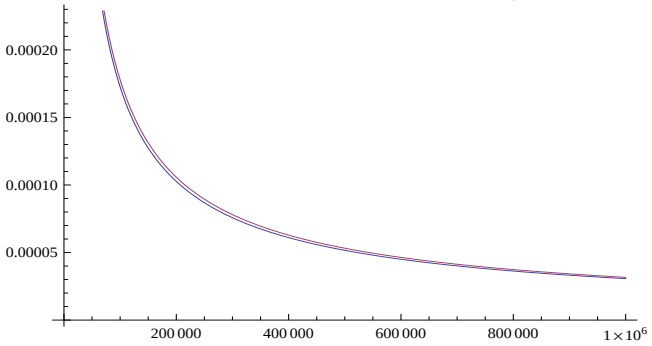

Figure 2.2: **Plot of** $\Pr(1/Q > x)$ **against** x **for** $1/Q \sim 1.5/\chi^{1.5}_2$**, (blue), together with** $x^{-0.75}$ **against** x **(red).**

3 Estimation via finite mixture sieves

Sancetta (2009) considers Bayesian semiparametric estimation of an elliptic density in a similar framework to that above, establishing weak conditions for posterior consistency. In this paper, we adopt a frequentist approach and define a sequence of approximations to equation (2.3)

$$\mathcal{S}_M = \left\{ f^M : f^M(y) := \sum_{s=1}^{M} \Lambda_s \phi(y|\Omega/q_s);\ M \in \mathbb{N} \right\}, \qquad (3.1)$$

where $\{\Lambda_s\}_{s=1}^M$ are weights such that $\Lambda_s \in [0, 1]\ \forall\ s \in \{1, \ldots, M\}$ and $\sum_{s=1}^M \Lambda_s = 1$, and the q_s are such that $1/q_s \in (0, \bar{x}(M)]$, where $\bar{x}(M) \to \infty$ as $M \to \infty$. Interest lies in finding

the optimal rate at which to allow M to grow with the sample size in order to achieve the optimal trade-off between approximation error and estimation error. Since we allow $\bar{x}$ to go to infinity, we permit one or more of the estimated component densities of the mixture to have infinite variance in the limit as the sample size goes to infinity.

3.1 Two-stage sieve estimation

Notice that estimating $f_n^M(y)$ requires estimation of $p(p-1)/2$ orbital eccentricity coefficients from $\Omega = [\Omega_{kl}]$. There is more than one way of doing this, and we shall make some specific suggestions below. We shall assume that whichever method is chosen obeys the following Condition.

Condition 5. *The sequence* $(\widehat{\Omega}_n)_{n\in\mathbb{N}}$ *satisfies* $|\widehat{\Omega}_n - \Omega| = O_p(n^{-1/2})$.

The above condition admits estimation of Ω by the Stahel-Donoho robust estimator of multivariate location and scatter, as discussed in Maronna and Yohai (1995). This estimator also takes care of estimation of μ, which is a requirement for re-centering at zero the X in equation (2.1) in the case that μ is unknown. As an alternative, the vector of univariate sample means or sample medians could be used to estimate μ and a different estimator for Ω may be used.

We advocate the following estimator based on a transformation of $p(p-1)/2$ estimators of the Kendall tau dependence measure. The following canonical transformation is valid for all members of the elliptic class (Lindskog et al., 2003):

$$\tau_{kl} = \frac{2}{\pi} \arcsin(\rho_{kl}), \tag{3.2}$$

where τ_{kl} is the Kendall tau dependence measure, defined as

$$\tau_{kl} := \Pr\left((Y_{i,k} - Y_{j,k})(Y_{i,l} - Y_{j,l}) > 0\right) - \Pr\left((Y_{i,k} - Y_{j,k})(Y_{i,l} - Y_{j,l}) < 0\right),$$

with $(Y_{j,k}, Y_{j,l})$ an independent copy of $(Y_{i,k}, Y_{i,l})$. In the case that the elements of Y have finite variances and $\Omega = [\Omega_{kl}] = [Cov(Y_k, Y_l)/\,Var(Y_k)\,Var(Y_l)]$, the standard estimator of the orbital eccentricity is the Pearson product moment correlation coefficient, derived as the maximum likelihood solution in the case that Y is multivariate normally distributed. This estimator is not robust, in the sense that it is not unaffected by departures of Y from normality. In particular, the Pearson estimator performs very poorly when Y is from a distribution with heavy tails (Pollard, 2000); by contrast, the sample version of Kendall's tau $\widehat{\tau}_{kl}$ is robust to heavy tailedness. $\widehat{\tau}_{kl}$ is the proportion of concordant pairs minus the

proportion of discordant pairs, i.e.

$$\begin{aligned}\widehat{\tau}_{k,l} &= \sum_{i=1}^{n}\sum_{j=i}^{n-1}\frac{\mathbb{1}\{(Y_{i,k}-Y_{j,k})(Y_{i,l}-Y_{j,l})>0\}-\mathbb{1}\{(Y_{i,k}-Y_{j,k})(Y_{i,l}-Y_{j,l})<0\}}{\mathbb{1}\{(Y_{i,k}-Y_{j,k})(Y_{i,l}-Y_{j,l})>0\}+\mathbb{1}\{(Y_{i,k}-Y_{j,k})(Y_{i,l}-Y_{j,l})<0\}} \\ &= \binom{n}{2}^{-1}\sum_{i,j<i}\left(\mathbb{1}\{(Y_{i,k}-Y_{j,k})(Y_{i,l}-Y_{j,l})>0\}-\mathbb{1}\{(Y_{i,k}-Y_{j,k})(Y_{i,l}-Y_{j,l})<0\}\right),\end{aligned}$$

which, in light of equation (3.2) provides a natural means of estimating the orbital eccentricity efficiently:

$$\widehat{\rho}_{kl} = sin\left(\frac{\pi}{2}\widehat{\tau}_{kl}\right). \tag{3.3}$$

Zhao et al. (1997) show that each of the $(p-1)p/2$ Kendall tau estimators satisfy $|\widehat{\tau}_{kl}-\tau_{kl}| = O_{a.s.}(n^{-1/2})$. Since $\sin(\frac{\pi}{2}\cdot)$ is a continuous mapping, Slutsky's theorem implies $\widehat{\Omega}_{kl} = \Omega_{kl} + O_p(n^{-1/2})\ \forall k \neq l$, hence Condition 5 is satisfied by the Kendall tau transform estimator.

We next turn to estimation of the weights. Equation (3.1), being a finite scale-mixture of normal densities, yields an incomplete data problem as it is not observed from which component density each of the data points are drawn. Since the true maximum likelihood solution is infeasible, we must rely on iterative methods that compute a sequence of mixture vectors $\Lambda_1, \ldots, \Lambda_j, \ldots$ for which some composite mixture $\sum_{s=1}^{M}\widehat{\Lambda}_s q_s \phi(y|\Omega/q_s)$ converges to the (unknown) maximum likelihood mixture $\sum_{s=1}^{M}\widehat{\Lambda}_s^{ML} q_s \phi(y|\Omega/q_s)$, where Λ_s^{ML} is the infeasible maximum likelihood solution and Λ_s is some composite of the mixture weights computed in an iterative algorithm, which is yet to be specified. Note that for typographical reasons, we make no notational distinction between scalar weights and vector weights; the distinction will be clear from the context and the indexing. The celebrated *Expectation Maximisation (EM) algorithm*, is one such iterative procedure; we refer to the seminal paper by Dempster et al. (1977) for details. Despite its theoretical ability to overcome the latent factor problem, the EM algorithm can converge quite slowly (Meng and van Dyk, 1997) and can be impractical, especially when processing large data sets and data streams (Cappé and Moulines, 2009). Concerns in the latter case arise from the fact that all the data must be made available at each iteration, implying large storage requirements and computational inefficiency. On-line variants allow previous estimates to be updated based on each new piece of information and are therefore attractive from a computational perspective.

We study the asymptotic properties of the mixture sieve estimator using the *Exponentiated Gradient algorithm* (henceforth termed *EG algorithm*) of Helmbold et al. (1997). This is a grid-based on-line algorithm for the unsupervised mixture proportions estimation problem, providing significant computational speed-ups over conventional EM. The EG algorithm considers a finite number, M, of mixture components, corresponding to fixed values of $(q_s)_{s=1}^{M}$ in the support of $\mathbb{P}$. This is in contrast to continuous support EM-type algorithms in which the conditional expectation of the log-likelihood is maximised with respect to the scale parameters, $(q_s)_{s=1}^{M}$, in addition to the mixing parameters, $(\Lambda_s)_{s=1}^{M}$. Clearly, such a grid-based

approach will have the disadvantage that the $(q_s)_{s=1}^M$ must be chosen a priori; we make the choice that minimises the approximation error, i.e. we choose these values to be separated by equal intervals for all $s \in 1, \ldots, M$. The case where equal intervals are chosen is also covered in Helmbold et al. (1997) and thus an explicit bound is available on the estimation error from making this choice.

Write $L_N(\Lambda)$ for the log mixture likelihood constructed from the first N sample points and evaluated at Λ, i.e.

$$L_N(\Lambda) = \frac{1}{N} \sum_{i=1}^{N} \ln \left(\sum_{s=1}^{M} \Lambda_s \phi(Y_i|\Omega/q_s) \right).$$

At the j^{th} iteration, the EG algorithm with *learning rate* η (termed EG_η algorithm) seeks to maximise with respect to Λ_j

$$F(\Lambda_j) = \eta \left(L_j(\Lambda_{j-1}) - \sum_s \nabla_s L_j(\Lambda_{j-1})(\Lambda_{j,s} - \Lambda_{j-1,s}) \right) - d(\Lambda_j, \Lambda_{j-1}) \tag{3.4}$$

subject to the constraint $\sum_{s=1}^M \Lambda_{j,s} = 1$, where $d(\Lambda_j, \Lambda_{j-1})$ is the Kullback-Leibler divergence between the two probability distributions Λ_j and Λ_{j-1}. Heuristically, the EG_η algorithm is easy to understand as a Taylor series approximation to $L_j(\Lambda_j)$ with a penalty to reflect the quality of the approximation. Maximising the Lagrangian corresponding to equation (3.4), amounts to solving the $M+1$ equations for the M elements of Λ_j:

$$\frac{F(\Lambda_j, \gamma)}{\partial \Lambda_{j,s}} = \eta \nabla_s L_j(\Lambda_{j-1}) - \frac{\partial d(\Lambda_j, \Lambda_{j-1})}{\partial \Lambda_j} + \gamma = 0$$

and

$$\sum_{s=1}^{M} \Lambda_{j,s} = 1.$$

Replacing $d(\Lambda_j, \Lambda_{j-1})$ with the Kullback-Leibler divergence yields the requirement

$$\eta \nabla_s L_j(\Lambda_{j-1}) - \left(\ln \left(\frac{\Lambda_{j,s}}{\Lambda_{j-1,s}} \right) + 1 \right) + \gamma = 0. \tag{3.5}$$

Solving equation (3.5) for the $\Lambda_{j,s}$ and imposing the normalisation constraint yields the analytic update

$$\Lambda_{j,s} = \frac{\Lambda_{j-1,s} \exp\{\eta \nabla_s L_j(\Lambda_{j-1})\}}{\sum_{t=1}^{M} \Lambda_{j-1,t} \exp\{\eta \nabla_t L_j(\Lambda_{j-1})\}},$$

which, given an initialisation vector Λ_0, gives a recursive estimate of the mixing weights $(\Lambda_s)_{s=1}^M$; we denote the EG estimates by $(\widehat{\Lambda}_s^{EG})_{s=1}^M$. At each iteration, the EG solution $\widehat{\Lambda}^{EG}$ is charged a loss,

$$-L_n(\widehat{\Lambda}^{EG}) = -\frac{1}{n} \sum_{i=1}^{n} \ln \left(\sum_{s=1}^{M} \widehat{\Lambda}_s^{EG} \psi(Y_i|q_s) \right).$$

The true (unknown) maximum likelihood solution, $\widehat{\Lambda}^{ML}$ (the one that would maximise the log likelihood if the labels were observed) suffers an analogous loss. It is instructive to consider the average additional loss incurred over $I = n$ iterations of the EG algorithm over that incurred by the unknown maximum likelihood solution, which is obtained from equation (9) of Helmbold et al. (1997) as

$$-\frac{1}{I}\sum_{j=1}^{I}\frac{1}{n}\sum_{i=1}^{n}\ln\left(\sum_{s=1}^{M}\widehat{\Lambda}_{j,s}^{EG}\psi(Y_i|q_s)\right) \leq -\frac{1}{n}\sum_{i=1}^{n}\ln\left(\sum_{s=1}^{M}\widehat{\Lambda}_{s}^{ML}\psi(Y_i|q_s)\right) + \frac{2\ln M}{I\eta} \tag{3.6}$$

where $\psi(y|q_s)$ is the s^{th} mixture density at y; here

$$\psi(y|q_s) := \phi(y|\Omega/q_s). \tag{3.7}$$

Define the density estimator

$$\widehat{f}_n^M(y) = \sum_{s=1}^{M}\widehat{\Lambda}_{I,s}^{EG}\phi(y|\widehat{\Omega}_{,n}/q_s), \tag{3.8}$$

where $(\widehat{\Lambda}_{I,s})_{s=1}^{M}$ is the average set of weights estimated in $I = n$ iterations of the EG algorithm with learning rate η. Note that (3.6) does not imply convergence of $\widehat{\Lambda}^{EG}$ to $\widehat{\Lambda}^{ML}$ at rate $(2\ln M)/I\eta$, but rather convergence in log likelihood of $\widehat{\Lambda}_{I,s}^{EG}$ to $\widehat{\Lambda}^{ML}$.

4 Asymptotic properties and practicalities

Using the EG_η algorithm to estimate the mixture proportions in equation (3.1), and replacing Ω with $\widehat{\Omega}$ in (3.7) gives rise to the bound in the following theorem.

Theorem 1. *Let Conditions 1 - 4 hold. Fix* $\{q_s = 1/x_s : s = 1,\ldots,M\}$ *where* $x_1 = M^{-\alpha/(1+\alpha)}/2$ *and* $x_s = x_{s-1} + M^{-\alpha/(1+\alpha)}$ *and let* $M := M(n)$ *be given by* $M = \left\lfloor cn^{\frac{1+\alpha}{2+4\alpha}}\right\rfloor$ *(where c is a positive constant and* $\lfloor x\rfloor$ *gives the largest integer less than or equal to x). Suppose that* $\widehat{f}_n^M(z)$ *is defined by (3.8) with the learning rate* $\eta = 2r\sqrt{\frac{2\ln M}{I}}$ *where r is a lower bound on the instances* $\psi(Y_i|q_s)$ *for all i and all s. Then,*

$$\sup_{y\in D}|f(y) - \widehat{f}_n^M(y)| = O_p\left(n^{-\frac{1+3\alpha}{4(1+2\alpha)}}\right) \tag{4.1}$$

for any compact subset D of $\mathbb{R}^p$ *such that* $0 \notin D$. *Allowing r to go to zero at rate* $\frac{\sqrt{2\ln M}}{2M\sqrt{n}}$ *delivers the same rate in equation (4.1).*

Although constants depend on p because of the estimation of $p(p-1)/2$ eccentricity parameters in Ω, for a fixed p, the rate of convergence of $\widehat{f}_n^M(y)$ does not depend on p; this is qualitatively the same result as that obtained in Liebscher (2005).

Remark. In a neighbourhood of $y = 0$, the proof strategy does not allow us to say anything about the convergence rate. In fact, the density f can be infinite and nondifferentiable at this point under our conditions. Under the strong additional condition that the mixing measure $\mathbb{P}$ is supported on $[\delta, \infty)$ where $\delta > 0$, the rate in equation (4.1) becomes uniform over the whole of $\mathbb{R}^p$, as the space of functions,

$$\mathcal{G} := \{g : q \mapsto \phi\left(y|\Omega/q\right);\ 1/q \in \mathcal{Q},\ y \in \mathbb{R}^p\},$$

is bounded for $\mathcal{Q} = [\delta, \infty)$. We note that in Ghosal and van der Vaart (2001), which studies univariate normal mixture density estimation, the assumption of compactly supported mixing measure is assumed in order to obtain rates that hold uniformly over $(-\infty, \infty)$. In this case rates of almost $n^{-1/2}$ are available. The case of noncompact support is considered by Genovese and Wasserman (2000) who establish rates of at most $n^{-1/4}$ when the tails of the mixing density are described by our Condition 4 (see section 5.2 of Genovese and Wasserman (2000) and the discussion surrounding equation (36) op. cit.).

Clearly, the rate of convergence of the sieve estimator depends rather significantly on the tail exponent α of the mixing measure. When the law of the mixing measure satisfies the least restrictive tail condition (smallest tail exponent), a bound of $O_p\left(n^{-1/4}\right)$ may be achieved. In the limit however, as the law of $1/Q$ satisfies a stronger tail condition, a rate of $O_p\left(n^{-3/8}\right)$ is achievable. This rate corresponds to the case in which Y is normally distributed in $\mathbb{R}^p$ (see the discussion in section 4.1 below).

4.1 A practical method for selection of the tuning parameter

Since the smoothing parameter required to attain the above rate depends on α, in order to reduce the potential inefficiency induced by unnecessarily taking α too small in condition 4, it would be preferable that the choice of α be data-driven. One of the advantages of our approach over that of Liebscher (2005) is that it does admit a cross-validatory choice of α. However, since cross validation is such a computationally intensive procedure, we also consider the possibility of direct estimation of α from realisations of $Y^T\Omega^{-1}Y \stackrel{d}{=} R^2$. Since the law of the inverse mixing random variable $1/Q$ (as determined by α) is related to the law of the random variable R^2, the tail behaviour of the two is linked in a way that is quantified in equation (2.5) in the case that Y is a student t random variable with $d = 2$ degrees of freedom. The tail behavior of $1/Q$ can be described in a similar way for any d.

Considering the case in which Y is distributed as a student t random variable with $d = \{2, 5, 10\}$ degrees of freedom, we display the probability density functions of $1/dQ$, which has an inverse χ^2 distribution with d degrees of freedom, and the corresponding density functions of R^2/p, which has an $F_{p,d}$ distribution (see e.g. Muirhead, 1982, section 15 and exercise 1.30).

As $d \to \infty$ the inverse χ^2_d density collapses to a spike at zero which corresponds to the density of $1/Q = d\chi_d^{-2}$ collapsing to a spike at 1, corresponding to normality of the random vector Y.

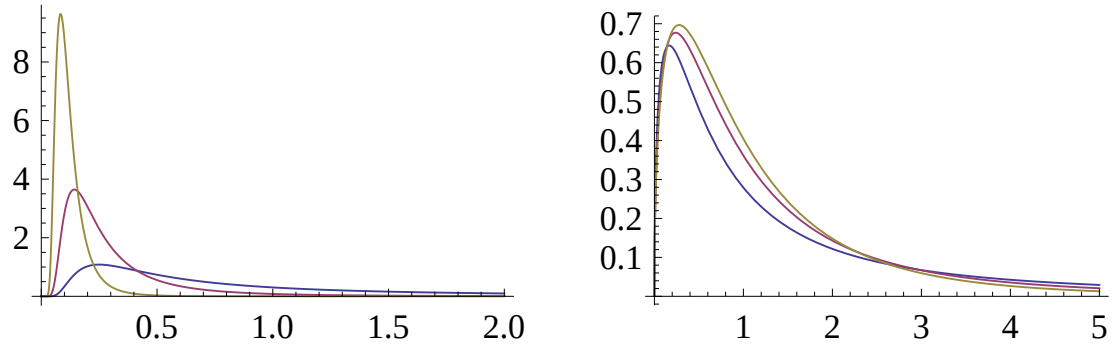

Figure 4.1: **Density function of** $1/dQ \overset{d}{=} \chi_d^{-2}$ **and the corresponding density function of** $R^2/p = (1/p)Y^T\Omega^{-1}Y \overset{d}{=} F_{p,d}$ **when** Y **is distributed as a student-t random variable with dimension** $p = 3$ **and** d **degrees of freedom (**$d = \{2, 5, 10\}$ **with corresponding curve** $\{$**blue**, **red**, **yellow**$\}$**).**

The discussion above motivates the use of a Hill estimator (Hill, 1975) of the tail exponent of R^2, whose realisations come from the same distribution as realisations of $Y^T\Omega^{-1}Y$, in order to gain insight into the tail behaviour of $1/Q$, whose realisations are inaccessible. Let $\widehat{R}^2$ be defined as the random variable obtained from the transformation $Y^T\widehat{\Omega}^{-1}Y$, where $\widehat{\Omega}$ is the estimator of the eccentricity matrix obtained from the canonical Kendall tau transform in equation (3.3). Let $\widehat{R}^2_{(k)}$ denote the k^{th} order statistic of this random variable $\widehat{R}^2$. Then a Hill estimator for the tail exponent of R^2 (based on a subsample of the K largest observations $\widehat{R}^2$) may be defined as

$$\widehat{\alpha}_{K,n}(R^2) = \frac{1}{K}\sum_{k=1}^{K}(\ln \widehat{R}^2_{(n-k)} - \ln \widehat{R}^2_{(n-K)}).$$

Based on 100 simulations and $n = 100000$ draws from the distribution of $1/Q$ and R^2 when Y is distributed as a trivariate student t random variable with d degrees of freedom, we find that $\widehat{\alpha}_{K,n}(R^2)$ exhibits a strong relationship with $\widehat{\alpha}_{K,n}(1/Q)$ (where we take $K = 0.05n = 5000$) (see Figure 4.2 below), the consistent, yet practically infeasible estimator of the tail exponent of the law of $1/Q$. Figure 4.2 depicts the average (over 100 simulations for each value of d) Hill estimate for the tail exponent of the inaccessible $1/Q$. This is plotted against the average Hill estimate for the tail exponent of R^2, which can be estimated easily and reliably from the data. The right panel of Figure 2 plots the average of $\widehat{\alpha}_{K,n}(1/Q)$ and the average of the exponential transformations of $\widehat{\alpha}_{K,n}(R^2)$, where this transformation is chosen with a view to making the relationship linear. Empirically the relationship seems fairly robust to changes in the dimension p (see Figure 4.4).

The right panels of Figures 4.2 and 4.3 both indicate that an estimator of the form

$$\widehat{\alpha}(1/Q) := c_1 + c_2 \exp\{\widehat{\alpha}_{K,n}(\widehat{R}^2)\} \tag{4.2}$$

may be reasonable for the tail exponent of $1/Q$. However, the constants c_1 and c_2 are not invariant to the dimension of the problem. We plot in Figure 4.4 the ordinary least squares estimates of the coefficients as a function of the dimension p, along with the coeffient of determination. We do not plot confidence bands for the coefficients c_1 and c_2 as the upper and lower confidence limits are indistinguishable from the estimated regression coefficients; it is worth noting that zero is not included in any of the confidence bands.

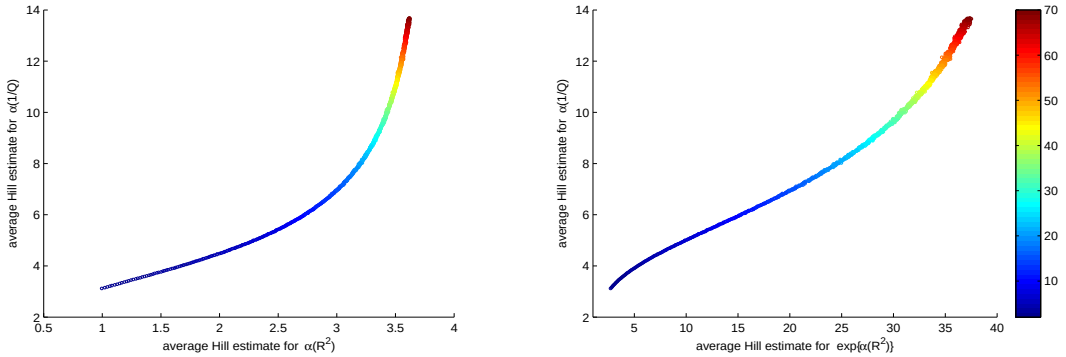

Figure 4.2: **Scatter plots of the average (over 100 simulations) of $\widehat{\alpha}_{K,n}(1/Q)$ against the average of $\widehat{\alpha}_{K,n}(R^2)$ (left) and the average of $\widehat{\alpha}_{K,n}(1/Q)$ against the average of $\exp\{\widehat{\alpha}_{K,n}(R^2)\}$ when Y is distributed as a trivariate student t random variable with d degrees of freedom. The colour indicates increasing d (from 2 to 70 in steps of 0.05).**

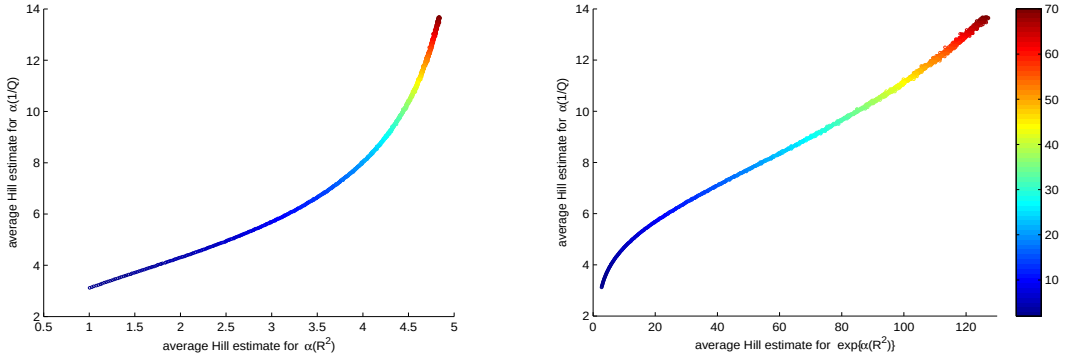

Figure 4.3: **Scatter plots of the average (over 100 simulations) of $\widehat{\alpha}_{K,n}(1/Q)$ against the average of $\widehat{\alpha}_{K,n}(R^2)$ (left) and the average of $\widehat{\alpha}_{K,n}(1/Q)$ against the average of $\exp\{\widehat{\alpha}_{K,n}(R^2)\}$ when Y is distributed as a student t random variable in $p = 5$ dimensions with d degrees of freedom. The colour indicates increasing d (from 2 to 70 in steps of 0.05).**

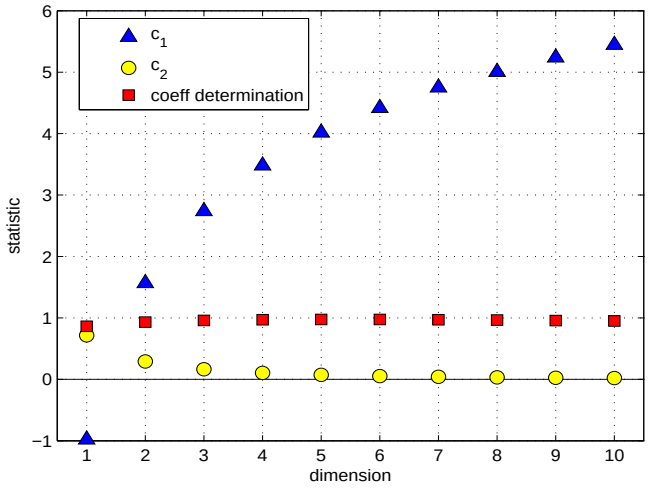

Figure 4.4: **Estimated regression coefficients c_1 and c_2, and the coefficient of determination from a regression of $\widehat{\alpha}(1/Q)$ on $\exp\{\widehat{\alpha}(R^2)\}$ within the class of p-dimensional t distributions. Statistics are plotted as a function of the dimension p.**

We emphasise that this is simply a rule of thumb method designed around a particular class of distributions; although it will not always work, it is a simple and reasonable approach for choosing the unknown parameter $\alpha(1/Q)$. More generally we can define a cross-validation procedure that would work for any distribution.

Although we do not dedicate further discussion to the properties of $\widehat{f}_n^M$ based on an estimator of $\alpha(1/Q)$, to illustrate that the rule of thumb approach of equation (4.2) provides reasonable results even outside the class of multivariate t distributions, we present in section 5 the MISE minimising choice of α and the corresponding ISE estimates when the data are drawn from a symmetric stable distribution with various tail parameters. We compare the resulting ISEs to those obtained from using the choice of α based on equation (4.2) with the constants c_1 and c_2 of Figure 4.4.

4.2 Discussion of results

We have already discussed our results in relation to those of Genovese and Wasserman (2000) for univariate normal mixtures. We now compare them with those of Liebscher (2005) who considered estimation of elliptically symmetric densities under smoothness conditions. He obtained the bound

$$O_p\left((n/ln(n))^{-\frac{k}{2k+1}}\right) \tag{4.3}$$

for his kernel-based nonparametric estimator (2.5). In the above display, k is the maximum order of derivative of the density generator, $g(\cdot)$, for which the derivative exists and is bounded on $\mathbb{R}_+$; it is assumed to be an even integer greater than or equal to 2. Note that

the rate in (4.3) is – up to a logarithmic factor – optimal in the sense of Stone (1980) for densities whose effective dimension is one. However, the bound is only valid uniformly over compact subsets of $\mathbb{R}^p$ that exclude the mean, rather than over the whole of $\mathbb{R}^p$ (to which Stone's minimax rate applies).

The validity of the rate of Liebscher (2005) is dependent on several conditions, which rule out cases that we are able to accommodate with our method. The discussion surrounding Figure 4.1. allows us to relate the condition of Liebscher (2005) that $Y^T\Omega^{-1}Y \stackrel{d}{=} R^2$ has finite $4+\epsilon$ moment, to the tails of the density of Y. Given that the $F_{p,d}$ distribution only has finite moments of order $d/2$, the assumption of finite $4+\epsilon$ moment rules out the case in which $Y^T\Omega^{-1}Y$ is distributed as an $F_{p,d}$ random variable with $d \leq 8$, which corresponds to Y having student t distribution with $d \leq 8$ degrees of freedom.

Our procedure has several practical advantages over that of Liebscher (2005). Aside from the practical advantages of online estimation over batch estimation, the approach of Liebscher (2005) requires that a suitably chosen transformation (so chosen to alleviate problems that would otherwise lead to significant bias in a neighbourhood of the mean) be applied to the data. This transformation is delicate and influences the order and choice of higher-order kernel (see Conditions $\mathcal{T}(p)$ and $\mathcal{K}(p)$ of the paper). The suggested transformations involve a tuning parameter that, due to the nature of the procedure, cannot be chosen by cross validation. This is in contrast to our approach, which does support a cross validatory choice of α or potentially even direct estimation of α along the lines discussed in section 4.1 above. Furthermore, Monte Carlo evidence illustrates the encouraging finite sample performance of our procedure, even when α is not chosen as the true one.

A final but important point to note is that the estimator of Liebscher (2005) requires a kernel of order k in order to achieve the k-dependent rate of display (4.3). In order to have high-order even moments equal to zero (his condition $\mathcal{K}(p)$), higher-order kernels need to be negative and highly oscillatory, which means the resulting estimate is not a proper density. This effect is gradually dampened as $n \to \infty$, but will be prominent in small and moderate sample sizes. Although such a symptom may be unproblematic if, for instance, we are interested in using the estimator for nonparametric regression, it is an undesirable feature in many other cases of interest. Our estimator has the advantage that it always yields an estimate that is a proper density.

5 Simulation studies

5.1 Analysis of the rule of thumb procedure

In this section, we analyse the extent to which the rule of thumb procedure of section 4.1, which was built around a particular family of distributions, may be suitable for other members of the elliptic class. We consider univariate symmetric stable distributions, whose

characteristic function is parameterised by a tail parameter $\gamma \in (0, 2]$ in addition to the location and scale parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$. Since the stable distribution does not possess a closed form solution for its density function, a symmetric stable random variable $Y \sim S(\gamma, 0, \sigma, \mu)$ is described by its characteristic function as

$$\mathbb{E}\exp\{iuY\} = \begin{cases} \exp\{-\sigma^\gamma |u|^\gamma + i\mu u\} & \gamma \neq 1 \\ \exp\{-\sigma^\gamma |u| + i\mu u\} & \gamma = 1 \, . \end{cases}$$

The density function of a stable random variable can be calculated numerically by the algorithm described by Nolan (1997) in the univariate case and using the method of Abdul-Hamid and Nolan (1998) in the multivariate case. Due to the lack of freely available software that support this distribution, we consider only the univariate case here, for which user-written software is publically available (Veillette, 2012).

For values of the tail parameter γ increasing from 0.3 to 2 in steps of size 0.1, we conduct a simulation study in which $n = 500$ symmetric stable random variables $Y \sim S(\gamma, 0, 1, 0)$ on each of 100 Monte Carlo replications. For values of the tuning parameter $\alpha \in \mathcal{A}$, where $\mathcal{A}$ is a finite set of elements increasing from 0.25 to 3 in steps of size 0.25, we construct an estimate of the density function, we also construct an estimate of the density function based on the Hill estimate $\widehat{\alpha}(\widehat{R}^2)$ and the rule of thumb procedure described in section 4.1. We plot in the left panel of Figure 5.1 below the symmetric stable density function for various values of γ. We also plot in the right panel of Figure 5.1 the estimates of $\alpha(1/Q)$ and the values of the tuning parameter in $\mathcal{A}$ that minimise the empirical MISE over the 100 Monte Carlo replications. In Figure 5.2 we plot the natural logarithm of the integrated square error in 100 simulations for each value of the tail parameter γ; the left panel corresponds to the MISE minimising choice of tuning parameter, and the right panel corresponds to the Hill estimate.

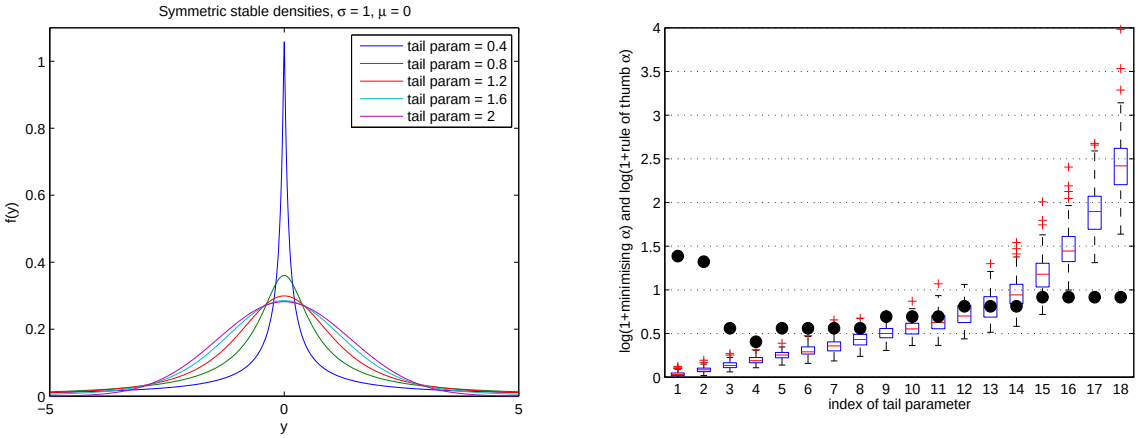

Figure 5.1: **Left panel: univariate symmetric stable density functions for different tail parameters γ with μ and σ fixed at 0 and 1 respectively; right panel: boxplots of the natural logarithm of $(1 + \alpha)$ with α chosen by the rule of thumb described in equation**

(4.2), **with c_1 and c_2 the least squares estimates of Figure 4.4. The black dots represent the α chosen by minimising the MISE over the 100 Monte Carlo replications.**

The right panel of Figure 5.1 illustrates that for γ in the range (0.6,1.6), the rule of thumb based on equation (4.2) performs reasonably well in the sense that the estimated $\widehat{\alpha}(1/Q)$ is close to the α that produces the lowest MISE over the 100 Monte Carlo replications. For large and small values of γ, the rule of thumb is less reliable. This is not surprising in light of the right panels of Figures 4.2 and 4.3, where the linearity of the relationship breaks down for very heavy and very light tails.

Viewing the right panel for Figure 5.1 together with Figure 5.2, we see that taking α too small is not detrimental to the performance of the estimator, whilst taking α too large results in too few mixtures being used to approximate the density of interest and causes the performance to deteriorate substantially.

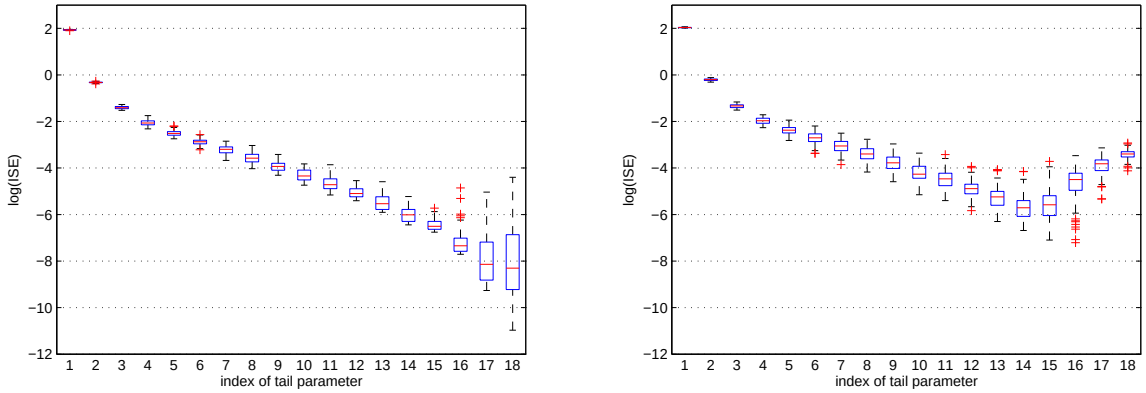

Figure 5.2: **Right panel: boxplots of the ISE over 100 simulations when the empirical MISE minimising α is chosen; boxplots of the ISE over 100 simulations when α is chosen by the rule of thumb described in equation (4.2), with c_1 and c_2 the least squares estimates of Figure .4**

5.2 Finite sample performance

We consider the relative performance of our sieve estimator in samples of size 100, 500, 1000 and 2000 when the data come from the bivariate and trivariate t distributions with 1.5, 2, 3 and 4 degrees of freedom and with correlation coefficients $\rho = 0.4$ in the $p = 2$ case and $\rho_{12} = 0.4$, $\rho_{13} = -0.7$, $\rho_{23} = 0.1$ in the $p = 3$ case.

Figures 5.3 and 5.4 depict the integrated squared errors (ISE) based on 1000 Monte Carlo replications for the kernel density estimator, the Liebscher (2005) estimator, and for several versions of the Kendall tau-transform Gaussian mixture sieve estimator, where the latter is computed using both the EG and EM algorithms over a fixed set of mixing points $\{1/q_s, s = 1, \ldots, M\}$. We examine the performance of the EG and EM mixture sieve estimators using

both the rule of thumb procedure and the unobservable theoretical value of α, with which the estimator is optimised. Letting α_d satisfy $\Pr(d\chi_d^{-2} > x) = x^{-\alpha_d}$, which is the relevant quantity $\Pr(1/Q > x)$ from Condition 4 for the case in which Y is a t distribution with d degrees of freedom, we have $\alpha_{1.5} \approx 0.75$, $\alpha_2 \approx 1$, $\alpha_3 \approx 1.52$ and $\alpha_4 \approx 2.05$. Since this quantity is unknown in practice, we consider the performance of the EG and EM variants of the mixture sieve estimator when these values of α are estimated using the rule of thumb procedure discussed in section 4.1. Letting α denote the true value of α (i.e. dropping the d subscript), and letting $\widehat{\alpha}$ denote the estimated value of α based on the rule of thumb procedure, we take respectively, $\epsilon = M^{-\alpha/1+\alpha}$, with $M = \left\lfloor n^{\frac{1+\alpha}{2+4\alpha}} \right\rfloor$ and $\epsilon = M^{-\widehat{\alpha}/1+\widehat{\alpha}}$, with $M = \left\lfloor n^{\frac{1+\widehat{\alpha}}{2+4\widehat{\alpha}}} \right\rfloor$, i.e. the optimal number of mixtures from Theorem 1 with $c = 1$. Letting $\{x_s, s = 1, \ldots, M\}$ be the values of $\{1/q_s, s = 1, \ldots, M\}$, we consider uniform spacings $x_j - x_{j-1}$ starting at $x_1 = \delta + \epsilon/2$, with δ some small constant that we take equal to 0.25, and $x_s = x_{s-1} + \epsilon \; \forall \; s\{1, \ldots, M\}$. We take initial weights $\{\Lambda_{0,s} : s = 1, \ldots, M\}$ all equal to $1/M$ and we set $\eta = 1.5$. Where applicable, we use the same choice of tuning parameters for the finite mixture sieve with EM update.

For the kernel density estimator we use an empirical bandwidth selector based on least squares cross validation (LSCV) (Wand and Jones, 1995, section 4.7). The empirical bandwidths and kernel density estimates were computed using the `ks` package (Duong, 2007) in `R`. Although we do not discuss this further, it is worth pointing out that an adaptive bandwidth kernel density estimator such as that of Abramson (1982) is likely to perform better on heavy tailed data such as these; we also refer to Sain and Scott (1996) for univariate adaptive kernel methods, and to Sain (2002) and Scott and Sain (2005) for multivariate adaptive techniques.

For the Liebscher estimator, we employ the transformation suggested in the example preceding Theorem 1 of Liebscher (2005). The transformation depends on a constant, a, which is not discussed in the paper; we take this constant equal to 1, but the estimator appears to be rather insensitive to this choice. The kernel estimator and sieve estimator can all produce estimates in dimensions higher than $p = 3$, using respectively the `ks` package and the code we wrote to compute the sieve estimates in the $p = 2$ and $p = 3$ cases (available upon request). The only modifications required are the fairly simple ones required to produce the $p > 3$ dimensional grid-array of evaluation points.

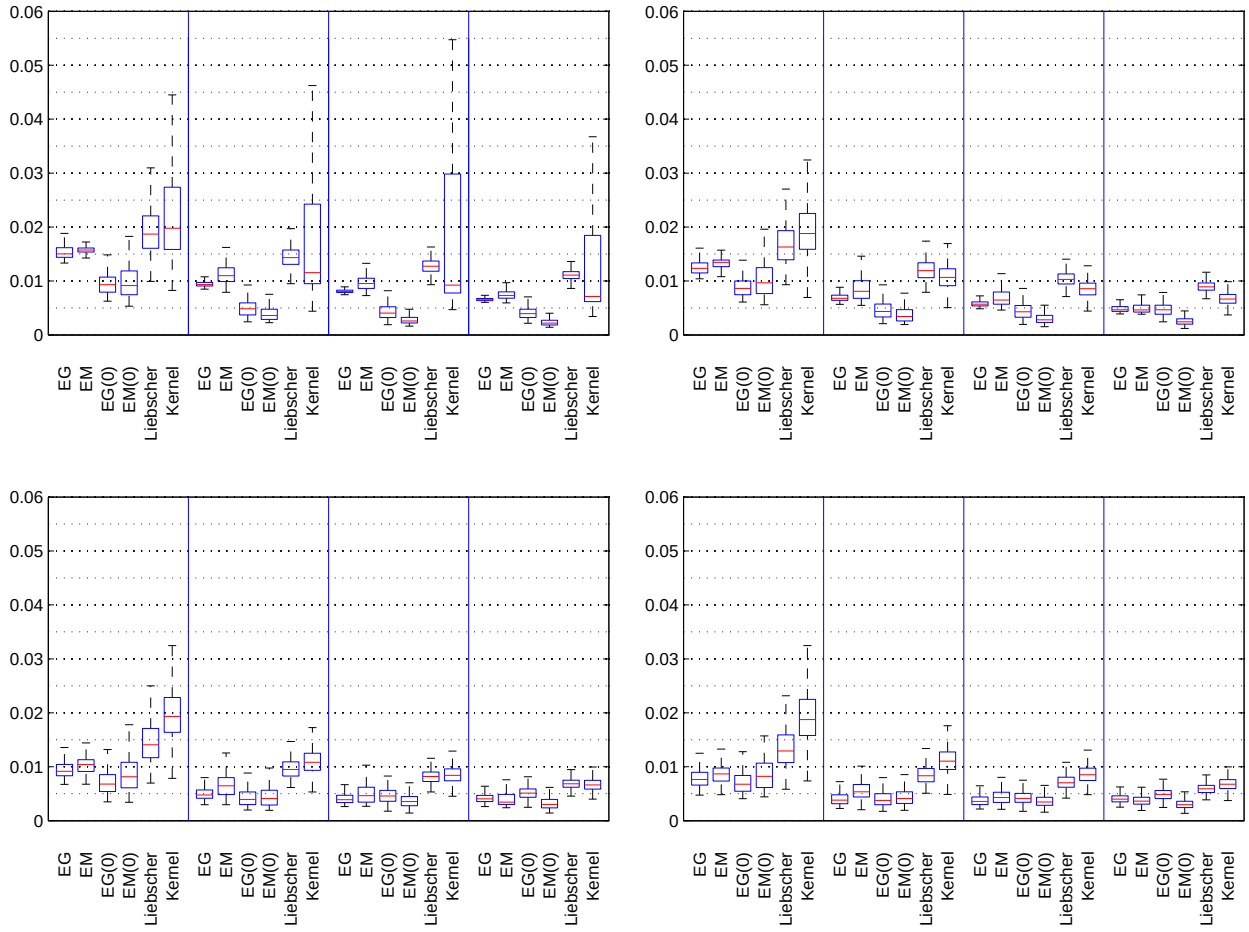

Figure 5.3: **Boxplots of the square root of the integrated squared error (divided by the number of points in the grid-array of evaluation points) based on 1000 MC replications using n observations from a bivariate t distribution with d degrees of freedom and with correlation coefficient $\rho = 0.4$. Top left panel: $d = 1.5$; top right panel: $d = 2$; bottom left panel: $d = 3$; bottom right panel: $d = 4$. In all panels, the blocks relate to the sample size: (from left to right) $n = 100, 500, 1000, 2000$.**

We see from Figures 5.3 and 5.4 that the performance of the kernel density estimator with empirical bandwidth selector is very unpredictable in small to moderate sample sizes, a problem that is amplified when $p = 3$. Except for in the trivariate case with 2 degrees of freedom and $n = 2000$ observations, where the kernel density estimator marginally outperforms the mixture sieve (EG), the Gaussian mixture sieve (EG and EM) estimates consistently yield a smaller median ISE estimate than do the Liebscher estimates and the kernel estimates based on empirically selected bandwidth. The differences are particularly prominent in the cases where the sample size is small and the true density is heavy tailed. The Liebscher estimator has a much smaller standard deviation than does the LSVC kernel in small sample sizes, and it marginally outperforms for all sample sizes in the $p = 2$ case when the tails of the true

density are not too heavy. The EM algorithm tends to slightly outperform the EG algorithm in terms of median integrated squared error in the $p = 3$ case. In the $p = 2$ case, the EG algorithm tends to slightly outperform EM when the α is estimated based on the rule of thumb procedure. It all scenarios, there is very little difference in performance between the EM and EG algorithms. Further figures are available in a longer version of this paper.

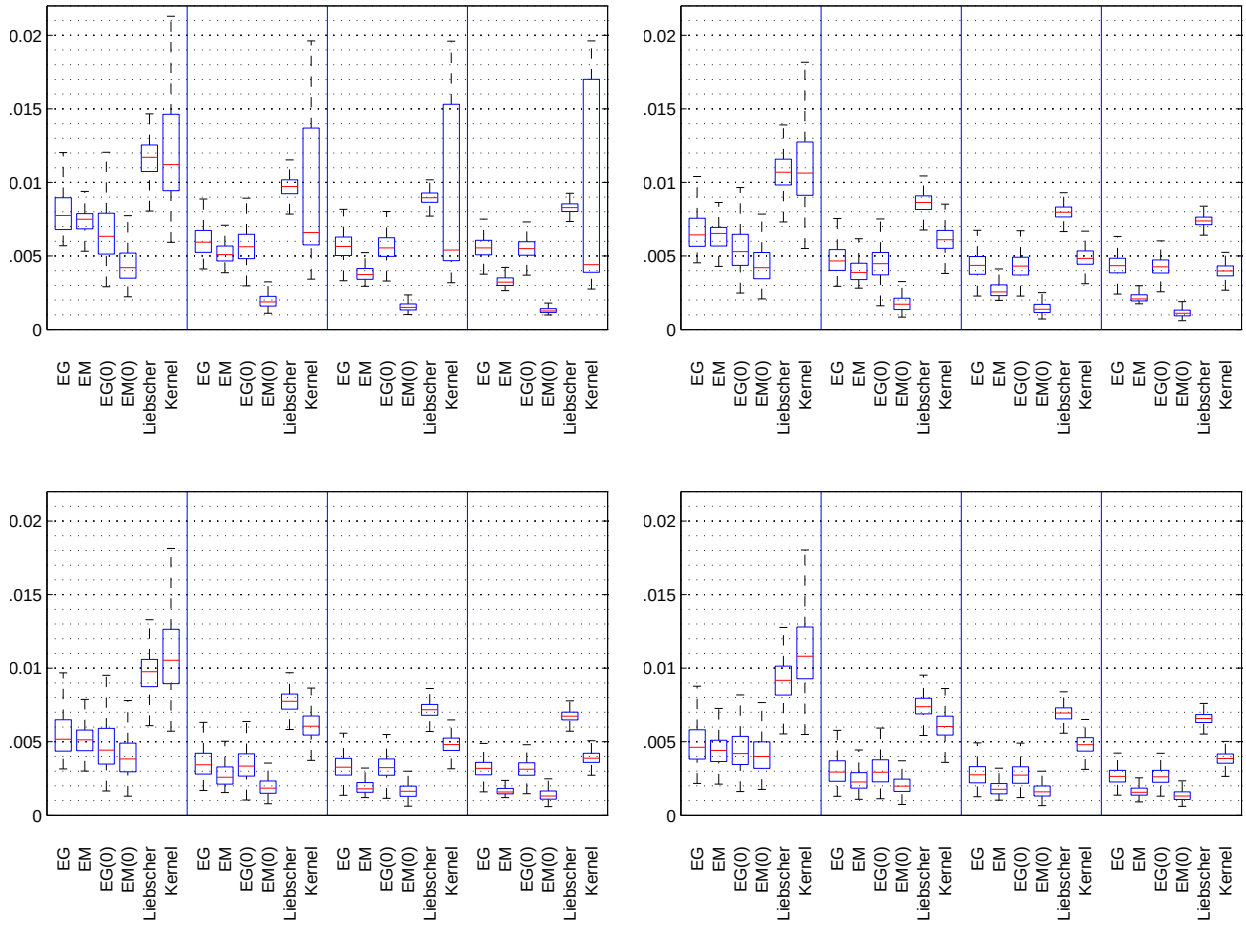

Figure 5.4: **Boxplots of the square root of the integrated squared error (divided by the number of points in the grid-array of evaluation points) based on 1000 MC replications using n observations from a trivariate t distribution with d degrees of freedom and with correlation coefficients $\rho_{12} = 0.4$, $\rho_{13} = -0.7$, $\rho_{23} = 0.1$. Top left panel: $d = 1.5$; top right panel: $d = 2$; bottom left panel: $d = 3$; bottom right panel: $d = 4$. In all panels, the blocks relate to the sample size: (from left to right) $n = 100, 500, 1000, 2000$.**

Because our proof strategy does not allow is to comment on the theoretical performance of our estimator in a neighbourhood of the origin, we provide in Figures 5.5 and 5.6 the root squared errors (over the same 1000 Monte Carlo replications) of each estimator at zero. While the Liebscher estimator behaves moderately well at zero in the $p = 2$ case (see Figure 5.5), when $p > 2$ the estimator is not defined at zero. This is due to the nature of

the transformations employed; referring to equation (2.5) of Liebscher (2005) we see that whenever $p > 2$, the component $z^{-p/2+1}$ is problematic at $z = 0$, and equation (2.5) results in a zero times infinity operation. For this reason the Liebscher estimator is not used as a comparison in Figure 5.6.

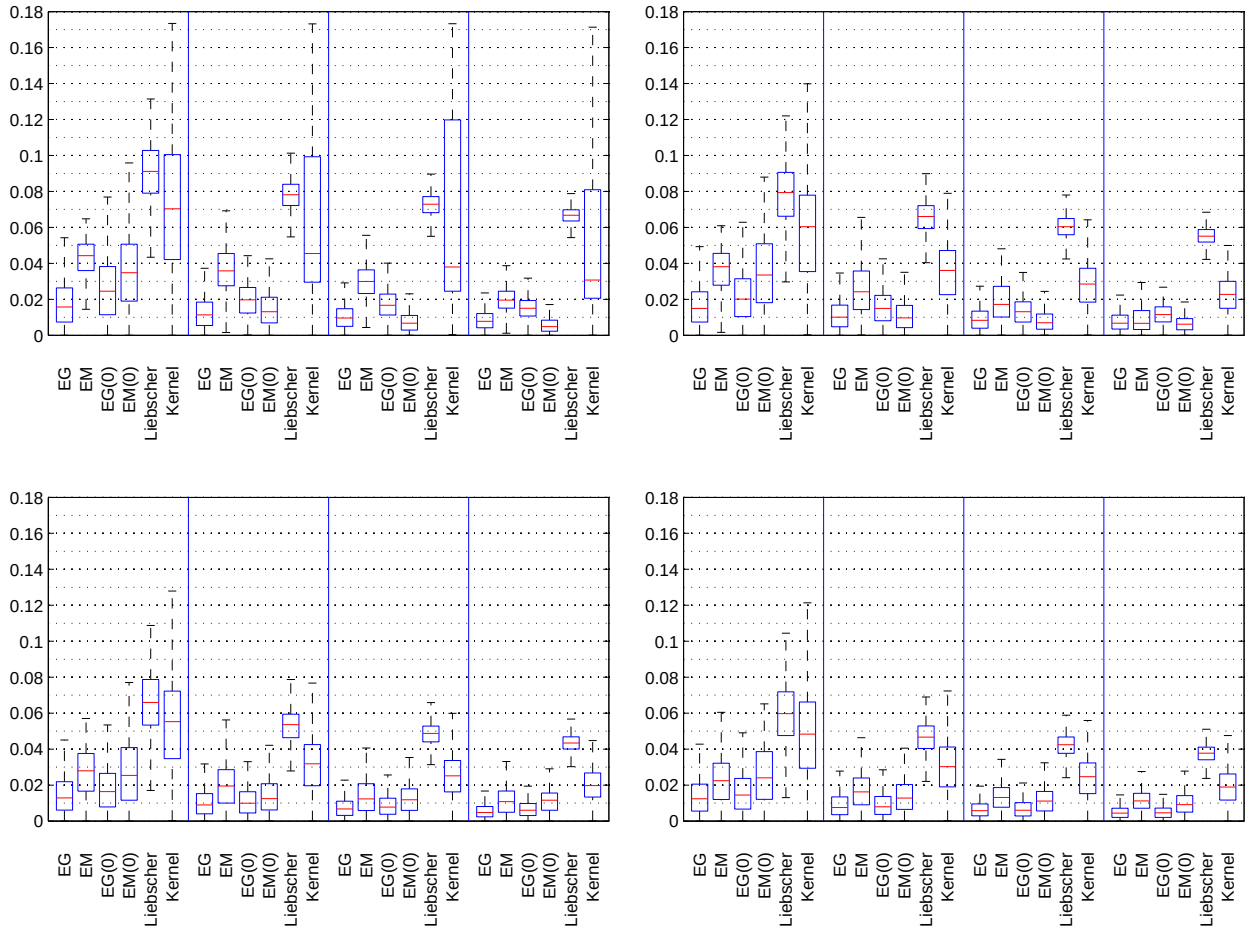

Figure 5.5: **Boxplots of the root squared error at zero based on 1000 MC replications using n observations from a bivariate t distribution with d degrees of freedom and with correlation coefficient $\rho = 0.4$. Top left panel: $d = 1.5$; top right panel: $d = 2$; bottom left panel: $d = 3$; bottom right panel: $d = 4$. In all panels, the blocks relate to the sample size: (from left to right) $n = 100, 500, 1000, 2000$.**

It is worth noting that the difference in computation time for the different estimators is substantial in large and moderate sample sizes, especially when the data are heavy tailed. On the same standard desktop computer (2.4GHz, 2GB RAM), the average computation time over ten different draws of $n = 2000$ observations from the trivariate t distribution with 1.5 degrees of freedom was 0.74 seconds for the EG mixture sieve (true α), 27.31 seconds for the EM mixture sieve (true α), and 10.32 seconds for the Liebscher estimator, whilst that for the LSVC kernel density estimator was over 10 minutes. It is comforting to note

that, the EG algorithm, an algorithm with such substantial practical advantages when it comes to handling large and growing datasets can perform so well, even in samples of small and moderate size. Since each iteration of the EG alogirthm with $p = 3$ only takes around 8/1000 seconds to compute, it is an attractive option for handling financial and other data generated on a tic-by-tic basis.

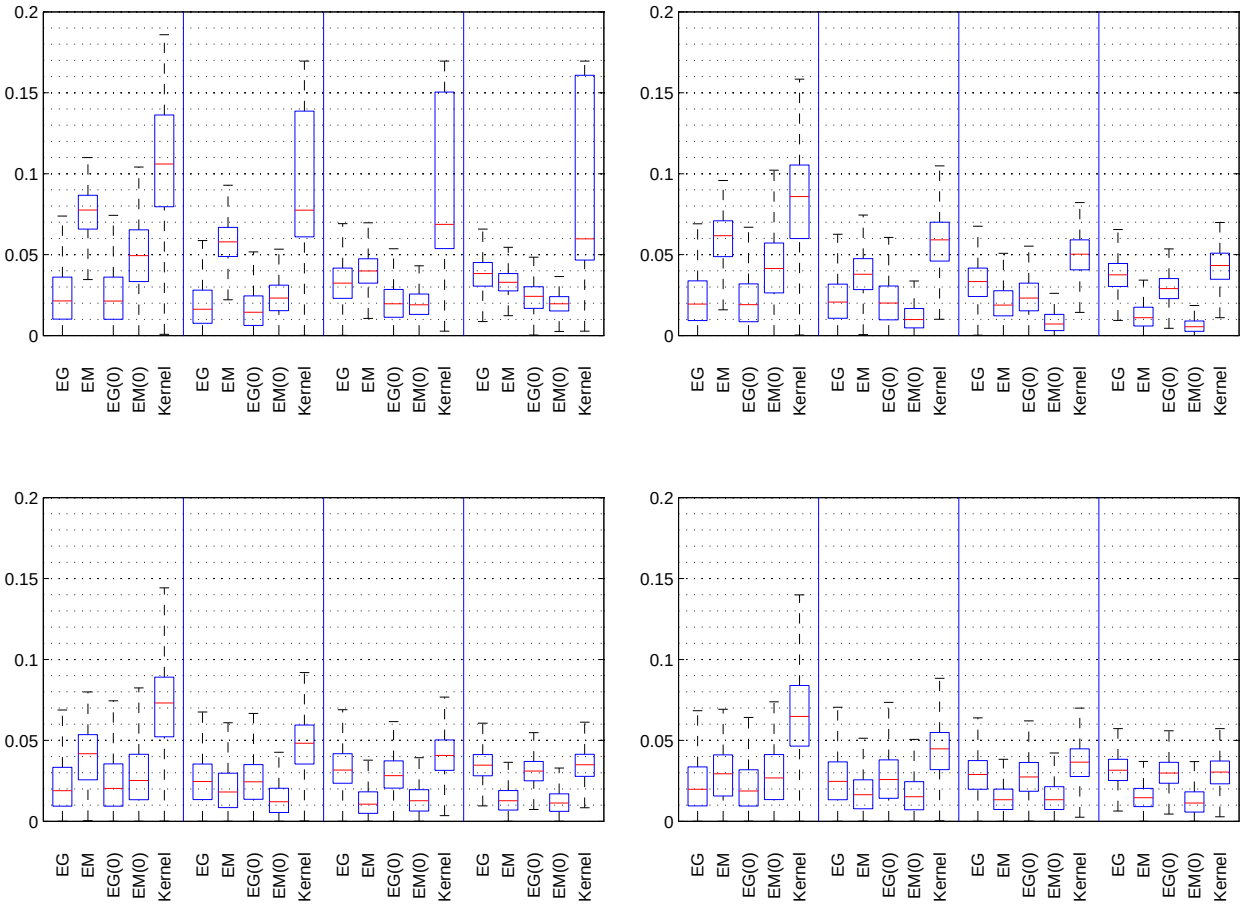

Figure 5.6: **Boxplots of the root squared error at zero based on 1000 MC replications using n observations from a trivariate t distribution with d degrees of freedom and with correlation coefficients $\rho_{12} = 0.4$, $\rho_{13} = -0.7$, $\rho_{23} = 0.1$. Top left panel: $d = 1.5$; top right panel: $d = 2$; bottom left panel: $d = 3$; bottom right panel: $d = 4$. In all panels, the blocks relate to the sample size: (from left to right) $n = 100, 500, 1000, 2000$.**

5.3 Sensitivity to the choice of c

A natural question that arises when considering Theorem 1 is how large a role the choice of c makes to the performance of the estimator. To address this question, we analyse the performance of the mixture sieve (EG) algorithm over a range of values of c. For c increasing from 0.5 to 2 in steps of size 0.1, we compute the integrated squared errors of the EG estimator

in 100 simulation experiments in which $n = 500$ observations were drawn from a bivariate t distribution with 1.5 degrees of freedom and correlation coefficient $\rho = 0.4$. η was taken as 1.5 and α was fixed at 0.75, which is the correct value of α to use when the data are drawn from a t distribution with 1.5 degrees of freedom. An analogous experiment is conducted for the 4 degrees of freedom case. The results are displayed in Figure 5.7.

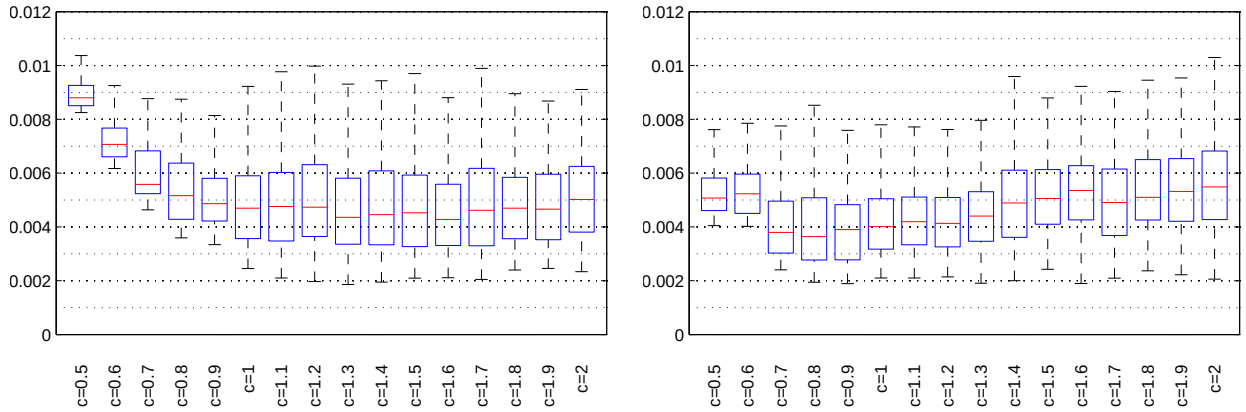

Figure 5.7: **Boxplots of the root integrated squared error (divided by the number of points in the grid-array of evaluation points) based on 100 MC replications using 500 observations from a bivariate t distribution with d degrees of freedom and with correlation coefficient $\rho = 0.4$. Left panel: $d = 1.5$; right panel: $d = 4$.**

As Figure 5.7 indicates for this example, in practice the estimator is rather insensitive to the choice of c as long as c is not taken too small. The reason that small c results in a deterioration in performance is that too few mixtures are being used to approximate the density. By contrast, when a large number of mixtures are used, the algorithm produces estimates of the mixing weight that are close to zero for many of the mixture components.

6 Application to binary classification of Wisconsin breast cancer data

We consider our elliptic density estimator in the context of binary classification using the well-known Wisconsin breast cancer (diagnostic) dataset used also in Liebscher (2005). This dataset is available on the UCI Machine Learning Repository website:
http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.
It consists of 30 real-valued continuous attributes based on the cell nuclei of 569 breast tumor patients, of which 212 instances are malignant and 357 instances are benign, along with a variable indicating whether the tumor was malignant or benign. The dataset is discussed in more detail in Street et al. (1993). Figure 6.1 presents a smoothed scatterplot matrix of three of these attributes (the area, smoothness and texture of the worst nuclei observations

for each patient) for the malignant and benign groups.

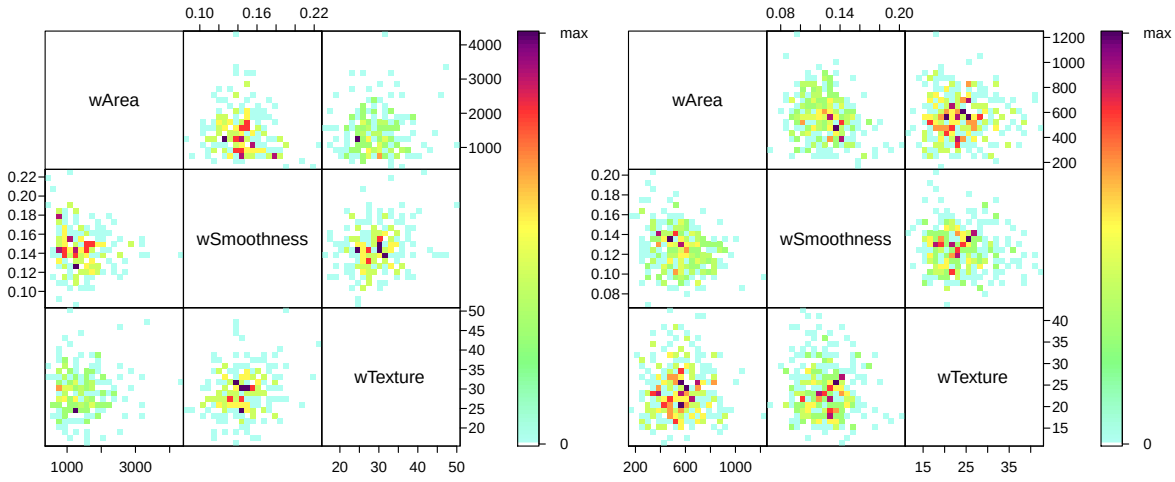

Figure 6.1: **Smoothed scatter plot matrices for malignant cases (left) and benign cases (right) for three of the attributes of the Wisconsin breast cancer data: worst area, worst smoothness and worst texture.**

Figure 6.2 plots the hill estimates of $\alpha(R)$ against the threshold, K (see section 4.1) for the benign and the malignant cases (standardised data). It suggests that the tails of the inverse mixing measure for the trivariate density of worst case area, texture and smoothness are much heavier in the case of a malignant tumour than in the case of a benign tumour. We consider the problem of classifying patients into the benign or malignant group based on observations on these three variables. To this end, we construct trivariate density estimates in the benign and malignant groups using the mixture sieve (EG) estimator with tuning parameter α chosen by the rule of thumb procedure in section 4.1. Using a threshold of $K = 65$ observations in equation (4.2) yields a $\widehat{\alpha}(1/Q)$ of 3.47 in the malignant case and of 4.17 in the benign case. Letting $\widehat{f}_n(y|M)$ denote the density at some evaluation point y in the malignant group and letting $\widehat{f}_n(y|B)$ denote the density at y in the benign group, we may construct a Bayes' classifier based on the estimated posterior probabilities,

$$\widehat{P}(M|y) = \frac{\widehat{f}_n^M(y|M)\widehat{P}(M)}{\widehat{f}_n(y|M)\widehat{P}(M) + \widehat{f}_n(y|B)\widehat{P}(B)} \quad \text{and} \quad \widehat{P}(B|y) = \frac{\widehat{f}_n^M(y|B)\widehat{P}(B)}{\widehat{f}_n(y|M)\widehat{P}(M) + \widehat{f}_n(y|B)\widehat{P}(B)} .$$

where $\widehat{P}(M)$ and $\widehat{P}(B)$ denote the estimated probability of being in the malignant and benign groups respectively; these quantities are obtained using the empirical proportions of malignant and benign cases. The relative magnitudes of $\widehat{P}(M|y_i)$ and $\widehat{P}(B|y_i)$ then determines whether individual i with attributes y_i is assigned to group M (malignant) or group B (benign). Based on these posterior probabilities, we test the performance of the classifier by comparing the output of the classifier to the true group labels. Our classifier misclassifies

only 69 of the patients in the malignant group, as compared with 79 misclassifications in the malignant group when the classification is based on the trivariate normal distribution. However, the trivariate normal model does classify the patients in the benign group better, with 33 misclassifications occurring rather than 43 in the case of the our classifier.

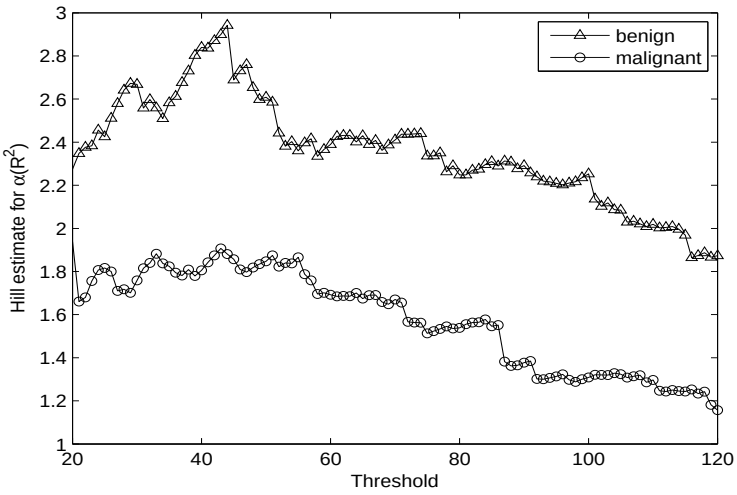

Figure 6.2: **hill estimates of $\alpha(R)$ against threshold, K, for malignant and benign cases.**

We have presented here one possible application of the methodology developed in this paper. A further application to systemic risk management using emerging markets exchange rates is presented in a longer version of this paper, which is available elsewhere.

7 Proofs

Proof of Theorem 1. As usual, decompose the left hand side of equation (4.1) into estimation error and approximation error,

$$\begin{aligned}\sup_{y\in D}|f(y)-\widehat{f}_n^M(y)| &\leq \sup_{y\in D}|f(y)-f^M(y)|+\sup_{y\in D}|f^M(y)-\widehat{f}_n^M(y)| \\ &= I_1+I_2,\end{aligned} \tag{7.1}$$

We will make use of linear functional notation throughout, i.e.

$$\mathbb{P}(A):=\int \mathbb{1}\{1/q\in A\}\mathbb{P}(d(1/q)) \quad \text{and} \quad \mathbb{P}h:=\int_{\mathcal{Q}} h(1/q)\mathbb{P}(d(1/q)).$$

The controls over I_1 and I_2 both rely on the following preliminary lemma

Lemma 1. *For a fixed $p<\infty$ and D any fixed compact subset of $\mathbb{R}^p$ that does not include the zero vector, the set of functions*

$$\mathcal{G}:=\{g: q\mapsto \psi(y|q);\ q\in\mathcal{Q},\ y\in D\}, \tag{7.2}$$

is a subset of $\mathcal{B}$, the set of bounded functions from A to $\mathbb{R}$, where A is an arbitrary set.

Proof. Let

$$\zeta(q) := q^{p/2} \quad \text{and} \quad \vartheta(q) := \exp\left\{-qy^T\Omega^{-1}y\right\},$$

and as before, let $w := y^T\Omega^{-1}y$. Then $\lim_{1/q\to 0}\zeta(q)\vartheta(q)$ is indeterminate. Transforming, we have

$$\lim_{1/q\to 0}\zeta(q)\vartheta(q) = \lim_{1/q\to 0}\frac{\zeta(q)}{1/\vartheta(q)},$$

which is zero for all w outside a small neighbourhood of $w = 0$ and $p < \infty$, hence $\mathcal{G} \subset \mathcal{B}$. Note that the above limit is infinity in a neighbourhood of $w = 0$, which is why a neighbourhood of $w = 0$ must be excluded. □

Control over I_1.

Definition 1 (Dudley (2002))**.** Let $\mathbb{P}_M$ and $\mathbb{P}$ be laws on $\mathcal{Q}$. The *Prohorov metric* is defined as

$$\rho(\mathbb{P}_M, \mathbb{P}) := \inf\left\{\epsilon > 0 : \mathbb{P}_M(A) \le \mathbb{P}(A^\epsilon) + \epsilon \text{ for all Borel sets A}\right\}, \tag{7.3}$$

where A is an arbitrary Borel set and $A^\epsilon := \{r \in \mathbb{R}^+ : d(r, A) < \epsilon\}$ is the "ϵ-enlargement" of A.

Lemma 2. *Let $\mathbb{P}_M \in \mathcal{M}$ with*

$$\mathcal{M} := \left\{\sum_{s=1}^{M}\Lambda_s\delta_{1/q_s} : \Lambda_1, \ldots, \Lambda_M \in \mathbb{Q}\cap[0,1],\ \sum_{s=1}^{M}\Lambda_s = 1, M = 1, 2, \ldots\right\}.$$

δ_{1/q_s} is the Dirac measure at $1/q_s$, i.e., for an arbitrary Borel set A, $\delta_{1/q_s}(A) = 1$ if $1/q_s \in A$, 0 otherwise. Under Condition 4,

$$\rho(\mathbb{P}, \mathbb{P}_M) = O\left(M^{-\alpha/(1+\alpha)}\right),$$

where $\rho(\cdot,\cdot)$ is the Prohorov metric of equation (7.3).

Proof. Let $\{1/q_s : s \in \mathbb{N}\}$ be equally spaced over $\mathcal{Q}$ and let $\mathcal{P}(\mathcal{Q})$ be the set of all probability measures on the Borel sets of $\mathcal{Q}$. Since $\mathcal{M}$ is dense in $\mathcal{P}(\mathcal{Q})$ under the weak topology (Parthasarathy, 1967, Theorem 6.3), there exists a choice of weights, $(\Lambda_s)_{s=1,\ldots M}$, such that the sequence of weighted discrete measures, $\mathbb{P}_M(A) := \sum_{s=1}^{M}\Lambda_s\delta_{1/q_s}(A)$ converges weakly to $\mathbb{P}(A)$ as $M \to \infty$. Furthermore, separability of $\mathcal{Q}$ means weak convergence of $\mathbb{P}_M$ to $\mathbb{P}$ is equivalent to convergence under the Prohorov metric (Dudley, 2002, Theorem 11.3.3). To establish the rate of convergence, we fix an $\epsilon > 0$ and introduce the function $h \in \mathcal{C}_b$,

$$h(1/q) = 0 \vee \left(1 - \frac{1}{2\epsilon}d(1/q, A)\right), \tag{7.4}$$

which is the same function introduced in Dudley (2002) Chapter 11.3. In the above definition $d(1/q, A) = \inf_{r\in A} d(1/q, r)$, so that $h(1/q) = 1$ if and only if $1/q \in A$, and $h(1/q) = 0$ as soon as $1/q$ is more than a 2ϵ away from the boundary of A. Hence $\mathbb{1}\{1/q \in A\} \leq h(1/q) \leq \mathbb{1}\{1/q \in A^{2\epsilon}\}$. For this fixed ϵ, there exists a $x = \epsilon M(\epsilon)$ such that $\mathbb{P}(1/Q > x(\epsilon)) \leq \epsilon$ by Condition 4. Cover $[0, x(\epsilon)]$ with disjoint open balls of radius $\epsilon/2$ around the $\{1/q_s : s = 1, \ldots, M(\epsilon)\}$, i.e. with B_s^ϵ satisfying $B_s^\epsilon \cap B_j^\epsilon = 0 \ \forall j \neq s$, and fix $\{\Lambda_s : s = 1, \ldots, M(\epsilon)\}$ such that $\sum_{s=1}^{M(\epsilon)} |\Lambda_s \delta_{1/q_s} - \mathbb{P}(B_s^\epsilon)| \leq \epsilon$. Then for the arbitrary Borel set A,

$$\begin{aligned}
\mathbb{P}_M(A) &\leq \int_{\mathcal{Q}} h(1/q)\mathbb{P}(d(1/q)) + \int_{\mathcal{Q}} h(1/q)|\mathbb{P}_M - \mathbb{P}|(d(1/q)) \\
&\leq \int_{\mathcal{Q}} \mathbb{1}\{1/q \in A^{2\epsilon}\}\mathbb{P}(d(1/q)) + \int_{\mathcal{Q}} h(1/q)|\mathbb{P}_M - \mathbb{P}|(d(1/q)) \\
&= \mathbb{P}(A^{2\epsilon}) + \left| \sum_{s=1}^{M(\epsilon)} \Lambda_s h(1/q_s) - \int_{\mathcal{Q}} h(1/q)\mathbb{P}(d(1/q)) \right| \\
&\leq \mathbb{P}(A^{2\epsilon}) + \sup_{r\in\mathcal{Q}} |h(r)| \sum_{s=1}^{M(\epsilon)} \left| \Lambda_s \delta_{1/q_s} - \mathbb{P}(B_s^\epsilon) \right| + \sup_{r\in\mathcal{Q}} |h(r)| \mathbb{P}\left(\left(\cup_{s=1}^{M(\epsilon)} B_s^\epsilon \right)^c \right) \\
&\leq \mathbb{P}(A^{2\epsilon}) + 2\epsilon.
\end{aligned}$$

Since $x = \epsilon M(\epsilon)$, Condition 4 implies that

$$\mathbb{P}(1/Q > \epsilon M(\epsilon)) \lesssim [\epsilon M(\epsilon)]^{-\alpha}$$

where $\lesssim$ means *less than up to a finite absolute constant.* Solving for ϵ gives

$$\epsilon = [\epsilon M(\epsilon)]^{-\alpha} = M(\epsilon)^{-\frac{\alpha}{1+\alpha}} \tag{7.5}$$

The infimum over the ϵ such that the condition $\mathbb{P}_M \leq \mathbb{P}(A^{2\epsilon}) + 2\epsilon$ holds is $\epsilon = M(\epsilon)^{-\alpha/(1+\alpha)}$ and thus the Prohorov metric tends to zero at a rate of $O(M^{-\alpha/(1+\alpha)})$ when $x = \epsilon M(\epsilon) = M^{\frac{1}{1+\alpha}}$. □

Corollary 1. *Under Conditions 1 and 4, with $\mathbb{P}_M$ as defined in Lemma 2 and $\{1/q_s : s = 1, \ldots, M\}$ equally spaced between $0 + \epsilon/2$ and $x - \epsilon/2 = M^{\frac{1}{1+\alpha}} - \epsilon/2$, where $\epsilon = M^{-\alpha/1+\alpha}$,*

$$I_1 = |\Omega|^{-1/2} \sup_{y\in D} \left| \int_0^\infty \psi(y|q)(\mathbb{P} - \mathbb{P}_M)(dq) \right| = O\left(M^{-\alpha/(1+\alpha)}\right).$$

Proof. By Jensen's inequality and the convexity of the supremum and the absolute value

$$\begin{aligned}
I_1 =& |\Omega|^{-1/2} \sup_{y\in D} \left| \int_0^\infty \psi(y|q)(\mathbb{P} - \mathbb{P}_M)(d(1/q)) \right| \\
\leq& |\Omega|^{-1/2} \int_0^\infty \sup_{y\in D} \psi(y|q)(d(1/q)) \int |\mathbb{P} - \mathbb{P}_M| \, (d(1/q)).
\end{aligned} \tag{7.6}$$

The right hand side of (7.6) converges at rate r_M if

$$\int_0^\infty \sup_{y\in D} \psi(y|q) d(1/q) < \infty \tag{7.7}$$

and $\mathbb{P}_M \to \mathbb{P}$ at rate r_M in some metric that metricises the topology of weak convergence. Lemma 1 ensures

$$\sup_{y\in D} \psi(y|q) < \infty \quad \forall q \in \mathbb{R}^+,$$

and since $\sup_{y\in D} \psi(y|q) \searrow 0$ for $1/q \to \infty$, the condition in equation 7.7 is satisfied. The rate r_M is provided by Lemma 2 alongside Theorem 11.3.3 of Dudley (2002). □

Control over I_2. Since we control the approximation error separately, we assume in what follows, that each Y_i is generated from a member of the set of densities $\{\psi(\cdot|q_s) : q_s \in \mathbb{R}^+,\ s = 1, \ldots, M\}$ and write the true log likelihood, i.e. the one that contains the latent mixture indicators $(Z_{i,s})$, as

$$P_n L(\Lambda) := \frac{1}{n} \sum_{i=1}^{n} \ln \left(\sum_{s=1}^{M} \Lambda_s |\Omega|^{1/2} Z_{i,s} \psi(Y_i|q_s) \right),$$

where

$$Z_{i,s} = \begin{cases} 1 & \text{if } Y_i \in \mathcal{Y}_s \\ 0 & \text{otherwise,} \end{cases}$$

and $\mathcal{Y}_s$ is the set of Y_i drawn from the s^{th} mixture component. Then the infeasible maximum likelihood estimator is

$$\widehat{\Lambda}^{ML} := \operatorname*{argsup}_{\Lambda\in\Gamma} P_n L(\Lambda) \quad \text{and} \quad \Lambda^0 := \operatorname*{argsup}_{\Lambda\in\Gamma} PL(\Lambda),$$

where

$$\Gamma = \left\{ \Lambda : \Lambda_s \in [0,1] \quad \forall s,\ \sum_{s=1}^{M} \Lambda_s = 1 \right\};$$

P_n is the empirical measure and P is the true one.

Let $\widehat{\Lambda}_I^{EG}$ denote the vector of average weights computed over $I = n$ iterations of the EG

algorithm. We decompose I_2 as

$$\begin{aligned} I_2 &= \sup_{y\in D}\left||\Omega|^{-\frac{1}{2}}\sum_{s=1}^{M}\Lambda_s^0\phi(y|\Omega/q_s) - |\widehat{\Omega}|^{-\frac{1}{2}}\sum_{s=1}^{M}\widehat{\Lambda}_{I,s}^{EG}\phi(y|\widehat{\Omega}_n/q_s)\right| \\ &\leq \sup_{y\in D}\left|\left(|\Omega|^{-1/2} - |\widehat{\Omega}|^{-1/2}\right)\sum_{s=1}^{M}\Lambda_s^0\phi(y|\Omega/q_s)\right| \\ &\quad + \sup_{y\in D}\left||\widehat{\Omega}|^{-\frac{1}{2}}\sum_{s=1}^{M}\Lambda_s^0\phi(y|\Omega/q_s) - |\widehat{\Omega}|^{-\frac{1}{2}}\sum_{s=1}^{M}\Lambda_s^0\phi(y|\widehat{\Omega}/q_s)\right| \\ &\quad + |\widehat{\Omega}|^{-\frac{1}{2}}\sup_{y\in D}\left|\sum_{s=1}^{M}\Lambda_s^0\phi(y|\Omega/q_s) - \sum_{s=1}^{M}\widehat{\Lambda}_{I,s}^{EG}\phi(y|\Omega/q_s)\right| \\ &= II_1 + II_2 + II_3. \end{aligned}$$

Remark. $|\cdot|$ denotes both determinant and absolute value here, with the context ensuring no ambiguity.

Control over II_1.

$$II_1 \leq \left|\left(|\Omega|^{-1/2} - |\widehat{\Omega}|^{-1/2}\right)\right|\sum_{s=1}^{M}\Lambda_s^0\sup_{y\in D}\phi(y|\Omega/q_s)$$

by convexity of the supremum. By Condition 5, $\widehat{\Omega}$ is root-n consistent for Ω; we show here that the determinant is also root-n consistent. The determinant of a p-dimensional matrix $X := (x_{jk})$ is given by $|X| = \sum_{j=1}^{p} c_{jk}x_{jk}$ for any $k \in \{1,\ldots,p\}$, where $C := (c_{jk})$ is a matrix of cofactors (Abadir and Magnus, 2005, Exercise 4.36) hence for any $k \in \{1,\ldots,p\}$

$$\left||\widehat{\Omega}| - |\Omega|\right| \leq \sum_{j=1}^{p}|c_{jk} - \widehat{c}_{jk}||\Omega_{jk}| + \sum_{j=1}^{p}|c_{jk}||\Omega_{jk} - \widehat{\Omega}_{jk}|$$

where

$$|c_{jk} - \widehat{c}_{jk}| = \left|(-1)^{j+k}\left(|\Omega^{jk}| - |\widehat{\Omega}^{jk}|\right)\right|$$

and Ω^{jk} is the $(p-1)$-dimensional matrix obtained by removing the j^{th} row and k^{th} column from Ω. By induction, $\left||\widehat{\Omega}| - |\Omega|\right| = O_p(n^{-1/2})$ with constants of order $p!$. The final result follows by Lemma 1.

Control over II_2. It suffices to control $M\max_s\sup_{y\in D}|\phi(y|\widehat{\Omega}/q_s) - \phi(y|\widehat{\Omega}/q_s)|$. Notice that

$$\begin{aligned} &\Pr\left(|\phi(y|\widehat{\Omega}/q_s) - \phi(y|\widehat{\Omega}/q_s)| > K\eta_n\right) \\ \leq\ &\Pr\left(|\phi(y|\widehat{\Omega}/q_s) - \phi(y|\widehat{\Omega}/q_s)| > K\eta_n, \|\widehat{\Omega} - \Omega\|_1 \leq K\delta_n\right) \\ &+ \Pr\left(\|\widehat{\Omega} - \Omega\|_1 > K\delta_n\right) \\ =\ &III_1 + III_2. \end{aligned}$$

III_2 is $O(1)$ with $\delta_n = n^{-1/2}$ and constant of the order $p(p-1)/2$ by Condition 5. For any fixed $y \in D$ and $s \in \{1, \ldots, M\}$ let $D\phi(\widehat{\Omega})$ denote the derivative of $\phi(y|\Omega/q_s)$ with respect to Ω evaluated at $\widehat{\Omega}$. Since

$$\begin{aligned} & \sup_{y \in D, q_s \in \{q_1, \ldots, q_M\}} \|D\phi(\widehat{\Omega})\|_1 \\ = & \sup_{y \in D, q_s \in \{q_1, \ldots, q_M\}} \left\| (2\pi)^{-p/2} \exp\left\{ -\frac{1}{2} q_s y^T \widehat{\Omega} y \right\} (-q_s \widehat{\Omega}^{-1} y \otimes \widehat{\Omega}^{-1} y) \right\|_1 \\ < & \infty, \end{aligned}$$

we have, through a Taylor expansion of $\phi(y|\Omega/q_s)$ around $\phi(y|\widehat{\Omega}/q_s)$,

$$\sup_{y \in D, q_s \in \{q_1, \ldots, q_M\}} \left| \phi(y|\Omega/q_s) - \phi(y|\widehat{\Omega}/q_s) \right| \lesssim \|\widehat{\Omega}_n - \Omega\|_1 \, .$$

where $\|\cdot\|_1$ is the l_1 norm. The right hand side is bounded by $K\delta_n = Kn^{-1/2}$ by the statement of the joint event. $III_1 = O_p(n^{-1/2})$ uniformly over $y \in D$ and $q_s \in \{q_1, \ldots, q_M\}$ which together with III_2 implies II_2 is $O_p(M/\sqrt{n})$.

Control over II_3. The control over II_3 relies on three preliminary lemmata, stated and proved below.

Lemma 3. *Suppose Conditions 1 - 4 hold. Then*

(i) $d(\widehat{\Lambda}^{ML}, \Lambda_0) \to_p 0$, *and*

(ii) $d(\widehat{\Lambda}^{ML}, \Lambda_0) = O_p(M/\sqrt{n})$,

where $d(\cdot, \cdot)$ *is Euclidean distance.*

Proof. To prove part (i) we verify the conditions of Theorem 5.7 of van der Vaart (1998). Let

$$\mathcal{L} := \left\{ L(\Lambda) : \Lambda \mapsto \ln(\sum_s \Lambda_s Z \psi(Y|q_s)) \; ; \; \Lambda \in \Gamma \right\} \subset \mathbb{L}_1(P),$$

and let $\mathbb{M}(\Lambda) = PL(\Lambda)$, $\mathbb{M}_n(\Lambda) = P_n L(\Lambda)$ so that $\mathbb{M}_n(\Lambda) - \mathbb{M}(\Lambda) = (P_n - P)L(\Lambda; Y, Z)$. By the identification established in Holzmann et al. (2006), the map $\Lambda \mapsto \mathbb{M}(\Lambda)$ has unique maximum at Λ^0, hence

$$\sup_{\Lambda \in \Gamma : d(\Lambda, \Lambda^0) \geq \delta} \mathbb{M}(\Lambda) < \mathbb{M}(\Lambda^0) \quad \forall \delta > 0.$$

It remains to show that

$$\sup_{\Lambda \in \Gamma} |\mathbb{M}_n(\Lambda) - \mathbb{M}(\Lambda)| \to_p 0. \tag{7.8}$$

Since $\Gamma \subset [0,1]^M$ is compact, it has a finite cover, hence we can easily construct brackets for $\mathcal{L}$ (see e.g. van de Geer, 2000, proof of Lemma 3.10). (7.8) follows by Lemma 3.1 of van de Geer (2000).

Our proof of part (ii) follows van der Vaart and Wellner (1996) Theorem 3.2.5. Re-write our goal in Lemma 3 (ii) as

$$\lim_{J\to\infty} \limsup_{n\to\infty} \Pr\left(r_n d(\widehat{\Lambda}^{ML}, \Lambda^0) > 2^J\right) = 0. \tag{7.9}$$

Through a Taylor series expansion of $\mathbb{M}(\Lambda)$ around Λ^0 we have

$$\mathbb{M}(\Lambda) = \mathbb{M}(\Lambda^0) + (\Lambda - \Lambda^0)^T \ddot{\mathbb{M}}(\Lambda^0)(\Lambda - \Lambda^0) + o(\|\Lambda - \Lambda^0\|_2^2) \qquad \forall \Lambda \in \Gamma,$$

where the Hessian matrix $\ddot{\mathbb{M}}(\Lambda^0)$ is negative definite, implying

$$\mathbb{M}(\Lambda^0) - \mathbb{M}(\Lambda) \geq C d^2(\Lambda^0, \Lambda) \quad \forall \Lambda \in \Gamma. \tag{7.10}$$

Next, by the definition of $\widehat{\Lambda}^{ML}$ as a maximiser,

$$\sup_{\Lambda\in\Gamma} \mathbb{M}_n(\Lambda) \leq \mathbb{M}_n(\widehat{\Lambda}^{ML}). \tag{7.11}$$

For $n, j \in \mathbb{N}$, we introduce the sequence of sets

$$S_{jn} := \left\{\Lambda \in \Gamma : 2^{j-1} < r_n \|\widehat{\Lambda}^{ML} - \Lambda^0\|_2 \leq 2^j\right\}, \quad j \in \mathbb{N},$$

where r_n is an divergent sequence of positive real numbers and, as before

$$\Gamma = \left\{\Lambda : \Lambda_s \in [0,1]\ \forall s,\ \sum_{s=1}^{M} \Lambda_s = 1\right\}.$$

Let $\mathfrak{K} := \bigcup_{j=j_0}^{\infty} S_{jn}$, where $j_0 > J$. For $\widehat{\Lambda}^{ML} \in \mathfrak{K}$

$$\begin{aligned}
& \mathbb{M}_n(\widehat{\Lambda}^{ML}) - \mathbb{M}_n(\Lambda^0) + \mathbb{M}(\Lambda^0) - \mathbb{M}(\widehat{\Lambda}^{ML}) \\
= \ & (P_n - P)L(\widehat{\Lambda}^{ML}) - (P_n - P)L(\Lambda^0) \\
\geq \ & C d^2(\Lambda^0, \widehat{\Lambda}^{ML}) \\
& \text{[by equations (7.10) and (7.11)]} \\
> \ & C(r_n^{-1} 2^{j_0 - 1})^2 \\
= \ & C r_n^{-1} 2^{2j_0 - 2}.
\end{aligned}$$

Letting $\mathbb{G}_n(\Lambda) := \sqrt{n}(P_n - P)L(\Lambda)$, introduce the event

$$B := \left\{\frac{1}{\sqrt{n}}\left(\mathbb{G}_n(\widehat{\Lambda}^{ML}) - \mathbb{G}_n(\Lambda^0)\right) > \frac{C 2^{2j_0-2}}{r_n^2}, \widehat{\Lambda}^{ML} \in \mathfrak{K}\right\},$$

and consider

$$
\begin{aligned}
& \Pr(r_n\|\widehat{\Lambda}^{ML}-\Lambda^0\|_2>2^J) \\
\leq\ & \Pr\left(B,\|\widehat{\Lambda}^{ML}-\Lambda^0\|_2<\eta\right)+\Pr\left(B,\|\widehat{\Lambda}^{ML}-\Lambda^0\|_2\geq\eta\right) \\
& \text{[by equation (7.11)]} \\
\leq\ & \Pr\left(B,\|\widehat{\Lambda}^{ML}-\Lambda^0\|_2<\eta\right)+\Pr(\|\widehat{\Lambda}^{ML}-\Lambda^0\|_2\geq\eta) \\
\leq\ & \Pr\left(\bigcup_{j\geq j_0:2^{j-1}<r_n\eta}\sup_{\Lambda\in S_{jn}}\frac{1}{\sqrt{n}}\left(\mathbb{G}_n(\Lambda)-\mathbb{G}_n(\Lambda^0)\right)>\frac{C2^{2j-2}}{r_n^2}\right)+\Pr(\|\widehat{\Lambda}^{ML}-\Lambda^0\|_2\geq\eta) \\
\leq\ & \sum_{j\geq j_0:2^{j-1}<r_n\eta}\Pr\left(\sup_{\Lambda\in S_{jn}}\frac{1}{\sqrt{n}}\left(\mathbb{G}_n(\Lambda)-\mathbb{G}_n(\Lambda^0)\right)>\frac{C2^{2j-2}}{r_n^2}\right)+\Pr(2\|\widehat{\Lambda}^{ML}-\Lambda^0\|_2\geq\eta) \\
& \text{[by Boole's inequality]} \\
=\ & IV_1+IV_2.
\end{aligned}
$$

IV_2 is $o(1)$ on $\mathfrak{K}$ by the consistency established in part (i). The object of interest is the largest r_n for which IV_1 is $O(1)$ for a fixed $M<\infty$. By Markov's inequality

$$
\begin{aligned}
IV_1 &= \sum_{j\geq j_0:2^{j-1}<r_n\eta}\Pr\left(\sup_{\Lambda\in S_{jn}}\frac{1}{\sqrt{n}}\left(\mathbb{G}_n(\Lambda)-\mathbb{G}_n(\Lambda^0)\right)>\frac{C2^{2j-2}}{r_n^2}\right) \\
&\leq \sum_{j\geq j_0:2^{j-1}<r_n\eta}\frac{\mathbb{E}\sup_{\Lambda\in S_{jn}}r_n^2\left|\mathbb{G}_n(\Lambda)-\mathbb{G}_n(\Lambda^0)\right|}{C\sqrt{n}2^{2j-2}},
\end{aligned}
$$

which requires a bound on the expected modulus of continuity of the empirical process $\mathbb{G}_n(\Lambda):=\sqrt{n}(P_n-P)L(\Lambda)$ over the $\Lambda\in\Gamma$ such that $\|\Lambda-\Lambda^0\|_2<2^jr_n^{-1}$. let $\dot{\mathbb{G}}_n(\Lambda)$ denote the gradient of $\mathbb{G}_n$ at Λ, and $\dot{\mathbb{G}}_n^s(\Lambda)$ the s^{th} element. Then

$$
\begin{aligned}
\dot{\mathbb{G}}_n^s(\Lambda)=\frac{1}{n}\sum_{i=1}^n & \frac{q_s^{p/2}Z_{i,s}\exp\left\{-\frac{1}{2}Y_i^T(\Omega/q_s)^{-1}Y_i\right\}}{\sum_{s=1}^M\Lambda_sq_s^{p/2}Z_{i,s}\exp\left\{-\frac{1}{2}Y_i^T(\Omega/q_s)^{-1}Y_i\right\}} \\
& - \frac{q_s^{p/2}\mathbb{E}\left[Z_{i,s}\exp\left\{-\frac{1}{2}Y_i^T(\Omega/q_s)^{-1}Y_i\right\}\right]}{\sum_{s=1}^M\Lambda_sq_s^{p/2}\mathbb{E}\left[Z_{i,s}\exp\left\{-\frac{1}{2}Y_i^T(\Omega/q_s)^{-1}Y_i\right\}\right]}.
\end{aligned}
\tag{7.12}
$$

The class of functions,

$$
\mathcal{G}_s:=\left\{g:q_s\mapsto\psi(y|q_s):\ q_s\in[q_1,q_M],\ y\in\mathbb{R}^p\right\},
$$

is bounded Lipschitz so, since $q_M<\infty$, $q_s^{p/2}V_{i,s}\exp\left\{-\frac{1}{2}Y_i^T(\Omega/q_s)^{-1}Y_i\right\}$ and its expectation are bounded. Furthermore, since

$$
\sum_{s=1}^M\Lambda_sq_s^{p/2}Z_{i,s}\exp\left\{-\frac{1}{2}Y_i^T(\Omega/q_s)^{-1}Y_i\right\}\geq\sup_s\Lambda_sq_s^{p/2}Z_{i,s}\exp\left\{-\frac{1}{2}Y_i^T(\Omega/q_s)^{-1}Y_i\right\}>0\ ,
$$

the supremum of equation (7.12) is bounded by some constant K, leading to

$$\begin{aligned} IV_1 &\leq \sum_{j\geq j_0: 2^{j-1}<r_n\eta} \frac{\mathbb{E}\sup_{\Lambda\in S_{jn}} r_n^2 \dot{\mathbb{G}}_n^T(\bar{\Lambda})\|\Lambda-\Lambda^0\|}{C\sqrt{n}2^{2j-2}} \\ &\leq \sum_{j\geq j_0: 2^{j-1}<r_n\eta} \frac{\mathbb{E}\sup_{\Lambda\in S_{jn}} r_n^2 M \max_{1\leq s\leq M} \dot{\mathbb{G}}_n^s(\bar{\Lambda}_s)\,|\Lambda_s-\Lambda_s^0|}{C\sqrt{n}2^{2j-2}} \\ &\leq \sum_{\sum_{l\in\mathbb{N}: 2^{l+J}\leq r_n\eta}} \frac{MKr_n^2 2^{l+J} r^{-1}}{C\sqrt{n}2^{2(l+J)-2}} = O\left(\frac{MK(1-2^2)r_n}{C2^{2J-2}\sqrt{n}}\right), \end{aligned}$$

where $\bar{\Lambda}$ lies in the convex hull of Λ and Λ^0. The above display is $O(1)$ with $r_n=\sqrt{n}$ for $J\to\infty$ and $M<\infty$, which proves the claim. □

Lemma 4 (ULLN for the Hessian)**.** *Under Conditions 1-3,*

$$\|\nabla_\Lambda\nabla_\Lambda L_n(\Lambda) - \mathbb{E}\nabla_\Lambda\nabla_\Lambda L_n(\Lambda)\| = O_p(n^{-1/2}).$$

Proof. By Compactness of Γ, introduce a finite cover for Γ, $\{\mathcal{N}(\Lambda^{(j)},\delta),\ j=1,\ldots,J\}$, $\Lambda^{(j)}\in\Gamma$. Introduce also

$$H_n^{(l,r)}(\Lambda) = [\nabla_\Lambda\nabla_\Lambda L_n(\Lambda)]_{l,r} - \mathbb{E}[\nabla_\Lambda\nabla_\Lambda L_n(\Lambda)]_{l,r},$$

where $[\nabla_\Lambda\nabla_\Lambda L_n(\Lambda)]_{l,r}$ is the $(l,r)^{th}$ element of the Hessian, given by

$$[\nabla_\Lambda\nabla_\Lambda L_n(\Lambda)]_{l,r} = -\frac{1}{n}\sum_{i=1}^n \frac{\psi(Y_i|q_l)\psi(Y_i|q_r)}{[\sum_k \Lambda_k\psi(Y_i|q_k)]^2}.$$

We use a familiar argument based on pointwise convergence and stochastic equicontinuity, stated here for ease of reference.

$$\begin{aligned} &\Pr\left(\sup_{\Lambda\in\Gamma}|[\nabla_\Lambda\nabla_\Lambda L_n(\Lambda)]_{l,r} - \mathbb{E}[\nabla_\Lambda\nabla_\Lambda L_n(\Lambda)]_{l,r}| > \frac{2C}{\sqrt{n}}\right) \\ \leq\ &\Pr\left(\max_{j\in\{1,\ldots,J\}}\sup_{\Lambda\in\mathcal{N}(\Lambda^{(j)},\delta)}\left(\left|H_n^{(l,r)}(\Lambda) - H_n^{(l,r)}(\Lambda^{(j)})\right| + \left|H_n^{(l,r)}(\Lambda^{(j)})\right|\right) > \frac{2C}{\sqrt{n}}\right) \\ \leq\ &\Pr\left(\max_{j\in\{1,\ldots,J\}}\sup_{\Lambda\in\mathcal{N}(\Lambda^{(j)},\delta)}\left|H_n^{(l,r)}(\Lambda) - H_n^{(l,r)}(\Lambda^{(j)})\right|\frac{C}{\sqrt{n}}\right) + \Pr\left(\bigcup_{j=1}^J\left\{\left|H_n^{(l,r)}(\Lambda^{(j)})\right| > \frac{C}{\sqrt{n}}\right\}\right) \\ \leq\ &\Pr\left(\max_{j\in\{1,\ldots,J\}}\sup_{\Lambda\in\mathcal{N}(\Lambda^{(j)},\delta)}\left|H_n^{(l,r)}(\Lambda) - H_n^{(l,r)}(\Lambda^{(j)})\right|\frac{C}{\sqrt{n}}\right) + \sum_{j=1}^J\Pr\left(\left|H_n^{(l,r)}(\Lambda^{(j)})\right| > \frac{C}{\sqrt{n}}\right). \end{aligned}$$

Hence the problem is one of showing:

(i) $\exists\ 0 < C < \infty$ such that $\sum_{j=1}^{J} \Pr\left(\left|H_n^{l,r}(\Lambda^{(j)})\right| > \frac{C}{\sqrt{n}}\right) < \epsilon\ \forall\ \epsilon > 0$;

(ii) $\exists\ 0 < C < \infty$ such that, as $\delta = \delta_n \searrow 0$ at rate $\sqrt{n}$,

$$\Pr\left(\sup_{\substack{\Lambda,\Lambda' \in \Gamma: \\ \|\Lambda - \Lambda'\| \leq \delta_n}} \left|H_n^{(l,r)}(\Lambda) - H_n^{(l,r)}(\Lambda')\right| > \frac{C}{\sqrt{n}}\right) < \epsilon\ \ \forall\ \epsilon > 0.$$

For pointwise convergence (i), Markov's inequality and the fact that $(Y_i)_{i=1}^n$ are i.i.d. implies

$$\begin{aligned} & \Pr\left(\left|[\nabla_\Lambda \nabla_\Lambda L_n(\Lambda)]_{l,r} - \mathbb{E}[\nabla_\Lambda \nabla_\Lambda L_n(\Lambda)]_{l,r}\right| > \frac{C}{\sqrt{n}}\right) \\ \leq\ & \frac{n\mathbb{E}\left\{\left|[\nabla_\Lambda \nabla_\Lambda L_n(\Lambda)]_{l,r} - \mathbb{E}[\nabla_\Lambda \nabla_\Lambda L_n(\Lambda)]_{l,r}\right|^2\right\}}{C^2} \\ =\ & \frac{n\mathbb{E}\left\{\left|h^{(l,r)}(Y_i;\Lambda) - \mathbb{E}h^{(l,r)}(Y_i;\Lambda)\right|^2\right\}}{nC^2}, \end{aligned}$$

where

$$h^{(l,r)}(Y_i;\Lambda) = -\frac{\psi(Y_i|q_l)\psi(Y_i|q_r)}{[\sum_k \Lambda_k \psi(Y_i|q_k)]^2}.$$

By Lemma 1, $h^{(l,r)}(Y_i;\Lambda) := h^{(l,r)}(Y_i;q_l,q_r,\Lambda)$ is bounded uniformly over $y \in D$ and $q_l, q_r \in \mathcal{Q}$, hence (i) is true. For (ii), we first note that, by the mean value theorem, for any $\Lambda, \Lambda' \in \Gamma$, there exists a $\bar{\Lambda}$ in the convex hull of $\{\Lambda, \Lambda'\}$ such that

$$h^{(l,r)}(Y_i,\Lambda) - h^{(l,r)}(Y_i,\Lambda') = \left[\nabla_\Lambda h^{(l,r)}(Y_i,\bar{\Lambda})\right]^T (\Lambda - \Lambda').$$

Averaging, taking absolute values of both sides and applying the Cauchy-Schwarz inequality gives

$$\left|n^{-1}\sum_{i=1}^n h^{(l,r)}(Y_i,\Lambda) - n^{-1}\sum_{i=1}^n h^{(l,r)}(Y_i,\Lambda')\right| \leq \sup_{\Lambda\in\Gamma}\left\|n^{-1}\sum_{i=1}^n \nabla_\Lambda h^{(l,r)}(Y_i,\bar{\Lambda})\right\| \|\Lambda - \Lambda'\|.$$

An analogous bound applies to $|\mathbb{E}h^{(l,r)}(Y_i,\Lambda) - \mathbb{E}h^{(l,r)}(Y_i,\Lambda')|$ and we have

$$\begin{aligned}
&\Pr\left(\sup_{\substack{\Lambda,\Lambda'\in\Gamma:\\ \|\Lambda-\Lambda'\|\leq\delta_n}} \left|H_n^{l,r}(\Lambda) - H_n^{l,r}(\Lambda')\right| \frac{C}{\sqrt{n}}\right) \\
\leq\ &\Pr\left(\sup_{\substack{\Lambda,\Lambda'\in\Gamma:\\ \|\Lambda-\Lambda'\|\leq\delta_n}} \left|n^{-1}\sum_{i=1}^n h^{(l,r)}(Y_i,\Lambda) - n^{-1}\sum_{i=1}^n h^{(l,r)}(Y_i,\Lambda')\right| > \frac{C}{2\sqrt{n}}\right) \\
+\ &\mathbb{1}\left\{\sup_{\substack{\Lambda,\Lambda'\in\Gamma:\\ \|\Lambda-\Lambda'\|\leq\delta_n}} \left|\mathbb{E}h^{(l,r)}(Y_i,\Lambda) - \mathbb{E}h^{(l,r)}(Y_i,\Lambda')\right| > \frac{C}{2\sqrt{n}}\right\} \\
\leq\ &\Pr\left(\sup_{\substack{\Lambda,\Lambda'\in\Gamma:\\ \|\Lambda-\Lambda'\|\leq\delta_n}} B_n\|\Lambda-\Lambda'\| > \frac{C}{2\sqrt{n}}\right) + \mathbb{1}\left\{\sup_{\substack{\Lambda,\Lambda'\in\Gamma:\\ \|\Lambda-\Lambda'\|\leq\delta_n}} B\|\Lambda-\Lambda'\| > \frac{C}{2\sqrt{n}}\right\},
\end{aligned}$$

where

$$B_n = \sup_{\Lambda\in\Gamma}\left\|n^{-1}\sum_{i=1}^n \nabla_\Lambda h^{(l,r)}(Y_i,\bar{\Lambda})\right\| \quad \text{and} \quad B = \sup_{\Lambda\in\Gamma}\left\|\nabla_\Lambda\left[\mathbb{E}g^{(l,r)}(Y_i,\bar{\Lambda})\right]\right\|.$$

Letting $\delta_n \searrow 0$ at rate $\sqrt{n}$, we require, for the existence of a $0 < C < \infty$ such that

$$\Pr\left(\sup_{\substack{\Lambda,\Lambda'\in\Gamma:\\ \|\Lambda-\Lambda'\|\leq\delta_n}} B_n\|\Lambda-\Lambda'\| > \frac{C}{2\sqrt{n}}\right) < \frac{\epsilon}{2}\ \ \forall\epsilon > 0,$$

that such a C ensures $\Pr(B_n > C/2) < \epsilon/2\ \forall\epsilon > 0$. Let $\xi^{l,r}(Y_i;\Lambda)$ be a vector with s^{th} element

$$[\xi^{(l,r)}(Y_i;\Lambda)]_s = \frac{2\psi(Y_i|q_l)\psi(Y_i|q_r)\psi(Y_i|q_s)}{[\sum_k|\Lambda_k\psi(Y_i|q_k)]^3}.$$

By Markov's inequality

$$\begin{aligned}
&\Pr(B_n > C/2) \\
\leq\ &\frac{4\mathbb{E}\sup_{\Lambda\in\Gamma}\left\|\frac{1}{n^2}[\sum_{i=1}^n \xi^{(l,r)}(Y_i;\Lambda)^2 + 2\sum_{i=1}^n\sum_{j<i}\xi^{(l,r)}(Y_i;\Lambda)\xi^{(l,r)}(Y_j;\Lambda)]\right\|}{C^2} \\
\leq\ &\frac{4\mathbb{E}\frac{1}{n^2}\sup_{\Lambda\in\Gamma}\left\|\sum_{i=1}^n \xi^{(l,r)}(Y_i;\Lambda)^2\right\| + 2\sum_{i=1}^n\sum_{j<i}\sup_{\Lambda\in\Gamma}\left\|\xi^{(l,r)}(Y_i;\Lambda)\xi^{(l,r)}(Y_j;\Lambda)]\right\|}{C^2} \\
&\text{[By Jensen's inequality]} \\
=\ &\frac{4}{C^2}\left[\frac{1}{n}\mathbb{E}\sup_{\Lambda\in\Gamma}\left\|\sum_{i=1}^n \xi^{(l,r)}(Y_i;\Lambda)^2\right\| + \frac{n(n-1)}{n^2}\sup_{\Lambda\in\Gamma}\left\|\xi^{(l,r)}(Y_i;\Lambda)\xi^{(l,r)}(Y_j;\Lambda)\right\|\right] \\
&\longrightarrow \mathbb{E}\sup_{\Lambda\in\Gamma}\left\|\xi^{(l,r)}(Y_i;\Lambda)\xi^{(l,r)}(Y_j;\Lambda)\right\| \quad \text{as } n\to\infty.
\end{aligned}$$

Existence of a finite C for which

$$\mathbb{1}\left\{\sup_{\substack{\Lambda,\Lambda'\in\Gamma:\\ \|\Lambda-\Lambda'\|\leq\delta_n}} B\|\Lambda-\Lambda'\| > \frac{C}{2\sqrt{n}}\right\} = 0$$

is ensured by the boundedness of B. This establishes (ii). □

The following lemma, which establishes the rate of convergence of the arithmetic mean to the geometric mean, is standard (see e.g. de Leeuw, 2011). The proof is reproduced here for convenience.

Lemma 5. *Let $(a)_{i=1}^m$ denote a collection of numbers and define,*

$$a_m := \frac{1}{m}\sum_{i=1}^m a_i \quad \textit{and} \quad g_m = [\prod_{i=1}^m a_i]^{1/m} .$$

a_m and g_m converge to the same limit at rate 2^m.

Proof.

$$x_{m+1} := \frac{a_{m+1}}{g_{m+1}} = \frac{a_m+g_m}{2\sqrt{a_m g_m}} = \frac{1}{2}\left(\sqrt{\frac{a_m}{g_m}}+\sqrt{\frac{g_m}{a_m}}\right) = \frac{1}{2}\left(\sqrt{x_m}+1/\sqrt{x_m}\right)$$

By the geometric-arithmetic mean inequality, we have $x_m \geq 1$, so $1/\sqrt{x_m} \leq 1$ and $\sqrt{x_m} \leq x_m$. Hence

$$x_{m+1}-1 = \frac{1}{2}\left(\sqrt{x_m}+\frac{1}{\sqrt{x_m}}-2\right) \leq \frac{1}{2}(x_m-1)$$

so

$$0 \leq (x_{m+1}-1) \leq \frac{1}{2}(x_m-1),$$

and the recursion gives

$$0 \leq (x_{m+1}-1) \leq \frac{1}{2^m}(x_0-1);$$

i.e. $x_m \to 1$ at rate 2^m. □

Corollary 2. *When η is chosen as $2r\sqrt{(2\ln M)/I}$, where r is the lower bound on the instances $\psi(Y_i|q_s)$, introduced in Theorem 1 of Helmbold et al. (1997),*

$$|\widehat{\Omega}|^{-1/2}\sup_{y\in D}\left|\sum_{s=1}^M \Lambda_s^0\psi(y|q_s) - \sum_{s=1}^M \widehat{\Lambda}_{I,s}^{EG}\psi(y|q_s)\right| = O_p\left(\frac{M}{\sqrt{n}}\right).$$

where $I = n$. We may also consider the asymptotics as we allow $r \searrow 0$. This yields the same rate if r is of order at most $\frac{\sqrt{2\ln M}}{2M\sqrt{n}}$ and yields

$$|\widehat{\Omega}|^{-1/2}\sup_{y\in D}\left|\sum_{s=1}^M \Lambda_s^0\psi(y|q_s) - \sum_{s=1}^M \widehat{\Lambda}_{I,s}^{EG}\psi(y|q_s)\right| = O_p\left(\frac{\sqrt{2\ln M}}{(2r)n}\right)$$

if r converges to zero faster.

Proof.

$$|\widehat{\Omega}|^{-1/2} \sup_{y\in D} \left| \sum_{s=1}^{M} \Lambda_s^0 \psi(y|q_s) - \sum_{s=1}^{M} \widehat{\Lambda}_{I,s}^{EG} \psi(y|q_s) \right| \leq |\widehat{\Omega}|^{-1/2} \left[\max_{s\in\{1\ldots,M\}} \sup_{y\in D} \psi(y|q_s) \right] \sum_{s=1}^{M} |\Lambda_s^0 - \widehat{\Lambda}_{I,s}^{EG}|.$$

The term in parenthesis is bounded by Lemma 1. For $\sum_{s=1}^{M} |\Lambda_s^0 - \widehat{\Lambda}_{I,s}^{EG}|$, consider, by the triangle inequality

$$\|\widehat{\Lambda}_I^{EG} - \Lambda^0\|_1 \leq \|\widehat{\Lambda}_I^{EG} - \widehat{\Lambda}^{ML}\|_1 + \|\widehat{\Lambda}^{ML} - \Lambda^0\|_1, \tag{7.13}$$

where $\|\cdot\|_1$ is the l_1-norm. An application of Lemma 3 gives $\|\widehat{\Lambda}^{ML} - \Lambda^0\|_1 = O_p(M/\sqrt{n})$. For the first term, mean value expand the partial derivative vector of the log likelihood around $\widehat{\Lambda}^{ML}$ to give

$$\nabla_\Lambda L_n(\widehat{\Lambda}_I^{EG}) = \nabla_\Lambda L_n(\widehat{\Lambda}^{ML}) + \nabla_\Lambda \nabla_\Lambda L_n(\bar{\Lambda})(\widehat{\Lambda}_I^{EG} - \widehat{\Lambda}^{ML}),$$

where $\bar{\Lambda}$ lies in the convex hull of $\widehat{\Lambda}_I^{EG}$ and $\widehat{\Lambda}^{ML}$. Inverting,

$$(\widehat{\Lambda}_I^{EG} - \widehat{\Lambda}^{ML}) = \left[\nabla_\Lambda \nabla_\Lambda L_n(\bar{\Lambda})\right]^{-1} \left(\nabla_\Lambda L_n(\widehat{\Lambda}_I^{EG}) - \nabla_\Lambda L_n(\widehat{\Lambda}^{ML}) \right),$$

and Chebyshev's inequality applied to

$$\left| \nabla_\Lambda L_n(\widehat{\Lambda}^{ML}) - \mathbb{E}\nabla_\Lambda L_n(\widehat{\Lambda}^{ML}) \right| \quad \text{and} \quad \left| \nabla_\Lambda L_n(\widehat{\Lambda}_I^{EG}) - \mathbb{E}\nabla_\Lambda L_n(\widehat{\Lambda}_I^{EG}) \right|$$

implies that, elementwise for all $s \in \{1, \ldots, M\}$,

$$\begin{aligned}
& \left| \nabla_\Lambda L_n(\widehat{\Lambda}_I^{EG}) - \nabla_\Lambda L_n(\widehat{\Lambda}^{ML}) \right| \\
= & \left| \nabla_\Lambda \mathbb{E} L_n(\widehat{\Lambda}_I^{EG}) - \nabla_\Lambda \mathbb{E} L_n(\widehat{\Lambda}^{ML}) \right| + O_p(n^{-1/2}) \\
= & \left| \nabla_\Lambda \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^{n} \ln \left(\frac{1}{I} \sum_{j=1}^{I} \sum_{k} \widehat{\Lambda}_{k,j}^{EG} \psi(Y_i|q_k) \right) \right) - \nabla_\Lambda \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^{n} \ln \left(\sum_{k} \widehat{\Lambda}_k^{ML} \psi(Y_i|q_k) \right) \right) \right| \\
& \qquad + O_p(n^{-1/2}) \\
= & \left| \nabla_\Lambda \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^{n} \ln \left[\prod_{j=1}^{I} \sum_{k} \widehat{\Lambda}_{k,j}^{EG} \psi(Y_i|q_k) \right]^{1/I} \right) - \nabla_\Lambda \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^{n} \ln \left(\sum_{k} \widehat{\Lambda}_k^{ML} \psi(Y_i|q_k) \right) \right) \right| \\
& \qquad + O_p(n^{-1/2}) + O_p(2^{-I}) \\
= & \left| \nabla_\Lambda \mathbb{E} \left[\frac{1}{I} \sum_{j=1}^{I} L_n(\widehat{\Lambda}_j^{EG}) \right] - \nabla_\Lambda \mathbb{E} L_n(\widehat{\Lambda}^{ML}) \right| + O_p(n^{-1/2}) + O_p(2^{-I}),
\end{aligned}$$

where the penultimate line follows by Lemma 5. Equation (3.6) and the negative semi-definiteness of the Hessian imply that, elementwise for all $s \in \{1, \ldots, M\}$,

$$\left| \nabla_\Lambda \mathbb{E} \left[\frac{1}{I} \sum_{j=1}^{I} L_n(\widehat{\Lambda}_j^{EG}) \right] - \nabla_\Lambda \mathbb{E} L_n(\widehat{\Lambda}^{ML}) \right| = O\left(\frac{\sqrt{2 \ln M}}{2r\sqrt{I}} \right),$$

where $I = n$ and r is a lower bound on the random variables $\psi(Y_i, q_s)$ introduced by Helmbold et al. (1997) (see Theorem 1 op. cit.). By Lemma 4 along with the continuity of the inverse, which ensures the applicability of Slutsky's theorem, $\left[\nabla_\Lambda \nabla_\Lambda L_n(\bar{\Lambda})\right]^{-1}$ converges to some finite constant at rate $\sqrt{n}$. We have

$$\begin{aligned}(\widehat{\Lambda}_I^{EG} - \widehat{\Lambda}^{ML}) &= \left[\nabla_\Lambda \nabla_\Lambda L_n(\bar{\Lambda})\right]^{-1} \left(\nabla_\Lambda L_n(\widehat{\Lambda}_I^{EG}) - \nabla_\Lambda L_n(\widehat{\Lambda}^{ML})\right) \\ &= O_p\left(\frac{1}{\sqrt{n}}\right) O_p\left(\max\left\{\frac{1}{\sqrt{n}}, \frac{\sqrt{2\ln M}}{2r\sqrt{I}}\right\}\right) = O_p\left(\frac{\sqrt{2\ln M}}{(2r)n}\right)\end{aligned}$$

since $I = n$. □

We conclude the proof of Theorem 1 by returning to equation (7.1). We see from the controls over II_1, II_2 and II_3 that II_3 dominates in the bound on I_2. Equalising the antagonistic approximation and estimation error terms and solving for M delivers the rate in Theorem 1

□

Acknowledgements: This work was partially funded by the ESRC. We thank Alessio Sancetta and Alexei Onatski for helpful discussions, as well as two anonymous referees, whose comments helped improve the paper.

References

Abadir, K. M. and Magnus, J. R. (2005). *Matrix algebra*, vol. 1 of *Econometric Exercises*. Cambridge University Press, Cambridge.

Abdul-Hamid, H. and Nolan, J. P. (1998). Multivariate stable densities as functions of one-dimensional projections. *J. Multivariate Anal.* **67** 80–89.

Abramson, I. S. (1982). On bandwidth variation in kernel estimates—a square root law. *Ann. Statist.* **10** 1217–1223.

Arnold, D. and Beyer, H.-G. (2003). On the effects of outliers on evolutionary optimisation. In *Intelligent Data Engineering and Automated Learning – IDEAL 2003*, vol. 2690 of *Lecture notes in computer science*. Springer, New York, 151–160.

Beale, E. M. L. and Mallows, C. L. (1959). Scale mixing of symmetric distributions with zero means. *Ann. Math. Statist.* **30** 1145–1151.

Berk, J. (1997). Necessary conditions for the CAPM. *Journal of Economic Theory* **73** 245 – 257.

Cambanis, S., Huang, S. and Simons, G. (1981). On the theory of elliptically contoured distributions. *J. Multivariate Anal.* **11** 368–385.

Cappé, O. and Moulines, E. (2009). Online EM algorithm for latent data models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 593–613.

Chmielewski, M. A. (1981). Elliptically symmetric distributions: a review and bibliography. *Internat. Statist. Rev.* **49** 67–74.

Cule, M. and Samworth, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Stat.* **4**.

Cule, M., Samworth, R. and Stewart, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *J. R. Stat. Soc. Ser. B Stat. Methodol.* .

de Leeuw, J. (2011). Rate of convergence for the arithmetic-geometric mean process. Tech. rep.
URL http://statistics.ucla.edu/preprints/uclastat-preprint-2008:20

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. With discussion.

Dudley, R. M. (2002). *Real analysis and probability*, vol. 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge. Revised reprint of the 1989 original.

Duong, T. (2007). *Package 'ks': Kernel smoothing*. R package version 1.4.11.
URL http://cran.r-project.org/web/packages/ks/index.html

Fang, K.-T., Kotz, S. and Ng, K. (1990). *Symmetric Multivariate and Related Distributions*, vol. 36 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Fix, E. and Hodges, J. (1951). Discriminatory analysis, nonparametric discrimination: consistency properties. Tech. rep.

Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28** 1105–1127.

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263.

Helmbold, D. P., Schapire, R. E., Singer, Y. and Warmuth, M. K. (1997). A comparison of new and old algorithms for a mixture estimation problem. In *Machine Learning*.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3** 1163–1174.

Holzmann, H., Munk, A. and Gneiting, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scand. J. Statist.* **33** 753–763.

Kano, Y. (1994). Consistency property of elliptical probability density functions. *J. Multivariate Anal.* **51** 139–147.

Kariya, T. and Eaton, M. L. (1977). Robust tests for spherical symmetry. *Ann. Statist.* **5** 206–215.

Liebscher, E. (2005). A semiparametric density estimator based on elliptical distributions. *J. Multivariate Anal.* **92** 205–225.

Lindskog, F., McNeil, A. and Schmock, U. (2003). Kendall's tau for elliptical distributions. In *Credit Risk. Measurement, Evaluation and Management*. Springer-Verlag, Heidelberg, 149–156.

Maronna, R. A. and Yohai, V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *J. Amer. Statist. Assoc.* **90** 330–341.

Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** 712–736.

Marsh, P. (2007). Constructing optimal tests on a lagged dependent variable. *Journal of Time Series Analysis* **28** 723–743.

Meng, X.-L. and van Dyk, D. (1997). The EM algorithm—an old folk-song sung to a fast new tune. *J. Roy. Statist. Soc. Ser. B* **59** 511–567. With discussion and a reply by the authors.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Nolan, J. P. (1997). Numerical calculation of stable densities and distribution functions. *Comm. Statist. Stochastic Models* **13** 759–774. Heavy tails and highly volatile phenomena.

Owen, J. and Rabinovitch, R. (1983). On the class of elliptical distributions and their applications to the theory of portfolio choice. *The Journal of Finance* **38**.

Parthasarathy, K. R. (1967). *Probability measures on metric spaces*. Probability and Mathematical Statistics, No. 3, Academic Press Inc., New York.

Pollard, D. (2000). Linear correlation estimation.

Rudolph, G. (1997). Local convergence rates of simple evolutionary algorithms with cauchy mutations. *IEEE Transactions on Evolutionary Computation* **1** 249 –258.

Sain, S. R. (2002). Multivariate locally adaptive density estimation. *Comput. Statist. Data Anal.* **39** 165–186.

Sain, S. R. and Scott, D. W. (1996). On locally adaptive density estimation. *J. Amer. Statist. Assoc.* **91** 1525–1534.

Sancetta, A. (2009). Bayesian semiparametric estimation of elliptic densities. URL http://www.sancetta.googlepages.com/academicpublications

Scott, D. W. and Sain, S. R. (2005). Multidimensional density estimation. In *Data Mining and Data Visualization* (E. W. C.R. Rao and J. Solka, eds.), vol. 24 of *Handbook of Statistics*. Elsevier, 229 – 261.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

Street, W. N., Wolberg, W. H. and Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis .

Stute, W. and Werner, U. (1991). Nonparametric estimation of elliptically contoured densities. In *Nonparametric functional estimation and related topics (Spetses, 1990)*, vol. 335 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* Kluwer Acad. Publ., Dordrecht, 173–190.

van de Geer, S. A. (2000). *Applications of empirical process theory*, vol. 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. (1998). *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics, Springer-Verlag, New York. With applications to statistics.

Veillette, M. (2012). *STBL: Alpha stable distributions for MATLAB*. Matlab Central File Exchange, retreived October 10, 2012.
URL http://www.mathworks.com/matlabcentral/fileexchange/37514

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*, vol. 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London.

Zhao, L., Bai, Z., Chao, C.-C. and Liang, W.-Q. (1997). Error bound in a central limit theorem of double-indexed permutation statistics. *Ann. Statist.* **25** 2210–2227.