A mixture model for multivariate extremes

M.-O. Boldi and A. C. Davison

Ecole Polytechnique Fédérale de Lausanne, Switzerland

Summary. The spectral density function plays a key role in fitting the tail of multivariate extremal data and so in estimating probabilities of rare events. This function satisfies moment constraints but unlike the univariate extreme value distributions has no simple parametric form. Parameterized subfamilies of spectral densities have been suggested for use in applications, and non-parametric estimation procedures have been proposed, but semiparametric models for multivariate extremes have hitherto received little attention. We show that mixtures of Dirichlet distributions satisfying the moment constraints are weakly dense in the class of all non-parametric spectral densities, and discuss frequentist and Bayesian inference in this class based on the EM algorithm and reversible jump Markov chain Monte Carlo simulation. We illustrate the ideas using simulated and real data.

Keywords: Adequacy; Air pollution data; Dirichlet distribution; EM algorithm; Multivariate extreme values; Oceanographic data; Reversible jump Markov chain Monte Carlo simulation; Spectral distribution

1. Introduction

The modelling of multivariate extreme values is increasingly important in applications ranging from finance to environmental science, and from oceanography to material science (Coles and Tawn, 1994; Heffernan and Tawn, 2004; Coles, 2001). A key aspect is the estimation of the joint tail of a multivariate distribution, which depends on the so-called spectral distribution function. This is reviewed by Beirlant *et al.* (2004), chapters 8 and 9, and Kotz and Nadarajah (2000), chapter 3. Both parametric models for the spectral distribution function and non-parametric approaches to estimating it have been suggested. However, most parametric models are insufficiently flexible for accurate modelling of tail behaviour in several dimensions, and non-parametric approaches have been developed only for two- or three-dimensional data. The curse of dimensionality is particularly burdensome in the context of extremes, because the analysis entails extrapolation beyond the sample tail in more than one dimension, and so there seems little hope that a fully non-parametric approach can be entirely successful. In this paper we propose a middle path—a semiparametric model based on mixtures of Dirichlet distributions—show that it may approximate the full class of non-parametric models arbitrarily well in the weak sense, discuss how it may be fitted to data and illustrate its applicability. A much fuller account may be found in Boldi (2004).

The spectral distribution function measures the dependence among extremes, the underlying theory for which may be based on two related convergence results. Let $X, X_1, \ldots, X_n \in \mathbb{R}^p$ be independent identically distributed random vectors whose distribution function F has unit Fréchet marginal distributions, i.e. $\exp(-1/x)$ for $x > 0$. If the renormalized maximum

$M_n = n^{-1} \max(X_1, \ldots, X_n)$, where here and below comparisons among vectors are done componentwise, converges in distribution to a non-degenerate distribution function G, then this has the form

$$G(x) = \lim_{n \to \infty} \{P(M_n \leqslant x)\} = \exp\left\{ -p \int_{\mathcal{S}_p} \max(w_1/x_1, \ldots, w_p/x_p) \, H(\mathrm{d}w) \right\}. \tag{1}$$

Here H is the spectral distribution function, a probability measure on the p-dimensional unit simplex $\mathcal{S}_p = \{w \in \mathbb{R}_+^p : \Sigma_{j=1}^p w_j = 1\}$ which satisfies the mean constraints

$$\int_{\mathcal{S}_p} w_j \, H(\mathrm{d}w) = p^{-1}, \qquad j = 1, \ldots, p. \tag{2}$$

These constraints are intrinsic to the model: as each of the X_j has a marginal unit Fréchet distribution, letting $x_i \to \infty$ for $i \neq j$ in equation (1) produces equation (2). Analogous constraints would be produced by transformation to any other set of known marginal distributions.

An alternative formulation may be based on the sample process $\{n^{-1}\delta_{X_j}\}_{j=1}^n$, where δ_x denotes the Dirac mass at $x \in \mathbb{R}^p$. The sample process converges in distribution as $n \to \infty$ to a Poisson process on $[0, \infty)^p \setminus \{0\}$ with intensity measure

$$\Lambda(\mathrm{d}x) = \frac{1}{r^2} \mathrm{d}r \times p \, H(\mathrm{d}w), \tag{3}$$

where we use the pseudopolar co-ordinate system (r, w). A datum $x = (x_1, \ldots, x_p)$ is extreme if its pseudoradius $r = \Sigma_{j=1}^p x_j$ is large, in which case the pseudoangle $w = x/r$—the relative attribution of r among the components of x—is independent of r and distributed according to H.

In practice these asymptotic results are applied by fitting a suitable distribution to the margins of an original data set, using this fit to transform the data margins to the unit Fréchet scale and making a further transformation to the pseudopolar scale; this yields $(r_1, w_1), \ldots, (r_n, w_n)$. A high threshold r_0 is selected, the set $\mathcal{I}_0 = \{i : r_i > r_0\}$ is defined and a Poisson process with intensity (3) is fitted to pairs $\{(r_i, w_i)\}_{i \in \mathcal{I}_0}$ that exceed this threshold. This last stage boils down to fitting H to $\{w_i\}_{i \in \mathcal{I}_0}$, treated as an independent random sample. The threshold r_0 is typically chosen informally, guided by probability plots based on equation (3).

The outline above pertains to models in which the degree of dependence between variables does not depend on the size of the underlying rare event. Our model does not encompass the more refined theory in which the degree of association may vary with rarity of the corresponding events has been developed by Ledford and Tawn (1996, 1997, 2003), but we shall see below that our approach can nevertheless be useful in assessing whether these more sophisticated ideas should be used in a given application.

The remainder of the paper is laid out as follows. Section 2 introduces our new model and outlines some of its properties. In Section 3 we discuss model fitting, focusing on a reversible jump Markov chain Monte Carlo algorithm for Bayesian inference, but with a few remarks on use of the EM algorithm. Section 4 gives results on simulated and real data, and Section 5 contains a brief discussion. Technical details are relegated to appendices.

2. Mixture model

The main difficulty in constructing spectral distribution functions is the need to satisfy condition (2), which imposes awkward constraints on the parameters of the model. One standard

density on the simplex $\mathcal{S}_p$ is the Dirichlet, which we parameterize in terms of the concentration parameter ν and mean vector $(\mu_1, \ldots, \mu_p)$ satisfying $\mu_1 + \ldots + \mu_p = 1$, i.e.

$$g(w; \mu, \nu) = \frac{\Gamma(\nu)}{\prod\limits_{j=1}^{p} \Gamma(\nu \mu_j)} \prod_{j=1}^{p} w_j^{\nu \mu_j - 1}, \qquad w \in \mathcal{S}_p, \mu_j, \nu > 0, \tag{4}$$

but imposition of condition (2) forces equation (4) to satisfy $\mu_1 = \ldots = \mu_p = 1/p$, leaving a symmetric distribution with just one parameter ν that is too inflexible for applications. No such difficulty applies to a mixture of k Dirichlet densities

$$
\begin{aligned}
h(w) &= \sum_{m=1}^{k} \pi_m \, g(w; \mu^{(m)}, \nu_m) \\
&= \sum_{m=1}^{k} \pi_m \frac{\Gamma(\nu_m)}{\prod\limits_{j=1}^{p} \Gamma(\nu_m \mu_j^{(m)})} \prod_{j=1}^{p} w_j^{\nu_m \mu_j^{(m)} - 1}, \qquad w \in \mathcal{S}_p,
\end{aligned}
\tag{5}
$$

where π_m represents the probability of the mth component of the mixture, and that component has mean vector $\mu^{(m)} \in \mathcal{S}_p$ with jth element $\mu_j^{(m)}$, the corresponding shape parameter being ν_m. For $j = 1, \ldots, p$ and $m = 1, \ldots, k$, we have

$$\pi_m \geqslant 0, \qquad \sum_{m=1}^{k} \pi_m = 1, \qquad \nu_m > 0, \qquad \mu_j^{(m)} \geqslant 0, \qquad \sum_{j=1}^{p} \mu_j^{(m)} = 1. \tag{6}$$

Below we use the term mixture model as shorthand for the density (5) subject to conditions (6). The constraints (2) are satisfied if

$$\sum_{m=1}^{k} \pi_m \mu_j^{(m)} = p^{-1}, \qquad j = 1, \ldots, p, \tag{7}$$

and in this case the model has $d = p(k-1) + k$ free parameters; this number varies linearly with the dimension p of the data and the number k of mixture components. For k fixed model (5) is parametric, but letting k vary yields a semiparametric model, in the sense that the parameter space has countable dimension. In practice k will typically be selected using the data.

The huge literature on mixture densities is reviewed by Redner and Walker (1984), Titterington *et al.* (1985) and McLachlan and Peel (2000). Use of Dirichlet components has several advantages: their analytical properties are simple and well known (Wilks, 1962), the family of finite mixtures of Dirichlet distributions is very rich, and its theoretical properties have been studied in detail in the context of non-parametric Bayesian inference. Any prior distribution function can be approximated in the weak sense by a sequence of finite mixtures of Dirichlet processes, so the distribution of any random probability vector can be approximated in the weak sense by a sequence of finite mixtures of Dirichlet distributions (Dalal, 1978; Dalal and Hall, 1980). In Appendix A we show that this property carries over to distribution functions on $\mathcal{S}_p$ subject to constraints (2), implying that any spectral distribution function may be weakly approximated by a mixture of Dirichlet distributions that satisfies constraints (2). Boldi (2004) gives empirical evidence for the practical usefulness of this for fairly complex simulated data sets of the sizes that are often met in applications. Moreover continuous functions of the joint tail model, such as quantiles, values at risk, and so forth, will also be approximated arbitrarily closely in the weak sense by fitting such mixtures.

3. Inference

A Bayesian approach to fitting model (5) may be based on the reversible jump algorithm (Green, 1995), which has been used extensively to fit mixtures with an unknown number of components (Richardson and Green, 1997). This algorithm generates a Markov chain whose stationary distribution is the required posterior: at each step, the algorithm proposes and accepts at random a new state for the parameters of model (5) subject to conditions (6) and (7). Three types of changes to the means $\mu^{(m)}$ and probabilities π_m are allowed: a 'split' move, according to which a component selected at random, with parameters $(\pi_0, \mu^{(0)})$ say, can be split into two other components with parameters $(\pi_1, \mu^{(1)})$ and $(\pi_2, \mu^{(2)})$, but holding the sums

$$\pi_0 \mu^{(0)} = \pi_1 \mu^{(1)} + \pi_2 \mu^{(2)},$$
$$\pi_0 = \pi_1 + \pi_2$$

fixed, a 'combine' move, which merges two components into one, under the same constraints, and a 'combine–split' move, which moves two randomly selected components while preserving their total probability and their weighted mean. The first two types of move change the dimension k of the model, whereas the third is a standard Metropolis–Hastings step. The treatment of ν_m is simpler since very few constraints must be satisfied. In each case the proposed new state is accepted with a probability given by computations in Green (1995). See Appendix B for details.

Fig. 1 shows a directed acyclic graph giving the structure of the model. The prior densities are determined by parameters $\delta_1, \ldots, \delta_k$ for the probabilities that are attached to the mixture components, λ for their number, and $\tilde{\nu}$ and σ for their shapes. Useful prior information is typically unavailable in the context of multivariate extremes, and it may be desirable to choose fairly uninformative priors; for example, setting $\delta_1 = \ldots = \delta_k = 1$ makes the prior density for π uniform on $\mathcal{S}_k$. The choice of λ can be used to penalize the complexity of the distribution; a referee has pointed out that using a uniform prior on $k \in \{1, \ldots, 20\}$ would be more tolerant of variation in k. The choice of hyperparameters for $\log(\nu)$ is more delicate and may require sensitivity analysis.
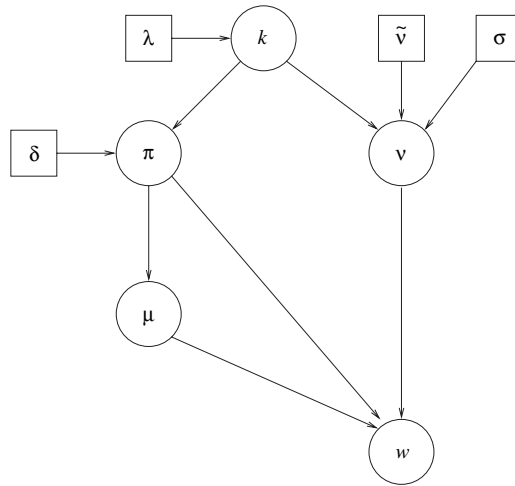

Fig. 1. Directed acyclic graph for the mixture model ($\square$, hyperparameters; $\bigcirc$, random quantities): given k, $\log(\nu)$ and π are of length k, and μ has dimension k_p ($k \sim$ Poisson(λ); $\log(\nu)|k \sim \mathcal{N}_k\{\log(\bar{\nu}), \sigma^2 I_k\}$; $\pi|k \sim$ Dirichlet$(\delta_1, \ldots, \delta_k)$; $\mu|(\pi, k) \sim F_\mu$; $w|(\nu, \pi, \mu, k) \sim \Sigma_{m-1}^k \pi_m g_m$); the distribution of F_μ incorporating the constraints linking π with μ is given in Appendix B, and g_m denotes the Dirichlet density with parameters $\mu^{(m)}$ and ν_m

Output analysis and convergence assessment are key aspects of Markov chain Monte Carlo methods (Brooks and Roberts, 1998). They are more awkward for reversible jump methods, because of the need to focus on quantities that are defined without respect to the model dimension. One such quantity is the fitted density, whose posterior mean may be estimated from the output from iterations $t = 1, \ldots, T$ by

$$\hat{E}\{h(w|k, \pi, \theta)|y\} = T^{-1} \sum_{t=1}^{T} \sum_{m=1}^{k_t} \pi_m^{(t)} h(w|\theta_m^{(t)}),$$

where k_t is the model dimension and $\pi_m^{(t)}$ and $\theta_m^{(t)}$ are the parameter values at iteration t. Another is the dependence measure between extremes given by

$$\xi = \int_{S_p} \min(w_1, \ldots, w_p) w_j \, H(\mathrm{d}w). \tag{8}$$

When $p = 2$ and under our model this equals

$$\xi = \int_0^1 \min(w, 1 - w) \, H(\mathrm{d}w)$$

$$= \sum_{m=1}^{k} \pi_m \{\mu_1^{(m)} \operatorname{beta}(0, \tfrac{1}{2}; \alpha_1^{(m)} + 1, \alpha_2^{(m)}) + \mu_2^{(m)} \operatorname{beta}(\tfrac{1}{2}, 1; \alpha_1^{(m)}, \alpha_2^{(m)} + 1)\}, \tag{9}$$

where $\alpha_j^{(m)} = \nu_m \mu_j^{(m)}$ for $j = 1, 2$, and $\operatorname{beta}(x, y; a, b)$ is the probability that a beta variable with parameters a and b lies in the interval (x, y), where $0 \leqslant x < y \leqslant 1$. The quantity ξ can be interpreted as a measure of the asymptotic dependence among the variables: it takes values $0 \leqslant \xi \leqslant \tfrac{1}{2}$, with $\xi = 0$ for independence and $\xi = \tfrac{1}{2}$ for perfect dependence.

For this study we used the convergence assessment techniques of Brooks and Giudici (2000), who proposed a three-plot diagnostic based on several independent chains with dispersed starting-points. This diagnostic compares estimates of the overall variance of a given quantity, η, say, its mean variance within models and its variance between models. If stationarity is achieved, then the between-chain and the within-chain variance estimates should have converged to a common value. The choice of η is critical since it must contain enough information to infer potential lack of convergence. In the next section we summarize numerical experiments that illustrate the performance of our fitting algorithms.

Mixture models are commonly fitted using an EM algorithm (Dempster *et al.*, 1977; Meng and Pedlow, 1992; McLachlan and Krishnan, 1997) to estimate the component parameters. Here the constraints (2) may be incorporated into the maximization step by using standard constrained optimization routines. Akaike's information criterion AIC has been used to select the number of components k for finite mixtures (McLachlan and Peel (2000), page 203). This commonly leads to overfitting in the regression context, where a corrected version such as AIC$_c$ (Hurvich and Tsai, 1989) may be preferable. The Bayesian information criterion BIC penalizes overfitting more sharply than does AIC and leads to consistent model selection when the true model lies within the fitted family (McQuarrie and Tsai, 1998; Burnham and Anderson, 2002). In the present context we take

$$\mathrm{AIC}(k) = -2\hat{l}_k + 2d,$$

$$\mathrm{AIC}_c(k) = \mathrm{AIC}(k) + 2d(d + 1)/(n - d - 1)$$

and

$$\text{BIC}(k) = -2\hat{l}_k + d \, \log(n),$$

where the number of parameters is $d = k + p(k-1)$, $\hat{l}_k$ is the maximized likelihood and n is the sample size.

4. Numerical examples

4.1. Simulated data

We first consider a data set of size $n = 500$ generated from the mixture model in dimension $p = 2$, with

$$k = 4, \qquad \pi = \begin{pmatrix} 0.5 \\ 0.25 \\ 0.125 \\ 0.125 \end{pmatrix}, \qquad \mu = \begin{pmatrix} 0.5 & 0.5 \\ 0.8 & 0.2 \\ 0.1 & 0.9 \\ 0.3 & 0.7 \end{pmatrix}, \qquad \nu = \begin{pmatrix} 0.9 \\ 20 \\ 1 \\ 50 \end{pmatrix}.$$

We fit our model using the reversible jump algorithm starting from $k = 1$, $\pi = 1$, $\nu = 1.18$ and $\mu = (0.5, 0.5)$. The output from 100 000 iterations after a 100 000-iteration burn-in period is shown in Fig. 2. The posterior median is close to the true density, which is contained in the 90% credibility interval, though the first mode lies to the right of the final estimate. The posterior distributions put mass on $k = 3, 4, 5, 6$, with corresponding posterior probabilities 0.12, 0.65, 0.22 and 0.01.

Three other chains with different initial values were also run: diagnostics based on equation (8), computed every 10 iterations, gave no evidence against convergence, though we found that mixing can be greatly improved by tuning the parameters of the algorithm.

It is natural to wonder about two causes of sensitivity to the prior specification: the prior on k, and those on the parameters conditional on k. The second is difficult to quantify because the
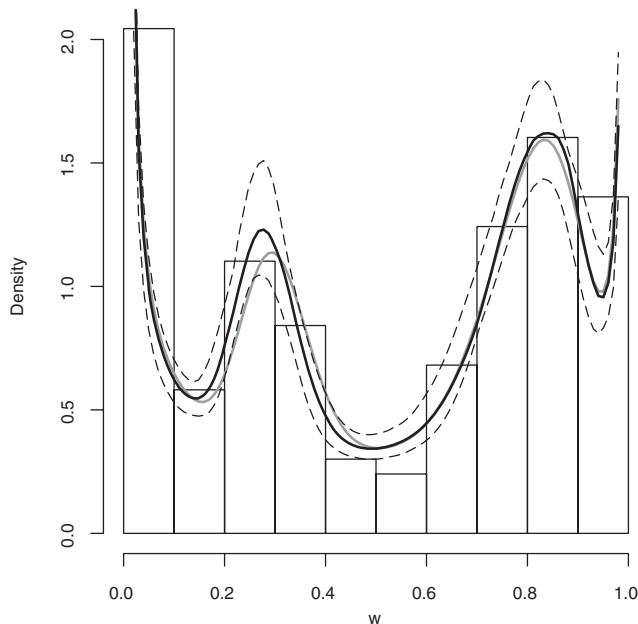

Fig. 2. Results of the reversible jump algorithm: histogram of the data, true density (———), posterior median (———) and 90% credibility interval (– – –)
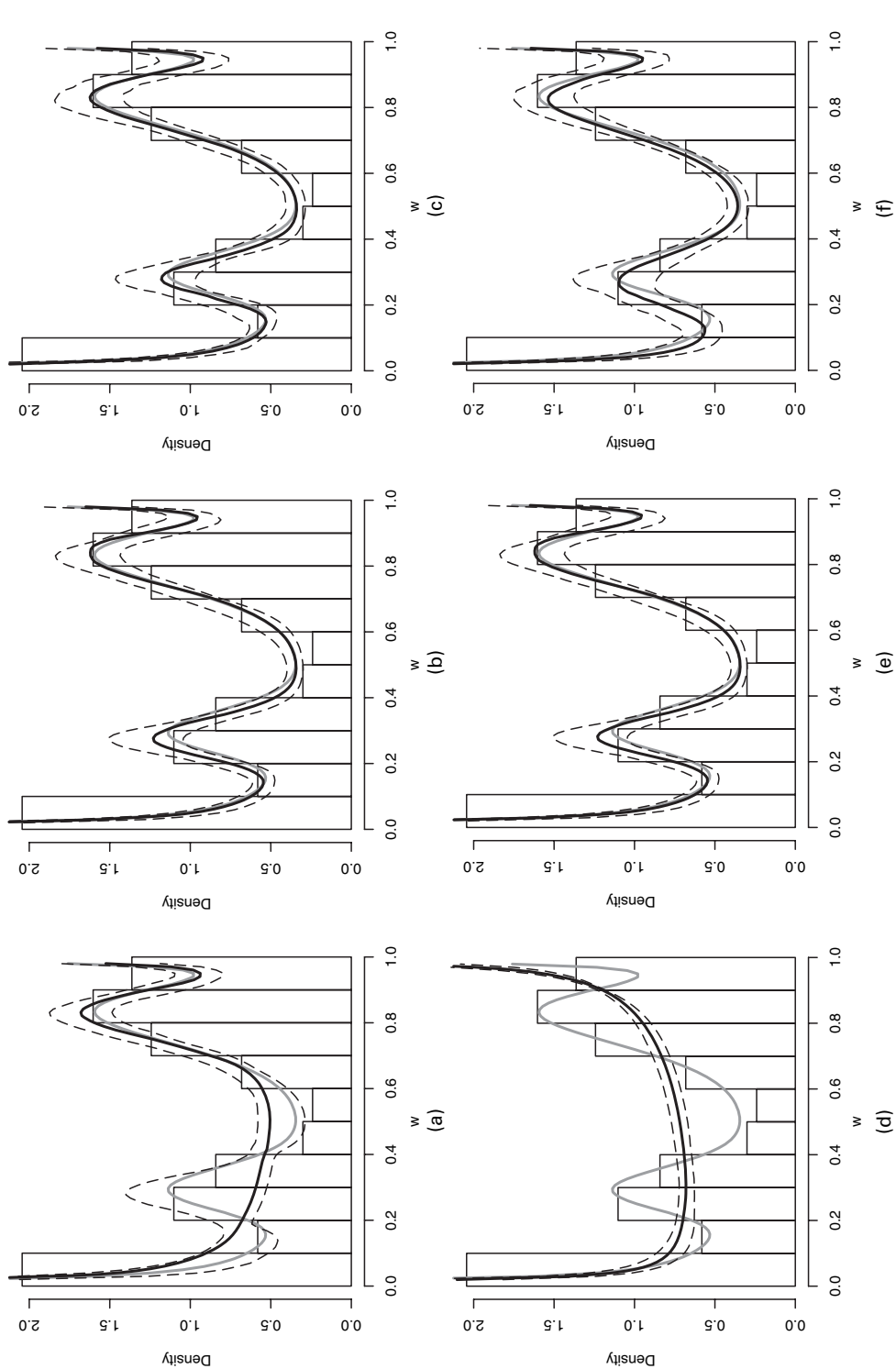
Fig. 3. Sensitivity of posterior results to prior distributions (each plot contains histograms of the data, the true density (———), posterior median (———) and 90% credibility intervals (– – –)): (a) prior on k, $\sigma = 10$, $\lambda = 1$; (b) prior on k, $\sigma = 10$, $\lambda = 2$; (c) prior on k, $\sigma = 10$, $\lambda = 8$; (d) prior on ν, $\lambda = 2$, $\sigma = 1$; (e) prior on ν, $\lambda = 2$, $\sigma = 10$; (f) prior on ν, $\lambda = 2$, $\sigma = 100$

influence of the prior varies with k—for example, the sensitivity of $\log(\nu)$ may inflate as k grows. We now briefly outline how changes to the Poisson prior on k and the normal distribution on $\log(\nu)$ alter the posterior, using the uniform prior on π provided by taking $\delta_1 = \ldots = \delta_p = 1$. To assess the sensitivity to the prior on k, three chains were run for 20 000 iterations, after a 30 000-iteration burn-in; in each we took $\log(\tilde{\nu}) = 0$ and $\sigma = 10$. The results are shown in Figs 3(a)–3(c). The final estimate when $\lambda = 1$ misses one component, though the credibility intervals contain the true density, but the fit improves when $\lambda = 2$ or $\lambda = 8$. Histograms of k indicate that the number of components remains around 4 or 5 even when $\lambda = 8$: taking a high λ need not entail an explosion in the number of mixture components.

For the sensitivity to the prior on $\log(\nu)$, three chains were run for 20 000 iterations, after a 30 000-iteration burn-in. The results are shown in Figs 3(d)–3(f). The influence of σ on the results is strong, and it is important that the prior for $\log(\nu)$ be sufficiently dispersed. For example, if the prior for $\log(\nu)$ is given a mean of 0 and unit standard deviation, the prior for $\log(\nu)$ misses $\log(20) = 2.996$ and $\log(50) = 3.91$, corresponding to the two central bumps, with a probability exceeding 95%, and the posterior shown in Fig. 3(f) misses these. A further 50 000 iterations gave no improvement, so the difficulty seems not to stem from a lack of convergence.

4.2. Oceanographic data

We first consider trivariate data from Coles and Tawn (1994) consisting of a sequence of hourly surge records for the years 1971–1977 for the port of Newlyn, Cornwall, and 3-hourly wave records from a ship. After initial transformations and analysis, the remaining series is of length 2894. Coles and Tawn (1994) assumed that these data were serially independent and used the Dirichlet model to account for the dependence structure between the three variables. They chose a threshold of $r_0 = \exp(3.3)$, giving 222 excesses, and took marginal thresholds of $u_1 = 6.59$, $u_2 = 11.6$ and $u_3 = 0.351$. The marginal and dependence parameters are simultaneously estimated by maximum likelihood, yielding estimates $\hat{\alpha} = (0.497, 0.985, 0.338)$ of the dependence parameters of the Dirichlet model, whose fit is shown in Fig. 4(a); the variables w_1, w_2 and w_3 are the wave height, surge and period respectively.

The mixture model was fitted with the reversible jump algorithm modified to allow simultaneous estimation of the marginal and dependence parameters. The resulting fitted density is shown in Fig. 4(b). Beyond the general impression that the mixture model gives a better fit, Fig. 4(b) suggests asymptotic independence of the surge and the period, (w_2, w_3).

Thus, although our model does not encompass asymptotic independence in the sense of Ledford and Tawn (1996, 1997, 2003), the results may suggest inadequacies with models that allow only asymptotic dependence.

4.3. Air quality data

The air quality data can be downloaded from `http://www.airquality.co.uk` and were analysed by Heffernan and Tawn (2004). They comprise daily series of monitoring measurements of ozone levels (O_3), nitrogen dioxide (NO_2), nitrogen oxide (NO) and particulate matter (PM_{10}), in the city centre of Leeds, UK, over 1994–1998. Levels of the gases are measured in parts per billion, and those of PM_{10} in micrograms per cubic metre. We follow Heffernan and Tawn (2004) in focusing on data for November–February, with missing values deleted and stationarity in time assumed. The margins are standardized to the unit Fréchet distribution by non-parametric transformation. The threshold is selected as $\exp(2.5)$ by an *ad hoc* method, leaving 247 extremes.

Fig. 5 shows results that were obtained by using the EM algorithm for $k = 1, 5, 10$ mixture components; BIC suggests that $k = 10$, and contour plots for fitted densities with $k > 10$ are
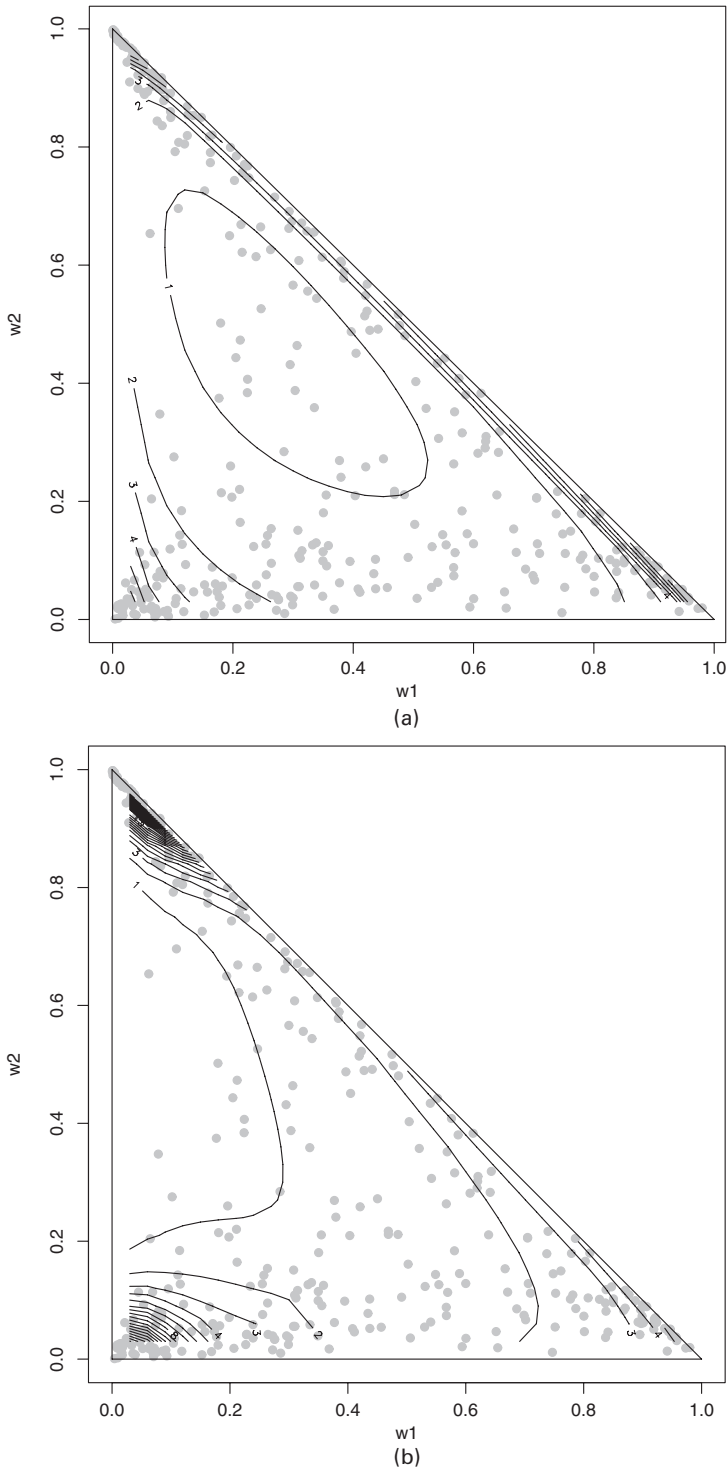
Fig. 4. Contours of estimated spectral densities for the oceanographic data *w* (●): (a) fitted Dirichlet model (Coles and Tawn 1994); (b) mean posterior density estimated from the Dirichlet mixture model
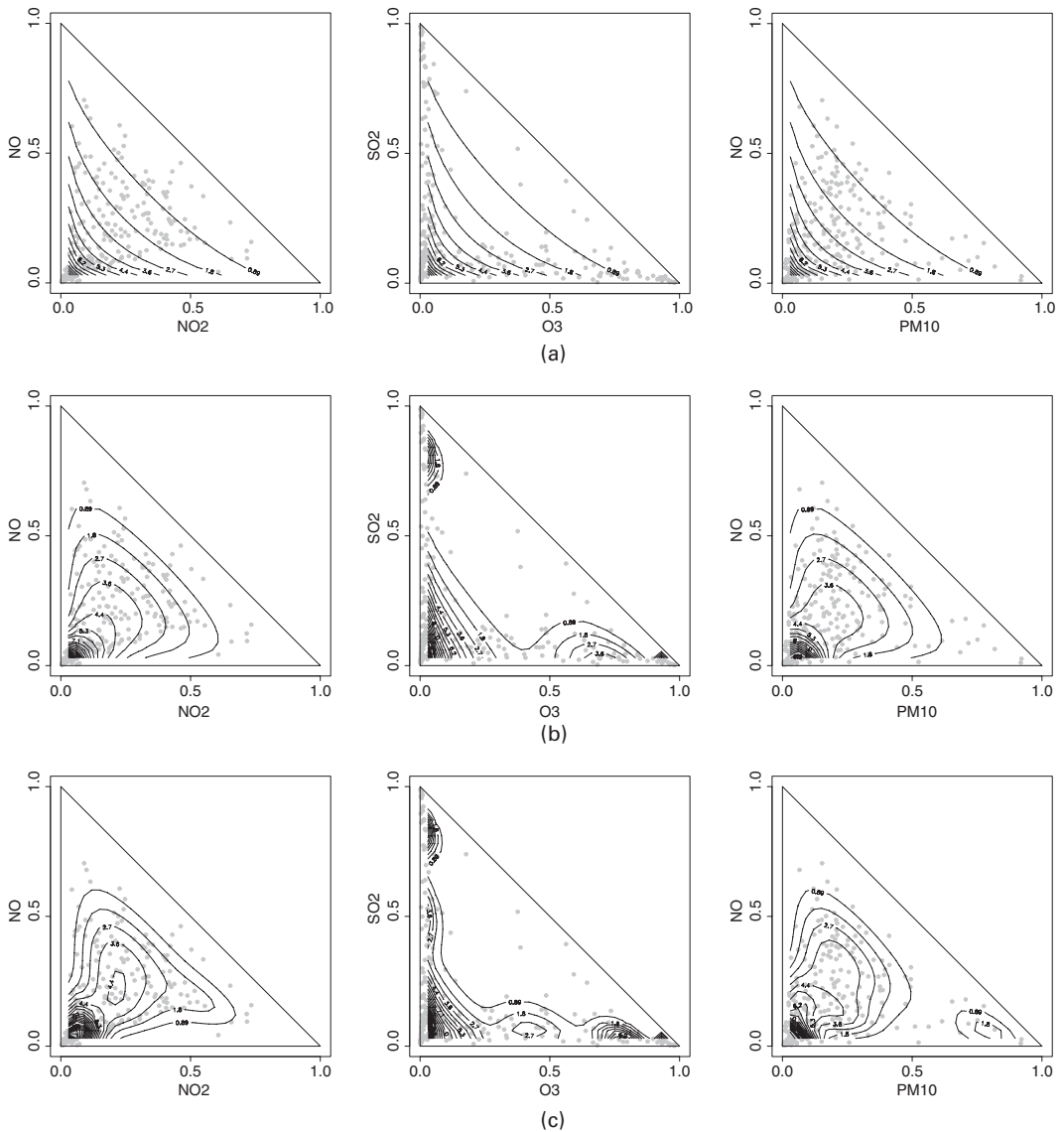
Fig. 5. Results of the EM algorithm used to fit the mixture model to air quality data (———, fitted bivariate density contour levels; ●, data): (a) $k = 1$; (b) $k = 5$; (c) $k = 10$

essentially the same as those for $k = 10$. The panels suggest dependence between NO_2 and NO and between NO and PM_{10}, and independence between O_3 and SO_2. Similar conclusions can be drawn from other bivariate marginal plots. The density contours also highlight more subtle dependences. For example, the plot of (NO_2, NO) shows clearly the strong dependence between these two components. Similarly the plot of (PM_{10}, NO) shows the dependence of this pair but, when the level of NO is very low, the level of PM_{10} might be quite high with a rather high probability. In other words, given a low level of NO, PM_{10} is independent of (NO_2, SO_2, O_3). Detailed interpretation of the conditional dependences between various components would require a lengthy substantive discussion, and we stop here: it is clear that the mixture model is both flexible and useful in several dimensions.

5.   Discussion

This paper introduces a semiparametric model for the spectral distribution in multivariate extremes, based on mixtures of Dirichlet distributions. This is a rich model which can be fitted to high dimensional data using the extensive existing knowledge about mixture models, and in principle it can approach any spectral distribution arbitarily well, in the weak sense. One practical drawback with the approach stems from the use of simulation algorithms, which may converge slowly unless they have been tuned. A second is that the number of parameters increases rapidly with the number of mixture components, so model complexity must be sharply penalized through an information criterion or a prior on the number of mixture components.

Our approach underlines that the spectral measure may be treated as the distribution of a random probability measure. Although not developed here, this viewpoint gives insight into the multivariate extremes of random measures (Boldi, 2004).

Acknowledgements

Appendix A: Adequacy of constrained Dirichlet mixtures

Let H represent a given spectral distribution function on the p-dimensional simplex $\mathcal{S}_p$; we write its mean in vector form as $\mu = \int w\, H(\mathrm{d}w)$, and note that $\mu = p^{-1}\mathbf{1}_p$, where $\mathbf{1}_p$ is the unit vector in $\mathbb{R}^p$.

Let $\mathcal{D}$ denote the set of mixtures of Dirichlet distributions on $\mathcal{S}_p$. The results of Dalal (1978) and Dalal and Hall (1980) on adequacy of such mixtures imply that there is a sequence of measures $\{D_n\} \subset \mathcal{D}$ such that D_n converges weakly to H, i.e. $\int f\, \mathrm{d}D_n \to \int f\, \mathrm{d}H$ as $n \to \infty$ for any bounded and continuous function $f : \mathcal{S}_p \to \mathbb{R}$. Thus in particular $\mu_j^n = \int_{\mathcal{S}_p} w_j\, D_n(\mathrm{d}w) \to p^{-1}$ for each j, and the corresponding mean vector $\mu_n \to \mu$ as $n \to \infty$.

Our strategy is to use the sequence $\{D_n\}$ to construct a sequence $\{H_n\} \subset \mathcal{D}$ of valid spectral distribution functions that converges weakly to H.

Let δ_j denote the vector with a 1 in the jth place and remaining components 0. Now $\mathcal{S}_p$ is the convex hull of $\delta_1, \ldots, \delta_p$, each of which can be regarded as having a degenerate Dirichlet distribution; we denote these distributions by $\delta_1', \ldots, \delta_p'$. As both μ and μ_n lie in the convex hull of the δ_j, it is possible to find probabilities $\{\pi_0^n, \pi_1^n, \ldots, \pi_p^n\}$ such that

$$\pi_0^n \mu_n + \sum_{j=1}^p \pi_j^n \delta_j = \mu, \qquad n = 1, 2, \ldots,$$

and, as $\pi_0^n = 1 - \Sigma_{j=1}^p \pi_j^n$, this equation may be rewritten as

$$\mu - \mu^n = \sum_{j=1}^n \pi_j^n (\delta_j - \mu^n).$$

As $\mu^n \to \mu$, for sufficiently large n the set $\{\delta_1 - \mu^n, \ldots, \delta_p - \mu^n\}$ spans $\mathcal{S}_p' = \{\omega \in \mathbb{R}^p : \Sigma_{j=1}^p \omega_j = 0\}$, which has dimension $p-1$. Thus for sufficiently large n we can choose the π_j^n such that at least one of $\pi_1^n, \ldots, \pi_p^n$ equals 0. It follows that there is a subsequence $\{n_m\} \subset \mathbb{N}$ and a j such that $\pi_j^{n_m} = 0$ for all m. Without loss of generality we may suppose that this is true of the original sequence, and that $\pi_1^n = 0$ for all n. If so, the first co-ordinate of $\mu - \mu^n$ is $-\Sigma_{j>1} \pi_j^n \mu_1^n$, and this converges to 0 as $n \to \infty$, whereas $\mu_1^n \to 1/p$. Thus $\Sigma_{j>1} \pi_j^n = \Sigma_j \pi_j^n \to 0$, and hence $\pi_0^n \to 1$. The distributions $H_n = \pi_0^n D_n + \pi_1^n \delta_1' + \ldots + \pi_p^n \delta_p'$ corresponding to this sequence are Dirichlet mixtures having mean μ and for which

$$\int f\, \mathrm{d}H_n = \pi_0^n \int f\, \mathrm{d}D_n + \pi_1^n\, f(\delta_1) + \ldots + \pi_p^n\, f(\delta_p) \to \int f\, \mathrm{d}H,$$

for any bounded continuous function f on $\mathcal{S}_p$: thus $\{H_n\} \subset \mathcal{D}$ is a sequence of measures all having mean μ which converges weakly to H. This implies that mixtures of Dirichlet distributions with a given mean are adequate for the class of distributions on $\mathcal{S}_p$ having the same mean.

The use of the degenerate Dirichlet distributions δ'_j above is not essential: the sequence μ^n will eventually lie inside a set that is bounded away from the edges of $\mathcal{S}_p$, and so the δ'_j could be replaced by non-degenerate Dirichlet distributions having means $ap^{-1}\mathbf{1}_p + (1-a)\delta_j$ for some small positive a, without changing the conclusion of the argument.

Appendix B: Reversible jump algorithm

We first describe the possible reversible jump moves. At each step, two move types are possible: an 'MCMC' move type updating parameters for a fixed k and a 'split–combine' move type that adds one component to or subtracts one component from the mixture. They are defined as follows.

(a) Split–combine: choose between split and combine with probability $p_s(k)$ and $p_c(k)$, such that $p_s(k) + p_c(k) = 1$ and $p_c(1) = 0$.
(b) MCMC: construct all possible random couples (m_{i_1}, m_{i_2}), $1 \leqslant i_1 < i_2 \leqslant k$, without replacement. Apply a combine and then a split move to each couple.

These move types decompose into the following moves.

(a) Split: choose m_0 in $1, \dots, k$ uniformly. Simulate v according to a beta distribution. Set $\pi_{m_1} = v\pi_{m_0}$ and $\pi_{m_2} = (1-v)\pi_{m_0}$. Simulate μ_{m_2} according to a Dirichlet distribution on $\mathcal{S}_k$. Set $\mu_{m_1} = \pi_{m_1}^{-1}(\pi_{m_0}\mu_{m_0} - \pi_{m_2}\mu_{m_2})$. Simulate $\log(\nu_{m_1})$ and $\log(\nu_{m_2})$ independently according to a normal with mean $\log(\nu_0)$.
(b) Combine: choose a couple (m_1, m_2) uniformly. Set $\pi_{m_0} = \pi_{m_1} + \pi_{m_2}$ and $\mu_{m_0} = \pi_{m_0}^{-1}(\pi_{m_1}\mu_{m_1} + \pi_{m_2}\mu_{m_2})$. Simulate $\log(\nu_{m_0})$ according to a normal with mean $\log(\nu_{m_1}) + \log(\nu_{m_2})$.

The condition for convergence of the chain to the target distribution, the reversibility of jumps, is satisfied since for each combine one can choose a split inverting it, and vice versa.

To obtain good mixing properties of the chain, the parameters of the proposals are selected according to the size of the move, 'big', 'medium' or 'small'. These parameters are such that a big move proposes a new state far away from the current one, etc.

We now give the acceptance probabilities for the moves described above. Consider a split of a component m_0 into two components labelled m_1 and m_2. The probability that this is accepted is the minimum of 1 and the acceptance ratio

$$\frac{g(\mu', \pi', \nu', k'|w)}{g(\mu, \pi, \nu, k|w)} \, \frac{q\{\log(\nu_{m_0})|\log(\nu'_{m_1}), \log(\nu'_{m_2})\}}{q\{\log(\nu'_{m_1})|\log(\nu_{m_0})\} \, q\{\log(\nu'_{m_2})|\log(\nu_{m_0})\}} \, \frac{v^{p-1}}{q(\mu'_{m_2}|\mu_{m_0}) \, q(v)\pi_{m_0}} \, \frac{2k}{(k+1)k},$$

where a prime indicates a proposal, g is the posterior density function and q is a generic term for a proposal density function. From left to right appear the ratio of posterior densities, the ratio for proposals $\nu_{m_0} \mapsto (\nu'_{m_1}, \nu'_{m_2})$, the ratio for the proposal $(\mu_{m_0}, \pi_{m_0}) \mapsto (\mu_{m_1}, \mu_{m_2}, \pi_{m_1}, \pi_{m_2})$ and the ratio due to the random choice of m_0 forwards and (m_1, m_2) backwards. The acceptance ratio for a combine move is the reciprocal of that for a split.

We now construct F_μ, the prior on μ given π appearing in Section 3. Let $\mu = \{\mu_j^{(m)}\}$. The prior density $f(\mu) = f(\mu_1^{(1)}, \dots, \mu_{p-1}^{(1)}, \mu_1^{(2)}, \dots, \mu_{p-1}^{(2)}, \dots, \mu_{p-1}^{(k-1)})$ is the product of successive conditionals, starting from the right,

$$f(\mu_{p-1}^{(k-1)}|\mu_1^{(1)}, \dots, \mu_{p-1}^{(1)}, \mu_1^{(2)}, \dots, \mu_{p-1}^{(2)}, \dots, \mu_{p-2}^{(k-1)}), \dots, f(\mu_1^{(1)}).$$

These conditionals are taken to be uniform on the largest interval such that constraints (2), i.e., for $i = 1, \dots, p-1$, $j = 1, \dots, k-1$,

$$I_i^{(m)} = \left[0, \min\left(1 - \sum_{j=1}^{i-1} \mu_j^{(m)}, \frac{p^{-1} - \sum_{l=1}^{m-1} \pi_l \mu_i^{(l)}}{\pi_m}\right)\right].$$

The prior density on μ is the inverse of the product of the lengths of all $I_i^{(m)}$.

References

Beirlant, J., Goegebeur, Y., Teugels, J. and Segers, J. (2004) *Statistics of Extremes: Theory and Applications*. New York: Wiley.

Boldi, M.-O. (2004) Mixture models for multivariate extremes. *PhD Thesis*. Ecole Polytechnique Fédérale de Lausanne, Lausanne.

Brooks, S. P. and Giudici, P. (2000) Markov chain Monte Carlo convergence assessment via two-way analysis of variance. *J. Computnl Graph. Statist.*, **9**, 266–285.

Brooks, S. P. and Roberts, G. O. (1998) Convergence assessment techniques for Markov chain Monte Carlo. *Statist. Comput.*, **8**, 319–335.

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference: a Practical Information-theoretic Approach*, 2nd edn. New York: Springer.

Coles, S. G. (2001) *An Introduction to Statistical Modeling of Extreme Values*. London: Springer.

Coles, S. G. and Tawn, J. A. (1994) Statistical methods for multivariate extremes: an application to structural design (with discussion). *Appl. Statist.*, **43**, 1–48.

Dalal, S. R. (1978) A note on the adequacy of mixtures of Dirichlet processes. *Sankhya* A, **40**, 185–191.

Dalal, S. R. and Hall, G. J. (1980) On approximating parametric Bayes models by nonparametric Bayes models. *Ann. Statist.*, **8**, 664–672.

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

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Heffernan, J. E. and Tawn, J. A. (2004) A conditional approach for multivariate extreme values (with discussion). *J. R. Statist. Soc.* B, **66**, 497–546.

Hurvich, C. M. and Tsai, C.-L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Kotz, S. and Nadarajah, S. (2000) *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press.

Ledford, A. W. and Tawn, J. A. (1996) Statistics for near independence in multivariate extreme values. *Biometrika*, **83**, 169–187.

Ledford, A. W. and Tawn, J. A. (1997) Modelling dependence within joint tail regions. *J. R. Statist. Soc.* B, **59**, 475–499.

Ledford, A. W. and Tawn, J. A. (2003) Diagnostics for dependence within time series extremes. *J. R. Statist. Soc.* B, **65**, 521–543.

McLachlan, G. J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*. New York: Wiley.

McLachlan, G. J. and Peel, D. (2000) *Finite Mixture Models*. New York: Wiley.

McQuarrie, A. D. R. and Tsai, C.-L. (1998) *Regression and Time Series Model Selection*. Singapore: World Scientific Press.

Meng, X.-L. and Pedlow, S. (1992) EM: a bibliographic review with missing articles. *Proc. Statist. Comput. Sect. Am. Statist. Ass.*, 24–27.

Redner, R. A. and Walker, H. F. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, **26**, 195–239.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Statist. Soc.* B, **59**, 731–792; correction, **60** (1998), 661.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Wilks, S. S. (1962) *Mathematical Statistics*. New York: Wiley.