Tilted Beta Extremal Mixtures

Samuel Sekarski

June 2020

Contents

List of Figures

Chapter 1

Introduction

Modelling extreme events is becoming more and more important, mostly in order to assess risks (financial, ecological, structural, ...). Modelling of univariate extremes is well documented and explored, using techniques such as block maxima, threshold exceedances and point processes. However, things become, as usual, more complicated in higher dimensions. Multivariate extremes suffer from problems that affect univariate extremes less, such as the curse of dimensionality and sparsity.

The most primitive way to deal with multivariate extremes is to study each component as a univariate process. However, this is limiting, as we could easily imaging that there is interdependence of the components, which we lose by considering the components independently. Another reason, as is stated in [2], is that the combination of the individual processes might be of more interest than each process individually.

Methods analogous to block maxima and threshold analysis exist for multivariate cases and we can find models for extreme multivariate events, but we do not have a characterization for the class of all the models. Theorem 8.1 from [2] defines a family of bivariate extreme value distributions (and can be generalized to general multivariate case) that arise as the limiting distribution for componentwise block maxima.

Here is Theorem 8.1 restated (for a bivariate process) for completeness:

Theorem 1 *Let $M_n^* = (\max_{i=1,\dots,n}\{X_i\}/n, \max_{i=1,\dots,n}\{Y_i\}/n)$ be the vector of rescaled componentwise maxima, where (X_i, Y_i) are independent vectors with standard Fréchet marginal distributions. Then if*

$$\mathbb{P}\{M_n^* \leq (x,y)\} \xrightarrow{d} G(x,y),$$

where G is a non-degenerate distribution function, then G has the form

$$G(x,y) = \exp\{-V(x,y)\}, \quad x > 0, y > 0$$

where

$$V(x,y) = 2\int_0^1 \max\left(\frac{w}{x}, \frac{1-w}{y}\right) d\nu(w)$$

and ν is a distribution on $[0,1]$ satisfying the mean constraint

$$\int_0^1 w d\nu(w) = 1/2.$$

The problem is that we don't know how to characterize ν. An approach is to try and approximate the class arbitrarily well, using parametric subfamilies or nonparametric methods and another way is to use nonparametric methods.

Boldi and Davison [3] approached the problem by using a semi-parametric model based on mixtures of Dirichlet distributions that weakly approximates the class of limit distributions.

In this project we will try to use mixtures of beta distributions that have been tilted using Theorem 2 from Coles and Tawn [1] to satisfy the mean constraints.

In Section 2 we will discuss how it is possible to tilt a distribution for it to satisfy the mean constraints, how to sample from a tilted distribution, and provide examples of tilted densities and sampling therefrom. In Section 3 we will explore how to fit a tilted distribution to some data, using maximum likelihood, fit for some artificially generated data and fit from some real world data, and assess the quality of the fits.

Chapter 2

Multivariate extremes

2.1 Basic setup

We will restrict ourselves to the two-dimensional case but some theorems and results will be stated for arbitrary D dimensions. The D-simplex on which our considered distributions will be defined is the set

$$S_D := \left\{ x \in \mathbb{R}_+^D : \sum_{i=1}^{D} x_i = 1 \right\}$$

When $D = 2$, that means that we only need to define a distribution on $x_1 \in [0,1]$, and $x_2 = 1 - x_1$ is completely determined by x_1. As mentioned in Section 1, we are going to consider distributions that are a mixture of K beta distributions:

$$\nu^*(x_1, x_2) = \prod_{k=1}^{K} \pi_k Beta(x_1; \alpha_k, \beta_k), \quad 0 < x_1 < 1,$$

with

$$\prod_{k=1}^{K} \pi_k = 1, \quad \pi_k \geq 0.$$

The mean of this distribution is

$$\mathbb{E}[X_1] = \sum_{k=1}^{K} \pi_k \frac{\alpha_k}{\alpha_k + \beta_k}, \quad (X_1, X_2) \sim \nu^*,$$

But in general this is not equal to $1/2$ and so this class of distributions does not satisfy Theorem 8.1. In section 2.2 we will see how to tilt a wide class of distributions to force the mean constraint $1/D$ to hold, and will apply it to our case.

2.2 Construction of angular distributions

The main tool for tilting distributions is Theorem 2 from the 1991 paper from Coles and Tawn [1], which we state again here for completeness:

Theorem 2 *If h^* is any positive function on S_D with finite first moments, then*

$$\nu(w) = (m^T w)^{-(D+1)} D^{-1} \left(\prod_{d=1}^{D} m_d \right) \nu^* \left(\frac{m_1 w_1}{m^T w}, \dots, \frac{m_D w_D}{m^T w} \right), \quad (w_1, \dots, w_D) \in S_D,$$

where

$$m_d = \int_{S_D} u_d \nu^*(u) du, \quad d = 1, \dots, D,$$

satisfies mean constraints $1/D$ and is therefore the density of a valid measure function ν.

To verify that the theorem holds, we need to verify that the Jacobian of the transformation

$$W_d = \frac{W_d^*/m_d}{\sum_{c=1}^{D} W_c^*/m_c}, \quad W_d^* = \frac{m_d W_d}{\sum_{c=1}^{D} m_c W_c}, \quad d = 1, \dots, D,$$

is $|\partial w^*/\partial w| = (m^T w)^{-D} \prod_{d=1}^{D} m_d$. To do this, we rewrite the first transformation as

$$w_s^* = \frac{m_d w_d}{m_D + \sum_{c=1}^{D-1} (m_c - m_D) w_c}, \quad d = 1, \dots, D-1.$$

by noting that $w = (w_1, \dots, w_D) \in S_D$. To simplify notation, we write $m^T w = m_D + \sum_{c=1}^{D-1} (m_c - m_D) w_c$. As such,

$$\begin{aligned} \partial w_d^*/\partial w_d &= m_d/(m^T w) - m_d w_d (m_d - m_D)/(m^T w)^2, \\ \partial w_d^*/\partial w_c &= -m_d w_d (m_c - m_D)/(m^T w)^2, \quad c \neq d. \end{aligned}$$

This defines a matrix that we can write as $A + ab^T$, where

$$\begin{aligned} A &= \operatorname{diag}(m_1, \dots, m_{D-1})/(m^T w), \\ a &= (m_1 w_1, \dots, m_{D-1} w_{D-1})^T, \\ b &= -(m_1 - m_D, \dots, m_{D-1} - m_D)^T/(m^T w)^2. \end{aligned}$$

Let's recall the determinant lemma: Let A be an invertible $p \times p$ matrix, and let a, b, be vectors of length p. Then $|A + ab^T| = |A|(1 + b^T A^{-1} a)$.

In our case, A is diagonal, so

$$|A| = (m^T w)^{-(D-1)} \prod_{c=1}^{D-1} m_c,$$

$$b^T A^{-1} a = -\frac{m^T w + m_D}{m^T w},$$

so

$$1 + b^T A^{-1} a = m_D / m^T w.$$

Therefore, $|\partial w^* / \partial w| = (m^T w)^{-D} \prod_{d=1}^{D} m_d$, so the variable $W = (W_1, \ldots, W_D)$ has the probability density function

$$f(w) = (m^T w)^{-D} \left(\prod_{d=1}^{D} m_d \right) \nu^* \left(\frac{m_1 w_1}{m^T w}, \ldots, \frac{m_D w_D}{m^T w} \right)$$

For every d, W_d^* / m_d has unit expectation, which leads to the following equality

$$1 = \mathbb{E}\left(\frac{W_d^*}{m_d} \right) = \mathbb{E}\left(\frac{W_d}{m^T W} \right) = \int_{S_{D-1}} w_d (m^T w)^{-(D+1)} \left(\prod_{d=1}^{D} m_d \right) \nu^* \left(\frac{m_1 w_1}{m^T w}, \ldots, \frac{m_D w_D}{m^T w} \right) \mathrm{d}w$$

by noting that $\sum_d w_d = 1$ and summing the previous equality over d, we get that

$$\nu(w) = D^{-1} (m^T w)^{-(D+1)} \left(\prod_{d=1}^{D} m_d \right) \nu^* \left(\frac{m_1 w_1}{m^T w}, \ldots, \frac{m_D w_D}{m^T w} \right).$$

which is a well defined density on S_{D-1} satisfying the mean constraint

$$\int_{S_{D-1}} w_d \nu(w) \mathrm{d}w = D^{-1}, \quad d = 1, \ldots, D.$$

Sampling from tilted density

As we have seen in the proof of theorem 2, to construct the appropriate density ν, we start with a density ν^*, apply a change of variable to it, then tilt the result by diving it by $D m^T w$, which is bounded, in order to sample from the tilted density, we can sample from the original density ν^*, transform the the samples, then apply the following Acceptance-Rejection step:

$$U \leq m_{\min} / m^T w$$

where $m_{\min} = \min_d m_d$

This is because a realisation W^* of ν^* that has been transformed has density f and not ν. But $m_{\min}/m^T W \leq 1$, so conditional on $W = w$ the event $U \leq m_{\min}/m^T w$ has probability $m_{\min}/m^T w$. Thus the marginal density of the W for which the event occurs is proportional to $f(w)/(m^T w)$ and has to be ν.

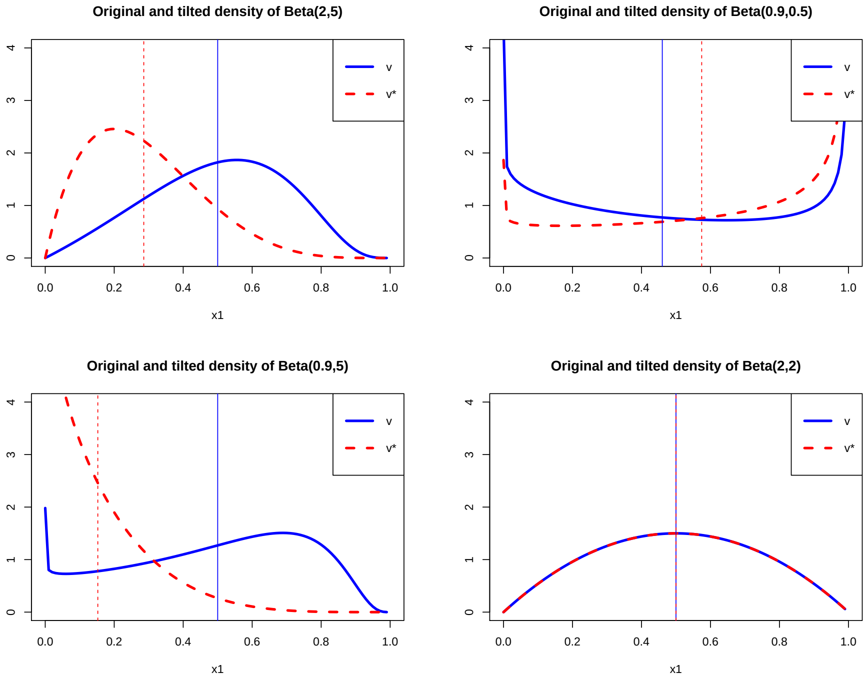


Figure 2.1: Graphs of beta distributions for various shape parameters, before and after tilting. The vertical lines represent the means. As we can see, although the original densities have various means, the tilted densities all have mean 0.5, and when the original distribution already has the correct mean, it is not tilted.

The acceptance probability is given by

$$\int \frac{m_{\min}}{m^T w} f(w) \mathrm{d}w = D m_{\min} \int \nu(w) \mathrm{d}w = D m_{\min}$$

Thus the number of accepted samples is proportional to m_{min}, so the algorithm is the most efficient when all the m_d are equal to $1/D$ which would mean our original density ν^* already satisfies the criteria and we would not have to run the algorithm at all.

Tests

Figure 2.2 some examples of tilted mixtures of beta distributions, sampled using the algorithm, and the theoretical tilted distributions using the formula.

2.3 Tilted mixtures are dense

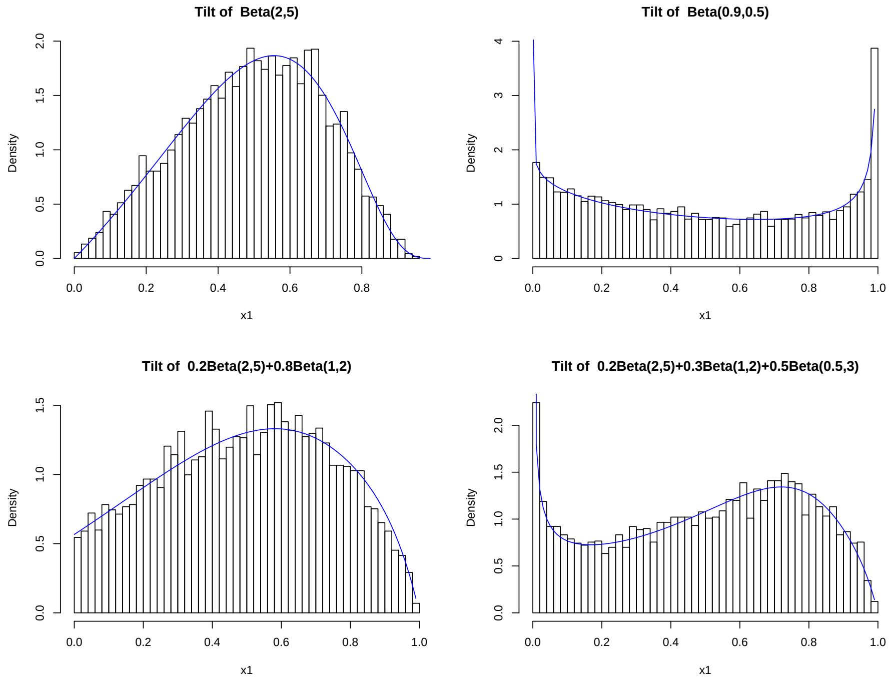


Figure 2.2: Histograms of 10^4 samples from various mixtures of beta distributions that have been tilted. The pdfs of the distributions are overlaid in blue. As we can see, the sampling algorithm for the tilted mixtures is correct.

Chapter 3

Statistical aspects

3.1 Likelihood fitting

We use the generic R optimiser *optim* to minimise the negative log-likelihood function. In order to use it we first reparameterized the parameters, in order to express the S_2 constraint in a way that the optimiser can understand.

$$\pi_k = \exp(\eta_k) / \left\{ 1 + \sum_{i=2}^{K} \exp(\eta_i) \right\}, \quad k = 2, \ldots, K,$$

and

$$\pi_1 = 1 / \left\{ 1 + \sum_{i=2}^{K} \exp(\eta_i) \right\}.$$

We also reparametrize the α and β parameters as

$$\alpha_k = \exp(\xi_k), \quad \beta_k = \exp(\zeta_k), \quad k = 1, \ldots, K,$$

to ensure that the parameters can take any values in $\mathbb{R}$. There are $3K - 1$ parameters to estimate, which can then be transformed back into the original parameterization after estimation. Initially, we consider K to be a fixed parameter, and for given data we will try to fit with various values of K, including the real value, to see the accuracy, and if values of K that are close, but not correct give a "good enough" estimation (for example we could use a smaller K than the true K and still have a model that is accurate "enough" for certain purposes. Then we shall try to consider K as a parameter to be estimated as well, using methods such as step-AIC/BIC to select the model.

3.2 Numerical experiments

As a measure of the error between the true density f and the estimated density based on n samples f_n, we will use the integrated square error, defined as

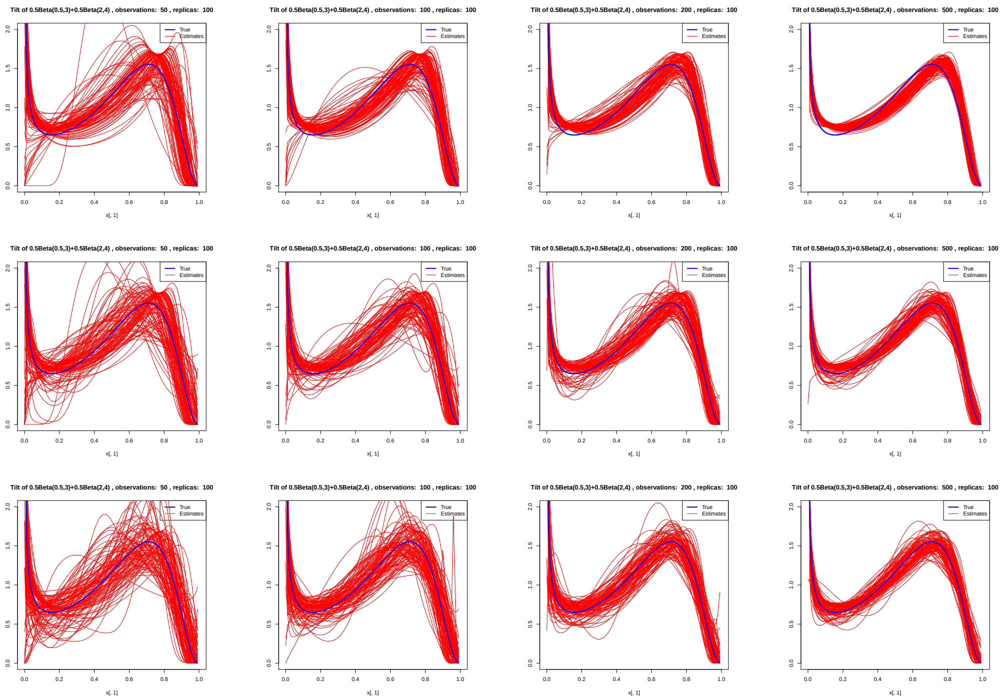


Figure 3.1: MLE fits to samples of size, from left to right, $n = 50$, $n = 100$, $n = 200$ and $n = 500$, with different number of betas in the mixture. From top to bottom: $K = 1$, $K = 2$ (which is the number that generated the data), and $K = 3$, using a *Nelder-Mead* optimiser each time, with a maximum of 500 iterations.

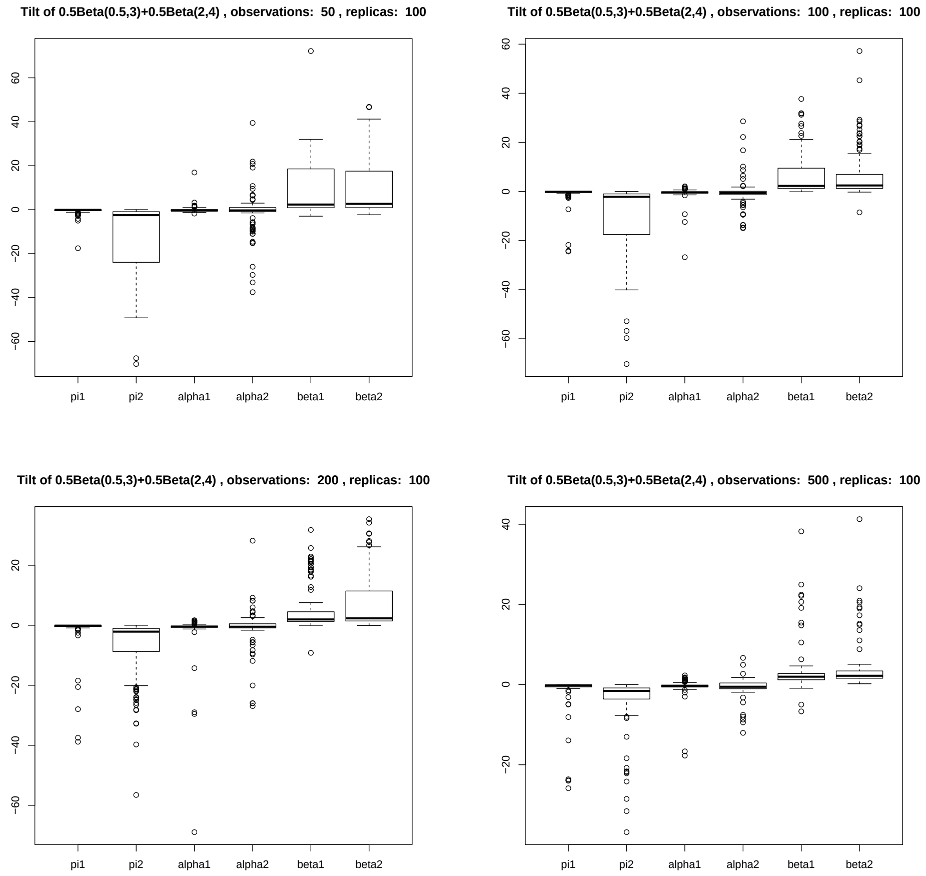
$$\mathrm{err}(f_n) := ||f_n - f||_2^2 = \int_0^1 (f_n(x) - f(x))^2 dx$$

In Section 3.2.1 we fitted a tilted beta mixture to data actually simulated from a tilted beta mixture, then in Section 3.2.2 we fitted a tilted beta mixture to data simulated from other distributions (i.e., not from a beta mixture).

For each test distribution, we simulate $R = 100$ samples of $n = 50$, $n = 100$, $n = 200$, and $n = 500$ independent observations. The Maximum Likelihood Estimation is done with ten different sets of initial parameters chosen uniformly at random between -0.5 and $+0.5$.

3.2.1 Data simulated from mixture

First we generated data using a tilted mix of two beta distributions, $0.5Beta(0.5, 3)+ 0.5Beta(2, 4)$, to test how well the maximum likelihood estimation would work. In Figure 3.1 we see that for samples with a high number of observations, even


Tilt of 0.5Beta(0.5,3)+0.5Beta(2,4) , observations: 50 , replicas: 100

Tilt of 0.5Beta(0.5,3)+0.5Beta(2,4) , observations: 100 , replicas: 100

Tilt of 0.5Beta(0.5,3)+0.5Beta(2,4) , observations: 200 , replicas: 100

Tilt of 0.5Beta(0.5,3)+0.5Beta(2,4) , observations: 500 , replicas: 100

Figure 3.2: Boxplot of log MLE estimates (Nelder-Mead, maxit 500), for $K = 2$ and sample of 50, 100, 200, 500 observations

estimating with a different number K of components, the results are good, which suggests that we could indeed use a step-AIC/BIC method, starting with a single tilted beta distribution, and keep estimating with more until the added accuracy is no longer significant. Looking at the case n=50, however, we see a lot of volatility and artefacts. Some of the estimations have clearly not converged in 500 iterations of the Nelder-Mead algorithm, even when the correct number K of betas is used.

Looking at the boxplots of the log parameter for $K = 2$, (Figure 3.2) we see that there are a lot of very large and very small values. This makes the analysis of the boxplots difficult. But looking at Figure 3.1 we can guess what is happening: as the algorithm fits relatively well with just just a single beta

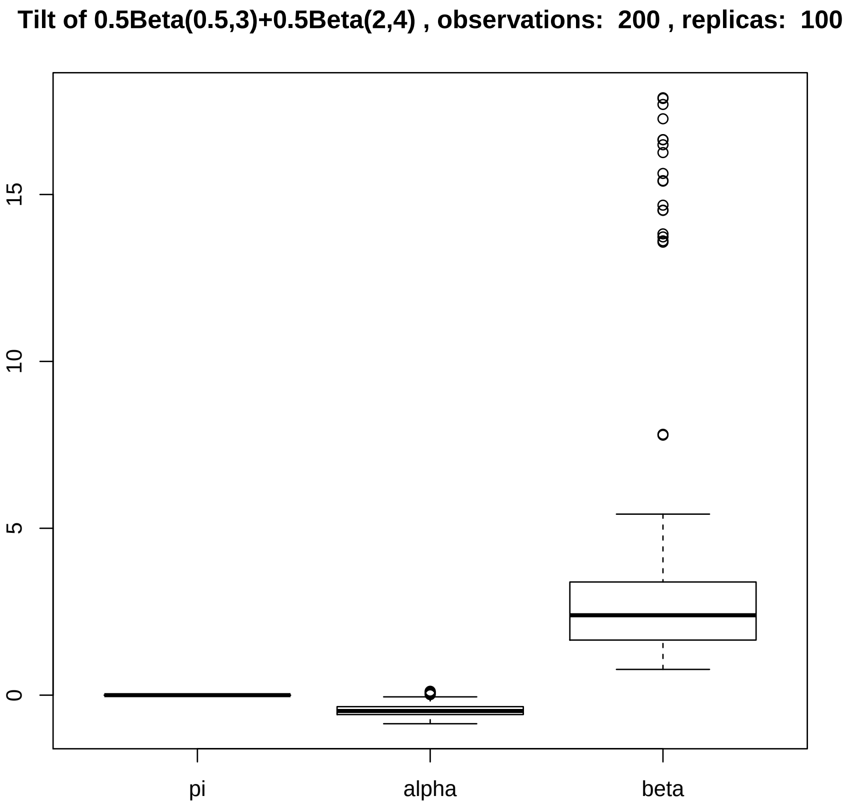


Figure 3.3: Boxplot of log MLE estimates (Nelder-Mead, maxit 500), for $K = 1$ and sample of 200 observations

distribution instead of two, all the cases with a very small π_i, is just the algorithm fitting one of the components very well, which sends the other component to insignificance. As for very small/large values for α_i and β_i, our hypothesis is that in the cases where the algorithm fitted one component really well, and the other component can take pretty much any value it wants, as it's contribution to the mixture has hardly any weight.

To try to confirm this, first we can look at a boxplot of one of the fits with $K = 1$, and indeed if we look at Figure 3.3 we do see that the parameters seem a lot better contained. The group of β outliers could be explained by the fact the original density we are working with increases asymptotically on the left side (the side controlled by the α parameters) so the relative importance of the

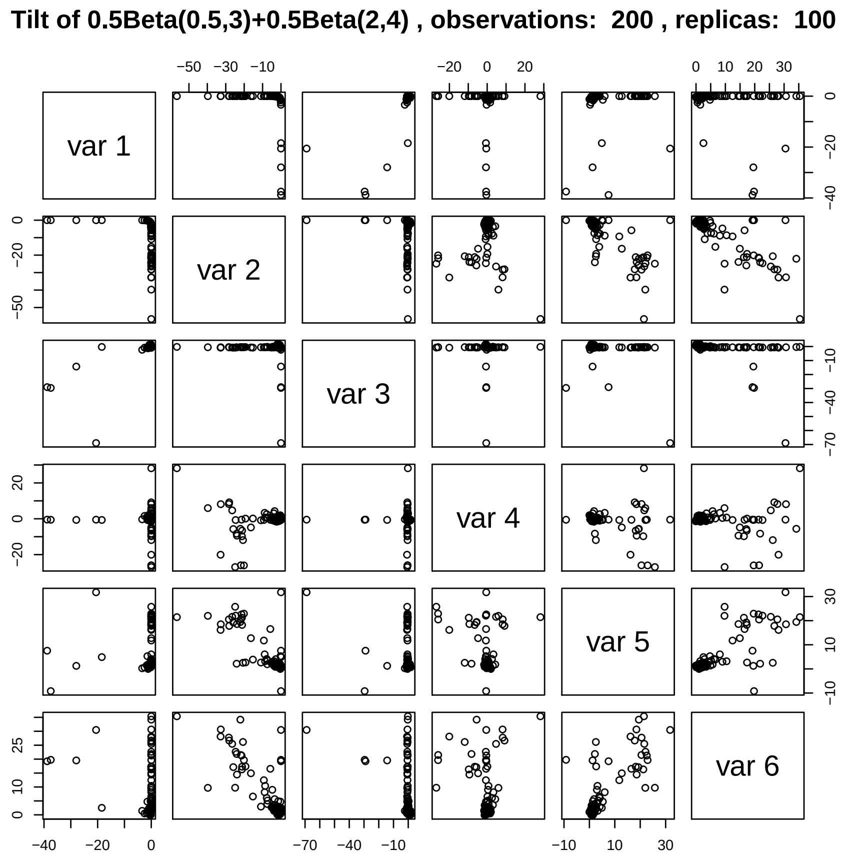


Figure 3.4: pairs plot of log MLE estimates (Nelder-Mead, maxit 500), for $K = 2$ and sample of 200 observations

right side (controlled by the β parameter) might be low.

Another check we can do is have a look at a pairs plot for $K = 2$, to see if low values of π_i correspond to very small/large values of α_i and β_i. In Figure 3.4, we have the pairs plot for $n = 200$ samples. Comparing Var2 (a π_i) vs Var6 (the corresponding β_i), we clearly see a clump down at the right corner, where stuff is happening according to plan, but also that as the weight of the component gets smaller, the value of the β parameter gets larger and more volatile. If we look at Var2 vs Var4 (the corresponding α_i), we also see a nice clump around 0 on the right side, where the weight of the component is still significant, and as the weight gets smaller, the α parameter gets more volatile, with a slight tendency for very small values.

To try and fight this phenomenon, we will try two things: the first is to use a constrained optimiser (L-BFGS-B) to keep all the log-parameters in a box of say, +/- 5, and the second is to first fit with a single component, and use those estimates, along with added random ones, as a starting point for a round of Nelder-Mead or BFGS (though, considering that the problem seems to be that the algorithms are overfitting one component in neglect of the other, this probably won't help, and method 1 might be the only course of action.)

3.2.2 Data simulated from other distributions

3.3 Numerical experiments

Much as in the previous section, we have fitted tilted beta mixtures to various simulated data, but this time we also estimate the marginal parameters.

3.4 Numerical example

Chapter 4

Conclusions

Bibliography

[1] S. Coles and J. A. Tawn, *Modelling Extreme Multivariate Events*, Journal of the Royal Statistical Society, 1991, pp. 381-382

[2] S. Coles, *An Introduction to Statistical Modelling of Extreme Values*, Springer Series in Statistics, 2001, p. 144

[3] M.-O. Boldi and A. C. Davison, *A mixture model for multivariate extremes*, Journal of the Royal Statistical Society Series B, 2007, pp.217-218

Code