Simulation Report about M-estimator

Zeyu Xiong

1 Theoretical explanation

Given d-dimensional sample $\mathbf{X} = (X_1, \ldots, X_d)$, let R_i^j denote the rank of X_{ij} among $X_{1j}, \ldots, X_{nj}$, $i = 1, \ldots, n$, $j = 1, \ldots, d$. For $k \in 1, \ldots, n$, then we can define the empirical tail dependence function estimator of l by

$$\begin{aligned}
\hat{l}_n(x) &= \hat{l}_{k,n}(x1, \ldots, x_d) \\
&:= \frac{1}{k}\sum_{i=1}^{n} \mathbf{1}\{R_i^1 > n + 0.5 - kx_1 \; or \ldots or \; R_i^d > n + 0.5 - kx_d\} \\
&= \frac{1}{k}\sum_{i=1}^{n} 1 - \mathbf{1}\{R_i^1 \le n + 0.5 - kx_1 \; and \ldots and \; R_i^d \le n + 0.5 - kx_d\} \\
&= \frac{n}{k} - \frac{1}{k}\sum_{i=1}^{n} 1\{x_1 \le \frac{n + 0.5 - R_i^1}{k} \; and \ldots and \; x_d \le \frac{n + 0.5 - R_i^d}{k}\}
\end{aligned} \tag{1.1}$$

The M-estimator is somehow similar to method of moment by defining the a integrable function vector $g = (g_1, \ldots, g_q)^T : [0,1]^d \to \mathrm{R}^q$ such that $\phi : \Theta \to \mathrm{R}^q$ defined by

$$\phi(\theta) := \int_{[0,1]^d} g(x) l(x; \Theta) \mathrm{d}x \tag{1.2}$$

is a homeomorphism between Θ and its image $\phi(\theta)$, where $\Theta \subset \mathrm{R}^p$ is the parametric family, and $q \ge p \ge 1$.

Let θ_0 denote the true parameter value, then the M-estimator $\hat{\theta}_n$ of θ_0 is defined as a minimizer of the following criterion function

$$\mathrm{Q}_{k,n}(\theta) = \|\phi(\theta) - \int g \hat{l}_n\|^2 = \sum_{m=1}^{q} \left(\int_{[0,1]^d} g_m(x) \Big(\hat{l}_n(x) - l(x;\theta) \Big) \mathrm{d}x \right)^2 \tag{1.3}$$

Intuitively speaking, $\phi(x)$ gives a measure of $l(x;\theta)$ using some simple functions $g(x)$ if we know the exact type of tail dependence function[1], and we can get the empirical measure of $hatl_n(x)$ using the same functions. Therefore, if the sample does come from the assumed assumption, naturally, we hope the distance between both measures to be as close as possible. Here, M-estimator uses frequently used squared error as the measure of distance so that it is convex which is easy to do minimization, it is also possible to use other measures.

[1] Usually, it is one of the assumption

2 Logistic Model

The theoretical stable tail dependence function of multivariate logistic distribution with standard Fréchet margins is given by

$$l(x_1,\dots,x_d;\theta) = (x_1^{1/\theta} + \cdots + x_d^{1/\theta})^\theta \tag{2.1}$$

where d is the dimensions, $x_1 > 0,\dots,x_d > 0$ and $\theta \in [0,1]$.

First, we observe how the M-estimator changes with difference choices of k and for difference function g, and we let $p = q = 1$ for simplicity. In detail, we look at 200 replications of samples of size $n = 1500$ and take the threshold parameters $k \in \{40, 80, \dots, 320\}$. In bivariate case, we compare $g(x_1,x_2) = 1$, $g(x_1,x_2) = x_1$, $g(x_2,x_2) = x_1 + x_2$, $g(x_1,x_2) = x_1^2 + x_2^2$ as choice for g. And we also include the censored MLE into comparison[2].

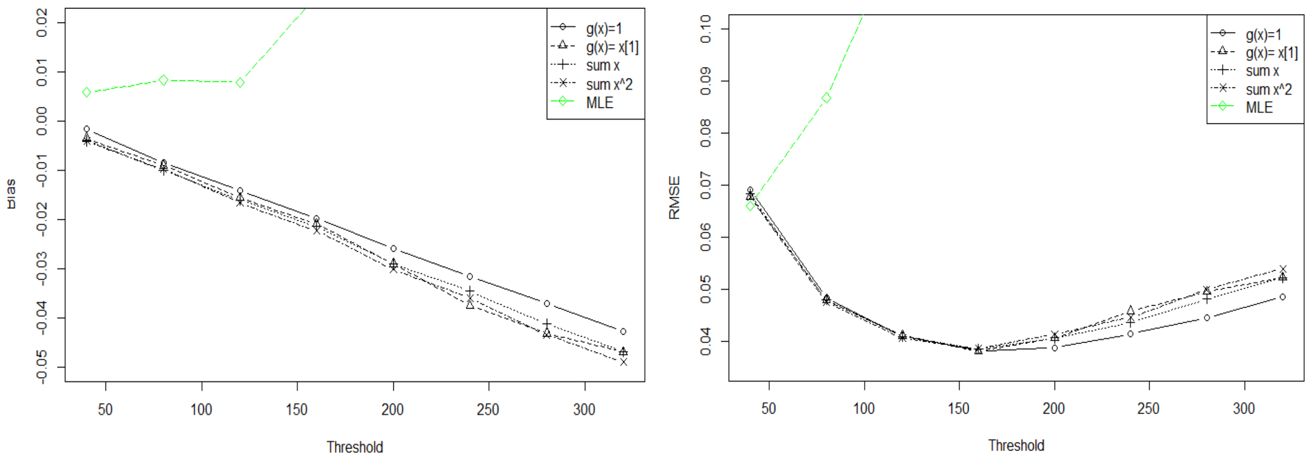

Figure 2.1: Logistic model: M-estimator for difference function g in dimension 2

All of the above choices for g result in similar finite-sample behavior of the estimator, but simpler function $g = 1$ leads to somewhat better performance. Because censured MLE doesn't converge well for some data, so the corresponding results are not charming. It turns out that censured MLE is very sensitive to the choice of threshold k, it gives bad result once k is too big. Therefore, by using smaller set of $k \in \{8, 21, 34, \dots, 100\}$, we are able to compare the performance of both methods. In general, M-estimator tends to underestimate the parameter, while censured MLE tends to overestimate the parameter, both of them are accurate enough. However, the RMSE of censured MLE grows quickly as we enlarge the threshold while M-estimator is constantly decreasing until the threshold reaches to 160.

Five-dimension Result We also show the result in five dimensions, but in a different format. We simulate 200 samples of size $n = 1500$ from five-dimensional logistic distribution function with $\theta_0 = 0.5$. The bias and the RMSE of this estimator are shown in the upper panels of figure 2.3.

Also, we consider the estimation of $l(1,1,1,1,1;\theta)$. From 2.1 it follows that $l(1,1,1,1,1;\theta) = 5^\theta$, so the estimated quantity is then $5^{\hat{\theta}}$. Since $\theta_0 = 0.5$, so the true quantity is $\sqrt{5}$. We also include the bias and RMSE of this quantity in the lower panel.

[2]Note that for some generated data, censured MLE method does not converge at all, so the result may be unconvincing.

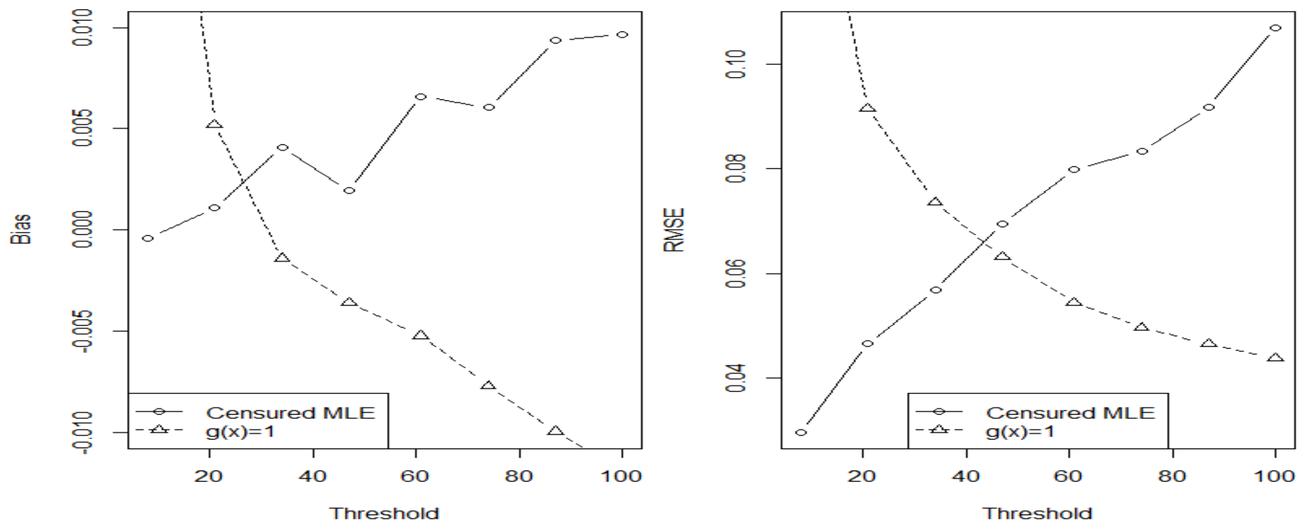

Figure 2.2: Logistic model: Comparison between M-estimator and Censured MLE

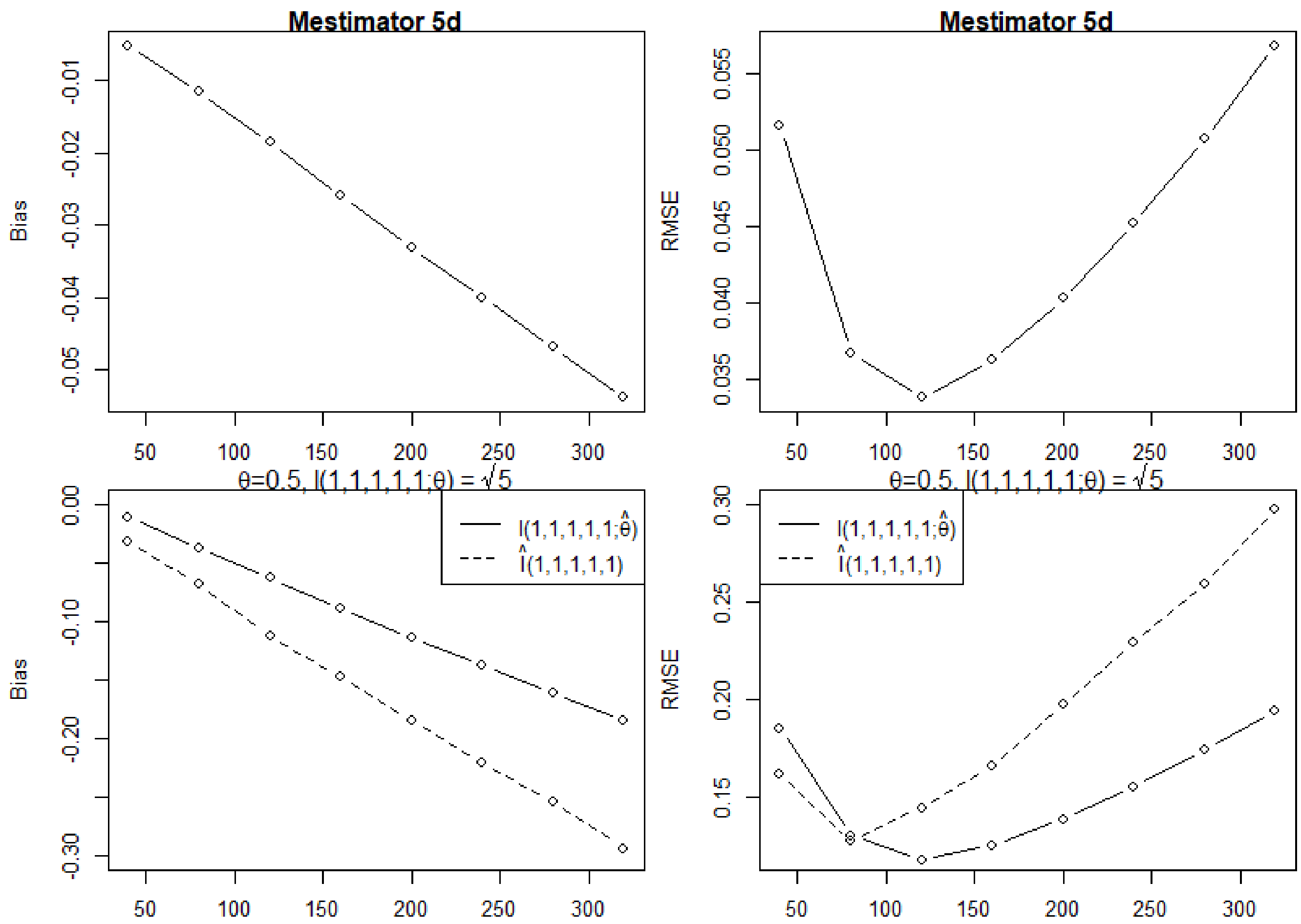

Figure 2.3: Logistic model: M-estimator in five-dimension

3 Factor Model

Consider the r-factor model, $r \in \mathbb{N}$, in dimension d: $X = (X_1, X_2, \ldots, X_d)$ and

$$X_j = \sum_{i=1}^{r} a_{ij} Z_i + \epsilon_j, \quad j \in \{1, \ldots, d\}$$

with Z_i independent Fréchet(υ) random variables, $\upsilon > 0$, and with a_{ij} non-negative constants. For purpose of studying tail properties, it is more convenient to consider the max factor model

$$X_j = \max_{i=1,\ldots,r} \{a_{ij} Z_i\}, \quad j \in \{1,\ldots,d\}$$

Note that they have the same tail dependence function l. Then in this case, its stable tail dependence function is given by

$$l(x_1,\ldots,x_d) = \sum_{i=1}^{r} \max_{j=1,\ldots,d} \{b_{ij} x_j\} \tag{3.1}$$

where $b_{ij} = a_{ij}/\sum_{i=1}^{r} a_{ij}$. Similarly, according to 1.3, we are able to estimate the coefficient of each factor composing $X_i, i = 1,\ldots,d$.

Simulation result: Four-dimensional model with two factors. We simulate 200 samples of size $n = 5000$ from a four-dimension model:

$$\begin{aligned} X_1 &= 0.2Z_1 \vee 0.8Z_2 \\ X_2 &= 0.5Z_1 \vee 0.5Z_2 \\ X_3 &= 0.7Z_1 \vee 0.3Z_2 \\ X_4 &= 0.9Z_1 \vee 0.1Z_2 \end{aligned}$$

with Z_1 and Z_2 being independent standard $Fréchet$ facotrs, and thus $\theta_0 = (0.2, 0.5, 0.7, 0.9)$.

In Figure 3.1, we show the bias and the RMSE of the M-estimator based on $q = 5$ moment equations, with auxiliary function $g_i(x) = x_i$ for $i = 1,2,3,4$ and $g_5 = 1$. Note that, since we have more parameters to estimate, it is reasonable to include more data into estimation, and correspondingly, the threshold should change as well. As we can see, this M-estimator performs fairly well when k is approximately 300, both bias and RMSE are in a reasonable range.

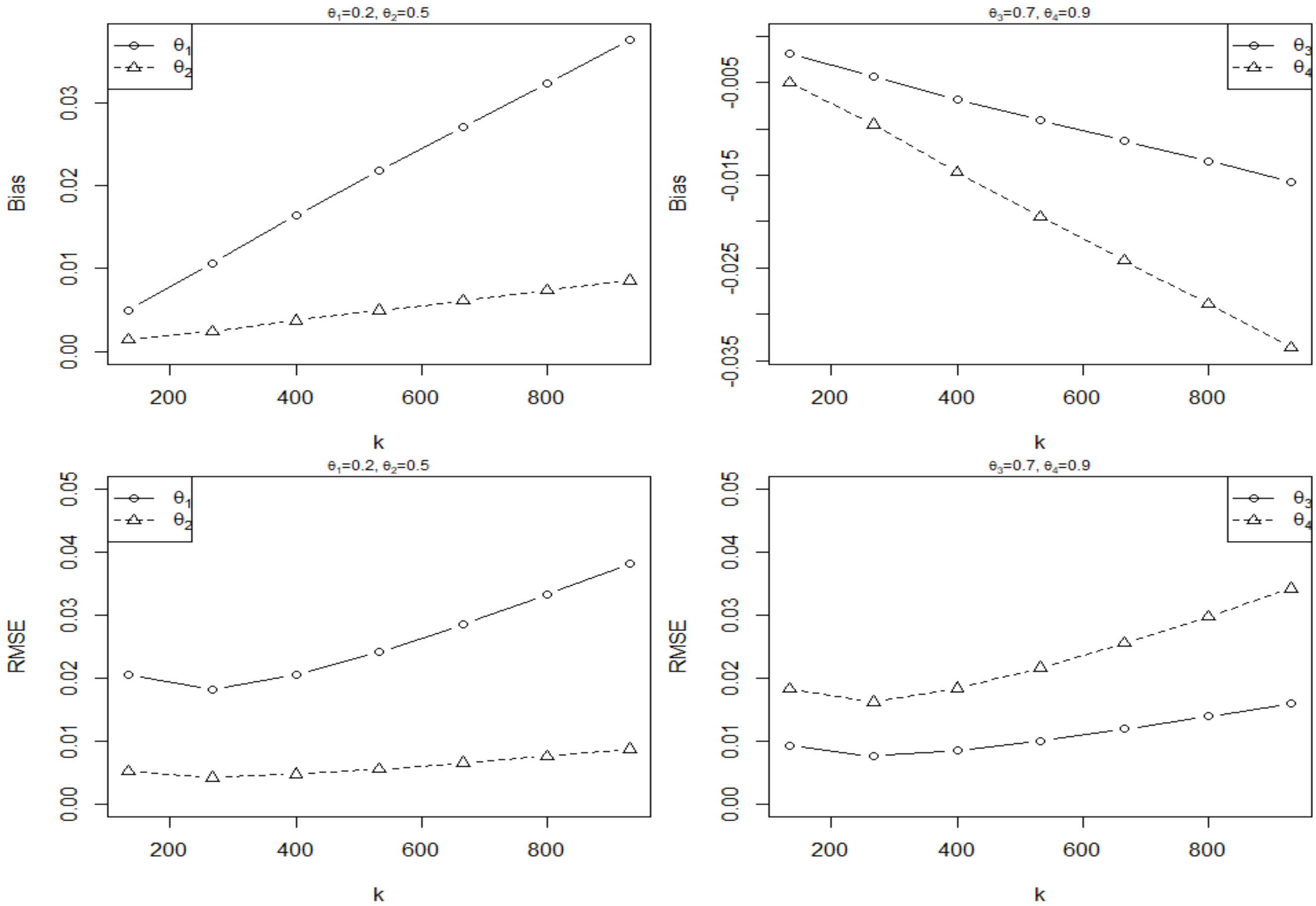

Figure 3.1: Four-dimensional 2-factor model, estimation of $theta = (0.2, 0.5, 0.7, 0.9)$