Let's try it.

```
library(evd)
# simulate data with alpha = 0.5
alpha <- 0.5
z <- rmvevd(500, dep = alpha, d = 5, mar = c(1, 1, 1))

## bivariate density
d_bivar <- function(z, alpha){
    #here "z" is a single observation (ordered pair)
    inside <- z[1]^(-1/alpha) + z[2]^(-1/alpha)
    one <- exp(-inside^alpha)
    two <- (z[1]*z[2])^(-1 / alpha - 1)
    three <- (1 / alpha - 1)*inside^(alpha - 2)
    four <- inside^(2 * alpha - 2)
    one*two*(three + four)
}

d_bivar(c(4, 5), alpha = alpha)
```

```
## [1] 0.003650963
```

```
dmvevd(c(4,5), dep = alpha, d = 2, mar = c(1,1,1))
```

```
## [1] 0.003650963
```

```
## estimate alpha
log_pair_lhood <- function(alpha, z) {
    #here "z" is bivariate matrix of observations
    inside <- z[, 1]^(-1 / alpha) + z[, 2]^(-1 / alpha)
    log_one <- -inside^alpha
    log_two <- (-1 / alpha - 1) * (log(z[, 1]) + log(z[,
2]))
    three <- (1 / alpha - 1) * inside^(alpha - 2)
    four <- inside^(2 * alpha - 2)
    contrib <- log_one + log_two + log(three + four)
    return(sum(contrib))
}
```

get all pairwise likelihoods and sum (only allows pairwise dependence)

```
        all_pairs_lhood <- function(alpha, z) {
expand.grid(dim1 = seq_len(ncol(z)),                          dim2 =
seq_len(ncol(z))) |>                    filter(dim1 < dim2) |>
rowwise() |>                  mutate(log_pair_lhood =
log_pair_lhood(alpha, cbind(z[, dim1], z[, dim2]))) |>
ungroup() |>                  summarise(res = sum(log_pair_lhood))
|>                    pull(res)          }                alpha_mple <-
optim(.2, lower = .01, upper = .99, all_pairs_lhood, z = z, method =
"Brent", hessian = TRUE, control = list(fnscale = -1))
        (ci_mple <- alpha_mple$par + c(-1.96, 1.96)*sqrt(-1 /
alpha_mple$hessian[1, 1]))
```

$\hat{\alpha}_{MPLE}$

← true $\alpha = 0.5$

```
## [1] 0.4954979 0.5182678
```

Seems ok?

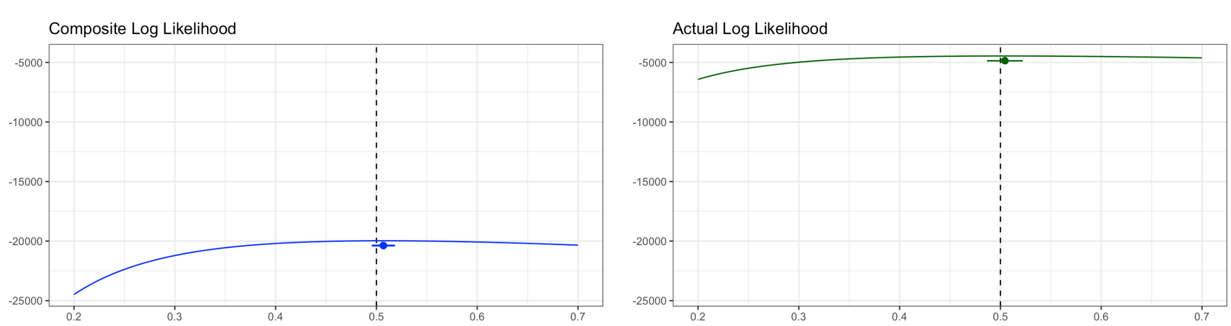
```
        ## checking coverage
        #checking coverage
        B <- 200
        coverage <- numeric(B)
        for(k in seq_len(B)) {
            z_k <- rmvevd(500, dep = .5, d = 5, mar = c(1, 1, 1))
            alpha_mple_k <- optim(.2, lower = .01, upper = .99,
all_pairs_lhood, z = z_k, method = "Brent", hessian = TRUE, control =
list(fnscale = -1))
            ci <- alpha_mple_k$par + c(-1.96, 1.96)*sqrt(-1 /
alpha_mple_k$hessian[1, 1])
            coverage[k] <- as.numeric(ci[1] < alpha & ci[2] > alpha)
        }
        mean(coverage)
```

generate data

calculate CI based on generated data k

does $CI_k$ contain the true $\alpha$?

```
## [1] 0.745
```




↑ this dsn has a sharper curve than this one

⟹ narrower interval ⟹ low empirical coverage it is.

So, it looks like the point estimate from the pairwise likelihood is ok, but we need to be able to get an appropriate measure of uncertainty.

CI:

Recall if $\hat{\theta}_{MLE}$ is the estimate from the correct model and $\underline{\theta}$ is the true value of the parameter, then

$$\sqrt{n}\left(\hat{\underline{\theta}}_{MLE} - \underline{\theta}\right) \xrightarrow{d} N\left(0, I(\underline{\theta})^{-1}\right)$$

So for fixed, large $n$, $\hat{\theta}_{MLE} \overset{\cdot}{\sim} N\left(\underline{\theta}, \frac{1}{n} I(\underline{\theta})^{-1}\right)$

where $\underline{I}(\underline{\theta}) = E\left[\left(\frac{\partial}{\partial \underline{\theta}^T} \log f(y_1, \underline{\theta})\right)\left(\frac{\partial}{\partial \underline{\theta}} \log f(y_1, \underline{\theta})\right)\right]$ "variance of the score"

If this is the correct model! $= E\left[-\frac{\partial^2}{\partial \underline{\theta}\, \partial \underline{\theta}^T} \log f(y_1, \underline{\theta})\right]$ "hessian of score contribution"

In practice, with the correct model,

$\frac{1}{n} I(\underline{\theta})^{-1} = \left[n I(\underline{\theta})\right]^{-1}$ & $n I(\underline{\theta})$ is approximated with $n \underline{I}(\hat{\underline{\theta}}_{MLE}) = -\frac{\partial^2 \ell(\hat{\underline{\theta}}_{MLE})}{\partial \underline{\theta}\, \partial \underline{\theta}^T}$

This is wrong in the misspecified case!

The proper adjustment is

(A.C. Davidson, Statistical Models, p. 147).

$$\hat{\underline{\theta}}_{EE} \overset{\cdot}{\sim} N\Big(\underline{\theta}, \underbrace{I(\underline{\theta})^{-1} K(\underline{\theta}) I(\underline{\theta})^{-1}}_{\text{"sandwich estimator"}}\Big)$$

"estimating equations"

where $K(\underline{\theta}) = n E\left[\left(\frac{\partial}{\partial \underline{\theta}^T} \log f_p(y_1; \underline{\theta})\right)\left(\frac{\partial}{\partial \underline{\theta}} \log f_p(y_1, \underline{\theta})\right)\right]$

and $I(\theta) = -n E\left[\frac{\partial^2}{\partial \underline{\theta}\, \partial \underline{\theta}^T} \log f_p(y_1; \underline{\theta})\right]$

where $f_p$ is the pairwise density.

the incorrectly specified model.

We will approach this from a more general discussion of M-estimator/estimating equations. (not just pairwise).

# 1 Introduction

There are 2 parts of a fully specified statistical model:

1. Systematic part (mean) used for answering the underlying scientific question
2. distributional assumptions about the random part of the model

$\Rightarrow$ likelihood inference.

We want to develop robust inference so that misspecification of 2 doesn't invalidate the inference.

$\Rightarrow$ we want to define an estimator of interest as the solution to some "estimating equations" equation, but it might not come from the derivative of a log likelihood.

M-estimators are solutions of the vector equation

$$\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{Y}_i, \boldsymbol{\theta}) = \mathbf{0},$$ i.e. if $\hat{\underline{\theta}}$ is an "M-estimator", $\sum_{i=1}^{n} \underline{\psi}(\underline{Y}_i, \hat{\underline{\theta}}) = \underline{0}$.

Notes:

$Y_i$ are independent (not necessarily i.d. e.g. regression).

$\underline{\theta}$ is a $b$-dim. parameter

$\underline{\psi}$ is a known $b \times 1$ function that does not depend on $i$ or $n$, but can depend on $\underline{x}_i$ for regression.

regression: above equation $\sum_{i=1}^{n} \underline{\psi}(Y_i, \underline{x}_i, \underline{\theta}) = \underline{0}$.

In the likelihood setting, what is $\boldsymbol{\psi}$?

$\underline{\psi}$ is the derivative of the log-likelihood (the score function).

---

There are 2 types of M-estimators.

1. $\rho$-type: solutions $\underline{\theta}$ which minimize $\sum_{i=1}^{n} \rho(Y_i, \underline{\theta})$
2. $\psi$-type: solutions $\underline{\theta}$ to $\sum_{i=1}^{n} \underline{\psi}(\underline{Y}_i, \underline{\theta}) = \underline{0}$.

Often an M-estimator is of both types, i.e. if $\rho$ has a continuous first derivative wrt $\underline{\theta}$, then an M-estimator of $\psi$-type is an M-estimator of $\rho$-type with $\underline{\psi}(y, \underline{\theta}) = \nabla_\theta \rho(y; \underline{\theta})$.

**Example:** Let $Y_1, \ldots, Y_n$ be independent, univariate random variables. Is $\theta = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ an M-estimator?

① $\psi$-type?

$$\theta = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\Rightarrow 0 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \theta = \sum_{i=1}^{n} \frac{1}{n}(Y_i - \theta) = \sum_{i=1}^{n}(Y_i - \theta) \Rightarrow \psi(Y_i, \theta) = Y_i - \theta$$

② $\rho$-type? What does the sample mean minimize?

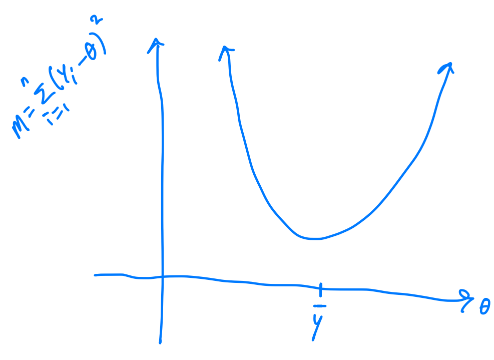
$$M = \sum_{i=1}^{n}(Y_i - \theta)^2 = \sum_{i=1}^{n}\rho(Y_i, \theta)$$

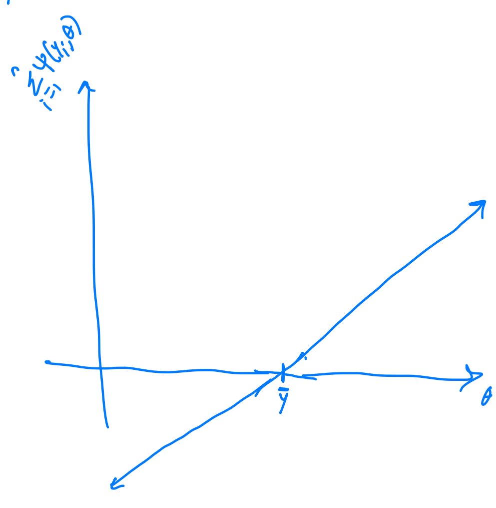
$$= \sum_{i=1}^{n} Y_i^2 - 2\theta\sum_{i=1}^{n} Y_i + n\theta^2$$

To minimize,

$$\frac{\partial M}{\partial \theta} = -2\sum_{i=1}^{n} Y_i + 2n\theta \overset{\text{set}}{=} 0$$

$$\theta = \frac{1}{n}\sum_{i=1}^{n} Y_i$$




We will mainly focus on $\psi$-type M-estimators because it is more straight forward to get our sandwich estimator matrix.

But can be useful to think of underlying $\rho$-type estimator you had to take derivative to get $\psi$-type.

MAD

**Example:** Consider the mean deviation from the sample mean, A measure of spread.

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \overline{Y}|.$$

Is this an M-estimator?

To calculate this $\hat{\theta}_1$ requires 2 steps:

1. Calculate $\bar{Y}$
2. Calculate MAD

$\Longrightarrow$ no single equation of the form $\sum_{i=1}^{n} \psi(Y_i, \theta) = 0$ can be formed.

But a system of equations of $\psi$-type can be written!

Let $\theta_2 = \bar{Y}$

$\psi_2(y, \theta_2) = y - \theta_2$

$\psi_1(y, \theta_1, \theta_2) = |y - \theta_2| - \theta_1$

So $\underline{\hat{\theta}} = (\hat{\theta}_1, \hat{\theta}_2)$ will solve

$$\sum_{i=1}^{n} \underline{\psi}(Y_i, \hat{\theta}_1, \hat{\theta}_2) = \begin{pmatrix} \sum_{i=1}^{n}\{|Y_i - \hat{\theta}_2| - \hat{\theta}_1\} \\ \sum_{i=1}^{n}(Y_i - \hat{\theta}_2) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Even though at first the MAD doesn't look like an M-estimator, with a little work we can write it as one.

# 2 Basic Approach (the theory).

M-estimators are solutions of the vector equation
$\hat{\underline{\theta}}$

$$\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{Y}_i, \boldsymbol{\theta}) = \mathbf{0}.$$

iid case.

but what are they estimating? Some true parameter $\underline{\theta}_0$, where

(*) $E_F[\underline{\Psi}(Y_1, \underline{\theta}_0)] = \int \underline{\Psi}(y; \underline{\theta}_0)\, dF(y) = \underline{0}$ where $Y_1 \sim F$.

**Example (Sample Mean, cont'd):** Recall we said $\theta = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is an M-estimator for $\psi(Y_i, \theta) = Y_i - \theta$. What is the true parameter?

The true parameter solves $\int (y - \theta_0)\, dF(y) = 0$

$$\Rightarrow \underbrace{\int y\, dF(y)}_{} = \theta_0$$

This agrees w/ the definition of the population mean.

Recall the 5-dimensional motivating example. We said the $\alpha$ which maximizes the pairwise likelihood seems like it would be a good estimator for $\alpha_0$.

We didn't show this. To do so, we would need to use (*).

Try this at home.