"Misspecified Models" "M-estimation"

Estimating Equations

Now we will consider "robustifying" inference so that misspecification does not invalidate our resulting inference.

Motivating **Example:** Consider the $\boldsymbol{Z} = (Z_1, \ldots, Z_5)^\top$ with cdf

$$F(\boldsymbol{z}; \alpha) = \exp\left\{ -\left(z_1^{-\frac{1}{\alpha}} + z_2^{-\frac{1}{\alpha}} + z_3^{-\frac{1}{\alpha}} + z_4^{-\frac{1}{\alpha}} + z_5^{-\frac{1}{\alpha}} \right)^{\alpha} \right\}, \quad \boldsymbol{z} \geq \mathbf{0}, \alpha \in (0,1].$$

If $\alpha = 1$ independence

$\alpha \to 0$ complete dependence ($Z_i = Z_j$ w.p. 1).

Marginal:

$$P(Z_i \leq z) = \exp\left[-(z^{-1/\alpha})^{\alpha} \right] = \exp(-z^{-1})$$

"Unit Frechet"

Comments:

1. F is "max-stable." → suitable for multivariate extreme value data

defn $[F(nz)]^n = F(z)$

$$\begin{aligned}
[F(nz)]^n &= \left(\exp\left[-\{ (nz_1)^{-1/\alpha} + \ldots + (nz_5)^{-1/\alpha} \}^{\alpha} \right] \right)^n \\
&= \left(\exp\left[-\{ n^{-1/\alpha} (z_1^{-1/\alpha} + \ldots + z_5^{-1/\alpha}) \}^{\alpha} \right] \right)^n \\
&= \left(\exp\left[-n^{-1} (z_1^{-1/\alpha} + \ldots + z_5^{-1/\alpha})^{\alpha} \right] \right)^n \\
&= \exp\left[-(z_1^{-1/\alpha} + \ldots + z_5^{-1/\alpha})^{\alpha} \right] \;//
\end{aligned}$$

2. $Z_1, \ldots, Z_5$ are exchangeable. order doesn't matter

$P(Z_1, \ldots, Z_5) = P(Z_3, Z_2, Z_4, Z_5, Z_1)$ etc.

Realistic? Maybe not.

But this gives us equal pairwise dependence ⇒ which can help reduce # parameters.

↳ and illustrate the concept of an estimating equation.

Let's consider the likelihood.

Suppose we observe $\underline{z}_i = (z_{i1}, \dots, z_{i5})^T$, $i = 1, \dots, n$ iid $\sim F$. We want to estimate α.

We need to find the density, i.e. $\dfrac{\partial^5 F}{\partial z_1 \cdots \partial z_5}$

$$\frac{\partial F}{\partial z_1} = \exp\left[-\left(z_1^{-1/\alpha} + \dots + z_5^{-1/\alpha}\right)^{\alpha}\right] \times \left\{-\alpha\left(z_1^{-1/\alpha} + \dots + z_5^{-1/\alpha}\right)^{\alpha-1}\right\} \times \left\{-\frac{1}{\alpha} z_1^{-\frac{1}{\alpha}-1}\right\}$$

$$\frac{\partial^2 F}{\partial z_1 \partial z_2} \overset{\text{product rule}}{=} \exp\left[-\left(z_1^{-1/\alpha} + \dots + z_5^{-1/\alpha}\right)^{\alpha}\right] \times \left\{-\alpha\left(z_1^{-1/\alpha} + \dots + z_5^{-1/\alpha}\right)^{\alpha-1}\right\}^2 \times \left\{-\frac{1}{\alpha} z_2^{-\frac{1}{\alpha}-1}\right\} \times \left\{-\frac{1}{\alpha} z_1^{-\frac{1}{\alpha}-1}\right\}$$
$$+ \exp\left[-\left(z_1^{-1/\alpha} + \dots + z_5^{-1/\alpha}\right)^{\alpha}\right] \times \left\{-\alpha(\alpha-1)\left(z_1^{-1/\alpha} + \dots + z_5^{-1/\alpha}\right)^{\alpha-2}\right\} \times \left\{-\frac{1}{\alpha} z_2^{-\frac{1}{\alpha}-1}\right\} \times \left\{-\frac{1}{\alpha} z_1^{-\frac{1}{\alpha}-1}\right\}$$

$$\frac{\partial^3 F}{\partial z_1 \partial z_2 \partial z_3} = \text{product rule on each of the 2 terms} \longrightarrow 4 \text{ terms.}$$

by the time we get to $\dfrac{\partial^5 F}{\partial z_1 \cdots \partial z_5}$ things are gross just to write the likelihood!

How about if we were to just use pairs of points to estimate α?

$$F_{Z_1 Z_2}(z_1, z_2) = \exp\left[-\left(z_1^{-1/\alpha} + z_2^{-1/\alpha}\right)^{\alpha}\right]$$

$$\frac{\partial^2 F}{\partial z_1 \partial z_2} = \exp\left[-\left(z_1^{-1/\alpha} + z_2^{-1/\alpha}\right)^{\alpha}\right] (z_1 z_2)^{-\frac{1}{\alpha}-1} \left\{ \left(\frac{1}{\alpha}-1\right)\left(z_1^{-1/\alpha} + z_2^{-1/\alpha}\right)^{\alpha-2} + \left(z_1^{-1/\alpha} + z_2^{-1/\alpha}\right)^{2\alpha-2} \right\}.$$

If we just used $(z_{1i}, z_{2i}), i = 1, \ldots, n$ would the likelihood based on the bivariate density be a good estimator for α?

Yes: unbiased

No: inefficient (not using all data).

What if we took all $\binom{5}{2} = 10$ pairs? (z_{1i}, z_{2i}), (z_{1i}, z_{3i}), ...

Yes: unbiased, efficient (using all data).

No: It's not the right likelihood!

Composite likelihood.

Let's try it.

```r
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
```

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

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

```r
## 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))
}

all_pairs_lhood <- function(alpha, z) {
```

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

```
    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]))
```

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

```
## 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

get MLE

create CI

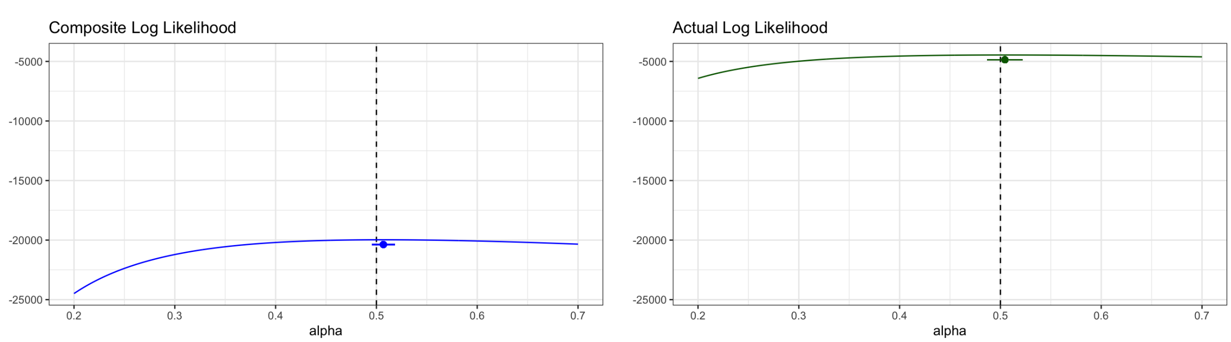
95%

← did CI contain truth?

← want to be close to .95

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

uh oh!



this has a sharper curve than this one

⇒ narrower interval !!

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.

it is

CI:

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

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

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

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\underline{\theta}^T}\log f(Y_1,\underline{\theta})\right)\left(\frac{\partial}{\partial\underline{\theta}}\log f(Y_1,\underline{\theta})\right)\right]$$ "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 models,

$\frac{1}{n} I(\underline{\theta})^{-1} = [nI(\underline{\theta})]^{-1}$ & $nI(\underline{\theta})$ approximated w/ $nI(\hat{\underline{\theta}}_{MLE}) = \dfrac{-\partial^2 \ell(\hat{\underline{\theta}}_{MLE})}{\partial\underline{\theta}\,\partial\underline{\theta}^T}$

This is wrong in the misspecified case!

The proper adjustment is

A.C. Davidson, Statistical models pg. 147.

$$\hat{\underline{\theta}}_{EE} \dot\sim N\Big(\underline{\theta}, \underbrace{I(\underline{\theta})^{-1} K(\underline{\theta}) I^{-1}(\underline{\theta})}\Big)$$

"estimating equation"

"Sandwich estimator"

bread meat bread

where $I(\theta) = -n\, E\left[\frac{\partial^2}{\partial\underline{\theta}\partial\underline{\theta}^T}\log f_p(Y_1,\underline{\theta})\right]$

$K(\theta) = n\, E\left[\left(\frac{\partial}{\partial\underline{\theta}^T}\log f_p(Y_1;\underline{\theta})\right)\left(\frac{\partial}{\partial\underline{\theta}}\log f_p(Y_1;\underline{\theta})\right)\right]$

where f_p is the pairwise density. (the incorrectly specified model).

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

1 Introduction

There are 2 parts of a fully specified statisical 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.

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

M-estimators are solutions of the vector equation

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

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

known $b \times 1$ function does not depend on n or i. $\quad$ b-dim parameter

Notes

Y_i are independent (not necessarily iid, e.g. regression).

For regression, ψ can depend on x_i

$$\sum_{i=1}^{n} \underline{\psi}(Y_i, x_i, \underline{\theta}) = \underline{0}.$$

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

ψ is the derivative of the log likelihood contribution (the score contribution).

There are 2 types of M-estimators:

(1) ψ-type: solutions $\underline{\theta}$ to $\sum_{i=1}^{n} \underline{\psi}(Y_i, \underline{\theta}) = 0$

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

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

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 ψ- 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 ρ - type? What does the sample mean minimize?

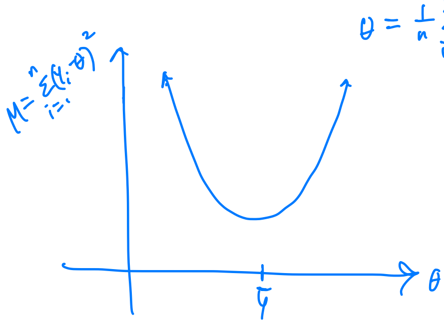
$$M = \sum_{i=1}^{n} (Y_i - \theta)^2 = \sum_{i=1}^{n} \rho(Y_i, \theta)$$

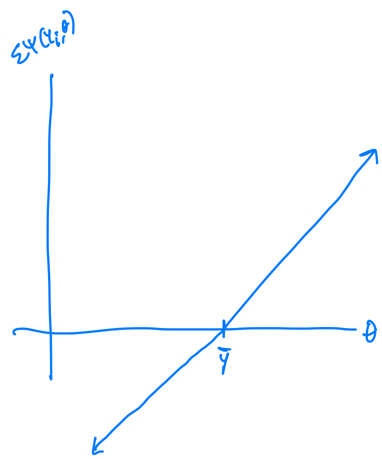
$$= \sum_{i=1}^{n} Y_i^2 - 2\theta \sum_{i=1}^{n} Y_i + n\theta^2$$

To minimize,

$$\frac{dM}{d\theta} = -2\sum_{i=1}^{n} Y_i + 2n\theta \stackrel{\text{set}}{=} 0$$

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





We will mainly focus on ψ-type M-estimators - because its more straightforward to get the sandwich estimator.

But it can be useful to think of an underlying ρ-type estimator.

Example: Consider the mean deviation (MAD) 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 $\hat{\theta}_1$, requires 2 steps:

① calculate $\overline{Y}$

② calculate MAD $\Rightarrow$ no single equation of the for $\sum_{i=1}^{n} \psi(Y_i, \theta) = 0$ can be found.

But a system of equations of ψ-type can be written.

let $\hat{\theta}_2 = \overline{Y}$

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

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

So $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$, 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 MAD doesn't look like an M-estimator, with a little work we can write it as one.