NMST 434 - Ex. session 6. M-estimators

$X_1,\dots,X_n$ random sample from F with unknown parameter $\theta\in\Theta\subset\mathbb{R}^p$.

$$\hat\theta = \underset{\theta\in\Theta}{\operatorname{argmin}}\ \frac{1}{n}\sum_{i=1}^{n}\rho(X_i,\theta) \qquad \text{M-estimator}$$

$$0_p = \sum_{i=1}^{n}\psi(X_i,\hat\theta) \qquad \text{Z-estimator} \quad \left(\text{often } \psi(x,t)=\frac{\partial\rho(x,\theta)}{\partial\theta}\right)$$

$\hat\theta$ estimates $\theta_X = \underset{\theta}{\operatorname{argmin}}\ E\rho(X_1,\theta)$ and $0_p = E\psi(X_1,\theta_X)$, respectively.

Under [Z0]-[Z6], $\Gamma(\theta)=E D_\psi(X_1,\theta)$, $\Sigma(\theta_X)=E\left(\psi(X_1,\theta_X)\psi(X_1,\theta_X)^T\right)$ from Theorem 9

$$\sqrt{n}(\hat\theta-\theta_X) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma^{-1}(\theta_X)\psi(X_i,\theta_X)+o_P(1), \text{ i.e. } \sqrt{n}(\hat\theta-\theta_X)\xrightarrow[n\to\infty]{d} N_p\left(0_p,\Gamma^{-1}(\theta_X)\Sigma(\theta_X)\Gamma^{-1}(\theta_X)\right).$$

Ex. $N(\mu,1)$ MLE $L(\mu)=c\cdot\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right)$, $\ell(\mu)=c'-\frac{1}{2}\sum(x_i-\mu)^2$, $\rho(x,t)=(x-t)^2$

i.e. $\psi(x,t)=\frac{\cancel{2}}{2}(x-t)$ and $\bar X$ is an M-estimator of $E(X_1-\theta_X)=0$, $\theta_X=EX=\mu$.
Now if the true distribution is not $N(\mu,1)$, we get $ED_\psi(X_1,\theta)=-1=\Gamma(\mu)$,
$\operatorname{var}\psi(X_1,\theta_X)=\operatorname{var}(X_1-\mu_X)=\operatorname{var}X_1$ and $\sqrt{n}(\bar X-\mu)\xrightarrow[n\to\infty]{d}N(0,\operatorname{var}X_1)$ no matter what the true distribution of X_1 is. (CLT.) By Example 37 this choice of $\theta_X=EX$ minimizes the Kullback-Leibler divergence of the family $\{N(\mu,1),\mu\in\mathbb{R}\}$ from the true distribution F of X_1, i.e. $\theta_X=\underset{\theta}{\operatorname{argmin}}\int_{\mathbb{R}}\log\left[\frac{f(x)}{f(x,\theta)}\right]\cdot f(x)\,d\mu(x)$. This holds generally for M-estimators given as MLE in systems of densities. $L(\theta)$ likelihood, $\rho(x,\theta)=-\ell(x,\theta)$, $\psi(x,\theta)=\frac{\partial\ell(x,\theta)}{\partial\theta}$.

Ex. $N(\mu,\sigma^2)$ $\ell(\mu,\sigma^2)=-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2$, $\theta=(\mu,\sigma^2)^T$, $\rho(x,\theta)=\frac{(x-\mu)^2}{\sigma^2}+\log\sigma^2$

$\psi(x,\theta)=\begin{pmatrix}x-\mu\\(x-\mu)^2-\sigma^2\end{pmatrix}$ solves $\hat\mu=\bar X$, $\hat\sigma^2=\frac{1}{n}\sum(X_i-\bar X)^2$, population $E(X-\mu)=0\Rightarrow\mu_X=EX$, $E(X-\mu)^2=\sigma^2\Rightarrow\sigma_X^2=\operatorname{var}X$

$$ED_\psi(X_1,\theta_X)=E\begin{pmatrix}-1&0\\-2(X_1-\mu_X)&-1\end{pmatrix}=-I_{2\times2}=\Gamma(\theta_X) \qquad \Sigma(\theta_X)=\begin{pmatrix}\sigma_X^2 & E(X-\mu_X)^3\\ E(X-\mu_X)^3 & E(X-\mu_X)^4-\sigma^4\end{pmatrix}$$

$E(X-\mu_X)((X-\mu_X)^2-\sigma_X^2)$
$=E(X-\mu_X)^3$
$E[(X-\mu_X)^2-\sigma^2]^2=E(X-\mu_X)^4-\sigma^4$

$$\sqrt{n}\left(\begin{pmatrix}\bar X\\ \hat\sigma^2\end{pmatrix}-\begin{pmatrix}\mu_X\\ \sigma_X^2\end{pmatrix}\right)\xrightarrow{d}N_2\left(0_2,\begin{pmatrix}\operatorname{var}X_1 & E(X_1-\mu_X)^3\\ E(X_1-\mu_X)^3 & E(X_1-\mu_X)^4-\sigma^4\end{pmatrix}\right) \quad \text{again CLT.} \quad (*)$$

$$=N_2\left(0_2,\begin{pmatrix}\operatorname{var}X_1 & 0\\ 0 & 2\sigma^4\end{pmatrix}\right) \quad \text{under normality.}$$

MLE: for the choice $\psi_2(x,\theta)=\begin{pmatrix}\frac{x-\mu}{\sigma^2}\\ \frac{(x-\mu)^2}{2\sigma^4}-\frac{1}{2\sigma^2}\end{pmatrix}$ one gets the same solution $\hat\mu=\bar X$, $\hat\sigma^2=\frac{1}{n}\sum(X_i-\bar X)^2$

but $\Gamma(\theta_X)=\begin{pmatrix}\frac{1}{\sigma^2}&0\\0&\frac{1}{2\sigma^4}\end{pmatrix}$, $\Sigma(\theta_X)=\begin{pmatrix}\frac{1}{\sigma^2}&\frac{E(X_1-\mu)^3}{2\sigma^3}\\ \frac{E(X_1-\mu)^3}{2\sigma^3}&\frac{E(X_1-\mu)^4-\sigma^4}{4\sigma^8}\end{pmatrix}$ (under normality $\Sigma(\theta_X)=\Gamma(\theta_X)=I(\theta_X)$ as follows from general theory) Fisher's inform. matrix

and one gets the sandwich estimator of the as. variance of $(\bar X,\hat\sigma^2)^T$

$\Gamma^{-1}(\theta_X)\Sigma(\theta_X)\Gamma^{-1}(\theta_X)$ with θ_X replaced by $\hat\theta$. This estimator holds true also under misspecification and is the same as in (*).

Ex: M estimators and Δ-theorem. Inference also about σ. Parameter (μ, σ^2, θ) for $\theta = \sqrt{\sigma^2}$.

Estimating function $\psi(\mu,\sigma^2,\theta) = \begin{pmatrix} x-\mu \\ (x-\mu)^2-\sigma^2 \\ \sqrt{\sigma^2}-\theta \end{pmatrix}$

$\theta_x = \sigma_x$ estimating equations $\Leftrightarrow$ MLE which is invariant to reparametrization

$$\Gamma(\mu,\sigma^2,\theta) = E\begin{pmatrix} -1 & 0 & 0 \\ -2(X-\mu) & -1 & 0 \\ 0 & \frac{1}{2\sqrt{\sigma^2}} & -1 \end{pmatrix} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & \frac{1}{2\sigma} & -1 \end{pmatrix}, \quad \Gamma^{-1}(\mu,\sigma^2,\theta) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & -1/2\sigma & -1 \end{pmatrix}$$

$\mu_i = E(X-\mu)^i$

$$\Sigma(\mu,\sigma^2,\theta) = \operatorname{var}\begin{pmatrix} X-\mu \\ (X-\mu)^2-\sigma^2 \\ \sigma-\theta \end{pmatrix} = \begin{pmatrix} \sigma^2 & \mu_3 & 0 \\ \mu_3 & \mu_4-\sigma^4 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\sqrt{n}\left(\begin{pmatrix} \hat\mu \\ \hat\sigma^2 \\ \hat\theta \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \\ \theta \end{pmatrix}\right) \xrightarrow{d} N_3\Big(0, \underbrace{\Gamma^{-1}(\mu,\sigma^2,\theta)\,\Sigma(\mu,\sigma^2,\theta)\,\Gamma^{-1}(\mu,\sigma^2,\theta)}_{\begin{pmatrix} \sigma^2 & \mu_3 & \mu_3/2\sigma \\ \mu_3 & \mu_4-\sigma^4 & (\mu_4-\sigma^4)/2\sigma \\ \mu_3/2\sigma & (\mu_4-\sigma^4)/2\sigma & (\mu_4-\sigma^4)/4\sigma^2 \end{pmatrix}}\Big)$$

|| See the Mathematica script
S_n as an M-estimator

the same result as obtained using the Δ-theorem

Influence function. Consider a general estimator $\hat\theta$ of $\theta \in \mathbb{R}^p$, $\theta = \theta(P)$ functional of the measure P. Let $t \in [0,1]$, x in the sample space and $P_t := (1-t)P \oplus t\,\delta_x$ mixture of P and δ_x. Influence function is of $\theta(P)$ is the Gâteaux (directional) derivative of $\theta(P)$ in direction δ_x (Dirac at x), i.e.
$$IF(x) = \lim_{\varepsilon\to 0} \frac{\theta(P_\varepsilon)-\theta(P_0)}{\varepsilon} = \frac{\partial}{\partial t}\theta_t\Big|_{t=0} \quad \text{for } \theta_t := \theta(P_t)$$
For ε small $\theta(P_\varepsilon) \approx \theta(P_0) + \varepsilon . IF(x)$. quantifies the effect of an infinitesimal change of P in direction δ_x on the (estimator) $\theta(P)$ (functional). Bounded IF implies resistance to outliers and measurement errors. $\longrightarrow$ Robustness

---

For M-estimators: $\theta(P)$ solves $E_P\psi(X,\theta(P)) = 0_p$, i.e. $E_{P_t}\psi(X,\theta_t) = 0 =$
$$= \int \psi(y,\theta_t)\,dP_t(y) = (1-t)\int\psi(y,\theta_t)\,dP(y) + t\int\psi(y,\theta_t)\,d\delta_x(y) = (1-t)\int\psi(y,\theta_t)\,dP(y) + t\,\psi(x,\theta_t)$$

$$/ \cdot \frac{\partial}{\partial t} \text{ gives } 0 = -\int \underbrace{\psi(y,\theta_t)}_{p\times 1}\,dP(y) + (1-t)\int \underbrace{\frac{\partial\psi(y,\theta)}{\partial\theta^T}\Big|_{\theta_t}}_{p\times p} dP(y) \cdot \underbrace{\frac{\partial\theta_t}{\partial t}}_{p\times 1} + \psi(x,\theta_t)$$
$$+ t\cdot\frac{\partial}{\partial\theta^T}\psi(x,\theta)\Big|_{\theta_t}\cdot\frac{\partial\theta_t}{\partial t} \qquad /\, t\to 0+$$

$$\frac{\partial\theta_t}{\partial t} = \lim_{\varepsilon\to 0}\frac{\theta_{t+\varepsilon}-\theta_t}{\varepsilon} \quad /t\to 0+ \quad \rightsquigarrow \quad \frac{\partial\theta_t}{\partial t}\Big|_{t=0} = IF(x)$$

under appropriate conditions $t\,\frac{\partial}{\partial\theta^T}\psi(x,\theta)\big|_{\theta_t}\frac{\partial\theta_t}{\partial t} \xrightarrow[t\to 0]{} 0$

$$0 = \underbrace{-\int\psi(y,\theta_0)\,dP(y)}_{0 \text{ from identification of } \theta_0 \text{, because } \theta_0 \text{ is given by } E_P\psi(X,\theta_0)=0} + \Gamma(\theta_0).\,IF(x) + \psi(x,\theta_0) + 0$$

$$IF(x) = -\Gamma^{-1}(\theta_0).\,\psi(x,\theta_0)$$

From Theorem 9 also $\hat\theta_n - \theta_x = \frac{1}{n}\sum_{i=1}^n IF(X_i) + o_P\left(\frac{1}{\sqrt{n}}\right)$ and a bounded IF, or equivalently bounded $\psi(x,\theta)$ in x results in resistant estimators.
Above $\psi(x,\theta) = (x-\mu)^j$, $j\geq 1$ $\rightsquigarrow$ unbounded IF, $\bar X$ and $\hat\sigma^2$ are affected by outliers and errors, choose other ψ functions

Ex. Robust location estimators   Ex. session 7

Cauchy distribution MLE $f(x,\theta) = \frac{1}{\pi}\frac{1}{1+(x-\mu)^2}$ $\ell(\mu) = c - \log(1+(x-\mu)^2)$

$\psi(x,\mu) = -\frac{\partial \ell(\mu)}{\partial \mu} = \frac{2(x-\mu)}{1+(x-\mu)^2}$ $\mu_X = \arg\min_\mu E \log(1+(X-\mu)^2)$

$\hat\mu = \arg\min_\mu \frac{1}{n}\sum_{i=1}^{n} \log(1+(X_i-\mu)^2)$ computable only numerically, non-convex minimization

$\Gamma(\mu) = E D_\psi(X,\mu) = E\frac{2(1-(x-\mu)^2)}{(1+(x-\mu)^2)^2} \overset{\text{if } X\sim \text{Cauchy}(\mu)}{=} \frac{1}{2}$

if $X\sim$ Cauchy $\sqrt{n}(\hat\mu - \mu_X) \xrightarrow{d} N(0,2)$

$\Sigma(\mu_X) = \text{var}\, \frac{2(X-\mu_X)}{1+(X-\mu_X)^2} = \frac{1}{2}$ if $X\sim$ Cauchy

$|\psi(x,\mu)| \le 1 \quad \forall x,\mu$ generalized to t-distributions

Laplace MLE $f(x,\theta) = \frac{1}{2}\exp(-|x-\theta|)$ $\ell(\theta) = c - |x-\theta|$

$\theta_X = \arg\min_\theta E|X-\theta|$ $\hat\theta = \arg\min \frac{1}{n}\sum |X_i-\theta|$ $\Rightarrow$ θ_X median X

To be always defined, $\rho(x,\theta) := |x-\theta| - |x|$. Then $|E\rho(X,\theta)| = |E(|X-\theta|-|X|)| \le$

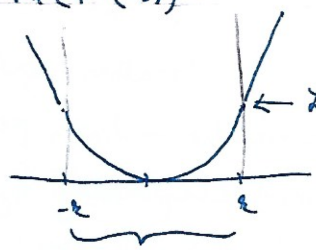
$\le E||X-\theta|-|X|| \le E|\theta| = \theta$ and $E\rho(X,\theta)$ is defined even if $E|X| = \infty$.

$\psi(x,\theta) = \frac{\partial}{\partial\theta}\rho(x,\theta) = \text{sgn}(x-\theta)$ if $x \ne \theta$ does not satisfy [25]

ρ is convex in θ $\Rightarrow$ Theorem 10 gives $\Gamma(\theta) = \frac{\partial^2}{\partial\theta^2} E\rho(X_1,\theta)\Big|_{\theta_X} = \dots = \frac{\partial(2F(\theta)-1)}{\partial\theta}\Big|_{\theta_X} = 2f(\theta_X)$

$\Sigma(\theta) = \text{var}\,\psi(X,\theta_X) = 1$ $\sqrt{n}(\hat\theta - \theta_X) \xrightarrow{d} N\left(0, \frac{1}{4f^2(F^{-1}(1/2))}\right)$

Huber's estimator $\rho_H(x,\theta) = \begin{cases} \frac{(x-\theta)^2}{2} & |x-\theta| \le k \\ k\left(|x-\theta| - \frac{k}{2}\right) & |x-\theta| > k \end{cases}$

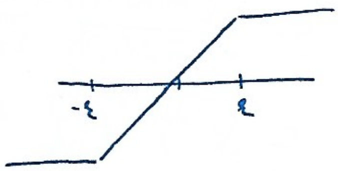


$\psi_H(x,\theta) = \begin{cases} -(x-\theta) & |x-\theta| \le k \\ -k\,\text{sgn}(x-\theta) & |x-\theta| > k \end{cases}$

$\Sigma(\theta_X) = \text{var}\,\psi_H(X,\theta_X) = E\,\psi_H(X,\theta_X)^2 =$

$\theta_X = \arg\min E\rho_H(X,\theta)$ for f symmetric the median

$= \int_{\theta_X-k}^{\theta_X+k} (X-\theta_X)^2\, dF(x) + k^2\int_{\theta_X+k}^{\infty} dF(x) + k^2\int_{-\infty}^{\theta_X-k} dF(x)$



$= \int_{\theta_X-k}^{\theta_X+k} (x-\theta_X)^2\, dF(x) + k^2\left(F(\theta_X-k) + 1 - F(\theta_X+k)\right)$

$\Gamma(\theta) = \frac{\partial^2}{\partial\theta^2} E\rho(X,\theta) = \frac{\partial^2}{\partial\theta^2}\left(\int_{\theta-k}^{\theta+k}\frac{(x-\theta)^2}{2}dF(x) + k\int_{-\infty}^{\theta-k}\left(-(x-\theta) - \frac{k}{2}\right)dF(x) + k\int_{k+\theta}^{\infty}\left((x-\theta)-\frac{k}{2}\right)dF(x)\right) =$

$(x-\theta) < -k$ $x < \theta - k$ $(x-\theta) > k$

$= \frac{\partial}{\partial\theta}\left[\cancel{\frac{k^2}{2}f(\theta+k)} - \cancel{\frac{k^2}{2}f(\theta-k)} + \int_{\theta-k}^{\theta+k} -(x-\theta)\,dF(x) + \cancel{k\cdot\frac{k}{2}f(\theta-k)} + k\cdot\int_{-\infty}^{\theta-k} 1\, dF(x) - \cancel{k\frac{k}{2}f(\theta+k)} - k\int_{k+\theta}^{\infty} dF(x)\right] =$

$\left(= \frac{\partial}{\partial\theta} E\psi_H(X,\theta)\right)$

Leibniz integral rule

$\frac{\partial}{\partial\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)\,dx = f(b(\theta),\theta)\frac{\partial b(\theta)}{\partial\theta} - f(a(\theta),\theta)\frac{\partial a(\theta)}{\partial\theta} + \int_{a(\theta)}^{b(\theta)}\frac{\partial f(x,\theta)}{\partial\theta}dx$

$dF(x) = f(x)\,dx$

$= \frac{\partial}{\partial\theta}\left[-\int_{\theta-k}^{\theta+k}(x-\theta)\,dF(x) + k F(\theta-k) - k\left(1-F(\theta+k)\right)\right] \overset{\text{Leibniz}}{=} -k\cancel{f(\theta+k)} - k\cancel{f(\theta-k)} + \int_{\theta-k}^{\theta+k} dF(x) + k\cancel{f(\theta-k)} + k\cancel{f(\theta+k)}$

$= F(\theta+k) - F(\theta-k)$

$\frac{\partial}{\partial\theta}\int_{\theta-k}^{\theta+k}(x-\theta)\,dF(x) = k\cdot f(\theta+k) - (-k)\, f(\theta-k) + \int_{\theta-k}^{\theta+k} -1\, dF(x)$

Theorem 10: $\sqrt{m}\left(\hat\theta-\theta_x\right) \xrightarrow{d} N\left(0, \Gamma^{-1}(\theta_x)\,\Sigma(\theta_x)\,\Gamma^{-1}(\theta_x)\right) = N\left(0, \dfrac{\int_{\theta_x-k}^{\theta_x+k}(x-\theta_x)^2 dF(x) + k^2\left(F(\theta_x-k)+1-F(\theta_x+k)\right)}{\left(F(\theta_x+k)-F(\theta_x-k)\right)^2}\right)$

M-estimation of location and scale.

$X_1,\dots,X_m \overset{iid}{\sim}$ location-scale model $F_{\mu,\sigma} = F\left(\frac{x-\mu}{\sigma}\right)$, $\mu\in\mathbb{R}$ location parameter, $\sigma>0$ scale parameter

MLE: f density of F, maximize $\sum_{i=1}^m \log f\left(\frac{x_i-\mu}{\sigma}\right) - m\log\sigma$, i.e. solve

$\frac{1}{m}\sum_{i=1}^m \psi_1\left(\frac{x_i-\mu}{\sigma}\right) = 0$ and $\frac{1}{m}\sum_{i=1}^m \psi_2\left(\frac{x_i-\mu}{\sigma}\right) = 0$ for $\psi_1(t) = -\frac{f'(t)}{f(t)}$, $\psi_2(t) = \psi_1(t)t-1$

Estimated parameters are identified by $E\,\psi_i\left(\frac{X-\mu}{\sigma}\right) = 0$ $i=1,2$, influence function is proportional to $\psi(x,\theta) = \begin{pmatrix}\psi_1\left(\frac{x-\mu}{\sigma}\right)\\ \psi_1\left(\frac{x-\mu}{\sigma}\right)\left(\frac{x-\mu}{\sigma}\right)-1\end{pmatrix}$, asymptotic distribution follows from Theorems 9, 10.

Ex. $N(\mu,\sigma^2)$ $f(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$ $\psi_1(t) = t$ $\psi_2(t) = t^2-1$ $\hat\mu = \bar X$ $\hat\sigma^2 = \frac{1}{m}\sum(X_i-\bar X)^2$

identify $E\,\frac{X-\mu_x}{\sigma_x} = 0 \Rightarrow \mu_x = EX$, $E\left(\frac{X-\mu_x}{\sigma_x}\right)^2 = 1 \Rightarrow \sigma_x^2 = \operatorname{var} X$

influence function and asympt. normality derived already.

Multivariate M-estimation of location and scatter.

$X_1,\dots,X_m \overset{iid}{\sim}$ location-scatter model with density $f(x) = |\Sigma|^{-1/2}\, g\left(\left[(x-\mu)'\Sigma^{-1}(x-\mu)\right]^{1/2}\right)$

for Σ a scatter matrix, pos. def symmetric, $\mu\in\mathbb{R}^d$ location parameter, g fixed function

X_1 is called elliptically symmetric with parameters g,μ,Σ.

MLE: $\ell_m(\mu,\Sigma) = -\frac{1}{2}m\log|\Sigma| + \sum_{i=1}^m \log g(\operatorname{md}(x_i))$, $\operatorname{md}(x_i) = \left[(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right]^{1/2}$

Mahalanobis distance of x_i from μ.

$V := \Sigma^{-1}$ and use Zehn invariance principle

$\frac{\partial \ell_m}{\partial\mu} = -\sum_{i=1}^m \frac{g'(\operatorname{md}(x_i))}{g(\operatorname{md}(x_i))}\,\frac{1\cdot 2}{2\operatorname{md}(x_i)}\, V(x_i-\mu) \overset{!}{=} 0$

$\frac{\partial \ell_m}{\partial V} = -\frac{m}{2}\left(V^{-1}\right)^T + \sum_{i=1}^m \frac{g'(\operatorname{md}(x_i))}{g(\operatorname{md}(x_i))}\,\frac{1}{2\operatorname{md}(x_i)}\cdot\frac{\partial}{\partial V}\overbrace{\operatorname{Tr}\left[(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right]}^{=\operatorname{Tr}\left(\Sigma^{-1}(x_i-\mu)(x_i-\mu)^T\right)}$

$= -\frac{m}{2}V^{-1} + \frac{1}{2}\sum_{i=1}^m \frac{g'(\operatorname{md}(x_i))}{g(\operatorname{md}(x_i))}\,\frac{1}{\operatorname{md}(x_i)}\cdot(x_i-\mu)(x_i-\mu)^T \overset{!}{=} 0$

We used $\frac{\partial}{\partial V}\log|V| = \left(V^{-1}\right)^T$ $\left[V = \begin{pmatrix}a & b\\ c & d\end{pmatrix}, \log|V| = \log(ad-bc), \frac{\partial}{\partial V}\log|V| = \frac{1}{|V|}\begin{pmatrix}d & -c\\ -b & a\end{pmatrix} = \left[\frac{1}{|V|}\begin{pmatrix}d & -b\\ -c & a\end{pmatrix}\right]^T = \left(V^{-1}\right)^T\right]$

$\frac{\partial}{\partial V}\operatorname{Tr}(V\cdot A) = \frac{\partial}{\partial V}\left(\sum_{i=1}^d (VA)_{ii}\right)$

$= \frac{\partial}{\partial V}\sum_{i=1}^d\sum_{j=1}^d v_{ij}a_{ji} = A$

Set $u(t) = -\frac{1}{t}\frac{\partial \log g(t)}{\partial t}$. Then

we solve the system

$0_d = \frac{1}{m}\sum_{i=1}^m u(\operatorname{md}(x_i))(x_i-\mu)$

$0_{d\times d} = \frac{1}{m}\sum_{i=1}^m\left[u(\operatorname{md}(x_i))(x_i-\mu)(x_i-\mu)^T - \Sigma\right]$

For $u: [0,\infty)\to[0,\infty)$ arbitrary these define M-estimators of location and scatter.

affine equivariance an estimator of location is affine equivariant if for any $A \in \mathbb{R}^{d\times d}$ non-singular and $b \in \mathbb{R}^d$ $\quad \mu_{AX+b} = A\mu_X + b \quad$ for μ_{AX+b} the estimator computed from data $\{AX_i+b\}_{i=1}^{m}$. An estimator of scatter $\Sigma = \Sigma_X$ is affine equivariant if $\quad \Sigma_{AX+b} = A\Sigma_X A^T$.

M-estimators of location and scatter are affine equivariant.

Proof: $(Ax+b-(A\mu+b))^T(A\Sigma A^T)^{-1}(Ax+b-(A\mu+b)) = (x-\mu)^T A^T (A^T)^{-1}\Sigma^{-1}A^{-1}A(x-\mu)$
$= (x-\mu)^T\Sigma^{-1}(x-\mu)$. Thus $E\,u(md(X))(X-\mu_X) = 0$ if and only if

$$E\,u(md(AX+b;A\mu_X+b, A\Sigma_X A'))(AX+b-(A\mu_X+b)) = A\,E\,u(md(X;\mu_X,\Sigma_X))(X-\mu_X) = 0$$

Similarly $E\,u(md(AX+b;A\mu_X+b, A\Sigma_X A'))(AX+b-(A\mu_X+b))(AX+b-(A\mu_X+b))^T -$
$- A\Sigma_X A' = A(E\,u(md(X;\mu_X,\Sigma_X))(X-\mu_X)(X-\mu_X)^T - \Sigma_X)A^T = 0$

and μ_X, Σ_X is an M-estimator for $X \Leftrightarrow A\mu_X+b, A\Sigma_X A^T$ is an M-estimator for $AX+b$ $\square$

In general, estimators of location without estimating scatter are not affine equivariant.

By affine equivariance it suffices to find M-estimators for $\mu=0$ and $\Sigma = I_d$.

Influence functions of M-estimators of location and scatter are again proportional to ψ-function $u(t).t$ and $u(t).t^2$. Ex. session 8.

(*) **Identification of parameters.** If the distribution of X_i is symmetric around 0 (WLOG $\mu=0, \Sigma=I_d$)
$E\,u(md(X))(X-\mu_X) = 0$ is solved by $\mu_X = 0$ $[md(X) = md(-X)] \Rightarrow \mu$ is correctly identified.
If $(X,Y) \overset{d}{=} (X,-Y)$ for $X = (X,Y)^T$ and a bivariate case also the non-diagonal elements in $E\,u(md(X))XX^T - \Sigma_X$ are 0 and for the diagonal terms to make sure that $E\,u(md(X))X_1^2 - 1 = 0$ one has to consider $\tilde{u}(t) = \frac{u(t)}{k}$ for $k = E\,u(md(X))X_1^2 = E\,u(\|X\|).X_1^2$. Only then all parameters are correctly identified.

Example: Cauchy distribution (or t-distribution with ν degrees of freedom)

$$g_\nu(t) = c.(\nu+t^2)^{-\frac{d+\nu}{2}}$$

$$\log g_\nu(t) = \log c - \frac{d+\nu}{2}\log(\nu+t^2)$$

$$-\frac{\partial \log g_\nu(t)}{\partial t} = \frac{d+\nu}{2}\cdot\frac{2t}{\nu+t^2}$$

$$u(t) = \frac{d+\nu}{\nu+t^2}$$

influence functions are proportional to $\frac{c t^j}{\nu+t^2}$, $j=1,2$ and are all bounded.
asymptotic normality follows from Theorem 9
efficiency at normal model can be computed numerically. (Mathematica)
In practice, solutions to the estimating equations are obtained iteratively, or numerically.

Identification of parameters for elliptically symmetric distributions

$X \sim f(x) = |\Sigma|^{-1/2} g(MD(x))$, $\quad MD(x) := \sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$ we say $X \sim Ell(\mu,\Sigma,g)$

Let $b\in\mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$ positive definite. Then $AX+b \sim Ell(A\mu+b, A\Sigma A^T, g)$.

Proof: Transformation of a random vector: $Y = AX+b$, $X = A^{-1}(Y-b)$, $|J| = |A^{-1}| = |A|^{-1}$

$$f_Y(y) = f_X(A^{-1}(y-b))|A|^{-1} = |A|^{-1}|\Sigma|^{-1/2} g\left(\sqrt{(A^{-1}(y-b)-\mu)^T\Sigma^{-1}(A^{-1}(y-b)-\mu)}\right) =$$

$$= |A\Sigma A^T|^{-1/2} g\left(\sqrt{(y-(A\mu+b))^T (A^{-1})^T\Sigma^{-1}A^{-1}(y-(A\mu+b))}\right) \sim Ell(A\mu+b, A\Sigma A^T, g) \quad \square$$

In particular, if $X \sim Ell(0, I, g) \Rightarrow AX \sim Ell(0, I, g)$ for any A orthogonal, i.e. $AA^T = I$.

X is invariant under orthogonal transformations (rotations) X is spherically symmetric.

Identification: By affine equivariance let $\mu = 0$, $\Sigma = I$, $X \sim Ell(0, I, g)$. Take $A =$

$$= \begin{pmatrix} -1 & & 0 \\ & -1 & \\ 0 & & \ddots \; -1 \end{pmatrix} \quad A \text{ is OG and } \quad AX = \begin{pmatrix} -X_1 \\ -X_2 \\ \vdots \\ -X_d \end{pmatrix} \overset{d}{=} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix} \Rightarrow X_1 \overset{d}{=} -X_1, \; X \overset{d}{=} -X.$$

Thus the first estimating equation of M-estimator gives $E\,u(\|X\|)(X-0) = E(u(\|X\|)(-X-0)) = 0$

$MD(X) = \sqrt{X' I^{-1} X} = \|X\|$ and $\mu_X = 0$ is correctly identified.

Further $E\,u(\|X\|)XX^T - I = \begin{cases} \text{if } j=k,\ j,k\text{-th element}, & E\,u(\|X\|)X_j^2 - 1 = * \\ \text{if } j \neq k, & E\,u(\|X\|)X_jX_k - 0 = E\,u(\|X\|)X_j(-X_k) = 0 \end{cases}$ ✓

$*$ for $* = 0$ one has to assume that $E\,u(\|X\|)X_j^2 = 1 \;\forall j$. Since $X_j \overset{d}{=} X_1$ (take $A =$ I with j-th and 1-st row flipped) it suffices that $E\,u(\|X\|)X_1^2 = 1 \Rightarrow$ modify and take the estimating equations with $k = E\,u(\|X\|)X_1^2$

$$0_d = E\,u(MD(X))(X-\mu)$$

$$0_{d\times d} = \frac{E\,u(MD(X))(X-\mu)(X-\mu)^T}{E\,u(\|X\|)X_1^2} - \Sigma = E\,u_2(MD(X))(X-\mu)(X-\mu)^T - \Sigma$$

for $u_2(t) = u(t)/k$.

Then both parameters are identified.

For $d=1$: $\quad 0 = E\,u\left(\frac{|X-\mu|}{\sigma}\right)(X-\mu), \quad 0 = \dfrac{E\,u\left(\frac{|X-\mu|}{\sigma}\right)(X-\mu)^2}{E\,u(|X|)X^2} - \sigma^2$

$MD(X) = \sqrt{\frac{(X-\mu)^2}{\sigma^2}} = \frac{|X-\mu|}{\sigma}$

M-estimators in regression.

Ex. 38: Misspecified linear model. $\binom{X_1}{Y_1}, \dots, \binom{X_n}{Y_n}$ iid, $Y_i|X_i \sim N(\beta' X_i, \sigma^2)$, MLE $\rightsquigarrow$ M-estimator for β of form $\rho(x,y;\beta) = (y - x'\beta)^2$

$\psi(x,y;\beta) = -2x(y-x'\beta) \Rightarrow E\,X(Y - X'\beta_x) = 0 \Leftrightarrow EXY = E[XX']\beta_x$

$\beta_x = (E[XX'])^{-1} EXY$

if $E[Y|X] = X'\beta \Rightarrow \beta_x = (E[XX'])^{-1} EXY = (E[XX'])^{-1} E\,E[XY|X] = \beta$ and β is identified

$E\,D_\psi(X,Y,\beta) = E\,2XX^T$

$E\,\psi(X,Y,\beta_x)^{\otimes 2} = 4\,E\,X(Y-X'\beta_x)(X(Y-X'\beta_x))^T = 4\,E\,E[X(Y-X'\beta_x)(Y-X'\beta_x)^T X^T|X] =$

$= 4\,E\,X\,\mathrm{diag}(\mathrm{var}_{Y|X})X' = 4\,E\,\sigma^2(X)XX^T$

$\sigma^2(x_i) = \mathrm{var}(Y|X=x_i)$

$$\sqrt{n}(\hat\beta - \beta) \xrightarrow{d} N_p\left(0_p, (EXX^T)^{-1}(E\sigma^2(X)XX^T)(EXX^T)^{-1}\right)$$

Sandwich estimator from linear regression.

Robust regression estimators

OLS: $\rho(x,y,\beta) = (y - x'\beta)^2$

$\psi(x,y,\beta) = x(y - x'\beta)$ — not robust, unbounded IF

LAD: $\rho(x,y,\beta) = |y - x'\beta|$

$\psi(x,y,\beta) = x \,\mathrm{sign}(y - x'\beta)$ — robust in Y, non-robust in X $\to$ median regression

Huber: $\rho(x,y,\beta) = \rho_H(y - x'\beta)$

$\psi(x,y,\beta) = \psi_H(y - x'\beta)\, x$ — robust in Y, non-robust in X $\to$ Huber's regression

Identification of parameters. (OLS and LAD)

Let $Y = \alpha + X'\beta + \varepsilon$ for ε independent of X. Let the form of X be correct, i.e. the M-estimator is given by $\psi(x,y,a,b) = x(y - a - x'b)$. Then the M-estimator identifies

OLS:

$$E X(Y - a - X'b) = E X(\alpha + X'\beta + \varepsilon - a - X'b) = 0$$

$$\alpha EX + EXX'\beta + EXE\varepsilon - aEX - EXX'b = 0$$

$$EX(\alpha + E\varepsilon - a) + (EXX')(\beta - b) = 0$$

and the choice $a = \alpha + E\varepsilon$, $b = \beta$ solves the equation.

LAD: for given x, $\rho(x,y,a,b)$ is minimized by $a + bx \approx$ conditional median of $Y|X$, which is $\alpha + x'\beta + \mathrm{median}(\varepsilon)$. This is a linear function of x $\Rightarrow$ choice $a = \alpha + \mathrm{med}\,\varepsilon$, $b = \beta$ identifies parameters.

~~E...~~ for X given, $E[|Y - m| \,|X]$ is minimized by $m = \mathrm{median}(Y|X)$

In LAD, we fit a line $m(X) = x'\beta$ that approximates $\mathrm{median}(Y|X)$ the best in the sense $E|Y - X'\beta|$ is minimized.

Example: $Y = \alpha + \beta X + \varepsilon$, $X \in \mathbb{R}$, X independent of ε. Misspecified model is $Y \sim bX$, no intercept in the model.

OLS: $0 = EX(Y - Xb) = EX(\alpha + \beta X + \varepsilon - Xb) = \alpha EX + \beta EX^2 + EXE\varepsilon - bEX^2 = 0$

$$b = \frac{1}{EX^2}(\alpha EX + \beta EX^2 + EXE\varepsilon) = \beta + \frac{\alpha + E\varepsilon}{EX^2}\cdot EX$$

[in R script $\alpha = 0$, $EX = 1/2$, $EX^2 = 1/3$, $E\varepsilon = 1$ (Exp(1)) and $b = 5/2$, $\beta = 1$]

conditional expectation: $E[Y|X] = \alpha + E\varepsilon + \beta X$

$$E[(Y - m(X))^2|X] = E[(Y - E[Y|X])^2|X] + E[(E[Y|X] - m(X))^2|X] + 2\underbrace{E[(Y - E[Y|X])(E[Y|X] - m(X))|X]}_{=0}$$

and for given X, $E\underbrace{\rho(X,Y,b)}_{(Y - m(X))^2}$ is minimized iff $m(X)$ is close in the L^2-norm (X) to $E(Y|X)$ $\Rightarrow$ it is enough to minimize $E((E[Y|X] - m(X))^2)$

Pythagorean theorem: ~~Hilbert~~ ~~Hilb~~ Hilbert space $L^2(X)$ given by $\|f\|_X^2 = E f(X)^2$

LAD: minimize $E|\alpha + \beta X + \varepsilon - bX| = EE[|\alpha + \beta X + \varepsilon - bX| \,|X]$

for any X the inner expectation is minimized by conditional median $\alpha + \beta X + \mathrm{med}\,\varepsilon$ but this is not a function of the form bX and because $L^1(X)$ is not Hilbert, no version Pythagorean theorem can be stated. We do not minimize $E|m(X) - \mathrm{med}(Y|X)|$

OLS: $\rho(x,z) := (z - m(x))^2$ minimize over functions $m \in \mathcal{M}$

$E\rho(X,Y) = E(Y - m(X))^2 = E\,E[(Y-m(X))^2 \mid X] = E\Big(E[(Y - E(Y|X))^2 \mid X] + E[(E[Y|X] - m(X))^2 \mid X]$

$+ 2E[(Y - E(Y|X))(E(Y|X) - m(X)) \mid X]\Big) = \underbrace{E((Y - E(Y|X))^2)}_{\text{does not depend on } m} + \underbrace{E(E(Y|X) - m(X))^2}_{= \| E(Y|X) - m \|_X^2}$

In OLS, always the closest function to $E[Y|X]$ in the $\|\cdot\|_X$-norm is specified.
This holds because of the Hilbert-space structure of $\|\cdot\|_{X,2}$, $L^2(X)$, $\|f\|_{X,2} = \sqrt{E f(X)^2}$
No such thing holds true in the for $L^1(X)$, LAD does not specify the closest function to median $(Y|X)$ in the $L^1(X)$-norm. $\|f\|_{X,1} = E|f(X)|$

NMST434 - exercise 6/7 misspecified Poisson

+ R code

Example ~~36~~ 39: $X_i = (X_{i1} \dots X_{ip})^T$ $\quad Y_i | X_i \sim Po(\lambda(x_i))$

$\lambda(x_i) = e^{\beta x_i}$ $\quad \beta = (\beta_1 - \beta_p)^T$

$$L(\beta) = \prod \frac{e^{-\lambda(x_i)} \lambda(x_i)^{y_i}}{y_i!} \qquad \ell(\beta) = -\sum \lambda(x_i) + \sum y_i \log \lambda(x_i) + c$$

$$\frac{\partial}{\partial \beta} \lambda(x_i) = x_i \lambda(x_i) \qquad U(\beta) = -\sum x_i \lambda(x_i) + \sum y_i x_i = \sum x_i (y_i - \lambda(x_i))$$

$\Rightarrow \psi(x, y, \beta) = x(y - e^{\beta' x})$ and β_x solves (the true est. parameter)

$$E \psi(X, Y, \beta_x) = 0$$

heteroscedasticity, or different distribution

If $Y_i | X_i$ is not Poisson, but still $E[Y|X] = \lambda(X)$ we have for $\beta_0 = \beta$

$$E \psi(X, Y, \beta_0) = E X(Y - \lambda(X)) = E X(Y - e^{\beta_0' X})$$

$$= E E[X(Y - e^{\beta_0' X}) | X] = 0 \quad \Rightarrow \quad \underline{\underline{\beta_0 = \beta_X}}$$

and we still estimate the correct parameter.

Theorem 9: $\Gamma(\beta_x) = E \frac{\partial}{\partial \beta'} \psi(X, Y, \beta)|_{\beta_x} = E XX' e^{\beta_x' X}$

$$\Sigma(\beta_x) = E X(Y - e^{\beta_x' X})^2 X' = E XX'(Y - e^{\beta_x' X})$$

$$\sqrt{n}(\hat{\beta} - \beta_x) \xrightarrow{d} N_p(0, \Gamma^{-1}(\beta_x) \Sigma(\beta_x) \Gamma^{-1}(\beta_x))$$

estimate Γ a Σ as usual from the observed quantities

$\Rightarrow$ Sandwich estimator in Poisson regression

If the true Poisson is true, avar is given by Fisher information

$\sim \Sigma$ and Γ (both correspond to $I(\beta)$)