Empirical Likelihood

Trial lecture for the degree of Philosophiae Doctor (PhD)

Martin Jullum

University of Oslo

martinju@math.uio.no

April 1, 2016

Problem setup

Basic setup

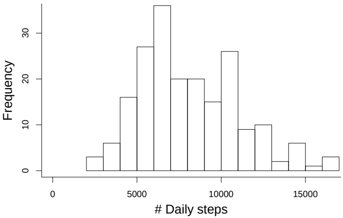
- I.i.d. data $Y_1, \ldots, Y_n$ stems from some d-dimensional distribution G_0
- Want to learn about some properties, structures or mechanisms of G_0
    - $\mu = T(G)$ of dimension p, with true unknown value $\mu_{\text{true}} = T(G_0)$

Problem setup

Basic setup

- I.i.d. data $Y_1, \ldots, Y_n$ stems from some d-dimensional distribution G_0
- Want to learn about some properties, structures or mechanisms of G_0
  - $\mu = T(G)$ of dimension p, with true unknown value $\mu_{\text{true}} = T(G_0)$

Examples

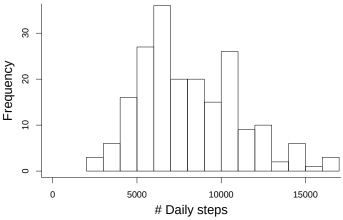


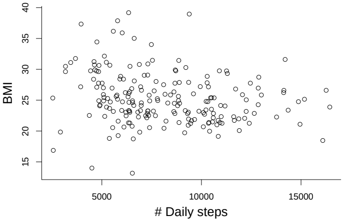
Problem setup

Basic setup

- I.i.d. data $Y_1, \ldots, Y_n$ stems from some d-dimensional distribution G_0
- Want to learn about some properties, structures or mechanisms of G_0
  - $\mu = T(G)$ of dimension p, with true unknown value $\mu_{\text{true}} = T(G_0)$

Examples



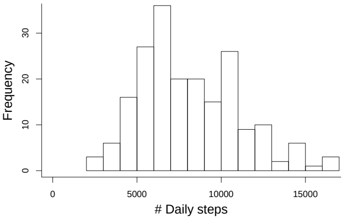


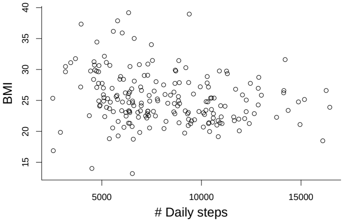
Problem setup

Basic setup

- I.i.d. data $Y_1, \ldots, Y_n$ stems from some d-dimensional distribution G_0
- Want to learn about some properties, structures or mechanisms of G_0
    - $\mu = T(G)$ of dimension p, with true unknown value $\mu_{\text{true}} = T(G_0)$

Examples





Typical goals

- Estimate and compute confidence intervals/regions (CR) for μ
- Perform hypothesis tests (HT): $H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \mu_0$

Classical parametric likelihood approach

- Restrict G_0 to some parametric family F_θ, indexed by $\theta \in \Theta$
  - Density/probability mass function $f(\cdot;\theta)$
- Examples
  - Assume data $Y_i \sim \mathsf{Pois}(\theta), \theta > 0$,
  - Assume data $Y_i \sim \mathsf{N}(\xi, \sigma^2), \theta = (\xi, \sigma) \in (\mathbb{R} \times \mathbb{R}^+)$
- Likelihood: $L(\theta) = \prod_{i=1}^n f(Y_i;\theta)$
  - Measures the 'chance' of sampling the observed data if $G_0 = F_\theta$, as a function of θ
- Maximum likelihood: $\widehat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$
- Estimate $\mu_{\mathrm{true}} = T(G_0)$ by $\widehat{\mu}_{\mathrm{pm}} = T(F_{\widehat{\theta}})$
- Trust $F_{\widehat{\theta}}$ for further inference

Further parametric inference

Asymptotic normality based inference

- $\sqrt{n}(\widehat{\mu}_{\text{pm}} - \mu_{\text{true}}) \to_L \mathrm{N}(0, v_{\text{pm}})$
- 95% CR: $C_{\mathsf{N},0.95} = \{\mu \,|\, \mu \in \widehat{\mu}_{\text{pm}} \pm 1.96\sqrt{v_{\text{pm}}}/\sqrt{n}\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\mathsf{N},0.95}$

Likelihood ratio (LR) based inference

- Likelihood ratio: $LR(\theta) = L(\theta)/L(\widehat{\theta})$
- Profile likelihood ratio: $\mathcal{R}(\mu) = \sup\{LR(\theta) \,|\, \mu = T(F_\theta), \theta \in \Theta\}$
- **Asymptotic theory (Wilks's theorem):** $-2\log\mathcal{R}(\mu_{\text{true}}) \to_L \chi^2_p$
- 95% CR: $C_{\mathsf{LR},0.95} = \{\mu \,|\, -2\log\mathcal{R}(\mu) \leq \chi^{2,0.95}_p\} = \{\mu \,|\, \mathcal{R}(\mu) \geq \exp(-\frac{1}{2}\chi^{2,0.95}_p)\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\mathsf{LR},0.95}$
- Neyman-Pearson: LR is the most powerful test

Possible problems

- Parametric model might be wrong $\Rightarrow$ biased inference, invalid conclusions

Further parametric inference

Asymptotic normality based inference

- $\sqrt{n}(\widehat{\mu}_{\text{pm}} - \mu_{\text{true}}) \to_L \mathrm{N}(0, v_{\text{pm}})$
- 95% CR: $C_{\mathsf{N},0.95} = \{\mu \,|\, \mu \in \widehat{\mu}_{\text{pm}} \pm 1.96\sqrt{v_{\text{pm}}}/\sqrt{n}\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\mathsf{N},0.95}$

Likelihood ratio (LR) based inference

- Likelihood ratio: $LR(\theta) = L(\theta)/L(\widehat{\theta})$
- Profile likelihood ratio: $\mathcal{R}(\mu) = \sup\{LR(\theta) \,|\, \mu = T(F_\theta), \theta \in \Theta\}$
- **Asymptotic theory (Wilks's theorem):** $-2\log\mathcal{R}(\mu_{\text{true}}) \to_L \chi^2_p$
- 95% CR: $C_{\mathsf{LR},0.95} = \{\mu \,|\, -2\log\mathcal{R}(\mu) \le \chi_p^{2,0.95}\} = \{\mu \,|\, \mathcal{R}(\mu) \ge \exp(-\frac{1}{2}\chi_p^{2,0.95})\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\mathsf{LR},0.95}$
- Neyman-Pearson: LR is the most powerful test

Possible problems

- Parametric model might be wrong $\Rightarrow$ biased inference, invalid conclusions

Further parametric inference

Asymptotic normality based inference

- $\sqrt{n}(\widehat{\mu}_{\text{pm}} - \mu_{\text{true}}) \to_L \text{N}(0, v_{\text{pm}})$
- 95% CR: $C_{\text{N},0.95} = \{\mu \,|\, \mu \in \widehat{\mu}_{\text{pm}} \pm 1.96\sqrt{v_{\text{pm}}}/\sqrt{n}\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\text{N},0.95}$

Likelihood ratio (LR) based inference

- Likelihood ratio: $LR(\theta) = L(\theta)/L(\widehat{\theta})$
- Profile likelihood ratio: $\mathcal{R}(\mu) = \sup\{LR(\theta) \,|\, \mu = T(F_\theta), \theta \in \Theta\}$
- **Asymptotic theory (Wilks's theorem):** $-2\log\mathcal{R}(\mu_{\text{true}}) \to_L \chi^2_p$
- 95% CR: $C_{\text{LR},0.95} = \{\mu \,|\, -2\log\mathcal{R}(\mu) \le \chi_p^{2,0.95}\} = \{\mu \,|\, \mathcal{R}(\mu) \ge \exp(-\frac{1}{2}\chi_p^{2,0.95})\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\text{LR},0.95}$
- Neyman-Pearson: LR is the most powerful test

Possible problems

- Parametric model might be wrong $\Rightarrow$ biased inference, invalid conclusions

Further parametric inference

Asymptotic normality based inference

- $\sqrt{n}(\widehat{\mu}_{\mathrm{pm}} - \mu_{\mathrm{true}}) \to_L \mathrm{N}(0, v_{\mathrm{pm}})$
- 95% CR: $C_{\mathsf{N},0.95} = \{\mu \,|\, \mu \in \widehat{\mu}_{\mathrm{pm}} \pm 1.96\sqrt{v_{\mathrm{pm}}}/\sqrt{n}\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\mathsf{N},0.95}$

Likelihood ratio (LR) based inference

- Likelihood ratio: $LR(\theta) = L(\theta)/L(\widehat{\theta})$
- Profile likelihood ratio: $\mathcal{R}(\mu) = \sup\{LR(\theta) \,|\, \mu = T(F_\theta), \theta \in \Theta\}$
- **Asymptotic theory (Wilks's theorem):** $-2\log\mathcal{R}(\mu_{\mathrm{true}}) \to_L \chi^2_p$
- 95% CR:
  $C_{\mathsf{LR},0.95} = \{\mu \,|\, -2\log\mathcal{R}(\mu) \leq \chi_p^{2,0.95}\} = \{\mu \,|\, \mathcal{R}(\mu) \geq \exp(-\frac{1}{2}\chi_p^{2,0.95})\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\mathsf{LR},0.95}$
- Neyman-Pearson: LR is the most powerful test

Possible problems

- Parametric model might be wrong $\Rightarrow$ biased inference, invalid conclusions

Further parametric inference

Asymptotic normality based inference

- $\sqrt{n}(\widehat{\mu}_{\text{pm}} - \mu_{\text{true}}) \to_L \mathrm{N}(0, v_{\text{pm}})$
- 95% CR: $C_{\mathsf{N},0.95} = \{\mu \,|\, \mu \in \widehat{\mu}_{\text{pm}} \pm 1.96\sqrt{v_{\text{pm}}}/\sqrt{n}\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\mathsf{N},0.95}$

Likelihood ratio (LR) based inference

- Likelihood ratio: $LR(\theta) = L(\theta)/L(\widehat{\theta})$
- Profile likelihood ratio: $\mathcal{R}(\mu) = \sup\{LR(\theta) \,|\, \mu = T(F_\theta), \theta \in \Theta\}$
- **Asymptotic theory (Wilks's theorem):** $-2\log\mathcal{R}(\mu_{\text{true}}) \to_L \chi^2_p$
- 95% CR:
  $C_{\mathsf{LR},0.95} = \{\mu \,|\, -2\log\mathcal{R}(\mu) \le \chi_p^{2,0.95}\} = \{\mu \,|\, \mathcal{R}(\mu) \ge \exp(-\frac{1}{2}\chi_p^{2,0.95})\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\mathsf{LR},0.95}$
- Neyman-Pearson: LR is the most powerful test

Possible problems

- Parametric model might be wrong $\Rightarrow$ biased inference, invalid conclusions

Further parametric inference

Asymptotic normality based inference

- $\sqrt{n}(\widehat{\mu}_{\rm pm} - \mu_{\rm true}) \to_L \mathrm{N}(0, v_{\rm pm})$
- 95% CR: $C_{\mathsf{N},0.95} = \{\mu \,|\, \mu \in \widehat{\mu}_{\rm pm} \pm 1.96\sqrt{v_{\rm pm}}/\sqrt{n}\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\mathsf{N},0.95}$

Likelihood ratio (LR) based inference

- Likelihood ratio: $LR(\theta) = L(\theta)/L(\widehat{\theta})$
- Profile likelihood ratio: $\mathcal{R}(\mu) = \sup\{LR(\theta) \,|\, \mu = T(F_\theta), \theta \in \Theta\}$
- **Asymptotic theory (Wilks's theorem):** $-2\log\mathcal{R}(\mu_{\rm true}) \to_L \chi^2_p$
- 95% CR:
  $C_{\mathsf{LR},0.95} = \{\mu \,|\, -2\log\mathcal{R}(\mu) \le \chi_p^{2,0.95}\} = \{\mu \,|\, \mathcal{R}(\mu) \ge \exp(-\frac{1}{2}\chi_p^{2,0.95})\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\mathsf{LR},0.95}$
- Neyman-Pearson: LR is the most powerful test

Possible problems

- Parametric model might be wrong $\Rightarrow$ biased inference, invalid conclusions

Nonparametric modelling

- Let the data speak for themselves, with few structural restrictions
- For i.i.d. data, one typically uses the empirical distribution function: $\widehat{G}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \leq y\}}$
  - Plug-in estimators $\widehat{\mu}_{\text{np}} = T(\widehat{G}_n)$

Asymptotic normality based inference

- $\sqrt{n}(\widehat{\mu}_{\text{np}} - \mu_{\text{true}}) \to_L \mathrm{N}(0, v_{\text{np}})$
- 95% CR: $C_{\mathsf{N},0.95} = \{\mu \,|\, \mu \in \widehat{\mu}_{\text{np}} \pm 1.96\sqrt{v_{\text{np}}}/\sqrt{n}\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\mathsf{N},0.95}$

Drawbacks of normal theory

- Always symmetric confidence regions
- No proper quantification of how likely each value of μ is

Nonparametric modelling

- Let the data speak for themselves, with few structural restrictions
- For i.i.d. data, one typically uses the empirical distribution function: $\widehat{G}_n(y) = n^{-1}\sum_{i=1}^n \mathbf{1}_{\{Y_i \le y\}}$
  - Plug-in estimators $\widehat{\mu}_{\rm np} = T(\widehat{G}_n)$

Asymptotic normality based inference

- $\sqrt{n}(\widehat{\mu}_{\rm np} - \mu_{\rm true}) \to_L \mathrm{N}(0, v_{\rm np})$
- 95% CR: $C_{\mathsf{N},0.95} = \{\mu \,|\, \mu \in \widehat{\mu}_{\rm np} \pm 1.96\sqrt{v_{\rm np}}/\sqrt{n}\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\mathsf{N},0.95}$

Drawbacks of normal theory

- Always symmetric confidence regions
- No proper quantification of how likely each value of μ is

Nonparametric modelling

- Let the data speak for themselves, with few structural restrictions
- For i.i.d. data, one typically uses the empirical distribution function: $\widehat{G}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \le y\}}$
  - Plug-in estimators $\widehat{\mu}_{\text{np}} = T(\widehat{G}_n)$

Asymptotic normality based inference

- $\sqrt{n}(\widehat{\mu}_{\text{np}} - \mu_{\text{true}}) \to_L \mathrm{N}(0, v_{\text{np}})$
- 95% CR: $C_{\mathsf{N},0.95} = \{\mu \,|\, \mu \in \widehat{\mu}_{\text{np}} \pm 1.96\sqrt{v_{\text{np}}}/\sqrt{n}\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{\mathsf{N},0.95}$

Drawbacks of normal theory

- Always symmetric confidence regions
- No proper quantification of how likely each value of μ is

Likelihood inference without parameters?

- Likelihood is a parametric concept, however...

Empirical likelihood (EL)

- Constructs nonparametric pseudo likelihoods and pseudo likelihood ratios
- A nonparametric analogue to Wilks's theorem:

$$-2\log\mathcal{R}^*(\mu_{\text{true}}) \to_L \chi^2_p$$

- Developed by Art B. Owen (1988, 1990, 1991)

Likelihood inference without parameters?

- Likelihood is a parametric concept, however...

Empirical likelihood (EL)

- Constructs nonparametric pseudo likelihoods and pseudo likelihood ratios
- A nonparametric analogue to Wilks's theorem:

$$-2 \log \mathcal{R}^*(\mu_{\text{true}}) \to_L \chi^2_p$$

- Developed by Art B. Owen (1988, 1990, 1991)

Likelihood inference without parameters?

- Likelihood is a parametric concept, however...

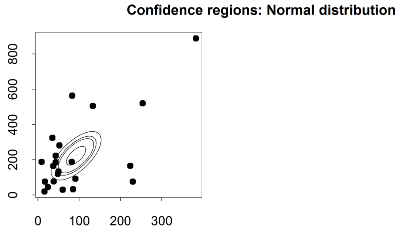
Empirical likelihood (EL)

- Constructs nonparametric pseudo likelihoods and pseudo likelihood ratios
- A nonparametric analogue to Wilks's theorem:

$$-2 \log \mathcal{R}^*(\mu_{\text{true}}) \to_L \chi^2_p$$

- Developed by Art B. Owen (1988, 1990, 1991)

Illustration: Typical behaviour



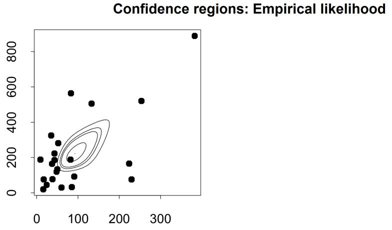
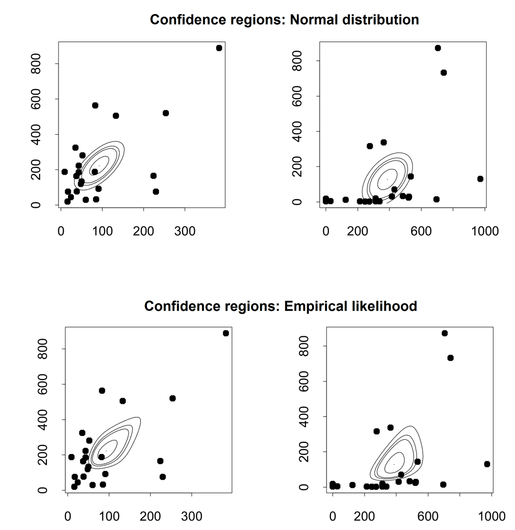


Illustration: Typical behaviour



Empirical likelihood (EL)

- Write $G(\{y\}) = \Pr_G(Y = y) = G(y) - G(y-)$
- A **'nonparametric likelihood':** $L^*(G) = \prod_{i=1}^{n} G(\{Y_i\})$
- $\widehat{G}_n$ maximises $L^*(G)$, thus sometimes referred to as the nonparametric maximum likelihood estimator
- A **'nonparametric likelihood ratio':** $LR^*(G) = L^*(G)/L^*(\widehat{G}_n)$
- A **'profile nonparametric likelihood ratio'** corresponding to μ:

$$\mathcal{R}^*(\mu) = \sup\{LR^*(G) \,|\, T(G) = \mu, G \in \mathcal{G}\},$$

for $\mathcal{G}$ some (slightly restricted) space of distributions

Empirical likelihood theorem (ELT)

Under certain conditions:

$$-2\log \mathcal{R}^*(\mu_{\text{true}}) \to_L \chi_p^2$$

- 95% CR: $C_{LR^*,0.95} = \{\mu \,|\, -2\log \mathcal{R}^*(\mu) \le \chi_p^{2,0.95}\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{LR^*,0.95}$

Empirical likelihood (EL)

- Write $G(\{y\}) = \mathrm{Pr}_G(Y = y) = G(y) - G(y-)$
- A **'nonparametric likelihood':** $L^*(G) = \prod_{i=1}^n G(\{Y_i\})$
- $\widehat{G}_n$ maximises $L^*(G)$, thus sometimes referred to as the nonparametric maximum likelihood estimator
- A **'nonparametric likelihood ratio':** $LR^*(G) = L^*(G)/L^*(\widehat{G}_n)$
- A **'profile nonparametric likelihood ratio'** corresponding to μ:

$$\mathcal{R}^*(\mu) = \sup\{LR^*(G) \,|\, T(G) = \mu, G \in \mathcal{G}\},$$

for $\mathcal{G}$ some (slightly restricted) space of distributions

Empirical likelihood theorem (ELT)

Under certain conditions:

$$-2\log\mathcal{R}^*(\mu_{\text{true}}) \to_L \chi_p^2$$

- 95% CR: $C_{LR^*,0.95} = \{\mu \,|\, -2\log\mathcal{R}^*(\mu) \le \chi_p^{2,0.95}\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{LR^*,0.95}$

Empirical likelihood (EL)

- Write $G(\{y\}) = \text{Pr}_G(Y = y) = G(y) - G(y-)$
- A **'nonparametric likelihood':** $L^*(G) = \prod_{i=1}^n G(\{Y_i\})$
- $\widehat{G}_n$ maximises $L^*(G)$, thus sometimes referred to as the nonparametric maximum likelihood estimator
- A **'nonparametric likelihood ratio':** $LR^*(G) = L^*(G)/L^*(\widehat{G}_n)$
- A **'profile nonparametric likelihood ratio'** corresponding to μ:

$$\mathcal{R}^*(\mu) = \sup\{LR^*(G) \,|\, T(G) = \mu, G \in \mathcal{G}\},$$

for $\mathcal{G}$ some (slightly restricted) space of distributions

Empirical likelihood theorem (ELT)

Under certain conditions:

$$-2\log \mathcal{R}^*(\mu_{\text{true}}) \to_L \chi_p^2$$

- 95% CR: $C_{LR^*,0.95} = \{\mu \,|\, -2\log \mathcal{R}^*(\mu) \leq \chi_p^{2,0.95}\}$
- HT: Reject H_0 on level 0.05 if $\mu_0 \notin C_{LR^*,0.95}$

Details of EL

- Let $w_i = w_i(G)$ be the weight which G puts on observation Y_i, $i = 1, \dots, n$
- If no ties: $w_i = G(\{Y_i\})$, if ties $G(\{Y_i\}) = \sum_{j:Y_i=Y_j} w_j$
- $w_i \geq 0, \sum_{i=1}^n w_i \leq 1$

If no ties

- $L^*(G) = \prod_{i=1}^n w_i, L^*(\widehat{G}_n) = \prod_{i=1}^n n^{-1}, LR^*(G) = \prod_{i=1}^n nw_i$, s.t.

$$\mathcal{R}^*(\mu) = \sup\{\prod_{i=1}^n nw_i \,|\, T(G) = \mu, G \in \mathcal{G}\}$$

If ties

- $\prod_{i=1}^n w_i$ is not unique
- $\prod_{i=1}^n w_i$ is maximised if jumps in G at tied observations are equally distributed $\Rightarrow L^*(G) \propto \prod_{i=1}^n w_i$
- $\Rightarrow LR^*(G)$ and $\mathcal{R}^*(\mu)$ remain unchanged

Details of EL

- Let $w_i = w_i(G)$ be the weight which G puts on observation Y_i, $i = 1, \dots, n$
- If no ties: $w_i = G(\{Y_i\})$, if ties $G(\{Y_i\}) = \sum_{j: Y_i = Y_j} w_j$
- $w_i \geq 0, \sum_{i=1}^n w_i \leq 1$

If no ties

- $L^*(G) = \prod_{i=1}^n w_i, L^*(\widehat{G}_n) = \prod_{i=1}^n n^{-1}, LR^*(G) = \prod_{i=1}^n nw_i$, s.t.

$$\mathcal{R}^*(\mu) = \sup\{\prod_{i=1}^n nw_i \,|\, T(G) = \mu, G \in \mathcal{G}\}$$

If ties

- $\prod_{i=1}^n w_i$ is not unique
- $\prod_{i=1}^n w_i$ is maximised if jumps in G at tied observations are equally distributed $\Rightarrow L^*(G) \propto \prod_{i=1}^n w_i$
- $\Rightarrow LR^*(G)$ and $\mathcal{R}^*(\mu)$ remain unchanged

Details of EL

- Let $w_i = w_i(G)$ be the weight which G puts on observation Y_i, $i = 1, \ldots, n$
- If no ties: $w_i = G(\{Y_i\})$, if ties $G(\{Y_i\}) = \sum_{j: Y_i = Y_j} w_j$
- $w_i \geq 0, \sum_{i=1}^n w_i \leq 1$

If no ties

- $L^*(G) = \prod_{i=1}^n w_i, L^*(\widehat{G}_n) = \prod_{i=1}^n n^{-1}, LR^*(G) = \prod_{i=1}^n nw_i$, s.t.

$$\mathcal{R}^*(\mu) = \sup\{\prod_{i=1}^n nw_i \,|\, T(G) = \mu, G \in \mathcal{G}\}$$

If ties

- $\prod_{i=1}^n w_i$ is not unique
- $\prod_{i=1}^n w_i$ is maximised if jumps in G at tied observations are equally distributed $\Rightarrow L^*(G) \propto \prod_{i=1}^n w_i$
- $\Rightarrow LR^*(G)$ and $\mathcal{R}^*(\mu)$ remain unchanged

Restricting the distribution space

- Consider $G_\varepsilon = (1-\varepsilon)\widehat{G}_n + \varepsilon\delta_y$, with

$$LR^*(G_\varepsilon) = L^*(G_\varepsilon)/L^*(\widehat{G}_n) = \frac{\prod_{i=1}^n (1-\varepsilon)/n}{\prod_{i=1}^n 1/n} = (1-\varepsilon)^n$$

- $LR^*(G_\varepsilon)$ can be arbitrarily close to 1, while G_ε may have an arbitrarily large/small mean
- Allowing such distributions gives CR for the mean covering all of $\mathbb{R}^p$
- One way out is to restrict $\mathcal{G}$ to distributions with domain equal to the convex hull of the data
- Any mean in the convex hull can be represented solely by a linear combinations of the sampled values
- $\Rightarrow$ no means are ruled out by requiring $\sum_i w_i = 1$
- Consequently, $\mathcal{R}^*(\mu)$ takes the form

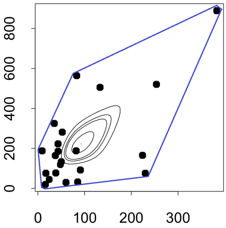
$$\mathcal{R}^*(\mu) = \sup\{\prod_{i=1}^n nw_i \,|\, T(G) = \mu, w_i \geq 0, \sum_{i=1}^n w_i = 1\}$$

Restricting the distribution space

- Consider $G_\varepsilon = (1-\varepsilon)\widehat{G}_n + \varepsilon\delta_y$, with

$$LR^*(G_\varepsilon) = L^*(G_\varepsilon)/L^*(\widehat{G}_n) = \frac{\prod_{i=1}^n (1-\varepsilon)/n}{\prod_{i=1}^n 1/n} = (1-\varepsilon)^n$$

- $LR^*(G_\varepsilon)$ can be arbitrarily close to 1, while G_ε may have an arbitrarily large/small mean
- Allowing such distributions gives CR for the mean covering all of $\mathbb{R}^p$
- One way out is to restrict $\mathcal{G}$ to distributions with domain equal to the convex hull of the data



- Any mean in the convex hull can be represented solely by a linear combinations of the sampled values
- $\Rightarrow$ no means are ruled out by requiring $\sum_i w_i = 1$
- Consequently, $\mathcal{R}^*(\mu)$ takes the form

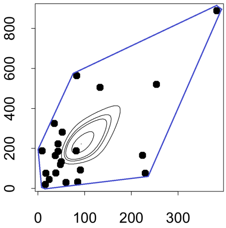
$$\mathcal{R}^*(\mu) = \sup\{\prod_{i=1}^n nw_i \,|\, T(G) = \mu, w_i \geq 0, \sum_{i=1}^n w_i = 1\}$$

Restricting the distribution space

- Consider $G_\varepsilon = (1-\varepsilon)\widehat{G}_n + \varepsilon\delta_y$, with

$$LR^*(G_\varepsilon) = L^*(G_\varepsilon)/L^*(\widehat{G}_n) = \frac{\prod_{i=1}^n (1-\varepsilon)/n}{\prod_{i=1}^n 1/n} = (1-\varepsilon)^n$$

- $LR^*(G_\varepsilon)$ can be arbitrarily close to 1, while G_ε may have an arbitrarily large/small mean
- Allowing such distributions gives CR for the mean covering all of $\mathbb{R}^p$
- One way out is to restrict $\mathcal{G}$ to distributions with domain equal to the convex hull of the data
- Any mean in the convex hull can be represented solely by a linear combinations of the sampled values
- $\Rightarrow$ no means are ruled out by requiring $\sum_i w_i = 1$
- Consequently, $\mathcal{R}^*(\mu)$ takes the form



$$\mathcal{R}^*(\mu) = \sup\{\prod_{i=1}^n nw_i \,|\, T(G) = \mu, w_i \geq 0, \sum_{i=1}^n w_i = 1\}$$

Computing EL for the univariate mean

- Consider estimation of a univariate mean: $\mu = T(G) = \int y \, \mathrm{d}G(y)$
- $\widehat{\mu}_{\mathrm{np}} = T(\widehat{G}_n) = \int y \, \mathrm{d}\widehat{G}_n(y) = n^{-1} \sum_{i=1}^n Y_i = \overline{Y}$
- Finding $\mathcal{R}^*(\mu) \Leftrightarrow$ solving the following optimisation problem:
  - Maximize $\prod_i nw_i$(or $\sum_i \log(nw_i)$) over $w_i \geq 0$ subject to $\sum_i w_i = 1$ and $\mu = \sum_{i=1}^n w_i Y_i$
- All relevant $\mu \in [\min Y_i, \max Y_i]$
- $\sum_i \log(nw_i)$ is strictly concave, convex set of weights
- Lagrange multiplier problem, optimizing:

$$\sum_{i=1}^n \log(nw_i) - n\lambda \sum_{i=1}^n (Y_i - \mu) + \gamma \left(\sum_{i=1}^n w_i - 1 \right), \tag{1}$$

  for λ and γ Lagrange multipliers
- One finds $\gamma = -n$ and λ the root of

$$n^{-1} \sum_{i=1}^n \frac{Y_i - \mu}{1 + \lambda(Y_i - \mu)} \tag{2}$$

Computing EL for the univariate mean

- Consider estimation of a univariate mean: $\mu = T(G) = \int y \, \mathrm{d}G(y)$
- $\widehat{\mu}_{\mathrm{np}} = T(\widehat{G}_n) = \int y \, \mathrm{d}\widehat{G}_n(y) = n^{-1} \sum_{i=1}^{n} Y_i = \overline{Y}$
- Finding $\mathcal{R}^*(\mu) \Leftrightarrow$ solving the following optimisation problem:
  - Maximize $\prod_i nw_i$(or $\sum_i \log(nw_i)$) over $w_i \geq 0$ subject to $\sum_i w_i = 1$ and $\mu = \sum_{i=1}^{n} w_i Y_i$
- All relevant $\mu \in [\min Y_i, \max Y_i]$
- $\sum_i \log(nw_i)$ is strictly concave, convex set of weights
- Lagrange multiplier problem, optimizing:

$$\sum_{i=1}^{n} \log(nw_i) - n\lambda \sum_{i=1}^{n} (Y_i - \mu) + \gamma \left(\sum_{i=1}^{n} w_i - 1 \right), \tag{1}$$

  for λ and γ Lagrange multipliers
- One finds $\gamma = -n$ and λ the root of

$$n^{-1} \sum_{i=1}^{n} \frac{Y_i - \mu}{1 + \lambda(Y_i - \mu)} \tag{2}$$

Computing EL for the univariate mean

- Consider estimation of a univariate mean: $\mu = T(G) = \int y \, \mathrm{d}G(y)$
- $\widehat{\mu}_{\mathrm{np}} = T(\widehat{G}_n) = \int y \, \mathrm{d}\widehat{G}_n(y) = n^{-1} \sum_{i=1}^n Y_i = \overline{Y}$
- Finding $\mathcal{R}^*(\mu) \Leftrightarrow$ solving the following optimisation problem:
  - Maximize $\prod_i nw_i$(or $\sum_i \log(nw_i)$) over $w_i \geq 0$ subject to $\sum_i w_i = 1$ and $\mu = \sum_{i=1}^n w_i Y_i$
- All relevant $\mu \in [\min Y_i, \max Y_i]$
- $\sum_i \log(nw_i)$ is strictly concave, convex set of weights
- Lagrange multiplier problem, optimizing:

$$\sum_{i=1}^n \log(nw_i) - n\lambda \sum_{i=1}^n (Y_i - \mu) + \gamma \left(\sum_{i=1}^n w_i - 1 \right), \tag{1}$$

  for λ and γ Lagrange multipliers
- One finds $\gamma = -n$ and λ the root of

$$n^{-1} \sum_{i=1}^n \frac{Y_i - \mu}{1 + \lambda(Y_i - \mu)} \tag{2}$$

Computing EL for the univariate mean

- Consider estimation of a univariate mean: $\mu = T(G) = \int y \, \mathrm{d}G(y)$
- $\widehat{\mu}_{\text{np}} = T(\widehat{G}_n) = \int y \, \mathrm{d}\widehat{G}_n(y) = n^{-1} \sum_{i=1}^n Y_i = \overline{Y}$
- Finding $\mathcal{R}^*(\mu) \Leftrightarrow$ solving the following optimisation problem:
  - Maximize $\prod_i nw_i$(or $\sum_i \log(nw_i)$) over $w_i \geq 0$ subject to $\sum_i w_i = 1$ and $\mu = \sum_{i=1}^n w_i Y_i$
- All relevant $\mu \in [\min Y_i, \max Y_i]$
- $\sum_i \log(nw_i)$ is strictly concave, convex set of weights
- Lagrange multiplier problem, optimizing:

$$\sum_{i=1}^n \log(nw_i) - n\lambda \sum_{i=1}^n (Y_i - \mu) + \gamma \left(\sum_{i=1}^n w_i - 1 \right), \tag{1}$$

  for λ and γ Lagrange multipliers
- One finds $\gamma = -n$ and λ the root of

$$n^{-1} \sum_{i=1}^n \frac{Y_i - \mu}{1 + \lambda(Y_i - \mu)} \tag{2}$$

Proving ELT for the univariate mean

- The proof for the χ^2-limit for the EL is non-trivial
- Taylor expand (2) on the previous slide around $\lambda = 0$ (corresponding to $\mu = \overline{Y}$) to find
$$\lambda \approx (\overline{Y} - \mu)/\widehat{\sigma}^2,$$
where $\widehat{\sigma}^2$ is the sample variance
- Inserting that approximation in (1), and another Taylor expansion of that expression shows that
$$-2 \log \mathcal{R}^*(\mu_{\text{true}}) \approx n(\overline{Y} - \mu_{\text{true}})^2/\widehat{\sigma}^2$$
- $n(\overline{Y} - \mu_{\text{true}})^2/\widehat{\sigma}^2 \to_L \chi^2_1$, since by CLT $\sqrt{n}(\overline{Y} - \mu_{\text{true}})/\widehat{\sigma} \to_L \mathrm{N}(0, 1)$

Proving ELT for the univariate mean

- The proof for the χ^2-limit for the EL is non-trivial
- Taylor expand (2) on the previous slide around $\lambda = 0$ (corresponding to $\mu = \overline{Y}$) to find
$$\lambda \approx (\overline{Y} - \mu)/\widehat{\sigma}^2,$$
where $\widehat{\sigma}^2$ is the sample variance
- Inserting that approximation in (1), and another Taylor expansion of that expression shows that
$$-2 \log \mathcal{R}^*(\mu_{\text{true}}) \approx n(\overline{Y} - \mu_{\text{true}})^2/\widehat{\sigma}^2$$
- $n(\overline{Y} - \mu_{\text{true}})^2/\widehat{\sigma}^2 \to_L \chi^2_1$, since by CLT $\sqrt{n}(\overline{Y} - \mu_{\text{true}})/\widehat{\sigma} \to_L \mathrm{N}(0, 1)$

Illustration: Stock data (Owen, 2001)

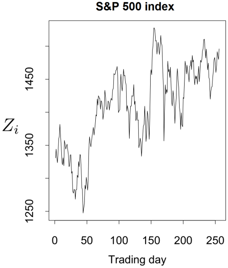
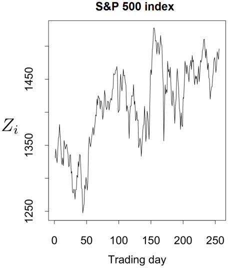


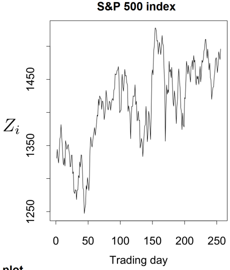
Illustration: Stock data (Owen, 2001)

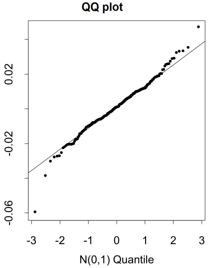


Volatility

$$\mu = \{255\,\mathrm{Var}(Y_i)\}^{1/2}$$
$$Y_i = \log(Z_i/Z_{i-1})$$

Illustration: Stock data (Owen, 2001)

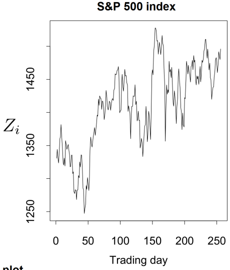




Volatility

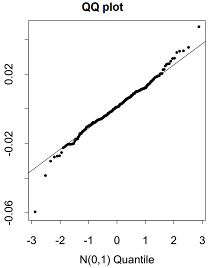
$$\mu = \{255 \operatorname{Var}(Y_i)\}^{1/2}$$
$$Y_i = \log(Z_i / Z_{i-1})$$

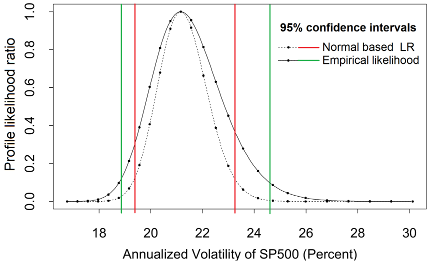
Illustration: Stock data (Owen, 2001)



Volatility

$$\mu = \{255\,\mathrm{Var}(Y_i)\}^{1/2}$$
$$Y_i = \log(Z_i/Z_{i-1})$$





Estimating equations

- A flexible way to represent interest quantities and statistics
- Specifies the interest parameter μ as the solution to

$$\mathrm{E}\{m(Y_i, \mu, \nu)\} = 0,$$

  where $m(Y_i, \mu, \nu) \in \mathbb{R}^r$ is a vector valued function of
    - vector valued data $Y_i \in \mathbb{R}^d, i = 1, \ldots, n,$
    - the interest quantity $\mu \in \mathbb{R}^p$,
    - a possible nuisance parameter $\nu \in \mathbb{R}^q$
- Also called Φ-type M-estimators or Z-estimators
- A wide range of the functionals $\mu = T(G)$ takes this form
- Examples:
    - Means: $\mathrm{E}(Y_{i1} - \mu_1) = 0, \mathrm{E}(Y_{i2} - \mu_2) = 0$
    - Event probabilities: $\mathrm{E}(1_{\{Y_i \in A\}} - \mu) = 0$ for some event A
    - Univariate quantiles: $\mathrm{E}(1_{\{Y_i \leq \mu\}} - \alpha) = 0$ for a quantile $\alpha \in (0, 1)$
    - Covariance: $\mathrm{E}(Y_{i1} - \nu_1) = 0, \mathrm{E}(Y_{i2} - \nu_2) = 0$, and $\mathrm{E}\{(Y_{i1} - \nu_1)(Y_{i2} - \nu_2) - \mu\} = 0$

Estimating equations

- A flexible way to represent interest quantities and statistics
- Specifies the interest parameter μ as the solution to

$$\mathrm{E}\{m(Y_i, \mu, \nu)\} = 0,$$

where $m(Y_i, \mu, \nu) \in \mathbb{R}^r$ is a vector valued function of
  - vector valued data $Y_i \in \mathbb{R}^d, i = 1, \ldots, n$,
  - the interest quantity $\mu \in \mathbb{R}^p$,
  - a possible nuisance parameter $\nu \in \mathbb{R}^q$
- Also called Φ-type M-estimators or Z-estimators
- A wide range of the functionals $\mu = T(G)$ takes this form
- Examples:
  - Means: $\mathrm{E}(Y_{i1} - \mu_1) = 0, \mathrm{E}(Y_{i2} - \mu_2) = 0$
  - Event probabilities: $\mathrm{E}(\mathbf{1}_{\{Y_i \in A\}} - \mu) = 0$ for some event A
  - Univariate quantiles: $\mathrm{E}(\mathbf{1}_{\{Y_i \leq \mu\}} - \alpha) = 0$ for a quantile $\alpha \in (0, 1)$
  - Covariance: $\mathrm{E}(Y_{i1} - \nu_1) = 0, \mathrm{E}(Y_{i2} - \nu_2) = 0$, and $\mathrm{E}\{(Y_{i1} - \nu_1)(Y_{i2} - \nu_2) - \mu\} = 0$

ELT for estimating equations

- For a given data set, μ_{true} is estimated by solving

$$n^{-1}\sum_{i=1}^{n} m(Y_i;\mu,\nu) = 0, \quad \text{for } \mu \text{ and } \nu$$

- Define then

$$\mathcal{R}^*(\mu,\nu) = \sup\{\prod_{i=1}^{n} nw_i \mid \sum_{i=1}^{n} w_i m(Y_i;\mu,\nu) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1\}$$

$$\mathcal{R}^*(\mu) = \sup_{\nu} \mathcal{R}^*(\mu,\nu)$$

Empirical likelihood theorem (ELT)

Under weak conditions:

$$-2\log \mathcal{R}^*(\mu_{\text{true}}) \to_L \chi_p^2$$

ELT for estimating equations

- For a given data set, μ_{true} is estimated by solving

$$n^{-1}\sum_{i=1}^{n} m(Y_i;\mu,\nu) = 0, \quad \text{for } \mu \text{ and } \nu$$

- Define then

$$\mathcal{R}^*(\mu,\nu) = \sup\{\prod_{i=1}^{n} nw_i \mid \sum_{i=1}^{n} w_i m(Y_i;\mu,\nu) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1\}$$

$$\mathcal{R}^*(\mu) = \sup_{\nu} \mathcal{R}^*(\mu,\nu)$$

Empirical likelihood theorem (ELT)

Under weak conditions:

$$-2\log\mathcal{R}^*(\mu_{\text{true}}) \rightarrow_L \chi^2_p$$

ELT for estimating equations

- For a given data set, μ_{true} is estimated by solving

$$n^{-1}\sum_{i=1}^{n} m(Y_i;\mu,\nu) = 0, \quad \text{for } \mu \text{ and } \nu$$

- Define then

$$\mathcal{R}^*(\mu,\nu) = \sup\{\prod_{i=1}^{n} nw_i \mid \sum_{i=1}^{n} w_i m(Y_i;\mu,\nu) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1\}$$

$$\mathcal{R}^*(\mu) = \sup_{\nu} \mathcal{R}^*(\mu,\nu)$$

Empirical likelihood theorem (ELT)

Under weak conditions:

$$-2\log \mathcal{R}^*(\mu_{\text{true}}) \to_L \chi^2_p$$

Auxiliary information

- Sometimes one has additional knowledge related to μ
- Easy to incorporate in the EL optimisation through estimating equations

Basic examples

- Coinciding mean and median (symmetric distribution):

$$\mathrm{E}(Y_i - \mu) = 0, \quad \text{and} \quad \mathrm{E}(\mathbf{1}_{\{Y_i \leq \mu\}} - 1/2) = 0$$

- Conditional mean of Y_{i1}, given that $Y_{i2} > 7$:

$$\mathrm{E}\{(Y_{i1} - \mu)\mathbf{1}_{\{Y_{i2} > 7\}}\} = 0$$

- Also possible to combine different data sources

Auxiliary information

- Sometimes one has additional knowledge related to μ
- Easy to incorporate in the EL optimisation through estimating equations

Basic examples

- Coinciding mean and median (symmetric distribution):

$$\mathrm{E}(Y_i - \mu) = 0, \quad \text{and} \quad \mathrm{E}(\mathbf{1}_{\{Y_i \leq \mu\}} - 1/2) = 0$$

- Conditional mean of Y_{i1}, given that $Y_{i2} > 7$:

$$\mathrm{E}\{(Y_{i1} - \mu)\mathbf{1}_{\{Y_{i2} > 7\}}\} = 0$$

- Also possible to combine different data sources

Auxiliary information

- Sometimes one has additional knowledge related to μ
- Easy to incorporate in the EL optimisation through estimating equations

Basic examples

- Coinciding mean and median (symmetric distribution):

$$\mathrm{E}(Y_i - \mu) = 0, \quad \text{and} \quad \mathrm{E}(\mathbf{1}_{\{Y_i \leq \mu\}} - 1/2) = 0$$

- Conditional mean of Y_{i1}, given that $Y_{i2} > 7$:

$$\mathrm{E}\{(Y_{i1} - \mu)\mathbf{1}_{\{Y_{i2} > 7\}}\} = 0$$

- Also possible to combine different data sources

Other data and situations

Regression

- Data $(Y_i, X_i), i = 1, \ldots, n$
- Trust e.g. $E(Y_i|X_i = x) = \beta_0 + \beta_1 x$, but no normality or homoscedasticity assumption
- EL based tests and confidence regions ($\widehat{\beta}_{LS}$ optimises $L^*(G)$)
  - Random covariate: Through estimating equations
  $$\mathrm{E}(\boldsymbol{Y} - \boldsymbol{X\beta}) = 0 \quad \text{and} \quad \mathrm{E}\{(\boldsymbol{Y} - \boldsymbol{X\beta})\boldsymbol{X}\} = 0$$
  - Fixed covariates: Requires a triangular ELT
- Similarly: Generalised linear models

Other situations

- Kernel based regression and density estimation, censored data, time series data...
- Nonstandard cases: Missing data, combining parametric and nonparametrics, growing number of estimating equations, high dimension low sample size, confidence distributions and goodness-of-fit tests

Other data and situations

Regression

- Data $(Y_i, X_i), i = 1, \ldots, n$
- Trust e.g. $E(Y_i|X_i = x) = \beta_0 + \beta_1 x$, but no normality or homoscedasticity assumption
- EL based tests and confidence regions ($\widehat{\beta}_{LS}$ optimises $L^*(G)$)
  - Random covariate: Through estimating equations

$$\mathrm{E}(\boldsymbol{Y} - \boldsymbol{X\beta}) = 0 \quad \text{and} \quad \mathrm{E}\{(\boldsymbol{Y} - \boldsymbol{X\beta})\boldsymbol{X}\} = 0$$

  - Fixed covariates: Requires a triangular ELT
- Similarly: Generalised linear models

Other situations

- Kernel based regression and density estimation, censored data, time series data...
- Nonstandard cases: Missing data, combining parametric and nonparametrics, growing number of estimating equations, high dimension low sample size, confidence distributions and goodness-of-fit tests

Coverage accuracy of the EL

- Recall $C_{\text{LR},1-\alpha} = \{\mu \,|\, -2\log\mathcal{R}^*(\mu) \le \chi_p^{2,1-\alpha}\}$
  - Based on the asymptotic result $-2\log\mathcal{R}^*(\mu_{\text{true}}) \rightarrow_L \chi_p^2$
- Coverage error $CE = \Pr(\mu_{\text{true}} \in C_{\text{LR},1-\alpha}) - (1-\alpha) \rightarrow 0$
- Under weak moment conditions $CE = O(n^{-1})$
- Typically $C_{\text{LR},1-\alpha}$ undercovers for finite n, i.e. $CE < 0$
- This might be corrected by replacing the threshold $\chi_p^{2,1-\alpha}$

Coverage accuracy of the EL

- Recall $C_{\mathsf{LR},1-\alpha} = \{\mu \,|\, -2\log \mathcal{R}^*(\mu) \leq \chi_p^{2,1-\alpha}\}$
  - Based on the asymptotic result $-2\log \mathcal{R}^*(\mu_{\text{true}}) \to_L \chi_p^2$
- Coverage error $CE = \Pr(\mu_{\text{true}} \in C_{\mathsf{LR},1-\alpha}) - (1-\alpha) \to 0$
- Under weak moment conditions $CE = O(n^{-1})$
- Typically $C_{\mathsf{LR},1-\alpha}$ undercovers for finite n, i.e. $CE < 0$
- This might be corrected by replacing the threshold $\chi_p^{2,1-\alpha}$

Correcting the coverage probability

F-distribution correction

- Use $\frac{p(n-1)}{n-p} F_{p,n-p}^{1-\alpha}$ as threshold
- Larger confidence regions
- Still $CE = O(n^{-1})$, but better for small samples

Bartlett correction

- Use as threshold
$$\left(1 - \frac{a}{n}\right)^{-1} \chi_p^{2,1-\alpha} \quad \text{or} \quad \left(1 + \frac{a}{n}\right) \chi_p^{2,1-\alpha},$$
for some proper constant a.
- a typically unknown
- $CE = O(n^{-2})$ even if we estimate a

Bootstrap correction

- Resample the data B times
- For $b = 1, \ldots, B$: Compute $r^{(b)} = \mathcal{R}^*(\widehat{\mu}_{\text{np}})$ using the data from resample b
- Use as threshold the $1 - \alpha$ empirical quantile of $\{r^{(b)}\}_{b=1,\ldots,B}$

Correcting the coverage probability

F-distribution correction

- Use $\frac{p(n-1)}{n-p} F_{p,n-p}^{1-\alpha}$ as threshold
- Larger confidence regions
- Still $CE = O(n^{-1})$, but better for small samples

Bartlett correction

- Use as threshold

$$\left(1 - \frac{a}{n}\right)^{-1} \chi_p^{2,1-\alpha} \quad \text{or} \quad \left(1 + \frac{a}{n}\right) \chi_p^{2,1-\alpha},$$

  for some proper constant a.
- a typically unknown
- $CE = O(n^{-2})$ even if we estimate a

Bootstrap correction

- Resample the data B times
- For $b = 1, \ldots, B$: Compute $r^{(b)} = \mathcal{R}^*(\widehat{\mu}_{\text{np}})$ using the data from resample b
- Use as threshold the $1 - \alpha$ empirical quantile of $\{r^{(b)}\}_{b=1,\ldots,B}$

Correcting the coverage probability

F-distribution correction

- Use $\frac{p(n-1)}{n-p} F_{p,n-p}^{1-\alpha}$ as threshold
- Larger confidence regions
- Still $CE = O(n^{-1})$, but better for small samples

Bartlett correction

- Use as threshold

$$\left(1 - \frac{a}{n}\right)^{-1} \chi_p^{2,1-\alpha} \quad \text{or} \quad \left(1 + \frac{a}{n}\right) \chi_p^{2,1-\alpha},$$

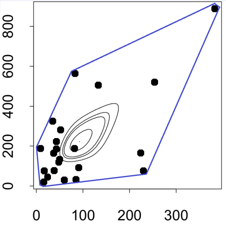
  for some proper constant a.
- a typically unknown
- $CE = O(n^{-2})$ even if we estimate a

Bootstrap correction

- Resample the data B times
- For $b = 1, \ldots, B$: Compute $r^{(b)} = \mathcal{R}^*(\widehat{\mu}_{\mathrm{np}})$ using the data from resample b
- Use as threshold the $1 - \alpha$ empirical quantile of $\{r^{(b)}\}_{b=1,\ldots,B}$

Mismatch problem

- For the mean: Only values within convex hull of the data are available
- Possibly the main reason for undercoverage



Penalized EL (Bartolucci, 2007)

- Removes the convex hull constraint and replaces it with penalisation term based on the Mahalanobis distance

Adjusted EL (Chen et. al, 2008)

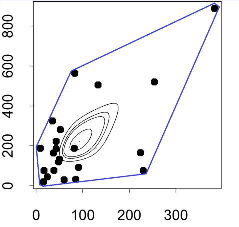
- Adds one (or two) cleverly placed pseudo-observation to the sample

Extended EL (Tsao and Wu, 2013)

- Uses $\mathcal{R}^{\circ}(\mu) = \mathcal{R}^{*}(h_n(\mu))$, where h_n is a shrinkage function ensuring μ is mapped to the proper domain
- Generally $\mathcal{R}^{*}(\mu) \leq \mathcal{R}^{\circ}(\mu)$
- Simulations studies suggests it is accurate for samples as small as 10

Mismatch problem

- For the mean: Only values within convex hull of the data are available
- Possibly the main reason for undercoverage



Penalized EL (Bartolucci, 2007)

- Removes the convex hull constraint and replaces it with penalisation term based on the Mahalanobis distance

Adjusted EL (Chen et. al, 2008)

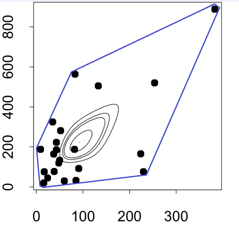
- Adds one (or two) cleverly placed pseudo-observation to the sample

Extended EL (Tsao and Wu, 2013)

- Uses $\mathcal{R}^{\circ}(\mu) = \mathcal{R}^{*}(h_n(\mu))$, where h_n is a shrinkage function ensuring μ is mapped to the proper domain
- Generally $\mathcal{R}^{*}(\mu) \leq \mathcal{R}^{\circ}(\mu)$
- Simulations studies suggests it is accurate for samples as small as 10

Mismatch problem

- For the mean: Only values within convex hull of the data are available
- Possibly the main reason for undercoverage



Penalized EL (Bartolucci, 2007)

- Removes the convex hull constraint and replaces it with penalisation term based on the Mahalanobis distance

Adjusted EL (Chen et. al, 2008)

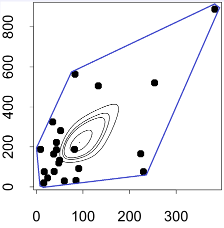
- Adds one (or two) cleverly placed pseudo-observation to the sample

Extended EL (Tsao and Wu, 2013)

- Uses $\mathcal{R}^{\circ}(\mu) = \mathcal{R}^{*}(h_n(\mu))$, where h_n is a shrinkage function ensuring μ is mapped to the proper domain
- Generally $\mathcal{R}^{*}(\mu) \leq \mathcal{R}^{\circ}(\mu)$
- Simulations studies suggests it is accurate for samples as small as 10

Mismatch problem

- For the mean: Only values within convex hull of the data are available
- Possibly the main reason for undercoverage



Penalized EL (Bartolucci, 2007)

- Removes the convex hull constraint and replaces it with penalisation term based on the Mahalanobis distance

Adjusted EL (Chen et. al, 2008)

- Adds one (or two) cleverly placed pseudo-observation to the sample

Extended EL (Tsao and Wu, 2013)

- Uses $\mathcal{R}^{\circ}(\mu) = \mathcal{R}^{*}(h_n(\mu))$, where h_n is a shrinkage function ensuring μ is mapped to the proper domain
- Generally $\mathcal{R}^{*}(\mu) \leq \mathcal{R}^{\circ}(\mu)$
- Simulations studies suggests it is accurate for samples as small as 10

Summary

- Nonparametric procedure for constructing CR and HT based on a 'likelihood approach'
- Alternative to bootstrap (and jackknife)
- Combines the reliability of nonparametrics with flexibility and effectivness of the likelihood approach
- Empirical likelihood theorem (ELT): Nonparametric analogue to Wilks's theorem
- CR and HT constructed by solving a (convex) optimisation problem
- Gives data-shaped CR
- 'Easy' to incorporate auxiliary information
- Efficiency and power comparable to parametric approaches
- The optimization problem is 'easy' for estimating equations in lower dimensions, but may be harder and computationally intensive in general
- Several adjustment routines and extensions
- `R`-packages:
  - `emplik`: EL for estimating equations (with censored data)
  - `eel`: Extended EL for estimating equations

Suggested further readings

- General methodology: Owen (2001), Empirical likelihood, Chapman & Hall
- Regression: Chen and Van Keilegom (2009), A review on empirical likelihood methods for regression, Test
- Extended EL: Tsao and Wu (2013), Extended empirical likelihood on the full parameter space, Annals of Statistics