Focused model selection and inference using robust estimators

Sam-Erik Walker (joint work with Nils Lid Hjort)

Dept. of Math., Univ. of Oslo, 23 May 2017

Outline

1. The FIC approach
2. FIC with robust elements
3. Some examples
4. Concluding remarks

The FIC approach (Jullum & Hjort, 2016)

- Assume data $y_1,\ldots,y_n \sim G$ i.i.d.
- A scalar focus parameter $\mu = \mu(G)$ is defined, e.g. a mean or a quantile, or more generally a smooth function of such quantities
- An estimate of μ based on the non-parametric alternative is $\hat{\mu}_{\text{np}} = \mu\left(\hat{G}_n\right)$, i.e. the focus parameter functional applied to the empirical distribution
- Estimates of μ can be based on one or more parametric models through $\hat{\mu}_{\text{pm}} = \mu\left(F\left(\cdot,\hat{\theta}_n\right)\right)$, where $\hat{\theta}_n$ are model parameters estimated using MLE

- The best alternative minimizes “squared bias + variance”

The FIC approach (Jullum & Hjort, 2016)

potentially non-robust

- Assume data $y_1,\ldots,y_n \sim G$ i.i.d.
- A scalar focus parameter $\mu = \mu(G)$ is defined, e.g. a mean or a quantile, or more generally a smooth function of such quantities
- An estimate of μ based on the non-parametric alternative is $\hat{\mu}_{\text{np}} = \mu\left(\hat{G}_n\right)$, i.e. the focus parameter functional applied to the empirical distribution
- Estimates of μ can be based on one or more parametric models through $\hat{\mu}_{\text{pm}} = \mu\left(F\left(\cdot,\hat{\theta}_n\right)\right)$, where $\hat{\theta}_n$ are model parameters estimated using MLE

potentially non-robust

- The best alternative minimizes "squared bias + variance"

FIC with robust elements

- When can it be reasonable to use robust statistics and estimators in the FIC approach?
  - When we suspect that the data might be contaminated by erroneous data/outliers, which could have a large impact on the focus parameter being estimated
  - When we suspect that the data generating distribution is heavy-tailed, so that the mean and higher-order moments might be suspicious to use, and where a relatively large portion of the data is in the tails of the distribution - while the focus parameter aims at characterizing properties of the more central (or head) part of the distribution
  - When we wish to investigate whether robust estimators might have an edge as compared with MLE in terms of squared bias + variance, even if robustness is not an issue

FIC with robust elements

- Replace the mean with e.g. an α-trimmed mean, where we remove a fraction $0 < \alpha \leq 0.5$ of the data on each side of the distribution before taking the mean
- Quantiles are robust if they are not too extreme, e.g. up to 0.8 (0.2) or 0.9 (0.1)
- Standard deviation may be replaced by robust measures of scale such as e.g. the MAD (Median Absolute Deviation from the median)
- A set of general robust estimators for parametric models is the class of Density Power Divergence (DPD) estimators, a.k.a. BHHJ (Basu et al., 1998)

FIC formulae revisited

$$\hat{b} = \hat{\mu}_{\text{pm}} - \hat{\mu}_{\text{np}}; \quad \hat{\kappa} = \hat{v}_{\text{pm}} + \hat{v}_{\text{np}} - 2\hat{v}_{\text{c}}; \quad \hat{V} = \begin{pmatrix} \hat{v}_{\text{np}} & \hat{v}_{\text{c}} \\ \hat{v}_{\text{c}} & \hat{v}_{\text{pm}} \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \text{IF}_{\text{np}}\left(y_i, \hat{G}_n\right) \\ \text{IF}_{\text{pm}}\left(y_i, \hat{G}_n\right) \end{pmatrix} \begin{pmatrix} \text{IF}_{\text{np}}\left(y_i, \hat{G}_n\right) \\ \text{IF}_{\text{pm}}\left(y_i, \hat{G}_n\right) \end{pmatrix}^t$$

$$\text{FIC}_{\text{np}} = 0^2 + \frac{\hat{v}_{\text{np}}}{n}; \quad \text{FIC}_{\text{pm}} = \max\left(0, \hat{b}^2 - \frac{\hat{\kappa}}{n}\right) + \frac{\hat{v}_{\text{pm}}}{n}$$

Master Lemma: $$\begin{pmatrix} \sqrt{n}\left(\hat{\mu}_{\text{np}} - \mu\right) \\ \sqrt{n}\left(\hat{\mu}_{\text{pm}} - \mu\right) \end{pmatrix} \rightarrow_d N\left(\begin{pmatrix} 0 \\ b \end{pmatrix}, \begin{pmatrix} v_{\text{np}} & v_{\text{c}} \\ v_{\text{c}} & v_{\text{pm}} \end{pmatrix}\right)$$

Robust nonparametric alternative: The trimmed mean

$$\mu_{\mathrm{np},\alpha} = \frac{1}{1-2\alpha}\int_{y_\alpha}^{y_{1-\alpha}} y dG; \quad y_\alpha = G^{-1}(\alpha); \quad y_{1-\alpha} = G^{-1}(1-\alpha); \quad \alpha = 0: \text{Mean}; \quad \alpha = 0.5: \text{Median}$$

$$\hat{\mu}_{\mathrm{np},\alpha} = \frac{1}{n-2m}\sum_{i=m+1}^{n-m} y_{(i)}; \quad m = \lfloor \alpha n \rfloor \text{ i.e. } \alpha n \text{ rounded down to nearest integer}$$

$$\mathrm{IF}_{\mathrm{np},\alpha}\left(y_i, \hat{G}_n\right) = \begin{cases} \dfrac{1}{1-2\alpha}\left\{\hat{G}_n^{-1}(\alpha) - \bar{w}_\alpha\right\} & \text{for} \quad y_i < \hat{G}_n^{-1}(\alpha) \\ \dfrac{1}{1-2\alpha}\left\{y_i - \bar{w}_\alpha\right\} & \text{for} \quad \hat{G}_n^{-1}(\alpha) \le y_i \le \hat{G}_n^{-1}(1-\alpha) \\ \dfrac{1}{1-2\alpha}\left\{\hat{G}_n^{-1}(1-\alpha) - \bar{w}_\alpha\right\} & \text{for} \quad y_i > \hat{G}_n^{-1}(1-\alpha) \end{cases}$$

$$\bar{w}_\alpha = (1-2\alpha)\hat{\mu}_{\mathrm{np},\alpha} + \alpha\left(\hat{G}_n^{-1}(\alpha) + \hat{G}_n^{-1}(1-\alpha)\right): \text{The } \alpha\text{-winsorized mean}$$

Robust parametric alternative: Trimmed mean using DPD

$$\hat{\mu}_{\mathrm{pm},\alpha} = \mu_\alpha\left(F_{\hat{\theta}_n}\right) = \frac{1}{1-2\alpha}\int_{y_\alpha}^{y_{1-\alpha}} y f\left(y,\hat{\theta}_n\right)dy; \quad y_\alpha = F^{-1}\left(\alpha,\hat{\theta}_n\right); \quad y_{1-\alpha} = F^{-1}\left(1-\alpha,\hat{\theta}_n\right)$$

$$\mathrm{IF}_{\mathrm{pm},\alpha}\left(y_i,\hat{G}_n\right) = \left(\frac{\partial\mu}{\partial\theta}\right)^t\Bigg|_{\hat{\theta}_n} \mathrm{IF}_{\mathrm{DPD}}\left(y_i,\hat{G}_n\right); \quad \text{with} \quad \frac{\partial\mu}{\partial\theta_j}\Bigg|_{\hat{\theta}_n} \approx \frac{\mu\left(F_{\hat{\theta}_{n,j}+\varepsilon}\right)-\mu\left(F_{\hat{\theta}_{n,j}-\varepsilon}\right)}{2\varepsilon}$$

$$\hat{\theta}_n = \arg\max_\theta \frac{1}{n}\sum_{i=1}^{n}\frac{a+1}{a} f\left(y_i,\hat{\theta}_n\right)^a - \int f\left(y_i,\hat{\theta}_n\right)^{1+a} dy; \quad a \geq 0; \quad a = 0 : \mathrm{MLE}$$

$$\mathrm{IF}_{\mathrm{DPD}}\left(y_i,\hat{G}_n\right) = \hat{J}^{-1}\left(f\left(y_i,\hat{\theta}_n\right)^a u\left(y_i,\hat{\theta}_n\right) - \xi\left(\hat{\theta}_n,a\right)\right); \quad \hat{J}: \text{Hessian matrix}$$

$$u\left(y,\hat{\theta}_n\right) = \nabla_\theta \log f\left(y,\hat{\theta}_n\right); \quad \xi\left(\hat{\theta}_n,a\right) = \int f\left(y,\hat{\theta}_n\right)^{1+a} u\left(y,\hat{\theta}_n\right)dy$$

Example 1: Estimation of location of a symmetric heavy-tailed distribution

- Cauchy distribution: $g\left(y;\xi,\gamma\right)=\left\{\pi\gamma\left(1+\left(\left(y-\xi\right)/\gamma\right)^{2}\right)\right\}^{-1};\ -\infty<y<\infty$
- The mean $\mu\left(G\right)=\int_{-\infty}^{+\infty}ydG$ is undefined, but the α-trimmed mean

$\mu_{\alpha}\left(G\right)=\left(1-2\alpha\right)^{-1}\int_{G^{-1}(\alpha)}^{G^{-1}(1-\alpha)}ydG$ exists, and is equal to ξ for all $0<\alpha\leq 0.5$

- We use $\mu_{\alpha}\left(\hat{G}_{n}\right)=\left(1-2m\right)^{-1}\sum_{i=m+1}^{n-m}y_{(i)};\quad m=\lfloor\alpha n\rfloor$ as the nonparametric alt.

- For the parametric alternatives we use $\mu_{\alpha}\left(F_{\hat{\theta}_{n}}\right)$ for the following models:
  (1) A narrow normal: $N\left(\xi,1\right)$
  (2) A normal: $N\left(\xi,\sigma^{2}\right)$
  (3) A non-standardized Student's t: $t\left(\xi,\gamma,\nu\right)$ with ξ,γ,ν as location, scale, df.

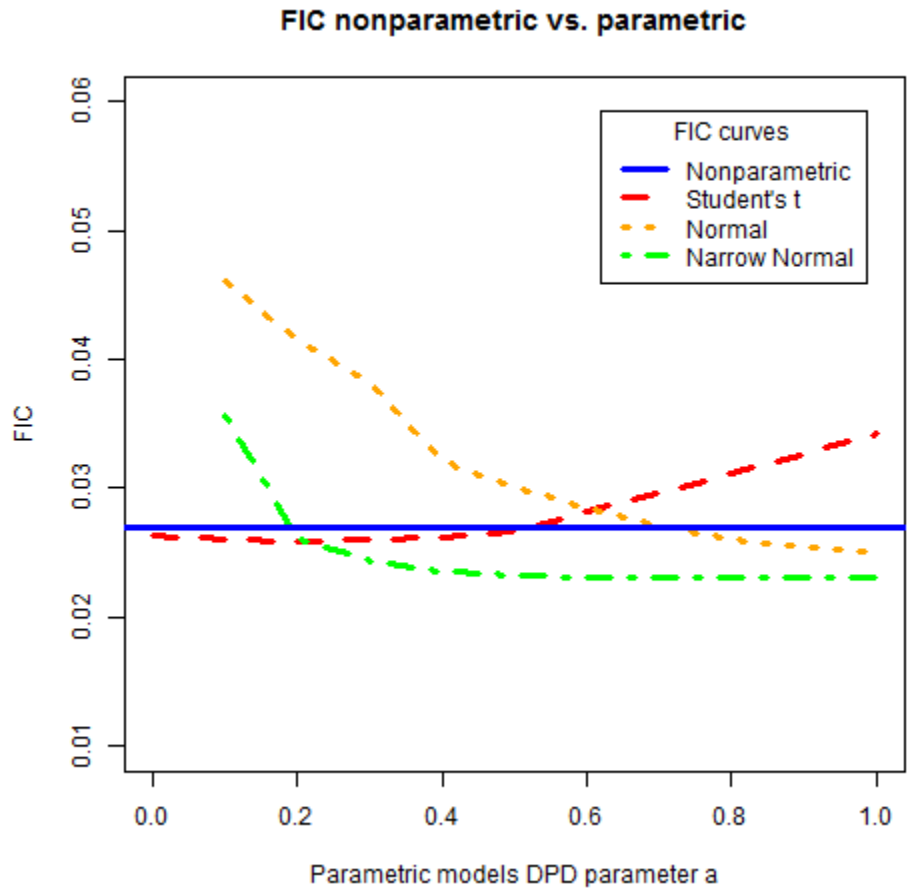
Example 1: Estimation of location of a symmetric heavy-tailed distribution

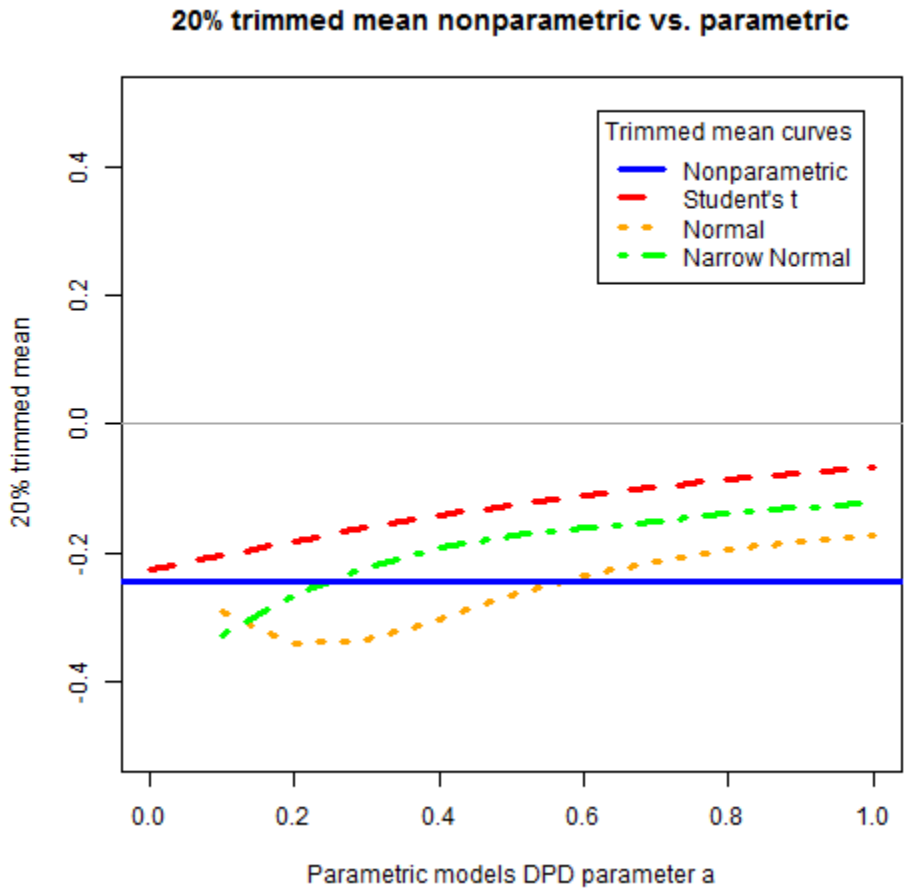
- To estimate the parameters in the parametric models given the data we use the Density Power Divergence (DPD) estimator, with robustness/efficiency tuning parameter a, from $a = 0.1$ to $a = 1$ in step of 0.1

- We try two degrees of trimming α: $\alpha = 0.2$ (most commonly used); and $\alpha = 0.38$ which is known to be optimally efficient for the estimation of the location parameter of a Cauchy distribution

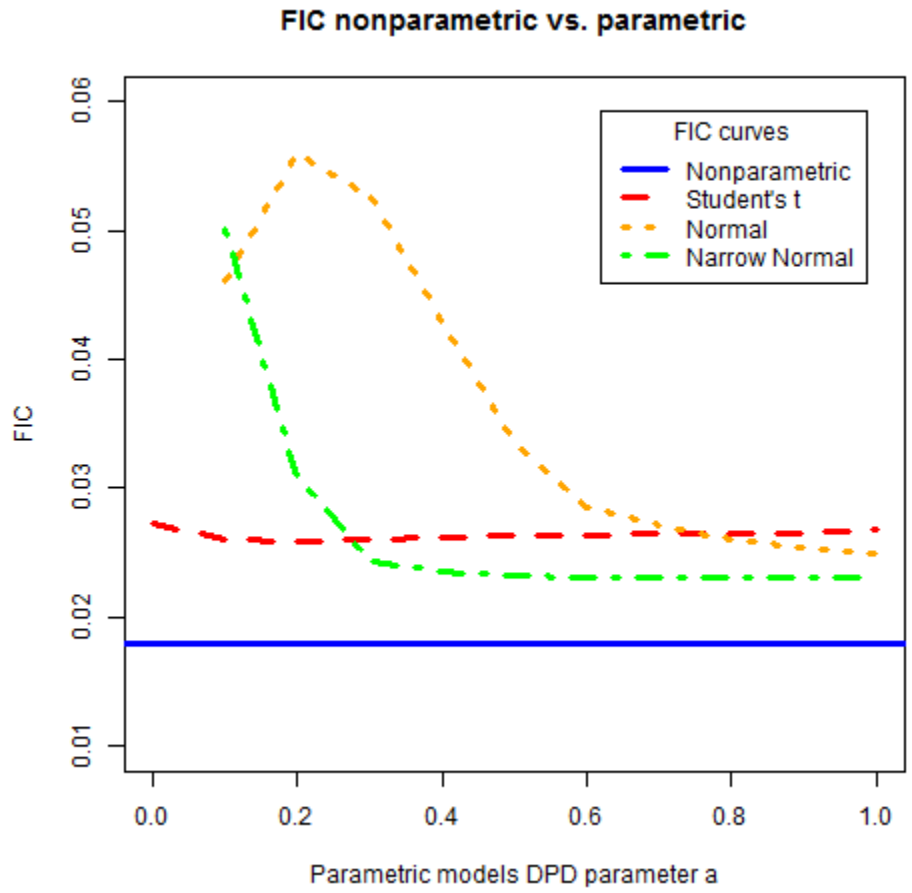
- We simulate n = 100 values from the Cauchy distribution with location and scale 0 and 1, and compare the nonparametric alternative to each of the parametric ones using the FIC criterion

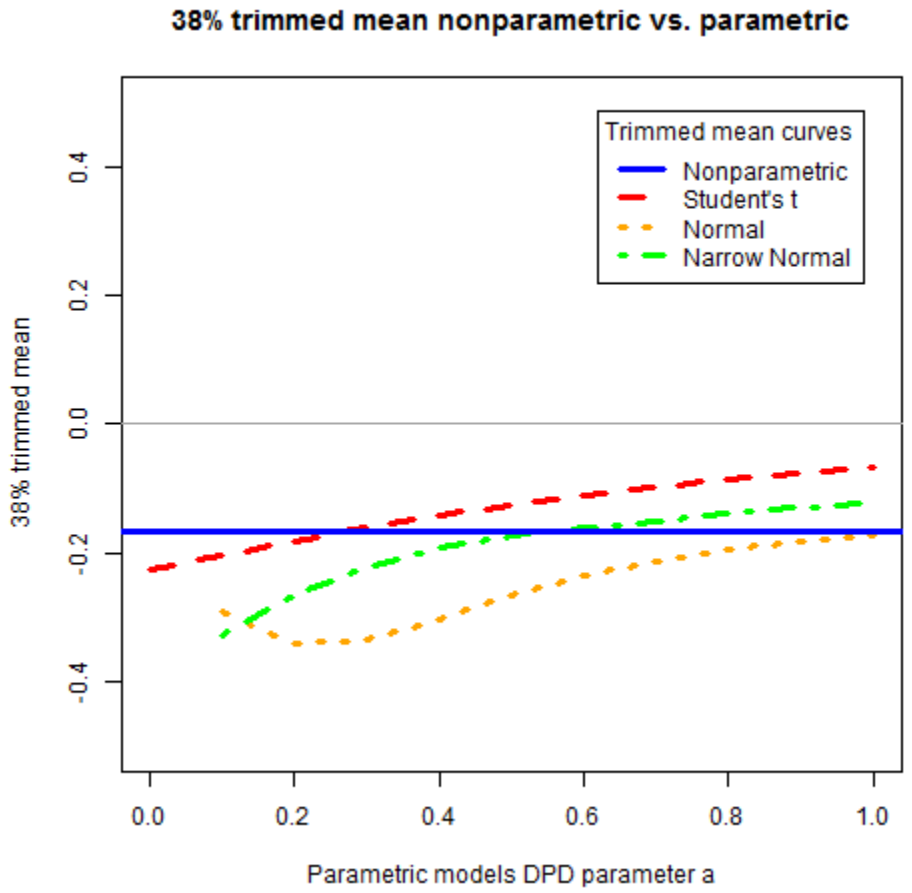
Example 1: Estimation of location of a symmetric heavy-tailed distribution: 20% trimmed mean



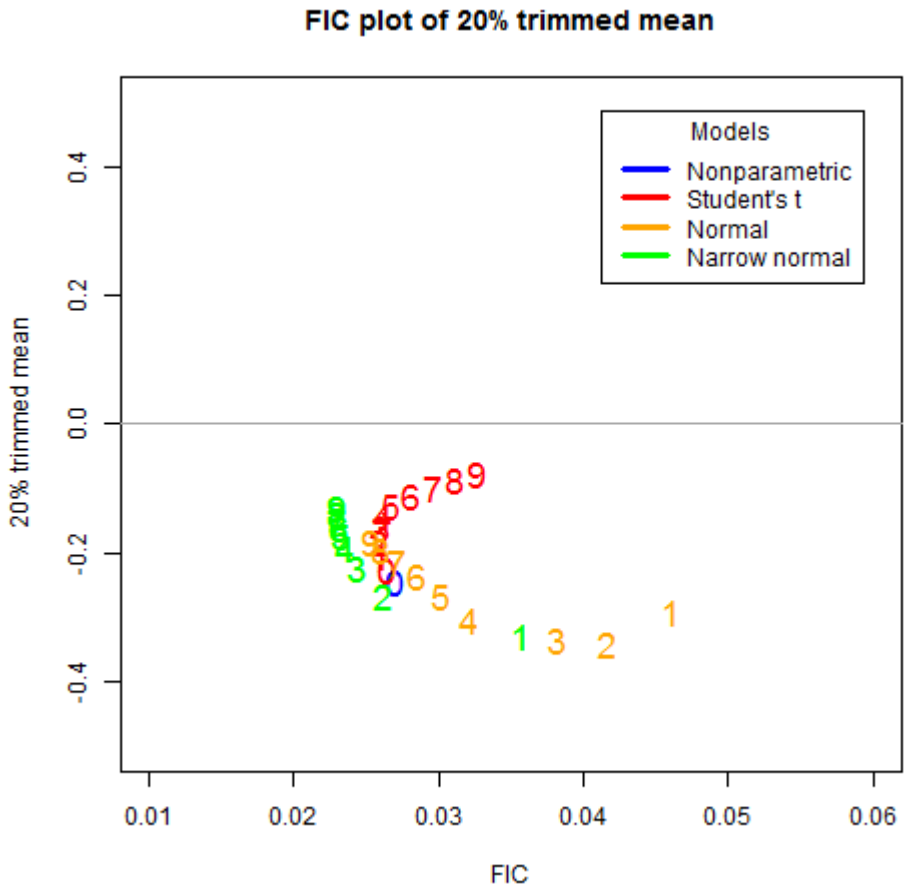


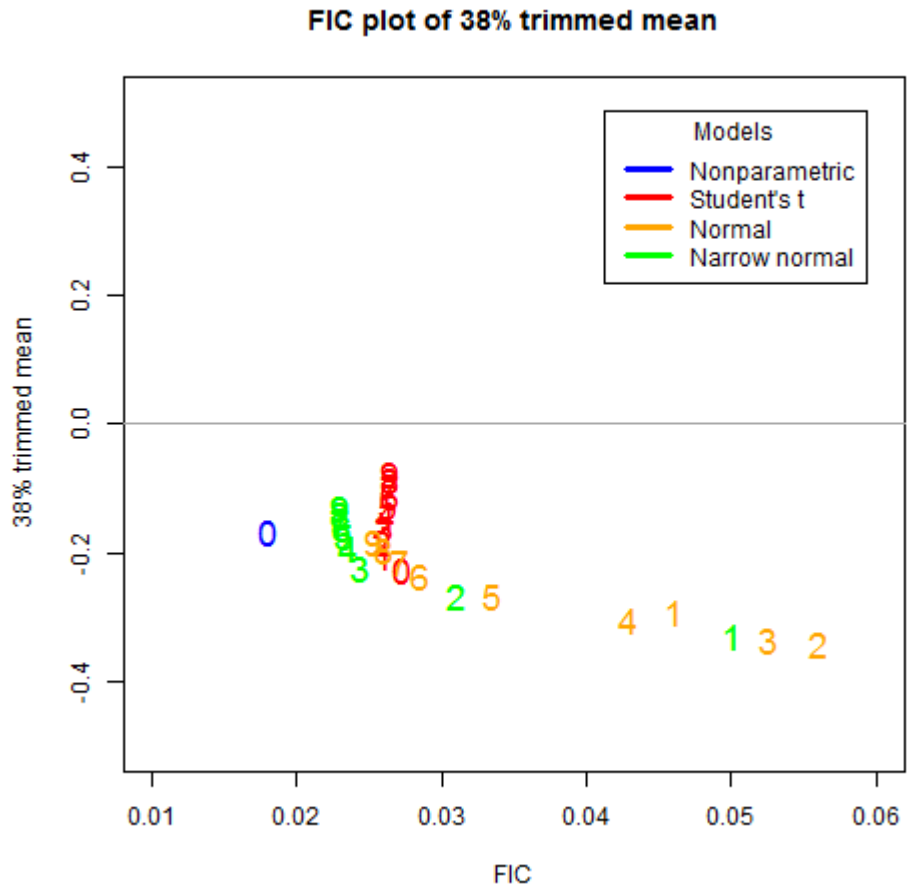
Example 1: Estimation of location of a symmetric heavy-tailed distribution: 38% trimmed mean





Example 1: Estimation of location of a symmetric heavy-tailed distribution: 20% and 38% trimmed mean





0: MLE; 1 = DPD a = 0.1; … 9 = DPD a = 0.9

Example 2: Health assessment questionnaire HAQ data fitted using Beta and Log-polynomial models

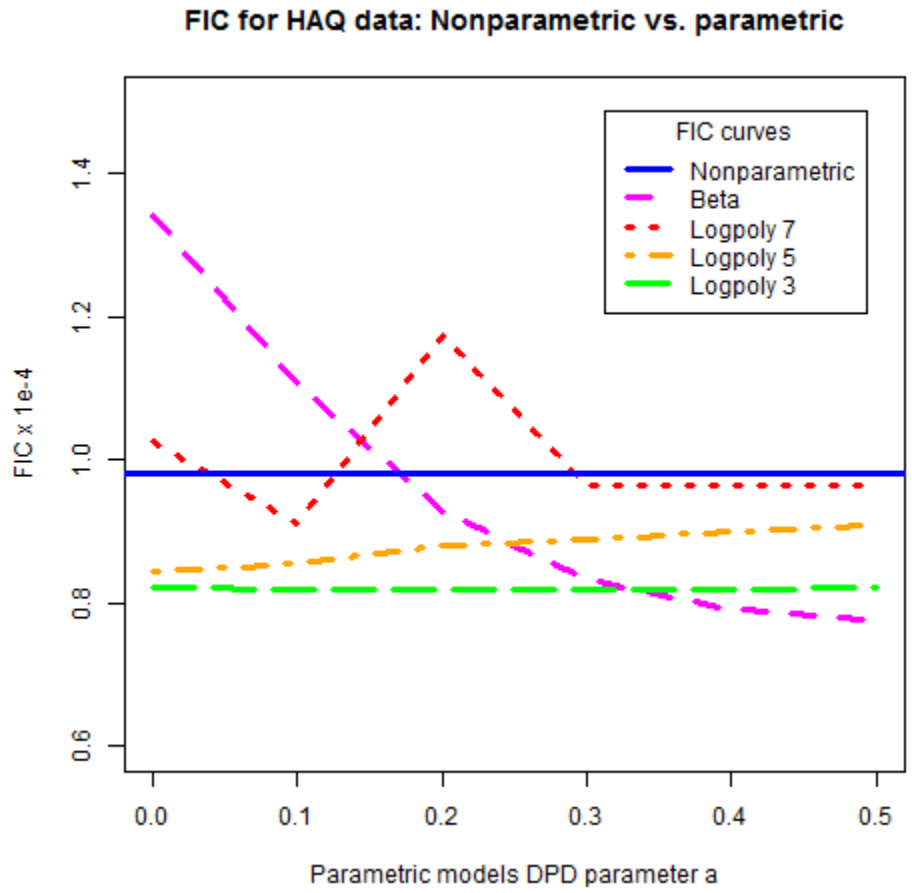
- Study involve $n = 799$ (originally 1018) patients from the Division for Women and Children at the Oslo University Hospital at Ullevål
- Data originally in the range [0,3], but mapped to the range (0,1) here using the transformation $y_i = \left(\left(HAQ_i / 3 \right) \left(n - 1 \right) + 0.5 \right) / n$ for $i = 1, \ldots, n$

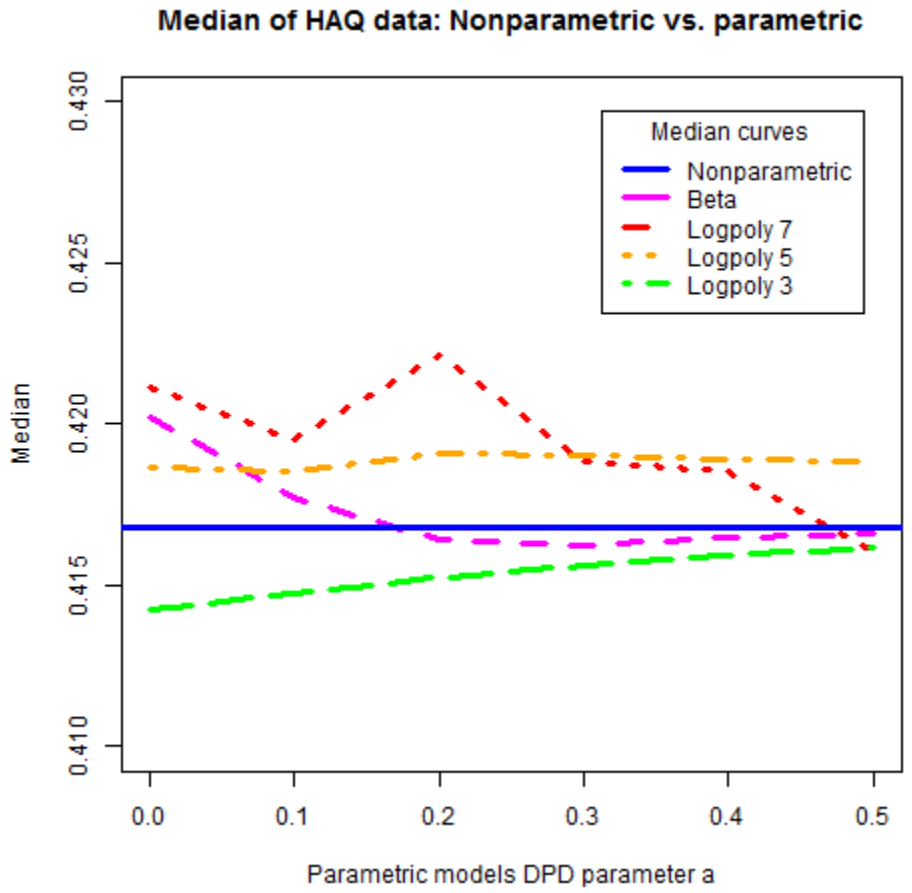
- Beta model: $f\left(y; \alpha, \beta \right) = \dfrac{\Gamma\left(\alpha + \beta \right)}{\Gamma\left(\alpha \right) \Gamma\left(\beta \right)} y^{\alpha - 1} \left(1 - y \right)^{\beta - 1}; \quad 0 < y < 1$

- Log-polynomial models:

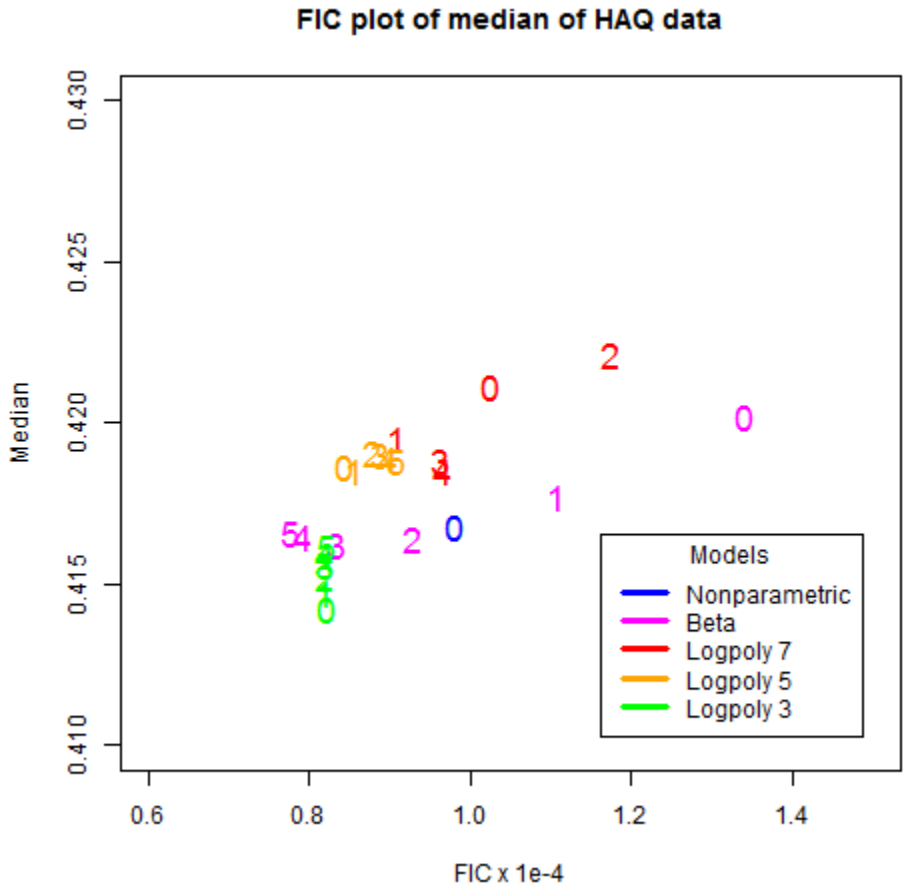
$$f\left(y; \theta_1, \ldots, \theta_p \right) = k\left(\theta_1, \ldots, \theta_p \right) \exp\left\{ \theta_1 y + \ldots + \theta_p y^p \right\}; \quad 0 < y < 1; \quad p = 1, 2, \ldots.$$

Example 2: Estimation of median of HAQ data using Beta and Log-polynomial models



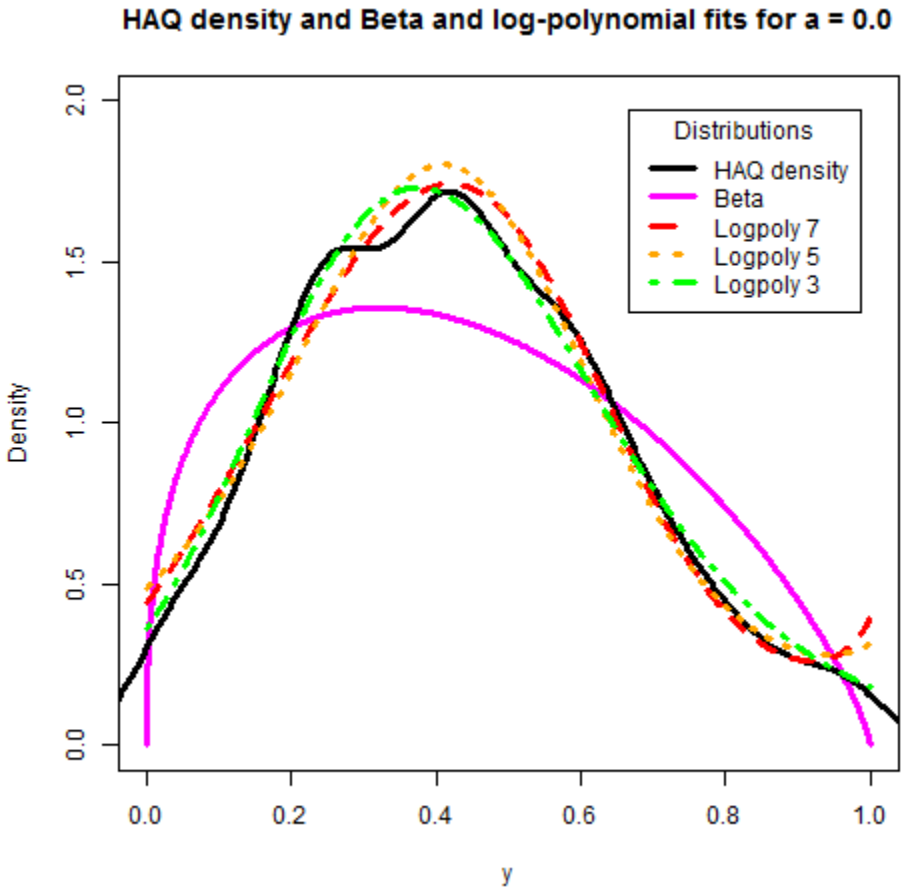


Example 2: Estimation of median of HAQ data using Beta and Log-polynomial models

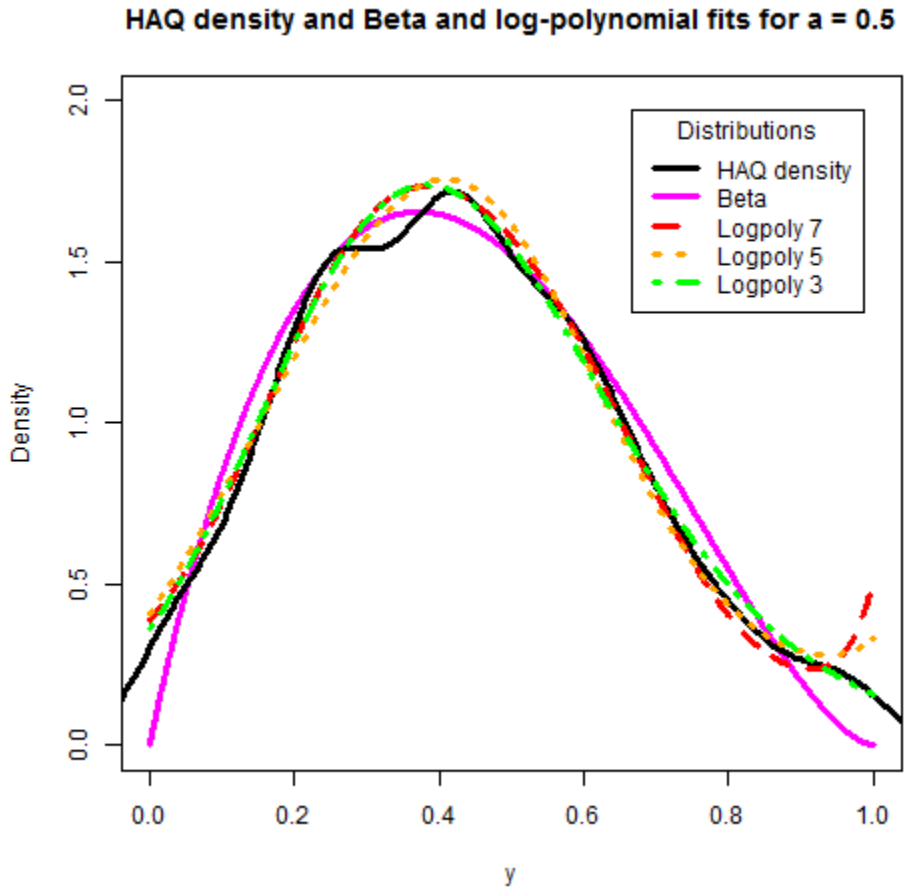


Estimators:
0: MLE
1: DPD a = 0.1
2: DPD a = 0.2
3: DPD a = 0.3
4: DPD a = 0.4
5: DPD a = 0.5

Example 2: Estimation of median of HAQ data using Beta and Log-polynomial models

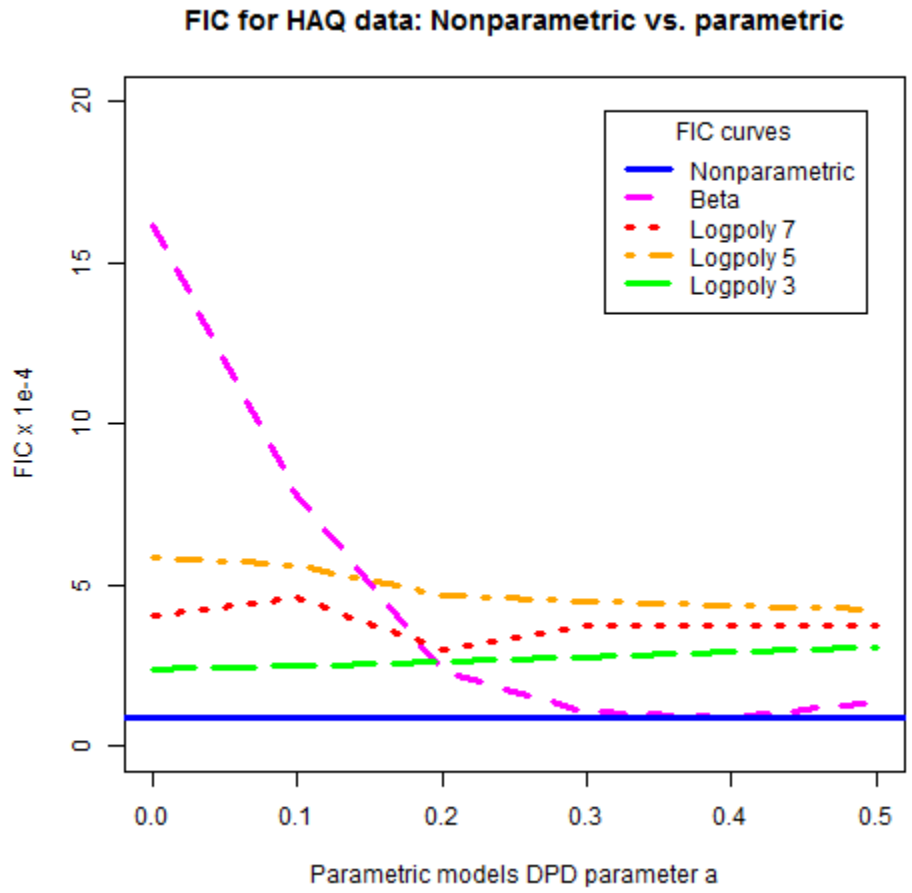


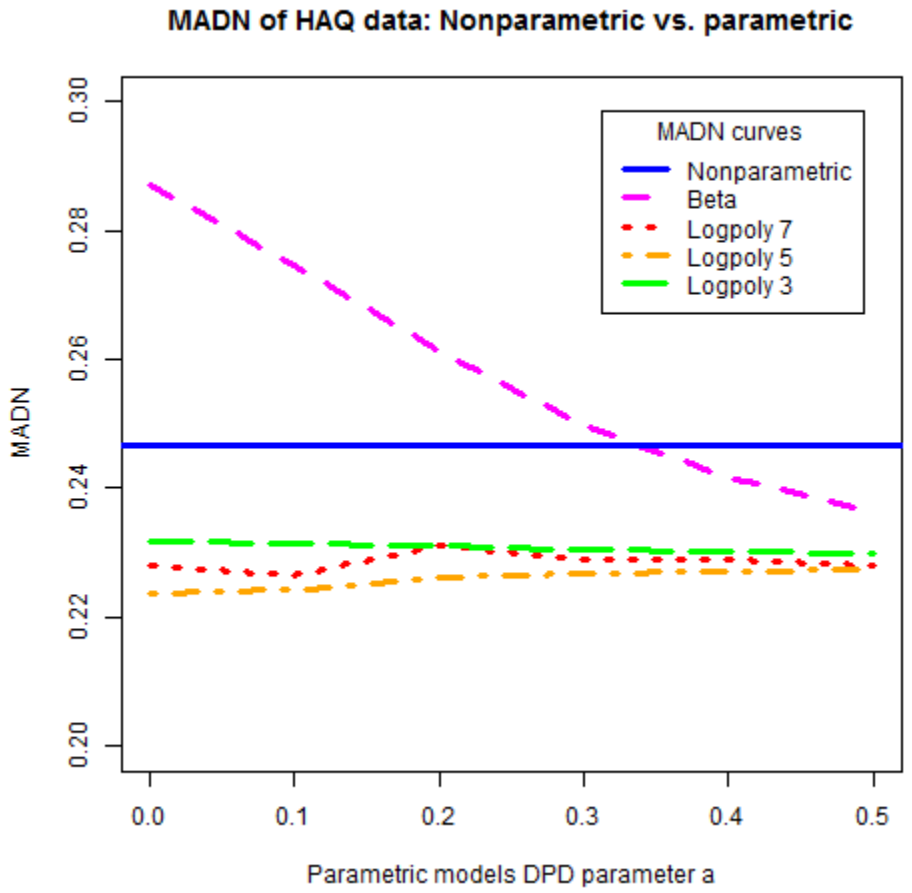
MLE



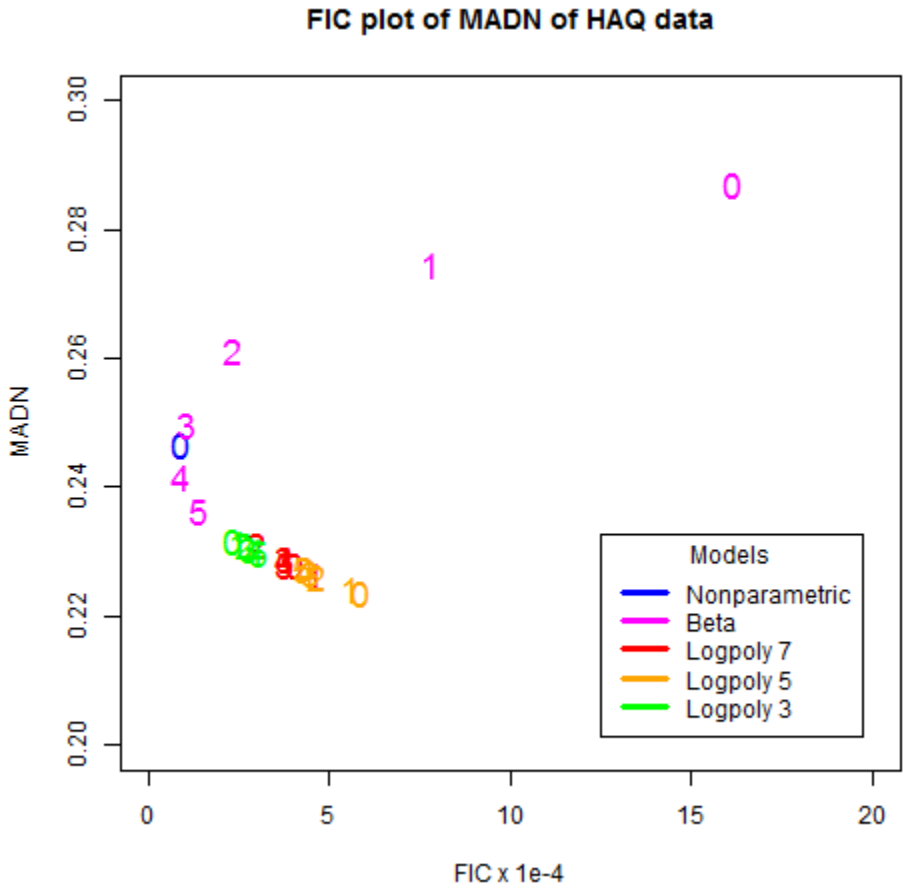
DPD with a = 0.5

Example 2: Estimation of MAD of HAQ data using Beta and Log-polynomial models





Example 2: Estimation of MAD of HAQ data using Beta and Log-polynomial models



Estimators:
0: MLE
1: DPD a = 0.1
2: DPD a = 0.2
3: DPD a = 0.3
4: DPD a = 0.4
5: DPD a = 0.5

Concluding remarks

- We have shown how focused model selection and inference using FIC with a non-parametric alternative can be extended to the use of robust statistics and estimators
- This will generally be of value when the data contains erroneous data/outliers which could have a large impact on the focus parameter being estimated
- It might also be beneficial to use robust estimators in situations where the data generating distribution is heavy-tailed, but where the focus parameter aims at characterizing properties of the more central part of the distribution
- More generally, robust estimators, such as the class of DPD estimators, might have an edge as compared with MLE in terms of FIC (squared bias + variance), even in situations where robustness might not be an issue

Building bridges?

- How to do bridge building between parametrics and nonparametrics for a focus parameter when robustness is an issue?

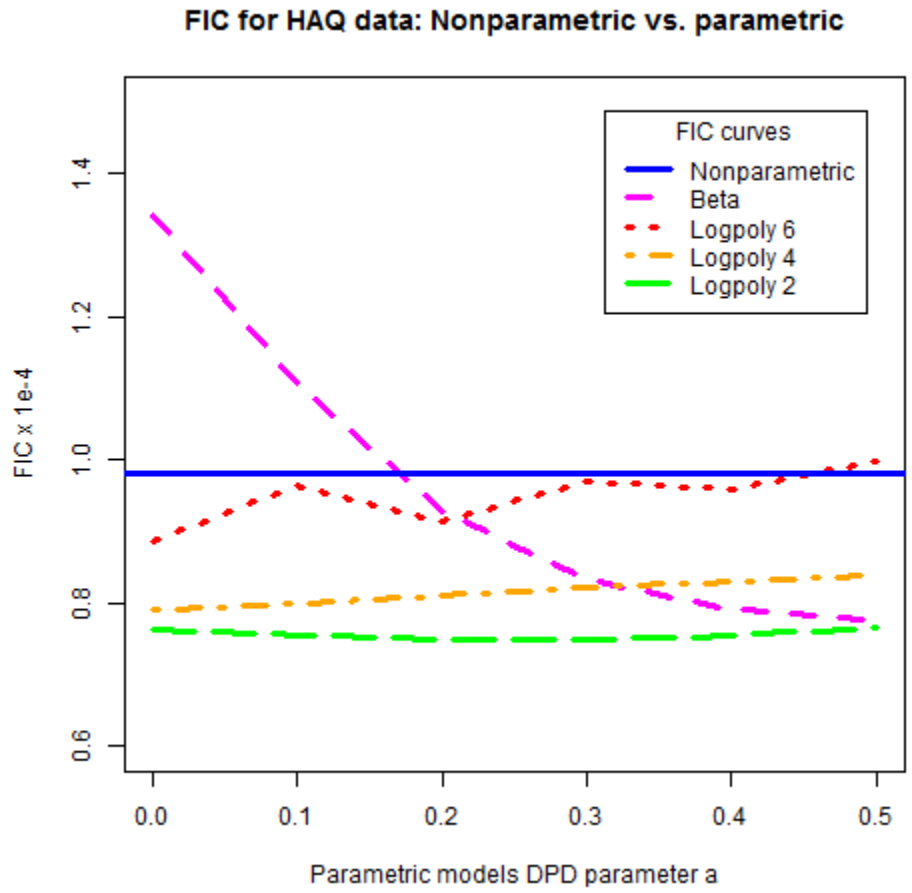
- One possibility is to combine various robust model estimates using e.g. the DPD estimator, with a given robust non-parametric alternative (statistic), using weights that depends on the calculated FIC scores (Claeskens & Hjort(2008))
- This will give highest (lowest) weights to the candidates with the lowest (highest) FIC score, and will provide an overall weighted estimate of the focus parameter

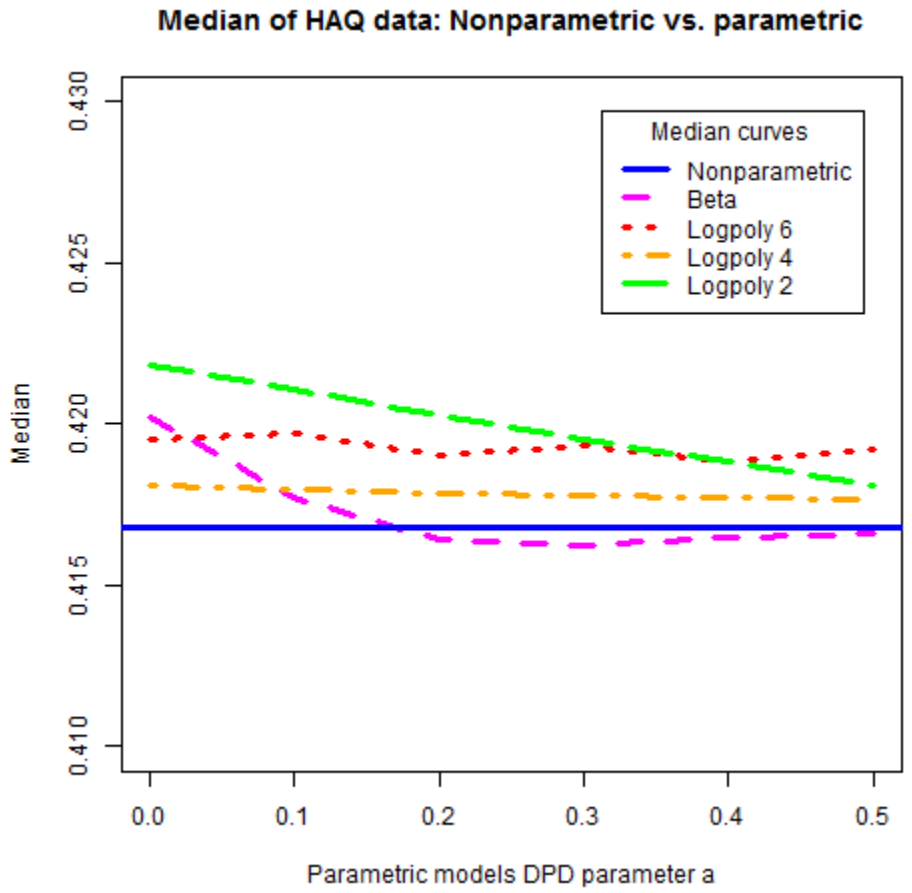
References

- Basu, A., Shioya H., Park C. (2011) Statistical Inference. The minimum distance approach. Chapman & Hall/CRC Press, Boca Raton, Florida.
- Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C. (1998) Robust and efficient estimation by minimising a density power divergence. *Biometrika* 85, 549-559.
- Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98:900-916.
- Claeskens, G. and Hjort, N. L. (2008). Model selection and model averaging. Cambridge University Press, Cambridge.
- Jullum, M. and Hjort, N.L. (2017). Parametric or nonparametric: The FIC approach. Statistica Sinica, Vol. 27.

Thanks for your attention!

Example 2: Estimation of median of HAQ data using Beta and Log-polynomial models





Example 2: Estimation of MAD of HAQ data using Beta and Log-polynomial models