Focused model selection and inference using robust estimators

Sam-Erik Walker, Nils Lid Hjort

Dept. of Mathematics, University of Oslo, Norway

BACKGROUND AND MOTIVATION

Selection of parametric models based on general information criteria like Akaike's information criterion (AIC), Bayesian information criterion (BIC), and similar, is a well-established practice within the statistical science. Over the years, more specific model selection criteria have also been developed, like the so-called focused information criterion, or FIC, where models are selected based on specifying a focus parameter – a certain parameter or function of parameters deemed most important in a given statistical setting. Up to now FIC has been mostly based on maximum likelihood estimator (MLE) methodology (Claeskens and Hjort, 2008; Jullum and Hjort, 2017). Such procedures are efficient under model conditions, but not robust.

In this work we extend the theory and application of the FIC to the use of robust estimators, such as the density power divergence (DPD) estimator (Basu et al., 1998), and to a newly developed maximum weighted likelihood (MWL) estimator (Hjort and Walker, 2018). By comparing with a robust nonparametric alternative, we may perform focused model selection and inference also in situations where the model might be misspecified, or where the data might be contaminated with atypical values or outliers.

BASIC FIC APPROACH

The basic FIC approach is defined as follows:

Focus parameter: $\mu = \mu(G)$

Parametric and nonparametric estimates: $\hat{\mu}_{\rm pm} = \mu\left(F_{\hat{\theta}}\right)$; $\hat{\mu}_{\rm np} = \mu\left(\hat{G}_n\right)$

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

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

FIC Master Lemma: $\begin{pmatrix} \sqrt{n}\left(\hat{\mu}_{\rm np} - \mu\right) \\ \sqrt{n}\left(\hat{\mu}_{\rm pm} - \mu_0\right) \end{pmatrix} \to_d N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} v_{\rm np} & v_{\rm c} \\ v_{\rm c} & v_{\rm pm} \end{pmatrix}\right)$

ROBUST FIC APPROACH

By a robust FIC approach we mean using a robust focus parameter such as e.g. a robust location measure such as a trimmed mean, median etc.; or a robust scale measure such as a trimmed standard deviation or the median of absolute deviations from the median (MAD); or other such robust measures. In addition we need to replace the non-robust MLE with robust estimators such as DPD or MWL. As long as the parametric and nonparametric estimators are asymptotically linear the basic FIC master lemma as given above holds and we can use FIC. This holds for most robust statistics and for the DPD and MWL estimators.

DENSITY POWER DIVERGENCE ESTIMATOR - DPD

The DPD estimator (Basu et al., 1998) is defined as follows:

- Assume data $y_1,\ldots,y_n \sim g(y)$ i.i.d., and let $f(y,\theta)$ be a given model density with parameters $\theta \in \Theta \subseteq \mathbb{R}^p$.
- The DPD estimator is

$$\hat{\theta}_n = \arg\max_\theta M_n(\theta) = \arg\max_\theta \left\{ \frac{1+a}{a}\frac{1}{n}\sum_{i=1}^{n} f(y_i,\theta)^a - \int f(z,\theta)^{1+a}\,dz \right\}$$

where $a \geq 0$ is a robustness vs. efficiency tuning parameter.

It is an M-estimator which is robust with a bounded influence function (B-robust) if the robustness vs. efficiency tuning parameter a > 0. Increasing a leads to increased robustness but typically less efficiency as compared with MLE. DPD approaches MLE in the limit when a -> 0.

MAXIMUM WEIGHTED LIKELIHOOD ESTIMATOR - MWL

The MWL estimator (Hjort and Walker, 2018) is defined as follows:

- Assume data $y_1,\ldots,y_n \sim g(y)$ i.i.d., and let $f(y,\theta)$ be a given model density with parameters $\theta \in \Theta \subseteq \mathbb{R}^p$.
- The MWL estimator is

Necessary penalty term for consistency

$$\hat{\theta}_n = \arg\max_\theta \ell_{{\rm w},n}(\theta \mid \tilde{\alpha}_n, \vec{y}) = \arg\max_\theta \left\{ \frac{1}{n}\sum_{i=1}^{n} w(y_i,\tilde{\alpha}_n)\log f(y_i,\theta) - \int w(y,\tilde{\alpha}_n) f(y,\theta)\,dy \right\}$$

where $w(y,\tilde{\alpha}_n)$ is an estimated weight function, with weight parameters $\tilde{\alpha}_n \in {\rm A} \subseteq \mathbb{R}^q$ estimated from the same data $y_1,\ldots,y_n$.

It is a two-step extremum estimator (Amemiya, 1985; Newey and McFadden, 1994), where in the first step the parameters of the weight function is estimated using e.g. DPD, and where the parameters of the model density is estimated in the second step applying the estimated weight function. MWL is an M-estimator if the weight function is predefined and independent of the data. If the weight function is constant and equal to one, MWL is identical to MLE.

Here we use MWL with a density threshold weight function defined as follows:

The density threshold weight function is

$$w(y,\tilde{\theta}_n) = \begin{cases} 1 & \text{if } f(y,\tilde{\theta}_n) \geq \varphi_\varepsilon \\ \exp\left(-0.5\left(\log f(y,\tilde{\theta}_n) - \log\varphi_\varepsilon\right)^2 / \sigma_w^2\right) & \text{if } f(y,\tilde{\theta}_n) < \varphi_\varepsilon, \end{cases}$$

where $\varphi_\varepsilon > 0$ is a threshold value defined so that $\int w(z,\tilde{\theta}_n) f(z,\tilde{\theta}_n)\,dz = 1-\varepsilon$, where ε is a small value, say $\varepsilon \leq 0.1$. This corresponds to approximately losing a fraction ε of the data by using this weight function. Here σ_w is set so that the weight is 0.01 when $f(y,\tilde{\theta}_n) = 0.01\varphi_\varepsilon$.

Like the density power weight function implicitly used by DPD, this is also a probabilistic weight function which down-weight data values which has low probability density under the model, but unlike DPD, weights will here be one for all other data points. Thus, there will be less down-weighting as compared with DPD. MWL will be robust with a bounded influence function (B-robust) if DPD with a > 0 is used as a weight function estimator in Step 1.

REAL CASE EXAMPLE: MAMMALS RATIO OF BODY VS. BRAIN WEIGHTS

We consider the "msleep" dataset, publicly available in the R package "ggplot2" (Wickham, 2009; R core team, 2018), which is an updated and expanded version of the "mammals sleep" dataset (Savage and West, 2007). It contains data on average brain weight (kg) and body weight (kg) of n = 56 species of mammals. Here we consider the ratio of body weight divided by brain weight for each of the n species as our data.

Fig. 1 shows a plot of these ratios against the data index (left panel), and as a histogram (right panel).

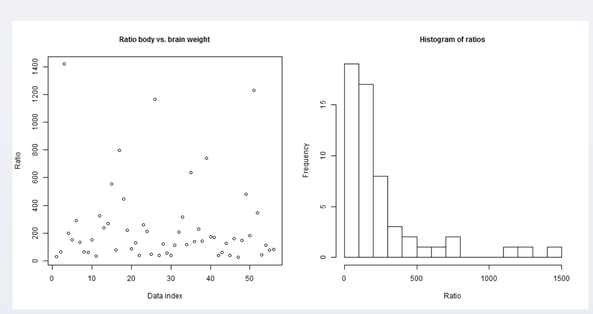

Fig. 1. Plot of ratios of body weights divided by brain weights (left) and as a histogram (right).

These data are viewed here as i.i.d. except perhaps for the three outlying data values with ratios above 1000, corresponding to Cow (1418.4), Brazilian Tapir (1227.8) and African Elephant (1164.9).

We use MAD (median of absolute deviations from the median) of the ratios as our focus parameter, which represents a robust scale estimator for these data. We perform a robust FIC analysis using four parametric candidate models: Exponential; Gamma; Weibull and Lognormal, combined with DPD and MWL estimators for a range of their respective tuning parameters, and compare FIC scores obtained with these with the FIC score obtained from the nonparametric alternative based on estimating MAD directly from the data.

Fig. 2 shows a FIC plot of MAD for the four parametric models using DPD (left panel) and MWL (right panel), with square root of FIC on the x-axis and estimated MAD values on the y-axis. For DPD, its tuning parameter a is in the range 0 (triangle) to 1.5 (square) in steps of 0.1 (points). For MLW, its tuning parameter ε is in the range 0 (triangle) to 0.1 (square) in steps of 0.01 (points), except for the Weibull model where ε ranges from 0 to 0.3 in steps of 0.03. For MWL we use DPD with a = 0.2 as weight function estimator in Step 1, which was found to give the overall smallest FIC scores in combination with ε > 0 in Step 2.

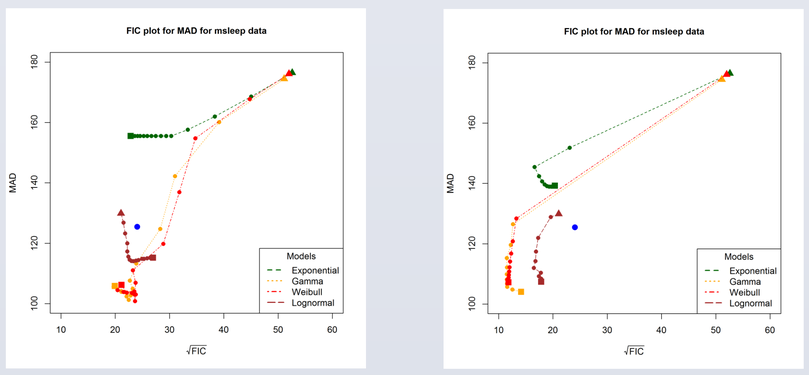

Fig. 2. FIC plot of MAD for the four candidate models based on DPD (left panel), and MWL (right panel).

For DPD (left panel), the smallest root FIC score 19.68 is obtained for the Gamma model using a = 1.4 with an estimated MAD of 105.59 (19.68). For MWL (right panel), the smallest root FIC score is 11.49 obtained for the Gamma model and MWL with a = 0.2 and ε = 0.06, with an estimated MAD of 108.17 (11.49). The nonparametric alternative gives an estimated MAD of 125.4 (24.08) but with a somewhat higher root FIC score of 24.08. The overall winner with the smallest FIC score is hence the Gamma model combined with the above best MWL estimator, with an estimated MAD of 108.17 with standard deviation 11.49, representing our best focused inference regarding the robust scale measure MAD for these mammals sleep data.

CONCLUDING REMARKS

In this work we have extended the theory and application of the FIC to the use of robust estimators, such as the density power divergence estimator (Basu et al. 1998), and to a newly developed maximum weighted likelihood estimator (Hjort and Walker, 2018).

By comparing with nonparametric alternatives, using FIC, we may perform model selection and inference in situations where the model might be misspecified, or where data might be contaminated with atypical values or outliers which could have a large impact on the estimated focus parameter.

ACKNOWLEDGEMENTS

This work was done as part of the FocuStat project at the Dept. of Mathematics, Univ. of Oslo (http://www.mn.uio.no/math/english/research/projects/focustat/), which is a 5 year project (2014 - 2018) funded by the Norwegian Research Council.

REFERENCES

Amemiya, T. (1985) Advanced Econometrics. Harvard University Press, Cambridge, MA.

Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998) Robust and efficient estimation by minimising a density power divergence. Biometrika, **85**: 549-559.

Claeskens, G., Hjort, N.L. (2008). Model selection and model averaging. Cambridge University Press, Cambridge.

Hjort, N. L. and Walker, S.-E. (2018) Estimation and model selection via weighted likelihoods. Manuscript. Department of Mathematics, University of Oslo.

Jullum, M. and Hjort, N. L. (2017) Parametric or nonparametric: The FIC approach. Statistica Sinica, **27**, 951-981.

Newey, W.K., McFadden, D. (1994) Large sample estimation and hypothesis testing, in *Handbook of Econometrics* 1994, Ch. 36, Vol. 4. Elsevier Science B.V.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Savage, V. M. and West, G. B. (2007) A quantitative, theoretical framework for understanding mammalian sleep. Proceedings of the National Academy of Sciences, **104** (3), 1051-1056.

Wickham, H. (2009) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.