Focused Regularised Likelihood

focusstat

FOCUS DRIVEN STATISTICAL
INFERENCE WITH COMPLEX DATA

Gudmund Horn Hermansen with Nils Lid Hjort

Norske Statistikermtet (i Fredrikstad)

June 15, 2017

Introduction, motivation and summary

Suppose that F_θ is our favourite (parametric) model, but that $Y_1, Y_2, \ldots Y_n$ are i.i.d. from a model with (true) distribution function G.

We do not necessarily assume that F_θ span the true G (model misspecification).

Furthermore, suppose that we are particularly concerned with some parameters $\psi_1, \ldots, \psi_r$ that are functionals of the (underlying) distribution, i.e.

$$\psi_j = \psi_j(G),$$

for $j = 1, \ldots, r$.

Some exampels are quantiles $\psi_j = G^{-1}(p_j)$ and $\psi_j = \Pr\{Y \in A_j\} = \int_{A_j} \mathrm{d}G(x)$.

Let $\ell_n(\theta)$ be the log-likelihood associated with F_θ and let

$$\widehat{\theta}_{\mathrm{ML}} = \arg\max_{\theta}\{\ell_n(\theta)\}$$

be the maximum likelihood estimator (MLE).

Under the model we estimate ψ_j by the plug-in principle, i.e. $\widehat{\psi}_{\mathrm{ML},j} = \psi_j(F_{\widehat{\theta}_{\mathrm{ML}}})$.

If our model F_θ is 'far' from the true G, then

$$\widehat{\psi}_{\mathrm{ML},j} = \psi_j(F_{\widehat{\theta}_{\mathrm{ML}}}) \doteq \psi_j(F_{\theta_0}) \neq \psi_j(G),$$

where θ_0 is the so-called least false parameter value.

Introduction, motivation and summary

To avoid problems related to misspecification, we may try a nonparametric construction, or considering a complex (bigger) parametric model.

However, suppose we want to keep the original parametric model.

If $\widehat{\psi}_{\mathrm{np},j}$ are nonparametric alternatives (e.g. empirical quantiles or probabilities) such that

$$\widehat{\psi}_{\mathrm{np},j} \doteq \psi_j(G)$$

we propose a strategy that 'solves' the issues with F_θ with respect to the ψ_j.

The idea is to penalise the MLE (under F_θ) if it (the model) is not able to match a nonparametric $\widehat{\psi}_{\mathrm{np},j}$ (i.e. if $\widehat{\psi}_{\mathrm{np},j}$ are far from $\widehat{\psi}_{\mathrm{pa},j}$).

The focused regularised likelihood estimator (FRLE) is defined as

$$\widehat{\theta}_\lambda = \arg\max_\theta \Big\{ \overbrace{\ell_n(\theta) - \tfrac{1}{2}\lambda n \sum_{j=1}^{r} w_j \{\widehat{\psi}_{j,\mathrm{np}} - \psi_j(\theta)\}^2}^{\text{log-FRL}} \Big\},$$

where $\psi_j(\theta)$ are regularisation/control parameters (under the model) and

- λ is a tuning parameter (where $\lambda = 0$ will reproduce the MLE)
- n makes sure that the regularisation will not be washed out
- w_j are weights
- $\widehat{\psi}_{j,\mathrm{np}}$ are alternative nonparametric estimators for the same ψ_j

Illustration: Focused estimation of survival

How long is a life? Here we consider a classical data set of life-lengths in Roman Egypt, collected by W. Spiegelberg in 1901 and analysed by Karl Pearson (1902).

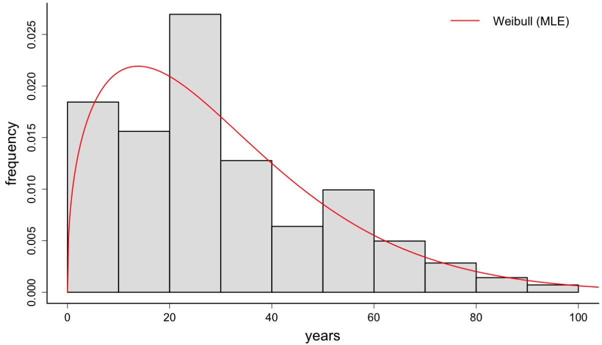

Figure: The age at death for 141 Egyptian mummies (82 men and 59 women) in the Roman period (around year 100 B.C.); see e.g. Claeskens & Hjort (2008) for details.

Assuming a Weibull is a reasonable model, suppose we are particularly interested in the estimates for

$$\psi_1 = \Pr\{0 \le Y < 15\}, \quad \psi_2 = \Pr\{15 \le Y < 30\} \quad \text{and} \quad \psi_3 = \Pr\{30 \le Y < 100\}.$$

Illustration: Focused estimation of survival

The FRLE is given by

$$\widehat{\theta}_\lambda = \arg\max_\theta \{\ell_n(\theta) - \tfrac{1}{2}\lambda n \sum_{j=1}^{3} w_i \{\widehat{\psi}_{\mathrm{np},j} - \psi_j(\theta)\}^2\}$$

where $w_j = 1/3$ and $\widehat{\psi}_{\mathrm{np},j}$ are the proportions with the corresponding life-lengths.

From data and the Weibull model (with $\lambda = 0$) we find

$$\widehat{\psi}_{\mathrm{np},1} = 0.22, \quad \widehat{\psi}_{\mathrm{np},2} = 0.37 \quad \text{and} \quad \widehat{\psi}_{\mathrm{np},3} = 0.41, \ \text{and}$$
$$\widehat{\psi}_{\mathrm{pa},1} = 0.28, \quad \widehat{\psi}_{\mathrm{pa},2} = 0.30 \quad \text{and} \quad \widehat{\psi}_{\mathrm{pa},3} = 0.43,$$

and with the FRLE we obtain (goodness-of-fit test)

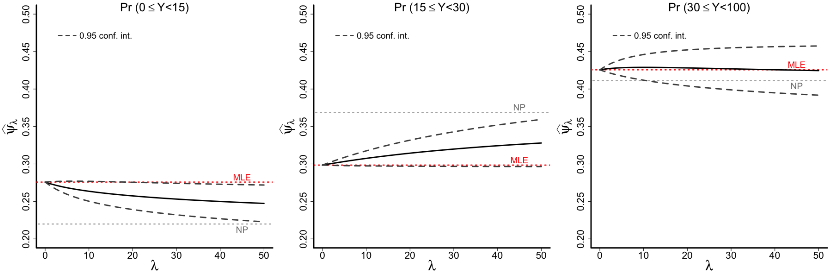

Figure: The effect of increasing λ on the control parameters.

The basic theory for i.i.d. data (part 1)

If we only consider one ψ and let $Y_1, Y_2, \dots Y_n$ be i.i.d. with distribution G (and density g), then

$$\widehat{\theta}_\lambda = \arg\max_\theta \Big\{ \ell_n(\theta) - \tfrac{1}{2}\lambda n\{\widehat{\psi}_{\mathrm{np}} - \psi(\theta)\}^2 \Big\},$$

in order to 'understand' the FRLE we will:

(1) find what $\widehat{\theta}_\lambda$ aims at and

(2) derive the limit behaviour of $\sqrt{n}(\widehat{\theta}_\lambda - \theta_\lambda)$.

The limit (1) is obtained by similar arguments as the MLE outside the model and

$$\widehat{\theta}_\lambda \to_{\mathrm{pr}} \theta_\lambda = \arg\min_\theta \{ \mathrm{KL}(g, f_\theta) + \tfrac{1}{2}\lambda\{\psi_{\mathrm{true}} - \psi(\theta)\}^2 \},$$

In order to derive the limt distribution in (2) we 'only' need to work with the scaled first and second derivative of log-FRL, i.e.

$$\text{(i) } \sqrt{n}U_{n,\lambda}(\theta) = \frac{\partial}{\partial\theta}\ell_{n,\lambda}(\theta)/\sqrt{n} \quad \text{and} \quad \text{(ii) } J_{n,\lambda}(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{t}}}\ell_{n,\lambda}(\theta)/n.$$

The weak limt (i) only depends on the joint limit of the original (scaled) score function and the non-parametric estimator.

If

$$\begin{pmatrix} \sqrt{n}[U_n(\theta_\lambda) + \lambda\{\psi_{\rm true} - \psi(\theta_\lambda)\}\psi^*(\theta_\lambda)] \\ \sqrt{n}\{\widehat\psi_{\rm np} - \psi_{\rm true}\} \end{pmatrix} \to_d \begin{pmatrix} U_\lambda \\ V \end{pmatrix} \sim \mathrm{N}_{p+1}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(\theta_\lambda) & c^{\rm t} \\ c & \tau^2 \end{pmatrix}\right),$$

with $\mathrm{Var}_g(U_n(\theta_\lambda)) \to K(\theta_\lambda)$, then

$$\sqrt{n}U_{n,\lambda}(\theta_\lambda) \to_d U_\lambda = U(\theta_\lambda) + \lambda V\psi^*(\theta_\lambda),$$

which is zero-mean normal with covariance matrix

$$K_\lambda = K(\theta_\lambda) + \lambda^2\tau^2\psi^*(\theta_\lambda)\psi^*(\theta_\lambda)^{\rm t} + 2\lambda c\psi^*(\theta_\lambda)^{\rm t}.$$

In order to establish the limit of (ii) we need

$$\begin{aligned} J_{n,\lambda}(\theta_\lambda) &= -n^{-1}\ell_n^{**}(\theta_\lambda) + \lambda[\psi^*(\theta_\lambda)\psi^*(\theta_\lambda)^{\rm t} + \{\widehat\psi_{\rm np} - \psi(\theta_\lambda)\}\psi^{**}(\theta_\lambda)] \\ &\to_{\rm pr} J(\theta_\lambda) + \lambda[\psi^*(\theta_\lambda)\psi^*(\theta_\lambda)^{\rm t} + \{\psi_{\rm true} - \psi(\theta_\lambda)\}\psi^{**}(\theta_\lambda)] \\ &= J(\theta_\lambda) + \lambda L = J_\lambda, \end{aligned}$$

where $J(\cdot)$ is the classical Fisher information matrix.

Now, by combining (i) and (ii) we obtain the weak limit (2) as

$$\sqrt{n}(\widehat\theta_\lambda - \theta_\lambda) \to_d J_\lambda^{-1}U_\lambda = (J(\theta_\lambda) + \lambda L)^{-1}\{U(\theta_\lambda) + \lambda V\psi^*(\theta_\lambda)\},$$

Focused Regularised Regression

Suppose

$$Y_i = \beta^{\mathrm{t}} x_i + \gamma^{\mathrm{t}} z_i + \sigma \epsilon_i$$

with covariates x_i and z_i, and where ϵ_i are i.i.d. standard normal errors and $\sigma > 0$.

Let the wide model be the model with mean $\beta^{\mathrm{t}} x + \gamma^{\mathrm{t}} z$ and the narrow model be $\beta^{\mathrm{t}} x + \gamma_0^{\mathrm{t}} z = \beta^{\mathrm{t}} x$ with $\gamma_0 = 0$.

Suppose we care about $\psi(x^*, z^*) = \mathrm{E}[Y^* \mid x^*, z^*]$ for a set of r important (x_j^*, z_j^*).

If the wide model play the part as the nonparametric component, then

$$\widehat{\beta}_\lambda = \arg\max_\beta \Big\{ \ell_n(\beta, \widehat{\sigma}(\beta)) - \tfrac{1}{2} \lambda n \sum_{j=1}^r w_j (\widehat{\beta}_{\mathrm{wide}} x_j^* + \widehat{\gamma}_{\mathrm{wide}} z_j^* - \beta x_j^*)^2 \Big\}$$

with $\widehat{\beta}_{\mathrm{wide}}$ and $\widehat{\gamma}_{\mathrm{wide}}$ fitted under the wide model.

This also motivates a Focused Regularised Least Squares Regression by

$$\widehat{\beta}_\lambda = \arg\min_\beta \{ (Y - X\beta)^{\mathrm{t}} (Y - X\beta) + \lambda n (\widehat{Y}^*_{\mathrm{wide}} - X^* \beta)^{\mathrm{t}} (\widehat{Y}^*_{\mathrm{wide}} - X^* \beta)/r \},$$

resulting in a explicit formulas and properties for $\widehat{\beta}_\lambda$ (not based on asymptotics).

Simulated data from $Y_i = \beta_0 + \beta_1 x_i + \gamma_1 z_i + \epsilon_i$, with $z_i = (0.5 - x_i)^2$ for $n = 50$, $\beta_0 = 0.5$, $\beta_1 = 2.0$, $\gamma_1 = 2.5$ and $\sigma = 1.2$.

Let $\psi = \mathrm{E}[Y^* \mid x^* = 0.1]$.

For optimal λ we compare root mean squared error (rmse) for $\mathrm{E}[Y^* \mid x^*]$ all x^*.

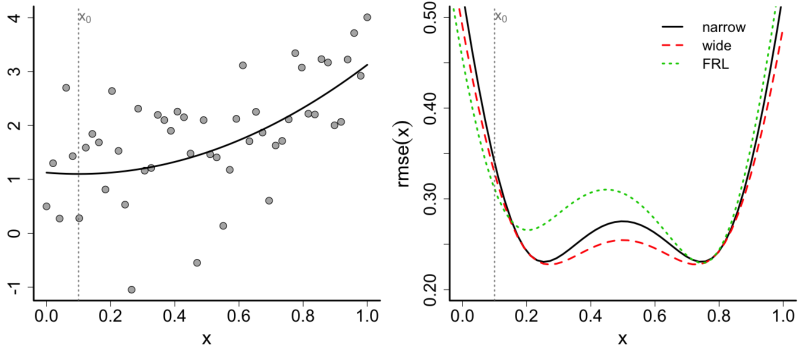

Figure: The difference in rmse for one $x_0 = 0.1$.

Simulated data from $Y_i = \beta_0 + \beta_1 x_i + \gamma_1 z_i + \epsilon_i$, with $z_i = (0.5 - x_i)^2$ for $n = 50$, $\beta_0 = 0.5$, $\beta_1 = 2.0$, $\gamma_1 = 2.5$ and $\sigma = 1.2$.

For each x let $\psi = \mathrm{E}[Y^* \mid x^* = x]$.

For optimal λ we compare root mean squared error (rmse) for $\mathrm{E}[Y^* \mid x^*]$ all x^*.

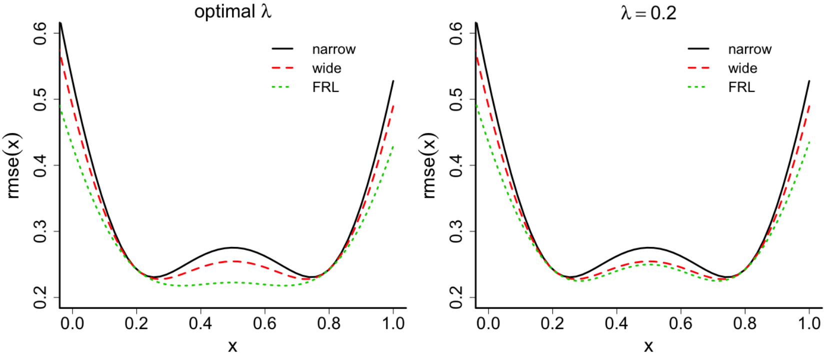

Figure: The difference in rmse for optimal and fixed λ for each possible x.

Concluding remarks

The FRLE push the MLE to match the nonparametric procedure with respect to the selected ψ_j.

Easy to implement and not too sensitive to fine tuning of λ.

There is a connection to a empirical Bayesian procedure.

We have also general and explicit formulas and asymptotic theory for a family of stationary Gaussian time series models.

Clear results under a so-called locally misspecified modelling framework, i.e. where

$$f_{\text{true}}(y) = f(y, \theta, \gamma_0 + \delta/\sqrt{n}),$$

with corresponding methodology for finding a good tuning parameter λ.

The FRL procedure may also be seen as:

- focused robust estimation, where we use e.g. empirical quantiles to correct for potential model misspecifications (borrowing strengths)
- robust focused inference, if $\mu(g)$ is especially important we may use $r+1$ control parameters $\psi_0 = \mu(g)$ and $\psi_1, \ldots, \psi_r$ for $r \geq 0$
- model selection, checking or testing via e.g. asymptotic confidence intervals for $\sqrt{n}(\widehat{\theta}_\lambda - \widehat{\theta}_{\text{ML}})$, for a given λ, under the parametric model
- robust double focused inference, applying the above strategy within the traditional FIC framework (i.e. focused model selection)