Focused Regularised Likelihood

Gudmund Horn Hermansen with Nils Lid Hjort

Godt Hjort

Merker meg at nesten alle estimerer de øyene
med mange ulike arter godt, mens den med 4
dårlig.

(fjerner man denne passer alt veldig fint, noe som
kan være med på å gi modellen troverdighet)

OK diskusjon.

Mye bra her, Gudmund. Ta hovedfag.
Nils

Suppose we have data from

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \sigma\epsilon_i = x_i^{\mathrm{t}}\beta + \sigma\epsilon_i,$$

with $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. standard normals, $\beta \in \mathbb{R}^{p+1}$ and $\sigma > 0$.

Suppose we have data from

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \sigma\epsilon_i = x_i^{\mathrm{t}}\beta + \sigma\epsilon_i,$$

with $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. standard normals, $\beta \in \mathbb{R}^{p+1}$ and $\sigma > 0$.

Ridge regression is a common regularised method for estimating β:

$$\widehat{\beta}_{\mathrm{ridge}} = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - x_i^{\mathrm{t}}\beta)^2 + \lambda \sum_{j=1}^{p+1} |\beta_j|^2 \right\}.$$

Note that:

- $\lambda = 0$ is the same as ordinary least squares regression
- increasing λ will 'shrink' the $\widehat{\beta}_j$-s toward zero

Introduction, Summary and Notation

Suppose we have data from

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \sigma\epsilon_i = x_i^{\mathrm{t}}\beta + \sigma\epsilon_i,$$

with $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. standard normals, $\beta \in \mathbb{R}^{p+1}$ and $\sigma > 0$.

Ridge regression is a common regularised method for estimating β:

$$\widehat{\beta}_{\mathrm{ridge}} = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - x_i^{\mathrm{t}}\beta)^2 + \lambda \sum_{j=1}^{p+1} |\beta_j|^2 \right\}.$$

Note that:

- $\lambda = 0$ is the same as ordinary least squares regression
- increasing λ will 'shrink' the $\widehat{\beta}_j$-s toward zero

The above is (essentially) the same as

$$\widehat{\theta}_{\lambda} = (\widehat{\beta}_{\lambda}, \widehat{\sigma}_{\lambda}) = \arg\max_{\beta,\sigma} \left\{ \ell_n(\beta,\sigma) + \lambda n \sum_{j=1}^{p+1} |\beta_j|^2 \right\},$$

where $\ell_n(\beta)$ is the log-likelihood corresponding to a Gaussian distribution.

The penalisation term is now scaled by n.

Introduction, Summary and Notation

We denote the 'true' data-generating distribution by G.

In general $Y_1, Y_2, \ldots, Y_n$ are i.i.d. from G.

Let F_θ be a parametric model, and $\ell_n(\theta)$ the corresponding log-likelihood.

We denote the 'true' data-generating distribution by G.

In general $Y_1, Y_2, \ldots, Y_n$ are i.i.d. from G.

Let F_θ be a parametric model, and $\ell_n(\theta)$ the corresponding log-likelihood.

Then the FRL estimator is

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

where λ is a tuning parameter and ψ a control parameter.

Introduction, Summary and Notation

We denote the 'true' data-generating distribution by G.

In general $Y_1, Y_2, \ldots, Y_n$ are i.i.d. from G.

Let F_θ be a parametric model, and $\ell_n(\theta)$ the corresponding log-likelihood.

Then the FRL estimator is

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

where λ is a tuning parameter and ψ a control parameter.

Effective control parameters are important characteristics of a distribution where we also have robust alternative estimators (non-parametric).

Example: A quantile with $0 \le p \le 1$:

$$\psi(\theta) = \psi(F_\theta, p) = F_\theta^{-1}(p) \quad \text{and} \quad \widehat{\psi} = \widehat{\psi}(p) = \widehat{G}_n^{-1}(p),$$

where $\widehat{G}_n$ is the empirical CDF.

We denote the 'true' data-generating distribution by G.

In general $Y_1, Y_2, \dots, Y_n$ are i.i.d. from G.

Let F_θ be a parametric model, and $\ell_n(\theta)$ the corresponding log-likelihood.

Then the FRL estimator is

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

where λ is a tuning parameter and ψ a control parameter.

Effective control parameters are important characteristics of a distribution where we also have robust alternative estimators (non-parametric).

Example: k-th moment:

$$\psi(\theta) = \psi(F_\theta, k) = \int y^k \, \mathrm{d}F_\theta(y) \quad \text{and} \quad \widehat{\psi} = \widehat{\psi}(k) = \frac{1}{n}\sum_{i=1}^n y_i^k.$$

We denote the 'true' data-generating distribution by G.

In general $Y_1, Y_2, \ldots, Y_n$ are i.i.d. from G.

Let F_θ be a parametric model, and $\ell_n(\theta)$ the corresponding log-likelihood.

Then the FRL estimator is

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

where λ is a tuning parameter and ψ a control parameter.

Effective control parameters are important characteristics of a distribution where we also have robust alternative estimators (non-parametric).

Example: A probability, e.g.

$$\psi(\theta) = \psi(F_\theta, q) = \int I(y > q)\,\mathrm{d}F_\theta(y) \quad \text{and} \quad \widehat{\psi}(q) = \frac{1}{n}\sum_{i=1}^{n} I(y_i > q).$$

In general, the log Focused Regularised Likelihood (log-FRL) is

$$\ell_{n,\lambda}(\theta) = \ell_{n,\lambda,\boldsymbol{w},\boldsymbol{\psi}}(\theta) = \ell_n(\theta) - \tfrac{1}{2}\lambda n \sum_{j=1}^{r} w_j \{\widehat{\psi}_j - \psi_j(\theta)\}^2,$$

where

- $\ell_n(\theta)$ is the log-likelihood corresponding to F_θ
- λ is a tuning parameter
- ψ_j are control or focus parameter, e.g. quantiles, moments, ...
- $\widehat{\psi}_j$ are non-parametric or robust alternative estimates for ψ_j
- $w_1, \ldots, w_r$ are weights with $w_1 + \cdots + w_r = 1$

Introduction, Summary and Notation

In general, the log Focused Regularised Likelihood (log-FRL) is

$$\ell_{n,\lambda}(\theta) = \ell_{n,\lambda,\boldsymbol{w},\boldsymbol{\psi}}(\theta) = \ell_n(\theta) - \tfrac{1}{2}\lambda n \sum_{j=1}^{r} w_j \{\widehat{\psi}_j - \psi_j(\theta)\}^2,$$

where

- $\ell_n(\theta)$ is the log-likelihood corresponding to F_θ
- λ is a tuning parameter
- ψ_j are control or focus parameter, e.g. quantiles, moments, ...
- $\widehat{\psi}_j$ are non-parametric or robust alternative estimates for ψ_j
- $w_1, \ldots, w_r$ are weights with $w_1 + \cdots + w_r = 1$

Note that:

- if $\lambda = 0$ we have $\widehat{\theta}_\lambda = \widehat{\theta}_{\mathrm{ML}}$
- increasing λ will 'push' $\psi_j(\widehat{\theta}_\lambda)$ to match $\widehat{\psi}_j$

We also need a set of 'standard' regularity assumptions to be true.

Why?

Control parameters can make the estimated model more robust.

Control as a focus parameter can improve the model where it is important.

Why?

Control parameters can make the estimated model more robust.

Control as a focus parameter can improve the model where it is important.

Analytic large sample theory for:

- standard models for i.i.d. data
- models with local misspecification
- regression models
- stationary time series (will not talk about this here)

Illustration: Measuring the Speed of Light

We can use FRL for robust estimation of a normal density.

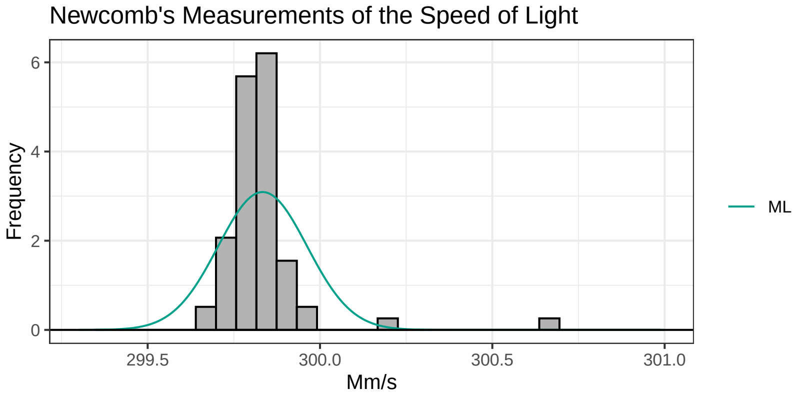
In particular if data contains outliers or is contaminated.

Illustration: Measuring the Speed of Light

We can use FRL for robust estimation of a normal density.

In particular if data contains outliers or is contaminated.

Suppose we want to model the data below with a $\mathrm{N}(\mu, \sigma^2)$.



Simon Newcomb speed of light measurements; see e.g. Stigler (1977) for details about the data.

Illustration: Measuring the Speed of Light

We will do this by adding some quantiles as control parameters:

$$\psi(\mu, \sigma, p) = \sigma \times \Phi^{-1}(p) + \mu \quad \text{and} \quad \hat{\psi} = \hat{\psi}(p) = \widehat{G}^{-1}(p)$$

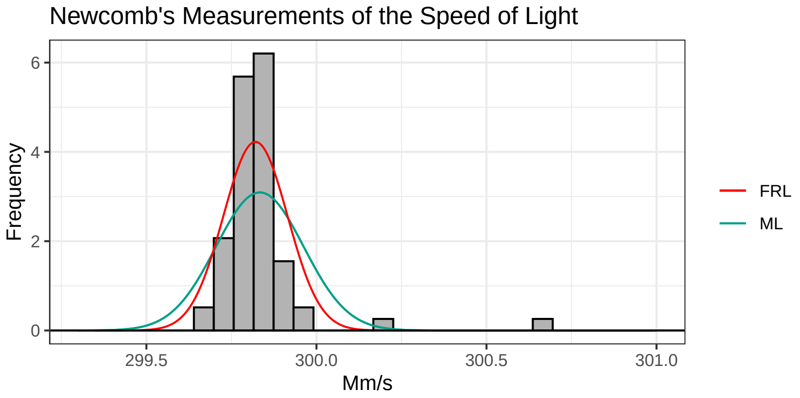
where $\widehat{G}$ is the empirical CDF, with $p = 0.1$, 0.5, $p = 0.90$ and $\lambda = 1000$.

Illustration: Measuring the Speed of Light

We will do this by adding some quantiles as control parameters:

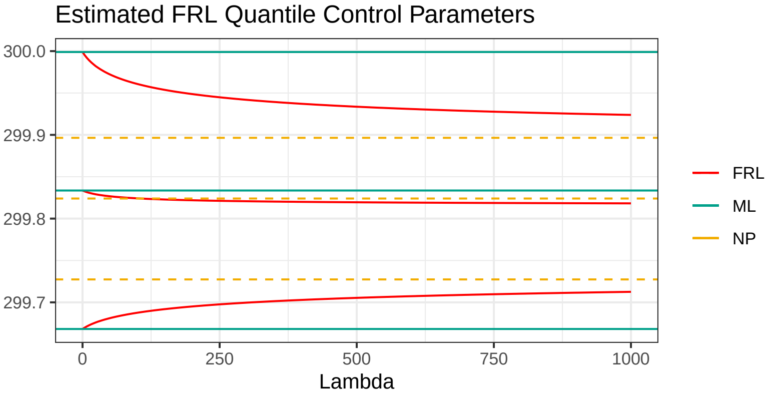
$$\psi(\mu, \sigma, p) = \sigma \times \Phi^{-1}(p) + \mu \quad \text{and} \quad \hat{\psi} = \hat{\psi}(p) = \widehat{G}^{-1}(p)$$

where $\widehat{G}$ is the empirical CDF, with $p = 0.1$, 0.5, $p = 0.90$ and $\lambda = 1000$.

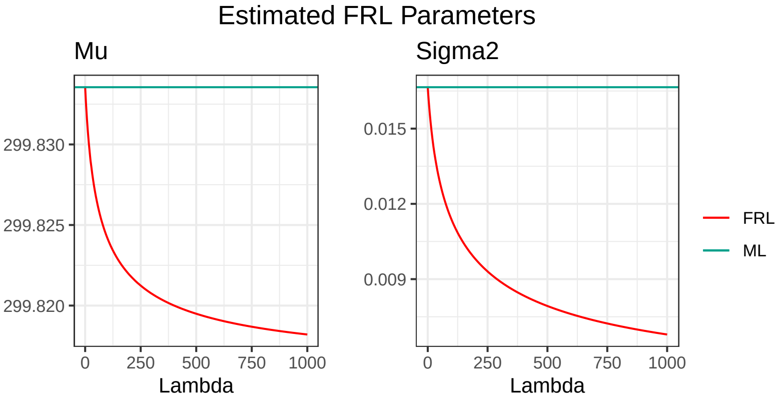


Simon Newcomb speed of light measurements; see e.g. Stigler (1977) for details about the data.

Increasing λ 'push' the estimated quantiles towards the empirical quantiles.



And, the estimated parametres move away from the ML estimates.



However, are these FRL estimates more precise?

Illustration: Simulated Data with Contamination

Simulated data with contamination (outliers).

Repeated simulations of independent $Y_1, \ldots, Y_{100}$ with $Y_i \sim \mathrm{N}(0, 1)$.

Add 4% contamination from a $\mathrm{N}(4, 0.5)$.

Again, we will use control parameters based on quantiles (0.1, 0.5 and 0.9).

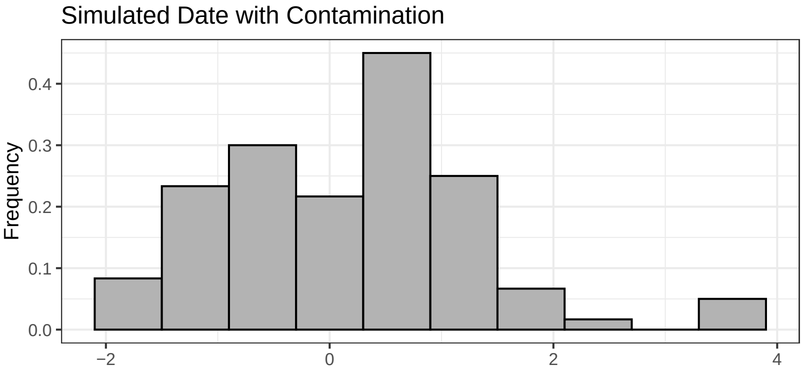


Illustration: Simulated Data with Contamination

Estimated quantiles are ‘pushed’ towards the empirical (and true) quantiles.

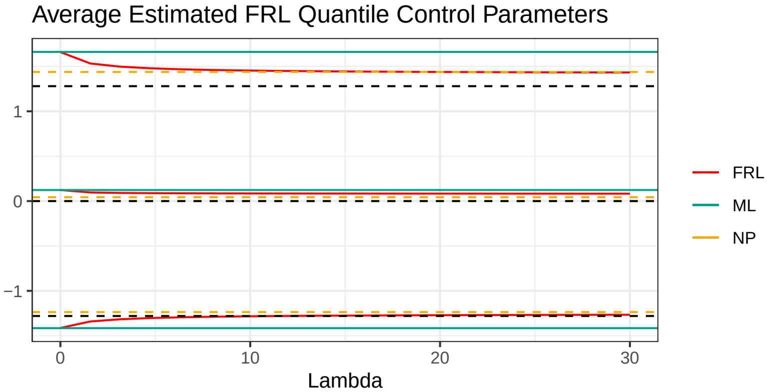
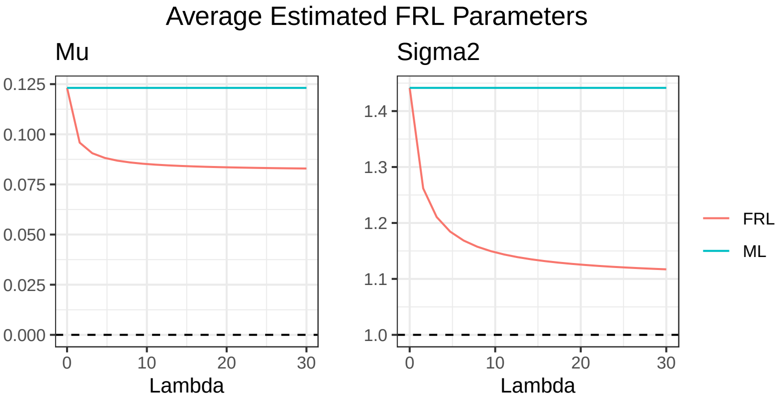
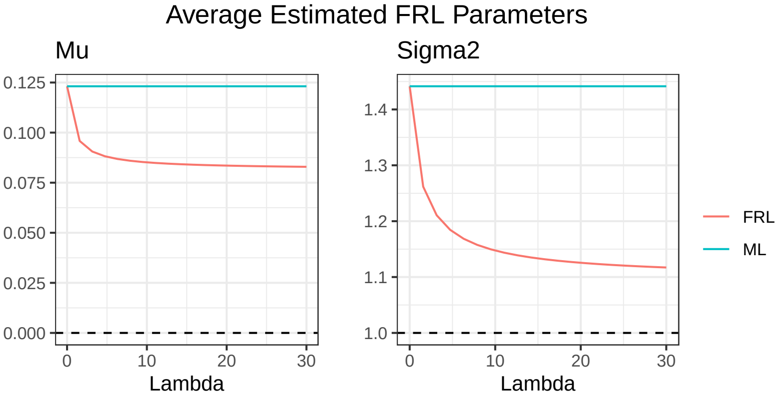


Illustration: Simulated Data with Contamination

And the estimated parameters move closer to the true values.



And the estimated parameters move closer to the true values.



Here, the median was used as a control parameter.

However, should we just use the non-parametric estimate(s)?

Illustration: Estimation of Location Parameter with Contamination

Same simulation setup, but with 'focus' on estimating a location parameter.

Illustration: Estimation of Location Parameter with Contamination

Same simulation setup, but with 'focus' on estimating a location parameter.

We frame this as estimating μ in a $\mathrm{N}(\mu, 1)$, two natural estimators are

$$\widehat{\mu}_{\mathrm{ML}} = \bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad \widehat{m} = \mathrm{median}(Y_1, \ldots, Y_n).$$

We consider $\mu = 0$ to be the 'true' target value.

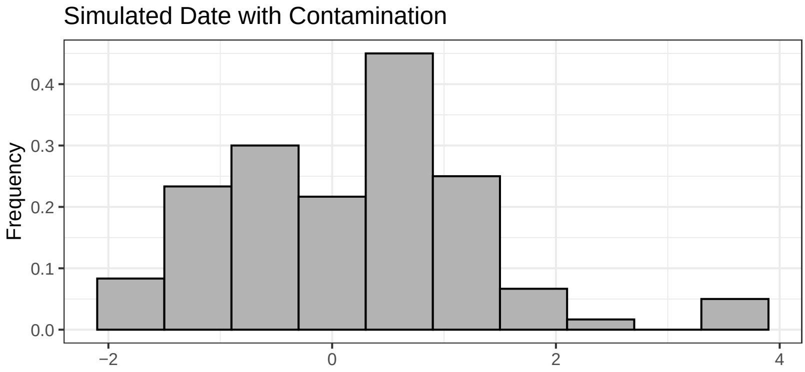


Illustration: Estimation of Location Parameter with Contamination

The FRL, with the median as control $\psi = \mu$ and $\widehat{\psi} = \widehat{m}$ is

$$\widehat{\mu}_\lambda = \arg\max_{\mu} \left\{ \ell_n(\mu) - \tfrac{1}{2}\lambda n\{\widehat{m} - \mu\}^2 \right\}$$

Illustration: Estimation of Location Parameter with Contamination

The FRL, with the median as control $\psi = \mu$ and $\widehat{\psi} = \widehat{m}$ is

$$\widehat{\mu}_\lambda = \arg\max_{\mu} \left\{ \ell_n(\mu) - \tfrac{1}{2}\lambda n\{\widehat{m} - \mu\}^2 \right\}$$

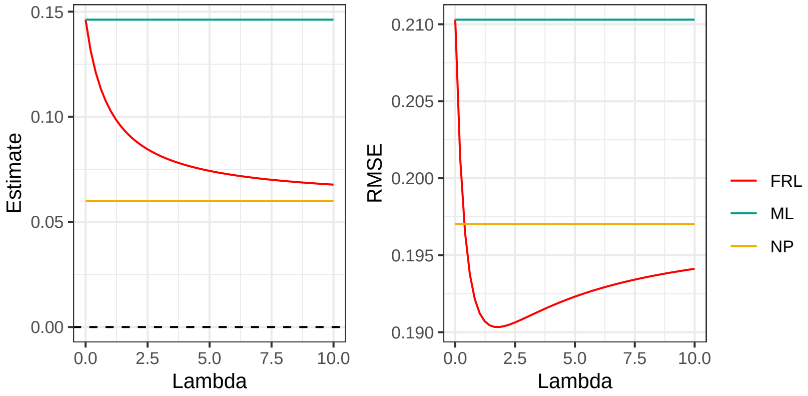
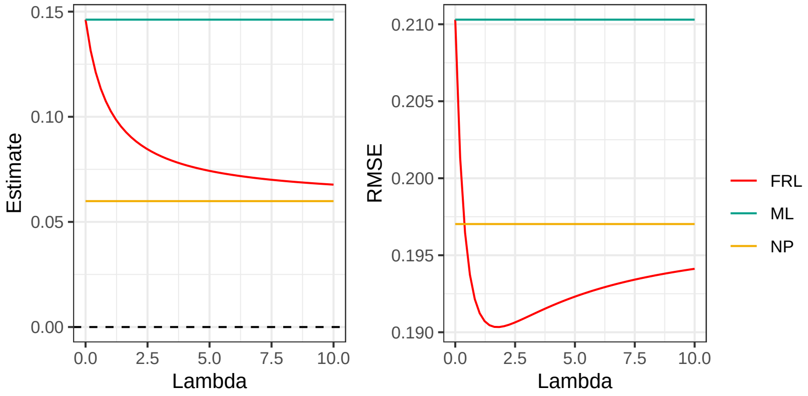


Illustration: Estimation of Location Parameter with Contamination

The FRL, with the median as control $\psi = \mu$ and $\widehat{\psi} = \widehat{m}$ is

$$\widehat{\mu}_\lambda = \arg\max_{\mu} \left\{ \ell_n(\mu) - \tfrac{1}{2}\lambda n \{\widehat{m} - \mu\}^2 \right\}$$



Note that this is a bias–variance trade–off game (with respect to RMSE).

How to determine the optimal λ?

Large Sample Theory - Illustration

We can use large-sample theory or the bootstrap to analyse the 'behaviour' of the FRL estimator, find optimal λ, compare it to the MLE, ...

A bootstrap approach seems to work well, will not focus on this here.

Large Sample Theory - Illustration

We can use large-sample theory or the bootstrap to analyse the 'behaviour' of the FRL estimator, find optimal λ, compare it to the MLE, ...

A bootstrap approach seems to work well, will not focus on this here.

We need the target of $\widehat{\theta}_\lambda$, say θ_λ, and the limit distribution of $\sqrt{n}(\widehat{\theta}_\lambda - \theta_\lambda)$.

Large Sample Theory - Illustration

We can use large-sample theory or the bootstrap to analyse the 'behaviour' of the FRL estimator, find optimal λ, compare it to the MLE, ...

A bootstrap approach seems to work well, will not focus on this here.

We need the target of $\widehat{\theta}_\lambda$, say θ_λ, and the limit distribution of $\sqrt{n}(\widehat{\theta}_\lambda - \theta_\lambda)$.

Again, consider estimating a location parameter, with competing estimators

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad \widehat{m} = \text{median}(Y_1, \ldots, Y_n),$$

with Y_i are i.i.d. and $Y_i \sim G$.

In order to fit the FRL framework, we view this as estimating μ in a $\mathrm{N}(\mu, 1)$.

Large Sample Theory - Illustration

We can use large-sample theory or the bootstrap to analyse the 'behaviour' of the FRL estimator, find optimal λ, compare it to the MLE, ...

A bootstrap approach seems to work well, will not focus on this here.

We need the target of $\widehat{\theta}_\lambda$, say θ_λ, and the limit distribution of $\sqrt{n}(\widehat{\theta}_\lambda - \theta_\lambda)$.

Again, consider estimating a location parameter, with competing estimators

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad \widehat{m} = \text{median}(Y_1, \ldots, Y_n),$$

with Y_i are i.i.d. and $Y_i \sim G$.

In order to fit the FRL framework, we view this as estimating μ in a $\mathrm{N}(\mu, 1)$.

With the median as the control parameter, i.e. $\psi(\mu) = \mu$ and $\widehat{\psi} = \widehat{m}$ as the robust alternative, then

$$\widehat{\mu}_\lambda = \arg\max_{\mu} \left\{\ell_n(\mu) - \tfrac{1}{2}\lambda n\{\widehat{m} - \mu\}^2\right\}$$

Large Sample Theory - Illustration

We can use large-sample theory or the bootstrap to analyse the 'behaviour' of the FRL estimator, find optimal λ, compare it to the MLE, ...

A bootstrap approach seems to work well, will not focus on this here.

We need the target of $\widehat{\theta}_\lambda$, say θ_λ, and the limit distribution of $\sqrt{n}(\widehat{\theta}_\lambda - \theta_\lambda)$.

Again, consider estimating a location parameter, with competing estimators

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad \widehat{m} = \text{median}(Y_1, \ldots, Y_n),$$

with Y_i are i.i.d. and $Y_i \sim G$.

In order to fit the FRL framework, we view this as estimating μ in a $\mathrm{N}(\mu, 1)$.

With the median as the control parameter, i.e. $\psi(\mu) = \mu$ and $\widehat{\psi} = \widehat{m}$ as the robust alternative, then

$$\widehat{\mu}_\lambda = \arg\max_{\mu} \left\{ \ell_n(\mu) - \tfrac{1}{2}\lambda n \{\widehat{m} - \mu\}^2 \right\}$$

and

$$\widehat{\mu}_\lambda = \frac{1}{1+\lambda}\bar{Y}_n + \frac{\lambda}{1+\lambda}\widehat{m}.$$

Large Sample Theory - Illustration

We know that $\bar{Y}_n \to_{\text{pr}} \mathrm{E}_g\, Y_1 = \mu_g$, and $\widehat{m} \to_{\text{pr}} m_g = G^{-1}(0.5)$.

Large Sample Theory - Illustration

We know that $\bar{Y}_n \to_{\text{pr}} \mathrm{E}_g\, Y_1 = \mu_g$, and $\widehat{m} \to_{\text{pr}} m_g = G^{-1}(0.5)$.

And, we can show that

$$\widehat{\mu} \to_{\text{pr}} \mu_\lambda = \arg\min_{\mu} \left\{ \mathrm{KL}(g, f_\mu) + \tfrac{1}{2}\lambda\{m - \mu\}^2 \right\} = \frac{1}{1+\lambda}\mu_g + \frac{\lambda}{1+\lambda} m_g,$$

where f_μ is the density of a $\mathrm{N}(\mu, 1)$.

Large Sample Theory - Illustration

We know that $\bar{Y}_n \to_{\rm pr} \mathrm{E}_g\, Y_1 = \mu_g$, and $\widehat{m} \to_{\rm pr} m_g = G^{-1}(0.5)$.

And, we can show that

$$\widehat{\mu} \to_{\rm pr} \mu_\lambda = \arg\min_{\mu} \left\{ \mathrm{KL}(g, f_\mu) + \tfrac{1}{2}\lambda\{m - \mu\}^2 \right\} = \frac{1}{1+\lambda}\mu_g + \frac{\lambda}{1+\lambda} m_g,$$

where f_μ is the density of a $\mathrm{N}(\mu, 1)$.

Moreover,

$$\sqrt{n}(\widehat{\mu} - \mu_\lambda) \to_d \Lambda \sim \mathrm{N}\left(0, \frac{1}{(1+\lambda)^2}\left[\sigma_g^2 + \frac{\lambda^2}{4g(m_g)^2} + \frac{\lambda \times \mathrm{E}_g\,|Y_1 - m_g|}{g(m_g)}\right]\right),$$

where $\sigma_g^2 = \mathrm{Var}_g(Y_1)$.

Large Sample Theory - Illustration

We know that $\bar{Y}_n \to_{\mathrm{pr}} \mathrm{E}_g\, Y_1 = \mu_g$, and $\widehat{m} \to_{\mathrm{pr}} m_g = G^{-1}(0.5)$.

And, we can show that

$$\widehat{\mu} \to_{\mathrm{pr}} \mu_\lambda = \arg\min_{\mu} \left\{ \mathrm{KL}(g, f_\mu) + \tfrac{1}{2}\lambda\{m-\mu\}^2 \right\} = \frac{1}{1+\lambda}\mu_g + \frac{\lambda}{1+\lambda} m_g,$$

where f_μ is the density of a $\mathrm{N}(\mu, 1)$.

Moreover,

$$\sqrt{n}(\widehat{\mu} - \mu_\lambda) \to_d \Lambda \sim \mathrm{N}\left(0, \frac{1}{(1+\lambda)^2}\left[\sigma_g^2 + \frac{\lambda^2}{4g(m_g)^2} + \frac{\lambda \times \mathrm{E}_g\,|Y_1 - m_g|}{g(m_g)}\right]\right),$$

where $\sigma_g^2 = \mathrm{Var}_g(Y_1)$.

There are analogous limit distribution results for both $\bar{Y}$ and the median $\widehat{m}$.

From this, we can extract (limit) bias, variance, RMSE, etc.

Large Sample Theory - Illustration

We know that $\bar{Y}_n \to_{\rm pr} {\rm E}_g\, Y_1 = \mu_g$, and $\widehat{m} \to_{\rm pr} m_g = G^{-1}(0.5)$.

And, we can show that

$$\widehat{\mu} \to_{\rm pr} \mu_\lambda = \arg\min_\mu \left\{{\rm KL}(g, f_\mu) + \tfrac{1}{2}\lambda\{m-\mu\}^2\right\} = \frac{1}{1+\lambda}\mu_g + \frac{\lambda}{1+\lambda}m_g,$$

where f_μ is the density of a ${\rm N}(\mu, 1)$.

Moreover,

$$\sqrt{n}(\widehat{\mu} - \mu_\lambda) \to_d \Lambda \sim {\rm N}\left(0, \frac{1}{(1+\lambda)^2}\left[\sigma_g^2 + \frac{\lambda^2}{4g(m_g)^2} + \frac{\lambda \times {\rm E}_g\,|Y_1 - m_g|}{g(m_g)}\right]\right),$$

where $\sigma_g^2 = {\rm Var}_g(Y_1)$.

There are analogous limit distribution results for both $\bar{Y}$ and the median $\widehat{m}$.

From this, we can extract (limit) bias, variance, RMSE, etc.

And, for example:

- we can derive an expression for the optimal value of λ
- compare estimators
- make asymptotic test and diagnostics tools/plots

Large Sample Theory - Summary

Consider one ψ and let $Y_1, Y_2, \dots Y_n$ be i.i.d. from G, then

$$\widehat{\theta}_\lambda = \arg\max_\theta \left\{ \ell_n(\theta) - \tfrac{1}{2}\lambda n\{\widehat{\psi} - \psi(\theta)\}^2 \right\}.$$

In order to 'understand' the FRL estimate we need to:

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

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

Large Sample Theory - Summary

Consider one ψ and let $Y_1, Y_2, \ldots Y_n$ be i.i.d. from G, then

$$\widehat{\theta}_\lambda = \arg\max_\theta \left\{ \ell_n(\theta) - \tfrac{1}{2}\lambda n \{\widehat{\psi} - \psi(\theta)\}^2 \right\}.$$

In order to 'understand' the FRL estimate we need to:

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

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

We obtain (1) by similar arguments as the MLE for a misspecified model:

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

Large Sample Theory - Summary

Consider one ψ and let $Y_1, Y_2, \ldots Y_n$ be i.i.d. from G, then

$$\widehat{\theta}_\lambda = \arg\max_\theta \left\{ \ell_n(\theta) - \tfrac{1}{2}\lambda n \{\widehat{\psi} - \psi(\theta)\}^2 \right\}.$$

In order to 'understand' the FRL estimate we need to:

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

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

We obtain (1) by similar arguments as the MLE for a misspecified model:

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

Similar for (2), where we can show

$$\sqrt{n}(\widehat{\theta}_\lambda - \theta_\lambda) \to_d \text{N}_p(0, [J(\theta_\lambda) + \lambda L]^{-1} K_\lambda [J(\theta_\lambda) + \lambda L]^{-1}),$$

Large Sample Theory - Summary

Consider one ψ and let $Y_1, Y_2, \dots Y_n$ be i.i.d. from G, then

$$\widehat{\theta}_\lambda = \arg\max_\theta \left\{ \ell_n(\theta) - \tfrac{1}{2}\lambda n \{\widehat{\psi} - \psi(\theta)\}^2 \right\}.$$

In order to 'understand' the FRL estimate we need to:

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

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

We obtain (1) by similar arguments as the MLE for a misspecified model:

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

Similar for (2), where we can show

$$\sqrt{n}(\widehat{\theta}_\lambda - \theta_\lambda) \to_d \mathrm{N}_p(0, [J(\theta_\lambda) + \lambda L]^{-1} K_\lambda [J(\theta_\lambda) + \lambda L]^{-1}),$$

where J is the Fisher information, $L = \dot{\psi}(\theta_\lambda)\dot{\psi}(\theta_\lambda)^{\mathrm{t}} + [\psi_{\mathrm{true}} - \psi(\theta_\lambda)]\ddot{\psi}(\theta_\lambda)$,

$$K_\lambda = K(\theta_\lambda) + 2\lambda c \dot{\psi}(\theta_\lambda)^{\mathrm{t}} + \lambda^2 \tau^2 \dot{\psi}(\theta_\lambda)\dot{\psi}(\theta_\lambda)^{\mathrm{t}}$$

and $K(\cdot)$, τ^2 and c are elements from the covariance matrix involving $\widehat{\psi}$ and the scaled score-function, $\dot{\psi}$ and $\ddot{\psi}$ are first and second order derivatives.

Models with Local Misspecification - Summary

Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. and $Y_i \sim F_{\theta_0, \gamma_0 + \delta/\sqrt{n}}$ and γ_0 is known.

Models with Local Misspecification - Summary

Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. and $Y_i \sim F_{\theta_0, \gamma_0 + \delta/\sqrt{n}}$ and γ_0 is known.

Models with local misspecification are useful for examining bias–variance trade–offs in a large-sample framework; Claeskens and Hjort (2008).

Models with Local Misspecification - Summary

Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. and $Y_i \sim F_{\theta_0, \gamma_0 + \delta/\sqrt{n}}$ and γ_0 is known.

Models with local misspecification are useful for examining bias–variance trade–offs in a large-sample framework; Claeskens and Hjort (2008).

Assume we are interested in estimating $\psi = \psi(\theta, \gamma)$ (focus parameter).

Models with Local Misspecification - Summary

Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. and $Y_i \sim F_{\theta_0, \gamma_0 + \delta/\sqrt{n}}$ and γ_0 is known.

Models with local misspecification are useful for examining bias–variance trade–offs in a large-sample framework; Claeskens and Hjort (2008).

Assume we are interested in estimating $\psi = \psi(\theta, \gamma)$ (focus parameter).

We can compare $\psi_{\text{narr}} = \psi(\widehat{\theta}_{\text{narr}}, \gamma_0)$ with $\psi_{\text{wide}} = \psi(\widehat{\theta}, \widehat{\gamma})$ in the limit, i.e.

$$\sqrt{n}(\widehat{\psi}_{\text{narr}} - \psi_{\text{true}}) \to_d N(\omega\delta, \tau_0^2)$$
$$\sqrt{n}(\widehat{\psi}_{\text{wide}} - \psi_{\text{true}}) \to_d N(0, \tau_0^2 + \omega^2\kappa^2)$$

with $\omega = J_{10} J_{00}^{-1} \dot{\psi}_\theta - \dot{\psi}_\gamma$ and $\tau_0^2 = \dot{\psi}_\theta^{\text{t}} J_{00}^{-1} \dot{\psi}_\theta$, and $\dot{\psi}_\cdot$ are partial derivatives.

Models with Local Misspecification - Summary

Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. and $Y_i \sim F_{\theta_0, \gamma_0 + \delta/\sqrt{n}}$ and γ_0 is known.

Models with local misspecification are useful for examining bias–variance trade–offs in a large-sample framework; Claeskens and Hjort (2008).

Assume we are interested in estimating $\psi = \psi(\theta, \gamma)$ (focus parameter).

We can compare $\psi_{\text{narr}} = \psi(\widehat{\theta}_{\text{narr}}, \gamma_0)$ with $\psi_{\text{wide}} = \psi(\widehat{\theta}, \widehat{\gamma})$ in the limit, i.e.

$$\sqrt{n}(\widehat{\psi}_{\text{narr}} - \psi_{\text{true}}) \to_d N(\omega\delta, \tau_0^2)$$
$$\sqrt{n}(\widehat{\psi}_{\text{wide}} - \psi_{\text{true}}) \to_d N(0, \tau_0^2 + \omega^2\kappa^2)$$

with $\omega = J_{10} J_{00}^{-1} \dot{\psi}_\theta - \dot{\psi}_\gamma$ and $\tau_0^2 = \dot{\psi}_\theta^{\mathrm{t}} J_{00}^{-1} \dot{\psi}_\theta$, and $\dot{\psi}_\cdot$ are partial derivatives.

If the FRL estimate is

$$\widehat{\theta}_\lambda = \arg\max_\theta \left\{ \ell_n(\theta) - \tfrac{1}{2}\lambda n \{\widehat{\psi}_{\text{wide}} - \psi(\theta, \gamma_0)\}^2 \right\},$$

and $\widehat{\psi}_\lambda = \psi(\widehat{\theta}_\lambda, \gamma_0)$

Models with Local Misspecification - Summary

Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. and $Y_i \sim F_{\theta_0, \gamma_0 + \delta/\sqrt{n}}$ and γ_0 is known.

Models with local misspecification are useful for examining bias–variance trade–offs in a large-sample framework; Claeskens and Hjort (2008).

Assume we are interested in estimating $\psi = \psi(\theta, \gamma)$ (focus parameter).

We can compare $\psi_{\text{narr}} = \psi(\widehat{\theta}_{\text{narr}}, \gamma_0)$ with $\psi_{\text{wide}} = \psi(\widehat{\theta}, \widehat{\gamma})$ in the limit, i.e.

$$\sqrt{n}(\widehat{\psi}_{\text{narr}} - \psi_{\text{true}}) \to_d N(\omega\delta, \tau_0^2)$$
$$\sqrt{n}(\widehat{\psi}_{\text{wide}} - \psi_{\text{true}}) \to_d N(0, \tau_0^2 + \omega^2\kappa^2)$$

with $\omega = J_{10} J_{00}^{-1} \dot{\psi}_\theta - \dot{\psi}_\gamma$ and $\tau_0^2 = \dot{\psi}_\theta^{\text{t}} J_{00}^{-1} \dot{\psi}_\theta$, and $\dot{\psi}_\cdot$ are partial derivatives.

If the FRL estimate is

$$\widehat{\theta}_\lambda = \arg\max_\theta \left\{ \ell_n(\theta) - \tfrac{1}{2} \lambda n \{\widehat{\psi}_{\text{wide}} - \psi(\theta, \gamma_0)\}^2 \right\},$$

and $\widehat{\psi}_\lambda = \psi(\widehat{\theta}_\lambda, \gamma_0)$ we can show that

$$\sqrt{n}(\widehat{\psi}_\lambda - \psi_{\text{true}}) \to_d N(\omega_\lambda \delta, \tau_\lambda^2),$$

with $J_\lambda = J_{00} + \lambda \dot{\psi}_\theta \dot{\psi}_\theta^{\text{t}}$ and

$$\omega_\lambda = (J_{01} + \lambda \dot{\psi}_\gamma \dot{\psi}_\theta) J_\lambda^{-1} \dot{\psi}_\theta - \dot{\psi}_\gamma \quad \text{and} \quad \tau_\lambda = \dot{\psi}_\theta^{\text{t}} [J_\lambda^{-1} + \lambda J_\lambda^{-1} [(I + \lambda \tau_0^2) \dot{\psi}_\theta \dot{\psi}_\theta^{\text{t}}] J_\lambda^{-1}] \dot{\psi}_\theta.$$

Models with Local Misspecification - Exponential or Weibull?

Let $Y_1, \ldots, Y_n$ be i.i.d. Weibull with parameters $\theta_0 = 0.34$ and $\gamma = 1 + \delta/\sqrt{n}$.

Note that $\delta = 0$ is the exponential distribution (narrow).

Comparing the exponential (narrow), Weibull (wide) and FRL at estimating

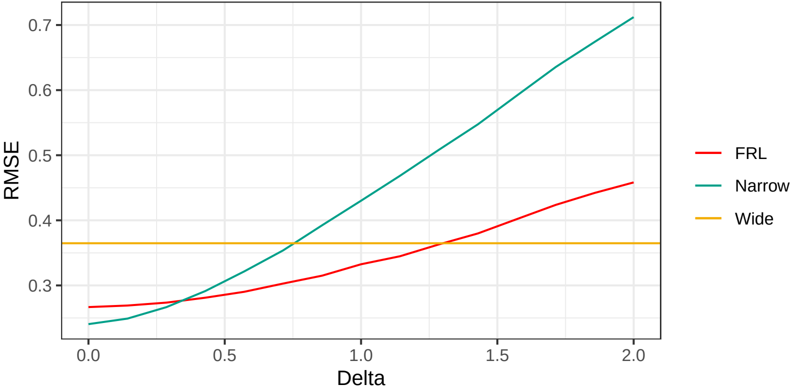
$$\psi(\theta, \gamma) = \Pr\{Y_1 > 1\}.$$

Let $Y_1, \dots, Y_n$ be i.i.d. Weibull with parameters $\theta_0 = 0.34$ and $\gamma = 1 + \delta/\sqrt{n}$.

Note that $\delta = 0$ is the exponential distribution (narrow).

Comparing the exponential (narrow), Weibull (wide) and FRL at estimating

$$\psi(\theta, \gamma) = \Pr\{Y_1 > 1\}.$$



Focused regularised regression

The general idea and framework is easily extended to regression models.

Focused regularised regression

The general idea and framework is easily extended to regression models.

A canonical FRL construction is then

$$\widehat{\theta}_\lambda = (\widehat{\beta}_\lambda, \widehat{\sigma}_\lambda) = \arg\max_{\beta,\sigma} \left\{ \ell_n(\beta,\sigma) - \tfrac{1}{2}\lambda n \frac{1}{|I|} \sum_{z_i \in I} \{\widehat{\mu}(z_i) - z_i^{\mathrm{t}}\beta\}^2 \right\},$$

where $\widehat{\mu}$ is an alternative estimate of the mean and I is a set of important and/or control 'individuals'.

Focused regularised regression

The general idea and framework is easily extended to regression models.

A canonical FRL construction is then

$$\widehat{\theta}_\lambda = (\widehat{\beta}_\lambda, \widehat{\sigma}_\lambda) = \arg\max_{\beta,\sigma} \left\{ \ell_n(\beta,\sigma) - \tfrac{1}{2}\lambda n \frac{1}{|I|} \sum_{z_i \in I} \{\widehat{\mu}(z_i) - z_i^{\mathrm{t}}\beta\}^2 \right\},$$

where $\widehat{\mu}$ is an alternative estimate of the mean and I is a set of important and/or control 'individuals'.

Example: A simple model combined with a sophisticated non-parametric model for $\widehat{\mu}(\cdot)$; e.g. to control for missing interactions.

Focused regularised regression

The general idea and framework is easily extended to regression models.

A canonical FRL construction is then

$$\widehat{\theta}_\lambda = (\widehat{\beta}_\lambda, \widehat{\sigma}_\lambda) = \arg\max_{\beta,\sigma} \left\{ \ell_n(\beta,\sigma) - \tfrac{1}{2}\lambda n \frac{1}{|I|} \sum_{z_i \in I} \{\widehat{\mu}(z_i) - z_i^{\mathrm{t}}\beta\}^2 \right\},$$

where $\widehat{\mu}$ is an alternative estimate of the mean and I is a set of important and/or control 'individuals'.

Example: A simple model combined with a sophisticated non-parametric model for $\widehat{\mu}(\cdot)$; e.g. to control for missing interactions.

Example: To integrate an estimated model with the output from a physical or mechanistic model; weather, hydrology, biology, ...

Focused regularised regression

The general idea and framework is easily extended to regression models.

A canonical FRL construction is then

$$\widehat{\theta}_\lambda = (\widehat{\beta}_\lambda, \widehat{\sigma}_\lambda) = \arg\max_{\beta,\sigma} \left\{ \ell_n(\beta,\sigma) - \tfrac{1}{2}\lambda n \frac{1}{|I|} \sum_{z_i \in I} \{\widehat{\mu}(z_i) - z_i^{\mathrm{t}}\beta\}^2 \right\},$$

where $\widehat{\mu}$ is an alternative estimate of the mean and I is a set of important and/or control 'individuals'.

Example: A simple model combined with a sophisticated non-parametric model for $\widehat{\mu}(\cdot)$; e.g. to control for missing interactions.

Example: To integrate an estimated model with the output from a physical or mechanistic model; weather, hydrology, biology, ...

Example: To integrate local data with external data where we do not have access to raw data, relying on an estimated $\widehat{\mu}(\cdot)$ for integration.

Focused regularised regression

The general idea and framework is easily extended to regression models.

A canonical FRL construction is then

$$\widehat{\theta}_\lambda = (\widehat{\beta}_\lambda, \widehat{\sigma}_\lambda) = \arg\max_{\beta,\sigma} \left\{ \ell_n(\beta,\sigma) - \tfrac{1}{2}\lambda n \frac{1}{|I|} \sum_{z_i \in I} \{\widehat{\mu}(z_i) - z_i^{\mathrm{t}}\beta\}^2 \right\},$$

where $\widehat{\mu}$ is an alternative estimate of the mean and I is a set of important and/or control 'individuals'.

Example: A simple model combined with a sophisticated non-parametric model for $\widehat{\mu}(\cdot)$; e.g. to control for missing interactions.

Example: To integrate an estimated model with the output from a physical or mechanistic model; weather, hydrology, biology, ...

Example: To integrate local data with external data where we do not have access to raw data, relying on an estimated $\widehat{\mu}(\cdot)$ for integration.

Example: To regularise a complex model, by penalising towards a simple model where data is sparse.

Simulated data with

$$Y_i = \mu(x_i) + \epsilon_i$$

for some smooth function $\mu(\cdot)$ and independent $\epsilon_1, \ldots, \epsilon_n$.

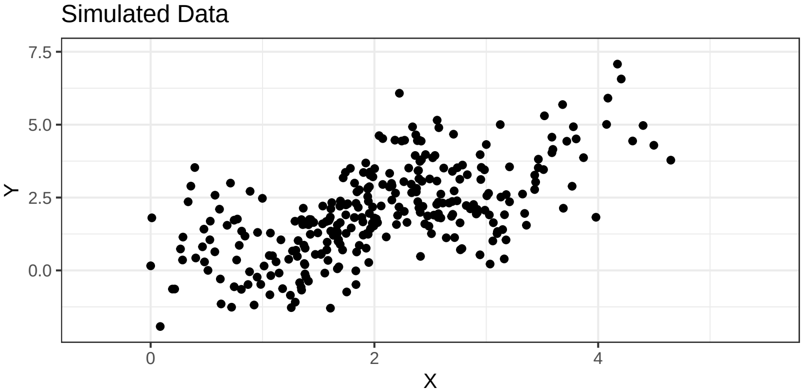
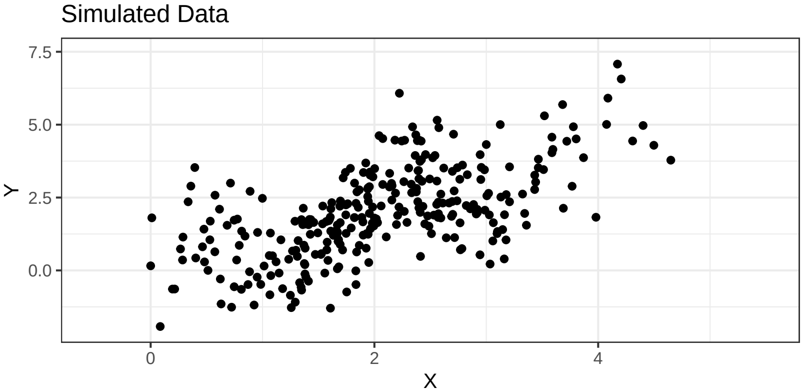


Illustration - Focused regularised regression

Simulated data with

$$Y_i = \mu(x_i) + \epsilon_i$$

for some smooth function $\mu(\cdot)$ and independent $\epsilon_1, \ldots, \epsilon_n$.



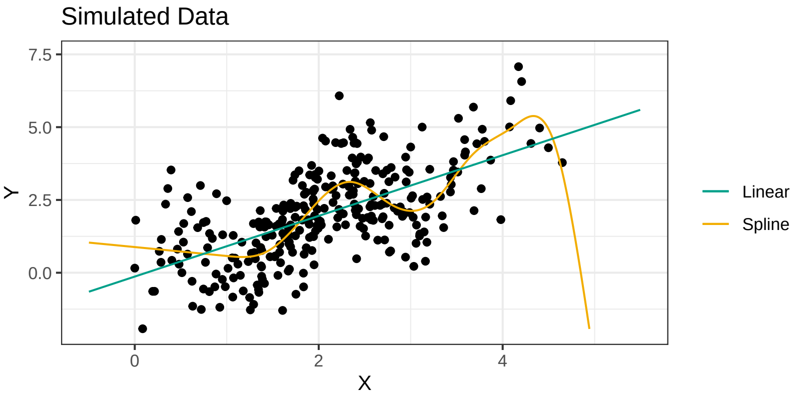
A model that captures both the overall trend and the detailed behaviour?

Useful for extrapolation.

Illustration - Focused regularised regression

A smooth spline effectively capture detailed behaviour where data is dense

Can use a simple linear model to capture the overall trend.



How to combine?

Illustration - Focused regularised regression

Inspired by the FRL setup

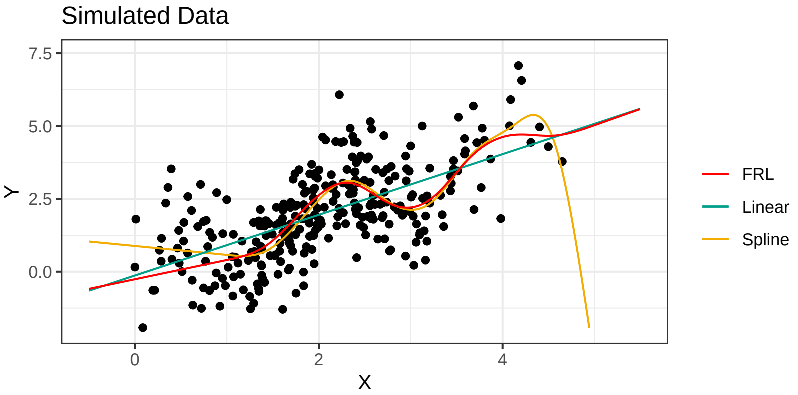
$$\widehat{\theta}_\lambda = (\widehat{\beta}_\lambda, \widehat{\sigma}_\lambda) = \arg\max_{\beta,\sigma} \left\{ \ell_n(s_\beta, \sigma) - \tfrac{1}{2}\lambda n \frac{1}{|I|} \sum_{z_i \in I} \{(\widehat{a} + \widehat{b} z_i) - s_\beta(z_i)\}^2 \right\},$$

where s_β is a smooth spline and I is a set of control points – some below $x = 0$ and some above $x = 4$.

Inspired by the FRL setup

$$\widehat{\theta}_\lambda = (\widehat{\beta}_\lambda, \widehat{\sigma}_\lambda) = \arg\max_{\beta,\sigma} \left\{ \ell_n(s_\beta, \sigma) - \tfrac{1}{2}\lambda n \frac{1}{|I|} \sum_{z_i \in I} \{(\widehat{a} + \widehat{b} z_i) - s_\beta(z_i)\}^2 \right\},$$

where s_β is a smooth spline and I is a set of control points – some below $x = 0$ and some above $x = 4$.



Concluding Remarks

Just do it.

A straightforward method for improving robustness of parametric models.

And can make inference more focused.

Large-sample theory justify the use in simple models.

Works well for regression models.

And stationary time series.

Bootstrapping techniques also works well (in simulated data examples).

Link to empirical likelihood and empirical Bayes.