Sequential learning of regression models through penalized estimation

Wessel N. van Wieringen
(joint work with Harald Binder)

Dept. of Epidemiology & Data Science, Amsterdam UMC
Dept. of Mathematics, Vrije Universiteit Amsterdam
Amsterdam, the Netherlands

IBC2022, Riga, 14.07.2022

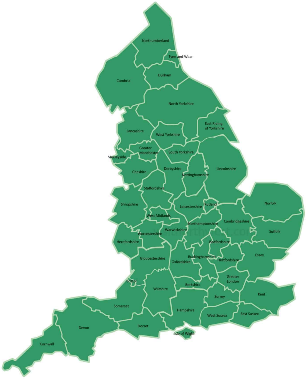
The Fingertips study

Epidemiological study into health.

Some details:

- Max. 153 counties and urban areas of England,
- Nine consecutive years, 2008–2016,
- Response: `suicide rate`,
- Max. 23 covariates, e.g. `alcohol`, `depression`, `homeless`, `unemployment`, `child neglect`, ...

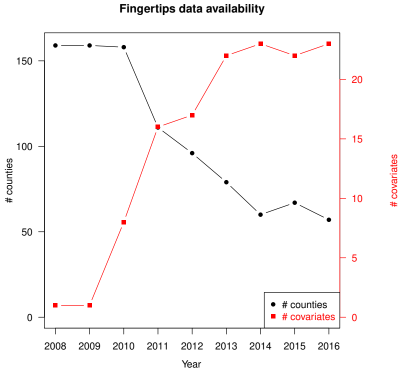
`https://fingertips.phe.org.uk/`



The Fingertips study

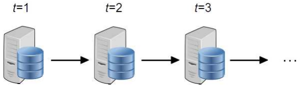
We wish to learn $Y = X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_{nn})$ ...

... but data arrive in batches
... of varying dimension and sample size,



Updated ridge regression

Let $\{\mathbf{Y}_t, \mathbf{X}_t\}_{t=1}^{\infty}$ be a sequence of data sets



of fixed dimension p but varying sample sizes $\{n_t\}_{t=1}^{\infty}$.

Fit $\mathbf{Y}_t = \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t$ with $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}_{n_t}, \sigma^2 \mathbf{I}_{n_t,n_t})$ for $t = 1, 2, \ldots$ by:

$$\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1}) \quad = \quad \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y}_t - \mathbf{X}_t \boldsymbol{\beta}\|_2^2 + \lambda_t \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1}, \hat{\boldsymbol{\beta}}_{t-2})\|_2^2,$$

with $\hat{\boldsymbol{\beta}}_0(\lambda_0) = \boldsymbol{\beta}_0 \in \mathbb{R}^p$.

Then,

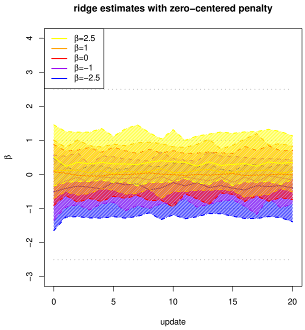
$$\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1}) \quad = \quad (\mathbf{X}_t^\top \mathbf{X}_t + \lambda_t \mathbf{I}_{pp})^{-1} [\mathbf{X}_t^\top \mathbf{Y}_t + \lambda_t \hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1}, \hat{\boldsymbol{\beta}}_{t-2})].$$

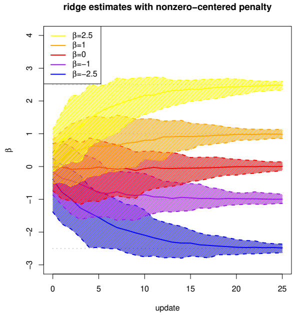
Comparison I

Data from $Y_t = X_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t$ with $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0_n, \sigma^2 I_{nn})$ for $t = 1, 2, \ldots$.

For each t:

- Regular ridge regression + 10-fold CV lambda tuning (*left*).
- Updated ridge regression + 10-fold CV lambda tuning (*right*).

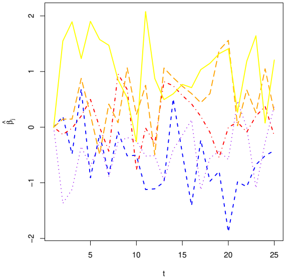




Markov chain

View the sequence $\{\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1})\}_{t=1}^{\infty}$ as a 1^{st} order Markov process with continuous state space $\mathbb{R}^p$ due to:

$$\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t) \,|\, \{\hat{\boldsymbol{\beta}}_{t'}(\lambda_{t'}, \hat{\boldsymbol{\beta}}_{t'-1})\}_{t'=0}^{t} \sim \hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t) \,|\, \hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1}).$$



The process is time-homogeneous:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{t+\tau+1}(\lambda_{t+\tau+1}, \hat{\boldsymbol{\beta}}_{t+\tau}) \;\; &| \, \hat{\boldsymbol{\beta}}_{t+\tau}(\lambda_{t+\tau}, \hat{\boldsymbol{\beta}}_{t+\tau-1}) = \boldsymbol{\beta}, \lambda_{t+\tau} = \lambda, \mathbf{X}_{t+\tau+1} = \mathbf{X} \\ \sim \;\; \hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t) \, &| \, \hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1}) = \boldsymbol{\beta}, \lambda_t = \lambda, \mathbf{X}_{t+1} = \mathbf{X}, \end{aligned}$$

for any $\tau \in \mathbb{N}$, $\mathbf{X} \in \mathcal{M}^{n,p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, and $\lambda > 0$.

Asymptotics

Theorem

Consider a sequence $\{\mathrm{X}_t, \mathrm{Y}_t = \mathrm{X}_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t\}_{t=1}^{\infty}$ *with* $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0_{n_t}, \sigma^2 \mathrm{I}_{n_t, n_t})$ *for all* t.

Let $\{\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1})\}_{t=1}^{\infty}$ *be the related estimator sequence initiated by* $\boldsymbol{\beta}_0$.

Assume $\cap_{t=T}^{\infty} \mathit{null}(\mathrm{X}_t) = 0_p$ *for* T *large.*

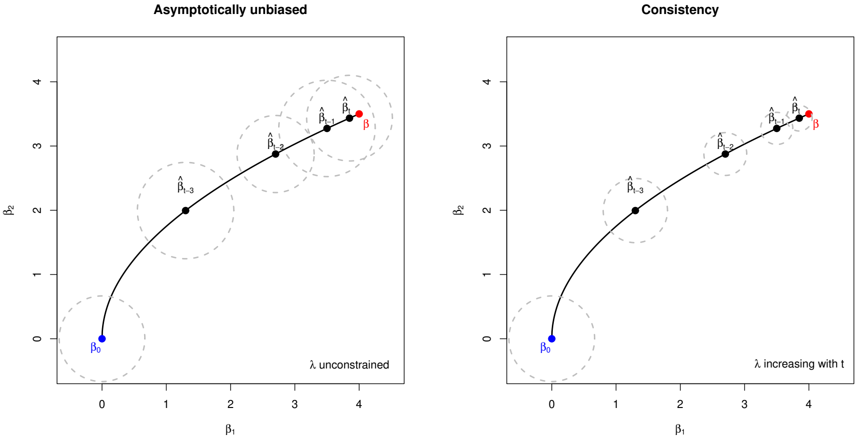
Then, for large $T \in \mathbb{N}$, $\lambda_t > 0$, *the estimator is unbiased, i.e.*

$$\lim_{t \to \infty} \mathbb{E}[\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t)] \quad = \quad \boldsymbol{\beta},$$

while, if $\lim_{t \to \infty} \sigma_{\varepsilon}^2 \, p \, d_1^2(\mathrm{X}_t) \, \lambda_t^{-2} = 0$ *with* $d_1(\mathrm{X}_t)$ *the largest singular value of* X_t, *the estimator is also consistent, i.e.*

$$\lim_{t \to \infty} P[\|\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1}) - \boldsymbol{\beta} \,\| \geq c] \quad \to \quad 0.$$

The role of λ



Constrained cross-validation

A heuristic regularization scheme to have a good *future* predictive performance, while not neglecting the *past*:

$$\lambda_t^{(\text{opt})} \quad = \quad \arg\min_{\lambda_t \in \mathcal{D}} K^{-1} \sum_{k=1}^{K} \| \mathbf{Y}_t^{(k)} - \mathbf{X}_t^{(k)} \hat{\boldsymbol{\beta}}_t^{(-k)}(\lambda_t) \|_2^2$$

with

$$\begin{aligned} \mathcal{D} \quad := \quad \Big\{ \lambda_t > 0 : (1 - f_t) K^{-1} \sum_{k=1}^{K} \sum_{\tau=1}^{t-1} \| \mathbf{Y}_\tau - \mathbf{X}_\tau \boldsymbol{\beta}_t^{(-\hat{k})}(\lambda_t) \quad \|_2^2 \\ \leq \quad \sum_{\tau=1}^{t-1} \| \mathbf{Y}_\tau - \mathbf{X}_\tau \hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1}) \|_2^2 \Big\}, \end{aligned}$$

where $f_t = n_t / (\sum_{\tau=1}^{t} n_\tau)$ provides leverage.

The constraint safeguards against ‘outlying’ novel data, and propogates the latest update of the estimator.

Comparison II

Mixed model to capture batch effect:

$$\mathbb{Y}_t = \mathbb{X}_t\boldsymbol{\beta} + \underbrace{\mathbb{Z}_t\mathbb{G}_t}_{\text{random batch effect}} + \mathbb{E}_t$$

where e.g. $\mathbb{Y}_t = (\mathsf{Y}_1^\top, \mathsf{Y}_2^\top, \ldots, \mathsf{Y}_t^\top)^\top$.

Theorem

Consider a sequence $\{\mathsf{X}_t, \mathsf{Y}_t = \mathsf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t\}_{t=1}^{\infty}$ *with* X_t *be orthonormal and* $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0_{n_t}, \sigma^2 \mathsf{I}_{n,n})$ *for all* t.

Let $\{\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1})\}_{t=1}^{\infty}$ *be the related estimator sequence initiated by* $\hat{\boldsymbol{\beta}}_1^{(me)}$.

If $\lambda_t > \sigma_\varepsilon(\sigma_\varepsilon^2 + \sigma_\gamma^2)^{-1/2} 2^{t/2} T^{1/2}$ *for* $1 \le t \le T$, *then*

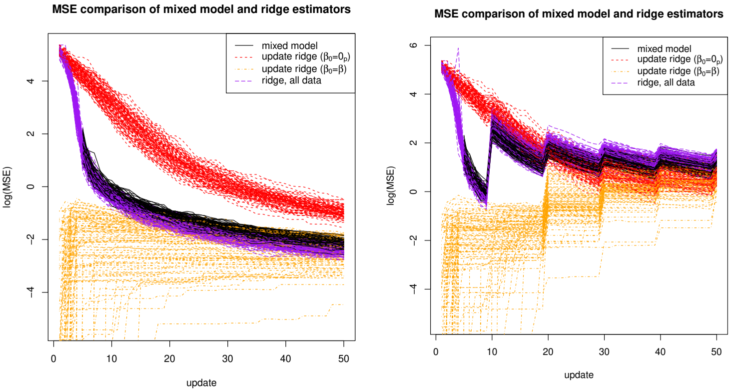
$$MSE[\hat{\boldsymbol{\beta}}_T(\lambda_T)] \quad < \quad MSE[\hat{\boldsymbol{\beta}}_T^{(me)}].$$

Comparison II

Left: $\mathrm{Y}_t = \mathrm{X}_t\boldsymbol{\beta} + \varepsilon_t$ for $t = 1, \ldots, 50$.

Right: $\mathrm{Y}_t = \mathrm{X}_t\boldsymbol{\beta} + \varepsilon_t$ for $t = 1, \ldots, 50$ s.t. $t \mod 10 \neq 0$,

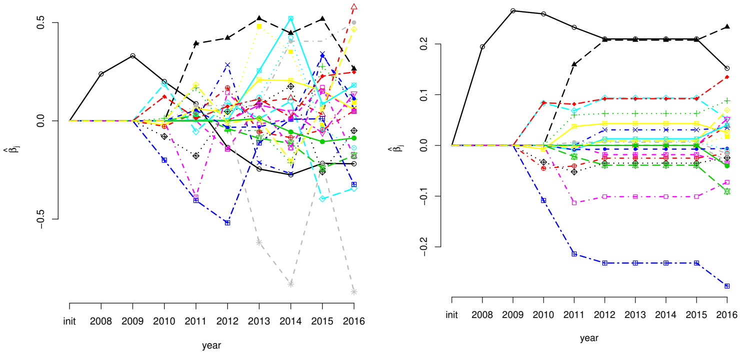
$\mathrm{Y}_t = \varepsilon_t$ for $t = 1, \ldots, 50$ s.t. $t \mod 10 = 0$.



The Fingertips study

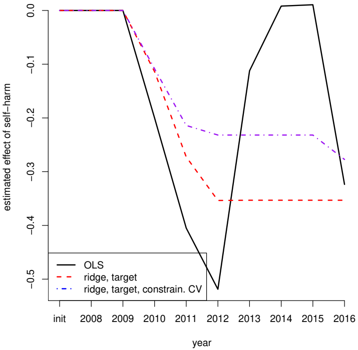
Estimate `suicide rate` $=$ `alcohol` $\times \beta_1 + \ldots +$ `error` by

- OLS (*left*),
- updated ridge regression + unconstrained K-fold CV.
- updated ridge regression + constrained K-fold CV (*right*).



The Fingertips study

Updated ridge regression estimator yields sign consistent estimates over time.



Conclusion

Frequentist version of Bayesian updating.

Also available for:
- logistic regression,
- Gaussian graphical model.

References:

[1] van Wieringen W.N., Binder, H. (2022). Sequential learning of regression models by penalized estimation. *Journal of Computational and Graphical Statistics*, accepted.

[2] van Wieringen, W. and Aflakparast, M. (2021). `porridge`: Ridge-Type Estimation of a Potpourri of Models. R package version 0.2.1, `https://CRAN.R-project.org/package=porridge`.