Adaptive L^q penalized estimation for diffusion processes in Yuima package

Francesco Iafrate
Joint work with A. De Gregorio

Department Statistical Sciences, Sapienza Univeristy, Rome

3rd Yuima workshop

June 26, 2019

Table of Contents

Penalized estimation: basic ideas I

- "Penalized estimator"

$$\hat{\theta} = \arg\min_{\theta \in \Theta} [L(\theta) + \lambda \|\theta\|_q^q] \tag{1}$$

- L: loss function (e.g. quadratic)
- $\|\theta\|_q^q = \sum_i \theta_i^q$: L^q metric on parameter space Θ
- $\lambda > 0$: tuning parameter
- why? Estimation + Model selection
- Desirable properties (oracle)
  1. Consistency
  2. Asymptotic Normality
  3. Selection consistency

Penalized estimation: basic ideas II

- $q = 1$: LASSO
  $q = 2$: ridge
  $q \to 0$: AIC-like criterion
  $0 < q \leq 1$
- LSA (Least Squares Approximation) estimator.

  Consider Taylor expansion around $\tilde{\theta}$, minimum for L, with Hessian G

$$L(\theta) \approx L(\tilde{\theta}) + (\theta - \tilde{\theta})^T \nabla L(\tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})^T G(\tilde{\theta})(\theta - \tilde{\theta}) \quad (2)$$

$$\hat{\theta}_{LSA} = \arg\min_{\theta \in \Theta} [(\theta - \tilde{\theta})^T G(\tilde{\theta})(\theta - \tilde{\theta}) + \lambda \|\theta\|_q^q] \quad (3)$$

- Adaptive LASSO: variable amount of shrinkage for each factor. Penalty term becomes $\lambda_0 \sum_{j=1}^{p} w_j \theta_j^q$, typically $w_j = 1/\tilde{\theta}_j^\delta$, $\delta > 0$.

L^q penalized LSA estimator I

Generalized penalization scheme: different "groups" of parameters can be penalized with different "norms"

Motivating example

SDE driven by Brownian motion, i.e. d-dimensional solution process $X := (X_t)_{t\geq 0}$ to the SDE

$$dX_t = b(X_t, \alpha)dt + \sigma(X_t, \beta)dW_t, \quad X_0 = x_0, \tag{4}$$

$b : \mathbb{R}^d \times \Theta_\alpha \to \mathbb{R}^d$, $\sigma : \mathbb{R}^d \times \Theta_\beta \to \mathbb{R}^d \otimes \mathbb{R}^r$: known functions (up to α and β)
Furthermore, $\alpha \in \Theta_\alpha \subset \mathbb{R}^{p_1}, \beta \in \Theta_\beta \subset \mathbb{R}^{p_2}, p_1, p_2 \in \mathbb{N}$, are unknown parameters

IDEA: consider different penalization schemes for diffusion and drift parameters

L^q penalized LSA estimator II

- Parameter of interest:
  $\theta := (\alpha, \beta) = (\alpha_1, ..., \alpha_{p_1}, \beta_1, ..., \beta_{p_2})' \in \Theta := \Theta_\alpha \times \Theta_\beta \subset \mathbb{R}^{p_1+p_2}$
- True value of the parameter:
  $\theta_0 := (\alpha_0, \beta_0) := (\alpha_{0,1}, ..., \alpha_{0,p_1}, \beta_{0,1}, ..., \beta_{0,p_2})'$
- some components of α_0 and β_0 are exactly zero: $p_1^0 := |\{j : \alpha_{0,j} \neq 0\}|$ and $p_2^0 := |\{j : \beta_{0,j} \neq 0\}|$.
- Initial estimator: e.g. qmle $\tilde{\theta}_n \in \arg\min_\theta(-H_n(\theta))$
- adaptive q-LASSO estimator $\hat{\theta}^{(q)}$, with $q := (q_1, q_2)$,

  $$\hat{\theta}_n^{(q)} = (\hat{\alpha}_n^{(q_1)}, \hat{\beta}_n^{(q_2)}) \in \arg\min_\theta \mathcal{F}_n^{(q)}(\theta) \tag{5}$$

  where

  $$\mathcal{F}_n^{(q)}(\theta) := (\theta - \tilde{\theta}_n)'\hat{G}_n(\tilde{\theta}_n)(\theta - \tilde{\theta}_n) + \sum_{j=1}^{p_1} \lambda_{n,j}|\alpha_j|^{q_1} + \sum_{k=1}^{p_2} \gamma_{n,k}|\beta_k|^{q_2} \tag{6}$$

  $\hat{G}_n(\tilde{\theta}_n) := \partial_\theta^2 L_n(\tilde{\theta}_n)$.
- $(\lambda_{n,j})_{j=1}^{p_1}$, $(\gamma_{n,k})_{k=1}^{p_2} > 0$: adaptive amount of the shrinkage for each element of α and β. (e.g. $\lambda_{n,j} = \lambda_{0,n}/\tilde{\alpha}_j^{\delta_1}, \gamma_{n,j} = \gamma_{0,n}/\tilde{\beta}_j^{\delta_2}$)

L^q penalized LSA estimator III

- $a_n := \max\{\lambda_{n,j}, j \le p_1^0\}$, $b_n := \max\{\gamma_{n,k}, k \le p_2^0\}$,
  $c_n := \min\{\lambda_{n,j}, j > p_1^0\}$ and $d_n := \min\{\gamma_{n,k}, k > p_2^0\}$.
- r_n, $s_n > 0$, tending to 0 as $n \to \infty$.

$$A_n := \begin{pmatrix} r_n \mathbf{I}_{p_1} & 0 \\ 0 & s_n \mathbf{I}_{p_2} \end{pmatrix}$$

Assumptions:

A1. There exists $p_i \times p_i$ positive definite symmetric matrix $G_i, i = 1, 2$, such that

$$A_n \hat{G}_n A_n \xrightarrow{p} \text{diag}(G_1, G_2).$$

A2. The estimator $\tilde{\theta}_n$ is consistent; i.e.

$$A_n^{-1}(\tilde{\theta}_n - \theta_0) = (r_n^{-1}(\tilde{\alpha}_n - \alpha_0), s_n^{-1}(\tilde{\beta}_n - \beta_0)) = O_p(1).$$

A3. The estimator $\tilde{\theta}_n$ is asymptotically normal; i.e.

$$A_n^{-1}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} N_{p_1+p_2}(0, \text{diag}(G_1^{-1}, G_2^{-1})).$$

L^q penalized LSA estimator IV

Main results.(De Gregorio, I., *ongoing*)

Theorem (Consistency)

Under the assumptions A1, A2 and by assuming that $r_n a_n = O_p(1)$ *and* $s_n b_n = O_p(1)$, *we have that* $A_n^{-1}(\hat{\theta}_n^{(q)} - \theta_0) = O_p(1)$.

Theorem (Selection consistency)

Under the assumptions A1, A2 and by assuming that $r_n a_n = O_p(1)$, $s_n b_n = O_p(1)$, $r_n^{(2-q_1)} c_n \xrightarrow{p} \infty$ *and* $s_n^{(2-q_2)} d_n \xrightarrow{p} \infty$ *we have that*

$$P(\hat{\alpha}_{n,j}^{(q_1)} = 0) \longrightarrow 1, \quad j = p_1^0 + 1, ..., p_1;$$

$$P(\hat{\beta}_{n,k}^{(q_2)} = 0) \longrightarrow 1, \quad k = p_2^0 + 1, ..., p_2.$$

Theorem (Asymptotic normality)

Under the assumptions A1-A3 and by assuming that $r_n a_n = o_p(1)$, $s_n b_n = o_p(1)$, $r_n^{(2-q_1)} c_n \xrightarrow{p} \infty$ *and* $s_n^{(2-q_2)} d_n \xrightarrow{p} \infty$, *we have that*

$$(r_n^{-1}(\hat{\alpha}_n^{(q_1)} - \alpha_0)_{\mathrm{S}^1}, s_n^{-1}(\hat{\beta}_n^{(q_2)} - \beta_0)_{\mathrm{R}^1}) \xrightarrow{d} N_{p_1^0 + p_2^0}(0, \mathit{diag}((G_1)_{\mathrm{S}^{11}}^{-1}, (G_2)_{\mathrm{R}^{11}}^{-1})).$$

see also Suzuki, Yoshida (2018), Masuda, Shimizu (2017).

Table of Contents

Beyond cross-validation I

- cross validation techniques: work for i.i.d. data

  Treat all observations as equivalent: randomly split sample in K folds, use $K-1$ for estimation, $K-$th for prediction.

  Not viable for dependent obs!

- Example from non–parametric literature.

$$\hat{b}_n(x) = \frac{1}{\Delta_n} \frac{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h_n^b}\right)(X_j - X_{j-1})}{\sum_{j=0}^{n} K\left(\frac{X_j - x}{h_n^b}\right)}$$

$$\hat{\sigma}_n(x) = \frac{1}{\Delta_n} \frac{\sum_{j=0}^{n} K\left(\frac{X_j - x}{h_n^a}\right)(X_j - X_{j-1})^2}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h_n^\sigma}\right)}$$

  Kernel estimation of drift and infinitesimal variance require selection of bandwidths h_n^b, h_n^σ.

- Bandi, Corradi, Moloche (2009) introduce a two step selection procedure based on residuals: choose tuning parameters so that they are approximately normally distributed.

Beyond cross-validation II

- Euler-Maruyama discretization of the solution of (4).

$$X_{t_{i+1}^n} = X_{t_i^n} + b(X_{t_i^n}, \alpha)\Delta_n + \sigma(X_{t_i^n}, \beta)\Delta W_{t_i^n} \tag{7}$$

t_i, Δ_n s.t. $n\Delta_n \to \infty$, $\Delta_n \to 0$ and $n\Delta_n^p \to 0$ as $n \to \infty$, $p \geq 2$

The "residuals" are then defined as

$$r_{t_i^n} = \Delta_n^{-1/2}\Sigma^{-1/2}(X_{t_i^n}, \beta)(X_{t_{i+1}^n} - X_{t_i^n} - \Delta_n b(X_{t_i^n}, \alpha)) \tag{8}$$

$i = 1, \ldots, n$, where $\Sigma = \sigma\sigma^T$.

- $r_{t_i^n}$ are approx. *i.i.d.* $\mathcal{N}_d(0_d, I_d)$.

Given a sample $\mathbf{X}_n = (X_{t_i^n})$ and estimates of the parameters $\hat{\alpha}$ and $\hat{\beta}$,the residuals can be estimated as

$$\hat{r}_{t_i} = \Delta_n^{-1/2}\Sigma^{-1/2}(X_{t_i^n}, \hat{\beta})(X_{t_{i+1}^n} - X_{t_i^n} - \Delta_n b(X_{t_i^n}, \hat{\alpha})) \tag{9}$$

Idea: find tuning parameters in such a way that the residuals fit best to a white noise scheme.

Beyond cross-validation III

- $\psi = (q_1, q_2, \lambda_{n,0}, \gamma_{n,0}, \delta_1, \delta_2)$ vector of tuning parameters varying in some suitable parameter space Ψ .
- choose ψ by optimizing some score function S penalizing tuning parameters producing correlated/non-gaussian residuals

$$\psi^* = \arg\min_{\psi \in \Psi} S(r_{t_1^n}(\psi), \ldots, r_{t_n^n}(\psi)) \tag{10}$$

- **Example 1.** penalty function for ψ can be the test statistic in a white noise hypothesis testing, e.g.Ljung-Box test statistic

$$Q_n(\ell) = n(n+2) \sum_{j=1}^{\ell} \frac{\hat{\rho}_j^2}{n-j} \tag{11}$$

where ℓ is the number of lags to be tested, $\hat{\rho}_j$ denotes the sample auto-correlations at lag j Under the null hypothesis that the observations are not correlated up to lag ℓ

Beyond cross-validation IV

- **Example 2.** penalty measuring the distance of the empirical distribution of the residuals from the Gaussian distribution function: equip the space of distribution functions with some norm $\|\cdot\|$.

$$\psi^* = \arg\min_{\psi \in \Psi} \|\hat{F}_n(r_{t_1^n}(\psi), \dots, r_{t_n^n}(\psi)) - Q_d\| \tag{12}$$

  where $\hat{F}_n$ denotes the empirical distribution function and Q_d denotes the distribution function of the $d-$ dimensional standard Gaussian distribution.

- sup norm: Kolmogorov – Smirnov test statistic.

Table of Contents

R implementation with Yuima I

The `qlasso` function performs an adaptive q-lasso estimate.

```
qlasso = function (yuima, q=c(1, 1), lambda0=c(1, 1),
          delta = c(.5, .5), opt_args = list(), select_tuning = T )
```

Arguments.

yuima a yuima object, contaning the data and the model specification.

q a numeric vector with two components, containing the q_1 and q_2 parameters.

lambda0 vector with two components, $(\lambda_{0,n}, \gamma_{0,n})$ coefficients of the adaptive weights.

delta vector with two components, δ_1, δ_2 exponents in the adaptive weights.

select_tuning logical value indicating whether to perform the tuning parameter search.

R implementation with Yuima II

opt_args list containing arguemnts to be supplied to optimization routine. Specifically

- **start_p, upper_p, lower_p** starting point, upper and lower bounds for the parameter search
- **method_p, method_h** optimization algorithms to be used for the parameter and tuning parameter selection, respectively

Value. Returns a list with both QMLE and LASSO estimates, their standard deviations and the tuning parameters.

Details. This function behaves much like the `yuima::lasso` function. From an initial guess of QML estimates, performs adaptive q-LASSO estimation using the Least Squares Approximation (LSA). If `select_tuning` is true the values supplied for the tuning parameters are passed to the optimizer as starting points for the optimal tuning parameters search.

R implementation with Yuima III

The following algorithm implements criterion (10).

- Step 0. Suppose a set of data points $\{x_{t_i^n}\}$ is given. Initialize the tuning parameter vector ψ with some value ψ_0. Fix a threshold $\epsilon > 0$.
- Until convergence is reached:
  - Step 1. Compute the current *q*-lasso estimates with the current value $\psi^{(k)}$ of the tuning parameters $\hat{\alpha}^{(k)} = \hat{\alpha}(\psi^{(k)})$, $\hat{\beta}^{(k)} = \hat{\beta}(\psi^{(k)})$.
  - Step 2. Compute the residuals $\{\hat{r}_{t_i}^{(k)}\}_i = \{\hat{r}_{t_i}(\psi^{(k)})\}_i$ as in fomula (9), with the current estimates of the parameters $\hat{\alpha}^{(k)}$ and $\hat{\beta}^{(k)}$.
  - Step 3. Evaluate the score of the current residuals $s^{(k)} = S(\hat{r}_{t_1}^{(k)}, \ldots, \hat{r}_{t_n}^{(k)})$
  - Step 4. If $|s^{(k)} - s^{(k-1)}| < \epsilon$ stop: convergence is reached. Set $\psi^* = \psi^{(k)}$ and return the optimal q-lasso estimates of the parameters $\alpha^* = \alpha^{(k)}$ and $\beta^* = \beta^{(k)}$. Otherwise move to some new point $\psi^{(k+1)}$ (chosen according to some optimization algorithm) and repeat Steps 1 to 4.

Noisy Data I

$$\begin{pmatrix} \mathrm{d}X_t^1 \\ \mathrm{d}X_t^2 \end{pmatrix} = \begin{pmatrix} -\theta_{2.1} X_t^1 - \theta_{2.2} \\ -\theta_{2.2} X_t^2 - \theta_{2.1} \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \theta_{1.1} & 1 \\ \theta_{1.2} & 1 \end{pmatrix} \begin{pmatrix} \mathrm{d}W_t^1 \\ \mathrm{d}W_t^2 \end{pmatrix} \tag{13}$$

$$X_0^1 = 1, X_0^2 = 1$$

$\theta_{1.1}^0 = 1.6, \theta_{1.2}^0 = 0, \theta_{2.1}^0 = 3.5, \theta_{2.2}^0 = 0.$

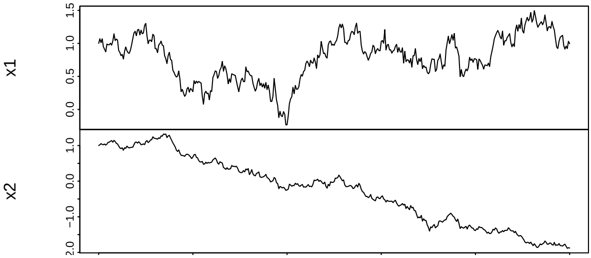

Figure: sample path of (X_1, X_2)

Noisy Data II

Add an exogenous measurement error to the observations $(Z^1, Z^2) \sim \mathcal{N}(0, \Sigma_Z)$

$$\Sigma_Z = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

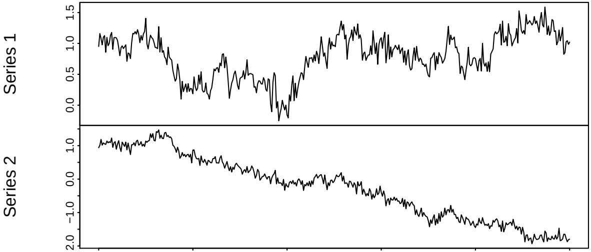

Figure: perturbed sample path of (X_1, X_2)

Noisy Data III

- Compare MSEs

	$\theta^0_{1.1}$	$\theta^0_{1.2}$	$\theta^0_{2.1}$	$\theta^0_{2.2}$
LASSO	1.009	1.754	1.365	4.262
qLASSO	0.772	1.752	1.809	3.012

Table: MSE tables for $n = 800$, $B = 1000$

- Relative efficiencies

$$e(\hat{\theta}^{(q)}, \theta^{(L)}; \theta^0_{1.2} = 0) = 1.001$$
$$e(\hat{\theta}^{(q)}, \theta^{(L)}; \theta^0_{2.2} = 0) = 1.415$$

CKLS Model I

CKLS: a large family of models

$$dX_t = (\theta_1 + \theta_2 X_t)dt + \theta_3 X_t^{\theta_4} \, dW_t \qquad (14)$$

	θ_1	θ_2	θ_4	See
Merton	Any	0	0	Merton (1973b)
Vasicek or Ornstein–Uhlenbeck	Any	Any	0	Vasicek (1977)
CIR or square root process	Any	Any	1/2	Cox et al. (1985)
Dothan	0	0	1	Dothan (1978)
Geometric BM or Black and Scholes	0	Any	1	Black and Scholes (1973)
Brennan and Schwartz	Any	Any	1	Brennan and Schwartz (1980)
CIR VR	0	0	3/2	Cox et al. (1980)
CEV	0	Any	Any	Cox (1996)

Figure: Family of CKLS processes and its embedded elements under different parametric specifications. (source: the YUIMA Book, p. 74)

CKLS Model II

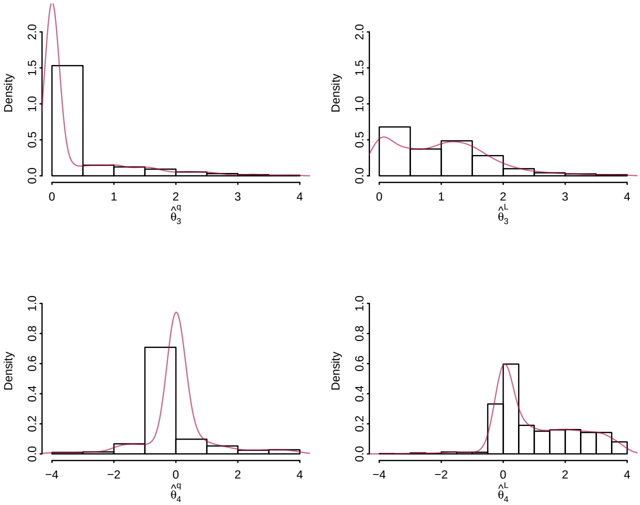

Figure: Empirical Density Estimates of $\hat{\theta}_j^{(q)}, \theta_j^{(L)}$, j = 3,4

CKLS Model III

- Compare MSEs

	θ_1^0	θ_2^0	θ_3^0	θ_4^0
LASSO	1.407	1.668	0.180	0.177
qLASSO	1.074	1.215	0.203	0.245

Table: MSE tables for $n = 400$, $B = 1000$

- Relative efficiencies

$$e(\hat{\theta}^{(q)}, \theta^{(L)}; \theta_1^0 = 0) = 1.311$$
$$e(\hat{\theta}^{(q)}, \theta^{(L)}; \theta_2^0 = 0) = 1.374$$

References

- De Gregorio A., Iafrate F., Adaptive L^q penalized estimation for diffusion processes, *ongoing*
- Suzuki T., Yoshida N., Penalized least squares approximation methods and their applications to stochastic processes, *preprint*, 2018
- Masuda H., Shimizu Y., Moment convergence in regularized estimation under multiple and mixed-rates asymptotics,*Math. Meth., Stat.*, 2017
- De Gregorio A., Iacus S. M., Adaptive LASSO-type estimation for diffusion processes, *Economet. Theory*, 2012
- Iacus S. M., Yoshida N., Simulation and Inference for Stochastic Processes with YUIMA, *Springer*, 2018