On LASSO type estimation for discretely observed diffusion processes

S.M. Iacus (University of Milan)
joint with
A. De Gregorio (University of Rome)

CREST and Sakigake International Symposium, Tokyo Institute of Technology, 15 Dec. 2010

Summary

- model selection is a data-driven procedure to select a statistical model from the data. It can be based, e.g., on information criteria (like AIC) or Lasso-type approach
- LASSO is a widely used statistical methodology for simultaneous estimation and variable selection.
- LASSO is also a shrinkage method which allows to select parsimonious models.
- we develop the LASSO method for discretely observed diffusion process, but first we recall some result in model selection based on information criteria for these models
- we conclude with some numerical and empirical evidence to corroborate the theoretical results

Information Criteria

Model selection via AIC

Consider the one-dimensional stochastic differential equation

$$\mathrm{d}X_t = b(X_t, \alpha)\mathrm{d}t + \sigma(X_t, \beta)\mathrm{d}W_t$$

$X_0 = x_0$, where the parameter $\theta = (\alpha, \beta)$ is such that $\theta \in \Theta_\alpha \times \Theta_\beta = \Theta$, $\Theta_\alpha \subset \mathbb{R}^p$, $\Theta_\beta \subset \mathbb{R}^q$, and Θ convex. As usual, $b(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$ are two known (up to α and β) regular functions such that a solution of SDE exists.

X is supposed to be ergodic and the asymptotic is $\Delta_n \to 0$, $n\Delta_n = T \to \infty$ and $n\Delta_n^2 \to 0$ as $n \to \infty$.

The aim is to try to identify the underlying continuous model on the basis of discrete observations using AIC (Akaike Information Criterion) statistics defined as (Akaike 1973,1974)

$$\mathrm{AIC} = -2\ell_n\left(\hat{\theta}_n^{(ML)}\right) + 2\dim(\Theta),$$

where $\hat{\theta}_n^{(ML)}$ is the true maximum likelihood estimator and $\ell_n(\theta)$ is the log-likelihood

Akaike's index idea is to penalize this value

$$-2\ell_n\left(\hat{\theta}_n^{(ML)}\right)$$

with the dimension of the parameter space

$$2\dim(\Theta)$$

Thus, as the number of parameter increases, the fit may be better, i.e. $-2\ell_n\left(\hat{\theta}_n^{(ML)}\right)$ decreases, at the cost of overspecification and $\dim(\Theta)$ compensate for this effect.

When comparing several models for a given data set, the models such that the AIC is lower is preferred.

Model selection via AIC

In order to calculate

$$\mathrm{AIC} = -2\ell_n\left(\hat{\theta}_n^{(ML)}\right) + 2\dim(\Theta),$$

we need to evaluate the exact value of the log-likelihood $\ell_n(\cdot)$ at point $\hat{\theta}_n^{(ML)}$.

Problem: for discretely observed diffusion processes the true likelihood function is not known in most cases

An approximate likelihood (local gaussian, hermite polynomial expansion, etc.) may be good for estimation purposes but not necessarily to obtain good estimates of $\ell_n\left(\hat{\theta}_n^{(ML)}\right)$. See, e.g. I. (2008) for a review.

Uchida and Yoshida (2005) considered the following approximation of the log-likelihood ℓ_n due to Dacunha-Castelle and Florens-Zmirou (1986)

$$u_n(\theta) = \sum_{k=1}^{n} u(\Delta_n, X_{i-1}, X_i, \theta)\,,$$

where

$$u(t,x,y,\theta) = -\frac{1}{2}\log(2\pi t) - \log\sigma(y,\beta) - \frac{S^2(x,y,\beta)}{2t} + H(x,y,\theta) + t\tilde{g}(x,y,\theta)\,,$$

with

$$S(x,y,\beta) = \int_x^y \frac{\mathrm{d}u}{\sigma(u,\beta)}\,, \quad H(x,y,\theta) = \int_x^y \frac{B(u,\theta)}{\sigma(u,\beta)}\mathrm{d}u$$

$$\tilde{g}(x,y,\theta) = -\frac{1}{2}\left\{C(x,\theta) + C(y,\theta) + \frac{1}{3}B(x,\theta)B(y,\theta)\right\}$$

$$C(x,\theta) = \frac{1}{2}B^2(x,\theta) + \frac{1}{2}B_x(x,\theta)\sigma(x,\beta), \quad B(x,\theta) = \frac{b(x,\alpha)}{\sigma(x,\beta)} - \frac{1}{2}\sigma_x(x,\beta)$$

Uchida and Yoshida (2005) proposed to use the previous approximation of the likelihood and instead of the true ML estimator, the approximated ML estimator (AML) of the local gaussian approximation is plugged into AIC. So the proposed AIC statistics is as follows

$$\mathrm{AIC} = -2u_n\left(\hat{\theta}_n^{(AML)}\right) + 2\dim(\Theta)\,.$$

This statistics is a proper approximation of the true AIC statistics, i.e. $\mathbb{E}\{u_n(\theta_0) - \ell_n(\theta_0)\} = o(1)$.

Notice again that if u_n does not properly approximate the true ℓ_n the AIC statistics is completely useless.

The **sde** package for the R statistical environment implements this AIC statistics as the function `sdeAIC`.

We compare three models

$$\mathrm{d}X_t = -\alpha_1(X_t - \alpha_2)\mathrm{d}t + \beta_1\sqrt{X_t}\mathrm{d}W_t \qquad \text{(true model)},$$
$$\mathrm{d}X_t = -\alpha_1(X_t - \alpha_2)\mathrm{d}t + \sqrt{\beta_1 + \beta_2 X_t}\mathrm{d}W_t \qquad \text{(competing model 1)},$$
$$\mathrm{d}X_t = -\alpha_1(X_t - \alpha_2)\mathrm{d}t + (\beta_1 + \beta_2 X_t)^{\beta_3}\mathrm{d}W_t \qquad \text{(competing model 2)},$$

We call the above models Mod1, Mod2 and Mod3.

We generate data from Mod1 with parameters

$$\mathrm{d}X_t = -(X_t - 10)\mathrm{d}t + 2\sqrt{X_t}\mathrm{d}W_t\,,$$

and initial value $X_0 = 8$. We use $n = 1000$ and $\Delta = 0.1$.

We test the performance of the AIC statistics for the three competing models

Simulation results. 1000 Monte Carlo replications

$$\mathrm{d}X_t = -(X_t - 10)\mathrm{d}t + 2\sqrt{X_t}\mathrm{d}W_t \quad \text{(true model)},$$

$$\mathrm{d}X_t = -\alpha_1(X_t - \alpha_2)\mathrm{d}t + \beta_1\sqrt{X_t}\mathrm{d}W_t \quad \text{(Model 1)}$$

$$\mathrm{d}X_t = -\alpha_1(X_t - \alpha_2)\mathrm{d}t + \sqrt{\beta_1 + \beta_2 X_t}\mathrm{d}W_t \quad \text{(Model 2)}$$

$$\mathrm{d}X_t = -\alpha_1(X_t - \alpha_2)\mathrm{d}t + (\beta_1 + \beta_2 X_t)^{\beta_3}\mathrm{d}W_t \quad \text{(Model 3)}$$

Model selection via AIC

Model 1 (true)	Model 2	Model 3
99.2 %	0.6 %	0.2 %

QMLE estimates under the different models

	α_1	α_2	β_1	β_2	β_3
Model 1	1.10	8.05	0.90		
Model 2	1.12	8.07	2.02	0.54	
Model 3	1.12	9.03	7.06	7.26	0.61

Shrinkage Estimation

The Least Absolute Shrinkage and Selection Operator (LASSO) is a useful and well studied approach to the problem of model selection and its major advantage is the simultaneous execution of both parameter estimation and variable selection (see Tibshirani, 1996; Knight and Fu, 2000, Efron *et al.*, 2004).

To simplify the idea: take a full specified regression model

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_k X_k$$

perform least squares estimation under L_1 constraints, i.e.

$$\hat{\theta} = \arg\min_{\theta} \left\{ (Y - \theta X)^T (Y - \theta X) + \sum_{i=1}^{k} |\theta_i| \right\}$$

model selection occurs when some of the θ_i are estimated as zeros.

LASSO estimation

In the AIC framework, one needs to evaluate the AIC statistics for all possible submodels (a HUGE number), compare the various AIC's and then choose the model with the smallest AIC.

Some heuristic methods like stepwise-regression are possible to reduce the number of models to consider though.

R has '`step`' in `base` and '`stepAIC`' in `MASS`.

An additional feature of the LASSO method is that it is shrinkage estimator (estimates with reduced standard errors)

Ok, but how does it work for diffusion processes? Why are diffusion processes so special?

Let X_t be a multidmensional diffusion process solution to

$$\mathrm{d}X_t = b(\alpha, X_t)\mathrm{d}t + \sigma(\beta, X_t)\mathrm{d}W_t$$

$$\alpha = (\alpha_1, ..., \alpha_p)' \in \Theta_p \subset \mathbb{R}^p, \quad p \geq 1$$

$$\beta = (\beta_1, ..., \beta_q)' \in \Theta_q \subset \mathbb{R}^q, \quad q \geq 1$$

$b : \Theta_p \times \mathbb{R}^d \to \mathbb{R}^d$, $\sigma : \Theta_q \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^m$ and $W_t, t \in [0, T]$, is a standard Brownian motion in $\mathbb{R}^m$.

We assume that the functions b and σ are known up to α and β.

We denote by $\theta = (\alpha, \beta) \in \Theta_p \times \Theta_q = \Theta$ the parametric vector and with $\theta_0 = (\alpha_0, \beta_0)$ its unknown true value.

For a matrix A, we denote by A^{-1} the inverse of A and let $\Sigma(\beta, x) = \sigma(\beta, x)\sigma(\beta, x)'$.

The sample path of X_t is observed only at $n+1$ equidistant discrete times t_i, such that $t_i - t_{i-1} = \Delta_n < \infty$ for $1 \leq i \leq n$ (with $t_0 = 0$ and $t_n = T$). We denote by $\mathbf{X}_n = \{X_{t_i}\}_{0 \leq i \leq n}$ our random sample with values in $\mathbb{R}^{(n+1)\times d}$.

The asymptotic scheme adopted in this talk is the following:

$T = n\Delta_n \to \infty$, $\Delta_n \to 0$ and $n\Delta_n^2 \to 0$ as $n \to \infty$.

This asymptotic framework is called *rapidly increasing design* and the condition $n\Delta_n^2 \to 0$ means that Δ_n shrinks to zero slowly.

Implications: the parameters β are $\sqrt{n}$ – consistent while the parameters α in the drift are only $\sqrt{n\Delta_n}$ – consistent. This requires a non trivial adaptation of the LASSO method.

Regularity conditions

$\mathcal{A}_1$. there exists a constant C such that

$$|b(\alpha_0, x) - b(\alpha_0, y)| + |\sigma(\beta_0, x) - \sigma(\beta_0, y)| \leq C|x - y|;$$

$\mathcal{A}_2$. $\inf_{\beta,x} \det(\Sigma(\beta, x)) > 0$;

$\mathcal{A}_3$. the process $X_t, t \in [0, T]$, is ergodic for every θ with invariant probability measure μ_θ;

$\mathcal{A}_4$. if the coefficients $b(\alpha, x) = b(\alpha_0, x)$ and $\sigma(\beta, x) = \sigma(\beta_0, x)$ for all x (μ_{θ_0}-almost surely), then $\alpha = \alpha_0$ and $\beta = \beta_0$;

$\mathcal{A}_5$. for all $m \geq 0$ and for all $\theta \in \Theta$, $\sup_t E|X_t|^m < \infty$;

$\mathcal{A}_6$. for every $\theta \in \Theta$, the coefficients $b(\alpha, x)$ and $\sigma(\beta, x)$ are five times differentiable with respect to x and the derivatives are bounded by a polynomial function in x, uniformly in θ;

$\mathcal{A}_7$. the coefficients $b(\alpha, x)$ and $\sigma(\beta, x)$ and all their partial derivatives respect to x up to order 2 are three times differentiable with respect to θ for all x in the state space. All derivatives with respect to θ are bounded by a polynomial function in x, uniformly in θ.

$\mathcal{A}_1$ ensures the existence and uniqueness of a solution to the SDE for the value $\theta_0 = (\alpha_0, \beta_0)$ of $\theta \in \Theta$, while $\mathcal{A}_4$ is the identifiability condition. From now on we assume that the conditions $\mathcal{A}_1 - \mathcal{A}_7$ hold.

Quasi-likelihood function

We can discretize the SDE

$$X_{t+\mathrm{d}t} - X_t = b(\alpha, X_t)\mathrm{d}t + \sigma(\beta, X_t)(W_{t+\mathrm{d}t} - W_t),$$

and the increments $X_{t+\mathrm{d}t} - X_t$ are then independent Gaussian random variables with mean $b(\alpha, X_t)\mathrm{d}t$ and variance-covariance matrix $\Sigma(\beta, x)\mathrm{d}t$. Therefore the transition density of the process can be written as a simple Gaussian density.

Quasi-likelihood function

$$\mathbb{H}_n(\mathbf{X}_n, \theta) = \frac{1}{2}\sum_{i=1}^{n}\left\{\log \det(\Sigma_{i-1}(\beta)) + \frac{1}{\Delta_n}(\Delta X_i - \Delta_n b_{i-1}(\alpha))'\Sigma_{i-1}^{-1}(\beta)(\Delta X_i - \Delta_n b_{i-1}(\alpha))\right\}$$

where $\Delta X_i = X_{t_i} - X_{t_{i-1}}$, $\Sigma_i(\beta) = \Sigma(\beta, X_{t_i})$ and $b_i(\alpha) = b(\alpha, X_{t_i})$.

This quasi-likelihood has been introduced by, e.g., Yoshida (1992), Genon-Catalot and Jacod (1993) and Kessler (1997) and used to obtain quasi-MLE estimators.

$\mathbb{H}_n$ plays the role of the negative log-likelihood ($-\ell_n$ of previous part of the talk) for this model. The quasi-MLE $\tilde{\theta}_n$ for this model is the solution of the following problem

$$\tilde{\theta}_n = (\tilde{\alpha}_n, \tilde{\beta}_n)' = \arg\min_{\theta} \mathbb{H}_n(\mathbf{X}_n, \theta)$$

Optimality properties of the QMLE estimator

Consider the matrix (of rates of convergence)

$$\varphi(n) = \begin{pmatrix} \frac{1}{n\Delta_n}\mathbf{I}_p & 0 \\ 0 & \frac{1}{n}\mathbf{I}_q \end{pmatrix}$$

where $\mathbf{I}_p$ and $\mathbf{I}_q$ are respectively the identity matrix of order p and q. Let

$$\mathcal{I}(\theta) = \begin{pmatrix} \Gamma_\alpha = [\mathcal{I}_b^{kj}(\alpha)]_{k,j=1,\dots,p} & 0 \\ 0 & \Gamma_\beta = [\mathcal{I}_\sigma^{kj}(\beta)]_{k,j=1,\dots,q} \end{pmatrix}$$

where

$$\mathcal{I}_b^{kj}(\alpha) = \int \frac{1}{\sigma^2(\beta,x)} \frac{\partial b(\alpha,x)}{\partial \alpha_k} \frac{\partial b(\alpha,x)}{\partial \alpha_j} \mu_\theta(dx)\,,$$

$$\mathcal{I}_\sigma^{kj}(\beta) = 2\int \frac{1}{\sigma^2(\beta,x)} \frac{\partial \sigma(\beta,x)}{\partial \beta_k} \frac{\partial \sigma(\beta,x)}{\partial \beta_j} \mu_\theta(dx)\,.$$

Optimality properties of the QMLE estimator

Lemma 1 (see e.g., Kessler, 1997)**.** *Let* $\Lambda_n(\theta) = \varphi(n)^{1/2}\ddot{\mathbb{H}}_n(\mathbf{X}_n, \theta)\varphi(n)^{1/2}$*. Under the conditions* $\mathcal{A}_1 - \mathcal{A}_7$*, and* $n\Delta_n \to \infty$*,* $n\Delta_n^2 \to 0$*,* $\Delta_n \to 0$ *as* $n \to \infty$*, the following two properties hold true*

i) for $\epsilon_n \to 0$*, as* $n \to \infty$*, then*

$$\Lambda_n(\theta_0) \xrightarrow{p} \mathcal{I}(\theta_0)$$

$$\sup_{||\theta|| \leq \epsilon_n} |\Lambda_n(\theta + \theta_0) - \Lambda_n(\theta_0)| = o_p(1)$$

ii) for each $\theta \in \Theta$*,* $\tilde{\theta}_n$ *is a consistent estimator of* θ *and asymptotically Gaussian, i.e.*

$$\varphi(n)^{-1/2}(\tilde{\theta}_n - \theta) \xrightarrow{d} N(0, \mathcal{I}(\theta)^{-1})$$

The classical adaptive LASSO objective function for the present model is then

$$\min_{\alpha,\beta} \left\{ H_n(\alpha,\beta) + \sum_{j=1}^{p} \lambda_{n,j}|\alpha_j| + \sum_{k=1}^{q} \gamma_{n,k}|\beta_k| \right\}$$

$\lambda_{n,j}$ and $\gamma_{n,k}$ are appropriate sequences representing an adaptive amount of shrinkage for each element of α and β.

Adaptiveness is essential to avoid the situation in which larger parameter are estimated with larger bias (up to missing consistency)

Unfortunately, the above is a non-linear optimization problem under L_1 constraints which might be numerically challenging to solve. Luckily, following Wang and Leng (2007), the minimization problem can be transformed into a quadratic minimization problem (under L_1 constraints) which is asymptotically equivalent to minimizing the original LASSO objective function.

Idea of Quadratic Approximation

By Taylor expansion of the original LASSO objective function, for θ around $\tilde{\theta}_n$ (the QMLE estimator)

$$\begin{aligned}\mathbb{H}_n(\mathbf{X}_n, \theta) &= \mathbb{H}_n(\mathbf{X}_n, \tilde{\theta}_n) + (\theta - \tilde{\theta}_n)'\dot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n) + \frac{1}{2}(\theta - \tilde{\theta}_n)'\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)(\theta - \tilde{\theta}_n) \\ &\quad + o_p(1) \\ &= \mathbb{H}_n(\mathbf{X}_n, \tilde{\theta}_n) + \frac{1}{2}(\theta - \tilde{\theta}_n)'\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)(\theta - \tilde{\theta}_n) + o_p(1)\end{aligned}$$

with $\dot{\mathbb{H}}_n$ and $\ddot{\mathbb{H}}_n$ the gradient and Hessian of $\mathbb{H}_n$ with respect to θ.

We define the adaptive LASSO estimator the solution to the quadratic problem under L_1 constraints

$$\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n) = \arg\min_{\theta} \mathcal{F}(\theta).$$

with

$$\mathcal{F}(\theta) = (\theta - \tilde{\theta}_n)\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)(\theta - \tilde{\theta}_n)' + \sum_{j=1}^{p} \lambda_{n,j}|\alpha_j| + \sum_{k=1}^{q} \gamma_{n,k}|\beta_k|$$

We will discuss adaptiveness later

LASSO cautions

- Adaptiveness: without adaptiveness, larger (true) parameters are estimated with more bias because of the penalization
- Speed of convergence: in diffusion models the speed of the parameters in the drift (α) and diffusion (β) are different (big difference w.r.t. i.i.d. models)
- Oracle property: the method should correctly estimate as zero the parameters which are truly zero

We will present formally the oracle property of the adaptive LASSO estimator.

Without loss of generality, we assume that the true model, indicated by $\theta_0 = (\alpha_0, \beta_0)$, has parameters α_{0j} and β_{0k} equal to zero for $p_0 < j \leq p$ and $q_0 < k \leq q$, while $\alpha_{0j} \neq 0$ and $\beta_{0k} \neq 0$ for $1 \leq j \leq p_0$ and $1 \leq k \leq q_0$.

Denote by $\theta^* = (\alpha^*, \beta^*)'$ the vector corresponding to the nonzero parameters, where $\alpha^* = (\alpha_1, ..., \alpha_{p_0})'$ and $\beta^* = (\beta_1, ..., \beta_{q_0})'$, while $\theta^\circ = (\alpha^\circ, \beta^\circ)'$ is the vector corresponding to the zero parameters where $\alpha^\circ = (\alpha_{p_0+1}, ..., \alpha_p)'$ and $\beta^\circ = (\beta_{q_0+1}, ..., \beta_q)'$.

Therefore,

$$\text{TRUE :} \quad \theta_0 = (\alpha_0, \beta_0)' = (\alpha_0^*, \alpha_0^\circ, \beta_0^*, \beta_0^\circ)'$$

$$\text{LASSO :} \quad \hat{\theta}_n = (\hat{\alpha}_n^*, \hat{\alpha}_n^\circ, \hat{\beta}_n^*, \hat{\beta}_n^\circ)'$$

$$\text{MLE :} \quad \tilde{\theta}_n = (\tilde{\alpha}_n^*, \tilde{\alpha}_n^\circ, \tilde{\beta}_n^*, \tilde{\beta}_n^\circ)'$$

Intuition behind adaptiveness

$\mathcal{C}_1$. $\frac{\mu_n}{\sqrt{n\Delta_n}} \to 0$ and $\frac{\nu_n}{\sqrt{n}} \to 0$ where $\mu_n = \max\{\lambda_{n,j}, 1 \leq j \leq p_0\}$ and $\nu_n = \max\{\gamma_{n,k}, 1 \leq k \leq q_0\}$;

$\mathcal{C}_2$. $\frac{\kappa_n}{\sqrt{n\Delta_n}} \to \infty$ and $\frac{\omega_n}{\sqrt{n}} \to \infty$ where $\kappa_n = \min\{\lambda_{n,j}, j > p_0\}$ and $\omega_n = \min\{\gamma_{n,k}, k > q_0\}$.

Assumption $\mathcal{C}_1$ implies that the maximal tuning coefficients μ_n and ν_n for the parameters α_j and β_k, with $1 \leq j \leq p_0$ and $1 \leq k \leq q_0$, tends to infinity slower than $\sqrt{n\Delta_n}$ and $\sqrt{n}$ respectively.

Analogously, we observe that $\mathcal{C}_2$ means that that the minimal tuning coefficient for the parameter α_j and β_k, with $j > p_0$ and $k > q_0$, tends to infinity faster than $\sqrt{n\Delta_n}$ and $\sqrt{n}$ respectively.

Theorem 2. *Under conditions $\mathcal{A}_1 - \mathcal{A}_7$ and $\mathcal{C}_1$, one has that*

$$||\hat{\alpha}_n - \alpha_0|| = O_p\left((n\Delta_n)^{-1/2}\right) \quad \textit{and} \quad ||\hat{\beta}_n - \beta_0|| = O_p\left(n^{-1/2}\right).$$

Theorem 3. *Under conditions $\mathcal{A}_1 - \mathcal{A}_7$ and $\mathcal{C}_2$, we have that*

$$P(\hat{\alpha}_n^{\circ} = 0) \to 1 \qquad \textit{and} \qquad P(\hat{\beta}_n^{\circ} = 0) \to 1. \tag{1}$$

From Theorem 2, we can conclude that the estimator $\hat{\theta}_n$ is consistent.

Theorem 3 says us that all the estimates of the zero parameters are correctly set equal to zero with probability tending to 1

Idea of the proof of Theorem 2

One as to prove the existence of a consistent local minimizer; this is implied by that fact that for an arbitrarily small $\varepsilon > 0$, there exists a sufficiently large constant C, such that

$$\lim_{n\to\infty} P\left\{\inf_{z\in\mathbb{R}^{p+q}:||z||=C} \mathcal{F}(\theta_0 + \varphi(n)^{1/2}z) > \mathcal{F}(\theta_0)\right\} > 1-\varepsilon, \tag{2}$$

with $z = (u, v)' = (u_1, ..., u_p, v_1, ..., v_q)'$. After some calculations, we obtain that

$$\mathcal{F}(\theta_0 + \varphi(n)^{1/2}z) - \mathcal{F}(\theta_0)$$

$$\geq z'\varphi(n)^{1/2}\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)\varphi(n)^{1/2}z + 2z'\varphi(n)^{1/2}\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)\varphi(n)^{1/2}\varphi(n)^{-1/2}(\theta_0 - \tilde{\theta}_n)$$

$$- \left[p_0\frac{\mu_n}{\sqrt{n\Delta_n}}||u|| + q_0\frac{\nu_n}{\sqrt{n}}||v||\right]$$

$$= \Xi_1 + \Xi_2 - \Xi_3$$

Idea of the proof of Theorem 2

Let $\tau_{min}(A)$ is the minimal eigenvalue of A. Then, Lemma 1, being $||z|| = C$, Ξ_1 is uniformly larger than $\tau_{min}(\varphi(n)^{1/2}\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)\varphi(n)^{1/2})C^2$ and

$$\tau_{min}(\varphi(n)^{1/2}\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)\varphi(n)^{1/2})C^2 \xrightarrow{p} C^2\tau_{min}(\mathcal{I}(\theta_0)).$$

Moreover, Lemma 1 also implies that

$$||\varphi(n)^{1/2}\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)\varphi(n)^{1/2}\varphi(n)^{-1/2}(\theta_0 - \tilde{\theta}_n)|| = O_p(1)$$

and then Ξ_2 is bounded and linearly dependent on C.

Therefore, for C sufficiently large, $\mathcal{F}(\theta_0 + \varphi(n)^{1/2}z) - \mathcal{F}(\theta_0)$ dominates $\Xi_1 + \Xi_2$ with arbitrarily large probability. Further, from the condition $\mathcal{C}_1$, one has that $\Xi_3 = o_p(1)$.

Strict convexity of $\mathcal{F}(\theta)$ implies that the consistent local minimum is the consistent global one.

Let $\mathcal{I}_0(\theta_0^*)$ the $(p_0 + q_0) \times (p_0 + q_0)$ submatrix of $\mathcal{I}(\theta)$ at point θ_0^* and introduce the following rate of convergence matrix

$$\varphi_0(n) = \begin{pmatrix} \frac{1}{n\Delta_n}\mathbf{I}_{p_0} & 0 \\ 0 & \frac{1}{n}\mathbf{I}_{q_0} \end{pmatrix}$$

Theorem 4 (Oracle property)**.** *Under conditions $\mathcal{A}_1 - \mathcal{A}_7$ and $\mathcal{C}_1 - \mathcal{C}_2$, we have that*

$$\varphi_0(n)^{-\frac{1}{2}}(\hat{\theta}_n^* - \theta_0^*) \xrightarrow{d} N(0, \mathcal{I}_0^{-1}(\theta_0^*)) \tag{3}$$

where θ_0^* is the subset of non-zero true parameters.

How to choose the adaptive sequences

Clearly, the theoretical and practical implications of our method rely to the specification of the tuning parameter $\lambda_{n,j}$ and $\gamma_{n,k}$.

The tuning parameters should be chosen as is Zou (2006) in the following way

$$\lambda_{n,j} = \lambda_0 |\tilde{\alpha}_{n,j}|^{-\delta_1}, \qquad \gamma_{n,k} = \gamma_0 |\tilde{\beta}_{n,j}|^{-\delta_2} \tag{4}$$

where $\tilde{\alpha}_{n,j}$ and $\tilde{\beta}_{n,k}$ are the unpenalized QML estimator of α_j and β_k respectively, $\delta_1, \delta_2 > 0$ and usually taken unitary. The asymptotic results hold under the additional conditions

$$\frac{\lambda_0}{\sqrt{n\Delta_n}} \to 0, \quad (n\Delta_n)^{\frac{\delta_1 - 1}{2}} \lambda_0 \to \infty$$

and

$$\frac{\gamma_0}{\sqrt{n}} \to 0, \quad n^{\frac{\delta_2 - 1}{2}} \gamma_0 \to \infty$$

as $n \to \infty$.

The “standard” approach adopted by Wang and Leng (2007) also holds when the diffusion process has the *same* parametric vector θ in both drift and diffusion coefficients.

In this context, we use the following objective function

$$(\theta - \tilde{\theta}_n)' \ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)(\theta - \tilde{\theta}_n) + \sum_{j=1}^{p} \lambda_{n,j} |\theta_j|,$$

where $\mathbb{H}_n$ can represent the quasi-likelihood function as well as an alternative contrast function (see, e.g., Aït-Sahalia, 2002, and Kessler and Sorensen 1999).

In order to establish the properties of the LASSO estimator, we have to consider a slightly different hypotheses and asymptotic setting, for example the mesh $\Delta_n = \Delta$ is fixed and $n \to \infty$.

Numerical Evidence

To show the oracle properties of the lasso, we consider the following 1-dimensional SDE

$$\mathrm{d}X_t = (\theta_1 - \theta_2 X_t)\mathrm{d}t + (\theta_3 + \theta_4 X_t)^{\theta_5}\mathrm{d}W_t, \quad X_0 = 1$$

We simulate 1000 trajectories of this process with true parameter vector $\theta = (\theta_1 = 1, \theta_2 = 0.1, \theta_3 = 0, \theta_4 = 2, \theta_5 = 0.5)$

In order to get as close as possible to the asymptotic scheme of this talk, we consider the following simulation setup: for a given number n of observations, we set $T = n^{\frac{1}{3}}$ (time horizon) and $\Delta_n = T/n$.

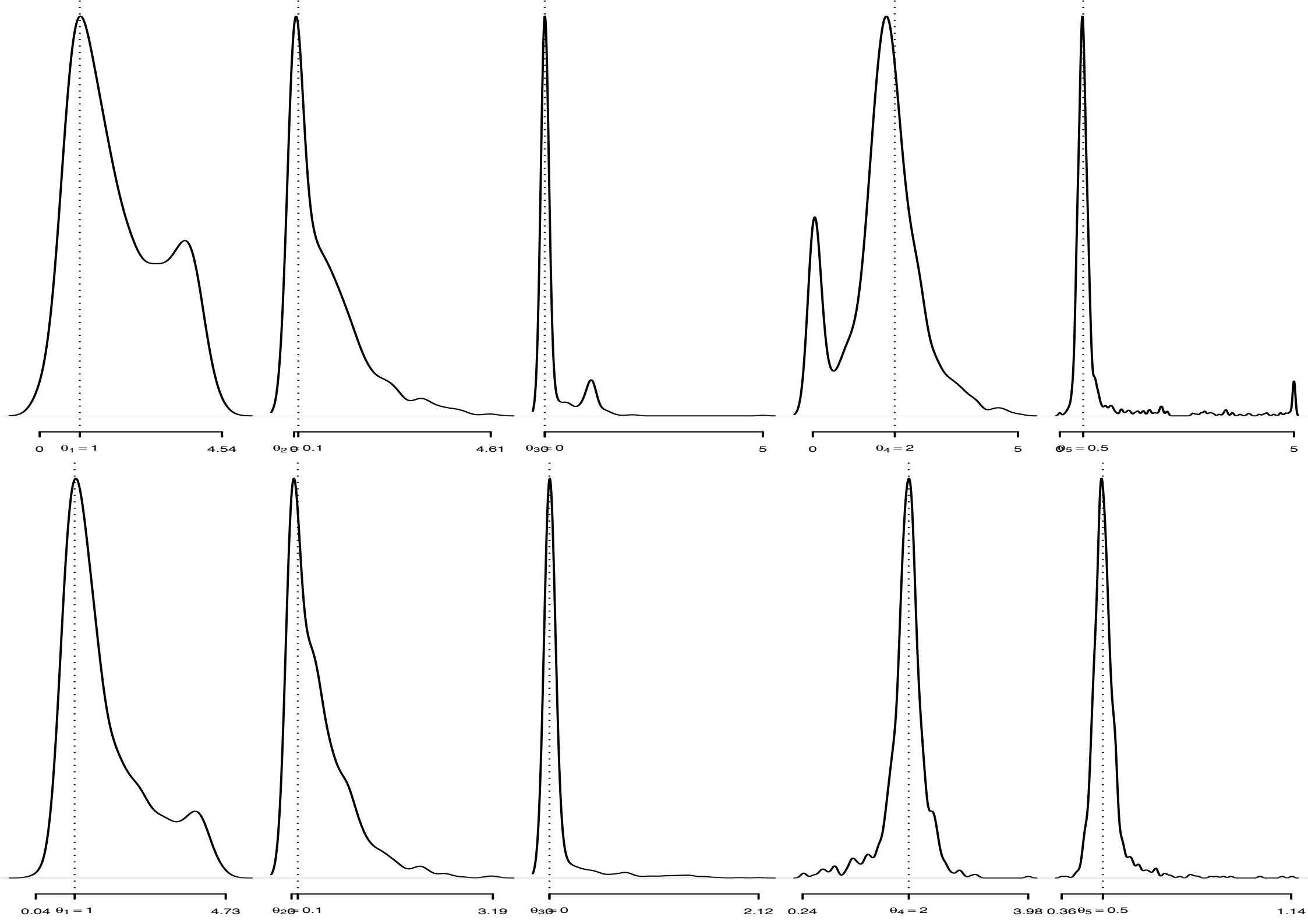
Then we take $n = 100$ and obtain $\Delta_n = 0.046$, while for $n = 1000$, we have that $\Delta_n = 0.01$.

We simulate 1000 trajectories of this process according to the second Milstein scheme

$$X_{t_{i+1}} = X_{t_i} + \left(b - \frac{1}{2}\sigma\sigma_x\right)\Delta_n + \sigma Z\sqrt{\Delta_n} + \frac{1}{2}\sigma\sigma_x\Delta_n Z^2 + \Delta_n^{\frac{3}{2}}\left(\frac{1}{2}b\sigma_x + \frac{1}{2}b_x\sigma + \frac{1}{4}\sigma^2\sigma_{xx}\right)Z + \Delta_n^2\left(\frac{1}{2}bb_x + \frac{1}{4}b_{xx}\sigma^2\right)$$

with $Z \sim N(0,1)$, b_x and b_{xx} (resp. σ_x and σ_{xx}) are the first and second partial derivative in x of the drift (resp. diffusion) coefficient. This scheme has weak second-order convergence and guarantees good numerical stability (see, Milstein, 1978)

Next plot shows the oracle property as n increases from $n = 100$ (up) to $n = 1000$ (bottom)


0
$\theta_1 = 1$
4.54
$\theta_2 = 0.1$
4.61
$\theta_3 = 0$
5
0
$\theta_4 = 2$
5
$\theta_5 = 0.5$
5
0.04
$\theta_1 = 1$
4.73
$\theta_2 = 0.1$
3.19
$\theta_3 = 0$
2.12
0.24
$\theta_4 = 2$
3.98
0.36
$\theta_5 = 0.5$
1.14

Results

	θ_1	θ_2	θ_3	θ_4	θ_5	% $\theta_3 = 0$
True	1.0	0.1	0.0	2.0	0.5	
Qmle: $n = 100$	2.58	1.04	0.27	1.89	0.75	
	(1.47)	(0.91)	(0.57)	(1.10)	(0.87)	
Lasso: $\lambda_0 = \gamma_0 = 1, n = 100$	1.92	0.69	0.17	1.69	0.78	78%
	(1.10)	(0.84)	(0.41)	(0.92)	(0.93)	
Lasso: $\lambda_0 = \gamma_0 = 5, n = 100$	0.70	0.11	0.14	1.30	0.79	87%
	(0.56)	(0.38)	(0.37)	(0.80)	(0.96)	
Qmle: $n = 1000$	2.07	0.56	0.11	1.90	0.52	
	(1.25)	(0.52)	(0.27)	(0.37)	(0.06)	
Lasso: $\lambda_0 = \gamma_0 = 1, n = 1000$	1.74	0.42	0.07	1.94	0.51	84%
	(1.01)	(0.49)	(0.25)	(0.35)	(0.06)	
Lasso: $\lambda_0 = \gamma_0 = 5, n = 1000$	0.93	0.11	0.05	1.94	0.51	91%
	(0.47)	(0.29)	(0.22)	(0.33)	(0.08)	

Monte Carlo standard errors in parentheses; 1000 Monte Carlo replications for each sample size

A multidimensional example

We consider this two dimensional geometric Brownian motion process solution to the stochastic differential equation

$$\begin{pmatrix} \mathrm{d}X_t \\ \mathrm{d}Y_t \end{pmatrix} = \begin{pmatrix} 1 - \mu_{11} X_t + \mu_{12} Y_t \\ 2 + \mu_{21} X_t - \mu_{22} Y_t \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \sigma_{11} X_t & -\sigma_{12} Y_t \\ \sigma_{21} X_t & \sigma_{22} Y_t \end{pmatrix} \begin{pmatrix} \mathrm{d}W_t \\ \mathrm{d}B_t \end{pmatrix}$$

with initial condition $(X_0 = 1, Y_0 = 1)$ and $W_t, t \in [0, T]$, and $B_t, t \in [0, T]$, are two independent Brownian motions.

This model is a classical model for pricing of basket options in mathematical finance.

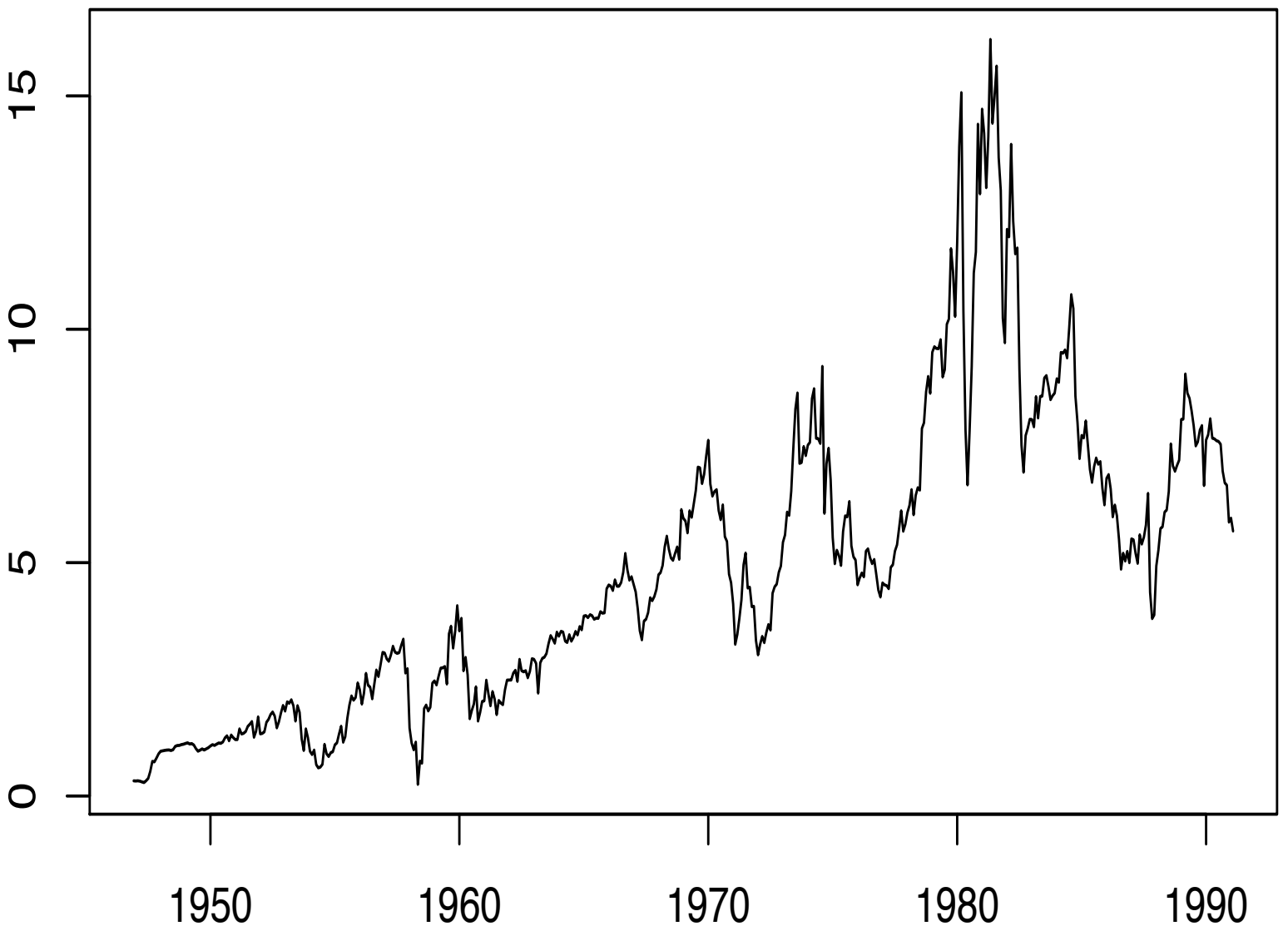
We assume that $\alpha = (\mu_{11} = 0.9, \mu_{12} = 0, \mu_{21} = 0, \mu_{22} = 0.7)'$ and $\beta = (\sigma_{11} = 0.3, \sigma_{12} = 0, \sigma_{21} = 0, \sigma_{22} = 0.2)'$, $\theta = (\alpha, \beta)$.

Results

	μ_{11}	μ_{12}	μ_{21}	μ_{22}	σ_{11}	σ_{12}	σ_{21}	σ_{22}
True	0.9	0.0	0.0	0.7	0.3	0.0	0.0	0.2
Qmle: $n = 100$	0.96	0.05	0.25	0.81	0.30	0.04	0.01	0.20
	(0.08)	(0.06)	(0.27)	(0.15)	(0.03)	(0.05)	(0.02)	(0.02)
Lasso: $\lambda_0 = \gamma_0 = 1, n = 100$	0.86	0.00	0.05	0.71	0.30	0.02	0.01	0.20
	(0.12)	(0.00)	(0.13)	(0.09)	(0.03)	(0.05)	(0.02)	(0.02)
% of times $\theta_i = 0$	0.0	99.9	80.2	0.0	0.3	67.2	66.7	0.1
Lasso: $\lambda_0 = \gamma_0 = 5, n = 100$	0.82	0.00	0.00	0.66	0.29	0.01	0.00	0.20
	(0.12)	(0.00)	(0.00)	(0.09)	(0.03)	(0.03)	(0.01)	(0.02)
% of times $\theta_i = 0$	0.0	100.0	99.9	0.0	0.4	86.9	89.7	0.2
Qmle: $n = 1000$	0.95	0.03	0.21	0.79	0.30	0.04	0.01	0.20
	(0.07)	(0.04)	(0.25)	(0.13)	(0.03)	(0.06)	(0.02)	(0.02)
Lasso: $\lambda_0 = \gamma_0 = 1, n = 1000$	0.88	0.00	0.08	0.73	0.30	0.02	0.01	0.20
	(0.08)	(0.00)	(0.16)	(0.09)	(0.03)	(0.05)	(0.01)	(0.02)
% of times $\theta_i = 0$	0.0	99.7	72.1	0.0	0.1	67.5	66.6	0.1
Lasso: $\lambda_0 = \gamma_0 = 5, n = 1000$	0.86	0.00	0.00	0.68	0.29	0.01	0.00	0.20
	(0.09)	(0.00)	(0.01)	(0.06)	(0.03)	(0.04)	(0.01)	(0.02)
% of times $\theta_i = 0$	0.0	100.0	99.4	0.0	0.2	87.8	89.9	0.2

Application to real data

Interest rates LASSO estimation examples



Interest rates LASSO estimation examples

LASSO estimation of the U.S. Interest Rates monthly data from 06/1964 to 12/1989. These data have been analyzed by many author including Nowman (1997), Aït-Sahalia (1996), Yu and Phillips (2001) and it is a nice application of LASSO.

Reference	Model	α	β	γ
Merton (1973)	$\mathrm{d}X_t = \alpha \mathrm{d}t + \sigma \mathrm{d}W_t$		0	0
Vasicek (1977)	$\mathrm{d}X_t = (\alpha + \beta X_t)\mathrm{d}t + \sigma \mathrm{d}W_t$			0
Cox, Ingersoll and Ross (1985)	$\mathrm{d}X_t = (\alpha + \beta X_t)\mathrm{d}t + \sigma \sqrt{X_t} \mathrm{d}W_t$			$1/2$
Dothan (1978)	$\mathrm{d}X_t = \sigma X_t \mathrm{d}W_t$	0	0	1
Geometric Brownian Motion	$\mathrm{d}X_t = \beta X_t \mathrm{d}t + \sigma X_t \mathrm{d}W_t$	0		1
Brennan and Schwartz (1980)	$\mathrm{d}X_t = (\alpha + \beta X_t)\mathrm{d}t + \sigma X_t \mathrm{d}W_t$			1
Cox, Ingersoll and Ross (1980)	$\mathrm{d}X_t = \sigma X_t^{3/2} \mathrm{d}W_t$	0	0	$3/2$
Constant Elasticity Variance	$\mathrm{d}X_t = \beta X_t \mathrm{d}t + \sigma X_t^{\gamma} \mathrm{d}W_t$	0		
CKLS (1992)	$\mathrm{d}X_t = (\alpha + \beta X_t)\mathrm{d}t + \sigma X_t^{\gamma} \mathrm{d}W_t$			

Interest rates LASSO estimation examples

Model	Estimation Method	α	β	σ	γ
Vasicek	MLE	4.1889	-0.6072	0.8096	–
CKLS	Nowman	2.4272	-0.3277	0.1741	1.3610
CKLS	Exact Gaussian (Yu & Phillips)	2.0069 (0.5216)	-0.3330 (0.0677)	0.1741	1.3610
CKLS	QMLE	2.0822 (0.9635)	-0.2756 (0.1895)	0.1322 (0.0253)	1.4392 (0.1018)
CKLS	QMLE + LASSO with mild penalization	1.5435 (0.6813)	-0.1687 (0.1340)	0.1306 (0.0179)	1.4452 (0.0720)
CKLS	QMLE + LASSO with strong penalization	0.5412 (0.2076)	0.0001 (0.0054)	0.1178 (0.0179)	1.4944 (0.0720)

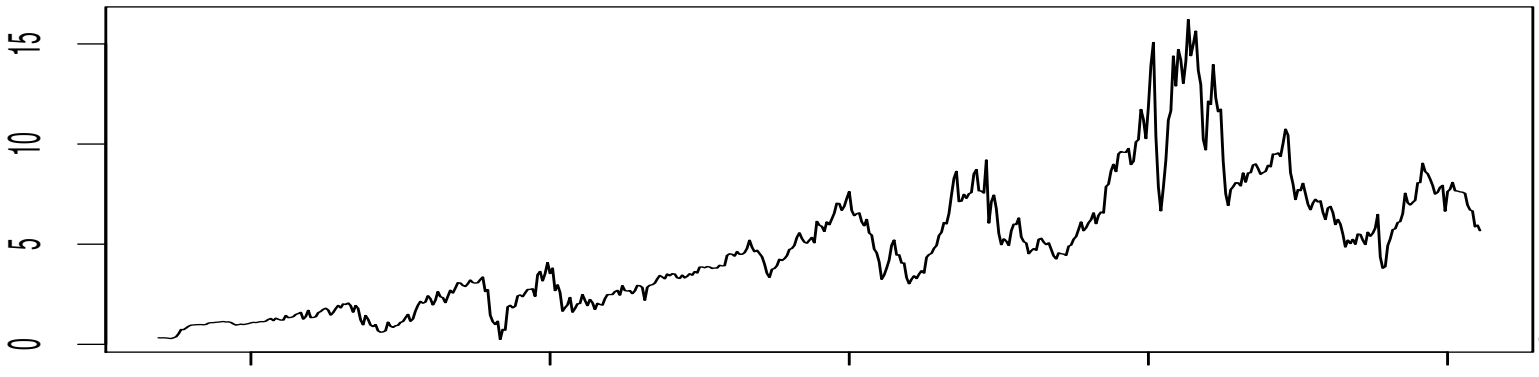
LASSO selected: Cox, Ingersoll and Ross (1980) model

$$\mathrm{d}X_t = \frac{1}{2}\mathrm{d}t + 0.12 \cdot X_t^{3/2}\mathrm{d}W_t$$

An example of Lasso estimation using `yuima` package. We make use of real data with CKLS model

$$\mathrm{d}X_t = (\alpha + \beta X_t)\mathrm{d}t + \sigma X_t^{\gamma}\mathrm{d}W_t$$

```
> library(Ecdat)
> data(Irates)
> rates <- Irates[,"r1"]
> plot(rates)
> require(yuima)
> X <- window(rates, start=1964.471, end=1989.333)
> mod <- setModel(drift="alpha+beta*x", diffusion=matrix("sigma*x^gamma",1,1))
> yuima <- setYuima(data=setData(X), model=mod)
```



Example of Lasso estimation


```
> lambda10 <- list(alpha=10, beta =10, sigma =10, gamma =10)
> start <- list(alpha=1, beta =-.1, sigma =.1, gamma =1)
> low <- list(alpha=-5, beta =-5, sigma =-5, gamma =-5)
> upp <- list(alpha=8, beta =8, sigma =8, gamma =8)
> lasso10 <- lasso(yuima, lambda10, start=start, lower=low, upper=upp,
   method="L-BFGS-B")

Looking for MLE estimates...
Performing LASSO estimation...

> round(lasso10$mle, 3)  # QMLE
 sigma  gamma  alpha   beta
 0.133  1.443  2.076 -0.263

> round(lasso10$lasso, 3)  # LASSO
sigma gamma alpha  beta
0.117 1.503 0.591 0.000
```

$$\mathrm{d}X_t = (\alpha + \beta X_t)\mathrm{d}t + \sigma X_t^{\gamma}\mathrm{d}W_t$$

$$\mathrm{d}X_t = 0.6\mathrm{d}t + 0.12X_t^{\frac{3}{2}}\mathrm{d}W_t$$

References

Aït-Sahalia, Y. (1996) Testing continuous-time models of the spot interest rate, *Rev. Financial Stud.*, **9**(2), 385–426.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle, *2nd International Symposium in Information Theory*, Petrov, B.N., Csaki, F., eds., Akademiai Kaido, Budapest, 267–281.

Akaike, H. (1974) A new look at the statistical model identification, *IEEE Trans. Autom. Control*, **AC-19**, 716–723.

Dacunha-Castelle, D., Florens-Zmirou, D. (1986) Estimation of the coefficients of a diffusion from discrete observations, *Stochastics*, **19**, 263–284.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004) Least angle regression, *The Annals of Statistics*, **32**, 407–489.

Knight, K., Fu, W. (2000) Asymptotics for lasso-type estimators, *Annals of Statistics*, **28**, 1536–1378.

Nowman, K. (1997) Gaussian estimation of single-factor continuous time models of the term structure of interest rates, *Journal of Finance*, **52**, 1695–1703.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso, *J. Roy. Statist. Soc. Ser. B*, **58**, 267–288.

Uchida, M., Yoshida, N. (2005) AIC for ergodic diffusion processes from discrete observations, preprint MHF 2005-12, March 2005, Faculty of Mathematics, Kyushu University, Fukuoka, Japan.

Yu, J., Phillips, P.C.B. (2001) Gaussian estimation of continuous time models of the short term interest rate, *Cowles Foundation Discussion Paper*, n. 1309. Available at `cowles.econ.yale.edu/P/cd/d13a/d1309.pdf`

Zou, H. (2006) The adaptive LASSO and its Oracle properties, *J. Amer. Stat. Assoc.*, **101**(476), 1418-1429.