Model Selection and Lasso for SDE models

Stefano M. Iacus (University of Milan)

Third YUIMA Workshop @ Brixen-Bressanone, 28-06-2019

Model selection

Idea

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 **true** 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

Uchida and Yoshida (2005) develop the AIC statistics defined as

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

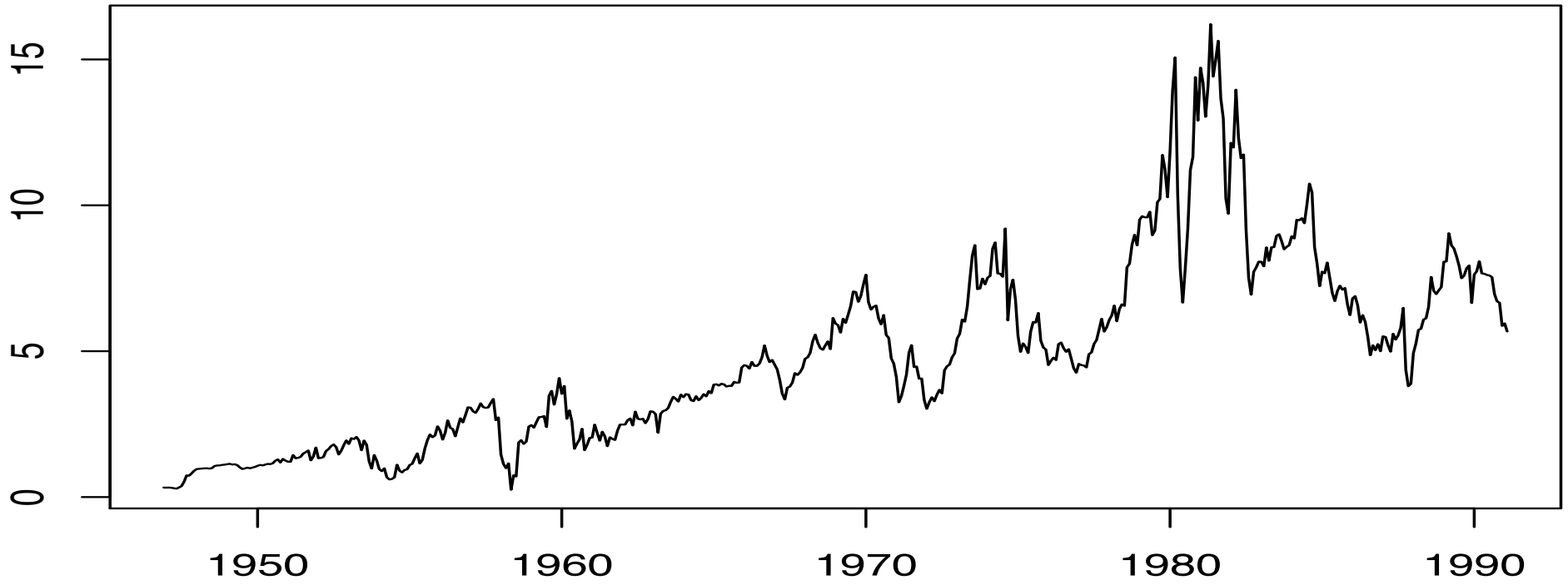
where $\hat{\theta}_n^{(QML)}$ is the quasi maximum likelihood estimator and $\tilde{\ell}_n$ the local Gaussian approximation of the true log-likelihood.

Here we proceed with the QLA approach as seen in the previous slides. It it then important to see what is the effect of the two different approximations on the results.

Exact vs quasi-likelihood analysis

The U.S. Interest Rates monthly data from 06/1964 to 12/1989

```
R> library(Ecdat)
R> library(sde)
R> data(Irates)
R> X <- Irates[,"r1"]
R> plot(X)
```



$$\mathrm{d}X_t = (\theta_1 - \theta_2 X_t)\mathrm{d}t + \theta_3\sqrt{X_t}\mathrm{d}W_t, \quad X_0 = x_0 > 0\,,$$

where $\theta_1, \theta_2, \theta_3 \in \mathbb{R}_+$. If $2\theta_1 > \theta_3^2$, the process is strictly positive; otherwise it is only nonnegative.The transition density $p_\theta(t, \cdot|x)$ follows a non-central χ^2 distribution,

$$p_\theta(t, y|x) = ce^{-u-v}\left(\frac{u}{v}\right)^{q/2} I_q(2\sqrt{uv}), \qquad x, y \in \mathbb{R}_+\,,$$

where

$$c = \frac{2\theta_2}{\theta_3^2(1 - e^{-\theta_2 t})}\,, \qquad q = \frac{2\theta_1}{\theta_3^2} - 1\,,$$
$$u = cxe^{-\theta_2 t}, \qquad v = cy\,.$$

Here $I_q(\cdot)$ is the modified Bessel function of the first kind of order q and $\Gamma(\cdot)$ is the Gamma function.

Optimization of the CIR likelihood

For CIR, MLE estimators do not exist in explicit form.

Numerical optimization is far from being easy because of the Bessel functions

$$I_q(x) = \sum_{k=0}^{\infty} \left(\frac{x}{2}\right)^{2k+q} \frac{1}{k!\Gamma(k+q+1)}, \quad x \in \mathbb{R},$$

Indeed, for real data (e.g. interest rate models) the parameters are in a region for which numerical method are not well behaved.

A number of papers deal with this purely numerical problem. The `sde` package implements some tricks for the different regions of the parameters' space. But it's only for 1-dimensional models.

CIR model

We try to fit the CIR model using the exact maximum likelihood estimation through the `sde` package

$$\mathrm{d}X_t = (\theta_1 - \theta_2 X_t)\mathrm{d}t + \theta_3\sqrt{X_t}\mathrm{d}B_t$$

```
R> CIR.loglik <- function(theta1,theta2,theta3) {
+  n <- length(X)
+  dt <- deltat(X)
+  -sum(dcCIR(x=X[-1], Dt=dt, x0=X[-n], theta=c(theta1,theta2,theta3),
+    log=TRUE))
+ }
R>
R> fit <-  mle(CIR.loglik, start=list(theta1=.1,  theta2=.1,theta3=.3),
+   method="L-BFGS-B",lower=rep(1e-3,3), upper=rep(1,3))
R> coef(fit)
   theta1    theta2    theta3
0.9194592 0.1654958 0.8255179
```

CIR model

We try to fit the CIR model using the quasi maximum likelihood using `yuima`

$$\mathrm{d}X_t = (\theta_1 - \theta_2 X_t)\mathrm{d}t + \theta_3 \sqrt{X_t}\mathrm{d}B_t$$

```
R> start <- list(theta1=1, theta2 =.1, theta3 =.3)
R> low <- list(theta1=1e-3, theta2 =1e-3, theta3 =1e-3)
R> upp <- list(theta1=3, theta2 =3, theta3 =3)
R> mod <- setModel(drift = "theta1-theta2*x", diffusion = "theta3*sqrt(x)")
R> yuima <- setYuima(data=setData(X), model=mod)
R> fit2 <- qmle(yuima,  start=start, lower=low, upper=upp,
+    method="L-BFGS-B")
R> coef(fit2)[names(coef(fit))] # QMLE
   theta1    theta2    theta3
0.8555436 0.1524043 0.8154104
```

compared to exact maximum likelihood estimation

```
R> coef(fit) # EXACT MLE
   theta1    theta2    theta3
0.9194592 0.1654958 0.8255179
```

CKLS model

We try to fit the CLKS model using the quasi maximum likelihood using `yuima`

$$\mathrm{d}X_t = (\theta_1 - \theta_2 X_t)\mathrm{d}t + \theta_3 {X_t}^{\theta_4}\mathrm{d}B_t$$

```
R> mod2 <- setModel(drift="theta1-theta2*x", diffusion=matrix("theta3*x^theta4",1,1))
R> start <- list(theta1=1, theta2 =.1, theta3 =.3, theta4=0.5)
R> low <- list(theta1=1e-3, theta2 =1e-3, theta3 =1e-3, theta4=.1)
R> upp <- list(theta1=3, theta2 =3, theta3 =3, theta4=2)
R> yuima <- setYuima(data=setData(X), model=mod2)
R> fit3 <- qmle(yuima,  start=start, lower=low, upper=upp,
+    method="L-BFGS-B")
```

the quasi-maximum likelihood estimation

```
R> coef(fit3)[sort(names(coef(fit3)))] # QMLE
   theta1    theta2    theta3    theta4
0.8863715 0.1591604 0.7151518 0.5929715
```

but which is the true model for these data?

Model selection in practice

A trivial example

We compare three models

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

We call the above models Mod1, Mod2 and Mod3.

We generate data from Mod1 with parameters

$$\mathrm{d}X_t = (10 - X_t)\mathrm{d}t + 0.3\sqrt{X_t}\mathrm{d}W_t\,,$$

and initial value $X_0 = 8$. We use $n = 1000$ and $\Delta = 0.1$, therefore $T = n\Delta = 100$.

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

Simulation results. 100 Monte Carlo replications

$$\mathrm{d}X_t = 1(10 - X_t)\mathrm{d}t + 0.3\sqrt{X_t}\mathrm{d}W_t \qquad \text{(true model)},$$

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

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

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

Model selection via AIC

Model 1 (true)	Model 2	Model 3
86 %	14 %	0 %

QMLE estimates under the different models

	$\alpha_1 = 1$	$\alpha_2 = 10$	$\beta_1 = 0.3$	β_2	β_3
Model 1	1.070	10.001	0.322		
Model 2	1.069	10.001	0.995	0.003	
Model 3	1.069	10.001	2.827	5.015	0.030

Simulation results. 100 Monte Carlo replications

Same analysis with $\Delta = 0.01$, but in this case $T = n\Delta = 10$, hence we loose performance in the estimation of the drift parameters

Model selection via AIC

Model 1 (true)	Model 2	Model 3
87 %	11 %	2 %

QMLE estimates under the different models

	$\alpha_1 = 1$	$\alpha_2 = 10$	$\beta_1 = 0.3$	β_2	β_3
Model 1	1.413	9.970	0.303		
Model 2	1.415	9.970	0.358	0.055	
Model 3	1.418	9.969	0.083	0.018	0.165

Sparse Estimation

Lasso: Least Absolute Selection and Shrinkage Operator

Lasso estimates (see Tibshirani, 1996; Knight and Fu, 2000, Efron *et al.*, 2004) minimize

$$RSS + \lambda \sum_{j=1}^{k} |\beta_j|.$$

The important difference with ridge regression is in the penalty part (l_1 vs l_2). This seemingly tiny difference makes qualitative gaps practically as well as theoretically.

The l_1 penalty causes some coefficients to be shrunken exactly to zero, i.e., the predictive model is **sparse**

Lasso performs both variable selection **and** shrinkage

The previous Lasso approach can be generalized further to l_q constraints (**Bridge** estimation), for some $q > 0$, i.e.

$$\hat{\beta} = \arg\min_{\beta} RSS + \lambda \sum_{i=1}^{k} |\beta_i|^q$$

Where Lasso is for $q = 1$, Ridge is for $q = 2$ and the limiting case $q = 0$ is OLS.

Notice that, in the limit as $q \to 0$, this procedure approximates AIC/BIC criteria as

$$\lim_{q \to 0} \sum_{i=1}^{k} |\beta_i|^q = \sum_{i=1}^{k} \mathbf{1}_{\{\beta_i \neq 0\}}$$

as the RHS amounts to the number of non-null parameters.

Adaptive Estimation

Oracle procedures

Let $\mathcal{A} = \{j : \beta_j \neq 0\}$ be the set of true non-zero coefficients in the standard regression model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

such that $|\mathcal{A}| = p_0 < p$. Denote by $\hat{\beta}(\delta)$ the estimates of an estimation procedure δ. Following Fan and Li (2001), we call δ an **oracle procedure** if $\hat{\beta}(\delta)$ (asymptotically) has the following oracle properties:

- Identifies the right subset model, $\{j : \hat{\beta}_j \neq 0\} = \mathcal{A}$
- Has the optimal estimation rate $\sqrt{n}(\hat{\beta}(\delta)_{\mathcal{A}} - \beta_{\mathcal{A}})$ converges in distribution to $N(0, \Sigma^*)$ where where Σ^* is the covariance matrix of the true subset/reduced model.

Remind that if all coefficients are non-zero, the MLE estimator satisfies

$$\sqrt{n}(\hat{\beta}^{ML} - \beta) \xrightarrow{d} N(0, \mathcal{I}^{-1}(\beta) =: \Sigma^*)$$

In the classic Lasso procedure, the main assumption are that

$$\frac{1}{n}X'X \to C$$

where C is A positive definite matrix. Let us re-order the coefficients β so that the true non-zero coefficients occupy the first positions $1, \ldots, p_0$. Then let

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

where C_{11} is $p_0 \times p_0$. Now let $\lambda = \lambda_n$ in the Lasso penalty function

$$\hat{\beta}_n = \arg\min_{\beta} \left(RSS + \lambda_n \sum_{j=1}^{p} |\beta_j| \right)$$

Lasso is not an Oracle procedure!

If λ_n is such that $\lim_{n\to\infty} \lambda_n/n = \lambda_0 \geq 0$, then Lemma 1 (Knight and Fu, 2000):

$$\hat{\beta}_n \xrightarrow{p} \arg\min_{\beta} V_1, \quad \text{with} \quad V_1(u) = (u-\beta)'C(u-\beta) + \lambda_0 \sum_{j=1}^{p} |u_j|$$

and if $\lim_{n\to\infty} \lambda_n/\sqrt{n} = \lambda_0 \geq 0$ then, Lemma 2 (Knight and Fu, 2000):

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \arg\min_{\beta} V_2$$

with

$$V_2(u) = -2u'W + u'Cu + \lambda_0 \sum_{j=1}^{p} \left(u_j \text{sign}(\beta_j) I_{\{\beta_j \neq 0\}} + |u_j| I_{\{\beta_j = 0\}} \right)$$

with $W = N(0, \sigma^2 C)$.

Lasso is not an Oracle procedure!

Lemma 1 shows that only if $\lambda_0 = 0$ the Lasso estimators are consistent.

Lemma 2 shows that Lasso can be $\sqrt{n}$-consistent under the same conditions. But in general bias remains.

Indeed, it is also possible to prove that

$$\lim_{n \to \infty} P(\mathcal{A}_n = \mathcal{A}) \leq c < 1$$

which means that the true set of non-zero coefficients is not correctly identified even asymptotically.

Adaptive Lasso addresses this problem.

Adaptive Lasso is an Oracle procedure!

Let $\tilde{\beta}$ be a $\sqrt{n}$-consistent estimator of β (e.g. OLS or MLE). Let $\gamma > 0$ and define $\tilde{w}_j = 1/|\tilde{\beta}|^\gamma, j = 1, \ldots, p$. The adaptive Lasso estimator is defined as follows

$$\hat{\beta} = \arg\min_{\beta} \left(RSS + \lambda_n \sum_{j=1}^{p} \tilde{w}_j |\beta_j| \right)$$

If $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{\frac{\gamma-1}{2}} \to \infty$, then (Zou, 2006), we have the **oracle** properties:

- consistent variable selection: $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$
- asymptotic normality: $\sqrt{n}(\hat{\beta}_{\mathcal{A}} - \beta_{\mathcal{A}}) \xrightarrow{d} N\left(0, \sigma^2 C_{11}^{-1}\right)$.

Application to SDEs

Let X_t be a 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.

Sampling scheme

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.

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 Lasoo 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

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

The `yuima` package implements the above optimization problem.

Adaptive Lasso properties

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 $\omega_n = \min\{\gamma_{n,k}, k > q_0\}$.

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.

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 1. *Under conditions standard regularity conditions 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 2. *Under conditions standard regularity conditions 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 1, we can conclude that the estimator $\hat{\theta}_n$ is consistent.

Theorem 2 says us that all the estimates of the zero parameters are correctly set equal to zero with probability tending to 1

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

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

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

$$\lambda_{n,j} = \lambda_0 |\tilde{\alpha}_{n,j}|^{-\delta_1}, \qquad \gamma_{n,k} = \gamma_0 |\tilde{\beta}_{n,j}|^{-\delta_2} \tag{3}$$

where $\tilde{\alpha}_{n,j}$ and $\tilde{\beta}_{n,k}$ are the unpenalized QML estimator of α_j and β_k respectively, $\delta_1, \delta_2 > 1$. 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 $n \to \infty$.

Numerical evidence of oracle properties

Oracle property

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 $\Delta_n = T/n$.

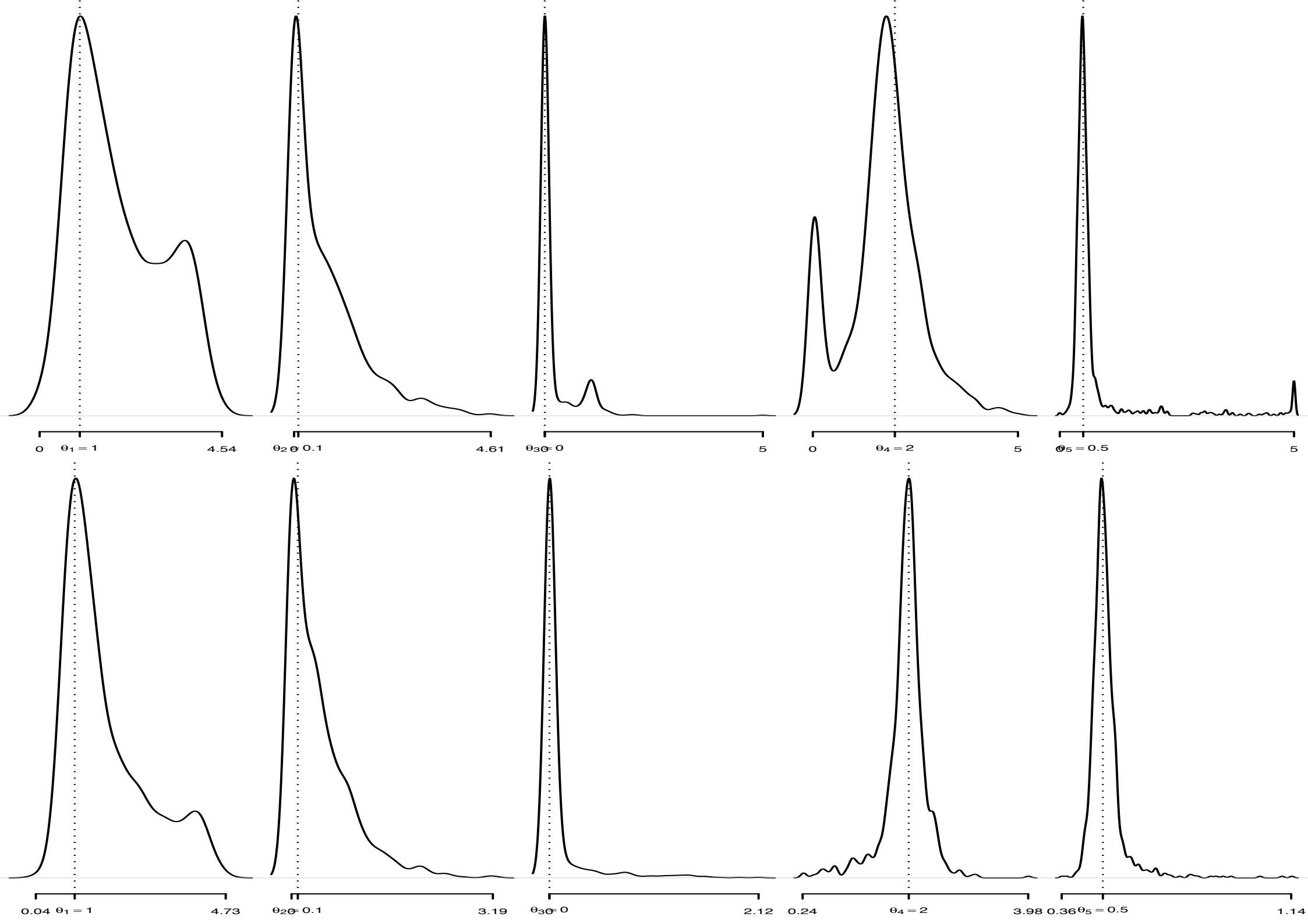
Then we take $n = 100$ and obtain $\Delta_n = 0.046$, while for $n = 1000$, we have that $\Delta_n = 0.01$.

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
θ1 = 1
4.54
θ2 = 0.1
4.61
θ3 = 0
5
0
θ4 = 2
5
θ5 = 0.5
5
0.04
θ1 = 1
4.73
θ2 = 0.1
3.19
θ3 = 0
2.12
0.24
θ4 = 2
3.98
0.36
θ5 = 0.5
1.14

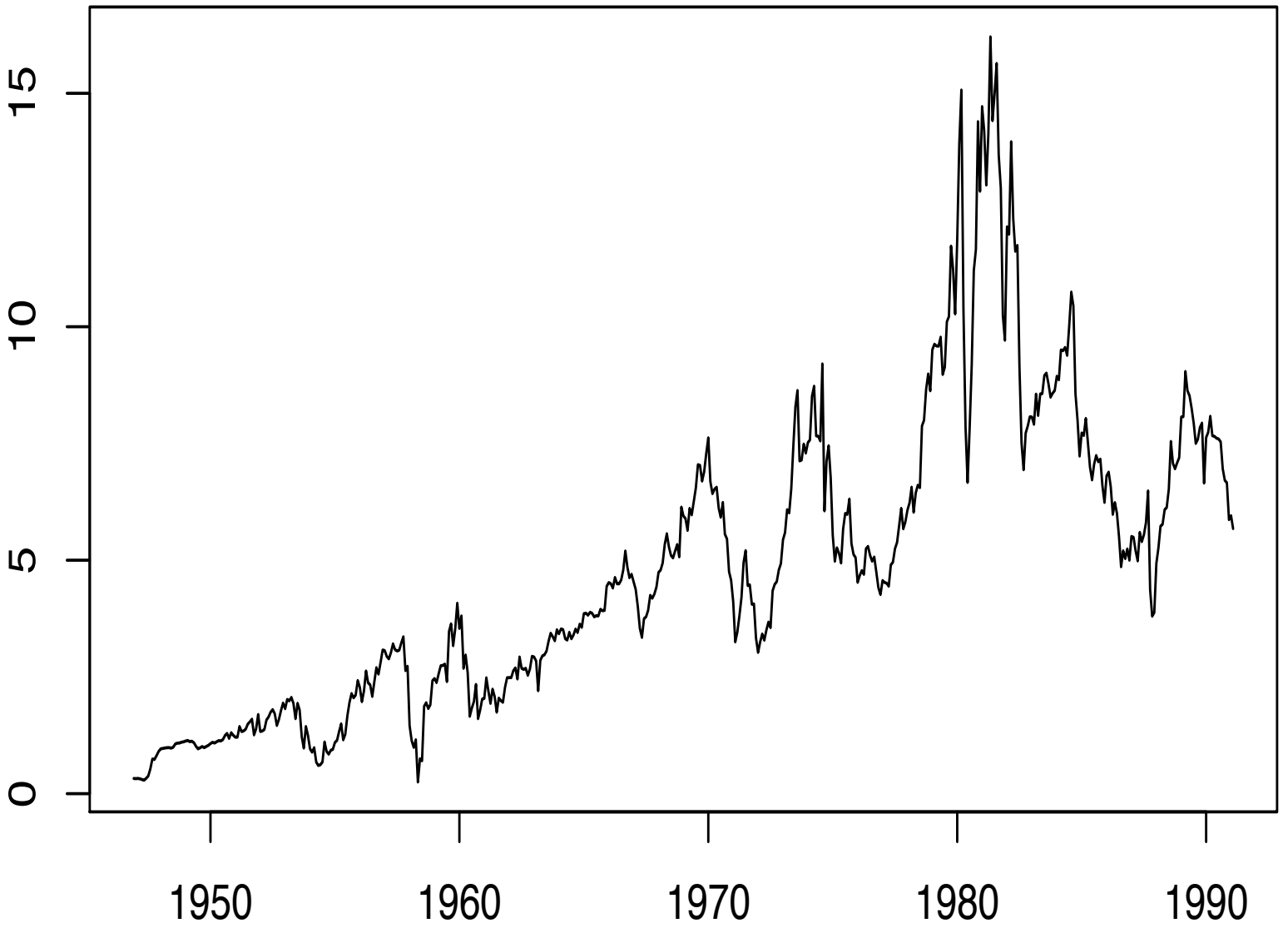
Results

	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

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)

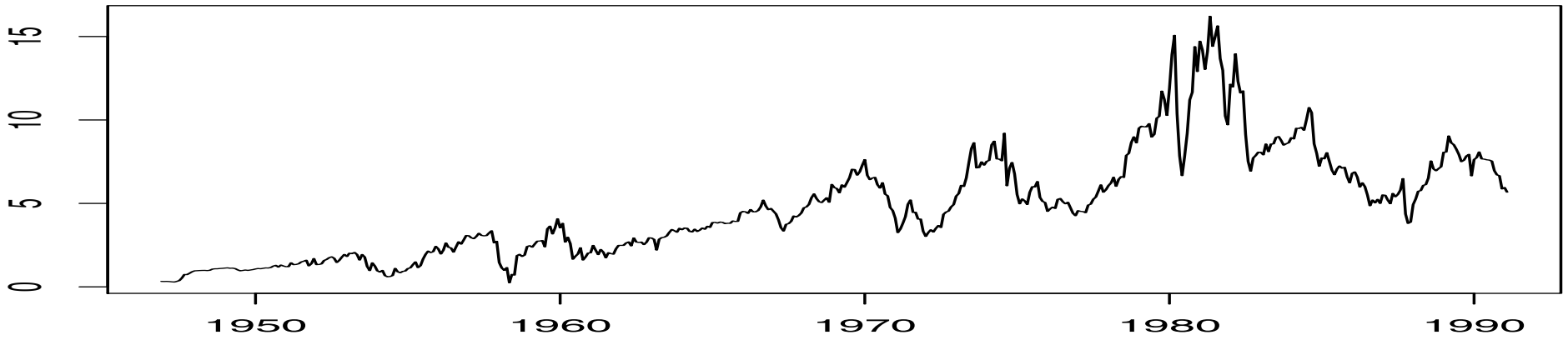
LASSO selected: Cox, Ingersoll and Ross (1980) model

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

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

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

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



Example of Lasso estimation

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

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

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

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

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

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

Sparsity and robustness in forecasting

Example: Lasso and QMLE in forecasting true data

We now try to show, through an experiment why a Lasso (or a regularized solution) is preferable to the full MLE solution when the target is the forecasting.

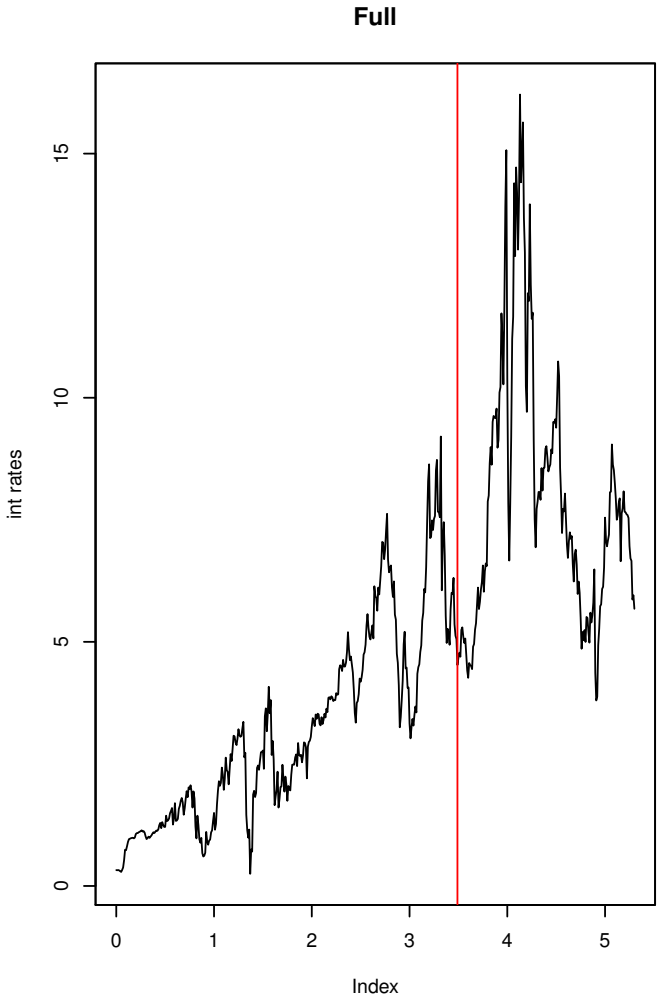
The experiment consists in trying to estimate the QMLE and LASSO solution on a subset of 350 observations from the previous real data consisting of a total of 531 observations. The remaining 181 observations are used as training set.

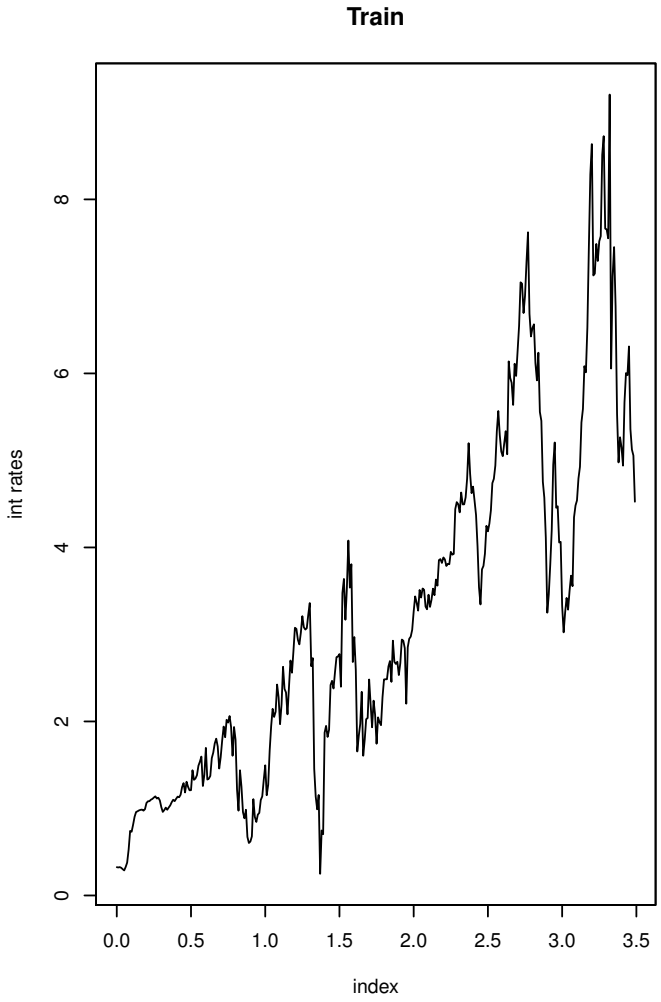
Once the QMLE and LASSO estimates are available, a number of simulations is run and then the forecasting MSE is calculated for each simulated trajectory, i.e.

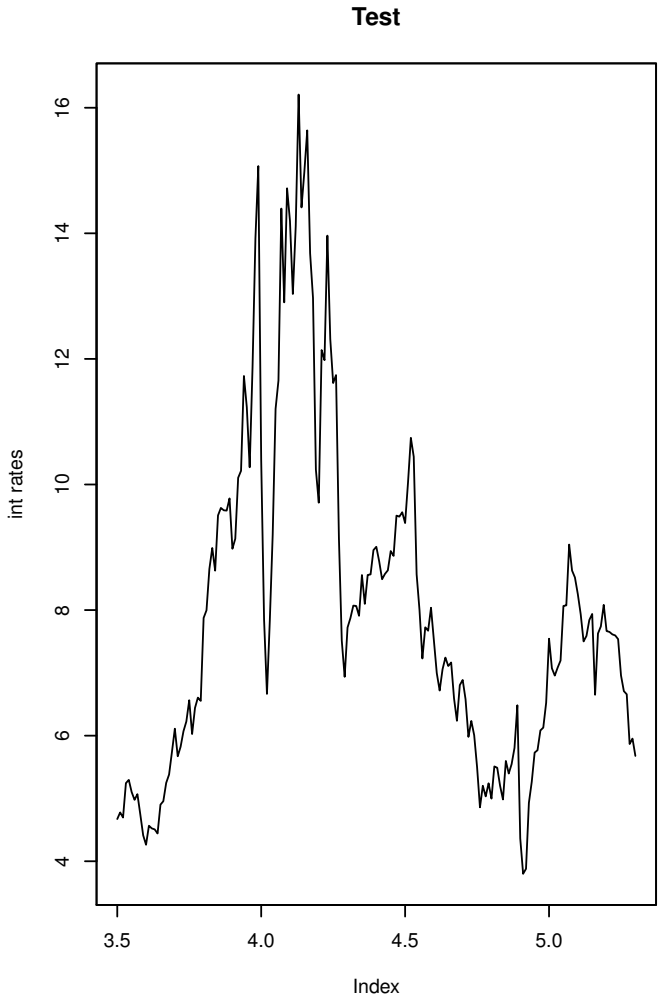
$$\text{fMSE}(LASSO) = \frac{1}{181} \sum_{i=351}^{531} \left(X_j^{LASSO} - X_j^{TRUE} \right)^2$$

$$\text{fMSE}(QMLE) = \frac{1}{181} \sum_{i=351}^{531} \left(X_j^{QMLE} - X_j^{TRUE} \right)^2$$

Example: Lasso and QMLE in forecasting true data







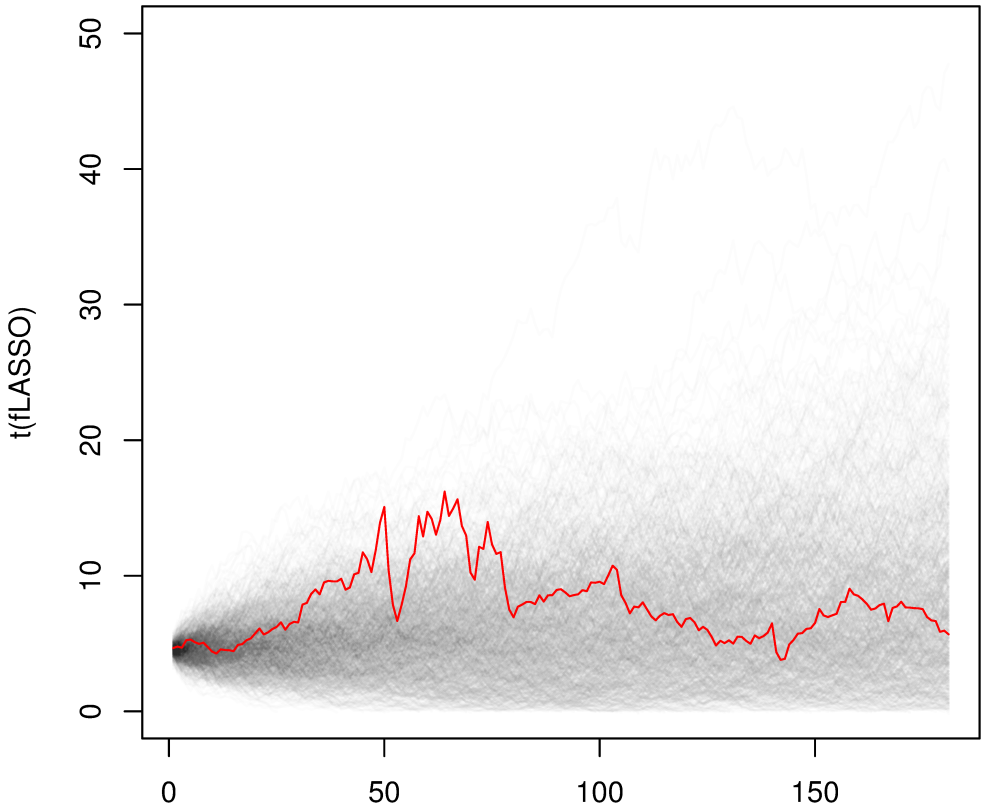
Example: simulation study with 1000 replications

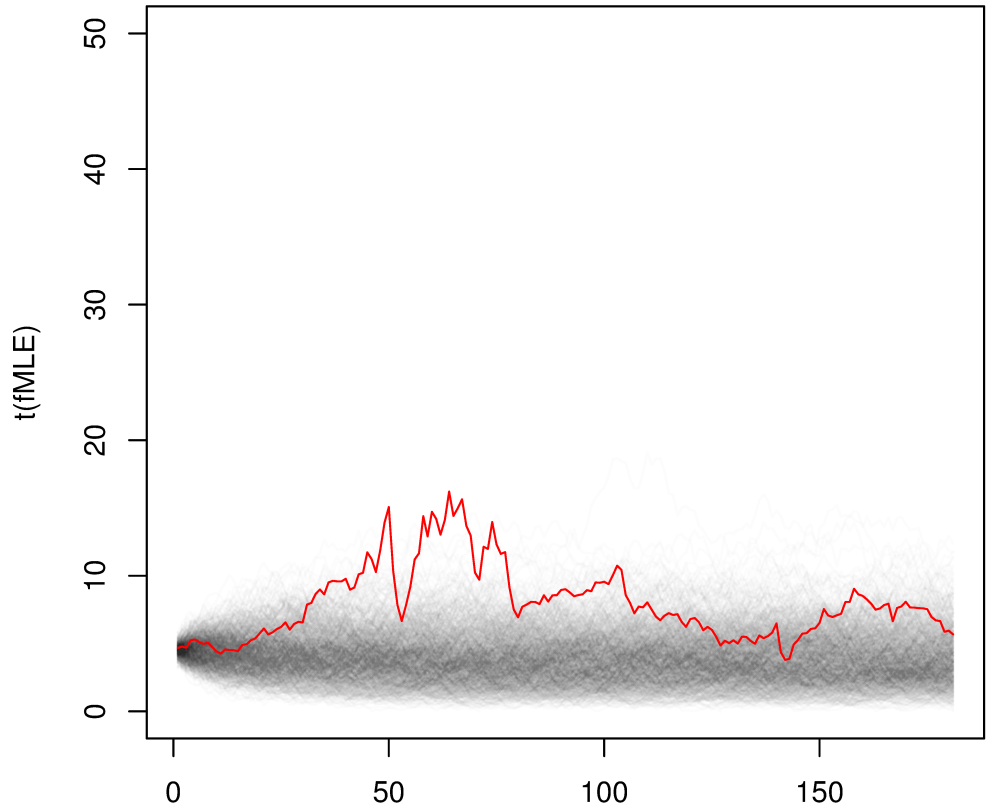
$$\text{fMSE}(LASSO) = \frac{1}{181} \sum_{i=351}^{531} \left(X_j^{LASSO} - X_j^{TRUE} \right)^2$$

$$\text{fMSE}(QMLE) = \frac{1}{181} \sum_{i=351}^{531} \left(X_j^{QMLE} - X_j^{TRUE} \right)^2$$

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
fMSE(QMLE)	5.43	20.19	26.55	26.84	33.12	51.38
fMSE(LASSO)	4.03	14.88	23.40	29.25	34.45	586.50

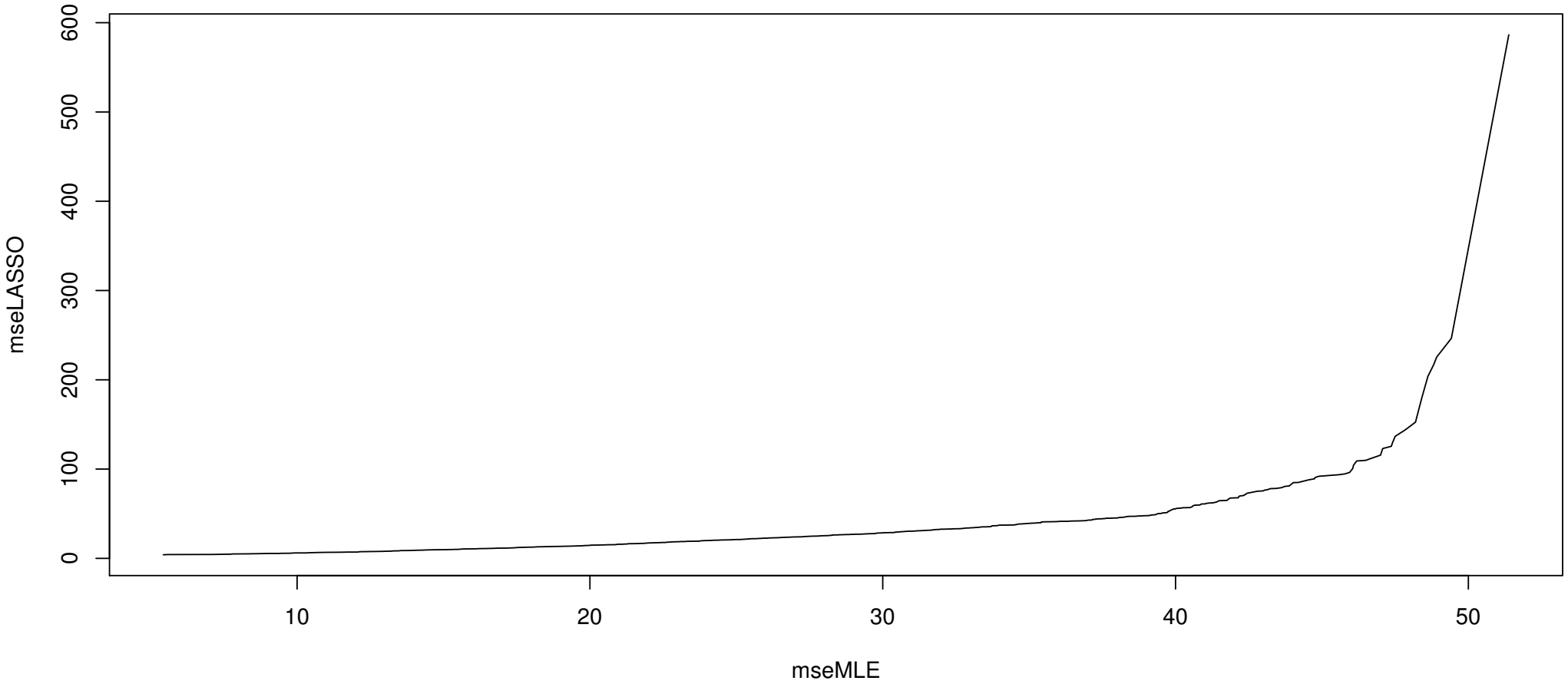
Example: simulation study with 1000 replications





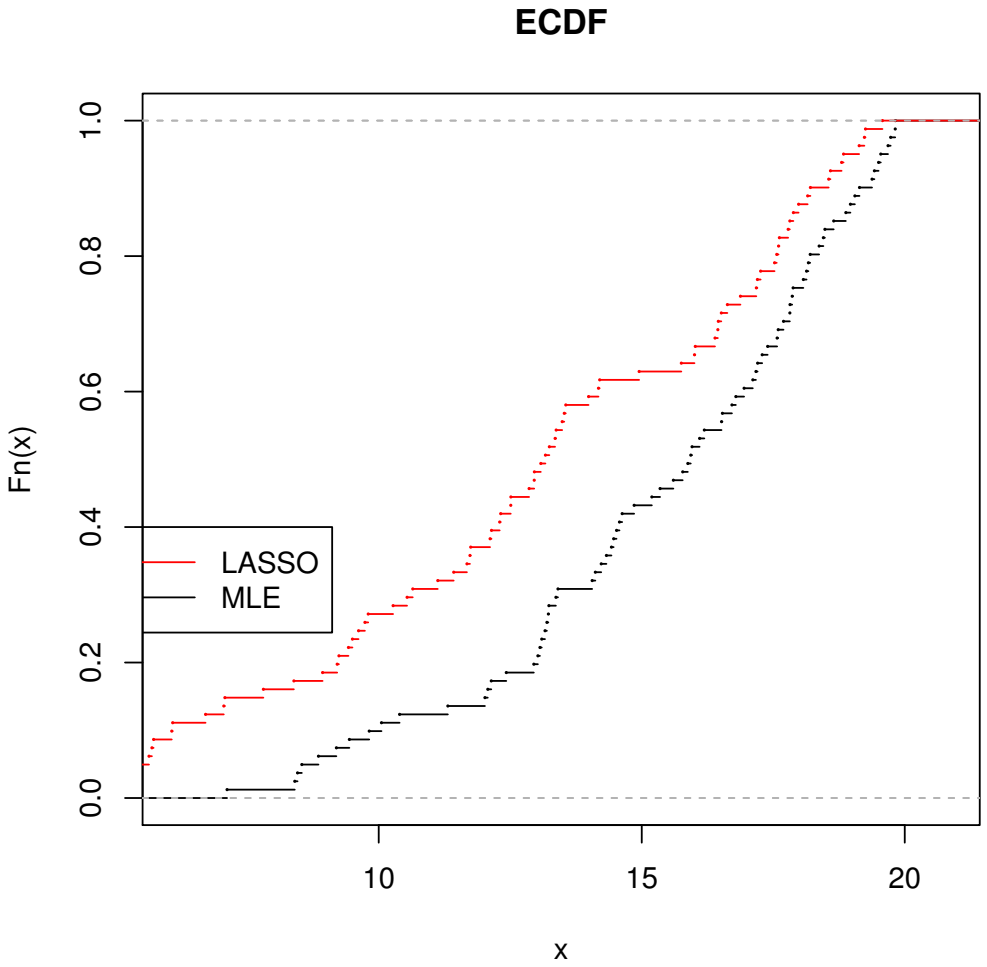
Example: simulation study with 1000 replications

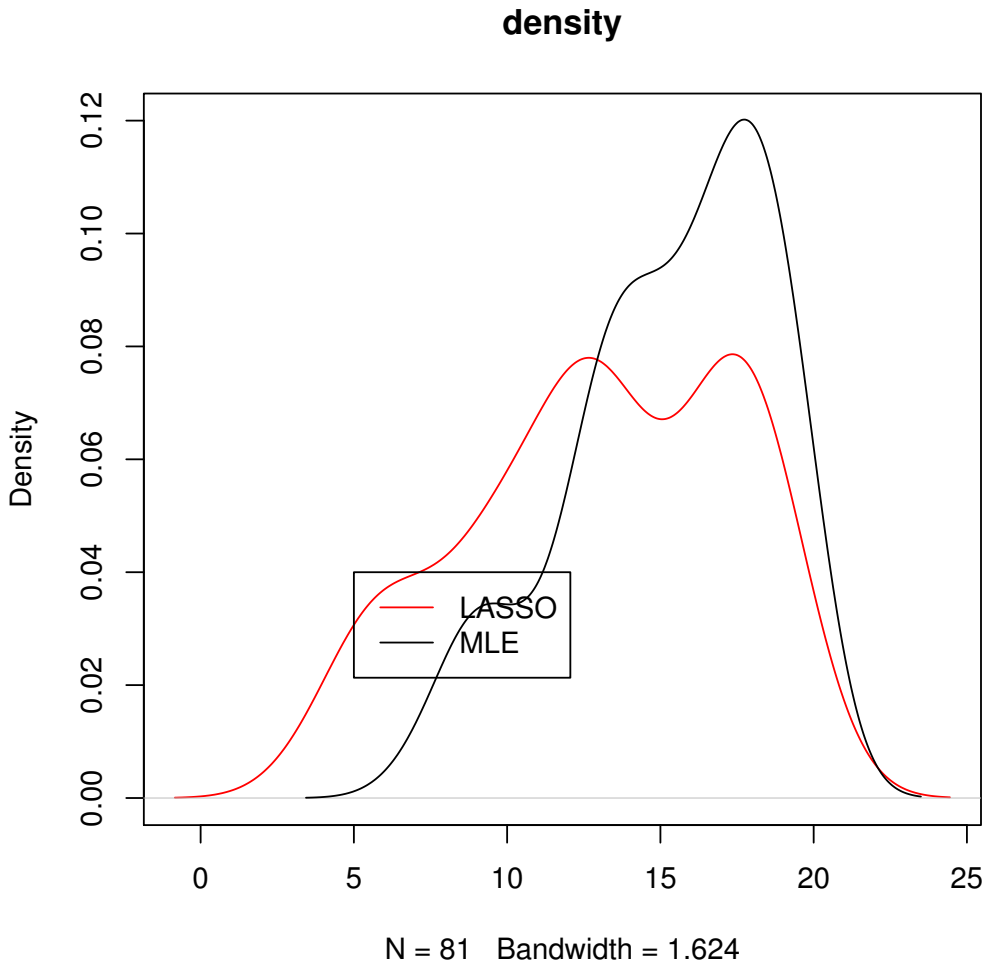
Simulation by simulation MSE(QMLE) $\geq$ MSE(LASSO)



Example: simulation study with 1000 replications

LASSO seems slightly better in this example





Model selection and causal inference (with Lasso)

A typical usage of Lasso in model selection is the case **causation** (closely related to Granger causation). For example, in a model like this

$$\begin{pmatrix} \mathrm{d}X_t \\ \mathrm{d}Y_t \end{pmatrix} = \begin{pmatrix} \kappa_0 + \mu_{11}X_t + \mu_{12}Y_t \\ \kappa_1 + \mu_{21}X_t + \mu_{22}Y_t \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \sigma_{11}X_t & \sigma_{12}Y_t \\ \sigma_{21}X_t & \sigma_{22}Y_t \end{pmatrix} \begin{pmatrix} \mathrm{d}W_t \\ \mathrm{d}B_t \end{pmatrix}$$

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.

The case of $\mu_{12} = 0$, $\mu_{21} = 0$, $\sigma_{12} = 0$, $\sigma_{21} = 0$ is of practical interest because the systems becomes

$$\begin{aligned} \mathrm{d}X_t = &\quad \kappa_0 + \mu_{11}X_t + \sigma_{11}X_t\mathrm{d}W_t \\ \mathrm{d}Y_t = &\quad \kappa_1 + \mu_{22}Y_t + \sigma_{22}Y_t\mathrm{d}B_t \end{aligned}$$

Of course this can be generalized to affine diffusion in higher dimension without imposing a specific correlation structure like in the above simple example.

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

References

References

- Azencott, R. (1982) Formule de taylor stochastique et développement asymptotique d'intégrales de feynmann, *Séminaire de Probabilités XVI; Supplément: Géometrie Différentielle Stochastique. Lecture Notes In Math*, **921**, 237–285.
- Dacunha-Castelle, D., Florens-Zmirou, D. (1986) Estimation of the coefficients of a diffusion from discrete observations, *Stochastics*, **19**, 263–284.
- De Gregorio, A., Iacus, S. M. (2012) Adaptive lasso-type estimation for multivariate diffusion processes, *Econometric Theory*, **28**, 838–860.
- Fan, J. (1997). Comments on "Wavelets in Statistics: A Review" by A. Antoniadis". *Journal of the Italian Statistical Association*, **6**, 131-138.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *JASA*, **96**, 1348-1360.
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004) Least angle regression, *The Annals of Statistics*, **32**, 407–489.
- Freidlin, M. I., Wentzell, A. D. (1998) *Random perturbations of dynamical systems. 2nd. ed.*, Springer-Verlag, New York.
- Geyer, C.J. (1994) On the asymptotics of constrained M-estimation, *Annals of Statistics*, **22**, 1993–2010.
- Geyer, C.J. (1996) On the asymptotics of convex stochastic optimization, available at `http://www.stat.umn.edu/PAPERS/preprints/convex.ps`.
- Iacus, S. M. (2000) Semiparametric estimation of the state of a dynamical system with small noise, *Stat. Infer. for Stoch. Proc.*, **3**, 277–288.
- Hastie, T., Tibshirani, R. and Friedman, J. (2008), *The Elements of Statistical Learning*, Springer Verlag, New York.
- Iacus, S. M., Kutoyants, Y. (2001) Semiparametric hypotheses testing for dynamical systems with small noise, *Mathematical Methods of Statistics*, **10**(1), 105–120.
- Kato, K. (2009) Asymptotics for argmin processes: Convexity arguments, *Journal of Multivariate Analysis*, **100**(8), 1816–1829.
- Kim, J., Pollard, D. (1990) Cube root asymptotics, *Annals of Statistics*, **18**, 191–219.
- Knight, K., Fu, W. (2000) Asymptotics for lasso-type estimators, *Annals of Statistics*, **28**, 1536–1378.
- Kunitomo, N., Takahashi, A. (2001) The asymptotic expansion approach to the valuation of interest rate contingent claims, *Mathematical Finance*, **11**(1), 117–151.
- Kutoyants, Y. (1984) *Parameter estimation for stochastic processes*, Heldermann, Berlin.
- Kutoyants, Y. (1991) Minimum distance parameter estimation for diffusion type observations, *C.R. Acad. Paris*, **312**, Sér. I, 637–642.
- Kutoyants, Y. (1994) *Identification of Dynamical Systems with Small Noise*, Kluwer Academic Publishers, Dordrecht, The Netherlands.
- Kutoyants, Y., Philibossian, P. (1994) On minimum L_1-norm estimates of the parameter of Ornstein-Uhlenbeck process, *Statistics and Probability Letters*, **20**(2), 117–123.
- Masuda, H., Shimizu, Y. (2016) Moment convergence in regularized estimation under multiple and mixed-rates asymptotics, `https://arxiv.org/abs/1406.6751`
- Takahashi, A., Yoshida, N. (2004) An asymptotic expansion scheme for optimal investment problems, *Stat. Inference Stoch. Process.*, **7**, 153–188.
- Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso, *J. Roy. Statist. Soc. Ser. B*, **58**, 267–288.
- Uchida, M. and Yoshida, N. (2004a) Asymptotic expansion for small diffusions applied to option pricing, *Statist. Infer. Stochast. Process*, **7**, 189–223.
- Uchida, M. and Yoshida, N. (2004b) Information criteria for small diffusions via the theory of Malliavin-Watanabe, *Statist. Infer. Stoch. Process*, **7**, 35–67.

References

- Park, T., and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, **103**(482), 681-686.
- Pollard, D. (1991) Asymptotics for least absolute deviation regression estimators, *Econometric Theory*, **7**, 186–199.
- Yoshida, N. (1992a) Asymptotic expansion for statistics related to small diffusions, *Journal of the Japan Statistical Society*, **22**, 139–159.
- Yoshida, N. (1992b) Asymp. expansion of maximum likelihood estimators for small diffusions via the theory of Malliavin-Watanabe, *P. Theory Rel.. Fields*, **92**, 275–311.
- Yoshida, N. (2003) Conditional expansions and their applications, *Stochastic Process. Appl.*, **107**, 53–81.
- Yoshida, N. (2011) Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations, *Ann. Inst. Statist. Math.*, **63**(3), 431–479.
- Yuan, M. and Lin, Y. (2006), Model Selection and Estimation in Regression with Grouped Variables, *JRSS, B*, **68**, 49-67.
- Zou, H. (2006) The adaptive LASSO and its Oracle properties, *J. Amer. Stat. Assoc.*, **101**(476), 1418-1429.
- Zou, H., Hastie, T. (2005) Regularization and variable selection via the elastic net, *J. R. Statist. Soc. B*, **67**(Part 2), 301-302.
- Zou, H., Zhang, E. (2009) On the adaptive elastic-net with a diverging number of parameters, *The Annals of Statistics*, **37**(4), 1733-1751.