Econometric Theory, 2012, Page 1 of 23.
doi:10.1017/S0266466611000806

ADAPTIVE LASSO-TYPE ESTIMATION FOR MULTIVARIATE DIFFUSION PROCESSES

ALESSANDRO DE GREGORIO
Sapienza University of Rome

STEFANO M. IACUS
University of Milan

The least absolute shrinkage and selection operator (LASSO) is a widely used statistical methodology for simultaneous estimation and variable selection. It is a shrinkage estimation method that allows one to select parsimonious models. In other words, this method estimates the redundant parameters as zero in the large samples and reduces variance of estimates. In recent years, many authors analyzed this technique from a theoretical and applied point of view. We introduce and study the adaptive LASSO problem for discretely observed multivariate diffusion processes. We prove oracle properties and also derive the asymptotic distribution of the LASSO estimator. This is a nontrivial extension of previous results by Wang and Leng (2007, *Journal of the American Statistical Association*, 102(479), 1039–1048) on LASSO estimation because of different rates of convergence of the estimators in the drift and diffusion coefficients. We perform simulations and real data analysis to provide some evidence on the applicability of this method.

1. INTRODUCTION

Model selection is an important issue in applied econometric analysis. For example, general regression models are used extensively by practitioners, and these are useful as long as the set of parameters (or covariates) is correctly specified. Therefore, correct model selection is crucial in the subsequent step of estimation. Model selection consists in setting some of the parameters to zero. As Caner (2009) noticed, models do not need to be nested, but one can rather construct a single large parametric model merging two orthogonal models and let the selection method choose one of the two models. A typical application is structural change models as explained in Andrews and Lu (2001).

Variable selection is particularly important when the true underlying model has a sparse representation. Correctly identifying significant predictors will

The authors would like to thank the editor and two anonymous referees for their comments and suggestions, which led to a substantial improvement of the earlier versions of the manuscript. Address correspondence to Stefano M. Iacus, Dipartimento di Scienze Economiche, Aziendali e Statistiche, Via Conservatorio 22, 20122 - Milan, Italy; e-mail: stefano.iacus@unimi.it.

improve the prediction performance of the fitted model (for an overview of feature selection, see Fan and Li, 2006).

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; and Efron, Hastie, Johnstone, and Tibshirani, 2004). The LASSO method could allow the dimensionality of the parameter space to change with the sample size; this is the main advantage of the LASSO approach over the classical information criterions (AIC, BIC, etc.). The LASSO method usually consists of the minimization of an L^2 norm under L^1 norm constraints on the parameters. Thus it usually implies a least squares or maximum likelihood approach plus constraints. The important property stating that the correct parameters are set to zero by the LASSO method under the true data generating model is called the oracle property (Fan and Li, 2001). As shown by Zou (2006), since the classical LASSO estimator uses the same amount of shrinkage for each parameter, the resulting model selection could be inconsistent. To overcome this drawback, it is possible to consider an adaptive amount of shrinkage for each parameter.

Originally, the LASSO procedure was introduced for linear regression problems, but in recent years this approach has been applied to time series analysis by several authors, mainly in the case of autoregressive models. For example, Wang, Li, and Tsai (2007) consider the problem of shrinkage estimation of regressive and autoregressive coefficients, while Nardi and Rinaldo (2011) consider penalized order selection in an AR(p) model. Furthermore, Caner and Knight (2010) show that econometricians can use a Bridge estimator to differentiate stationarity from unit root type of nonstationarity and select the optimal lag in autoregression (AR) series as well. The vector autoregression (VAR) case was considered in Hsu, Hung, and Chang (2008). Furthermore, Caner (2009) studied the LASSO method for general generalized method of moments (GMM) estimator also in the case of time series, and Knight (2008) extended the LASSO approach to nearly singular designs. More recently, Liao (2010) introduced a set of nuisance parameters in possibly misspecified moment conditions. The author shows that the LASSO-type techniques can simultaneously achieve consistent moment selection and efficient estimation in GMM with weakly dependent data. In the reduced rank error correction model, Liao and Phillips (2010) use the LASSO-type approach to perform cointegration rank selection, lagged differences selection, and efficient estimation simultaneously. They show that consistent cointegration rank selection can be achieved even if the parameters of the model are inconsistently estimated due to the weakly dependent innovations.

In this paper we consider the LASSO approach for discretely observed diffusion processes solution to stochastic differential equations. In recent years, there has been increasing interest around continuous time models specified by stochastic differential equations in economics. In particular, the econometric literature for these models evolves accordingly in order to produce correct statistical inference.

Multivariate diffusion models like the ones considered in this paper have been proposed in finance (Sundaresan, 2000), macroeconomics (Bergstrom, 1990; McCrorie and Chambers, 2006), and macro-finance (Piazzesi, 2009).

In the context of discretely observed diffusion processes, the likelihood function is not usually known in closed form. In this paper we use the quasi likelihood Gaussian approximation as proposed by many authors (e.g., Yoshida, 1992; Genon-Catalot and Jacod, 1993; Kessler, 1997).

For diffusion processes, the LASSO method requires some additional care because the rates of convergence of the estimators of parameters in the drift and the diffusion coefficient are different. We point out that the usual model selection strategy based on AIC (see Uchida and Yoshida, 2005) usually depends on the properties of the estimators but also on the method used to approximate the likelihood. Indeed, Akaike information criteria (AIC) requires a very precise calculation of the likelihood function to avoid bias (see Iacus, 2008). In contrast, the present LASSO approach depends solely on the properties of the estimator and so the problem of likelihood approximation is not particularly compelling. It is worth mentioning that model selection for continuous time diffusion processes was considered earlier in Uchida and Yoshida (2001) by means of information criteria.

The paper is organized as follows: Section 2 introduces the model and the regularity assumptions and states the problem of LASSO estimation for discretely sampled diffusion processes. Section 3 proves consistency and oracle properties of the LASSO estimator. Section 4 contains a Monte Carlo analysis and one application to real financial data. The conclusions of this work are summarized in Section 5. Proofs are collected in Section 6.

2. THE LASSO PROBLEM FOR DIFFUSION MODELS

We begin by introducing the reference model and the basic notations. Let X_t, $t \in [0, T]$, $0 < T < \infty$, be a d-dimensional diffusion process solution of the multivariate stochastic differential equation

$$\mathrm{d}X_t = b(\alpha, X_t)\mathrm{d}t + \sigma(\beta, X_t)\mathrm{d}W_t, \qquad X_0 = x_0, \tag{2.1}$$

where $\alpha = (\alpha_1, ..., \alpha_p)' \in \Theta_p \subset \mathbb{R}^p$, $p \geq 1$, $\beta = (\beta_1, ..., \beta_q)' \in \Theta_q \subset \mathbb{R}^q$, $q \geq 1$, are $p \times 1$ and $q \times 1$ vectors, respectively, $b : \Theta_p \times \mathbb{R}^d \to \mathbb{R}^d$, $\sigma : \Theta_q \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^m$, 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 the parameters α and β. We denote by $\theta = (\alpha, \beta) \in \Theta_p \times \Theta_q = \Theta$ the $(p + q) \times 1$ 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 by $|A|^2 = \mathrm{tr}(AA')$, i.e., the sum of squares of the elements of A. Furthermore we use the notation $\Sigma(\beta, x) = \sigma(\beta, x)\sigma(\beta, x)'$. 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 paper is $T = n\Delta_n \to \infty$, $\Delta_n \to 0$ 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.

We need some assumptions on the regularity of the process $X_t, t \in [0, T]$.

Assumption 1. There exists a constant C such that

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

Assumption 2. $\inf_{\beta,x} \det(\Sigma(\beta, x)) > 0$.

Assumption 3. The process $X_t, t \in [0, T]$, is ergodic for $\theta = \theta_0$ with invariant probability measure μ_{θ_0}.

Assumption 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 $\beta = \beta_0$.

Assumption 5. For all $m \geq 0$ and for all $\theta \in \Theta$, $\sup_t \mathrm{E}|\mathrm{X_t}|^{\mathrm{m}} < \infty$.

Assumption 6. For every $\theta \in \Theta$, the coefficients $b(\alpha, x)$ and $\sigma(\beta, x)$ are five times continuously differentiable with respect to x and the derivatives are bounded by a polynomial function in x, uniformly in θ.

Assumption 7. The coefficients $b(\alpha, x)$ and $\sigma(\beta, x)$ and all their partial derivatives with respect to x up to order 2 are three times continuously 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.

We observe that Assumption 1 ensures the existence and uniqueness of a solution to (2.1) for the value $\theta_0 = (\alpha_0, \beta_0)$ of $\theta \in \Theta$, while Assumption 4 is the identifiability condition. Hereafter, we assume that Assumptions 1–7 hold. These conditions are equivalent to the ones in Uchida and Yoshida (2005) and Kessler (1997) for what concerns the regularity of the model. In order to introduce the LASSO problem, we consider the negative quasi-loglikelihood function $\mathbb{H}_n : \mathbb{R}^{(n+1)\times d} \times \Theta \to \mathbb{R}$,

$$\mathbb{H}_n(\mathbf{X}_n, \theta) = \frac{1}{2} \sum_{i=1}^{n} \left\{ \log\det(\Sigma_{i-1}(\beta)) + \frac{1}{\Delta_n} (\Delta X_i - \Delta_n b_{i-1}(\alpha))' \Sigma_{i-1}^{-1}(\beta) \right.$$
$$\left. \times (\Delta X_i - \Delta_n b_{i-1}(\alpha)) \right\}, \tag{2.2}$$

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 used by, e.g., Yoshida (1992), Genon-Catalot and Jacod (1993), and Kessler (1997) to estimate stochastic differential equations because the true transition probability density for $X_t, t \in [0, T]$, does not have a closed form expression. The function (2.2) is obtained by discretization of the continuous time stochastic differential equation (2.1) by Euler-Maruyama scheme, that is,

$$X_{t_i} - X_{t_{i-1}} \cong b(\alpha, X_{t_{i-1}})\Delta_n + \sigma(\beta, X_{t_{i-1}})(W_{t_i} - W_{t_{i-1}}),$$

and the increments $(X_{t_i} - X_{t_{i-1}})$ are conditionally independent Gaussian random variables for $i = 1, ..., n$.

We denote by $\dot{\mathbb{H}}_n(\mathbf{X}_n, \theta)$ the vector of the first derivatives with respect to θ and by $\ddot{\mathbb{H}}_n(\mathbf{X}_n, \theta)$ the Hessian matrix. Let $\tilde{\theta}_n : \mathbb{R}^{(n+1)\times d} \to \Theta$ be the quasi-maximum likelihood estimator (QMLE) of $\theta \in \Theta$, based on (2.2), that is

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

We consider the matrix

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

where $\mathbf{I}_p$ and $\mathbf{I}_q$ are respectively the identity matrix of order p and q. This matrix plays the role of the rate of convergence in the estimation problem for the stochastic differential equation (2.1).

The regularity conditions of Assumptions 1–7 imply the following fundamental results, which have a crucial role in the proofs.

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 Assumptions 1–7, the following two properties hold true.*

(i) $\Lambda_n(\theta_0) \xrightarrow{p} \mathcal{I}(\theta_0)$, $\sup_{||\theta|| \le \epsilon_n} |\Lambda_n(\theta + \theta_0) - \Lambda_n(\theta_0)| = o_p(1)$, *for* $\epsilon_n \to 0$ *as* $n \to \infty$;

(ii) $\tilde{\theta}_n$ *is a consistent estimator of* θ_0 *and asymptotically Gaussian with rate of convergence given by* $\varphi(n)^{-1/2}$; *i.e.,*

$$\varphi(n)^{-1/2}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} N(0, \mathcal{I}(\theta_0)^{-1}),$$

where $\mathcal{I}(\theta_0)$ *is the positive definite and invertible Fisher information matrix at* θ_0 *given by*

$$\mathcal{I}(\theta_0) = \begin{pmatrix} \Gamma_\alpha = [\mathcal{I}_b^{kj}(\alpha_0)]_{k,j=1,...,p} & 0 \\ 0 & \Gamma_\beta = [\mathcal{I}_\sigma^{kj}(\beta_0)]_{k,j=1,...,q} \end{pmatrix}$$

where

$$\mathcal{I}_b^{kj}(\alpha_0) = \int \left(\frac{\partial b(\alpha_0, x)}{\partial \alpha_k}\right)' \Sigma^{-1}(\beta_0, x) \left(\frac{\partial b(\alpha_0, x)}{\partial \alpha_j}\right) \mu_{\theta_0}(dx),$$

$$\mathcal{I}_\sigma^{kj}(\beta_0) = 2\int tr\left[\frac{\partial \Sigma(\beta_0, x)}{\partial \beta_k}\Sigma^{-1}(\beta_0, x)\frac{\partial \Sigma(\beta_0, x)}{\partial \beta_j}\Sigma^{-1}(\beta_0, x)\right] \mu_{\theta_0}(dx).$$

The classical adaptive LASSO objective function, in this case, should be given by

$$\mathbb{H}_n(\mathbf{X}_n, \theta) + \sum_{j=1}^{p} \lambda_{n,j}|\alpha_j| + \sum_{k=1}^{q} \gamma_{n,k}|\beta_k|, \tag{2.3}$$

where $\lambda_{n,j}$ and $\gamma_{n,k}$ assume real positive values representing an adaptive amount of the shrinkage for each element of α and β. The LASSO estimator is the minimizer of the objective function (2.3). Usually, this is a nonlinear optimization problem under L_1 constraints, which might be numerically challenging to solve. Nevertheless, using the approach of Wang and Leng (2007), we can consider a different objective function with respect to (2.3). Indeed, we define the adaptive LASSO-type estimator $\hat{\theta}_n : \mathbb{R}^{(n+1)\times d} \to \Theta$ as the solution of

$$\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n) = \arg\min_{\theta} \mathcal{F}(\theta), \tag{2.4}$$

where

$$\mathcal{F}(\theta) = (\theta - \tilde{\theta}_n)' \ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)(\theta - \tilde{\theta}_n) + \sum_{j=1}^{p} \lambda_{n,j}|\alpha_j| + \sum_{k=1}^{q} \gamma_{n,k}|\beta_k|. \tag{2.5}$$

Then, (2.5) leads to a minimum distance criterion plus the penalty terms and it is much easier to solve numerically than (2.3); nevertheless, the solutions of the two objective functions are equivalent. Although the asymptotic properties of the LASSO-type estimator have been established in Wang and Leng, the extension to discretely observed diffusion processes is nontrivial because the estimators for the drift and diffusion parameters have two different rates of convergence. For this reason, in the objective function (2.5), two sets of L_1 constraints and weighting sequences ($\lambda_{n,j}$ and $\gamma_{n,k}$) are required to take into account the different rates of convergence.

Remark 1. The approach adopted by Wang and Leng (2007) also holds when the diffusion process (2.1) has the same parametric vector θ in both drift and diffusion coefficients. In this context, we use the 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 Sørensen, 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 $n \to \infty$.

3. ORACLE PROPERTIES

As argued by Fan and Li (2001), a good selection procedure should have the so-called oracle properties:

(i) consistently estimates null parameters as zero and vice versa;
(ii) has the optimal estimation rate and converges to a Gaussian random variable $N(0, \Sigma)$, where Σ is the covariance matrix of the true subset model.

The aim of this section is to prove that the adaptive LASSO-type estimator $\hat{\theta}_n$ has good behavior in the oracle sense.

As shown by Zou (2006), the classical LASSO estimation cannot be as efficient as the oracle, and the selection results could be inconsistent, whereas its adaptive version has the oracle 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 $1 \leq k \leq q_0$. To study the asymptotic properties of the LASSO-type estimator $\hat{\theta}_n$, we consider the following conditions.

Condition 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 $\nu_n = \max\{\gamma_{n,k}, 1 \leq k \leq q_0\}$;

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

Condition 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 $1 \leq k \leq q_0$, tend to infinity slower than $\sqrt{n\Delta_n}$ and $\sqrt{n}$, respectively. Analogously, we observe that Condition 2 means that the minimal tuning coefficients for the parameter α_j and β_k, with $j > p_0$ and $k > q_0$, tend to infinity faster than $\sqrt{n\Delta_n}$ and $\sqrt{n}$, respectively.

THEOREM 1. *Under Assumptions 1–7 and Condition 1, one has that*

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

For the sake of simplicity, we 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, $\theta_0 = (\alpha_0, \beta_0) = (\alpha_0^*, \alpha_0^\circ, \beta_0^*, \beta_0^\circ)$ and $\hat{\theta}_n = (\hat{\alpha}_n^*, \hat{\alpha}_n^\circ, \hat{\beta}_n^*, \hat{\beta}_n^\circ)$, $\tilde{\theta}_n = (\tilde{\alpha}_n^*, \tilde{\alpha}_n^\circ, \tilde{\beta}_n^*, \tilde{\beta}_n^\circ)$.

THEOREM 2. *Under Assumptions 1–7 and Conditions 1–2, we have that*

$$P(\hat{\alpha}_n^\circ = 0) \to 1 \quad \text{and} \quad P(\hat{\beta}_n^\circ = 0) \to 1. \tag{3.1}$$

From Theorem 1, we can conclude that the estimator $\hat{\theta}_n$ is consistent. Furthermore, Theorem 2 says that all the estimates of the zero parameters are correctly set equal to zero with probability tending to 1. In other words, the model selection procedure is consistent and the true subset model is correctly indentified with probability tending to 1.

To complete our program, we derive the asymptotic distribution of $\hat{\theta}_n^* = (\hat{\alpha}_n^*, \hat{\beta}_n^*)$. Hence, we indicate by $\mathcal{I}_0(\theta_0^*)$ the $(p_0 + q_0) \times (p_0 + q_0)$ submatrix of $\mathcal{I}(\theta_0)$; that is,

$$\mathcal{I}_0(\theta_0^*) = \begin{pmatrix} \Gamma_\alpha^{**} = [\mathcal{I}_b^{kj}(\alpha_0)]_{k,j=1,...,p_0} & 0 \\ 0 & \Gamma_\beta^{**} = [\mathcal{I}_\sigma^{kj}(\beta_0)]_{k,j=1,...,q_0} \end{pmatrix},$$

and introduce the rate of convergence matrix

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

The next result establishes that the estimator $\hat{\theta}_n^*$ is efficient as well as the oracle estimator.

THEOREM 3 (Oracle property). *Under Assumptions 1–7 and Conditions 1–2, we have that*

$$\varphi_0(n)^{-1/2}(\hat{\theta}_n^* - \theta_0^*) \xrightarrow{d} N(0, \mathcal{I}_0^{-1}(\theta_0^*)). \tag{3.2}$$

Clearly, the theoretical and practical implications of our method rely on the specification of the tuning parameter $\lambda_{n,j}$ and $\gamma_{n,k}$. As observed in Wang and Leng (2007), these values could be obtained by means of Bayes information criteria (BIC) instead of other model selection criteria like generalized cross-validation (GCV) or Akaike information criteria (AIC). As observed in Wang et al. (2007), GCV and AIC are asymptotically equivalent and are inconsistent in terms of model selection. Unfortunately, this solution is computationally heavy and then impracticable. Therefore, the tuning parameters should be chosen as in Zou (2006) in the following way,

$$\lambda_{n,j} = \lambda_0 |\tilde{\alpha}_{n,j}|^{-\delta_1}, \qquad \gamma_{n,k} = \gamma_0 |\tilde{\beta}_{n,j}|^{-\delta_2}, \tag{3.3}$$

where $\tilde{\alpha}_{n,j}$ and $\tilde{\beta}_{n,k}$ are the unpenalized estimators of α_j and β_k, respectively, $\delta_1, \delta_2 > 0$ and usually taken unitary. Since $\tilde{\alpha}_{n,j}$ and $\tilde{\beta}_{n,k}$ are consistent estimators (see Lemma 1), we have that under the conditions

$$\frac{\lambda_0}{\sqrt{n\Delta_n}} \to 0, \qquad (n\Delta_n)^{\frac{\delta_1-1}{2}}\lambda_0 \to \infty, \quad \text{and} \quad \frac{\gamma_0}{\sqrt{n}} \to 0, \qquad n^{\frac{\delta_2-1}{2}}\gamma_0 \to \infty$$

as $n \to \infty$, Conditions 1 and 2 hold. Then the asymptotic results listed in the above theorems are valid.

4. PERFORMANCE OF THE LASSO METHOD FOR SMALL SAMPLE SIZE

In this section we perform a small Monte Carlo analysis to check whether the LASSO method is able to select a specified model also in small samples. We also apply the method to a benchmark data set often used in the literature of model selection. The asymptotic framework of this paper is not completely realized in the simulated and real data experiments (i.e., Δ_n is not so small and T does not diverge), nevertheless we test what happens outside the theoretical framework. The simulations are done using the sde package (see Iacus, 2008) for the R statistical environment.

As Wang, Phillips, and Yu (2011) pointed out, when estimating the drift parameters, bias may arise even for large samples and for linear diffusions. Although our simulation setup considers these aspects, when reading the results one should keep in mind that estimates of the drift parameters are necessarily more biased than estimates of the diffusion parameters.

In both cases, we do not pretend to give extensive analysis of the method, because the previous theorems already prove the asymptotic validity of the LASSO approach for diffusion processes. Instead, we just want to show some evidence on simulated and real data to give the feeling of the applicability of the method.

To solve the LASSO problem, we make use of the "L-BFGS-B" optimizer by Byrd, Lu, Nocedal, and Zhu (1995), which allows for box constrained optimization. Indeed, we use a zero lower bound in the "L-BFGS-B" optimizer, and often the solution results in some coefficients estimated exactly as zero.

4.1. A Simulation Experiment: One-Dimensional Case

We reproduce the experimental design similar to Uchida and Yoshida (2005). Therefore, we consider a parametric diffusion process solution of the stochastic differential equation

$$\mathrm{d}X_t = (\theta_1 - \theta_2 X_t)\mathrm{d}t + (\theta_3 + \theta_4 X_t)^{\theta_5}\mathrm{d}W_t, \qquad X_0 = 1\,. \tag{4.1}$$

We simulate 1,000 trajectories of this process with true parameter vector $\theta = (\theta_1 = 1, \theta_2 = 0.1, \theta_3 = 0, \theta_4 = 2, \theta_5 = 0.5)$ using the second Milstein scheme, i.e., the data are simulated according to

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

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

In order to get as close as possible to the asymptotic scheme of this paper, we consider the following simulation setup: for a given number n of observations, we set $T = n^{1/3}$ (time horizon) and $\Delta_n = T/n$. Then, choosing $n = 100$, we obtain $\Delta_n = 0.046$, while for $n = 1{,}000$, we have that $\Delta_n = 0.01$.

We perform both QMLE and LASSO estimation using the objective function $\mathcal{F}(\theta)$, the quasi-likelihood estimator, and the Hessian matrix obtained by the function (2.2), particularized for the present model (4.1). For the penalization terms, we set $\lambda_0 = \gamma_0$ to 1 and 5 and $\delta_1 = \delta_2$ equal to 1 and 2 in (3.3).

Table 1 reports average values of quasi-maximum likelihood (QML) and LASSO estimates over 1,000 Monte Carlo replications as well as the Monte Carlo

TABLE 1. LASSO and QML estimates of the model (4.1) for different sample sizes and penalization terms

	θ_1	θ_2	θ_3	θ_4	θ_5	% $\theta_3 = 0$
True	1.0	0.1	0.0	2.0	0.5	
Sample size: $n = 100$						
QMLE:	2.58	1.04	0.27	1.89	0.75	
	(1.47)	(0.91)	(0.57)	(1.10)	(0.87)	
LASSO: $\lambda_0 = \gamma_0 = 1, \delta_1 = \delta_2 = 1$	1.92	0.69	0.17	1.69	0.78	78%
	(1.10)	(0.84)	(0.41)	(0.92)	(0.93)	
LASSO: $\lambda_0 = \gamma_0 = 1, \delta_1 = \delta_2 = 2$	1.32	0.47	0.22	1.74	0.68	80%
	(0.68)	(0.30)	(0.12)	(0.51)	(0.19)	
LASSO: $\lambda_0 = \gamma_0 = 5, \delta_1 = \delta_2 = 1$	0.70	0.11	0.14	1.30	0.79	87%
	(0.56)	(0.38)	(0.37)	(0.80)	(0.96)	
LASSO: $\lambda_0 = \gamma_0 = 5, \delta_1 = \delta_2 = 2$	0.74	0.15	0.19	1.80	0.64	87%
	(0.54)	(0.14)	(0.05)	(0.46)	(0.12)	
Sample size: $n = 1{,}000$						
QMLE:	2.07	0.56	0.11	1.90	0.52	
	(1.25)	(0.52)	(0.27)	(0.37)	(0.06)	
LASSO: $\lambda_0 = \gamma_0 = 1, \delta_1 = \delta_2 = 1$	1.74	0.42	0.07	1.94	0.51	84%
	(1.01)	(0.49)	(0.25)	(0.35)	(0.06)	
LASSO: $\lambda_0 = \gamma_0 = 1, \delta_1 = \delta_2 = 2$	1.30	0.29	0.06	1.97	0.51	92%
	(0.46)	(0.16)	(0.03)	(0.16)	(0.03)	
LASSO: $\lambda_0 = \gamma_0 = 5, \delta_1 = \delta_2 = 1$	0.93	0.11	0.05	1.94	0.51	91%
	(0.47)	(0.29)	(0.22)	(0.33)	(0.08)	
LASSO: $\lambda_0 = \gamma_0 = 5, \delta_1 = \delta_2 = 2$	0.90	0.10	0.04	2.00	0.50	96%
	(0.40)	(0.07)	(0.01)	(0.15)	(0.02)	

Notes: Average values over 1,000 Monte Carlo replications. In parentheses: Monte Carlo standard errors. The column % $\theta_3 = 0$ represents the number of times, in percentage over 1,000 replications, that the parameter θ_3 is exactly estimated as zero.

standard deviations. As emerges from Table 1, on average the estimates get better as the sample size increases and the shrinkage of the parameters, for a given sample size, is higher for higher values of λ_0 and γ_0 but particularly as $\delta_1 = \delta_2$ increases. This is also expected from the theory.

Figures 1 and 2 report the density estimation of the estimates of the parameters $\theta_i, i = 1, \ldots, 5$ against their theoretical true values for $n = 100$ (top) and $n = 1{,}000$ (bottom). These distributions are obtained using the estimates obtained from the 1,000 Monte Carlo replications. Figures 1 and 2 indicate that all parameters are correctly estimated most of the time and, in particular, the parameter θ_3 is often estimated as zero (Table 1 also reports the percentage of times θ_3 is estimated as zero). Notice that, as expected, the precision of the estimates grows with sample

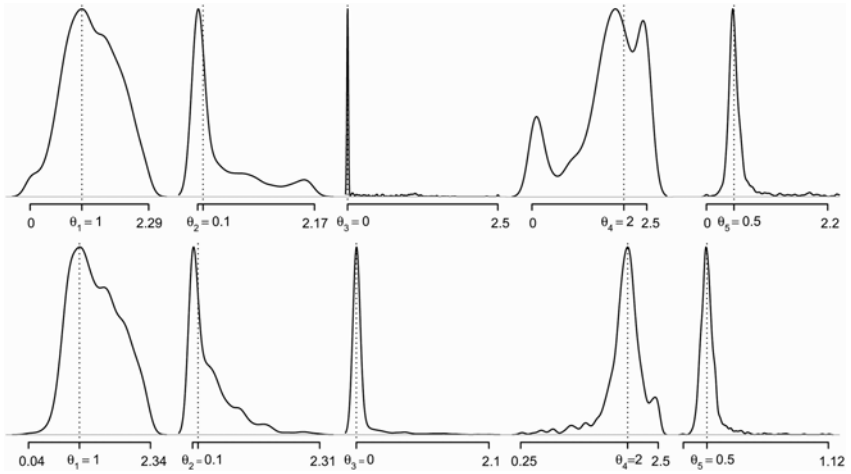

FIGURE 1. Density estimation of the LASSO-type estimates of the parameters of the process $\mathrm{d}X_t = (\theta_1 - \theta_2 X_t)\mathrm{d}t + (\theta_3 + \theta_4 X_t)^{\theta_5}\mathrm{d}W_t$ over 1,000 Monte Carlo replications. True values ($\theta_1 = 1, \theta_2 = 0.1, \theta_3 = 0, \theta_4 = 2, \theta_5 = 0.5$) represented as vertical dotted lines. Upper panel: sample size of $n = 100$. Bottom panel: sample size $n = 1{,}000$. Penalization in both cases: $\lambda_0 = \gamma_0 = 1$, $\delta_1 = \delta_2 = 1$.

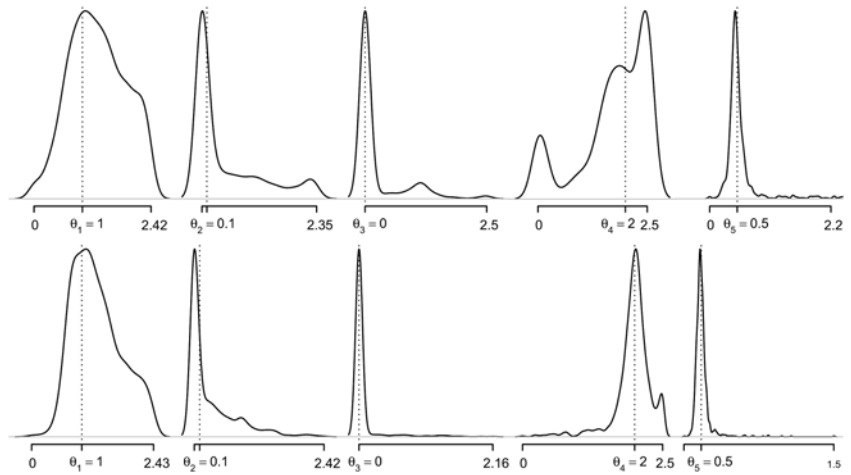

FIGURE 2. Density estimation of the LASSO-type estimates of the parameters of the process $\mathrm{d}X_t = (\theta_1 - \theta_2 X_t)\mathrm{d}t + (\theta_3 + \theta_4 X_t)^{\theta_5}\mathrm{d}W_t$ over 1,000 Monte Carlo replications. True values ($\theta_1 = 1, \theta_2 = 0.1, \theta_3 = 0, \theta_4 = 2, \theta_5 = 0.5$) represented as vertical dotted lines. Upper panel: sample size of $n = 100$. Bottom panel: sample size $n = 1{,}000$. Penalization in both cases: $\lambda_0 = \gamma_0 = 1$, $\delta_1 = \delta_2 = 2$.

size as well as the smoothness of the estimated density. From Figures 1 and 2 one can see that, for small sample sizes, the estimated densities present bimodality for some parameters. A similar effect has been noticed previously by Leeb and Potscher (2005, Fig. 2). The explanation for this bimodality effect seems to be related to the fact that the convergence result to the Gaussian limit is only pointwise for each fixed vector of the parameter space, and this convergence is not uniform, not even locally. Thus, for certain values of the parameters the convergence to the Gaussian limit is slower and hard to obtain in finite samples. In our case, the effect is highly reduced when $\delta_1 = \delta_2$ move from 1 (Figure 1) to 2 (Figure 2). We think that the comparison of these two plots is very instructive to understand the behavior of the LASSO estimates for different penalizations and against the quasi-maximum likelihood estimates.

4.2. A Simulation Experiment: Multidimensional Case

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

$$\begin{pmatrix} \mathrm{d}X_t \\ \mathrm{d}Y_t \end{pmatrix} = \begin{pmatrix} 1-\mu_{11}X_t+\mu_{12}Y_t \\ 2+\mu_{21}X_t-\mu_{22}Y_t \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \sigma_{11}X_t & -\sigma_{12}Y_t \\ \sigma_{21}X_t & \sigma_{22}Y_t \end{pmatrix} \begin{pmatrix} \mathrm{d}W_t \\ \mathrm{d}B_t \end{pmatrix}, \tag{4.2}$$

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. Model (4.2) 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 $\beta = (\sigma_{11} = 0.3, \sigma_{12} = 0, \sigma_{21} = 0, \sigma_{22} = 0.2)'$, $\theta = (\alpha, \beta)$.

We consider the same simulation scheme as in the previous section. Notice that we have eight parameters in the model but only four are nonzero. Table 2 shows good behavior of the LASSO-type estimator even in the multidimensional case.

As before, as sample size increases the estimators get better and better and the oracle property for sparse system reveals as well; i.e., the LASSO estimate is able to shrink toward zero the parameters μ_{12}, μ_{21}, σ_{12}, and σ_{21}. Table 2 also reports the number of times each coefficient is estimated as zero by the LASSO method. Again, on average we notice better performance when $\delta_1 = \delta_2$ passes from 1 to 2 for a given sample size n and constants γ_0, λ_0.

In this particular experiment, we also test the post-LASSO estimation as in Belloni and Chernozhukov (2011). This means that we run the QMLE procedure on the model selected by the LASSO method, which is the true model. Table 2 shows the performance of the QML estimator in this case for sample sizes $n = 100$ and $n = 1{,}000$. It is quite clear that the QML estimator gains in perfomance, suggesting that the post-LASSO approach is a reasonable one.

TABLE 2. LASSO and QML estimates of the model (4.2) for sample sizes of 100 and 1,000 and penalization terms

	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
Sample size $n = 100$								
QMLE:	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: $\lambda_0 = \gamma_0 = 1, \delta_1 = \delta_2 = 1$	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: $\lambda_0 = \gamma_0 = 1, \delta_1 = \delta_2 = 2$	0.86	0.00	0.03	0.70	0.30	0.00	0.00	0.20
	(0.10)	(0.00)	(0.05)	(0.08)	(0.02)	(0.00)	(0.00)	(0.01)
% of times $\theta_i = 0$	0.4	100.0	92.0	0.0	0.0	99.4	99.1	0.1
LASSO: $\lambda_0 = \gamma_0 = 5, \delta_1 = \delta_2 = 1$	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
LASSO: $\lambda_0 = \gamma_0 = 5, \delta_1 = \delta_2 = 2$	0.81	0.00	0.00	0.66	0.29	0.00	0.00	0.19
	(0.10)	(0.00)	(0.00)	(0.07)	(0.02)	(0.00)	(0.00)	(0.01)
% of times $\theta_i = 0$	0.0	100.0	100.0	0.0	0.5	99.5	99.4	0.5
Post-LASSO QMLE:	0.88	–	–	0.70	0.35	–	–	0.22
	(0.16)	–	–	(0.10)	(0.04)	–	–	(0.02)
Sample size $n = 1{,}000$								
QMLE:	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: $\lambda_0 = \gamma_0 = 1, \delta_1 = \delta_2 = 1$	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: $\lambda_0 = \gamma_0 = 1, \delta_1 = \delta_2 = 2$	0.88	0.00	0.05	0.71	0.30	0.00	0.00	0.20
	(0.07)	(0.00)	(0.06)	(0.06)	(0.02)	(0.00)	(0.00)	(0.01)
% of times $\theta_i = 0$	0.0	100.0	86.4	0.0	0.3	99.7	98.9	0.3
LASSO: $\lambda_0 = \gamma_0 = 5, \delta_1 = \delta_2 = 1$	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
LASSO: $\lambda_0 = \gamma_0 = 5, \delta_1 = \delta_2 = 2$	0.86	0.00	0.00	0.68	0.29	0.01	0.00	0.19
	(0.07)	(0.00)	(0.00)	(0.04)	(0.02)	(0.00)	(0.00)	(0.01)
% of times $\theta_i = 0$	0.0	100.0	99.9	0.0	0.3	99.7	99.6	0.3
Post-LASSO QMLE:	0.89	–	–	0.70	0.35	–	–	0.22
	(0.11)	–	–	(0.07)	(0.04)	–	–	(0.02)

Notes: Average values over 1,000 Monte Carlo replications. In parentheses: Monte Carlo standard errors.

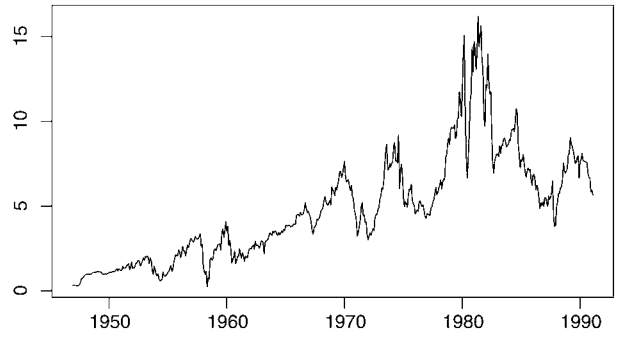

FIGURE 3. U.S. interest rate monthly data from 06/1964 to 12/1989.

4.3. An Example of Use in the Problem of Identification of the Term Structure of Interest Rates

In this section we reanalyze the U.S. interest rates monthly data from 06/1964 to 12/1989 for a total of 307 observations (see Figure 3). These data have been analyzed by many authors, including Aït-Sahalia (1996), Nowman (1997), and Yu and Phillips (2001), to mention a few. We do not pretend to give the definitive answer on the subject, but just to analyze the effect of the model selection via the LASSO in a real application.

The data used for this application were taken from the R package Ecdat by Croissant (2006). The different authors all try to fit a version of the so-called CKLS model (from Chan, Karolyi, Longstaff, and Sanders, 1992) that is the solution X_t of the stochastic differential equation

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

This model encompasses several other models depending on the number of non-null parameters, as Table 3 shows. This makes clear why the model selection on the CKLS model is quite appealing.

Our application of the LASSO method is reported in Table 4, along with the results from Yu and Phillips (2001) just for comparison.

Although we have proven that asymptotically the LASSO provides consistent estimates with the oracle properties, for finite sample size this is not always the case, as mentioned by several authors. In this application, we estimate the parameters using quasi-likelihood method (QMLE in the table) in the first stage, then set the penalties as in (3.3) and run the LASSO optimization. We estimate the CKLS parameters via the LASSO using mild penalties (i.e., $\lambda_0 = \gamma_0 = 1$ in (3.3)) and strong penalties (i.e., $\lambda_0 = \gamma_0 = 10$). Very strong penalties suggest that the model does not contain the term β, and in both cases the LASSO estimation suggests $\gamma = 3/2$ for $\delta_1 = \delta_2 = 1$ and therefore a model quite close to Cox, Ingersoll, and Ross (1980). For $\delta_1 = \delta_2 = 2$ we obtain $\gamma = 1.79$, which is different from

TABLE 3. The family of one-factor short-term interest rate models seen as special cases of the general CKLS model

Reference	Model	α	β	γ
Merton (1973)	$dX_t = \alpha dt + \sigma dW_t$		0	0
Vasicek (1977)	$dX_t = (\alpha + \beta X_t)dt + \sigma dW_t$			0
Cox, Ingersoll, and Ross (1985)	$dX_t = (\alpha + \beta X_t)dt + \sigma \sqrt{X_t} dW_t$			1/2
Dothan (1978)	$dX_t = \sigma X_t dW_t$	0	0	1
Geometric Brownian motion	$dX_t = \beta X_t dt + \sigma X_t dW_t$	0		1
Brennan and Schwartz (1980)	$dX_t = (\alpha + \beta X_t)dt + \sigma X_t dW_t$			1
Cox, Ingersoll, and Ross (1980)	$dX_t = \sigma X_t^{3/2} dW_t$	0	0	3/2
Constant elasticity variance	$dX_t = \beta X_t dt + \sigma X_t^{\gamma} dW_t$	0		
CKLS (1992)	$dX_t = (\alpha + \beta X_t)dt + \sigma X_t^{\gamma} dW_t$			

TABLE 4. Model selection on the CKLS model for the U.S. interest rates data

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	2.0069	−0.3330	0.1741	1.3610
		(0.5216)	(0.0677)		
CKLS	QMLE	2.0822	−0.2756	0.1322	1.4392
		(0.9635)	(0.1895)	(0.0253)	(0.1018)
CKLS	QMLE + LASSO	1.5435	−0.1687	0.1306	1.4452
	with mild penalization $\delta = 1$	(0.6813)	(0.1340)	(0.0179)	(0.0720)
CKLS	QMLE + LASSO	0.5412	0.0001	0.1178	1.4944
	with strong penalization $\delta = 1$	(0.2076)	(0.0054)	(0.0179)	(0.0720)
CKLS	QMLE + LASSO	0.7982	−0.0009	0.1145	1.5139
	with mild penalization $\delta = 2$	(0.2164)	(0.0110)	(0.0181)	(0.0726)
CKLS	QMLE + LASSO	0.7465	−0.0001	0.0431	1.7941
	with strong penalization $\delta = 2$	(0.2106)	(0.0038)	(0.0181)	(0.0726)
CIR('80)	Post-LASSO QMLE	0.8072	–	0.1297	1.4555
		(0.2959)	–	(0.0241)	(0.0994)

Notes: Table taken from Yu and Phillips (2001) and updated with LASSO results. Standard errors in parentheses when available.

any known model in Table 3. Still, in most cases, the parameter β is estimated as zero, indicating that the LASSO method suggests a lack of mean reversion term as expected from the plot of the time series (see Figure 3). The post-LASSO estimation seems to confirm the exponent γ of about 1.5. Being a shrinkage estimator,

the LASSO estimates have very low standard error compared to the other cases. As said, this application has been done to show the applicability of the LASSO method, and we do not pretend to draw in-depth conclusions from this empirical evidence, which is out of our competence.

5. CONCLUSIONS

In this paper, for multivariate diffusion processes defined by stochastic differential equations, we introduce the LASSO methodology, which permits us to perform model selection and estimation simultaneously. This approach is particularly useful because it leads us to discard the redundant parameters overcoming the AIC drawbacks. Indeed, AIC requires a very precise calculation of the likelihood function to avoid bias. Furthermore, we show that the LASSO estimator enjoys the oracle property, that is, it selects the true subset model and estimates the nonredundant parameters efficiently, as if the true model was known. We point out that in our case the LASSO procedure is not a trivial extension of the classical LASSO one because there are two different rates of convergence for the estimators of drift and diffusion coefficients. In regard to this fact, the objective function has two different constraints on the vector parameters α and β.

In view of the growing importance of multivariate diffusion processes in econometrics, we believe that the technique analyzed in this paper is useful for applied researchers. For example, restricting the attention to a model with few and noredundant parameters, one may be able to improve the predictive performance of the estimated model.

6. PROOFS

Proof of Theorem 1. Following Fan and Li (2001), the existence of a consistent local minimizer is implied by that fact that for an arbitrarily small $\varepsilon > 0$, there exists a sufficiently large constant C, such that

$$\lim_{n\to\infty} P\left\{\inf_{z\in\mathbb{R}^{p+q}:|z|=C} \mathcal{F}\big(\theta_0+\varphi(n)^{1/2}z\big) > \mathcal{F}(\theta_0)\right\} > 1-\varepsilon, \tag{6.1}$$

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

$$\begin{aligned}\mathcal{F}\big(\theta_0+\varphi(n)^{1/2}z\big)-\mathcal{F}(\theta_0) &= z'\varphi(n)^{1/2}\ddot{\mathbb{H}}_n\big(\mathbf{X}_n,\tilde{\theta}_n\big)\varphi(n)^{1/2}z \\ &\quad + 2z'\varphi(n)^{1/2}\ddot{\mathbb{H}}_n(\mathbf{X}_n,\tilde{\theta}_n)\varphi(n)^{1/2}\varphi(n)^{-1/2}\big(\theta_0-\tilde{\theta}_n\big) \\ &\quad + \left(\sum_{j=1}^{p}\lambda_{n,j}\left|\alpha_{0j}+\frac{u_j}{\sqrt{n\,\Delta_n}}\right| - \sum_{j=1}^{p}\lambda_{n,j}\left|\alpha_{0j}\right|\right) \\ &\quad + \left(\sum_{k=1}^{q}\gamma_{n,k}\left|\beta_{0k}+\frac{v_j}{\sqrt{n}}\right| - \sum_{j=1}^{q}\gamma_{n,k}\left|\beta_{0k}\right|\right)\end{aligned}$$

$$
\begin{aligned}
&= z'\varphi(n)^{1/2}\ddot{\mathbb{H}}_n\bigl(\mathbf{X}_n,\tilde{\theta}_n\bigr)\varphi(n)^{1/2}z \\
&\quad + 2z'\varphi(n)^{1/2}\ddot{\mathbb{H}}_n\bigl(\mathbf{X}_n,\tilde{\theta}_n\bigr)\varphi(n)^{1/2}\varphi(n)^{-1/2}\bigl(\theta_0-\tilde{\theta}_n\bigr) \\
&\quad + \left(\sum_{j=1}^{p}\lambda_{n,j}\left|\alpha_{0j}+\frac{u_j}{\sqrt{n\Delta_n}}\right| - \sum_{j=1}^{p_0}\lambda_{n,j}\left|\alpha_{0j}\right|\right) \\
&\quad + \left(\sum_{k=1}^{q}\gamma_{n,k}\left|\beta_{0k}+\frac{v_j}{\sqrt{n}}\right| - \sum_{j=1}^{q_0}\gamma_{n,k}\left|\beta_{0k}\right|\right) \\
&\geq z'\varphi(n)^{1/2}\ddot{\mathbb{H}}_n\bigl(\mathbf{X}_n,\tilde{\theta}_n\bigr)\varphi(n)^{1/2}z \\
&\quad + 2z'\varphi(n)^{1/2}\ddot{\mathbb{H}}_n\bigl(\mathbf{X}_n,\tilde{\theta}_n\bigr)\varphi(n)^{1/2}\varphi(n)^{-1/2}\bigl(\theta_0-\tilde{\theta}_n\bigr) \\
&\quad + \sum_{j=1}^{p_0}\lambda_{n,j}\left(\left|\alpha_{0j}+\frac{u_j}{\sqrt{n\Delta_n}}\right| - \left|\alpha_{0j}\right|\right) \\
&\quad + \sum_{k=1}^{q_0}\gamma_{n,k}\left(\left|\beta_{0k}+\frac{v_j}{\sqrt{n}}\right| - \gamma_{n,k}\left|\beta_{0k}\right|\right) \\
&\geq z'\varphi(n)^{1/2}\ddot{\mathbb{H}}_n\bigl(\mathbf{X}_n,\tilde{\theta}_n\bigr)\varphi(n)^{1/2}z \\
&\quad + 2z'\varphi(n)^{1/2}\ddot{\mathbb{H}}_n\bigl(\mathbf{X}_n,\tilde{\theta}_n\bigr)\varphi(n)^{1/2}\varphi(n)^{-1/2}\bigl(\theta_0-\tilde{\theta}_n\bigr) \\
&\quad - \left[p_0\frac{\mu_n}{\sqrt{n\Delta_n}}|u| + q_0\frac{\nu_n}{\sqrt{n}}|v|\right] \\
&= \Xi_1+\Xi_2-\Xi_3.
\end{aligned}
$$

Now, it is clear that from Condition 1 that one has $\Xi_3 = o_p(1)$. Furthermore, by Lemma 1, being that $|z| = C$, 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

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

where $\tau_{\min}(A)$ is the minimal eigenvalue of A. Moreover, Lemma 1 implies that

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

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. This implies (6.1) and the proof is completed by noticing that $\mathcal{F}(\theta)$ is strictly convex, which implies that the consistent local minimum is the consistent global one. Then the statements of Theorem 1 hold. ■

Proof of Theorem 2. Theorem 2 is proved using contradiction. Let us assume that for some $j = p_0+1,\dots,p$ and $k = q_0+1,\dots,q$ there are $\hat{\alpha}_{n,j}\neq 0$ and

$\hat{\beta}_{n,j} \neq 0$. In view of the Karush-Kuhn-Tucker (KKT) optimality conditions, we have that

$$\frac{1}{\sqrt{n\Delta_n}} \frac{\partial \mathcal{F}(\theta)}{\partial \alpha_j}\bigg|_{\theta=\hat{\theta}_n} = 2\frac{1}{\sqrt{n\Delta_n}} \ddot{\mathbb{H}}_n^{(j)}\big(\mathbf{X}_n, \tilde{\theta}_n\big)\varphi(n)^{-1/2}\varphi(n)^{1/2}\big(\hat{\theta}_n - \tilde{\theta}_n\big) + \frac{\lambda_{n,j}}{\sqrt{n\Delta_n}}\operatorname{sgn}(\hat{\alpha}_{n,j}) = 0, \tag{6.2}$$

where $\ddot{\mathbb{H}}_n^{(j)}$ denotes the jth row of $\ddot{\mathbb{H}}_n$. The first term of (6.2) is $O_p(1)$, while $\frac{\lambda_{n,j}}{\sqrt{n\Delta_n}} \geq \frac{\kappa_n}{\sqrt{n\Delta_n}} \to \infty$. By taking into account the proof of Theorem 1, we have that $\hat{\theta}_n$ is a minimizer of $\mathcal{F}(\theta)$, and this leads to a contradiction. Therefore, $P(\hat{\alpha}_{n,j} = 0) \to 1$. Similarly for the estimators of the coefficients β_k, $k = q_0 + 1, ..., q$. Indeed, by KKT optimality conditions we have that

$$\frac{1}{\sqrt{n}} \frac{\partial \mathcal{F}(\theta)}{\partial \beta_k}\bigg|_{\theta=\hat{\theta}_n} = 2\frac{1}{\sqrt{n}} \ddot{\mathbb{H}}_n^{(k)}\big(\mathbf{X}_n, \tilde{\theta}_n\big)\varphi(n)^{-1/2}\varphi(n)^{1/2}\big(\hat{\theta}_n - \tilde{\theta}_n\big) + \frac{\lambda_{n,j}}{\sqrt{n}}\operatorname{sgn}\big(\hat{\beta}_{n,j}\big) = 0.$$

By the same arguments adopted above, we get that $P(\hat{\beta}_{n,k} = 0) \to 1$. ■

Proof of Theorem 3. Before starting the proof, it is necessary to introduce some notations. Let

$$\ddot{\mathbb{H}}_n(\tilde{\theta}_n) = \begin{pmatrix} \hat{\Gamma}_{\alpha}^{**} & \hat{\Gamma}_{\alpha}^{*\circ} & \hat{\Gamma}_{\alpha\beta}^{**} & \hat{\Gamma}_{\alpha\beta}^{*\circ} \\ \hat{\Gamma}_{\alpha}^{\circ *} & \hat{\Gamma}_{\alpha}^{\circ\circ} & \hat{\Gamma}_{\alpha\beta}^{\circ *} & \hat{\Gamma}_{\alpha\beta}^{\circ\circ} \\ \hat{\Gamma}_{\beta\alpha}^{**} & \hat{\Gamma}_{\beta\alpha}^{\circ *} & \hat{\Gamma}_{\beta}^{**} & \hat{\Gamma}_{\beta}^{*\circ} \\ \hat{\Gamma}_{\beta\alpha}^{*\circ} & \hat{\Gamma}_{\beta\alpha}^{\circ\circ} & \hat{\Gamma}_{\beta}^{\circ *} & \hat{\Gamma}_{\beta}^{\circ\circ} \end{pmatrix},$$

where

- $\hat{\Gamma}_{\alpha}^{**}$ is the $p_0 \times p_0$ matrix with elements $[\ddot{\mathbb{H}}_n]_{hi}$, $h, i = 1, ..., p_0$,
- $\hat{\Gamma}_{\alpha}^{*\circ}$ is the $p_0 \times (p - p_0)$ matrix with elements $[\ddot{\mathbb{H}}_n]_{hi}$, $h = 1, ..., p_0$, $i = p_0 + 1, ..., p$, and $\hat{\Gamma}_{\alpha}^{\circ *} = (\hat{\Gamma}_{\alpha}^{*\circ})'$,
- $\hat{\Gamma}_{\alpha}^{\circ\circ}$ is the $(p - p_0) \times (p - p_0)$ matrix with elements $[\ddot{\mathbb{H}}_n]_{hi}$, $h, i = p_0 + 1, ..., p$,
- $\hat{\Gamma}_{\beta}^{**}$ is the $q_0 \times q_0$ matrix with elements $[\ddot{\mathbb{H}}_n]_{hi}$, $h, i = p + 1, ..., p + q_0$,
- $\hat{\Gamma}_{\beta}^{*\circ}$ is the $q_0 \times (q - q_0)$ matrix with elements $[\ddot{\mathbb{H}}_n]_{hi}$, $h = p + 1, ..., p + q_0$, $i = p + q_0 + 1, ..., p + q$, and $\hat{\Gamma}_{\beta}^{\circ *} = (\hat{\Gamma}_{\beta}^{*\circ})'$,
- $\hat{\Gamma}_{\beta}^{\circ\circ}$ is the $(q - q_0) \times (q - q_0)$ matrix with elements $[\ddot{\mathbb{H}}_n]_{hi}$, $h, i = p + q_0 + 1, ..., p + q$,

- $\hat{\Gamma}^{**}_{\alpha\beta}$ is the $p_0 \times q_0$ matrix with elements $[\ddot{\mathbb{H}}_n]_{hi}$, $h = 1, ..., p_0$, $i = p + 1, ..., p + q_0$,
- $\hat{\Gamma}^{*\circ}_{\alpha\beta}$ is the $p_0 \times (q - q_0)$ matrix with elements $[\ddot{\mathbb{H}}_n]_{hi}$, $h = 1, ..., p_0$, $i = p + q_0 + 1, ..., p + q$,
- $\hat{\Gamma}^{\circ\circ}_{\alpha\beta}$ is the $(p - p_0) \times (q - q_0)$ matrix with elements $[\ddot{\mathbb{H}}_n]_{hi}$, $h = p_0 + 1, ..., p$, $i = p + q_0 + 1, ..., p + q$,
- $\hat{\Gamma}^{\circ *}_{\alpha\beta}$ is the $(p - p_0) \times q_0$ matrix with elements $[\ddot{\mathbb{H}}_n]_{hi}$, $h = p_0 + 1, ..., p$, $i = p + 1, ..., p + q_0$,

and finally,

$$\begin{pmatrix} \hat{\Gamma}^{**}_{\beta\alpha} & \hat{\Gamma}^{\circ *}_{\beta\alpha} \\ \hat{\Gamma}^{*\circ}_{\beta\alpha} & \hat{\Gamma}^{\circ\circ}_{\beta\alpha} \end{pmatrix} = \begin{pmatrix} \hat{\Gamma}^{**}_{\alpha\beta} & \hat{\Gamma}^{*\circ}_{\alpha\beta} \\ \hat{\Gamma}^{\circ *}_{\alpha\beta} & \hat{\Gamma}^{\circ\circ}_{\alpha\beta} \end{pmatrix}'.$$

Furthermore, by Lemma 1,

$$\frac{1}{n\Delta_n} \begin{pmatrix} \hat{\Gamma}^{**}_{\alpha} & \hat{\Gamma}^{*\circ}_{\alpha} \\ \hat{\Gamma}^{\circ *}_{\alpha} & \hat{\Gamma}^{\circ\circ}_{\alpha} \end{pmatrix} \xrightarrow{p} \Gamma_\alpha = \begin{pmatrix} \Gamma^{**}_{\alpha} & \Gamma^{*\circ}_{\alpha} \\ \Gamma^{\circ *}_{\alpha} & \Gamma^{\circ\circ}_{\alpha} \end{pmatrix},$$

where

- $\Gamma^{**}_{\alpha} = [\mathcal{I}^{kj}_b(\alpha_0)]_{k,j}$, is a $p_0 \times p_0$ matrix, with $k, j = 1, \ldots, p_0$,
- $\Gamma^{*\circ}_{\alpha} = [\mathcal{I}^{kj}_b(\alpha_0)]_{k,j}$, is a $p_0 \times (p - p_0)$ matrix, with $k = 1, \ldots, p_0$; $j = p_0 + 1, \ldots, p$, and $\Gamma^{\circ *}_{\alpha} = (\Gamma^{*\circ}_{\alpha})'$,
- $\Gamma^{\circ\circ}_{\alpha} = [\mathcal{I}^{kj}_b(\alpha_0)]_{k,j}$, is a $(p - p_0) \times (p - p_0)$ matrix, with $k, j = p_0 + 1, \ldots, p$,

and

$$\frac{1}{n} \begin{pmatrix} \hat{\Gamma}^{**}_{\beta} & \hat{\Gamma}^{*\circ}_{\beta} \\ \hat{\Gamma}^{\circ *}_{\beta} & \hat{\Gamma}^{\circ\circ}_{\beta} \end{pmatrix} \xrightarrow{p} \Gamma_\beta = \begin{pmatrix} \Gamma^{**}_{\beta} & \Gamma^{*\circ}_{\beta} \\ \Gamma^{\circ *}_{\beta} & \Gamma^{\circ\circ}_{\beta} \end{pmatrix},$$

where

- $\Gamma^{**}_{\beta} = [\mathcal{I}^{kj}_\sigma(\beta_0)]_{k,j}$, is a $q_0 \times q_0$ matrix, with $k, j = 1, \ldots, q_0$,
- $\Gamma^{*\circ}_{\beta} = [\mathcal{I}^{kj}_\sigma(\beta_0)]_{k,j}$, is a $q_0 \times (q - q_0)$ matrix, with $k = 1, \ldots, q_0$; $j = q_0 + 1, \ldots, q$, and $\Gamma^{\circ *}_{\beta} = (\Gamma^{*\circ}_{\beta})'$,
- $\Gamma^{\circ\circ}_{\beta} = [\mathcal{I}^{kj}_\sigma(\beta_0)]_{k,j}$, is a $(q - q_0) \times (q - q_0)$ matrix, with $k, j = q_0 + 1, \ldots, q$.

Finally, by Lemma 1 we have that

$$\frac{1}{n\sqrt{\Delta_n}} \begin{pmatrix} \hat{\Gamma}^{**}_{\alpha\beta} & \hat{\Gamma}^{*\circ}_{\alpha\beta} \\ \hat{\Gamma}^{\circ *}_{\alpha\beta} & \hat{\Gamma}^{\circ\circ}_{\alpha\beta} \end{pmatrix} \xrightarrow{p} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{6.3}$$

From Theorem 2 and (2.4), it follows that estimator $\hat{\theta}_n^*$ globally minimizes the objective function

$$\begin{aligned}\mathcal{F}_0(\theta^*) &= \big(\alpha^* - \tilde{\alpha}_n^*, -\tilde{\alpha}_n^\circ, \beta^* - \tilde{\beta}_n^*, -\tilde{\beta}_n^\circ\big)' \ddot{\mathbb{H}}_n(\tilde{\theta}_n) \big(\alpha^* - \tilde{\alpha}_n^*, -\tilde{\alpha}_n^\circ, \beta^* - \tilde{\beta}_n^*, -\tilde{\beta}_n^\circ\big) \\ &\quad + \sum_{j=1}^{p} \lambda_{n,j}|\alpha_j| + \sum_{k=1}^{q} \gamma_{n,k}|\beta_k| \\ &= \big(\alpha^* - \tilde{\alpha}_n^*\big)' \hat{\Gamma}_\alpha^{**} \big(\alpha^* - \tilde{\alpha}_n^*\big) - 2\big(\alpha^* - \tilde{\alpha}_n^*\big)' \hat{\Gamma}_\alpha^{*\circ} \tilde{\alpha}_n^\circ \\ &\quad + \big(\tilde{\alpha}_n^\circ\big)' \hat{\Gamma}_\alpha^{\circ\circ} \tilde{\alpha}_n^\circ + \sum_{j=1}^{p} \lambda_{n,j}|\alpha_j| \\ &\quad + 2\big(\beta^* - \tilde{\beta}_n^*\big)' \hat{\Gamma}_{\beta\alpha}^{**} \big(\alpha^* - \tilde{\alpha}_n^*\big) - 2\big(\tilde{\beta}_n^\circ\big)' \hat{\Gamma}_{\beta\alpha}^{*\circ} \big(\alpha^* - \tilde{\alpha}_n^*\big) \\ &\quad + 2\big(\tilde{\beta}_n^\circ\big)' \hat{\Gamma}_{\beta\alpha}^{\circ\circ} \tilde{\alpha}_n^\circ + \big(\beta^* - \tilde{\beta}_n^*\big)' \hat{\Gamma}_\beta^{**} \big(\beta^* - \tilde{\beta}_n^*\big) - 2\big(\beta^* - \tilde{\beta}_n^*\big)' \hat{\Gamma}_\beta^{*\circ} \tilde{\beta}_n^\circ \\ &\quad + \big(\tilde{\beta}_n^\circ\big)' \hat{\Gamma}_\beta^{\circ\circ} \tilde{\beta}_n^\circ + \sum_{k=1}^{q} \gamma_{n,k}|\beta_k| - 2\big(\beta^* - \tilde{\beta}_n^*\big)' \hat{\Gamma}_{\beta\alpha}^{\circ *} \tilde{\alpha}_n^\circ.\end{aligned}$$

Hence, the following normal equations hold:

$$\begin{aligned}0 = \frac{1}{2} \frac{\partial \mathcal{F}_0(\theta)}{\partial \alpha^*}\bigg|_{\theta^* = \hat{\theta}_n^*} &= \hat{\Gamma}_\alpha^{**} \big(\hat{\alpha}_n^* - \tilde{\alpha}_n^*\big) - \hat{\Gamma}_\alpha^{*\circ} \tilde{\alpha}_n^\circ \\ &\quad + \hat{\Gamma}_{\alpha\beta}^{**} \big(\hat{\beta}_n^* - \tilde{\beta}_n^*\big) - \hat{\Gamma}_{\alpha\beta}^{*\circ} \tilde{\beta}_n^\circ + A\big(\hat{\alpha}_n^*\big), \qquad \textbf{(6.4)}\end{aligned}$$

$$\begin{aligned}0 = \frac{1}{2} \frac{\partial \mathcal{F}_0(\theta)}{\partial \beta^*}\bigg|_{\theta^* = \hat{\theta}_n^*} &= \hat{\Gamma}_\beta^{**} \big(\hat{\beta}_n^* - \tilde{\beta}_n^*\big) - \hat{\Gamma}_\beta^{*\circ} \tilde{\beta}_n^\circ \\ &\quad + \hat{\Gamma}_{\beta\alpha}^{**} \big(\hat{\alpha}_n^* - \tilde{\alpha}_n^*\big) - \hat{\Gamma}_{\beta\alpha}^{\circ *} \tilde{\alpha}_n^\circ + B\big(\hat{\beta}_n^*\big), \qquad \textbf{(6.5)}\end{aligned}$$

where $A(\hat{\alpha}_n^*)$ and $B(\hat{\beta}_n^*)$ are, respectively, p_0 and q_0 vectors with jth and kth components given by $\frac{1}{2}\lambda_{n,j}\operatorname{sgn}(\hat{\alpha}_{n,j}^*)$ and $\frac{1}{2}\gamma_{n,k}\operatorname{sgn}(\hat{\beta}_{n,j}^*)$. From (6.4), by simple calculations, we have that

$$\begin{aligned}\sqrt{n\Delta_n}\big(\hat{\alpha}_n^* - \alpha_0^*\big) &= \sqrt{n\Delta_n}\big(\tilde{\alpha}_n^* - \alpha_0^*\big) + \left(\frac{1}{n\Delta_n}\hat{\Gamma}_\alpha^{**}\right)^{-1} \frac{1}{n\Delta_n} \hat{\Gamma}_\alpha^{*\circ}\big(\sqrt{n\Delta_n}\tilde{\alpha}_n^\circ\big) \\ &\quad - \left(\frac{1}{n\Delta_n}\hat{\Gamma}_\alpha^{**}\right)^{-1} \left(\frac{1}{n\sqrt{\Delta_n}}\hat{\Gamma}_{\alpha\beta}^{**}\right) \sqrt{n}\big(\hat{\beta}_n^* - \tilde{\beta}_n^*\big) \\ &\quad + \left(\frac{1}{n\Delta_n}\hat{\Gamma}_\alpha^{**}\right)^{-1} \left(\frac{1}{n\sqrt{\Delta_n}}\hat{\Gamma}_{\alpha\beta}^{*\circ}\right) \big(\sqrt{n}\tilde{\beta}_n^\circ\big) \\ &\quad - \left(\frac{1}{n\Delta_n}\hat{\Gamma}_\alpha^{**}\right)^{-1} \frac{A(\hat{\alpha}_n^*)}{\sqrt{n\Delta_n}}.\end{aligned}$$

By Condition 1 we have that $\frac{A(\hat{\alpha}_n^*)}{\sqrt{n\Delta_n}} = o_p(1)$, and (6.3) leads to $\frac{1}{n\sqrt{\Delta_n}}\hat{\Gamma}_{\alpha\beta}^{*\circ} = o_p(1)$ and $\frac{1}{n\sqrt{\Delta_n}}\hat{\Gamma}_{\alpha\beta}^{**} = o_p(1)$. Thus, we can write via Theorems 1 and 2,

$$\sqrt{n\Delta_n}(\hat{\alpha}_n^* - \alpha_0^*) = \sqrt{n\Delta_n}(\tilde{\alpha}_n^* - \alpha_0^*) + \left(\frac{1}{n\Delta_n}\hat{\Gamma}_{\alpha}^{**}\right)^{-1}\left(\frac{1}{n\Delta_n}\hat{\Gamma}_{\alpha}^{*\circ}\right)\left(\sqrt{n\Delta_n}\tilde{\alpha}_n^{\circ}\right) + o_p(1).$$

Notice that

$$\Gamma_{\alpha}^{-1} = \begin{pmatrix} \left(\Gamma_{(\alpha\alpha)}^{**}\right)^{-1} & -\left(\Gamma_{(\alpha\alpha)}^{**}\right)^{-1}\Gamma_{\alpha}^{*\circ}\left(\Gamma_{\alpha}^{\circ\circ}\right)^{-1} \\ -\left(\Gamma_{\alpha}^{\circ\circ}\right)^{-1}\Gamma_{\alpha}^{\circ *}\left(\Gamma_{(\alpha\alpha)}^{**}\right)^{-1} & \left(\Gamma_{\alpha}^{\circ\circ}\right)^{-1} + \left(\Gamma_{\alpha}^{\circ\circ}\right)^{-1}\Gamma_{\alpha}^{\circ *}\left(\Gamma_{(\alpha\alpha)}^{**}\right)^{-1}\Gamma_{\alpha}^{*\circ}\left(\Gamma_{\alpha}^{\circ\circ}\right)^{-1} \end{pmatrix},$$

where $\left(\Gamma_{(\alpha\alpha)}^{**}\right)^{-1} = \left(\Gamma_{\alpha}^{**} - \Gamma_{\alpha}^{*\circ}\left(\Gamma_{\alpha}^{\circ\circ}\right)^{-1}\Gamma_{\alpha}^{\circ *}\right)^{-1}$, then $\left(\Gamma_{(\alpha\alpha)}^{**}\right)^{-1}\Gamma_{\alpha}^{*\circ} = -\left(\Gamma_{\alpha}^{*\circ}\right)^{-1}\Gamma_{\alpha}^{\circ\circ}$. Therefore, since $\left(\frac{1}{n\Delta_n}\hat{\Gamma}_{\alpha}^{**}\right)^{-1}\left(\frac{1}{n\Delta_n}\hat{\Gamma}_{\alpha}^{*\circ}\right) \xrightarrow{p} \left(\Gamma_{(\alpha\alpha)}^{**}\right)^{-1}\Gamma_{\alpha}^{*\circ}$, we have that

$$\sqrt{n\Delta_n}\left(\hat{\alpha}_n^* - \alpha_0^*\right) = \sqrt{n\Delta_n}\left(\tilde{\alpha}_n^* - \alpha_0^*\right) - \left(\Gamma_{\alpha}^{*\circ}\right)^{-1}\Gamma_{\alpha}^{\circ\circ}\left(\sqrt{n\Delta_n}\tilde{\alpha}_n^{\circ}\right) + o_p(1),$$

and by means of Lemma 1(ii), we derive that

$$\sqrt{n\Delta_n}\left(\hat{\alpha}_n^* - \alpha_0^*\right) \xrightarrow{d} N\left(0, \left(\Gamma_{(\alpha\alpha)}^{**}\right)^{-1}\right).$$

Similarly, from (6.5) we obtain that

$$\sqrt{n}\left(\hat{\beta}_n^* - \beta_0^*\right) = \sqrt{n}\left(\tilde{\beta}_n^* - \beta_0^*\right) + \left(\frac{1}{n}\hat{\Gamma}_{\beta}^{**}\right)^{-1}\left(\frac{1}{n}\hat{\Gamma}_{\beta}^{*\circ}\right)\sqrt{n}\tilde{\beta}_n^{\circ} + o_p(1),$$

and then $\sqrt{n}\left(\hat{\beta}_n^* - \beta_0^*\right)$ converges in distribution to $N\left(0, \left(\Gamma_{(\beta\beta)}^{**}\right)^{-1}\right)$, with $\left(\Gamma_{(\beta\beta)}^{**}\right)^{-1} = \left(\Gamma_{\beta}^{**} - \Gamma_{\beta}^{*\circ}\left(\Gamma_{\beta}^{\circ\circ}\right)^{-1}\Gamma_{\beta}^{\circ *}\right)^{-1}$. This concludes the proof. ■

REFERENCES

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

Aït-Sahalia, Y. (2002) Maximum likelihood estimation of discretely sampled diffusions: A closed-form likelihood approximation approach. *Econometrica* 70, 223–262.

Andrews, D.W.K. & B. Lu (2001) Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics* 101, 123–165.

Belloni, A. & V. Chernozhukov (2011) *l*1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics* 39(1), 82–130.

Bergstrom, A.R. (1990) *Continuous Time Econometric Modelling*. Oxford University Press.

Brennan, M.J. & E. Schwartz (1980) Analyzing convertible securities. *Journal of Financial and Quantitative Analysis* 15(4), 907–929.

Byrd, R.H., P. Lu, J. Nocedal, & C. Zhu (1995) A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing* 16, 1190–1208.

Caner, M. (2009) LASSO-type GMM estimator. *Econometric Theory* 25, 270–290.

Caner, M. & K. Knight (2011) An Alternative to Unit Root Tests: Bridge Estimators Differentiate between Nonstationary versus Stationary Models and Select Optimal Lag. Working paper, Michigan State University. Available at http://econ.msu.edu/seminars/docs/Caner%20paper.pdf.

Chan, K.C., G.A. Karolyi, F.A. Longstaff, & A.B. Sanders (1992) An empirical investigation of alternative models of the short-term interest rate. *Journal of Finance* 47, 1209–1227.

Cox, J.C., J.E. Ingersoll, & S.A. Ross (1980) An analysis of variable rate loan contracts. *Journal of Finance* 35(2), 389–403.

Cox, J.C., J.E. Ingersoll, & S.A. Ross (1985) A theory of the term structure of interest rates. *Econometrica* 53, 385–408.

Croissant, Y. (2006) *Ecdat: Data sets for econometrics*. R package v.0.1-5. Available at www.r-project. org.

Dothan, U.L. (1978) On the term structure of interest rates. *Journal of Financial Economics* 6, 59–69.

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

Fan, J. & R. Li (2001) Variable selection via nonconcave penalized likelihood and its Oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.

Fan, J. & R. Li (2006) Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *Proceedings of the Madrid International Congress of Mathematicians*. European Mathematical Society.

Genon-Catalot, V. & J. Jacod (1993) On the estimation of the diffusion coefficient for multidimensional diffusion processes. *Annales de l'Institut Henri Poincaré* 29, 119–151.

Hsu, N.-J., H.-L. Hung, & Y.-M. Chang (2008) Subset selection for vector autoregressive processes using the Lasso. *Computational Statistics & Data Analysis* 52, 3645–3657.

Iacus, S.M. (2008) *Simulation and Inference for Stochastic Differential Equations*. Springer.

Kessler, M. (1997) Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics* 24, 211–229.

Kessler, M. & M. Sørensen (1999) Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli* 5, 299–314.

Knight, K. (2008) Shrinkage estimation for nearly singular designs. *Econometric Theory* 24, 323–337.

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

Leeb, H. & B. Potscher (2005) Model selection and inference: Facts and fiction. *Econometric Theory* 21(1), 21–59.

Liao, Z. (2010) Adaptive GMM Shrinkage Estimation with Consistent Moment Selection. Working paper, Yale University.

Liao, Z. & P.C.B. Phillips (2010) Automated Estimation of Vector Error Correction Models. Working paper, Yale University.

McCrorie, J.R & M.J. Chambers (2006) Granger causality and the sampling of economic processes. *Journal of Econometrics* 132(2), 311–336.

Merton, R.C. (1973) Theory of rational option pricing. *Bell Journal Economics and Management Science* 4(1), 141–183.

Milstein, G.N. (1978) A method of second-order accuracy integration of stochastic differential equations. *Theory of Probability and Its Applications* 23, 396–401.

Nardi, Y. & A. Rinaldo (2011) Autoregressive processes modeling via the Lasso procedure. *Journal of Multivariate Analysis* 102, 528–549.

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

Piazzesi, M. (2009) Affine term structure models. In Y. Aït-Sahalia & L. Hansen (eds.), *Handbook for Financial Econometrics*. North-Holland.

Sundaresan, S.M. (2000) Continuous time models in finance: A review and an assessment. *Journal of Finance* 55, 1569–1622.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*. 58, 267–288.

Uchida, M. & N. Yoshida (2001) Information criteria in model selection for mixing processes. *Statistical Inference for Stochastic Processes* 4, 73–98.

Uchida, M. & N. Yoshida (2005) AIC for Ergodic Diffusion Processes from Discrete Observations. Preprint MHF 2005-12, Kyushu University.

Vasicek, O. (1977) An equilibrium characterization of the term structure. *Journal of Financial Economics* 5, 177–188.

Wang, H. & C. Leng (2007) Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association* 102(479), 1039–1048.

Wang, H., G. Li & C.-L. Tsai (2007) Regression coefficient and autoregressive order shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*. 169(1), 63–78.

Wang, X., P.C.B. Phillips & J. Yu (2011) Bias in estimating linear multivariate diffusions. *Journal of Econometrics* 161, 228–245.

Yoshida, N. (1992) Estimation for diffusion processes from discrete observation. *Journal of Multivariate Analysis* 41(2), 220–242.

Yu, J. & P.C.B. Phillips (2001) Gaussian estimation of continuous time models of the short term interest rate. *Econometrics Journal* 4, 210–224.

Zou, H. (2006) The adaptive LASSO and its Oracle properties. *Journal of the American Statistical Association* 101(476), 1418–1429.