Annals of the Institute of Statistical Mathematics (2021) 73:921–951
https://doi.org/10.1007/s10463-020-00769-w

Regularized bridge-type estimation with multiple penalties

Alessandro De Gregorio[1] · **Francesco Iafrate**[1]

Received: 8 February 2020 / Revised: 21 September 2020 / Accepted: 2 October 2020
Published online: 9 November 2020

Abstract
The aim of this paper is to introduce an adaptive penalized estimator for identifying the true reduced parametric model under the sparsity assumption. In particular, we deal with the framework where the unpenalized estimator of the structural parameters needs simultaneously multiple rates of convergence (i.e., the so-called mixed-rates asymptotic behavior). We introduce a bridge-type estimator by taking into account penalty functions involving ℓ^q norms ($0 < q \leq 1$). We prove that the proposed regularized estimator satisfies the oracle properties. Our approach is useful for the estimation of stochastic differential equations in the parametric sparse setting. More precisely, under the high-frequency observation scheme, we apply our methodology to an ergodic diffusion and introduce a procedure for the selection of the tuning parameters. Furthermore, the paper contains a simulation study as well as a real data prediction in order to assess about the performance of the proposed bridge estimator.

Keywords High-frequency scheme · Oracle properties · Multidimensional diffusion processes · Prediction accuracy · Penalized estimation · Quasi-likelihood function

1 Introduction

Statistical learning ensures high prediction accuracy and discovers relevant predictive variables. Furthermore, variable selection is particularly important when the true underlying model has a sparse representation. Identifying significant variables will enhance the prediction performance of the fitted model. A possible way to address this issue is represented by the stepwise and subset selection procedures. Nevertheless, the main drawbacks of this approach are the computational

✉ Alessandro De Gregorio
alessandro.degregorio@uniroma1.it

Francesco Iafrate
francesco.iafrate@uniroma1.it

[1] Department of Statistical Sciences, "Sapienza" University of Rome, P.le Aldo Moro, 5, 00185 Rome, Italy

complexity and the variability. In the last twenty years, the penalized statistical methods became very popular in the variable selection framework (see, e.g., Hastie et al. 2015, Hastie et al. 2009).

Let us consider the classical linear regression model $y = \mathbf{X}\theta + \varepsilon$, where y is the response vector and $\mathbf{X}$ is the $n \times p$ predictor matrix of standardized variables, $\theta := (\theta_1, \dots, \theta_p)' \in \mathbb{R}^p$ is the parametric vector and ε is a Gaussian vector with independent components with mean zero. The least absolute shrinkage and selection operator (LASSO), introduced in Tibshirani (1996), is a useful and well studied approach to the problem of model selection. Its major advantage is the simultaneous execution of both parameter estimation and variable selection. The LASSO estimator is obtained by solving the ℓ_1 penalized least squares problem

$$\hat{\theta}(\text{LASSO}) = \arg\min_{\theta}\{|y - \mathbf{X}\theta|^2 + \lambda||\theta||_1\} \tag{1}$$

where $\lambda \geq 0$ (tuning parameter), $|\cdot|$ is the euclidean distance and $||\theta||_1 = \sum_{j=1}^{p} |\theta_j|$, or, equivalently, by dealing with an unpenalized optimization problem with a constraint

$$\hat{\theta}(\text{LASSO}) = \arg\min_{\theta}\{|y - \mathbf{X}\theta|^2\}, \quad \text{subject to} \quad ||\theta||_1 \leq t, \tag{2}$$

with $t \geq 0$. Let us notice that the ℓ_1 penalty admits some singularities. For this reason, some coefficients are shrunken to zero.

The bridge estimation generalized the LASSO approach by using ℓ^q norm (see, e.g., Frank et al. 1993; Knight and Fu 2000) as follows

$$\hat{\theta}(\text{bridge}) = \arg\min_{\theta}\left\{|y - \mathbf{X}\theta|^2 + \lambda||\theta||_q\right\}, \tag{3}$$

where $||\theta||_q = \sum_{j=1}^{p} |\theta_j|^q, q > 0$. The estimator (3) becomes LASSO for $q = 1$, and Ridge for $q = 2$. Notice that, AIC or BIC criterion can be viewed as limiting cases of bridge estimation as $q \to 0$;, i.e.,

$$\lim_{q \to 0} \sum_{j=1}^{p} |\theta_j|^q = \sum_{j=1}^{p} 1_{\theta_j \neq 0}.$$

There are other possible approaches: for instance, in Fan and Li (2001), the authors proposed a shrinkage procedure based on the smoothly clipped deviation (SCAD) penalty term.

The regularization methods 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.) which use a fixed penalty on the size of a model.

As argued in Fan and Li (2001), Fan and Peng (2004) and Fan and Li (2006), a good selection procedure should have (asymptotically) the so-called oracle properties:

(i) consistently estimates null parameters as zero; i.e., the selection procedure identifies the right subset model;
(ii) has the optimal estimation rate and converges to a Gaussian random variable $N(0, \Sigma_0)$ where Σ_0 is the covariance matrix of the true subset model.

As shown in Zou (2006), since LASSO procedure assigns the same amount of penalization to each parameter, it does not represent an oracle procedure. For this reason, the author introduced the following adaptive LASSO estimator

$$\hat{\theta}_n(\text{AdaLASSO}) = \arg\min_{\theta} \left\{ |y - \mathbf{X}\theta|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\theta_j| \right\},$$

where $\lambda_n > 0$ and $(\hat{w}_j)_{j=1}^p$ are suitable data-driven weights. Usually, $\hat{w}_j = 1/|\hat{\theta}_j(\text{ols})|^\gamma, \gamma > 0$. As the sample size grows, the weights for zero-coefficient regressors increase to infinity, whereas the weights for nonzero-coefficient predictors tend to a finite constant. Thus, we are able to estimate consistently null parameters and $\hat{\theta}_n(\text{AdaLASSO})$ is oracle.

Originally, the LASSO procedure was introduced for linear regression problems, but, in the recent years, this approach has been applied to different fields of stochastic processes. In Wang et al. (2007), the problem of shrinkage estimation of regressive and autoregressive coefficients has been dealt with, while in Nardi and Rinaldo (2011) penalized order selection AR(p) models are studied. Furthermore, in Caner and Knight (2013) is shown that the bridge estimator can be used to differentiate stationarity from unit root type of non-stationarity and to select the optimal lag in AR series as well. For other issues on penalized estimation problems for time series analysis, the reader can also consult Basu and Michailidis (2015).

Very recently, regularized estimators have been applied to multidimensional diffusion processes and point processes and represent a new research topic in the field of statistics for stochastic processes. For instance, in the high-frequency framework, the reader can consult (De Gregorio and Iacus 2012; Masuda and Shimizu 2017; Suzuki and Yoshida 2019; Kinoshita and Yoshida 2019) where the authors used penalized selection procedure for discovering the underlying true model. In De Gregorio and Iacus (2018) and Gaïffas and Matulewicz (2019), LASSO and bridge estimators have been applied to continuously observed stochastic differential equations.

In this paper, we address the estimation problem for a sparse parametric model where different rates of convergence must be considered simultaneously for the asymptotic identification of the vector of structural parameters (i.e., mixed-rates asymptotics). In particular, we introduce a bridge-type estimator by means of the least squares approximation approach developed in Wang and Leng (2007). The main idea is represented by an objective function with multiple adaptive penalty functions involving ℓ^q norms ($0 < q \leq 1$). We will show that the estimator obtained by minimizing the above objective function satisfies the oracle properties.

Our motivating example is represented by a multidimensional ergodic diffusion process solution to a stochastic differential equation; i.e.,

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

Under the high-frequency observation scheme, $\alpha \in \mathbb{R}^{p_1}$ and $\beta \in \mathbb{R}^{p_2}$ admit optimal estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$ having two different asymptotic rates (i.e., infill asymptotic). We will assume the sparsity condition for the above diffusion process; i.e., some coefficients of the true values of α and β are exactly equal to zero. We will apply our methodology to this framework. Actually, it would be possible to deal with other types of random models described by stochastic differential equations: for instance, small diffusions (see, e.g., Sørensen and Uchida 2003; Gloter and Sørensen 2009), and diffusions with jumps (see, e.g., Shimizu and Yoshida 2006; Clément and Gloter 2019; Masuda 2019). Such as random processes define parametric models with two groups of parameters having two different asymptotic rates.

Furthermore, there are several econometric models where the exact evaluation of the structural parameters requires estimators with mixed-rates asymptotics. For instance, in Antoine and Renault (2012) the asymptotic theory of GMM inference is extended; the main goal of this work is to allow sample counterparts of the estimating equations to converge at (multiple) rates, different from the usual square-root of the sample size. Moreover, in Antoine and Renault (2012) are provided some econometric examples where the mixed-rates behavior arises. In Lee (2004), the theoretical properties of the maximum likelihood estimator and the quasi-maximum likelihood estimator for the spatial autoregressive model are investigated. The rates of convergence of those estimators may depend on some general features of the spatial weights matrix of the model. When each unit can be influenced by many neighbors, irregularity of the information matrix may occur and various components of the estimators may have different rates of convergence.

The paper is organized as follows. In Sect. 2, we introduce the estimation parametric problem. The setup is that of a parametric model with unpenalized estimators that asymptotically behave well under multiple rates of convergence. In this setting, we introduce a regularized estimator by means of the least squares approximation approach developed in Wang and Leng (2007). Section 3 contains the discussion of the oracle properties of the introduced estimator. Section 4 is devoted to the application of our methodology to diffusion processes related to stochastic differential equations. Furthermore, a selection procedure for tuning parameters, based on the standardized residuals of the discretized sample path, is proposed. In order to evaluate the performance of our estimator, in Sect. 5 a simulation study on a linear diffusion process is carry on to select the true underlying model. In the same section, we test the prediction accuracy of our methodology for four financial time series of daily closing stock prices of major tech companies: Google, Amazon, Apple and Microsoft. Besides, in the framework of ergodic diffusions, we compare the bridge estimator introduced in this paper and the disjoint method developed in Suzuki and Yoshida (2019). All the proofs of the oracle properties are collected in the last section.

2 Adaptive bridge-type estimation with multiple penalties

Let us introduce the shrinking estimator in a general setup. We deal with a parameter of interest $\theta := (\theta^1, \dots, \theta^m)'$, where $\theta^i := (\theta^i_1, \dots, \theta^i_{p_i}), p_i \in \mathbb{N}, i = 1, \dots, m$. Furthermore, $\theta \in \Theta := \Theta_1 \times \cdots \times \Theta_m \subset \mathbb{R}^{\mathfrak{p}}, \mathfrak{p} := \sum_{i=1}^m p_i$, where Θ_i is a bounded convex subset of $\mathbb{R}^{p_i}$. We denote by $\theta_0 := (\theta^1_0, \dots, \theta^m_0)'$, where $\theta^i_0 := (\theta^i_{0,1}, \dots, \theta^i_{0,p_i}), i = 1, \dots, m$, the true value of θ.

Assume that there exists a loss function $\theta \mapsto \mathfrak{L}_n(\theta)$ and

$$\tilde{\theta}_n = (\tilde{\theta}^1_n, \dots, \tilde{\theta}^m_n)' \in \arg\min_{\theta} \mathfrak{L}_n(\theta).$$

Usually, $\mathfrak{L}_n(\theta)$ is a (negative) log-likelihood function or the sum of squared residuals. Furthermore, suppose that $\tilde{\theta}_n$ admits a mixed-rates asymptotic behavior in the sense of Radchenko (2008); that is for the asymptotic estimation of $\theta^i_0, i = 1, \dots, m$, is necessary to consider simultaneously different rates of convergence for $\tilde{\theta}^i_n, i = 1, \dots, m$.

We assume that θ_0 is sparse (i.e., some components of θ_0 are exactly zero). Let $p^0_i := |\{j : \theta^i_{0,j} \neq 0\}|, i = 1, \dots, m$, and $\mathfrak{p}^0 := \sum_{i=1}^m p^0_i$. Therefore, our target is the identification of the true model θ_0 by exploiting a multidimensional random sample $(X_n)_n$ on the probability space $(\Omega, \mathcal{F}, P)$.

In order to carry out simultaneously estimation and variable selection, we use a penalized approach involving suitable shrinking terms. Since we have to take into account the multiple asymptotic behavior of the non-regularized estimator $\tilde{\theta}_n$, we suggest to penalize different sets of parameters with different norms. Therefore, the adaptive objective function with weighted ℓ^{q_i} penalties should be given by

$$\mathfrak{L}_n(\theta) + \left[\sum_{j=1}^{p_1} \lambda^1_{n,j} |\theta^1_j|^{q_1} + \ldots + \sum_{j=1}^{p_m} \lambda^m_{n,j} |\theta^m_j|^{q_m} \right], \quad q_i \in (0, 1], i = 1, \dots, m, \tag{4}$$

where $(\lambda^i_{n,j})_{n \geq 1}, j = 1, \dots, p_i, i = 1, \dots, m$, are sequences of real positive random variable representing an adaptive amount of the shrinkage for each element of θ^i. The bridge-type estimator is the minimizer of the objective function (4), which reduce to the LASSO-type estimator if $q_i = 1$ for any i. This is a nonlinear optimization problem which might be numerically challenging to solve. By resorting the least squares approximation approach developed in Wang and Leng (2007) and Suzuki and Yoshida (2019), we can replace (4) with a more tractable objective function. Indeed, if $\mathfrak{L}_n$ is twice differentiable with respect to θ, we have

$$\begin{aligned} \mathfrak{L}_n(\theta) &\simeq \mathfrak{L}_n(\tilde{\theta}_n) + (\theta - \tilde{\theta}_n)' \nabla_\theta \mathfrak{L}_n(\tilde{\theta}_n) + \frac{1}{2}(\theta - \tilde{\theta}_n)' \ddot{\mathfrak{L}}_n(\tilde{\theta}_n)(\theta - \tilde{\theta}_n) \\ &= \mathfrak{L}_n(\tilde{\theta}_n) + \frac{1}{2}(\theta - \tilde{\theta}_n)' \ddot{\mathfrak{L}}_n(\tilde{\theta}_n)(\theta - \tilde{\theta}_n), \end{aligned}$$

where $\ddot{\mathfrak{L}}_n$ represents the Hessian matrix. Therefore, we may minimize instead of (4) the following objective function with multiple penalty terms

$$(\theta-\tilde{\theta}_n)'\ddot{\mathfrak{L}}_n(\tilde{\theta}_n)(\theta-\tilde{\theta}_n)+\left[\sum_{j=1}^{p_1}\lambda_{n,j}^1|\theta_j^1|^{q_1}+\cdots+\sum_{j=1}^{p_m}\lambda_{n,j}^m|\theta_j^m|^{q_m}\right]. \tag{5}$$

The gain of (5) is twofold: it reduces the computational complexity of (4); furthermore, the least squares term allows to unify many different types of penalized objective functions.

Now, inspired by (5), we are able to define the adaptive penalized estimator studied in this paper.

Definition 1 Let $\hat{G}_n$ be a $\mathfrak{p}\times\mathfrak{p}$ almost surely positive definite symmetric random matrix depending on n. We define the adaptive bridge-type estimator $\hat{\theta}_n : \mathbb{R}^{(n+1)\times d}\to\overline{\Theta}$ as follows

$$\hat{\theta}_n=(\hat{\theta}_n^1,\dots,\hat{\theta}_n^m)'\in\arg\min_{\theta\in\overline{\Theta}}\mathcal{F}_n(\theta) \tag{6}$$

where

$$\mathcal{F}_n(\theta):=(\theta-\tilde{\theta}_n)'\hat{G}_n(\theta-\tilde{\theta}_n)+\left[\sum_{j=1}^{p_1}\lambda_{n,j}^1|\theta_j^1|^{q_1}+\cdots+\sum_{j=1}^{p_m}\lambda_{n,j}^m|\theta_j^m|^{q_m}\right], \tag{7}$$

with $q_i\in(0,1]$, $i=1,\dots,m$.

Clearly, (7) reduces to (5) if $\hat{G}_n:=\ddot{\mathfrak{L}}_n(\tilde{\theta}_n)$.

The estimator (6) coincides with the estimator introduced in Suzuki and Yoshida (2019) for $m=1$ (actually, they are slightly different because we will require different asymptotic conditions on the matrix $\hat{G}_n$). Our scope is to generalize the approach developed in Suzuki and Yoshida (2019), in order to extend the bridge-type methodology to statistical parametric models with multiple rates of convergence. Thus, for this reason, the objective function (7) involves different norms, one for each set of parameters. For instance, in the case of ergodic diffusions the shrinking estimator (6) is theoretical equivalent (see Theorem 4 below) to its counterpart studied in Suzuki and Yoshida (2019), Sect. 7.1. Furthermore, when we apply the bridge-type estimation procedure, it is necessary to work in the finite sample size setting. Therefore, our methodology, based on the joint estimation, is able to take into account the cross-correlations between the variables of the model, by means of the random matrix $\hat{G}_n$. This last issue is a crucial point in the statistical learning, where the correct identification of the dependent variables improves the performance of the fitted model. These features could be lost if we split the penalized estimation of the parameters.

3 Oracle properties

For the sake of simplicity, hereafter, we assume $\theta^i_{0,j} \neq 0, j = 1, \dots, p^0_i$, for any $i = 1, \dots, m$. We deal with $r^i_n, i = 1, \dots, m$, representing sequences of positive numbers tending to 0 as $n \to \infty$. $\mathbf{I}_m$ stands for the identity matrix of size m. Furthermore, we introduce the following matrices

$$A_n := \text{diag}(r^1_n \mathbf{I}_{p_1}, \dots, r^m_n \mathbf{I}_{p_m}).$$

We main assumptions in the paper are the following ones.

A1. Let $\hat{\mathfrak{D}}_n := A_n \hat{G}_n A_n$. There exists a $\mathfrak{p} \times \mathfrak{p}$ positive definite symmetric random matrix G such that

$$\hat{\mathfrak{D}}_n \xrightarrow{p} G.$$

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

$$A_n^{-1}(\tilde{\theta}_n - \theta_0) = \left(\frac{1}{r^1_n}(\tilde{\theta}^1_n - \theta^1_0), \dots, \frac{1}{r^m_n}(\tilde{\theta}^m_n - \theta^m_0) \right)' = O_p(1).$$

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

$$A_n^{-1}(\tilde{\theta}_n - \theta_0) = \left(\frac{1}{r^1_n}(\tilde{\theta}^1_n - \theta^1_0), \dots, \frac{1}{r^m_n}(\tilde{\theta}^m_n - \theta^m_0) \right)' \xrightarrow{d} N_{\mathfrak{p}}(0, \mathfrak{F}),$$

where $\mathfrak{F} := \Gamma^{-1}$ and Γ is a $\mathfrak{p} \times \mathfrak{p}$ positive definite symmetric matrix.

The conditions A2 and A3 reveal the mixed-rates asymptotic behavior of the estimator $\tilde{\theta}_n$. Actually, A3 could be replaced with a stronger condition involving the stable convergence to a mixed normal random variable (see, e.g., Suzuki and Yoshida 2019).

Let us denote by $a^i_n := \max\{\lambda^i_{n,j}, j \leq p^0_i\}$, $b^i_n := \min\{\lambda^i_{n,j}, j > p^0_i\}$; we introduce the following conditions.

B1. $r^i_n a^i_n = O_p(1)$ for any $i = 1, \dots, m$.
B2. $r^i_n a^i_n = o_p(1)$ for any $i = 1, \dots, m$.
B3. $(r^i_n)^{2-q_i} b^i_n \xrightarrow{p} \infty$, for any $i = 1, \dots, m$.

The main goal of this section is to argue on the theoretical features of the regularized statistical procedure arising from Definition 1; i.e., we are able to prove that the estimator $\hat{\theta}_n$ is asymptotically oracle.

Theorem 1 (Consistency). *Assume A1, A2 and B1, then*

$$A_n^{-1}(\hat{\theta}_n - \theta_0) = O_p(1).$$

For the vectors $x^i = (x_1, \dots, x_{p_i})'$, we deal with the following notation: $x^i_\star := (x_1, \dots, x_{p_i^0})'$, $x^i_\bullet := (x_{p_i^0+1}, \dots, x_{p_i})'$.

Theorem 2 (Selection consistency). *If the assumptions A1, A2, B1 and B3 are satisfied, we have that*

$$P(\hat{\theta}^i_{n\bullet} = 0) \longrightarrow 1, \quad i = 1, \dots, m,$$

as $n \longrightarrow \infty$.

Theorem 2 allows to claim that with probability tending to 1, all of the zero parameters must be estimated as 0. Theorem 1 leads to the consistency of the estimators of the nonzero coefficients. Both theorems imply that the bridge estimator (6) identifies the true model consistently.

In what follows, we adopt the following notation: let M be a partitioned $p_i \times p_j$ matrix, $1 \le i, j \le m$,

$$M = \begin{pmatrix} M_{\star\star} & M_{\star\bullet} \\ M_{\bullet\star} & M_{\bullet\bullet} \end{pmatrix},$$

where the blocks are given by:

- $M_{\star\star} = (m_{ij})_{1 \le i \le p_i^0, 1 \le j \le p_j^0}$ is a $p_i^0 \times p_j^0$ matrix;
- $M_{\star\bullet} = (m_{ij})_{1 \le i \le p_i^0, p_j^0 < j \le p_j}$, is a $p_i^0 \times (p_j - p_j^0)$ matrix;
- $M_{\bullet\star} = (m_{ij})_{p_i^0 < i \le p_i, 1 \le j \le p_j^0}$ is a $(p_i - p_i^0) \times p_j^0$ matrix;
- $M_{\bullet\bullet} = (m_{ij})_{p_i^0 < i \le p_i, p_j^0 < j \le p_j}$ is a $(p_i - p_i^0) \times (p_j - p_j^0)$ matrix.

Moreover, we take into account the following assumption representing a special case of the condition A1.

C1. There exist $p_i \times p_i$ positive definite symmetric random matrices $G^{ii}, i = 1, 2, \dots, m$, such that

$$\hat{\mathfrak{D}}_n \xrightarrow{p} G := \operatorname{diag}(G^{11}, G^{22}, \dots, G^{mm}).$$

Let us assume C1 and introduce the following $\mathfrak{p}^0 \times \mathfrak{p}$ matrix

$$\mathfrak{G} := \begin{pmatrix} \mathfrak{G}_1 & 0 & 0 & \cdots & 0 \\ 0 & \mathfrak{G}_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mathfrak{G}_m \end{pmatrix},$$

where $\mathfrak{G}_i := (\mathbf{I}_{p_i^0} \ (G^{ii}_{\star\star})^{-1} G^{ii}_{\star\bullet}), i = 1, 2, \dots, m$. Now, we are able to prove the asymptotic normality of the bridge-type estimator and its efficiency with respect to the true subset model.

Theorem 3 (Asymptotic normality). *Under the assumptions C1, A2, B2 and B3, we have that*

$$\left(\frac{1}{r_n^1}(\hat{\theta}_n^1 - \theta_0^1)_\star, \dots, \frac{1}{r_n^m}(\hat{\theta}_n^m - \theta_0^m)_\star\right)' - \mathfrak{G}\left\{A_n^{-1}(\tilde{\theta}_n - \theta_0)\right\} \xrightarrow{p} 0, \tag{8}$$

as $n \longrightarrow \infty$. *Furthermore, adding A3, we obtain*

$$\left(\frac{1}{r_n^1}(\hat{\theta}_n^1 - \theta_0^1)_\star, \dots, \frac{1}{r_n^m}(\hat{\theta}_n^m - \theta_0^m)_\star\right)' \xrightarrow{d} N_{\mathfrak{p}^0}\left(0, \mathfrak{G}\,\mathfrak{I}\,\mathfrak{G}'\right), \tag{9}$$

as $n \longrightarrow \infty$, *and if* $G = \Gamma$, *one has that*

$$\mathfrak{G}\,\mathfrak{I}\,\mathfrak{G}' := \operatorname{diag}\left((\Gamma_{\star\star}^{11})^{-1}, (\Gamma_{\star\star}^{22})^{-1}, \dots, (\Gamma_{\star\star}^{mm})^{-1}\right).$$

Remark 3.1 A possible reasonable choice of the sequences of adaptive amounts is the following one (see Zou 2006)

$$\lambda_{n,j}^i = \frac{\alpha_n^i}{|\tilde{\theta}_{n,j}^i|^{\delta_i}}, \quad i = 1, 2, \dots, m, \tag{10}$$

where $(\alpha_n^i)_{n\geq 1}$ represents a sequence of positive real numbers satisfying the following conditions

$$r_n^i \alpha_n^i \longrightarrow 0, \quad (r_n^i)^{2-q_i-\delta_i}\alpha_n^i \longrightarrow \infty, \tag{11}$$

with $\delta_i > 1 - q_i$. Under the conditions (11), the assumptions B1-B3 fulfill.

Remark 3.2 It is worth to mention that the estimator $\hat{\theta}_n$ is oracle when the dimension $\mathfrak{p}$ and the sparsity dimension $\mathfrak{p}^0$ are finite and fixed. We are not considering the high-dimensional setting; i.e., $\mathfrak{p} \to \infty$ (and simultaneously $\mathfrak{p}^0 \to \infty$) as $n \to \infty$. Penalized statistics when number of parameters diverges has been studied, for instance, in Fan and Peng (2004). The study of the bridge-type estimator (6) in the high-dimensional case represents a future research topic.

4 Application to stochastic differential equations

4.1 Ergodic diffusions

Let $(\Omega, \mathcal{F}, \mathbf{F} = (\mathcal{F}_t)_{t\geq 0}, P)$ be a filtered complete probability space.

Let us consider a d-dimensional solution process $X := (X_t)_{t\geq 0}$ to the following stochastic differential equation (SDE)

$$\mathrm{d}X_t = b(X_t, \alpha)\mathrm{d}t + \sigma(X_t, \beta)\mathrm{d}W_t, \quad X_0 = x_0, \tag{12}$$

where x_0 is a deterministic initial point, $b : \mathbb{R}^d \times \Theta_\alpha \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \times \Theta_\beta \to \mathbb{R}^d \otimes \mathbb{R}^r$ are Borel known functions (up to α and β) and

$(W_t)_{t\geq 0}$ is a r-dimensional standard $\mathcal{F}_t$-Brownian motion. Furthermore, $\alpha \in \Theta_\alpha \subset \mathbb{R}^{p_1}, \beta \in \Theta_\beta \subset \mathbb{R}^{p_2}, p_1, p_2 \in \mathbb{N}$, are unknown parameters where $\Theta_\alpha, \Theta_\beta$ are compact convex sets. Let $\theta := (\alpha, \beta)' \in \Theta := \Theta_\alpha \times \Theta_\beta$ and denote by $\theta_0 := (\alpha_0, \beta_0)'$ the true value of θ. Let us assume that $\theta_0 \in \text{Int}(\Theta)$ and $0 \in \mathbb{R}^{p_1+p_2}$ belongs to Θ. The stochastic differential equation X represents a sparse parametric model; that is θ_0 has a sparse representation.

The sample path of X is observed only at $n+1$ equidistant discrete times t_i^n, such that $t_i^n - t_{i-1}^n = \Delta_n < \infty$ for $i = 1, \dots, n$, (with $t_0^n = 0$). Therefore, the data are the discrete observations of the sample path of X, that we represent by $\mathbf{X}_n := (X_{t_i^n})_{0\leq i\leq n}$. Let p an integer with $p \geq 2$, the asymptotic scheme adopted in this paper is the following: $n\Delta_n \longrightarrow \infty$, $\Delta_n \longrightarrow 0$ and $n\Delta_n^p \longrightarrow 0$ as $n \to \infty$ and there exists $\epsilon \in (0, (p-1)/p)$ such that $n^\epsilon \leq n\Delta_n$ for large n.

X satisfies some mild regularity conditions (see, e.g., Kessler 1997; Yoshida 2011). For instance, the functions b and σ are smooth, $\Sigma(x, \beta) := \sigma\sigma'(x, \beta)$ is supposed invertible and X is an ergodic diffusion; i.e., there exists a unique invariant probability measure $\mu = \mu_{\theta_0}$ such that for any bounded measurable function $g : \mathbb{R}^d \to \mathbb{R}$, the $\frac{1}{T}\int_0^T g(X_t)\mathrm{d}t \longrightarrow \int_{\mathbb{R}^d} g(x)\mu(\mathrm{d}x)$.

We are interested to the estimation of θ as well as the correct identification of the zero coefficients by using the data $\mathbf{X}_n$. For this reason, we apply the bridge-type estimator (6) in this setting. We assume that an initial estimator $\tilde{\theta}_n := (\tilde{\alpha}_n, \tilde{\beta}_n)' : \mathbb{R}^{(n+1)\times d} \longrightarrow \overline{\Theta}$ of θ satisfies the following asymptotic properties:

(i) $\tilde{\alpha}_n$ is $\sqrt{n\Delta_n}$-consistent while $\tilde{\beta}_n$ is $\sqrt{n}$-consistent; i.e., $(\sqrt{n\Delta_n}(\tilde{\alpha}_n - \alpha_0), \sqrt{n}(\tilde{\beta}_n - \beta_0))' = O_p(1)$;

(ii) $\tilde{\theta}_n$ is asymptotically normal; i.e

$$(\sqrt{n\Delta_n}(\tilde{\alpha}_n - \alpha_0), \sqrt{n}(\tilde{\beta}_n - \beta_0))' \xrightarrow{d} N_{p_1+p_2}(0, \text{diag}((\Gamma^{11})^{-1}, (\Gamma^{22})^{-1})),$$

where

$$\Gamma^{11} := \int_{\mathbb{R}^d} \big(\partial_\alpha b(\alpha_0, x)\big)' \Sigma^{-1}(\beta_0, x)\partial_\alpha b(\alpha_0, x)\mu(\mathrm{d}x)\,,$$

$$\Gamma^{22} := \frac{1}{2}\int_{\mathbb{R}^d} \text{tr}\big[(\partial_\beta \Sigma)\Sigma^{-1}(\partial_\beta \Sigma)\Sigma^{-1}(\beta_0, x)\big]\mu(\mathrm{d}x),$$

where $\partial_\alpha := (\frac{\partial}{\partial_{\alpha_1}}, \dots, \frac{\partial}{\partial_{\alpha_{p_1}}})', \partial_\beta := (\frac{\partial}{\partial_{\beta_1}}, \dots, \frac{\partial}{\partial_{\beta_{p_2}}})'$. We assume the integrability and the non-degeneracy of Γ^{11} and Γ^{22}.

Therefore, from (i) and (ii) emerge that the estimator $\tilde{\theta}_n$ works in a mixed-rates asymptotic regime with two different rates, $\sqrt{n\Delta_n}$ and $\sqrt{n}$, for the two groups of parameters α and β. The assumptions A2 and A3 hold by setting $A_n = \text{diag}(1/(\sqrt{n\Delta_n})\mathbf{I}_{p_1}, 1/\sqrt{n}\mathbf{I}_{p_2})$.

Let $q_1, q_2 \in (0, 1]$, the objective function (7) becomes

$$\mathcal{F}_n(\theta) := (\theta - \tilde{\theta}_n)'\hat{G}_n(\theta - \tilde{\theta}_n) + \sum_{j=1}^{p_1} \lambda_{n,j}|\alpha_j|^{q_1} + \sum_{k=1}^{p_2} \gamma_{n,k}|\beta_k|^{q_2}, \tag{13}$$

where $\hat{G}_n$ is a $(p_1+p_2)\times(p_1+p_2)$ matrix assumed to be symmetric and positive definite and such that $A_n\hat{G}_nA_n \xrightarrow{p} \text{diag}(\Gamma^{11},\Gamma^{22})$ (condition C1 fulfills). In this framework, we consider sequences $\lambda_{n,j}$ and $\gamma_{n,k}$ as in (10); i.e., they turn out as follows

$$\lambda_{n,j}=\frac{\lambda_{n,0}}{|\tilde{\alpha}_{n,j}|^{\delta_1}},\qquad \gamma_{n,k}=\frac{\gamma_{n,0}}{|\tilde{\beta}_{n,k}|^{\delta_2}}\qquad j=1,,\dots,p_1,\ k=1,\dots,p_2,$$

where the exponents δ_1 and δ_2 are such that $\delta_i>1-q_i,1,2$. We assume that $\lambda_{n,0}$ and $\gamma_{n,0}$ are deterministic sequences satisfying the conditions (11); that is

$$\frac{\lambda_{n,0}}{\sqrt{n\Delta_n}}\longrightarrow 0,\quad (n\Delta_n)^{\frac{\delta_1-2+q_1}{2}}\lambda_{n,0}\longrightarrow\infty,$$

and

$$\frac{\gamma_{n,0}}{\sqrt{n}}\longrightarrow 0,\quad n^{\frac{\delta_2-2+q_2}{2}}\gamma_{n,0}\longrightarrow\infty,$$

as $n\to\infty$. Finally, the bridge-type estimator for the stochastic differential equation (12) becomes

$$\hat{\theta}_n:=(\hat{\alpha}_n,\hat{\beta}_n)'\in\arg\min_{\theta\in\overline{\Theta}}\mathcal{F}_n(\theta). \tag{14}$$

In the literature appeared different estimators for ergodic diffusions satisfying the asymptotic properties (i) and (ii). For instance, the quasi-maximum-likelihood estimator, the quasi-Bayesian estimator (see, Yoshida 1992; Kessler 1997; Yu and Phillips 2001; Yoshida 2011; Uchida and Yoshida 2012, Uchida and Yoshida 2014) and the hybrid multistep estimator (see Kamatani and Uchida 2015).

For $p=2$, a suitable loss function could be the negative quasi-log-likelihood function

$$\begin{aligned}\ell_n(\mathbf{X}_n,\theta):=\frac{1}{2}\sum_{i=1}^{n}\Big\{&\log\det(\Sigma(X_{t_i^n},\beta))\\&+\frac{1}{\Delta_n}(X_{t_i^n}-X_{t_{i-1}^n}-\Delta_n b(X_{t_i^n},\alpha))'\Sigma^{-1}(X_{t_i^n},\beta)(X_{t_i^n}-X_{t_{i-1}^n}-\Delta_n b(X_{t_i^n},\alpha))\Big\}.\end{aligned} \tag{15}$$

Therefore, a possible choice of $\hat{G}_n$ is the Hessian matrix $\ddot{\ell}_n(\mathbf{X}_n,\theta)$ and

$$\tilde{\theta}_n^{QL}\in\arg\min_{\overline{\Theta}}\ell_n(\mathbf{X}_n,\theta)$$

represents the quasi-maximum likelihood estimator (QMLE). For more details on the quasi-likelihood analysis for stochastic differential equations, the reader can consult, for instance, Kessler (1997) and Yoshida (2011).

Now, we are able to present the oracle properties for the bridge-type estimator introduced in the framework of sparse ergodic diffusions. We also argue about the boundedness of the estimator which is useful for the moment convergence. Under

the assumptions of Sect. 3, the next theorem concerns these issues in the case of initial estimator equals to $\tilde{\theta}_n^{QL}$. Clearly, the statement of the theorem holds true also if $\tilde{\theta}_n$ is the quasi-Bayesian estimator or the hybrid multistep estimator.

Theorem 4 *If $\tilde{\theta}_n = \tilde{\theta}_n^{QL}$, the bridge-type estimator (14) has the following properties:*

- *(Consistency)* $(\sqrt{n\Delta_n}(\hat{\alpha}_n - \alpha_0), \sqrt{n}(\hat{\beta}_n - \beta_0)) = O_p(1)$;
- *(Selection consistency)* $P(\hat{\alpha}_{n\bullet} = 0) \longrightarrow 1$ *and* $P(\hat{\beta}_{n\bullet} = 0) \longrightarrow 1$;
- *(Asymptotic normality)*

$$(\sqrt{n\Delta_n}(\tilde{\alpha}_n - \alpha_0)_\star, \sqrt{n}(\tilde{\beta}_n - \beta_0)_\star) \xrightarrow{d} N_{p_1^0+p_2^0}(0, \mathrm{diag}((\Gamma_{\star\star}^{11})^{-1}, (\Gamma_{\star\star}^{22})^{-1}));$$

- *(Uniform L^q-boundedness) if* $\sup_n \mathbb{E}[||\hat{\mathfrak{D}}_n||^q] < \infty$ *and* $\sup_n \mathbb{E}[||\hat{\mathfrak{D}}_n^{-1}||^q] < \infty$ *for all* $q \geq 1$, *we have that*

$$\sup_n \mathbb{E}[|A_n^{-1}(\hat{\theta}_n - \theta_0)|^q] < \infty.$$

Theorem 4 represents a generalization of the results previously obtained in De Gregorio and Iacus (2012).

For the identification of the true model, it is also possible to use AIC criterion as discussed in Uchida (2011). Nevertheless, as pointed out in Iacus and Yoshida (2018), it is necessary to specify some parametric models.

4.2 Tuning parameter selection

Consider a diffusion process as in (12). In the linear regression problem, the regularized estimates are obtained by choosing the tuning parameters by means of a cross-validation procedure. Nevertheless, For this reason this technique doesn't apply because the dependency structure of the data. In our framework, we propose a data-driven technique for choosing the tuning parameters for the penalized estimation problem (14).

Consider the Euler discretization of the solution of (12)

$$X_{t_{i+1}^n} = X_{t_i^n} + b(X_{t_i^n}, \alpha)\Delta_n + \sigma(X_{t_i^n}, \beta)(W_{t_{i+1}^n} - W_{t_i^n}) \tag{16}$$

where t_i and Δ_n are specified as above. The "standardized residuals" are then defined as

$$r_{t_i^n} = \Delta_n^{-1/2}\Sigma^{-1/2}(X_{t_i^n}, \beta)(X_{t_{i+1}^n} - X_{t_i^n} - \Delta_n b(X_{t_i^n}, \alpha)) \qquad i = 1, \ldots, n. \tag{17}$$

The residuals $r_{t_i^n}$ are $N_d(0_d, \mathbf{I}_d)$ and conditionally independent. The idea is to find the tuning parameters in such a way that the residuals fit best to a white noise scheme.

Given a series of observations $\mathbf{X}_n$ and some estimates of the parameters $\hat{\alpha}$ and $\hat{\beta}$ the residuals can be estimated as

$$\hat{r}_{t_i^n} = \Delta_n^{-1/2}\Sigma^{-1/2}(X_{t_i^n}, \hat{\beta})(X_{t_{i+1}^n} - X_{t_i^n} - \Delta_n b(X_{t_i^n}, \hat{\alpha})) \qquad i = 1, \dots, n. \tag{18}$$

Let $\psi := (q_1, q_2, \lambda_n, \gamma_n, \delta_1, \delta_2)$ be the vector of tuning parameters varying in some suitable parameter space $\Psi \subset \mathbb{R}^6$. Besides, the penalized estimates will depend on ψ and consequently also the residuals will. We set $\hat{r}_{t_i^n} = \hat{r}_{t_i^n}(\psi)$ to stress this fact. We can choose a desirable value for ψ by optimizing some score function which penalizes tuning parameters producing residuals which deviate most from the hypothesis of being uncorrelated. More formally let $S : \mathbb{R}^{nd} \mapsto \mathbb{R}^+$ be such a score function which takes in input the d-dimensional residuals and returns a low score if the residuals appear to be incorrelated and a high value otherwise (low score is better). We choose the optimal value of the tuning parameter vector ψ^* as

$$\psi^* = \arg\min_{\psi \in \Psi} S(r_{t_1^n}(\psi), \dots, r_{t_n^n}(\psi)). \tag{19}$$

The penalty function can be the test statistic in a white noise hypothesis testing. In the numerical simulations, we consider the Ljung–Box test statistic defined as

$$Q_\ell(r_n) = n(n+2)\sum_{j=1}^{\ell} \frac{\hat{\rho}_j^2(r_n)}{n-j} \tag{20}$$

where $r_n = (r_{t_i^n})_{i=0}^n$ is a vector of residuals, ℓ is the number of lags to be tested, $\hat{\rho}_j$ denotes the sample auto-correlations of the residuals at lag j

$$\hat{\rho}_j^2(r_n) = \frac{\frac{1}{n-j}\sum_{i=1}^{n-j}(r_{t_i^n} - \bar{r}_n)(r_{t_{i+j}^n} - \bar{r}_n)}{\frac{1}{n}\sum_{i=1}^{n}(r_{t_i^n} - \bar{r}_n)^2,} \tag{21}$$

where $\bar{r}_n = n^{-1}\sum_{i=1}^n r_{t_i^n}$ and n is the number of observations. Under the hypothesis that the residuals are not correlated up to lag ℓ, Q_ℓ is asymptotically distributed as a χ^2_ℓ.

A similar approach can be adapted if one wants to tune the tuning parameter in order to optimize the fit of the residuals to the Gaussian distribution. This idea was introduced in Bandi et al. (2009) in the context of bandwidth selection for nonparametric estimates of the drift and diffusion coefficients. One can consider a penalty measuring the distance of the empirical distribution of the residuals from the Gaussian distribution function (such as the Kolmogorov–Smirnov test statistic). More formally, equip the space of distribution functions with some norm $\|\cdot\|$. The parameter ψ^* can then be chosen as

$$\psi^* = \arg\min_{\psi \in \Psi} \|\hat{F}_n(r_{t_1^n}(\psi), \dots, r_{t_n^n}(\psi)) - P_d\|, \tag{22}$$

where $\hat{F}_n$ denotes the empirical distribution function of the residuals and P_d denotes the distribution function of the d–dimensional standard Gaussian distribution. In

the case where the sup norm is chosen one recovers the Kolmogorov–Smirnov test statistic.

Two or more criteria can be combined in order in such a way to minimize simultaneously over multiple score functions. In the following, we consider together the criteria based on (20) and (22) in order to seek residuals which are uncorrelated and Gaussian, in the sense that they minimize both two scores. The tuning parameters are then computed as the solution of the following optimization problem

$$\psi^* = \arg\min_{\psi\in\Psi}[Q_{\ell^2}(r_n(\psi)) + \|\hat{F}_n(r_n(\psi)) - P_d\|]. \tag{23}$$

4.3 Algorithmic implementation

The following algorithm implements criterion (23).

- Step 0. Suppose a set of data points $\mathbf{X}_n$ is given. Initialize the tuning parameter vector ψ with some value ψ_0. Fix a threshold $\epsilon > 0$.
- Until convergence is reached:
  - *Step 1* Compute the current bridge estimates with the current value $\psi^{(k)}$ of the tuning parameters $\hat{\alpha}^{(k)} = \hat{\alpha}(\psi^{(k)})$, $\hat{\beta}^{(k)} = \hat{\beta}(\psi^{(k)})$.
  - *Step 2* Compute the residuals $(\hat{r}^{(k)}_{t_i^n})_{i=0}^n := (\hat{r}_{t_i^n}(\psi^{(k)}))_{i=0}^n$ as in formula (18), with the current estimates of the parameters $\hat{\alpha}^{(k)}$ and $\hat{\beta}^{(k)}$.
  - *Step 3* Evaluate the score of the current residuals $s^{(k)} = S(\hat{r}^{(k)}_{t_1^n}, \dots, \hat{r}^{(k)}_{t_n^n})$.
  - *Step 4* If $|s^{(k)} - s^{(k-1)}| < \epsilon$ stop: convergence is reached. Set $\psi^* = \psi^{(k)}$ and return the optimal bridge estimates of the parameters $\alpha^* = \alpha^{(k)}$ and $\beta^* = \beta^{(k)}$. Otherwise move to some new point $\psi^{(k+1)}$ (chosen according to some optimization algorithm) and repeat Steps 1 to 4.

5 Numerical results

5.1 Simulation study

Consider a multivariate diffusion process $X = (X_t)_{t\geq 0}$ driven by the SDE

$$\begin{cases} \mathrm{d}X_t^{(1)} = (\alpha_{10} + \alpha_{11}X_t^{(1)} + \alpha_{12}X_t^{(2)} + \alpha_{13}X_t^{(3)})\mathrm{d}t + (\beta_{10} + \beta_{11}X_t^{(1)} + \beta_{12}X_t^{(2)} + \beta_{13}X_t^{(3)})\mathrm{d}W_t^{(1)} \\ \mathrm{d}X_t^{(2)} = (\alpha_{20} + \alpha_{21}X_t^{(1)} + \alpha_{22}X_t^{(2)} + \alpha_{23}X_t^{(3)})\mathrm{d}t + (\beta_{20} + \beta_{21}X_t^{(1)} + \beta_{22}X_t^{(2)} + \beta_{23}X_t^{(3)})\mathrm{d}W_t^{(2)} \\ \mathrm{d}X_t^{(3)} = (\alpha_{30} + \alpha_{31}X_t^{(1)} + \alpha_{32}X_t^{(2)} + \alpha_{33}X_t^{(3)})\mathrm{d}t + (\beta_{30} + \beta_{31}X_t^{(1)} + \beta_{32}X_t^{(2)} + \beta_{33}X_t^{(3)})\mathrm{d}W_t^{(3)} \end{cases} \tag{24}$$

which can be written in compact matrix notation as

$$\mathrm{d}X_t = (\alpha_0 + AX_t)\mathrm{d}t + \mathrm{diag}(\beta_0 + BX_t)\,\mathrm{d}W_t \tag{25}$$

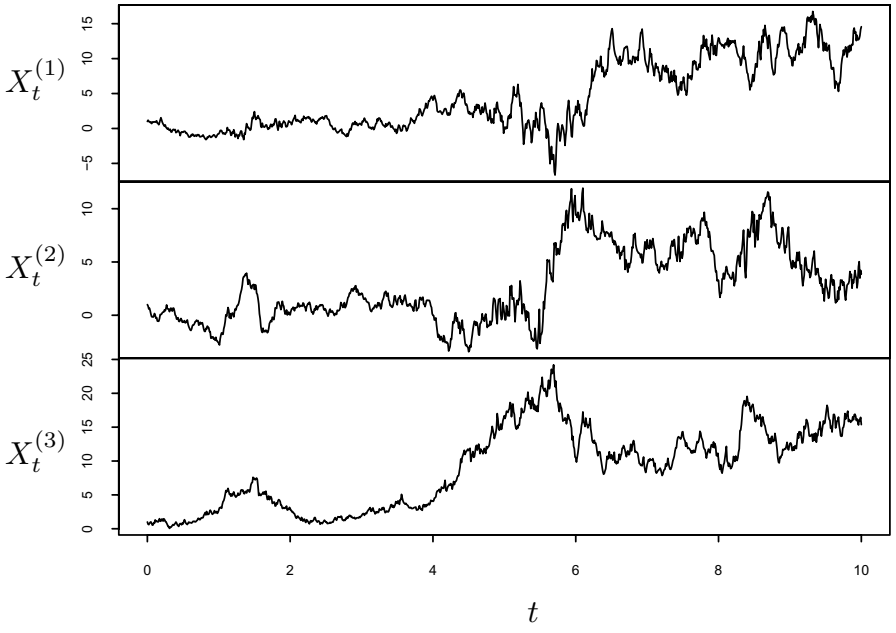

Fig. 1 A sample path of the solution to (25)

where $X_t = (X_t^{(1)}, X_t^{(2)}, X_t^{(3)})'$, $W_t = (W_t^{(1)}, W_t^{(2)}, W_t^{(3)})'$, with $W_t^{(i)}, i = 1, 2, 3$, independent Brownian motions, $\alpha_0 = (\alpha_{10}, \alpha_{20}, \alpha_{30})'$, $\beta_0 = (\beta_{10}, \beta_{20}, \beta_{30})'$, $A = (\alpha_{ij})_{1\le i,j\le 3}$, $B = (\beta_{ij})_{1\le i,j\le 3}$. A sample path of the solution to (25) is represented in Fig. 1.

In our simulation, we set several parameters to zero:

$$\alpha_{21} = \alpha_{31} = \alpha_{32} = \alpha_{33} = \beta_{11} = \beta_{21} = \beta_{31} = \beta_{22} = \beta_{32} = 0.$$

The true values for the nonzero parameters are displayed in Table 1. In particular, this choice of the parameters implies that model (25) can be interpreted in terms of Granger causality. The idea is that the three components of X_t are correlated, but the value of $X_t^{(2)}$ effects $X_t^{(1)}$, and $X_t^{(3)}$ influences both $X_t^{(1)}$ and $X_t^{(2)}$ and not vice versa. Formally, in the continuous time setting we have the following definition of non-causality (see McCrorie and Chambers 2006). Suppose $Z_t = (Y_t^{(1)}, Y_t^{(2)}, W_t)'$ is an n–dimensional process, where $Y_t^{(1)}, Y_t^{(2)}$ and W have dimension n_1, n_2 and n_3, respectively, with $n_1 + n_2 + n_3 = n$. We say that $Y_t^{(1)}$ does *not* Granger cause $Y_t^{(2)}$ if

$$\forall t, h \ge 0, \qquad \mathbb{E}(Y_{t+h}^{(2)} | \mathcal{I}_t) = \mathbb{E}(Y_{t+h}^{(2)} | \mathcal{I}_t - \mathcal{Y}_t^1), \tag{26}$$

where $\mathcal{I}_t = \sigma(Z_s, s \le t)$ and $\mathcal{I}_t - \mathcal{Y}_t^1 = \sigma((Y_s^{(2)}, W_s), s \le t)$. Clearly, in the model (24), $(X_t^{(1)}, X_t^{(2)})$ does not Granger cause $X_t^{(3)}$, by setting $Y_t^{(1)} = (X_t^{(1)}, X_t^{(2)})$ and $Y_t^{(2)} = X_t^{(3)}$ in definition (26), but the contrary is not true. Analogously, $X_t^{(1)}$ does not Granger cause $X_t^{(2)}$.

Table 1 Summary of the results of the simulation study

Par.	Bridge		LASSO		QMLE		True value
	Avg. (St. Err)	$\widehat{MSE}$	Avg. (St. Err)	$\widehat{MSE}$	Avg. (St. Err)	$\widehat{MSE}$	
(A) $n = 500$							
α_{10}	− 1.4576 (1.2738)	1.6234	− 1.6417 (1.386)	1.9397	− 1.6737 (1.3784)	1.9289	− 1.5
α_{11}	− 0.9376 (0.8198)	0.9879	− 2.0592 (0.9213)	1.161	− 2.0692 (0.9044)	1.1414	− 1.5
α_{12}	0.2629 (0.5437)	0.5327	1.0277 (0.8829)	0.8561	1.0459 (0.8748)	0.8523	0.75
α_{13}	0.776 (0.6678)	0.4463	1.0245 (0.7355)	0.616	1.0431 (0.7283)	0.616	0.75
α_{20}	− 0.3547 (0.6353)	1.7152	− 1.7171 (1.3249)	1.8013	− 1.7309 (1.3206)	1.796	− 1.5
α_{21}	0.1177 (0.3624)	0.1451	0.0286 (0.6292)	0.3964	0.0229 (0.6321)	0.3998	0
α_{22}	− 1.8366 (0.7923)	0.7406	− 2.1356 (0.8421)	1.1127	− 2.1412 (0.8399)	1.1161	− 1.5
α_{23}	0.6343 (0.4934)	0.2566	0.9938 (0.6388)	0.4672	1.01 (0.64)	0.4769	0.75
α_{30}	1.1661 (0.6206)	0.4964	2.5298 (1.1192)	2.3122	2.5465 (1.1169)	2.3417	1.5
α_{31}	0.0007 (0.4084)	0.1667	− 0.0315 (0.6012)	0.3621	− 0.0333 (0.6062)	0.3684	0
α_{32}	− 0.0064 (0.4386)	0.1923	− 0.0286 (0.6702)	0.4497	− 0.0269 (0.674)	0.4546	0
α_{33}	− 0.0689 (0.3356)	0.1173	− 0.2664 (0.5313)	0.3531	− 0.2732 (0.5281)	0.3534	0
β_{10}	1.3649 (0.6112)	0.3915	1.4397 (0.5957)	0.3583	1.4935 (0.5859)	0.3431	1.5
β_{11}	0.0521 (0.3491)	0.1245	0.0398 (0.339)	0.1164	0.0561 (0.3749)	0.1436	0
β_{12}	0.6217 (0.6415)	0.4604	0.6536 (0.7094)	0.5672	0.7554 (0.6808)	0.5894	0.4
β_{13}	0.6152 (0.5373)	0.3348	0.6254 (0.5772)	0.3837	0.6761 (0.5781)	0.4102	0.4
β_{20}	1.3751 (0.4769)	0.2429	1.457 (0.4639)	0.2169	1.5149 (0.4549)	0.2071	1.5
β_{21}	− 0.003 (0.2767)	0.0765	− 0.0136 (0.272)	0.0741	− 0.0062 (0.3039)	0.0923	0
β_{22}	0.0118 (0.2698)	0.0729	0.0039 (0.2566)	0.0658	0.0143 (0.2927)	0.0858	0
β_{23}	0.5612 (0.4646)	0.2417	0.5251 (0.5008)	0.2663	0.5854 (0.5123)	0.2966	0.4
β_{30}	1.4165 (0.5219)	0.2791	1.4205 (0.5028)	0.259	1.4754 (0.4937)	0.2442	1.5
β_{31}	0.0554 (0.3063)	0.0968	0.0467 (0.2939)	0.0885	0.058 (0.3226)	0.1074	0
β_{32}	0.0155 (0.3137)	0.0986	0.0146 (0.2896)	0.084	0.018 (0.3241)	0.1053	0
β_{33}	0.5599 (0.4971)	0.2725	0.5322 (0.5108)	0.2783	0.5907 (0.5191)	0.3056	0.4
(B) $n = 1000$							
α_{10}	− 1.5139 (1.2947)	1.676	− 1.6591 (1.3822)	1.9353	− 1.6783 (1.372)	1.9137	− 1.5
α_{11}	− 0.9119 (0.7743)	0.9453	− 2.0197 (0.9178)	1.1122	− 2.0419 (0.8857)	1.0778	− 1.5
α_{12}	0.2774 (0.5604)	0.5373	1.0386 (0.8749)	0.8485	1.0519 (0.8651)	0.8393	0.75
α_{13}	0.7671 (0.6585)	0.4338	0.9992 (0.7405)	0.6104	1.0133 (0.7236)	0.5927	0.75
α_{20}	− 0.3403 (0.6324)	1.7447	− 1.6824 (1.3291)	1.7991	− 1.6921 (1.3239)	1.789	− 1.5
α_{21}	0.1122 (0.3708)	0.15	0.0288 (0.6019)	0.3631	0.0296 (0.5981)	0.3585	0
α_{22}	− 1.8588 (0.7865)	0.7471	− 2.0806 (0.8562)	1.07	− 2.0909 (0.8389)	1.0526	− 1.5
α_{23}	0.6528 (0.4718)	0.232	0.9831 (0.6509)	0.4779	0.9988 (0.6416)	0.4734	0.75
α_{30}	1.1897 (0.585)	0.4384	2.5817 (1.0944)	2.3672	2.6016 (1.0786)	2.3765	1.5
α_{31}	0.007 (0.3917)	0.1534	− 0.0177 (0.5583)	0.312	− 0.0189 (0.5562)	0.3096	0
α_{32}	0.0022 (0.4454)	0.1984	− 0.0169 (0.6417)	0.412	− 0.0135 (0.642)	0.4122	0
α_{33}	− 0.0749 (0.3483)	0.1269	− 0.2345 (0.5209)	0.3262	− 0.2435 (0.512)	0.3213	0
β_{10}	1.3518 (0.674)	0.4761	1.4854 (0.6191)	0.3834	1.525 (0.6155)	0.3793	1.5
β_{11}	0.0389 (0.3635)	0.1336	0.0282 (0.3483)	0.122	0.0395 (0.3771)	0.1437	0
β_{12}	0.5937 (0.6152)	0.4159	0.7112 (0.6889)	0.5712	0.7741 (0.6874)	0.6123	0.4

Table 1 (continued)

Par.	Bridge		LASSO		QMLE		True value
	Avg. (St. Err)	$\widehat{MSE}$	Avg. (St. Err)	$\widehat{MSE}$	Avg. (St. Err)	$\widehat{MSE}$	
β_{13}	0.5835 (0.5317)	0.3163	0.6615 (0.5822)	0.4072	0.702 (0.593)	0.4427	0.4
β_{20}	1.3247 (0.5453)	0.328	1.4639 (0.5058)	0.2571	1.5015 (0.5093)	0.2593	1.5
β_{21}	0.0117 (0.311)	0.0968	− 0.0022 (0.2984)	0.089	0.0021 (0.3255)	0.1059	0
β_{22}	0.0137 (0.3313)	0.1099	0.0119 (0.2975)	0.0886	0.0253 (0.3294)	0.1091	0
β_{23}	0.5722 (0.5054)	0.285	0.5925 (0.5377)	0.3261	0.635 (0.5541)	0.3622	0.4
β_{30}	1.3948 (0.5406)	0.3032	1.4599 (0.4897)	0.2414	1.5036 (0.4846)	0.2348	1.5
β_{31}	0.0178 (0.2985)	0.0894	0.0153 (0.2981)	0.0891	0.0174 (0.3235)	0.105	0
β_{32}	0.004 (0.3178)	0.101	0.0035 (0.2968)	0.0881	0.0053 (0.3252)	0.1057	0
β_{33}	0.5406 (0.4914)	0.2611	0.5723 (0.5241)	0.3043	0.6183 (0.5464)	0.3461	0.4
(C) $n = 10000$							
α_{10}	− 1.5743(0.3311)	0.1096	− 1.4107(0.2665)	0.0723	− 1.4919(0.2817)	0.0799	− 1.5
α_{11}	− 0.6301(0.6105)	1.1287	− 1.7535(0.4647)	0.2797	− 1.7576(0.4825)	0.2989	− 1.5
α_{12}	0.0909(0.2492)	0.4964	0.605(0.3613)	0.1513	0.6291(0.3919)	0.168	0.75
α_{13}	1.0865(0.5839)	0.4535	1.2614(0.4958)	0.5069	1.2625(0.5173)	0.5299	0.75
α_{20}	− 0.5559(0.2808)	0.1064	− 2.1026(0.3989)	0.1631	− 2.1682(0.4172)	0.1795	− 1.5
α_{21}	0.1623(0.3033)	0.1181	0.0336(0.4542)	0.207	0.0476(0.459)	0.2127	0
α_{22}	− 1.6534(0.5847)	0.3648	− 1.7702(0.4627)	0.2867	− 1.7595(0.4739)	0.2917	− 1.5
α_{23}	0.8658(0.4269)	0.1953	1.7437(0.5375)	1.2757	1.744(0.5296)	1.2682	0.75
α_{30}	1.5828(0.3598)	0.1294	1.7036(0.2788)	0.0776	1.7197(0.2945)	0.0867	1.5
α_{31}	0.0221(0.2168)	0.0474	0.0157(0.2268)	0.0516	0.0067(0.2444)	0.0597	0
α_{32}	0.0033(0.2153)	0.0463	0.004(0.213)	0.0453	0.0085(0.2293)	0.0526	0
α_{33}	− 0.1163(0.2848)	0.0945	− 0.1964(0.3236)	0.1431	− 0.1817(0.3486)	0.1544	0
β_{10}	1.441(0.268)	0.0728	1.4335(0.2691)	0.0735	1.4754(0.2701)	0.0737	1.5
β_{11}	0.0116(0.0966)	0.0094	0.0071(0.0944)	0.009	0.007(0.1052)	0.0111	0
β_{12}	0.253(0.1999)	0.0615	0.3061(0.171)	0.038	0.3123(0.1709)	0.0369	0.4
β_{13}	0.6054(0.1768)	0.0734	0.6014(0.1801)	0.0729	0.6131(0.2053)	0.0875	0.4
β_{20}	1.8843(0.2431)	0.06	1.8948(0.2583)	0.0678	1.9126(0.2682)	0.0732	1.5
β_{21}	− 0.007(0.0702)	0.005	− 0.0055(0.073)	0.0053	− 0.0057(0.0768)	0.0059	0
β_{22}	− 0.0011(0.0509)	0.0026	− 0.0023(0.0694)	0.0048	− 0.0031(0.069)	0.0048	0
β_{23}	0.8396(0.1905)	0.2295	0.8414(0.1907)	0.2312	0.8432(0.1987)	0.2358	0.4
β_{30}	0.9928(0.2341)	0.0649	0.9846(0.2369)	0.0665	1.0082(0.256)	0.0752	1.5
β_{31}	− 0.0054(0.0512)	0.0026	− 0.0065(0.0651)	0.0043	− 0.0062(0.0647)	0.0042	0
β_{32}	− 0.0014(0.062)	0.0038	0.0002(0.0574)	0.0033	− 0.0023(0.0721)	0.0052	0
β_{33}	0.4085(0.1261)	0.0159	0.4105(0.1361)	0.0186	0.4168(0.1677)	0.0284	0.4

Remark 5.1 It is worth observing that the model (25) may not satisfy some assumption such a, for instance, the ergodicity (see Remark 1 in Uchida and Yoshida 2012, for a sufficient condition). Nevertheless, it would be possible to modify slightly the

SDE (25), in order to guarantee that the assumptions fulfill. For instance, each diffusion term in the equations (24) could be replaced with the following function

$$\sigma_i(x,\beta) = \begin{cases} \beta_{i0} + \sum_{j=1}^{3} \beta_{ij} x_j, & |\beta_{i0} + \sum_{j=1}^{3} \beta_{ij} x_j| < M, \\ M, & \text{otherwise}, \end{cases} \qquad i = 1,2,3,$$

where $M > 0$ is a positive constant sufficiently large. Therefore,

$$\Sigma(x,\beta) = \text{diag}(\sigma_1^2(x,\beta), \sigma_2^2(x,\beta), \sigma_3^2(x,\beta))$$

turns out to be bounded, which is a required condition for the ergodicity.

The aim of this simulation is the ability to recover the true model from the full model which contains a number of unnecessary relations between the variables. Moreover, we want to verify that the multiple-penalties bridge estimation technique, together with the tuning parameter calibration procedure described above, is able to identify the relevant relations among many. As a benchmark comparison, we juxtapose the results of the bridge estimation with those of the LASSO method and, with the un-penalized QMLE. In order to compute the bridge estimates, we will use as initial estimator $\tilde{\theta}_n^{QL}$ (with $p = 2$) and $\hat{G}_n = \ddot{\ell}(\mathbf{X}_n, \theta)$.

We simulated $N = 10^3$ trajectories from model (25) over a long time interval and a fine grid, according to a high-frequency sampling scheme by setting $\Delta_n = n^{-1/3}$. We tested our model with increasing sample sizes equal to $n = 500$, $n = 1000$ and $n = 10000$, in order to approach the asymptotic regime. The simulation was carried out in the context of the YUIMA framework (see Iacus and Yoshida 2018), which provides the tools for simulating sample paths of SDEs, performing quasi-maximum likelihood and adaptive LASSO estimation.

Let $\hat{\alpha}_{n,ij}^{(k)}$ and $\hat{\beta}_{n,ij}^{(k)}$ denote the estimate obtained at replication k for the drift parameter α_{ij} and for the diffusion parameter β_{ij}, whose true values are denoted by $\alpha_{0,ij}$ and $\beta_{0,ij}$, respectively. The performance of each estimation technique is evaluated by computing the empirical mean square errors

$$\begin{aligned} \widehat{MSE}_{1ij} &= \frac{1}{N}\sum_{k=1}^{N}(\hat{\alpha}_{n,ij}^{(k)} - \alpha_{0,ij})^2 \qquad 1 \le i \le 3,\ 0 \le j \le 3 \\ \widehat{MSE}_{2ij} &= \frac{1}{N}\sum_{k=1}^{N}(\hat{\beta}_{n,ij}^{(k)} - \beta_{0,ij})^2 \qquad 1 \le i \le 3,\ 0 \le j \le 3 \end{aligned} \tag{27}$$

and by means of the empirical selection frequencies, i.e., number of times that a null parameter was estimated as zero. The quantity (27) and the selection frequencies are computed for each of the estimation methods we want to compare. The numerical results are summarized in Tables 1 and 2, respectively.

With respect to the experiment we conducted, we can draw the following conclusions. In particular we focus on model selection and thus identification of "true" causal relations in the context of Granger causality.

Table 2 Empirical selection frequencies for the zero parameters

Par.	$n = 500$			$n = 1000$			$n = 10000$		
	Bridge	LASSO	QMLE	Bridge	LASSO	QMLE	Bridge	LASSO	QMLE
α_{21}	0.382	0.058	0.023	0.398	0.080	0.041	0.566	0.195	0.159
α_{31}	0.401	0.065	0.031	0.391	0.093	0.044	0.613	0.368	0.326
α_{32}	0.357	0.056	0.025	0.368	0.086	0.038	0.647	0.410	0.355
α_{33}	0.404	0.053	0.030	0.402	0.076	0.045	0.487	0.229	0.205
β_{11}	0.501	0.676	0.169	0.605	0.691	0.324	0.948	0.967	0.962
β_{21}	0.646	0.803	0.232	0.720	0.793	0.442	0.963	0.981	0.973
β_{22}	0.588	0.772	0.172	0.685	0.778	0.359	0.981	0.985	0.976
β_{31}	0.629	0.780	0.217	0.760	0.805	0.470	0.985	0.987	0.983
β_{32}	0.610	0.771	0.178	0.713	0.791	0.392	0.983	0.987	0.983

(1) *The bridge procedure can boost uniformity in model identification.* By looking at the selection frequencies (Table 2), we immediately note the different behavior of the estimators for the drift and diffusion parameters: each technique performs worse on the α_{ij}'s. The selection of the diffusion parameters is generally more accurate, almost reaching consistency for $n = 10000$. This agrees with the results of Theorem 4: from a theoretical point of view, we expect a much slower convergence rate for the drift parameters. We see that the selection probability of the un-penalized estimator is generally lower than that of penalized techniques, as expected. The LASSO and the bridge both show high selection probabilities for the diffusion parameters, with the LASSO slightly better. But the main difference lies in the behavior of the drift parameters estimators: the bridge has a selection probability several times higher than the LASSO, especially for moderate sample sizes. That is the bridge estimator has a generally higher accuracy in the selection of the *whole* model, losing a bit of accuracy on the diffusion parameters group but gaining a much greater improvement for the drift parameters group. To illustrate this point, we chose one representative of the drift parameters and one of the diffusion parameters. In Fig. 2, we show the empirical distributions of the three types of estimators of the chosen representatives (i.e., the estimators of α_{21} and β_{21}), for each sample size. The results of the empirical mean squared errors and selection probabilities allow us to conclude that the adaptive bridge penalization has a good performance uniformly with respect to the several parameter groups.

(2) *The bridge estimator is less sensible to a poor initial guess.* In the cases where the LASSO estimator gives its better results the quasi-maximum likelihood estimator is more concentrated around zero too. This means that the initial estimates for the adaptive procedure were usually quite good and produced higher weights for the null components of the vector parameters. The bridge estimator on the other hand is able to identify the zero parameters with a higher frequency also when the distribution of the QMLE is more spread apart. This suggests that the

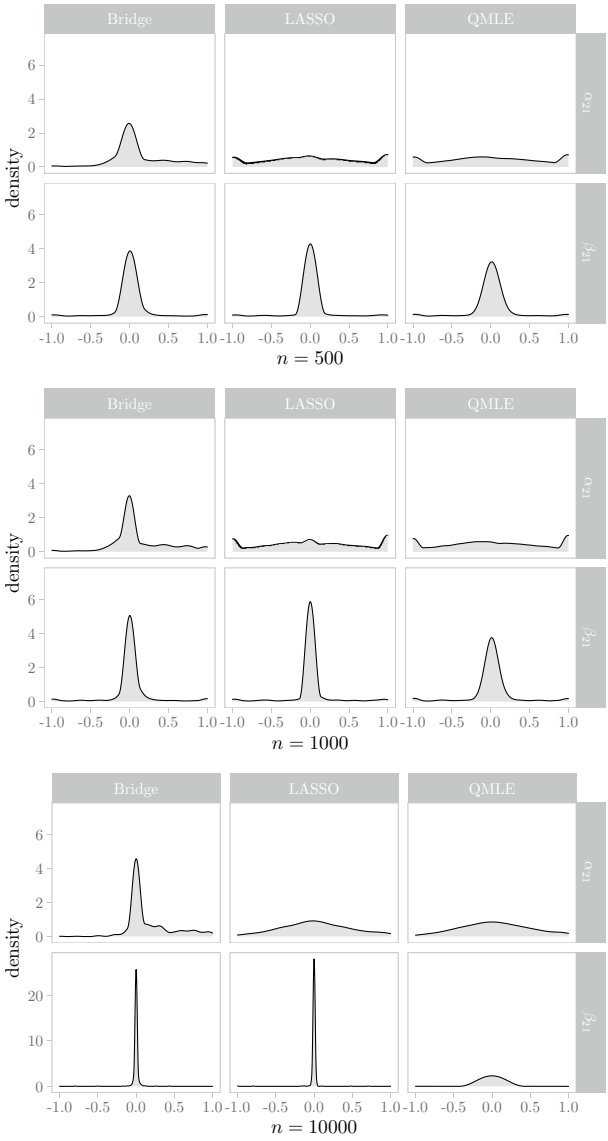

Fig. 2 Comparison of the distributions of the estimators of two of the null parameters, α_{21} and β_{21}: Bridge (column 1), LASSO (column 2) and QMLE (column 3). Note that the plots on the last line are on a different scale

performance of the bridge estimator is, at least in this example, less subordinate to a good performance of the initial estimate with respect to the LASSO procedure.

(3) *The automatic tuning parameter selection criterion is conservative*. The initial values for the tuning parameters were set to

$\lambda_{n,0} = \gamma_{n,0} = 2, \delta_1 = \delta_2 = 1, q_1 = q_2 = .9$. It is worth noticing that the automatic tuning parameter choice procedure does only a fine adjustments around the initial value supplied. This means that the user can choose the magnitude of penalization he desires and the algorithm will tweak it to better fit the data. The results of the tuning parameter selection procedure obtained in our simulation study, for $n = 1000$, are summarized in Table 3.

5.2 Real data prediction

One of the main purposes of the regularized procedures is to improve the predictive capability of the model. In fact, usually when we apply of statistical learning techniques we are not interested on the estimates of the parameters as much as we are in the ability of the model to provide accurate predictions for future outcomes. Bearing this in mind, we tested our model on real data in a predictive study. We fitted a model of the form (25) with four components on a financial time series of daily closing stock prices of four major tech companies, Google, Amazon, Apple and Microsoft, which will be denoted by X_1, X_2, X_3, X_4, respectively. The time series consists of $n = 3283$ observations starting Jan. 3, 2007. The goal is to predict the evolution of the price over a year long period. The training data consist of 3031 observations, until 16-01-2019. The test data are made of the last year of observations, from 17-01-2019 to 16-01-2020. The data have been downloaded by using the service Yahoo Finance and imported into R by using the library `quantmod`.

The scheme of the experiment is as follows. We first fitted the model (25) on our training set by using the adaptive bridge and LASSO methods by using the QMLE as initial estimator. Then, we performed parametric bootstrap to obtain simulations of the series for the time period corresponding to the test data. In order to assess the performance of the estimators, we computed predictive mean square errors and predictive confidence bands. Let n_{te} be the number of observation in the test set $(x_{t_i})_{i=1}^{n_{te}}$ and N the number of simulations performed. The corresponding predicted value is denoted by $(\hat{x}_{t_i}^{(k)})_{i=1}^{n_{te}}, k = 1, \ldots, N$. The predictive mean square error is computed as

$$\widehat{MSE}_p = \frac{1}{n_{te} \cdot N} \sum_{i=1}^{n_{te}} \sum_{k=1}^{N} (x_{t_i} - \hat{x}_{t_i}^{(k)})^2. \tag{28}$$

The error bands are computed as the quantiles of the predicted values at each time instant. In this case, we show 80% and 95% confidence bands.

The results are summarized in Table 4. It compares the results obtained with the bridge and LASSO technique over $N = 10^4$ simulations for each of the stocks considered. The table also reports the result obtained with the unpenalized QMLE as a benchmark. The tuning parameters have been set to $\lambda_0 = \gamma_0 = 10, \delta_1 = \delta_2 = 2.5$ for both the LASSO and the bridge estimator and $q_i, i = 1, 2, ,$ was chosen to be 0.9. We did not use any tuning parameter selection technique in this predictive study. We adopted the rule to set to zero all the parameters for which the absolute

Table 3 Summary of the tuning parameters obtained by implementing the adjustment procedure (23) for $n = 1000$

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St. Dev.
q_1	0.001	0.89	0.94	0.94	0.98	1	0.05
q_2	0.76	0.9	0.93	0.94	0.98	1	0.05
λ_0	0.1	1.69	1.99	2.06	2.27	7.37	0.93
γ_0	0.1	1.56	1.98	1.96	2.04	10	0.87
δ_1	0.1	0.92	1	1.07	1.27	2	0.32
δ_2	0.1	0.71	0.99	0.92	1	2	0.34

Table 4 Comparison of predictive mean square errors over $N = 10^4$ simulations

Series	$\widehat{MSE}_p$		
	Bridge	LASSO	QMLE
X_1	4.64	9.47	10.78
X_2	2.75	5.96	6.09
X_3	4.68	9.34	10.52
X_4	8.12	11.4	11.72

value of the estimate was below a certain threshold ϵ, thus obtaining a reduced model. We then ran the parametric bootstrap simulations with the reduced model for each technique. Initially, the full model had 40 parameters, 11 of which were estimated as zero by the bridge estimator and 9 by the LASSO, having set $\epsilon = 10^{-3}$.

(1) *Bridge estimator can achieve better predictive capability*. Results of Table 4 show that the predictive MSE of the bridge estimator is smaller on all the three data series. The bridge estimator was able to produce improvements on the predictive error of 57%, 54%, 55% and 31% for the three series with respect to the unpenalized estimator. The LASSO improved the predictive capability of the model as well, but in this case the reduction was modest, 12%, 2% and 11%, 2%, respectively, for the three series. We arrive to the same conclusion by comparing the confidence bands depicted in Fig. 3. The bands obtained with the bridge estimator, on the first line, are narrower, leading to less uncertainty in the prediction.

(2) *Bridge estimator resulted reliable over longer time periods*. By looking at Fig. 4, we see how the predictive mean square error changes over time. At first, the two errors are similar, but at later times the LASSO error grows at a higher rate than the error of the bridge estimator, reaching values sometimes even double in the last part of the trajectory. This means that in the case under scrutiny the bridge estimator allows to obtain predictions for longer time periods, thanks to the slower growth of its error rate.

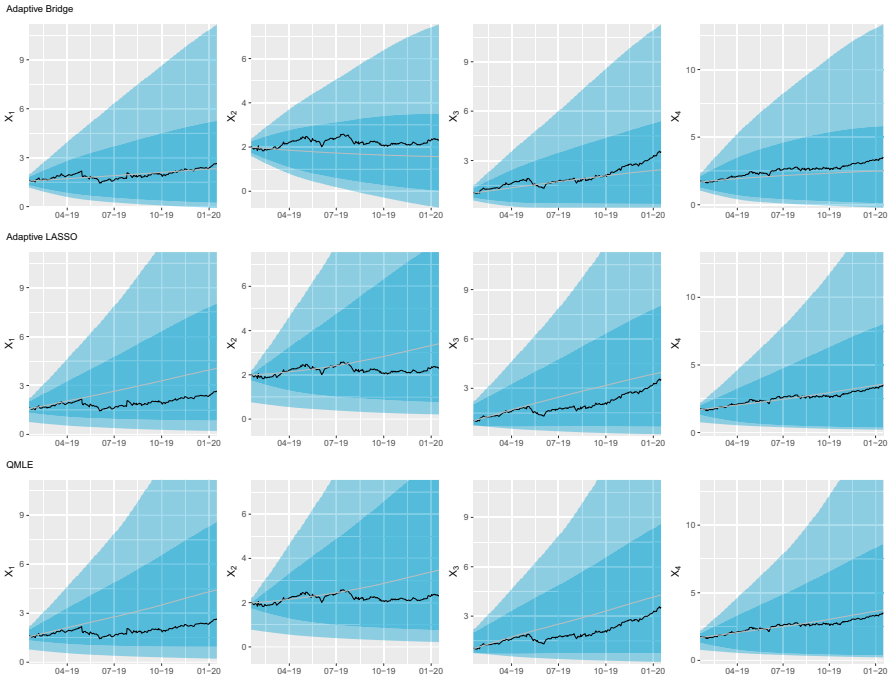

Fig. 3 Predictive MSE obtained with the bridge (row 1), LASSO (row 2) and QML estimators (row 3) for each of the four data series. The darker band depicts the 80% quantiles, the lighter band the 95% quantiles

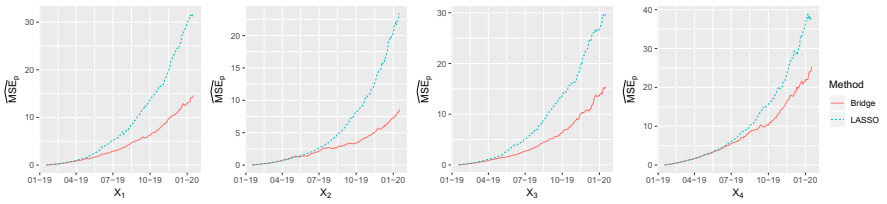

Fig. 4 Comparison of the predictive MSE over time

5.3 Comparison with disjoint estimation

In Suzuki and Yoshida (2019), the authors introduced the penalized estimator for ergodic diffusions (12) defined as follows by

$$\hat{\alpha}_n^{(q_1)} \in \arg\min_{\theta\in\bar{\Theta}_1} Q_{1,n}^{(q_1)}(\alpha), \qquad \hat{\beta}_n^{(q_2)} \in \arg\min_{\theta\in\bar{\Theta}_2} Q_{2,n}^{(q_2)}(\beta), \tag{29}$$

where

Table 5 Selection probabilities for the parameters with true value equal to zero

Sel. Prob.	α_{12}	α_{21}	β_{11}	β_{22}
Joint	0.987	0.998	0.093	0.199
Disjoint	0.997	1.000	0.095	0.218

Table 6 Mean square error for each parameter

MSE	α_{11}	α_{12}	α_{21}	α_{22}	β_{11}	β_{22}
Joint	0.913	0.004	0.000	0.664	7.824	1.901
Disjoint	0.967	0.001	0.000	0.752	42.700	1.081

$$Q_{1,n}^{(q_1)}(\alpha) = (\alpha - \tilde{\alpha}_n)' \hat{G}_{1,n}(\alpha - \tilde{\alpha}_n) + \sum_{i=1}^{p_1} \kappa_{i,n}^1 |\alpha_i|^{q_1},$$

$$Q_{2,n}^{(q_2)}(\beta) = (\beta - \tilde{\beta}_n)' \hat{G}_{2,n}(\beta - \tilde{\beta}_n) + \sum_{i=1}^{p_2} \kappa_{i,n}^2 |\beta_i|^{q_2},$$

and $\hat{G}_{j,n}$ are $p_j \times p_j$ random matrices satisfying suitable regularity conditions, $(\tilde{\alpha}_n, \tilde{\beta}_n)'$ is an initial unpenalized estimator and $\kappa_{i,n}^j, j = 1, 2$ represents suitable adaptive weights (see Suzuki and Yoshida 2019, for details). The main difference between this estimator and (14) is that the former estimates each parameter group separately: in the following, we refer to (29) as *disjoint* estimator, while we call *joint* estimator (14).

In this section, we compare the performances of the joint and disjoint estimators on a simple model. Consider the following SDE

$$\begin{pmatrix} \mathrm{d}X_{1,t} \\ \mathrm{d}X_{2,t} \end{pmatrix} = \begin{pmatrix} -\alpha_{11} X_{1,t}^3 + \alpha_{12}(\sin X_{2,t} + 2) \\ +\alpha_{21}(\cos X_{1,t} + 2) - \alpha_{22} X_{2,t} \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \beta_{11} & 1 \\ 1 & \beta_{22} \end{pmatrix} \begin{pmatrix} \mathrm{d}W_{1,t} \\ \mathrm{d}W_{2,t} \end{pmatrix}, \quad (30)$$

$0 \le t \le T = 10$, $(X_{1,0}, X_{2,0})' = (1, 1)'$, where $(\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22})' \in [0, 10]^4$ and $(\beta_{11}, \beta_{22})' \in [0, 10]^2$. This model is closely related to the one considered (Uchida and Yoshida 2012). We simulated $N = 10^3$ sample paths from this model, each with $n = 10^3$ equally spaced data points, with true parameter value $(\alpha_{11}^*, \alpha_{12}^*, \alpha_{21}^*, \alpha_{22}^*)' = (1, 0, 0, 1)'$, $(\beta_{11}^*, \beta_{22}^*)' = (0, 0)'$.

We compared the performances of the joint and disjoint estimation techniques for the bridge estimator with $q_1 = q_2 = 0.9$, with tuning parameters $\lambda_1 = \lambda_2 = 10$, $\delta_1 = \delta_2 = 2.5$, and QMLE as initial estimator.

Table 5 shows the empirical selection probabilities for the zero parameters, while Table 6 shows the empirical mean squared errors (up to the third decimal digit).

We see that whereas the selection probabilities are practically equivalent up to some Monte Carlo sample error, the mean square errors can be significantly higher for the disjoint method. Therefore, the joint estimator turns out to have a better performance with respect to the disjoint one, at least for the SDE (30).

6 Proofs

Proof of Theorem 1 For the proof of this theorem, we were inspired from the proof of Theorem 1 in Suzuki and Yoshida (2019). Let us start by observing that

$$
\begin{aligned}
0 &\geq \mathcal{F}_n(\hat{\theta}_n) - \mathcal{F}_n(\theta_0) \\
&= (\hat{\theta}_n - \tilde{\theta}_n)' \hat{G}_n (\hat{\theta}_n - \tilde{\theta}_n) + \sum_{i=1}^{m} \sum_{j=1}^{p_i} \lambda_{n,j}^i |\hat{\theta}_{n,j}^i|^{q_i} \\
&\quad - \left((\theta_0 - \tilde{\theta}_n)' \hat{G}_n (\theta_0 - \tilde{\theta}_n) + \sum_{i=1}^{m} \sum_{j=1}^{p_i} \lambda_{n,j}^i |\theta_{0,j}^i|^{q_i} \right) \\
&= (\hat{\theta}_n - \theta_0)' \hat{G}_n (\hat{\theta}_n - \theta_0) + 2(\hat{\theta}_n - \theta_0)' \hat{G}_n (\theta_0 - \tilde{\theta}_n) + \sum_{i=1}^{m} \sum_{j=1}^{p_i} \lambda_{n,j}^i \left(|\hat{\theta}_{n,j}^i|^{q_i} - |\theta_{0,j}^i|^{q_i} \right).
\end{aligned}
$$

Let $K_i := \max_{1 \leq j \leq p_i^0} |\theta_{0,j}|^{q_i - 1}, i = 1, \ldots, m$. By exploiting the same arguments adopted in the proof of Theorem 1 in Suzuki and Yoshida (2019), we can write down

$$
\sum_{j=1}^{p_i} \lambda_{n,j}^i \left(|\hat{\theta}_{n,j}^i|^{q_i} - |\theta_{0,j}^i|^{q_i} \right) \geq -p_i^0 K_i a_n^i |\hat{\theta}_n^i - \theta_0^i|, \quad i = 1, \ldots, m.
$$

Let $||\cdot||$ be a matrix norm. We get

$$
\begin{aligned}
0 &\geq (\hat{\theta}_n - \theta_0)' \hat{G}_n (\hat{\theta}_n - \theta_0) + 2(\hat{\theta}_n - \theta_0)' \hat{G}_n (\theta_0 - \tilde{\theta}_n) \\
&\quad - \sum_{i=1}^{m} p_i^0 K_i r_n^i a_n^i |(r_n^i)^{-1} (\hat{\theta}_n^i - \theta_0^i)| \\
&\geq (\hat{\theta}_n - \theta_0)' \hat{G}_n (\hat{\theta}_n - \theta_0) + 2(\hat{\theta}_n - \theta_0)' \hat{G}_n (\theta_0 - \tilde{\theta}_n) \\
&\quad - \left(\sum_{i=1}^{m} p_i^0 K_i r_n^i a_n^i \right) |A_n^{-1} (\hat{\theta}_n - \theta_0)| \\
&\geq [A_n^{-1} (\hat{\theta}_n - \theta_0)]' \hat{\mathfrak{D}}_n [A_n^{-1} (\hat{\theta}_n - \theta_0)] + 2[A_n^{-1} (\hat{\theta}_n - \theta_0)]' \hat{\mathfrak{D}}_n [A_n^{-1} (\theta_0 - \tilde{\theta}_n)] \\
&\quad - \left(\sum_{i=1}^{m} p_i^0 K_i r_n^i a_n^i \right) |A_n^{-1} (\hat{\theta}_n - \theta_0)|.
\end{aligned}
$$

Let $\rho_{\min}(M)$ and $\rho_{\max}(M)$ be the minimum and maximum eigenvalue, respectively, of a matrix M. We have

$$
\begin{aligned}
[A_n^{-1} (\hat{\theta}_n - \theta_0)]' \hat{\mathfrak{D}}_n [A_n^{-1} (\hat{\theta}_n - \theta_0)] &\geq \rho_{\min}(\hat{\mathfrak{D}}_n) |A_n^{-1} (\hat{\theta}_n - \theta_0)|^2 \\
&\geq ||\hat{\mathfrak{D}}_n^{-1}||^{-1} |A_n^{-1} (\hat{\theta}_n - \theta_0)|^2,
\end{aligned}
$$

where the last step follows from $||\hat{\mathfrak{D}}_n^{-1}|| \geq \rho_{\max}(\hat{\mathfrak{D}}_n^{-1}) = 1/\rho_{\min}(\hat{\mathfrak{D}}_n)$. Furthermore,

$$
\begin{aligned}
[A_n^{-1}(\hat{\theta}_n - \theta_0)]' \hat{\mathfrak{D}}_n [A_n^{-1}(\theta_0 - \tilde{\theta}_n)] &\geq -|[A_n^{-1}(\hat{\theta}_n - \theta_0)]' \hat{\mathfrak{D}}_n [A_n^{-1}(\theta_0 - \tilde{\theta}_n)]| \\
&\geq -|A_n^{-1}(\hat{\theta}_n - \theta_0)| \, |\hat{\mathfrak{D}}_n [A_n^{-1}(\theta_0 - \tilde{\theta}_n)]| \\
&\geq -|A_n^{-1}(\hat{\theta}_n - \theta_0)| \, ||\hat{\mathfrak{D}}_n|| \, |A_n^{-1}(\tilde{\theta}_n - \theta_0)|.
\end{aligned}
$$

Hence, we have proved that

$$
\begin{aligned}
0 \geq ||\hat{\mathfrak{D}}_n^{-1}||^{-1} |A_n^{-1}(\hat{\theta}_n - \theta_0)|^2 - 2||\hat{\mathfrak{D}}_n|| (|A_n^{-1}(\hat{\theta}_n - \theta_0)| \, |A_n^{-1}(\tilde{\theta}_n - \theta_0)|) \\
- \left(\sum_{i=1}^{m} p_i^0 K_i r_n^i a_n^i \right) |A_n^{-1}(\hat{\theta}_n - \theta_0)|.
\end{aligned}
$$

Therefore, from the assumptions

$$
|A_n^{-1}(\hat{\theta}_n - \theta_0)| \leq ||\hat{\mathfrak{D}}_n^{-1}|| \left[2||\hat{\mathfrak{D}}_n|| \, |A_n^{-1}(\tilde{\theta}_n - \theta_0)| + \sum_{i=1}^{m} p_i^0 K_i r_n^i a_n^i \right] = O_p(1), \tag{31}
$$

which concludes the proof. □

Proof of Theorem 2 By taking into account the standard approach based on the Karush–Kuhn–Tucker (KKT) conditions, we are able to prove the selection consistency property of the bridge-type estimator (6). Let us assume that $\hat{\theta}_{n,j}^i \neq 0$ for some $j = p_i^0 + 1, \ldots, p_i$. Let us note that

$$
\hat{G}_n = \begin{pmatrix} \hat{G}_n^1 \\ \hat{G}_n^2 \\ \vdots \\ \hat{G}_n^m \end{pmatrix}
$$

where $\hat{G}_n^i$ is a $p_i \times \mathfrak{p}$ random matrix, for $i = 1, 2, \ldots, m$. Furthermore,

$$
r_n^i \frac{\partial}{\partial \theta_j^i} \mathcal{F}_n(\theta) \Big|_{\theta = \hat{\theta}_n} = 2 r_n^i \hat{G}_n^i(j) A_n A_n^{-1}(\hat{\theta}_n - \tilde{\theta}_n) + r_n^i q_i \lambda_{n,j}^i |\hat{\theta}_{n,j}^i|^{q_i - 1} \operatorname{sgn}(\hat{\theta}_{n,j}^i) = 0, \tag{32}
$$

for $j = p_i^0 + 1, \ldots, p_i$ and $i = 1, \ldots, m$.

By $\hat{G}_n^i(j)$, we denote the j-th row of $\hat{G}_n^i$. From (32), one has

$$
|2 r_n^i \hat{G}_n^i(j) A_n A_n^{-1}(\hat{\theta}_n - \tilde{\theta}_n)| \, |r_n^i \hat{\theta}_{n,j}^i|^{1 - q_i} = q_i (r_n^i)^{2 - q_i} \lambda_{n,j}^i \geq q_i (r_n^i)^{2 - q_i} b_n^i.
$$

By Theorem 1 and the assumptions, we have that

$$
\underbrace{|2 r_n^i \hat{G}_n^i(j) A_n A_n^{-1}(\hat{\theta}_n - \tilde{\theta}_n)|}_{O_p(1)} \underbrace{|r_n^i \hat{\theta}_{n,j}^i|^{1 - q_i}}_{o_p(1)} = o_p(1),
$$

while $q_i (r_n^i)^{2 - q_i} b_n^i \xrightarrow{p} \infty$. Therefore, for any for $j = p_i^0 + 1, \ldots, p_i$, it turns out that

$$P\left(\hat{\theta}^i_{n,j} \neq 0\right) \leq P\left(|2r^i_n \hat{G}^i_n(j) A_n A_n^{-1}(\hat{\theta}_n - \tilde{\theta}_n)|\, |r^i_n \hat{\theta}^i_{n,j}|^{1-q_i} \geq q_i (r^i_n)^{2-q_i} b^i_n\right) \longrightarrow 0,$$

as $n \longrightarrow \infty$. □

Proof of Theorem 3 In order to simplify the reading of the proof, we drop the dependence from n; then we set $\hat{\theta} := \hat{\theta}_n, \tilde{\theta} := \tilde{\theta}_n$ and $\hat{G} := \hat{G}_n$. We will use an approach similar to that developed in the proof of Theorem 3 in Suzuki and Yoshida (2019). Let us rewrite $\hat{G}$ as a partitioned matrix

$$\hat{G} = \begin{pmatrix} \hat{G}^{11} & \hat{G}^{12} & \cdots & \hat{G}^{1m} \\ \hat{G}^{21} & \hat{G}^{22} & \cdots & \hat{G}^{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{G}^{m1} & \hat{G}^{m2} & \cdots & \hat{G}^{mm} \end{pmatrix}$$

where the blocks are given by

$$\hat{G}^{ij} = \begin{pmatrix} \hat{G}^{ij}_{\star\star} & \hat{G}^{ij}_{\star\bullet} \\ \hat{G}^{ij}_{\bullet\star} & \hat{G}^{ij}_{\bullet\bullet} \end{pmatrix}, \quad 1 \leq i, j \leq m.$$

We observe that

$$\begin{aligned}
\mathcal{F}_n(\theta) &= (\theta - \tilde{\theta})' \hat{G} (\theta - \tilde{\theta}) + \sum_{i=1}^{m} \sum_{j=1}^{p_i} \lambda^i_{n,j} |\theta^i_j|^{q_i} \\
&= \sum_{i=1}^{m} (\theta^i - \tilde{\theta}^i)'_\star \hat{G}^{ii}_{\star\star} (\theta^i - \tilde{\theta}^i)_\star + \sum_{i=1}^{m} (\theta^i - \tilde{\theta}^i)'_\bullet \hat{G}^{ii}_{\bullet\bullet} (\theta^i - \tilde{\theta}^i)_\bullet \\
&\quad + 2 \sum_{i=1}^{m} (\theta^i - \tilde{\theta}^i)'_\star \hat{G}^{ii}_{\star\bullet} (\theta^i - \tilde{\theta}^i)_\bullet + 2 \sum_{i=1}^{m} \sum_{j>i} \Big[(\theta^i - \tilde{\theta}^i)'_\star \hat{G}^{ij}_{\star\star} (\theta^j - \tilde{\theta}^j)_\star \\
&\quad + (\theta^i - \tilde{\theta}^i)'_\star \hat{G}^{ij}_{\star\bullet} (\theta^j - \tilde{\theta}^j)_\bullet + (\theta^i - \tilde{\theta}^i)'_\bullet \hat{G}^{ij}_{\bullet\star} (\theta^j - \tilde{\theta}^j)_\star + (\theta^i - \tilde{\theta}^i)'_\bullet \hat{G}^{ij}_{\bullet\bullet} (\theta^j - \tilde{\theta}^j)_\bullet \Big] \\
&\quad + \sum_{i=1}^{m} \sum_{j=1}^{p^0_i} \lambda^i_{n,j} |\theta^i_j|^{q_i} + \sum_{i=1}^{m} \sum_{j=p^0_i+1}^{p_i} \lambda^i_{n,j} |\theta^i_j|^{q_i}.
\end{aligned}$$

By setting $\check{\theta} := (\theta^1_\star, 0, \theta^2_\star, 0, \ldots, \theta^m_\star, 0)' \in \mathbb{R}^{\mathfrak{p}}$, we have

$$\begin{aligned}
\mathcal{F}_n(\check{\theta}) &= \sum_{i=1}^{m} (\theta^i - \tilde{\theta}^i)'_\star \hat{G}^{ii}_{\star\star} (\theta^i - \tilde{\theta}^i)_\star + \sum_{i=1}^{m} (\tilde{\theta}^i)'_\bullet \hat{G}^{ii}_{\bullet\bullet} \tilde{\theta}^i_\bullet - 2 \sum_{i=1}^{m} (\theta^i - \tilde{\theta}^i)'_\star \hat{G}^{ii}_{\star\bullet} \tilde{\theta}^i_\bullet \\
&\quad + 2 \sum_{i=1}^{m} \sum_{j>i} \Big[(\theta^i - \tilde{\theta}^i)'_\star \hat{G}^{ij}_{\star\star} (\theta^j - \tilde{\theta}^j)_\star - (\theta^i - \tilde{\theta}^i)'_\star \hat{G}^{ij}_{\star\bullet} \tilde{\theta}^j_\bullet - (\tilde{\theta}^i_\bullet)' \hat{G}^{ij}_{\bullet\star} (\theta^j - \tilde{\theta}^j)_\star + (\tilde{\theta}^i_\bullet)' \hat{G}^{ij}_{\bullet\bullet} \tilde{\theta}^j_\bullet \Big] \\
&\quad + \sum_{i=1}^{m} \sum_{j=1}^{p^0_i} \lambda^i_{n,j} |\theta^i_j|^{q_i}.
\end{aligned}$$

Let $B_n^i := \{\min_{1\le j\le p_i^0} |\hat{\theta}_j^i| > 0,\ \hat{\theta}_\bullet^i = 0,\ \det(\hat{G}_{\star\star}^{ii}) > 0\}$, by Theorem 1–2 follows $P(\cap_{i=1}^m B_n^i) \longrightarrow 1$. We observe that, if $\cap_{i=1}^m B_n^i$ holds, then $\mathcal{F}_n(\hat{\theta}) = \min_{\check{\theta}\in\mathbb{R}_0^{\mathfrak{p}}} \mathcal{F}_n(\check{\theta})$, where $\mathbb{R}_0^{\mathfrak{p}} := \{\theta \in \mathbb{R}^{\mathfrak{p}} : \theta_\bullet^i = 0, i = 1, \dots, m\}$. This remark implies on B_n^i

$$0 = \frac{1}{2}\frac{\partial}{\partial\theta_\star^i}\mathcal{F}_n(\theta)\Big|_{\theta=\hat{\theta}} = \hat{G}_{\star\star}^{ii}(\hat{\theta}^i - \tilde{\theta}^i)_\star - \hat{G}_{\star\bullet}^{ii}\tilde{\theta}_\bullet^i + \sum_{j>i}\left[\hat{G}_{\star\star}^{ij}(\hat{\theta}^j - \tilde{\theta}^j)_\star - \hat{G}_{\star\bullet}^{ij}\tilde{\theta}_\bullet^j\right] + Z(\hat{\theta}^i)$$

where $Z(\hat{\theta}^i) := (\frac{1}{2}q_i\lambda_{n,1}^i|\hat{\theta}_1^i|^{q_i-1}\mathrm{sgn}(\hat{\theta}_1^i), \dots, \frac{1}{2}q_i\lambda_{n,p_i^0}^i|\hat{\theta}_{p_i^0}^i|^{q_i-1}\mathrm{sgn}(\hat{\theta}_{p_i^0}^i))'$. Therefore,

$$(\hat{\theta}^i - \theta_0^i)_\star = (\tilde{\theta}^i - \theta_0^i)_\star + (\hat{G}_{\star\star}^{ii})^{-1}\hat{G}_{\star\bullet}^{ii}\tilde{\theta}_\bullet^i - \sum_{j>i}\left[(\hat{G}_{\star\star}^{ii})^{-1}\hat{G}_{\star\star}^{ij}(\hat{\theta}^j - \tilde{\theta}^j)_\star - (\hat{G}_{\star\star}^{ii})^{-1}\hat{G}_{\star\bullet}^{ij}\tilde{\theta}_\bullet^j\right] - (\hat{G}_{\star\star}^{ii})^{-1}Z(\hat{\theta}^i).$$

Let $\hat{\mathfrak{G}}_i := (\mathbf{I}_{p_i^0}\ \frac{1}{(r_n^i)^2}(\hat{G}_{\star\star}^{ii})^{-1}(r_n^i)^2\hat{G}_{\star\bullet}^{ii}) \xrightarrow{p} \mathfrak{G}_i$. Hence,

$$\begin{aligned}
&\frac{1}{r_n^i}(\hat{\theta}^i - \theta_0^i)_\star - \mathfrak{G}_i\left\{\frac{1}{r_n^i}(\tilde{\theta}^i - \theta_0^i)\right\}\\
&= \mathbf{1}_{B_n^i}\left\{\frac{1}{r_n^i}(\tilde{\theta}^i - \theta_0^i)_\star + \frac{1}{r_n^i}(\hat{G}_{\star\star}^{ii})^{-1}\hat{G}_{\star\bullet}^{ii}\tilde{\theta}_\bullet^i - \sum_{j>i}\left[\frac{1}{r_n^i}(\hat{G}_{\star\star}^{ii})^{-1}\hat{G}_{\star\star}^{ij}(\hat{\theta}^j - \tilde{\theta}^j)_\star - \frac{1}{r_n^i}(\hat{G}_{\star\star}^{ii})^{-1}\hat{G}_{\star\bullet}^{ij}\tilde{\theta}_\bullet^j\right]\right.\\
&\quad \left. - \frac{1}{r_n^i}(\hat{G}_{\star\star}^{ii})^{-1}Z(\hat{\theta}^i) - \mathfrak{G}_i\left\{\frac{1}{r_n^i}(\tilde{\theta}^i - \theta_0^i)\right\}\right\} + \mathbf{1}_{(B_n^i)^c}\left\{\frac{1}{r_n^i}(\hat{\theta}^i - \theta_0^i)_\star - \mathfrak{G}_i\left\{\frac{1}{r_n^i}(\tilde{\theta}^i - \theta_0^i)\right\}\right\}\\
&= \mathbf{1}_{B_n^i}\left\{(\hat{\mathfrak{G}}_i - \mathfrak{G}_i)\left\{\frac{1}{r_n^i}(\tilde{\theta}^i - \theta_0^i)\right\} - \sum_{j>i}\left[\frac{1}{r_n^i}(\hat{G}_{\star\star}^{ii})^{-1}\hat{G}_{\star\star}^{ij}(\hat{\theta}^j - \tilde{\theta}^j)_\star - \frac{1}{r_n^i}(\hat{G}_{\star\star}^{ii})^{-1}\hat{G}_{\star\bullet}^{ij}\tilde{\theta}_\bullet^j\right]\right.\\
&\quad \left. - \frac{1}{r_n^i}(\hat{G}_{\star\star}^{ii})^{-1}Z(\hat{\theta}^i)\right\} + \mathbf{1}_{(B_n^i)^c}\left\{\frac{1}{r_n^i}(\hat{\theta}^i - \theta_0^i)_\star - \mathfrak{G}_i\left\{\frac{1}{r_n^i}(\tilde{\theta}^i - \theta_0^i)\right\}\right\}\\
&= \mathbf{1}_{B_n^i}\left\{(\hat{\mathfrak{G}}_i - \mathfrak{G}_i)\left\{\frac{1}{r_n^i}(\tilde{\theta}^i - \theta_0^i)\right\} + o_p(1)\right\} + \mathbf{1}_{(B_n^i)^c}\left\{\frac{1}{r_n^i}(\hat{\theta}^i - \theta_0^i)_\star - \mathfrak{G}_i\left\{\frac{1}{r_n^i}(\tilde{\theta}^i - \theta_0^i)\right\}\right\},
\end{aligned}$$

where the last step holds because:

$$\begin{aligned}
&\frac{1}{(r_n^i)^2}(\hat{G}_{\star\star}^{ii})^{-1}r_n^i r_n^j\hat{G}_{\star\star}^{ij}\frac{1}{r_n^j}(\hat{\theta}^j - \tilde{\theta}^j)_\star\,\mathbf{1}_{B_n^i} = o_p(1)O_p(1) = o_p(1);\\
&\frac{1}{(r_n^i)^2}(\hat{G}_{\star\star}^{ii})^{-1}r_n^i r_n^j\hat{G}_{\star\bullet}^{ij}\frac{1}{r_n^j}\tilde{\theta}_\bullet^j\,\mathbf{1}_{B_n^i} = o_p(1);\\
&\frac{1}{(r_n^i)^2}(\hat{G}_{\star\star}^{ii})^{-1}r_n^i Z(\hat{\theta}^i)\,\mathbf{1}_{B_n^i} = o_p(1).
\end{aligned}$$

Finally,

$$\frac{1}{r_n^i}(\hat{\theta}^i - \theta_0^i)_\star - \mathfrak{G}_i\left\{\frac{1}{r_n^i}(\tilde{\theta}^i - \theta_0^i)\right\} \xrightarrow{p} 0,$$

and then the result (8) holds.

By adding the assumption A3, (9) is a trivial consequence of (8). Furthermore, if $G = \Gamma = \text{diag}(\Gamma^{11}, \Gamma^{22}, \dots, \Gamma^{mm})$, we get

$$\mathfrak{T} = \text{diag}(\mathfrak{T}^{11}, \mathfrak{T}^{22}, \dots, \mathfrak{T}^{mm})$$

and

$$\mathfrak{G}\,\mathfrak{T}\,\mathfrak{G}' = \text{diag}(\mathfrak{G}_1\,\mathfrak{T}^{11}\,\mathfrak{G}_1', \mathfrak{G}_2\,\mathfrak{T}^{22}\,\mathfrak{G}_2', \dots, \mathfrak{G}_m\,\mathfrak{T}^{mm}\,\mathfrak{G}_m'),$$

where $\mathfrak{T}^{ii} := (\Gamma^{ii})^{-1}, i = 1, 2, \dots, .m$. By exploiting the blockwise inversion of Γ^{ii}, we recall that

$$\mathfrak{T}^{ii} = \begin{pmatrix} \mathfrak{T}^{ii}_{\star\star} & -\mathfrak{T}^{ii}_{\star\star}\Gamma^{ii}_{\star\bullet}(\Gamma^{ii}_{\bullet\bullet})^{-1} \\ -(\Gamma^{ii}_{\bullet\bullet})^{-1}\Gamma^{ii}_{\bullet\star}\mathfrak{T}^{ii}_{\star\star} & (\Gamma^{ii}_{\bullet\bullet})^{-1} + (\Gamma^{ii}_{\bullet\bullet})^{-1}\Gamma^{ii}_{\bullet\star}\mathfrak{T}^{ii}_{\star\star}\Gamma^{ii}_{\star\bullet}(\Gamma^{ii}_{\bullet\bullet})^{-1} \end{pmatrix} \tag{33}$$

where

$$\mathfrak{T}^{ii}_{\star\star} = (\Gamma^{ii}_{\star\star} - \Gamma^{ii}_{\star\bullet}(\Gamma^{ii}_{\bullet\bullet})^{-1}\Gamma^{ii}_{\bullet\star})^{-1}. \tag{34}$$

By taking into account (33) and (34), we obtain

$$\begin{aligned}
\mathfrak{G}_i\,\mathfrak{T}^{ii}\,\mathfrak{G}_i' &= \mathfrak{T}^{ii}_{\star\star} - (\Gamma^{ii}_{\star\star})^{-1}\Gamma^{ii}_{\star\bullet}(\Gamma^{ii}_{\bullet\bullet})^{-1}\Gamma^{ii}_{\bullet\star}\mathfrak{T}^{ii}_{\star\star} - \mathfrak{T}^{ii}_{\star\star}\Gamma^{ii}_{\star\bullet}(\Gamma^{ii}_{\bullet\bullet})^{-1}\Gamma^{ii}_{\bullet\star}(\Gamma^{ii}_{\star\star})^{-1} \\
&\quad + (\Gamma^{ii}_{\star\star})^{-1}\Gamma^{ii}_{\star\bullet}(\Gamma^{ii}_{\bullet\bullet})^{-1}\Gamma^{ii}_{\bullet\star}(\Gamma^{ii}_{\star\star})^{-1} \\
&\quad + (\Gamma^{ii}_{\star\star})^{-1}\Gamma^{ii}_{\star\bullet}(\Gamma^{ii}_{\bullet\bullet})^{-1}\Gamma^{ii}_{\bullet\star}\mathfrak{T}^{ii}_{\star\star}\Gamma^{ii}_{\star\bullet}(\Gamma^{ii}_{\bullet\bullet})^{-1}\Gamma^{ii}_{\bullet\star}(\Gamma^{ii}_{\star\star})^{-1} \\
&= \mathfrak{T}^{ii}_{\star\star}\left[\Gamma^{ii}_{\star\star} - \Gamma^{ii}_{\star\bullet}(\Gamma^{ii}_{\bullet\bullet})^{-1}\Gamma^{ii}_{\bullet\star}\right](\Gamma^{ii}_{\star\star})^{-1} + (\Gamma^{ii}_{\star\star})^{-1}\Gamma^{ii}_{\star\bullet}(\Gamma^{ii}_{\bullet\bullet})^{-1}\Gamma^{ii}_{\bullet\star}(\Gamma^{ii}_{\star\star})^{-1} \\
&\quad - (\Gamma^{ii}_{\star\star})^{-1}\Gamma^{ii}_{\star\bullet}(\Gamma^{ii}_{\bullet\bullet})^{-1}\Gamma^{ii}_{\bullet\star}\mathfrak{T}^{ii}_{\star\star}\left[\Gamma^{ii}_{\star\star} - \Gamma^{ii}_{\star\bullet}(\Gamma^{ii}_{\bullet\bullet})^{-1}\Gamma^{ii}_{\bullet\star}\right](\Gamma^{ii}_{\star\star})^{-1} \\
&= (\Gamma^{ii}_{\star\star})^{-1},
\end{aligned}$$

which concludes the proof. □

Proof of Theorem 4 The consistency, the selection consistency and the asymptotic normality are consequences of Theorem 1–3. From (31) and the Cauchy–Schwarz inequality, we derive the following bound

$$\begin{aligned}
\mathbb{E}|A_n^{-1}(\hat{\theta}_n - \theta_0)|^q &\le 2^{2q-1}\mathbb{E}\left[||\hat{\mathfrak{D}}_n^{-1}||^q\,||\hat{\mathfrak{D}}_n||^q\,|A_n^{-1}(\tilde{\theta}_n^{QL} - \theta_0)|^q\right] \\
&\quad + 2^{(q-1)(m-1)}\sum_{i=1}^{m}\mathbb{E}\left[p_i^0 K_i r_n^i a_n^i\right]^q \\
&\le 2^{2q-1}\sqrt{\mathbb{E}||\hat{\mathfrak{D}}_n^{-1}||^{2q}}\left(\mathbb{E}||\hat{\mathfrak{D}}_n||^{4q}\right)^{1/4}\left(\mathbb{E}|A_n^{-1}(\tilde{\theta}_n^{QL} - \theta_0)|^{4q}\right)^{1/4} \\
&\quad + 2^{(q-1)(m-1)}\sum_{i=1}^{m}\mathbb{E}\left[p_i^0 K_i r_n^i a_n^i\right]^q.
\end{aligned}$$

From the assumptions, the polynomial-type large deviation result (25) and Proposition 1 in Yoshida (2011), we obtain the uniform L^q-boundedness of the estimator. □

Acknowledgements We would like to thank the Associate Editor and the Referees for their insightful remarks which led to a substantial improvement of the first version of the paper.

References

Antoine, B., Renault, E. (2012). Efficient minimum distance estimation with multiple rates of convergence. *Journal of Econometrics*, *170*(2), 350–367.

Bandi, F., Corradi, V., Moloche, G. (2009). Bandwidth selection for continuous-time Markov processes. Unpublished paper.

Basu, S., Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, *43*(4), 1535–1567.

Caner, M., Knight, K. (2013). An alternative to unit root tests: bridge estimators differentiate between nonstationary versus stationary models and select optimal lag. *Journal of Statistical Planning and Inference*, *143*(4), 691–715.

Clément, E., Gloter, A. (2019). Estimating functions for SDE driven by stable Lévy processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, *55*(3), 1316–1348.

De Gregorio, A., Iacus, S. M. (2012). Adaptive LASSO-type estimation for multivariate diffusion processes. *Econometric Theory*, *28*(4), 838–860.

De Gregorio, A., Iacus, S. M. (2018). On penalized estimation for dynamical systems with small noise. *The Electronic Journal Statistics*, *12*(1), 1614–1630.

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

Fan, J., Li, R. (2006). Statistical Challenges With High Dimensionality: Feature Selection in Knowledge Discovery. In *Proceedings of the Madrid international congress of mathematicians*, Madrid.

Fan, J., Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*(3), 928–961.

Frank, L. E., Friedman, J. H., Silverman, B. W. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 109–135.

Gaïffas, S., Matulewicz, G. (2019). Sparse inference of the drift of a high-dimensional Ornstein-Uhlenbeck process. *Journal of Multivariate Analysis*, *169*, 1–20.

Gloter, A., Sørensen, M. (2009). Estimation for stochastic differential equations with a small diffusion coefficient. *Stochastic Processes and their Applications*, *119*(3), 679–699.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The elements of statistical learning. Data mining, inference, and prediction* 2nd ed. Springer Series in Statistics. New York: Springer.

Hastie, T., Tibshirani, R., Wainwright, M. (2015). *Statistical learning with sparsity. The LASSO and generalizations.* Monographs on Statistics and Applied Probability, 143. Boca Raton: CRC Press.

Iacus, S.M., Yoshida N. (2018). *Simulation and inference for stochastic processes with YUIMA. A comprehensive R framework for SDEs and other stochastic processes.* Use R!. Cham: Springer.

Kamatani, K., Uchida, M. (2015). Hybrid multi-step estimators for stochastic differential equations based on sampled data. *Statistical Inference for Stochastic Processes*, *18*(2), 177–204.

Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics.*, *24*(2), 211–229.

Kinoshita Y., Yoshida N. (2019). Penalized quasi likelihood estimation for variable selection. https://arxiv.org/abs/1910.12871.

Knight, K., Fu, W. (2000). Asymptotics for LASSO-type estimators. *The Annals of Statistics*, *28*(5), 1536–1378.

Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, *72*(6), 1899–1925.

Masuda, H. (2019). Non-Gaussian quasi-likelihood estimation of SDE driven by locally stable Lévy process. *Stochastic Processes and their Applications*, *129*(3), 1013–1059.
Masuda, H., Shimizu, Y. (2017). Moment convergence in regularized estimation under multiple and mixed-rates asymptotics. *Mathematical Methods of Statistics*, *26*(2), 81–110.
McCrorie, J. R., Chambers, M. J. (2006). Granger causality and the sampling of economic processes. *Journal of Econometrics*, *132*(2), 311–336.
Nardi, Y., Rinaldo, A. (2011). Autoregressive process modeling via the LASSO procedure. *Journal of Multivariate Analysis*, *102*(3), 528–549.
Radchenko, P. (2008). Mixed-rates asymptotics. *The Annals of Statistics*, *36*(1), 287–309.
Shimizu, Y., Yoshida, N. (2006). Estimation of parameters for diffusion processes with jumps from discrete observations. *Statistical Inference for Stochastic Processes*, *9*(3), 227–277.
Sørensen, M., Uchida, M. (2003). Small-diffusion asymptotics for discretely sampled stochastic differential equations. *Bernoulli*, *9*(6), 1051–1069.
Suzuki, T., Yoshida, N. (2019). Penalized least squares approximation methods and their applications to stochastic processes. To appear in *Japanese Journal of Statistics and Data Science*, https://arxiv.org/abs/1811.09016.
Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*: *Series B*, *58*(1), 267–288.
Uchida, M. (2011). Contrast-based information criterion for ergodic diffusion processes from discrete observations. *Annals of the Institute of Statistical Mathematics*, *62*(1), 161–187.
Uchida, M., Yoshida, N. (2012). Adaptive estimation of an ergodic diffusion process based on sampled data. *Stochastic Processes and their Applications*, *122*(8), 2885–2924.
Uchida, M., Yoshida, N. (2014). Adaptive Bayes type estimators of ergodic diffusion processes from discrete observations. *Statistical Inference for Stochastic Processes*, *17*(2), 181–219.
Wang, H., Leng, C. (2007). Unified LASSO estimation by Least Squares Approximation. *Journal of American Statistical Association*, *102*(479), 1039–1048.
Wang, H., Li, G., Tsai, C.-L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B*, *169*(1), 63–78.
Yoshida, N. (1992). Estimation for diffusion processes from discrete observation. *Journal of Multivariate Analysis*, *41*(2), 220–242.
Yoshida, N. (2011). Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. *Annals of the Institute of Statistical Mathematics*, *63*(3), 431–479.
Yu, J., Phillips, P. C. B. (2001). Gaussian estimation of continuous time models of the short term interest rate. *The Econometrics Journal*, *4*(2), 210–224.
Zou, H. (2006). The adaptive LASSO and its Oracle properties. *Journal of American Statistical Association*, *101*(476), 1418–1429.