Electronic Journal of Statistics
Vol. 12 (2018) 1614–1630
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1436

On penalized estimation for dynamical systems with small noise

Alessandro De Gregorio

Department of Statistical Sciences
University of Rome "La Sapienza"
P.le Aldo Moro 5, 00185- Rome, Italy
e-mail: alessandro.degregorio@uniroma1.it

and

Stefano Maria Iacus

Department of Economics, Management and Quantitative Methods
University of Milan
Via Conservatorio 7, 20122 - Milan, Italy
e-mail: stefano.iacus@unimi.it

Abstract: We consider a dynamical system with small noise for which the drift is parametrized by a finite dimensional parameter. For this model, we consider minimum distance estimation from continuous time observations under l^p-penalty imposed on the parameters in the spirit of the Lasso approach, with the aim of simultaneous estimation and model selection. We study the consistency and the asymptotic distribution of these Lasso-type estimators for different values of p. For $p = 1$, we also consider the adaptive version of the Lasso estimator and establish its oracle properties.

MSC 2010 subject classifications: Primary 62M05, 62J07; secondary 62F12.

Keywords and phrases: Dynamical systems, lasso estimation, model selection, inference for stochastic processes, diffusion-type processes, oracle properties.

Received February 2018.

Contents

1. Introduction

Usually ordinary differential equation models are the result of averaging and/or neglecting some details of an original system without modeling a complex system with a huge number of degrees of freedom or tuning parameters. Introducing noise is therefore a way to approximate closer the reality of observable complex systems. It is then natural to think of the noise as small, for example when one is considering the dynamics of macroscopic quantities, i.e. averages of quantities of interest over a whole population or in the case of signal that travels through a perturbed medium, etcetera.

Dynamical systems with small perturbations have been indeed widely studied in [1] and [8]. Applications of small diffusion processes to mathematical finance and option pricing have been considered in [43], [24], [34], [41] and references therein. Examples from biology and life sciences include [17], [2], [6].

Model selection is an important aspect in the above applied fields although sometimes neglected. What occurs for dynamical systems with small noise, is not so different from what happens in ordinary least squares (OLS) model estimation. Indeed, linear regression models are used extensively by many practitioners but, once estimated, these models are useful as long as the set of parameter (or covariates) is correctly specified. Therefore, the model selection step is an important part of the analysis.

To introduce the idea of Lasso-type estimation we begin with linear models and OLS. In this framework model selection occurs when some of the regression parameters are estimated as zero. Different models are compared in terms of information criteria like AIC/BIC or hypotheses testing. The advantage of the Lasso-type approach over AIC/BIC is that statistical 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 to choose one of the two models [3].

Variable selection becomes particularly important when the true underlying model has a sparse representation. Correctly identifying significant predictors will improve the prediction performance of the fitted model (for an overview of feature selection see [7]).

Considered the linear regression model $Y_i = x_i^T\beta + \varepsilon_i$, with x_i a vector of covariates, β a vector of $q > 0$ parameters and ε_i i.i.d. Gaussian random variables. [23] proposed the following l^p-penalized estimator for β

$$\hat\beta_n = \arg\min_u \left(\sum_{i=1}^n (Y_i - x_i^T u)^2 + \lambda_n \sum_{j=1}^q |u_j|^p \right) \tag{1.1}$$

for some $p > 0$ and $\lambda_n \to 0$ as $n \to \infty$. The family of estimators $\hat\beta_n$ solution to (1.1) is a generalization of the Ridge estimators which corresponds to the case $p = 2$ (see [5]). The original Lasso estimators proposed in [35] are obtained setting $p = 1$ while OLS is the case $\lambda_n = 0$, not considered here. The link between Lasso-type estimation and model selection is also due to the fact that, in the limit as $p \to 0$, this procedure approximate the AIC or BIC selection methods (which correspond to $p = 0$ with $\lambda_n > 0$), i.e.

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

which amounts to the number of non-null parameters in the model. Here $\mathbf{1}_A$ the indicator function for set A.

As said, the estimators solutions to (1.1) are attractive because with them it is possible to perform estimation and model selection in a single step, i.e. the procedure does not need to estimate different models and compare them later with information criteria as the dimension of the space of the parameters does not change; just some of the components of the vector β_j are assumed to be zero. In non-linear models a preliminary simple reparametrization (e.g. $\beta \mapsto \beta' - \beta$) is needed to interpret this approach in terms of model selection.

In this work, we extend the problem in (1.1) to the class of diffusion-type processes with small noise solution to the stochastic differential equation $\mathrm{d}X_t = S_t(\theta, X)\mathrm{d}t + \varepsilon \mathrm{d}W_t$, $t \in [0, T]$, by replacing least squares estimation with minimum distance estimation. The asymptotic is considered as $\varepsilon \to 0$ for fixed $0 < T < \infty$ with $\theta \in \Theta \subset \mathbb{R}^q$ a q-dimensional parameter.

Since the seminal works of [25, 26, 27] and [42], statistical inference for continuously observed small diffusion processes is well developed today (see, e.g., [28, 18, 19, 44, 40]) but the Lasso problem has not been considered so far. Although here we consider only continuous time observations, it is worth mentioning that there is also a growing literature on parametric inference for discretely observed small diffusion processes (see, e.g., [9, 14, 31, 32, 33, 36, 37, 38, 39, 11, 12]) to which this Lasso problem can be extended. Adaptive Lasso-type estimation for ergodic diffusion processes sampled at discrete time has been studied in [4], while for continuous time ergodic diffusion processes shrinkage estimation has been considered in [29].

This paper is organized as follows. In Section 2, we introduce the model, the assumptions and the statement of the problem. In Section 3, we study the consistency of the estimators and derive their asymptotic distribution for different values of p. For $p = 1$, we also consider the case of adaptive Lasso estimation that is meant to control asymptotic bias. We are also able to prove that the adaptive estimation represents an oracle procedure.

2. The Lasso-type problem for dynamical systems with small noise

Let us assume that on the probability space $(\Omega, \mathcal{F}, P)$, with the filtration $\{\mathcal{F}_t, 0 \leq t \leq T\}$ (where each $\mathcal{F}_t, 0 \leq t \leq T$, is augmented by sets from $\mathcal{F}$ having zero P-measure), is given a Wiener process $\{W_t, \mathcal{F}_t, 0 \leq t \leq T\}$. Let $X = \{X_t, 0 \leq t \leq T\}$ be a real valued diffusion-type process solution to the following stochastic differential equation

$$\mathrm{d}X_t = S_t(\theta, X)\mathrm{d}t + \varepsilon \mathrm{d}W_t, \quad \varepsilon \in (0, 1], \tag{2.1}$$

with non random initial condition $X_0 = x_0$, where $S_t(\cdot, X)$ is a known measurable non-anticipative functional (see, e.g., [13]). The parameter $\theta \in \Theta \subset \mathbb{R}^q$, where Θ is a bounded, open and convex set, is supposed to be unknown.

Let $(C[0,T], \mathcal{B}[0,T])$ be the measurable space of continuous functions x_t on $[0,T]$ with σ-algebra $\mathcal{B}[0,T] = \sigma\{x_t, 0 \leq t \leq T\}$. Moreover, $P_\theta^{(\varepsilon)}$ denotes the law induced by the process X in $(C[0,T], \mathcal{B}[0,T])$ when the true parameter is θ. We denote by $u = (u_1, \ldots, u_q)^T$ the (transposed) vector $u \in \mathbb{R}^q$ and the true value of θ by θ^*. Let $||\cdot|| = ||\cdot||_{L_2(\mu)}$ be the L_2-norm with respect to some finite measure μ on $[0,T]$, i.e.

$$||f||^2 = \int_0^T f^2(t)\mu(\mathrm{d}t).$$

We suppose that the trend coefficient in (2.1) is of integral type, i.e.

$$S_t(\theta, X) = V(\theta, t, X) + \int_0^t K(\theta, t, s, X_s)\mathrm{d}s,$$

where $V(\theta, t, x)$ and $K(\theta, t, s, x)$ are known measurable, non-anticipative functionals such that (2.1) has a strong unique solution. For example, the usual conditions (1.34) and (1.35) in [27] and Theorem 4.6 in [13] about Lipschitz behavior and linear growth are sufficient.

Assumption 1. *For all $t \in [0,T]$, $\theta \in \Theta$ and $X_t, Y_t \in C[0,T]$*

$$\begin{aligned}
&|V(\theta,t,X_t) - V(\theta,t,Y_t)| + \int_0^t |K(\theta,t,s,X_s) - K(\theta,t,s,Y_s)|\mathrm{d}s \\
&\leq L_1 \int_0^t |X_s - Y_s|\mathrm{d}H_s + L_2|X_t - Y_t|, \\
&|V(\theta,t,X_t)| + \int_0^t |K(\theta,t,s,X_s)|\mathrm{d}s \leq L_1 \int_0^t (1+|X_s|)\mathrm{d}H_s + L_2(1+|X_t|),
\end{aligned}$$

where L_1 and L_2 are positive constants and H_s is a nondecreasing right-continuous function, $0 \leq H_t \leq H_0, H_0 > 0$.

Assumption 1 implies that all the probability measures $P_\theta^{(\varepsilon)}, \theta \in \Theta$, are equivalent (see Theorem 7.7 in [13]). The asymptotic in this model is considered as $\varepsilon \to 0$ and $0 < T < \infty$ fixed.

We will also write $x(\theta) = \{x_t(\theta), 0 \leq t \leq T\}$ to denote all the solution of the limiting dynamical system

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = V(\theta,t,x_t) + \int_0^t K(\theta,t,s,x_s)\mathrm{d}s, \quad x_0,$$

where $x_t = x_t(\theta)$. We assume that, for all $0 \leq t \leq T$ and for each $\theta \in \Theta$, the random element X_t and $x_t(\theta)$ belong to $L_2(\mu)$. Furthermore, we suppose that the functionals $V(\theta,t,x)$ and $K(\theta,t,s,x)$ have bounded first derivative with respect to θ.

Let $x^{(1)}(\theta^*) = \{x_t^{(1)}(\theta^*), 0 \le t \le T\}$ be the Gaussian process solution to

$$\mathrm{d}x_t^{(1)} = \left(V_x(\theta^*, t, x_t(\theta^*))x_t^{(1)} + \int_0^t K_x(\theta^*, t, s, x_s(\theta^*))x_s^{(1)}\mathrm{d}s \right) \mathrm{d}t + \mathrm{d}W_t, \quad (2.2)$$
$$x_0^{(1)} = 0,$$

where $x_t^{(1)} = x_t^{(1)}(\theta^*)$, $V_x(\theta, t, x) = \frac{\partial}{\partial x}V(\theta, t, x)$ and $K_x(\theta, t, s, x) = \frac{\partial}{\partial x}K(\theta, t, s, x)$. The process $x^{(1)}(\theta^*)$ plays a central role in the definition of the asymptotic distribution of the estimators in the theory of dynamical systems with small noise. We need in addition the following assumptions.

Assumption 2. *The stochastic process X is differentiable in ε at the point $\varepsilon = 0$ in the following sense: for all $\nu > 0$*

$$\lim_{\varepsilon \to 0} P_{\theta^*}^{(\varepsilon)} \left(||\varepsilon^{-1}(X - x(\theta^*)) - x^{(1)}(\theta^*)|| > \nu \right) = 0$$

where $x^{(1)}(\theta^*) = \{x_t^{(1)}(\theta^*), 0 \le t \le T\}$ is from (2.2) with bounded coefficients $V_x(\theta^*, t, \cdot)$ and $K_x(\theta^*, t, s, \cdot)$.

We further denote by $\dot{x}_t(\theta)$ the q-dimensional vector of partial derivatives of $x_t(\theta)$ with respect to θ_j, $j = 1, \ldots, q$, i.e., $\dot{x}_t(\theta) = (\frac{\partial}{\partial \theta_1}x_t(\theta), \ldots, \frac{\partial}{\partial \theta_q}x_t(\theta))^T$, and $\dot{x}_t(\theta^*)$ satisfies the systems of equations

$$\begin{aligned} \frac{\mathrm{d}\dot{x}_t(\theta^*)}{\mathrm{d}t} &= [V_x(\theta^*, t, x_t(\theta^*))\dot{x}_t(\theta^*) + \dot{V}(\theta^*, t, x_t(\theta^*)) \\ &\quad + \int_0^t (\dot{K}(\theta^*, t, s, x_s(\theta^*)) + K_x(\theta^*, t, s, x_s(\theta^*))\dot{x}_s(\theta^*))\mathrm{d}s]\mathrm{d}t, \\ \dot{x}_0(\theta^*) &= 0, \end{aligned}$$

where the point corresponds to the differentiation on θ; i.e.

$$\dot{V}(\theta, t, x_t(\theta)) = \left(\frac{\partial}{\partial \theta_1}V(\theta, t, x_t(\theta)), ..., \frac{\partial}{\partial \theta_q}V(\theta, t, x_t(\theta)) \right)^T$$

and

$$\dot{K}(\theta, t, s, x_s(\theta)) = \left(\frac{\partial}{\partial \theta_1}K(\theta, t, s, x_s(\theta)), ..., \frac{\partial}{\partial \theta_q}K(\theta, t, s, x_s(\theta)) \right)^T.$$

Assumption 3. *The deterministic dynamical system $x_t(\theta)$ is $L_2(\mu)$-differentiable in θ at the point θ^*; i.e.*

$$||x(\theta^* + h) - x(\theta^*) - h^T\dot{x}(\theta^*))|| = o(|h|)$$

where $h \in \mathbb{R}^q$.

Assumption 4. *The matrix*

$$\mathcal{I}(\theta^*) = \int_0^T \dot{x}_t(\theta^*)\dot{x}_t^T(\theta^*)\mu(\mathrm{d}t)$$

is positive definite and nonsingular.

2.1. The Lasso-type estimator

We introduce a constrained minimum distance estimator for θ for the model (2.1). The asymptotic properties of the unconstrained minimum distance estimators in the i.i.d. framework have been established in [15, 16]. Later [26, 27] and [28] studied in details the properties of such estimators for diffusion processes with small noise. Information criteria for this model have been studied in [40], while here we study the Lasso-type approach.

To define the Lasso-type estimator the following penalized contrast function has to be considered

$$Z_\varepsilon(u) = ||X - x(u)|| + \lambda_\varepsilon \sum_{j=1}^q |u_j|^p, \tag{2.3}$$

where $p > 0$, $u \in \Theta$ and $\lambda_\varepsilon > 0$ is a real sequence. In analogy to (1.1), we introduce the Lasso-type estimator $\hat{\theta}^\varepsilon : C[0,T] \to \bar{\Theta}$ for θ, defined as

$$\hat{\theta}^\varepsilon = \arg\min_{\theta \in \bar{\Theta}} Z_\varepsilon(\theta), \tag{2.4}$$

where $\bar{\Theta}$ is the closure of Θ.

The following example explains well the spirit of the Lasso procedure. We consider a linear small diffusion-type process X given by

$$\mathrm{d}X_t = \sum_{j=1}^q \theta_j A_j(t, X)\mathrm{d}t + \varepsilon \mathrm{d}W_t, \quad 0 \le t \le T.$$

By applying the estimator (2.4), some parameters θ_j will be set equal to 0 and this implies a simultaneous estimation and selection of the model. Therefore, the Lasso methodology is particularly useful in the random dynamical systems framework where a sparse representation of the drift term emerges (i.e. some components of θ are exactly zero) and we are interested in identifying the true model.

3. Asymptotic properties of the Lasso-type estimator

The additional l_p-penalization term in the contrast function (2.3) modifies the traditional properties of the minimum distance estimator. The analysis should be performed for the different values of p.

3.1. Consistency

Let us introduce the following functions

$$g_{\theta^*}^\varepsilon(\nu) = \inf_{|\theta - \theta^*| \ge \nu} \left\{ ||x(\theta) - x(\theta^*)|| + \lambda_\varepsilon \sum_{j=1}^q |\theta_j|^p \right\},$$

$$h^{\varepsilon}_{\theta^*}(\nu) = \inf_{|\theta-\theta^*|<\nu} \left\{ ||x(\theta) - x(\theta^*)|| + \lambda_\varepsilon \sum_{j=1}^{q} |\theta_j|^p \right\}$$

where $|\theta - \theta^*| \geq \nu$ (resp. $< \nu$) is to be intended componentwise, for all $\nu > 0$. We need the following identifiability-type condition.

Assumption 5. *For every $\nu > 0$, we assume that*

$$g^{\varepsilon}_{\theta^*}(\nu) > h^{\varepsilon}_{\theta^*}(\nu).$$

Theorem 1. *Suppose Assumption 1 and Assumption 5 are fulfilled and assume $\lambda_\varepsilon = O(\varepsilon)$ as $\varepsilon \to 0$. Then $\hat{\theta}^\varepsilon$ in (2.4) is a uniformly consistent estimator of θ^*; i.e. for any $\nu > 0$*

$$\lim_{\varepsilon\to 0} \sup_{\theta^*\in\Theta} P^{(\varepsilon)}_{\theta^*} \left(|\hat{\theta}^\varepsilon - \theta^*| \geq \nu \right) = 0.$$

Proof. By definition of $\hat{\theta}^\varepsilon$, for any $\nu > 0$, we have that

$$\left\{ \omega : |\hat{\theta}^\varepsilon - \theta^*| \geq \nu \right\} = \left\{ \omega : \inf_{|\theta-\theta^*|<\nu} Z_\varepsilon(\theta) > \inf_{|\theta-\theta^*|\geq\nu} Z_\varepsilon(\theta) \right\}$$

Moreover,

$$Z_\varepsilon(\theta) \leq ||X - x(\theta^*)|| + ||x(\theta) - x(\theta^*)|| + \lambda_\varepsilon \sum_{j=1}^{q} |\theta_j|^p,$$

$$Z_\varepsilon(\theta) \geq ||x(\theta) - x(\theta^*)|| - ||X - x(\theta^*)|| + \lambda_\varepsilon \sum_{j=1}^{q} |\theta_j|^p.$$

Then, from the above inequality, we get

$$\begin{aligned} P^{(\varepsilon)}_{\theta^*} \left(|\hat{\theta}^\varepsilon - \theta^*| \geq \nu \right) &= P^{(\varepsilon)}_{\theta^*} \left(\inf_{|\theta-\theta^*|<\nu} Z_\varepsilon(\theta) > \inf_{|\theta-\theta^*|\geq\nu} Z_\varepsilon(\theta) \right) \\ &\leq P^{(\varepsilon)}_{\theta^*} \left(||X - x(\theta^*)|| + \frac{h^{\varepsilon}_{\theta^*}(\nu)}{2} > \frac{g^{\varepsilon}_{\theta^*}(\nu)}{2} \right) \end{aligned}$$

Since (see Lemma 1.13, in [27])

$$||X - x(\theta^*)|| \leq C\varepsilon \sup_{0\leq t\leq T} |W_t|, \quad P^{(\varepsilon)}_{\theta^*} - \text{a.s.},$$

where $C = C(L_1, L_2, K_0, T)$ is a positive constant, under Assumption 5, we get

$$\begin{aligned} \sup_{\theta^*\in\Theta} P^{(\varepsilon)}_{\theta^*} \left(|\hat{\theta}^\varepsilon - \theta^*| \geq \nu \right) &\leq P^{(\varepsilon)}_{\theta^*} \left(C\varepsilon \sup_{0\leq t\leq T} |W_t| > \frac{1}{2} \inf_{\theta^*\in\Theta} \{ g^{\varepsilon}_{\theta^*}(\nu) - h^{\varepsilon}_{\theta^*}(\nu) \} \right) \\ &\leq 2\exp\left\{ -\frac{(\inf_{\theta^*\in\Theta}\{g^{\varepsilon}_{\theta^*}(\nu) - h^{\varepsilon}_{\theta^*}(\nu)\})^2}{8TC^2\varepsilon^2} \right\} \to 0. \end{aligned}$$

In the above, we made use of the following estimate for $N > 0$

$$P\left(\sup_{0\le t\le T}|W_t| > N\right) \le 4P\left(W_T > N\right) \le 2e^{-\frac{N^2}{2T}},$$

see, e.g., [27], and observed that

$$g^{\varepsilon}_{\theta^*}(\nu) - h^{\varepsilon}_{\theta^*}(\nu) \to \inf_{|\theta-\theta^*|\ge\nu} ||x(\theta) - x(\theta^*)|| > 0, \quad \varepsilon \to 0. \qquad \square$$

From the proof of the consistency of (2.4) it is clear that the speed of convergence of $\hat{\theta}^{\varepsilon}$ depends on the asymptotic rate of λ_ε. The rate of convergence of λ_ε also affects the asymptotic distribution of the estimator.

Remark 1. *It is possible to define other types of Lasso-type estimators modifying the metric in (2.3); i.e. by considering, for instance, the sup-norm and the L_1-norm. Hence, if $\{X_t, 0 \le t \le T\}$ and $\{x_t(\theta), 0 \le t \le T\}, \theta \in \Theta$, are elements of the space $C[0,T]$ or $L_1(\mu)$, we can introduce the Lasso estimator*

$$\breve{\theta}^{\varepsilon} = \arg\min_{\theta\in\bar{\Theta}} \left\{ \sup_{0\le t\le T} |X_t - x_t(\theta)| + \lambda_\varepsilon \sum_{j=1}^{q} |u_j|^p \right\}$$

or

$$\breve{\breve{\theta}}^{\varepsilon} = \arg\min_{\theta\in\bar{\Theta}} \left\{ \int_0^T |X_t - x_t(\theta)|\mu(\mathrm{d}t) + \lambda_\varepsilon \sum_{j=1}^{q} |u_j|^p \right\},$$

respectively. The estimators $\breve{\theta}^{\varepsilon}$ and $\breve{\breve{\theta}}^{\varepsilon}$ are uniformly consistent and the proof follows by the same steps adopted to prove Theorem 1. Clearly, it is necessary to redefine the functions $g^{\varepsilon}_{\theta^*}$ and $h^{\varepsilon}_{\theta^*}$ appearing in Assumption 5 by replacing $L_2(\mu)$-norm with sup-norm (for $\breve{\theta}^{\varepsilon}$) or $L_1(\mu)$-norm (for $\breve{\breve{\theta}}^{\varepsilon}$).

3.2. Asymptotic distribution

In order to study the asymptotic distribution of the Lasso-type estimator we need to distinguish the different cases for p. We start with the case of $p \ge 1$. We denote by "$\to_d$" the convergence in distribution and we denote by ζ the following Gaussian random vector

$$\zeta = \int_0^T x^{(1)}_t(\theta^*)\dot{x}_t(\theta^*)\mu(\mathrm{d}t); \tag{3.1}$$

i.e. $\zeta \sim N_q(\mathbf{0}, \sigma^2)$ where

$$\sigma^2 = \int_0^T \int_0^T \dot{x}_t(\theta^*)\dot{x}_s(\theta^*)^T E[x^{(1)}_t(\theta^*)x^{(1)}_s(\theta^*)]\mu(\mathrm{d}t)\mu(\mathrm{d}s),$$

(see also Lemma 2.13 in [27]). The next two theorems have been inspired from Theorem 2 and Theorem 3 in [23]. Nevertheless, in our case the convexity argument adopted in [23] does not work.

Theorem 2. *Suppose that Assumption 1–Assumption 5 are fulfilled, ζ is defined as in (3.1), $p \geq 1$ and $\varepsilon^{-1}\lambda_\varepsilon \to \lambda_0 \geq 0$. Then*

$$\varepsilon^{-1}(\hat{\theta}^\varepsilon - \theta^*) \to_d \arg\min_u V(u)$$

where

$$V(u) = -2u^T\zeta + u^T\mathcal{I}(\theta^*)u + p\lambda_0 \sum_{j=1}^q u_j \operatorname{sgn}(\theta_j^*)|\theta_j^*|^{p-1}$$

for $p > 1$ and

$$V(u) = -2u^T\zeta + u^T\mathcal{I}(\theta^*)u + \lambda_0 \sum_{j=1}^q \left(|u_j|\mathbf{1}_{\{\theta_j^*=0\}} + u_j \operatorname{sgn}(\theta_j^*)\mathbf{1}_{\{\theta_j^*\neq 0\}}\right)$$

if $p = 1$.

Proof. Let $u \in \mathbb{R}^q$ and introduce the random function

$$\begin{aligned} V_\varepsilon(u) &= \frac{1}{\varepsilon^2}\Bigg(||X - x(\theta^* + \varepsilon u)||^2 - ||X - x(\theta^*)||^2 \\ &\quad + \lambda_\varepsilon \sum_{j=1}^q \left\{|\theta_j^* + \varepsilon u_j|^p - |\theta_j^*|^p\right\}\Bigg), \end{aligned} \tag{3.2}$$

which is minimized at the point $u = \varepsilon^{-1}(\hat{\theta}^\varepsilon - \theta^*)$ by definition of $\hat{\theta}^\varepsilon$. By exploiting Assumption 2-Assumption 4, we get

$$\begin{aligned} &\frac{1}{\varepsilon^2}\left\{||X - x(\theta^* + \varepsilon u)||^2 - ||X - x(\theta^*)||^2\right\} \\ &= \frac{1}{\varepsilon^2}\left\{||X - x(\theta^*) - \varepsilon u^T \dot{x}(\theta^*)||^2 - ||X - x(\theta^*)||^2\right\} + o_\varepsilon(1) \\ &= u^T||\dot{x}(\theta^*)||^2 u - 2u^T||\varepsilon^{-1}(X - x(\theta^*))\dot{x}(\theta^*)|| + o_\varepsilon(1) \\ &\xrightarrow[\varepsilon\to 0]{P_{\theta^*}^{(\varepsilon)}} u^T\mathcal{I}(\theta^*)u - 2u^T\zeta, \end{aligned} \tag{3.3}$$

where $\xrightarrow{P_{\theta^*}^{(\varepsilon)}}$ stands for the convergence in probability and ζ is from (3.1). For the term in (3.2)

$$\frac{\lambda_\varepsilon}{\varepsilon^2}\sum_{j=1}^q \left\{|\theta_j^* + \varepsilon u_j|^p - |\theta_j^*|^p\right\}$$

we have to distinguish the case $p = 1$ and $p > 1$. Let $p > 1$, then

$$\frac{\lambda_\varepsilon}{\varepsilon^2}\sum_{j=1}^q \left\{|\theta_j^* + \varepsilon u_j|^p - |\theta_j^*|^p\right\}$$

$$= \frac{\lambda_\varepsilon}{\varepsilon}\sum_{j=1}^{q} u_j \frac{|\theta_j^* + \varepsilon u_j|^p - |\theta_j^*|^p}{\varepsilon u_j} \underset{\varepsilon\to 0}{\longrightarrow} p\lambda_0 \sum_{j=1}^{q} u_j \mathrm{sgn}(\theta_j^*)|\theta_j^*|^{p-1} \tag{3.4}$$

If $p = 1$, then by similar arguments, we have

$$\frac{\lambda_\varepsilon}{\varepsilon^2}\sum_{j=1}^{q}\left\{|\theta_j^* + \varepsilon u_j| - |\theta_j^*|\right\} \underset{\varepsilon\to 0}{\longrightarrow} \lambda_0 \sum_{j=1}^{q}\left(|u_j|\mathbf{1}_{\{\theta_j^*=0\}} + u_j \mathrm{sgn}(\theta_j^*)\mathbf{1}_{\{\theta_j^*\neq 0\}}\right). \tag{3.5}$$

Notice that $V_\varepsilon(u)$ is not convex in u and then we have to consider the convergence in distribution on the topology induced by the uniform metric on compact sets; i.e. we deal with the convergence in distribution of $V_\varepsilon(u)$ on the space of the continuous functions topologized by the distance $\rho(y_1, y_2) = \sup_{u\in K}|y_1(u) - y_2(u)|$, where K is a compact subset of $\mathbb{R}^d$. From (3.3), (3.4) and (3.5) follows the convergence of the finite-dimensional distributions

$$(V_\varepsilon(u_1), ..., V_\varepsilon(u_k)) \to_d (V(u_1), ..., V(u_k))$$

for any $u_i \in \mathbb{R}^d, i = 1, ..., k$. The tightness of $V_\varepsilon(u)$ is implied by

$$\sup_{\varepsilon\in(0,1]} E\left[\sup_{u\in K}\left|\frac{\mathrm{d}}{\mathrm{d}u}V_\varepsilon(u)\right|\right] < \infty$$

which follows from the regularity conditions on $\{X_t, 0 \le t \le T\}$ and $\{x_t(\theta), 0 \le t \le T\}$. Indeed it is not hard to prove that

$$\lim_{h\to 0}\limsup_{\varepsilon\to 0} E[w(V_\varepsilon(u), h)\wedge 1] \le \lim_{h\to 0} h \sup_{\varepsilon\in(0,1]} E\left[\sup_{u\in K}\left|\frac{\mathrm{d}}{\mathrm{d}u}V_\varepsilon(u)\right|\right] = 0,$$

where $w(y, h) = \sup\{\rho(y(u), y(v)) : |u - v| \le h\}$, with y a continuous function on compact sets and $h > 0$. Therefore by Theorem 16.5 in [20], we conclude that

$$V_\varepsilon(u) \to_d V(u)$$

uniformly on u. Since $\arg\min_u V(u)$ is unique ($P_{\theta^*}^{(\varepsilon)}$–a.s.), to prove that

$$\arg\min_u V_\varepsilon(u) = \varepsilon^{-1}(\hat\theta^\varepsilon - \theta^*) \to_d \arg\min_u V(u),$$

we can use Theorem 2.7 in [21]. Hence, it is sufficient to show that $\arg\min_u V_\varepsilon(u) = O_{P_{\theta^*}^{(\varepsilon)}}(1)$. We observe that

$$V_\varepsilon(u) = V_\varepsilon^l(u) + o_\varepsilon(1)$$

where

$$V_\varepsilon^l(u) = \frac{1}{\varepsilon^2}\Big\{u^T||\dot x(\theta^*)||^2 u - 2u^T||\varepsilon^{-1}(X - x(\theta^*))\dot x(\theta^*)||$$

$$+ \lambda_\varepsilon \sum_{j=1}^{q} \left\{ |\theta_j^* + \varepsilon u_j|^p - |\theta_j^*|^p \right\} \Bigg\}$$

is a convex function. Since for each $a \in \mathbb{R}$ and $\delta > 0$, there exists a compact set $K_{a,\delta}$ such that (see, [22])

$$\limsup_{\varepsilon \to 0} P_{\theta^*}^{(\varepsilon)} \left(\inf_{u \notin K_{a,\delta}} V_\varepsilon(u) \leq a \right) \leq \delta,$$

then $\arg\min_u V_\varepsilon(u) = O_{P_{\theta^*}^{(\varepsilon)}}(1)$. □

In the case $0 < p < 1$, a different rate of convergence must be imposed on the sequence λ_ε.

Theorem 3. *Suppose that Assumption 1–Assumption 4 hold, ζ defined as in (3.1), $0 < p < 1$ and $\lambda_\varepsilon/\varepsilon^{2-p} \to \lambda_0 \geq 0$. Then*

$$\varepsilon^{-1}(\hat{\theta}^\varepsilon - \theta^*) \to_d \arg\min_u V(u)$$

where

$$V(u) = -2u^T\zeta + u^T\mathcal{I}(\theta^*)u + \lambda_0 \sum_{j=1}^{q} |u_j|^p \mathbf{1}_{\{\theta_j^*=0\}}.$$

Proof. We proceed analogously to the proof of Theorem 2. As before we start with $V_\varepsilon(u)$ from (3.2). The first part of the expression in $V_\varepsilon(u)$ converges in distribution to $-2u^T\zeta + u^T\mathcal{I}(\theta^*)u$ as in Theorem 2. For the second term, we need to distinguish the two cases $\theta_k^* = 0$ or $\theta_k^* \neq 0$. By assumptions we have that $\lambda_\varepsilon/\varepsilon^{2-p} \to \lambda_0$ and hence necessarily $\lambda_\varepsilon/\varepsilon \to 0$.

Consider first the case $\theta_k^* \neq 0$. We have that

$$\frac{\lambda_\varepsilon}{\varepsilon} u_k \left(\frac{|\theta_k^* + \varepsilon u_k|^p - |\theta_k^*|^p}{\varepsilon u_k} \right) \to 0.$$

Conversely, if $\theta_k^* = 0$ we have that

$$\frac{\lambda_\varepsilon}{\varepsilon^2} \sum_{j=1}^{q} \left(|\theta_j^* + \varepsilon u_j|^p - |\theta_j^*|^p \right) \to \lambda_0 \sum_{j=1}^{q} |u_j|^p \mathbf{1}_{\{\theta_j^*=0\}}$$

So, by means of the same arguments adopted in the proof of Theorem 2, we can prove that $V_\varepsilon(u) \to_d V(u)$ uniformly on u. Following [21], the final step consists in showing that $\arg\min V_\varepsilon = O_{P_{\theta^*}^{(\varepsilon)}}(1)$ and so $\arg\min V_\varepsilon \to_d \arg\min V$. Indeed,

$$V_\varepsilon(u) \geq \frac{1}{\varepsilon^2} \left(||X - x(\theta^* + \varepsilon u)||^2 - ||X - x(\theta^*)||^2 \right) - \frac{\lambda_\varepsilon}{\varepsilon^2} \sum_{j=1}^{q} |\varepsilon u_j|^p$$

and for all u and ε sufficiently small, $\delta > 0$, we have

$$V_\varepsilon(u) \geq \frac{1}{\varepsilon^2} \left(||X - x(\theta^* + \varepsilon u)||^2 - ||X - x(\theta^*)||^2 \right) - (\lambda_0 + \delta) \sum_{j=1}^{q} |u_j|^p = V_\varepsilon^\delta(u).$$

The term $|u_j|^p$ grows slower than the the first normed terms in $V_\varepsilon^\delta(u)$, so $\arg\min_u V_\varepsilon^\delta(u) = O_{P_{\theta^*}^{(\varepsilon)}}(1)$ and, in turn, $\arg\min_u V_\varepsilon(u)$ is also $O_{P_{\theta^*}^{(\varepsilon)}}(1)$. The uniqueness of $\arg\min_u V(u)$ completes the proof. □

Remark 2. *If $\lambda_0 = 0$, from the above theorems we immediately obtain that*

$$\varepsilon^{-1}(\hat\theta^\varepsilon - \theta^*) \to_d \arg\min_u V(u) = \mathcal{I}^{-1}(\theta^*)\zeta,$$

where $\mathcal{I}^{-1}(\theta^*)\zeta \sim N_q(\mathbf{0}, \mathcal{I}^{-1}(\theta^*)\sigma^2\mathcal{I}^{-1}(\theta^*))$.

4. Adaptive version of the penalized estimator

Theorem 3 shows that, if $p < 1$, one can estimate the nonzero parameters $\theta_j^* \neq 0$ at the usual rate without introducing asymptotic bias due to the penalization and, at the same time, shrink the estimates of the null $\theta_j^* = 0$ parameters toward zero with positive probability.

On the contrary, if $p \geq 1$ non zero parameters are estimated with some asymptotic bias if $\lambda_0 > 0$. This is a well known result in the literature [23], [45] and has been indeed considered in [4] for ergodic diffusion models with discrete observations.

In this section we consider only the case for $p = 1$, i.e. the real Lasso estimator. Furthermore, we deal with an adaptive version of the Lasso estimator for the diffusion-type process (2.1).

To state the results we need to rearrange the elements of the vector parameters θ in this way. Suppose that $q_0 \leq q$ values of θ^* are not null, than we reorder θ^* as follows: $\theta^* = (\theta_1^*, \dots, \theta_{q_0}^*, \theta_{q_0+1}^*, \dots, \theta_q^*)^T$, where we denoted by $\theta_k^* = 0$, $k = q_0 + 1, \dots, q$, the null parameters. We now need to modify the optimization function by introducing one adaptive sequence for each of the parameters θ_j; i.e.

$$\tilde Z_\epsilon(u) = ||X - x(u)|| + \sum_{j=1}^q \lambda_{\varepsilon,j}|u_j|, \tag{4.1}$$

and, as in the above, the *adaptive* Lasso-type estimator is the solution to

$$\tilde\theta^\varepsilon = (\tilde\theta_1^\varepsilon, ..., \tilde\theta_q^\varepsilon) = \arg\min_{\theta\in\bar\Theta} \tilde Z_\varepsilon(\theta). \tag{4.2}$$

We now need to slightly modify the rate of convergence of the new sequences $\{\lambda_{\varepsilon,j}, j = 1, \dots, q\}$.

Assumption 6. *Let*

$$\kappa_\varepsilon = \min_{j>q_0} \lambda_{\varepsilon,j} \qquad \textit{and} \qquad \gamma_\varepsilon = \max_{1\leq j\leq q_0} \lambda_{\varepsilon,j}.$$

Then the following convergence must hold

$$\frac{\kappa_\varepsilon}{\varepsilon} \to \infty \qquad \textit{and} \qquad \frac{\gamma_\varepsilon}{\varepsilon} \to 0.$$

Let

$$\dot{x}_t^1(\theta) = \left(\frac{\partial}{\partial\theta_1}x_t(\theta), \dots, \frac{\partial}{\partial\theta_{q_0}}x_t(\theta)\right)^T,$$

and

$$\mathcal{I}_{11}(\theta) = \int_0^T \dot{x}_t^1(\theta)\dot{x}_t^1(\theta)^T\mu(\mathrm{d}t), \quad (q_0 \times q_0 \text{ matrix}).$$

Let η be a Gaussian random vector defined as follows

$$\eta = \int_0^T x_t^{(1)}(\theta^*)\dot{x}_t^1(\theta^*)\mu(\mathrm{d}t) \sim N_{q_0}(\mathbf{0}, \sigma_1^2), \tag{4.3}$$

where

$$\sigma_1^2 = \int_0^T \int_0^T \dot{x}_t^1(\theta^*)\dot{x}_s^1(\theta^*)^T E[x_t^{(1)}(\theta^*)x_s^{(1)}(\theta^*)]\mu(\mathrm{d}t)\mu(\mathrm{d}s).$$

The estimator $\tilde{\theta}^\varepsilon$ enjoys asymptotically the oracle properties. Indeed, a good fitting procedure should have the following (asymptotically) properties: (i) consistently estimates null parameters as zero and vice versa; i.e. identifies the right subset model; (ii) has the optimal estimation (prediction) rate and converges to a Gaussian random variable with covariance matrix of the true subset model. In our framework, it is reasonable to require that the estimator $\tilde{\theta}^\varepsilon$ defines an oracle procedure. Indeed, for instance, as observed at the end of the Section 2.1, the diffusion-type processes can have a sparse representation and then it is useful to identify consistently the true model.

Theorem 4 (Oracle properties)**.** *Suppose Assumption 1–Assumption 6 are fulfilled. Then, as $\varepsilon \to 0$, the following results hold.*

(i) Consistency in variable selection; i.e.

$$P_{\theta^*}^{(\varepsilon)}(\tilde{\theta}_k^\varepsilon = 0) \longrightarrow 1, \quad k = q_0+1, \dots, q;$$

(ii) Asymptotic normality; i.e.

$$\varepsilon^{-1}(\tilde{\theta}_1^\varepsilon - \theta_1^*, ..., \tilde{\theta}_{q_0}^\varepsilon - \theta_{q_0}^*)^T \longrightarrow_d \mathcal{I}_{11}^{-1}(\theta^*)\eta,$$

where $\mathcal{I}_{11}^{-1}(\theta^*)\eta \sim N_{q_0}(\mathbf{0}, \mathcal{I}_{11}^{-1}(\theta^*)\,\sigma_1^2\,\mathcal{I}_{11}^{-1}(\theta^*))$.

Proof. (i) We briefly outline the proof. The proof is by contradiction. Let us assume that for one $j = q_0+1, \dots, q$ the adaptive Lasso estimator for $\theta_j^* = 0$ is $\tilde{\theta}_j^\varepsilon \neq 0$. By taking into account the Karush-Kuhn-Tucker (KKT) optimality conditions, we have

$$\frac{1}{\varepsilon}\frac{\partial}{\partial u_j}\tilde{Z}_\epsilon(u)\bigg|_{u=\tilde{\theta}^\varepsilon} = \frac{1}{\varepsilon}\left(\frac{\partial}{\partial u_j}||X - x(u)||\bigg|_{u=\tilde{\theta}^\varepsilon}\right) + \frac{\lambda_{\varepsilon,j}}{\varepsilon}\mathrm{sgn}(\tilde{\theta}_j^\varepsilon) = 0.$$

The first term is $O_{P_{\theta^*}^{(\varepsilon)}}(1)$ by Assumption 2 and the fact that $\tilde{\theta}^\varepsilon$ is the solution of (4.2). For the second term we have that $\frac{\lambda_{\varepsilon,j}}{\varepsilon} \geq \frac{\kappa_\varepsilon}{\varepsilon} \to \infty$ by Assumption 6.

(ii) Let

$$\tilde{V}_\varepsilon(u) = \frac{1}{\varepsilon^2}\left(||X - x(\theta^* + \varepsilon u)||^2 - ||X - x(\theta^*)||^2 + \sum_{j=1}^{q} \lambda_{\varepsilon,j} \left\{ |\theta_j^* + \varepsilon u_j| - |\theta_j^*| \right\} \right)$$
$$= u^T ||\dot{x}(\theta^*)||^2 u - 2u^T ||\varepsilon^{-1}(X - x(\theta^*))\dot{x}(\theta^*)|| + o_\varepsilon(1)$$
$$+ \sum_{j=1}^{q} \frac{\lambda_{\varepsilon,j}}{\varepsilon} \left\{ \frac{|\theta_j^* + \varepsilon u_j| - |\theta_j^*|}{\varepsilon} \right\} \tag{4.4}$$

From Assumption 6, since

$$u_j \frac{|\theta_j^* + \varepsilon u_j| - |\theta_j^*|}{u_j \varepsilon} \xrightarrow[\varepsilon \to 0]{} u_j \mathrm{sgn}(\theta_j^*),$$

for $j = 1, ..., q_0$, we have that

$$\sum_{j=1}^{q_0} \frac{\lambda_{\varepsilon,j}}{\varepsilon} \left\{ \frac{|\theta_j^* + \varepsilon u_j| - |\theta_j^*|}{\varepsilon} \right\} \leq \frac{\gamma_\varepsilon}{\varepsilon} \sum_{j=1}^{q_0} \left\{ u_j \frac{|\theta_j^* + \varepsilon u_j| - |\theta_j^*|}{u_j \varepsilon} \right\} \xrightarrow[\varepsilon \to 0]{} 0,$$

while for $\theta_j^* = 0, j = q_0 + 1, ..., q$, one has that $\sum_{j=q_0+1}^{q} \frac{\lambda_{\varepsilon,j}}{\varepsilon} |u_j| \xrightarrow[\varepsilon \to 0]{} \infty$. Therefore, it is not possible to use the topology of the uniform converge on compact sets. Nevertheless, we can define the convergence of $\tilde{V}_\varepsilon$ via epi-convergence in distribution; i.e. from Lemma 4.1 in [10], follows that $\tilde{V}_\varepsilon(u) \to_d \tilde{V}(u)$ for every u, where

$$\tilde{V}(u) = \begin{cases} u_1^T \mathcal{I}_{11}(\theta^*) u_1 - 2u_1^T \eta, & \text{if } u_{q_0+1} = ... = u_q = 0, \\ \infty, & \text{otherwise,} \end{cases}$$

and $u_1 = (u_1, ..., u_{q_0})^T$ and the previous convergence is considered on the space of extended functions $\mathbb{R}^q \to [-\infty, +\infty]$ with a suitable metric. For more details on the epi-convergence see [10], [22] and [30]. Since the minimum point of $\tilde{V}_\varepsilon(u)$ is given by $\varepsilon^{-1}(\tilde{\theta}^\varepsilon - \theta^*)$ and $\arg\min_u \tilde{V}(u) = (\mathcal{I}_{11}^{-1}(\theta^*)\eta, \mathbf{0})^T$ is $P_{\theta^*}-$unique, from Theorem 4.4 in [10] follows the result (ii).

□

Now let $\tilde{\theta}^\varepsilon$ be any consistent estimator of θ^*, for example, the unconstrained minimum distance estimator or the maximum likelihood estimator [27]. Then, as suggested in [45], for any constant $\lambda_0 > 0$ and $\delta > 1$, it is sufficient to choose the sequences $\lambda_{\varepsilon,j}$ as follows

$$\lambda_{\varepsilon,j} = \frac{\lambda_0}{|\tilde{\theta}^\varepsilon|^\delta}. \tag{4.5}$$

If $\lambda_0/\varepsilon \to 0$ and $\varepsilon^{\delta-1}\lambda_0 \to \infty$ as $\varepsilon \to 0$, then Assumption 6 is satisfied. Usually values of $\delta = 1.5$ or $\delta = 2$ are common in adaptive Lasso estimation. The idea of choosing weights as in (4.5) is to exploit the ability of consistent estimators to give an initial guess of how large is a parameter, and then using Lasso approach to shrink adaptively the penalty function in order to avoid bias for true large parameters.

Acknowledgments

We thank an anonymous referee for his/her comments which allowed to clarify some parts of the paper.

References

[1] Azencott, R. (1982). Formule de Taylor stochastique et développement asymptotique d'intégrales de Feynman. *Seminar on Probability, XVI, Supplement, pp. 237–285, Lecture Notes in Math., 921*. Springer, Berlin-New York. MR658728

[2] Bressloff, P. C. (2014). *Stochastic processes in cell biology, interdisciplinary applied mathematics, 41*. Springer-Verlag, New York. MR3244328

[3] Caner, M. (2009). Lasso-type GMM estimator. *Econometric Theory* **25** 270–290. MR2472053

[4] De Gregorio, A. and Iacus, S. M. (2012). Adaptive LASSO-type estimation for multivariate diffusion processes. *Econometric Theory* **28** 838-860. MR2959127

[5] Efron, B., Hastie, T. and Tibshirani, R. (2004). Least angle regression. With discussion, and a rejoinder by the authors. *The Annals of Statistics* **32** 407–499. MR2060166

[6] Ermentrout, G. B. and Terman, D. H. (2010). *Mathematical Foundations of Neurosciences, Interdisciplinary Applied Mathematics 35*. Springer-Verlag, New York. MR2674516

[7] Fan, J. and li, R. (2006). Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. *Proceedings of the International Congress of Mathematicians, Vol. III, 595–622, Eur. Math. Soc., Zurich*. MR2275698

[8] Freidlin, M. I. and Wentzell, A. D. (1998). *Random perturbations of dynamical systems*. Springer-Verlag, New York. MR1652127

[9] Genon-Catalot, V. (1990). Maximum contrast estimation for diffusion processes from discrete observations. *Statistics* **21** 99–116. MR1056065

[10] Geyer, C. J. (1994). On the asymptotics of constrained m-estimation. *The Annals of Statistics* **22** 1993–2010. MR3131286

[11] Gloter, A. and Sørensen, M. (2009). Estimation for stochastic differential equations with a small diffusion coefficient. *Stochastic Processes and their Applications* **119** 679–699. MR2500255

[12] Guy, R., Larédo, C. F. and Vergu, E. (2014). Parametric inference for discretely observed multidimensional diffusions with small diffusion coefficient. *Stochastic Processes and their Applications* **124** 51–80. MR3131286

[13] Lipster, R. S. and Shiryaev, A. N. (2001). *Statistics for Random Processes I: General Theory*. Springer-Verlag, New York. MR3244328

[14] Larédo, C. F. (1990). A sufficient condition for asymptotic sufficiency of incomplete observations of a diffusion process. *The Annals of Statistics* **18** 1158–1171. MR1062703

[15] Millar, P. W. (1983). The minimax principle in asymptotic statistical theory. *Eleventh Saint Flour probability summer school—1981 (Saint Flour, 1981), 75–265, Lecture Notes in Mathematics, 976.* Springer, Berlin. MR0722983

[16] Millar, P. W. (1984). A general approach to the optimality of the minimum distance estimators. *Transactions of the American Mathematical Society* **286** 377–418. MR0756045

[17] Murray, J. D. (2002). *Mathematical Biology I. An introduction.* Springer-Verlag, New York. MR1908418

[18] Iacus, S. M. (2000). Semiparametric estimation of the state of a dynamical system with small noise. *Statistical Inference for Stochastic Processes* **3** 277–288. MR1819400

[19] Iacus, S. M. and Kutoyants, Yu. (2001). Semiparametric hypotheses testing for dynamical systems with small noise. *Mathematical Methods of Statistics* **10** 105–120. MR1841810

[20] Kallenberg, O. (2002). *Foundations of Modern Probability.* Springer-Verlag, New York. MR1876169

[21] Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics* **18** 191–219. MR1041391

[22] Knight, K. (1999). Epi-convergence in distribution and stochastic equi-semicontinuity. *Unpublished manuscript.*

[23] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* **28** 1356–1378. MR1805787

[24] Kunitomo, N. and Takahashi, A. (2001). The asymptotic expansion approach to the valuation of interest rate contingent claims. *Mathematical Finance* **11** 117–151. MR1807851

[25] Kutoyants, Yu. (1984). *Parameter estimation for stochastic processes.* Heldermann Verlag, Berlin. MR0777685

[26] Kutoyants, Yu. (1991). Minimum distance parameter estimation for diffusion type observations. *Comptes Rendus de l'Académie des Sciences. Série I. Mathématique*, **312**, 637–642. MR1101048

[27] Kutoyants, Yu. (1994). *Identification of dynamical systems with small noise.* Kluwer Academic Publishers Group, Dordrecht. MR1332492

[28] Kutoyants, Yu. and Pilibossian, P. (1994). On minimum uniform metric estimate of parameters of diffusion-type processes. *Stochastic Processes and their Applications* **51** 259–267. MR1288291

[29] Nkurunziza, S. (2012). Shrinkage strategies in some multiple multi-factor dynamical systems. *ESAIM: Probability and Statistics* **16** 139–150. MR2946124

[30] Rockafellar, R. T. and Wets, R. J. B. (1998). *Variational Analysis.* Springer-Verlag, Berlin. MR1491362

[31] Sørensen, M. (1997). Small dispersion asymptotics for diffusion martingale estimating functions. *Department of Statistics and Operations Research, University of Copenaghen, Preprint No. 2000-2.*

[32] Sørensen, M. (2012). Estimating functions for diffusion-type processes. In M. Kessler, A. Lindner, and M. Sørensen, editors, *Statistical Methods for*

Stochastic Differential Equations. Proceedings of the Second International Symposium on Information Theory, pages 1–107. CRC Press, Chapmann and Hall. MR2976982

[33] Sørensen, M. and Uchida, M. (2003). Small diffusion asymptotics for discretely sampled stochastic differential equations. *Bernoulli* **9** 1051–1069. MR2046817

[34] Takahashi, A. and Yoshida, N. (2004). An asymptotic expansion scheme for optimal investment problems. *Statistical Inference for Stochastic Processes* **7** 153–188. MR2061183

[35] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological.* **58** 267–288. MR1379242

[36] Uchida, M. (2003). Estimation for dynamical systems with small noise from discrete observations. *Journal of the Japan Statistical Society* **33** 157–167. MR2039892

[37] Uchida, M. (2004). Estimation for discretely observed small diffusions based on approximate martingale estimating functions. *Scandinavian Journal of Statistics* **31** 553–566. MR22101539

[38] Uchida, M. (2006). Martingale estimating functions based on eigenfunctions for discretely observed small diffusions. *Bulletin of Informatics and Cybernetics* **38** 1–13. MR2312660

[39] Uchida, M. (2008). Approximate martingale estimating functions for stochastic differential equations with small noises. *Stochastic Processes and their Applications* **118** 1076–1721. MR2442376

[40] Uchida, M. and Yoshida, N. (2004). Information criteria for small diffusions via the theory of Malliavin-Watanabe. *Statistical Inference for Stochastic Processes* **7** 35–67. MR2041908

[41] Uchida, M. and Yoshida, N. (2004). Asymptotic expansion for small diffusions applied to option pricing. *Statistical Inference for Stochastic Processes* **7** 189–223. MR2111290

[42] Yoshida, Y. (1992). Asymptotic expansion of maximum likelihood estimators for small diffusions via the theory of Malliavin-Watanabe. *Probability Theory and Related Fileds* **92** 275–311. MR1165514

[43] Yoshida, Y. (1992). Asymptotic expansion for statistics related to small diffusions. *Journal of the Japan Statistical Society* **22** 139–159. MR1212246

[44] Yoshida, Y. (2003). Conditional expansions and their applications. *Stochastic Processes and their Applications* **107** 53–81. MR1995921

[45] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469