Statistical Inference for Stochastic Processes (2025) 28:22
https://doi.org/10.1007/s11203-025-09341-w

Adaptive Elastic-Net estimation for sparse diffusion processes

Alessandro De Gregorio[1] · Dario Frisardi[1] · Stefano Iacus[2] · Francesco Iafrate[3]

Received: 1 August 2025 / Accepted: 30 November 2025 / Published online: 12 December 2025

Abstract
Penalized estimation methods for diffusion processes and dependent data have recently gained significant attention due to their effectiveness in handling stochastic systems. In this work, we introduce an adaptive Elastic-Net estimator for ergodic diffusion processes observed under high-frequency sampling schemes. Our method combines the least squares approximation of the quasi-likelihood with adaptive ℓ_1 and ℓ_2 regularization. This approach allows to enhance prediction accuracy and interpretability while effectively recovering the sparse underlying structure of the model. In the spirit of recent research trends, we provide finite-sample guarantees for the (block-diagonal) estimator's performance by deriving high-probability non-asymptotic bounds for the ℓ_2 estimation error. These results complement the established oracle properties in the high-frequency asymptotic regime with mixed convergence rates, ensuring consistent selection of the relevant interactions and achieving optimal rates of convergence. Furthermore, we utilize our results to analyze one-step-ahead predictions, offering non-asymptotic control over the ℓ_1 prediction error. The performance of our method is evaluated through simulations and real data applications, demonstrating its effectiveness, particularly in scenarios with strongly correlated variables.

Keywords Discrete observations · Ergodic diffusion processes · Non-asymptotic bounds · Oracle properties · Pathwise optimization · Prediction error · Regularized estimation

✉ Alessandro De Gregorio
alessandro.degregorio@uniroma1.it

Dario Frisardi
dario.frisardi@uniroma1.it

Stefano Iacus
siacus@iq.harvard.edu

Francesco Iafrate
francesco.iafrate@uni-hamburg.de

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

[2] Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, CGIS Knafel Building, Room K350, Cambridge, MA 02138, USA

[3] Department of Mathematics, University of Hamburg, Bundestr. 55, 20146 Hamburg, Germany

Contents

1 Introduction

Nowadays, the regularization methods are very useful for a comprehensive understanding of the underlying parametric model which generally is supposed sparse; that is some coefficients are exactly zero. The main idea is to perform simultaneously the selection of the true model and to estimate the parameters. Penalized estimators of a parameter θ are generally defined as follows:

$$\hat{\theta}_n \in \arg\min_{\theta\in\overline{\Theta}}\{\mathfrak{L}_n(\theta) + p(\theta)\} \tag{1}$$

where $\mathfrak{L}_n(\theta)$ is a contrast function (negative log-likelihood or sum of squared residuals) based on a random sample of size n, Θ is the parameter space and $p(\theta)$ is the penalty function. If $p(\theta) = \lambda|\theta|^q, \lambda > 0$, where $q \in (0, 1]$, we have the Bridge estimator introduced in Frank and Friedman (1993), which for $q = 1$ reduces to LASSO (Least Absolute Shrinkage Selection Operator) studied in Tibshirani (1996) for linear regression. Furthermore, a good selection procedure should satisfy the so-called oracle properties, as discussed in Fan and Li (2001); Fan and Peng (2004); Fan and Li (2006).

Penalized estimation for stochastic processes is a quite recent research topic in the field of statistical learning for random complex system and discrete-time dependent data, and in particular, shrinkage estimators have been applied to multidimensional ergodic diffusion processes:

$$\mathrm{d}X_t = b(X_t, \alpha)\mathrm{d}t + \sigma(X_t, \beta)\mathrm{d}W_t, \quad X_0 = x_0, \tag{2}$$

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 Brownian motion.

In the high frequency setting, regularized estimation problems for discretely observed low-dimensional sparse stochastic differential equations (2) (i.e., the dimensions of the unknown parameters α and β are fixed) have been dealt with in De Gregorio and Iacus (2012), Masuda and Shimizu (2017), Suzuki and Yoshida (2020), Kinoshita and Yoshida (2019) and De Gregorio and Iafrate (2021). The authors used penalized selection procedures (1) based on LASSO (ℓ_1 constraints) and Bridge type penalties (ℓ_2 constraints). The asymptotic oracle properties of the regularized estimators are derived: selection consistency (i.e., consistently estimates null parameters as zero) and optimal rate asymptotic normality of the true subset model. In De Gregorio and Iacus (2018), the LASSO estimator has been analyzed for diffusion processes with small noise observed at continuous time.

Recently, some papers dealt with penalized methods for high-dimensional diffusion processes where the number of parameters as well as the dimension of the model is large; for instance, the reader can consult Gaïffas and Matulewicz (2019), Fujimori (2019), Ciołek et al. (2020), Ciolek et al. (2022), Dexheimer and Strauch (2024) and Amorino et al. (2025). Usually in this framework, the statistical analysis focuses on regularized estimators for the drift term and some non-asymptotic oracle bounds are derived.

The Elastic-Net procedure (see Zou and Hastie (2005) and Zou and Zhang (2009)) is a regularization method for a linear regression model developed to enhance the performance of both LASSO and Ridge regression techniques. It uses a combination of ℓ_1 and ℓ_2 penalties in the regularization term; namely the estimator (1) involves $p(\theta) = \lambda_1|\theta| + \lambda_2|\theta|^2$. The Elastic-Net allows to handle multicollinearity, more effectively than LASSO regression, with the capability of selection of groups of correlated predictors together and then improving the prediction of the underlying model (usually LASSO procedures select one predictor of the group). For an insightful discussion of this method see Zou and Hastie (2005).

In this paper, we introduce an adaptive Elastic-Net estimator for ergodic diffusion processes observed under high-frequency sampling schemes. Up to our knowledge, this is the first attempt to study the Elastic-Net problem for the statistical analysis of stochastic differential equations. Our method combines the least squares approximation introduced in Wang and Leng (2007) with the adaptive ℓ_1 and ℓ_2 regularization; namely, let $\theta = (\alpha, \beta)$, the objective function appearing in (1) becomes:

$$\mathfrak{L}_n(\theta) + L_n(\theta) + R_n(\theta),$$

where $L_n(\theta)$ is the adaptive LASSO penalty, while $R_n(\theta)$ is the Ridge term. The idea is to replace the quasi-log-likelihood function of (2) with its second order Taylor expansion and adding the adaptive ℓ_1 and ℓ_2 penalties. It is worth to mention that a crucial issue is the choice of the initial non-penalized estimator (i.e. the point where the approximation takes place), which has to satisfy some standard properties. By means of this approach the estimator obtained from the minimization problem (1) enhances prediction accuracy and interpretability while effectively recovering the sparse underlying structure of the model. In this way, it is possible to overcome some limitations of the LASSO regularization, especially when we consider multivariate diffusion process having correlated groups of variables. A part of our analysis is devoted to prove some asymptotic properties of the Elastic-Net estimator.

Furthermore, we introduce a block-diagonal Elastic-Net estimator, which is asymptotic equivalent to the original estimator, and prove that non-asymptotic bounds for the estimation error (under suitable assumptions on the contrast function). Our problem is addressed within a low-dimensional framework. Nevertheless, it's worth mentioning that our non-asymptotic results offer insights into the analysis of potential high-dimensional scenarios in this setting. Moreover, we obtain finite sample guarantees for the mean absolute error of predictions at a future time. Our theoretical findings are supported by empirical results that show that introducing a flexible Elastic-Net penalty leads to better model selection and more accurate forecasting when compared to LASSO regularization.

The paper is organized as follows. In Section 2, we introduce the sparse parametric diffusion model and the underlying asymptotic regime. Furthermore, by means of the least squares approximations approach, the adaptive Elastic-Net estimator for stochastic differential equations observed at discrete times is defined. In Section 3 the theoretical features of the estimator are discussed. In particular, we get the oracle properties of the estimator, which are crucial in order to define a reasonable selection procedure. Non-asymptotic bounds, involving the parametric dimension under suitable assumptions, are introduced in Section 4 for a block-diagonal Elastic-Net estimator. This latter is asymptotically equivalent to the estimator

defined in Section 2. Section 5 is devoted to the analysis of the prediction error; in particular, the mean absolute error is studied and its non-asymptotic bounds discussed. In Section 6-7, we assess the performance of the adaptive Elastic-Net estimator, by means of a numerical analysis on simulated data and an application of our methodology to well-being real data (comparing Italy and Japan during the Covid outbreak). The optimization path algorithm is briefly described. All the proofs are collected in the last section.

2 Definition of Elastic-Net estimator for diffusions sampled at discrete times

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):

$$\mathrm{d}X_t = b(X_t, \alpha)\mathrm{d}t + \sigma(X_t, \beta)\mathrm{d}W_t, \quad X_0 = x_0, \tag{3}$$

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 $\mathbf{F}$–Brownian motion.

We assume that $\alpha \in \Theta_\alpha \subset \mathbb{R}^{\mathsf{p}}$, $\beta \in \Theta_\beta \subset \mathbb{R}^{\mathsf{q}}$, $\mathsf{p}, \mathsf{q} \in \mathbb{N}$, are unknown parameters and Θ_α, Θ_β are compact convex sets. In this framework, the parameter of interest is $\theta{:=}(\alpha, \beta)^\intercal$ and the parametric space is given by $\Theta{:=}\Theta_\alpha \times \Theta_\beta \subset \mathbb{R}^{\mathsf{m}}$, where $\mathsf{m}{:=}\mathsf{p} + \mathsf{q}$.

The true value of θ is indicated by $\theta_0{:=}(\alpha_0, \beta_0)^\intercal \in \mathbb{R}^{\mathsf{m}}$. Furthermore, $\theta_0 \in \mathrm{Int}(\Theta)$ and $0 \in \mathbb{R}^{\mathsf{m}}$ belongs to Θ. $P{:=}P_{\theta_0}$ stands for the probability law of X.

We assume that the stochastic differential equation X represents a sparse parametric model. The sparsity condition of the parametric model is introduced by imposing that some coefficients in θ_0 are exactly zero; i.e. $\mathsf{p}^0{:=}\big|\{j : \alpha_{0,j} \neq 0\}\big|$, $\mathsf{q}^0{:=}\big|\{h : \beta_{0,h} \neq 0\}\big|$ and $\mathsf{m}^0{:=}\mathsf{p}^0 + \mathsf{q}^0$.

Let us recall some notations. Let $\langle A, x^{\otimes 2}\rangle = \mathrm{tr}(Ax^{\otimes 2}) = x^\intercal Ax$, where $x^{\otimes 2} = xx^\intercal$, $x \in \mathbb{R}^{\mathsf{m}}$ $A \in \mathbb{R}^{\mathsf{m}} \otimes \mathbb{R}^{\mathsf{m}}$. We denote by $\tau_{\min}(M)$ and $\tau_{\max}(M)$ the smallest and the largest eigenvalues, respectively, of a symmetric matrix M. Furthermore, the Euclidean norm of the vector x is indicated by $|x|$ and $||A|| = \sqrt{\tau_{\max}(AA^\intercal)}$. We sometimes adopt the notation $[n]$ to denote the set $\{1, 2, \ldots, n\}$.

Hereafter, we suppose that X satisfies the following regularity conditions.

(Existence and uniqueness) There exists a constant C such that

$$\sup_{\alpha\in\Theta_\alpha} |b(x,\alpha) - b(y,\alpha)| + \sup_{\beta\in\Theta_\beta} \|\sigma(x,\beta) - \sigma(y,\beta)\| \leq C|x-y|, \quad x, y \in \mathbb{R}^d.$$

(Non-degeneracy) There exists $\tau > 0$ such that $\tau^{-1} \leq \tau_{\min}(\Sigma(x,\beta))$, uniformly in x and β.

(Ergodicity) The process X is ergodic for $\theta = \theta_0$; that is, let $\mu = \mu_{\theta_0}$ be the invariant probability measure of X_t, then

$$\frac{1}{T}\int_0^T g(X_t)\mathrm{d}t \xrightarrow{p} \int_{\mathbb{R}^d} g(x)\mu(\mathrm{d}x),$$

as $T \longrightarrow \infty$, for any bounded measurable function g.

The sample path of X is observed at $n+1$ equidistant discrete times t_i^n, such that $t_i^n - t_{i-1}^n = \Delta_n < \infty$ for $i = 1, \ldots, n$ with $t_0^n = 0$. The collected data are the discrete observations of

the sample path of X, that we represent by $\mathbf{X}_n := (X_{t_i^n})_{0 \le i \le n}$. The main goal of the paper is to estimate θ and simultaneously select the true subset model from the sample $\mathbf{X}_n$.

The asymptotic scheme adopted in this paper is the following

$$n\Delta_n \longrightarrow \infty, n\Delta_n^2 \longrightarrow 0 \text{ and } \Delta_n \longrightarrow 0$$

as $n \to \infty$. We have high frequency sampling and long-run data. Actually, it is possible to weaken the above asymptotic condition by imposing $n\Delta_n^k \longrightarrow 0, k > 2$. In this setting, some estimators for ergodic diffusions have been proposed in Kessler (1997), Uchida and Yoshida (2012) and in Iguchi and Beskos (2025) for degenerate SDEs.

In order to introduce the Elastic-Net estimation methodology for diffusion processes, we take into account the same approach developed in Wang and Leng (2007), De Gregorio and Iacus (2012), Suzuki and Yoshida (2020) and De Gregorio and Iafrate (2021). Let us start by dealing with a generic contrast function $\mathfrak{L}_n(\theta) := \mathfrak{L}_n(\theta; \mathbf{X}_n)$ such that the estimator $\tilde{\theta}_n := \tilde{\theta}_n(\mathbf{X}_n)$ of θ is given by:

$$\tilde{\theta}_n := (\tilde{\alpha}_n, \tilde{\beta}_n)^\intercal \in \arg\min_\theta \mathfrak{L}_n(\theta). \tag{4}$$

If $\mathfrak{L}_n(\theta)$ is twice differentiable with respect to θ, it is approximated by Taylor expansion as follows:

$$\mathfrak{L}_n(\theta) \simeq \mathfrak{L}_n(\tilde{\theta}_n) + \frac{1}{2}\langle \ddot{\mathfrak{L}}_n(\tilde{\theta}_n), (\theta - \tilde{\theta}_n)^{\otimes 2} \rangle$$

where $\ddot{\mathfrak{L}}_n$ represents the Hessian matrix of the second derivatives with respect to θ. Therefore, we simplify the optimization problem by minimizing $\langle \ddot{\mathfrak{L}}_n(\tilde{\theta}_n), (\theta - \tilde{\theta}_n)^{\otimes 2} \rangle$ instead of $\mathfrak{L}_n(\theta)$. This remark inspires the following definition.

Usually, $\tilde{\theta}_n$ minimizes a loss function; nevertheless, we will not assume the existence of a loss function. Let $\hat{G}_n$ be an $\mathsf{m} \times \mathsf{m}$ almost surely positive definite symmetric random matrix depending on n. We introduce the following objective function:

$$\begin{aligned} \mathcal{F}_n(\theta; \tilde{\theta}_n) &= \langle \hat{G}_n, (\theta - \tilde{\theta}_n)^{\otimes 2} \rangle + L_n(\theta) + R_n(\theta) \\ &= |\hat{G}_n^{1/2}(\theta - \tilde{\theta}_n)|^2 + L_n(\theta) + R_n(\theta) \end{aligned} \tag{5}$$

where $\hat{G}_n^{1/2}$ is the symmetric matrix such that $\hat{G}_n^{1/2}\hat{G}_n^{1/2} = \hat{G}_n$ and $L_n(\theta)$ and $R_n(\theta)$ are functions of the parameter of interest and represent the penalization term of Elastic-Net in the objective function. In particular, $L_n(\theta)$ represents the LASSO component of the penalization term:

$$L_n(\theta) = |\alpha|_{1,\kappa_n} + |\beta|_{1,\pi_n} \tag{6}$$

involving the weighted ℓ_1 norms

$$|\alpha|_{1,\kappa_n} = \sum_{j=1}^{\mathsf{p}} \kappa_{n,j}|\alpha_j|$$

and

$$|\beta|_{1,\pi_n} = \sum_{h=1}^{\mathsf{q}} \pi_{n,h}|\beta_h|,$$

where $\kappa_{n,j}$, $j = 1, ..., \mathsf{p}$, and $\pi_{n,h}$, $h = 1, ..., \mathsf{q}$, are sequence of real positive random variable depending on the data and representing the amount of shrinkage for α_j and β_h, respectively.

Remark 1 Usually, the adaptive weights are chosen as follows (Zou (2006))

$$\frac{\lambda_{1,n}}{|\tilde{\alpha}_{n,j}|^{\delta_1}}, \quad j = 1, \dots, \mathsf{p}, \tag{7}$$

$$\frac{\gamma_{1,n}}{|\tilde{\beta}_{n,h}|^{\delta_2}}, \quad h = 1, \dots, \mathsf{q}, \tag{8}$$

where $\lambda_{1,n}$ and $\gamma_{1,n}$ are suitable sequences of positive numbers and $\delta_1, \delta_2 > 0$. In this case we write down

$$L_n(\theta) = \lambda_{1,n}|\alpha|_{1,\kappa_n} + \gamma_{1,n}|\beta|_{1,\pi_n},$$

where the weights appearing in the norms are given by

$$\kappa_{n,j} = \frac{1}{|\tilde{\alpha}_{n,j}|^{\delta_1}}, \quad j = 1, \dots, \mathsf{p}, \tag{9}$$

$$\pi_{n,h} = \frac{1}{|\tilde{\beta}_{n,h}|^{\delta_2}}, \quad h = 1, \dots, \mathsf{q}. \tag{10}$$

The idea is that $\tilde{\theta}_n$ is consistent and then this choice allows to assign higher weights to the zero coefficients.

The Ridge component of the Elastic-Net penalty is:

$$R_n(\theta) = \lambda_{2,n}|\alpha|^2 + \gamma_{2,n}|\beta|^2 \tag{11}$$

where $\lambda_{2,n}$ and $\gamma_{2,n}$ are sequences of positive real values, depending on n.

We observe that if $\tilde{\theta}_n$ coincides with (4), by setting $\hat{G}_n = \ddot{\mathfrak{L}}_n(\tilde{\theta}_n)$, the first term in $\mathcal{F}_n$ is exactly the least squares approximation of the loss function $\mathfrak{L}_n(\theta)$.

The Elastic-Net estimator $\hat{\theta}_n := \hat{\theta}_n(\mathbf{X}_n) : \mathbb{R}^{(n+1)\times d} \longrightarrow \overline{\Theta}$ is defined by minimizing the objective function (5); i.e.

$$\hat{\theta}_n := (\hat{\alpha}_n, \hat{\beta}_n)^{\intercal} \in \arg\min_{\theta \in \overline{\Theta}} \mathcal{F}_n(\theta; \tilde{\theta}_n). \tag{12}$$

The primary advantage of our approach lies in addressing a penalized optimization problem that involves a convex cost function. This characteristic enhances the feasibility of the estimation procedure both theoretically and computationally.

Remark 2 The adaptive LASSO estimator $\hat{\theta}_n(\text{LASSO})$ for stochastic differential equations studied in De Gregorio and Iacus (2012) is defined as follows

$$\hat{\theta}_n(\text{LASSO}) \in \arg\min_{\theta} \left\{ |\hat{G}_n^{1/2}(\theta - \tilde{\theta}_n)|^2 + L_n(\theta) \right\}$$

or equivalently

$$\hat{\theta}_n(\text{LASSO}) \in \arg\min_{\theta} \left\{ \langle \hat{G}_n, \theta^{\otimes 2} \rangle - 2\tilde{\theta}_n^{\intercal} \hat{G}_n \theta + L_n(\theta) \right\}. \tag{13}$$

Clearly, the stationary points of $\mathcal{F}_n(\theta; \tilde{\theta}_n)$ coincides with those of the function

$$\langle \hat{G}_n + C(\lambda_{2,n}, \gamma_{2,n}), \theta^{\otimes 2} \rangle - 2\tilde{\theta}_n^{\intercal} \hat{G}_n \theta + L_n(\theta),$$

where $C(\lambda_{2,n}, \gamma_{2,n}) := \text{diag}(\lambda_{2,n}\mathbf{I}_{\mathsf{p}}, \gamma_{2,n}\mathbf{I}_{\mathsf{q}})$, where $\mathbf{I}_{\mathsf{g}}$ stands for the identity matrix of size g. Therefore, the Elastic-Net estimator becomes:

$$\hat{\theta}_n \in \arg\min_{\theta} \left\{ \langle \hat{G}_n + C(\lambda_{2,n}, \gamma_{2,n}), \theta^{\otimes 2} \rangle - 2\tilde{\theta}_n^{\intercal} \hat{G}_n \theta + L_n(\theta) \right\};$$

i.e. $\hat{\theta}_n$ is a stabilized version of (13) where the Ridge adjustment is taken into account by the perturbation of $\hat{G}_n$ with the rates $\lambda_{2,n}$ and $\gamma_{2,n}$.

3 Theoretical properties of Elastic-Net estimator

Without loss of generality, we set $\alpha_{0,j} = 0$ and $\beta_{0,h} = 0$ for $j = \mathsf{p}^0 + 1, \ldots, \mathsf{p}$ and $h = \mathsf{q}^0 + 1, \ldots, \mathsf{q}$ and denote by $A_n := \text{diag}\left(\frac{1}{\sqrt{n\Delta_n}}\mathbf{I}_{\mathsf{p}}, \frac{1}{\sqrt{n}}\mathbf{I}_{\mathsf{q}}\right)$.

We introduce the following assumptions.

P1. The initial estimator $\tilde{\theta}_n := \left(\tilde{\alpha}_n, \tilde{\beta}_n\right)^{\intercal} : \mathbb{R}^{(n+1)\times d} \longrightarrow \overline{\Theta}$ of θ is consistent in a mixed-rates asymptotic regime, $\tilde{\alpha}_n$ is $\sqrt{n\Delta_n}$-consistent while $\tilde{\beta}_n$ is $\sqrt{n}$-consistent:

$$\left(\sqrt{n\Delta_n}(\tilde{\alpha}_n - \alpha_0), \sqrt{n}(\tilde{\beta}_n - \beta_0)\right)^{\intercal} = O_p(1).$$

P2. $\tilde{\theta}_n$ is asymptotically normal; i.e.

$$\left((\sqrt{n\Delta_n}(\tilde{\alpha}_n - \alpha_0),\ \sqrt{n}(\tilde{\beta}_n - \beta_0)\right)^{\intercal} \xrightarrow{d} N_{\mathsf{m}}\left(0, \Gamma(\theta_0)^{-1}\right),$$

where $\Gamma(\theta_0) := \text{diag}\left(\Gamma^{\alpha\alpha}(\theta_0), \Gamma^{\beta\beta}(\beta_0)\right)$ is the asymptotic Fisher information and

$$\Gamma^{\alpha\alpha}(\theta_0) := \left[\int_{\mathbb{R}^d} \partial_{\alpha_i} b^{\intercal}(x, \alpha_0)\Sigma^{-1}(x, \beta_0)\partial_{\alpha_j} b(x, \alpha_0)\mu(\mathrm{d}x)\right]_{i,j=1}^{\mathsf{p}},$$

$$\Gamma^{\beta\beta}(\beta_0) := \left[\frac{1}{2}\int_{\mathbb{R}^d} \text{tr}\left[(\partial_{\beta_i}\Sigma)\Sigma^{-1}(\partial_{\beta_j}\Sigma)\Sigma^{-1}(x, \beta_0)\right]\mu(\mathrm{d}x)\right]_{i,j=1}^{\mathsf{q}}.$$

P3. Let G be a $\mathsf{m} \times \mathsf{m}$ matrix assumed to be symmetric and positive definite and such that:

$$\hat{D}_n := A_n \hat{G}_n A_n \xrightarrow{p} G.$$

P4. There exist two positive definite symmetric random matrices $G^{\alpha\alpha} \in \mathbb{R}^{\mathsf{p}} \otimes \mathbb{R}^{\mathsf{p}}$ and $G^{\beta\beta} \in \mathbb{R}^{\mathsf{q}} \otimes \mathbb{R}^{\mathsf{q}}$ such that:

$$\hat{D}_n \xrightarrow{p} G := \text{diag}(G^{\alpha\alpha}, G^{\beta\beta}).$$

Remark 3 The conditions P1 and P2 reveal the mixed-rates asymptotic nature of a good estimation procedure for ergodic diffusion processes. This is the main reason leading to introduce in the penalty terms L_n and R_n two different norms for each group of parameters.

Example 3.1 (Quasi-likelihood estimation) The maximum quasi-likelihood estimator, indicated by $\tilde{\theta}_n^{(\text{QL})}$, is obtained by maximizing the contrast function given by the quasi-log-likelihood function

$$\ell_n(\theta) := -\frac{1}{2}\sum_{i=1}^{n}\left\{\log\det(\Sigma(X_{t_{i-1}^n}, \beta)) + \frac{1}{\Delta_n}\langle\Sigma^{-1}(X_{t_{i-1}^n}, \beta), (X_{t_i^n} - X_{t_{i-1}^n} - \Delta_n b(X_{t_{i-1}^n}, \alpha))^{\otimes 2}\rangle\right\}; \tag{14}$$

i.e. $\tilde{\theta}_n^{(\text{QL})} \in \arg\max_\theta \ell_n(\theta)$. The terms appearing in the sum (14) represent a local-Gaussian approximation of the transition density of X between two observations $X_{t_{i-1}^n}$ and $X_{t_i^n}$, arising from the Euler–Maruyama discretization scheme (see Kloeden and Platen (1992)). This approach was originally introduced in Florens-Zmirou (1989) with $\sigma(x, \beta) = \beta$. Therefore,

we can assume that $\mathfrak{L}_n(\theta) = -\ell_n(\theta)$ and a possible choice of $\hat{G}_n$ is the Hessian matrix $-\ddot{\ell}_n(\tilde{\theta}_n^{(\mathrm{QL})})$ evaluated in $\tilde{\theta}_n^{(\mathrm{QL})}$. Under some mild conditions, the estimator $\tilde{\theta}_n^{(\mathrm{QL})}$ and $\ddot{\ell}_n(\tilde{\theta}_n^{(\mathrm{QL})})$ behave properly and the properties P1, P2 and P3, P4 are satisfied. In particular

$$\hat{D}_n = A_n(-\ddot{\ell}_n(\tilde{\theta}_n^{(\mathrm{QL})}))A_n \xrightarrow{p} \Gamma(\theta_0)$$

(actually uniformly in θ). The maximum quasi-likelihood estimation for the ergodic stochastic differential equations has been developed, for instance, in Yoshida (2011) by means of a polynomial-type inequality for suitable statistical random fields. In this setting, P1, P2 and P3 hold under the following additional assumptions on the diffusion. For $l \geq 1$ and $m \geq 1$, let $f(x,\theta) \in C_{\uparrow}^{l,m}(\mathbb{R}^d \times \Theta, \mathbb{R})$ be a space such that $f(x,\theta)$ is continuously differentiable with respect to x up to order l for all θ, $f(x,\theta)$ and all its x-derivatives up to order l are m times continuously differentiable with respect to θ and $f(x,\theta)$ and all derivatives are of polynomial growth in x uniformly in θ.

(Smoothness)
$b \in C_{\uparrow}^{0,4}(\mathbb{R}^d \times \Theta_\alpha, \mathbb{R}^d)$ and $\sigma \in C_{\uparrow}^{2,4}(\mathbb{R}^d \times \Theta_\beta, \mathbb{R}^d \otimes \mathbb{R}^r)$.
(Mixing) There exists a positive constant a such that

$$v_X(u) \leq \frac{e^{-au}}{a}, \quad u > 0$$

where

$$v_X(u) = \sup_{t \geq 0} \sup_{\substack{A \in \sigma\{X_r : r \leq t\} \\ B \in \sigma\{X_r : r \geq t+u\}}} |P(A \cap B) - P(A)P(B)|.$$

(Uniform boundedness) $\sup_t E[|X_t|^k] < \infty$ for all $k > 0$.
(Identifiability) $b(x,\alpha) = b(x,\alpha_0)$ for μ_{θ_0} *a.s.* all $x \Rightarrow \alpha = \alpha_0$,
$\Sigma(x,\beta) = \Sigma(x,\beta_0)$ for μ_{θ_0} *a.s.* all $x \Rightarrow \beta = \beta_0$

The above assumptions entail, in particular that the ergodic property holds true. Such property, representing a kind of law of large numbers is required for the consistent estimation of the parameter α appearing in the drift term. Mixing and moment conditions are hard to verify in practice. Nevertheless, it is known that they hold under the following conditions (see Pardoux and Veretennikov (2001), Veretennikov (1988)): Σ is bounded and there exist positive constants λ_-, λ_+ and Λ such that for all β

$$0 < \lambda_- \leq \langle \Sigma(x,\beta)x/|x|, x/|x| \rangle \leq \lambda_+, \quad \frac{\mathrm{Tr}(\Sigma(x,\beta))}{d} \leq \Lambda$$

and for all α

$$\langle b(x,\alpha), x/|x| \rangle \leq -r|x|^a, \quad |x| \geq M_0,$$

with $M_0 \geq 0$, $a \geq -1$ and $r > 0$.

Other estimators for ergodic diffusions satisfying the asymptotic properties P1, P2 and P3 appeared in the literature. For instance, the quasi-Bayesian estimator is studied in Yoshida (2011), while the hybrid multistep estimator has been introduced in Kamatani and Uchida (2015). An approach based on the approximate martingale estimating functions is discussed in Sørensen (2024). If $n\Delta_n^k \longrightarrow 0, k \geq 2$, estimators defined by local-Gaussian contrast functions involving higher-order expansions of the conditional moments of X_t, have been proposed in Kessler (1997) and Uchida and Yoshida (2012). It seems possible to relax the

non-degeneracy assumption on the diffusion matrix, and subsequently develop our procedure for degenerate SDEs by using the estimators introduced in Ditlevsen and Samson (2019), Gloter and Yoshida (2021) and Iguchi and Beskos (2025).

Let us introduce $a_n := \max\{\kappa_{n,j};\ j \leq \mathsf{p}^0\}$, $b_n := \min\{\kappa_{n,j};\ j > \mathsf{p}^0\}$, $c_n := \max\{\pi_{n,h};\ h \leq \mathsf{q}^0\}$ and $d_n := \min\{\pi_{n,h};\ h > \mathsf{q}^0\}$. We introduce the following conditions.

A1. $\frac{a_n}{\sqrt{n\Delta_n}} = O_p(1)$ and $\frac{c_n}{\sqrt{n}} = O_p(1)$.
A2. $\frac{\lambda_{2,n}}{\sqrt{n\Delta_n}} = O(1)$ and $\frac{\gamma_{2,n}}{\sqrt{n}} = O(1)$,
A3. $\frac{a_n}{\sqrt{n\Delta_n}} = o_p(1)$ and $\frac{c_n}{\sqrt{n}} = o_p(1)$.
A4. $\frac{b_n}{\sqrt{n\Delta_n}} \xrightarrow{p} \infty$ and $\frac{d_n}{\sqrt{n}} \xrightarrow{p} \infty$.[1]

We denote $\alpha_\star := (\alpha_1, \ldots, \alpha_{\mathsf{p}^0})^\intercal$, $\beta_\star := (\beta_1, \ldots, \beta_{\mathsf{q}^0})^\intercal$ and $\alpha_\bullet := (\alpha_{\mathsf{p}^0+1}, \ldots, \alpha_{\mathsf{p}})^\intercal$, $\beta_\bullet := (\beta_{\mathsf{q}^0+1}, \ldots, \beta_{\mathsf{q}})^\intercal$. For the discussion of the next result we need to introduce the following notation on a block matrix $M \in \mathbb{R}^u \otimes \mathbb{R}^v$, where $u, v \in \{\mathsf{p}, \mathsf{q}\}$,

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

where for $u^0, v^0 \in \{\mathsf{p}^0, \mathsf{q}^0\}$,

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

Furthermore, under P4 we introduce the following $\mathsf{m}^0 \times \mathsf{m}$ matrix

$$\mathfrak{G} := \begin{pmatrix} \mathfrak{G}_\alpha & 0 \\ 0 & \mathfrak{G}_\beta \end{pmatrix},$$

where $\mathfrak{G}_\alpha := (\mathbf{I}_{\mathsf{p}^0} \ \ (G^{\alpha\alpha}_{\star\star})^{-1} G^{\alpha\alpha}_{\star\bullet})$ and $\mathfrak{G}_\beta := (\mathbf{I}_{\mathsf{q}^0} \ \ (G^{\beta\beta}_{\star\star})^{-1} G^{\beta\beta}_{\star\bullet})$.

Now, we are able to prove the crucial asymptotic Oracle properties for the estimator (12).

Theorem 1 (Oracle properties) *The adaptive Elastic-Net estimator $\hat{\theta}_n$ satisfies the following properties.*

i) (Consistency) Let us assume that P1, P3, A1 and A2 hold. Then

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

(ii) (Selection consistency) Under the assumptions P1, P3, A1, A2 and A4, we have that:

$$P(\hat{\alpha}_{n\bullet} = 0) \longrightarrow 1 \text{ and } P(\hat{\beta}_{n\bullet} = 0) \longrightarrow 1$$

as $n \longrightarrow \infty$.

(iii) (Asymptotic normality) Let us assume that P2, P4, A3 and A4 are fulfilled. Then

$$\left(\sqrt{n\Delta_n}(\hat{\alpha}_n - \alpha_0)_\star, \sqrt{n}(\hat{\beta}_n - \beta_0)_\star\right)^\intercal \xrightarrow{d} N_{\mathsf{m}^0}\left(0, \mathfrak{G}\Gamma(\theta_0)^{-1}\mathfrak{G}\right)$$

[1] A sequence of random variables X_n converges to ∞ in probability, we write $X_n \xrightarrow{p} \infty$ if $P(X_n > K) \to 1$ for every $K > 0$.

as $n \longrightarrow \infty$. *Furthermore if* $G = \Gamma(\theta_0)$

$$\mathfrak{G}\Gamma(\theta_0)^{-1}\mathfrak{G} = \text{diag}\left(\Gamma_{\star\star}^{\alpha\alpha}(\theta_0)^{-1}, \Gamma_{\star\star}^{\beta\beta}(\beta_0)^{-1}\right).$$

We recall that a sequence of random variables $(X_n)_{n\geq 1}$ is uniformly L^r-bounded if and only if $\sup_{n\geq 1}\mathbb{E}\left[|X_n|^r\right] < \infty$, for all $r \geq 1$.

Theorem 2 (Uniform L^r-boundness) *By assuming A2 and* $\left(\hat{D}_n\right)_{n\geq 1}$, $\left(\hat{D}_n^{-1}\right)_{n\geq 1}$, $\left(A_n^{-1}(\tilde{\theta}_n - \theta_0)\right)_{n\geq 1}$, $\left(\frac{a_n}{\sqrt{n\Delta_n}}\right)_{n\geq 1}$, $\left(\frac{c_n}{\sqrt{n}}\right)_{n\geq 1}$ *uniformly* L^r*-bounded, then*

$$\sup_{n\geq 1}\mathbb{E}\left[|A_n^{-1}(\hat{\theta}_n - \theta_0)|^r\right] < \infty, \quad r \geq 1. \tag{15}$$

Remark 4 The assumptions required in Theorem 2 are not strong since, for instance, the maximum quasi-likelihood and Bayesian estimators are uniformly L^r-bounded (see Yoshida (2011)). Furthermore, we can chose the weights properly, in order to get L^r-bounded coefficients a_n and c_n; for instance by setting

$$\kappa_{n,j} = \frac{\lambda_1}{|\tilde{\alpha}_{n,j}|^{\delta_1} + a_n}, \quad \pi_{n,h} = \frac{\gamma_1}{|\tilde{\beta}_{n,h}|^{\delta_2} + b_n}$$

or

$$\kappa_{n,j} = \frac{\lambda_1}{(|\tilde{\alpha}_{n,j}| \vee a_n)^{\delta_1}}, \quad \pi_{n,h} = \frac{\gamma_1}{(|\tilde{\beta}_{n,h}| \vee b_n)^{\delta_2}}$$

with $a_n, b_n > 0$ and such that $a_n, b_n \downarrow 0$ and $\lambda_1, \gamma_1 > 0$.

The L^r estimates are useful, for example, in the asymptotic decision theory, where the estimator is efficient if it attains the Hajek's minimax bound, or for the validation of the information criterion.

4 Block-diagonal estimator and non-asymptotic bounds

In this section, we derive non-asymptotic results for the Elastic-Net estimator in the case where the $\hat{G}_n$ information matrix is taken to be block-diagonal. This is a mild restriction, since in most cases we have that the off-diagonal blocks of the scaled information matrix $\hat{D}_n^{\alpha\beta}$ converge to zero in probability. This is the case in the notable example of the hessian matrix in quasi-likelihood estimation, when evaluated at $\tilde{\theta}_n^{(QL)}$.

Specifically, let us assume $\hat{G}_n = \text{diag}(\hat{G}_n^{\alpha\alpha}, \hat{G}_n^{\beta\beta})$ which satisfies condition P4. In this case, the objective function can be decomposed as follows

$$\mathcal{F}_n(\theta; \tilde{\theta}_n) = \mathcal{F}_{1,n}(\alpha; \tilde{\alpha}_n) + \mathcal{F}_{2,n}(\beta; \tilde{\beta}_n)$$

where

$$\begin{aligned}\mathcal{F}_{1,n}(\alpha; \tilde{\alpha}_n) &= \langle \hat{G}_n, (\alpha - \tilde{\alpha}_n)^{\otimes 2}\rangle + |\alpha|_{1,\kappa_n} + \lambda_{2,n}|\alpha|^2 \\ &= |(\hat{G}_n^{\alpha\alpha})^{1/2}(\alpha - \tilde{\alpha}_n)|^2 + |\alpha|_{1,\kappa_n} + \lambda_{2,n}|\alpha|^2\end{aligned}$$

and

$$\mathcal{F}_{2,n}(\beta; \tilde{\beta}_n) = \langle \hat{G}_n^{\beta\beta}, (\beta - \tilde{\beta}_n)^{\otimes 2}\rangle + |\beta|_{1,\pi_n} + \gamma_{2,n}|\beta|^2$$

$$= |(\hat{G}_n^{\beta\beta})^{1/2}(\beta - \tilde{\beta}_n)|^2 + |\beta|_{1,\pi_n} + \gamma_{2,n}|\beta|^2$$

Furthermore, $\hat{\alpha}_n \in \arg\min_\alpha \mathcal{F}_{1,n}(\alpha; \tilde{\alpha}_n)$ and $\hat{\beta}_n \in \arg\min_\beta \mathcal{F}_{2,n}(\beta; \tilde{\beta}_n)$.

The following theorem provides non-asymptotic bounds for the Elastic-Net estimator.

Theorem 3 (Non-asymptotic bounds) *Let us assume* $\hat{G}_n = diag(\hat{G}_n^{\alpha\alpha}, \hat{G}_n^{\beta\beta})$ *satisfying P4 and consider* $\lambda_{1,n}, \lambda_{2,n}, \gamma_{1,n}$ *and* $\gamma_{2,n}$ *defined as in* (7), (8) *and* (11). *The error bounds of the Elastic-Net estimators* $\hat{\alpha}_n$ *and* $\hat{\beta}_n$ *are given by:*

$$|\hat{\alpha}_n - \alpha_0| \leq \frac{2}{n\Delta_n \tau_{\min}(\hat{D}_n^{\alpha\alpha}) + \lambda_{2,n}} \left(\lambda_{2,n}|\alpha_0| + n\Delta_n \tau_{\max}(\hat{D}_n^{\alpha\alpha})|\tilde{\alpha}_n - \alpha_0| + \lambda_{1,n}|\kappa_n|\right), \quad a.s., \tag{16}$$

and

$$|\hat{\beta}_n - \beta_0| \leq \frac{2}{n\tau_{\min}(\hat{D}_n^{\beta\beta}) + \gamma_{2,n}} \left(\gamma_{2,n}|\beta_0| + n\tau_{\max}(\hat{D}_n^{\beta\beta})|\tilde{\beta}_n - \beta_0| + \gamma_{1,n}|\pi_n|\right), \quad a.s. \tag{17}$$

Remark 5 If $\frac{\lambda_{2,n}}{\sqrt{n\Delta_n}} \longrightarrow 0$ and $\frac{\gamma_{2,n}}{\sqrt{n}} \longrightarrow 0$, one has immediately that P1 leads to

$$A_n(\hat{G}_n + C(\lambda_{2,n}, \gamma_{2,n}))A_n \xrightarrow{p} G.$$

Therefore, Remark 2 implies that the Oracle properties shown in De Gregorio and Iacus (2012) and De Gregorio and Iafrate (2021) hold true for $\hat{\theta}_n$.

In the previous theorem the properties of the estimator depend on the error of the initial estimator, which in particular depends implicitly on the dimension, with the advantage of requiring very little assumptions. In particular, we observe that the bounds appearing in Theorem 3 contain the factors

$$\frac{2}{n\Delta_n \tau_{\min}(\hat{D}_n^{\alpha\alpha}) + \lambda_{2,n}}, \quad \frac{2}{n\tau_{\min}(\hat{D}_n^{\beta\beta}) + \gamma_{2,n}}.$$

The tuning parameter $\lambda_{2,n}$ appears in both denominators, a point which reveals that the Ridge penalty is crucial to avoid the explosion of the upper bounds.

We recall that our estimator is analyzed within a low-dimensional setting. Nevertheless, in the following theorem we provide some insights into potential high-dimensional features of the estimator by means of a high-probability result. We need to introduce some extra assumptions in order to control the finite sample behavior of the contrast and the Hessian matrix. Moreover, we state the following results in the case where the initial estimator is the quasi-likelihood estimator.

Dealing with the non-asymptotic properties of the estimator, we state the following assumptions, which we need in order to control the finite-sample behavior of the contrast function and information matrix. Denote by $\partial_\theta \ell_n$ and $\partial^2_{\theta,\theta} \ell_n$, respectively, the gradient and Hessian matrix of the quasi-log-likelihood function ℓ_n, defined according to (14), and by $\partial_\theta \overline{\ell_n} = A_n \partial_\theta \ell_n$, $\partial^2_{\theta\theta} \overline{\ell_n} = A_n \partial^2_{\theta\theta} \ell_n A_n$ their scaled version.

We need the following two conditions.

A5(r). (*Regular contrast*) The negative quasi-likelihood and its θ-derivatives ℓ_n, $\partial_\theta \ell_n$, $\partial^2_{\theta\theta} \ell_n$ can be extended continuously to the boundary of Θ and, for $r > 0$, there exist a sequence of positive a.s. random variables ξ_n and $\mu > 0$ s.t.

$$(i) \quad \max_{i \in [\mathsf{p}+\mathsf{q}]} \sup_{\theta : |\theta_0 - \theta| \leq r} |\partial_{\theta_i} \overline{\ell_n}(\theta)| \leq \xi_n,$$

$$(ii) \quad \inf_{v\in\mathbb{R}^{\mathsf{p}+\mathsf{q}}:|v|=1} \inf_{\theta:|\theta-\tilde{\theta}_n|\le r} v^\top \partial^2_{\theta\theta}\overline{\ell_n}(\theta)v \ge \mu$$

for all n, P_{θ_0}–a.s..

A6. (*Regular information*) The random matrix $\hat{D}_n$ satisfies

$$\tau_1 \le \tau_{\min}(\hat{D}_n) \le \tau_{\max}(\hat{D}_n) \le \tau_2$$

for some $\tau_1, \tau_2 > 0$, for all n, P_{θ_0}–a.s..

Remark 6 Condition A5(r)-(ii) can be seen as a finite sample version of the so-called identifiability condition (see, e.g., Ciolek et al. (2022),Amorino et al. (2025)).

Lemma 4 (Non-asymptotic bound for QMLE) *Let $\tilde{\theta}_n^{(QL)} = (\tilde{\alpha}_n^{(QL)}, \tilde{\beta}_n^{(QL)})$ be the QL estimator. Under A5(r), on the event $\{|\tilde{\theta}_n^{(QL)} - \theta_0| \le r\}$, we have*

$$\sqrt{n\Delta_n}|\tilde{\alpha}_n^{(QL)} - \alpha_0| \le \frac{2\xi_n}{\mu}\sqrt{\mathsf{p}}, \qquad \sqrt{n}|\tilde{\beta}_n^{(QL)} - \beta_0| \le \frac{2\xi_n}{\mu}\sqrt{\mathsf{q}}. \tag{18}$$

Theorem 5 (Non-asymptotic bounds II) *Suppose that $\tilde{\theta}_n = \tilde{\theta}_n^{(QL)}$. For any $r > 0$, under the assumptions of Theorem 3, and, in addition, under A5($r/n\Delta_n$) - A6, the Elastic-Net estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$ satisfies the bounds*

$$|\hat{\alpha}_n - \alpha_0| \le \frac{2}{n\Delta_n\tau_1 + \lambda_{2,n}}\left(\lambda_{2,n}|\alpha_0| + \frac{\tau_2\xi_n}{\mu}\sqrt{n\Delta_n\,\mathsf{p}} + \lambda_{1,n}|\kappa_n|\right), \tag{19}$$

and

$$|\hat{\beta}_n - \beta_0| \le \frac{2}{n\tau_1 + \gamma_{2,n}}\left(\gamma_{2,n}|\beta_0| + \frac{\tau_2\xi_n}{\mu}\sqrt{n\,\mathsf{q}} + \gamma_{1,n}|\pi_n|\right) \tag{20}$$

with probability at least $1 - C_L/r^L$, for some $L > 0$.

Remark 7 In the previous theorem, the parameter $r > 0$ plays the role of a tuning parameter for the above inequalities. In particular, it controls the finite-sample regularity of the contrast in a neighborhood of the maximum likelihood estimator. For fixed n, the larger the desired probability, the larger has to be the neighborhood in which A5($r/n\Delta_n$) holds. At the same time, as n increases, the regularity requirement is less restrictive. Note also that this property holds asymptotically, in view of the fact that, under mild regularity assumptions, the quasi-likelihood satisfies an asymptotically quadratic approximation (see Yoshida (2011)).

Furthermore, we can deal with a scenario closer to the high-dimensional setting by assuming $\mathsf{p} = \mathsf{p}_n = O((n\Delta_n)^{\nu_1})$, $\nu_1 \in (0, 1)$, and $\mathsf{q} = \mathsf{q}_n = O(n^{\nu_2})$, $\nu_2 \in (0, 1)$. In this case the parameters tend to be large as n increases and by exploiting the bounds (19) and (20) (and condition A3), it is possible to conclude that $\hat{\alpha}_n$ and $\hat{\beta}_n$ are consistent.

5 Prediction error

We now turn our attention to the analysis of the prediction error. Our aim is to obtain finite sample guarantees for the mean absolute error of predictions at a future time. Assume that

A7. The drift function is Lipschitz continuous w.r.t. α with constant possibly depending on x and uniformly bounded in $L^2(P)$; i.e.

$$|b(x,\alpha_1) - b(x,\alpha_2)| \le C(x)|\alpha_1 - \alpha_2| \quad \forall \alpha_1, \alpha_2 \in \Theta_\alpha$$

and $D = \sup_t \mathbb{E}C^2(X_t) < \infty$.

Remark 8 Assumption A7 is satisfied, for example, by a linear drift or when the drift is of the form $b(x,\alpha) = \sum_j \alpha_j b_j(x)$ for Lipschitz functions b_j. We cannot just assume that $C(x)$ is finite a.s., as it would fail even in the linear drift case.

The construction of the predictor is based on the Euler–Maruyama approximation of X, starting from X_t, that is:

$$\tilde{X}_{t+h} = \tilde{X}_{t+h}(W,\theta) = X_t + b(X_t,\alpha)h + \sigma(X_t,\beta)\Delta W_h \tag{21}$$

where $\Delta W_h = W_{t+h} - W_t$. By standard arguments, it can be shown that

$$\mathbb{E}[|\tilde{X}_{T+h} - X_{T+h}|^2 | \mathcal{F}_T] \le Ah$$

where the constant A depends on the Lipschitz constant of b and σ and the moments of X_0, but not on h (see Kloeden and Platen (1992), Th 10.2.2). Our goal is to obtain a prediction for the future value X_{T+h}, but we cannot directly exploit (21) since, in general, we do not have access to the underlying Brownian motion W nor to the true parameter θ_0. For this reason, we introduce an independent Brownian motion W' and, given a $\mathcal{F}_T$-measurable estimator $\hat{\theta}$, we define the one-step predictor:

$$\hat{X}_{T+h} = \mathbb{E}[\tilde{X}_{T+h}(W',\hat{\theta})|\mathcal{F}_T] = X_T + b(X_T,\hat{\alpha})h\,. \tag{22}$$

Given a sample $\{X_{t_i}, i \in [n]\}$ and the Elastic-Net estimator satisfying the assumptions of Theorem 3, we give an error bound for the prediction error associated with the estimator $\hat{X}_{T_n+h}$, where $T_n = n\Delta_n$ is fixed.

Theorem 6 (Non-Asymptotic prediction error bounds) *Under the assumptions of Theorem 3, A6 and A7, there exist positive constants C_1, C_2 and C_3, such that*

$$\begin{aligned}\mathsf{MAE}(\hat{X}_{T_n+h}) &= \mathbb{E}|X_{T_n+h} - \hat{X}_{T_n+h}| \\ &\le \sqrt{h}C_1 + hC_2 + \frac{C_3 h}{T_n\tau_1 + \lambda_{2,n}}\left(\lambda_{2,n}|\alpha_0| + \tau_2 T_n\sqrt{\mathbb{E}|\tilde{\alpha}_n - \alpha_0|^2} + \lambda_{1,n}\sqrt{\mathbb{E}|\kappa_n|^2}\right).\end{aligned} \tag{23}$$

The intuition behind the previous result is as follows. Suppose h is small, so that $\sqrt{h}+h \lesssim \sqrt{h}$, and that n is large, so that, heuristically, we can write $\mathbb{E}T_n|\tilde{\alpha}_n - \alpha_0|^2 \sim \mathsf{p}$. Considering adaptive weights as in Remark 4, by taking $a_n \sim \sqrt{T_n}$, $\delta_1 = 1$, one has that $E|\kappa_n| \lesssim \sqrt{T_n}$. Equation (23) essentially tells us that the prediction mean absolute error behaves as

$$\mathsf{MAE}(\hat{X}_{T_n+h}) \lesssim \sqrt{h} + \frac{h\sqrt{\mathsf{p}}}{\sqrt{T_n}} + O(T_n^{-1/2})\,. \tag{24}$$

The first term plays the role of an *irreducible* error related to the forecasting task, showing that the predictions degrade with the forecasting horizon. The second source of error is related to the estimation task: it improves with larger sample sizes, it worsens in higher dimensions.

6 Elastic-Net Optimization and Numerical Analysis

6.1 Optimization Algorithm and Coefficient Path

To obtain an explicit solution in an optimization problem where the objective function (5) can be decomposed as $\mathcal{F}_n(\theta; \tilde{\theta}_n) = g + h$, with $g = \langle \hat{G}_n, (\theta - \tilde{\theta}_n)^{\otimes 2} \rangle + R_n(\theta)$ being convex and differentiable, and $h = L_n(\theta)$ being convex but not differentiable, iterative proximal gradient algorithms can be used. Specifically, accelerated gradient methods use a weighted combination of the current estimate and the previous gradient directions, and the proximal map function is iteratively applied till convergence of the algorithm.

In particular, in the objective function (5) we can consider the adaptive ℓ_1-penalty as the function h in the above decomposition. It is possible to use the coordinate descent iterative algorithm where, at iteration t, the update is:

$$\hat{\theta}_k^{t+1} = \arg\min_{\theta_k} \mathcal{F}_n \left(\hat{\theta}_1^t, \ldots, \hat{\theta}_{k-1}^t, \theta_k, \hat{\theta}_{k+1}^t, \ldots, \hat{\theta}_{\mathsf{p}+\mathsf{q}}^t \right) . \tag{25}$$

Fixing the sample size n, the coefficients in (7), (8) and (11) no longer depend on n. Specifically, we consider the penalization constant $\lambda \cdot \gamma$ for the LASSO component (instead of using $\lambda_{1,n}$ and $\gamma_{1,n}$), while, similarly, we consider $\lambda(1-\gamma)$ for the Ridge component (instead of $\lambda_{2,n}$ and $\gamma_{2,n}$), where $\lambda > 0$ and $\gamma \in (0, 1]$. The tuning term γ controls the relative proportion between the regularization penalties ℓ_1 and ℓ_2; i.e. $\gamma = 1$, it reduces to LASSO.

Let w_k denote the adaptive weight linked to the parameter θ_k, defined according to (7) and (8), which represents an element of the diagonal matrix $W = \mathrm{diag}(\kappa_{n,1}, \ldots, \kappa_{n,\mathsf{p}}, \pi_{n,1}, \ldots, \pi_{n,\mathsf{q}})$. The coordinate descent updates take the form:

$$\hat{\theta}_k^t = \frac{1}{1 + \frac{\lambda}{\hat{G}_{(k,k)}}(1-\gamma)} \mathcal{S}_{\frac{\lambda\gamma}{\hat{G}_{(k,k)}} w_k} \left(\tilde{\theta}_k - \frac{1}{\hat{G}_{(k,k)}} \sum_{i \neq k} \hat{G}_{(k,i)} (\hat{\theta}_i^{t-1} - \tilde{\theta}_i) \right) \tag{26}$$

or equivalently

$$\hat{\theta}_k^t = \frac{1}{\hat{G}_{(k,k)} + \lambda(1-\gamma)} \mathcal{S}_{\lambda\gamma w_k} \left(\hat{G}_{(k,k)} \hat{\theta}_k^{t-1} - \hat{G}_{k\cdot} (\hat{\theta}^{t-1} - \tilde{\theta}) \right) \tag{27}$$

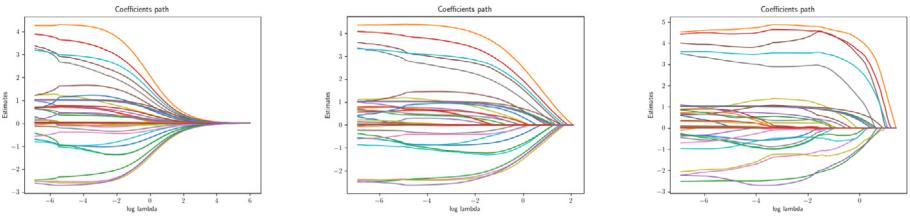
for $k = 1, \ldots, \mathsf{p}+\mathsf{q}$, $t \geq 1$, where $\hat{G}_{(i,k)}$ is the element in the i-th row and k-th column of $\hat{G}_n$ and $\mathcal{S}_\mu(z) := \mathrm{sign}(z)(z-\mu)_+$ is the soft-thresholding operator.

From this expression, it is immediate to check that the largest λ value that is necessary to consider, $\lambda_{\max}$, is:

$$\inf\{\lambda > 0 : \hat{\theta}(\lambda) = 0\} \leq \|(\gamma W)^{-1} \hat{G} \tilde{\theta}\|_\infty =: \lambda_{\max} .$$

A more efficient algorithm for the computation of the solution path can be obtained by applying Proximal Accelerated Gradient Descent algorithms. In the LASSO case, a popular algorithm of this type is the so-called FISTA(Beck and Teboulle (2009)). See Hastie et al. (2015) for an introduction to the subject. In order to obtain the solution of (5), we apply Algorithm 1 in De Gregorio and Iafrate (2024), which is based on the proximal map for the Elastic-Net problem, given by

$$\mathrm{prox}_{s,\lambda,w}^{EN} = \arg\min_u \left\{ \frac{1}{2}(x-u)^2 + \lambda\gamma w|u| + \frac{\lambda(1-\gamma)}{2} u^2 \right\}$$



(A) $\gamma = 0.01$ (B) $\gamma = 0.5$ (C) $\gamma = 1$

Fig. 1 Elastic-Net coefficient paths as the tuning parameter λ for different values of the mixing parameter γ

$$= \frac{1}{1+\lambda s(1-\gamma)} \text{sign}(x)(|x| - \lambda\gamma s w)_+$$
$$= \frac{1}{1+\lambda s(1-\gamma)} \mathcal{S}_{\lambda\gamma s w}(x) \tag{28}$$

where s denotes the stepsize.

Improvements beyond the reported algorithm can be achieved by leveraging block-wise Proximal Alternating Minimization (referred to in the literature as PALM, see Bolte et al. (2014)), which updates one block of parameters at each step (see De Gregorio and Iafrate (2024), for further insight regarding pathwise solution algorithms).

In a diffusion process, as in (3), the classical cross-validation technique for tuning parameter selection λ is not applicable due to the data dependency structure. For this reason, a data-driven technique is used, based on an iterative algorithm that considers the Euler discretization of the solution of (3) and, given estimates of the parameters, a score function of the current estimated residuals (see De Gregorio and Iafrate (2021)). Another criterion for choosing λ is to use the AIC method, penalizing the log-likelihood function (14).

Figure 1 shows the solution path computed by the coordinate descent updates (26) for an Ornstein–Uhlenbek model

$$\mathrm{d}X_t = -BX_t\mathrm{d}t + A\mathrm{d}W_t$$

where A, B are parameter SPD matrices in $\mathbb{R}^{d\times d}$, with $d = 5$.

Among all the λ values that make up the solution path for the Elastic-Net, the optimal value, λ_{opt}, is the largest λ in the sequence such that the loss is within a median range of the minimum, thereby ensuring a more stable selection. These algorithms are implemented in the Python library `sdelearn` (see Iafrate (2024)).

6.2 Stochastic regression model

In this subsection we test our findings by applying our method to synthetic data. We evaluate the model selection and predictive capabilities of our Adaptive Elastic-Net method, and we compare it with LASSO and non-regularized (quasi-likelihood estimator). Adaptive estimators are based on the quasi-likelihood initial estimator, as in Remark 3.1.

This next study shows the model selection capability of Elastic-Net in the presence of correlation between the coordinates of a multivariate diffusion process. Consider a stochastic regression model, i.e. a $(d+1)$-dimensional process $(Y, X_1, X_2, \ldots, X_d)$ that satisfies a

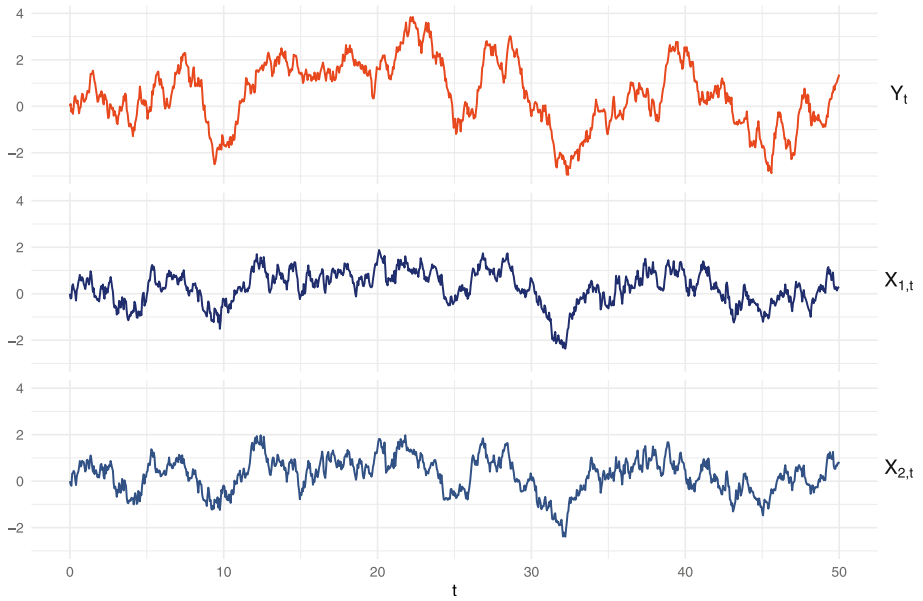


Fig. 2 A sample path of the process (Y, X_1, X_2) that satisfies a SDE with linear drift (29) and diffusion matrix (30)

SDE (3) with linear drift

$$b(y, x_1, , \ldots, x_d) = \begin{pmatrix} \sum_{j=1}^{d} \alpha_j x_j - \alpha_{d+1} y \\ \alpha_{01} - \alpha_{11} x_1 \\ \vdots \\ \alpha_{0d} - \alpha_{1d} x_d \end{pmatrix} \tag{29}$$

and diffusion matrix

$$A = \begin{bmatrix} \beta_0 & 0 & \cdots & 0 \\ 0 & \beta_{11} & \cdots & \beta_{1d} \\ 0 & \vdots & & \vdots \\ 0 & \beta_{d1} & \cdots & \beta_{dd} \end{bmatrix}. \tag{30}$$

Matrix A controls the correlation between the regressors. We discuss separately the case $d = 2$ from the setting $d > 2$.

Case $d = 2$. The parameters used in the simulation are as follows: $(\alpha_1, \alpha_2, \alpha_3) = (1, 1, 1)$, $(\alpha_{01}, \alpha_{02}, \alpha_{11}, \alpha_{22}) = (0, 0, 1, 1)$, $\beta_0 = 1$ and

$$\begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} = \begin{bmatrix} 0.8487 & 0.5316 \\ 0.5316 & 0.8487 \end{bmatrix} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}^{\frac{1}{2}}.$$

The β_{ij} values are chosen so that the noise correlation between X_1 and X_2 is $\rho = 0.9$. A sample path of (Y, X_1, X_2) is shown in Figure 2.

Table 1 Stochastic Regression Simulation Results

n	Δ_n	T	**E-Net Accuracy**	**E-Net Selection**	**LASSO Accuracy**	**LASSO Selection**
250	0.1	25	0.90	0.73	0.88	0.27
1000	0.05	50	0.93	0.89	0.92	0.6
10 000	0.01	100	0.97	0.98	0.97	0.87

We assess the performance of the model using two criteria. *Accuracy* is the proportion of regression coefficients that are correctly either included or excluded, while *selection* is the probability that the fitted model includes every predictor whose true coefficient is non-zero. In $d = 2$, both the true coefficients, α_1 for X_1 and α_2 for X_2, are assumed to be non-zero.

The simulation setting is as follows. We repeated $N = 10^3$ simulations from the process (Y, X_1, X_2) with linear drift (29) and diffusion matrix (30), with increasing sample sizes equal to $n = 250$, $n = 1000$ and $n = 10\,000$, according to a high-frequency sampling scheme in order to approach the asymptotic regime. The mixing parameter γ is set equal to 0.5. The optimal lambda value, λ_{opt}, is chosen based on an AIC criterion (where the classical log-likelihood is replaced with (14)). Results are rounded up to the second decimal point. In this study the diffusion parameters are fixed, we focus on the drift estimation in order to isolate the impact of correlation on variable selection, in analogy with linear regression.

Results are shown in Table 1. We see that Elastic-Net regularization is capable of correctly dealing with the coefficients of the correlated variables, by including both of them, while simultaneously selecting the correct model. The LASSO instead tends to select only one of the two variables, with a lower percentage of model identification. This effect is most apparent for a moderate sample size.

Let $\hat{\alpha}_n^{(k)}$ denote the estimate obtained at simulation k for the drift parameter whose true values are denoted by α_0. We evaluate the performance of the Elastic-Net estimator LASSO estimator by computing the empirical mean square errors:

$$\widehat{\mathsf{MSE}}_j = \frac{1}{N}\sum_{k=1}^{N}(\hat{\alpha}_{n,j}^{(k)} - \alpha_{0,j})^2 \qquad \text{for } j = 1, \ldots, \mathsf{p}\,. \tag{31}$$

The results obtained with $N = 10^3$ simulations, as presented in Table 2, demonstrate that the estimates generated by Elastic-Net method exhibit superior performance in terms of empirical mean square error compared to those produced by the LASSO across all three scenarios. It is noteworthy that the estimates derived using the Elastic-Net exhibit a lower mean square error for the parameters α_1 and α_2, which, as shown in (29), are associated with the correlated variables, particularly in the case of small sample sizes.

To provide a clear view of the differences in the estimates obtained with Elastic-Net and LASSO, Figure 3 shows the empirical distributions of the parameters α_1 and α_2.

In adherence to the theoretical framework, which asserts that LASSO tends to select only one of the correlated variables, unlike Elastic-Net, which does not shrink the parameters associated to correlated variables to zero, and in alignment with the findings depicted in Table 1 and Table 2, the plots show that Elastic-Net exhibits lower percentages of parameter shrinkage across all three scenarios. This phenomenon is especially discernible in scenarios involving smaller sample sizes.

Table 2 provides a summary of the results obtained from the simulation study.

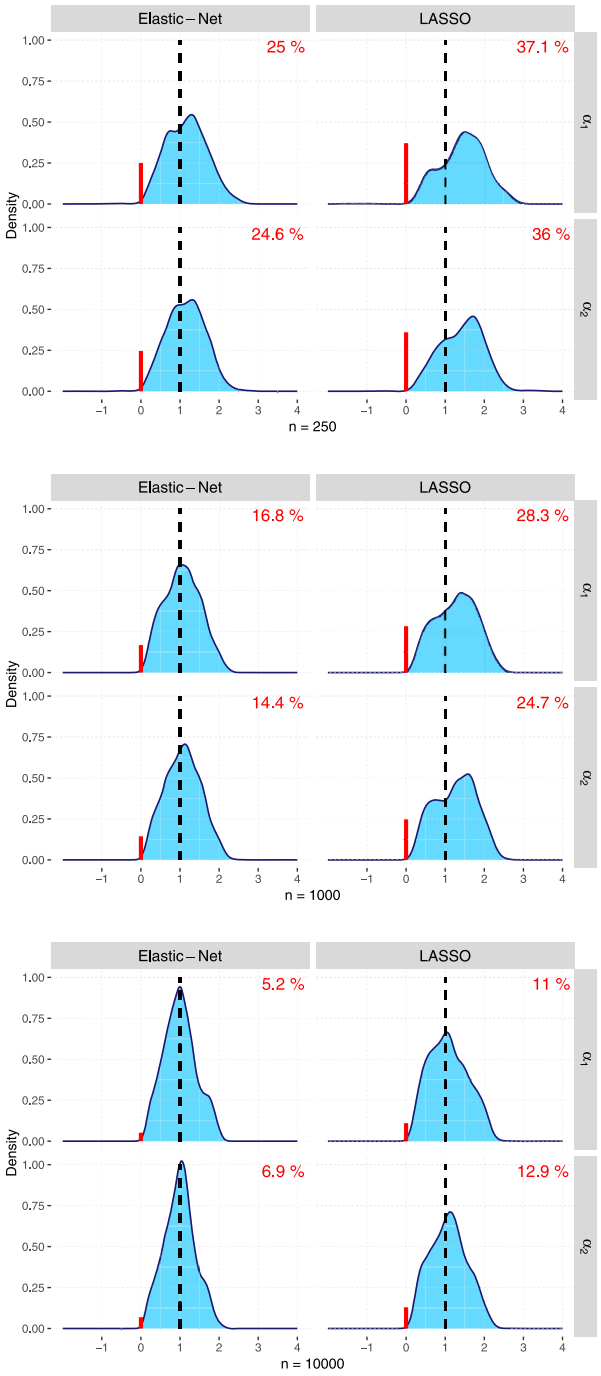


Fig. 3 Distribution of parameters α_1 and α_2, with increasing sample size. For each parameter-method combination, the plot shows in red the proportion of parameters estimated as zero by the method under consideration, while in blue the density of parameters estimated as non-zero. The true parameter values are shown with black dashed lines. The integration of the continuous and discrete components in each plot is normalized to 1

Table 2 Summary of estimated parameters across different sample sizes n.

(a) $n = 250$

Par.	True	Elastic-Net					LASSO				
	Value	$\widehat{MSE}$	$Avg.$ $(Std.Err.)$	$q_{0.25}$	Me	$q_{0.75}$	$\widehat{MSE}$	$Avg.$ $(St.Err.)$	$q_{0.25}$	Me	$q_{0.75}$
α_1	1	0.469	0.876 (0.674)	0	0.911	1.392	0.705	0.909 (0.835)	0	0.943	1.640
α_2	1	0.436	0.867 (0.647)	0.167	0.922	1.369	0.664	0.909 (0.810)	0	0.945	1.639
α_3	1	0.052	0.910 (0.210)	0.765	0.893	1.039	0.060	0.926 (0.234)	0.759	0.907	1.069
α_{01}	0	0.011	0.004 (0.103)	0	0	0	0.007	0.001 (0.086)	0	0	0
α_{02}	0	0.011	0.002 (0.104)	0	0	0	0.007	0 (0.085)	0	0	0
α_{11}	1	0.060	0.929 (0.235)	0.768	0.906	1.070	0.073	0.912 (0.256)	0.731	0.889	1.070
α_{12}	1	0.058	0.928 (0.230)	0.768	0.913	1.071	0.070	0.911 (0.249)	0.741	0.884	1.068

(b) $n = 1000$

Par.	True	Elastic-Net					LASSO				
	Value	$\widehat{MSE}$	$Avg.$ $(Std.Err.)$	$q_{0.25}$	Me	$q_{0.75}$	$\widehat{MSE}$	$Avg.$ $(St.Err.)$	$q_{0.25}$	Me	$q_{0.75}$
α_1	1	0.346	0.885 (0.577)	0.432	0.939	1.314	0.531	0.911 (0.724)	0	0.963	1.528
α_2	1	0.323	0.919 (0.563)	0.525	0.974	1.334	0.502	0.947 (0.707)	0.251	1.003	1.550
α_3	1	0.028	0.929 (0.153)	0.820	0.916	1.022	0.031	0.947 (0.167)	0.831	0.935	1.048
α_{01}	0	0.001	0 (0.038)	0	0	0	0.001	0.001 (0.037)	0	0	0
α_{02}	0	0.001	0 (0.036)	0	0	0	0.001	0 (0.035)	0	0	0

Table 2 continued

Par.	True	(b) $n = 1000$ Elastic-Net					LASSO				
	Value	$\widehat{MSE}$	$Avg.$ $(Std.Err.)$	$q_{0.25}$	Me	$q_{0.75}$	$\widehat{MSE}$	$Avg.$ $(St.Err.)$	$q_{0.25}$	Me	$q_{0.75}$
α_{11}	1	0.027	0.938 (0.153)	0.834	0.926	1.038	0.031	0.942 (0.166)	0.825	0.928	1.050
α_{12}	1	0.028	0.938 (0.155)	0.832	0.925	1.029	0.031	0.942 (0.168)	0.825	0.927	1.044
α_1	1	0.214	0.947 (0.459)	0.643	0.958	1.244	0.331	0.967 (0.575)	0.539	0.980	1.400
α_2	1	0.212	0.922 (0.454)	0.634	0.964	1.207	0.324	0.945 (0.567)	0.524	0.994	1.341
α_3	1	0.012	0.950 (0.098)	0.884	0.940	1.013	0.012	0.970 (0.104)	0.898	0.961	1.036
α_{01}	0	$< 10^{-3}$	0 (0.020)	0	0	0	$< 10^{-3}$	0 (0.021)	0	0	0
α_{02}	0	$< 10^{-3}$	0.001 (0.022)	0	0	0	0.001	0.001 (0.023)	0	0	0
α_{11}	1	0.013	0.959 (0.107)	0.885	0.956	1.027	0.013	0.974 (0.113)	0.895	0.970	1.045
α_{12}	1	0.012	0.962 (0.104)	0.889	0.963	1.030	0.012	0.977 (0.109)	0.901	0.973	1.048

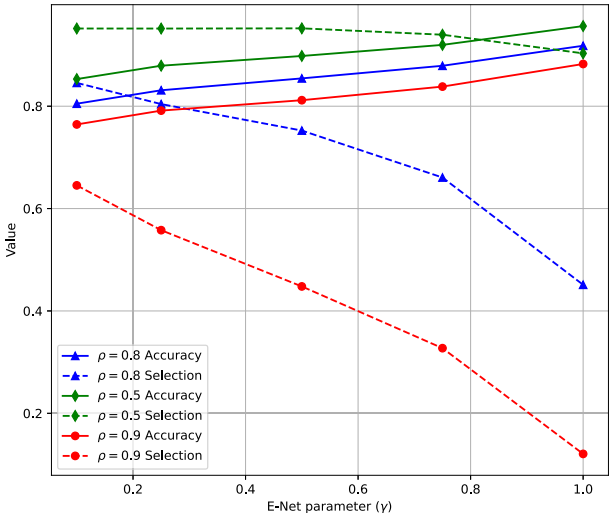


Fig. 4 Accuracy and selection for different values of the Elastic net parameter and correlation parameter

Case $d > 2$. In this scenario, we consider the case where some of the coefficients of processes $X_1, \ldots, X_d$ are zero. We assume the following correlation structure: $\Sigma = (\Sigma_{ij})_{i,j} = (\rho^{|i-j|})_{i,j}$ for $\rho = 0.5, 0.8, 0.9$. The drift parameters used in the simulation are

$$(\alpha_1, \ldots, \alpha_{\lfloor d/2 \rfloor}, \alpha_{\lfloor d/2 \rfloor + 1}, \ldots, \alpha_d, \alpha_{d+1}) = (1/d, \ldots, 1/d, 0, \ldots, 0, 2),$$

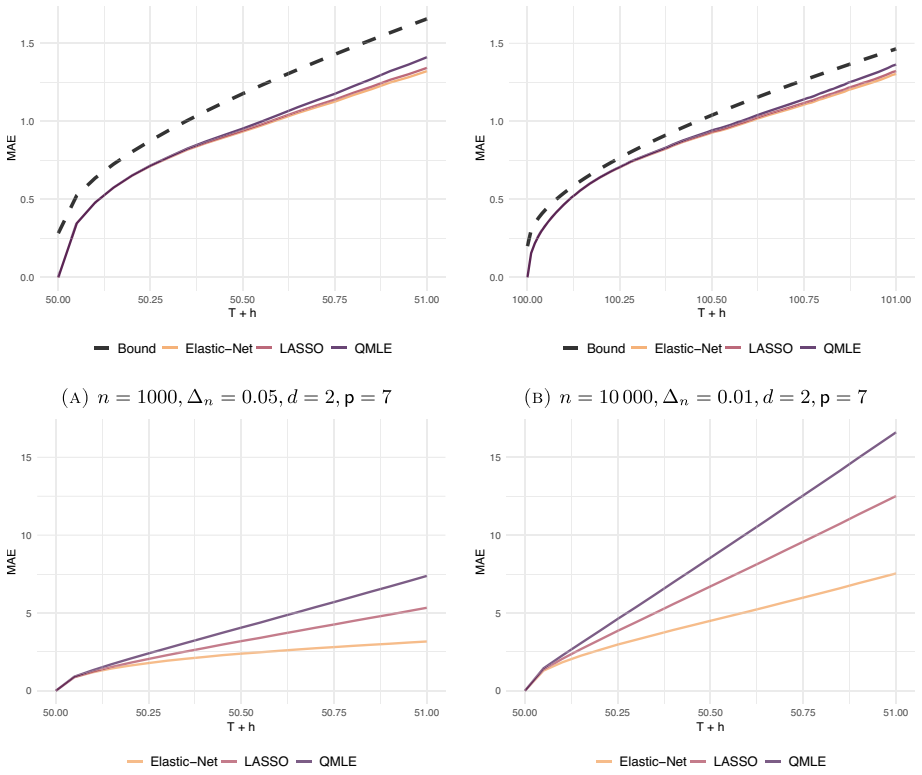
with $\alpha_{0,j} = 1, \alpha_{1,j} = 1$, $j \in 1, \ldots, d$. Results are shown in figure Figure 4.

Effect of the Elastic-Net parameter. From Figure 4 we see how the Elastic-Net parameter γ controls the trade-off in variable selection. For values of γ closer to 1, i.e. the LASSO case, the probability of correctly selecting the model decreases drastically due to the presence of correlated processes that the model cannot properly distinguish, especially for larger values of ρ. Conversely for smaller values of the Elastic-Net parameter γ this effect can be controlled. The cost is a smaller "accuracy", due to a higher number of variables falsely included in the model.

Forecasting performance. We now turn our attention to assessing the predictive efficacy of the Elastic-Net estimator. To this end, we conduct $N = 10^4$ simulations from the model specified in (29) and (30), considering the scenarios outlined in Table 1. We observe the data up to time $T_n = T$, while our goal is to predict future values up to time $T + 1$, with a step size of $\Delta_n \in \{0.05, 0.01\}$. For each simulation k, we estimate $X^{(k)}_{T+h}$ at time $T + h$, denoted by $\hat{X}^{(k)}_{T+h}$, is performed using the one-step-ahead prediction as described in (22), with h varying from 0 to 1, i.e. $h \in \{0, \Delta_n, 2\Delta_n, \ldots, 1\}$. We can calculate the empirical mean absolute prediction error for the Elastic-Net and LASSO estimators as follows:

$$\widehat{\mathsf{MAE}}_k = \frac{1}{N} \sum_{k=1}^{N} |X^{(k)}_{T+h} - \hat{X}^{(k)}_{T+h}| \qquad \text{with } h \in \{0, \Delta_n, 2\Delta_n, \ldots, 1\}. \tag{32}$$

Similarly, the non-asymptotic prediction error bound, at each time point $T + h$, can be computed in a straightforward manner up to constant factors, as described in (24). This allows for comparison with the empirical mean absolute prediction error calculated for the Elastic-Net, LASSO and QMLE estimators. The results are shown in the graphs in Figure



(A) $n = 1000, \Delta_n = 0.05, d = 2, \mathsf{p} = 7$

(B) $n = 10\,000, \Delta_n = 0.01, d = 2, \mathsf{p} = 7$

(C) $n = 1000, \Delta_n = 0.05, d = 16, \mathsf{p} = 49$

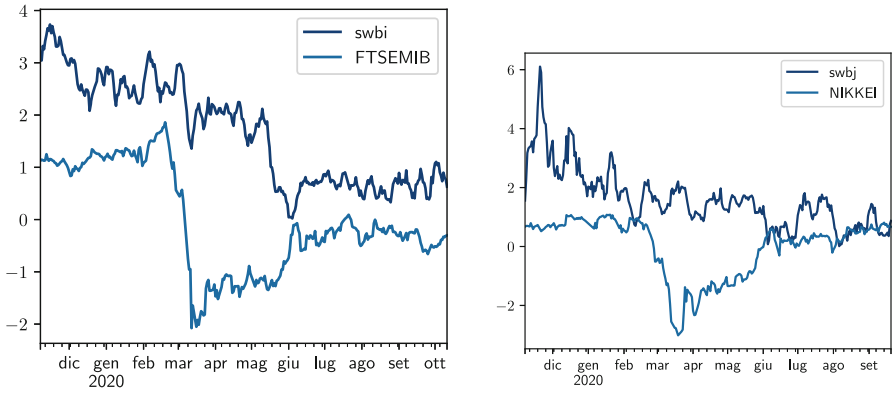
(D) $n = 1000, \Delta_n = 0.05, d = 33, \mathsf{p} = 100$

Fig. 5 Comparison of empirical mean absolute prediction errors for Elastic-Net, LASSO and QMLE estimators with respect of variation in the prediction horizon h, sample size n and the number of parameters. The graphs in the same row share the same scale on the y-axis. In Figure 5a and Figure 5b, the non-asymptotic prediction error bound, calculated according to (24), up to constant factors, is also represented by dashed line

5. In the first two graphs, Figure 5a and Figure 5b, the number of parameters is the same, $\mathsf{p} = 7$, while the sample size n and the step size Δ_n vary between the two figures. The approximate prediction error bound (dashed line), calculated according to (24), thus up to constant factors, correctly predicts the shape of the empirical mean absolute prediction error (32) for the Elastic-Net Estimator. It can be noted that for very large sample sizes the error curves align, consistently with the expected asymptotic behaviour of each estimator. Figures 5c and 5d, show the result for the same sample size n, and increasing number of parameters p. It can be observed that the gap in predictive performance widens, with the Elastic-Net estimator being more stable as p and the prediction horizon h increase. Our results show that the introduction of an Elastic-Net tunable penalty can considerably improve our ability to forecast a correlated dynamical system, while the LASSO penalty proves to be too rigid.

7 Real data application: well-being data analysis

In this section, we use our technique to discover association dependency patterns in real-world scenarios. Motivated by Carpi et al. (2022), we apply our methodology to the analysis



(A) Italy (B) Japan

Fig. 6 Response variable (subjective well being), and one predictor (country's economic index)

of subjective well-being. In that study, the authors examine the effect of the COVID-19 pandemic on subjective well-being (SWB), assessed through Twitter data from Japan and Italy. The study analyzes various data sources, including climate and air quality, COVID-19 cases and deaths, survey data on symptoms, Google search trends, policy measures, mobility patterns, economic indicators, and proxies for health and stress. We refer to the original study for a complete description of the dataset. In Figure 6 we show the response variable and one of the predictors included in the model.

We split the data in shorter 1-month or 3-months time windows, and we carry out an Elastic-Net analysis on each time frame. Such a framework, called Dynamic Elastic Net, allows to mitigate the effect of potential non-stationarity in the data (see Carpi et al. (2022)). The goal is to identify which variables contribute to the SWB and to analyze how the association patterns change over time. We finally compare the variables in the Importance-Frequency space in order to assess their overall relative importance.

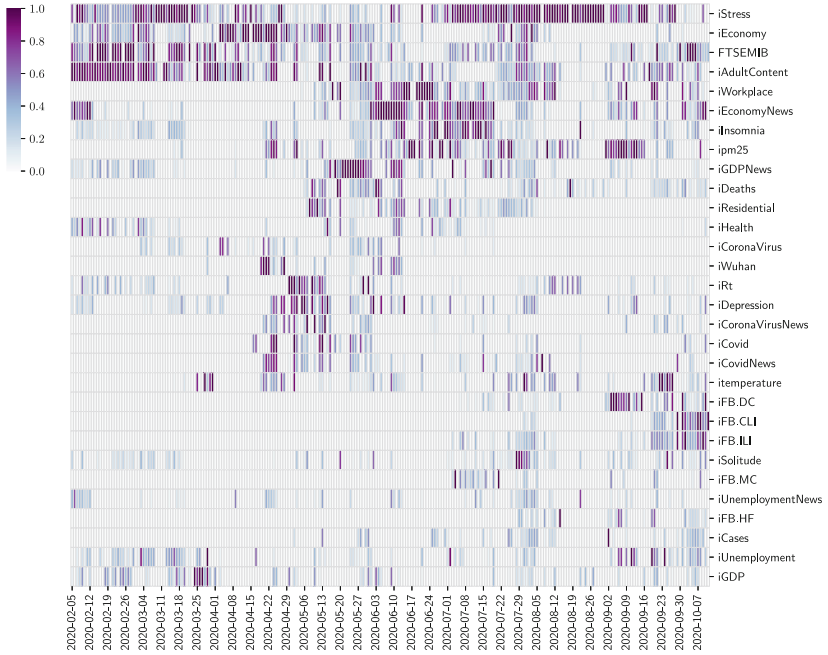
We treat the SWB variable as a response Y in a stochastic regression model and the other variables as predictors $X_1, \ldots, X_p$. We then consider a regression-type SDE of the form

$$\mathrm{d}Y_t = \left(\mu_Y - \alpha_{Y,0} Y_t - \sum_{j=1}^{p} \alpha_{Y,j} X_{j,t}\right) \mathrm{d}t + \sigma_Y \sqrt{Y_t} \mathrm{d}W_{1,t} \tag{33}$$

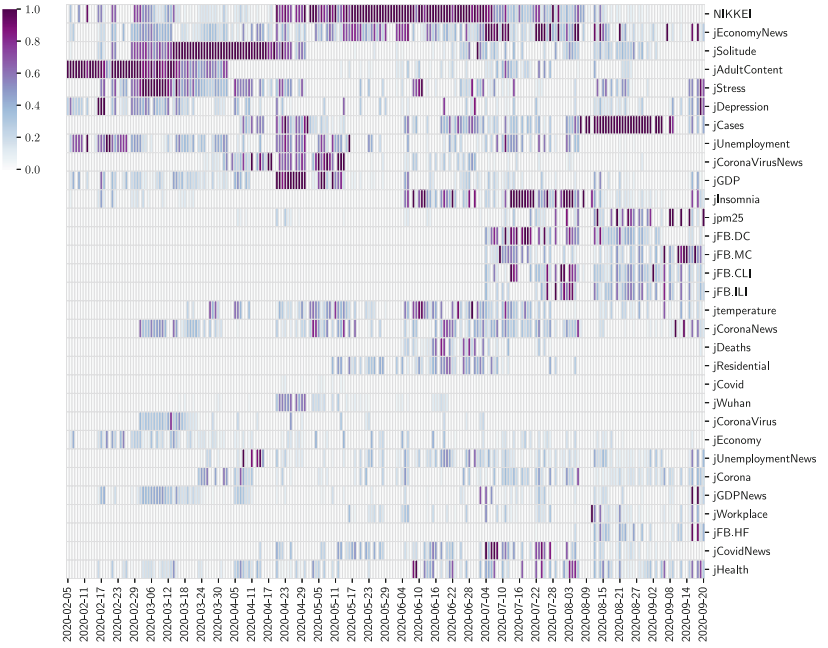
$$\mathrm{d}X_{j,t} = \left(\mu_j - \alpha_{X,j} X_{j,t}\right) \mathrm{d}t + \sigma_j \mathrm{d}W_{j+1,t} \tag{34}$$

where $W_1, \ldots, W_{j+1}$ are independent Brownian motions and X_j are Vasicek SDE. We use Elastic-Net estimation in order to select the true non-zero coefficients in the model. More in detail, we compute an initial quasi-likelihood estimate for all the parameters in the model, and then we apply our adaptive estimator by restricting our attention to the parameters $\alpha_{Y,j}$, $j = 1, \ldots p$. We repeat this process for subsequent 30 and 90 days windows, and we normalize the absolute value of the coefficients in order to have comparable results over different sets of predictors. We run our model on both Italy and Japan data. The results for a 90 days time frame are shown in Figure 7.

Finally, we compare the results in the Importance-Frequency space. The frequency is given by the number of times that a variable is selected over the different time windows. The importance component is computed by first ranking the selected variables in each time



(A) Italy



(B) Japan

Fig. 7 Normalized coefficients $\alpha_{Y,j}$, 90 days time window

window, and then averaging the results (normalized in 0-100). The results are shown in Figure 8. Variables in the upper right corner have a stronger effect on the response, in the sense that they are selected a larger number of times and, whenever they are selected, they tend to have the largest effect. Vice versa variables placed in the bottom left corner are comparatively less important.

We note how in Italy, Figure 8 (A), in the shorter time frame (31 days), variables related to mobility restrictions (residential) or to the pandemic spread (deaths, Rt) tend to be more important than in the long time. In the long run (90 days) psychological stress and economy have a larger impact. When comparing Italy and Japan, Figure 8 (B), we note that in Japan the number of Covid cases (jCases) have a larger effect, while in Italy deaths (iDeaths) are more impactful. Economy variables are generally more relevant in Japan.

8 Proofs

Proof of Theorem 1 In the proof we use the same approach adopted in De Gregorio and Iafrate (2021). In the present setting it is necessary to handle the Ridge penalty.

For the sake of simplicity, some of the dependency on n, indicated as a subscript, will not be taken into account in the following steps (i.e. $\hat{\theta} = \hat{\theta}_n$, $\tilde{\theta} = \tilde{\theta}_n$, $A = A_n$, $\hat{G} = \hat{G}_n$, $\hat{D} = \hat{D}_n$, $\kappa_j = \kappa_{n,j}$, $\pi_j = \pi_{n,j}$, $\lambda_2 = \lambda_{2,n}$, $\gamma_2 = \gamma_{2,n}$).

(i) Let us start with the consistency. We can write down

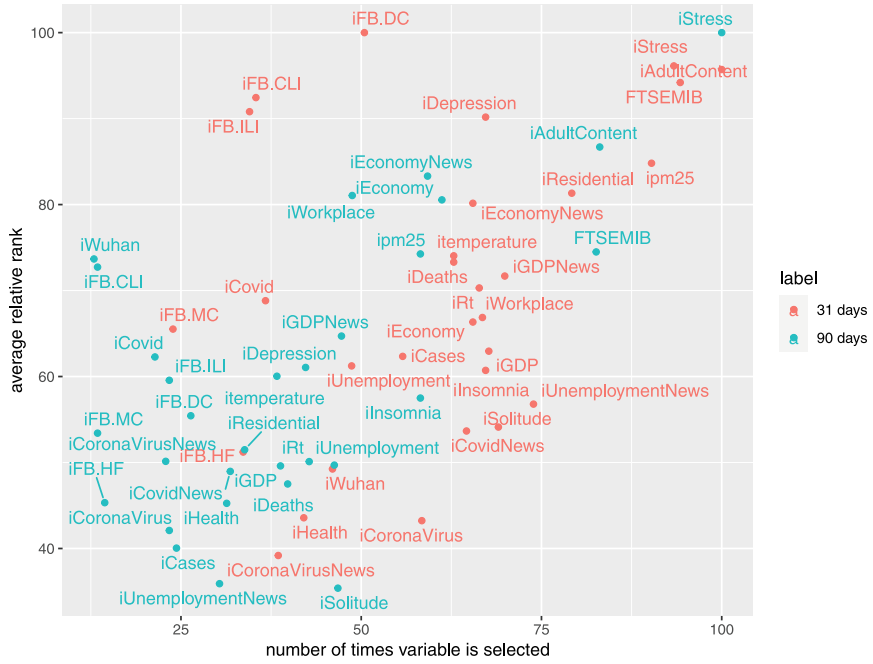
$$
\begin{aligned}
0 \geq & \mathcal{F}_n(\hat{\theta}, \tilde{\theta}) - \mathcal{F}_n(\theta_0, \tilde{\theta}) \\
= & \langle \hat{G}, (\hat{\theta} - \theta_0)^{\otimes 2} \rangle + 2(\hat{\theta} - \theta_0)^{\intercal} \hat{G}(\theta_0 - \tilde{\theta}) \\
& + |\hat{\alpha}|_{1,\kappa} - |\alpha_0|_{1,\kappa} + |\hat{\beta}|_{1,\pi} - |\beta_0|_{1,\pi} + \lambda_2(|\hat{\alpha}|^2 - |\alpha_0|^2) + \gamma_2(|\hat{\beta}|^2 - |\beta_0|^2) \\
\geq & \langle \hat{G}, (\hat{\theta} - \theta_0)^{\otimes 2} \rangle + 2(\hat{\theta} - \theta_0)^{\intercal} \hat{G}(\theta_0 - \tilde{\theta}) \\
& + \sum_{j=1}^{\mathsf{p}^0} \kappa_j (|\hat{\alpha}_j| - |\alpha_{0,j}|) + \sum_{h=1}^{\mathsf{q}^0} \pi_h (|\hat{\beta}_h| - |\beta_{0,h}|) \\
& + \lambda_2 \sum_{j=1}^{\mathsf{p}^0} (|\hat{\alpha}_j|^2 - |\alpha_{0,j}|^2) + \gamma_2 \sum_{h=1}^{\mathsf{q}^0} (|\hat{\beta}_h|^2 - |\beta_{0,h}|^2).
\end{aligned}
$$

From the well-known inequality $||x| - |y|| \leq |x - y|$, immediately it follows

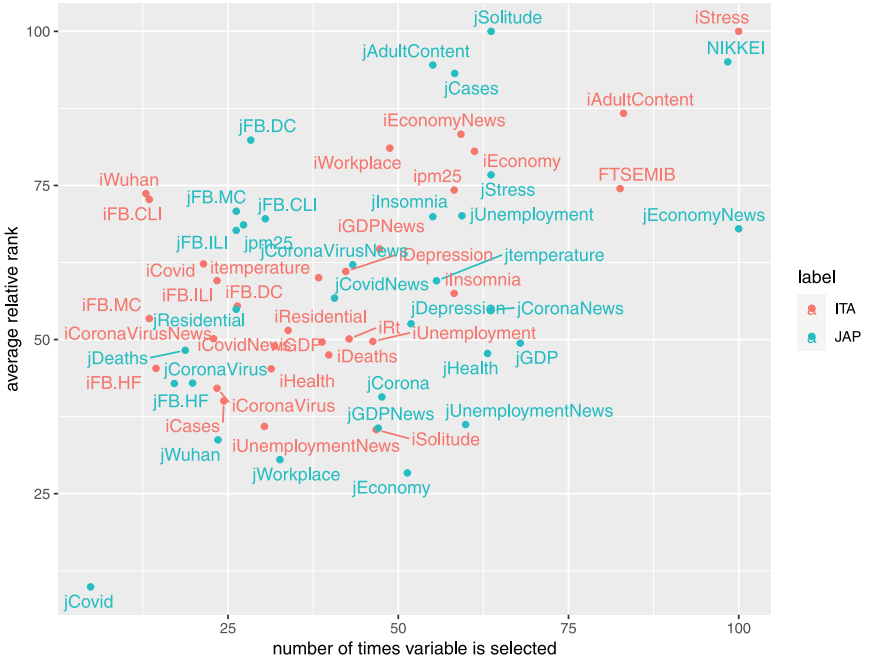
$$
\begin{aligned}
\sum_{j=1}^{\mathsf{p}^0} \kappa_j (|\hat{\alpha}_j| - |\alpha_{0,j}|) &\geq -\mathsf{p}^0 a_n |\hat{\alpha} - \alpha_0|, \\
\sum_{h=1}^{\mathsf{q}^0} \pi_h (|\hat{\beta}_h| - |\beta_{0,h}|) &\geq -\mathsf{q}^0 c_n |\hat{\beta} - \beta_0|.
\end{aligned}
$$

Given a convex function f, the following inequalities hold:

$$
f(y) - f(x) \geq \nabla f(x)^{\intercal}(y - x) \Rightarrow |y|^2 - |x|^2 \geq 2x^{\intercal}(y - x), \quad x, y \in \mathbb{R}^d.
$$



(A) Italy data, comparison between 31 and 90 days period



(B) 90 days time frame, comparison between Italy and Japan results

Fig. 8 Variables representation in the frequency-importance space

By applying this last inequality in the penalty part related to the Ridge component and the Cauchy–Schwarz inequality $|x^\intercal y| \leq |x||y|$, we have:

$$\sum_{j=1}^{\mathsf{p}^0}(|\hat{\alpha}_j|^2 - |\alpha_{0,j}|^2) \geq -2\mathsf{p}^0|\alpha_0||\hat{\alpha} - \alpha_0|,$$

$$\sum_{h=1}^{\mathsf{q}^0}(|\hat{\beta}_h|^2 - |\beta_{0,h}|^2) \geq -2\mathsf{q}^0|\beta_0||\hat{\beta} - \beta_0|.$$

Therefore, we get:

$$\begin{aligned}
0 \geq & \langle \hat{G}, (\hat{\theta} - \theta_0)^{\otimes 2}\rangle + 2(\hat{\theta} - \theta_0)^\intercal \hat{G}(\theta_0 - \tilde{\theta}) \\
& - \mathsf{p}^0 \frac{a_n}{\sqrt{n\Delta_n}}|\sqrt{n\Delta_n}(\hat{\alpha} - \alpha_0)| - \mathsf{q}^0 \frac{c_n}{\sqrt{n}}|\sqrt{n}(\hat{\beta} - \beta_0)| \\
& - 2\frac{\lambda_2}{\sqrt{n\Delta_n}}\mathsf{p}^0|\alpha_0||\sqrt{n\Delta_n}(\hat{\alpha} - \alpha_0)| - 2\frac{\gamma_2}{\sqrt{n}}\mathsf{q}^0|\beta_0||\sqrt{n}(\hat{\beta} - \beta_0)| \\
\geq & \langle \hat{D}, (A^{-1}(\hat{\theta} - \theta_0))^{\otimes 2}\rangle + 2[A^{-1}(\hat{\theta} - \theta_0)]^\intercal \hat{D}[A^{-1}(\theta_0 - \tilde{\theta})] \\
& - \left(\mathsf{p}^0 \frac{a_n}{\sqrt{n\Delta_n}} + \mathsf{q}^0 \frac{c_n}{\sqrt{n}} + 2\mathsf{p}^0 \frac{\lambda_2}{\sqrt{n\Delta_n}}|\alpha_0| + 2\mathsf{q}^0 \frac{\gamma_2}{\sqrt{n}}|\beta_0|\right)|A^{-1}(\hat{\theta} - \theta_0)|
\end{aligned}$$

and by using the same arguments in the proof of Theorem 1 in De Gregorio and Iafrate (2021), one has;

$$\begin{aligned}
|A^{-1}(\hat{\theta} - \theta_0)| \leq & \|\hat{D}^{-1}\| \left(2\|\hat{D}\||A^{-1}(\tilde{\theta} - \theta_0)| + \mathsf{p}^0 \frac{a_n}{\sqrt{n\Delta_n}} + \mathsf{q}^0 \frac{c_n}{\sqrt{n}}\right. \\
& \left. + 2\mathsf{p}^0 \frac{\lambda_2}{\sqrt{n\Delta_n}}|\alpha_0| + 2\mathsf{q}^0 \frac{\gamma_2}{\sqrt{n}}|\beta_0|\right). \qquad (35)
\end{aligned}$$

Given that the right-hand side of the inequality is $O_p(1)$ according to the statement, we obtain $A^{-1}\left(\hat{\theta} - \theta_0\right) = O_p(1)$.

(ii) In order to prove the selection consistency, we observe that $\hat{G} = \begin{pmatrix} \hat{G}_\alpha \\ \hat{G}_\beta \end{pmatrix}$ where $\hat{G}_\alpha$ and $\hat{G}_\beta$ are random matrices of dimension $\mathsf{p} \times \mathsf{m}$ and $\mathsf{q} \times \mathsf{m}$, respectively.

Let us assume $\hat{\alpha} \notin \partial\Theta_\alpha$ and $\hat{\alpha}_j \neq 0$ for some $j = \mathsf{p}^0 + 1, \ldots, \mathsf{p}$. From the Karush–Kuhn–Tucker (KKT) condition:

$$\frac{1}{\sqrt{n\Delta_n}} \frac{\partial}{\partial \alpha_j} \mathcal{F}(\theta, \tilde{\theta})\Big|_{\theta=\hat{\theta}} = 0 \qquad (36)$$

we derive

$$\begin{aligned}
\left|\frac{2}{\sqrt{n\Delta_n}} \hat{G}_\alpha(j) A A^{-1}(\hat{\theta} - \tilde{\theta})\right| &= \left|\frac{\kappa_j}{\sqrt{n\Delta_n}} \mathrm{sgn}(\hat{\alpha}_j) + 2\frac{\lambda_2}{\sqrt{n\Delta_n}}\hat{\alpha}_j\right| \\
&\geq \left|\frac{b_n}{\sqrt{n\Delta_n}}\right| - \frac{\lambda_2}{\sqrt{n\Delta_n}}|\hat{\alpha}_j|
\end{aligned}$$

for some $j = \mathsf{p}^0 + 1, \ldots, \mathsf{p}$, with $\hat{G}_\alpha(j)$ the j-th row of $\hat{G}_\alpha$.

By consistency and the assumptions, we have that $\left|\frac{2}{\sqrt{n\Delta_n}}\hat{G}_\alpha(j)AA^{-1}(\hat{\theta}-\tilde{\theta})\right| = O_p(1)$ and $\frac{\lambda_2}{\sqrt{n\Delta_n}}|\hat{\alpha}_j| = O_p(1)$, while $\frac{b_n}{\sqrt{n\Delta_n}} \xrightarrow{p} \infty$. Therefore, for any $j = \mathsf{p}^0+1,\ldots,\mathsf{p}$:

$$P\left(\hat{\alpha}_j \neq 0, \hat{\alpha} \notin \partial\Theta_\alpha\right) \leq P\left(\left|2\frac{1}{\sqrt{n\Delta_n}}\hat{G}_\alpha(j)A_nA_n^{-1}(\hat{\theta}-\tilde{\theta})\right| + \frac{\lambda_2}{\sqrt{n\Delta_n}}|\hat{\alpha}_j| \geq \frac{b_n}{\sqrt{n\Delta_n}}\right) \longrightarrow 0 \tag{37}$$

as $n \longrightarrow \infty$. By means of the same arguments in Suzuki and Yoshida (2020), we get $P(\hat{\alpha} \in \partial\Theta_\alpha) \longrightarrow 0$. Therefore

$$P\left(\hat{\alpha}_\bullet \neq 0\right) \leq P(\hat{\alpha} \in \partial\Theta_\alpha) + \sum_{j=\mathsf{p}^0+1}^{\mathsf{p}} P\left(\hat{\alpha}_j \neq 0, \hat{\alpha} \notin \partial\Theta_\alpha\right) \longrightarrow 0.$$

Similar steps lead to $P(\hat{\beta}_\bullet \neq 0) \longrightarrow 0$.

(iii) It is possible to adapt to our case the proof of Theorem 3 in De Gregorio and Iafrate (2021). We have two groups of parameters and two more terms arising from Ridge regularization. It is useful to rewrite $\hat{G}$ as a block matrix partitioned as follows:

$$\hat{G} = \begin{pmatrix} \hat{G}^{\alpha\alpha} & \hat{G}^{\alpha\beta} \\ \hat{G}^{\beta\alpha} & \hat{G}^{\beta\beta} \end{pmatrix}$$

where

$$\hat{G}^{\alpha\alpha} := \begin{pmatrix} \hat{G}^{\alpha\alpha}_{\star\star} & \hat{G}^{\alpha\alpha}_{\star\bullet} \\ \hat{G}^{\alpha\alpha}_{\bullet\star} & \hat{G}^{\alpha\alpha}_{\bullet\bullet} \end{pmatrix},\ \hat{G}^{\alpha\beta} := \begin{pmatrix} \hat{G}^{\alpha\beta}_{\star\star} & \hat{G}^{\alpha\beta}_{\star\bullet} \\ \hat{G}^{\alpha\beta}_{\bullet\star} & \hat{G}^{\alpha\beta}_{\bullet\bullet} \end{pmatrix},\ \hat{G}^{\beta\alpha} := \begin{pmatrix} \hat{G}^{\beta\alpha}_{\star\star} & \hat{G}^{\beta\alpha}_{\star\bullet} \\ \hat{G}^{\beta\alpha}_{\bullet\star} & \hat{G}^{\beta\alpha}_{\bullet\bullet} \end{pmatrix},\ \hat{G}^{\beta\beta} := \begin{pmatrix} \hat{G}^{\beta\beta}_{\star\star} & \hat{G}^{\beta\beta}_{\star\bullet} \\ \hat{G}^{\beta\beta}_{\bullet\star} & \hat{G}^{\beta\beta}_{\bullet\bullet} \end{pmatrix}.$$

We can rewrite the objective function:

$$\begin{aligned}
\mathcal{F}_n(\theta) &= (\theta-\tilde{\theta})^\intercal \hat{G}(\theta-\tilde{\theta}) + \sum_{j=1}^{\mathsf{p}}\kappa_j|\alpha_j| + \sum_{h=1}^{\mathsf{q}}\pi_h|\beta_h| + \lambda_n\sum_{j=1}^{\mathsf{p}}|\alpha_j|^2 + \gamma_n\sum_{h=1}^{\mathsf{q}}|\beta_h|^2 \\
&= (\alpha-\tilde{\alpha})_\star^\intercal \hat{G}^{\alpha\alpha}_{\star\star}(\alpha-\tilde{\alpha})_\star + (\beta-\tilde{\beta})_\star^\intercal \hat{G}^{\beta\beta}_{\star\star}(\beta-\tilde{\beta})_\star \\
&\quad + (\alpha-\tilde{\alpha})_\bullet^\intercal \hat{G}^{\alpha\alpha}_{\bullet\bullet}(\alpha-\tilde{\alpha})_\bullet + (\beta-\tilde{\beta})_\bullet^\intercal \hat{G}^{\beta\beta}_{\bullet\bullet}(\beta-\tilde{\beta})_\bullet \\
&\quad + 2\left[(\alpha-\tilde{\alpha})_\star^\intercal \hat{G}^{\alpha\alpha}_{\star\bullet}(\alpha-\tilde{\alpha})_\bullet + (\beta-\tilde{\beta})_\star^\intercal \hat{G}^{\beta\beta}_{\star\bullet}(\beta-\tilde{\beta})_\bullet\right] \\
&\quad + 2\Big[(\alpha-\tilde{\alpha})_\star^\intercal \hat{G}^{\alpha\beta}_{\star\star}(\beta-\tilde{\beta})_\star + (\alpha-\tilde{\alpha})_\star^\intercal \hat{G}^{\alpha\beta}_{\star\bullet}(\beta-\tilde{\beta})_\bullet \\
&\quad + (\alpha-\tilde{\alpha})_\bullet^\intercal \hat{G}^{\alpha\beta}_{\bullet\star}(\beta-\tilde{\beta})_\star + (\alpha-\tilde{\alpha})_\bullet^\intercal \hat{G}^{\alpha\beta}_{\bullet\bullet}(\beta-\tilde{\beta})_\bullet\Big] \\
&\quad + \sum_{j=1}^{\mathsf{p}}\kappa_j|\alpha_j| + \sum_{h=1}^{\mathsf{q}}\pi_h|\beta_h| + \lambda_n\sum_{j=1}^{\mathsf{p}}|\alpha_j|^2 + \gamma_n\sum_{h=1}^{\mathsf{q}}|\beta_h|^2.
\end{aligned}$$

By setting $\breve{\theta} := (\alpha_\star, 0, \beta_\star, 0)^\intercal \in \mathbb{R}^{\mathsf{m}}$, we have

$$\begin{aligned}
\mathcal{F}_n(\breve{\theta}) &= (\alpha-\tilde{\alpha})_\star^\intercal \hat{G}^{\alpha\alpha}_{\star\star}(\alpha-\tilde{\alpha})_\star + (\beta-\tilde{\beta})_\star^\intercal \hat{G}^{\beta\beta}_{\star\star}(\beta-\tilde{\beta})_\star \\
&\quad + (\tilde{\alpha}_\bullet)^\intercal \hat{G}^{\alpha\alpha}_{\bullet\bullet}\tilde{\alpha}_\bullet + (\tilde{\beta}_\bullet)^\intercal \hat{G}^{\beta\beta}_{\bullet\bullet}\tilde{\beta}_\bullet \\
&\quad - 2\left[(\alpha-\tilde{\alpha})_\star^\intercal \hat{G}^{\alpha\alpha}_{\star\bullet}\tilde{\alpha}_\bullet + (\beta-\tilde{\beta})_\star^\intercal \hat{G}^{\beta\beta}_{\star\bullet}\tilde{\beta}_\bullet\right] \\
&\quad + 2\left[(\alpha-\tilde{\alpha})_\star^\intercal \hat{G}^{\alpha\beta}_{\star\star}(\beta-\tilde{\beta})_\star - (\alpha-\tilde{\alpha})_\star^\intercal \hat{G}^{\alpha\beta}_{\star\bullet}\tilde{\beta}_\bullet - (\tilde{\alpha}_\bullet)^\intercal \hat{G}^{\alpha\beta}_{\bullet\star}(\beta-\tilde{\beta})_\star + (\tilde{\alpha}_\bullet)^\intercal \hat{G}^{\alpha\beta}_{\bullet\bullet}\tilde{\beta}_\bullet\right]
\end{aligned}$$

$$+\sum_{j=1}^{\mathsf{p}^0}\kappa_j|\alpha_j|+\sum_{h=1}^{\mathsf{q}^0}\pi_h|\beta_h|+\lambda_n\sum_{j=1}^{\mathsf{p}^0}|\alpha_j|^2+\gamma_n\sum_{h=1}^{\mathsf{q}^0}|\beta_h|^2.$$

Let us consider

$$B_n^{\alpha}:=\left\{\min_{1\le j\le \mathsf{p}^0}|\hat{\alpha}_j|>0,\ \hat{\alpha}_\bullet=0,\ \det(\hat{G}_{\star\star}^{\alpha\alpha})>0\right\},$$

$$B_n^{\beta}:=\left\{\min_{1\le h\le \mathsf{q}^0}|\hat{\beta}_h|>0,\ \hat{\beta}_\bullet=0,\ \det(\hat{G}_{\star\star}^{\beta\beta})>0\right\}.$$

By Theorem 1 (i)-(ii), $P(B_n^{\alpha}\cap B_n^{\beta})\longrightarrow 1$. If $B_n^{\alpha}\cap B_n^{\beta}$ holds, then $\mathfrak{F}_n(\hat{\theta})=\min_{\breve{\theta}\in\mathbb{R}_0^{\mathsf{m}}}\mathcal{F}_n(\breve{\theta})$, where $\mathbb{R}_0^{\mathsf{m}}:=\{\theta\in\mathbb{R}^{\mathsf{m}}:\alpha_\bullet=0,\beta_\bullet=0\}$.

Then, on B_n^{α}:

$$0=\frac{1}{2}\frac{\partial}{\partial\alpha_\star}\mathcal{F}_n(\theta)\Big|_{\theta=\hat{\theta}_n}=\hat{G}_{\star\star}^{\alpha\alpha}(\hat{\alpha}-\tilde{\alpha})_\star-\hat{G}_{\star\bullet}^{\alpha\alpha}\tilde{\alpha}_\bullet+\hat{G}_{\star\star}^{\alpha\beta}(\hat{\beta}-\tilde{\beta})_\star-\hat{G}_{\star\bullet}^{\alpha\beta}\tilde{\beta}_\bullet+Z(\hat{\alpha})+U(\hat{\alpha})$$

where $Z(\hat{\alpha})=\left(\frac{1}{2}\kappa_1\mathrm{sgn}(\hat{\alpha}_1),\ldots,\frac{1}{2}\kappa_{\mathsf{p}^0}\mathrm{sgn}(\hat{\alpha}_{\mathsf{p}^0})\right)^{\mathsf{T}}$ and $U(\hat{\alpha})=\left(\lambda_2\hat{\alpha}_1,\ldots,\lambda_2\hat{\alpha}_{\mathsf{p}^0}\right)^{\mathsf{T}}$. Then, adding and subtracting the true value α_0:

$$\begin{aligned}(\hat{\alpha}-\alpha_0)_\star&=(\tilde{\alpha}-\alpha_0)_\star+(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}\hat{G}_{\star\bullet}^{\alpha\alpha}\tilde{\alpha}_\bullet-(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}\hat{G}_{\star\star}^{\alpha\beta}(\hat{\beta}-\tilde{\beta})_\star\\&\quad+(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}\hat{G}_{\star\bullet}^{\alpha\beta}\tilde{\beta}_\bullet-(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}Z(\hat{\alpha})-(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}U(\hat{\alpha}).\end{aligned}$$

Let us introduce a $\mathsf{m}^0\times\mathsf{m}$ matrix

$$J:=\begin{pmatrix}J_\alpha & 0\\ 0 & J_\beta\end{pmatrix},$$

where $J_\alpha:=(\mathbf{I}_{\mathsf{p}^0}\ (\Gamma_{\star\star}^{\alpha\alpha})^{-1}\Gamma_{\star\bullet}^{\alpha\alpha})$ and $J_\beta:=(\mathbf{I}_{\mathsf{q}^0}\ (\Gamma_{\star\star}^{\beta\beta})^{-1}\Gamma_{\star\bullet}^{\beta\beta})$.

Let $\hat{J}_\alpha:=\left(\mathbf{I}_{\mathsf{p}^0}\ n\Delta_n(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}\frac{1}{n\Delta_n}\hat{G}_{\star\bullet}^{\alpha\alpha}\right)\xrightarrow{p}J_\alpha$. Hence multiplying by $\sqrt{n\Delta_n}$ and adding and subtracting $J_\alpha\left\{\sqrt{n\Delta_n}(\tilde{\alpha}-\alpha_0)\right\}$:

$$\begin{aligned}&\sqrt{n\Delta_n}(\hat{\alpha}-\alpha_0)_\star-J_\alpha\left\{\sqrt{n\Delta_n}(\tilde{\alpha}-\alpha_0)\right\}\\&=\mathbb{1}_{B_n^{\alpha}}\Big\{\sqrt{n\Delta_n}(\tilde{\alpha}-\alpha_0)_\star+\sqrt{n\Delta_n}(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}\hat{G}_{\star\bullet}^{\alpha\alpha}\tilde{\alpha}_\bullet-\sqrt{n\Delta_n}(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}\hat{G}_{\star\star}^{\alpha\beta}(\hat{\beta}-\tilde{\beta})_\star\\&\quad+\sqrt{n\Delta_n}(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}\hat{G}_{\star\bullet}^{\alpha\beta}\tilde{\beta}_\bullet-\sqrt{n\Delta_n}(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}Z(\hat{\alpha})-\sqrt{n\Delta_n}(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}U(\hat{\alpha})\\&\quad-J_\alpha\left\{\sqrt{n\Delta_n}(\tilde{\alpha}-\alpha_0)\right\}\Big\}\\&\quad+\mathbb{1}_{(B_n^{\alpha})^c}\left\{\sqrt{n\Delta_n}(\hat{\alpha}-\alpha_0)_\star-J_\alpha\left\{\sqrt{n\Delta_n}(\tilde{\alpha}-\alpha_0)\right\}\right\}\\&=\mathbb{1}_{B_n^{\alpha}}\Big\{(\hat{J}_\alpha-J_\alpha)\left\{\sqrt{n\Delta_n}(\tilde{\alpha}-\alpha_0)\right\}-\sqrt{n\Delta_n}(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}\hat{G}_{\star\star}^{\alpha\beta}(\hat{\beta}-\tilde{\beta})_\star\\&\quad+\sqrt{n\Delta_n}(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}\hat{G}_{\star\bullet}^{\alpha\beta}\tilde{\beta}_\bullet\\&\quad-\sqrt{n\Delta_n}(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}Z(\hat{\alpha})-\sqrt{n\Delta_n}(\hat{G}_{\star\star}^{\alpha\alpha})^{-1}U(\hat{\alpha})\Big\}\\&\quad+\mathbb{1}_{(B_n^{\alpha})^c}\left\{\sqrt{n\Delta_n}(\hat{\alpha}-\alpha_0)_\star-J_\alpha\left\{\sqrt{n\Delta_n}(\tilde{\alpha}-\alpha_0)\right\}\right\}\\&=\mathbb{1}_{B_n^{\alpha}}\left\{(\hat{J}_\alpha-J_\alpha)\left\{\sqrt{n\Delta_n}(\tilde{\alpha}-\alpha_0)\right\}\right\}+o_p(1)\end{aligned}$$

$$+ \mathbb{1}_{(B_n^\alpha)^c} \left\{ \sqrt{n\Delta_n}(\hat{\alpha} - \alpha_0)_\star - J_\alpha \left\{ \sqrt{n\Delta_n}(\tilde{\alpha} - \alpha_0) \right\} \right\}$$

where the last step holds because:

$$n\Delta_n (\hat{G}_{\star\star}^{\alpha\alpha})^{-1} \frac{1}{\sqrt{n\Delta_n}} \frac{1}{\sqrt{n}} \hat{G}_{\star\star}^{\alpha\beta} \sqrt{n}(\hat{\beta} - \tilde{\beta})_\star \mathbb{1}_{B_n^\alpha} = o_p(1) O_p(1) = o_p(1)$$

$$n\Delta_n (\hat{G}_{\star\star}^{\alpha\alpha})^{-1} \frac{1}{\sqrt{n\Delta_n}} \frac{1}{\sqrt{n}} \hat{G}_{\star\bullet}^{\alpha\beta} \sqrt{n} \tilde{\beta}_\bullet \mathbb{1}_{B_n^\alpha} = o_p(1)$$

$$n\Delta_n (\hat{G}_{\star\star}^{\alpha\alpha})^{-1} \frac{1}{\sqrt{n\Delta_n}} Z(\hat{\alpha}) \mathbb{1}_{B_n^\alpha} = o_p(1)$$

$$n\Delta_n (\hat{G}_{\star\star}^{\alpha\alpha})^{-1} \frac{1}{\sqrt{n\Delta_n}} U(\hat{\alpha}) \mathbb{1}_{B_n^\alpha} = o_p(1).$$

Finally,

$$\sqrt{n\Delta_n}(\hat{\alpha} - \alpha_0)_\star - J_\alpha \left\{ \sqrt{n\Delta_n}(\tilde{\alpha} - \alpha_0) \right\} \xrightarrow{p} 0.$$

The same steps can be done considering on B_n^β, with different rates and the result holds.

Adding the property P2 and following the blockwise inversion of $\Gamma^{\alpha\alpha}$ and $\Gamma^{\beta\beta}$, exploited in the proof of Theorem 3 in De Gregorio and Iafrate (2021), we get

$$\left(\sqrt{n\Delta_n}(\hat{\alpha} - \alpha_0)_\star,\ \sqrt{n}(\hat{\beta} - \beta_0)_\star \right)^\intercal \xrightarrow{d} N_{\mathsf{m}^0} \left(0, \operatorname{diag} \left((\Gamma_{\star\star}^{\alpha\alpha})^{-1}, (\Gamma_{\star\star}^{\beta\beta})^{-1} \right) \right).$$

□

Proof of Theorem 2 Starting from (35), by definition of convexity we have the following inequalities, for any $r \geq 1$:

$$\begin{aligned}
|A_n^{-1}(\hat{\theta}_n - \theta_0)|^r \leq & \Big| 2\|\hat{D}_n^{-1}\| \|\hat{D}_n\| |A_n^{-1}(\tilde{\theta}_n - \theta_0)| \\
& + \|\hat{D}_n^{-1}\| \left(\mathsf{p}^0 \frac{a_n}{\sqrt{n\Delta_n}} + \mathsf{q}^0 \frac{c_n}{\sqrt{n}} + 2\mathsf{p}^0 \frac{\lambda_{2,n}}{\sqrt{n\Delta_n}} |\alpha_0| + 2\mathsf{q}^0 \frac{\gamma_{2,n}}{\sqrt{n}} |\beta_0| \right) \Big|^r \\
\leq & \left| \frac{1}{2} \left| 2 \cdot 2\|\hat{D}_n^{-1}\| \|\hat{D}_n\| |A_n^{-1}(\tilde{\theta}_n - \theta_0)| \right| \right. \\
& \left. + \frac{1}{2} \left| 2 \cdot \|\hat{D}_n^{-1}\| \left(\mathsf{p}^0 \frac{a_n}{\sqrt{n\Delta_n}} + \mathsf{q}^0 \frac{c_n}{\sqrt{n}} + 2\mathsf{p}^0 \frac{\lambda_{2,n}}{\sqrt{n\Delta_n}} |\alpha_0| + 2\mathsf{q}^0 \frac{\gamma_{2,n}}{\sqrt{n}} |\beta_0| \right) \right| \right|^r \\
\leq & \frac{1}{2} \left| 2 \cdot 2\|\hat{D}_n^{-1}\| \|\hat{D}_n\| |A_n^{-1}(\tilde{\theta}_n - \theta_0)| \right|^r \\
& + \frac{1}{2} \left| 2 \cdot \|\hat{D}_n^{-1}\| \left(\mathsf{p}^0 \frac{a_n}{\sqrt{n\Delta_n}} + \mathsf{q}^0 \frac{c_n}{\sqrt{n}} + 2\mathsf{p}^0 \frac{\lambda_{2,n}}{\sqrt{n\Delta_n}} |\alpha_0| + 2\mathsf{q}^0 \frac{\gamma_{2,n}}{\sqrt{n}} |\beta_0| \right) \right|^r \\
= & 2^{2r-1} \|\hat{D}_n^{-1}\|^r \|\hat{D}_n\|^r |A_n^{-1}(\tilde{\theta}_n - \theta_0)|^r \\
& + 2^{r-1} \|\hat{D}_n^{-1}\|^r \left| \mathsf{p}^0 \frac{a_n}{\sqrt{n\Delta_n}} + \mathsf{q}^0 \frac{c_n}{\sqrt{n}} + 2\mathsf{p}^0 \frac{\lambda_{2,n}}{\sqrt{n\Delta_n}} |\alpha_0| + 2\mathsf{q}^0 \frac{\gamma_{2,n}}{\sqrt{n}} |\beta_0| \right|^r .
\end{aligned}$$

Applying the same previous steps, from the second part of the inequality it follows:

$$
\begin{aligned}
&\left|\mathrm{p}^0\frac{a_n}{\sqrt{n\Delta_n}}+\mathrm{q}^0\frac{c_n}{\sqrt{n}}+2\mathrm{p}^0\frac{\lambda_{2,n}}{\sqrt{n\Delta_n}}|\alpha_0|+2\mathrm{q}^0\frac{\gamma_{2,n}}{\sqrt{n}}|\beta_0|\right|^r\\
&\le 2^{r-1}\left|\mathrm{p}^0\frac{a_n}{\sqrt{n\Delta_n}}\right|^r+2^{r-1}\left|\mathrm{q}^0\frac{c_n}{\sqrt{n}}+2\mathrm{p}^0\frac{\lambda_{2,n}}{\sqrt{n\Delta_n}}|\alpha_0|+2\mathrm{q}^0\frac{\gamma_{2,n}}{\sqrt{n}}|\beta_0|\right|^r\\
&\le 2^{r-1}\left|\mathrm{p}^0\frac{a_n}{\sqrt{n\Delta_n}}\right|^r+2^{r-1}\left(2^{r-1}\left|\mathrm{q}^0\frac{c_n}{\sqrt{n}}\right|^r+2^{r-1}\left|2\mathrm{p}^0\frac{\lambda_{2,n}}{\sqrt{n\Delta_n}}|\alpha_0|+2\mathrm{q}^0\frac{\gamma_{2,n}}{\sqrt{n}}|\beta_0|\right|^r\right)\\
&\le 2^{r-1}\left|\mathrm{p}^0\frac{a_n}{\sqrt{n\Delta_n}}\right|^r+2^{2(r-1)}\left|\mathrm{q}^0\frac{c_n}{\sqrt{n}}\right|^r\\
&\quad+2^{2(r-1)}\left(2^{r-1}\left|2\mathrm{p}^0\frac{\lambda_{2,n}}{\sqrt{n\Delta_n}}|\alpha_0|\right|^r+2^{r-1}\left|2\mathrm{q}^0\frac{\gamma_{2,n}}{\sqrt{n}}|\beta_0|\right|^r\right)\\
&=2^{r-1}\left|\mathrm{p}^0\frac{a_n}{\sqrt{n\Delta_n}}\right|^r+2^{2(r-1)}\left|\mathrm{q}^0\frac{c_n}{\sqrt{n}}\right|^r+2^{4r-3}\left|\mathrm{p}^0\frac{\lambda_{2,n}}{\sqrt{n\Delta_n}}|\alpha_0|\right|^r+2^{4r-3}\left|\mathrm{q}^0\frac{\gamma_{2,n}}{\sqrt{n}}|\beta_0|\right|^r.
\end{aligned}
$$

Then

$$
\begin{aligned}
&2^{r-1}\|\hat{D}_n^{-1}\|^r\left|\mathrm{p}^0\frac{a_n}{\sqrt{n\Delta_n}}+\mathrm{q}^0\frac{c_n}{\sqrt{n}}+2\mathrm{p}^0\frac{\lambda_{2,n}}{\sqrt{n\Delta_n}}|\alpha_0|+2\mathrm{q}^0\frac{\gamma_{2,n}}{\sqrt{n}}|\beta_0|\right|^r\\
&\le\|\hat{D}_n^{-1}\|^r\left(2^{2(r-1)}\left|\mathrm{p}^0\frac{a_n}{\sqrt{n\Delta_n}}\right|^r+2^{3(r-1)}\left|\mathrm{q}^0\frac{c_n}{\sqrt{n}}\right|^r+2^{5r-4}\left|\mathrm{p}^0\frac{\lambda_{2,n}}{\sqrt{n\Delta_n}}|\alpha_0|\right|^r\right.\\
&\quad\left.+2^{5r-4}\left|\mathrm{q}^0\frac{\gamma_{2,n}}{\sqrt{n}}|\beta_0|\right|^r\right).
\end{aligned}
$$

Finally from the Cauchy–Schwarz inequality

$$
\begin{aligned}
\mathbb{E}\left[|A_n^{-1}(\hat{\theta}_n-\theta_0)|^r\right]&\le 2^{2r-1}\sqrt{\mathbb{E}\left[\|\hat{D}_n^{-1}\|^{2r}\right]}\left(\mathbb{E}\left[\|\hat{D}_n\|^{4r}\right]\right)^{\frac{1}{4}}\left(\mathbb{E}\left[|A_n^{-1}(\hat{\theta}_n-\theta_0)|^{4r}\right]\right)^{\frac{1}{4}}\\
&\quad+\|\hat{D}_n^{-1}\|^r\left(2^{2(r-1)}\mathbb{E}\left[\left|\mathrm{p}^0\frac{a_n}{\sqrt{n\Delta_n}}\right|^r\right]+2^{3(r-1)}\mathbb{E}\left[\left|\mathrm{q}^0\frac{c_n}{\sqrt{n}}\right|^r\right]\right.\\
&\quad\left.+2^{5r-4}\mathbb{E}\left[\left|\mathrm{p}^0\frac{\lambda_{2,n}}{\sqrt{n\Delta_n}}|\alpha_0|\right|^r\right]+2^{5r-4}\mathbb{E}\left[\left|\mathrm{q}^0\frac{\gamma_{2,n}}{\sqrt{n}}|\beta_0|\right|^r\right]\right).
\end{aligned}
$$

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

Proof of Lemma 4 Let $\tilde{\theta}_n=(\tilde{\alpha}_n,\tilde{\beta}_n)$ be the quasi likelihood estimator. By Taylor expansion, the Cauchy–Schwartz inequality and A5(r)-(i) we have that

$$
\begin{aligned}
|\ell_n(\tilde{\alpha}_n,\tilde{\beta}_n)-\ell_n(\alpha_0,\tilde{\beta}_n)|&\le\left|\int_0^1\left\langle\partial_\alpha\ell_n(\alpha_0+u(\tilde{\alpha}_n-\alpha_0),\tilde{\beta}_n),\tilde{\alpha}_n-\alpha_0\right\rangle\mathrm{d}u\right|\\
&\le\int_0^1|\partial_\alpha\overline{\ell_n}(\alpha_0+u(\tilde{\alpha}_n-\alpha_0),\tilde{\beta}_n)|\mathrm{d}u\,|\sqrt{n\Delta_n}(\tilde{\alpha}_n-\alpha_0)|\\
&\le\xi_n\sqrt{\mathrm{p}}|\sqrt{n\Delta_n}(\tilde{\alpha}_n-\alpha_0)|.
\end{aligned}
$$

Similarly, under the assumption that $|\tilde{\theta}_n-\theta_0|\le r$ and by A5$(r)-(ii)$ we have that

$$
\begin{aligned}
|\ell_n(\alpha_0,\tilde{\beta}_n)-\ell_n(\tilde{\alpha}_n,\tilde{\beta}_n)|&=\left|\int_0^1(1-u)\left\langle\partial_{\alpha\alpha}^2\ell_n(\tilde{\alpha}_n+u(\alpha_0-\tilde{\alpha}_n),\tilde{\beta}_n),(\tilde{\alpha}_n-\alpha_0)^{\otimes 2}\right\rangle\mathrm{d}u\right|\\
&\ge\frac{\mu}{2}|\sqrt{n\Delta_n}(\tilde{\alpha}_n-\alpha_0)|^2.
\end{aligned}
$$

By combining the two inequalities above we get the first claim. The proof for the second inequality is analogous. □

Proof of Theorem 3 By setting $B = \hat{G}_n^{1/2}$, $y = \hat{G}_n^{1/2}\tilde{\theta}_n$, $\epsilon = \hat{G}_n^{1/2}(\tilde{\theta}_n - \theta_0)$, $\lambda_i = \lambda_{i,n}$, $\gamma_i = \gamma_{i,n}$, $i = 1, 2$, the adaptive Elastic-Net estimator can be rewritten as the solution of a linear regression-type minimization problem; i.e.

$$\hat{\theta}_n = \hat{\theta}_n(\lambda_1, \gamma_1, \lambda_2, \gamma_2) = \arg\min_{\theta} \left\{ |y - B\theta|^2 + \lambda_1 |\alpha|_{1,\kappa_n} + \gamma_1 |\beta|_{1,\pi_n} + \theta^{\intercal} C(\lambda_2, \gamma_2)\theta \right\}$$

where $C(\lambda_2, \gamma_2) := \mathrm{diag}(\lambda_2 \mathbf{I}_{\mathsf{p}}, \gamma_2 \mathbf{I}_{\mathsf{q}})$.

Then the Ridge estimator ($\lambda_1 = \gamma_1 = 0$) is given by

$$\hat{\theta}_n(\lambda_2, \gamma_2) = (\hat{B}^{\top} B + C(\lambda_2, \gamma_2))^{-1} B^{\top} y = (\hat{G}_n + C(\lambda_2, \gamma_2))^{-1} \hat{G}_n \tilde{\theta}_n = (\hat{\alpha}_n(\lambda_2), \hat{\beta}_n(\gamma_2))^{\intercal}$$

where $\hat{\alpha}_n(\lambda_2)$ and $\hat{\beta}_n(\gamma_2)$ are the Ridge estimators related to the sub-optimization problems $\mathcal{F}_{1,n}$ and $\mathcal{F}_{2,n}$, respectively; i.e.

$$\begin{aligned} \hat{\alpha}_n(\lambda_2) &= (\hat{G}_n^{\alpha\alpha} + \lambda_2 \mathbf{I}_{\mathsf{p}})^{-1} \hat{G}_n^{\alpha\alpha} \tilde{\alpha}_n, \\ \hat{\beta}_n(\gamma_2) &= (\hat{G}_n^{\beta\beta} + \gamma_2 \mathbf{I}_{\mathsf{q}})^{-1} \hat{G}_n^{\beta\beta} \tilde{\beta}_n. \end{aligned}$$

By following the same steps as in Theorem 3.1 in Zou and Zhang (2009) we have that:

$$\begin{aligned} |\hat{\theta}_n - \theta_0|^2 &\le 2|\hat{\theta}_n - \hat{\theta}_n(\lambda_2, \gamma_2)|^2 + 2|\hat{\theta}_n(\lambda_2, \gamma_2) - \theta_0|^2 \\ &= 2|\hat{\alpha}_n - \hat{\alpha}_n(\lambda_2)|^2 + 2|\hat{\alpha}_n(\lambda_2) - \alpha_0|^2 \\ &\quad + 2|\hat{\beta}_n - \hat{\beta}_n(\gamma_2)|^2 + 2|\hat{\beta}_n(\gamma_2) - \beta_0|^2. \end{aligned}$$

Since

$$\hat{\alpha}_n(\lambda_2) - \alpha_0 = (\hat{G}_n^{\alpha\alpha} + \lambda_2 \mathbf{I}_{\mathsf{p}})^{-1} \hat{G}_n^{\alpha\alpha} (\tilde{\alpha}_n - \alpha_0) - (\hat{G}_n^{\alpha\alpha} + \lambda_2 \mathbf{I}_{\mathsf{p}})^{-1} \lambda_2 \alpha_0$$

we have that

$$\begin{aligned} |\hat{\alpha}_n(\lambda_2) - \alpha_0|^2 &\le 2|(\hat{G}_n^{\alpha\alpha} + \lambda_2 \mathbf{I}_{\mathsf{p}})^{-1} \lambda_2 \alpha_0|^2 + 2|(\hat{G}_n^{\alpha\alpha} + \lambda_2 \mathbf{I}_{\mathsf{p}})^{-1} \hat{G}_n^{\alpha\alpha} (\tilde{\alpha}_n - \alpha_0)|^2 \\ &\le \frac{2}{(\tau_{\min}(\hat{G}_n^{\alpha\alpha}) + \lambda_2)^2} \left(\lambda_2^2 |\alpha_0|^2 + |\hat{G}_n^{\alpha\alpha} (\tilde{\alpha}_n - \alpha_0)|^2 \right) \\ &\le \frac{2}{(\tau_{\min}(\hat{G}_n^{\alpha\alpha}) + \lambda_2)^2} \left(\lambda_2^2 |\alpha_0|^2 + \tau_{\max}((\hat{G}_n^{\alpha\alpha})^2) |\tilde{\alpha}_n - \alpha_0|^2 \right). \end{aligned} \tag{38}$$

Furthermore, by means of the same arguments adopted in Zou and Zhang (2009), pag.1744, we get

$$\begin{aligned} (\tau_{\min}(\hat{G}_n^{\alpha\alpha}) + \lambda_2) |\hat{\alpha}_n - \hat{\alpha}_n(\lambda_2)|^2 &\le (\hat{\alpha}_n - \hat{\alpha}_n(\lambda_2))^{\intercal} (\hat{G}_n^{\alpha\alpha} + \lambda_2 \mathbf{I}_{\mathsf{p}}) (\hat{\alpha}_n - \hat{\alpha}_n(\lambda_2)) \\ &\le \lambda_1 |\kappa_n| |\hat{\alpha}_n - \hat{\alpha}_n(\lambda_2)| \end{aligned}$$

and then

$$|\hat{\alpha}_n - \hat{\alpha}_n(\lambda_2)| \le \frac{\lambda_1 |\kappa_n|}{\tau_{\min}(\hat{G}_n^{\alpha\alpha}) + \lambda_2}. \tag{39}$$

Therefore, from (38) and (39), we obtain

$$|\hat{\alpha}_n - \alpha_0|^2 \le \frac{4}{(\tau_{\min}(\hat{G}_n^{\alpha\alpha}) + \lambda_2)^2} \left(\lambda_2^2 |\alpha_0|^2 + \tau_{\max}((\hat{G}_n^{\alpha\alpha})^2) |\tilde{\alpha}_n - \alpha_0|^2 + \lambda_1^2 |\kappa_n|^2 \right). \tag{40}$$

For two real symmetric positive definite matrices A and B, we have the following inequalities involving the minimum and maximum eigenvalues $\frac{1}{\tau_{\min}(AB)} \leq \frac{1}{\tau_{\min}(A)}\frac{1}{\tau_{\min}(B)}$ and $\tau_{\max}(AB) \leq \tau_{\max}(A)\tau_{\max}(B)$. Therefore from (40), we get

$$|\hat{\alpha}_n - \alpha_0|^2 \leq \frac{4}{(n\Delta_n \tau_{\min}(\hat{D}_n^{\alpha\alpha}) + \lambda_2)^2}\left(\lambda_2^2|\alpha_0|^2 + (n\Delta_n \tau_{\max}(\hat{D}_n^{\alpha\alpha})|\tilde{\alpha}_n - \alpha_0|)^2 + \lambda_1^2|\kappa_n|^2\right). \tag{41}$$

By similar steps we can write down

$$\begin{aligned}|\hat{\beta}_n - \beta_0|^2 &\leq \frac{4}{(\tau_{\min}(\hat{G}_n^{\beta\beta}) + \gamma_2)^2}\left(\gamma_2^2|\beta_0|^2 + \tau_{\max}((\hat{G}_n^{\beta\beta})^2)|\tilde{\beta}_n - \beta_0|^2 + \gamma_1^2|\pi_n|^2\right)\\ &\leq \frac{4}{(n\tau_{\min}(\hat{D}_n^{\beta\beta}) + \gamma_2)^2}\left(\gamma_2^2|\beta_0|^2 + (n\tau_{\max}(\hat{D}_n^{\beta\beta})|\tilde{\beta}_n - \beta_0|)^2 + \gamma_1^2|\pi_n|^2\right).\end{aligned}$$

□

Proof of Theorem 5 We prove the result for $\hat{\alpha}_n$, the proof for $\hat{\beta}_n$ is analogous. From Theorem 3, inequality (16) holds true, i.e.

$$|\hat{\alpha}_n - \alpha_0| \leq \frac{2}{n\Delta_n \tau_{\min}(\hat{D}_n^{\alpha\alpha}) + \lambda_{2,n}}\left(\lambda_{2,n}|\alpha_0| + n\Delta_n \tau_{\max}(\hat{D}_n^{\alpha\alpha})|\tilde{\alpha}_n - \alpha_0| + \lambda_{1,n}|\kappa_n|\right). \tag{42}$$

On the event $\{|A_n^{-1}(\tilde{\theta}_n - \theta_0)| \leq r\}$, which implies $\{|\tilde{\theta}_n - \theta_0| \leq r/n\Delta_n\}$ Lemma 4 with A5$(r/n\Delta_n)$ gives

$$\sqrt{n\Delta_n}|\tilde{\alpha}_n - \alpha_0| \leq \frac{2\xi_n}{\mu}\sqrt{\mathsf{p}}.$$

By combining the two inequalities and by using A6, inequality (19) immediately follows. It remains to show that it holds with the desired probability. This is a consequence of the polynomial large deviation results in Yoshida (2011) (see e.g. formula 2.14 in Yoshida (2022)). □

Proof of Theorem 6 Let R be a function on $\Theta \times \mathbb{R} \times \mathbb{R}^d$ for which there exists a positive constant C such that $R(\theta, u, x) \leq u(1 + ||x||)^C$ for all θ, u, x. We have that

$$\begin{aligned}\mathsf{MAE}(\hat{X}_{T_n+h}) &= \mathbb{E}|X_{T_n+h} - \hat{X}_{T_n+h}|\\ &= \mathbb{E}|X_{T_n+h} - X_{T_n} - hb(X_{T_n}, \alpha_0) - h(b(X_{T_n}, \hat{\alpha}_n) - b(X_{T_n}, \alpha_0))|\\ &\leq \left(\mathbb{E}\left[\mathbb{E}[|X_{T_n+h} - X_{T_n} - hb(X_{T_n}, \alpha_0)|^2|\mathcal{F}_{T_n}]\right]\right)^{\frac{1}{2}} + h\mathbb{E}[C(X_{T_n})|\hat{\alpha} - \alpha_0|]\\ &\leq \left(\mathbb{E}[h\,\mathrm{tr}(\Sigma(X_{t_n}, \beta_0)) + R(\theta_0, h^2, X_{T_n})]\right)^{\frac{1}{2}} + h\left(\mathbb{E}C^2(X_{T_n})\mathbb{E}|\hat{\alpha}_n - \alpha_0|^2\right)^{\frac{1}{2}}\\ &\leq C_1\sqrt{h} + C_2 h + Dh\left(\mathbb{E}|\hat{\alpha}_n - \alpha_0|^2\right)^{\frac{1}{2}},\end{aligned}$$

where we applied the Cauchy–Schwartz inequality and in the second step we applied Lemma 7 in Kessler (1997), adapted to the multidimensional case. The conclusion follows by Theorem 3. □

Acknowledgements The authors would like to thank both reviewers for their comments, which led to improving the previous version of the paper.

Author Contributions A.D.G, D.F., S.I. and F.I.: Conceptualization (equal); Formal analysis (equal); Methodology (equal); Supervision (equal); Visualization (equal); Writing - original draft (equal).

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

Data Availability No datasets were generated or analysed during the current study.

Code Availability The code for the numerical analysis presented in this paper is publicly available at online repository Iafrate and Frisardi (2024).

Declarations

Competing interests The authors declare no competing interests.

References

Amorino C, Pina F, Podolskij M (2025) Sampling effects on lasso estimation of drift functions in high-dimensional diffusion processes. Electron J Stat 19(2):5068–5116

Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J Imag Sci 2(1):183–202

Bolte J, Sabach S, Teboulle M (2014) Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math Program 146(1):459–494

Carpi T, Hino A, Iacus SM, Porro G (2022) The impact of covid-19 on subjective well-being: evidence from twitter data. J Data Sci 21(4):761–780

Ciolek G, Marushkevych D, Podolskij M (2022) On lasso estimator for the drift function in diffusion models. *arXiv preprint* arXiv:2209.05974

Ciołek G, Marushkevych D, Podolskij M (2020) On Dantzig and Lasso estimators of the drift in a high dimensional Ornstein-Uhlenbeck model. Electron J Stat 14(2):4395–4420

De Gregorio A, Iacus SM (2012) Adaptive lasso-type estimation for multivariate diffusion processes. Economet Theor 28(4):838–860

De Gregorio A, Iacus SM (2018) On penalized estimation for dynamical systems with small noise. Electron J Stat 12(1):1614–1630

De Gregorio A, Iafrate F (2021) Regularized bridge-type estimation with multiple penalties. Ann Inst Stat Math 73(5):921–951

De Gregorio A, Iafrate F (2024) Pathwise optimization for bridge-type estimators and its applications, https://arxiv.org/abs/2412.04047

Dexheimer N, Strauch C (2024) On Lasso and Slope drift estimators for Lévy-driven Ornstein-Uhlenbeck processes. Bernoulli 30(1):88–116

Ditlevsen S, Samson A (2019) Hypoelliptic diffusions: filtering and inference from complete and partial observations. J R Stat Soc Ser B Stat Methodol 81(2):361–384

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96(456):1348–1360

Fan J, Li R (2006) Statistical challenges with high dimensionality: feature selection in knowledge discovery. *arXiv preprint* math/0602133

Fan J, Peng H (2004) Nonconcave penalized likelihood with a diverging number of parameters. Ann Stat 32(3):928–961
Florens-Zmirou D (1989) Approximate discrete-time schemes for statistics of diffusion processes. Stat J Theor Appl Stat 20(4):547–557
Frank LE, Friedman JH (1993) A statistical view of some chemometrics regression tools. Technometrics 35(2):109–135
Fujimori K (2019) The dantzig selector for a linear model of diffusion processes. Stat Infer Stoch Process 22:475–498
Gaïffas S, Matulewicz G (2019) Sparse inference of the drift of a high-dimensional ornstein-uhlenbeck process. J Multivar Anal 169:1–20
Gloter A, Yoshida N (2021) Adaptive estimation for degenerate diffusion processes. Electron J Stat 15(1):1424–1472
Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity. Monogr Stat Appl Probab 143(143):8
Iafrate F (2024) SDELearn: a Python package for SDE Modeling. https://github.com/fiafrate/sdelearn. Version 0.1.2
Iafrate F, Frisardi D (2024) Elastic net estimation for ergodic diffusion processes. https://github.com/fiafrate/enet-sde
Iguchi Y, Beskos A (2025) Parameter inference for hypo-elliptic diffusions under a weak design condition. Electron J Stat 19(1):1337–1369
Kamatani K, Uchida M (2015) Hybrid multi-step estimators for stochastic differential equations based on sampled data. Stat Infer Stoch Process 18(2):177–204
Kessler M (1997) Estimation of an ergodic diffusion from discrete observations. Scand J Stat 24(2):211–229
Kinoshita Y, Yoshida N (2019) Penalized quasi likelihood estimation for variable selection. *arXiv preprint* arXiv:1910.12871
Kloeden PE, Platen E (1992) Numerical solution of stochastic differential equations. Springer, Berlin Heidelberg
Masuda H, Shimizu Y (2017) Moment convergence in regularized estimation under multiple and mixed-rates asymptotics. Math Methods Statist 26:81–110
Pardoux É, Veretennikov Y (2001) On the poisson equation and diffusion approximation. I. Ann Probab 29(3):1061–1085
Sørensen M (2024) Efficient estimation for ergodic diffusion processes sampled at high frequency. *arXiv preprint* arXiv:2401.04689
Suzuki T, Yoshida N (2020) Penalized least squares approximation methods and their applications to stochastic processes. Japanese J Stat Data Sci 3(2):513–541
Tibshirani R (1996) Regression selection and shrinkage via the lasso. J Roy Stat Soc B 58(1):267–288
Uchida M, Yoshida N (2012) Adaptive estimation of an ergodic diffusion process based on sampled data. Stoch Process Appl 122(8):2885–2924
Veretennikov AY (1988) Bounds for the mixing rate in the theory of stochastic equations. Theory Probab Appl 32(2):273–281
Wang H, Leng C (2007) Unified lasso estimation by least squares approximation. J Am Stat Assoc 102(479):1039–1048
Yoshida N (2011) Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. Ann Inst Stat Math 63(3):431–479
Yoshida N (2022) Quasi-likelihood analysis for nonlinear stochastic processes. Econom Stat 25(1):43–60
Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 67(2):301–320
Zou H, Zhang HH (2009) On the adaptive elastic-net with a diverging number of parameters. Ann Stat 37(4):1733