Singular ridge regression with homoscedastic residuals: generalization error with estimated parameters

Lyudmila Grigoryeva[1], Juan-Pablo Ortega[2,3]

Abstract

This paper characterizes the conditional distribution properties of the finite sample ridge regression estimator and uses that result to evaluate total regression and generalization errors that incorporate the inaccuracies committed at the time of parameter estimation. The paper provides explicit formulas for those errors. Unlike other classical references in this setup, our results take place in a fully singular setup that does not assume the existence of a solution for the non-regularized regression problem. In exchange, we invoke a conditional homoscedasticity hypothesis on the regularized regression residuals that is crucial in our developments.

Key Words: ridge regression, singular regression, training error, testing error, generalization error.

1 Introduction

The ridge regression [Tikh 43, Tikh 63, Hoer 70] has been introduced as a regularization method to estimate the unknown matrix B in the linear regression model

$$\mathbf{y} = B^\top \mathbf{x} + \boldsymbol{\varepsilon}, \tag{1.1}$$

when the covariance matrix of the explanatory variables $\Sigma_{\mathbf{x}} := \mathrm{Cov}(\mathbf{x}, \mathbf{x}) \in \mathbb{S}_p$ is singular or poorly conditioned and, therefore, the standard least squares estimator is ill-defined. The ridge regularization method produces an estimate $\widehat{B}_\lambda$ (λ is a regularization strength that we introduce later on) whose properties are always formulated in terms of the coefficient B that is assumed to solve the least squares problem. Nevertheless, there are many applications in physics, engineering, or machine learning (see, for instance, [Gema 92]) in which an underlying linear model like (1.1) with a B that solves the least squares problem does not exist and only a regularized version of it with parameter B_λ is available. In this paper we place ourselves in that fully singular situation and nevertheless, we manage to establish the conditional distribution properties of the finite sample ridge regression estimator $\widehat{B}_\lambda$ directly in terms of B_λ without using the least squares coefficient B in (1.1) whose existence is not needed. These results require that certain hypotheses on the homoscedasticity of the regularized regression residuals hold.

[1]Department of Mathematics and Statistics. Universität Konstanz. Box 146. D-78457 Konstanz. Germany. Lyudmila.Grigoryeva@uni-konstanz.de

[2]Universität Sankt Gallen. Bodanstrasse 6. CH-9000 Sankt Gallen. Switzerland. Juan-Pablo.Ortega@unisg.ch

[3]Centre National de la Recherche Scientifique (CNRS). France.

The second main contribution of this paper consists of using the properties of this estimator to write down explicit expressions that evaluate the regression (also called training) and the generalization (also called testing) errors committed by a regularized regression model. We recall that the regression or training error is the one committed by the regularized regression model when calculated with the finite sample that has been used to obtain the estimate $\widehat{B}_\lambda$ of the regression coefficient B_λ; for the generalization or testing error we keep $\widehat{B}_\lambda$ and we compute the error committed by the corresponding regression model using another sample that may have different size or even different statistical properties. In both cases our error formulas incorporate the error committed at the time of parameter estimation using the finite training sample.

The paper is structured in two main sections. The first one is Section 2 that contains a description of our setup and the hypotheses that we invoke. It also presents the properties of the ridge estimator in the singular framework that we have chosen to work on. Section 3 contains the results on the evaluation of the training and testing errors. All along the paper we illustrate our developments with various classical examples that give an idea of the scope of our results. All the proofs of the results and the technical details are included in the appendices in Section 4.

Notation and conventions: Column vectors are denoted by bold lower or upper case symbol like $\mathbf{v}$ or $\mathbf{V}$. We write $\mathbf{v}^\top$ to indicate the transpose of $\mathbf{v}$. Given a vector $\mathbf{v} \in \mathbb{R}^n$, we denote its entries by v_i, with $i \in \{1, \ldots, n\}$; we also write $\mathbf{v} = (v_i)_{i\in\{1,\ldots,n\}}$. The symbols $\mathbf{i}_n$ and $\mathbf{0}_n$ stand for the vectors of length n consisting of ones and zeros, respectively. We denote by $\mathbb{M}_{n,m}$ the space of real $n \times m$ matrices with $m, n \in \mathbb{N}$. When $n = m$, we use the symbol $\mathbb{M}_n$ to refer to the space of square matrices of order n. Given a matrix $A \in \mathbb{M}_{n,m}$, we denote its components by A_{ij} and we write $A = (A_{ij})$, with $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots m\}$. If A and B are two matrices with the same number of rows, we denote by $(A||B)$ the matrix resulting from their horizontal concatenation. We write $\mathbb{I}_n$ to denote the identity matrix of dimension n. We use $\mathbb{S}_n$ to indicate the subspace $\mathbb{S}_n \subset \mathbb{M}_n$ of symmetric matrices, that is, $\mathbb{S}_n = \left\{A \in \mathbb{M}_n \mid A^\top = A\right\}$. The symbol $||A||_{\text{Frob}}$ denotes the Frobenius norm of $A \in \mathbb{M}_{m,n}$ defined as $\|A\|^2_{\text{Frob}} := \text{trace}\left(A^T A\right)$ [Meye 00]. The symbols $\text{E}[\cdot]$ and $\text{Cov}(\cdot, \cdot)$ denote the expectation and the covariance of random variables, respectively. Given a random variable X the symbols $\text{E}_X[\cdot]$ and $\text{Cov}_X(\cdot, \cdot)$ denote the conditional expectation and the conditional covariance with respect to X, respectively. Given a random vector $\mathbf{x}$ and a random matrix X, we will use the symbols $\boldsymbol{\mu}_\mathbf{x}$ and M_X to denote the mean of $\mathbf{x}$ and X, respectively. Let Z be a m by n matrix random variable; the symbol $Z \sim \text{MN}_{m,n}(M_Z, U_Z, V_Z)$ indicates that Z is distributed according to the matrix valued normal distribution with mean matrix $M_Z \in \mathbb{M}_{m,n}$ and scale matrices $U_Z \in \mathbb{S}_m$ and $V_Z \in \mathbb{S}_n$ (see [Gupt 00] for additional definitions and properties).

Glossary of symbols. p is the number of explanatory variables (dimension of $\mathbf{x}$), q is the number of dependent variables (dimension of $\mathbf{y}$), N is the dimension of the (random) sample size, λ is the regularization strength.
$\boldsymbol{\mu}_\mathbf{x} := \text{E}[\mathbf{x}]$, $\boldsymbol{\mu}_\mathbf{y} := \text{E}[\mathbf{y}]$, $\Sigma_\mathbf{x} := \text{Cov}(\mathbf{x}, \mathbf{x}) \in \mathbb{S}_p$, $\Sigma_\mathbf{y} := \text{Cov}(\mathbf{y}, \mathbf{y}) \in \mathbb{S}_q$, $\Sigma_\mathbf{xy} := \text{Cov}(\mathbf{x}, \mathbf{y}) \in \mathbb{M}_{p,q}$.
$B_\lambda := \left(\Sigma_\mathbf{x} + \lambda \mathbb{I}_p\right)^{-1} \Sigma_\mathbf{xy}$ is the ridge regression matrix.
$\boldsymbol{\varepsilon}_\lambda := \mathbf{y} - B_\lambda^\top \mathbf{x}$ are the ridge regression residuals.
$\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda} := \text{E}[\boldsymbol{\varepsilon}_\lambda] = \boldsymbol{\mu}_\mathbf{y} - B_\lambda^\top \boldsymbol{\mu}_\mathbf{x}$, $\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}} := \text{E}_\mathbf{x}[\boldsymbol{\varepsilon}_\lambda] = \text{E}_\mathbf{x}[\mathbf{y}] - B_\lambda^\top \mathbf{x}$.
$\Sigma_{\boldsymbol{\varepsilon}_\lambda} := \text{Cov}(\boldsymbol{\varepsilon}_\lambda, \boldsymbol{\varepsilon}_\lambda)$, $\Sigma_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}} := \text{Cov}_\mathbf{x}(\boldsymbol{\varepsilon}_\lambda, \boldsymbol{\varepsilon}_\lambda)$.
$X := (\mathbf{x}_1||\mathbf{x}_2||\ldots||\mathbf{x}_N) \in \mathbb{M}_{p,N}$, $Y := (\mathbf{y}_1||\mathbf{y}_2||\ldots||\mathbf{y}_N) \in \mathbb{M}_{q,N}$, and $E_\lambda := (\boldsymbol{\varepsilon}_{\lambda,1}||\ldots||\boldsymbol{\varepsilon}_{\lambda,N}) \in \mathbb{M}_{q,N}$.
$A_N := \mathbb{I}_N - \dfrac{1}{N}\mathbf{i}_N\mathbf{i}_N^\top$.
$\widehat{B}_\lambda := (XA_NX^\top + \lambda N\mathbb{I}_p)^{-1}XA_NY^\top$ is the ridge regression matrix estimator.
$Z_\lambda := R_\lambda XA_NX^\top = \mathbb{I}_p - \lambda NR_\lambda$.
$R_\lambda := (XA_NX^\top + \lambda N\mathbb{I}_p)^{-1}$.
For $XA_NX^\top$ invertible:

$\widehat{B} := \widehat{B}_0 = (XA_NX^\top)^{-1}XA_NY^\top$ is the ordinary least squares matrix estimator.
$\widehat{B}_\lambda = Z_\lambda \widehat{B}$.
$Z_\lambda := R_\lambda XA_NX^\top = \mathbb{I}_p - \lambda N R_\lambda = \left(\mathbb{I}_p + \lambda N (XA_NX^\top)^{-1}\right)^{-1}$.

Acknowledgments: We acknowledge partial financial support of the French ANR "BIPHOPROC" project (ANR-14-OHRI-0002-02).

2 Multivariate ridge regressions and the ridge estimator

2.1 The setup

Let $\mathbf{x}$ and $\mathbf{y}$ be respectively $\mathbb{R}^p$ and $\mathbb{R}^q$-valued random variables defined on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$. All along this paper we work under the following assumption:

(A1) *The random variables $\mathbf{x}, \mathbf{y}$ belong to $L^2(\Omega, \mathcal{F}, \mathbb{P})$, that is their first and second order moments exist and are finite, that is, $\mathrm{E}[\mathbf{x}] < \infty$, $\mathrm{E}[\mathbf{y}] < \infty$, $\mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right] < \infty$, $\mathrm{E}\left[\mathbf{y}\mathbf{y}^\top\right] < \infty$. Additionally, the Cauchy-Schwarz inequality implies the finiteness of the covariance between $\mathbf{x}$ and $\mathbf{y}$, that is, $\mathrm{Cov}(\mathbf{x}, \mathbf{y}) < \infty$.*

In the sequel we denote $\boldsymbol{\mu}_\mathbf{x} := \mathrm{E}[\mathbf{x}] \in \mathbb{R}^p$, $\boldsymbol{\mu}_\mathbf{y} := \mathrm{E}[\mathbf{y}] \in \mathbb{R}^q$ and use the following notation for the second order central moments: $\Sigma_\mathbf{x} := \mathrm{Cov}(\mathbf{x}, \mathbf{x}) \in \mathbb{S}_p$, $\Sigma_\mathbf{y} := \mathrm{Cov}(\mathbf{y}, \mathbf{y}) \in \mathbb{S}_q$, and $\Sigma_{\mathbf{xy}} := \mathrm{Cov}(\mathbf{x}, \mathbf{y}) \in \mathbb{M}_{p,q}$.

The multivariate ridge regression optimization problem. Consider the ridge penalized least squares problem and define:

$$B_\lambda := \underset{B \in \mathbb{M}_{p,q}}{\arg\min} \left(\mathrm{trace}\left(\mathrm{E}\left[(B^\top\mathbf{x} - \mathbf{y})(B^\top\mathbf{x} - \mathbf{y})^\top\right]\right) + \lambda \|B\|^2_{\mathrm{Frob}}\right), \tag{2.1}$$

where $\lambda \in \mathbb{R}^+$ is regularization strength and $||\cdot||^2_{\mathrm{Frob}}$ denotes the squared Frobenius norm of the coefficient matrix B. The optimization problem (2.1) does not contain an intercept which does not imply any loss of generality since its presence can be accommodated by replacing $B \in \mathbb{M}_{p,q}$ and $\mathbf{x} \in \mathbb{R}^p$ in (2.1) by $\tilde{B} \in \mathbb{M}_{p+1,q}$ and $\tilde{\mathbf{x}} \in \mathbb{R}^{p+1}$, respectively, that are constructed by vertical concatenation of the intercept vector $\mathbf{b}_0^\top \in \mathbb{R}^q$ with B and of 1 with $\mathbf{x} \in \mathbb{R}^p$, respectively. In this case the penalization summand in (2.1) has to be replaced by $\lambda \|\tilde{B}\|^{2^*}_{\mathrm{Frob}}$ where $||\tilde{B}||^{2^*}_{\mathrm{Frob}}$ would denote the squared Frobenius norm of the $p \times q$ lower submatrix of $\tilde{B}$.

In the presence of the assumption **(A1)** the multivariate ridge regression problem (2.1) has a unique solution $B_\lambda \in \mathbb{M}_{p,q}$ given by

$$B_\lambda := \left(\Sigma_\mathbf{x} + \lambda \mathbb{I}_p\right)^{-1} \Sigma_{\mathbf{xy}}. \tag{2.2}$$

We will refer to B_λ as the ridge regression matrix. We emphasize that B_λ always exists even if the covariance matrix $\Sigma_\mathbf{x}$ is singular or ill-conditioned, provided that $\lambda > 0$.

The regression residuals. Given the random variables $\mathbf{x}, \mathbf{y}$ that satisfy the hypothesis **(A1)**, we can define, for each $\lambda \in \mathbb{R}^+$, the regression residuals using the corresponding and uniquely determined ridge regression matrix B_λ in (2.2) as:

$$\boldsymbol{\varepsilon}_\lambda := \mathbf{y} - B_\lambda^\top \mathbf{x}. \tag{2.3}$$

The residuals $\boldsymbol{\varepsilon}_\lambda \in \mathbb{R}^q$ in (2.3) are a $\mathbb{R}^q$-valued distributed random variable with mean $\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda} \in \mathbb{R}^q$ and covariance matrix $\Sigma_{\boldsymbol{\varepsilon}_\lambda} \in \mathbb{S}^q$, that is,

$$\boldsymbol{\varepsilon}_\lambda \sim D(\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda}, \Sigma_{\boldsymbol{\varepsilon}_\lambda}), \text{ with } \boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda} := \boldsymbol{\mu}_\mathbf{y} - B_\lambda^\top \boldsymbol{\mu}_\mathbf{x},\ \Sigma_{\boldsymbol{\varepsilon}_\lambda} := \Sigma_\mathbf{y} - B_\lambda^\top \Sigma_{\mathbf{xy}} - \Sigma_{\mathbf{xy}} B_\lambda + B_\lambda^\top (\Sigma_\mathbf{x} + \boldsymbol{\mu}_\mathbf{x}\boldsymbol{\mu}_\mathbf{x}^\top) B_\lambda. \tag{2.4}$$

In the following sections we will also use the conditional expectation and variance of the residuals $\boldsymbol{\varepsilon}_\lambda \in \mathbb{R}^q$ with respect to $\mathbf{x}$, that we denote by

$$\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}} := \mathrm{E}_\mathbf{x}[\boldsymbol{\varepsilon}_\lambda] = \mathrm{E}_\mathbf{x}[\mathbf{y}] - B_\lambda^\top \mathbf{x} \quad \text{and} \quad \Sigma_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}} := \mathrm{Cov}_\mathbf{x}(\boldsymbol{\varepsilon}_\lambda, \boldsymbol{\varepsilon}_\lambda), \tag{2.5}$$

respectively. Recall that by the laws of total expectation and variance, these moments and conditional moments are related by

$$\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda} = \mathrm{E}\left[\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}}\right] \quad \text{and} \quad \Sigma_{\boldsymbol{\varepsilon}_\lambda} = \mathrm{E}\left[\Sigma_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}}\right] + \mathrm{Cov}\left(\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}}, \boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}}\right).$$

We emphasize that in our treatment of the ridge regression problem we do not assume, as it is customary in the literature, that there is an underlying stochastically perturbed linear functional relation between $\mathbf{x}$ and $\mathbf{y}$. The aim of the error formulas that we present later on in Section 3 is indeed the quantitative evaluation of the inaccuracy committed when the actual functional link between those two random variables is approximated by a linear model obtained via the solution of the regularized regression problem (2.1). More explicitly, we proceed by first presenting the two random variables $\mathbf{x}$ and $\mathbf{y}$ that satisfy hypothesis **(A1)** and by solving, for a fixed penalization strength λ, the ridge penalized least squares problem (2.1); as we see later on, this strategy allows for the definition of the regression residuals $\boldsymbol{\varepsilon}_\lambda$ and, when they satisfy certain homoscedasticity conditions, it can be used to characterize the statistical properties of the ridge estimator in fully singular situations ($\Sigma_{\mathbf{x}}$ is not invertible) and to evaluate the regression (also called training) and the generalization (also called testing) errors.

The different elements and developments in the paper, as well as the reach of the hypotheses that are invoked, are illustrated with four running examples. Example A contains the standard linear multivariate regression model under the Gauss-Markov hypotheses; we will show in the context of this example that the classical results in [Hoer 70] can be obtained as a particular case of those in this paper. Examples B and C consider fully nonlinear functional relations between $\mathbf{x}$ and $\mathbf{y}$ that are subjected to additive and multiplicative stochastic disturbances, respectively. A particular case of Example B will be worked out in Section 3.4 in order to illustrate some of the consequences of the error formulas in Section 3.

► **Example A. Linear multivariate regression model**. Consider a multivariate linear regression model of the form

$$\mathbf{y} = B^\top \mathbf{x} + \boldsymbol{\varepsilon}, \tag{2.6}$$

where $\mathbf{x}$ and $\mathbf{y}$ are $\mathbb{R}^p$ and $\mathbb{R}^q$-valued random variables, respectively, subjected to the hypothesis **(A1)**. The term $\boldsymbol{\varepsilon}$ is a $\mathbb{R}^q$-valued random variable with zero mean and covariance matrix $\Sigma_{\boldsymbol{\varepsilon}}$, that is, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}_q, \Sigma_{\boldsymbol{\varepsilon}})$. We place ourselves in a Gauss-Markov setup by assuming that $\boldsymbol{\varepsilon}$ and $\mathbf{x}$ are independent random variables. In these conditions it is easy to show that for any $\lambda \in \mathbb{R}^+$, the ridge regression matrix B_λ in (2.2) and the linear regression coefficient matrix B in (2.6) are related by

$$B_\lambda = (\Sigma_{\mathbf{x}} + \lambda \mathbb{I}_p)^{-1}\Sigma_{\mathbf{x}} B, \;\; \text{or, equivalently,} \;\; B_\lambda - B = -\lambda(\Sigma_{\mathbf{x}} + \lambda \mathbb{I}_p)^{-1}, \tag{2.7}$$

where we used in (2.2) that $\Sigma_{\mathbf{xy}} = \Sigma_{\mathbf{x}} B$. Since the ridge residuals are determined by the equality $\mathbf{y} = B_\lambda^\top \mathbf{x} + \boldsymbol{\varepsilon}_\lambda$ and, at the same time, $\mathbf{x}$ and $\mathbf{y}$ are related by (2.6), we have that

$$\boldsymbol{\varepsilon}_\lambda := \mathbf{y} - B_\lambda^\top \mathbf{x} = (B - B_\lambda)^\top \mathbf{x} + \boldsymbol{\varepsilon}.$$

The unconditional first and the second moments of the ridge residuals $\boldsymbol{\varepsilon}_\lambda$ can be obtained in a straightforward way out of the relations in (2.4) as follows:

$$\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda} = (B - B_\lambda)^\top \boldsymbol{\mu}_{\mathbf{x}}, \tag{2.8}$$

$$\Sigma_{\boldsymbol{\varepsilon}_\lambda} = \Sigma_{\boldsymbol{\varepsilon}} + (B - B_\lambda)^\top \Sigma_{\mathbf{x}} (B - B_\lambda), \tag{2.9}$$

where we used that $\Sigma_{\mathbf{y}} = \Sigma_{\boldsymbol{\varepsilon}} + B^\top \Sigma_{\mathbf{x}} B$ and that $\Sigma_{\mathbf{yx}} = B^\top \Sigma_{\mathrm{x}}$. The conditional counterparts of (2.8)-(2.9) can be obtained with the help of (2.5):

$$\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}} = (B - B_\lambda)^\top \mathbf{x}, \tag{2.10}$$

$$\Sigma_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}} = \mathrm{E}_{\mathbf{x}}\left[\boldsymbol{\varepsilon}_\lambda \boldsymbol{\varepsilon}_\lambda^\top\right] - \boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}} \boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}}^\top = \Sigma_{\boldsymbol{\varepsilon}}. \;\blacktriangleleft \tag{2.11}$$

▶ **Example B. Nonlinear one-dimensional model with additive stochastic disturbance term.** Consider the following data generating process for the variable y

$$y = f(\xi) + \varepsilon, \tag{2.12}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a continuous function, $\varepsilon \sim D_\varepsilon(0, \sigma_\varepsilon^2)$ represents a disturbance term and ξ is distributed with zero mean, variance σ_ξ^2, and probability density function g_ξ. The random variables ε and ξ are assumed to be independent and subjected to **(A1)**.

We now approximate the model (2.12) using a univariate ridge regression in which the p explanatory variables are the first p elements $\{\phi_i\}_{i\in\{1,\dots,p\}}$ of a given ordered generating set of the function space to which the function f in (2.12) belongs, all of them evaluated at ξ. More explicitly, consider an approximate model of the following form:

$$y = \mathbf{B}_\lambda^\top \mathbf{x} + \varepsilon_\lambda, \;\; \varepsilon_\lambda \sim D_{\varepsilon_\lambda}(\mu_{\varepsilon_\lambda}, \sigma_{\varepsilon_\lambda}^2), \tag{2.13}$$

where the vector of regressors $\mathbf{x} \in \mathbb{R}^p$ is constructed as

$$\mathbf{x} := \left(\widetilde{\xi}_1, \dots, \widetilde{\xi}_p\right)^\top, \;\; \text{with} \;\; x_i = \widetilde{\xi}_i := \frac{\xi_i - \mathrm{E}\left[\xi_i\right]}{\sigma(\xi_i)}, \;\; \xi_i = \phi_i(\xi), \;\; i \in \{1, \dots, p\}. \tag{2.14}$$

We emphasize that for each $i \in \{1, \dots, p\}$, the explanatory variable x_i is constructed out of the standardized version of the evaluation $\xi_i = \phi_i(\xi)$ of the generating function associated ϕ_i using the standard deviation:

$$\sigma(\xi_i) := (\mathrm{E}\left[\xi_i^2\right] - \mathrm{E}\left[\xi_i\right]^2)^{1/2}, \;\; \text{for each} \;\; i \in \{1, \dots, p\}. \tag{2.15}$$

We now provide all the elements associated to the approximate model (2.13). First, (2.2) determines, for any regularization strength $\lambda \in \mathbb{R}^+$, the ridge regression matrix B_λ once the central moments $\Sigma_{\mathbf{x}} \in \mathbb{S}_p$ and $\mathbf{\Sigma}_{\mathbf{x}y} \in \mathbb{R}^p$ have been computed. It is easy to see that

$$(\Sigma_{\mathbf{x}})_{i,j} = \mathrm{E}\left[\widetilde{\xi}_i \widetilde{\xi}_j\right] = \frac{1}{\sigma(\xi_i)\sigma(\xi_j)} \int_{-\infty}^{\infty} \left(\phi_i(z) - \mathrm{E}\left[\xi_i\right]\right)\left(\phi_j(z) - \mathrm{E}\left[\xi_j\right]\right) g_\xi(z) dz, \;\; i, j \in \{1, \dots, p\}, \tag{2.16}$$

$$(\mathbf{\Sigma}_{\mathbf{x}y})_i = \mathrm{E}\left[\widetilde{\xi}_i f(\xi)\right] = \frac{1}{\sigma(\xi_i)} \int_{-\infty}^{\infty} \left(\phi_i(z) - \mathrm{E}\left[\xi_i\right]\right) f(z) g_\xi(z) dz, \;\; i \in \{1, \dots, p\}, \tag{2.17}$$

with $\sigma(\xi_i)$ as in (2.15) and where g_ξ is the probability density function of ξ. We now study the defining features of the ridge regression residuals ε_λ and their moments. First, by (2.13) and (2.12), the residuals ε_λ are given by

$$\varepsilon_\lambda = f(\xi) - \mathbf{B}_\lambda^\top \mathbf{x} + \varepsilon.$$

This expression can be used to explicitly compute the unconditional and conditional first and second moments of ε_λ; indeed, by (2.4):

$$\mu_{\varepsilon_\lambda} = \mathrm{E}\left[f(\xi)\right] = \int_{-\infty}^{\infty} f(z) g_\xi(z) dz, \tag{2.18}$$

where we used that $\mathrm{E}\left[\mathbf{x}\right] = 0$ by construction in (2.14). Analogously, (2.4) yields

$$\sigma_{\varepsilon_\lambda}^2 = \sigma_y^2 - \mathbf{B}_\lambda^\top \mathbf{\Sigma}_{\mathbf{x}y} - \mathbf{\Sigma}_{\mathbf{x}y}^\top \mathbf{B}_\lambda + \mathbf{B}_\lambda^\top \Sigma_{\mathbf{x}} \mathbf{B}_\lambda - \mu_{\varepsilon_\lambda}^2, \tag{2.19}$$

where

$$\sigma_y^2 = \mathrm{E}\left[f(\xi)^2\right] + \sigma_\varepsilon^2$$

with

$$\mathrm{E}\left[f(\xi)^2\right] = \int_{-\infty}^{\infty} f(z)^2 g_\xi(z) dz. \tag{2.20}$$

Regarding the conditional moments, by (2.5) we have

$$\begin{aligned}
\mu_{\varepsilon_\lambda|\mathbf{x}} &= f(\xi) - \mathbf{B}_\lambda^\top \mathbf{x}, \\
\sigma^2_{\varepsilon_\lambda|\mathbf{x}} &= \mathrm{E}_{\mathbf{x}}\left[\varepsilon_\lambda^2\right] - \mu^2_{\varepsilon_\lambda|\mathbf{x}} \\
&= f(\xi)^2 + \sigma_\varepsilon^2 + \mathbf{B}_\lambda^\top \mathbf{x}\mathbf{x}^\top \mathbf{B}_\lambda - 2f(\xi)\mathbf{B}_\lambda^\top \mathbf{x} - f(\xi)^2 - \mathbf{B}_\lambda^\top \mathbf{x}\mathbf{x}^\top \mathbf{B}_\lambda + 2f(\xi)\mathbf{B}_\lambda^\top \mathbf{x} = \sigma_\varepsilon^2. \blacktriangleleft
\end{aligned} \tag{2.21}$$

▶ **Example C. Non-linear one-dimensional model with mutiplicative stochastic disturbance term**. Consider the following data generating process for the variable y

$$y = f(\xi)\varepsilon, \tag{2.22}$$

where $\varepsilon \sim D_\varepsilon(\mu_\varepsilon, \sigma_\varepsilon^2)$ is a disturbance term and ξ is distributed with zero mean, variance σ_ξ^2, and probability density function g_ξ. As in Examples A and B, the random variables ε and ξ are assumed to be independent. In the conditions of this example we choose to approximate the model (2.22) by the same linear regression model (2.13) with the covariates in (2.14) that are used in Example B. Notice that in this case, the relations (2.15)-(2.17) in Example B hold true. As for the ridge regression residuals ε_λ, the choice (2.22) for the data generating process implies that

$$\varepsilon_\lambda = f(\xi)\varepsilon - \mathbf{B}_\lambda^\top \mathbf{x}.$$

It is easy to verify, using the independence of ξ and ε, that the unconditional first moment of the ridge regression residuals is

$$\mu_{\varepsilon_\lambda} = \mathrm{E}\left[f(\xi)\right]\mu_\varepsilon = \mu_\varepsilon \int_{-\infty}^{\infty} f(z) g_\xi(z) dz,$$

As to the unconditional second moment, the relation (2.19) holds true with σ_y^2 determined by

$$\sigma_y^2 = \mathrm{E}\left[f(\xi)^2\varepsilon^2\right] - \mu^2_{\varepsilon_\lambda} = \mathrm{E}\left[f(\xi)^2\right](\sigma_\varepsilon^2 + \mu_\varepsilon^2) - \mu^2_{\varepsilon_\lambda} \tag{2.23}$$

and with $\mathrm{E}\left[f(\xi)^2\right]$ as in (2.20). The conditional first and second moments are determined by

$$\begin{aligned}
\mu_{\varepsilon_\lambda|\mathbf{x}} =& f(\xi)\mu_\varepsilon - \mathbf{B}_\lambda^\top \mathbf{x}, \\
\sigma^2_{\varepsilon_\lambda|\mathbf{x}} =& \mathrm{E}_{\mathbf{x}}\left[\varepsilon_\lambda^2\right] - \mu^2_{\varepsilon_\lambda|\mathbf{x}} = f(\xi)^2(\sigma_\varepsilon^2 + \mu_\varepsilon^2) - 2\mu_\varepsilon f(\xi)\mathbf{B}_\lambda^\top \mathbf{x} + \mathbf{B}_\lambda^\top \mathbf{x}\mathbf{x}^\top \mathbf{B}_\lambda \\
& - f(\xi)^2\mu_\varepsilon^2 + 2\mu_\varepsilon f(\xi)\mathbf{B}_\lambda^\top \mathbf{x} - \mathbf{B}_\lambda^\top \mathbf{x}\mathbf{x}^\top \mathbf{B}_\lambda = f(\xi)^2\sigma_\varepsilon^2,
\end{aligned} \tag{2.24}$$

respectively. ◀

2.2 The finite sample ridge estimator

We now study the finite sample properties of the ridge estimator. We start by considering two random samples of size N, that is, $\{\mathbf{x}_i\}_{i\in\{1,\dots,N\}}$ and $\{\mathbf{y}_i\}_{i\in\{1,\dots,N\}}$. These samples are constituted by N different $\mathbb{R}^p$-valued (respectively, $\mathbb{R}^q$-valued), mutually independent, and identically distributed random variables $\mathbf{x}_i \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ (respectively, $\mathbf{y}_i \in L^2(\Omega, \mathcal{F}, \mathbb{P})$), $i \in \{1, \dots, N\}$. Notice that we require that all these random variables satisfy the hypothesis **(A1)**. We now horizontally concatenate the elements of each of these two random samples and construct the corresponding random matrices $X \in \mathbb{M}_{p,N}$ and $Y \in \mathbb{M}_{q,N}$, respectively, that is $X := (\mathbf{x}_1||\mathbf{x}_2||\dots||\mathbf{x}_N)$ and $Y := (\mathbf{y}_1||\mathbf{y}_2||\dots||\mathbf{y}_N)$.

The finite sample estimator $\widehat{B}_\lambda$ of the ridge regression matrix B_λ is constructed by plugging in (2.2) the natural estimators for the second order moments $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{xy}}$, that is,

$$\widehat{\Sigma}_{\mathbf{x}} := \frac{1}{N} X A_N X^\top, \tag{2.25}$$

$$\widehat{\Sigma}_{\mathbf{xy}} := \frac{1}{N} X A_N Y^\top, \tag{2.26}$$

with

$$A_N := \mathbb{I}_N - \frac{1}{N} \mathbf{i}_N \mathbf{i}_N^\top. \tag{2.27}$$

The expressions (2.25)-(2.26) substituted in (2.2) yield

$$\widehat{B}_\lambda := (X A_N X^\top + \lambda N \mathbb{I}_p)^{-1} X A_N Y^\top, \tag{2.28}$$

which can be subsequently evaluated for any given realization $\mathcal{X}$ and $\mathcal{Y}$ of the random matrices X and Y.

▶ **Example A (continued)**. We derive the particular expression of the ridge regression matrix estimator in the conditions implied by (2.6). We first recall that the ordinary least squares estimator $\widehat{B}$ of B is given by (2.28) with zero regularization strength, that is $\widehat{B} = \widehat{B}_0$. More explicitly,

$$\widehat{B} := \widehat{B}_0 = \widehat{\Sigma}_{\mathbf{x}}^{-1} \widehat{\Sigma}_{\mathbf{xy}} = (X A_N X^\top)^{-1} X A_N Y^\top. \tag{2.29}$$

It is important to underline that the ordinary least squares estimator (2.29) is available only when the sample covariance matrix $\widehat{\Sigma}_{\mathbf{x}}$ in (2.25) is invertible which, as we already pointed out in the introduction, is one of its main deficiencies.

The following lemma, whose proof is in Appendix 4.1, provides the relation between the ridge regression $\widehat{B}_\lambda$ and the ordinary least squares estimator $\widehat{B}$. We obviously restrict ourselves to the case in which the evaluation of (2.29) is feasible, that is, when $\widehat{\Sigma}_{\mathbf{x}}$ is invertible. The relations that we now state are not new and are well established in the literature [Hoer 70]. We provide them for future reference and in order to complete Example A.

Lemma 2.1 *In the conditions of Example A the following statements hold true:*

(i) *The relation between the finite sample ridge estimator $\widehat{B}_\lambda$ of B_λ and the finite sample ordinary least squares estimator $\widehat{B}$ of B is given by*

$$\widehat{B}_\lambda = Z_\lambda \widehat{B} \tag{2.30}$$

where

$$Z_\lambda := R_\lambda X A_N X^\top, \tag{2.31}$$

$$R_\lambda := (X A_N X^\top + \lambda N \mathbb{I}_p)^{-1} \tag{2.32}$$

with A_N as in (2.27). Expression (2.30) is the finite sample based analogue of the first relation in (2.7).

(ii) *The multiplier Z_λ in (2.31) can be equivalently expressed as*

$$Z_\lambda = \mathbb{I}_p - \lambda N R_\lambda, \tag{2.33}$$

or

$$Z_\lambda = \left(\mathbb{I}_p + \lambda N (X A_N X^\top)^{-1}\right)^{-1}, \tag{2.34}$$

with R_λ in (2.33) defined as in (2.32).

Remark 2.2 The relations $Z_\lambda = \mathbb{I}_p - \lambda N R_\lambda$ in (2.33) and $Z_\lambda = R_\lambda X A_N X^\top$ in (2.31) hold in a general context beyond the specific conditions of Example A and even if $X X^\top$ is singular or ill-conditioned. They are used later on in the paper. ◀

2.3 The sample ridge residuals and the homoscedasticity hypothesis

Consider now the $\mathbb{R}^q$-valued random residuals $\{\boldsymbol{\varepsilon}_{\lambda,i}\}_{i\in\{1,\ldots,N\}}$ obtained using the random samples $\{\mathbf{x}_i\}_{i\in\{1,\ldots,N\}}$ and $\{\mathbf{y}_i\}_{i\in\{1,\ldots,N\}}$, for each $i\in\{1,\ldots,N\}$ via the assignment $\boldsymbol{\varepsilon}_{\lambda,i}:=\mathbf{y}_i-B_\lambda\mathbf{x}_i$. Similarly to the procedure used in the construction of the random matrices X and Y we horizontally concatenate the entries $\boldsymbol{\varepsilon}_{\lambda,i}\in\mathbb{R}^q$, $i\in\{1,\ldots,N\}$, and we obtain the random matrix $E_\lambda\in\mathbb{M}_{q,N}$ given by $E_\lambda:=(\boldsymbol{\varepsilon}_{\lambda,1}||\ldots||\boldsymbol{\varepsilon}_{\lambda,N})$ and that satisfies the relation $E_\lambda=Y-B_\lambda^\top X$. We denote the unconditional first moment of E_λ by $M_{E_\lambda}\in\mathbb{M}_{q,N}$ and note that

$$M_{E_\lambda}:=\mathrm{E}\left[E_\lambda\right]=\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda}\mathbf{i}_N^\top, \tag{2.35}$$

where $\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda}$ is given by (2.4). We analogously specify the conditional expectation of E_λ given a random matrix X as

$$M_{E_\lambda|X}:=\mathrm{E}_X\left[E_\lambda\right]=(\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_{\lambda,1}|\mathbf{x}_1}||\ldots||\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_{\lambda,N}|\mathbf{x}_N}), \tag{2.36}$$

where each $\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_{\lambda,i}|\mathbf{x}_i}$, $i\in\{1,\ldots,N\}$ is given by (2.5).

The relations that we just provided for the conditional and unconditional first order moments of the residuals random matrix E_λ can be used to establish the expressions of their corresponding second-order counterparts. In general, one needs to use (2.4) and (2.5) in order to determine the second order moments of the residuals $\{\boldsymbol{\varepsilon}_{\lambda,i}\}_{i\in\{1,\ldots,N\}}$ and this implies that, in principle, there exist N different conditional covariance matrices $\Sigma_{\boldsymbol{\varepsilon}_{\lambda,i}|\mathbf{x}_i}$, $i\in\{1,\ldots,N\}$, defined for each $\boldsymbol{\varepsilon}_{\lambda,i}$ and given some corresponding random sample element $\mathbf{x}_i$. We consider in what follows a particular situation in which all these second order conditional moments are constant and equal for all $i\in\{1,\ldots,N\}$. As we discuss later on in the text, this conditional homoscedasticity property is natural and is satisfied in many standard situations. In particular, we will show that linear and nonlinear models with additive noise (examples A and B) have conditionally homoscedastic residuals unlike the case with multiplicative noise (Example C) that in general does not satisfy this property.

Lemma 2.3 *In the conditions that were just introduced consider the ridge residuals* $\{\boldsymbol{\varepsilon}_{\lambda,i}\}_{i\in\{1,\ldots,N\}}\sim D(\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda},\Sigma_{\boldsymbol{\varepsilon}_\lambda})$ *and suppose that they are homoscedastic, that is, we assume that the conditional covariance matrices* $\Sigma_{\boldsymbol{\varepsilon}_{\lambda,i}|\mathbf{x}_i}$, $i\in\{1,\ldots,N\}$, *in* (2.11) *are constant and equal to some matrix* $\Sigma^\lambda_{\boldsymbol{\varepsilon}|\mathbf{x}}\in\mathbb{S}_q^+$. *Then the following relations hold true:*

(i) *For the conditional second-order raw moments:*

$$\mathrm{E}_X\left[E_\lambda^\top E_\lambda\right]-M_{E_\lambda|X}^\top M_{E_\lambda|X}=\mathrm{trace}(\Sigma^\lambda_{\boldsymbol{\varepsilon}|\mathbf{x}})\mathbb{I}_N, \tag{2.37}$$

$$\mathrm{E}_X\left[E_\lambda E_\lambda^\top\right]-M_{E_\lambda|X}M_{E_\lambda|X}^\top=N\Sigma^\lambda_{\boldsymbol{\varepsilon}|\mathbf{x}}, \tag{2.38}$$

(ii) *For the unconditional second-order raw moments:*

$$\mathrm{E}\left[E_\lambda^\top E_\lambda\right]-M_{E_\lambda}^\top M_{E_\lambda}=\mathrm{trace}(\Sigma_{\boldsymbol{\varepsilon}_\lambda})\mathbb{I}_N, \tag{2.39}$$

$$\mathrm{E}\left[E_\lambda E_\lambda^\top\right]-M_{E_\lambda}M_{E_\lambda}^\top=N\Sigma_{\boldsymbol{\varepsilon}_\lambda}. \tag{2.40}$$

In these relations, the mean M_{E_λ} *and conditional mean* $M_{E_\lambda|X}$ *matrices are given by* (2.35) *and* (2.36)*, respectively.*

The proof of this lemma is not provided since it is straightforward. Lemma 2.3 has an important implication under an additional normality assumption for the ridge residuals that we state in the following corollary.

Corollary 2.4 *Suppose that the hypotheses of Lemma 2.3 are satisfied and that, additionally, the conditional residuals are independent and normally distributed, that is,* $\boldsymbol{\varepsilon}_{\lambda,i}|\mathbf{x}_i \sim \mathrm{IN}(\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_{\lambda,i}|\mathbf{x}_i}, \Sigma^{\lambda}_{\boldsymbol{\varepsilon}|\mathbf{x}})$*, for each* $i \in \{1, \ldots, N\}$*. Then, the random matrix of residuals* E_λ *conditional on* X *are distributed according to a matrix normal distribution with conditional mean* $M_{E_\lambda|X}$*, and* $\Sigma^{\lambda}_{\boldsymbol{\varepsilon}|\mathbf{x}}$ *and* $\mathbb{I}_N$ *as scale matrices, that is,*

$$E_\lambda|X \sim \mathrm{MN}(M_{E_\lambda|X}, \Sigma^{\lambda}_{\boldsymbol{\varepsilon}|\mathbf{x}}, \mathbb{I}_N). \tag{2.41}$$

We now go back to the examples that we considered in Section 2.1 and verify if the corresponding ridge residuals satisfy the homoscedasticity property. Notice that since Example A can be seen as the particular case of Example B, we state the results exclusively for Examples B and C.

▶ **Examples B and C (continued).** First, in the case of the general model with an additive stochastic disturbance term introduced in Example B, the homoscedasticity follows from the expression in (2.21) of the conditional variance of the ridge residuals. This observation has important implications and we hence frame it in the following proposition.

Proposition 2.5 *Let* X *and* Y *be random sample matrices generated by the regression model (2.13) in Example B that assumes that* y *is an additive stochastic perturbation of some continuous function of some random variable* ξ*. If, additionally, the stochastic perturbation and the random variable* ξ *are independent, then the ridge regression residuals of (2.13) are always homoscedastic.*

We emphasize that, as a corollary of this proposition, it follows that the ridge regression residuals of the linear regression model in Example A are also homoscedastic. This can be directly obtained out of the relation (2.11).

Regarding the model with a multiplicative stochastic perturbation introduced in Example C, relation (2.24) shows that, in this case, the ridge residuals are not in general homoscedastic. ◀

2.4 The properties of the ridge regression estimator

The results presented in Lemma 2.3 and its Corollary 2.4 allow us to take an important step. Indeed, we now see how, in the presence of the conditional normality and homoscedasticity hypotheses on the ridge residuals that we introduced above, we can spell out the properties of the ridge regression matrix estimator $\widehat{B}_\lambda$ introduced in (2.28), conditional on the covariates sample X. As we already pointed out in the introduction, our results generalize the classical ones in [Hoer 70] by formulating the properties of $\widehat{B}_\lambda$ directly in terms of the solution B_λ of the ridge regularized regression problem and without assuming the existence of a least squares matrix coefficient B or, equivalently, the regularity of the covariance matrix $\Sigma_{\mathbf{x}}$ of the covariates, that in many applications is just not available.

Theorem 2.6 *Consider the random samples* $\{\mathbf{x}_i\}_{i\in\{1,\ldots,N\}}$ *and* $\{\mathbf{y}_i\}_{i\in\{1,\ldots,N\}}$ *of length* N *that consist of the random variables* $\mathbf{x}_i \in \mathbb{R}^p$*,* $\mathbf{y}_i \in \mathbb{R}^q$*,* $i \in \{1, \ldots, N\}$*, respectively, that satisfy the hypothesis* **(A1)***. Let* $X \in \mathbb{R}_{p,N}$ *and* $Y \in \mathbb{M}_{q,N}$ *be the random matrices constructed by horizontal concatenation of all the corresponding entries of* $\{\mathbf{x}_i\}_{i\in\{1,\ldots,N\}}$ *and* $\{\mathbf{y}_i\}_{i\in\{1,\ldots,N\}}$*, respectively. Additionally, let* $\{\boldsymbol{\varepsilon}_{\lambda,i}\}_{i\in\{1,\ldots,N\}}$ *be the associated residuals of the ridge regression with regularization strength* $\lambda \in \mathbb{R}^+$ *defined by* $\boldsymbol{\varepsilon}_{\lambda,i} := \mathbf{y}_i - B_\lambda \mathbf{x}_i$ *and that we assume to be conditionally homoscedastic. Let* E_λ *be the random matrix obtained by horizontal concatenation of the ridge residuals conditioned on* X *and suppose that it is normally distributed as in (2.41) with conditional mean* $M_{E_\lambda|X}$ *and some constant scale matrix* $\Sigma^{\lambda}_{\boldsymbol{\varepsilon}|\mathbf{x}}$*. Finally, let* $\widehat{B}_\lambda$ *be the ridge estimator of the ridge regression matrix* B_λ *based on the random sample matrices* X *and* Y*. Then*

$$(\widehat{B}_\lambda - B_\lambda)|X \sim \mathrm{MN}(-\lambda N R_\lambda B_\lambda + R_\lambda X A_N M^{\top}_{E_\lambda|X}, Z_\lambda R_\lambda, \Sigma^{\lambda}_{\boldsymbol{\varepsilon}|\mathbf{x}}), \tag{2.42}$$

where the conditional mean $M_{E_\lambda|X}$ is given in (2.36), the conditional covariance matrix $\Sigma^{\lambda}_{\boldsymbol{\varepsilon}|\mathbf{x}}$ is provided in point **(i)** *of Lemma 2.3 and, additionally,*

$$R_\lambda := (XA_NX^\top + \lambda N\mathbb{I}_p)^{-1}, \tag{2.43}$$

$$Z_\lambda := R_\lambda XA_NX^\top = \mathbb{I}_p - \lambda NR_\lambda \tag{2.44}$$

with A_N as in (2.27).

▶ **Example A (continued)**. We now show how the classical results on the ridge estimator properties in Section 4a of [Hoer 70] that assume the existence of a least squares estimate follow as a corollary of Theorem 2.6. The proof of the following corollary is provided in Appendix 4.3.

Corollary 2.7 (Section 4a in [Hoer 70]) *Consider the $\mathbb{R}^q$-valued linear regression $Y = B^\top X + E$ obtained out of random samples of length N that satisfy the relation (2.6) considered in Example A. Given that, by hypothesis, the error term satisfies $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}_q, \Sigma_{\boldsymbol{\varepsilon}})$, it is easy to see that the residual random matrix $E \in \mathbb{M}_{q,N}$ obtained by horizontal concatenation, is matrix normal distributed as $E \sim \mathrm{MN}(\mathbb{O}_q, \Sigma_{\boldsymbol{\varepsilon}}, \mathbb{I}_N)$. The following statements hold:*

(i) *The ridge regression matrix estimator is distributed as*

$$(\widehat{B}_\lambda - B)|X \sim \mathrm{MN}(-\lambda NR_\lambda B, Z_\lambda R_\lambda, \Sigma_{\boldsymbol{\varepsilon}}), \tag{2.45}$$

where R_λ and Z_λ are defined in (2.43) and in (2.44), respectively.

(ii) *If the matrix $\widehat{\Sigma}_{\mathbf{x}}$ defined in (2.25) is invertible, then (2.45) can be rewritten as*

$$(\widehat{B}_\lambda - B)|X \sim \mathrm{MN}(-\lambda NR_\lambda B, Z_\lambda (XA_NX^\top)^{-1} Z_\lambda, \Sigma_{\boldsymbol{\varepsilon}}), \tag{2.46}$$

with Z_λ defined as in (2.44) or, equivalently, as in (2.34). ◀

3 Evaluation of the ridge regression and generalization errors

In this section we use the properties of the ridge estimator that we spelled out in Theorem 2.6 in order to write down explicit expressions that allow for the evaluation of the regression (also called training) and the generalization (also called testing) errors committed by a regularized regression model whose coefficient $\widehat{B}_\lambda$ has been estimated using a finite sample of a given size. The regression or training error is the one committed by the regression model when evaluated with the sample that has been used to obtain $\widehat{B}_\lambda$; for the generalization or testing error, we keep $\widehat{B}_\lambda$ and we evaluate the error committed by the corresponding regression model using another sample that may have different size or even different statistical properties. As we will see, in both cases our error formulas incorporate the error committed at the time of parameter estimation.

These two errors are of much importance in specific applications involving a finite sample framework since their relative values give indications about the pertinence of various modeling choices. Indeed, a typical behavior in a regression model is that an increase in the number of covariates makes smaller the training error. The testing error follows the same trend up to a point in which it starts increasing; that turning point in the testing error has to do with the appearance of overfitting in the training, that is, the overabundance of regressors is fitting not only the underlying deterministic model but also the noise that is perturbing it. As we will see in the example that we work out in Section 3.4, the formulas presented in this section can be used to avoid this phenomenon at the time of deciding on the model architecture.

All along this section we use the same hypotheses and notation as in Theorem 2.6 unless explicitly stated otherwise.

3.1 The characteristic errors of the regression model

We define the characteristic error as the one committed by the regression model without taking into account estimation errors, that is, the characteristic error is evaluated assuming that the model parameters are known. We separately address two instances of the characteristic error. First, we consider the standard unconditional characteristic error corresponding to the ridge model with a given regularization strength $\lambda \in \mathbb{R}^+$. This error corresponds to the expected regression error for arbitrary realizations of the explanatory and dependent variables $\mathbf{x}$ and $\mathbf{y}$, respectively; it is sometimes referred to in the literature as the irreducible error (see for example [Hast 13]). Second, we assume that the dependent variables $\mathbf{x}$ are fixed and we compute the characteristic error conditional on those values.

Characteristic (irreducible) error. It is given by:

$$\mathrm{MSE}^{\lambda}_{\mathrm{char}} := \mathrm{E}\left[(\mathbf{y} - B_\lambda \mathbf{x})^\top (\mathbf{y} - B_\lambda \mathbf{x})\right] = \mathrm{E}\left[\boldsymbol{\varepsilon}_\lambda^\top \boldsymbol{\varepsilon}_\lambda\right] = \mathrm{trace}(\Sigma_{\boldsymbol{\varepsilon}_\lambda} + \boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda}\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda}^\top), \tag{3.1}$$

where $\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda}$ and $\Sigma_{\boldsymbol{\varepsilon}_\lambda}$ are given in (2.4).

Conditional characteristic error. It is the characteristic error associated to the ridge regression model conditional on the covariates value $\mathbf{x}$:

$$\mathrm{MSE}^{\lambda}_{\mathrm{char}|\mathbf{x}} := \mathrm{E}_{\mathbf{x}}\left[(\mathbf{y} - B_\lambda \mathbf{x})^\top (\mathbf{y} - B_\lambda \mathbf{x})\right] = \mathrm{E}_{\mathbf{x}}\left[\boldsymbol{\varepsilon}_\lambda^\top \boldsymbol{\varepsilon}_\lambda\right] = \mathrm{trace}(\Sigma_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}} + \boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}}\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}}^\top), \tag{3.2}$$

with $\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}}$ and $\Sigma_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}}$ as in (2.5).

We now evaluate these errors for the examples that we consider along the paper.

▶ **Example A (continued)**. The expressions for the characteristic and the conditional characteristic errors in (3.1) and in (3.2) for the standard regression model in (2.6) can be made explicit by using the particular form of the covariance matrix $\Sigma_{\boldsymbol{\varepsilon}_\lambda}$ provided in (2.9) and of the corresponding conditional covariance $\Sigma_{\boldsymbol{\varepsilon}_\lambda|\mathbf{x}}$ given in (2.11). Indeed:

$$\mathrm{MSE}^{\lambda}_{\mathrm{char}} = \mathrm{trace}(\Sigma_{\boldsymbol{\varepsilon}} + (B - B_\lambda)^\top(\Sigma_{\mathbf{x}} + \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}^\top)(B - B_\lambda)) = \mathrm{trace}(\Sigma_{\boldsymbol{\varepsilon}} + (B - B_\lambda)^\top \mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right](B - B_\lambda)), \tag{3.3}$$

and for the conditional characteristic error:

$$\mathrm{MSE}^{\lambda}_{\mathrm{char}|\mathbf{x}} = \mathrm{trace}(\Sigma_{\boldsymbol{\varepsilon}} + (B - B_\lambda)^\top \mathbf{x}\mathbf{x}^\top (B - B_\lambda)). \tag{3.4}$$

From these expressions it is obvious that the relative magnitudes of these characteristic and conditional characteristic errors depend on the specific value of the random variable $\mathbf{x} \in \mathbb{R}^p$ used in the conditioning. ◀

▶ **Example B (continued)**. In this particular instance the characteristic error is:

$$\mathrm{MSE}^{\lambda}_{\mathrm{char}} = \sigma_\varepsilon^2 + \mathrm{E}\left[f(\xi)^2\right] + \mathbf{B}_\lambda^\top \Sigma_{\mathbf{x}} \mathbf{B}_\lambda - 2\mathbf{B}_\lambda^\top \boldsymbol{\Sigma}_{\mathbf{x}y}, \tag{3.5}$$

where $\mathrm{E}\left[f(\xi)^2\right]$, $\boldsymbol{\Sigma}_{\mathbf{x}y}$, and $\boldsymbol{\Sigma}_{\mathbf{x}}$ are given in (2.20), (2.17), and (2.16), respectively. Additionally, the expression of the conditional characteristic error is

$$\mathrm{MSE}^{\lambda}_{\mathrm{char}|\mathbf{x}} = \sigma_\varepsilon^2 + f(\xi)^2 + \mathbf{B}_\lambda^\top \mathbf{x}\mathbf{x}^\top \mathbf{B}_\lambda - 2f(\xi)\mathbf{B}_\lambda^\top \mathbf{x}. \tag{3.6}$$

It can be easily seen from (3.5) and (3.6) that as in Example A, any ordering between these two errors is possible depending on the value of the random variable $\mathbf{x} \in \mathbb{R}^p$. ◀

▶ **Example C (continued)**. The characteristic error is:

$$\mathrm{MSE}^{\lambda}_{\mathrm{char}} = \sigma_\varepsilon^2 \mathrm{E}\left[f(\xi)^2\right] + \mathbf{B}_\lambda^\top \Sigma_{\mathbf{x}} \mathbf{B}_\lambda - 2\mathbf{B}_\lambda^\top \boldsymbol{\Sigma}_{\mathbf{x}y}, \tag{3.7}$$

where $\mathrm{E}\left[f(\xi)^2\right]$, $\mathbf{\Sigma}_{\mathbf{x}y}$, $\mathbf{\Sigma}_{\mathbf{x}}$, and $\mu_{\varepsilon_\lambda}$ are given in (2.20), (2.17), (2.16), and (2.18), respectively. The expression of the conditional characteristic error is provided by:

$$\mathrm{MSE}^{\lambda}_{\mathrm{char}|\mathbf{x}} = \sigma^2_{\varepsilon} f(\xi)^2 + f(\xi)^2 \mu^2_{\varepsilon} - 2\mu_{\varepsilon} f(\xi) \mathbf{B}^{\top}_{\lambda} \mathbf{x} + \mathbf{B}^{\top}_{\lambda} \mathbf{x}\mathbf{x}^{\top} \mathbf{B}_{\lambda}. \tag{3.8}$$

As in Examples A and B, any ordering between these two errors is possible depending on the value of the random variable $\mathbf{x} \in \mathbb{R}^p$. ◀

3.2 The conditional total training error

In this section we provide an explicit expression for the total error committed when using a multivariate ridge regularized regression model with a coefficient matrix that has been estimated using a finite sample. In that situation there are two sources of error: first, the characteristic error associated to the regression model and second, the estimation error on the ridge regression matrix that appears due to the finiteness of the estimation sample. We first focus on the total regression or training error in which the error is evaluated using the same sample that has been earlier exploited to obtain an estimate of the the ridge regression matrix. The expression for the error that we provide is conditional on the covariates sample used at the time of estimation.

Consider the random matrices $X \in \mathbb{M}_{p,N}$, $Y \in \mathbb{M}_{q,N}$ constructed according to the prescriptions in Section 2.2 and let $\widehat{B}_\lambda \in \mathbb{M}_{p,q}$ be the estimator of B_λ based on X and Y that we introduced in (2.28). The conditional total training error is defined as

$$\mathrm{MSE}^{\lambda}_{\mathrm{training}|X} := \frac{1}{N} \mathrm{trace}\left(\mathrm{E}_X \left[(Y - \widehat{B}^{\top}_{\lambda} X)^{\top} (Y - \widehat{B}^{\top}_{\lambda} X) \right] \right). \tag{3.9}$$

The following theorem, whose proof is given in Appendix 4.4, provides an explicit expression for the conditional total training error.

Theorem 3.1 *Suppose that we are in the hypotheses of Theorem 2.6, that is, we consider a ridge regression between square summable explanatory and dependent variables that exhibits conditionally normal and homoscedastic residuals. Then, the total training error conditional on a random sample X of length N of the covariates is:*

$$\begin{aligned}
&\mathrm{MSE}^{\lambda}{}_{\mathrm{training}|X} = \\
&\quad = \mathrm{trace}(\Sigma^{\lambda}_{\boldsymbol{\epsilon}|\mathbf{x}}) + \frac{1}{N} \Big\{ \mathrm{trace}(\Sigma^{\lambda}_{\boldsymbol{\epsilon}|\mathbf{x}}) \mathrm{trace}\left(Z_\lambda (R_\lambda X X^{\top} - 2\mathbb{I}_p) \right) + \mathrm{trace}(M_{(\widehat{B}_\lambda - B_\lambda)} M^{\top}_{(\widehat{B}_\lambda - B_\lambda)} X X^{\top}) \\
&\qquad - 2\ \mathrm{trace}\Big((X^{\top} R_\lambda X A_N - \frac{1}{2}\mathbb{I}_N) M^{\top}_{E_\lambda|X} M_{E_\lambda|X} + (X^{\top} R_\lambda X A_N - \mathbb{I}_N) X^{\top} B_\lambda \boldsymbol{\mu}_{E_\lambda|X} \Big) \Big\},
\end{aligned} \tag{3.10}$$

or, equivalently,

$$\begin{aligned}
&\mathrm{MSE}^{\lambda}{}_{\mathrm{training}|X} = \\
&\quad = \mathrm{trace}(\Sigma^{\lambda}_{\boldsymbol{\epsilon}|\mathbf{x}}) + \frac{1}{N} \Big\{ \mathrm{trace}(\Sigma^{\lambda}_{\boldsymbol{\epsilon}|\mathbf{x}}) \mathrm{trace}\left(Z_\lambda (R_\lambda X X^{\top} - 2\mathbb{I}_p) \right) + \mathrm{trace}\left(M_{(\widehat{B}_\lambda - B_\lambda)} M^{\top}_{(\widehat{B}_\lambda - B_\lambda)} X X^{\top} \right) \\
&\qquad - 2\ \mathrm{trace}((X^{\top} R_\lambda X A_N - \frac{1}{2}\mathbb{I}_N) M^{\top}_{E_\lambda|X} M_{E_\lambda|X}) \Big\} + 2\lambda \mathrm{trace}(X^{\top} R_\lambda B_\lambda M_{E_\lambda|X}),
\end{aligned} \tag{3.11}$$

where

$$M_{(\widehat{B}_\lambda - B_\lambda)} := -\lambda N R_\lambda B_\lambda + R_\lambda X A_N M^{\top}_{E_\lambda|X} \tag{3.12}$$

is the bias of the estimator $\widehat{B}_\lambda$ of B_λ in (2.42) and where

$$Z_\lambda = R_\lambda X A_N X^{\top} = \mathbb{I}_p - \lambda N R_\lambda, \tag{3.13}$$

$$R_\lambda = \left(X A_N X^{\top} + \lambda N \mathbb{I}_p \right)^{-1}. \tag{3.14}$$

▶ **Example A (continued).** In the conditions of the linear regression model (2.6) of Example A, the conditional total training error in (3.9) can be made more explicit by using the expressions for the conditional covariance matrix $\Sigma^{\lambda}_{\boldsymbol{\varepsilon}|\mathbf{x}}$ in (2.11) that shows that the homoscedasticity hypothesis is satisfied in this setup. Moreover, as we saw in Corollary 2.7, the properties of the estimator $\widehat{B}_\lambda$ can be formulated in this case in terms of the matrix B. The proof of the following result is provided in Appendix 4.5.

Corollary 3.2 *In the conditions of Example A, the conditional total training error* (3.9) *is given by the expression:*

$$\mathrm{MSE}^{\lambda}_{\mathrm{training}|X} = \mathrm{trace}(\Sigma_{\boldsymbol{\varepsilon}}) + \frac{1}{N}\mathrm{trace}(\Sigma_{\boldsymbol{\varepsilon}})\mathrm{trace}\left(Z_\lambda(R_\lambda XX^\top - 2\mathbb{I}_p)\right) + \lambda^2 N \mathrm{trace}(R_\lambda BB^\top R_\lambda XX^\top), \tag{3.15}$$

where Z_λ is defined as in Corollary 2.7.

Expression (3.15) has been formulated for the first time in Proposition 3.3 of [Grig 16] in a machine learning context.◀

3.3 The conditional total testing or generalization error

As we already pointed out in the introduction to this section, the minimization of the training error in the construction of a model does not suffice to guarantee its good out-of-sample performance due to the possible appearance of overfitting phenomena. This makes necessary the evaluation of the so called testing or generalization error that measures the performance of a given model that has been trained on a finite sample that is independent and may even have different statistical properties from the one that is used for the testing. In specific machine learning and forecasting applications it is mainly the testing error that needs to be used in order to decide on questions related to model architecture.

We proceed in the following way in order to carry this out: we assume that the training and testing are carried out on two different sets of samples; those labeled with a one (respectively, two) are the training (respectively, testing) samples. More specifically, consider two pairs of random samples of sizes N_1 and N_2. The first pair $\left\{\mathbf{x}_i^{(1)}\right\}$, $\left\{\mathbf{y}_i^{(1)}\right\}$, with elements $\mathbf{x}_i^{(1)} \in \mathbb{R}^p$, $\mathbf{y}_i^{(1)} \in \mathbb{R}^q$, $i \in \{1, \ldots, N_1\}$, will be used for training and the second one, denoted by $\left\{\mathbf{x}_j^{(2)}\right\}$, $\left\{\mathbf{y}_j^{(2)}\right\}$ with $\mathbf{x}_j^{(2)} \in \mathbb{R}^p$, $\mathbf{y}_j^{(2)} \in \mathbb{R}^q$, $j \in \{1, \ldots, N_2\}$, will be used for testing. Suppose that all the elements in these random samples satisfy hypothesis **(A1)** and hence have finite first and second order moments. Additionally, assume that ridge residuals $\left\{\boldsymbol{\varepsilon}_{\lambda,i}^{(1)}\right\}_{i\in\{1,\ldots,N_1\}}$ associated to the training pair are conditionally normal and homoscedastic in the sense of Lemma 2.3, that is, their corresponding conditional covariance matrices are constant and equal to some matrix $\Sigma^{\lambda,(1)}_{\boldsymbol{\varepsilon}|\mathbf{x}} \in \mathbb{S}_q$. This hypothesis can be stated as $\boldsymbol{\varepsilon}_{\lambda,i}^{(1)}|\mathbf{x}_i^{(1)} \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\varepsilon}_{\lambda,i}^{(1)}|\mathbf{x}_i^{(1)}}, \Sigma^{\lambda,(1)}_{\boldsymbol{\varepsilon}|\mathbf{x}})$ for each $i \in \{1, \ldots, N_1\}$.

Consider now the random matrices $X_1 \in \mathbb{M}_{p,N_1}$, $Y_1 \in \mathbb{M}_{q,N_1}$, $X_2 \in \mathbb{M}_{p,N_2}$, $Y_2 \in \mathbb{M}_{q,N_2}$ obtained by horizontal concatenation of the elements in the random samples $\left\{\mathbf{x}_i^{(1)}\right\}_{i\in\{1,\ldots,N_1\}}$, $\left\{\mathbf{y}_i^{(1)}\right\}_{i\in\{1,\ldots,N_1\}}$, $\left\{\mathbf{x}_i^{(2)}\right\}_{i\in\{1,\ldots,N_2\}}$, and $\left\{\mathbf{y}_i^{(2)}\right\}_{i\in\{1,\ldots,N_2\}}$ respectively. We call (X_1, Y_1) and (X_2, Y_2) the training and the testing samples, respectively.

Let now $\widehat{B}_\lambda$ be the ridge regression matrix estimator of B_λ constructed using a training sample (X_1, Y_1). In these conditions we define the conditional total testing ridge error as:

$$\mathrm{MSE}^{\lambda}_{\mathrm{testing}|X_1} = \frac{1}{N_2}\mathrm{trace}\left(\mathrm{E}_{X_1}\left[\left(Y_2 - \widehat{B}_\lambda^\top X_2\right)^\top \left(Y_2 - \widehat{B}_\lambda^\top X_2\right)\right]\right). \tag{3.16}$$

This expression measures the mean square error committed when training is carried out with a fixed explanatory variables training sample X_1 of length N_1 and testing is implemented with any other testing sample (X_2, Y_2) of length N_2. On other words, we fix X_1, and we average the square errors committed when varying Y_1 and the testing samples (X_2, Y_2). The proof of the following result is provided in Appendix 4.6.

Theorem 3.3 *Consider two pairs of random samples* $\left\{\mathbf{x}_i^{(1)}\right\}$, $\left\{\mathbf{y}_i^{(1)}\right\}$ *and* $\left\{\mathbf{x}_j^{(2)}\right\}$, $\left\{\mathbf{y}_j^{(2)}\right\}$, *with* $\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)} \in \mathbb{R}^p$, $\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)} \in \mathbb{R}^q$, *of sizes* N_1 *and* N_2, *respectively. The first pair will be used for training and the second one for testing. Suppose that all the elements in these random samples satisfy hypothesis* **(A1)** *and that the ridge residuals* $\left\{\boldsymbol{\varepsilon}_{\lambda,i}^{(1)}\right\}_{i\in\{1,\dots,N_1\}}$ *associated to the training pair are conditionally normal and homoscedastic in the sense of Lemma 2.3, that is, their corresponding conditional covariance matrices are constant and equal to some matrix* $\Sigma_{\boldsymbol{\varepsilon}|\mathbf{x}}^{\lambda,(1)} \in \mathbb{S}_q$. *Consider now the random matrices* $X_1 \in \mathbb{M}_{p,N_1}$ *and* $Y_1 \in \mathbb{M}_{q,N_1}$ *obtained by horizontal concatenation of the elements in the training samples. Under these hypotheses, the conditional total testing error introduced in (3.16) can be written as:*

$$\begin{aligned}\mathrm{MSE}_{\mathrm{testing}|X_1}^{\lambda} = {} & \operatorname{trace}\left(\Sigma_{\mathbf{y}}^{(2)} + \boldsymbol{\mu}_{\mathbf{y}}^{(2)}\boldsymbol{\mu}_{\mathbf{y}}^{(2)\top}\right) - 2\operatorname{trace}\left(\left(\Sigma_{\mathbf{xy}}^{(2)} + \boldsymbol{\mu}_{\mathbf{x}}^{(2)}\boldsymbol{\mu}_{\mathbf{y}}^{(2)\top}\right) M_{\widehat{B}_\lambda}^{\top}\right) \\ & + \operatorname{trace}\left[\left(\Sigma_{\mathbf{x}}^{(2)} + \boldsymbol{\mu}_{\mathbf{x}}^{(2)}\boldsymbol{\mu}_{\mathbf{x}}^{(2)\top}\right)\left(\operatorname{trace}\left(\Sigma_{\boldsymbol{\varepsilon}|\mathbf{x}}^{\lambda,(1)}\right) Z_\lambda R_\lambda + M_{\widehat{B}_\lambda} M_{\widehat{B}_\lambda}^{\top}\right)\right], \end{aligned} \tag{3.17}$$

where

$$Z_\lambda := R_\lambda X_1 A_{N_1} X_1^{\top}, \tag{3.18}$$

$$R_\lambda := \left(X_1 A_{N_1} X_1^{\top} + \lambda N_1 \mathbb{I}_p\right)^{-1}, \tag{3.19}$$

$$M_{\widehat{B}_\lambda} := B_\lambda - \lambda N_1 R_\lambda B_\lambda + R_\lambda X_1 A_{N_1} M_{E_\lambda|X}^{(1)\top}, \tag{3.20}$$

$$\Sigma_{\mathbf{y}}^{(2)} = \mathrm{Cov}(\mathbf{y}^{(2)}, \mathbf{y}^{(2)}), \tag{3.21}$$

$$\Sigma_{\mathbf{x}}^{(2)} = \mathrm{Cov}(\mathbf{x}^{(2)}, \mathbf{x}^{(2)}), \tag{3.22}$$

$$\Sigma_{\mathbf{xy}}^{(2)} = \mathrm{Cov}(\mathbf{x}^{(2)}, \mathbf{y}^{(2)}). \tag{3.23}$$

with

$$A_{N_1} := \mathbb{I}_{N_1} - \frac{1}{N_1}\mathbf{i}_{N_1}\mathbf{i}_{N_1}^{\top}. \tag{3.24}$$

► **Example A (continued).** As we already pointed out when working with the training error, the homoscedasticity hypothesis is satisfied in the setup of Example A with $\Sigma_{\boldsymbol{\varepsilon}|\mathbf{x}}^{\lambda,(1)} = \Sigma_{\boldsymbol{\varepsilon}}^{(1)}$ and the properties of the estimator $\widehat{B}_\lambda$ can be formulated in this case in terms of the matrix B. These observations lead us to the following corollary of Theorem 3.3 that provides an explicit formula for the generalization error in this setup.

Corollary 3.4 *In the conditions of Example A the conditional total testing error is determined by the following expression:*

$$\begin{aligned}\mathrm{MSE}_{\mathrm{testing}|X_1}^{\lambda} = {} & \operatorname{trace}\left(\Sigma_{\mathbf{y}}^{(2)} + \boldsymbol{\mu}_{\mathbf{y}}^{(2)}\boldsymbol{\mu}_{\mathbf{y}}^{(2)\top}\right) - 2\operatorname{trace}\left(\left(\Sigma_{\mathbf{xy}}^{(2)} + \boldsymbol{\mu}_{\mathbf{x}}^{(2)}\boldsymbol{\mu}_{\mathbf{y}}^{(2)\top}\right) B^{\top} Z_\lambda^{\top}\right) \\ & + \operatorname{trace}\left[\left(\Sigma_{\mathbf{x}}^{(2)} + \boldsymbol{\mu}_{\mathbf{x}}^{(2)}\boldsymbol{\mu}_{\mathbf{x}}^{(2)\top}\right)\left(\operatorname{trace}(\Sigma_{\boldsymbol{\varepsilon}}^{(1)}) Z_\lambda R_\lambda + Z_\lambda B B^{\top} Z_\lambda^{\top}\right)\right], \end{aligned} \tag{3.25}$$

with R_λ, $\Sigma_{\mathbf{y}}^{(2)}$, $\Sigma_{\mathbf{x}}^{(2)}$, *and* $\Sigma_{\mathbf{xy}}^{(2)}$ *as in (3.19), (3.21), (3.22), and (3.23), respectively.* ◄,

3.4 Numerical illustration

The goal of this section is showing how to compute in an explicit example the total training and testing error formulas that we provided in Theorems 3.1 and 3.3 and to show how they can be used at the time of deciding on a modeling architecture that minimizes overfitting and generalization errors. On other words, we will see how the theoretical results that we just provided can help in finding the optimal trade-off between modeling complexity and out-of-sample performance.

The underlying model. Consider a particular instance of Example B where the dependent variable y is generated by the additive stochastic perturbation of a deterministic model of the form

$$y = f(\xi) + \varepsilon, \tag{3.26}$$

with

$$f(\xi) = e^{\xi} - k, \text{ with } k = \frac{1}{2}(e - e^{-1}). \tag{3.27}$$

The term $\varepsilon \sim \mathrm{N}(0, \sigma_{\varepsilon}^2)$ in the nonlinear model (3.26) represents the disturbance term. Additionally, we assume that ξ is uniformly distributed in the interval $[-1, 1]$, that is, $\xi \sim \mathcal{U}(-1, 1)$ and hence has probability density function $g_{\xi}(z) = \frac{1}{2} I_{[-1,1]}(z)$. As it was assumed in the general conditions of Example B, we suppose that ε and ξ are independent random variables and we require that $\mathrm{E}[f(\xi)] = 0$, which is automatically guaranteed by the choice of k in (3.27).

A polynomial approximation and the corresponding linear regression model. We now construct a polynomial approximation of (3.26) and, as in Example B, we formulate this procedure as a linear univariate regularized regression model of the form

$$y = \mathbf{B}_{\lambda}^{\top} \mathbf{x} + \varepsilon_{\lambda}, \tag{3.28}$$

where the covariates $\mathbf{x}$ are adequately standardized powers of increasing order of the random variable ξ. More specifically, we set:

$$\mathbf{x} := \left(\widetilde{\xi}, \widetilde{\xi^2}, \ldots, \widetilde{\xi^p}\right)^{\top}, \text{ with } x_i = \widetilde{\xi^i} := \frac{\xi^i - \mathrm{E}\left[\xi^i\right]}{\sigma(\xi^i)}, \ i \in \{1, \ldots, p\}, \tag{3.29}$$

Given that $\xi \sim \mathcal{U}(-1, 1)$, we have:

$$\mathrm{E}\left[\xi^i\right] = \frac{(-1)^i + 1}{2(i+1)}, \text{ and } \sigma\left(\xi^i\right) = \left(\frac{1}{2i+1} - \frac{(-1)^i + 1}{2(i+1)^2}\right)^{1/2}. \tag{3.30}$$

Once a regularization strength $\lambda \in \mathbb{R}^+$ has been fixed, the corresponding regularized regression vector $\mathbf{B}_{\lambda}$ is given by (2.2), that is, $\mathbf{B}_{\lambda} := (\Sigma_{\mathbf{x}} + \lambda \mathbb{I}_p)^{-1} \mathbf{\Sigma}_{\mathbf{x}y}$. In this relation $\Sigma_{\mathbf{x}} \in \mathbb{S}_p$ and $\mathbf{\Sigma}_{\mathbf{x}y} \in \mathbb{R}^p$ can be explicitly computed using the fact that $\xi \sim \mathcal{U}(-1, 1)$. Indeed, for $i, j \in \{1, \ldots, p\}$ we have:

$$(\Sigma_{\mathbf{x}})_{i,j} = \frac{1}{\sigma(\xi^i)\sigma(\xi^j)} \left(\mathrm{E}\left[\xi^{i+j}\right] - \mathrm{E}\left[\xi^i\right] \mathrm{E}\left[\xi^j\right]\right), \tag{3.31}$$

$$(\mathbf{\Sigma}_{\mathbf{x}y})_i = \frac{1}{2\sigma(\xi^i)} \Big((-1)^i \sum_{s=0}^{i} \frac{i!}{s!} ((-1)^s e - e^{-1}) - 2k \mathrm{E}\left[\xi^i\right] \Big), \tag{3.32}$$

with $\mathrm{E}\left[\xi^i\right]$ and $\sigma(\xi^i)$ as in (3.30). In order to complete the model specification in (3.28) we determine the regression residuals ε_{λ} that are given by

$$\varepsilon_{\lambda} = e^{\xi} - k - \mathbf{B}_{\lambda}^{\top} \mathbf{x} + \varepsilon. \tag{3.33}$$

The general results in Example B provide the following expressions for the first and the second unconditional moments of ε_λ

$$\mu_{\varepsilon_\lambda} = 0, \tag{3.34}$$
$$\sigma^2_{\varepsilon_\lambda} = \sigma^2_y - \mathbf{B}_\lambda^\top \mathbf{\Sigma}_{\mathbf{x}y} - \mathbf{\Sigma}_{\mathbf{x}y}^\top \mathbf{B}_\lambda + \mathbf{B}_\lambda^\top \Sigma_{\mathbf{x}} \mathbf{B}_\lambda, \tag{3.35}$$

where

$$\sigma^2_y = \frac{1}{2}(1 - e^{-2}) + \sigma^2_\varepsilon. \tag{3.36}$$

The conditional training and testing errors. We now use the formulas in Theorems 3.1 and 3.3 in order to asses the performance of the approximating linear regression model (3.28) as a function of its complexity or, more specifically, its ability to reproduce the underlying nonlinear relation between ξ and y. We will carry this out in two different situations that put the accent in the overfitting and in the lack of generalization power that the model may incur when the selected number of covariates is too high. These features will be detected by studying the evolution of the relative values of the conditional training and testing errors as a function of the model parsimony.

Regarding the conditional total training error, as we explained earlier, the training sample is generated using the underlying model in (3.26)-(3.27) with $\xi \sim \mathcal{U}(-1, 1)$. The sample length is denoted by N_1. In order to evaluate the conditional total training error in Theorem 3.1, the following expressions for the conditional first and the second moments of the ridge residuals are required

$$\mu_{\varepsilon_\lambda|\mathbf{x}} = e^\xi - k - \mathbf{B}_\lambda^\top \mathbf{x}, \tag{3.37}$$
$$\sigma^2_{\varepsilon_\lambda|\mathbf{x}} = \sigma^2_\varepsilon. \tag{3.38}$$

As to the conditional total testing error, we consider a general case in which the random testing sample $(y^{(2)}, \mathbf{x}^{(2)})$ of length N_2 is generated using the underlying model (3.26)-(3.27) but, this time around, we allow the support of the random variable used to generate the dependent variable $y^{(2)}$ to be an arbitrary closed interval, that is,

$$y^{(2)} = e^\zeta - k + \varepsilon, \tag{3.39}$$

where k is given in (3.27), $\varepsilon \sim \mathrm{N}(0, \sigma^2_\varepsilon)$, and $\zeta \sim \mathcal{U}(a, b)$ with $a, b \in \mathbb{R}$ such that $a < b$. At the same time, the covariates sample $\mathbf{x}^{(2)}$ is constructed as in (3.29), that is,

$$\mathbf{x}^{(2)} := \left(\widetilde{\zeta}, \widetilde{\zeta^2}, \ldots, \widetilde{\zeta^p}\right)^\top, \text{ with } (\mathbf{x}^{(2)})_i = \widetilde{\zeta^i} := \frac{\zeta^i - \mathrm{E}\left[\xi^i\right]}{\sigma(\xi^i)}, \ i \in \{1, \ldots, p\}, \tag{3.40}$$

where $\mathrm{E}\left[\xi^i\right]$ and $\sigma\left(\xi^i\right)$ for $\xi \sim \mathcal{U}(-1, 1)$ are defined in (3.30). The sample matrix $X^{(2)} \in \mathbb{M}_{p,N_2}$ and vector $\mathbf{y}^{(2)\top} \in \mathbb{R}^{N_2}$ are obtained by horizontal concatenation of the N_2 realizations of $(y^{(2)}, \mathbf{x}^{(2)})$. In order to evaluate the conditional total testing error in (3.17), we need the first and second order moments of the random testing sample. First, we notice that since $\xi^{(2)} \sim \mathcal{U}(a, b)$, we can conclude that:

$$(\boldsymbol{\mu}_{\mathbf{x}}^{(2)})_i = \frac{1}{\sigma\left(\xi^i\right)} \left(\mathrm{E}\left[\zeta^i\right] - \mathrm{E}\left[\xi^i\right]\right), \ i \in \{1, \ldots, p\}, \tag{3.41}$$

where

$$\mathrm{E}\left[\zeta^i\right] = \frac{b^{i+1} - a^{i+1}}{(i+1)(b-a)} \tag{3.42}$$

and where for each $i \in \{1, \ldots, p\}$, $\mathrm{E}\left[\xi^i\right]$ and $\sigma\left(\xi^i\right)$ are given by (3.30). Additionally, the first order moment of $y^{(2)}$ is given by

$$\mu_y^{(2)} = \frac{1}{b-a}(e^b - e^a) - k. \tag{3.43}$$

The second order moments of the testing sample are:

$$(\Sigma_{\mathbf{x}}^{(2)})_{i,j} = \frac{1}{(b-a)\sigma(\xi^i)\sigma(\xi^j)} \int_a^b \left(z^i - \mathrm{E}\left[\xi^i\right]\right)\left(z^j - \mathrm{E}\left[\xi^j\right]\right) dz - (\boldsymbol{\mu}_{\mathbf{x}}^{(2)})_i (\boldsymbol{\mu}_{\mathbf{x}}^{(2)})_j, \quad i,j \in \{1,\ldots,p\}\,, \tag{3.44}$$

$$(\boldsymbol{\Sigma}_{\mathbf{x}y}^{(2)})_i = \frac{1}{(b-a)\sigma(\xi^i)} \int_a^b \left(z^i - \mathrm{E}\left[\xi^i\right]\right)(e^z - k) dz - (\boldsymbol{\mu}_{\mathbf{x}}^{(2)})_i \mu_y^{(2)}, \quad i \in \{1,\ldots,p\}\,, \tag{3.45}$$

$$(\sigma_y^2)^{(2)} = \frac{1}{2(b-a)}(e^{2b} - e^{2a}) - \frac{1}{(b-a)^2}(e^b - e^a)^2 + \sigma_\varepsilon^2. \tag{3.46}$$

These relations allow for the explicit evaluation of the conditional total testing error that we now numerically study for two particular choices for the values of a and b that determine the support of the distribution of ζ.

Numerical illustration. Case I. Overfitting. Consider first the case in which $[a,b] = [-1,1]$ and hence $\zeta = \xi$. This means that the testing sample is generated in the same way as the training one. In this setup, an increase in the number of covariates p amounts to an increase in the degree of the polynomial that we are using to fit the sample. When this degree is too high, the fit may lead to a match with the noise in the sample rather than with the underlying deterministic signal that we are trying to model. Figure 1 shows the evolution of the conditional training and total errors for this case (for a given training covariates sample X_1 and a fixed regularization strength reported in the caption). As we anticipated, the conditional total training error is a decreasing function with p but, in turn, the generalization error decreases only up to a point and when overfitting to the training sample starts to occur, we observe a monotonous increase in the testing error. In this case, this turning point takes place for $p = 9$, a value that seems to offer the optimal tradeoff between the quality of training and out-of-sample performance. Another way to state this observation is that, with this noise level, the best polynomial approximation (in the mean square sense) of the the exponential function in the interval $[-1,1]$ is obtained when using polynomials of degree 9.

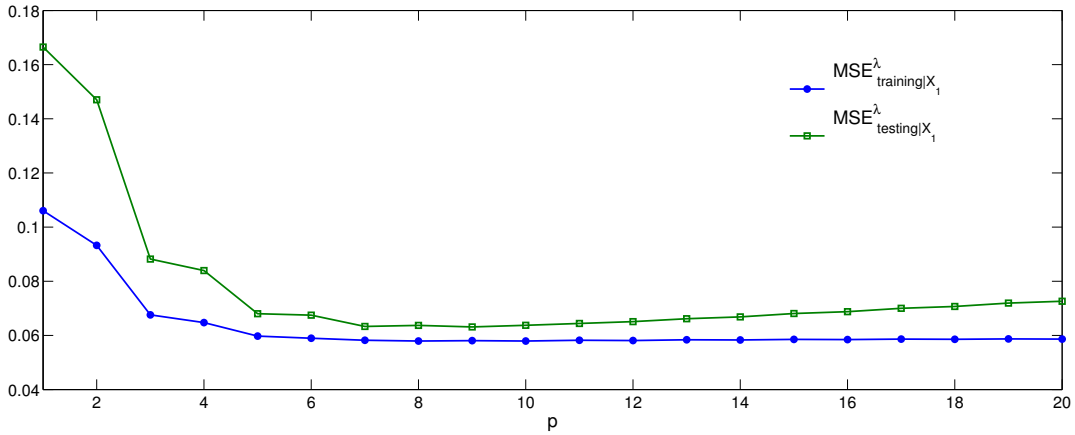

Figure 1: Conditional total training and testing errors committed by the regularized ($\lambda = 0.9$) polynomial approximation of the nonlinear underlying model (3.26)-(3.27) as a function of the number of covariates p or, equivalently, the degree of the polynomial approximation. The lengths of the training and testing samples are $N_1 = 20$ and $N_2 = 20$, respectively, and $\zeta, \xi \sim \mathcal{U}(-1,1)$.

Numerical illustration. Case II. Generalization performance. In this case we allow for the

support of distribution for ζ used to generate the testing sample to be different from the one that defines the construction of the training sample. Indeed, we use $\xi \sim \mathcal{U}(-1,1)$ and $\zeta \sim \mathcal{U}(1,2)$ and hence the training and testing are carried out in the different intervals where ξ and ζ are defined. Equivalently, in this situation the testing will be measuring the ability of the approximate model to reckon the values of the underlying model in an interval different from the one in which the training has taken place; that is why we explicitly talk in this case of generalization performance. Figure 2 shows the evolution of the conditional training and total errors for this case (for a given training covariates sample X_1 and a fixed regularization strength reported in the caption). Again, the conditional total training error is a decreasing function of p. As to the testing error in this case, its absolute value is much higher than in Case I and its minimum is already attained for $p = 4$, that is, in this case a fourth order polynomial provides the best tradeoff in the modeling of the exponential function.

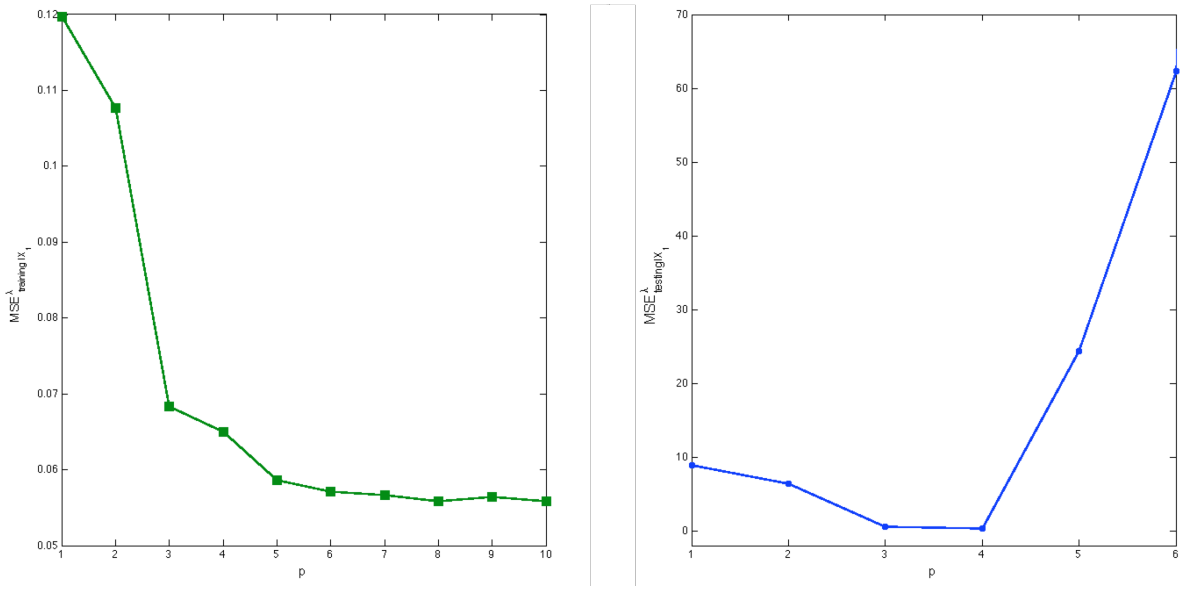

Figure 2: Conditional total training and testing errors committed by the regularized ($\lambda = 1$) polynomial approximation of the nonlinear underlying model (3.26)-(3.27) as a function of the number of covariates p or, equivalently, the degree of the polynomial approximation. The lengths of the training and testing samples are $N_1 = 30$ and $N_2 = 40$, respectively. The random variables $\xi \sim \mathcal{U}(-1,1)$ and $\zeta \sim \mathcal{U}(1,2)$ have been used in the dependent variables generating model.

4 Appendices

4.1 Proof of Lemma 2.1

Part (i). By (2.28) we have that $\widehat{B}_\lambda := (XA_NX^\top + \lambda N\mathbb{I}_p)^{-1}XA_NY^\top$. At the same time the relation (2.29) yields $XA_NY^\top = (XA_NX^\top)\widehat{B}$, which in (2.28) gives $\widehat{B}_\lambda := (XA_NX^\top + \lambda N\mathbb{I}_p)^{-1}(XA_NX^\top)\widehat{B}$. It is easy to see that the last expression is equivalent to **(i)** as required.
Part (ii). First, in order to establish (2.33), we notice that it follows from the definition of R_λ in (2.32) that $XA_NX^\top = R_\lambda^{-1} - \lambda N\mathbb{I}_p$ and substituting this relation into (2.31) in **(i)** yields

$$Z_\lambda = R_\lambda XA_NX^\top = R_\lambda\left(R_\lambda^{-1} - \lambda N\mathbb{I}_p\right) = \mathbb{I}_p - \lambda NR_\lambda,$$

as required. Second, we show that (2.34) holds. Notice that by rewriting the relation (2.31) using the definition of R_λ in (2.32) as

$$Z_\lambda = \left((XA_NX^\top)^{-1}XA_NX^\top + \lambda N(XA_NX^\top)^{-1}\right)^{-1} = \left(\mathbb{I}_p + \lambda N(XA_NX^\top)^{-1}\right)^{-1},$$

we establish that it is equivalent to (2.34), as required. ■

4.2 Proof of Theorem 2.6

We first notice that by definition of R_λ in (2.43), namely $R_\lambda = (XA_NX^\top + \lambda N\mathbb{I}_p)^{-1}$, the relation (2.28) can be rewritten as $\widehat{B}_\lambda = R_\lambda XA_NY^\top$. In this expression we use that $Y = B_\lambda^\top X + E_\lambda$ and that, by the same definition of R_λ in (2.43), $XAX^\top = R_\lambda^{-1} - \lambda N\mathbb{I}_p$. Taking both observations into account, we obtain that $\widehat{B}_\lambda = R_\lambda XA_NY^\top = R_\lambda XA_N(X^\top B_\lambda + E_\lambda^\top) = R_\lambda XA_NX^\top B_\lambda + R_\lambda XA_NE_\lambda^\top = B_\lambda - \lambda NR_\lambda B_\lambda + R_\lambda XA_NE_\lambda^\top$ or, equivalently,

$$\widehat{B}_\lambda - B_\lambda + \lambda NR_\lambda B_\lambda = R_\lambda XA_NE_\lambda^\top. \tag{4.1}$$

We now point out that the conditional homoscedasticity hypothesis on the residuals implies by Corollary 2.7 that $E_\lambda|X \sim \mathrm{MN}(M_{E_\lambda|X}, \Sigma^\lambda_{\boldsymbol{\epsilon}|\mathbf{x}}, \mathbb{I}_N)$ and hence by Theorem 2.3.10 in [Gupt 00] we have that

$$(\widehat{B}_\lambda - B_\lambda + \lambda NR_\lambda B_\lambda)|X \sim \mathrm{MN}(R_\lambda XA_NM_{E_\lambda|X}^\top, R_\lambda XA_NA_N^\top X^\top R_\lambda, \Sigma^\lambda_{\boldsymbol{\epsilon}|\mathbf{x}}). \tag{4.2}$$

Since $A_NA_N^\top = A_N$, the statement in (2.42) follows. ■

4.3 Proof of Corollary 2.7

As $\widehat{B}_\lambda - B = (\widehat{B}_\lambda - B_\lambda) + (B_\lambda - B)$, then from (2.42) in Theorem 2.6 follows that

$$(\widehat{B}_\lambda - B)|X \sim \mathrm{MN}(B_\lambda - B - \lambda NR_\lambda B_\lambda + R_\lambda XA_NM_{E_\lambda|X}^\top, R_\lambda XA_NX^\top R_\lambda, \Sigma^\lambda_{\boldsymbol{\epsilon}|\mathbf{x}}). \tag{4.3}$$

First, notice that by (2.10) and by (2.36) we can conclude that $M_{E_\lambda|X}^\top = X^\top(B - B_\lambda)$. Consequently,

$$\begin{aligned} B_\lambda - B - \lambda NR_\lambda B_\lambda + R_\lambda XA_NX^\top(B - B_\lambda) &= B_\lambda - B - \lambda NR_\lambda B_\lambda + R_\lambda(R_\lambda^{-1} - \lambda N\mathbb{I}_p)(B - B_\lambda) \\ &= B_\lambda - B - \lambda NR_\lambda B_\lambda + B - B_\lambda - \lambda NR_\lambda B + \lambda NR_\lambda B_\lambda \\ &= -\lambda NR_\lambda B, \end{aligned} \tag{4.4}$$

where we used the definition of R_λ in (2.43). Second, we rewrite the first scale matrix in (4.3) as follows

$$R_\lambda XA_NX^\top R_\lambda = R_\lambda(R_\lambda^{-1} - \lambda N\mathbb{I}_p)R_\lambda = (\mathbb{I}_p - \lambda NR_\lambda)R_\lambda = Z_\lambda R_\lambda. \tag{4.5}$$

At the same time the second scale matrix in (4.3) in the conditions of the statement is equal to $\Sigma_{\boldsymbol{\epsilon}}$ and hence together with (4.4) and (4.5) this yields (2.45).

Now, in order to show (2.46), we first use the definition of Z_λ in (2.33) and rewrite (4.5) as

$$R_\lambda XA_NX^\top R_\lambda = Z_\lambda R_\lambda = \frac{1}{\lambda N}Z_\lambda(\mathbb{I}_p - Z_\lambda) = \frac{1}{\lambda N}\left(Z_\lambda - Z_\lambda^2\right) = \frac{1}{\lambda N}Z_\lambda\left(Z_\lambda^{-1} - \mathbb{I}_p\right)Z_\lambda. \tag{4.6}$$

As by hypothesis $XA_NX^\top$ is invertible, then by (2.34) it holds that $(XA_NX^\top)^{-1} = \frac{1}{\lambda N}\left(Z_\lambda^{-1} - \mathbb{I}_p\right)$ and the expression (4.5) becomes

$$R_\lambda XA_NX^\top R_\lambda = Z_\lambda(XA_NX^\top)^{-1}Z_\lambda. \tag{4.7}$$

Consequently, the relations (4.3) and (4.4) together with (4.7) guarantee (2.46), as required. ■

4.4 Proof of Theorem 3.1

We start by using the definition of the conditional total training error provided in (3.9). First, we define $D_\lambda := \widehat{B}_\lambda - B_\lambda$ and subtract and add B_λ to $\widehat{B}_\lambda$ in (3.9) which results in

$$\begin{aligned}\mathrm{MSE}^{\lambda}_{\mathrm{training}|X} &= \frac{1}{N}\mathrm{trace}\left(\mathrm{E}_X\left[(Y-(\widehat{B}_\lambda - B_\lambda + B_\lambda)^\top X)^\top (Y-(\widehat{B}_\lambda - B_\lambda + B_\lambda)X)\right]\right)\\ &= \frac{1}{N}\mathrm{trace}\left(\mathrm{E}_X\left[(E_\lambda-(\widehat{B}_\lambda - B_\lambda)^\top X)^\top (E_\lambda-(\widehat{B}_\lambda - B_\lambda)^\top X)\right]\right)\\ &= \frac{1}{N}\mathrm{trace}\left(\mathrm{E}_X\left[(E_\lambda - D_\lambda^\top X)^\top (E_\lambda - D_\lambda^\top X)\right]\right) = \frac{1}{N}\mathrm{trace}(\mathrm{E}_X\left[E_\lambda^\top E_\lambda\right])\\ &+ \frac{1}{N}\mathrm{trace}(X^\top \mathrm{E}_X\left[D_\lambda D_\lambda^\top\right]X) - \frac{2}{N}\mathrm{trace}\left(X^\top \mathrm{E}_X\left[D_\lambda E_\lambda\right]\right). \end{aligned} \tag{4.8}$$

We now study separately the three summands in (4.8). The first one can be computed in a straightforward way. Indeed, by **(i)** in Lemma 2.3 we can immediately write

$$\frac{1}{N}\mathrm{trace}(\mathrm{E}_X\left[E_\lambda^\top E_\lambda\right]) = \mathrm{trace}(\Sigma^\lambda_{\boldsymbol{\varepsilon}|\mathbf{x}}) + \frac{1}{N}\mathrm{trace}(M_{E_\lambda|X}^\top M_{E_\lambda|X}). \tag{4.9}$$

As for the second summand, we notice first that, in the hypotheses of the theorem, the properties of the estimator $\widehat{B}_\lambda$ are as provided in Theorem 2.6, that is,

$$D_\lambda|X \sim \mathrm{MN}(M_{D_\lambda}, R_\lambda X A_N X^\top R_\lambda, \Sigma^\lambda_{\boldsymbol{\varepsilon}|\mathbf{x}}), \tag{4.10}$$

with

$$M_{D_\lambda} := -\lambda N R_\lambda B_\lambda + R_\lambda X A_N M_{E_\lambda|X}^\top \tag{4.11}$$

and

$$R_\lambda := (X A_N X^\top + \lambda N \mathbb{I}_p)^{-1}. \tag{4.12}$$

At the same time, by Theorem 2.3.10 in [Gupt 00] (see also Lemma 6.3 in [Grig 16]) we have that D_λ satisfies the following relation

$$\mathrm{E}_X\left[D_\lambda D_\lambda^\top\right] = \mathrm{trace}(\Sigma^\lambda_{\boldsymbol{\varepsilon}|\mathbf{x}}) R_\lambda X A_N X^\top R_\lambda + M_{D_\lambda} M_{D_\lambda}^\top. \tag{4.13}$$

Using this expression, we see that the second summand in (4.8) can be written as

$$\begin{aligned}\frac{1}{N}\mathrm{trace}\left(X^\top \mathrm{E}_X\left[D_\lambda D_\lambda^\top\right]X\right) &= \frac{1}{N}\mathrm{trace}\left(\mathrm{E}_X\left[D_\lambda D_\lambda^\top\right]XX^\top\right)\\ &= \frac{1}{N}\mathrm{trace}\left[\left(\mathrm{trace}(\Sigma^\lambda_{\boldsymbol{\varepsilon}|\mathbf{x}}) R_\lambda X A_N X^\top R_\lambda + M_{D_\lambda} M_{D_\lambda}^\top\right)XX^\top\right]\\ &= \frac{1}{N}\mathrm{trace}(\Sigma^\lambda_{\boldsymbol{\varepsilon}|\mathbf{x}}) Z_\lambda R_\lambda X X^\top + \frac{1}{N}\mathrm{trace}(M_{D_\lambda} M_{D_\lambda}^\top XX^\top), \end{aligned} \tag{4.14}$$

where

$$Z_\lambda := R_\lambda X A_N X^\top. \tag{4.15}$$

Finally, we can rewrite the third summand in (4.8) by using the definition of the ridge regression matrix estimator $\widehat{B}_\lambda$ in (2.28) and using R_λ in (4.12):

$$\begin{aligned}X^\top \mathrm{E}_X\left[D_\lambda E_\lambda\right] =& X^\top \mathrm{E}_X\left[\left(\widehat{B}_\lambda - B_\lambda\right)E_\lambda\right] = X^\top \mathrm{E}_X\left[\widehat{B}_\lambda E_\lambda\right] - X^\top B_\lambda \mathrm{E}_X\left[E_\lambda\right]\\ =& X^\top R_\lambda X A_N \mathrm{E}_X\left[Y^\top E_\lambda\right] - X^\top B_\lambda \mathrm{E}_X\left[E_\lambda\right] = X^\top R_\lambda X_N A_N \mathrm{E}_X\left[X^\top B_\lambda E_\lambda + E_\lambda^\top E_\lambda\right]\\ &- X^\top B_\lambda \mathrm{E}_X\left[E_\lambda\right] = X^\top R_\lambda X A_N (X^\top B_\lambda \mathrm{E}_X\left[E_\lambda\right] + \mathrm{E}_X\left[E_\lambda^\top E_\lambda\right]) - X^\top B_\lambda \mathrm{E}_X\left[E_\lambda\right].\end{aligned}$$

In this expression we now use the part **(i)** in Lemma 2.3 and hence conclude that

$$-\frac{2}{N}\text{trace}\left(\mathrm{E}_X\left[X^\top D_\lambda E_\lambda\right]\right) = -\frac{2}{N}\text{trace}\Big(X^\top R_\lambda X A_N (X^\top B_\lambda M_{E_\lambda|X} + \text{trace}(\Sigma^\lambda_{\boldsymbol{\varepsilon}|\mathbf{x}})\mathbb{I}_N + M^\top_{E_\lambda|X} M_{E_\lambda|X}) - X^\top B_\lambda M_{E_\lambda|X}\Big) = -\frac{2}{N}\Big(\text{trace}(\Sigma^\lambda_{\boldsymbol{\varepsilon}|\mathbf{x}})\text{trace}(Z_\lambda) + \text{trace}(X^\top R_\lambda X A_N M^\top_{E_\lambda|X} M_{E_\lambda|X}) + \text{trace}\big(\left(X^\top R_\lambda X A_N - \mathbb{I}_N\right) X^\top B_\lambda M_{E_\lambda|X}\big)\Big). \tag{4.16}$$

Using (4.9), (4.14), and (4.16) in (4.8) yields the expression (3.10). At the same time, in (3.10) the last term can be rewritten as

$$-\frac{2}{N}\ \text{trace}\left((X^\top R_\lambda X A_N - \mathbb{I}_N) X^\top B_\lambda M_{E_\lambda|X}\right) = -\frac{2}{N}\ \text{trace}\left(X^\top (Z_\lambda - \mathbb{I}_p) B_\lambda M_{E_\lambda|X}\right) = 2\lambda \text{trace}(X^\top R_\lambda B_\lambda M_{E_\lambda|X}), \tag{4.17}$$

by using (2.33). The substitution of this expression in (3.10) provides the relation (3.11). ■

4.5 Proof of Corollary 3.2

In order to prove (3.15), we just determine all the ingredients required for the evaluation of (3.9) in the conditions of Example A and we then show that the result follows. First, notice that by (2.11)

$$\Sigma^\lambda_{\boldsymbol{\varepsilon}|\mathbf{x}} = \Sigma_{\boldsymbol{\varepsilon}} \tag{4.18}$$

and that by (2.10)

$$M_{E_\lambda|X} = (B - B_\lambda)^\top X. \tag{4.19}$$

Additionally,

$$M^\top_{E_\lambda|X} M_{E_\lambda|X} = X^\top (B - B_\lambda)(B - B_\lambda)^\top X = X^\top B B^\top X - X^\top B B_\lambda^\top X - X^\top B_\lambda B^\top X + X^\top B_\lambda B_\lambda^\top X. \tag{4.20}$$

We now proceed by pointing out that $M_{(\widehat{B}_\lambda - B_\lambda)} = M_{\widehat{B}_\lambda} + B_\lambda$ and by using that in the conditions of Example A, the mean matrix $M_{\widehat{B}_\lambda}$ is given by $M_{\widehat{B}_\lambda} = Z_\lambda B$, with Z_λ as in (2.44). We hence have that

$$M_{(\widehat{B}_\lambda - B_\lambda)} = B_\lambda + Z_\lambda B = B_\lambda + (\mathbb{I}_p - \lambda N R_\lambda) B = B + B_\lambda - \lambda N R_\lambda B. \tag{4.21}$$

Additionally, we compute

$$M_{(\widehat{B}_\lambda - B_\lambda)} M^\top_{(\widehat{B}_\lambda - B_\lambda)} = (B + B_\lambda - \lambda N R_\lambda B)(B + B_\lambda - \lambda N R_\lambda B)^\top = B B^\top - B B_\lambda^\top - B_\lambda B^\top + B_\lambda B_\lambda^\top - \lambda N B B^\top R_\lambda - \lambda N R_\lambda B B^\top + \lambda N R_\lambda B B_\lambda^\top + \lambda N B_\lambda B^\top R_\lambda + \lambda^2 N^2 R_\lambda B B^\top R_\lambda. \tag{4.22}$$

We now substitute these expressions in the five summands that make formula (3.11) and that we analyze one. First,

$$\text{trace}(\Sigma^\lambda_{\boldsymbol{\varepsilon}|\mathbf{x}}) = \text{trace}(\Sigma_{\boldsymbol{\varepsilon}}), \tag{4.23}$$

$$\frac{1}{N}\text{trace}(\Sigma^\lambda_{\boldsymbol{\varepsilon}|\mathbf{x}})\text{trace}\left(Z_\lambda (R_\lambda X X^\top - 2\mathbb{I}_p)\right) = \frac{1}{N}\text{trace}(\Sigma_{\boldsymbol{\varepsilon}})\text{trace}\left(Z_\lambda (R_\lambda X X^\top - 2\mathbb{I}_p)\right), \tag{4.24}$$

$$\frac{1}{N}\text{trace}\left(M_{(\widehat{B}_\lambda - B_\lambda)}M_{(\widehat{B}_\lambda - B_\lambda)}^\top XX^\top\right) = \frac{1}{N}\text{trace}((BB^\top - BB_\lambda^\top - B_\lambda B^\top + B_\lambda B_\lambda^\top - \lambda NBB^\top R_\lambda - \lambda NR_\lambda BB^\top + \lambda NR_\lambda BB_\lambda^\top + \lambda NB_\lambda B^\top R_\lambda + \lambda^2 N^2 R_\lambda BB_\lambda^\top R_\lambda)XX^\top), \tag{4.25}$$

$$-\frac{2}{N}\text{trace}((X^\top R_\lambda XA_N - \frac{1}{2}\mathbb{I}_N)M_{E_\lambda|X}^\top M_{E_\lambda|X}) = -\frac{2}{N}\text{trace}((X^\top R_\lambda XA_N - \frac{1}{2}\mathbb{I}_N)(X^\top BB^\top X - X^\top BB_\lambda^\top X - X^\top B_\lambda B^\top X + X^\top B_\lambda B_\lambda^\top X)) = \frac{1}{N}\text{trace}((-BB^\top + BB_\lambda^\top + B_\lambda B^\top - B_\lambda B_\lambda^\top + 2\lambda NR_\lambda BB^\top - 2\lambda NR_\lambda BB_\lambda^\top - 2\lambda NR_\lambda B_\lambda B^\top + 2\lambda NR_\lambda B_\lambda B_\lambda^\top)XX^\top), \tag{4.26}$$

and finally,

$$2\lambda\text{trace}\left(X^\top R_\lambda B_\lambda M_{E_\lambda|X}\right) = 2\lambda\text{trace}\left(X^\top R_\lambda B_\lambda (B - B_\lambda)^\top X\right) = 2\lambda\text{trace}\left((R_\lambda BB^\top - R_\lambda B_\lambda B_\lambda^\top)XX^\top\right), \tag{4.27}$$

respectively.

Gathering the terms (4.23)-(4.27), we obtain that

$$\begin{aligned}\text{MSE}_{\text{training}|X}^\lambda =& \text{trace}(\Sigma_{\boldsymbol{\varepsilon}}) + \frac{1}{N}\text{trace}(\Sigma_{\boldsymbol{\varepsilon}})\text{trace}\left(Z_\lambda(R_\lambda XX^\top - 2\mathbb{I}_p)\right) \\ &+ \frac{1}{N}\text{trace}((BB^\top - BB_\lambda^\top - B_\lambda B^\top + B_\lambda B_\lambda^\top - \lambda NBB^\top R_\lambda - \lambda NR_\lambda BB^\top \\ &+ \lambda NB_\lambda B^\top R_\lambda + \lambda NR_\lambda BB_\lambda^\top + \lambda^2 N^2 R_\lambda BB^\top R_\lambda - BB^\top + BB_\lambda^\top + B_\lambda B^\top - B_\lambda B_\lambda^\top \\ &+ 2\lambda NR_\lambda BB^\top - 2\lambda NR_\lambda BB_\lambda^\top - 2\lambda NR_\lambda B_\lambda B^\top + 2\lambda NR_\lambda B_\lambda B_\lambda^\top + 2\lambda NR_\lambda B_\lambda B^\top \\ &- 2\lambda NR_\lambda B_\lambda B_\lambda^\top)XX^\top) \\ =& \text{trace}(\Sigma_{\boldsymbol{\varepsilon}}) + \frac{1}{N}\text{trace}(\Sigma_{\boldsymbol{\varepsilon}})\text{trace}\left(Z_\lambda(R_\lambda XX^\top - 2\mathbb{I}_p)\right) + \lambda^2 N\text{trace}(R_\lambda BB^\top R_\lambda XX^\top),\end{aligned}$$

which coincides with (3.15), as required. ■

4.6 Proof of Theorem 3.3

In order to prove (3.17), we first rewrite (3.16) as

$$\begin{aligned}\text{MSE}_{\text{testing}|X_1}^\lambda &= \frac{1}{N_2}\text{E}_{X_1}\left[\left(Y_2 - \widehat{B}_\lambda^\top X_2\right)^\top\left(Y_2 - \widehat{B}_\lambda^\top X_2\right)\right] \\ &= \frac{1}{N_2}\left[\text{trace}\left(\text{E}_{X_1}\left[Y_2^\top Y_2\right]\right) - 2\,\text{trace}\left(\text{E}_{X_1}\left[Y_2^\top \widehat{B}_\lambda^\top X_2\right]\right) + \text{trace}\left(\text{E}_{X_1}\left[X_2^\top \widehat{B}_\lambda \widehat{B}_\lambda^\top X_2\right]\right)\right].\end{aligned} \tag{4.28}$$

In order to evaluate this expression, we study separately all its terms. We start with the first summand and use that the pairs of random samples (X_1, Y_1) and (X_2, Y_2) used for training and testing, respectively, are independent from each other. We hence rewrite the first term in (4.28) as

$$\frac{1}{N_2}\text{trace}\left(\text{E}_{X_1}\left[Y_2^\top Y_2\right]\right) = \frac{1}{N_2}\text{trace}\left(\text{E}\left[Y_2^\top Y_2\right]\right) = \text{trace}\left[\Sigma_{\mathbf{y}}^{(2)} + \boldsymbol{\mu}_{\mathbf{y}}^{(2)}\boldsymbol{\mu}_{\mathbf{y}}^{(2)\top}\right] \tag{4.29}$$

with

$$\Sigma_{\mathbf{y}}^{(2)} = \mathrm{Cov}(\mathbf{y}^{(2)}, \mathbf{y}^{(2)}), \tag{4.30}$$

$$\boldsymbol{\mu}_{\mathbf{y}}^{(2)} = \mathrm{E}\left[\mathbf{y}^{(2)}\right]. \tag{4.31}$$

For the second summand we obtain

$$\begin{aligned} -\frac{2}{N_2}\mathrm{trace}\left(\mathrm{E}_{X_1}\left[Y_2^\top \widehat{B}_\lambda^\top X_2\right]\right) &= -\frac{2}{N_2}\mathrm{trace}\left(\mathrm{E}_{X_1}\left[X_2 Y_2^\top \widehat{B}_\lambda^\top\right]\right) = -\frac{2}{N_2}\mathrm{trace}\left(\mathrm{E}\left[X_2 Y_2^\top\right]\mathrm{E}_{X_1}\left[\widehat{B}_\lambda^\top\right]\right) \\ &= -2\mathrm{trace}\left(\left(\Sigma_{\mathbf{xy}}^{(2)} + \boldsymbol{\mu}_{\mathbf{x}}^{(2)}\boldsymbol{\mu}_{\mathbf{y}}^{(2)\top}\right) M_{\widehat{B}_\lambda}^\top\right), \end{aligned} \tag{4.32}$$

with

$$M_{\widehat{B}_\lambda} := B_\lambda - \lambda N_1 R_\lambda B_\lambda + R_\lambda X_1 A_{N_1} M_{E_\lambda|X}^{(1)\top}, \tag{4.33}$$

$$\Sigma_{\mathbf{xy}}^{(2)} = \mathrm{Cov}(\mathbf{x}^{(2)}, \mathbf{y}^{(2)}), \tag{4.34}$$

$$\boldsymbol{\mu}_{\mathbf{x}}^{(2)} = \mathrm{E}\left[\mathbf{x}^{(2)}\right]. \tag{4.35}$$

In (4.32) we used the properties of the ridge regression matrix estimator $\widehat{B}_\lambda$ provided in Theorem 2.6. Finally, we rewrite the third term as

$$\begin{aligned} \frac{1}{N_2}\mathrm{trace}\left(\mathrm{E}_{X_1}\left[X_2^\top \widehat{B}_\lambda \widehat{B}_\lambda^\top X_2\right]\right) &= \frac{1}{N_2}\mathrm{trace}\left(\mathrm{E}_{X_1}\left[X_2 X_2^\top \widehat{B}_\lambda \widehat{B}_\lambda^\top\right]\right) = \frac{1}{N_2}\mathrm{trace}\left(\mathrm{E}\left[X_2 X_2^\top\right]\mathrm{E}_{X_1}\left[\widehat{B}_\lambda \widehat{B}_\lambda^\top\right]\right) \\ &= \mathrm{trace}\left[\left(\Sigma_{\mathbf{x}}^{(2)} + \boldsymbol{\mu}_{\mathbf{x}}^{(2)}\boldsymbol{\mu}_{\mathbf{x}}^{(2)\top}\right)\left(\mathrm{trace}\left(\Sigma_{\boldsymbol{\varepsilon}|\mathbf{x}}^{\lambda,(1)}\right) Z_\lambda R_\lambda + M_{\widehat{B}_\lambda} M_{\widehat{B}_\lambda}^\top\right)\right], \end{aligned} \tag{4.36}$$

with

$$\Sigma_{\mathbf{x}}^{(2)} = \mathrm{Cov}(\mathbf{x}^{(2)}, \mathbf{x}^{(2)}), \tag{4.37}$$

and where $\boldsymbol{\mu}_{\widehat{B}_\lambda}$ is given in (4.33). In this expression we again used Theorem 2.6 and the property (4.13). Substituting expressions (4.29)-(4.36) into (4.28) immediately yields (3.17), as required. ■

References

[Gema 92] S. Geman, E. Bienenstock, and R. Doursat. "Neural Networks and the Bias/Variance Dilemma". *Neural Computation*, Vol. 4, No. 1, pp. 1–58, jan 1992.

[Grig 16] L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. "Nonlinear memory capacity of parallel time-delay reservoir computers in the processing of multidimensional signals". *To appear in Neural Computation*, 2016.

[Gupt 00] A. Gupta and D. Nagar. *Matrix Variate Distributions*. Chapman and Hall/CRC, 2000.

[Hami 94] J. D. Hamilton. *Time series analysis*. Princeton University Press, Princeton, NJ, 1994.

[Hast 13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Verlag, second Ed., 2013.

[Hoer 70] A. E. Hoerl and R. W. Kennard. "Ridge regression: biased estimation for nonorthogonal problems". *Technometrics*, Vol. 12, No. 1, pp. 55–67, 1970.

[Meye 00] C. Meyer. *Matrix Analysis and Applied Linear Algebra Book and Solutions Manual*. Society for Industrial and Applied Mathematics, 2000.

[Tikh 43] A. N. Tikhonov. "On the stability of inverse problems". *Dokl. Akad. Nauk SSSR*, Vol. 39, No. 5, pp. 195–198, 1943.

[Tikh 63] A. N. Tikhonov. "Solution of incorrectly formulated problems and the regularization method". *Dokl. Akad. Nauk SSSR*, Vol. 151, pp. 501–504, 1963.