Machine Learning (2021) 110:1549–1584
https://doi.org/10.1007/s10994-021-05976-x

Optimal data collection design in machine learning: the case of the fixed effects generalized least squares panel data model

Giorgio Gnecco[1] · Federico Nutarelli[1] · Daniela Selvi[2]

Received: 9 June 2020 / Revised: 15 January 2021 / Accepted: 5 April 2021 /
Published online: 8 June 2021

Abstract

This work belongs to the strand of literature that combines machine learning, optimization, and econometrics. The aim is to optimize the data collection process in a specific statistical model, commonly used in econometrics, employing an optimization criterion inspired by machine learning, namely, the generalization error conditioned on the training input data. More specifically, the paper is focused on the analysis of the conditional generalization error of the Fixed Effects Generalized Least Squares (FEGLS) panel data model, i.e., a linear regression model with applications in several fields, able to represent unobserved heterogeneity in the data associated with different units, for which distinct observations related to the same unit are corrupted by correlated measurement errors. The framework considered in this work differs from the classical FEGLS model for the additional possibility of controlling the conditional variance of the output variable given the associated unit and input variables, by changing the cost per supervision of each training example. Assuming an upper bound on the total supervision cost, i.e., the cost associated with the whole training set, the trade-off between the training set size and the precision of supervision (i.e., the reciprocal of the conditional variance of the output variable) is analyzed and optimized. This is achieved by formulating and solving in closed form suitable optimization problems, based on large-sample approximations of the generalization error associated with the FEGLS estimates of the model parameters, conditioned on the training input data. The results of the analysis extend to the FEGLS case and to various large-sample approximations of its conditional generalization error the ones obtained by the authors in recent works for simpler linear regression models. They highlight the importance of how the precision of supervision scales with respect to the cost per training example in determining the optimal trade-off between training set size and precision. Numerical results confirm the validity of the theoretical findings.

Keywords Fixed effects generalized least squares panel data model · First-order serial covariance · Conditional generalization error · Large-sample approximations · Optimal training set size

Editor: Bart Baesens.

Extended author information available on the last page of the article

1 Introduction

In various applications in economics, engineering, medicine, physics, and several other fields, one has often the need of approximating a function, based on a finite set of input/output noisy examples. This belongs to the typical class of problems investigated by supervised machine learning (Hastie et al. 2009; Vapnik 1998). In some cases, the noise variance of the output variable can be decreased, at least to some extent, by making the cost of each supervision larger. As an example, observations could be acquired by using more precise measurement devices (then, likely, having also larger acquisition cost). Similarly, each supervision could be made by an expert (also in this case, a larger cost would be expected by increasing the level of expertise). In all these situations, it can be useful to optimize the trade-off between the training set size and the precision of supervision. In the conference work Gnecco and Nutarelli (2019), this kind of analysis was conducted by proposing a modification of the classical linear regression model, in which one has the additional possibility to control the conditional variance of the output variable given the associated input variables, by changing the time (hence, the cost) dedicated to provide a label to each training input example, and fixing an upper bound on the time available for the supervision of the whole training set. Based on a large-sample approximation of the output of the ordinary least squares regression algorithm, it was shown in that work that the optimal choice of the supervision time per example highly depends on how the precision of supervision scales with respect to the cost per training example. The analysis was refined in Gnecco and Nutarelli (2019), where a related optimization problem, based on the analysis of the output produced by a different regression algorithm (namely, weighted least squares) was considered, obtaining similar results at optimality, for a model in which distinct training examples are possibly associated with different supervision times. Finally, in the conference work Gnecco and Nutarelli (2020), the analysis of the optimal trade-off between training set size and precision of supervision was extended to a more general linear model of the input/output relationship, namely, the fixed effects panel data model. In this model, observations associated with different units (individuals) depend also on additional constants, one for each unit, which make it possible to include, in the input/output relationship, unobserved heterogeneity in the data. Moreover, each unit is observed along another dimension, which is typically time. This kind of model (and its variations) is commonly applied in the analysis of data in both microeconomics and macroeconomics (Arellano 2004; Cameron and Trivedi 2005; Wooldridge 2002), where each unit may represent, for instance, a firm, or a country. It is also applied in biostatistics (Härdle et al. 2007) and sociology (Frees 2004). An important engineering application of the model (and of its variations) is in the calibration of sensors (Reeve 1988, Sect. 4.1).

In order to increase the applicability of the analysis carried out in our previous conference work (Gnecco and Nutarelli 2020), in this paper we extend it thoroughly in the following directions. First, Gnecco and Nutarelli (2020) investigated only the case in which the measurements errors of observations associated with the same unit are mutually independent. In this paper, we extend such analysis to the case of dependent measurement errors. Moreover, differently from Gnecco and Nutarelli (2020), we confirm the validity of the obtained theoretical results numerically. Further, in Gnecco and Nutarelli (2020), the optimal trade-off between training set size and precision of supervision was analyzed only for a fixed number of units, assuming that the number of observations associated with the same unit is large enough to justify a large-sample approximation with respect to the number of observations. In the last part of this work, we consider additionally the cases of a large-sample approximation with respect to the number of units, and of a large-sample approximation with respect to both the number of units and the number of observations per unit.

In line with the results of the theoretical analyses made in (Gnecco and Nutarelli 2019, 2019, 2020) for simpler linear regression models, we show that, also for the more applicable fixed effects generalized least squares panel data model, the following holds in general: when the precision of each supervision (i.e., the reciprocal of the conditional variance of the output variable, given the associated unit and input variables) increases less than proportionally versus an increase of the supervision cost per training example, the minimum (large-sample approximation of the) generalization error (conditioned on the training input data) is obtained in correspondence of the smallest supervision cost per example (hence, of the largest number of examples); when that precision increases more than proportionally versus an increase of the supervision cost per example, the optimal supervision cost per example is the largest one (which corresponds to the smallest number of examples). Differently from (Gnecco and Nutarelli 2019, 2019, 2020), in the analysis made in the present work, the number of training examples can be varied either by increasing the number of observations per unit, or the number of units, or both. In summary, the results of the analyses made in (Gnecco and Nutarelli 2019, 2019, 2020) and, for a different and more complex regression model, in this paper, highlight that increasing the training set size is not always beneficial, if a smaller number of more reliable data can be collected. Hence, not only the quantity of data, but of course, also their quality matters. This looks particularly relevant when the data collection process can be designed before data are actually collected.

The paper is structured as follows. Section 2 provides a background on the fixed effects generalized least squares panel data model. Section 3 presents the analysis of its conditional generalization error, and of the large-sample approximation of the latter with respect to time. Section 4 formulates and solves an optimization problem we propose in order to provide an optimal trade-off between training set size and precision of supervision for the fixed effects generalized least squares panel data model, using the large-sample approximation above. Section 5 presents some numerical results, which validate the theoretical ones. Finally, Sect. 6 discusses some possible applications and extensions of the theoretical results obtained in the work. Some technical proofs and remarks about the extension of the analysis made in the paper to other large-sample settings are reported in the Appendices.

2 Background

In this section, we recall some basic facts about the following Fixed Effects Generalized Least Squares (FEGLS) panel data model (see, e.g., Wooldridge 2002, Chapter 10). Specifically, we refer to the following model:

$$y_{n,t} := \eta_n + \underline{\beta}' \underline{x}_{n,t}, \quad \text{for } n = 1, \dots, N, t = 1, \dots, T, \tag{1}$$

where the outputs $y_{n,t} \in \mathbb{R}$ are scalars, whereas the inputs $\underline{x}_{n,t}$ ($n = 1, \dots, N, t = 1, \dots, T$) are column vectors in $\mathbb{R}^p$, and are modeled as random vectors. The superscript $'$ denotes transposition. The parameters of the model are the individual constants η_n ($n = 1, \dots, N$), one for each unit, and the column vector $\underline{\beta} \in \mathbb{R}^p$. The constants η_n are also called fixed effects. Eq. (1) represents a balanced panel data model, in which each unit n is associated with the same number T of outputs, each one at a different time t. The model represents the case in which the input/output relationship is linear, and different units, which are observed at the times $t = 1, \dots, T$, are associated with possibly different constants.

Note that the outputs $y_{n,t}$ are actually unavailable; only their noisy measurements $z_{n,t}$ can be obtained, which are assumed to be generated according to the following additive noise model:

$$z_{n,t} := y_{n,t} + \varepsilon_{n,t}\,, \text{ for } n = 1, \dots, N, t = 1, \dots, T\,, \tag{2}$$

where, for any n, the $\varepsilon_{n,t}$ are identically distributed and possibly dependent random variables, having mean 0, and are further independent from all the $\underline{x}_{n,t}$. For any two units $n \neq m$ and any two time instants $t_1, t_2 \in \{1, \dots, T\}$, ε_{n,t_1} and ε_{m,t_2} are assumed to be independent. Hence, only the possibility of temporal dependence for the measurement errors associated with the same unit is considered in the following, in line with several works in the literature (see, e.g., Bhargava et al. (1982) and (Wooldridge 2002, Section 10.5.5)).

For each unit n, let $X_n \in \mathbb{R}^{T \times p}$ be the matrix whose rows are the transposes of the $\underline{x}_{n,t}$; further, let $\underline{z}_n \in \mathbb{R}^T$ be the column vector which collects the noisy measurements $z_{n,t}$, and $\underline{\varepsilon}_n \in \mathbb{R}^T$ the column vector which collects the measurement noises $\varepsilon_{n,t}$. The input/corrupted output pairs $(\underline{x}_{n,t}, \underline{z}_{n,t})$, for $n = 1, \dots, N$, $t = 1, \dots, T$, are used to train the FEGLS model, i.e., to estimate its parameters.

The following first-order serial covariance form is assumed (see, e.g., Bhargava et al. (1982) and Wooldridge (2002, Section 10.5.5)) for the (unconditional) covariance matrix of the vector of measurement noises associated with the n-th unit[1], where $\sigma > 0$ and $\rho \in (-1, 1)$ hold (here, $\mathbb{E}$ denotes the expectation operator):

$$\Lambda := \sigma^2 \Psi := \mathrm{Var}\big(\underline{\varepsilon}_n\big) = \mathbb{E}\{\underline{\varepsilon}_n \underline{\varepsilon}_n'\} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-2} & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{T-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{T-1} & \rho^{T-2} & \cdots & \rho^2 & \rho & 1 \end{bmatrix} \in \mathbb{R}^{T \times T}\,, \tag{3}$$

which is a symmetric and positive-definite matrix. In other words, the measurement noise is assumed to be generated by a first-order autoregressive ($AR(1)$) process (Ruud 2000, Section 25.2). In the particular case of uncorrelated ($\rho = 0$) and independent measurement noises, one obtains the model considered in Gnecco and Nutarelli (2020).

Let the matrix $Q_T \in \mathbb{R}^{T \times T}$ be defined as

$$Q_T := I_T - \frac{1}{T} \underline{1}_T \underline{1}_T'\,, \tag{4}$$

where $I_T \in \mathbb{R}^{T \times T}$ is the identity matrix, and $\underline{1}_T \in \mathbb{R}^T$ a column vector whose elements are all equal to 1. One can check that Q_T is a symmetric and idempotent matrix (i.e., $Q_T' = Q_T = Q_T^2$), and its eigenvalues are 0 with multiplicity 1, and 1 with multiplicity $T - 1$. Hence, for each unit n, one can define

[1] An important implication of first-order serial covariance in noise terms is the unreliability of classical test statistics, based on the assumption of uncorrelated noises (see, e.g., Im et al. (1999)). To deal with this issue, the usual approach adopted in the literature consists in explicitly taking into account the form (3) for the covariance matrix of the zero-mean vector of measurement noises associated with each unit.

$$\ddot{X}_n := Q_T X_n = \begin{bmatrix} \underline{x}_{n,1} - \frac{1}{T}\sum_{t=1}^{T} \underline{x}_{n,t} \\ \underline{x}_{n,2} - \frac{1}{T}\sum_{t=1}^{T} \underline{x}_{n,t} \\ \cdots \\ \underline{x}_{n,T} - \frac{1}{T}\sum_{t=1}^{T} \underline{x}_{n,t} \end{bmatrix}, \tag{5}$$

$$\ddot{\underline{z}}_n := Q_T \underline{z}_n = \begin{bmatrix} z_{n,1} - \frac{1}{T}\sum_{t=1}^{T} z_{n,t} \\ z_{n,2} - \frac{1}{T}\sum_{t=1}^{T} z_{n,t} \\ \cdots \\ z_{n,T} - \frac{1}{T}\sum_{t=1}^{T} z_{n,t} \end{bmatrix}, \tag{6}$$

and

$$\ddot{\underline{\varepsilon}}_n := Q_T \underline{\varepsilon}_n = \begin{bmatrix} \varepsilon_{n,1} - \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{n,t} \\ \varepsilon_{n,2} - \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{n,t} \\ \cdots \\ \varepsilon_{n,T} - \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{n,t} \end{bmatrix}, \tag{7}$$

which represent, respectively, the matrix of time de-meaned training inputs, the vector of time de-meaned corrupted training outputs, and the vector of time de-meaned measurements noises. The goal of time de-meaning is to obtain a derived dataset where the fixed effects are removed, making it possible to estimate first the vector β, then - turning back to the original dataset - the fixed effects η_n. The covariance matrix $\mathbb{E}\{\ddot{\underline{\varepsilon}}_n \ddot{\underline{\varepsilon}}_n'\}$ has the expression

$$\Omega := \sigma^2 \Phi := \mathrm{Var}\left(\ddot{\underline{\varepsilon}}_n \ddot{\underline{\varepsilon}}_n'\right) = \mathbb{E}\{\ddot{\underline{\varepsilon}}_n \ddot{\underline{\varepsilon}}_n'\} = Q_T \mathbb{E}\{\underline{\varepsilon}_n \underline{\varepsilon}_n'\} Q_T' = Q_T \Lambda Q_T' = \sigma^2 Q_T \Psi Q_T', \tag{8}$$

which is symmetric and positive semi-definite, and has rank $T-1 < T$ (Wooldridge 2002). Although this deficient rank prevents the application of the most usual approach to Generalized Least Squares (GLS) estimation, based on the inversion of the covariance matrix Ω (which in this case cannot be inverted), one can still apply GLS by projecting Eqs. (1) and (2) onto the orthogonal complement L of the vector $\underline{1}_T$ by using Q_T, then solving a standard GLS problem on L (Aitken 1936). This is formally obtained by replacing the inverse of the covariance matrix with its Moore-Penrose pseudoinverse[2] (denoted by Ω^+), as made in the context of FEGLS estimation in Kiefer (1980, Im et al. (1999). More precisely, assuming the invertibility of the matrix $\sum_{n=1}^{N} \ddot{X}_n' \Omega^+ \ddot{X}_n$ (see Remark 3.1 for a justification of this assumption), the FEGLS estimate of $\underline{\beta}$ is

[2] It is recalled here from Strang (1993) that the Moore-Penrose pseudoinverse M^+ of a matrix $M \in \mathbb{R}^{T \times T}$ inverts a special restriction of the linear application represented by the matrix M, whose domain and codomain are restricted, respectively, to the row space of M and to the column space of M (which coincide in the case of a symmetric matrix). For a matrix $M \in \mathbb{R}^{T \times T}$ with singular value decomposition $M = U\Sigma V'$ (where $U, V \in \mathbb{R}^{T \times T}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{T \times T}$ is a diagonal matrix whose non-zero entries are the singular values of M), the singular value decomposition of its Moore-Penrose pseudoinverse is $M^+ = U\Sigma^+ V'$ (where $\Sigma^+ \in \mathbb{R}^{T \times T}$ is a diagonal matrix whose non-zero entries are the reciprocals of the singular values of M). Finally, in the particular case in which M is symmetric and positive semi-definite (e.g., when it is a covariance matrix), its singular values coincide with its positive eigenvalues, $U \in \mathbb{R}^{T \times T}$ is an orthogonal matrix whose columns are its eigenvectors, $V' = U^{-1}$, and M^+ is symmetric and positive semi-definite. The concept of Moore-Penrose pseudoinversion can be extended to the cases of rectangular matrices and matrices with complex entries, but such extensions are not needed in this work.

$$\hat{\underline{\beta}}_{FEGLS} = \left(\sum_{n=1}^{N} \ddot{X}_n' \Omega^{+} \ddot{X}_n \right)^{-1} \left(\sum_{n=1}^{N} \ddot{X}_n' \Omega^{+} \ddot{\underline{z}}_n \right). \tag{9}$$

The estimate $\hat{\underline{\beta}}_{FEGLS}$ in (9) can be interpreted as the GLS estimate of $\underline{\beta}$ obtained by replacing the original input/corrupted output training data with their de-meaned versions reported above. It is worth observing that the training input/corrupted output pairs $\left(\underline{x}_{n,t}, z_{n,t}\right)$ $(n = 1, \dots, N, t = 1, \dots, T)$ are all used to estimate $\underline{\beta}$.

Remark 2.1 Another commonly used approach to deal with the issue above is to drop one of the time periods from the analysis, in order to get an invertible covariance matrix. It can be rigorously proved (see, e.g. (Im et al. 1999, Theorem 4.3)) that this second approach is equivalent to the one based on the Moore-Penrose pseudoinverse (producing exactly the same FEGLS estimate), and that it does not matter which time period is dropped, as the resulting GLS estimator has always the same form. Therefore, dropping the last row of Q_T, one gets the matrix $\tilde{Q}_T \in \mathbb{R}^{(T-1)\times T}$, from which one obtains the matrix $\tilde{\ddot{X}}_n := \tilde{Q}_T X_n \in \mathbb{R}^{(T-1)\times p}$, the column vector $\tilde{\ddot{\underline{z}}}_n := \tilde{Q}_T \underline{z}_n \in \mathbb{R}^{T-1}$, and the column vector $\tilde{\ddot{\underline{\varepsilon}}}_n := \tilde{Q}_T \underline{\varepsilon}_n \in \mathbb{R}^{T-1}$. Moreover, denoting by $\tilde{X}_n \in \mathbb{R}^{(T-1)\times p}$, $\tilde{\underline{z}}_n \in \mathbb{R}^{T-1}$, and $\tilde{\underline{\varepsilon}}_n \in \mathbb{R}^{T-1}$ the matrix and the vectors obtained by removing the last row, respectively, from X_n, $\underline{z}_n$, and $\underline{\varepsilon}_n$, one gets

$$\tilde{\Omega} := \mathbb{E}\{\tilde{\ddot{\underline{\varepsilon}}}_n \tilde{\ddot{\underline{\varepsilon}}}_n'\} = \tilde{Q}_T \mathbb{E}\{\underline{\varepsilon}_n \underline{\varepsilon}_n'\} \tilde{Q}_T' = \tilde{Q}_T \Lambda \tilde{Q}_T', \tag{10}$$

which is, differently from Ω, an invertible matrix, with inverse $\tilde{\Omega}^{-1} = (\tilde{Q}_T \Lambda \tilde{Q}_T')^{-1}$. The resulting FEGLS estimate is

$$\hat{\underline{\beta}}_{FEGLS}^{alt} = \left(\sum_{n=1}^{N} \tilde{\ddot{X}}_n' \tilde{\Omega}^{-1} \tilde{\ddot{X}}_n \right)^{-1} \left(\sum_{n=1}^{N} \tilde{\ddot{X}}_n' \tilde{\Omega}^{-1} \tilde{\ddot{\underline{z}}}_n \right). \tag{11}$$

(see, e.g., Wooldridge (2002)). The FEGLS estimate $\hat{\underline{\beta}}_{FEGLS}$ and the alternative one $\hat{\underline{\beta}}_{FEGLS}^{alt}$ are actually identical (Im et al. 1999, Theorem 4.3). This equivalence is obtained by expressing such estimates in terms of the original variables before de-meaning, then exploiting the proof of (Im et al. 1999, Theorem 4.3), which shows that $Q_T' \Omega^{+} Q_T = \tilde{Q}_T' \tilde{\Omega}^{-1} \tilde{Q}_T$ (this still holds if an observation different from the last one is dropped, and $\tilde{Q}_T$ is redefined accordingly).

The FEGLS estimates of the η_n (also called fixed effects residuals (Wooldridge 2002)) are

$$\hat{\eta}_{n,FEGLS} := \frac{1}{T} \sum_{t=1}^{T} \left(z_{n,t} - \hat{\underline{\beta}}_{FEGLS}' \underline{x}_{n,t} \right). \tag{12}$$

They are obtained by subtracting the estimate $\hat{\underline{\beta}}_{FEGLS}' \underline{x}_{n,t}$ of $\underline{\beta}' \underline{x}_{n,t}$ from each corrupted output $z_{n,t}$, then performing an empirical average, limiting to training data associated with the unit n. The FEGLS estimates reported in Eq. (12) are motivated by the fact that the η_n are constants, whereas the $\varepsilon_{n,t}$ have mean 0.

By taking expectations, it readily follows from their definitions that the estimates (9) and (12) are conditionally unbiased with respect to the training input data $\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}$, i.e., that

$$\mathbb{E}\left\{\left(\underline{\hat{\beta}}_{FEGLS}-\underline{\beta}\right)\middle|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}=\underline{0}_p\,, \tag{13}$$

where $\underline{0}_p \in \mathbb{R}^p$ is a column vector whose elements are all equal to 0, and, for any $i = 1,\dots,N$,

$$\mathbb{E}\left\{\left(\hat{\eta}_{i,FEGLS}-\eta_i\right)\middle|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}=0\,. \tag{14}$$

Finally, the covariance matrix of $\underline{\hat{\beta}}_{FEGLS}$, conditioned on the training input data, is

$$\mathrm{Var}\left(\underline{\hat{\beta}}_{FEGLS}\middle|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right)=\left(\sum_{n=1}^{N}\ddot{X}_n'\Omega^{+}\ddot{X}_n\right)^{-1}\,. \tag{15}$$

3 Conditional generalization error and its large-sample approximation

The goal of this section is to analyze the generalization error associated with the FEGLS estimates (9) and (12), conditioned on the training input data, by providing its large-sample approximation. Then, in Sect. 4, the resulting expression is optimized, after choosing suitable models for the standard deviation σ of the measurement noise and for the time horizon, which is chosen in such a way it satisfies a suitable budget constraint.

First, we express the generalization error or expected risk for the i-th unit ($i = 1,\dots,N$), conditioned on the training input data, by

$$R_i\left(\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right):=\mathbb{E}\left\{\left(\hat{\eta}_{i,FEGLS}+\underline{\hat{\beta}}_{FEGLS}'\underline{x}_i^{test}-\eta_i-\underline{\beta}'\underline{x}_i^{test}\right)^2\middle|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}, \tag{16}$$

where $\underline{x}_i^{test} \in \mathbb{R}^p$ is independent from the training data. It is the expected mean squared error of the prediction of the output associated with a test input, conditioned on the training input data.

As shown in Appendix 1, we can express the conditional generalization error (16) as follows, highlighting its dependence on σ^2:

$$\begin{aligned}
&R_i\left(\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right)\\
&=\frac{\sigma^2}{T^2}\underline{1}_T'X_i\left(\sum_{n=1}^{N}\ddot{X}_n'\Phi^{+}\ddot{X}_n\right)^{-1}X_i'\underline{1}_T+\frac{\sigma^2}{T^2}\underline{1}_T'\Psi\underline{1}_T\\
&\quad-\frac{2\sigma^2}{T^2}\underline{1}_T'X_i\left(\sum_{n=1}^{N}\ddot{X}_n'\Phi^{+}\ddot{X}_n\right)^{-1}\ddot{X}_i'\Phi^{+}Q_T\Psi\underline{1}_T+\sigma^2\mathbb{E}\left\{\left(\underline{x}_i^{test}\right)'\left(\sum_{n=1}^{N}\ddot{X}_n'\Phi^{+}\ddot{X}_n\right)^{-1}\underline{x}_i^{test}\middle|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}\\
&\quad-\frac{2\sigma^2}{T}\underline{1}_T'X_i\left(\sum_{n=1}^{N}\ddot{X}_n'\Phi^{+}\ddot{X}_n\right)^{-1}\mathbb{E}\left\{\underline{x}_i^{test}\right\}+\frac{2\sigma^2}{T}\left(Q_T\Psi\underline{1}_T\right)'\Phi^{+}\ddot{X}_i\left(\sum_{n=1}^{N}\ddot{X}_n'\Phi^{+}\ddot{X}_n\right)^{-1}\mathbb{E}\left\{\underline{x}_i^{test}\right\},
\end{aligned} \tag{17}$$

where some computations (reported in Appendix 1) show that

$$\underline{1}_T' \Psi \underline{1}_T = T + 2T\left(\frac{1-\rho^T}{1-\rho} - 1\right) - \frac{2\rho}{1-\rho}\big(-(T-1)\rho^{T-1} + \rho^{T-2} + \rho^{T-3} + \cdots + 1\big)\,, \tag{18}$$

and

$$\underline{v}_T := Q_T \Psi \underline{1}_T$$
$$= \begin{bmatrix} \cancel{1} + \rho + \rho^2 + \rho^3 + \rho^4 + \cdots & + \rho^{T-1} - \cancel{1} - 2\frac{1-\rho^T}{1-\rho} + 2 + \frac{2\rho[-(T-1)\rho^{T-1}+\rho^{T-2}+\rho^{T-3}+\cdots+1]}{T(1-\rho)} \\ \rho + \cancel{1} + \rho + \rho^2 + \rho^3 + \cdots & + \rho^{T-2} - \cancel{1} - 2\frac{1-\rho^T}{1-\rho} + 2 + \frac{2\rho[-(T-1)\rho^{T-1}+\rho^{T-2}+\rho^{T-3}+\cdots+1]}{T(1-\rho)} \\ \rho^2 + \rho + \cancel{1} + \rho + \rho^2 + \cdots & + \rho^{T-3} - \cancel{1} - 2\frac{1-\rho^T}{1-\rho} + 2 + \frac{2\rho[-(T-1)\rho^{T-1}+\rho^{T-2}+\rho^{T-3}+\cdots+1]}{T(1-\rho)} \\ \cdots & \cdots \qquad \cdots \\ \rho^{T-3} + \rho^{T-4} + \cdots + \rho + \cancel{1} + \rho & + \rho^2 - \cancel{1} - 2\frac{1-\rho^T}{1-\rho} + 2 + \frac{2\rho[-(T-1)\rho^{T-1}+\rho^{T-2}+\rho^{T-3}+\cdots+1]}{T(1-\rho)} \\ \rho^{T-2} + \rho^{T-3} + \cdots + \rho^2 + \rho + \cancel{1} & + \rho - \cancel{1} - 2\frac{1-\rho^T}{1-\rho} + 2 + \frac{2\rho[-(T-1)\rho^{T-1}+\rho^{T-2}+\rho^{T-3}+\cdots+1]}{T(1-\rho)} \\ \rho^{T-1} + \rho^{T-2} + \cdots + \rho^3 + \rho^2 + \rho & + \cancel{1} - \cancel{1} - 2\frac{1-\rho^T}{1-\rho} + 2 + \frac{2\rho[-(T-1)\rho^{T-1}+\rho^{T-2}+\rho^{T-3}+\cdots+1]}{T(1-\rho)} \end{bmatrix}. \tag{19}$$

Next, we obtain a large-sample approximation of the conditional generalization error (17) with respect to T, for a fixed number of units N. Such an approximation is useful, e.g., in the application of the model to macroeconomics data, for which it is common to investigate the case of a large horizon T.

Under mild conditions (e.g., if for the unit i the $\underline{x}_{i,t}$ are mutually independent, identically distributed, and have finite moments up to the order 4), the following convergences in probability[3] hold (their proofs are reported in Appendix 2):

$$\underset{T\to+\infty}{\text{plim}} \frac{1}{T}\underline{1}_T' X_i = \left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)'\,, \tag{20}$$

$$\underset{T\to+\infty}{\text{plim}} \frac{1}{T}\ddot{X}_i' \Phi^+ Q_T \Psi \underline{1}_T = \underline{0}_p\,. \tag{21}$$

Similarly, if for each fixed unit n the $\underline{x}_{n,t}$ are mutually independent, identically distributed[4], and have finite moments up to the order 4, and one makes the additional assumption (whose validity is discussed extensively in Appendix 2) that

$$\lim_{T\to\infty} \|\Phi^+ - Q_T \Psi^{-1} Q_T'\|_2 = 0 \tag{22}$$

(where, for a symmetric matrix $M \in \mathbb{R}^{T\times T}$, $\|M\|_2 = \max_{t=1,\ldots,T} |\lambda_t(M)|$ denotes its spectral norm), then also the following convergence in probability holds:

$$\underset{T\to+\infty}{\text{plim}} \frac{1}{T}\sum_{n=1}^{N} \ddot{X}_n' \Phi^+ \ddot{X}_n = A_N\,, \tag{23}$$

[3] We recall that a sequence of random real matrices M_T, $T = 1, 2, \ldots$, converges in probability to the real matrix M if, for every $\varepsilon > 0$, $\text{Prob}\big(\|M_T - M\| > \varepsilon\big)$ (where $\|\cdot\|$ is an arbitrary matrix norm) tends to 0 as T tends to $+\infty$. In this case, we write $\underset{T\to+\infty}{\text{plim}}\, M_T = M$.

[4] This does not exclude the possibility for the $\underline{x}_{n,t}$ and $\underline{x}_{m,t}$ associated with different units n and m to be dependent/not identically distributed.

where

$$A_N = A_N' := \frac{1+\rho^2}{1-\rho^2}\sum_{n=1}^{N} \mathbb{E}\left\{\left(\underline{x}_{n,1} - \mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)\left(\underline{x}_{n,1} - \mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)'\right\} \tag{24}$$

is a symmetric and positive semi-definite matrix. In the following, its positive definiteness (hence, its invertibility) is also assumed.

Remark 3.1 The existence of the probability limit (23) and the assumed positive definiteness of the matrix A_N guarantee that the invertibility of the matrix $\sum_{n=1}^{N} \ddot{X}_n' \Phi^+ \ddot{X}_n$ (see the invertibility assumption before Eq. (9)) holds with probability near 1 for large T.

When (20), (21), and (23) hold, inserting such probability limits in Eq. (17), one gets the following large-sample approximation of the conditional generalization error (17) with respect to T:

$$\begin{aligned}(17) \simeq & \frac{\sigma^2}{T}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)' A_N^{-1}\mathbb{E}\left\{\underline{x}_{i,1}\right\} + \frac{\sigma^2}{T}\frac{1+\rho}{1-\rho} \\ & + \frac{\sigma^2}{T}\mathbb{E}\left\{\left(\underline{x}_i^{test}\right)' A_N^{-1}\underline{x}_i^{test}\right\} - 2\frac{\sigma^2}{T}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)' A_N^{-1}\mathbb{E}\left\{\underline{x}_i^{test}\right\} \\ = & \frac{\sigma^2}{T}\left(\frac{1+\rho}{1-\rho} + \mathbb{E}\left\{\left\| A_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\} - \underline{x}_i^{test}\right)\right\|_2^2\right\}\right),\end{aligned} \tag{25}$$

where, for a vector $\underline{v} \in \mathbb{R}^p$, $\|\underline{v}\|_2$ denotes its l_2 (Euclidean) norm, and $A_N^{-\frac{1}{2}}$ is the principal square root (i.e., the symmetric and positive definite square root) of the symmetric and positive definite matrix A_N^{-1}. Eq. (25) is obtained taking into account that, as a consequence of the Continuous Mapping Theorem (Florescu 2015, Theorem 7.33), the probability limit of the product of two random variables equals the product of their probability limits, when the latter two exist. By doing this, the third and sixth terms of Eq. (17) cancel out due to Eq. (21), whereas the second term is computed using Eq. (18).

Interestingly, the large-sample approximation (25) has the form $\frac{\sigma^2}{T}K_i$, where

$$K_i := \left(\frac{1+\rho}{1-\rho} + \mathbb{E}\left\{\left\| A_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\} - \underline{x}_i^{test}\right)\right\|_2^2\right\}\right) \tag{26}$$

is a positive constant (possibly, a different constant for each unit i). This simplifies the analysis of the trade-off between training set size and precision of supervision performed in the next section, since one does not need to compute the exact expression of K_i to find the optimal trade-off.

In Appendix 3, an extension of the analysis made above is presented, by considering, respectively, the case of large N, and the one in which both N and T are large.

4 Optimal trade-off between training set size and precision of supervision for the fixed effects generalized least squares panel data model under the large-sample approximation

In this section, we are interested in optimizing the large-sample approximation (25) of the conditional generalization error when the variance σ^2 is modeled as a decreasing function of the supervision cost per example c, and there is an upper bound C on the total supervision cost NTc associated with the whole training set. In the analysis, N is fixed, and T is chosen as $\left\lfloor \frac{C}{Nc} \right\rfloor$. Moreover, the supervision cost per example c is allowed to take values on the interval $[c_{\min}, c_{\max}]$, where $0 < c_{\min} < c_{\max}$, so that the resulting T belongs to $\left\{ \left\lfloor \frac{C}{Nc_{\max}} \right\rfloor, \dots, \left\lfloor \frac{C}{Nc_{\min}} \right\rfloor \right\}$. In the following, C is supposed to be sufficiently large, so that the large-sample approximation (25) can be assumed to hold for every $c \in [c_{\min}, c_{\max}]$.

Consistently with (Gnecco and Nutarelli 2019, 2019, 2020), we adopt the following model for the variance σ^2, as a function of the supervision cost per example c:

$$\sigma^2(c) = kc^{-\alpha}\,, \tag{27}$$

where $k, \alpha > 0$. For $0 < \alpha < 1$, the precision of each supervision is characterized by "decreasing returns of scale" with respect to its cost because, if one doubles the supervision cost per example c, then the precision $1/\sigma^2(c)$ becomes less than two times its initial value (or equivalently, the variance $\sigma^2(c)$ becomes more than one half its initial value). Conversely, for $\alpha > 1$, there are "increasing returns of scale" because, if one doubles the supervision cost per example c, then the precision $1/\sigma^2(c)$ becomes more than two times its initial value (or equivalently, the variance $\sigma^2(c)$ becomes less than one half its initial value). The case $\alpha = 1$ is intermediate and refers to "constant returns of scale". In all the cases above, the precision of each supervision increases by increasing the supervision cost per example c. Finally, it is worth observing that, according to the model (3) for the covariance matrix of the vector of measurement noises, the correlation coefficient between successive measurement noises does not depend on c.

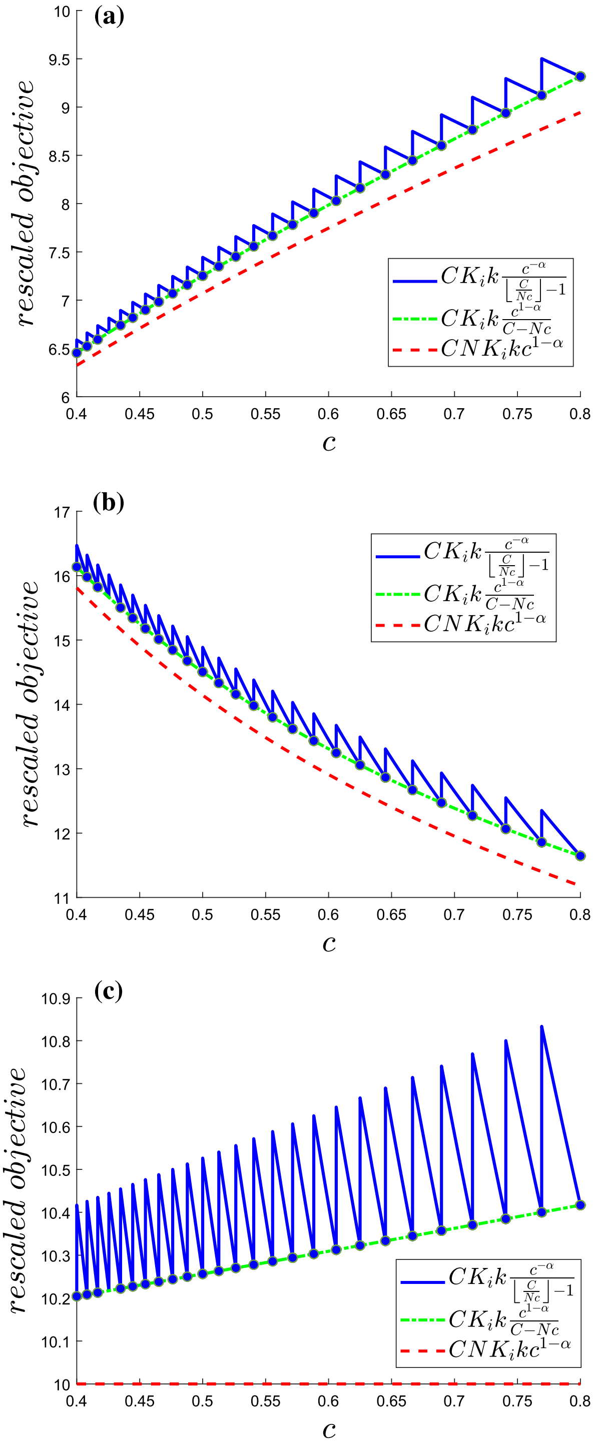
Concluding, under the assumptions above, the optimal trade-off between the training set size and the precision of supervision for the fixed effects generalized least squares panel data model is modeled by the following optimization problem:

$$\underset{c \in [c_{\min}, c_{\max}]}{\text{minimize}}\, K_i k \frac{c^{-\alpha}}{\left\lfloor \frac{C}{Nc} \right\rfloor - 1}\,. \tag{28}$$

By a similar argument as in the proof of Gnecco and Nutarelli (2019, Proposition 3.2), when C is sufficiently large, the objective function $CK_i k \frac{c^{-\alpha}}{\left\lfloor \frac{C}{Nc} \right\rfloor - 1}$ of the optimization problem (28), rescaled by the multiplicative factor C, can be approximated, with a negligible error in the maximum norm on $[c_{\min}, c_{\max}]$, by $NK_i kc^{1-\alpha}$. In order to illustrate this issue, Fig. 1 shows the behavior of the rescaled objective functions $CK_i k \frac{c^{-\alpha}}{\left\lfloor \frac{C}{Nc} \right\rfloor - 1}$ and $NK_i kc^{1-\alpha}$ for the three cases $0 < \alpha = 0.5 < 1$, $\alpha = 1.5 > 1$, and $\alpha = 1$ (the values of the other parameters are $k = 0.5$, $K_i = 2$, $N = 10$, $C = 200$, $c_{\min} = 0.4$, and $c_{\max} = 0.8$). The additional approximation $CNK_i k \frac{c^{1-\alpha}}{C - Nc}$ (which differs negligibly from $NK_i kc^{1-\alpha}$ for large C) is also reported in the figure.

Concluding, under the approximation above, one can replace the optimization problem (28) with

Fig. 1 Plots of the rescaled objective functions $CK_ik\frac{c^{-\alpha}}{\left\lfloor \frac{C}{Nc} \right\rfloor - 1}$, $CNK_ikc^{1-\alpha}$, $CNK_ik\frac{c^{1-\alpha}}{C-Nc}$, for $\alpha = 0.5$ (**a**), $\alpha = 1.5$ (**b**), and $\alpha = 1$ (**c**). The values of the other parameters are reported in the text



$$\underset{c \in [c_{\min}, c_{\max}]}{\text{minimize}} \, N K_i k c^{1-\alpha} , \tag{29}$$

whose optimal solutions c° have the following expressions:

1. if $0 < \alpha < 1$ (“decreasing returns of scale”): $c^\circ = c_{\min}$;
2. if $\alpha > 1$ (“increasing returns of scale”): $c^\circ = c_{\max}$;
3. if $\alpha = 1$ (“constant returns of scale”): $c^\circ =$ any cost c in the interval $[c_{\min}, c_{\max}]$.

In summary, the results of the analysis show that, in the case of “decreasing returns of scale”, “many but bad” examples are associated with a smaller conditional generalization error than “few but good” ones. The opposite occurs for “increasing returns of scale”, whereas the case of “constant returns of scale” is intermediate. These results are qualitatively in line with the ones obtained in (Gnecco and Nutarelli 2019, 2019, 2020) for simpler linear regression problems, to which different regression algorithms were applied (respectively, ordinary least squares, weighted least squares, and fixed effects ordinary least squares). This depends on the fact that, in all these cases, the large-sample approximation of the conditional generalization error has the same functional form $\frac{\sigma^2}{T} K_i$ (although different positive constants K_i are involved in the various cases).

One can observe that, in order to discriminate among the three cases of the analysis reported above, one does not need to know the exact values of the constants ρ, k, K_i, and N. Moreover, to discriminate between the first two cases, it is not necessary to know the exact value of the positive constant α (indeed, it suffices to know if α belongs, respectively, to the interval (0, 1) or the one $(1, +\infty)$). Interestingly, no precise knowledge of the probability distributions of the input examples (one for each unit) is needed. In particular, different probability distributions may be associated with different units, without affecting the results of the analysis. Finally, the same conclusions as above are reached if the objective function in (29) is replaced by the summation of the large-sample approximation of the conditional generalization error over all the N units. In that case, the constant K_i in (29) is replaced by $K := \sum_{i=1}^{N} K_i$.

5 Numerical results

In this section, the theoretical results obtained in the paper are tested through simulations. For each c, the following empirical approximation of the summation of the generalization error over all the units, conditioned on the training input data, is adopted. It is based on $\mathcal{N}^{tr}$ training sets and N_i^{test} test examples for each unit i ($i = 1, \ldots, N$), hence on a total number $N^{test} = \sum_{i=1}^{N} N_i^{test}$ of test examples:

$$\begin{aligned} &\sum_{i=1}^{N} \mathbb{E}\left\{ \left(\hat{\eta}_{i,FEGLS} + \underline{\hat{\beta}}'_{FEGLS} \underline{x}_i^{test} - \eta_i - \underline{\beta}' \underline{x}_i^{test} \right)^2 \Big| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T} \right\} \\ &\simeq \frac{1}{N^{test}} \sum_{i=1}^{N} \sum_{h=1}^{N_i^{test}} \frac{1}{\mathcal{N}^{tr}} \sum_{j=1}^{\mathcal{N}^{tr}} \left(\hat{\eta}_{i,FEGLS}^{j} + \left(\underline{\hat{\beta}}_{FEGLS}^{j} \right)' \underline{x}_{i,h}^{test} - \eta_i - \underline{\beta}' \underline{x}_{i,h}^{test} \right)^2 . \end{aligned} \tag{30}$$

In Eq. (30), $(\underline{x}_{i,h}^{test}, y_{i,h}^{test})$ is the h-th generated test example for the unit i, and $\underline{\hat{\beta}}_{FEGLS}^{j}$ is the estimate of the vector $\underline{\beta}$ obtained using the j-th generated training set. Similarly, $\hat{\eta}_{i,FEGLS}^{j}$ is the estimate of the individual constant η_i obtained using the j-th generated training set. For each choice of c, all the $\mathcal{N}^{tr}$ generated training sets share the same training input data matrices X_n, but differ in the random choice of the measurement noise. The number of rows in each matrix X_n is increased when c is reduced from $c_{\max}$ to $c_{\min}$, by increasing the number of observations T. For a fair comparison, when doing this, the rows already present in each matrix X_n are kept fixed. Finally, the same test examples (generated independently from the training sets) are used to assess the performance of the fixed effects generalized least squares estimates for different costs per example c.

For the simulations, we choose $N = 20$ for the number of units, $p = 5$ for the number of features, $c_{\min} = 2$, $c_{\max} = 4$, $\mathcal{N}^{tr} = 100$ for the number of training sets, $N_i^{test} = 50$ for the number of test examples per unit (hence the total number of test examples is $N^{test} = 1000$). The number of training examples per unit is $T = 50$ for $c = c_{\min}$, and $T = 25$ for $c = c_{\max}$. In this way, the (upper bound on the) total supervision cost is $C = 2000$ for both cases. Without loss of generality, the constant k in the model (27) of the variance of the supervision cost is assumed to be equal to 1. The components of the parameter vector $\underline{\beta}$ are generated randomly and independently according to a uniform distribution on $[-\overline{1}, 1]$, obtaining

$$\underline{\beta} = [-0.8562, 0.6837, 0.2640, -0.0038, -0.0598]' . \tag{31}$$

Similarly, the fixed effects η_n (for $n = 1, \ldots, N$) are generated randomly and independently according to a uniform distribution on $[-1, 1]$, obtaining the vector

$$\begin{aligned}\underline{\eta} = \Big[& -0.2330, -0.2779, -0.0434, -0.9707, 0.6848, 0.0720, -0.2033, -0.6877, 0.5967, -0.7895, \\ & 0.6500, 0.9717, 0.9673, -0.1443, -0.4211, 0.3109, 0.5189, 0.4709, 0.4414, -0.8382 \Big]' \in \mathbb{R}^N .\end{aligned} \tag{32}$$

For both training and test sets, the input data associated with each unit are generated as realizations of a multivariate Gaussian distribution with mean $\underline{0}$ and covariance matrix

$$\mathrm{Var}\left(\underline{x}_{n,t}\right) = \mathrm{Var}\left(\underline{x}_i^{test}\right) = \begin{bmatrix} 1.4016 & 0.8086 & 1.2594 & 0.9866 & 0.6206 \\ 0.8086 & 0.9988 & 0.9518 & 1.2044 & 0.5003 \\ 1.2594 & 0.9518 & 1.9087 & 1.5945 & 0.7120 \\ 0.9866 & 1.2044 & 1.5945 & 1.9089 & 0.8294 \\ 0.6206 & 0.5003 & 0.7120 & 0.8294 & 0.4776 \end{bmatrix}, \tag{33}$$

which is symmetric and positive definite. This covariance matrix has been generated by setting $\mathrm{Var}\left(\underline{x}_{n,t}\right) = \mathrm{Var}\left(\underline{x}_i^{test}\right) = A_x A_x'$, where the elements of $A_x \in \mathbb{R}^{p \times p}$ have been randomly and independently generated according to a uniform probability density on the interval [0,1]. The parameter ρ in the covariance matrix (3) of the zero-mean vector of measurement noises (modeled in the simulations by a multivariate Gaussian distribution) is chosen to be equal to 0.3. As a robustness check, the whole procedure is repeated 100 times.

The results of the analysis are displayed in Tables 1 (for $\alpha = 0.5$), 2 (for $\alpha = 1.5$), and 3 (for $\alpha = 1$). Each table reports the results obtained in each repetition for $c = c_{\min}$ and $c = c_{\max}$. The total simulation time (for a MATLAB 9.4 implementation of the procedure) is of about 501 sec on a notebook with a 2.30 GHz Intel(R) Core(TM) i5-4200U CPU and 6 GB of RAM. A statistical analysis of the elements of the tables leads to the following conclusions:

1. for $\alpha = 0.5$ (Table 1), the application of a one-sided Wilcoxon matched-pairs signed-rank test (Barlow 1989, Sect. 9.2.3) rejects the null hypothesis that the difference between the approximated performance index from Eq. (30) for $c = c_{\max}$ and the one for $c = c_{\min}$ has a symmetric distribution around its median and that median is smaller than or equal to 0 (p-value=$1.9780 \cdot 10^{-18}$, significance level set to 0.05);
2. for $\alpha = 1.5$ (Table 2), the application of a one-sided Wilcoxon matched-pairs signed-rank test rejects the null hypothesis that the same difference as above has a symmetric distribution around its median and that median is larger than or equal to 0 (p-value=$1.9780 \cdot 10^{-18}$, significance level set to 0.05);
3. for $\alpha = 1$ (Table 3), the application of a two-sided Wilcoxon matched-pairs signed-rank test fails to reject the null hypothesis that the same difference as above has a symmetric distribution around its median and that median is equal to 0 (p-value=0.4453, significance level set to 0.05).

Concluding, the tables show that the simulation results are in perfect agreement with the theoretical ones, leading to the same conclusions. Interestingly, this holds even though relatively small values for T have been chosen for the simulations.

6 Discussion and possible extensions

Up to the authors' knowledge, the analysis and the optimization, made in the present article, of the conditional generalization error in regression as a trade-off between training set size and precision of supervision, has been carried out only rarely in the literature. In particular, the authors believe that it was never addressed for the fixed effects generalized least squares panel data model. Nevertheless, the methodology used in the present article is similar to the one exploited in the context of the optimization of sample survey design, in which some parameters of the design are optimized to minimize, e.g., the sampling variance (see, for instance, the classical solution provided by Neyman allocation for optimal stratified sampling design, in case the dataset has a fixed size (Groves et al. 2004). It also shares some similarity to the approach used in Nguyen et al. (2009) - in a context, however, in which linear regression is marginally involved, since only arithmetic averages of measurement results are considered - for the optimal design of measurement devices. Finally, the present article can also be related to some recent works which, according to an emerging trend in the literature, combine methods from machine learning, optimization, and econometrics (Athey and Imbens 2016; Bargagli Stoffi and Gnecco 2018, 2020; Varian 2014) (e.g., the generalization error - which is typically considered in machine learning, and optimized by solving suitable optimization problems - is not investigated in the classical analysis of the fixed effects generalized least squares panel data model (Wooldridge 2002, Chapter 10)). In this way, the interaction between machine learning and optimization—which appears commonly in the literature (Bennett and Parrado-Hernández 2006; Bianchini et al. 1998; Gnecco et al. 2013; Gori 2017; Özöğür-Akyüz et al. 2011; Sra et al. 2011)—is extended to the econometrics field.

For what concerns practical applications, the theoretical results obtained in the analysis made in this work could be applied to the acquisition design of fixed effects panel data in both microeconometrics and macroeconometrics (Greene 2003, Chapter 13). A semi-artificial validation on existing datasets could also be performed by inserting artificial noise

Table 1 For $\alpha = 0.5$: values of the approximated performance index from Eq. (30) for the 100 repetitions of the simulation procedure

(Repetition number) Approximated performance index from Eq. (30)					
$c = c_{\min}$	(1) $5.461 \cdot 10^{-4}$	(2) $6.127 \cdot 10^{-4}$	(3) $5.710 \cdot 10^{-4}$	(4) $5.638 \cdot 10^{-4}$	(5) $5.792 \cdot 10^{-4}$
	(6) $5.535 \cdot 10^{-4}$	(7) $6.084 \cdot 10^{-4}$	(8) $5.756 \cdot 10^{-4}$	(9) $5.976 \cdot 10^{-4}$	(10) $5.669 \cdot 10^{-4}$
	(11) $5.496 \cdot 10^{-4}$	(12) $5.562 \cdot 10^{-4}$	(13) $6.043 \cdot 10^{-4}$	(14) $5.762 \cdot 10^{-4}$	(15) $6.391 \cdot 10^{-4}$
	(16) $6.071 \cdot 10^{-4}$	(17) $6.007 \cdot 10^{-4}$	(18) $5.925 \cdot 10^{-4}$	(19) $5.600 \cdot 10^{-4}$	(20) $6.244 \cdot 10^{-4}$
	(21) $5.791 \cdot 10^{-4}$	(22) $5.590 \cdot 10^{-4}$	(23) $5.230 \cdot 10^{-4}$	(24) $5.339 \cdot 10^{-4}$	(25) $5.342 \cdot 10^{-4}$
	(26) $5.713 \cdot 10^{-4}$	(27) $5.824 \cdot 10^{-4}$	(28) $5.941 \cdot 10^{-4}$	(29) $5.727 \cdot 10^{-4}$	(30) $5.798 \cdot 10^{-4}$
	(31) $5.954 \cdot 10^{-4}$	(32) $5.819 \cdot 10^{-4}$	(33) $5.640 \cdot 10^{-4}$	(34) $5.779 \cdot 10^{-4}$	(35) $5.824 \cdot 10^{-4}$
	(36) $5.608 \cdot 10^{-4}$	(37) $5.565 \cdot 10^{-4}$	(38) $5.527 \cdot 10^{-4}$	(39) $5.981 \cdot 10^{-4}$	(40) $5.395 \cdot 10^{-4}$
	(41) $5.944 \cdot 10^{-4}$	(42) $6.110 \cdot 10^{-4}$	(43) $5.540 \cdot 10^{-4}$	(44) $5.490 \cdot 10^{-4}$	(45) $5.771 \cdot 10^{-4}$
	(46) $6.150 \cdot 10^{-4}$	(47) $5.492 \cdot 10^{-4}$	(48) $5.921 \cdot 10^{-4}$	(49) $5.552 \cdot 10^{-4}$	(50) $5.810 \cdot 10^{-4}$
	(51) $5.731 \cdot 10^{-4}$	(52) $6.018 \cdot 10^{-4}$	(53) $6.140 \cdot 10^{-4}$	(54) $5.836 \cdot 10^{-4}$	(55) $5.530 \cdot 10^{-4}$
	(56) $5.866 \cdot 10^{-4}$	(57) $5.661 \cdot 10^{-4}$	(58) $5.938 \cdot 10^{-4}$	(59) $5.795 \cdot 10^{-4}$	(60) $5.979 \cdot 10^{-4}$
	(61) $5.966 \cdot 10^{-4}$	(62) $5.882 \cdot 10^{-4}$	(63) $5.687 \cdot 10^{-4}$	(64) $5.718 \cdot 10^{-4}$	(65) $6.014 \cdot 10^{-4}$
	(66) $5.774 \cdot 10^{-4}$	(67) $5.872 \cdot 10^{-4}$	(68) $5.566 \cdot 10^{-4}$	(69) $5.678 \cdot 10^{-4}$	(70) $5.845 \cdot 10^{-4}$
	(71) $5.531 \cdot 10^{-4}$	(72) $5.446 \cdot 10^{-4}$	(73) $5.700 \cdot 10^{-4}$	(74) $6.055 \cdot 10^{-4}$	(75) $5.727 \cdot 10^{-4}$
	(76) $6.240 \cdot 10^{-4}$	(77) $5.616 \cdot 10^{-4}$	(78) $5.876 \cdot 10^{-4}$	(79) $6.031 \cdot 10^{-4}$	(80) $5.869 \cdot 10^{-4}$
	(81) $6.142 \cdot 10^{-4}$	(82) $5.764 \cdot 10^{-4}$	(83) $5.530 \cdot 10^{-4}$	(84) $5.901 \cdot 10^{-4}$	(85) $5.795 \cdot 10^{-4}$
	(86) $5.794 \cdot 10^{-4}$	(87) $5.818 \cdot 10^{-4}$	(88) $5.674 \cdot 10^{-4}$	(89) $5.512 \cdot 10^{-4}$	(90) $5.887 \cdot 10^{-4}$
	(91) $5.716 \cdot 10^{-4}$	(92) $6.050 \cdot 10^{-4}$	(93) $5.423 \cdot 10^{-4}$	(94) $5.883 \cdot 10^{-4}$	(95) $5.705 \cdot 10^{-4}$
	(96) $5.665 \cdot 10^{-4}$	(97) $5.732 \cdot 10^{-4}$	(98) $5.462 \cdot 10^{-4}$	(99) $5.896 \cdot 10^{-4}$	(100) $5.875 \cdot 10^{-4}$
$c = c_{\max}$	(1) $8.347 \cdot 10^{-4}$	(2) $8.193 \cdot 10^{-4}$	(3) $7.994 \cdot 10^{-4}$	(4) $8.132 \cdot 10^{-4}$	(5) $8.198 \cdot 10^{-4}$
	(6) $8.281 \cdot 10^{-4}$	(7) $7.627 \cdot 10^{-4}$	(8) $7.891 \cdot 10^{-4}$	(9) $8.281 \cdot 10^{-4}$	(10) $8.268 \cdot 10^{-4}$
	(11) $8.277 \cdot 10^{-4}$	(12) $8.097 \cdot 10^{-4}$	(13) $8.396 \cdot 10^{-4}$	(14) $8.187 \cdot 10^{-4}$	(15) $8.614 \cdot 10^{-4}$
	(16) $8.461 \cdot 10^{-4}$	(17) $8.299 \cdot 10^{-4}$	(18) $8.477 \cdot 10^{-4}$	(19) $8.171 \cdot 10^{-4}$	(20) $8.422 \cdot 10^{-4}$
	(21) $7.651 \cdot 10^{-4}$	(22) $8.068 \cdot 10^{-4}$	(23) $7.859 \cdot 10^{-4}$	(24) $8.034 \cdot 10^{-4}$	(25) $8.479 \cdot 10^{-4}$
	(26) $7.741 \cdot 10^{-4}$	(27) $7.839 \cdot 10^{-4}$	(28) $8.243 \cdot 10^{-4}$	(29) $7.620 \cdot 10^{-4}$	(30) $7.543 \cdot 10^{-4}$
	(31) $8.296 \cdot 10^{-4}$	(32) $8.280 \cdot 10^{-4}$	(33) $8.299 \cdot 10^{-4}$	(34) $8.115 \cdot 10^{-4}$	(35) $8.372 \cdot 10^{-4}$
	(36) $8.085 \cdot 10^{-4}$	(37) $8.362 \cdot 10^{-4}$	(38) $8.357 \cdot 10^{-4}$	(39) $8.585 \cdot 10^{-4}$	(40) $7.864 \cdot 10^{-4}$
	(41) $8.572 \cdot 10^{-4}$	(42) $8.098 \cdot 10^{-4}$	(43) $7.839 \cdot 10^{-4}$	(44) $7.941 \cdot 10^{-4}$	(45) $7.923 \cdot 10^{-4}$
	(46) $8.157 \cdot 10^{-4}$	(47) $8.743 \cdot 10^{-4}$	(48) $8.239 \cdot 10^{-4}$	(49) $8.181 \cdot 10^{-4}$	(50) $8.134 \cdot 10^{-4}$
	(51) $8.727 \cdot 10^{-4}$	(52) $8.600 \cdot 10^{-4}$	(53) $7.804 \cdot 10^{-4}$	(54) $8.078 \cdot 10^{-4}$	(55) $7.901 \cdot 10^{-4}$
	(56) $7.954 \cdot 10^{-4}$	(57) $7.811 \cdot 10^{-4}$	(58) $8.182 \cdot 10^{-4}$	(59) $8.339 \cdot 10^{-4}$	(60) $8.384 \cdot 10^{-4}$
	(61) $8.143 \cdot 10^{-4}$	(62) $8.129 \cdot 10^{-4}$	(63) $8.210 \cdot 10^{-4}$	(64) $8.319 \cdot 10^{-4}$	(65) $8.468 \cdot 10^{-4}$
	(66) $7.811 \cdot 10^{-4}$	(67) $8.211 \cdot 10^{-4}$	(68) $7.470 \cdot 10^{-4}$	(69) $8.128 \cdot 10^{-4}$	(70) $8.399 \cdot 10^{-4}$
	(71) $8.600 \cdot 10^{-4}$	(72) $8.537 \cdot 10^{-4}$	(73) $8.524 \cdot 10^{-4}$	(74) $8.117 \cdot 10^{-4}$	(75) $8.372 \cdot 10^{-4}$
	(76) $7.895 \cdot 10^{-4}$	(77) $8.114 \cdot 10^{-4}$	(78) $8.161 \cdot 10^{-4}$	(79) $8.537 \cdot 10^{-4}$	(80) $8.159 \cdot 10^{-4}$
	(81) $7.802 \cdot 10^{-4}$	(82) $8.178 \cdot 10^{-4}$	(83) $7.546 \cdot 10^{-4}$	(84) $7.922 \cdot 10^{-4}$	(85) $8.380 \cdot 10^{-4}$
	(86) $8.011 \cdot 10^{-4}$	(87) $8.541 \cdot 10^{-4}$	(88) $7.823 \cdot 10^{-4}$	(89) $8.026 \cdot 10^{-4}$	(90) $7.652 \cdot 10^{-4}$
	(91) $7.600 \cdot 10^{-4}$	(92) $7.859 \cdot 10^{-4}$	(93) $8.102 \cdot 10^{-4}$	(94) $8.599 \cdot 10^{-4}$	(95) $8.773 \cdot 10^{-4}$
	(96) $8.397 \cdot 10^{-4}$	(97) $8.105 \cdot 10^{-4}$	(98) $7.885 \cdot 10^{-4}$	(99) $8.061 \cdot 10^{-4}$	(100) $8.208 \cdot 10^{-4}$

Table 2 For $\alpha = 1.5$: values of the approximated performance index from Eq. (30) for the 100 repetitions of the simulation procedure

(Repetition number) Approximated performance index from Eq. (30)					
$c = c_{\min}$	(1) $2.966 \cdot 10^{-4}$	(2) $3.008 \cdot 10^{-4}$	(3) $2.901 \cdot 10^{-4}$	(4) $2.950 \cdot 10^{-4}$	(5) $3.152 \cdot 10^{-4}$
	(6) $3.021 \cdot 10^{-4}$	(7) $2.611 \cdot 10^{-4}$	(8) $2.922 \cdot 10^{-4}$	(9) $2.814 \cdot 10^{-4}$	(10) $3.028 \cdot 10^{-4}$
	(11) $2.881 \cdot 10^{-4}$	(12) $2.937 \cdot 10^{-4}$	(13) $3.060 \cdot 10^{-4}$	(14) $3.077 \cdot 10^{-4}$	(15) $2.853 \cdot 10^{-4}$
	(16) $3.060 \cdot 10^{-4}$	(17) $2.952 \cdot 10^{-4}$	(18) $3.100 \cdot 10^{-4}$	(19) $2.876 \cdot 10^{-4}$	(20) $2.881 \cdot 10^{-4}$
	(21) $2.888 \cdot 10^{-4}$	(22) $3.084 \cdot 10^{-4}$	(23) $2.902 \cdot 10^{-4}$	(24) $2.902 \cdot 10^{-4}$	(25) $2.866 \cdot 10^{-4}$
	(26) $3.024 \cdot 10^{-4}$	(27) $2.866 \cdot 10^{-4}$	(28) $3.019 \cdot 10^{-4}$	(29) $2.921 \cdot 10^{-4}$	(30) $2.817 \cdot 10^{-4}$
	(31) $2.862 \cdot 10^{-4}$	(32) $2.828 \cdot 10^{-4}$	(33) $2.891 \cdot 10^{-4}$	(34) $2.842 \cdot 10^{-4}$	(35) $3.034 \cdot 10^{-4}$
	(36) $2.991 \cdot 10^{-4}$	(37) $2.870 \cdot 10^{-4}$	(38) $2.848 \cdot 10^{-4}$	(39) $2.837 \cdot 10^{-4}$	(40) $2.974 \cdot 10^{-4}$
	(41) $2.864 \cdot 10^{-4}$	(42) $2.724 \cdot 10^{-4}$	(43) $2.921 \cdot 10^{-4}$	(44) $2.991 \cdot 10^{-4}$	(45) $2.861 \cdot 10^{-4}$
	(46) $2.857 \cdot 10^{-4}$	(47) $2.887 \cdot 10^{-4}$	(48) $2.958 \cdot 10^{-4}$	(49) $2.985 \cdot 10^{-4}$	(50) $2.858 \cdot 10^{-4}$
	(51) $2.923 \cdot 10^{-4}$	(52) $2.698 \cdot 10^{-4}$	(53) $2.881 \cdot 10^{-4}$	(54) $3.008 \cdot 10^{-4}$	(55) $3.043 \cdot 10^{-4}$
	(56) $2.842 \cdot 10^{-4}$	(57) $2.781 \cdot 10^{-4}$	(58) $2.746 \cdot 10^{-4}$	(59) $2.819 \cdot 10^{-4}$	(60) $2.848 \cdot 10^{-4}$
	(61) $2.753 \cdot 10^{-4}$	(62) $3.010 \cdot 10^{-4}$	(63) $3.004 \cdot 10^{-4}$	(64) $2.805 \cdot 10^{-4}$	(65) $2.921 \cdot 10^{-4}$
	(66) $2.919 \cdot 10^{-4}$	(67) $2.947 \cdot 10^{-4}$	(68) $2.944 \cdot 10^{-4}$	(69) $2.960 \cdot 10^{-4}$	(70) $2.964 \cdot 10^{-4}$
	(71) $2.808 \cdot 10^{-4}$	(72) $2.940 \cdot 10^{-4}$	(73) $2.874 \cdot 10^{-4}$	(74) $2.851 \cdot 10^{-4}$	(75) $2.796 \cdot 10^{-4}$
	(76) $3.049 \cdot 10^{-4}$	(77) $2.885 \cdot 10^{-4}$	(78) $2.849 \cdot 10^{-4}$	(79) $2.711 \cdot 10^{-4}$	(80) $3.004 \cdot 10^{-4}$
	(81) $2.872 \cdot 10^{-4}$	(82) $2.908 \cdot 10^{-4}$	(83) $2.835 \cdot 10^{-4}$	(84) $2.779 \cdot 10^{-4}$	(85) $2.812 \cdot 10^{-4}$
	(86) $3.044 \cdot 10^{-4}$	(87) $2.736 \cdot 10^{-4}$	(88) $2.848 \cdot 10^{-4}$	(89) $2.815 \cdot 10^{-4}$	(90) $2.931 \cdot 10^{-4}$
	(91) $2.824 \cdot 10^{-4}$	(92) $2.923 \cdot 10^{-4}$	(93) $2.897 \cdot 10^{-4}$	(94) $2.872 \cdot 10^{-4}$	(95) $3.016 \cdot 10^{-4}$
	(96) $2.714 \cdot 10^{-4}$	(97) $2.807 \cdot 10^{-4}$	(98) $2.887 \cdot 10^{-4}$	(99) $2.838 \cdot 10^{-4}$	(100) $2.903 \cdot 10^{-4}$
$c = c_{\max}$	(1) $2.040 \cdot 10^{-4}$	(2) $2.029 \cdot 10^{-4}$	(3) $2.038 \cdot 10^{-4}$	(4) $2.021 \cdot 10^{-4}$	(5) $2.012 \cdot 10^{-4}$
	(6) $2.110 \cdot 10^{-4}$	(7) $2.030 \cdot 10^{-4}$	(8) $2.063 \cdot 10^{-4}$	(9) $2.064 \cdot 10^{-4}$	(10) $1.967 \cdot 10^{-4}$
	(11) $2.159 \cdot 10^{-4}$	(12) $2.019 \cdot 10^{-4}$	(13) $2.146 \cdot 10^{-4}$	(14) $2.027 \cdot 10^{-4}$	(15) $2.007 \cdot 10^{-4}$
	(16) $2.088 \cdot 10^{-4}$	(17) $1.979 \cdot 10^{-4}$	(18) $1.950 \cdot 10^{-4}$	(19) $2.023 \cdot 10^{-4}$	(20) $2.055 \cdot 10^{-4}$
	(21) $1.983 \cdot 10^{-4}$	(22) $2.081 \cdot 10^{-4}$	(23) $1.954 \cdot 10^{-4}$	(24) $2.213 \cdot 10^{-4}$	(25) $2.053 \cdot 10^{-4}$
	(26) $1.971 \cdot 10^{-4}$	(27) $2.031 \cdot 10^{-4}$	(28) $2.037 \cdot 10^{-4}$	(29) $1.976 \cdot 10^{-4}$	(30) $2.057 \cdot 10^{-4}$
	(31) $2.140 \cdot 10^{-4}$	(32) $2.043 \cdot 10^{-4}$	(33) $2.086 \cdot 10^{-4}$	(34) $2.087 \cdot 10^{-4}$	(35) $2.006 \cdot 10^{-4}$
	(36) $2.044 \cdot 10^{-4}$	(37) $1.967 \cdot 10^{-4}$	(38) $2.063 \cdot 10^{-4}$	(39) $1.953 \cdot 10^{-4}$	(40) $2.143 \cdot 10^{-4}$
	(41) $2.108 \cdot 10^{-4}$	(42) $2.105 \cdot 10^{-4}$	(43) $2.010 \cdot 10^{-4}$	(44) $1.970 \cdot 10^{-4}$	(45) $2.009 \cdot 10^{-4}$
	(46) $2.050 \cdot 10^{-4}$	(47) $1.948 \cdot 10^{-4}$	(48) $1.946 \cdot 10^{-4}$	(49) $2.093 \cdot 10^{-4}$	(50) $2.043 \cdot 10^{-4}$
	(51) $2.093 \cdot 10^{-4}$	(52) $2.036 \cdot 10^{-4}$	(53) $2.183 \cdot 10^{-4}$	(54) $2.022 \cdot 10^{-4}$	(55) $2.127 \cdot 10^{-4}$
	(56) $2.028 \cdot 10^{-4}$	(57) $2.020 \cdot 10^{-4}$	(58) $2.015 \cdot 10^{-4}$	(59) $2.028 \cdot 10^{-4}$	(60) $1.989 \cdot 10^{-4}$
	(61) $2.079 \cdot 10^{-4}$	(62) $2.199 \cdot 10^{-4}$	(63) $2.053 \cdot 10^{-4}$	(64) $2.127 \cdot 10^{-4}$	(65) $1.990 \cdot 10^{-4}$
	(66) $2.061 \cdot 10^{-4}$	(67) $1.983 \cdot 10^{-4}$	(68) $2.156 \cdot 10^{-4}$	(69) $2.073 \cdot 10^{-4}$	(70) $2.074 \cdot 10^{-4}$
	(71) $2.100 \cdot 10^{-4}$	(72) $2.024 \cdot 10^{-4}$	(73) $2.021 \cdot 10^{-4}$	(74) $1.989 \cdot 10^{-4}$	(75) $1.912 \cdot 10^{-4}$
	(76) $2.109 \cdot 10^{-4}$	(77) $2.043 \cdot 10^{-4}$	(78) $2.112 \cdot 10^{-4}$	(79) $2.015 \cdot 10^{-4}$	(80) $2.096 \cdot 10^{-4}$
	(81) $1.924 \cdot 10^{-4}$	(82) $2.071 \cdot 10^{-4}$	(83) $2.197 \cdot 10^{-4}$	(84) $2.173 \cdot 10^{-4}$	(85) $1.996 \cdot 10^{-4}$
	(86) $2.125 \cdot 10^{-4}$	(87) $1.978 \cdot 10^{-4}$	(88) $2.088 \cdot 10^{-4}$	(89) $2.011 \cdot 10^{-4}$	(90) $1.946 \cdot 10^{-4}$
	(91) $2.006 \cdot 10^{-4}$	(92) $2.156 \cdot 10^{-4}$	(93) $2.069 \cdot 10^{-4}$	(94) $2.018 \cdot 10^{-4}$	(95) $2.015 \cdot 10^{-4}$
	(96) $1.904 \cdot 10^{-4}$	(97) $1.983 \cdot 10^{-4}$	(98) $2.132 \cdot 10^{-4}$	(99) $1.933 \cdot 10^{-4}$	(100) $2.050 \cdot 10^{-4}$

Table 3 For $\alpha = 1$: values of the approximated performance index from Eq. (30) for the 100 repetitions of the simulation procedure

(Repetition number) Approximated performance index from Eq. (30)					
$c = c_{\min}$	(1) $4.251 \cdot 10^{-4}$	(2) $4.155 \cdot 10^{-4}$	(3) $4.141 \cdot 10^{-4}$	(4) $4.162 \cdot 10^{-4}$	(5) $4.243 \cdot 10^{-4}$
	(6) $4.104 \cdot 10^{-4}$	(7) $4.018 \cdot 10^{-4}$	(8) $4.273 \cdot 10^{-4}$	(9) $4.224 \cdot 10^{-4}$	(10) $3.956 \cdot 10^{-4}$
	(11) $3.973 \cdot 10^{-4}$	(12) $4.068 \cdot 10^{-4}$	(13) $4.238 \cdot 10^{-4}$	(14) $4.102 \cdot 10^{-4}$	(15) $4.283 \cdot 10^{-4}$
	(16) $4.567 \cdot 10^{-4}$	(17) $4.224 \cdot 10^{-4}$	(18) $4.123 \cdot 10^{-4}$	(19) $4.362 \cdot 10^{-4}$	(20) $3.970 \cdot 10^{-4}$
	(21) $4.310 \cdot 10^{-4}$	(22) $4.298 \cdot 10^{-4}$	(23) $4.240 \cdot 10^{-4}$	(24) $4.399 \cdot 10^{-4}$	(25) $3.957 \cdot 10^{-4}$
	(26) $4.226 \cdot 10^{-4}$	(27) $4.144 \cdot 10^{-4}$	(28) $4.060 \cdot 10^{-4}$	(29) $4.025 \cdot 10^{-4}$	(30) $4.106 \cdot 10^{-4}$
	(31) $4.057 \cdot 10^{-4}$	(32) $4.060 \cdot 10^{-4}$	(33) $4.056 \cdot 10^{-4}$	(34) $4.183 \cdot 10^{-4}$	(35) $4.200 \cdot 10^{-4}$
	(36) $4.170 \cdot 10^{-4}$	(37) $3.990 \cdot 10^{-4}$	(38) $3.959 \cdot 10^{-4}$	(39) $4.103 \cdot 10^{-4}$	(40) $3.995 \cdot 10^{-4}$
	(41) $3.829 \cdot 10^{-4}$	(42) $4.041 \cdot 10^{-4}$	(43) $4.009 \cdot 10^{-4}$	(44) $3.815 \cdot 10^{-4}$	(45) $4.128 \cdot 10^{-4}$
	(46) $3.976 \cdot 10^{-4}$	(47) $4.249 \cdot 10^{-4}$	(48) $4.076 \cdot 10^{-4}$	(49) $4.253 \cdot 10^{-4}$	(50) $4.222 \cdot 10^{-4}$
	(51) $4.130 \cdot 10^{-4}$	(52) $4.011 \cdot 10^{-4}$	(53) $3.998 \cdot 10^{-4}$	(54) $4.047 \cdot 10^{-4}$	(55) $3.960 \cdot 10^{-4}$
	(56) $4.235 \cdot 10^{-4}$	(57) $4.157 \cdot 10^{-4}$	(58) $3.909 \cdot 10^{-4}$	(59) $4.221 \cdot 10^{-4}$	(60) $4.455 \cdot 10^{-4}$
	(61) $4.051 \cdot 10^{-4}$	(62) $4.077 \cdot 10^{-4}$	(63) $4.405 \cdot 10^{-4}$	(64) $4.106 \cdot 10^{-4}$	(65) $4.192 \cdot 10^{-4}$
	(66) $4.111 \cdot 10^{-4}$	(67) $4.183 \cdot 10^{-4}$	(68) $4.279 \cdot 10^{-4}$	(69) $4.099 \cdot 10^{-4}$	(70) $4.367 \cdot 10^{-4}$
	(71) $4.060 \cdot 10^{-4}$	(72) $4.016 \cdot 10^{-4}$	(73) $4.279 \cdot 10^{-4}$	(74) $4.080 \cdot 10^{-4}$	(75) $4.153 \cdot 10^{-4}$
	(76) $4.172 \cdot 10^{-4}$	(77) $4.084 \cdot 10^{-4}$	(78) $4.060 \cdot 10^{-4}$	(79) $4.187 \cdot 10^{-4}$	(80) $3.963 \cdot 10^{-4}$
	(81) $4.148 \cdot 10^{-4}$	(82) $4.097 \cdot 10^{-4}$	(83) $4.233 \cdot 10^{-4}$	(84) $3.991 \cdot 10^{-4}$	(85) $4.167 \cdot 10^{-4}$
	(86) $4.090 \cdot 10^{-4}$	(87) $4.176 \cdot 10^{-4}$	(88) $3.991 \cdot 10^{-4}$	(89) $4.027 \cdot 10^{-4}$	(90) $3.870 \cdot 10^{-4}$
	(91) $4.060 \cdot 10^{-4}$	(92) $4.177 \cdot 10^{-4}$	(93) $4.061 \cdot 10^{-4}$	(94) $4.133 \cdot 10^{-4}$	(95) $4.022 \cdot 10^{-4}$
	(96) $4.105 \cdot 10^{-4}$	(97) $3.803 \cdot 10^{-4}$	(98) $4.141 \cdot 10^{-4}$	(99) $4.171 \cdot 10^{-4}$	(100) $4.176 \cdot 10^{-4}$
$c = c_{\max}$	(1) $3.911 \cdot 10^{-4}$	(2) $4.069 \cdot 10^{-4}$	(3) $4.036 \cdot 10^{-4}$	(4) $4.297 \cdot 10^{-4}$	(5) $4.113 \cdot 10^{-4}$
	(6) $4.192 \cdot 10^{-4}$	(7) $4.082 \cdot 10^{-4}$	(8) $3.914 \cdot 10^{-4}$	(9) $4.029 \cdot 10^{-4}$	(10) $4.308 \cdot 10^{-4}$
	(11) $3.915 \cdot 10^{-4}$	(12) $3.739 \cdot 10^{-4}$	(13) $4.075 \cdot 10^{-4}$	(14) $4.111 \cdot 10^{-4}$	(15) $4.265 \cdot 10^{-4}$
	(16) $4.352 \cdot 10^{-4}$	(17) $3.934 \cdot 10^{-4}$	(18) $4.044 \cdot 10^{-4}$	(19) $4.112 \cdot 10^{-4}$	(20) $4.258 \cdot 10^{-4}$
	(21) $4.306 \cdot 10^{-4}$	(22) $4.179 \cdot 10^{-4}$	(23) $4.095 \cdot 10^{-4}$	(24) $4.189 \cdot 10^{-4}$	(25) $4.228 \cdot 10^{-4}$
	(26) $4.413 \cdot 10^{-4}$	(27) $3.976 \cdot 10^{-4}$	(28) $4.134 \cdot 10^{-4}$	(29) $4.166 \cdot 10^{-4}$	(30) $4.121 \cdot 10^{-4}$
	(31) $3.866 \cdot 10^{-4}$	(32) $4.440 \cdot 10^{-4}$	(33) $4.050 \cdot 10^{-4}$	(34) $4.129 \cdot 10^{-4}$	(35) $3.934 \cdot 10^{-4}$
	(36) $3.944 \cdot 10^{-4}$	(37) $4.066 \cdot 10^{-4}$	(38) $4.045 \cdot 10^{-4}$	(39) $4.115 \cdot 10^{-4}$	(40) $3.973 \cdot 10^{-4}$
	(41) $4.002 \cdot 10^{-4}$	(42) $4.248 \cdot 10^{-4}$	(43) $4.134 \cdot 10^{-4}$	(44) $4.302 \cdot 10^{-4}$	(45) $4.222 \cdot 10^{-4}$
	(46) $4.121 \cdot 10^{-4}$	(47) $3.946 \cdot 10^{-4}$	(48) $4.139 \cdot 10^{-4}$	(49) $4.183 \cdot 10^{-4}$	(50) $4.245 \cdot 10^{-4}$
	(51) $3.962 \cdot 10^{-4}$	(52) $4.204 \cdot 10^{-4}$	(53) $4.183 \cdot 10^{-4}$	(54) $3.930 \cdot 10^{-4}$	(55) $4.206 \cdot 10^{-4}$
	(56) $4.044 \cdot 10^{-4}$	(57) $3.754 \cdot 10^{-4}$	(58) $4.247 \cdot 10^{-4}$	(59) $4.185 \cdot 10^{-4}$	(60) $4.007 \cdot 10^{-4}$
	(61) $4.564 \cdot 10^{-4}$	(62) $4.174 \cdot 10^{-4}$	(63) $4.094 \cdot 10^{-4}$	(64) $3.944 \cdot 10^{-4}$	(65) $4.266 \cdot 10^{-4}$
	(66) $4.352 \cdot 10^{-4}$	(67) $4.042 \cdot 10^{-4}$	(68) $4.281 \cdot 10^{-4}$	(69) $4.168 \cdot 10^{-4}$	(70) $3.093 \cdot 10^{-4}$
	(71) $4.074 \cdot 10^{-4}$	(72) $4.007 \cdot 10^{-4}$	(73) $4.096 \cdot 10^{-4}$	(74) $3.968 \cdot 10^{-4}$	(75) $3.932 \cdot 10^{-4}$
	(76) $4.066 \cdot 10^{-4}$	(77) $4.213 \cdot 10^{-4}$	(78) $4.040 \cdot 10^{-4}$	(79) $4.300 \cdot 10^{-4}$	(80) $4.091 \cdot 10^{-4}$
	(81) $3.901 \cdot 10^{-4}$	(82) $4.161 \cdot 10^{-4}$	(83) $4.812 \cdot 10^{-4}$	(84) $4.039 \cdot 10^{-4}$	(85) $3.857 \cdot 10^{-4}$
	(86) $4.078 \cdot 10^{-4}$	(87) $4.267 \cdot 10^{-4}$	(88) $4.233 \cdot 10^{-4}$	(89) $3.985 \cdot 10^{-4}$	(90) $3.902 \cdot 10^{-4}$
	(91) $4.110 \cdot 10^{-4}$	(92) $4.045 \cdot 10^{-4}$	(93) $3.997 \cdot 10^{-4}$	(94) $4.170 \cdot 10^{-4}$	(95) $4.249 \cdot 10^{-4}$
	(96) $4.005 \cdot 10^{-4}$	(97) $3.942 \cdot 10^{-4}$	(98) $4.254 \cdot 10^{-4}$	(99) $4.215 \cdot 10^{-4}$	(100) $4.193 \cdot 10^{-4}$

with variance expressed as in Eq. (27), possibly with the inclusion of an additional constant term in that variance, to model the case of the original dataset before the insertion of the artificial noise. As a possible extension, one could investigate and optimize the trade-off between training set size and precision of supervision for the unbalanced FEGLS case (in which different units are associated with possibly different numbers of observations)[5], for the situation in which some parameters of the noise model have to be estimated either from the data or from a subset of the data, and for the case of a non-zero correlation of measurement errors for the observations associated with different units. Such developments could open the way to the application of the proposed framework to real-world problems, e.g., in econometrics. Another possible extension concerns the replacement in the investigation of the fixed effects panel data model with the random effects one (Greene 2003, Chapter 13), which is also commonly applied to deal with the analysis of economic data, and differs from the fixed effects panel data model in that its parameters are random variables[6]. In the present analysis, however, a possible advantage of the fixed effects panel data model is that it also makes it possible to get estimates of the individual constants η_n (see Eq. (12)), which appear in the expression (16) of the conditional generalization error. Finally, another possible extension involves the case of dynamic panel data models (Cameron and Trivedi 2005, Chapter 21).

Appendix 1: proofs of Eqs. (17), (18), and (19)

To simplify the notation, here and in the next appendices, Q_T will be often replaced by the shorthand Q. Also the dependence of Ψ and Φ and other matrices/vectors on T will be typically omitted in the notation, apart from a few cases (e.g., in part of Appendix 2).

Proof of Eq. (17) For $n = 1, \dots, N$, let $\underline{\eta}_n \in \mathbb{R}^T$ be the column vector whose elements are all equal to η_n. Using the expressions (1), (2), (6), and (9) respectively of $y_{n,t}$, $z_{n,t}$, $\underline{\ddot{z}}_n$, and $\hat{\underline{\beta}}_{FEGLS}$, and

$$Q\underline{\eta}_n = \underline{0}_T\,, \tag{A.1}$$

one can write the term $\hat{\underline{\beta}}_{FEGLS} - \underline{\beta}$ as follows:

[5] The unbalanced case in which all the measurement errors are uncorrelated - therefore, the FEGLS model is replaced in the analysis by the simpler Fixed Effects (FE) model - is the subject of our recent work Gnecco et al. (2020).

[6] If the additional assumptions of the random effects model hold, then both the fixed and the random effects estimates are consistent, but the latter is more efficient than the former. However, if they do not hold, then the random effects model provides inconsistent estimates (Greene 2003, Chapter 13).

$$\begin{aligned}\underline{\hat{\beta}}_{FEGLS}-\underline{\beta}=&\left(\sum_{n=1}^{N}\ddot{X}_n'\Omega^{+}\ddot{X}_n\right)^{-1}\left(\sum_{n=1}^{N}\ddot{X}_n'\Omega^{+}\underline{\ddot{z}}_n\right)-\underline{\beta}\\ =&\left(\sum_{n=1}^{N}\ddot{X}_n'\Omega^{+}\ddot{X}_n\right)^{-1}\left(\sum_{n=1}^{N}\ddot{X}_n'\Omega^{+}Q\left(\underline{\eta}_n+X_n\underline{\beta}+\underline{\varepsilon}_n\right)\right)-\underline{\beta}\\ =&\left(\sum_{n=1}^{N}\ddot{X}_n'\Omega^{+}\ddot{X}_n\right)^{-1}\sum_{n=1}^{N}\ddot{X}_n'\Omega^{+}\underline{\ddot{\varepsilon}}_n\,.\end{aligned}\tag{A.2}$$

In the following, to simplify the notation, we set $B:=\left(\sum_{n=1}^{N}\ddot{X}_n'\Omega^{+}\ddot{X}_n\right)^{-1}$ and $\underline{b}:=\sum_{n=1}^{N}\ddot{X}_n'\Omega^{+}\underline{\ddot{\varepsilon}}_n$.

Now, we expand the conditional generalization error (16) as follows:

$$\begin{aligned}R_i\left(\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right)=&\,\mathbb{E}\left\{\left(\hat{\eta}_{i,FEGLS}+\underline{\hat{\beta}}_{FEGLS}'\underline{x}_i^{test}-\eta_i-\underline{\beta}'\underline{x}_i^{test}\right)^2\Big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}\\ =&\,\mathbb{E}\left\{(\hat{\eta}_{i,FEGLS}-\eta_i)^2\big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}+\mathbb{E}\left\{\left(\left(\underline{\hat{\beta}}_{FEGLS}-\underline{\beta}\right)'\underline{x}_i^{test}\right)^2\Big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}\\ &+2\mathbb{E}\left\{\left(\hat{\eta}_{i,FEGLS}-\eta_i\right)\left(\underline{\hat{\beta}}_{FEGLS}-\underline{\beta}\right)'\underline{x}_i^{test}\big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}.\end{aligned}\tag{A.3}$$

Exploiting the conditional unbiasedness of $\hat{\eta}_{i,FEGLS}$, and the expressions (1) of $y_{n,t}$, (2) of $z_{n,t}$, and (12) of $\hat{\eta}_{i,FEGLS}$ (with the index n replaced by the index i), one gets

$$\begin{aligned}&\mathbb{E}\left\{(\hat{\eta}_{i,FEGLS}-\eta_i)^2\big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}\\ &=\mathbb{E}\left\{\left(\frac{1}{T}\sum_{t=1}^{T}\left(\left(\underline{\beta}-\underline{\hat{\beta}}_{FEGLS}\right)'\underline{x}_{i,t}+\varepsilon_{i,t}\right)\right)^2\Big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}.\end{aligned}\tag{A.4}$$

Then, taking into account (A.2) and (A.4), one gets

$$\begin{aligned}(A.3)=&\,\mathbb{E}\left\{\left(\frac{1}{T}\underline{1}_T'\left(-X_iB\underline{b}+\underline{\varepsilon}_i\right)\right)^2\Big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}+\mathbb{E}\left\{\left(\underline{x}_i^{test}\right)'B\underline{b}\left(B\underline{b}\right)'\underline{x}_i^{test}\big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}\\ &+2\mathbb{E}\left\{\left(\frac{1}{T}\underline{1}_T'\left(-X_iB\underline{b}+\underline{\varepsilon}_i\right)\right)\left(B\underline{b}\right)'\underline{x}_i^{test}\big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}.\end{aligned}\tag{A.5}$$

Expanding the square in the first term in the expression above, and splitting its last term in two parts, one obtains

$$\begin{aligned}(A.5)=&\,\mathbb{E}\left\{\frac{1}{T^2}\underline{1}_T'X_iB\underline{b}\underline{b}'B'X_i'\underline{1}_T\big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}+\mathbb{E}\left\{\frac{1}{T^2}\underline{1}_T'\underline{\varepsilon}_i\underline{\varepsilon}_i'\underline{1}_T\big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}\\ &-2\mathbb{E}\left\{\frac{1}{T^2}\underline{1}_T'X_iB\underline{b}\underline{\varepsilon}_i'\underline{1}_T\big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}+\mathbb{E}\left\{\left(\underline{x}_i^{test}\right)'B\underline{b}\underline{b}'B'\underline{x}_i^{test}\big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}\\ &-2\mathbb{E}\left\{\frac{1}{T}\underline{1}_T'X_iB\underline{b}\underline{b}'B'\underline{x}_i^{test}\big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}+2\mathbb{E}\left\{\frac{1}{T}\underline{1}_T'\underline{\varepsilon}_i\underline{b}'B'\underline{x}_i^{test}\big|\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\}.\end{aligned}\tag{A.6}$$

In order to simplify the expressions above, one exploits the following properties:

$$\mathbb{E}\left\{\underline{\ddot{\varepsilon}}_n \underline{\ddot{\varepsilon}}'_m\right\} = \underline{0}_{T\times T} \tag{A.7}$$

for $n \neq m$ (being $\underline{0}_{T\times T} \in \mathbb{R}^{T\times T}$ the matrix whose elements are all equal to 0), and

$$\mathbb{E}\left\{\underline{\ddot{\varepsilon}}_n \underline{\ddot{\varepsilon}}'_n\right\} = \Omega = \sigma^2 \Phi\,. \tag{A.8}$$

Then, expanding $\underline{b}$ and exploiting also the facts that $\Omega^+\Omega\Omega^+ = \Omega$, $B = B'$, the matrix Ω^+ is symmetric and deterministic, and all the $\ddot{X}_n$ are known once all the $\underline{x}_{n,t}$ are given, one gets

$$\begin{aligned}
\mathbb{E}\left\{B\underline{b}\underline{b}'B' \,\middle|\, \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} =& \mathbb{E}\left\{B\left(\sum_{n=1}^{N} \ddot{X}'_n \Omega^+ \underline{\ddot{\varepsilon}}_n \sum_{m=1}^{N} \underline{\ddot{\varepsilon}}'_m \Omega^+ \ddot{X}_m\right)B' \,\middle|\, \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right\} \\
=& B\left(\sum_{n=1}^{N} \ddot{X}'_n \Omega^+ \Omega \Omega^+ \ddot{X}_n\right)B' \\
=& BB^{-1}B' \\
=& \sigma^2\left(\sum_{n=1}^{N} \ddot{X}'_n \Phi^+ \ddot{X}_n\right)^{-1}\,.
\end{aligned} \tag{A.9}$$

Finally, inserting (A.9) in (A.6), expanding the expressions of B and $\underline{b}$, and recalling that $\mathbb{E}\left\{\underline{\varepsilon}_i \underline{\varepsilon}'_i\right\} = \sigma^2 \Psi$, one obtains Eq. (17). □

Proof of Eq. (18) The expression $\underline{1}'_T \Psi \underline{1}_T$ in Eq. (18) is the summation of all the elements of the matrix Ψ. Now, the element $\rho^0 = 1$ appears in that summation T times, whereas the generic element ρ^t (for $t = 1, \ldots, T-1$) appears $2(T-t)$ times. Hence,

$$\underline{1}'_T \Psi \underline{1}_T = \left(T + \sum_{t=1}^{T-1} 2(T-t)\rho^t\right) = \left(T + 2T\sum_{t=1}^{T-1}\rho^t - 2\sum_{t=1}^{T-1} t\rho^t\right)\,. \tag{A.10}$$

Then, Eq. (18) is obtained from (A.10) by exploiting the following well-known expressions for the partial sums of the geometric series, and of its derivative:

$$\sum_{t=1}^{T-1}\rho^t = \frac{1-\rho^T}{1-\rho} - 1\,, \tag{A.11}$$

and

$$\sum_{t=1}^{T-1} t\rho^t = \rho\frac{d\left(\sum_{t=1}^{T-1}\rho^t\right)}{d\rho} = \frac{\rho}{1-\rho}(-(T-1)\rho^{T-1} + \rho^{T-2} + \rho^{T-3} + \ldots + 1)\,, \tag{A.12}$$

where the right-hand side in Eq. (A.12) has been obtained by simplifying common factors in the numerator and the denominator. □

Proof of Eq. (19) We compute $Q\Psi\underline{1}_T$, as follows:

$$\begin{aligned}Q\Psi\underline{1}_T =&\Big(I_T - \frac{1}{T}\underline{1}_T\underline{1}_T{}'\Big)\Psi\underline{1}_T\\ =&\Psi\underline{1}_T - \frac{1}{T}(\underline{1}_T'\Psi\underline{1}_T)\underline{1}_T\\ =&\Psi\underline{1}_T - \left[1+2\left(\frac{1-\rho^T}{1-\rho}-1\right)-\frac{2\rho}{T(1-\rho)}\big(-(T-1)\rho^{T-1}+\rho^{T-2}+\rho^{T-3}+\cdots+1\big)\right]\underline{1}_T\,,\end{aligned} \tag{A.13}$$

where the expression above of $(\underline{1}_T'\Psi\underline{1}_T)\underline{1}_T$ comes from Eq. (18). Finally, Eq. (19) is obtained from Eq. (A.13) by expanding the elements of $\Psi\underline{1}_T$, then simplifying the resulting expressions. □

Appendix 2: proofs of the probability limits in Section 3

In the following, Eqs. (20) and (21) are derived under the common assumption that, for the unit i, the $\underline{x}_{i,t}$ are mutually independent, identically distributed, and have finite moments up to the order 4. To derive Eq. (23), one makes the similar assumption that, for each fixed unit n, the $\underline{x}_{n,t}$ are mutually independent, identically distributed, and have finite moments up to the order 4, together with the additional assumption $\lim\limits_{T\to\infty}\|\Phi^+ - Q\Psi^{-1}Q'\|_2 = 0$ reported in Eq. (22). The validity of this last assumption is discussed extensively at the end of this appendix. Eqs. (20), (21), and (23) could be derived under more general conditions, but such possible extension is out of the scope of the paper.

Proof of Eq. (20) Eq. (20) simply replaces the empirical average of the transposes of the $\underline{x}_{n,t}$ (which is $\frac{1}{T}\underline{1}_T'X_i$) with their common expected value $\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)'$, and follows from Chebyschev's weak law of large numbers (Ruud 2000, Sect. 13.4.2). □

Proof of Eq. (21) In order to prove Eq. (21), it is convenient to introduce (recalling the definition of $\underline{u}_T$ provided in Eq. (19)) the vector

$$\underline{v}_T := Q'\Phi^+Q\Psi\underline{1}_T = Q'\Phi^+\underline{u}_T\,, \tag{A.14}$$

since the argument of the probability limit in Eq. (21) can be written as follows:

$$\frac{1}{T}\ddot{X}_i'\Phi^+Q\Psi\underline{1}_T = \frac{1}{T}X_i'Q'\Phi^+Q\Psi\underline{1}_T = \frac{1}{T}X_i'\underline{v}_T\,. \tag{A.15}$$

In other words, the T elements of each row of X_i' are summed with (different and deterministic) weights $v_{T,t}$ (the components of $\underline{v}_T$), for $t = 1,\ldots,T$, then their weighted sum is divided by T. This suggests the application of a suitable form of the law of large numbers, which holds in this case: specifically, the one provided in (Bai et al. 1997, Theorem 2.1)). In view of the next application of that theorem, first we investigate the following properties of the various terms involved in Eqs. (A.14) and (A.15).

(i) *The Euclidean norm of the vector $\underline{u}_T$ is bounded from above as follows, for $K_u > 0$ independent from T:*

$$\|\underline{u}_T\|_2 \leq K_u \sqrt{T}\,. \tag{A.16}$$

This follows from the fact that the absolute values of all the components of the vector $\underline{u}_T$, whose expression is reported in Eq. (19), are bounded from above by a sufficiently large $K_u > 0$, which is independent from T.

(ii) *All the eigenvalues of the matrix* $\boldsymbol{\Psi}$ *belong to the interval* $\left[\frac{1-\rho^2}{1+\rho^2+2\rho}, \frac{1-\rho^2}{1+\rho^2-2\rho}\right] \subset (0, +\infty)$. This result follows by observing that $\boldsymbol{\Psi}$ is a symmetric Toeplitz matrix[7] (Gray 2006). Then, by (Gray 2006, Lemma 4.1), all the eigenvalues of $\boldsymbol{\Psi}$ belong to the interval $[m_f, M_f]$, where m_f and M_f are respectively the minimum and the maximum of the function

$$f(\lambda) := \sum_{k=-\infty}^{+\infty} \rho^{|k|} e^{\imath k \lambda} \tag{A.19}$$

on the interval $[0, 2\pi]$, and $\imath$ is the imaginary unit. By inverting the Fourier series above as in (Gilgen 2006, Eqs. (7.77–7.79)), one gets

$$f(\lambda) = \frac{1-\rho^2}{1+\rho^2-2\rho\cos(\lambda)}\,, \tag{A.20}$$

from which one gets $m_f = \frac{1-\rho^2}{1+\rho^2+2\rho} > 0$ and $M_f = \frac{1-\rho^2}{1+\rho^2-2\rho} < +\infty$ since $\rho \in (-1, 1)$, which concludes the proof of item ii).

(iii) *The matrix* $\boldsymbol{\Phi}$ *has* 0 *as eigenvalue with multiplicity* 1, *and an associated eigenvector is* $\underline{1}_T$.

This result follows from the characterization of the eigenvalues of a symmetric matrix $M \in \mathbb{R}^{T \times T}$ as the stationary values of its Rayleigh quotient $\frac{\underline{x}' M \underline{x}}{\underline{x}' \underline{x}}$ (with $\underline{x} \in \mathbb{R}^T$ and $\underline{x} \neq \underline{0}_T$) (Parlett 1998, Chapter 1), the invertibility of the matrix $\boldsymbol{\Psi}$, and the fact that Q has eigenvalue 0 with multiplicity 1, and associated eigenvector $\underline{1}_T$. Hence, for $\underline{x} \neq \underline{0}_T$, $\frac{\underline{x}' M \underline{x}}{\underline{x}' \underline{x}} = 0$ if and only $\underline{x}$ is proportional to $\underline{1}_T$.

(iv) *All the other eigenvalues of* $\boldsymbol{\Phi}$ *belong to the interval* $\left[\frac{1-\rho^2}{1+\rho^2+2\rho}, \frac{1-\rho^2}{1+\rho^2-2\rho}\right] \subset (0, +\infty)$.

This follows again from the characterization of the eigenvalues of a symmetric matrix as the stationary values of its Rayleigh quotient, and also from Courant-Fisher's maxmin theorem (Parlett 1998, Theorem 10.2.1) and from item ii). Indeed, by ordering the (real) eigenvalues of $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ respectively as $\lambda_1(\boldsymbol{\Psi}) \leq \lambda_2(\boldsymbol{\Psi}) \leq \ldots \lambda_T(\boldsymbol{\Psi})$ and $\lambda_1(\boldsymbol{\Phi}) \leq \lambda_2(\boldsymbol{\Phi}) \leq \ldots \lambda_T(\boldsymbol{\Phi})$, and recalling that $\underline{x}' \boldsymbol{\Phi} \underline{x} = \underline{x}' Q \boldsymbol{\Psi} Q' \underline{x} = \underline{x}' \boldsymbol{\Psi} \underline{x}$ for any $\underline{x} \in \mathbb{R}^T$ orthogonal to $\underline{1}_T$, one gets

[7] We recall that a matrix $M \in \mathbb{R}^{T \times T}$ is a symmetric Toeplitz matrix if it has the form

$$M = \begin{bmatrix} m_0 & m_1 & m_2 & \cdots & m_{T-2} & m_{T-1} \\ m_1 & m_0 & m_1 & m_2 & \cdots & m_{T-2} \\ m_2 & m_1 & m_0 & m_1 & \cdots & m_{T-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ m_{T-1} & m_{T-2} & \cdots & m_2 & m_1 & m_0 \end{bmatrix}, \tag{A.18}$$

where $m_0, m_1, \ldots, m_{T-1} \in \mathbb{R}$.

$$\lambda_1(\varPsi) = \min_{\underline{x}\in\mathbb{R}^T, \underline{x}\neq\underline{0}_T} \frac{\underline{x}'\boldsymbol{\varPsi}\underline{x}}{\underline{x}'\underline{x}} \leq \min_{\underline{x}\in\mathbb{R}^T, \underline{x}\neq\underline{0}_T, \underline{x}\perp\underline{1}_T} \frac{\underline{x}'\boldsymbol{\varPsi}\underline{x}}{\underline{x}'\underline{x}} = \min_{\underline{x}\in\mathbb{R}^T, \underline{x}\neq\underline{0}_T, \underline{x}\perp\underline{1}_T} \frac{\underline{x}'\boldsymbol{\varPhi}\underline{x}}{\underline{x}'\underline{x}} = \lambda_2(\boldsymbol{\varPhi}) \quad \text{(A.17)}$$

from Courant-Fisher's maxmin theorem, whereas

$$\lambda_T(\varPsi) = \max_{\underline{x}\in\mathbb{R}^T, \underline{x}\neq\underline{0}_T} \frac{\underline{x}'\boldsymbol{\varPsi}\underline{x}}{\underline{x}'\underline{x}} \geq \max_{\underline{x}\in\mathbb{R}^T, \underline{x}\neq\underline{0}_T} \frac{\underline{x}'\boldsymbol{\varPhi}\underline{x}}{\underline{x}'\underline{x}} = \lambda_T(\boldsymbol{\varPhi}) \quad \text{(A.21)}$$

is obtained by expressing $Q'\underline{x}$ by using a basis of orthonormal eigenvectors $\underline{e}_t$ of $\varPsi$, associated with the respective eigenvalues $\lambda_t(\varPsi)$, $t = 1, \dots, T$. Indeed, for some coefficients α_t, $t = 1, \dots, T$ (depending on $\underline{x}$), one has

$$Q'\underline{x} = \sum_{t=1}^{T} \alpha_t \underline{e}_t \,, \quad \text{(A.22)}$$

hence

$$\underline{x}'\boldsymbol{\varPhi}\underline{x} = \underline{x}' Q \varPsi Q' \underline{x} = \sum_{t=1}^{T} \lambda_t(\varPsi)\alpha_t^2 \,, \quad \text{(A.23)}$$

whereas

$$\underline{x}'\underline{x} \geq \underline{x}' Q Q' \underline{x} = \sum_{t=1}^{T} \alpha_t^2 \,. \quad \text{(A.24)}$$

Then, one gets

$$\frac{\underline{x}'\boldsymbol{\varPhi}\underline{x}}{\underline{x}'\underline{x}} \leq \lambda_T(\varPsi) \quad \text{(A.25)}$$

for any $\underline{x} \neq \underline{0}_T$.

(v) *All the non-zero eigenvalues of the matrix* $\boldsymbol{\varPhi}^+$ *belong to the interval* $\left[\frac{1+\rho^2-2\rho}{1-\rho^2}, \frac{1+\rho^2+2\rho}{1-\rho^2}\right] \subset (0, +\infty)$.

This follows from items iii) and iv) and the relation between the singular value decomposition of a symmetric positive semi-definite matrix and the singular value decomposition of its Moore-Penrose pseudoinverse, which has been reported in footnote 2.

(vi) *The Euclidean norm of the vector* $\underline{v}_T$ *is bounded from above as follows, for* $K_v > 0$ *independent from* T:

$$\|\underline{v}_T\|_2 \leq K_v \sqrt{T} \,. \quad \text{(A.26)}$$

This is obtained by combining the definition of $\underline{v}_T$ provided in Eq. (A.14) with items i) and v), and the fact that the eigenvalue with maximus modulus of $Q = Q'$ is 1. A possible expression for K_v is $K_v = \frac{1+\rho^2+2\rho}{1-\rho^2} K_u$.

(vii) *The following holds:*

$$\limsup_{T\to\infty} \frac{1}{\sqrt{T}} \|\underline{v}_T\|_2 \leq K_v < +\infty\,. \tag{A.27}$$

This is obtained immediately from item vi).

To conclude the proof of Eq. (21), we first consider the case in which, for the unit i, all the $\underline{x}_{i,t}$ have mean $\underline{0}_p$. Later, this additional assumption is removed.

(viii) *Proof of Eq.* (21) *when all the* $\underline{x}_{i,t}$ *have mean* $\underline{0}_p$.

Item vii) and the fact that all the elements of each row of X_i' have 0 mean, are independent, identically distributed, and their moments up to the order 4 are finite allow one to apply (Bai et al. 1997, Theorem 2.1), getting the following result, where, for $r = 1, \ldots, p$, $(X_i'\underline{v}_T)_r$ denotes the r-th component of $X_i'\underline{v}_T$:

$$\limsup_{T\to\infty} \frac{|(X_i'\underline{v}_T)_r|}{T^{\frac{3}{4}}(\log T)^{\frac{1}{4}}} = 0 \text{ almost surely.}^{8} \tag{A.28}$$

This, combined with the inequalities

$$0 \leq \liminf_{T\to\infty} \frac{|(X_i'\underline{v}_T)_r|}{T} \leq \limsup_{T\to\infty} \frac{|(X_i'\underline{v}_T)_r|}{T} \leq \limsup_{T\to\infty} \frac{|(X_i'\underline{v}_T)_r|}{T^{\frac{3}{4}}(\log T)^{\frac{1}{4}}} \tag{A.29}$$

and the fact that almost sure convergence implies convergence in probability (Rao 1973), shows that, for all $r = 1, \ldots, p$, one has

$$\operatorname*{plim}_{T\to+\infty} \frac{1}{T}(X_i'\underline{v}_T)_r = 0\,. \tag{A.30}$$

To conclude, one gets Eq. (21) from Eqs. (A.15) and (A.30), by exploiting the fact that, for a sequence of random matrices with fixed dimension, element-wise convergence in probability implies convergence in probability of the whole sequence (Lee 2010).

(ix) *Proof of Eq.* (21) *when all the* $\underline{x}_{i,t}$ *have the same mean* $\underline{m} \in \mathbb{R}^p$.

We set $\bar{\underline{x}}_{i,t} := \underline{x}_{i,t} - \underline{m}$, in such a way that the $\bar{\underline{x}}_{i,t}$ have mean $\underline{0}_p$. Similarly, we set $\bar{X}_i = X_i - \underline{1}_T\underline{m}'$. Since $\ddot{X}_i = QX_i = Q(\bar{X}_i + \underline{1}_T\underline{m}') = Q\bar{X}_i = \ddot{\bar{X}}_i$, one reduces the analysis to the one made in item viii).

Remark 6.1 As a variation of item ii), a simpler argument (not based on the theory of Toeplitz matrices) can be used to prove that all the eigenvalues of the matrix Ψ belong to the interval $\left[1 - \frac{2\rho}{1-\rho}, 1 + \frac{2\rho}{1-\rho}\right]$. This result follows by seeing the matrix Ψ as a perturbation of the identity matrix, then applying Gershgorin's circle theorem (Gerschgorin 1931). Indeed, all the eigenvalues of Ψ (which are non-negative since Ψ is symmetric and positive semidefinite) belong to the union of the T Gershgorin's circles $\mathcal{C}_i$ $(i = 1, \ldots, T)$ in the complex plane, which have the same center 1 and respective radii $\sum_{j=1,\ldots,T, j\neq i} |\Psi_{ij}|$. The latter radii can be bounded from above by $\frac{2\rho}{1-\rho}$, which follows from a geometric series argument based on Eq. (A.11). We have preferred to use in the main text the argument based on Toeplitz

[8] We recall that a sequence of random real variables b_T, $T = 1, 2, \ldots$, converges almost surely to $b \in \mathbb{R}$ if $\text{Prob}(\lim_{T\to+\infty} b_T = b) = 1$

matrices, since imposing $\left[1-\frac{2\rho}{1-\rho}, 1+\frac{2\rho}{1-\rho}\right] \subset (0,+\infty)$ requires the additional assumption $\rho < \frac{1}{3}$ (instead, $\left[\frac{1-\rho^2}{1+\rho^2+2\rho}, \frac{1-\rho^2}{1+\rho^2-2\rho}\right] \subset (0,+\infty)$ holds for any $\rho \in (-1,1)$). Moreover, such argument produces an even better estimate (when $\sum_{j=1,\dots,T, j\neq i} |\Psi_{ij}|$ is replaced by its upper bound $\frac{2\rho}{1-\rho}$), since $1-\frac{2\rho}{1-\rho} < \frac{1-\rho^2}{1+\rho^2+2\rho}$ and $1+\frac{2\rho}{1-\rho} = \frac{1-\rho^2}{1+\rho^2-2\rho}$, hence $\left[\frac{1-\rho^2}{1+\rho^2+2\rho}, \frac{1-\rho^2}{1+\rho^2-2\rho}\right] \subset \left[1-\frac{2\rho}{1-\rho}, 1+\frac{2\rho}{1-\rho}\right]$.

Proof of Eq. (23) To make the reading easier, the proof of Eq. (23) is divided into several steps.

(i) *The following holds:*

$$Q'\Phi^+Q = \Phi^+ . \tag{A.31}$$

This is obtained as follows. First, since the matrix $Q = Q'$ is idempotent, one gets

$$Q'\Phi^+Q = \Phi^+Q \tag{A.32}$$

by (Maciejewski and Klein 1985, Appendix). Additionally, since Moore-Penrose pseudoinversion commutes with transposition (Barata and Hussein 2012), we get

$$\left(\Phi^+Q\right)' = Q'\left(\Phi^+\right)' = Q'\left(\Phi'\right)^+ = Q'\Phi^+ = \Phi^+ , \tag{A.33}$$

where the last step follows again by (Maciejewski and Klein 1985, Appendix) and by the symmetry of Φ. Transposing Eq. (A.33) and combining it with the symmetry of Φ^+ and with Eq. (A.32), we get Eq. (A.31).

(ii) *The following decomposition holds*:

$$\ddot{X}_n'\Phi^+\ddot{X}_n = X_n'Q'\Phi^+QX_n = X_n'\Phi^+X_n = X_n'\left[\Phi^+ - Q\Psi^{-1}Q'\right]X_n + X_n'Q\Psi^{-1}Q'X_n . \tag{A.34}$$

This is obtained straightforwardly, by applying item i) to get the second equality.

(iii) *Under the assumption* $\lim_{T\to\infty} \|\Phi^+ - Q\Psi^{-1}Q'\|_2 = 0$ *stated in Eq.* (22), *the following probability limit holds:*

$$\operatorname*{plim}_{T\to+\infty} \frac{1}{T} X_n'\left[\Phi^+ - Q\Psi^{-1}Q'\right]X_n = 0_{p\times p} . \tag{A.35}$$

This is obtained as follows. Denoting by $\epsilon > 0$ an upper bound on the spectral norm of the matrix $\Phi^+ - Q\Psi^{-1}Q'$ and by $\underline{c}_{n,h}$ the h-th column of X_n, the absolute value of the element in position (h,k) of the matrix $X_n'\left[\Phi^+ - Q\Psi^{-1}Q'\right]X_n$ can be bounded from above as follows:

$$|(X_n'\left[\Phi^+ - Q\Psi^{-1}Q'\right]X_n)_{h,k}| \leq \epsilon \|\underline{c}_{n,h}\|_2 \|\underline{c}_{n,k}\|_2 \leq \frac{1}{2}\epsilon(\|\underline{c}_{n,h}\|^2 + \|\underline{c}_{n,k}\|_2^2) , \tag{A.36}$$

where Cauchy-Schwarz inequality has been applied, together with the elementary inequality $|a||b| \leq \frac{a^2+b^2}{2}$, for $a, b \in \mathbb{R}$. Since by the assumption $\lim_{T\to\infty} \|\Phi^+ - Q\Psi^{-1}Q'\|_2 = 0$ stated in Eq. (22) one can make ϵ tend to 0 as T tends to $+\infty$, and both $\|\underline{c}_{n,h}\|_2^2$ and $\|\underline{c}_{n,k}\|_2^2$ are summations of T independent and identically distributed random variables with finite mean and finite second order moments, by applying Chebyschev's weak law of large numbers, one gets

$$\operatorname*{plim}_{T\to+\infty} \frac{1}{T} |(X_n' \left[\boldsymbol{\Phi}^+ - Q\Psi^{-1}Q'\right] X_n)_{h,k}| = 0\,. \tag{A.37}$$

Finally, one gets Eq. (A.35) from Eq. (A.37), since for a sequence of random matrices with fixed dimension, element-wise convergence in probability implies convergence in probability of the whole sequence (Lee 2010).

(iv) *The following probability limit holds:*

$$\operatorname*{plim}_{T\to+\infty} \frac{1}{T} X_n' Q\Psi^{-1}Q' X_n = \frac{1+\rho^2}{1-\rho^2}\mathbb{E}\left\{ \left(\underline{x}_{n,1} - \mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)\left(\underline{x}_{n,1} - \mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)' \right\}. \tag{A.38}$$

This is obtained as follows. Exploiting the symmetry of Q and the following Cholesky factorization (see, e.g. (Ruud 2000, Sect. 19.2))

$$\Psi^{-1} = \left(C_{\text{Chol}}^{-1}\right)' C_{\text{Chol}}^{-1}\,, \tag{A.39}$$

where

$$C_{\text{Chol}}^{-1} = \frac{1}{\sqrt{1-\rho^2}} \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & \cdots & 0 \\ -\rho & 1 & 0 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 0 & -\rho & 1 \end{bmatrix}, \tag{A.40}$$

one gets

$$X_n' Q\Psi^{-1}Q' X_n = X_n' Q'\Psi^{-1}Q X_n = \underline{\ddot{x}}_{n,1}\underline{\ddot{x}}_{n,1}' + \sum_{t=2}^{T} \left(\frac{1}{\sqrt{1-\rho^2}} (\underline{\ddot{x}}_{n,t} - \rho \underline{\ddot{x}}_{n,t-1}) \right) \left(\frac{1}{\sqrt{1-\rho^2}} (\underline{\ddot{x}}_{n,t} - \rho \underline{\ddot{x}}_{n,t-1}) \right)'. \tag{A.41}$$

Hence, from Eq. (A.41) one gets

$$\begin{aligned} &\operatorname*{plim}_{T\to+\infty} \frac{1}{T} \sum_{n=1}^{N} X_n' Q\Psi^{-1}Q' X_n \\ &= \operatorname*{plim}_{T\to+\infty} \frac{1}{T} \left[\underline{\ddot{x}}_{n,1}\underline{\ddot{x}}_{n,1}' + \sum_{t=2}^{T} \left(\frac{1}{\sqrt{1-\rho^2}} (\underline{\ddot{x}}_{n,t} - \rho \underline{\ddot{x}}_{n,t-1}) \right) \left(\frac{1}{\sqrt{1-\rho^2}} (\underline{\ddot{x}}_{n,t} - \rho \underline{\ddot{x}}_{n,t-1}) \right)' \right]. \end{aligned} \tag{A.42}$$

Now, we compute the probability limit in the right-hand side of Eq. (A.42) by considering separately the following various terms.

(iv.a) The following holds:

$$\operatorname*{plim}_{T\to+\infty} \frac{1}{T} \underline{\ddot{x}}_{n,1}\underline{\ddot{x}}_{n,1}' = 0_{p\times p}\,. \tag{A.43}$$

This is obtained by applying directly Chebyschev's inequality (Ruud 2000, Section D.2), since each element of the matrix $\underline{\ddot{x}}_{n,1}\underline{\ddot{x}}_{n,1}'$ has finite mean and finite second order moments.

(iv.b) Similarly, since the addition of a finite number of terms like the one reported in Eq. (A.43) does not change the probability limit, one gets

$$\begin{aligned}\operatorname*{plim}_{T\to+\infty}\frac{1}{T}\sum_{t=2}^{T}\ddot{\underline{x}}_{n,t}\ddot{\underline{x}}'_{n,t} &= \operatorname*{plim}_{T\to+\infty}\frac{1}{T}\sum_{t=2}^{T}\ddot{\underline{x}}_{n,t-1}\ddot{\underline{x}}'_{n,t-1}\\ &= \operatorname*{plim}_{T\to+\infty}\frac{1}{T}\sum_{t=1}^{T}\ddot{\underline{x}}_{n,t}\ddot{\underline{x}}'_{n,t}\\ &= \operatorname*{plim}_{T\to+\infty}\frac{1}{T}X'_nQ'QX_n\\ &= \mathbb{E}\left\{\left(\underline{x}_{n,1}-\mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)\left(\underline{x}_{n,1}-\mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)'\right\},\end{aligned} \tag{A.44}$$

where the last equality is obtained by exploiting the eigendecomposition of $Q'Q$ (which, combined with the assumptions on the $\underline{x}_{n,t}$, shows that each element in position (h, k) of the matrix $X'_nQ'QX_n$ is the summation of $T-1$ independent random variables with mean $(\mathbb{E}\{(\underline{x}_{n,1}-\mathbb{E}\{\underline{x}_{n,1}\})(\underline{x}_{n,1}-\mathbb{E}\{\underline{x}_{n,1}\})'\})_{h,k}$ and the same finite variance), then applying Chebyschev's weak law of large numbers.

(iv.c) Moreover,

$$\begin{aligned}&\operatorname*{plim}_{T\to+\infty}\frac{1}{T}\sum_{t=2}^{T}\ddot{\underline{x}}_{n,t}\ddot{\underline{x}}'_{n,t-1}\\ &\quad= \operatorname*{plim}_{T\to+\infty}\frac{1}{T}\sum_{t=2}^{T}\left(\underline{x}_{n,t}-\frac{\sum_{\tau=1}^{T}\underline{x}_{n,\tau}}{T}\right)\left(\underline{x}_{n,t-1}-\frac{\sum_{\tau=1}^{T}\underline{x}_{n,\tau}}{T}\right)'\\ &\quad= \operatorname*{plim}_{T\to+\infty}\frac{1}{T}\sum_{t=2}^{T}\left[\underline{x}_{n,t}\underline{x}'_{n,t-1}-\underline{x}_{n,t}\frac{\sum_{\tau=1}^{T}\underline{x}'_{n,\tau}}{T}-\frac{\sum_{\tau=1}^{T}\underline{x}_{n,\tau}}{T}\underline{x}'_{n,t-1}+\frac{\sum_{\tau=1}^{T}\underline{x}_{n,\tau}\sum_{\tau=1}^{T}\underline{x}'_{n,\tau}}{T^2}\right]\\ &\quad= \mathbb{E}\left\{\underline{x}_{n,1}\right\}\mathbb{E}\left\{\underline{x}'_{n,1}\right\}-\mathbb{E}\left\{\underline{x}_{n,1}\right\}\mathbb{E}\left\{\underline{x}'_{n,1}\right\}-\mathbb{E}\left\{\underline{x}_{n,1}\right\}\mathbb{E}\left\{\underline{x}'_{n,1}\right\}+\mathbb{E}\left\{\underline{x}_{n,1}\right\}\mathbb{E}\left\{\underline{x}'_{n,1}\right\}\\ &\quad= 0_{p\times p},\end{aligned} \tag{A.45}$$

where the second-last equality comes from the fact that the probability limit of the product of two factors equals the product of the probability limits of the two factors, when the latter probability limits exist (this is a consequence of the Continuous Mapping Theorem (Florescu 2015, Theorem 7.33)), and from the assumptions on the $\underline{x}_{n,t}$.

(iv.d) Finally, Eq. (A.38) is obtained by combining items iv.a), iv.b), and iv.c), and taking into account the constant factors in Eq. (A.42).

(v) *Final part of the proof of Eq.* (23).

To conclude, one gets Eq. (23) by combining Eqs. (A.34), (A.35), and (A.38), then summing over N.

Discussion of the validity of the assumption $\lim_{T\to\infty}\|\Phi^+-Q\Psi^{-1}Q'\|_2=0$

First, we prove a related result. In the following, for a matrix $M \in \mathbb{R}^{T \times T}$, $\|M\|_{HS} = \sqrt{\frac{1}{T}\sum_{i,j=1}^{T} M_{i,j}^2}$ denotes its Hilbert-Schmidt norm (Gray 2006, Eq. (2.17)), which is a scaled version of its Frobenius norm $\|M\|_F = \sqrt{\sum_{i,j=1}^{T} M_{i,j}^2}$.

The following holds:

$$\lim_{T\to\infty} \|\Phi^+ - Q\Psi^{-1}Q'\|_{HS} = 0\,. \tag{A.46}$$

Eq. (A.46) is derived by combining several steps, which are listed next, together with pointers to some theoretical results available in the literature that are directly applied for their proofs, and checks of the assumptions of such results in the context of our analysis. In the following, for a better clarity of exposition of this part, the dependence of Ψ and other matrices on T is highlighted by including the subscript T in the notation.

(i) $\lim_{T\to\infty} \|\Psi_T - C_T\|_{HS} = 0$, where C_T is a suitable symmetric and positive definite circulant matrix[9] approximation of the symmetric Toeplitz matrix Ψ_T (application of (Gray 2006, Lemma 4.6) to the circulant matrix approximation C_T of Ψ_T coming from (Gray 2006, Eq. (4.32)), where C_T is also symmetric and positive definite due to the symmetry and positive definiteness of the Toeplitz matrix Ψ_T; the application itself of (Gray 2006, Lemma 4.6) is made possible in this case by the convergence of $1 + 2\sum_{k=1}^{+\infty} |\rho|^k$).

(ii) $\lim_{T\to\infty} \|\Phi_T - Q_T C_T Q_T'\|_{HS} = 0$ (definition of Φ_T as $\Phi_T = Q_T \Psi_T Q_T'$; combination of item i) with (Gray 2006, Lemma 2.3) and the fact that $\|Q_T\|_2 = \max_{t=1,\ldots,T} |\lambda_t(Q_T)| = 1$).

(iii) $\lim_{T\to\infty} \|\Phi_T^+ - (Q_T C_T Q_T')^+\|_{HS} = 0$ (combination of item ii) with (Wedin 1973, Theorem 4.1), made possible by the fact that Φ_T and $Q_T C_T Q_T'$ have the same rank $T-1$, and the spectral norm of Φ_T^+ and the one of $(Q_T C_T Q_T')^+$ are uniformly bounded with respect to T, due respectively to item iv) in the proof of Eq. (21) and to the characterization of the eigenvalues of C_T provided in (Gray 2006, Eq. (4.34)), combined with $m_f = \frac{1-\rho^2}{1+\rho^2+2\rho} > 0$.

(iv) $\lim_{T\to\infty} \|\Psi_T^{-1} - C_T^{-1}\|_{HS} = 0$ (application of (Gray 2006, Theorem 5.2 (b)) to the function $f(\lambda)$ reported in Eq. (A.19)).

(v) $\lim_{T\to\infty} \|Q_T \Psi_T^{-1} Q_T' - Q_T C_T^{-1} Q_T'\|_{HS} = 0$ (obtained likewise item ii)).

(vi) $Q_T C_T^{-1} Q_T' = (Q_T C_T Q_T')^+$ (obtained by exploiting the following facts: since C_T is a symmetric matrix, it has the factorization $C_T = U_T \Sigma_T U_T'$ for an orthogonal matrix $U_T \in \mathbb{R}^{T\times T}$ and a diagonal matrix $\Sigma_T \in \mathbb{R}^{T\times T}$ containing its eigenvalues, which are positive; since C_T is also a circular matrix, one can choose one column of U_T to be proportional to $\underline{1}_T$, since this is an eigenvector of C_T (Gray 2006, Theorem 3.1); Q_T represents the orthogonal projection of $\mathbb{R}^T$ onto its subspace L orthogonal to $\underline{1}_T$; as a consequence of the facts above, one can easily check that $Q_T C_T^{-1} Q_T'$ satisfies all the defining properties[10] (Barata and Hussein 2012) of the Moore-Penrose pseudoinverse

[9] We recall that a symmetric circulant matrix is a symmetric Toeplitz matrix (see footnote 7) with $m_t = m_{T-t}$, for $t = 1, \ldots, T-1$ (Gray 2006).

[10] For example, $(Q_T C_T Q_T')(Q_T C_T^{-1} Q_T')(Q_T C_T Q_T') = (Q_T C_T Q_T')$ is checked by using a common basis of eigenvectors $\underline{e}_i \in \mathbb{R}^T$ (for $i = 1, \ldots, T$) of Q_T and C_T, i.e., by showing that $(Q_T C_T Q_T')(Q_T C_T^{-1} Q_T')(Q_T C_T Q_T')\underline{e}_i = (Q_T C_T Q_T')\underline{e}_i$ for each such eigenvector.

of $Q_T C_T Q'_T$, hence $Q_T C_T^{-1} Q'_T = (Q_T C_T Q'_T)^+$ by the uniqueness of the Moore-Penrose pseudoinverse (Barata and Hussein 2012).

(vii) Finally, Eq. (A.46) is obtained by combining items iii), v), and vi).

Remark 6.2 One can easily check (e.g., by a numerical study for selected values of T and of the parameter ρ in $\boldsymbol{\Psi}$) that, in general, the stronger result $\boldsymbol{\Phi}^+ = Q\boldsymbol{\Psi}^{-1}Q'$ does not hold. This depends on the fact that, given two matrices $M_1, M_2 \in \mathbb{R}^{T\times T}$, typically $(M_1 M_2)^+ \neq M_2^+ M_1^+$, apart from particular cases (Dattorro 2019, Eq. (E.0.0.0.1)).

Eq. (22), i.e., $\lim_{T\to\infty} \|\boldsymbol{\Phi}^+ - Q\boldsymbol{\Psi}^{-1}Q'\|_2 = 0$, represents a stronger convergence requirement on $\boldsymbol{\Phi}^+ - Q\boldsymbol{\Psi}^{-1}Q'$ with respect to the convergence result provided by Eq. (A.46), which has been proved above. This depends on the fact that, for any matrix $M \in \mathbb{R}^{T\times T}$, one has $\|M\|_2 \geq \|M\|_{HS}$ (Gray 2006, Eq. (2.19)). The validity of Eq. (22) has been assumed to complete the proof of Eq. (23) - specifically, of part of item iii) therein - since a similar argument based on Eq. (A.46) would be not enough to complete that proof. Although we are currently unable to provide a formal proof of Eq. (22) - which is why it has been reported here as an assumption - its validity is strongly supported by the numerical results shown in Fig. 2, where the spectral norm error $\|\boldsymbol{\Phi}^+ - Q\boldsymbol{\Psi}^{-1}Q'\|_2$ is reported for several choices of T and ρ (similar results are obtained for a wider range of values of T and other values of ρ). The difficulty in getting a proof of Eq. (22) depends on the fact that the vector $\underline{1}_T$ is not an eigenvector of $\boldsymbol{\Psi}$ (although it is an eigenvector of its circulant matrix approximation). Hence, there is no guarantee a priori that all the elements of an orthonormal basis of eigenvectors of $\boldsymbol{\Psi}$ have nonzero orthogonal projections onto $\underline{1}_T$ (indeed, it can be easily checked numerically - e.g., by finding a basis of eigenvectors of $\boldsymbol{\Psi}$ for a few choices of T, then computing such orthogonal projections - that they are typically nonzero)[11]. This suggests, as a possible way to proceed in the proof, to investigate theoretically if such orthogonal projections converge uniformly to 0 as T tends to $+\infty$. In any case, in this appendix we have reported Eq. (A.46) together with its proof, because such equation is obviously related to Eq. (22), and because a proof of the latter could be obtained by combining Eq. (A.46) with specific properties of the matrix $\boldsymbol{\Psi}$. It is also worth mentioning that such proof would be not necessarily based on the use of a circulant matrix approximation of $\boldsymbol{\Psi}$, then of $\boldsymbol{\Psi}^{-1}$ (for which negative results on the spectral norm approximation error are known, unless a related but restricted notion of finite-term strong convergence is considered (Sun 2003, Theorem 1)).

Appendix 3: other large-sample approximations of the conditional generalization error, and associated optimization problems

In the following, we report some notes about how the analysis made in Sects. 3 and 4 can be modified if one considers, respectively, the case of large N (whose application is potentially of interest in microeconometrics), and the one in which both N and T are large. For simplicity, we limit this extension of the analysis to the case $\rho = 0$, for which one obtains

[11] Otherwise, if $\underline{1}_T$ were an eigenvector of $\boldsymbol{\Psi}$, one could proceed likewise in item vi) of the proof of Eq. (A.46).

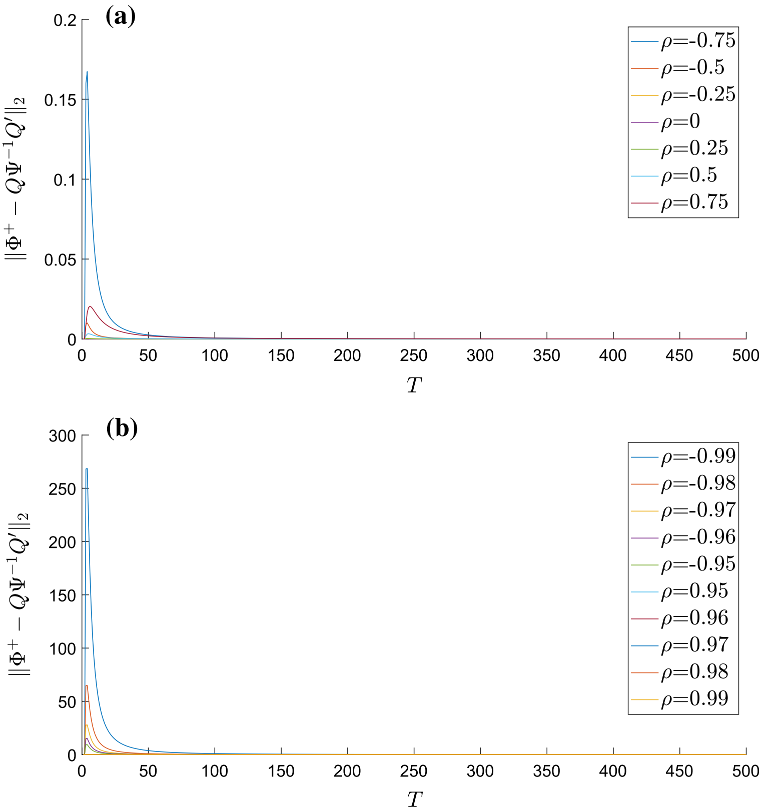


Fig. 2 Spectral norm error $\|\boldsymbol{\Phi}^+ - Q\boldsymbol{\Psi}^{-1}Q'\|_2$ as a function of T for **a** several choices of $\rho \in (-1, 1)$, and **b** other choices of $\rho \in (-1, 1)$ near either -1 or 1

the simplified expressions $\boldsymbol{\Psi} = I_T$ and $\boldsymbol{\Phi} = Q\boldsymbol{\Psi}Q' = QQ' = I_T - \frac{1}{T}\underline{1}_T\underline{1}'_T$. Then, one gets $Q'\boldsymbol{\Phi}^+Q = \boldsymbol{\Phi}^+ = QQ' = Q'Q$ by combining Eq. (A.31) and the relation between the singular value decomposition of a matrix and the singular value decomposition of its Moore-Penrose pseudoinverse.

First, we consider the case in which N is large. Assuming stationarity and mutual independence of different observations associated with the same unit, computations of the elements of the matrix

$$\ddot{X}'_n\boldsymbol{\Phi}^+\ddot{X}_n = \ddot{X}'_nQ'\boldsymbol{\Phi}^+Q\ddot{X}_n = \ddot{X}'_nQ'Q\ddot{X}_n = \ddot{X}'_n\ddot{X}_n \tag{A.47}$$

show that

$$
\begin{aligned}
&\mathbb{E}\left\{\ddot{X}_n'\ddot{X}_n\right\} \\
&= \mathbb{E}\left\{\sum_{t=1}^{T}\left(\underline{x}_{n,t} - \frac{\sum_{\tau=1}^{T}\underline{x}_{n,\tau}}{T}\right)\left(\underline{x}_{n,t} - \frac{\sum_{\tau=1}^{T}\underline{x}_{n,\tau}}{T}\right)'\right\} \\
&= \mathbb{E}\left\{\sum_{t=1}^{T}\left(\frac{T-1}{T}\underline{x}_{n,t} - \frac{\sum_{\tau=1\,\ldots,T,\tau\neq t}\underline{x}_{n,\tau}}{T}\right)\left(\frac{T-1}{T}\underline{x}_{n,t} - \frac{\sum_{\tau=1\,\ldots,T,\tau\neq t}\underline{x}_{n,\tau}}{T}\right)'\right\} \\
&= T\Bigg(\frac{(T-1)^2}{T^2}\mathbb{E}\left\{\underline{x}_{1,t}\underline{x}_{1,t}'\right\} - 2\frac{(T-1)^2}{T^2}\mathbb{E}\left\{\underline{x}_{1,t}\right\}\mathbb{E}\left\{\underline{x}_{1,t}'\right\} \\
&\qquad + \frac{(T-1)}{T^2}\mathbb{E}\left\{\underline{x}_{1,t}\underline{x}_{1,t}'\right\} + \frac{(T-1)(T-2)}{T^2}\mathbb{E}\left\{\underline{x}_{1,t}\right\}\mathbb{E}\left\{\underline{x}_{1,t}'\right\}\Bigg) \\
&= T\left(\frac{(T-1)T}{T^2}\mathbb{E}\left\{\underline{x}_{1,t}\underline{x}_{1,t}'\right\} - \frac{(T-1)T}{T^2}\mathbb{E}\left\{\underline{x}_{1,t}\right\}\mathbb{E}\left\{\underline{x}_{1,t}'\right\}\right) \\
&= (T-1)\mathbb{E}\left\{\left(\underline{x}_{n,1} - \mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)\left(\underline{x}_{n,1} - \mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)'\right\}.
\end{aligned}
\tag{A.48}
$$

Under mild technical conditions (e.g., under the additional assumption that mutual independence extends to all the $\underline{x}_{n,t}$, including those associated with different units, and that all the $\underline{x}_{n,t}$ are identically distributed[12] and have finite moments up to the order 4), from Eq. (A.48) one gets, applying Chebyschev's weak law of large numbers likewise in part of Appendix 2,

$$
\operatorname*{plim}_{N\to+\infty}\frac{1}{N(T-1)}\sum_{n=1}^{N}\ddot{X}_n'\ddot{X}_n = A\,, \tag{A.49}
$$

where

$$
A = A' := \mathbb{E}\left\{\left(\underline{x}_{1,1} - \mathbb{E}\left\{\underline{x}_{1,1}\right\}\right)\left(\underline{x}_{1,1} - \mathbb{E}\left\{\underline{x}_{1,1}\right\}\right)'\right\} \tag{A.50}
$$

is a symmetric and positive semi-definite matrix. Likewise for what concerns A_N in Sect. 3, the positive definiteness of A is also assumed in the following.

When (A.49) holds and $\rho = 0$, using also Eqs. (18) and (19) and the property $Q\underline{1}_T = \underline{0}_T$, one gets the following large-sample approximation with respect to N for the conditional generalization error (17), where the dependence on N has been highlighted:

[12] This assumption could be relaxed in order to apply in the analysis another suitable form of the weak law of large numbers, valid for the case of dependent/not identically distributed random variables.

$$
\begin{aligned}
(17) \simeq & \frac{1}{N}\frac{\sigma^2}{(T-1)T^2}\underline{1}'_T X_i A^{-1} X'_i \underline{1}_T + \frac{\sigma^2}{T} \\
& + \frac{1}{N}\frac{\sigma^2}{T-1}\mathbb{E}\left\{ \left(\underline{x}_i^{test}\right)' A^{-1} \underline{x}_i^{test} \right\} - \frac{2}{N}\frac{\sigma^2}{(T-1)T}\underline{1}'_T X_i A^{-1} \mathbb{E}\left\{\underline{x}_i^{test}\right\} \\
= & \frac{\sigma^2}{T} + \frac{1}{N}\frac{\sigma^2}{T-1}\mathbb{E}\left\{ \left\| A^{-\frac{1}{2}} \left(\frac{1}{T}\left(\underline{1}'_T X_i\right)' - \underline{x}_i^{test} \right) \right\|_2^2 \right\},
\end{aligned} \tag{A.51}
$$

where $A^{-\frac{1}{2}}$ is the principal square root of the symmetric and positive definite matrix A^{-1}.

Second, we consider the case in which both N and T are large. In this case, (A.49) is replaced by

$$
\underset{N,T \to +\infty}{\operatorname{plim}} \frac{1}{N(T-1)} \sum_{n=1}^{N} \ddot{X}'_n \ddot{X}_n = A\,, \tag{A.52}
$$

for the same matrix A as above.

When (20) and (A.52) hold and $\rho = 0$, the conditional generalization error (17) has the following large-sample approximation with respect to N and T:

$$
\begin{aligned}
(17) \simeq & \frac{1}{N}\frac{\sigma^2}{T-1}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)' A^{-1} \mathbb{E}\left\{\underline{x}_{i,1}\right\} + \frac{\sigma^2}{T} \\
& + \frac{1}{N}\frac{\sigma^2}{T-1}\mathbb{E}\left\{ \left(\underline{x}_i^{test}\right)' A^{-1} \underline{x}_i^{test} \right\} - \frac{2}{N}\frac{\sigma^2}{T-1}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)' A^{-1} \mathbb{E}\left\{\underline{x}_i^{test}\right\} \\
= & \frac{\sigma^2}{T} + \frac{\sigma^2}{N(T-1)}\mathbb{E}\left\{ \left\| A^{-\frac{1}{2}} \left(\mathbb{E}\left\{\underline{x}_{i,1}\right\} - \underline{x}_i^{test} \right) \right\|_2^2 \right\}.
\end{aligned} \tag{A.53}
$$

Starting from the large-sample approximations (A.51) and (A.53) for the conditional generalization error, and adopting the model (27) for the variance σ^2, two optimization problems similar to (28) can be stated and solved. For simplicity, in the following we make some approximations in the analysis of their optimal solutions.

In the first problem, one optimizes the corresponding large-sample approximation of the conditional generalization error with respect to N (or equivalently, with respect to c, as in (28)), whereas T is fixed. More precisely, for C sufficiently large (in such a way that the large-sample approximation (A.51) can be assumed to hold for every $c \in [c_{\min}, c_{\max}]$) and under the approximation $NTc \simeq C$ at optimality[13], setting

$$
K'_i := \mathbb{E}\left\{ \left\| A^{-\frac{1}{2}} \left(\frac{1}{T}\left(\underline{1}'_T X_i\right)' - \underline{x}_i^{test} \right) \right\|_2^2 \right\}, \tag{A.54}
$$

the first optimization problem can be written as

[13] This follows from the fact that the large-sample approximation (A.51) of the conditional generalization error is a decreasing function of N, for each fixed choice of the measurement noise variance σ^2, hence for each choice of c.

$$\begin{aligned}&\underset{c\in[c_{\min},c_{\max}]}{\text{minimize}}\left(\frac{kc^{-\alpha}}{T}+K_i'\frac{kc^{-\alpha}}{\frac{C}{c}\left(1-\frac{1}{T}\right)}\right)\\&=\underset{c\in[c_{\min},c_{\max}]}{\text{minimize}}\,\frac{1}{T}\left(kc^{-\alpha}+\frac{T^2K_i'k}{C(T-1)}c^{1-\alpha}\right),\end{aligned}\tag{A.55}$$

whose optimal solution c° has the following expression:

1. if $0<\alpha<1$ ("decreasing returns of scale") and
   (a) $c^\star := \frac{C(T-1)\alpha}{K_i'T^2(1-\alpha)} \in [c_{\min}, c_{\max}]$: $c^\circ = c^\star$;
   (b) $c^\star < c_{\min}$: $c^\circ = c_{\min}$;
   (c) $c^\star > c_{\max}$: $c^\circ = c_{\max}$;
2. if $\alpha > 1$ ("increasing returns of scale"): $c^\circ = c_{\max}$;
3. if $\alpha = 1$ ("constant returns of scale"): $c^\circ = c_{\max}$.

The analysis of the second problem (whose optimization variables are c, N, and T, as the large-sample approximation of the conditional generalizarion error with respect to both N and T is optimized) is slightly more involved, since it is formulated in terms of a larger number of optimization variables. Nevertheless, solving such problem can be reduced to solving, for each c, an optimization subproblem in which the same objective function is minimized with respect to the pair (N, T). In this problem, admissible such pairs have to satisfy the constraint $NTc \leq C$, and also two additional lower bounds $N \geq N_{\min} > 0$ and on $T \geq T_{\min} > 0$, under which the large-sample approximation made in (A.53) can be assumed to hold. More precisely, for C sufficiently large and under the approximations $T-1 \simeq T$ and $NTc \simeq C$ at optimality, setting

$$K_i'' := \mathbb{E}\left\{\left\|A^{-\frac{1}{2}}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}-\underline{x}_i^{test}\right)\right\|_2^2\right\},\tag{A.56}$$

the second optimization problem can be written as

$$\begin{aligned}&\underset{c\in[c_{\min},c_{\max}]}{\text{minimize}}\left(\frac{kc^{-\alpha}}{T}+K_i''\frac{kc^{-\alpha}}{\frac{C}{c}}\right)\\&\text{s. t. }\frac{C}{Tc}\geq N_{\min}, T\geq T_{\min}\,,\\&=\underset{c\in[c_{\min},c_{\max}]}{\text{minimize}}\left(\frac{k}{T}c^{-\alpha}+\frac{K_i''k}{C}c^{1-\alpha}\right)\\&\text{s. t. }\frac{C}{Tc}\geq N_{\min}, T\geq T_{\min}\,,\end{aligned}\tag{A.57}$$

whose optimal solutions c° have the following expressions (the optimal T is $T^\circ \simeq \frac{C}{N_{\min}c^\circ}$):

1. if $0<\alpha<1$ ("decreasing returns of scale"): $c^\circ = c_{\min}$;

2. if $\alpha > 1$ ("increasing returns of scale"): $c^\circ = c_{\max}$;
3. if $\alpha = 1$ ("constant returns of scale"): $c^\circ =$ any cost c in the interval $[c_{\min}, c_{\max}]$.

Funding Open access funding provided by Scuola IMT Alti Studi Lucca within the CRUI-CARE Agreement. G. Gnecco and F. Nutarelli are members of GNAMPA (Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni) - INdAM (Istituto Nazionale di Alta Matematica). The work was partially supported by the 2020 Italian project "Trade-off between Number of Examples and Precision in Variations of the Fixed-Effects Panel Data Model", funded by INdAM-GNAMPA.

Availability of data and material The data used for the simulation results are available upon request to the corresponding author.

Code availability The MATLAB code written to get the simulation results is available upon request to the corresponding author.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest/competing interests.

References

Aitken A. C. (1936). On least-squares and linear combinations of observations, Proceedings of the Royal Society of Edinburgh, 55, pp. 42-48.

Arellano, M. (2004). *Panel data econometrics*. Oxford: Oxford University Press.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences, 113,* 7353–7360.

Bai, Z., Cheng, P. E., & Zhang, C.-H. (1997). An extension of the Hardy-Littlewood strong law. *Statistica Sinica, 7,* 923–928.

Barata, J. C. A., & Hussein, M. S. (2012). The Moore-Penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics, 42,* 146–165.

Bargagli Stoffi, F. J., & Gnecco, G. (2018). Estimating heterogeneous causal effects in the presence of irregular assignment mechanisms, in Proceedings of the 5th IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA 2018), Turin, Italy, pp. 1-10.

Bargagli Stoffi, F. J., & Gnecco, G. (2020). Causal tree with instrumental variable: An extension of the causal tree framework to irregular assignment mechanisms. *International Journal of Data Science and Analytics, 9,* 315–337.

Barlow, R. J. (1989). *Statistics: A guide to the use of statistical methods in the physical sciences*. Hoboken: Wiley.

Bennett, K. P., & Parrado-Hernández, E. (2006). The interplay of optimization and machine learning research. *Journal of Machine Learning Research, 7,* 1265–1281.

Bhargava, A., Franzini, L., & Narendranathan, W. (1982). Serial correlation and the fixed effects model. *Review of Economic Studies, 49,* 533–549.

Bianchini, M., Frasconi, P., Gori, M., & Maggini, M. (1998). Optimal learning in artificial neural networks: A theoretical view. In: Leondes, C. T. (ed.), Neural networks systems, techniques and applications, (pp. 1-51).

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge: Cambridge University Press.

Dattorro, J. (2019). Convex analysis † Euclidean distance geometry, *Mεβoo* Publishing, https://web.stanford.edu/group/SOL/Books/0976401304.pdf.

Florescu, I. (2015). *Probability and stochastic processes*. Hoboken: Wiley.

Frees, E. W. (2004). *Longitudinal and panel data: Analysis and applications in the social sciences*. Cambridge: Cambridge University Press.

Gerschgorin, S. (1931). Über die abgrenzung der eigenwerte einer matrix. *Izvestija Akademii Nauk SSSR, Serija Matematika, 7,* 749–754.

Gilgen, H. (2006). *Univariate time series in geosciences: Theory and examples*. Berlin: Springer.

Gnecco, G., Gori, M., & Sanguineti, M. (2013). Learning with boundary conditions. *Neural Computation, 25,* 1029–1106.

Gnecco G., & Nutarelli, F. (2019). On the trade-off between number of examples and precision of supervision in regression problems, in Proceedings of the 4th International Conference of the International Neural Network Society on Big Data and Deep Learning (INNS BDDL 2019), Sestri Levante, Italy, pp. 1-6.

Gnecco, G., & Nutarelli, F. (2019). On the trade-off between number of examples and precision of supervision in machine learning problems. *Optimization Letters*. https://doi.org/10.1007/s11590-019-01486-x.

Gnecco, G., & Nutarelli, F. (2020). Optimal trade-off between sample size and precision of supervision for the fixed effects panel data model. In Proceedings of the 5th International Conference on machine Learning, Optimization & Data science (LOD 2019), Certosa di Pontignano (Siena), Italy. Lecture Notes in Computer Science, vol. 11943, pp. 1-12.

Gnecco, G., Nutarelli, F., & Selvi, D. (2020). Optimal trade-off between sample size, precision of supervision, and selection probabilities for the unbalanced fixed effects panel data model. *Soft Computing, 24,* 15937–15949.

Gori, M. (2017). *Machine learning: A constraint-based approach*. Burlington: Morgan Kaufmann.

Gray, R. M. (2006). *Toeplitz and circulant matrices: A review*. Delft: Now Publishers.

Greene, W. H. (2003). *Econometric analysis*. Delhi: Pearson Education.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken: Wiley-Interscience.

Härdle, W., Mori, Y., & Vieu, P. (2007). *Statistical methods for biostatistics and related fields*. Berlin: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Berlin: Springer.

Im, K. S., Ahn, S. C., Schmidt, P., & Wooldridge, J. M. (1999). Efficient estimation of panel data models with strictly exogenous explanatory variables. *Journal of Econometrics, 93,* 177–201.

Lee, M.-J. (2010). *Micro-econometrics: Methods of moments and limited dependent variables*. Berlin: Springer.

Kiefer, N. M. (1980). Estimation of fixed effect models for time series of cross-sections with arbitrary intertemporal covariance. *Journal of Econometrics, 14,* 195–202.

Maciejewski, A. A., & Klein, C. A. (1985). Obstacle avoidance for kinematically redundant manipulators in dynamically varying environments. *International Journal of Robotics Research, 4,* 109–116.

Nguyen, H. T., Kosheleva, O., Kreinovich, V., & Ferson, S. (2009). Trade-off between sample size and accuracy: Case of measurements under interval uncertainty. *International Journal of Approximate Reasoning, 50,* 1164–1176.

Özöğür-Akyüz, S., Ünay, D., & Smola, A. (2011). Guest editorial: Model selection and optimization in machine learning. *Machine Learning, 85,* 1–12.

Parlett, B. N. (1998). *The symmetric eigenvalue problem*. Philadelphia: SIAM.

Rao, C. P. (1973). *Linear statistical inference and its applications*. Hoboken: Wiley.

Reeve, C. P. (1988). A new statistical model for the calibration of force sensors, NBS Technical Note 1246, National Bureau of Standards, pp. 1-41.

Ruud, P. A. (2000). *An introduction to classical econometric theory*. Oxford: Oxford University Press.

Sra, S., Nowozin, S., & Wright, S. J. (Eds.). (2011). *Optimization for machine learning*. Cambridge: MIT Press.

Strang, G. (1993). The fundamental theorem of linear algebra. *The American Mathematical Monthly, 100,* 848–855.

Sun, F.-W. (2003). On the convergence of the inverses of Toeplitz matrices and its applications. *IEEE Transactions in Information Theory, 40,* 180–190.

Vapnik, V. N. (1998). *Statistical learning theory*. Hoboken: Wiley-Interscience.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28,* 3–38.

Wedin, P. A. (1973). Perturbation theory for pseudo-inverses. *BIT, 13,* 217–232.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge: MIT Press.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Giorgio Gnecco[1] · **Federico Nutarelli**[1] · **Daniela Selvi**[2]

✉ Giorgio Gnecco
giorgio.gnecco@imtlucca.it

Federico Nutarelli
federico.nutarelli@imtlucca.it

Daniela Selvi
daniela.selvi@unifi.it

[1] AXES Research Unit, IMT - School for Advanced Studies, Piazza San Francesco, 19 - 55100 Lucca, Italy

[2] Dipartimento di Ingegneria Industriale (DIEF), Università di Firenze, Via di Santa Marta, 3 - 50139 Firenze, Italy