Soft Computing (2020) 24:15937–15949
https://doi.org/10.1007/s00500-020-05317-5

FOUNDATIONS

Optimal trade-off between sample size, precision of supervision, and selection probabilities for the unbalanced fixed effects panel data model

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

Published online: 25 September 2020

Abstract
This paper is focused on the unbalanced fixed effects panel data model. This is a linear regression model able to represent unobserved heterogeneity in the data, by allowing each two distinct observational units to have possibly different numbers of associated observations. We specifically address the case in which the model includes the additional possibility of controlling the conditional variance of the output given the input and the selection probabilities of the different units per unit time. This is achieved by varying the cost associated with the supervision of each training example. Assuming an upper bound on the expected total supervision cost and fixing the expected number of observed units for each instant, we analyze and optimize the trade-off between sample size, precision of supervision (the reciprocal of the conditional variance of the output) and selection probabilities. This is obtained by formulating and solving a suitable optimization problem. The formulation of such a problem is based on a large-sample upper bound on the generalization error associated with the estimates of the parameters of the unbalanced fixed effects panel data model, conditioned on the training input dataset. We prove that, under appropriate assumptions, in some cases "many but bad" examples provide a smaller large-sample upper bound on the conditional generalization error than "few but good" ones, whereas in other cases the opposite occurs. We conclude discussing possible applications of the presented results, and extensions of the proposed optimization framework to other panel data models.

Keywords Unbalanced fixed effects panel data model · Noise variance control · Generalization error · Large-sample approximation · Optimal sample size and selection probabilities

1 Introduction

In many situations involving economics, engineering, physics, and other fields, it is required to approximate a function on the basis of a finite set of input–output noisy examples. This belongs to the typical class of problems addressed by supervised machine learning (Vapnik 1998). In some cases, the output noise variance can be reduced to some extent, by increasing the cost of each supervision. For example, devices with higher precision could be used to acquire measurements, or experts could be involved in the data analysis procedure. However, in the presence of a budget constraint, increasing the cost of each supervision could reduce the total number of available labeled examples. In such cases, the investigation of an optimal trade-off between the sample size and the precision of supervision plays a key role. In Gnecco and Nutarelli (2019a), this analysis was carried out by employing the classical linear regression model, suitably modified in order to include the additional possibility of controlling the conditional variance of the output given the input. Specifically, this was pursued by varying the time (hence, the cost) dedicated to the supervision of each training example, and

Communicated by A. Di Nola.

✉ 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 degli Studi di Firenze, Via di Santa Marta, 3 - 50139 Firenze, Italy

fixing an upper bound on the total available supervision time. Based on a large-sample approximation of the output of the ordinary least squares regression algorithm, it was shown therein that the optimal choice of the supervision time per example is highly dependent on the noise model. The analysis was refined in Gnecco and Nutarelli (2019b)[1], where an additional algorithm (weighted least squares) was considered, and shown to produce similar results at optimality as the ordinary least squares algorithm, for a model in which different training examples are possibly associated with different supervision times.

In this work, we analyze the optimal trade-off between sample size, precision of supervision, and selection probabilities for a more general linear model of the input–output relationship, which is the unbalanced fixed effects panel data model. The (either balanced or unbalanced) fixed effects model is commonly applied in the econometric analysis of microeconomic and macroeconomic data (Andreß et al. 2013; Arellano 2004; Cameron and Trivedi 2005; Wooldridge 2002), where each unit may represent, e.g., a firm, or a country. It is also applied, among other fields, in biostatistics (Härdle et al. 2007), educational research (Sherron et al. 2000), engineering (Reeve 1988; Yu et al. 2018; Zeifman 2015), neuroscience (Friston et al. 1999), political science (Bell and Jones 2014), and sociology (Frees 2004). In a fixed effects panel data model, observations related to different observational units (individuals) are associated with possibly different constants, which are able to represent unobserved heterogeneity in the data. Moreover, the same unit is observed along another dimension, which is typically time. In the unbalanced case, at each instant, different units may be not observed with some positive probability (possibly unit-dependent), resulting in a possibly unbalanced panel. In this framework, the balanced case corresponds to the situation in which the number of observations is the same for all the units.

The present work extends significantly the analysis of our previous conference article (Gnecco and Nutarelli 2020) to the unbalanced fixed effects panel data model, which is more general than the balanced case considered therein, and leads to an optimization problem that is more complex to investigate. Indeed, in Gnecco and Nutarelli (2020), all the units are always selected at each instant, therefore the selection probabilities do not appear as optimization variables in the corresponding model. Moreover, theoretical arguments are reported in much more details in the current work.

The results that will be presented in this paper concerning the unbalanced fixed effects panel data model are consistent with those of Gnecco and Nutarelli (2020) for the balanced case, and those of Gnecco and Nutarelli (2019a, b) concerning simpler linear regression models. Specifically, we show that, also for the unbalanced fixed effects panel data model, the following holds. When the precision of the supervision increases less than proportionally with respect to the supervision cost per example, the minimum (large-sample upper bound on the) generalization error (conditioned on the training input dataset) is obtained in correspondence of the smallest supervision cost per example. As a consequence of the problem formulation, this corresponds to the choice of the largest number of examples. Instead, when the precision of the supervision increases more than proportionally with respect to the supervision cost per example, the optimal supervision cost per example is the largest one. Again, as a consequence of the problem formulation, this corresponds to the choice of the smallest number of examples. The structure of the optimal selection probabilities is also investigated, under the constraint of a constant expected number of observed units for each instant. In summary, the results of the theoretical analyses performed, for different regression models of increasing complexity, in Gnecco and Nutarelli (2019a, b, 2020), and in this paper highlight that, in some circumstances, collecting a smaller number of more reliable data is preferable than increasing the size of the sample set. This looks particularly relevant when one is given a certain flexibility in designing the data collection process.

Up to our knowledge, the analysis and the optimization of the trade-off between sample size, precision of supervision, and selection probabilities in regression has been carried out rarely in the machine-learning literature. Nevertheless, the approach applied in this paper resembles the one used in the optimization of sample survey design, where some of the design parameters are optimized to minimize the sampling variance (Groves et al. 2004). Such an approach is also similar to the one exploited in Nguyen et al. (2009) for the optimization of the design of measurement devices. In that framework, however, linear regression is marginally involved, since only arithmetic averages of measurement results are considered therein. The search for optimal sample designs can be also performed by the Optimal Computing Budget Allocation (OCBA) method (Chen and Lee 2010). Differently from that approach, however, our analysis provides the optimal design a priori, i.e., before actually collecting the data. Our work can also be related to recent literature dealing with the joint application of machine learning, optimization, and econometrics (Varian 2014; Athey and Imbens 2016; Bargagli Stoffi and Gnecco 2018, 2019; Crane-Droesch 2017). For instance, the generalization error—which is typically investigated by machine learning, and optimized by solving suitable optimization problems—is not addressed in the classical analysis of the either balanced or unbalanced fixed effects panel data model (Wooldridge 2002, Chapters 10 and

[1] A short abstract version of Gnecco and Nutarelli (2019b) was presented at the session "Optimization in Machine Learning" of the International Conference on Optimization and Decision Science (ODS 2019), see Gnecco and Nutarelli (2019c).

17). Finally, an advantage of the approach considered in this work with respect to other possible ones grounded on Statistical Learning Theory (SLT) (Vapnik 1998) is that, being based on a large-sample approximation, it provides bounds on the conditional generalization error that do not need any a-posteriori evaluation of empirical risks.

The paper is structured as follows. Section 2 provides a background on the unbalanced fixed effects panel data model. Section 3 presents the analysis of its conditional generalization error, and of the large-sample upper bound on the latter with respect to time. Section 4 formulates and solves the optimization problem modeling the trade-off between sample size, precision of supervision, and selection probabilities for the unbalanced fixed effects panel data model, using the large-sample upper bound above. Finally, Sect. 5 discusses some possible applications and extensions of the theoretical results obtained in the work.

2 Background

We recall some basic facts about the following (static) unbalanced fixed effects panel data model (see, e.g., (Wooldridge 2002, Chapters 10 and 17)). Let $n = 1, \ldots, N$ denote observational units and, for each n, let $t = 1, \ldots, T_n$ be time instants. Moreover, let the inputs $\boldsymbol{x}_{n,t}$ ($n = 1, \ldots, N, t = 1, \ldots, T_n$) to the model be random column vectors in $\mathbb{R}^p$ and, for each $n = 1, \ldots, N$ and $t = 1, \ldots, T_n$, let the output $y_{n,t} \in \mathbb{R}$ be a scalar. The parameters of the model are some individual constants η_n ($n = 1, \ldots, N$), one for each unit, and a column vector $\boldsymbol{\beta} \in \mathbb{R}^p$. The (noise-free) input–output relationship is expressed as follows:

$$y_{n,t} := \eta_n + \boldsymbol{\beta}' \boldsymbol{x}_{n,t}, \quad n = 1, \ldots, N, \quad t = 1, \ldots, T_n. \tag{1}$$

Equation (1) represents an unbalanced panel data model, which can be applied in the following two situations:

- distinct units n are associated with possibly different numbers T_n of data collected at each time instant $t = 1, \ldots, T_n$ over a whole observation period $T \geq \max_{n=1}^{N} T_n$;
- the observations related to the same unit are associated with a subsequence $\{t_1, t_2, \ldots, t_{T_n}\}$ of the sequence $\{1, 2, \ldots, T\}$.

In the next sections, we focus on the second situation. To avoid burdening the notation by introducing an additional index, we still indicate, also in this case, by $\{1, 2, \ldots, T_n\}$ the subsequence $\{t_1, t_2, \ldots, t_{T_n}\}$. A possible way to get different numbers of observations T_n for distinct units consists in associating to each unit n a scalar $q_n \in (0, 1]$, which denotes the (positive) probability that n is observed at any time t. Selections for different units are supposed to be mutually independent. For simplicity, for each unit, selections at different times are also assumed to be mutually independent. For a total observation time T, denoting by $\mathbb{E}$ the expectation operator, the expected number of observations for each unit n is $\mathbb{E}\{T_n\} = q_n T$. The balanced case, which was considered in the analysis of Gnecco and Nutarelli (2020), corresponds to the situation $q_n = 1$ for each n.

Let $\{\varepsilon_{n,t}\}_{n=1,\ldots,N,\, t=1,\ldots,T_n}$ be a collection of mutually independent and identically distributed random variables, having mean 0 and the same variance σ^2. Moreover, let all the $\varepsilon_{n,t}$ be independent also from all the $\boldsymbol{x}_{n,t}$. It is assumed that noisy measurements $\tilde{y}_{n,t}$ of the outputs $y_{n,t}$ are available; specifically, the following additive noise model is considered:

$$\tilde{y}_{n,t} = y_{n,t} + \varepsilon_{n,t}, \quad n = 1, \ldots, N, \quad t = 1, \ldots, T_n. \tag{2}$$

The input–output pairs $(\boldsymbol{x}_{n,t}, \tilde{y}_{n,t})$ for $n = 1, \ldots, N$, $t = 1, \ldots, T_n$, are used to train the model, i.e., to estimate its parameters. In the following, for $n = 1, \ldots, N$, let $\boldsymbol{X}_n \in \mathbb{R}^{T_n, p}$ denote the matrix whose rows are the transposes of the $\boldsymbol{x}_{n,t}$; $\tilde{\boldsymbol{y}}_n$ be the column vector that collects the noisy measurements $\tilde{y}_{n,t}$; $\boldsymbol{I}_{T_n} \in \mathbb{R}^{T_n \times T_n}$ denote the identity matrix; $\boldsymbol{1}_{T_n} \in \mathbb{R}^{T_n}$ be the column vector whose elements are all equal to 1; and

$$\boldsymbol{Q}_n := \boldsymbol{I}_{T_n} - \frac{1}{T_n} \boldsymbol{1}_{T_n} \boldsymbol{1}_{T_n}' \tag{3}$$

be a symmetric and idempotent matrix, i.e., such that $\boldsymbol{Q}_n' = \boldsymbol{Q}_n = \boldsymbol{Q}_n^2$. Hence, for each unit n,

$$\boldsymbol{Q}_n \boldsymbol{X}_n = \begin{bmatrix} \boldsymbol{x}_{n,1} - \frac{1}{T_n} \sum_{t=1}^{T_n} \boldsymbol{x}_{n,t} \\ \boldsymbol{x}_{n,2} - \frac{1}{T_n} \sum_{t=1}^{T_n} \boldsymbol{x}_{n,t} \\ \cdots \\ \boldsymbol{x}_{n,T_n} - \frac{1}{T_n} \sum_{t=1}^{T_n} \boldsymbol{x}_{n,t} \end{bmatrix}, \tag{4}$$

and

$$\boldsymbol{Q}_n \tilde{\boldsymbol{y}}_n = \begin{bmatrix} \tilde{y}_{n,1} - \frac{1}{T_n} \sum_{t=1}^{T_n} \tilde{y}_{n,t} \\ \tilde{y}_{n,2} - \frac{1}{T_n} \sum_{t=1}^{T_n} \tilde{y}_{n,t} \\ \cdots \\ \tilde{y}_{n,T_n} - \frac{1}{T_n} \sum_{t=1}^{T_n} \tilde{y}_{n,t} \end{bmatrix} \tag{5}$$

represent, respectively, the matrix of time de-meaned training inputs, and the vector of time de-meaned corrupted training outputs. The aim of time de-meaning is to generate another dataset that does not include the fixed effects, making it possible to estimate first the vector $\boldsymbol{\beta}$, then—going back to the original dataset—the fixed effects η_n.

Assuming in the following the invertibility of the matrix $\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n \boldsymbol{X}_n$ (see the next Remark 3.2 for a mild condition ensuring this), the fixed effects estimate of $\boldsymbol{\beta}$ for the

unbalanced case is

$$\hat{\boldsymbol{\beta}}_{\mathrm{FE}} := \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n \tilde{\boldsymbol{y}}_n\right) = \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \tilde{\boldsymbol{y}}_n\right). \tag{6}$$

The unbalanced Fixed Effects (FE) estimates of the η_n, for $n = 1, \ldots, N$, are

$$\hat{\eta}_{n,\mathrm{FE}} := \frac{1}{T_n} \sum_{t=1}^{T_n} \left(\tilde{y}_{n,t} - \hat{\boldsymbol{\beta}}_{\mathrm{FE}}' \boldsymbol{x}_{n,t}\right). \tag{7}$$

Let $\mathbf{0}_p \in \mathbb{R}^p$ be the column vector whose elements are all equal to 0. By taking expectations and recalling the respective definitions and the fact that the measurement errors have 0 mean, it follows that the estimates (6) and (7) are conditionally unbiased with respect to the training input dataset $\{\boldsymbol{X}_n\}_{n=1}^N$, i.e.,

$$\mathbb{E}\left\{\left(\hat{\boldsymbol{\beta}}_{\mathrm{FE}} - \boldsymbol{\beta}\right) | \{\boldsymbol{X}_n\}_{n=1}^N\right\} = \mathbf{0}_p, \tag{8}$$

and, for any $i = 1, \ldots, N$,

$$\mathbb{E}\left\{\left(\hat{\eta}_{i,\mathrm{FE}} - \eta_i\right) | \{\boldsymbol{X}_n\}_{n=1}^N\right\} = 0. \tag{9}$$

Finally, the covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$, conditioned on the training input dataset, is

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{FE}} | \{\boldsymbol{X}_n\}_{n=1}^N\right) = \sigma^2 \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} = \sigma^2 \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}. \tag{10}$$

3 Large-sample upper bound on the conditional generalization error

This section analyzes the generalization error associated with the FE estimates (6) and (7), conditioned on the training input dataset, by providing its large-sample approximation, and a related large-sample upper bound on it. Then, in the next section, the resulting expression is optimized, after choosing a suitable model for the variance σ^2 of the measurement noise, and imposing appropriate constraints.

Let $\boldsymbol{x}_i^{\mathrm{test}} \in \mathbb{R}^p$ be a random test vector, which is assumed to have finite mean and finite covariance matrix, and to be independent from the training data. We express the generalization error for the i-th unit ($i = 1, \ldots, N$), conditioned on the training input dataset, as follows[2]:

$$\mathbb{E}\left\{\left(\hat{\eta}_{i,\mathrm{FE}} + \hat{\boldsymbol{\beta}}_{\mathrm{FE}}' \boldsymbol{x}_i^{\mathrm{test}} - \eta_i - \boldsymbol{\beta}' \boldsymbol{x}_i^{\mathrm{test}}\right)^2 | \{\boldsymbol{X}_n\}_{n=1}^N\right\}. \tag{11}$$

The conditional generalization error (11) represents the expected mean squared error of the prediction of the output associated with a test input, conditioned on the training input dataset.

For $n = 1, \ldots, N$, let $\boldsymbol{\varepsilon}_n \in \mathbb{R}^{T_n}$ be the column vector whose elements are the $\varepsilon_{n,t}$; $\boldsymbol{\eta}_n \in \mathbb{R}^{T_n}$ be the column vector whose elements are all equal to η_n; and $\mathbf{0}_{T_n \times T_n} \in \mathbb{R}^{T_n \times T_n}$ be a matrix whose elements are all equal to 0. Noting that

$$\mathbb{E}\left\{\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_m'\right\} = \mathbf{0}_{T_n \times T_n}, \quad \text{for } n \neq m, \tag{12}$$

$$\mathbb{E}\left\{\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n'\right\} = \sigma^2 \boldsymbol{I}_{T_n}, \tag{13}$$

$$\boldsymbol{Q}_n' \boldsymbol{Q}_n = \boldsymbol{Q}_n, \tag{14}$$

$$\boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{Q}_n' \boldsymbol{Q}_n = \boldsymbol{Q}_n' \boldsymbol{Q}_n, \tag{15}$$

$$\boldsymbol{Q}_n \boldsymbol{\eta}_n = \boldsymbol{Q}_n' \boldsymbol{\eta}_n = \mathbf{0}_{T_n}, \tag{16}$$

and

$$\boldsymbol{Q}_n \mathbf{1}_{T_n} = \boldsymbol{Q}_n' \mathbf{1}_{T_n} = \mathbf{0}_{T_n}, \tag{17}$$

we can express the conditional generalization error (11) as follows, highlighting its dependence on σ^2 and T_i (see "Appendix 1" for the details):

$$\begin{aligned}
&\mathbb{E}\left\{\left(\hat{\eta}_{i,\mathrm{FE}} + \hat{\boldsymbol{\beta}}_{\mathrm{FE}}' \boldsymbol{x}_i^{\mathrm{test}} - \eta_i - \boldsymbol{\beta}' \boldsymbol{x}_i^{\mathrm{test}}\right)^2 | \{\boldsymbol{X}_n\}_{n=1}^N\right\} \\
&= \frac{\sigma^2}{T_i^2} \mathbf{1}_{T_i}' \boldsymbol{X}_i \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{X}_i' \mathbf{1}_{T_i} \\
&\quad + \frac{\sigma^2}{T_i} \\
&\quad + \mathbb{E}\left\{\sigma^2 \left(\boldsymbol{x}_i^{\mathrm{test}}\right)' \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{x}_i^{\mathrm{test}} | \{\boldsymbol{X}_n\}_{n=1}^N\right\} \\
&\quad - 2\mathbb{E}\left\{\frac{\sigma^2}{T_i} \mathbf{1}_{T_i}' \boldsymbol{X}_i \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{x}_i^{\mathrm{test}} | \{\boldsymbol{X}_n\}_{n=1}^N\right\}.
\end{aligned} \tag{18}$$

Next, we obtain a large-sample approximation of the conditional generalization error (18) with respect to T, for a fixed number N of units[3].

[2] See the next Remark 3.3 for a justification of the choice of the conditioned generalization error for the analysis, instead of its unconditional version.

[3] Such an approximation is useful, e.g., in the application of the model to macroeconomics data, for which it is common to investigate the

For $n = 1, \ldots, N$, let the symmetric and positive semi-definite matrices $\boldsymbol{A}_n \in \mathbb{R}^{p \times p}$ be defined as

$$\boldsymbol{A}_n = \boldsymbol{A}_n' := \mathbb{E}\left\{\left(\boldsymbol{x}_{n,1} - \mathbb{E}\left\{\boldsymbol{x}_{n,1}\right\}\right)\left(\boldsymbol{x}_{n,1} - \mathbb{E}\left\{\boldsymbol{x}_{n,1}\right\}\right)'\right\} \quad (19)$$

In the following, the positive definiteness (hence, the invertibility) of each matrix A_n is assumed. This is a quite mild condition because it is associated with the fact that, with positive probability, the random vectors $\boldsymbol{x}_{n,1} - \mathbb{E}\left\{\boldsymbol{x}_{n,1}\right\}$ do not belong to any given subspace of $\mathbb{R}^p$ with dimension smaller than p (so, they are effectively p-dimensional random vectors).

Under mild conditions (e.g., if the $\boldsymbol{x}_{n,t}$ are mutually independent, identically distributed, and have finite moments up to the order 4), the following convergences in probability[4] hold:

$$\begin{aligned}
&\text{plim}_{T\to+\infty} \frac{1}{T_i} \mathbf{1}_{T_i}' \boldsymbol{X}_i \\
&= \text{plim}_{T\to+\infty} \frac{1}{T_i} \sum_{t=1}^{T_i} \boldsymbol{x}_{i,t}' \\
&= \left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}\right)', \qquad (20)
\end{aligned}$$

and

$$\begin{aligned}
&\text{plim}_{T\to+\infty} \frac{1}{T} \sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n \\
&= \text{plim}_{T\to+\infty} \sum_{n=1}^{N} \frac{T_n}{T} \frac{1}{T_n} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n \\
&= \boldsymbol{A}_N, \qquad (21)
\end{aligned}$$

where

$$\boldsymbol{A}_N = \boldsymbol{A}_N' := \sum_{n=1}^{N} q_n \boldsymbol{A}_n \quad (22)$$

which is the weighted summation, with positive weights q_n, of the symmetric and positive definite matrices $\boldsymbol{A}_n$, hence it is also a symmetric and positive definite matrix.

Remark 3.1 Equations (20) and (21) follow from the extension of Chebyschev's weak law of large numbers (Ruud 2000, Section 13.4.2) to the case of the summation of a random number of mutually independent random variables (Révész 1968, Theorem 10.1), combined with other technical results. First, for each $n = 1, \ldots, N$, convergence in probability of $\frac{1}{T_n} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n$ to $\boldsymbol{A}_n$ is proved element-wise, by applying (Révész 1968, Theorem 10.1). Then, one exploits the fact that, as a consequence of the Continuous Mapping Theorem (Florescu 2015, Theorem 7.33), the probability limit of the product of two random variables (in this case, $\frac{T_n}{T}$ and each element of $\frac{1}{T_n} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n$) equals the product of their probability limits, when the latter two exist (which is the case for $\frac{T_n}{T}$ and each element of $\frac{1}{T_n} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n$). Finally, one applies the fact that, for a random matrix, element-wise convergence in probability implies convergence in probability of the whole random matrix (Lee 2010).

Remark 3.2 The existence of the probability limit (21) and the positive definiteness of the matrix $\boldsymbol{A}_N$ guarantee that the invertibility of the matrix

$$\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n \boldsymbol{X}_n = \sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n \quad (23)$$

(see Sect. 2) holds with probability close to 1 for large T. Due to the generalization of Slutsky's theorem reported in (Greene 2003, Theorem D.14)[5], under the stated assumptions also the sequence of random matrices

$$\left(\frac{1}{T} \sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \quad (24)$$

converges in probability to $\boldsymbol{A}_N^{-1}$. This is needed to obtain the next large-sample approximation (25) of the conditional generalization error.

Remark 3.3 We point out that the conditional generalization error (11) is investigated in this work, instead of its unconditional version because, in general, probability limits and expectations cannot be inverted in order. This could prevent the application of (Greene 2003, Theorem D.14) (or of similar results about probability limits) when performing a similar analysis for the unconditional generalization error.

Let $\|\cdot\|_2$ denote the l_2-norm, and $\boldsymbol{A}_N^{-\frac{1}{2}}$ be the principal square root (i.e., the symmetric and positive definite square root) of the symmetric and positive definite matrix $\boldsymbol{A}_N^{-1}$. When (20) and (21) hold, from (18) and the assumed independence of $\boldsymbol{x}_i^{\text{test}}$ from all the other random vectors we get

case of a large horizon T. The case of finite T and large N is of more interest for microeconometrics (Cameron and Trivedi 2005), and will be investigated in future research.

[4] We recall that a sequence of random real matrices $\boldsymbol{M}_T$ of the same dimension, $T = 1, 2, \ldots$, converges in probability to the real matrix $\boldsymbol{M}$ if, for every $\varepsilon > 0$, Prob $(\|\boldsymbol{M}_T - \boldsymbol{M}\| > \varepsilon)$ (where $\|\cdot\|$ is an arbitrary matrix norm) tends to 0 as T tends to $+\infty$. In this case, one writes $\text{plim}_{T\to+\infty} \boldsymbol{M}_T = \boldsymbol{M}$.

[5] It states that, given a sequence of random real square matrices $\boldsymbol{M}_T$ of the same dimension, $T = 1, 2, \ldots$, if $\text{plim}_{T\to+\infty} \boldsymbol{M}_T = \boldsymbol{B}$ and $\boldsymbol{B}$ is invertible, then also $\text{plim}_{T\to+\infty} \boldsymbol{M}_T^{-1} = \boldsymbol{B}^{-1}$.

the following large-sample approximation (with respect to T) for the conditional generalization error (11):

$$\begin{aligned}
&\mathbb{E}\left\{\left(\hat{\eta}_{i,\mathrm{FE}}+\hat{\boldsymbol{\beta}}'_{\mathrm{FE}}\boldsymbol{x}_i^{\mathrm{test}}-\eta_i-\boldsymbol{\beta}'\boldsymbol{x}_i^{\mathrm{test}}\right)^2\big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
&\simeq\frac{\sigma^2}{T}\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}\right)'\boldsymbol{A}_N^{-1}\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}\\
&\quad+\frac{\sigma^2}{q_iT}\\
&\quad+\frac{\sigma^2}{T}\mathbb{E}\left\{\left(\boldsymbol{x}_i^{\mathrm{test}}\right)'\boldsymbol{A}_N^{-1}\boldsymbol{x}_i^{\mathrm{test}}\right\}\\
&\quad-2\frac{\sigma^2}{T}\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}\right)'\boldsymbol{A}_N^{-1}\mathbb{E}\left\{\boldsymbol{x}_i^{\mathrm{test}}\right\}\\
&=\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\mathbb{E}\left\{\left\|\boldsymbol{A}_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{\mathrm{test}}\right)\right\|_2^2\right\}\right).
\end{aligned}\tag{25}$$

In the following, we denote, for a generic symmetric matrix $\boldsymbol{A}\in\mathbb{R}^{s\times s}$, by $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$, respectively, its minimum and maximum eigenvalue. Starting from the large-sample approximation (25), the following steps can be proved (see "Appendix 2" for the details):

$$\begin{aligned}
&\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\mathbb{E}\left\{\left\|\boldsymbol{A}_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{\mathrm{test}}\right)\right\|_2^2\right\}\right)\\
&\le\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\lambda_{\max}(\boldsymbol{A}_N^{-1})\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{\mathrm{test}}\right)\right\|_2^2\right\}\right)\\
&=\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\frac{1}{\lambda_{\min}(\boldsymbol{A}_N)}\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{\mathrm{test}}\right)\right\|_2^2\right\}\right)\\
&\le\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\frac{1}{\sum_{n=1}^N q_n\lambda_{\min}(\boldsymbol{A}_n)}\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{\mathrm{test}}\right)\right\|_2^2\right\}\right).
\end{aligned}\tag{26}$$

We refer to the inequality

$$\begin{aligned}
&\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\mathbb{E}\left\{\left\|\boldsymbol{A}_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{\mathrm{test}}\right)\right\|_2^2\right\}\right)\\
&\le\frac{\sigma^2}{T}\left(\frac{1}{q_i}+\frac{1}{\sum_{n=1}^N q_n\lambda_{\min}(\boldsymbol{A}_n)}\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{\mathrm{test}}\right)\right\|_2^2\right\}\right)
\end{aligned}\tag{27}$$

as the large-sample upper bound on the conditional generalization error. Interestingly, its right-hand side is expressed in the separable form $\frac{\sigma^2}{T}K_i(\{q_n\}_{n=1}^N)$, where

$$\begin{aligned}
&K_i(\{q_n\}_{n=1}^N)\\
&:=\left(\frac{1}{q_i}+\frac{1}{\sum_{n=1}^N q_n\lambda_{\min}(\boldsymbol{A}_n)}\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{\mathrm{test}}\right)\right\|_2^2\right\}\right)
\end{aligned}\tag{28}$$

depends only on the q_n. As shown in the next section, this simplifies the analysis of the trade-off between sample size, precision of supervision, and selection probabilities performed therein, since one does not need to compute the exact expression of the function $K_i(\{q_n\}_{n=1}^N)$ to find the optimal trade-off with respect to a suitable subset of optimization variables.

4 Optimal trade-off between sample size, precision of supervision, and selection probabilities

In this section, we are interested in optimizing the large-sample upper bound (27) of the conditional generalization error when the variance σ^2 is modeled as a decreasing function of the supervision cost per example c, and a given upper bound $C>0$ is imposed on the expected total supervision cost $\sum_{n=1}^N q_nTc$ associated with the whole training set. For large T, this upper bound practically coincides with the total supervision cost $\sum_{n=1}^N T_nc$. This follows by an application of Chebyschev's weak law of large numbers.

Remark 4.1 In our previous conference work (Gnecco and Nutarelli 2020), the large-sample approximation (25) was optimized, instead of (27). This was motivated by the fact that all the selection probabilities q_n were fixed to 1, implying that both q_i and A_N, hence also the term

$$\left(\frac{1}{q_i}+\mathbb{E}\left\{\left\|\boldsymbol{A}_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\}-\boldsymbol{x}_i^{\mathrm{test}}\right)\right\|_2^2\right\}\right),\tag{29}$$

were constant therein.

In the following analysis of the optimal trade-off, N is kept fixed; furthermore, one imposes the constraints

$$q_{n,\min}\le q_n\le q_{n,\max},\quad n=1,\ldots,N,\tag{30}$$

for some given $q_{n,\min}\in(0,1)$ and $q_{n,\max}\in[q_{n,\min},1]$, and

$$\sum_{n=1}^N q_n=\bar{q}N,\tag{31}$$

for some given $\bar{q}\in\left[\frac{\sum_{n=1}^N q_{n,\min}}{N},\frac{\sum_{n=1}^N q_{n,\max}}{N}\right]\subseteq(0,1]$. In Eq. (31), $\sum_{n=1}^N q_n$ represents the expected number of observed units for each instant, which is fixed. Moreover, T is chosen as $\left\lfloor\frac{C}{\bar{q}Nc}\right\rfloor$. Finally, 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 $\left\{\left\lfloor\frac{C}{\bar{q}Nc_{\max}}\right\rfloor,\ldots,\left\lfloor\frac{C}{\bar{q}Nc_{\min}}\right\rfloor\right\}$. In the following, C is supposed to be sufficiently large, so that the large-sample upper bound (27) can be assumed to hold for every $c\in[c_{\min},c_{\max}]$ and every $q_n\in[q_{n,\min},q_{n,\max}]$ (for $n=1,\ldots,N$).

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

$$\sigma^2(c) = kc^{-\alpha}, \tag{32}$$

where $k, \alpha > 0$. For $0 < \alpha < 1$, if one doubles the supervision cost per example c, then the precision $1/\sigma^2(c)$ (i.e., the reciprocal of the conditional variance of the output) becomes less than two times its initial value (or equivalently, the variance $\sigma^2(c)$ becomes more than one half its initial value). This case is referred to as "decreasing returns of scale" in the precision of each supervision. Conversely, for $\alpha > 1$, 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). This case is referred to as "increasing returns of scale" in the precision of each supervision. Finally, 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.

Summarizing, under the assumptions above, the optimal trade-off between sample size, precision of supervision, and selection probabilities for the unbalanced fixed effects panel data model is modeled by the following optimization problem:

$$\begin{aligned} &\underset{\substack{c\in[c_{\min},c_{\max}],\\ q_n\in[q_{n,\min},q_{n,\max}],\\ n=1,\ldots,N}}{\text{minimize}} \quad K_i(\{q_n\}_{n=1}^N)k\frac{c^{-\alpha}}{\left\lfloor \frac{C}{\bar{q}Nc}\right\rfloor} \\ &\text{s.t.} \sum_{n=1}^{N} q_n = \bar{q}N. \end{aligned} \tag{33}$$

By a similar argument as in the proof of (Gnecco and Nutarelli 2019b, Proposition 3.2), which refers to an analogous function approximation problem, when C is sufficiently large, the objective function $CK_i(\{q_n\}_{n=1}^N)k\frac{c^{-\alpha}}{\left\lfloor \frac{C}{\bar{q}Nc}\right\rfloor}$ of the optimization problem (33), rescaled by the multiplicative factor C, can be approximated, with a negligible error in the maximum norm on $[c_{\min}, c_{\max}] \times \Pi_{n=1}^N [q_{n,\min}, q_{n,\max}]$, by $\bar{q}NK_i(\{q_n\}_{n=1}^N)kc^{1-\alpha}$. Figure 1 shows the behavior of the rescaled objective functions

$$CK_i(\{q_n\}_{n=1}^N)k\frac{c^{-\alpha}}{\left\lfloor \frac{C}{\bar{q}Nc}\right\rfloor} \tag{34}$$

and

$$\bar{q}NK_i(\{q_n\}_{n=1}^N)kc^{1-\alpha} \tag{35}$$

for the three cases $0 < \alpha = 0.5 < 1$, $\alpha = 1.5 > 1$, and $\alpha = 1$. The values of the other parameters are $k = 0.5$,

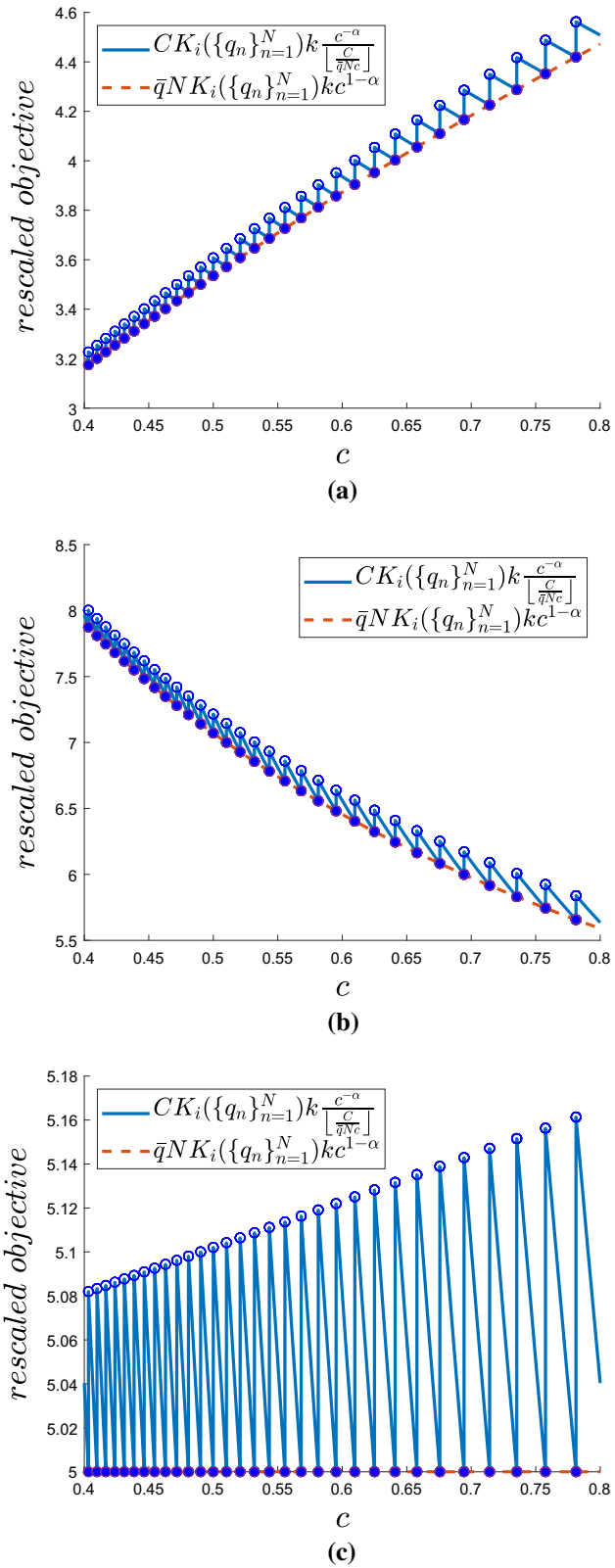

Fig. 1 Plots of the rescaled objective functions $CK_i(\{q_n\}_{n=1}^N)k\frac{c^{-\alpha}}{\left\lfloor \frac{C}{\bar{q}Nc}\right\rfloor}$ and $\bar{q}NK_i(\{q_n\}_{n=1}^N)kc^{1-\alpha}$ for $\alpha = 0.5$ (a), $\alpha = 1.5$ (b), and $\alpha = 1$ (c). The values chosen for the other parameters are detailed in the text

$\bar{q} = 0.5$, $K_i(\{q_n\}_{n=1}^N) = 2$ (which can be assumed to hold for a fixed choice of the set of the q_n), $N = 10$, $C = 125$, $c_{\min} = 0.4$, and $c_{\max} = 0.8$. One can show by standard calculus that, for $C \to +\infty$ and the q_n fixed to constant values, the number of discontinuity points of the rescaled objective function $C K_i(\{q_n\}_{n=1}^N) k \frac{c^{-\alpha}}{\left[\frac{C}{\bar{q}Nc}\right]}$ tends to infinity, whereas the amplitude of its oscillations above the lower envelope $\bar{q} N K_i(\{q_n\}_{n=1}^N) k c^{1-\alpha}$ tends to 0 uniformly with respect to $c \in [c_{\min}, c_{\max}]$.

Concluding, under the approximation above, one can replace the optimization problem (33) with

$$\begin{aligned} &\underset{\substack{c \in [c_{\min}, c_{\max}],\\ q_n \in [q_{n,\min}, q_{n,\max}],\\ n=1,\ldots,N}}{\text{minimize}} \quad \bar{q} N K_i(\{q_n\}_{n=1}^N) k c^{1-\alpha} \\ &\text{s.t.} \sum_{n=1}^N q_n = \bar{q} N. \end{aligned} \tag{36}$$

Such optimization problem appears in a separable form, in which one can optimize separately the variable c and the variables q_n, for $n = 1, \ldots, N$. In particular, the optimal solutions c° have the following expressions:

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

In summary, the results of this part of the analysis show that, in the case of "decreasing returns of scale", "many but bad" examples are associated with a smaller large-sample upper bound on the 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 (2020) for the balanced case and in Gnecco and Nutarelli (2019a, b) for simpler linear regression problems, to which the ordinary/weighted least squares algorithms were applied. This depends on the fact that, in all these cases, the large-sample approximation of the conditional generalization error (or its large-sample upper bound) has the functional form $\frac{\sigma^2}{T} K_i$, where K_i is either a constant, or depends on optimization variables related to neither σ nor T.

One can observe that, in order to discriminate among the three cases of the analysis reported above, it is not needed to know the exact values of the constants k and N, neither the expression of K_i as a function of the q_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 interval $(1, +\infty)$. Finally, for this part of the analysis, knowledge of the probability distributions of the input examples associated with the different units is limited to the determination of the expressions of the constants $\lambda_{\min}(\boldsymbol{A}_n)$ involved in the optimization of the variables q_n.

Assuming that the constant terms $\lambda_{\min}(\boldsymbol{A}_n)$ and $\mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\} - \boldsymbol{x}_i^{\text{test}}\right)\right\|_2^2\right\}$ are known, optimal q_n° can be derived as follows. First, note that, for each fixed admissible choice of q_i, the optimization of the other q_n can be restated as follows:

$$\begin{aligned} &\underset{\substack{q_n \in [q_{n,\min}, q_{n,\max}],\\ n=1,\ldots,N, n \neq i}}{\text{maximize}} \left(\lambda_{\min}(\boldsymbol{A}_i) q_i + \sum_{n=1,\ldots,N, n \neq i} \lambda_{\min}(\boldsymbol{A}_n) q_n \right) \\ &\text{s.t.} \sum_{n=1,\ldots,N, n \neq i} q_n = \bar{q} N - q_i. \end{aligned} \tag{37}$$

More precisely, an admissible choice for q_i is one for which $q_i \in \left[\hat{q}_{i,\min}, \hat{q}_{i,\max}\right]$, where

$$\hat{q}_{i,\min} := \max\{q_{i,\min}, \bar{q} N - \sum_{n=1,\ldots,N, n \neq i} q_{n,\max}\}, \tag{38}$$

and

$$\hat{q}_{i,\max} := \min\{q_{i,\max}, \bar{q} N - \sum_{n=1,\ldots,N, n \neq i} q_{n,\min}\}. \tag{39}$$

The optimization problem (37) is a linear programming one, which can be reduced to a continuous knapsack problem (Martello and Toth 1990, Section 2.2.1), after a rescaling of all its optimization variables and of their respective bounds. It is well known that, due to its particular structure, such a problem can be solved by the following greedy algorithm, which is divided into three steps (for simplicity of exposition, we assume that all the $\lambda_{\min}(\boldsymbol{A}_n)$ are different from each other):

1. first, the variables q_n are re-ordered according to decreasing values of the associated $\lambda_{\min}(\boldsymbol{A}_n)$. So, let $\breve{q}_n := q_{\pi(n)}$ and $\breve{\boldsymbol{A}}_n := \boldsymbol{A}_{\pi(n)}$, where the function $\pi : \{1, \ldots, N\} \to \{1, \ldots, N\}$ is a permutation satisfying $\lambda_{\min}(\breve{\boldsymbol{A}}_m) < \lambda_{\min}(\breve{\boldsymbol{A}}_n)$ for every $m \geq n$. Let also $\breve{i} = \pi(i)$;
2. starting from $\breve{q}_n = \breve{q}_{n,\min}$ for every $n \neq \breve{i}$, the first variable $\breve{q}_1$ (if $\breve{i} \neq 1$) is increased until either the constraint $\sum_{n=1,\ldots,N, n \neq \breve{i}} \breve{q}_n = \bar{q} N - \breve{q}_{\breve{i}}$, or the constraint $\breve{q}_1 = \breve{q}_{1,\max}$, is met; if $\breve{i} = 1$, then the procedure is applied to the second variable $\breve{q}_2$;
3. step 2 is repeated for the successive variables (excluding $\breve{q}_{\breve{i}}$), terminating the first time the constraint $\sum_{n=1,\ldots,N, n \neq \breve{i}} \breve{q}_n = \bar{q} N - \breve{q}_{\breve{i}}$ is met (this surely occurs, since q_i is admissible).

The resulting optimal q_n° (for $n = 1, \ldots, N$ with $n \neq i$) are parametrized by the remaining variable q_i. Then, the optimal

value of the objective function of the optimization problem (37) is a real-valued function of q_i which, in the following, is denoted by $f_i(q_i)$. It follows from the procedure above that $f_i(q_i)$ is a continuous and piece-wise affine function of q_i, with piece-wise constant slopes $\lambda_{\min}(\check{\boldsymbol{A}}_{\check{i}}) - \lambda_{\min}(\check{\boldsymbol{A}}_{n(q_i)})$, where the choice of the index n is a function of q_i, and is such that $\lambda_{\min}(\check{\boldsymbol{A}}_{n(q_i)})$ is a nonincreasing function of q_i. Hence, $f_i(q_i)$ is concave, and is nondecreasing for $q_i \leq \bar{q}N - \sum_{n=1}^{\check{i}-1} \check{q}_{n,\max}$, where $\check{q}_{n,\max} := q_{\pi(n),\max}$, and nonincreasing otherwise.

Exploiting the results above, the optimal value of q_i for the original optimization problem (36) is obtained by solving the following optimization problem:

$$\underset{q_i \in [\hat{q}_{i,\min}, \hat{q}_{i,\max}]}{\text{minimize}} \left(\frac{1}{q_i} + \frac{\mathbb{E}\left\{ \left\| \left(\mathbb{E}\left\{ \boldsymbol{x}_{i,1} \right\} - \boldsymbol{x}_i^{\text{test}} \right) \right\|_2^2 \right\}}{f_i(q_i)} \right). \tag{40}$$

This is a convex optimization problem, since the function $\frac{1}{q_i}$ is convex, whereas the function

$$\frac{\mathbb{E}\left\{ \left\| \left(\mathbb{E}\left\{ \boldsymbol{x}_{i,1} \right\} - \boldsymbol{x}_i^{\text{test}} \right) \right\|_2^2 \right\}}{f_i(q_i)} \tag{41}$$

is of the form $h(f_i)$, where f_i is concave and h is convex and nonincreasing, so $h(f_i)$ is convex (Boyd and Vandenberghe 2004, Section 3.2). After solving the optimization problem (40), the optimal values of the other q_n for the original optimization problem (36) are obtained as a consequence of the three steps detailed above.

It follows from the reasoning above that the structure of the optimal solutions q_n° is as follows. First, there exists a threshold $\bar{\lambda}^\circ > 0$ such that

(i) for any $n \neq i$ with $\lambda_{\min}(\boldsymbol{A}_n) > \bar{\lambda}^\circ$, q_n° is equal to its maximum admissible value $q_{n,\max}$;
(ii) for any $n \neq i$ with $\lambda_{\min}(\boldsymbol{A}_n) < \bar{\lambda}^\circ$, q_n° is equal to its minimum admissible value $q_{n,\min}$;
(iii) for at most one unit $n \neq i$ (for which $\lambda_{\min}(\boldsymbol{A}_n) = \bar{\lambda}^\circ$, provided that there exists one value of n for which this condition holds), q_n° belongs to the interior of the interval $[q_{n,\min}, q_{n,\max}]$.

Moreover,

(iv) if $\left(\bar{q}N - \sum_{n=1}^{\check{i}-1} \check{q}_{n,\max} \right) \geq \hat{q}_{i,\max}$, then

$$q_i^\circ = \hat{q}_{i,\max}, \tag{42}$$

and

$$\bar{\lambda}^\circ \in (0, \lambda_{\min}(\boldsymbol{A}_i)); \tag{43}$$

(v) if $\left(\bar{q}N - \sum_{n=1}^{\check{i}-1} \check{q}_{n,\max} \right) < \hat{q}_{i,\max}$, then

$$q_i^\circ \in \left[\left(\bar{q}N - \sum_{n=1}^{\check{i}-1} \check{q}_{n,\max} \right), \hat{q}_{i,\max} \right], \tag{44}$$

and

$$\bar{\lambda}^\circ > \lambda_{\min}(\boldsymbol{A}_i). \tag{45}$$

Finally, it is worth observing that the structure highlighted above for the optimal solutions q_n° and c° (the latter reported under Eq. (36)), which is valid for any fixed value of $\bar{q}$, can be useful to solve the modification of the optimization problem (36) obtained in case the constraint (31) is replaced by

$$\bar{q}_{\min} N \leq \sum_{n=1}^{N} q_n \leq \bar{q}_{\max} N, \tag{46}$$

for some given $\bar{q}_{\min}, \bar{q}_{\max} \in (0, 1]$, with $\bar{q}_{\min} < \bar{q}_{\max}$.

5 Conclusions

In this paper, the optimal trade-off between sample size, precision of supervision, and selection probabilities, has been studied with specific reference to a quite general linear model of input–output relationship representing unobserved heterogeneity in the data, namely the unbalanced fixed effects panel data model. First, we have analyzed its conditional generalization error, then we have minimized a large-sample upper bound on it with respect to some of its parameters. We have proved that, under suitable assumptions, "many but bad" examples provide a smaller upper bound on the conditional generalization error than "few but good" ones, whereas in other cases the opposite occurs. The choice between "many but bad" and "few but good" examples plays an important role when better supervision implies higher costs.

The theoretical results obtained in this work could be applied to the acquisition design of unbalanced panel data related to several fields, such as biostatistics, econometrics, educational research, engineering, neuroscience, political science, and sociology. Moreover, the analysis of the large-sample case could be extended to deal with large N, or with both large N and T. These cases would be of interest for their potential applications in microeconometrics (Cameron and Trivedi 2005). Another possible extension concerns the introduction, in the noise model, of a subset of not controllable parameters (beyond the controllable one, i.e., the

noise variance), which could be estimated from a subset of training data. As a final extension, one could investigate and optimize the trade-off between sample size and precision of supervision (and possibly, also selection probabilities) for the random effects panel data model (Greene 2003, Chapter 13). This is also commonly applied in the analysis of economic data, and differs from the fixed effects panel data model in that its parameters are considered as random variables. In the present context, however, a possible advantage of the fixed effects panel data model is that it also allows one to obtain estimates of the individual constants η_n (see Eq. (7)), which appear in the expression (11) of the conditional generalization error. Moreover, the application of the random effects model to the unbalanced case requires stronger assumptions than the ones needed for the application of the fixed effects model (Wooldridge 2002, Chapter 17).

Acknowledgements 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.

Funding Open access funding provided by Scuola IMT Alti Studi Lucca within the CRUI-CARE Agreement.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval/informed consent This article does not contain any studies with human participants or animals performed by any of the authors.

Appendix 1: Proof of Equation (18)

First, we expand the conditional generalization error (11) as follows:

$$
\begin{aligned}
&\mathbb{E}\left\{\left(\hat{\eta}_{i,\mathrm{FE}}+\hat{\boldsymbol{\beta}}_{\mathrm{FE}}'\boldsymbol{x}_i^{\mathrm{test}}-\eta_i-\boldsymbol{\beta}'\boldsymbol{x}_i^{\mathrm{test}}\right)^2\big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
&=\mathbb{E}\left\{\left(\left(\hat{\eta}_{i,\mathrm{FE}}-\eta_i\right)+\left(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i^{\mathrm{test}}\right)^2\big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
&=\mathbb{E}\left\{(\hat{\eta}_{i,\mathrm{FE}}-\eta_i)^2|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
&+\mathbb{E}\left\{\left(\left(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i^{\mathrm{test}}\right)^2\big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
&+2\mathbb{E}\left\{\left(\hat{\eta}_{i,\mathrm{FE}}-\eta_i\right)\left(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i^{\mathrm{test}}\big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}. \qquad (47)
\end{aligned}
$$

Exploiting the conditional unbiasedness of $\hat{\eta}_{i,\mathrm{FE}}$, and the expressions (1) of $y_{n,t}$, (2) of $\tilde{y}_{n,t}$, and (7) of $\hat{\eta}_{i,\mathrm{FE}}$ (with the index n replaced by the index i), one gets

$$
\begin{aligned}
&\mathbb{E}\left\{(\hat{\eta}_{i,\mathrm{FE}}-\eta_i)^2|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
&=\mathbb{E}\left\{\left(\frac{1}{T_i}\left(\sum_{t=1}^{T_i}\left(\eta_i+\boldsymbol{\beta}'\boldsymbol{x}_{i,t}+\varepsilon_{i,t}-\hat{\boldsymbol{\beta}}_{\mathrm{FE}}'\boldsymbol{x}_{i,t}\right)\right)-\eta_i\right)^2\right.\\
&\quad\left.\big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
&=\mathbb{E}\left\{\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\left(\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\mathrm{FE}}\right)'\boldsymbol{x}_{i,t}+\varepsilon_{i,t}\right)\right)^2\big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}. \qquad (48)
\end{aligned}
$$

It follows from Eq. (48) that Eq. (47) can be re-written as

$$
\begin{aligned}
&\mathbb{E}\left\{\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\left(\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\mathrm{FE}}\right)'\boldsymbol{x}_{i,t}+\varepsilon_{i,t}\right)\right)^2\big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
&+\mathbb{E}\left\{(\boldsymbol{x}_i^{\mathrm{test}})'\left(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}-\boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i^{\mathrm{test}}\big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}\\
&+2\mathbb{E}\left\{\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\left(\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\mathrm{FE}}\right)'\boldsymbol{x}_{i,t}+\varepsilon_{i,t}\right)\right)\right.\\
&\quad\left.\left(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i^{\mathrm{test}}\big|\{\boldsymbol{X}_n\}_{n=1}^N\right\}. \qquad (49)
\end{aligned}
$$

Using the expression (6) of $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$, and Eq. (16), one can simplify the term $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}-\boldsymbol{\beta}$ above as follows:

$$
\begin{aligned}
&\hat{\boldsymbol{\beta}}_{\mathrm{FE}}-\boldsymbol{\beta}\\
&=\left(\sum_{n=1}^N\boldsymbol{X}_n'\boldsymbol{Q}_n\boldsymbol{X}_n\right)^{-1}\left(\sum_{n=1}^N\boldsymbol{X}_n'\boldsymbol{Q}_n\tilde{\boldsymbol{y}}_n\right)-\boldsymbol{\beta}
\end{aligned}
$$

$$= \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n \left(\boldsymbol{\eta}_n + \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}_n\right)\right) - \boldsymbol{\beta}$$
$$= \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right). \tag{50}$$

Then, Eq. (49) becomes

$$\begin{aligned}
&\mathbb{E}\Bigg\{\Bigg(\frac{\mathbf{1}_{T_i}'}{T_i}\Bigg(-\boldsymbol{X}_i \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \\
&\quad \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) + \boldsymbol{\varepsilon}_i\Bigg)\Bigg)^2 \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Bigg\} \\
&+ \mathbb{E}\Bigg\{ \left(\boldsymbol{x}_i^{\text{test}}\right)' \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) \\
&\quad \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right)\right)' \boldsymbol{x}_i^{\text{test}} \\
&\quad \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Bigg\} \\
&+ 2\mathbb{E}\Bigg\{\Bigg(\frac{\mathbf{1}_{T_i}'}{T_i}\Bigg(-\boldsymbol{X}_i \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \\
&\quad \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) + \boldsymbol{\varepsilon}_i\Bigg)\Bigg) \\
&\quad \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right)\right)' \boldsymbol{x}_i^{\text{test}} \\
&\quad \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Bigg\}.
\end{aligned} \tag{51}$$

Expanding the square in the first term in the expression above, and splitting its last term in two parts, one obtains the following expression for Eq. (51):

$$\begin{aligned}
&\mathbb{E}\Bigg\{\frac{\mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i^2} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) \\
&\quad \left(\sum_{n=1}^{N} \boldsymbol{\varepsilon}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \\
&\quad \boldsymbol{X}_i' \mathbf{1}_{T_i} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Bigg\} \\
&+ \mathbb{E}\left\{\frac{\mathbf{1}_{T_i}' \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' \mathbf{1}_{T_i}}{T_i^2} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\right\} \\
&- 2\mathbb{E}\Bigg\{\frac{\mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i^2} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \\
&\quad \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) \boldsymbol{\varepsilon}_i' \mathbf{1}_{T_i} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Bigg\} \\
&+ \mathbb{E}\Bigg\{ \left(\boldsymbol{x}_i^{\text{test}}\right)' \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \\
&\quad \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) \left(\sum_{n=1}^{N} \boldsymbol{\varepsilon}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \\
&\quad \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \boldsymbol{x}_i^{\text{test}} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Bigg\} \\
&- 2\mathbb{E}\Bigg\{\frac{\mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i^2} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \\
&\quad \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) \left(\sum_{n=1}^{N} \boldsymbol{\varepsilon}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \\
&\quad \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \boldsymbol{x}_i^{\text{test}} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Bigg\} \\
&+ 2\mathbb{E}\Bigg\{\frac{\mathbf{1}_{T_i}' \boldsymbol{\varepsilon}_i}{T_i} \left(\sum_{n=1}^{N} \boldsymbol{\varepsilon}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \\
&\quad \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \boldsymbol{x}_i^{\text{test}} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Bigg\}.
\end{aligned} \tag{52}$$

In order to simplify the various terms contained in Eq. (52), one observes that, due to Eqs. (12), (13), and (15), one gets

$$\begin{aligned}
&\mathbb{E}\Bigg\{\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) \\
&\quad \left(\sum_{n=1}^{N} \boldsymbol{\varepsilon}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \\
&\quad \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Bigg\} \\
&= \mathbb{E}\Bigg\{\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \\
&\quad \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \\
&\quad \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Bigg\} \\
&= \sigma^2 \mathbb{E}\Bigg\{\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \\
&\quad \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)
\end{aligned}$$

$$\left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Big\}$$
$$= \sigma^2 \mathbb{E}\Big\{ \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right) \left(\left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1}\right)' \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Big\}$$
$$= \sigma^2 \mathbb{E}\Big\{ \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Big\}, \tag{53}$$

and

$$\mathbb{E}\Big\{ \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n\right) \boldsymbol{\varepsilon}_i' \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Big\}$$
$$= \mathbb{E}\Big\{ \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{X}_i' \boldsymbol{Q}_i' \boldsymbol{Q}_i \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Big\}$$
$$= \sigma^2 \mathbb{E}\Big\{ \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{X}_i' \boldsymbol{Q}_i' \boldsymbol{Q}_i \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Big\}. \tag{54}$$

Then, by an application of the two equations just derived above, one obtains the following equivalent expression for Eq. (52):

$$\frac{\sigma^2 \mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i^2} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{X}_i' \mathbf{1}_{T_i}$$
$$+ \frac{\sigma^2}{T_i}$$
$$- 2\frac{\sigma^2 \mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i^2} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{X}_i' \boldsymbol{Q}_i' \boldsymbol{Q}_i \mathbf{1}_{T_i}$$
$$+ \mathbb{E}\Big\{ \sigma^2 \left(\boldsymbol{x}_i^{\text{test}}\right)' \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{x}_i^{\text{test}} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Big\}$$
$$- 2\mathbb{E}\left\{ \frac{\sigma^2 \mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{x}_i^{\text{test}} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\right\}$$
$$+ 2\mathbb{E}\left\{ \frac{\sigma^2 \mathbf{1}_{T_i}' \boldsymbol{Q}_i' \boldsymbol{Q}_i \boldsymbol{X}_i}{T_i} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{x}_i^{\text{test}} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\right\}, \tag{55}$$

where, in some cases, the conditional expectations of deterministic matrices (and of random matrices, like $\boldsymbol{X}_i$, that become known once the set of conditioning matrices $\{\boldsymbol{X}_n\}_{n=1}^{N}$ has been fixed) have been replaced by the matrices themselves. Finally, exploiting Eq. (17), one can get rid of the third and sixth terms in Eq. (55), which then becomes

$$\frac{\sigma^2 \mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i^2} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{X}_i' \mathbf{1}_{T_i}$$
$$+ \frac{\sigma^2}{T_i}$$
$$+ \mathbb{E}\Big\{ \sigma^2 \left(\boldsymbol{x}_i^{\text{test}}\right)' \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{x}_i^{\text{test}} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\Big\}$$
$$- 2\mathbb{E}\left\{ \frac{\sigma^2 \mathbf{1}_{T_i}' \boldsymbol{X}_i}{T_i} \left(\sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{Q}_n' \boldsymbol{Q}_n \boldsymbol{X}_n\right)^{-1} \boldsymbol{x}_i^{\text{test}} \Big| \{\boldsymbol{X}_n\}_{n=1}^{N}\right\}, \tag{56}$$

which is Eq. (18).

Appendix 2: Proof of Equation (26)

The first inequality

$$\frac{\sigma^2}{T}\left(\frac{1}{q_i} + \mathbb{E}\left\{\left\| \boldsymbol{A}_N^{-\frac{1}{2}} \left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\} - \boldsymbol{x}_i^{\text{test}}\right)\right\|_2^2\right\}\right)$$
$$\leq \frac{\sigma^2}{T}\left(\frac{1}{q_i} + \lambda_{\max}(\boldsymbol{A}_N^{-1}) \mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\} - \boldsymbol{x}_i^{\text{test}}\right)\right\|_2^2\right\}\right) \tag{57}$$

in Eq. (26) is obtained by exploiting the definition of induced l_2-matrix norm, i.e.,

$$\|\boldsymbol{A}_N^{-\frac{1}{2}}\|_2 = \sup_{\boldsymbol{x} \in \mathbb{R}^p, \|\boldsymbol{x}\|_2 \neq 0} \frac{\|\boldsymbol{A}_N^{-\frac{1}{2}} \boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}, \tag{58}$$

and the fact that, being $\boldsymbol{A}_N^{-\frac{1}{2}}$ symmetric, one has

$$\|\boldsymbol{A}_N^{-\frac{1}{2}}\|_2^2 = \lambda_{\max}^2(\boldsymbol{A}_N^{-\frac{1}{2}}) = \lambda_{\max}(\boldsymbol{A}_N^{-1}). \tag{59}$$

Then, the equality

$$\frac{\sigma^2}{T}\left(\frac{1}{q_i} + \lambda_{\max}(\boldsymbol{A}_N^{-1}) \mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\} - \boldsymbol{x}_i^{\text{test}}\right)\right\|_2^2\right\}\right)$$
$$= \frac{\sigma^2}{T}\left(\frac{1}{q_i} + \frac{1}{\lambda_{\min}(\boldsymbol{A}_N)} \mathbb{E}\left\{\left\|\left(\mathbb{E}\left\{\boldsymbol{x}_{i,1}\right\} - \boldsymbol{x}_i^{\text{test}}\right)\right\|_2^2\right\}\right). \tag{60}$$

follows from the relationship $\lambda_{\min}(\boldsymbol{A}_N) = \frac{1}{\lambda_{\max}(\boldsymbol{A}_N^{-1})}$.

Finally, the last inequality in Eq. (26) is obtained by exploiting Weyl's inequalities (Bhatia 1997, Theorem III.2.1) for the eigenvalues of the sum of symmetric matrices, as detailed in the following remark.

Remark 5.1 Given any pair of symmetric matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{s\times s}$, let their eigenvalues and those of $\boldsymbol{C} := \boldsymbol{A} + \boldsymbol{B}$ be ordered nondecreasingly (with possible repetitions in case of multiplicity larger than 1) as

$$\begin{aligned}
&\lambda_1(\boldsymbol{A}) \le \lambda_2(\boldsymbol{A}) \le \ldots \le \lambda_k(\boldsymbol{A}) \le \ldots \le \lambda_s(\boldsymbol{A}),\\
&\lambda_1(\boldsymbol{B}) \le \lambda_2(\boldsymbol{B}) \le \ldots \le \lambda_k(\boldsymbol{B}) \le \ldots \le \lambda_s(\boldsymbol{B}),\\
&\lambda_1(\boldsymbol{C}) \le \lambda_2(\boldsymbol{C}) \le \ldots \le \lambda_k(\boldsymbol{C}) \le \ldots \le \lambda_s(\boldsymbol{C}). \qquad (61)
\end{aligned}$$

Then, Weyl's inequalities, in their simplest form, state that, for every $k = 1, \ldots, s$, one has

$$\lambda_k(\boldsymbol{A}) + \lambda_1(\boldsymbol{B}) \le \lambda_k(\boldsymbol{C}) \le \lambda_k(\boldsymbol{A}) + \lambda_s(\boldsymbol{B}). \qquad (62)$$

Hence, $\lambda_{\min}(\boldsymbol{C}) \ge \lambda_{\min}(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{B})$. Similarly, for any $\mu_1, \mu_2 \ge 0$, when $\boldsymbol{A}$ and $\boldsymbol{B}$ are also positive semi-definite (as in the case of the matrices $\boldsymbol{A}$ defined in Eq. (19)), one gets

$$\lambda_{\min}(\mu_1\boldsymbol{A} + \mu_2\boldsymbol{B}) \ge \mu_1\lambda_{\min}(\boldsymbol{A}) + \mu_2\lambda_{\min}(\boldsymbol{B}). \qquad (63)$$

Finally, Eq. (63) extends directly to the case of a weighted summation (with non-negative weights) of symmetric and positive semi-definite matrices, proving the last inequality in Eq. (26).

References

Andreß H-J, Golsch K, Schmidt AW (2013) Applied panel data analysis for economic and social surveys. Springer, Berlin

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

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. Proc Nat Acad Sci 113:7353–7360

Bhatia R (1997) Matrix analysis. Springer, Berlin

Bargagli Stoffi FJ, Gnecco G (2019) Causal tree with instrumental variable: an extension of the causal tree framework to irregular assignment mechanisms. Int J Data Sci Anal. https://doi.org/10.1007/s41060-019-00187-z

Bargagli Stoffi FJ, 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

Bell A, Jones K (2014) Explaining fixed effects: random effects modeling of time-series cross-sectional and panel data. Polit Sci Res Methods 3:133–153

Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, Cambridge

Cameron AC, Trivedi PK (2005) Microeconometrics: methods and applications. Cambridge University Press, Cambridge

Chen C-H, Lee LH (2010) Stochastic simulation optimization: an optimal computing budget allocation. World Scientific, Singapore

Crane-Droesch A (2017) Semiparametric panel data using neural networks. In: Proceedings of the 2017 annual meeting of the agricultural and applied economics association, Chicago, USA

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

Frees EW (2004) Longitudinal and panel data: analysis and applications in the social sciences. Cambridge University Press, Cambridge

Friston KJ, Holmes AP, Worsley KJ (1999) How many subjects constitute a study? Neuroimage 10:1–5

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. Optim Lett. https://doi.org/10.1007/s11590-019-01486-x

Gnecco G, Nutarelli F (2019) On the optimal trade-off between sample size and precision of supervision. In: Program of the international conference on optimization and decision science (ODS 2019), Genoa, Italy, p 27

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, vol 11943 of Lecture notes in computer science, pp 531–542

Greene WH (2003) Econometric analysis. Pearson Education, London

Groves RM, Fowler FJ Jr, Couper MP, Lepkowski JM, Singer E, Tourangeau R (2004) Survey methodology. Wiley, Hoboken

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

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

Martello S, Toth P (1990) Continuous knapsack problems: algorithms and computer implementations. Wiley, Hoboken

Nguyen HT, Kosheleva O, Kreinovich V, Ferson S (2009) Trade-off between sample size and accuracy: case of measurements under interval uncertainty. Int J Approx Reason 50:1164–1176

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

Révész P (1968) The laws of large numbers. Academic Press, Cambridge

Ruud PA (2000) An introduction to classical econometric theory. Oxford University Press, Oxford

Sherron T, Allen J, Shumacker, RE (2000) A fixed effects panel data model: mathematics achievement in the U.S. In: Proceedings of the annual meeting of the American educational research association, New Orleans, USA

Vapnik VN (1998) Statistical learning theory. Wiley, Hoboken

Varian HR (2014) Big data: new tricks for econometrics. J Econ Perspect 28:3–38

Wooldridge JM (2002) Econometric analysis of cross section and panel data. MIT Press, Cambridge

Yu Q, Qin Y, Liu P, Ren G (2018) A panel data model-based multi-factor predictive model of highway electromechanical equipment faults. IEEE Trans Intell Transp Syst 9:3039–3045

Zeifman M (2015) Measurement and verification of energy saving in smart building technologies. In: Proceedings of the IEEE symposium on signal processing applications in smart buildings (GlobalSIP 2015), Orlando, USA, pp 343–347