CORESETS FOR REGRESSIONS WITH PANEL DATA

By

Lingxiao Huang, K. Sudhir, and Nisheeth Vishnoi

November 2021

COWLES FOUNDATION DISCUSSION PAPER NO. 2309

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Coresets for Regressions with Panel Data[*]

Lingxiao Huang
Huawei TCS Lab

K. Sudhir
Yale University

Nisheeth K. Vishnoi
Yale University

October 29, 2021

Abstract

This paper introduces the problem of coresets for regression problems to panel data settings. We first define coresets for several variants of regression problems with panel data and then present efficient algorithms to construct coresets of size that depend polynomially on $1/\varepsilon$ (where ε is the error parameter) and the number of regression parameters – independent of the number of individuals in the panel data or the time units each individual is observed for. Our approach is based on the Feldman-Langberg framework in which a key step is to upper bound the "total sensitivity" that is roughly the sum of maximum influences of all individual-time pairs taken over all possible choices of regression parameters. Empirically, we assess our approach with synthetic and real-world datasets; the coreset sizes constructed using our approach are much smaller than the full dataset and coresets indeed accelerate the running time of computing the regression objective.

[*] A version of this paper in NeurIPS format has been published in the 2020 NeurIPS Proceedings. When citing, please cite the NeurIPS paper as follows: "Huang, L., Sudhir, K., & Vishnoi, N. K., "Coresets for Regressions with Panel Data,", Advances in Neural Information Processing Systems, 2020.

Contents

1 Introduction

Panel data, represented as $X \in \mathbb{R}^{N\times T\times d}$ and $Y \in \mathbb{R}^{N\times T}$ where N is the number of entities/individuals, T is the number of time periods and d is the number of features is widely used in statistics and applied machine learning. Such data track features of a cross-section of entities (e.g., customers) longitudinally over time. Such data are widely preferred in supervised machine learning for more accurate prediction and unbiased inference of relationships between variables relative to cross-sectional data (where each entity is observed only once) [28, 6].

The most common method for inferring relationships between variables using observational data involves solving regression problems on panel data. The main difference between regression on panel data when compared to cross-sectional data is that there may exist correlations within observations associated with entities over time periods. Consequently, the regression problem for panel data is the following optimization problem over regression variables $\beta \in \mathbb{R}^d$ and the covariance matrix Ω that is induced by the abovementioned correlations: $\min_{\beta\in\mathbb{R}^d,\Omega\in\mathbb{R}^{T\times T}} \sum_{i\in[N]}(y_i - X_i\beta)^\top \Omega^{-1}(y_i - X_i\beta)$. Here $X_i \in \mathbb{R}^{T\times d}$ denotes the observation matrix of entity i whose t-th row is x_{it} and Ω is constrained to have largest eigenvalue at most 1 where $\Omega_{tt'}$ represents the correlation between time periods t and t'. This regression model is motivated by the random effects model (Eq. (1) and Appendix A), common in the panel data literature [27, 24, 23]. A common way to define the correlation between observations is an autocorrelation structure $\mathsf{AR}(q)$ [25, 35] whose covariance matrix Ω is induced by a vector $\rho \in \mathbb{R}^q$ (integer $q \geq 1$). This type of correlation results in the generalized least-squares estimator (GLSE), where the parameter space is $\mathcal{P} = R^{d+q}$.

As the ability to track entities on various features in real-time has grown, panel datasets have grown massively in size. However, the size of these datasets limits the ability to apply standard learning algorithms due to space and time constraints. Further, organizations owning data may want to share only a subset of data with others seeking to gain insights to mitigate privacy or intellectual property related risks. Hence, a question arises: *can we construct a smaller subset of the panel data on which we can solve the regression problems with performance guarantees that are close enough to those obtained when working with the complete dataset?*

One approach to this problem is to appeal to the theory of "coresets." Coresets, proposed in [1], are weighted subsets of the data that allow for fast approximate inference for a large dataset by solving the problem on the smaller coreset. Coresets have been developed for a variety of unsupervised and supervised learning problems; for a survey, see [43]. But, thus, far coresets have been developed only for ℓ_2-regression cross-sectional data [18, 36, 8, 15, 33]; no coresets have been developed for regressions on panel data – an important limitation, given their widespread use and advantages.

Roughly, a coreset for cross-sectional data is a weighted subset of observations associated with entities that approximates the regression objective for every possible choice of regression parameters. An idea, thus, is to construct a coreset for each time period (cross-section) and output their union as a coreset for panel data. However, this union contains at least T observations which is undesirable since T can be large. Further, due to the covariance matrix Ω, it is not obvious how to use this union to approximately compute regression objectives. With panel data, one needs to consider both how to sample entities, and within each entity how to sample observations across time. Moreover, we also need to define how to compute regression objectives on such a coreset consisting of entity-time pairs.

Our contributions. We initiate the study of coresets for versions of ℓ_2-regression with panel data, including the ordinary least-squares estimator (OLSE; Definition 2.2), the generalized least-squares estimator (GLSE; Definition 2.3), and a clustering extension of GLSE (GLSE$_k$; Definition 2.4) in which all entities are partitioned into k clusters and each cluster shares the same regression parameters.

Overall, we formulate the definitions of coresets and propose efficient construction of ε-coresets of sizes independent of N and T. Our key contributions are:

1. We give a novel formulation of coresets for GLSE (Definition 3.3) and GLSE$_k$ (Definition 3.4). We represent the regression objective of GLSE as the sum of NT sub-functions w.r.t. entity-time pairs, which enables us to define coresets similar to the case of cross-sectional data. For GLSE$_k$, the regression objective cannot be similarly decomposed due to the min operations in Definition 2.4. To deal with this issue, we define the regression objective on a coreset S by including min operations.

2. Our coreset for OLSE is of size $O(\min\{\varepsilon^{-2}d, d^2\})$ (Theorems B.1 and B.2), based on a reduction to

coreset for ℓ_2-regression with cross-sectional data.

3. Our coreset for GLSE consists of at most $\tilde{O}(\varepsilon^{-2} \max\{q^4 d^2, q^3 d^3\})$ points (Theorem 4.1), independent of N and T as desired.

4. Our coreset for GLSE$_k$ is of size poly$(M, k, q, d, 1/\varepsilon)$ (Theorem 5.2) where M upper bounds the gap between the maximum individual regression objective of OLSE and the minimum one (Definition 5.1). We provide a matching lower bound $\Omega(N)$ (Theorem 5.4) for $k, q, d \leq 2$, indicating that the coreset size should contain additional factors than $k, q, d, 1/\varepsilon$, justifying the M-bounded assumption.

Our coresets for GLSE/GLSE$_k$ leverage the Feldman-Langberg (FL) framework [21] (Algorithms 1 and 2). The ρ variables make the objective function of GLSE non-convex in contrast to the cross-sectional data setting where objective functions are convex. Thus, bounding the "sensitivity" (Lemma 4.4) of each entity-time pair for GLSE, which is a key step in coreset construction using the FL framework, becomes significantly difficult. We handle this by upper-bounding the maximum effect of ρ, based on the observation that the gap between the regression objectives of GLSE and OLSE with respect to the same $\beta \in \mathbb{R}^d$ is always constant, which enables us to reduce the problem to the cross-sectional setting. For GLSE$_k$, a key difficulty is that the clustering centers are *subspaces* induced by regression vectors, instead of *points* as in Gaussian mixture models or k-means. Hence, it is unclear how GLSE$_k$ can be reduced to projective clustering used in Gaussian mixture models; see [20]. To bypass this, we consider observation vectors of an individual as one entity and design a two-staged framework in which the first stage selects a subset of individuals that captures the min operations in the objective function and the second stage applies our coreset construction for GLSE on each selected individuals. As in the case of GLSE, bounding the "sensitivity" (Lemma 5.8) of each entity for GLSE$_k$ is a key step at the first stage. Towards this, we relate the total sensitivity of entities to a certain "flexibility" (Lemma 5.7) of each individual regression objective which is, in turn, shown to be controlled by the M-bounded assumption (Definition 5.1).

We implement our GLSE coreset construction algorithm and test it on synthetic and real-world datasets while varying ε. Our coresets perform well relative to uniform samples on multiple datasets with different generative distributions. Importanty, the relative performance is robust and better on datasets with outliers. The maximum empirical error of our coresets is always below the guaranteed ε unlike with uniform samples. Further, for comparable levels of empircal error, our coresets perform much better than uniform sampling in terms of sample size and coreset construction speed.

1.1 Related work

With panel data, depending on different generative models, there exist several ways to define ℓ_2-regression [27, 24, 23], including the pooled model, the fixed effects model, the random effects model, and the random parameters model. In this paper, we consider the random effects model (Equation (1)) since the number of parameters is independent of N and T (see Section A for more discussion).

For cross-sectional data, there is more than a decade of extensive work on coresets for regression; e.g., ℓ_2-regression [18, 36, 8, 15, 33], ℓ_1-regression [11, 47, 12], generalized linear models [31, 40] and logistic regression [44, 31, 42, 49]. The most relevant for our paper is ℓ_2-regression (least-squares regression), which admits an ε-coreset of size $O(d/\varepsilon^2)$ [8] and an accurate coreset of size $O(d^2)$ [33].

With cross-sectional data, coresets have been developed for a large family of problems in machine learning and statistics, including clustering [21, 22, 30], mixture model [37], low rank approximation [16], kernel regression [53] and logistic regression [42]. We refer interested readers to recent surveys [41, 19]. It is interesting to investigate whether these results can be generalized to panel data.

There exist other variants of regression sketches beyond coreset, including weighted low rank approximation [13], row sampling [17], and subspace embedding [47, 39]. These methods mainly focus on the cross-sectional setting. It is interesting to investigate whether they can be adapted to the panel data setting that with an additional covariance matrix.

2 ℓ_2-regression with panel data

We consider the following generative model of ℓ_2-regression: for $(i,t) \in [N] \times [T]$,

$$y_{it} = x_{it}^\top \beta_i + e_{it}, \tag{1}$$

where $\beta_i \in \mathbb{R}^d$ and $e_{it} \in \mathbb{R}$ is the error term drawn from a normal distribution. Sometimes, we may include an additional entity or individual specified effect $\alpha_i \in \mathbb{R}$ so that the outcome can be represented by $y_{it} = x_{it}^\top \beta_i + \alpha_i + e_{it}$. This is equivalent to Equation (1) by appending an additional constant feature to each observation x_{it}.

Remark 2.1 *Sometimes, we may not observe individuals for all time periods, i.e., some observation vectors x_{it} and their corresponding outcomes y_{it} are missing. One way to handle this is to regard those missing individual-time pairs as $(x_{it}, y_{it}) = (0,0)$. Then, for any vector $\beta \in \mathbb{R}^d$, we have $y_{it} - x_{it}^\top \beta = 0$ for each missing individual-time pairs.*

As in the case of cross-sectional data, we assume there is no correlation between individuals. Using this assumption, the ℓ_2-regression function can be represented as follows: for any regression parameters $\zeta \in \mathcal{P}$ ($\mathcal{P}$ is the parameter space), $\psi(\zeta) = \sum_{i \in [N]} \psi_i(\zeta)$, where ψ_i is the individual regression function. Depending on whether there is correlation within individuals and whether β_i is unique, there are several variants of ψ_i. The simplest setting is when all β_is are the same, say $\beta_i = \beta$, and there is no correlation within individuals. This setting results in the ordinary least-squares estimator (OLSE); summarized in the following definition.

Definition 2.2 (Ordinary least-squares estimator (OLSE)) *For an ordinary least-squares estimator (OLSE), the parameter space is $\mathbb{R}^d$ and for any $\beta \in \mathbb{R}^d$ the individual objective function is*

$$\psi_i^{(O)}(\beta) := \textstyle\sum_{t \in [T]} \psi_{it}^{(O)}(\beta) = \sum_{t \in [T]} (y_{it} - x_{it}^\top \beta)^2.$$

Consider the case when β_i are the same but there may be correlations between time periods within individuals. A common way to define the correlation is called autocorrelation $\mathsf{AR}(q)$ [25, 35], in which there exists $\rho \in B^q$, where $q \geq 1$ is an integer and $B^q = \{x \in \mathbb{R}^q : \|x\|_2 < 1\}$, such that

$$e_{it} = \textstyle\sum_{a=1}^{\min\{t-1,q\}} \rho_a e_{i,t-a} + N(0,1). \tag{2}$$

This autocorrelation results in the generalized least-squares estimator (GLSE).

Definition 2.3 (Generalized least-squares estimator (GLSE)) *For a generalized least-squares estimator (GLSE) with $\mathsf{AR}(q)$ (integer $q \geq 1$), the parameter space is $\mathbb{R}^d \times B^q$ and for any $\zeta = (\beta, \rho) \in \mathbb{R}^d \times B^q$ the individual objective function is $\psi_i^{(G,q)}(\zeta) := \sum_{t \in [T]} \psi_{it}^{(G,q)}(\zeta)$ equal to*

$$(1 - \|\rho\|_2^2)(y_{i1} - x_{i1}^\top \beta)^2 + \textstyle\sum_{t=2}^T \left((y_{it} - x_{it}^\top \beta) - \sum_{j=1}^{\min\{t-1,q\}} \rho_j (y_{i,t-j} - x_{i,t-j}^\top \beta) \right)^2.$$

The main difference from OLSE is that a sub-function $\psi_{it}^{(G,q)}$ is not only determined by a single observation (x_{it}, y_{it}); instead, the objective of $\psi_{it}^{(G,q)}$ may be decided by up to $q+1$ contiguous observations $(x_{i,\max\{1,t-q\}}, y_{i,\max\{1,t-q\}}), \ldots, (x_{it}, y_{it})$.

Motivated by k-means clustering [48], we also consider a generalized setting of GLSE, called GLSE$_k$ ($k \geq 1$ is an integer), in which all individuals are partitioned into k clusters and each cluster corresponds to the same regression parameters with respect to some GLSE.

Definition 2.4 (GLSE$_k$: an extention of GLSE) *Let $k, q \geq 1$ be integers. For a GLSE$_k$, the parameter space is $\left(\mathbb{R}^d \times B^q\right)^k$ and for any $\zeta = (\beta^{(1)}, \ldots, \beta^{(k)}, \rho^{(1)}, \ldots, \rho^{(k)}) \in \left(\mathbb{R}^d \times B^q\right)^k$ the individual objective function is $\psi_i^{(G,q,k)}(\zeta) := \min_{l \in [k]} \psi_i^{(G,q)}(\beta^{(l)}, \rho^{(l)})$.*

GLSE$_k$ is a basic problem with applications in many real-world fields; as accounting for *unobserved heterogeneity* in panel regressions is critical for unbiased estimates [3, 26]. Note that each individual selects regression parameters $(\beta^{(l)}, \rho^{(l)})$ ($l \in [k]$) that minimizes its individual regression objective for GLSE. Note that GLSE$_1$ is exactly GLSE. Also note that GLSE$_k$ can be regarded as a generalized version of clustered linear regression [4], in which there is no correlation within individuals.

3 Our coreset definitions for panel data

In this section, we show how to define coresets for regression on panel data, including OLSE and GLSE. Due to the additional autocorrelation parameters, it is not straightforward to define coresets for GLSE as in the cross-sectional setting. One way is to consider all observations of an individual as an indivisible group and select a collection of individuals as a coreset. However, this construction results in a coreset of size depending on T, which violates the expectation that the coreset size should be independent of N and T. To avoid a large coreset size, we introduce a generalized definition: coresets of a query space, which captures the coreset definition for OLSE and GLSE.

Definition 3.1 (Query space [21, 9]) *Let $\mathcal{X}$ be a index set together with a weight function $u : \mathcal{X} \to \mathbb{R}_{\geq 0}$. Let $\mathcal{P}$ be a set called queries, and $\psi_x : \mathcal{P} \to \mathbb{R}_{\geq 0}$ be a given loss function w.r.t. $x \in \mathcal{X}$. The total cost of $\mathcal{X}$ with respect to a query $\zeta \in \mathcal{P}$ is $\psi(\zeta) := \sum_{x\in\mathcal{X}} u(x) \cdot \psi_x(\zeta)$. The tuple $(\mathcal{X}, u, \mathcal{P}, \psi)$ is called a query space. Specifically, if $u(x) = 1$ for all $x \in \mathcal{X}$, we use $(\mathcal{X}, \mathcal{P}, \psi)$ for simplicity.*

Intuitively, ψ represents a linear combination of weighted functions indexed by $\mathcal{X}$, and $\mathcal{P}$ represents the ground set of ψ. Due to the separability of ψ, we have the following coreset definition.

Definition 3.2 (Coresets of a query space [21, 9]) *Let $(\mathcal{X}, u, \mathcal{P}, \psi)$ be a query space and $\varepsilon \in (0, 1)$ be an error parameter. An ε-coreset of $(\mathcal{X}, u, \mathcal{P}, \psi)$ is a weighted set $S \subseteq \mathcal{X}$ together with a weight function $w : S \to \mathbb{R}_{\geq 0}$ such that for any $\zeta \in \mathcal{P}$, $\psi_S(\zeta) := \sum_{x\in S} w(x) \cdot \psi_x(\zeta) \in (1 \pm \varepsilon) \cdot \psi(\zeta)$.*

Here, ψ_S is a computation function over the coreset that is used to estimate the total cost of $\mathcal{X}$. By Definitions 2.2 and 2.3, the regression objectives of OLSE and GLSE can be decomposed into NT sub-functions. Thus, we can apply the above definition to define coresets for OLSE and GLSE. Note that OLSE is a special case of GLSE for $q = 0$. Thus, we only need to provide the coreset definition for GLSE. We let $u = 1$ and $\mathcal{P} = \mathbb{R}^d \times B^q$. The index set of GLSE has the following form:

$$Z^{(G,q)} = \left\{ z_{it} = \left(x_{i,\max\{1,t-q\}}, y_{i,\max\{1,t-q\}}, \ldots x_{it}, y_{it} \right) : (i,t) \in [N] \times [T] \right\},$$

where each element z_{it} consists of at most $q+1$ observations. Also, for every $z_{it} \in Z^{(G,q)}$ and $\zeta = (\beta, \rho) \in \mathcal{P}$, the cost function ψ_{it} is: if $t = 1$, $\psi_{it}^{(G,q)}(\zeta) = (1 - \|\rho\|_2^2) \cdot (y_{i1} - x_{i1}^\top \beta)^2$; and if $t \neq 1$, $\psi_{it}^{(G,q)}(\zeta) = \left((y_{it} - x_{it}^\top \beta) - \sum_{j=1}^{\min\{t-1,q\}} \rho_j (y_{i,t-j} - x_{i,t-j}^\top \beta) \right)^2$. Thus, $(Z^{(G,q)}, \mathcal{P}, \psi^{(G,q)})$ is a query space of GLSE.[1] Then by Definition 3.2, we have the following coreset definition for GLSE.

Definition 3.3 (Coresets for GLSE) *Given a panel dataset $X \in \mathbb{R}^{N\times T\times d}$ and $Y \in \mathbb{R}^{N\times T}$, a constant $\varepsilon \in (0, 1)$, integer $q \geq 1$, and parameter space $\mathcal{P}$, an ε-coreset for GLSE is a weighted set $S \subseteq [N] \times [T]$ together with a weight function $w : S \to \mathbb{R}_{\geq 0}$ such that for any $\zeta = (\beta, \rho) \in \mathcal{P}$,*

$$\psi_S^{(G,q)}(\zeta) := \sum_{(i,t)\in S} w(i,t) \cdot \psi_{it}^{(G,q)}(\zeta) \in (1 \pm \varepsilon) \cdot \psi^{(G,q)}(\zeta).$$

The weighted set S is exactly an ε-coreset of the query space $(Z^{(G,q)}, \mathcal{P}, \psi^{(G,q)})$. Note that the number of points in this coreset S is at most $(q+1) \cdot |S|$. Specifically, for OLSE, the parameter space is $\mathbb{R}^d$ since $q = 0$, and the corresponding index set is $Z^{(O)} = \{z_{it} = (x_{it}, y_{it}) : (i,t) \in [N] \times [T]\}$. Consequently, the query space of OLSE is $(Z^{(O)}, \mathbb{R}^d, \psi^{(O)})$.

Coresets for GLSE$_k$ Due to the min operation in Definition 2.4, the objective function $\psi^{(G,q,k)}$ can only be decomposed into sub-functions $\psi_i^{(G,q,k)}$ instead of individual-time pairs. Then let $u = 1$, $\mathcal{P}^k = \left(\mathbb{R}^d \times B^q\right)^k$, and $Z^{(G,q,k)} = \{z_i = (x_{i1}, y_{i1}, \ldots, x_{iT}, y_{iT}) : i \in [N]\}$. We can regard $(Z^{(G,q,k)}, \mathcal{P}^k, \psi^{(G,q,k)})$ as a query space of GLSE$_k$. By Definition 3.2, an ε-coreset of $(Z^{(G,q,k)}, \mathcal{P}^k, \psi^{(G,q,k)})$ is a subset $I_S \subseteq [N]$ together with a weight function $w' : I_S \to \mathbb{R}_{\geq 0}$ such that for any $\zeta \in \mathcal{P}^k$,

$$\sum_{i\in I_S} w'(i) \cdot \psi_i^{(G,q,k)}(\zeta) \in (1 \pm \varepsilon) \cdot \psi^{(G,q,k)}(\zeta). \tag{3}$$

[1] Here, we slightly abuse the notation by using $\psi_{it}^{(G,q)}(\zeta)$ instead of $\psi_{z_{it}}^{(G,q)}(\zeta)$.

However, each $z_i \in Z^{(G,q,k)}$ consists of T observations, and hence, the number of points in this coreset S is $T \cdot |S|$. To avoid the size dependence of T, we propose a new coreset definition for GLSE$_k$. The intuition is to further select a subset of time periods to estimate $\psi_i^{(G,q,k)}$.

Given $S \subseteq [N] \times [T]$, we denote $I_S := \{i \in [N] : \exists t \in [T], s.t., (i,t) \in S\}$ as the collection of individuals that appear in S. Moreover, for each $i \in I_S$, we denote $J_{S,i} := \{t \in [T] : (i,t) \in S\}$ to be the collection of observations for individual i in S.

Definition 3.4 (Coresets for GLSE$_k$) *Given a panel dataset $X \in \mathbb{R}^{N\times T\times d}$ and $Y \in \mathbb{R}^{N\times T}$, constant $\varepsilon \in (0,1)$, integer $k, q \geq 1$, and parameter space $\mathcal{P}^k$, an ε-coreset for GLSE$_k$ is a weighted set $S \subseteq [N] \times [T]$ together with a weight function $w : S \to \mathbb{R}_{\geq 0}$ such that for any $\zeta = (\beta^{(1)}, \ldots, \beta^{(k)}, \rho^{(1)}, \ldots, \rho^{(k)}) \in \mathcal{P}^k$,*

$$\psi_S^{(G,q,k)}(\zeta) := \sum_{i\in I_S} \min_{l\in[k]} \sum_{t\in J_{S,i}} w(i,t) \cdot \psi_{it}^{(G,q)}(\beta^{(l)}, \rho^{(l)}) \in (1 \pm \varepsilon) \cdot \psi^{(G,q,k)}(\zeta).$$

The key is to incorporate min operations in the computation function $\psi_S^{(G,q,k)}$ over the coreset. Similar to GLSE, the number of points in such a coreset S is at most $(q+1) \cdot |S|$.

4 Coresets for GLSE

In this section, we show how to construct coresets for GLSE. We let the parameter space be $\mathcal{P}_\lambda = \mathbb{R}^d \times B^q_{1-\lambda}$ for some constant $\lambda \in (0,1)$ where $B^q_{1-\lambda} = \left\{\rho \in \mathbb{R}^q : \|\rho\|_2^2 \leq 1-\lambda\right\}$. The assumption of the parameter space $B^q_{1-\lambda}$ for ρ is based on the fact that $\|\rho\|_2^2 < 1$ $(\lambda \to 0)$ is a stationary condition for $\mathsf{AR}(q)$ [35].

Theorem 4.1 (Coresets for GLSE) *There exists a randomized algorithm that, for a given panel dataset $X \in \mathbb{R}^{N\times T\times d}$ and $Y \in \mathbb{R}^{N\times T}$, constants $\varepsilon, \delta, \lambda \in (0,1)$ and integer $q \geq 1$, with probability at least $1-\delta$, constructs an ε-coreset for GLSE of size*

$$O\left(\varepsilon^{-2}\lambda^{-1}qd\left(\max\left\{q^2 d, qd^2\right\} \cdot \log\frac{d}{\lambda} + \log\frac{1}{\delta}\right)\right)$$

and runs in time $O(NTq + NTd^2)$.

Note that the coreset in the above theorem contains at most $(q+1) \cdot O\left(\varepsilon^{-2}\lambda^{-1}qd\left(\max\left\{q^2d, qd^2\right\} \cdot \log\frac{d}{\lambda} + \log\frac{1}{\delta}\right)\right)$ points (x_{it}, y_{it}), which is independent of both N and T. Also note that if both λ and δ are away from 0, e.g., $\lambda = \delta = 0.1$ the number of points in the coreset can be further simplified: $O\left(\varepsilon^{-2}\max\left\{q^4d^2, q^3d^3\right\} \cdot \log d\right) = \mathrm{poly}(q, d, 1/\varepsilon)$.

4.1 Algorithm for Theorem 4.1

We summarize the algorithm of Theorem 4.1 in Algorithm 1, which takes a panel dataset (X, Y) as input and outputs a coreset S of individual-time pairs. The main idea is to use importance sampling (Lines 6-7) leveraging the Feldman-Langberg (FL) framework [21, 9]. The key new step appears in Line 5, which computes a sensitivity function s for GLSE that defines the sampling distribution. Also note that the construction of s is based on another function $s^{(O)}$ (Line 4), which is actually a sensitivity function for OLSE that has been studied in the literature [8].

4.2 Useful notations and useful facts for Theorem 4.1

Feldman and Langberg [21] show how to construct coresets by importance sampling and the coreset size has been improved by [9].

Theorem 4.2 (FL framework [21, 9]) *Let $\varepsilon, \delta \in (0,1)$. Let* dim *be an upper bound of the pseudo-dimension. Suppose $s : [N] \times [T] \to \mathbb{R}_{\geq 0}$ is a sensitivity function satisfying that for any $(i,t) \in [N] \times [T]$, $s(i,t) \geq \sup_{\zeta \in \mathcal{P}_\lambda} \frac{\psi_{it}^{(G,q)}(\zeta)}{\psi^{(G,q)}(\zeta)}$, and $\mathcal{G} := \sum_{(i,t)\in[N]\times[T]} s(i,t)$. Let $S \subseteq \mathcal{X}$ be constructed by taking*

$$O\left(\varepsilon^{-2}\mathcal{G}(\dim \cdot \log \mathcal{G} + \log(1/\delta))\right)$$

```
Algorithm 1: CGLSE: Coreset construction of GLSE
Require: $X \in \mathbb{R}^{N\times T\times d}$, $Y \in \mathbb{R}^{N\times T}$, constant $\varepsilon, \delta, \lambda \in (0,1)$, integer $q \geq 1$ and parameter space $\mathcal{P}_\lambda$.
Ensure: a subset $S \subseteq [N]\times[T]$ together with a weight function $w : S \to \mathbb{R}_{\geq 0}$.
1: $M \leftarrow O\left(\varepsilon^{-2}\lambda^{-1}qd\left(\max\left\{q^2 d, qd^2\right\}\cdot\log\frac{d}{\lambda}+\log\frac{1}{\delta}\right)\right)$.
2: Let $Z \in \mathbb{R}^{NT\times(d+1)}$ be whose $(iT-T+t)$-th row is $z_{it} = (x_{it}, y_{it}) \in \mathbb{R}^{d+1}$ for $(i,t)\in[N]\times[T]$.
3: Compute $A \subseteq \mathbb{R}^{NT\times d'}$ whose columns form a unit basis of the column space of $Z$.
4: For each $(i,t)\in[N]\times[T]$, $s^{(O)}(i,t) \leftarrow \|A_{iT-T+t}\|_2^2$.
5: For each pair $(i,t)\in[N]\times[T]$, $s(i,t) \leftarrow \min\left\{1, 2\lambda^{-1}\left(s^{(O)}(i,t)+\sum_{j=1}^{\min\{t-1,q\}} s^{(O)}(i,t-j)\right)\right\}$.
6: Pick a random sample $S \subseteq [N]\times[T]$ of $M$ pairs, where each $(i,t)\in S$ is selected with probability
   $\frac{s(i,t)}{\sum_{(i',t')\in[N]\times[T]} s(i',t')}$.
7: For each $(i,t)\in S$, $w(i,t) \leftarrow \frac{\sum_{(i',t')\in[N]\times[T]} s(i',t')}{M\cdot s(i,t)}$.
8: Output $(S, w)$.
```

samples, where each sample $x \in \mathcal{X}$ is selected with probability $\frac{s(x)}{\mathcal{G}}$ and has weight $w(x) := \frac{\mathcal{G}}{|S|\cdot s(x)}$. Then, with probability at least $1-\delta$, S is an ε-coreset for GLSE.

Here, the sensitivity function s measures the maximum influence for each $x_{it} \in X$.

Note that the above is an importance sampling framework that takes samples from a distribution proportional to sensitivities. The sample complexity is controlled by the total sensitivity $\mathcal{G}$ and the pseudo-dimension dim. Hence, to apply the FL framework, we need to upper bound the pseudo-dimension and construct a sensitivity function.

4.3 Proof of Theorem 4.1

Algorithm 1 applies the FL framework (Feldman and Langberg [21]) that constructs coresets by importance sampling and the coreset size has been improved by [9]. The key is to verify the "pseudo-dimension" (Lemma 4.3) and "sensitivities" (Lemma 4.4) separately; summarized as follows.

Upper bounding the pseudo-dimension. We have the following lemma that upper bounds the pseudo-dimension of $(Z^{(G,q)}, \mathcal{P}_\lambda, \psi^{(G,q)})$.

Lemma 4.3 (Pseudo-dimension of GLSE) *The pseudo-dimension of any query space $(Z^{(G,q)}, u, \mathcal{P}_\lambda, \psi^{(G,q)})$ over weight functions $u : [N]\times[T] \to \mathbb{R}_{\geq 0}$ is at most $O\left((q+d)qd\right)$.*

The proof can be found in Section 4.4. The main idea is to apply the prior results [2, 52] which shows that the pseudo-dimension is polynomially dependent on the number of regression parameters ($q+d$ for GLSE) and the number of operations of individual regression objectives ($O(qd)$ for GLSE). Consequently, we obtain the bound $O\left((q+d)qd\right)$ in Lemma 4.3.

Constructing a sensitivity function. Next, we show that the function s constructed in Line 5 of Algorithm 1 is indeed a sensitivity function of GLSE that measures the maximum influence for each $x_{it} \in X$; summarized by the following lemma.

Lemma 4.4 (Total sensitivity of GLSE) *Function $s : [N]\times[T] \to \mathbb{R}_{\geq 0}$ of Algorithm 1 satisfies that for any $(i,t) \in [N]\times[T]$, $s(i,t) \geq \sup_{\zeta\in\mathcal{P}} \frac{\psi_{it}^{(G,q)}(\zeta)}{\psi^{(G,q)}(\zeta)}$ and $\mathcal{G} := \sum_{(i,t)\in[N]\times[T]} s(i,t) = O(\lambda^{-1}qd)$. Moreover, the construction time of function s is $O(NTq + NTd^2)$.*

Intuitively, if the sensitivity $s(i,t)$ is large, e.g., close to 1, $\psi_{it}^{(G,q)}$ must contribute significantly to the objective with respect to some parameter $\zeta \in \mathcal{P}_\lambda$. The sampling ensures that we are likely to include such pair (i,t) in the coreset for estimating $\psi(\zeta)$. Due to the fact that the objective function of GLSE is non-convex which is different from OLSE, bounding the sensitivity of each individual-time pair for GLSE becomes significantly difficult. To handle this difficulty, we develop a reduction of sensitivities from GLSE to OLSE (Line 5 of Algorithm 1), based on the relations between $\psi^{(G,q)}$ and $\psi^{(O)}$, i.e., for any $\zeta = (\beta, \rho) \in \mathcal{P}_\lambda$ we prove that $\psi_i^{(G,q)}(\zeta) \geq \lambda \cdot \psi_i^{(O)}(\beta)$ and $\psi_{it}^{(G,q)}(\zeta) \leq 2 \cdot \left(\psi_{it}^{(O)}(\beta) + \sum_{j=1}^{\min\{t-1,q\}} \psi_{i,t-j}^{(O)}(\beta)\right)$. The first inequality follows from the fact that the smallest eigenvalue of Ω_ρ^{-1} (the inverse covariance matrix induced by ρ) is at least λ. The intuition of the second inequality is from the form of function $\psi_{it}^{(G,q)}$, which relates to $\min\{t, q+1\}$ individual-time pairs, say $(x_{i,\min\{1,t-q\}}, y_{i,\min\{1,t-q\}}), \ldots, (x_{it}, y_{it})$. Combining these two inequalities, we obtain a relation between the sensitivity function s for GLSE and the sensitivity function $s^{(O)}$ for OLSE, based on the following observation: for any $\zeta = (\beta, \rho) \in \mathcal{P}_\lambda$,

$$\begin{aligned}
\frac{\psi_{it}^{(G,q)}(\zeta)}{\psi^{(G,q)}(\zeta)} &\leq \frac{2 \cdot \left(\psi_{it}^{(O)}(\beta) + \sum_{j=1}^{\min\{t-1,q\}} \psi_{i,t-j}^{(O)}(\beta)\right)}{\lambda \cdot \psi^{(O)}(\beta)} \\
&\leq 2\lambda^{-1} \cdot \left(s^{(O)}(i,t) + \sum_{j=1}^{\min\{t-1,q\}} s^{(O)}(i,t-j)\right) \\
&= s(i,t).
\end{aligned}$$

which leads to the construction of s in Line 5 of Algorithm 1. Then it suffices to construct $s^{(O)}$ (Lines 2-4 of Algorithm 1), which reduces to the cross-sectional data setting and has total sensitivity at most $d+1$ (Lemma 4.7). Consequently, we conclude that the total sensitivity $\mathcal{G}$ of GLSE is $O(\lambda^{-1} qd)$ by the definition of s.

Now we are ready to prove Theorem 4.1.

Proof: [Proof of Theorem 4.1] By Lemma 4.4, the total sensitivity $\mathcal{G}$ is $O(\lambda^{-1} qd)$. By Lemma 4.3, we let $\dim = O\left((q+d)qd\right)$. Pluging the values of $\mathcal{G}$ and dim in the FL framework [21, 9], we prove for the coreset size. For the running time, it costs $O(NTq + NTd^2)$ time to compute the sensitivity function s by Lemma 4.4, and $O(NTd)$ time to construct an ε-coreset. This completes the proof. □

4.4 Proof of Lemma 4.3: Upper bounding the pseudo-dimension

Our proof idea is similar to that in [37]. For preparation, we need the following lemma which is proposed to bound the pseudo-dimension of feed-forward neural networks.

Lemma 4.5 (Restatement of Theorem 8.14 of [2]) *Let $(\mathcal{X}, u, \mathcal{P}, f)$ be a given query space where $f_x(\zeta) \in \{0,1\}$ for any $x \in \mathcal{X}$ and $\zeta \in \mathcal{P}$, and $\mathcal{P} \subseteq \mathbb{R}^m$. Suppose that f can be computed by an algorithm that takes as input the pair $(x, \zeta) \in \mathcal{X} \times \mathcal{P}$ and returns $f_x(\zeta)$ after no more than l of the following operations:*

- *the arithmetic operations $+, -, \times$, and $/$ on real numbers.*
- *jumps conditioned on $>, \geq, <, \leq, =$, and $\neq$ comparisons of real numbers, and*
- *output 0,1.*

Then the pseudo-dimension of $(\mathcal{X}, u, \mathcal{P}, f)$ is at most $O(ml)$.

Note that the above lemma requires that the range of functions f_x is $[0,1]$. We have the following lemma which can help extend this range to $\mathbb{R}$.

Lemma 4.6 (Restatement of Lemma 4.1 of [52]) *Let $(\mathcal{X}, u, \mathcal{P}, f)$ be a given query space. Let $g_x : \mathcal{P} \times \mathbb{R} \to \{0,1\}$ be the indicator function satisfying that for any $x \in \mathcal{X}$, $\zeta \in \mathcal{P}$ and $r \in \mathbb{R}$,*

$$g_x(\zeta, r) = I\left[u(x) \cdot f(x, \zeta) \geq r\right].$$

Then the pseudo-dimension of $(\mathcal{X}, u, \mathcal{P}, f)$ is precisely the pseudo-dimension of the query space $(\mathcal{X}, u, \mathcal{P} \times \mathbb{R}, g_f)$.

Now we are ready to prove Lemma 4.3.

Proof: [Proof of Lemma 4.3] Fix a weight function $u : [N] \times [T] \to \mathbb{R}_{\geq 0}$. For every $(i,t) \in [N] \times [T]$, let $g_{it} : \mathcal{P}_\lambda \times \mathbb{R}_{\geq 0} \to \{0,1\}$ be the indicator function satisfying that for any $\zeta \in \mathcal{P}_\lambda$ and $r \in \mathbb{R}_{\geq 0}$,

$$g_{it}(\zeta, r) := I\left[u(i,t) \cdot \psi_{it}^{(G,q)}(\zeta) \geq r\right].$$

We consider the query space $(Z^{(G,q)}, u, \mathcal{P}_\lambda \times \mathbb{R}_{\geq 0}, g)$. By the definition of $\mathcal{P}_\lambda$, the dimension of $\mathcal{P}_\lambda \times \mathbb{R}_{\geq 0}$ is $m = q + 1 + d$. By the definition of $\psi_{it}^{(G,q)}$, g_{it} can be calculated using $l = O(qd)$ operations, including $O(qd)$ arithmetic operations and a jump. Pluging the values of m and l in Lemma 4.5, the pseudo-dimension of $(Z^{(G,q)}, u, \mathcal{P}_\lambda \times \mathbb{R}_{\geq 0}, g)$ is $O\left((q+d)qd\right)$. Then by Lemma 4.6, we complete the proof. □

4.5 Proof of Lemma 4.4: Bounding the total sensitivity

We prove Lemma 4.4 by relating sensitivities between GLSE and OLSE. For preparation, we give the following lemma that upper bounds the total sensitivity of OLSE. Given two integers $a, b \geq 1$, denote $T(a,b)$ to be the computation time of a column basis of a matrix in $\mathbb{R}^{a \times b}$. For instance, a column basis of a matrix in $\mathbb{R}^{a \times b}$ can be obtained by computing its SVD decomposition, which costs $O(\min\left\{a^2 b, ab^2\right\})$ time by [14].

Lemma 4.7 (Total sensitivity of OLSE) *Function $s^{(O)} : [N] \times [T] \to \mathbb{R}_{\geq 0}$ of Algorithm 1 satisfies that for any $(i,t) \in [N] \times [T]$,*

$$s^{(O)}(i,t) \geq \sup_{\beta \in \mathbb{R}^d} \frac{\psi_{it}^{(O)}(\beta)}{\psi^{(O)}(\beta)}, \tag{4}$$

and $\mathcal{G}^{(O)} := \sum_{(i,t) \in [N] \times [T]} s^{(O)}(i,t)$ satisfying $\mathcal{G}^{(O)} \leq d+1$. Moreover, the construction time of function $s^{(O)}$ is $T(NT, d+1) + O(NTd)$.

Proof: The proof idea comes from [51]. By Line 3 of Algorithm 1, $A \subseteq \mathbb{R}^{NT \times d'}$ is a matrix whose columns form a unit basis of the column space of Z. We have $d' \leq d+1$ and hence $\|A\|_2^2 = d' \leq d+1$. Moreover, for any $(i,t) \in [N] \times [T]$ and $\beta' \in \mathbb{R}^{d'}$, we have

$$\|\beta'\|_2^2 \leq \|A\beta'\|_2^2,$$

Then by Cauchy-Schwarz and orthonormality of A, we have that for any $(i,t) \in [N] \times [T]$ and $\beta' \in \mathbb{R}^{d+1}$,

$$|z_{it}^\top \beta'|^2 \leq \|A_{iT-T+t}\|_2^2 \cdot \|Z\beta'\|_2^2, \tag{5}$$

where A_{iT-T+t} is the $(iT - T + t)$-th row of A.

For each $(i,t) \in [N] \times [T]$, we let $s^{(O)}(i,t) := \|A_{iT-T+t}\|_2^2$. Then $\mathcal{G}^{(O)} = \|A\|_2^2 = d' \leq d+1$. Note that constructing A costs $T(NT, d+1)$ time and computing all $\|A_{iT-T+t}\|_2^2$ costs $O(NTd)$ time.

Thus, it remains to verify that $s^{(O)}(i,t)$ satisfies Inequality (4). For any $(i,t) \in [N] \times [T]$ and $\beta \in \mathbb{R}^d$, letting $\beta' = (\beta, -1)$, we have

$$\begin{aligned}
\psi_{it}^{(O)}(\beta) = &\ |z_{it}^\top \beta'|^2 && \text{(Defn. of } \psi_{it}^{(O)}) \\
\leq &\ \|A_{iT-T+t}\|_2^2 \cdot \|Z\beta'\|_2^2 && \text{(Ineq. (5))} \\
= &\ \|A_{iT-T+t}\|_2^2 \cdot \psi^{(O)}(\beta). && \text{(Defn. of } \psi^{(O)})
\end{aligned}$$

This completes the proof. □

Now we are ready to prove Lemma 4.4.

Proof: [Proof of Lemma 4.4] For any $(i,t) \in [N] \times [T]$, recall that $s(i,t)$ is defined by

$$s(i,t) := \min\left\{1, 2\lambda^{-1} \cdot \left(s^{(O)}(i,t) + \sum_{j=1}^{\min\{t-1,q\}} s^{(O)}(i,t-j)\right)\right\}.$$

We have that

$$\begin{aligned}\sum_{(i,t)\in[N]\times[T]} s(i,t) \leq & \sum_{(i,t)\in[N]\times[q]} 2\lambda^{-1} \times \left(s^{(O)}(i,t) + \textstyle\sum_{j=1}^{\min\{t-1,q\}} s^{(O)}(i,t-j)\right) \quad \text{(by definition)} \\ \leq & \; 2\lambda^{-1} \cdot \textstyle\sum_{(i,t)\in[N]\times[T]} (1+q)\cdot s^{(O)}(i,t) \\ \leq & \; 2\lambda^{-1}(q+1)(d+1). \qquad \text{(Lemma 4.7)}\end{aligned}$$

Hence, the total sensitivity $\mathcal{G} = O(\lambda^{-1}qd)$. By Lemma 4.7, it costs $T(NT, d+1) + O(NTd)$ time to construct $s^{(O)}$. We also know that it costs $O(NTq)$ time to compute function s. Since $T(NT, d+1) = O(NTd^2)$, this completes the proof for the running time.

Thus, it remains to verify that $s(i,t)$ satisfies that

$$s(i,t) \geq \sup_{\zeta\in\mathcal{P}} \frac{\psi^{(G,q)}_{it}(\zeta)}{\psi^{(G,q)}(\zeta)}.$$

Since $\sup_{\beta\in\mathbb{R}^d} \frac{\psi^{(O)}_{it}(\beta)}{\psi^{(O)}(\beta)} \leq 1$ always holds, we only need to consider the case that

$$s(i,t) = 2\lambda^{-1}\cdot\left(s^{(O)}(i,t) + \textstyle\sum_{j=1}^{\min\{t-1,q\}} s^{(O)}(i,t-j)\right).$$

We first show that for any $\zeta = (\beta,\rho)\in\mathcal{P}_\lambda$,

$$\psi^{(G,q)}(\zeta) \geq \lambda\cdot\psi^{(O)}(\beta). \tag{6}$$

Given an autocorrelation vector $\rho\in\mathbb{R}^q$, the induced covariance matrix Ω_ρ satisfies that $\Omega_\rho^{-1} = P_\rho^\top P_\rho$ where

$$P_\rho = \begin{bmatrix} \sqrt{1-\|\rho\|_2^2} & 0 & 0 & \ldots & \ldots & \ldots & 0 \\ -\rho_1 & 1 & 0 & \ldots & \ldots & \ldots & 0 \\ -\rho_2 & -\rho_1 & 1 & \ldots & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & -\rho_q & \ldots & -\rho_1 & 1 \end{bmatrix}. \tag{7}$$

Then by Equation (7), the smallest eigenvalue of P_ρ satisfies that

$$\begin{aligned} \lambda_{\min} = & \sqrt{1-\|\rho\|_2^2} \quad (\text{Defn. of } P_\rho) \\ \geq & \sqrt{\lambda}. \qquad (\rho\in B^q_{1-\lambda}) \end{aligned} \tag{8}$$

Also we have

$$\begin{aligned} \psi^{(G,q)}(\zeta) = & \sum_{i\in[N]} (y_i - X_i\beta)^\top \Omega_\rho^{-1}(y_i - X_i\beta) \quad \text{(Program (GLSE))} \\ = & \sum_{i\in[N]} \|P_\rho(y_i - X_i\beta)\|_2^2 \qquad (P_\rho^\top P_\rho = \Omega_\rho^{-1}) \\ \geq & \sum_{i\in[N]} \lambda\cdot\|(y_i - X_i\beta)\|_2^2 \qquad \text{(Ineq. (8))} \\ = & \; \lambda\cdot\psi^{(O)}(\beta), \qquad (\text{Defns. of } \psi^{(O)}) \end{aligned}$$

which proves Inequality (6). We also claim that for any $(i,t)\in[N]\times[T]$,

$$\psi^{(G,q)}_{it}(\zeta) \leq 2\cdot\left(\psi^{(O)}_{it}(\beta) + \textstyle\sum_{j=1}^{\min\{t-1,q\}} \psi^{(O)}_{i,t-j}(\beta)\right). \tag{9}$$

This trivially holds for $t = 1$. For $t \geq 2$, this is because

$$\begin{aligned}
& \psi_{it}^{(G,q)}(\zeta) \\
= \quad & \left((y_{it} - x_{it}^\top \beta) - \textstyle\sum_{j=1}^{\min\{t-1,q\}} \rho_j \cdot (y_{i,t-j} - x_{i,t-j}^\top \beta)\right)^2 && (t \geq 2) \\
\leq \quad & \left(1 + \textstyle\sum_{j=1}^{\min\{t-1,q\}} \rho_j^2\right) \times \left((y_{it} - x_{it}^\top \beta)^2 + \textstyle\sum_{j=1}^{\min\{t-1,q\}} (y_{i,t-j} - x_{i,t-j}^\top \beta)^2\right) && \text{(Cauchy-Schwarz)} \\
= \quad & 2 \cdot \left(\psi_{it}^{(O)}(\beta) + \textstyle\sum_{j=1}^{\min\{t-1,q\}} \psi_{i,t-j}^{(O)}(\beta)\right). && (\|\rho\|_2^2 \leq 1)
\end{aligned}$$

Now combining Inequalities (6) and (9), we have that for any $\zeta = (\beta, \rho) \in \mathcal{P}_\lambda$,

$$\begin{aligned}
\frac{\psi_{it}^{(G,q)}(\zeta)}{\psi^{(G,q)}(\zeta)} \leq \quad & \frac{2 \cdot \left(\psi_{it}^{(O)}(\beta) + \sum_{j=1}^{\min\{t-1,q\}} \psi_{i,t-j}^{(O)}(\beta)\right)}{\lambda \cdot \psi^{(O)}(\beta)} \\
\leq \quad & 2\lambda^{-1} \cdot \left(s^{(O)}(i,t) + \textstyle\sum_{j=1}^{\min\{t-1,q\}} s^{(O)}(i, t-j)\right) \\
= \quad & s(i,t).
\end{aligned}$$

This completes the proof. □

5 Coresets for GLSE$_k$

Following from Section 4, we assume that the parameter space is $\mathcal{P}_\lambda^k = (\mathbb{R}^d \times B_{1-\lambda}^q)^k$ for some given constant $\lambda \in (0,1)$. Given a panel dataset $X \in \mathbb{R}^{N \times T \times d}$ and $Y \in \mathbb{R}^{N \times T}$, let $Z^{(i)} \in \mathbb{R}^{T \times (d+1)}$ denote a matrix whose t-th row is $(x_{it}, y_{it}) \in \mathbb{R}^{d+1}$ for all $t \in [T]$ $(i \in [N])$. Assume there exists constant $M \geq 1$ such that the input dataset satisfies the following property.

Definition 5.1 (M-bounded dataset) *Given $M \geq 1$, we say a panel dataset $X \in \mathbb{R}^{N \times T \times d}$ and $Y \in \mathbb{R}^{N \times T}$ is M-bounded if for any $i \in [N]$, the condition number of matrix $(Z^{(i)})^\top Z^{(i)}$ is at most M, i.e., $\max_{\beta \in \mathbb{R}^d} \frac{\psi_i^{(O)}(\beta)}{\|\beta\|_2^2 + 1} \leq M \cdot \min_{\beta \in \mathbb{R}^d} \frac{\psi_i^{(O)}(\beta)}{\|\beta\|_2^2 + 1}$.*

If there exists $i \in [N]$ and $\beta \in \mathbb{R}^d$ such that $\psi_i^{(O)}(\beta) = 0$, we let $M = \infty$. Specifically, if all $(Z^{(i)})^\top Z^{(i)}$ are identity matrix whose eigenvalues are all 1, i.e., for any β, $\psi_i^{(O)}(\beta) = \|\beta\|_2^2 + 1$, we can set $M = 1$. Another example is that if $n \gg d$ and all elements of $Z^{(i)}$ are independently and identically distributed standard normal random variables, then the condition number of matrix $(Z^{(i)})^\top Z^{(i)}$ is upper bounded by some constant with high probability (and constant in expectation) [10, 46], which may also imply $M = O(1)$. The main theorem is as follows.

Theorem 5.2 (Coresets for GLSE$_k$) *There exists a randomized algorithm that given an M-bounded ($M \geq 1$) panel dataset $X \in \mathbb{R}^{N \times T \times d}$ and $Y \in \mathbb{R}^{N \times T}$, constant $\varepsilon, \lambda \in (0,1)$ and integers $q, k \geq 1$, with probability at least 0.9, constructs an ε-coreset for GLSE$_k$ of size*

$$O\left(\varepsilon^{-4} \lambda^{-2} M k^2 \max\left\{q^7 d^4, q^5 d^6\right\} \cdot \log \frac{Mq}{\lambda} \log \frac{Mkd}{\lambda}\right)$$

and runs in time $O(NTq + NTd^2)$.

Similar to GLSE, this coreset for GLSE$_k$ ($k \geq 2$) contains at most

$$(q+1) \cdot O\left(\varepsilon^{-4} \lambda^{-2} M k^2 \max\left\{q^7 d^4, q^5 d^6\right\} \cdot \log \frac{Mq}{\lambda} \log \frac{kd}{\lambda}\right)$$

points (x_{it}, y_{it}), which is independent of both N and T when M is constant. Note that the size contains an addtional factor M which can be unbounded. Our algorithm is summarized in Algorithm 2 and we outline Algorithm 2 and discuss the novelty in the following.

Remark 5.3 *Algorithm 2 is a two-staged framework, which captures the* min *operations in GLSE$_k$.*

```
Algorithm 2: $\mathsf{CGLSE_k}$: Coreset construction of GLSE$_k$
Require: an $M$-bounded (constant $M \geq 1$) panel dataset $X \in \mathbb{R}^{N\times T\times d}$ and $Y \in \mathbb{R}^{N\times T}$, constant
    $\varepsilon, \lambda \in (0,1)$, integers $k, q \geq 1$ and parameter space $\mathcal{P}_\lambda^k$.
Ensure: a subset $S \subseteq [N] \times [T]$ together with a weight function $w : S \to \mathbb{R}_{\geq 0}$.
    % Constructing a subset of individuals
 1: $\Gamma \leftarrow O\left(\varepsilon^{-2}\lambda^{-1}Mk^2 \max\left\{q^4d^2, q^3d^3\right\} \cdot \log \frac{Mq}{\lambda}\right)$.
 2: For each $i \in [N]$, let matrix $Z^{(i)} \in \mathbb{R}^{T\times(d+1)}$ be whose $t$-th row is $z_t^{(i)} = (x_{it}, y_{it}) \in \mathbb{R}^{d+1}$.
 3: For each $i \in [N]$, construct the SVD decomposition of $Z^{(i)}$ and compute
        $$u_i := \lambda_{\max}((Z^{(i)})^\top Z^{(i)}) \text{ and } \ell_i := \lambda_{\min}((Z^{(i)})^\top Z^{(i)}).$$
 4: For each $i \in [N]$, $s^{(O)}(i) \leftarrow \frac{u_i}{u_i + \sum_{i' \neq i} \ell_{i'}}$.
 5: For each $i \in [N]$, $s(i) \leftarrow \min\left\{1, \frac{2(q+1)}{\lambda} \cdot s^{(O)}(i)\right\}$.
 6: Pick a random sample $I_S \subseteq [N]$ of size $M$, where each $i \in I_S$ is selected w.p. $\frac{s(i)}{\sum_{i' \in [N]} s(i')}$.
 7: For each $i \in I_S$, $w'(i) \leftarrow \frac{\sum_{i' \in [N]} s(i')}{\Gamma \cdot s(i)}$.
    % Constructing a subset of time periods for each selected individual
 8: For each $i \in I_S$, apply $\mathbf{CGLSE}(X_i, y_i, \frac{\varepsilon}{3}, \frac{1}{20\Gamma}, \lambda, q)$ and construct $J_{S,i} \subseteq [T]$ together with a weight
    function $w^{(i)} : J_{S,i} \to \mathbb{R}_{\geq 0}$.
 9: Let $S \leftarrow \{(i,t) \in [N] \times [T] : i \in I_S, t \in J_{S,i}\}$.
10: For each $(i,t) \in S$, $w(i,t) \leftarrow w'(i) \cdot w^{(i)}(t)$.
11: Output $(S, w)$.
```

First stage. *We construct an $\frac{\varepsilon}{3}$-coreset $I_S \subseteq [N]$ together with a weight function $w' : I_S \to \mathbb{R}_{\geq 0}$ of the query space $(Z^{(G,q,k)}, \mathcal{P}^k, \psi^{(G,q,k)})$, i.e., for any $\zeta \in \mathcal{P}^k$*

$$\sum_{i \in I_S} w'(i) \cdot \psi_i^{(G,q,k)}(\zeta) \in (1 \pm \varepsilon) \cdot \psi^{(G,q,k)}(\zeta).$$

The idea is similar to Algorithm 1 except that we consider N sub-functions $\psi_i^{(G,q,k)}$ instead of NT. In Lines 2-4 of Algorithm 2, we first construct a sensitivity function $s^{(O)}$ of OLSE$_k$. The definition of $s^{(O)}$ captures the impact of min *operations in the objective function of OLSE$_k$ and the total sensitivity of $s^{(O)}$ is guaranteed to be upper bounded by Definition 5.1. The key is showing that the maximum influence of individual i is at most $\frac{u_i}{u_i + \sum_{j \neq i} \ell_j}$ (Lemma 5.7), which implies that the total sensitivity of $s^{(O)}$ is at most M. Then in Line 5, we construct a sensitivity function s of GLSE$_k$, based on a reduction from $s^{(O)}$ (Lemma 5.8).*

Second stage. *In Line 8, for each $i \in I_S$, apply* $\mathbf{CGLSE}(X_i, y_i, \frac{\varepsilon}{3}, \frac{1}{20 \cdot |I_S|}, \lambda, q)$ *and construct a subset $J_{S,i} \subseteq [T]$ together with a weight function $w^{(i)} : J_{S,i} \to \mathbb{R}_{\geq 0}$. Output $S = \{(i,t) \in [N] \times [T] : i \in I_S, t \in J_{S,i}\}$ together with a weight function $w : S \to \mathbb{R}_{\geq 0}$ defined as follows: for any $(i,t) \in S$, $w(i,t) := w'(i) \cdot w^{(i)}(t)$.*

We also provide a lower bound theorem which shows that the size of a coreset for GLSE$_k$ can be up to $\Omega(N)$. It indicates that the coreset size should contain additional factors than $k, q, d, 1/\varepsilon$, which reflects the reasonability of the M-bounded assumption.

Theorem 5.4 (Size lower bound of GLSE$_k$) *Let $T = 1$ and $d = k = 2$ and $\lambda \in (0,1)$. There exists $X \in \mathbb{R}^{N\times T\times d}$ and $Y \in \mathbb{R}^{N\times T}$ such that any 0.5-coreset for GLSE$_k$ should have size $\Omega(N)$.*

5.1 Proof overview

We first give a proof overview for summarization.

Proof overview of Theorem 5.2. For GLSE$_k$, we propose a two-staged framework (Algorithm 2): first sample a collection of individuals and then run **CGLSE** on every selected individuals. By Theorem 4.1, each subset $J_{S,i}$ at the second stage is of size $\mathrm{poly}(q,d)$. Hence, we only need to upper bound the size of I_S at the first stage. By a similar argument as that for GLSE, we can define the pseudo-dimension of GLSE$_k$ and upper bound it by $\mathrm{poly}(k,q,d)$, and hence, the main difficulty is to upper bound the total sensitivity of GLSE$_k$. We show that the gap between the individual regression objectives of GLSE$_k$ and OLSE$_k$ (GLSE$_k$ with $q=0$) with respect to the same $(\beta^{(1)},\ldots,\beta^{(k)})$ is at most $\frac{2(q+1)}{\lambda}$, which relies on $\psi_i^{(G,q)}(\zeta)\geq\lambda\cdot\psi_i^{(O)}(\beta)$ and an observation that for any $\zeta=(\beta^{(1)},\ldots,\beta^{(k)},\rho^{(1)},\ldots,\rho^{(k)})\in\mathcal{P}^k$, $\psi_i^{(G,q,k)}(\zeta)\leq 2(q+1)\cdot\min_{l\in[k]}\psi_i^{(O)}(\beta^{(l)})$. Thus, it suffices to provide an upper bound of the total sensitivity for OLSE$_k$. We claim that the maximum influence of individual i is at most $\frac{u_i}{u_i+\sum_{j\neq i}\ell_j}$ where u_i is the largest eigenvalue of $(Z^{(i)})^\top Z^{(i)}$ and ℓ_j is the smallest eigenvalue of $(Z^{(j)})^\top Z^{(j)}$. This fact comes from the following observation: $\min_{l\in[k]}\|Z^{(i)}(\beta^{(l)},-1)\|_2^2\leq\frac{u_i}{\ell_j}\cdot\min_{l\in[k]}\|Z^{(j)}(\beta^{(l)},-1)\|_2^2$, and results in an upper bound M of the total sensitivity for OLSE$_k$ since $\sum_{i\in[N]}\frac{u_i}{u_i+\sum_{j\neq i}\ell_j}\leq\frac{\sum_{i\in[N]}u_i}{\sum_{j\in[N]}\ell_j}\leq M$.

Proof overview of Theorem 5.4. For GLSE$_k$, we provide a lower bound $\Omega(N)$ of the coreset size by constructing an instance in which any 0.5-coreset should contain observations from all individuals. Note that we consider $T=1$ which reduces to an instance with cross-sectional data. Our instance is to let $x_{i1}=(4^i,\frac{1}{4^i})$ and $y_{i1}=0$ for all $i\in[N]$. Then letting $\zeta^{(i)}=(\beta^{(1)},\beta^{(2)},\rho^{(1)},\rho^{(2)})$ where $\beta^{(1)}=(\frac{1}{4^i},0)$, $\beta^{(2)}=(0,4^i)$ and $\rho^{(1)}=\rho^{(2)}=0$, we observe that $\psi^{(G,q,k)}(\zeta^{(i)})\approx\psi_i^{(G,q,k)}(\zeta^{(i)})$. Hence, all individuals should be contained in the coreset such that regression objectives with respect to all $\zeta^{(i)}$ are approximately preserved.

5.2 Proof of Theorem 5.2: Upper bound for GLSE$_k$

The proof of Theorem 5.2 relies on the following two theorems. The first theorem shows that I_S of Algorithm 2 is an $\frac{\varepsilon}{3}$-coreset of $\left(Z^{G,q,k},\mathcal{P}_\lambda^k,\psi^{(G,q,k)}\right)$. The second one is a reduction theorem that for each individual in I_S constructs an ε-coreset $J_{S,i}$.

Theorem 5.5 (Coresets of $\left(Z^{G,q,k},\mathcal{P}_\lambda^k,\psi^{(G,q,k)}\right)$) *For any given M-bounded observation matrix $X\in\mathbb{R}^{N\times T\times d}$ and outcome matrix $Y\in\mathbb{R}^{N\times T}$, constant $\varepsilon,\delta,\lambda\in(0,1)$ and integers $q,k\geq 1$, with probability at least 0.95, the weighted subset I_S of Algorithm 2 is an $\frac{\varepsilon}{3}$-coreset of the query space $\left(Z^{G,q,k},\mathcal{P}_\lambda^k,\psi^{(G,q,k)}\right)$, i.e., for any $\zeta=(\beta^{(1)},\ldots,\beta^{(k)},\rho^{(1)},\ldots,\rho^{(k)})\in\mathcal{P}_\lambda^k$,*

$$\sum_{i\in I_S}w'(i)\cdot\psi_i^{(G,q,k)}(\zeta)\in(1\pm\frac{\varepsilon}{3})\cdot\psi^{(G,q,k)}(\zeta). \tag{10}$$

Moreover, the construction time of I_S is

$$N\cdot\mathsf{SVD}(T,d+1)+O(N).$$

We defer the proof of Theorem 5.5 later.

Theorem 5.6 (Reduction from coresets of $\left(Z^{G,q,k},\mathcal{P}_\lambda^k,\psi^{(G,q,k)}\right)$ to coresets for GLSE$_k$) *Suppose that the weighted subset I_S of Algorithm 2 is an $\frac{\varepsilon}{3}$-coreset of the query space $\left(Z^{G,q,k},\mathcal{P}_\lambda^k,\psi^{(G,q,k)}\right)$. Then with probability at least 0.95, the output (S,w) of Algorithm 2 is an ε-coreset for GLSE$_k$.*

Proof: [Proof of Theorem 5.6] Note that S is an ε-coreset for GLSE$_k$ if Inequality (10) holds and for all $i\in[N]$, $J_{S,i}$ is an $\frac{\varepsilon}{3}$-coreset of $\left((Z^{(i)})^{(G,q)},\mathcal{P}_\lambda,\psi^{(G,q)}\right)$. By condition, we assume Inequality (10) holds. By Line 6 of Algorithm 2, the probability that every $J_{S,i}$ is an $\frac{\varepsilon}{3}$-coreset of $\left((Z^{(i)})^{(G,q)},\mathcal{P}_\lambda,\psi^{(G,q)}\right)$ is at least

$$1-\Gamma\cdot\frac{1}{20\Gamma}=0.95,$$

which completes the proof. □

Observe that Theorem 5.2 is a direct corollary of Theorems 5.5 and 5.6.

Proof: Combining Theorems 5.5 and 5.6, S is an ε-coreset of $\left(Z^{G,q,k}, \mathcal{P}_\lambda^k, \psi^{(G,q,k)}\right)$ with probability at least 0.9. By Theorem 4.1, the size of S is

$$\Gamma \cdot O\left(\varepsilon^{-2}\lambda^{-1}qd\left(\max\left\{q^2d, qd^2\right\}\cdot\log\frac{d}{\lambda}+\log\frac{\Gamma}{\delta}\right)\right),$$

which satisfies Theorem 5.2 by pluging in the value of Γ.

For the running time, it costs $N \cdot \mathsf{SVD}(T, d+1)$ to compute I_S by Theorem 5.5. Moreover, by Line 3 of Algorithm 2, we already have the SVD decomposition of $Z^{(i)}$ for all $i \in [N]$. Then it only costs $O\left(T(q+d)\right)$ to apply **CGLSE** for each $i \in I_S$ in Line 8 of Algorithm 2. Then it costs $O\left(NT(q+d)\right)$ to construct S. This completes the proof of the running time. □

Proof of Theorem 5.5: I_S is a coreset of $\left(Z^{(G,q,k)}, \mathcal{P}_\lambda^k, \psi^{(G,q,k)}\right)$. It remains to prove Theorem 5.5. Note that the construction of I_S applies the Feldman-Langberg framework. The analysis is similar to Section 4 in which we provide upper bounds for both the total sensitivity and the pseudo-dimension.

We first discuss how to bound the total sensitivity of $(Z^{(G,q,k)}, \mathcal{P}^k, \psi^{(G,q,k)})$. Similar to Section 4.5, the idea is to first bound the total sensitivity of $(Z^{(G,0,k)}, \mathcal{P}^k, \psi^{(G,0,k)})$ – we call it the query space of OLSE$_k$ whose covariance matrices of all individuals are identity matrices.

Lemma 5.7 (Total sensitivity of OLSE$_k$) *Function $s^{(O)} : [N] \to \mathbb{R}_{\geq 0}$ of Algorithm 2 satisfies that for any $i \in [N]$,*

$$s^{(O)}(i) \geq \sup_{\beta^{(1)},\ldots,\beta^{(k)}\in\mathbb{R}^d} \frac{\min_{l\in[k]}\psi_i^{(O)}(\beta^{(l)})}{\sum_{i'\in[N]}\min_{l\in[k]}\psi_{i'}^{(O)}(\beta^{(l)})}, \tag{11}$$

and $\mathcal{G}^{(O)} := \sum_{i\in[N]} s^{(O)}(i)$ satisfying that $\mathcal{G}^{(O)} = O(M)$. Moreover, the construction time of function $s^{(O)}$ is

$$N \cdot \mathsf{SVD}(T, d+1) + O(N).$$

Proof: For every $i \in [N]$, recall that $Z^{(i)} \in \mathbb{R}^{T\times(d+1)}$ is the matrix whose t-th row is $z_t^{(i)} = (x_{it}, y_{it}) \in \mathbb{R}^{d+1}$ for all $t \in [T]$. By definition, we have that for any $\beta \in \mathbb{R}^d$,

$$\psi_i^{(O)}(\beta) = \|Z^{(i)}(\beta, -1)\|_2^2.$$

Thus, by the same argument as in Lemma 4.7, it suffices to prove that for any matrix sequences $Z^{(1)}, \ldots, Z^{(N)} \in \mathbb{R}^{T\times(d+1)}$,

$$s^{(O)}(i) \geq \sup_{\beta^{(1)},\ldots,\beta^{(k)}\in\mathbb{R}^d} \frac{\min_{l\in[k]}\|Z^{(i)}(\beta^{(l)},-1)\|_2^2}{\sum_{i'\in[N]}\min_{l\in[k]}\|Z^{(i')}(\beta^{(l)},-1)\|_2^2}. \tag{12}$$

For any $\beta^{(1)}, \ldots, \beta^{(k)} \in \mathbb{R}^d$ and any $i \neq j \in [N]$, letting $l^\star = \arg\min_{l\in[k]}\|Z^{(j)}(\beta^{(l)}, -1)\|_2^2$, we have

$$\begin{aligned}
& \min_{l\in[k]}\|Z^{(i)}(\beta^{(l)},-1)\|_2^2 & \\
\leq\ & \|Z^{(i)}(\beta^{(l^\star)},-1)\|_2^2 & \\
\leq\ & u_i\cdot(\|\beta^{(l^\star)}\|_2^2+1) & \text{(Defn. of } u_i) \\
\leq\ & \frac{u_i}{\ell_j}\cdot\|Z^{(j)}(\beta^{(l^\star)},-1)\|_2^2 & \text{(Defn. of } \ell_i) \\
=\ & \frac{u_i}{\ell_j}\cdot\min_{l\in[k]}\|Z^{(j)}(\beta^{(l)},-1)\|_2^2. & \text{(Defn. of } l^\star)
\end{aligned}$$

Thus, we directly conclude that

$$
\begin{aligned}
& \frac{\min_{l\in[k]} \|Z^{(i)}(\beta^{(l)},-1)\|_2^2}{\sum_{i'\in[N]} \min_{l\in[k]} \|Z^{(i')}(\beta^{(l)},-1)\|_2^2} \\
\leq \quad & \frac{\min_{l\in[k]} \|Z^{(i)}(\beta^{(l)},-1)\|_2^2}{\left(1+\sum_{i'\neq i}\frac{\ell_{i'}}{u_i}\right)\cdot \min_{l\in[k]} \|Z^{(i)}(\beta^{(l)},-1)\|_2^2} \\
= \quad & \frac{u_i}{u_i+\sum_{i'\neq i}\ell_{i'}} \\
= \quad & s^{(O)}(i).
\end{aligned}
$$

Hence, Inequality (12) holds. Moreover, since the input dataset is M-bounded, we have

$$
\mathcal{G}^{(O)} \leq \sum_{i\in[N]} \frac{u_i}{\sum_{i'\in[N]}\ell_{i'}} \leq M,
$$

which completes the proof of correctness.

For the running time, it costs $N\cdot \mathsf{SVD}(T,d+1)$ to compute SVD decompositions for all $Z^{(i)}$. Then it costs $O(N)$ time to compute all u_i and ℓ_i, and hence costs $O(N)$ time to compute sensitivity functions $s^{(O)}$. Thus, we complete the proof. □

Note that by the above argument, we can also assume

$$
\sum_{i\in[N]} \frac{u_i}{u_i+\sum_{i'\neq i}\ell_{i'}} \leq M,
$$

which leads to the same upper bound for the total sensitivity $\mathcal{G}^{(O)}$. Now we are ready to upper bound the total sensitivity of $(Z^{(G,q,k)},\mathcal{P}^k,\psi^{(G,q,k)})$.

Lemma 5.8 (Total sensitivity of GLSE$_k$) *Function $s:[N]\rightarrow\mathbb{R}_{\geq 0}$ of Algorithm 2 satisfies that for any $i\in[N]$,*

$$
s(i) \geq \sup_{\zeta\in\mathcal{P}^k_\lambda} \frac{\psi_i^{(G,q,k)}(\zeta)}{\psi^{(G,q,k)}(\zeta)}, \tag{13}
$$

and $\mathcal{G}:=\sum_{i\in[N]} s(i)$ satisfying that $\mathcal{G}=O(\frac{qM}{\lambda})$. Moreover, the construction time of function s is

$$
N\cdot\mathsf{SVD}(T,d+1)+O(N).
$$

Proof: Since it only costs $O(N)$ time to construct function s if we have $s^{(O)}$, we prove the construction time by Lemma 5.7.

Fix $i\in[N]$. If $s(i)=1$ in Line 4 of Algorithm 2, then Inequality (13) trivially holds. Then we assume that $s(i)=\frac{2(q+1)}{\lambda}\cdot s^{(O)}(i)$. We first have that for any $i\in[N]$ and any $\zeta\in\mathcal{P}^k_\lambda$,

$$
\begin{aligned}
& \psi_i^{(G,q,k)}(\zeta) & \\
= \quad & \min_{l\in[k]} \textstyle\sum_{t\in[T]}\psi_{it}^{(G,q)}(\beta^{(l)},\rho^{(l)}) & \text{(Defn. 2.4)} \\
\geq \quad & \min_{l\in[k]} \textstyle\sum_{t\in[T]}\lambda\cdot\psi_{it}^{(O)}(\beta^{(l)}) & \text{(Ineq. (6))} \\
= \quad & \lambda\cdot\min_{l\in[k]}\psi_i^{(O)}(\beta^{(l)}). & \text{(Defn. of } \psi_i^{(O)})
\end{aligned}
$$

which directly implies that

$$
\psi^{(G,q,k)}(\zeta) \geq \lambda\cdot\sum_{i'\in[N]}\min_{l\in[k]}\psi_{i'}^{(O)}(\beta^{(l)}). \tag{14}
$$

We also note that for any $(i,t) \in [N] \times [T]$ and any $(\beta, \rho) \in \mathcal{P}_\lambda$,

$$\begin{aligned}
& \psi_{it}^{(G,q)}(\beta,\rho) \\
\leq \quad & \left((y_{it} - x_{it}^\top \beta) - \textstyle\sum_{j=1}^{\min\{t-1,q\}} \rho_j \cdot (y_{i,t-j} - x_{i,t-j}^\top \beta)\right)^2 && \text{(Defn. of } \psi_{it}^{(G,q)}\text{)} \\
\leq \quad & (1 + \textstyle\sum_{j=1}^{\min\{t-1,q\}} \rho_j^2) \times \left((y_{it} - x_{it}^\top \beta)^2 + \textstyle\sum_{j=1}^{\min\{t-1,q\}} (y_{i,t-j} - x_{i,t-j}^\top \beta)^2\right) && \text{(Cauchy-Schwarz)} \\
\leq \quad & 2\left((y_{it} - x_{it}^\top \beta)^2 + \textstyle\sum_{j=1}^{\min\{t-1,q\}} (y_{i,t-j} - x_{i,t-j}^\top \beta)^2\right). && (\|\rho\|_2^2 \leq 1)
\end{aligned}$$

Hence, we have that

$$\frac{1}{2} \cdot \psi_{it}^{(G,q)}(\beta,\rho) \leq (y_{it} - x_{it}^\top \beta)^2 + \sum_{j=1}^{\min\{t-1,q\}} (y_{i,t-j} - x_{i,t-j}^\top \beta)^2. \tag{15}$$

This implies that

$$\begin{aligned}
& \psi_i^{(G,q,k)}(\zeta) \\
= \quad & \min_{l \in [k]} \textstyle\sum_{t \in [T]} \psi_{it}^{(G,q)}(\beta^{(l)}, \rho^{(l)}) && \text{(Defn. 2.4)} \\
\leq \quad & \min_{l \in [k]} \textstyle\sum_{t \in [T]} 2 \times \left((y_{it} - x_{it}^\top \beta)^2 + \textstyle\sum_{j=1}^{\min\{t-1,q\}} (y_{i,t-j} - x_{i,t-j}^\top \beta)^2\right) && \text{(Ineq. (15))} \\
\leq \quad & 2(q+1) \cdot \min_{l \in [k]} \textstyle\sum_{t \in [T]} \psi_{it}^{(O)}(\beta^{(l)}) \\
= \quad & 2(q+1) \cdot \min_{l \in [k]} \psi_i^{(O)}(\beta^{(l)}). && \text{(Defn. of } \psi_i^{(O)}\text{)}
\end{aligned} \tag{16}$$

Thus, we have that for any $i \in [N]$ and $\zeta \in \mathcal{P}_\lambda^k$,

$$\begin{aligned}
\frac{\psi_i^{(G,q,k)}(\zeta)}{\psi^{(G,q,k)}(\zeta)} \leq \quad & \frac{2(q+1) \cdot \min_{l \in [k]} \psi_i^{(O)}(\beta^{(l)})}{\lambda \cdot \sum_{i \in [N]} \min_{l \in [k]} \psi_i^{(O)}(\beta^{(l)})} && \text{(Ineqs. (14) and (16))} \\
\leq \quad & \frac{2(q+1)}{\lambda} \cdot s^{(O)}(i) && \text{(Lemma 5.7)} \\
= \quad & s(i), && \text{(by assumption)}
\end{aligned}$$

which proves Inequality (13). Moreover, we have that

$$\mathcal{G} = \sum_{i \in [N]} s(i) \leq \frac{2(q+1)}{\lambda} \cdot \mathcal{G}^{(O)} = O(\frac{qM}{\lambda}),$$

where the last inequality is from Lemma 5.7. We complete the proof. □

Next, we upper bound the pseudo-dimension of GLSE$_k$. The proof is similar to that of GLSE by applying Lemmas 4.5 and 4.6.

Lemma 5.9 (Pseudo-dimension of GLSE$_k$) *The pseudo-dimension of any query space $(Z^{(G,q,k)}, u, \mathcal{P}_\lambda^k, \psi^{(G,q,k)})$ over weight functions $u : [N] \to \mathbb{R}_{\geq 0}$ is at most*

$$O\left(k^2 q^2 (q+d) d^2\right).$$

Proof: The proof idea is similar to that of Lemma 4.3. Fix a weight function $u : [N] \to \mathbb{R}_{\geq 0}$. For every $i \in [N]$, let $g_i : \mathcal{P}_\lambda^k \times \mathbb{R}_{\geq 0} \to \{0,1\}$ be the indicator function satisfying that for any $\zeta = (\beta^{(1)}, \ldots, \beta^{(k)}, \rho^{(1)}, \ldots, \rho^{(k)}) \in \mathcal{P}_\lambda^k$ and $r \in \mathbb{R}_{\geq 0}$,

$$g_i(\zeta, r) := \qquad I\left[u(i) \cdot \psi_i^{(G,q,k)}(\zeta) \geq r\right]$$

$$= \qquad I\left[\forall l \in [k],\ u(i)\cdot \sum_{t\in[T]} \psi_{it}^{(G,q)}(\beta^{(l)},\rho^{(l)}) \geq r\right].$$

We consider the query space $(Z^{(G,q,k)}, u, \mathcal{P}_\lambda^k \times \mathbb{R}_{\geq 0}, g)$. By the definition of $\mathcal{P}_\lambda^k$, the dimension of $\mathcal{P}_\lambda^k \times \mathbb{R}_{\geq 0}$ is $m = k(q+d)+1$. Also note that for any $(\beta,\rho)\in\mathcal{P}_\lambda$, $\psi_{it}^{(G,q)}(\beta,\rho)$ can be represented as a multivariant polynomial that consists of $O(q^2d^2)$ terms $\rho_{c_1}^{b_1}\rho_{c_2}^{b_2}\beta_{c_3}^{b_3}\beta_{c_4}^{b_4}$ where $c_1,c_2\in[q]$, $c_3,c_4\in[d]$ and $b_1,b_2,b_3,b_4\in\{0,1\}$. Thus, g_i can be calculated using $l = O(kq^2d^2)$ operations, including $O(kq^2d^2)$ arithmetic operations and k jumps. Pluging the values of m and l in Lemma 4.5, the pseudo-dimension of $(Z^{(G,q,k)}, u, \mathcal{P}_\lambda^k\times\mathbb{R}_{\geq 0}, g)$ is $O\left(k^2q^2(q+d)d^2\right)$. Then by Lemma 4.6, we complete the proof. □

Combining with the above lemmas and Theorem 4.2, we are ready to prove Theorem 5.5.

Proof: [Proof of Theorem 5.5] By Lemma 5.8, the total sensitivity $\mathcal{G}$ of $(Z^{(G,q,k)}, \mathcal{P}_\lambda^k, \psi^{(G,q,k)})$ is $O(\frac{qM}{\lambda})$. By Lemma 5.9, we can let $\dim = O\left(k^2(q+d)q^2d^2\right)$ which is an upper bound of the pseudo-dimension of every query space $(Z^{(G,q,k)}, u, \mathcal{P}_\lambda^k, \psi^{(G,q,k)})$ over weight functions $u:[N]\to\mathbb{R}_{\geq 0}$. Pluging the values of $\mathcal{G}$ and dim in Theorem 4.2, we prove for the coreset size.

For the running time, it costs $N\cdot \mathsf{SVD}(T,d+1)+O(N)$ time to compute the sensitivity function s by Lemma 5.8, and $O(N)$ time to construct I_S. This completes the proof. □

5.3 Proof of Theorem 5.4: Lower bound for GLSE$_k$

Actually, we prove a stronger version of Theorem 5.4 in the following. We show that both the coreset size and the total sensitivity of the query space $(Z^{(G,q,k)}, u, \mathcal{P}_\lambda^k, \psi^{(G,q,k)})$ may be $\Omega(N)$, even for the simple case that $T=1$ and $d=k=2$.

Theorem 5.10 (Size and sensitivity lower bound of GLSE$_k$) *Let $T=1$ and $d=k=2$ and $\lambda\in(0,1)$. There exists an instance $X\in\mathbb{R}^{N\times T\times d}$ and $Y\in\mathbb{R}^{N\times T}$ such that the total sensitivity*

$$\sum_{i\in[N]} \sup_{\zeta\in\mathcal{P}_\lambda^k} \frac{\psi_i^{(G,q,k)}(\zeta)}{\psi^{(G,q,k)}(\zeta)} = \Omega(N).$$

and any 0.5-coreset of the query space $(Z^{(G,q,k)}, u, \mathcal{P}_\lambda^k, \psi^{(G,q,k)})$ should have size $\Omega(N)$.

Proof: We construct the same instance as in [49]. Concretely, for $i\in[N]$, let $x_{i1} = (4^i, \frac{1}{4^i})$ and $y_{i1}=0$. We claim that for any $i\in[N]$,

$$\sup_{\zeta\in\mathcal{P}_\lambda^k} \frac{\psi_i^{(G,q,k)}(\zeta)}{\psi^{(G,q,k)}(\zeta)} \geq \frac{1}{2}. \tag{17}$$

If the claim is true, then we complete the proof of the total sensitivity by summing up the above inequality over all $i\in[N]$. Fix $i\in[N]$ and consider the following $\zeta = (\beta^{(1)},\beta^{(2)},\rho^{(1)},\rho^{(2)})\in\mathcal{P}_\lambda^k$ where $\beta^{(1)} = (\frac{1}{4^i}, 0)$, $\beta^{(2)} = (0, 4^i)$ and $\rho^{(1)}=\rho^{(2)}=0$. If $j\leq i$, we have

$$\begin{aligned}\psi_j^{(G,q,k)}(\zeta) &= \min_{l\in[2]}(y_{i1} - x_{i1}^\top \beta^{(l)})^2 \\ &= \min\left\{\frac{1}{16^{j-i}}, \frac{1}{16^{i-j}}\right\} \\ &= \frac{1}{16^{i-j}}.\end{aligned}$$

Similarly, if $j>i$, we have

$$\psi_j^{(G,q,k)}(\zeta) = \min\left\{\frac{1}{16^{j-i}}, \frac{1}{16^{i-j}}\right\} = \frac{1}{16^{j-i}}.$$

By the above equations, we have

$$\psi^{(G,q,k)}(\zeta) = \sum_{j=1}^{i} \frac{1}{16^{i-j}} + \sum_{j=i+1}^{N} \frac{1}{16^{j-i}} < \frac{5}{4}. \tag{18}$$

Combining with the fact that $\psi_i^{(G,q,k)}(\zeta) = 1$, we prove Inequality (17).

For the coreset size, suppose $S \subseteq [N]$ together with a weight function $w : S \to \mathbb{R}_{\geq 0}$ is a 0.5-coreset of the query space $(Z^{(G,q,k)}, u, \mathcal{P}_\lambda^k, \psi^{(G,q,k)})$. We only need to prove that $S = [N]$. Suppose there exists some $i^\star \in S$ with $w(i^\star) > 2$. Letting $\zeta = (\beta^{(1)}, \beta^{(2)}, \rho^{(1)}, \rho^{(2)})$ where $\beta^{(1)} = (\frac{1}{4^{i^\star}}, 0)$, $\beta^{(2)} = (0, 4^{i^\star})$ and $\rho^{(1)} = \rho^{(2)} = 0$, we have that

$$\begin{aligned}
\sum_{i \in S} w(i) \cdot \psi_i^{(G,q,k)}(\zeta) > \quad & w(i^\star) \cdot \psi_{i^\star}^{(G,q,k)}(\zeta) & \\
> \quad & 2 & (w(i^\star) > 2 \text{ and Defns. of } \zeta) \\
> \quad & (1 + \frac{1}{2}) \cdot \frac{5}{4} & \\
> \quad & (1 + \frac{1}{2}) \cdot \psi^{(G,q,k)}(\zeta), & (\text{Ineq. (18)})
\end{aligned}$$

which contradicts with the assumption of S. Thus, we have that for any $i \in S$, $w(i) \leq 2$. Next, by contradiction assume that $i^\star \notin S$. Again, letting $\zeta = (\beta^{(1)}, \beta^{(2)}, \rho^{(1)}, \rho^{(2)})$ where $\beta^{(1)} = (\frac{1}{4^{i^\star}}, 0)$, $\beta^{(2)} = (0, 4^{i^\star})$ and $\rho^{(1)} = \rho^{(2)} = 0$, we have that

$$\begin{aligned}
\sum_{i \in S} w(i) \cdot \psi_i^{(G,q,k)}(\zeta) \leq \quad & 2 \left(\psi^{(G,q,k)}(\zeta) - \psi_{i^\star}^{(G,q,k)}(\zeta)\right) & \\
& (w(i) \leq 2) & \\
\leq \quad & 2(\frac{5}{4} - 1) & (\text{Ineq. (18)}) \\
\leq \quad & (1 - \frac{1}{2}) \cdot 1 & \\
\leq \quad & (1 - \frac{1}{2}) \cdot \psi^{(G,q,k)}(\zeta), &
\end{aligned}$$

which contradicts with the assumption of S. This completes the proof.

□

6 Empirical results

We implement our coreset algorithms for GLSE, and compare the performance with uniform sampling on synthetic datasets and a real-world dataset. The experiments are conducted by PyCharm on a 4-Core desktop CPU with 8GB RAM.[2]

Datasets. We experiment using **synthetic** datasets with $N = T = 500$ ($250k$ observations), $d = 10$, $q = 1$ and $\lambda = 0.2$. For each individual $i \in [N]$, we first generate a mean vector $\overline{x}_i \in \mathbb{R}^d$ by first uniformly sampling a unit vector $x'_i \in \mathbb{R}^d$, and a length $\tau \in [0, 5]$, and then letting $\overline{x}_i = \tau x'_i$. Then for each time period $t \in [T]$, we generate observation x_{it} from a multivariate normal distribution $N(\overline{x}_i, \|\overline{x}_i\|_2^2 \cdot I)$ [50].[3] Next, we generate outcomes Y. First, we generate a regression vector $\beta \in \mathbb{R}^d$ from distribution $N(0, I)$. Then we generate an autoregression vector $\rho \in \mathbb{R}^q$ by first uniformly sampling a unit vector $\rho' \in \mathbb{R}^q$ and a length $\tau \in [0, 1 - \lambda]$, and then letting $\rho = \tau \rho'$. Based on ρ, we generate error terms e_{it} as in Equation (2). To assess performance robustness in the presence of outliers, we simulate another dataset replacing $N(0, I)$ in Equation (2) with the heavy tailed **Cauchy**(0,2) distribution [38]. Finally, the outcome $y_{it} = x_{it}^\top \beta + e_{it}$ is the same as Equation (1).

[2] Codes are in `https://github.com/huanglx12/Coresets-for-regressions-with-panel-data`.

[3] The assumption that the covariance of each individual is proportional to $\|\overline{x}_i\|_2^2$ is common in econometrics. We also fix the last coordinate of x_{it} to be 1 to capture individual specific fixed effects.

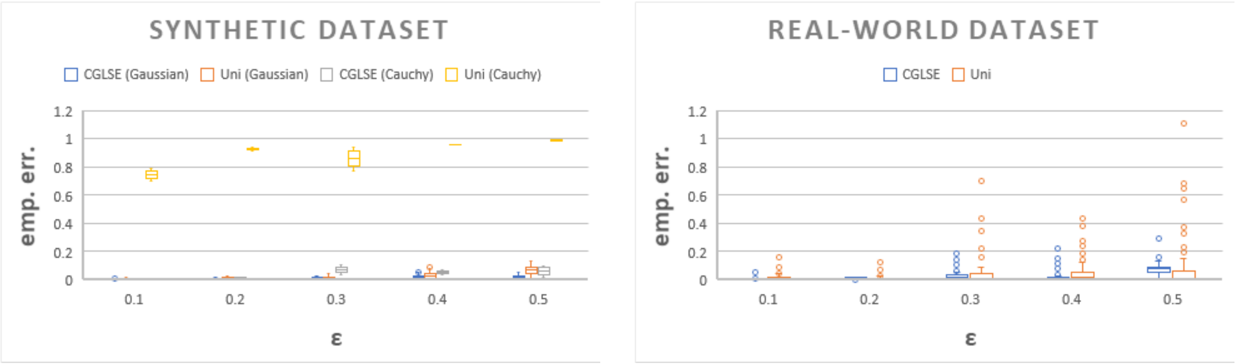

Figure 1: Boxplots of empirical errors for GLSE w.r.t. varying ε. **Uni** has higher average and maximum empirical errors than **CGLSE**.

We also experiment on a **real-world** dataset involving the prediction of monthly profits from customers for a credit card issuer as a function of demographics, past behaviors, and current balances and fees. The panel dataset consisted of 250k observations: 50 months of data ($T = 50$) from 5000 customers ($N = 5000$) with 11 features ($d = 11$). We set $q = 1$ and $\lambda = 0.2$.

Baseline and metrics. As a baseline coreset, we use uniform sampling (**Uni**), perhaps the simplest approach to construct coresets: Given an integer Γ, uniformly sample Γ individual-time pairs $(i, t) \in [N] \times [T]$ with weight $\frac{NT}{\Gamma}$ for each.

Given regression parameters ζ and a subset $S \subseteq [N] \times [T]$, we define the *empirical error* as $\left|\frac{\psi_S^{(G,q)}(\zeta)}{\psi^{(G,q)}(\zeta)} - 1\right|$. We summarize the empirical errors $e_1, \ldots, e_n$ by maximum, average, standard deviation (std) and root mean square error (RMSE), where RMSE= $\sqrt{\frac{1}{n}\sum_{i\in[n]} e_i^2}$. By penalizing larger errors, RMSE combines information in both average and standard deviation as a performance metric,. The running time for solving GLSE on dataset X and our coreset S are T_X and T_S respectively. T_C is the running time for coreset S construction .

Simulation setup. We vary $\varepsilon = 0.1, 0.2, 0.3, 0.4, 0.5$ and generate 100 independent random tuples $\zeta = (\beta, \rho) \in \mathbb{R}^{d+q}$ (the same as described in the generation of the synthetic dataset). For each ε, we run our algorithm **CGLSE** and **Uni** to generate coresets. We guarantee that the total number of sampled individual-time pairs of **CGLSE** and **Uni** are the same. We also implement IRLS [32] for solving GLSE. We run IRLS on both the full dataset and coresets and record the runtime.

Results. Table 1 summarizes the accuracy-size trade-off of our coresets for GLSE for different error guarantees ε. The maximum empirical error of **Uni** is always larger than that of our coresets (1.16-793x). Further, there is no error guarantee with **Uni**, but errors are always below the error guarantee with our coresets. The speed-up with our coresets relative to full data ($\frac{T_X}{T_C+T_S}$) in solving GLSE is 1.2x-108x. To achieve the maximum empirical error of .294 for GLSE in the real-world data, only 1534 individual-time pairs (0.6%) are necessary for **CGLSE**. With **Uni**, to get the closest maximum empirical error of 0.438, at least 2734 individual-time pairs) (1.1%) is needed; i.e.., **CGLSE** achieves a smaller empirical error with a smaller sized coreset. Though **Uni** may sometimes provide lower average error than **CGLSE**, it *always* has higher RMSE, say 1.2-745x of **CGLSE**. When there are outliers as with Cauchy, our coresets perform even better on all metrics relative to **Uni**. This is because **CGLSE** captures tails/outliers in the coreset, while **Uni** does not. Figure 1 presents the boxplots of the empirical errors.

7 Conclusion, limitations, and future work

This paper initiates a theoretical study of coreset construction for regression problems with panel data. We formulate the definitions of coresets for several variants of ℓ_2-regression, including OLSE, GLSE, and GLSE$_k$. For each variant, we propose efficient algorithms that construct a coreset of size independent of both N and T, based on the FL framework. Our empirical results indicate that our algorithms can accelerate the evaluation time and perform significantly better than uniform sampling.

Table 1: performance of ε-coresets for GLSE w.r.t. varying ε. We report the maximum/average/standard deviation/RMSE of the empirical error w.r.t. the 100 tuples of generated regression parameters for our algorithm **CGLSE** and **Uni**. Size is the # of sampled individual-time pairs, for both **CGLSE** and **Uni**. T_C is construction time (seconds) of our coresets. T_S and T_X are the computation time (seconds) for GLSE over coresets and the full dataset respectively. "Synthetic (G)" and "Synthetic (C)" represent synthetic datasets with Gaussian errors and Cauchy errors respectively.

	ε	max. emp. err.		avg./std./RMSE of emp. err.		size	T_C	$T_C + T_S$	T_X (s)
		CGLSE	**Uni**	**CGLSE**	**Uni**				
synthetic (G)	0.1	**.005**	.015	.001/.001/.002	.007/.004/.008	116481	2	372	458
	0.2	**.018**	.029	.006/.004/.008	.010/.007/.013	23043	2	80	458
	0.3	**.036**	.041	.011/.008/.014	.014/.010/.017	7217	2	29	458
	0.4	**.055**	.086	.016/.012/.021	.026/.020/.032	3095	2	18	458
	0.5	**.064**	.130	.019/.015/.024	.068/.032/.075	1590	2	9	458
synthetic (C)	0.1	**.001**	.793	.000/.000/.001	.744/.029/.745	106385	2	1716	4430
	0.2	**.018**	.939	.013/.003/.014	.927/.007/.927	21047	2	346	4430
	0.3	**.102**	.937	.072/.021/.075	.860/.055/.862	6597	2	169	4430
	0.4	**.070**	.962	.051/.011/.053	.961/.001/.961	2851	2	54	4430
	0.5	**.096**	.998	.060/.026/.065	.992/.004/.992	472	2	41	4430
real-world	0.1	**.029**	.162	.005/.008/.009	.016/.026/.031	50777	3	383	2488
	0.2	**.054**	.154	.017/.004/.017	.012/.024/.026	13062	3	85	2488
	0.3	**.187**	.698	.039/.038/.054	.052/.106/.118	5393	3	24	2488
	0.4	**.220**	.438	.019/.033/.038	.050/.081/.095	2734	3	20	2488
	0.5	**.294**	1.107	.075/.038/.084	.074/.017/.183	1534	3	16	2488

For GLSE$_k$, our coreset size contains a factor M, which may be unbounded and result in a coreset of size $\Omega(N)$ in the worst case. In practice, if M is large, each sensitivity $s(i)$ in Line 5 of Algorithm 2 will be close or even equal to 1. In this case, I_S is drawn from all individuals via uniform sampling which weakens the performance of Algorithm 2 relative to **Uni**. Future research should investigate whether a different assumption than the M-bound can generate a coreset of a smaller size.

There are several directions for future work. Currenly, q and d have a relatively large impact on coreset size; future work needs to reduce this effect. This will advance the use of coresets for machine learning, where d is typically large, and q is large in high frequency data. This paper focused on coreset construction for panel data with ℓ_2-regression. The natural next steps would be to construct coresets with panel data for other regression problems, e.g., ℓ_1-regression, generalized linear models and logistic regression, and beyond regression to other supervised machine learning algorithms.

Broader impact. In terms of broader impact on practice, many organizations have to routinely outsource data processing to external consultants and statisticians. But a major practical challenge for organizations in doing this is to minimize issues of data security in terms of exposure of their data for potential abuse. Further, minimization of such exposure is considered as necessary due diligence by laws such as GDPR and CCPA which mandates firms to minimize security breaches that violate the privacy rights of the data owner [45, 34]. Coreset based approaches to sharing data for processing can be very valuable for firms in addressing data security and to be in compliance with privacy regulations like GDPR and CCPA.

Further, for policy and managerial decision making in economics, social sciences and management, obtaining unbiased estimates of the regression relationships from observational data is critical. Panel data is a critical ingredient for obtaining such unbiased estimates. As ML methods are being adopted by many social scientists [5], ML scholars are becoming sensitive to these issues and our work in using coreset methods for panel data can have significant impact for these scholars.

A practical concern is that coresets constructed and shared for one purpose or model may be used by the data processor for other kinds of models, which may lead to erroneous conclusions. Further, there is also the potential for issues of fairness to arise as different groups may not be adequately represented in the coreset without incorporating fairness constraints [29]. These issues need to be explored in future research.

Acknowledgements

This research was conducted when LH was at Yale and was supported in part by an NSF CCF-1908347 grant.

References

[1] Pankaj K Agarwal, Sariel Har-Peled, and Kasturi R Varadarajan. Approximating extent measures of points. *Journal of the ACM (JACM)*, 51(4):606–635, 2004.

[2] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University press, 2009.

[3] Manuel Arellano. *Panel Data Econometrics*. 2002.

[4] Bertan Ari and H Altay Güvenir. Clustered linear regression. *Knowledge-Based Systems*, 15(3):169–175, 2002.

[5] Susan Athey. Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 5–6, 2015.

[6] Badi Baltagi. *Econometric analysis of panel data*. John Wiley & Sons, 2008.

[7] Joshua Batson, Daniel A Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012.

[8] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal coresets for least-squares regression. *IEEE transactions on information theory*, 59(10):6880–6892, 2013.

[9] Vladimir Braverman, Dan Feldman, and Harry Lang. New frameworks for offline and streaming coreset constructions. *CoRR*, abs/1612.00889, 2016.

[10] Zizhong Chen and Jack J. Dongarra. Condition numbers of gaussian random matrices. *SIAM Journal on Matrix Analysis and Applications*, 27(3):603–620, 2005.

[11] Kenneth L Clarkson. Subgradient and sampling algorithms for l_1 regression. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 257–266. Society for Industrial and Applied Mathematics, 2005.

[12] Kenneth L Clarkson, Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, Xiangrui Meng, and David P Woodruff. The fast cauchy transform and faster robust linear regression. *SIAM Journal on Computing*, 45(3):763–810, 2016.

[13] Kenneth L. Clarkson and David P. Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63:1 – 45, 2017.

[14] Alan Kaylor Cline and Inderjit S Dhillon. Computation of the singular value decomposition. *Citeseer*, 2006.

[15] Michael B Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 181–190. ACM, 2015.

[16] Michael B Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1758–1777. SIAM, 2017.

[17] Michael B. Cohen and Richard Peng. L_p row sampling by lewis weights. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 183–192, 2015.

[18] Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Sampling algorithms for l_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136. Society for Industrial and Applied Mathematics, 2006.

[19] Dan Feldman. Core-sets: An updated survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 10(1), 2020.

[20] Dan Feldman, Zahi Kfir, and Xuan Wu. Coresets for gaussian mixture models of any shape. *arXiv preprint arXiv:1906.04895*, 2019.

[21] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 569–578. ACM, 2011. https://arxiv.org/abs/1106.1379.

[22] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1434–1453. SIAM, 2013.

[23] Edward W Frees et al. *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge University Press, 2004.

[24] William E Griffiths, George G Judge, R Carter Hill, Helmut Lütkepohl, and Tsoung-Chao Lee. *The Theory and Practice of Econometrics*. Wiley, 1985.

[25] John N Haddad. A simple method for computing the covariance matrix and its inverse of a stationary autoregressive process. *Communications in Statistics-Simulation and Computation*, 27(3):617–623, 1998.

[26] Charles N. Halaby. Panel models in sociological research: Theory into practice. *Review of Sociology*, 30(1):507–544, 2004.

[27] Daniel Hoechle. Robust standard errors for panel regressions with cross-sectional dependence. *The stata journal*, 7(3):281–312, 2007.

[28] Cheng Hsiao. Analysis of panel data. *Analysis of Panel Data, by Cheng Hsiao, pp. 382. ISBN 0521818559. Cambridge, UK: Cambridge University Press, February 2003.*, page 382, 2003.

[29] Lingxiao Huang, Shaofeng Jiang, and Nisheeth K. Vishnoi. Coresets for clustering with fairness constraints. In *Advances in Neural Information Processing Systems*, pages 7587–7598, 2019.

[30] Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in euclidean spaces: Importance sampling is nearly optimal. In *STOC 2020: 52nd Annual ACM Symposium on Theory of Computing*, pages 1416–1429, 2020.

[31] Jonathan Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable Bayesian logistic regression. In *Advances in Neural Information Processing Systems*, pages 4080–4088, 2016.

[32] Murray Jorgensen. Iteratively reweighted least squares. *Encyclopedia of Environmetrics*, 3, 2006.

[33] Ibrahim Jubran, Alaa Maalouf, and Dan Feldman. Fast and accurate least-mean-squares solvers. In *Advances in Neural Information Processing Systems*, pages 8305–8316, 2019.

[34] T Tony Ke and K Sudhir. Privacy rights and data security: Gdpr and personal data driven markets. *Available at SSRN 3643979*, 2020.

[35] James P LeSage. The theory and practice of spatial econometrics. *University of Toledo. Toledo, Ohio*, 28(11), 1999.

[36] Mu Li, Gary L Miller, and Richard Peng. Iterative row sampling. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 127–136. IEEE, 2013.

[37] Mario Lucic, Matthew Faulkner, Andreas Krause, and Dan Feldman. Training Gaussian mixture models at scale via coresets. *The Journal of Machine Learning Research*, 18(1):5885–5909, 2017.

[38] Ping Ma, Michael W. Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 91–99, 2014.

[39] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 91–100, 2013.

[40] Alejandro Molina, Alexander Munteanu, and Kristian Kersting. Core dependency networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[41] Alexander Munteanu and Chris Schwiegelshohn. Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms. *Künstliche Intell.*, 32(1):37–53, 2018.

[42] Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David Woodruff. On coresets for logistic regression. In *Advances in Neural Information Processing Systems*, pages 6561–6570, 2018.

[43] Jeff M Phillips. Coresets and sketches. *arXiv preprint arXiv:1601.00617*, 2016.

[44] Sashank J Reddi, Barnabás Póczos, and Alexander J Smola. Communication efficient coresets for empirical loss minimization. In *UAI*, pages 752–761, 2015.

[45] Supreeth Shastri, Melissa Wasserman, and Vijay Chidambaram. The seven sins of personal-data processing systems under {GDPR}. In *11th {USENIX} Workshop on Hot Topics in Cloud Computing (HotCloud 19)*, 2019.

[46] Lin Shi, Taibin Gan, Hong Zhu, and Xianming Gu. The exact distribution of the condition number of complex random matrices. *The Scientific World Journal*, 2013(2013):729839–729839, 2013.

[47] Christian Sohler and David P. Woodruff. Subspace embeddings for the l_1-norm with applications. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 755–764, 2011.

[48] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. Cluster analysis: basic concepts and algorithms. *Introduction to data mining*, 8:487–568, 2006.

[49] Elad Tolochinsky and Dan Feldman. Generic coreset for scalable learning of monotonic kernels: Logistic regression, sigmoid and more. *arXiv: Learning*, 2018.

[50] Yung Liang Tong. *The multivariate normal distribution.* Springer Science & Business Media, 2012.

[51] Kasturi Varadarajan and Xin Xiao. On the sensitivity of shape fitting problems. In *32nd International Conference on Foundations of Software Technology and Theoretical Computer Science*, page 486, 2012.

[52] Mathukumalli Vidyasagar. *A theory of learning and generalization.* Springer-Verlag, 2002.

[53] Yan Zheng and Jeff M Phillips. Coresets for kernel regression. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 645–654. ACM, 2017.

A Discussion of the generative model (1)

In this section, we discuss the equivalence between the generative model (1) and the random effects estimator. In random effects estimators, there exist additional individual specific effects $\alpha_i \in \mathbb{R}$, i.e.,

$$y_{it} = x_{it}^\top \beta_i + \alpha_i + e_{it}, \tag{19}$$

and we assume that all individual effects are drawn from a normal distribution, i.e.,

$$\alpha_i \sim N(\mu, \sigma_0^2), \qquad \forall i \in [N].$$

where $\mu \in \mathbb{R}$ is the mean and $\sigma_0^2 \in \mathbb{R}_{\geq 0}$ is the covariance of an unknown normal distribution. By Equation (19), for any $i \in [N]$, we let $\alpha_i = \mu + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma_0^2)$. Then Equation (19) can be rewritten as

$$y_{it} = x_{it}^\top \beta_i + \mu + (\varepsilon_i + e_{it}).$$

Let $\Omega \in \mathbb{R}^{T\times T}$ denote the covariance matrix among error terms e_{it}. Next, we simplify $\varepsilon_i + e_{it}$ by e'_{it}. Consequently, error terms e'_{it} satisfy that

$$\begin{aligned}
&\mathsf{Exp}[e'_{it}] = 0, && \forall (i,t) \in [N] \times [T];\\
&\mathsf{Cov}(e'_{it}, e'_{i't'}) = 0 && \forall i \neq i'\\
&\mathsf{Cov}(e'_{it}, e'_{it'}) = \Omega_{tt'} + \sigma_0^2 = \Omega'_{tt'} && \forall i \in [N], t, t' \in [T].
\end{aligned}$$

By this assumption, a random effects estimator can be defined by the following:

$$\min_{\beta,\Omega} \sum_{i\in[N]} (y_i - X_i\beta_i - \mu\cdot\mathbf{1})^\top (\Omega')^{-1} (y_i - X_i\beta_i - \mu\cdot\mathbf{1}).$$

Thus, we verify that the random effects estimator is equivalent to the generative model (1).

B Existing results and approaches for OLSE

We note that finding an ε-coreset of X for OLSE can be reduced to finding an ε-coreset for least-squares regression with cross-sectional data. For completeness, we summarize the following theorems for OLSE whose proofs mainly follow from the literature.

Theorem B.1 (ε-Coresets for OLSE [8]) *There exists a deterministic algorithm that for any given observation matrix $X \in \mathbb{R}^{N\times T\times d}$, outcome matrix $Y \in \mathbb{R}^{N\times T}$, a collection $\mathcal{B} \subseteq \mathbb{R}^d$ and constant $\varepsilon \in (0,1)$, constructs an ε-coreset of size $O(d/\varepsilon^2)$ of OLSE, with running time $T_{SVD} + O(NTd^3/\varepsilon^2)$ where T_{SVD} is the time needed to compute the left singular vectors of a matrix in $\mathbb{R}^{NT\times(d+1)}$.*

Theorem B.2 (Accurate coresets for OLSE [33]) *There exists a deterministic algorithm that for any given observation matrix $X \in \mathbb{R}^{N\times T\times d}$, outcome matrix $Y \in \mathbb{R}^{N\times T}$, a collection $\mathcal{B} \subseteq \mathbb{R}^d$, constructs an accurate coreset of size $O(d^2)$ of OLSE, with running time $O(NTd^2 + d^8 \log(NT/d))$.*

B.1 Proof of Theorem B.1

We first prove Theorem B.1 and propose the corresponding algorithm that constructs an ε-coreset. Recall that $\mathcal{B} \subseteq \mathbb{R}^d$ denotes the domain of possible vectors β.

Proof: [Proof of Theorem B.1] Construct a matrix $A \in \mathbb{R}^{NT\times d}$ by letting the $(iT - T + t)$-th row of A be x_{it} for $(i,t) \in [N] \times [T]$. Similarly, construct a vector $\mathbf{b} \in \mathbb{R}^{NT}$ by letting $\mathbf{b}_{iT-T+t} = y_{it}$. Then for any $\beta \in \mathcal{B}$, we have

$$\psi^{(O)}(\beta) = \|A\beta - \mathbf{b}\|_2^2.$$

Thus, finding an ε-coreset of X of OLSE is equivalent to finding a row-sampling matrix $S \in \mathbb{R}^{m \times NT}$ whose rows are basis vectors $e_{i_1}^\top, \ldots, e_{i_m}^\top$ and a rescaling matrix $W \in \mathbb{R}_{\geq 0}^{m \times m}$ that is a diagonal matrix such that for any $\beta \in \mathcal{B}$,

$$\|WS\left(A\beta - \mathbf{b}\right)\|_2^2 \in (1 \pm \varepsilon) \cdot \|A\beta - \mathbf{b}\|_2^2.$$

By Theorem 1 of [8], we only need $m = O(d/\varepsilon^2)$ which completes the proof of correctness. Note that Theorem 1 of [8] only provides a theoretical guarantee of a weak-coreset which only approximately preserves the optimal least-squares value. However, by the proof of Theorem 1 of [8], their coreset indeed holds for any $\beta \in \mathbb{R}^d$.

The running time also follows from Theorem 1 of [8], which can be directly obtained by the algorithm stated below. □

Algorithm in [8]. We then introduce the approach of [8] as follows. Suppose we have inputs $A \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$.

1. Compute the SVD of $Y = [A, b] \in \mathbb{R}^{n \times (d+1)}$. Let $Y = U\Sigma V^\top$ where $U \in \mathbb{R}^{n \times (d+1)}, \Sigma \in \mathbb{R}^{(d+1)\times(d+1)}$ and $V \in \mathbb{R}^{(d+1)\times(d+1)}$.

2. By Lemma 2 of [8] which is based on Theorem 3.1 of [7], we deterministically construct sampling and rescaling matrices $S \in \mathbb{R}^{m \times n}$ and $W \in \mathbb{R}^{m \times m}$ ($m = O(d/\varepsilon^2)$) such that for any $y \in \mathbb{R}^{d+1}$,

$$\|WSUy\|_2^2 \in (1 \pm \varepsilon) \cdot \|Uy\|_2^2.$$

   The construction time is $O(nd^3/\varepsilon^2)$.

3. Output S and W.

B.2 Proof of Theorem B.2

Next, we prove Theorem B.2 and propose the corresponding algorithm that constructs an accurate coreset.

Proof: [Proof of Theorem B.2] The proof idea is similar to that of Theorem B.1. Again, we construct a matrix $A \in \mathbb{R}^{NT \times d}$ by letting the $(iT - T + t)$-th row of A be x_{it} for $(i, t) \in [N] \times [T]$. Similarly, construct a vector $\mathbf{b} \in \mathbb{R}^{NT}$ by letting $\mathbf{b}_{iT-T+t} = y_{it}$. Then for any $\beta \in \mathcal{B}$, we have

$$\psi^{(O)}(\beta) = \|A\beta - \mathbf{b}\|_2^2.$$

Thus, finding an ε-coreset of X of OLSE is equivalent to finding a row-sampling matrix $S \in \mathbb{R}^{m \times NT}$ whose rows are basis vectors $e_{i_1}^\top, \ldots, e_{i_m}^\top$ and a rescaling matrix $W \in \mathbb{R}_{\geq 0}^{m \times m}$ that is a diagonal matrix such that for any $\beta \in \mathcal{B}$,

$$\|WS\left(A\beta - \mathbf{b}\right)\|_2^2 = \|A\beta - \mathbf{b}\|_2^2.$$

By Theorem 3.2 of [33], we only need $m = (d+1)^2 + 1 = O(d^2)$. Moreover, we can construct matrices W and S in $O(NTd^2 + d^8 \log(NT/d))$ time by applying $n = NT$, and $k = 2(d+1)$ in Theorem 3.2 of [33]. □

Main approach in [33]. Suppose we have inputs $A \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$. Let $A' = [A, \mathbf{b}] \in \mathbb{R}^{n \times (d+1)}$ For any $\beta \in \mathbb{R}^d$, we let $\beta' = (\beta, -1) \in \mathbb{R}^{d+1}$ and have that

$$\|A\beta - \mathbf{b}\|_2^2 = \|A'\beta'\|_2^2 = (\beta')^\top (A')^\top A'\beta'.$$

The main idea of [33] is to construct a sub-matrix $C \in \mathbb{R}^{\left((d+1)^2+1\right)\times(d+1)}$ of A' whose rows are of the form $w_i \cdot (a_i, \mathbf{b}_i)^\top$ for some $i \in [n]$ and $w_i \geq 0$, such that $C^\top C = (A')^\top A'$. Then we have for any $\beta \in \mathbb{R}^d$,

$$\|C\beta'\|_2^2 = (\beta')^\top C^\top C\beta' = (\beta')^\top (A')^\top A'\beta' = \|A\beta - \mathbf{b}\|_2^2.$$

By the definition of C, there exists a row-sampling matrix S and a rescaling matrix W such that $C = WSA'$.

We then discuss why such a sub-matrix C exists. The main observation is that $(A')^\top A' \in \mathbb{R}^{(d+1)\times(d+1)}$ and

$$(A')^\top A' = \sum_{i \in [n]} (a_i, \mathbf{b}_i) \cdot (a_i, \mathbf{b}_i)^\top.$$

Thus, $\frac{1}{n} \cdot (A')^\top A'$ is inside the convex hull of n matrices $(a_i, \mathbf{b}_i) \cdot (a_i, \mathbf{b}_i)^\top \in \mathbb{R}^{(d+1)\times(d+1)}$. By the Caratheodory's Theorem, there must exist at most $(d+1)^2+1$ matrices $(a_i, \mathbf{b}_i) \cdot (a_i, \mathbf{b}_i)^\top$ whose convex hull also contains $\frac{1}{n} \cdot (A')^\top A'$. Then $\frac{1}{n} \cdot (A')^\top A'$ can be represented as a linear combination of these matrices, and hence, the sub-matrix $C \in \mathbb{R}^{\left((d+1)^2+1\right)\times(d+1)}$ exists.

Algorithm 1 of [33] shows how to directly construct such a matrix C. However, the running time is $O(n^2 d^2)$ which is undesirable. To accelerate the running time, Jubran et al. [33] apply the following idea.

1. For each $i \in [n]$, set $p_i \in \mathbb{R}^{(d+1)^2}$ as the concatenation of the $(d+1)^2$ entries of $(a_i, \mathbf{b}_i) \cdot (a_i, \mathbf{b}_i)^\top$. Let P be the collection of these points p_i. Then our objective is reduced to finding a subset $S \subseteq P$ of size $(d+1)^2+1$ such that the convex hull of S contains $\overline{P} = \frac{1}{n} \cdot \sum_{i\in[n]} p_i$.

2. Compute a balanced partition $P_1, \ldots, P_k$ of P into $k = 3(d+1)^2$ clusters of roughly the same size. By the Caratheodory's Theorem, there must exist at most $(d+1)^2+1$ partitions P_i such that the convex hull of their union contains $\overline{P}$. The main issue is how to these partitions P_i efficiently.

3. To address this issue, Jubran et al. [33] compute a sketch for each partition P_i including its size $|P_i|$ and the weighted mean
$$u_i := \frac{1}{|P_i|} \cdot \sum_{j\in P_i} p_j.$$
The construction of sketches costs $O(nd^2)$ time. The key observation is that there exists a set S of at most $(d+1)^2+1$ points u_i such that the convex hull of their union contains $\overline{P}$ by the Caratheodory's Theorem. Moreover, the corresponding partitions P_i of these u_i are what we need – the convex hull of $\bigcup_{i\in[n]:u_i\in S} P_i$ contains $\overline{P}$. Note that the construction of S costs $O\left(k^2\left((d+1)^2\right)^2\right) = O(d^8)$ time. Overall, it costs $O(nd^2+d^8)$ time to obtain the collection $\bigcup_{i\in[n]:u_i\in S} P_i$ whose convex hull contains $\overline{P}$.

4. We repeat the above procedure over $\bigcup_{i\in[n]:u_i\in S} P_i$ until obtaining an accurate coreset of size $(d+1)^2+1$. By the value of k, we note that
$$\left|\bigcup_{i\in[n]:u_i\in S} P_i\right| \leq n/2,$$
i.e., we half the size of the input set by an iteration. Thus, there are at most $\log(n/d)$ iterations and the overall running time is
$$\sum_{i=0}^{\log n} \frac{O(nd^2)}{2^i} + O(d^8)\cdot\log(n/d) = O\left(nd^2 + d^8\log(n/d)\right).$$