

OPTIMAL SAMPLING DESIGNS FOR MULTI-DIMENSIONAL STREAMING TIME SERIES WITH APPLICATION TO POWER GRID SENSOR DATA

BY RUI XIE¹, SHUYANG BAI^{2,*} AND PING MA^{2,†}

¹Department of Statistics and Data Science, University of Central Florida, Orlando, FL 32826 USA, rui.xie@ucf.edu

²Department of Statistics, University of Georgia, Athens, GA 30602 USA, *bsy9142@uga.edu; †pingma@uga.edu

The Internet of Things (IoT) system generates massive high-speed temporally correlated streaming data and is often connected with online inference tasks under computational or energy constraints. Online analysis of these streaming time series data often faces a trade-off between statistical efficiency and computational cost. One important approach to balance this trade-off is sampling, where only a small portion of the sample is selected for the model fitting and update. Motivated by the demands of dynamic relationship analysis of IoT system, we study the data-dependent sample selection and online inference problem for a multi-dimensional streaming time series, aiming to provide low-cost real-time analysis of high-speed power grid electricity consumption data. Inspired by D-optimality criterion in design of experiments, we propose a class of online data reduction methods that achieve an optimal sampling criterion and improve the computational efficiency of the online analysis. We show that the optimal solution amounts to a strategy that is a mixture of Bernoulli sampling and leverage score sampling. The leverage score sampling involves auxiliary estimations that have a computational advantage over recursive least squares updates. Theoretical properties of the auxiliary estimations involved are also discussed. When applied to European power grid consumption data, the proposed leverage score based sampling methods outperform the benchmark sampling method in online estimation and prediction. The general applicability of the sampling-assisted online estimation method is assessed via simulation studies.

1. Introduction. In the era of Internet of things (IoT), the prevalence of sensor networks, wearable devices, and power grid networks has led to an enormous amount of data streams being automatically collected every second, or even every millisecond. Examples range from security monitoring in power grids (Li et al., 2019) to traffic monitoring in smart traffic system (Nellore and Hancke, 2016), from health surveillance through smart wearable devices (Islam et al., 2015) to soil condition sensors in precision agriculture (Mat et al., 2016). These IoT data streams carry rich and time-sensitive information on the targeted subjects or systems, offering an unprecedented potential for real-time monitoring, forecasting and control. This potential, however, has not yet been sufficiently exploited, because the computing infrastructure still lags far behind the exponential growth of data sources. For instance, a network layer of IoT system deployed in a smart power grid usually consists of a large number of low bandwidth, low energy, low processing power nodes for communication using WiFi, 3G, 4G or power line communication technologies, rendering sophisticated real-time analytics infeasible (Jaradat et al., 2015). The IoT sensor data streams can arrive at a very high speed, which will accumulate a large quantity of data to be analyzed in a short period of time. For real-time tasks in IoT applications, such as prediction or dynamic information flow tracking, the inference speed may lag behind the data arriving speed, especially for complex inference tasks. To conquer the data overflow challenge in large-scale IoT applications, we aim to provide a sampling solution with reliable inference performance and low computation costs. We

Keywords and phrases: Streaming data, Sampling design, Multi-dimensional time series, D-optimality.

use publicly available power grid electricity consumption data as a motivating example to illustrate challenges and potential solutions.

EXAMPLE. Open Power System Data: Time series of electricity consumption

Electricity consumption, measured by electricity loads over time in a power grid system, is a typical type of streaming data that arrives at a high speed. In the smart IoT application, electricity consumption recorded from smart meters are capable of observing data at very high frequency, such as 25 kHz sampling frequency for the sinusoidal voltage signal (Jumar et al., 2020). In a power system, real-time feedback of electricity loads, which includes prediction and model fitting, is important in optimizing energy consumption patterns (Marangoni and Tavoni, 2021). The real-time inference helps effectively lower energy consumption by reducing energy demand and leveling off the usage peaks. The accurate real-time prediction will also benefit effective scheduling and decision-making in the power system (Xu et al., 2021). Therefore, online analysis of the electricity load time series plays an important role in practice.

In 2020, the electricity consumption data in the United States has projected to use a total of 1,000 million terabytes (TB) of storage (Siddik et al., 2021; Shehabi et al., 2016). There are emerging needs for novel approaches to analyzing such massive data, especially the real-time analysis of massive data streams. In our study, we shall work with the data streams from publicly available power system data platform (Open Power System Data, 2020), given the fact that energy data is often subject to restrictive terms of use. The Open Power System Data (2020) consists of time series of electricity consumption (load) for 37 European countries with hourly resolution. The selected time series are recorded from 2006-01-01, 00:00:00 Coordinated Universal Time (UTC) to 2018-12-14, 23:00:00 UTC, which results in 113,544 time points. Electricity power consumption data from different countries are reported through different platforms in the Open Power System Data. We use the actual load of ENTSO-E power statistics from 19 countries as the variables of interest. Figure 1 displays the dynamic

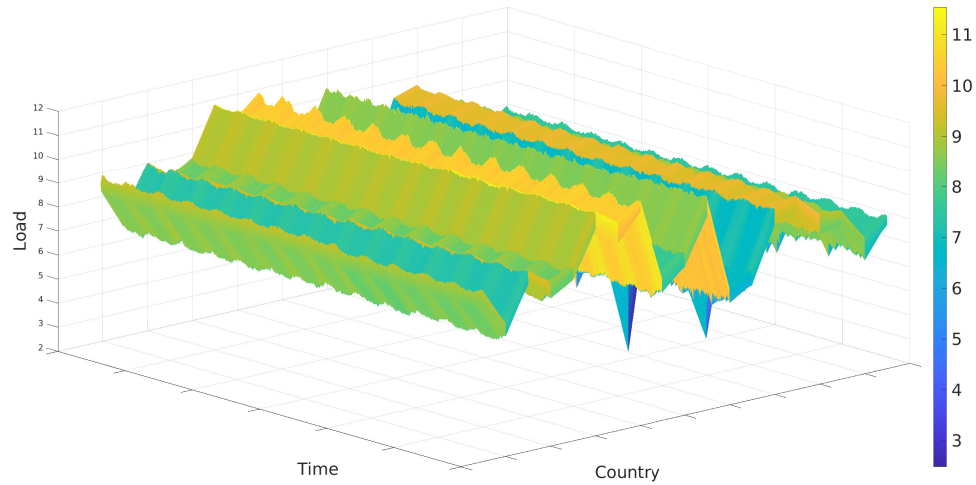


Fig 1: Load curves from 2006 to 2018 for 19 European countries.

evolution of the load curves from 2006 to 2018 for those 19 countries.

The online analysis of the IoT systems usually starts from sampling (Zhou and Saad, 2019) or filtering (Akbar et al., 2017) algorithms to optimize the data collection process from the IoT sensors, which results in massive volumes of regularly spaced multidimensional time series data streams, the latter being a well-developed statistical subject (Hamilton, 1994; Lütkepohl, 2005). Analyzing those large-scale high-frequency data streams poses challenges in real-time inference of multidimensional time series model, where the computational cost is in the order of K^3n^3 with n samples of K -dimensional data. Real-time calculations, including inference and prediction, are usually critical for decision making and resource optimization in IoT applications.

Methodologically speaking, *online analysis* of classical multidimensional time series is often carried out under the framework of dynamic linear model or state space model (West and Harrison, 1997; Petris et al., 2009). With dynamic linear models, online statistical computations including estimation and forecasting can often be aided by the efficient Kalman filter recursive algorithms (Kalman, 1960; Kalman and Bucy, 1961). However, for applications in many IoT systems, the stringent computational resource often cannot afford to perform typical dynamic linear model computations such as likelihood optimization or Bayesian sampling fast enough to achieve real-time analysis. In fact, even restricting to linear systems, for which much less costly online estimation algorithms (e.g., recursive least squares) can be applied, a real-time inference utilizing the whole data stream can still be challenging (Gabel et al., 2015; Berberidis et al., 2016; Hooi et al., 2018).

For smart power grid analysis, due to the limited accessibility of the data, various sampling techniques, including simple random sampling and Latin hypercube sampling (Cai et al., 2013), are used to generate synthetic data points for Monte-Carlo simulation or static power system model training (Balduin et al., 2022). The leverage score based sampling technique was used for monitoring cyberattacks in IoT system but not for the online inference (Li et al., 2019). To the best of our knowledge, the literature on sampling or data reduction method for online analysis of large-scale dynamic power grid systems is still lacking. Such a situation calls for the development of new approaches.

Broadly speaking, the problem described above belongs to a contemporary research direction that considers the trade-off between statistical efficiency and computational cost (Jordan, 2013). One natural and important approach to balance this trade-off is *sampling*. In this approach, a small portion of the full observed data is chosen as a surrogate to carry out computations of interest. Sampling, or more generally *data reduction*, has been considered in various studies as a means to reduce computational cost. The majority of these studies aim to achieve certain numerical approximation accuracy, which has sparked the popularity of notions such as sketching (e.g., Drineas et al. (2012); Liberty (2013); Woodruff (2014); Zhang et al. (2017)) and coresets (e.g., Agarwal et al. (2005); Dasgupta et al. (2009); Feldman et al. (2013)).

Recently in the context of big data, some statistical investigations of sampling methods with an emphasis on *statistical estimation efficiency* have emerged (Ma et al., 2015; Wang et al., 2018, 2019; Ting and Brochu, 2018; Yu et al., 2020; Meng et al., 2020; Ma et al., 2020; Wang et al., 2021). The aforementioned investigations are all carried out under the setup of independent samples and offline computation where the whole data set is available from the beginning. For online analysis of streaming time series exhibiting temporal dependence, relevant statistical research is still largely lacking. Exceptions are the work Xie et al. (2019) that considered online estimation of a Gaussian vector autoregression (VAR) model assisted by the so-called leverage score sampling (LSS), and the work Eshragh et al. (2022) that applied leverage score-based sampling for the analysis of large-scale univariate time series data. Leverage score sampling proposed in Xie et al. (2019) selects samples based on thresholding a leverage score of the lagged covariate in vector autoregression. However, a number of

questions regarding this sampling method were left unaddressed. Xie et al. (2019) showed that LSS achieves an asymptotic efficiency superior to a naive Bernoulli sampler. However, its optimality over other possible sampling methods was not clarified. Furthermore, the sample selection rule of LSS focuses exclusively on high-leverage covariate points, which in practice might lead to a lack of good fit over low-leverage region and sensitivity to outliers. In addition in Xie et al. (2019), the auxiliary estimations involved in implementing LSS were not justified.

The overall objective of this paper is to develop methodologies for performing online analysis on high-speed multidimensional time series, where we apply them to provide a real-time solution for massive power grid sensor data inference in IoT applications. On the methodology side, we propose a computationally efficient online sampling method called relaxed-LSS, which can be applied to linear multivariate time series models including extraneous variables and enjoys more robustness compared with LSS. The relaxed-LSS and the time series model are used for online inference and prediction of the IoT power consumption data. On the theory side, we establish the D-optimality of asymptotic estimation efficiency of LSS and more generally relaxed-LSS in a broad class of online sampling methods. We also establish consistency properties incorporating some auxiliary estimations in the sampling methods.

We organize the article as follows. In Section 2, we consider a framework that covers a large class of models subsuming the Gaussian Vector Autoregression considered in Xie et al. (2019). In Section 3, we formulate a class of online sampling methods within the framework of Section 2 and show that LSS is the optimal choice among them for asymptotic estimation efficiency. We then proceed to propose a novel relaxed version of LSS which keeps a proportion of low-leverage covariate points. Section 4 addresses the auxiliary estimations and practical implementation involved in the sampling algorithm, including an online estimate of an inverse covariance matrix which is performed sparsely for computational advantage. Section 5 considers an application to the electricity consumption time series in Open Power System Data. Section 6 includes the simulation studies of LSS and its relaxed version. The conclusion and future work are presented in Section 7. The proofs of the theoretical results are collected in Supplementary Material.

2. Model Assumptions. By default, all vectors are column vectors. Throughout this paper, $\|\cdot\|$ denotes the matrix operator norm with respect to the Euclidean norm (so it is the Euclidean norm for vectors).

Suppose $(\mathbf{y}_t)$ is the K -dimensional time series of interest. For notional convenience, we allow t to vary within the whole integer set $\mathbb{Z}$, whereas the starting point of t will become clear in a context. We shall model $(\mathbf{y}_t)$ as a $\mathbb{R}^K$ -valued ergodic (strictly) stationary process with finite variance. Suppose, in addition, that we have a $\mathbb{R}^p$ -valued ergodic stationary process $(\mathbf{x}_t)$ with finite variance serving as explanatory variables for $(\mathbf{y}_t)$. We impose the following stationarity assumption, which will help simplify the formulation of the basic results. Let $\boldsymbol{\mu}_X = E[\mathbf{x}_t]$, $\boldsymbol{\mu}_Y = E[\mathbf{y}_t]$, we shall assume the following linear system model for the centered time series,

$$(1) \quad \mathbf{y}_t - \boldsymbol{\mu}_Y = B'(\mathbf{x}_t - \boldsymbol{\mu}_X) + \mathbf{e}_t,$$

where B is a $p \times K$ coefficient matrix and B' stands for its transpose, $(\mathbf{e}_t)$ is a $\mathbb{R}^K$ -valued ergodic stationary noise process satisfying the martingale difference property,

$$(2) \quad E[\mathbf{e}_t | \mathcal{F}_t] = \mathbf{0},$$

where the filtration $\mathcal{F}_t = \sigma(\mathbf{x}_s, \mathbf{e}_{s-1}, s \leq t)$, and the conditional expectation in (2) is taken component-wise. We also assume a constant conditional covariance matrix for the error process:

$$(3) \quad E[\mathbf{e}_{t_1} \mathbf{e}_{t_2}' | \mathcal{F}_t] = \Omega 1_{\{t_1=t_2\}}, \quad t_1, t_2 \in \mathbb{Z}, \quad t_1, t_2 > t$$

for some non-singular covariance matrix Ω . Assume, in addition, that $(\mathbf{x}_t; \mathbf{e}_t)$ is jointly stationary. In literature (e.g., [Box et al. \(2015\)](#), Chapter 14)), the model (1) is often replaced by one with the means absorbed into an intercept term in $\mathbf{x}_t$. Our formulation here singles out the estimation of the means which is computationally trivial, and facilitates the development of the optimality results.

It is worth pointing out that the components of the explanatory variable process $(\mathbf{x}_t)$ are allowed to contain lagged values of $(\mathbf{y}_t)$. For instance, we may set

$$(4) \quad \mathbf{x}_t = (\mathbf{y}'_{t-1}, \dots, \mathbf{y}'_{t-p_1}, \mathbf{v}'_{t-1}, \dots, \mathbf{v}'_{t-p_2})' \in \mathbb{R}^{K(p_1+p_2)},$$

where $(\mathbf{v}_t)$ is a stationary process extraneous to $(\mathbf{y}_t)$, where each $\mathbf{v}_t \in \mathbb{R}^K$. The model (1) with the specification (4) is often known as a VARX model (e.g., [Lütkepohl \(2005\)](#), Chapter 10)), which becomes the well-known vector autoregression model if the extraneous variables are absent. The VARX model is more commonly expressed as

$$(5) \quad \mathbf{y}_t = \sum_{i=1}^{p_1} \Phi_i \mathbf{y}_{t-i} + \sum_{j=1}^{p_2} \Psi_j \mathbf{v}_{t-j} + \mathbf{e}_t$$

for some $K \times K$ coefficient matrices Φ_i 's and Ψ_j 's, where Φ_i 's must satisfy appropriate constraints to allow the existence of stationary solution of (5) (e.g., [Hamilton \(1994\)](#), Section 10.1)).

The linear system model (1) and its variants have been applied in various IoT contexts for real-time analysis of streaming time series: anomaly detection in streaming environmental sensor data ([Hill and Minsker, 2010](#)); tracking causal interactions between brain regions based on MEG sensor streams ([Michalareas et al., 2013](#)); design of energy-efficient operation of low-power wireless medical sensors ([Anagnostopoulos et al., 2014](#)); traffic forecasting in large urban areas based on road network sensors ([Schimbinschi et al., 2017](#)).

For the development of optimal online sampling theory in Section 3, we shall impose an additional assumption on the distributional shape of the covariate $\mathbf{x}_t$. Recall a p -dimensional elliptical (contoured) distribution $EC_p(\boldsymbol{\mu}, \Sigma, \nu)$ (cf. [Fang et al. \(1990\)](#)) is specified by the following three components: a vector $\boldsymbol{\mu} \in \mathbb{R}^p$ called *location*, a $p \times p$ non-negative definite matrix Σ called *scatter*, which coincides with the covariance matrix when the latter exists, and a probability distribution ν on $\mathbb{R}_+$. In particular, $EC_p(\boldsymbol{\mu}, \Sigma, \nu)$ denotes the distribution of random vector $\boldsymbol{\mu} + \xi \Sigma^{1/2} \mathbf{S}$, where $\Sigma^{1/2}$ is a symmetric square root of Σ , the random variable ξ follows the distribution ν and is called the *generating variate*, and $\mathbf{S}$ is the uniform distribution on the p -dimensional unit sphere $\{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\| = 1\}$ which is independent of ξ . A multivariate normal distribution is included in this family as a special case. For the rest of the paper, we shall assume that for each fixed $t \in \mathbb{Z}$, the covariate vector

$$(6) \quad \mathbf{x}_t \sim EC_p(\boldsymbol{\mu}_X, \Sigma, \nu),$$

where the distribution ν is absolutely continuous and the scatter Σ is also the positive definite covariance matrix of $\mathbf{x}_t$. Hence the distribution of $\mathbf{x}_t$ is absolutely continuous.

3. Optimal Online Sampling. To develop the main ideas, throughout this section we shall also assume for simplicity that

$$(7) \quad \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y = \mathbf{0}.$$

In practice, online estimation of these means can be achieved efficiently with a negligible computational cost compared to that of estimating B . See also Section 4 below for more details.

3.1. *A Class of Online Samplers.* When a sample stream $(\mathbf{y}_t; \mathbf{x}_t)_{t=1}^n$ from the whole stream $(\mathbf{y}_t; \mathbf{x}_t)_{t=1}^\infty$ satisfying (1) is observed and $n \gg Kp$, the least square estimator

$$\hat{B}_{n,LS} = \arg \min_B \sum_{t=1}^n \|\mathbf{y}_t - B' \mathbf{x}_t\|^2$$

may be used to estimate B . In the context of online estimation, the computation of $\hat{B}_{n,LS}$ can be implemented in a recursive manner based on the Sherman–Morrison inversion formula (Sherman and Morrison, 1950), an algorithm often known as the *recursive least squares* (Plackett, 1950). However, when data stream arrives at an overwhelmingly fast rate with high dimension (Kp is large), real-time update of $\hat{B}_{n,LS}$ can still be challenging given a limited computational capacity in the context of streaming data (see Section 1).

The basic idea we shall propose is simple: instead of updating the estimation along the whole stream of data $(\mathbf{y}_t; \mathbf{x}_t)_{t=1}^\infty$, we shall skip some time points t and only update the estimate along a subset $I \subset \mathbb{Z}_+$. The problem becomes: how should I be selected, or how do we select samples among $(\mathbf{y}_t; \mathbf{x}_t)_{t=1}^\infty$?

It is useful to note that our problem bears similarity to the study of *optimal designs* (e.g., Pukelsheim (1993); Papalambros and Wilde (2000)). Generally speaking, an optimal design aims at achieving an optimal estimation precision with a fixed sample size (number of experimental runs). In optimal design for linear regression with response $\mathbf{y}$ and design matrix X , i.e., $\mathbf{y} = X\beta + \epsilon$, where random error ϵ has mean zero and constant variance $\sigma^2 I$, one often considers optimizing the matrix $(X'X)^{-1}$ (recall $\text{var}[\hat{\beta}] = \sigma^2(X'X)^{-1}$, where $\hat{\beta}$ is the least square estimator) based on a certain criterion, e.g., the determinant (D-optimality). The optimization is typically over a finite available set of candidate covariate points (treatment runs) with the rows of X (number of runs) fixed. In other words, given a limited sample size, one needs to select appropriate covariate points among available ones to optimize $(X'X)^{-1}$.

Motivated by this insight drawn from optimal design, we propose to base our online sample selection criterion on the covariate $\mathbf{x}_t$ as well. We consider the following class of samplers: at each time point t , the selection of sample unit $(\mathbf{y}_t; \mathbf{x}_t)$ is determined solely by $\mathbf{x}_t \in \mathbb{R}^p$ up to a randomization. This is made precise as follows.

DEFINITION 3.1. Suppose there is a measurable function $s : \mathbb{R}^p \rightarrow [0, 1]$, called the sampling function. A sampling method (or sampler) $\mathcal{S}(s)$ is defined as follows: conditioning on $(\mathbf{y}_t; \mathbf{x}_t)_{t=1}^\infty$, for each $t \in \mathbb{Z}_+$, the sample $(\mathbf{y}_t; \mathbf{x}_t)$ is selected independently with probability $s(\mathbf{x}_t)$.

The sampler $\mathcal{S}(s)$ can be alternatively described as follows: let $(U_t)_{t=1}^\infty$ be i.i.d. $\text{Uniform}(0, 1)$ random variables which are independent of $(\mathbf{y}_t; \mathbf{x}_t)_{t=1}^\infty$. The selected index set I is given by

$$I = \{t \in \mathbb{Z}_+ : U_t \leq s(\mathbf{x}_t)\}.$$

The (unconditional) sampling rate of $\mathcal{S}(s)$ is given by

$$(8) \quad q := \text{pr}(U_t \leq s(\mathbf{x}_t)) = E[s(\mathbf{x}_t)] \in [0, 1].$$

When a stream of total length n passes through, the size of selected sample is given by a random number

$$N := \sum_{t=1}^n 1_{\{U_t \leq s(\mathbf{x}_t)\}} = |I|.$$

By the ergodic theorem (Kallenberg, 2002, Theorem 10.6), we have

$$(9) \quad \frac{N}{n} \rightarrow q$$

a.s. as $n \rightarrow \infty$.

As an example, a constant sampling function $s(\mathbf{x}) \equiv q$ corresponds to the *Bernoulli sampler*, that is, each index t is independently selected with probability q regardless of $\mathbf{x}_t$.

For an *ideal* stationary system, one may argue that there is no need to employ the sampling-assisted approach for online estimation. Indeed, if the regression coefficients in B in (1) stays invariant along the whole data stream, then the older data provides exactly the same information as the newer data about B . Hence one may use the available computational capacity to update the estimate up to the best speed even if it cannot catch up with the newest received data. However, in the practice of real-time analysis of streaming time series, this approach should be avoided since it fails to reflect the latest information about the streams. In fact, the usefulness of sampling-assisted online estimation is necessarily tied to *non-stationarity*, be it for the detection of departure from stationarity, or for the predictive modeling of non-stationary time series. For example, for predictive modeling, one may propose a time-varying version of (1) and estimate time-varying B by weighted least squares (e.g., Zhou and Wu (2010); Zhang and Wu (2012)). Since such a system can be locally viewed as stationary, the foundation laid in this paper will still be highly relevant in that context.

3.2. D-Optimality of Leverage Score Sampler. The next step is to formulate optimality among the class $\{S(s)\}$. Note that unlike a conventional setup in optimal design, for online estimation the number of rows of the design matrix X keeps increasing. So we propose to formulate the optimization, in a sense, on an asymptotic version of $(X^T X)^{-1}$. Let $\hat{B}_{n,I}$ be the least squares estimator of B using only $(\mathbf{y}_t; \mathbf{x}_t)_{t \in I}$, namely,

$$(10) \quad \hat{B}_{n,I} = \left(\sum_{t \in I} \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left(\sum_{t \in I} \mathbf{x}_t \mathbf{y}_t' \right) = \left(\sum_{t=1}^n 1_{\{U_t \leq s(\mathbf{x}_t)\}} \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left(\sum_{t=1}^n 1_{\{U_t \leq s(\mathbf{x}_t)\}} \mathbf{x}_t \mathbf{y}_t' \right).$$

We have the following asymptotic normality result.

THEOREM 3.2. *Under the assumptions in Section 2, suppose in addition that*

$$(11) \quad \Gamma(s) = E[s(\mathbf{x}_t) \mathbf{x}_t \mathbf{x}_t']$$

is invertible. Then as the total stream size $n \rightarrow \infty$,

$$(12) \quad \sqrt{N}(\text{vec}(\hat{B}_{n,I}) - \text{vec}(B)) \sim \sqrt{nq}(\text{vec}(\hat{B}_{n,I}) - \text{vec}(B)) \xrightarrow{d} \mathcal{N}(\mathbf{0}, P(s)^{-1}),$$

where $\text{vec}(B)$ denotes the vectorization of matrix B by stacking the columns (from left to right) of B into a single column, and the asymptotic precision matrix is

$$(13) \quad P(s) = \Omega^{-1} \otimes (q^{-1} \Gamma(s)),$$

where $\otimes$ denotes the Kronecker product and Ω is the covariance matrix defined in (3).

An extension of Theorem 3.2 incorporating auxiliary estimates for the means and the sampling function is stated in Theorem 4.1 below.

It is now natural to consider the optimization of the precision (information) matrix $P(s)$ under the constraint that the sampling rate $q = E[s(\mathbf{x}_t)]$ is fixed at a value in the interval $(0, 1)$. In general, one cannot expect to optimize $P(s)$ with respect to the Loewner order, namely, there is no $P(s)$ which is optimal in every direction (e.g., Pukelsheim (1993, Chapter 4)). Instead, one may consider the optimization of a suitable scalar function of $P(s)$, and most popularly, the determinant, which leads to the so-called D-optimality (e.g., Pukelsheim (1993, Chapter 9)). By a property of Kronecker product, $\det(P(s))$ is proportional to $\det(\Gamma(s))^K$, and hence we need to maximize $\det(\Gamma(s))$.

THEOREM 3.3 (D-optimality). *If the distribution of $\mathbf{x}_t$ is elliptical as specified by (6), then the following constrained optimization problem:*

$$(14) \quad \arg \max_s \det(P(s)) = \arg \max_s \det(\Gamma(s)) \quad \text{subject to} \quad E[s(\mathbf{x}_t)] = q \in (0, 1],$$

where the maximization is over all measurable $s : \mathbb{R}^p \rightarrow [0, 1]$, has solution

$$(15) \quad s(\mathbf{x}) = 1_{\{\mathbf{x}'\Sigma^{-1}\mathbf{x} > r\}},$$

where $r > 0$ is chosen so that

$$(16) \quad \text{pr}(\mathbf{x}_t'\Sigma^{-1}\mathbf{x}_t > r) = q.$$

This solution is almost everywhere unique with respect to the distribution of $\mathbf{x}_t$.

Theorem 3.3 is a special case of Theorem 3.4 below, which is provided in Section 3.3. We note that the optimality problem in Theorem 3.3 can be regarded as a special case of the general optimality problem formulated in Pronzato (2006); Pronzato and Wang (2020). Instead of obtaining an explicit solution as in (15), the aforementioned studies focused on stochastic approximation algorithms and were in general constrained to the setup of i.i.d. trials. The explicit solution in Theorem 3.3, although obtained under a more restrictive framework compared to Pronzato (2006); Pronzato and Wang (2020), enables one to formulate explicit and efficient online sampling algorithms which can be applied in the streaming time series context.

The tail quantile r in (16) may not be unique if $\text{pr}(\mathbf{x}_t'\Sigma^{-1}\mathbf{x}_t > r)$ is not strictly decreasing with respect to r (until reaching zero). One may eliminate this non-uniqueness by assuming that the density of ν in (6) is positive.

Besides D-optimality, one may consider some other optimality criteria, e.g., A-optimality which minimizes $\text{tr}(P(s)^{-1})$, E-optimality which maximizes the minimum eigenvalue of $P(s)$, T-optimality which maximizes $\text{tr}(P(s))$, etc (e.g., Pukelsheim (1993, Chapter 9)). Different criteria may lead to different optimal solutions of the sampling function s . We have chosen D-optimality since the solution has intuitively appealing interpretation as well as computational advantage.

Theorem 3.3 leads to the following simple deterministic sampling rule: $(\mathbf{y}_t; \mathbf{x}_t)$ is selected if $\mathbf{x}_t'\Sigma^{-1}\mathbf{x}_t$ exceeds the threshold r . Such a sampling rule was proposed in Xie et al. (2019) for the special case of Gaussian vector autoregression model, and was termed *leverage score sampling (LSS)* since

$$\ell_t = \mathbf{x}_t'\Sigma^{-1}\mathbf{x}_t,$$

can be interpreted as a *leverage score* of the $\mathbf{x}_t$ which marks the influence of this covariate point (cf. Seber and Lee (2012, Section 10.2)). In Xie et al. (2019) the superiority of LSS over a Bernoulli sampler was established under the same sampling rate q in terms of the Loewner order. Theorem 3.3 above takes a further step to show that under the D-optimality criterion, LSS is the best choice among the $\mathcal{S}(s)$ class in Definition 3.1 which includes the Bernoulli sampler.

To implement LSS in practice, one needs auxiliary estimations of the means of $\mathbf{y}_t$ and $\mathbf{x}_t$ (recall that we have assumed centering (7)), the calculation of leverage score ℓ_t as well as the q -quantile of ℓ_t . These issues will be discussed in a more general context in Section 4 below.

3.3. *Relaxed Leverage Score Sampling (Relaxed-LSS).* The hard thresholding found in (15) employed in LSS selects exclusively high-leverage $\mathbf{x}_t$ points. There are two potential risks of removing low-leverage design points in a linear regression: (a) if the linear regression relationship between the response and the covariates fails to hold in the low-leverage design region, then one would not be able to detect this departure; (b) relying on high-leverage points may make the regression particularly susceptible to the influence of outliers. Recall that the influence of an outlier is a combined effect of the magnitude of the regression residual and the leverage of the design point, as measured by Cook's distance (Cook, 1977) for instance.

These potential problems suggest us to relax LSS by including a reasonable fraction of low-leverage points. This motivates the following optimization modified from that in Theorem 3.3:

$$(17) \quad \arg \max_s \det(\Gamma(s)) \quad \text{subject to} \quad E[s(\mathbf{x}_t)] = q \in (0, 1], \quad s(\mathbf{x}) \geq q_0,$$

where s is the sampling function in Definition 3.1, and $q_0 \in [0, q]$. In other words, we include the additional constraint that there is a base sampling rate q_0 regardless of the design point $\mathbf{x}_t$.

THEOREM 3.4. *Under the same conditions as Theorem 3.3, the optimization problem (17) has solution*

$$(18) \quad s(\mathbf{x}) = q_0 + (1 - q_0)1_{\{\mathbf{x}'\Sigma^{-1}\mathbf{x} > r\}},$$

where r is chosen so that

$$(19) \quad \text{pr}(\mathbf{x}_t'\Sigma^{-1}\mathbf{x}_t > r) = \frac{q - q_0}{1 - q_0}.$$

This solution is almost everywhere unique with respect to the distribution of $\mathbf{x}_t$.

Theorem 3.4 includes Theorem 3.3 as a special case when $q_0 = 0$. This solution has the following interpretation: at each time point t , with a small probability q_0 , the sample unit $(\mathbf{y}_t; \mathbf{x}_t)$ will be selected regardless of the leverage score ℓ_t of $\mathbf{x}_t$; with a large probability $1 - q_0$, the hard thresholding of ℓ_t is applied to select sample. Such a strategy is also called rejection control in Monte Carlo computing (Liu, 2004). With a fractional of small-leverage points included, one may perform model diagnostics, e.g., using regression residuals, to assess the goodness of fit (Lütkepohl, 2005, Chapter 13) over low-leverage design region.

4. Auxiliary Estimates, Algorithm and Practical Implementation.

4.1. *Practical Implementation and Computational Complexity.* The Algorithm 1 summarizes the relaxed-LSS-assisted online estimation procedure involving some of the ingredients discussed below.

Incremental estimation of means

In practice, the means $\mu_Y = E[\mathbf{y}_t]$ and $\mu_X = E[\mathbf{x}_t]$ are often not zero and unknown. Since the computational cost of updating the means μ_X and μ_Y is minimal and having them estimated accurately is important for ensuring the consistency of $\hat{B}_{n,I}$, we propose to update them frequently, even at every step. We suggest starting the LSS-assisted online estimation using initialized means based on a pilot stream sample, and then have them updated incrementally.

Real-time calculation of leverage score

An effective real-time calculation of leverage score requires less computational effort than updating the least square. In practice, the inverse covariance matrix Σ^{-1} is unknown. To gain the computational advantage, a real-time calculation of leverage score approach is used by incorporating a crude estimate of precision matrix. A crude estimate of Σ^{-1} will affect only the efficiency of the leverage score sampling but not the consistency of $\widehat{B}_{n,I}$. We hence propose to use pilot estimation $P_0 := \widehat{\Sigma}_{n_0}^{-1}$ based on pilot data of size n_0 as a crude estimation of leverage score, i.e.

$$(20) \quad \ell_t \approx \mathbf{x}_t' P_0 \mathbf{x}_t,$$

so that we can update the leverage score with one vector-matrix multiplication, which has cost $O(p^2)$ per time point; or, alternatively, we update P_0 sparsely in time, which has total cost $O(cp^2)$ with $c \ll n$, given n as the total length of the observed stream size.

Computational Complexity

The running time for the LSS-assisted online estimation depends on both the time to calculate the leverage scores and the time to update the model estimation using sampled data. For an observed stream of size n , calculating leverage scores requires total $O(cp^2)$ time, and updating least squares estimation for sampled data requires $O((qn)p^2)$ time with sampling rate $q \ll 1$. So the total computational complexity of Algorithm 1 is $O((qn + c)p^2)$, where $(qn + c) \ll n$. The computational complexity of LSS or relaxed-LSS sampling methods hence is lower than the recursive least squares methods, where the latter needs to update the least squares estimates at every time point resulting in total $O(np^2)$ time. For the Bernoulli sampler (i.e., sampling function $s(\mathbf{x}) \equiv q$), the running time is trivial. To gain a further computational advantage in practice, one may use an efficient approximate computation of the leverage scores to perform relaxed-LSS algorithm (cf. Ma et al. (2015)). The approximation error analysis of the leverage scores can be found in Drineas et al. (2012) and Gittens and Mahoney (2013).

Determination of the threshold r

Another important practical issue is the determination of the threshold r . If each $\mathbf{x}_t$ is Gaussian, it can be shown that the sampling rate

$$(21) \quad pr(\ell_t > r) = pr(\chi_p^2 > r),$$

where χ_p^2 denotes a Chi-square distribution with p degrees of freedom. Hence for a pre-determined sampling rate $q \in (0, 1)$, one can choose r based on (21). In reality, data often exhibit heavier-tailed fluctuation compared to Gaussian. So one may start the sampling using a threshold r determined by (21), and then replace it with the empirical quantile of ℓ_t observed thus far. It is also possible to perform a pre-estimation of the tail quantile r based on a pilot sample of $\mathbf{x}_t' \Sigma^{-1} \mathbf{x}_t$ (with Σ^{-1} also pre-estimated).

4.2. Auxiliary Estimates. Below we formulate a general asymptotic result extending Theorem 3.2 which incorporates auxiliary estimates. Suppose the assumptions in Section 2 hold, but now we do not assume (7). We shall suppose that for some $\alpha > 2$,

$$(22) \quad E\|\mathbf{x}_t\|^\alpha < \infty, \quad t \in \mathbb{Z}_+,$$

as well as

$$(23) \quad \sup_t t^{1/2} [E\|\widehat{\boldsymbol{\mu}}_{t,X} - \boldsymbol{\mu}_X\|^\alpha]^{1/\alpha} < \infty, \quad \sup_t t^{1/2} [E\|\widehat{\boldsymbol{\mu}}_{t,Y} - \boldsymbol{\mu}_Y\|^\alpha]^{1/\alpha} < \infty.$$

The latter two relations are moderate strengthening of the usual root- n consistency. If $\widehat{\boldsymbol{\mu}}_{t,X}$ and $\widehat{\boldsymbol{\mu}}_{t,Y}$ are sample means up to time t and $E\|\mathbf{x}_t\|^\alpha < \infty$, $E\|\mathbf{y}_t\|^\alpha < \infty$, then (23) holds if a

strong mixing condition holds for $(\mathbf{x}_t)$ and $(\mathbf{y}_t)$ (Yokoyama, 1980), or if they are short-range dependent linear processes (Surgailis et al., 2012, Proposition 4.4.3).

Next, we assume that there is a family of sampling functions $s = s_q$ indexed by the sampling rate q , which satisfies

$$(24) \quad \lim_{q \rightarrow 0} \frac{q^{1-2/\alpha} \lambda_{\max}(\Gamma(s))^{1/2}}{\lambda_{\min}(\Gamma(s))} = 0,$$

where $q = E[s(\mathbf{x}_t)]$ is the sampling rate and $\lambda_{\max}(\Gamma(s))$ and $\lambda_{\min}(\Gamma(s))$ stand for the largest and the smallest eigenvalue of $\Gamma(s)$ respectively. Condition (24), roughly speaking, ensures that the magnitude of $\Gamma(s)$ decays slower than a power of q as $q \rightarrow 0$. This will be verified for the sampling function corresponding to the relaxed-LSS method in Lemma 4.2 below.

We suppose that $\hat{s}_t \in [0, 1]$ is an estimate of the plugged-in sampling function $s(\mathbf{x}_t - \boldsymbol{\mu}_X)$ based on the data stream observed up to time t such that as $t \rightarrow \infty$ (q fixed),

$$(25) \quad \hat{s}_t - s(\mathbf{x}_t - \boldsymbol{\mu}_X) \xrightarrow{P} 0.$$

This consistency condition will also be verified in Lemma 4.2 below for the auxiliary estimates involved in the relaxed-LSS method.

THEOREM 4.1. *Suppose that the conditions (22), (23), (24) and (25) hold. Let the estimator of B based on stream up to time n be as*

$$\hat{B}_{n,s} = \left(\sum_{t=1}^n \tilde{\mathbf{x}}_t \tilde{\mathbf{x}}_t' 1_{\{U_t \leq \hat{s}_t\}} \right)^{-1} \left(\sum_{t=1}^n \tilde{\mathbf{x}}_t \tilde{\mathbf{y}}_t' 1_{\{U_t \leq \hat{s}_t\}} \right),$$

where U_t 's are as in Definition 3.1, and $\tilde{\mathbf{x}}_t = \mathbf{x}_t - \hat{\boldsymbol{\mu}}_{t,X}$, $\tilde{\mathbf{y}}_t = \mathbf{y}_t - \hat{\boldsymbol{\mu}}_{t,Y}$. Then we have the decomposition

$$(26) \quad \sqrt{N}(\hat{B}_{n,s} - B) = M_n + R_n,$$

where as the total stream size $n \rightarrow \infty$,

$$(27) \quad \text{vec}(M_n) \xrightarrow{d} \mathcal{N}(\mathbf{0}, P(s)^{-1})$$

as $n \rightarrow \infty$ with $P(s)$ as in (13) but with $\Gamma(s)$ in (11) redefined as

$$(28) \quad \Gamma(s) = E[s(\mathbf{x}_t - \boldsymbol{\mu}_X)(\mathbf{x}_t - \boldsymbol{\mu}_X)(\mathbf{x}_t - \boldsymbol{\mu}_X)'],$$

which we assume to be non-singular. The term R_n satisfies for any $\delta > 0$,

$$(29) \quad \lim_{q \rightarrow 0} \limsup_n pr(\|P(s)^{1/2} \text{vec}(R_n)\| > \delta) = 0.$$

The double limit in (29) says that when the sampling rate is small, as typically desired in practice, the term R_n is negligible compared to M_n . Note that the same double limit with R_n replaced by M_n will not be zero due to (27).

The following lemma shows that the conditions (24) and (25) are satisfied in the context of relaxed-LSS, and hence justify the procedure in Algorithm 1.

LEMMA 4.2.

- (a) *Suppose $s(\mathbf{x}) = s_q(\mathbf{x}) = q_0 + (1 - q_0)1_{\{\mathbf{x}'\Sigma^{-1}\mathbf{x} > r\}}$ (q_0 and r depend on q) is the sampling function of the relaxed-LSS as in Theorem 3.4, where $q_0 = q_0(q)$ satisfies that for some constant $c \in (0, 1)$*

$$(30) \quad q_0 \leq cq.$$

Algorithm 1: Relaxed-LSS-Assisted Online Estimation of Stationary Linear System

Initialization:Choose a sampling rate $q \in (0, 1)$;Choose a base sampling rate $q_0 < q$;Initialize the estimates of μ_X , μ_Y , Σ^{-1} and B based on a pilot sample;**Online Estimation:****while** New sample $(y_t; x_t)$ at time t arrives **do** Update μ_X and μ_Y ; Calculate ℓ_t based on P_0 in (20); **if** Bernoulli(q_0) random number = 1 **then** Update the estimates of B and Ω with new centered sample $(y_t - \mu_Y; x_t - \mu_X)$; **else** **if** $\ell_t > r$ **then** Update the estimates of B and Ω with new centered sample $(y_t - \mu_Y; x_t - \mu_X)$; **end** **end** Update r based on (35) (or use (21) when the sample size is small).**end**

When $\alpha \in (2, 4]$, assume in addition that for some constant $c > 0$ and $\beta \in (\alpha, 4\alpha/(4 - \alpha))$ (right boundary understood as $+\infty$ when $\alpha = 4$), we have

$$(31) \quad \nu(x, \infty) > cx^{-\beta}$$

for all sufficient large x , where ν is as in (6). Then the condition (24) holds.

- (b) Suppose that $\hat{\Sigma}_n^{-1}$ is a consistent estimate of Σ^{-1} based on the stream observed up to time n , which can be realized by

$$(32) \quad \hat{\Sigma}_n^{-1} = \left(\frac{1}{M_n} \sum_{t=1}^n J_t(x_t - \hat{\mu}_{t,X})(x_t - \hat{\mu}_{t,X})' \right)^{-1},$$

with a desirable small update rate $u \in (0, 1)$, which corresponds to i.i.d. Bernoulli(u) random variables (J_t) independent of everything else, where $\hat{\mu}_{t,X}$ and $\hat{\mu}_{t,Y}$ be estimators of μ_X , μ_Y based on the stream observed up to time t , respectively.

Define the leverage score incorporating the auxiliary estimates as

$$(33) \quad \tilde{\ell}_t = \tilde{x}_t' \hat{\Sigma}_t^{-1} \tilde{x}_t.$$

Let q_0 be the base sampling rate for relaxed-LSS in Section 3.3. Suppose that $\hat{r}_n$ is a consistent estimate of

$$(34) \quad r(q, q_0) := \inf \left\{ r \geq 0 : \text{pr}((x_t - \mu_X)' \Sigma^{-1} (x_t - \mu_X) \leq r) \geq \frac{q - q_0}{1 - q_0} \right\}$$

based on the stream observed up to time n , which can be realized by

$$(35) \quad \hat{r}_n = \inf \left\{ r \geq 0 : \frac{1}{n} \sum_{t=1}^n 1_{\{\tilde{\ell}_t \leq r\}} \geq \frac{q - q_0}{1 - q_0} \right\}.$$

Let the estimated plugged-in sample function be

$$(36) \quad \hat{s}_t = q_0 + (1 - q_0) 1_{\{\tilde{\ell}_t > \hat{r}_t\}}.$$

Then the condition (25) holds.

Examining the proof reveals that the condition (30) can be relaxed by allowing a power of q . We omit such a generalization here for simplicity. We also note that the assumption (31) is not stringent. In particular, suppose the tail probability $\nu(x, \infty)$ is regularly varying with index $-\gamma$ (see Bingham et al. (1989)), which roughly speaking, says $\nu(x, \infty)$ behaves like $x^{-\gamma}$ as $x \rightarrow \infty$, $\gamma > 2$. This regular variation assumption includes common heavy-tailed distributions such as Pareto distributions and t -distributions. Then in view of Potter’s bound (Bingham et al., 1989, Theorem 1.5.6), one can find α and β satisfying $2 < \alpha < \gamma < \beta < 4\alpha/(4 - \alpha)$ so that the conditions (22) and (31) both hold.

5. The Open Power System Data Application. In this section, we apply our LSS and relaxed-LSS methods and benchmark Bernoulli method to the [Open Power System Data \(2020\)](#) for real-time inference on the dynamics of power grid load profiles and one-step ahead prediction. We compare the proposed LSS-based methods to the benchmark sampling method and “full sample” estimation and demonstrate the strengths of real-time estimation and prediction of the proposed methods.

5.1. Open Power System Data. Geographically, the Open Power System Data consist 37 European countries that cover European Union and neighboring countries. The data measures the total load (in Terawatt hour, TWh) for a country, control area or bidding zone. The total load is a power statistic, which is defined as, roughly speaking, the total power generated or imported minus the power being consumed at power plant, stored or exported. More specifically, the data reported by ENTSO-E Transparency Platform are used due to its high efficiency in data reporting, which results in a subset of 19 countries. The selected multivariate time series streams are recorded from 2006-01-01, 00:00:00 Coordinated Universal Time (UTC) to 2018-12-14, 23:00:00 UTC. There are total 113,544 time points been observed in the 19-dimensional electricity load stream, which are complete without missing values.

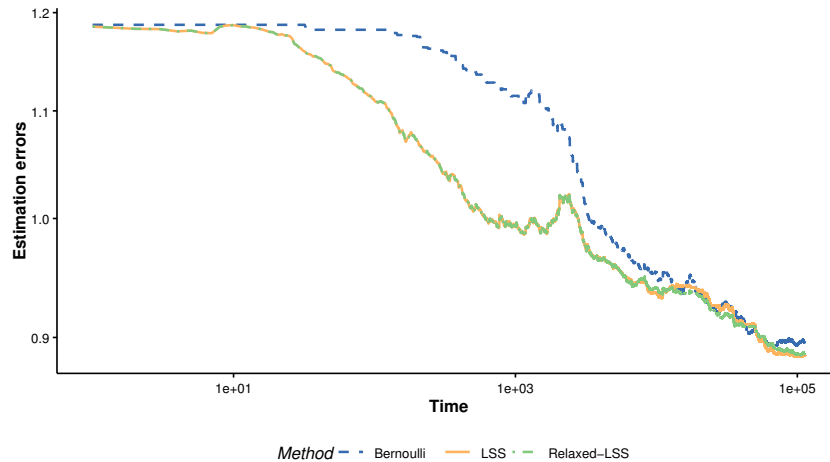


Fig 2: Comparison of estimation errors for each of the time points of power load data.

5.2. Seasonal VARX modeling and accuracy measurements. Electricity loads exhibit strong seasonality compared to loads one day (24 hours) earlier since it was aggregated

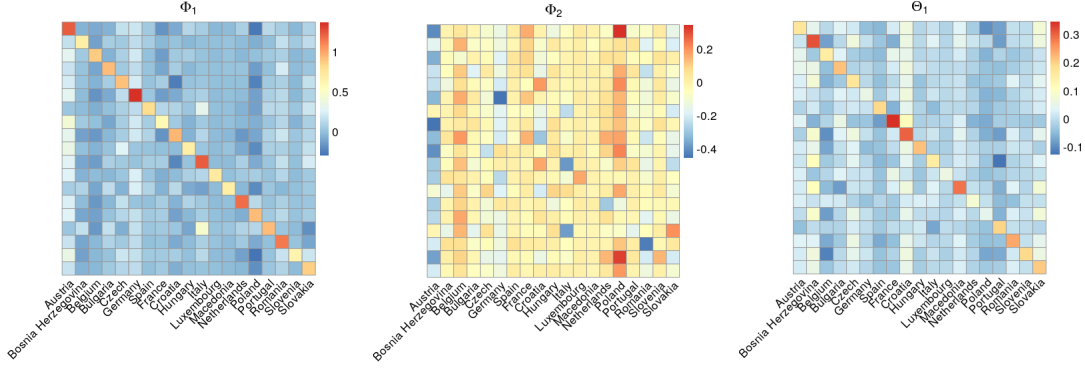


Fig 3: Visualization of the estimated coefficient matrices Φ_1 , Φ_2 and Θ_1 at time $t = 113,424$ using relaxed-LSS method.

to hourly temporal resolution. We consider the seasonal vector autoregression model with centering:

$$(37) \quad \mathbf{y}_t = \sum_{i=1}^{p_1} \Phi_i \mathbf{y}_{t-i} + \sum_{i=1}^{p_2} \Theta_i \mathbf{y}_{t-24i} + \mathbf{e}_t,$$

where we choose the model order as $p_1 = 2$, $p_2 = 1$ incorporating the daily seasonality. The weather, electricity prices, and clean energy generation and capacities can be added in to this seasonal VARX model as exogenous variables. However, due to the large portion and complicated missing pattern in the database, we did not include them in the real-time analysis. Our goals are to estimate the model parameters Φ 's and Θ 's in real time, which represent the dynamic dependence of the electricity loads, and the one-hour ahead forecasting (prediction) of electricity loads, which is crucial for effective scheduling and energy management in power grids.

We denote the $\hat{B}_t$ as the real time estimation of model parameter matrix B at time point t , and $\hat{B}_{\text{Full}}$ is the offline model parameter estimation based on the entire dataset as a substitute of the unknown population B . We use the relative errors to measure the parameter matrix estimation accuracy $\|\hat{B}_t - \hat{B}_{\text{Full}}\|_F / \|\hat{B}_{\text{Full}}\|_F$.

At time point t , if the sample unit $(\mathbf{y}_t; \mathbf{x}_t)$ is sampled, we will update the estimation $\hat{B}_t$ using least squares; otherwise the estimation from previous time point will be retained as the current estimation. The results are demonstrated in Figure 2, where the x-axis represents the time of updates.

For the pointwise prediction accuracy, we compute the one-step (one-hour) ahead relative prediction errors for each of the time points as: $\|\hat{\mathbf{y}}_{t+1} - \mathbf{y}_{t+1}\| / \|\mathbf{y}_{t+1}\|$. We visualized the prediction errors in Figure 4, where x-axis represents hourly time.

5.3. Results. We compare our LSS and relaxed-LSS methods to the benchmark Bernoulli sampling method for parameter matrix estimation accuracy and prediction accuracy. For all three methods, we use the same pilot data of size 500 to estimate the model order and initial values for model parameter B and precision matrix P_0 . The sampling rate is $q = 0.05$ for all three methods, and the base sampling rate for relaxed-LSS is $q_0 = 0.025$. The update rate for the inverse covariance matrix is 0.025. We handle the data in a streaming fashion and take data samples and estimate the model in real time.

As an illustration, Figure 3 depicts the estimated coefficient matrices Φ_1 , Φ_2 and Θ_1 at time $t = 113,424$, which is the last update, using relaxed-LSS method. The cross-correlations

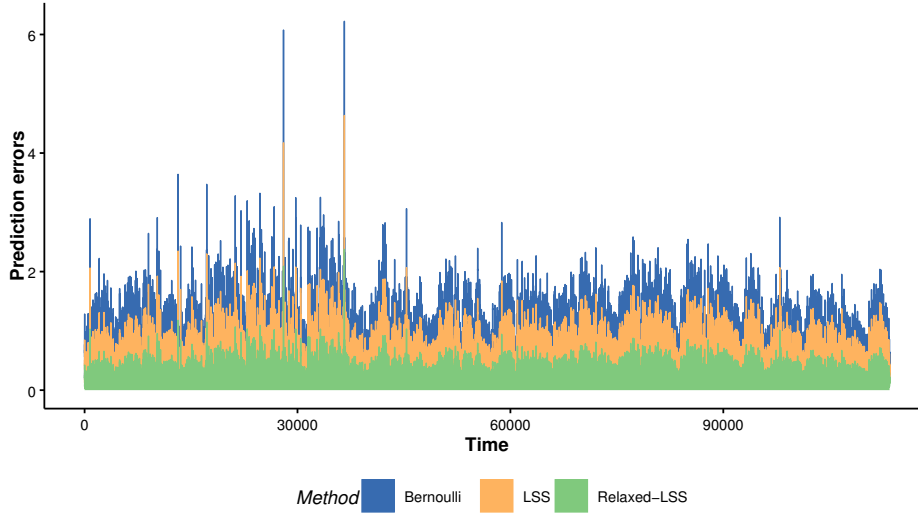


Fig 4: Comparison of one-step ahead prediction errors for each of the time point of power load data.

at lag 1 (coefficient matrix Φ_1) between Italy and Portugal, Austria and France, as well as Italy and Spain are higher than others, which reflect their positive interdependence in electricity consumption. For the cross-correlations at lag 2 (coefficient matrix Φ_2), Austria and several countries are negatively correlated, and Poland and several countries are positively correlated with magnitudes higher than others. We also notice that Germany has higher auto-correlation at lag 1 (positively) and lag 2 (negatively). France, Bosnia Herzegovina, Croatia and Macedonia have higher daily-seasonal correlation than other countries. It is also worth noting that those auto and cross correlations are dynamically evolving so that the Figure 3 is just a snapshot of the relationships at the time point. The computational cost of online estimation of those model coefficient matrices could be higher if other variables of the power grid system are included in the model, such as electricity prices, wind and solar power generation and capacities. Those additional variables are available at different temporal resolutions or time frames on Open Power System Data platform. The need for real-time inference under complicated models and high computing costs reinforce the necessity of data reduction methods in analyzing IoT sensor streams.

In terms of estimation accuracy, LSS and relaxed-LSS methods behave similarly and outperform the Bernoulli sampling most of the time during updates. Overall, estimation errors for all three methods are decreasing along the update time, which suggests that the online estimates converge to the “full sample” estimate. However, in practice, one cannot afford to wait for the “full sample” due to the need for real-time monitoring in the power grid. Thus, given limited access to the data under computational constraint, the faster it converges to the “full sample” estimate the better. Especially, given the same initial values, LSS and relaxed-LSS methods achieve better estimation accuracy at the early stage of updates than those of the Bernoulli method. Eventually, LSS, relaxed-LSS and even Bernoulli sampling estimates are close to the “full sample” estimate since a large enough amount of samples are used for updating in all three sampling methods.

It is worth mentioning that there are a few sudden increases of estimation errors shortly after time at 1000. We believe they are caused by the abnormal points observed in dimension 12 Luxembourg on 2011-03-27 at 22:00:00 UTC and in dimension 13 Macedonia on 2010-03-28 at 01:00:00 UTC. Those abnormal points were high leverage score points and thus

were sampled by the LSS and relaxed-RSS methods. The corresponding model estimates deviated from the “full sample” estimates, which lead to sudden increases in estimation errors but were soon corrected by new data points. This phenomenon reflects the advantage of LSS-based sampling methods in capturing the influential or abnormal data points during stream monitoring. It is an important feature that LSS can be applied in the online monitoring of the dynamic dependence of the power grid system for security or online decision-making purposes. However, the phenomenon also reflects the limitation of the leverage based sampling method in lacking of robustness. The online estimation based on leverage score sampling is sensitive to the changes in underlying data generating processes or outliers.

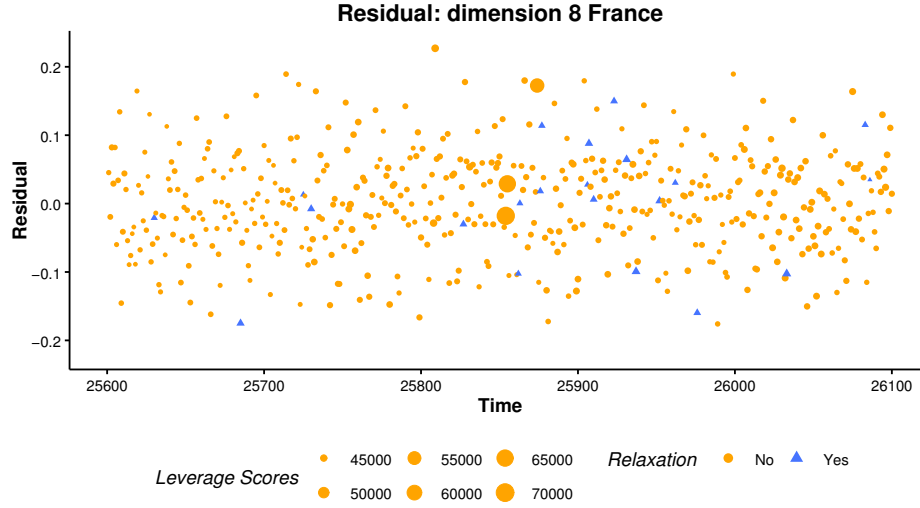


Fig 5: Residual plot from regression fits for France, which is calculated based on the final estimate of relaxed-LSS along all historical data. A filled circle denotes a point selected by the leverage score thresholding rule with the size of the filled circle indicating the leverage score, and a triangle point indicates a point selected by the Bernoulli sampling rule.

In addition to theoretically verified properties of LSS-based methods in parameter estimation, we also consider the prediction as a practical performance measurement. The accurate prediction of the electricity loads is crucial for power grid system in resource allocation and demand management. Prediction is also an important task for general IoT sensor data stream analysis due to similar reasons. For prediction accuracy, the relaxed-LSS method consistently achieves the smallest prediction errors among the three compared methods; while Bernoulli method always delivers the largest prediction errors. The superiority of relaxed-LSS over LSS in prediction may be due to its inclusion of low-leverage covariate points. We also note the two peaks in prediction errors before and after time 30000, which are also caused by the abnormal points in Macedonia and Luxembourg.

Lastly, to examine the goodness of model fitting with relaxed-LSS sampling, we display the regression residuals from a segment of a selected sample of size 500 from France in Figure 5, which are calculated based on the final estimate of relaxed-LSS on all historical data. In Figure 5, a filled circle denotes a point selected by the leverage score thresholding rule with the size of the filled circle indicating the leverage score, and a triangle indicates the low leverage score point. The residual plot shows that the linear relation proposed in the relaxed-LSS sampler holds equally well for both the points selected by a high leverage score and by Bernoulli sampling.

6. Simulation Studies for general applicability. Although the proposed relaxed-LSS demonstrates better performance in inference and prediction on Open Power System Data, we also conduct simulation studies to evaluate the effectiveness of LSS and its relaxed version in general settings. We compare the LSS, relaxed-LSS and the Bernoulli sampling method, which corresponds to $s(x) \equiv q$ in Definition 3.1. We generate the multi-dimensional time series which follows the VARX model defined in (5) with $K = 10$, $p_1 = 1$, and $p_2 = 1$. We consider the multivariate Gaussian and the multivariate StudentT distributions for noise process (e_t) and extraneous process (v_t) .

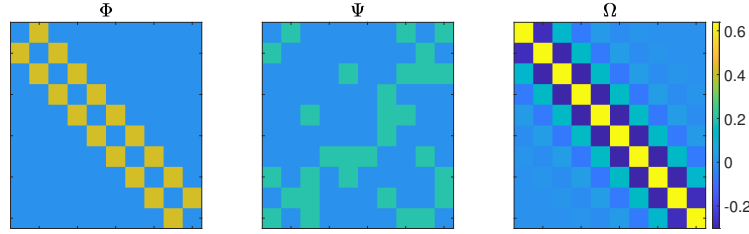


Fig 6: Visualization of the coefficient matrices Φ and Ψ , and covariance matrix for the error process Ω .

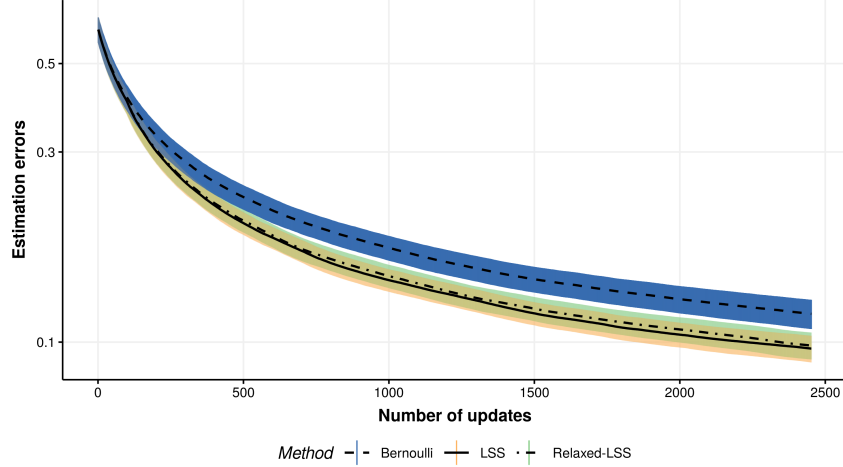


Fig 7: Comparison of estimation errors for simulated 10 dimensional time series under Gaussian setting. Visualized for the first 2,500 updates with underlying time series of length 25,000. The average and one standard deviation error bar of the estimation errors are plotted over 1,000 independent replicates.

6.1. Multivariate Gaussian Distribution Case. We first consider the Gaussian case where the noise process (e_t) follows the i.i.d. multivariate Gaussian distribution $N(\mathbf{0}, \Omega)$ and the extraneous process (v_t) follows i.i.d. multivariate Gaussian distribution $N(\mathbf{0}, I_{Kp_2})$,

where $I_{Kp_2} \in \mathbb{R}^{Kp_2}$ is the identity matrix. We follow [Qiu et al. \(2015\)](#) to generate the coefficient matrices Φ and Ψ , and the covariance matrix for the error process Ω , see Figure 6 for visualization.

We generate the 10-dimensional time series of length $n = 25,000$. The sampling rate is $q = 0.1$ and base sampling rate for relaxed-LSS is $q_0 = 0.05$. The update rate for the inverse covariance matrix is 0.1. A pilot sample of size 100 is used to calculate initial estimations. We define the estimation error as $\|\hat{B}_\tau - B\|_F / \|B\|_F$, where $\hat{B}_\tau$ is the τ th update of the estimation of model parameter matrix B (here we use τ to denote the number of updates to distinguish from the notation of time point t) and $\|\cdot\|_F$ denotes the Frobenius norm.

Figure 7 displays the average and one standard deviation error bar of the estimation errors at each update for each of the compared methods over 1,000 independent replicates. Focusing on the first 2,500 updates for each of the compared methods, we observe that LSS method achieve smallest estimation errors compared to relaxed-LSS and Bernoulli methods as predicted by our optimality theory. We note that the estimation errors for both LSS and relaxed-LSS are significantly smaller than those of Bernoulli method.

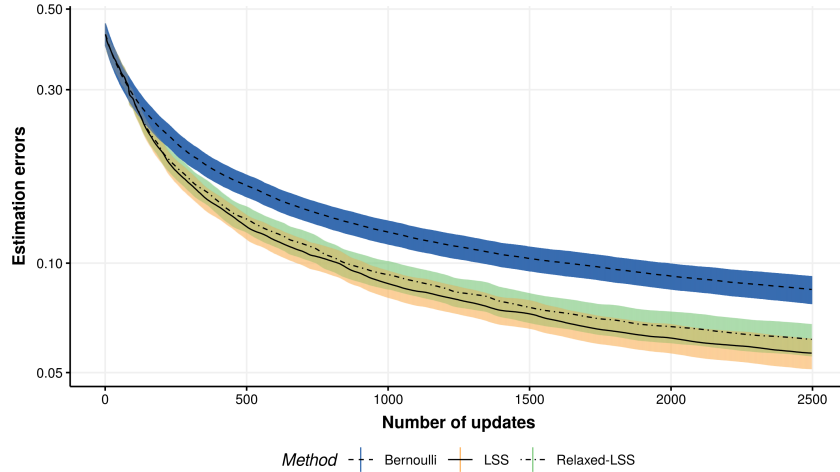


Fig 8: Comparison of estimation errors for simulated 10 dimensional time series under `StudentT` setting. Visualized for the first 2,500 updates with underlying time series of length 25,000. The average and one standard deviation error bar of the estimation errors are plotted over 1,000 independent replicates.

6.2. Multivariate Student T Distribution Case. We next consider the `StudentT` case. The noise process (e_t) follows the i.i.d. multivariate student t distribution with location $\mathbf{0}$, degree of freedom 3 and scale matrix Ω , and the extraneous process (v_t) follows i.i.d. multivariate student t distribution with location $\mathbf{0}$, degree of freedom 3 and scale matrix I_{Kp_2} .

We generate the time series following [Rémillard et al. \(2012\)](#); [Qiu et al. \(2015\)](#). Particularly, we first simulate an initial observation $\mathbf{y}_1$, extraneous variables $\mathbf{v}_1, \dots, \mathbf{v}_{t-1}$, and innovations $\mathbf{e}_2, \dots, \mathbf{e}_t$. After generating $(\mathbf{y}'_1, \mathbf{e}'_2, \dots, \mathbf{e}'_t)'$ and $(\mathbf{v}'_1, \dots, \mathbf{v}'_{t-1})'$, one can form $(\mathbf{y}'_1, \mathbf{y}'_2, \dots, \mathbf{y}'_t)'$ recursively following the iterative algorithm in [Rémillard et al. \(2012\)](#). The model coefficient matrices Φ and Ψ , and the covariance matrix Ω are the same as the `Gaussian` case. We generate the multi-dimensional time series of length $t = 25,000$ and set the

sampling rate $q = 0.1$, the base sampling rate $q_0 = 0.05$ and the update rate for the inverse covariance matrix is 0.1. The pilot sample size is 100.

Figure 8 presents the average and one standard deviation error bar of estimation errors against number of estimate updates, which are based on 1,000 independent replicates. Similar to the Gaussian case, the LSS method achieves the best performance as predicted by our optimality theory; the LSS and relaxed-LSS methods are comparable; both methods are significantly better than the Bernoulli method in terms of estimation errors, where the wider difference compared to the Gaussian case is due to the fact that the t -distribution has a heavier tail and hence generates a larger leverage effect.

7. Conclusion. We introduced a class of online sample selection (sampling) methods for large scale streaming time series with application in the online analysis of electricity power grid data. We provide a solution to online statistical inference of high speed multidimensional time series streams under computational constraint. The proposed methods were motivated by optimal designs in design of experiments and were applied to the high temporal resolution data streams in power grid system as an example of IoT sensor network data stream analysis. The proposed methods were based on a relaxed version of leverage score sampling and achieved an optimality criterion. Therefore, the proposed methods enjoyed the optimality in online sampling theoretically and improved the computational efficiency of the online analysis. The elliptical distributed synthetic data and electricity consumption real data analysis demonstrated the effectiveness of the proposed sampling methods. Our proposed relaxed-LSS method provides online analysis of electricity loads through data selection without loss of identification of electricity consumption patterns and flexibility. Our work was based on the stationary linear multivariate time series models for streaming data modeling, which serves as a foundation for tackling the more involved non-stationary case. We shall leave the study of sampling of non-stationary data to future work.

Acknowledgments. We thank the Editor, Associate Editor, and two anonymous reviewers for many valuable comments and suggestions.

Funding. This work is partially supported by NIH awards R01MD018025, R03AG069799, NSF awards DMS-1903226, DMS-1925066 and DMS-2124493.

REFERENCES

- Agarwal, P. K., S. Har-Peled, and K. R. Varadarajan (2005). Geometric approximation via coresets. *Combinatorial and Computational Geometry* 52, 1–30.
- Akbar, A., A. Khan, F. Carrez, and K. Moessner (2017). Predictive analytics for complex IoT data streams. *IEEE Internet of Things Journal* 4(5), 1571–1582.
- Anagnostopoulos, C., S. Hadjiefthymiades, A. Katsikis, and I. Maglogiannis (2014). Autoregressive energy-efficient context forwarding in wireless sensor networks for pervasive healthcare systems. *Personal and Ubiquitous Computing* 18(1), 101–114.
- Balduin, S., E. Veith, and S. Lehnhoff (2022). Sampling strategies for static powergrid models. *arXiv preprint arXiv:2204.09053*.
- Berberidis, D., V. Kekatos, and G. B. Giannakis (2016). Online censoring for large-scale regressions with application to streaming big data. *IEEE Transactions on Signal Processing* 64(15), 3854–3867.
- Bingham, N. H., C. M. Goldie, and J. L. Teugels (1989). *Regular Variation*, Volume 27. Cambridge university press.
- Box, G. E. P., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2015). *Time Series Analysis: Forecasting and Control* (5 ed.). Wiley.
- Cai, D., D. Shi, and J. Chen (2013). Probabilistic load flow computation with polynomial normal transformation and latin hypercube sampling. *IET generation, transmission & distribution* 7(5), 474–482.
- Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics* 19(1), 15–18.

- Dasgupta, A., P. Drineas, B. Harb, R. Kumar, and M. W. Mahoney (2009). Sampling algorithms and coresets for ℓ_p regression. *SIAM Journal on Computing* 38(5), 2060–2078.
- Drineas, P., M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff (2012). Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research* 13(1), 3475–3506.
- Eshragh, A., F. Roosta, A. Nazari, and M. W. Mahoney (2022). LSAR: Efficient leverage score sampling algorithm for the analysis of big time series data. *Journal of Machine Learning Research* 23, 1–36.
- Fang, K.-T., S. Kotz, and K. W. Ng (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall.
- Feldman, D., M. Schmidt, and C. Sohler (2013). 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*, pp. 1434–1453. Society for Industrial and Applied Mathematics.
- Gabel, M., D. Keren, and A. Schuster (2015). Monitoring least squares models of distributed streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 319–328. ACM.
- Gittens, A. and M. Mahoney (2013). Revisiting the nystrom method for improved large-scale machine learning. In *International Conference on Machine Learning*, pp. 567–575. PMLR.
- Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press. Princeton, NJ.
- Hill, D. J. and B. S. Minsker (2010). Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software* 25(9), 1014–1022.
- Hooi, B., H. A. Song, A. Pandey, M. Jereminov, L. Pileggi, and C. Faloutsos (2018). Streamcast: Fast and online mining of power grid time sequences. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pp. 531–539. SIAM.
- Islam, S. R., D. Kwak, M. H. Kabir, M. Hossain, and K.-S. Kwak (2015). The Internet of things for health care: a comprehensive survey. *IEEE Access* 3, 678–708.
- Jaradat, M., M. Jarrah, A. Boussehlam, Y. Jararweh, and M. Al-Ayyoub (2015). The Internet of energy: smart sensor networks and big data management for smart grid. *Procedia Computer Science* 56, 592–597.
- Jordan, M. I. (2013). On statistics, computation and scalability. *Bernoulli* 19(4), 1378–1390.
- Jumar, R., H. Maaß, B. Schäfer, L. R. Gorjão, and V. Hagenmeyer (2020). Database of power grid frequency measurements. *arXiv preprint arXiv:2006.01771*.
- Kallenberg, O. (2002). *Foundations of Modern Probability* (2 ed.). Springer Science & Business Media.
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* 82(1), 35–45.
- Kalman, R. E. and R. S. Bucy (1961). New results in linear filtering and prediction theory. *Journal of Basic Engineering* 83(1), 95–108.
- Li, F., R. Xie, Z. Wang, L. Guo, J. Ye, P. Ma, and W. Song (2019). Online distributed IoT security monitoring with multidimensional streaming big data. *IEEE Internet of Things Journal* 7(5), 4387–4394.
- Liberty, E. (2013). Simple and deterministic matrix sketching. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 581–588. ACM.
- Liu, J. S. (2004). *Monte Carlo Strategies in Scientific Computing*. Springer.
- Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media.
- Ma, P., M. W. Mahoney, and B. Yu (2015). A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research* 16(1), 861–911.
- Ma, P., X. Zhang, X. Xing, J. Ma, and M. Mahoney (2020). Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 1026–1035. PMLR.
- Marangoni, G. and M. Tavoni (2021). Real-time feedback on electricity consumption: evidence from a field experiment in Italy. *Energy Efficiency* 14(1), 1–17.
- Mat, I., M. R. M. Kassim, A. N. Harun, and I. M. Yusoff (2016). IoT in precision agriculture applications using wireless moisture sensor network. In *2016 IEEE Conference on Open Systems (ICOS)*, pp. 24–29. IEEE.
- Meng, C., R. Xie, A. Mandal, X. Zhang, W. Zhong, and P. Ma (2020). Lowcon: A design-based subsampling approach in a misspecified linear model. *Journal of Computational and Graphical Statistics* 0, 1–32.
- Michalareas, G., J.-M. Schoffelen, G. Paterson, and J. Gross (2013). Investigating causality between interacting brain areas with multivariate autoregressive models of MEG sensor data. *Human Brain Mapping* 34(4), 890–913.
- Nellore, K. and G. P. Hancke (2016). A survey on urban traffic management system using wireless sensor networks. *Sensors* 16(2), 157.
- Open Power System Data (2020). *Data package time series. Version 2020-10-06*. (Primary data from various sources, https://doi.org/10.25832/time_series/2020-10-06).
- Papalambros, P. Y. and D. J. Wilde (2000). *Principles of Optimal Design: Modeling and Computation*. Cambridge University Press.

- Petris, G., S. Petrone, and P. Campagnoli (2009). Dynamic linear models. In *Dynamic Linear Models with R*, pp. 31–84. Springer.
- Plackett, R. L. (1950). Some theorems in least squares. *Biometrika* 37(1/2), 149–157.
- Pronzato, L. (2006). On the sequential construction of optimum bounded designs. *Journal of Statistical Planning and Inference* 136(8), 2783–2804.
- Pronzato, L. and H. Wang (2020). Sequential online subsampling for thinning experimental designs. *arXiv preprint arXiv:2004.00792*.
- Pukelsheim, F. (1993). *Optimal Design of Experiments*, Volume 50. SIAM.
- Qiu, H., S. Xu, F. Han, H. Liu, and B. Caffo (2015). Robust estimation of transition matrices in high dimensional heavy-tailed vector autoregressive processes. In *International Conference on Machine Learning*, pp. 1843–1851.
- Rémillard, B., N. Papageorgiou, and F. Soustra (2012). Copula-based semiparametric models for multivariate time series. *Journal of Multivariate Analysis* 110, 30–42.
- Schimbinschi, F., L. Moreira-Matias, V. X. Nguyen, and J. Bailey (2017). Topology-regularized universal vector autoregression for traffic forecasting in large urban areas. *Expert Systems with Applications* 82, 301–316.
- Seber, G. A. and A. J. Lee (2012). *Linear Regression Analysis*, Volume 329. John Wiley & Sons.
- Shehabi, A., S. Smith, D. Sartor, R. Brown, M. Herrlin, J. Koomey, E. Masanet, N. Horner, I. Azevedo, and W. Lintner (2016). United states data center energy usage report.
- Sherman, J. and W. J. Morrison (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics* 21(1), 124–127.
- Siddik, M. A. B., A. Shehabi, and L. Marston (2021). The environmental footprint of data centers in the united states. *Environmental Research Letters* 16(6), 064017.
- Surgailis, D., H. L. Koul, and L. Giraitis (2012). *Large Sample Inference for Long Memory Processes*. World Scientific Publishing Company.
- Ting, D. and E. Brochu (2018). Optimal subsampling with influence functions. In *Advances in Neural Information Processing Systems*, pp. 3654–3663.
- Wang, H., M. Yang, and J. Stufken (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association* 114(525), 393–405.
- Wang, H., R. Zhu, and P. Ma (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association* 113(522), 829–844.
- Wang, L., J. Elmstedt, W. K. Wong, and H. Xu (2021). Orthogonal subsampling for big data linear regression. *The Annals of Applied Statistics* 15(3), 1273 – 1290.
- West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models* (2nd ed.). Springer Science & Business Media.
- Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science* 10(1–2), 1–157.
- Xie, R., Z. Wang, S. Bai, P. Ma, and W. Zhong (2019). Online decentralized leverage score sampling for streaming multidimensional time series. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2301–2311.
- Xu, X., Y. Chen, Y. Goude, and Q. Yao (2021). Day-ahead probabilistic forecasting for french half-hourly electricity loads and quantiles for curve-to-curve regression. *Applied Energy* 301, 117465.
- Yokoyama, R. (1980). Moment bounds for stationary mixing sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 52(1), 45–57.
- Yu, J., H. Wang, M. Ai, and H. Zhang (2020). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association* (just-accepted), 1–29.
- Zhang, K., C. Liu, J. Zhang, H. Xiong, E. Xing, and J. Ye (2017). Randomization or condensation?: Linear-cost matrix sketching via cascaded compression sampling. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 615–623. ACM.
- Zhang, T. and W. B. Wu (2012). Inference of time-varying regression models. *The Annals of Statistics* 40(3), 1376–1402.
- Zhou, B. and W. Saad (2019). Joint status sampling and updating for minimizing age of information in the internet of things. *IEEE Transactions on Communications* 67(11), 7468–7482.
- Zhou, Z. and W. B. Wu (2010). Simultaneous inference of linear models with time varying coefficients. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(4), 513–531.