Proceedings of Machine Learning Research vol 168:1–12, 2022 4th Annual Conference on Learning for Dynamics and Control

Time varying regression with hidden linear dynamics

Ali Jadbabaie† JADBABAI@MIT.EDU
Horia Mania HMANIA@MIT.EDU
Devavrat Shah DEVAVRAT@MIT.EDU
Suvrit Sra SUVRIT@MIT.EDU
LIDS, Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Editors: R. Firoozi, N. Mehr, E. Yel, R. Antonova, J. Bohg, M. Schwager, M. Kochenderfer

Abstract

We revisit a model for time-varying linear regression that assumes the unknown parameters evolve according to a linear dynamical system. Counterintuitively, we show that when the underlying dynamics are stable the parameters of this model can be estimated from data by combining just two ordinary least squares estimates. We offer a finite sample guarantee on the estimation error of our method and discuss certain advantages it has over Expectation-Maximization (EM), which is the main approach proposed by prior work.

Keywords: Linear dynamical systems, time series, time-varying regression, system identification

1. Introduction

The distribution of labels given the covariates changes over time in a variety of applications of regression. Some example domains where such problems arise include economics, marketing, fashion, and supply chain optimization, where market properties evolve over time. Motivated by such problems, we revisit a model for time-varying linear regression that assumes the unknown parameters evolve according to a linear dynamical system.

One way to account for distribution change in linear regression is to assume that the unknown model parameters change slowly with time Bamieh and Giarre (2002); Kalaba and Tesfatsion (1989); Zhang et al. (2012). While this assumption simplifies the problem and makes it tractable, it misses on exploiting additional structure available and it also fails to model periodicity (e.g., due to seasonality) present in some problems. As an alternative, we are interested in a dynamic model previously studied by Chow (1981), Carraro (1984), and Shumway et al. (1988). Given data $\{(x_t, y_t)\}_{t=0}^{T-1}$ we assume the label y_t is a linear function of the features x_t with unknown parameters β_t that evolve according to linear dynamics:

$$\begin{aligned} \beta_{t+1} &= A_\star \beta_t + w_t, \\ y_t &= x_t^\top \beta_t + \epsilon_t, \end{aligned} \tag{1}$$

where w_t is *process noise* and ϵ_t is *observation noise*. The hidden states β_t are not observed and the parameter matrix $A_\star$ is unknown. In this work the features x_t are d-dimensional vectors and the observations y_t are scalar, but our results can be extended to vector valued observations. In some

. † The author names appear in the alphabetical order of their last names.

applications, such as modeling economic trends, one may be interested in recovering the unknown parameters $A_\star$, which would offer insights into long term trends. When $A_\star$ and the distributions of w_t and ϵ_t are known prediction is also possible via the Kalman filter. Therefore, our main goal is to estimate $A_\star$ from data $\{(x_t, y_t)\}_{t=0}^{T-1}$.

Estimating $A_\star$ in (1) is a system identification problem of an unactuated linear system that has a time-varying observation map. Up until now there has been no simple estimator for this problem that has a finite sample guarantee. Counterintuitively, we show that when $A_\star$ is stable it can be estimated from data by combining just two ordinary least squares estimates, one that regresses y_t^2 on $x_t x_t^\top$ and one that regresses $y_t y_{t+1}$ on $x_{t+1} x_t^\top$. When the hidden dynamics are stable these two regressions estimate the covariances $\Sigma_\infty := \mathbb{E}\beta_t\beta_t^\top$ and $A_\star\Sigma_\infty = \mathbb{E}\beta_{t+1}\beta_t^\top$, from which we can obtain an estimate of $A_\star$. For this reason we call this approach the covariance method (CM).

In addition to its simplicity, CM also admits a finite sample guarantee on the estimation error. In contrast, the statistical performance of Expectation-Maximization (EM), which is a classical estimation method used by prior work Shumway and Stoffer (1982), is not well understood theoretically. We prove the following guarantee for CM.

Main result (informal) *Let Σ_w be the covariance of w_t, σ_ϵ be the standard deviation of ϵ_t, and d be dimension of x_t. Suppose $A_\star$ has spectral radius $\rho = \rho(A_\star)$ smaller than one. Then, given a trajectory of length T, CM produces with probability at least $1-\delta$ an estimate $\widehat{A}$ such that*

$$\|\widehat{A} - A_\star\|_F \lesssim \frac{1 + \|A_\star\|}{\lambda_{\min}(\Sigma_\infty)} \sqrt{\frac{d^4}{T\delta}\left(\sigma_\epsilon^4 + \frac{\|\Sigma_\infty\|^2}{1-\rho^2}\right)}.$$

This results shows that our method estimates $A_\star$ at a $\mathcal{O}(T^{-1/2})$ rate without the need of a good initialization as in the case of EM. Some dependence on $\lambda_{\min}(\Sigma_\infty)$ is to be expected because when Σ_∞ is rank deficient the states β_t do not span all of $\mathbb{R}^d$, in which case no method would be able to identify $A_\star$ fully. Unsurprisingly, the upper bound on the estimation error blows up as ρ approaches one. The main idea of CM is to estimate Σ_∞, which itself blows up as ρ approaches one.

We discuss CM in detail and the formal statement of our main result in Section 2. In Section 3 we present a meta result that highlights the main idea of our analysis. Then, in order to ground the comparison between our method and EM in Section 5 we discuss background on stable linear systems and their identification in the fully observed case, followed in Section 6 by a discussion of EM. We conclude with a discussion of some open questions.

1.1. Notation.

Before moving forward we introduce some useful notation. We use $y_{s:t}$ to denote the sequence of observations $\{y_s, y_{s+1}, \ldots, y_{t-1}\}$, with a similar notation for x and β. The bold symbols $\mathbf{x}$, $\mathbf{y}$, and $\boldsymbol{\beta}$ are used to denote the entire trajectories $x_{0:T}$, $y_{0:T}$, and $\beta_{0:T}$. The norm $\|\cdot\|$ denotes the ℓ_2 norm for vectors and the operator norm for matrices. We use $\|\cdot\|_F$ to denote the Frobenius norm. We use c to denote universal constants that may change from line to line.

1.2. What makes problem (1) challenging?

Classical methods for linear system identification, such as the eigensystem realization algorithm Juang and Pappa (1985) and methods based on ordinary least squares Sarkar et al. (2021); Simchowitz et al. (2019); Tsiamis and Pappas (2019), were designed for the identification of time invariant systems of the form:

$$\beta_{t+1} = A_\star \beta_t + B_\star u_t + w_t \quad (2)$$
$$y_t = C_\star \beta_t + \epsilon_t. \quad (3)$$

Two factors make the estimation of such systems easier: the time invariance and the actuation of the system through u_t. We discuss why either of these two factors make the estimation easier.

Time-invariant and actuated systems. When the dynamics are actuated and time invariant one can regress y_t on a history of inputs $(u_{t-1}, u_{t-2}, \ldots)$ in order to estimate the Markov parameters $C_\star A_\star^j B_\star$ with $j \in \{0, 1, \ldots, r\}$ for some r. Then, one can apply the Ho-Kalman algorithm to extract estimates of $A_\star$, $B_\star$, and $C_\star$ from the estimated Markov parameters. Sarkar et al. (2021) and Simchowitz et al. (2019) use this high-level approach to offer guarantees on the estimation of systems of the form (3). The success of such an estimation strategy is guaranteed by Oymak and Ozay (2021), who bounded the error of the estimates produced by the Ho-Kalman algorithm as a function of the estimation error of the Markov parameters.

Time-varying and actuated systems. In our setting the "$C_\star$" matrix (observer matrix) is time-varying which implies time-varying Markov parameters. Since the time-varying observer matrices are assumed known in our setting, if we were able to actuate the dynamical system, we could estimate the Markov parameters $A_\star^j B_\star$ and then use the Ho-Kalman algorithm to recover the $A_\star$ and $B_\star$ matrices. However, (1) is unactuated.

Moreover, Majji et al. (2010) showed that if one can collect multiple trajectories from the time-varying dynamics by restarting the system, then one can recover all A_t, B_t, C_t. If only one trajectory were available and A_t, B_t, and C_t were time-varying and unknown, then identification would impossible because we would have only one data point for each time step.

Time-invariant and unactuated systems. This setting corresponds to having $B_\star = 0$ in (3) and can be handled by regressing y_{t+1} on a history of past observations $y_t, y_{t-1}, \ldots y_{t-h}$ for some h. This approach is successful because one can show that there exists a matrix M such that

$$y_{t+1} = M \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-h} \end{bmatrix} + \zeta_t,$$

for some small ζ_t Tsiamis and Pappas (2019). However, in the time-varying setting there is no time-invariant matrix M that satisfies this identity. Even if feasible, such an approach would not use the fact that the observer matrices $x_t^\top$ are known.

Time-varying and unactuated systems. Our problem is both time-varying and unactuated and the methods previously discussed cannot be extended easily to address it. As discussed previously, in the time-invariant and unactuated case a simple estimator can solve the problem when both $A_\star$ and $C_\star$ are unknown. While our problem is time-varying, the observer matrices are given. Hence, only $A_\star$ must be estimated. Our problem has fewer unknown parameters than in the time-invariant case and yet the methods discussed thus far are not applicable.

2. The covariance method (CM)

In this section we propose and analyze a method for estimating $A_\star$ that first estimates the covariance of the hidden states β_t sampled from the stationary distribution. To estimate the covariance matrix our method applies ordinary least squares to regress y_t^2 on $x_t x_t^\top$. Interestingly, the need to estimate the covariance of random vectors that can only be observed through linear measurements arises in the cryo-EM heterogeneity problem Katsevich et al. (2015). We employ the same solution to this covariance estimation problem and offer a finite sample guarantee for this approach.

Assumptions. In order to estimate the covariance of the stationary distribution we need the stationary distribution to exist and have a finite covariance. The dynamics $\beta_{t+1} = A_\star \beta_t + w_t$ have a stationary distribution with finite covariance if and only if the matrix $A_\star$ is stable (i.e., $\rho(A_\star) < 1$). We also assume the covariates x_t are sampled i.i.d. according to some distribution. For simplicity we assume $x_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, but our analysis can be generalized to features with a sub-Gaussian distribution and a more general covariance structure. We also assume $w_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_w)$ and $\epsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$, but our analysis can be extended to any independent noise processes with bounded fourth moments.

When $A_\star$ is stable, the linear system mixes exponentially fast to a distribution with covariance Σ_∞ (the distribution is Gaussian when w_t are Gaussian), where Σ_∞ is the unique positive semidefinite solution P of the Lyapunov equation $P = APA^\top + \Sigma_w$. Given that the convergence to stationarity is fast we also assume that the initial state β_0 is sampled from the stationary distribution $\mathcal{N}(0, \Sigma_\infty)$. Hence, the marginal distribution of all β_t is $\mathcal{N}(0, \Sigma_\infty)$.

The method. Given our assumptions, we observe that there exist random variables ζ_t and ξ_t whose means are zero conditional on x_t and x_{t+1} and that satisfy the identities:

$$y_t^2 = x_t^\top \Sigma_\infty x_t + \sigma_\epsilon^2 + \zeta_t \quad \text{and} \quad y_t y_{t+1} = x_{t+1}^\top A \Sigma_\infty x_t + \xi_t, \tag{4}$$

where the noise terms ζ_t and ξ_t are defined by

$$\begin{aligned}
\zeta_t &= x_t^\top (\beta_t \beta_t^\top - \Sigma_\infty) x_t + 2\epsilon_t x_t^\top \beta_t + (\epsilon^2 - \sigma_\epsilon^2), \\
\xi_t &= x_{t+1}^\top A \left(\beta_t \beta_t^\top - \Sigma_\infty\right) x_t + \epsilon_{t+1} x_t^\top \beta_t + x_t^\top \beta_t w_t^\top x_{t+1} + \epsilon_t x_{t+1}^\top A \beta_t + \epsilon_t \epsilon_{t+1} + \epsilon_t w_t^\top x_{t+1}.
\end{aligned}$$

Since Σ_∞ and $A\Sigma_\infty$ enter (4) linearly, we estimate them with ordinary least squares. Concretely, our method performs the following steps[1]:

$$\begin{aligned}
\widehat{\Sigma}_\infty &\in \operatorname*{argmin}_M \sum_{t=0}^{T-1} (y_t^2 - x_t^\top M x_t - \sigma_\epsilon^2)^2, \\
\widehat{M} &\in \operatorname*{argmin}_M \sum_{t=0}^{T-1} (y_t y_{t+1} - x_t^\top M x_{t+1})^2, \\
\widehat{A} &= \widehat{M} \widehat{\Sigma}_\infty^{-1}.
\end{aligned} \tag{5}$$

To offer a guarantee on the estimation error $\widehat{A} - A_\star$ we first analyze the error in the estimation of Σ_∞ and $A\Sigma_\infty$. It is known that ordinary least squares is a consistent estimator for the recovery of covariance matrices from squared linear measurements such as (4) when the measurements are

1. For the analysis we assume σ_ϵ^2 is given, but in practice one would run OLS with an intercept also to account for σ_ϵ^2.

i.i.d. Katsevich et al. (2015). However, in our case the measurements y_t are not independent. Moreover, we wish to quantify the estimation rate. Analyzing the performance of OLS applied to (4) is challenging because ζ_t and ξ_t are heavy tailed, they are dependent on the covariates x_t, and they are dependent across time due to the dynamics. By using Chebyshev's inequality we circumvent these issues, yielding a medium probability guarantee.

Before we state our main result we review a consequence of Gelfand's formula for stable matrices. For any stable matrix $A_\star$ and any $\gamma \in (\rho(A_\star), 1)$ there exists a finite τ such that $\|A_\star^k\| \leq \tau\gamma^k$ for all $k \geq 0$. We denote $\tau(A_\star, \gamma) := \sup\{\|A_\star^k\|\gamma^{-k} : k \geq 0\}$ (see Tu et al., 2017, for more details).

Theorem 1 *Suppose* $x_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$, $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_w)$, *and* $\epsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. *Moreover, assume* $\rho(A_\star) < 1$ *and let* $\gamma \in (\rho(A_\star), 1)$. *Then, with probability* $1 - \delta$ *the covariance method produces estimates* $\widehat{\Sigma}_\infty$ *and* $\widehat{M}$ *that satisfy*

$$\|\widehat{\Sigma}_\infty - \Sigma_\infty\|_F \leq \sqrt{\frac{cd^2}{T\delta}\left[\sigma_\epsilon^4 + \frac{\tau(A_\star,\gamma)^2}{1-\gamma^2}\|\Sigma_\infty\|^2\left(d^2 + \log\left(\frac{2T}{\delta}\right)^2\right)\right]}, \tag{6a}$$

$$\|\widehat{M} - A\Sigma_\infty\|_F \leq \sqrt{\frac{cd^2}{T\delta}\left[\sigma_\epsilon^4 + \frac{\tau(A_\star,\gamma)^3}{1-\gamma^2}\|\Sigma_\infty\|^2\left(d^2 + \log\left(\frac{2T}{\delta}\right)^2\right)\right]}, \tag{6b}$$

whenever $T \geq cd^4 \log(c\frac{d^2}{\delta})$. *The term* c *a universal constant that changes from line to line (the only difference between the two upper bounds is the exponent of* τ*).*

The proof of this result is based on the meta-analysis for linear regression with heavy tailed and dependent measurements presented in Section 3. Theorem 1 shows that as the number T of data points collected grows the estimation error decays at a rate $\mathcal{O}(T^{-1/2})$. Theorem 1 also suggests that the closer to instability the matrix $A_\star$ is the more data our method requires to recover Σ_∞ and $A_\star\Sigma_\infty$. This relationship is unsurprising since Σ_∞ and the variances of ζ_t and ξ_t in (4) approach infinity as $\rho(A_\star)$ approaches one. Nonetheless, in the case of time-invariant linear dynamical systems with partial observations it is possible to estimate $A_\star$ even when it is marginally unstable (i.e. $\rho(A_\star) = 1$) Simchowitz et al. (2019); Tsiamis and Pappas (2019).

Despite its merits, Theorem 1 is likely not capturing the right dependence of problem (1) on the state dimension d, the probability of failure δ, and the variances σ_ϵ^2 and Σ_∞. For example, we are trying to estimate d^2 unknown parameters so we expect to need a number of scalar measurements that scales as d^2, not as d^4. We obtain a $1/\delta$ dependence on δ instead of $\log(1/\delta)$ because the proof relies on Chebyshev's inequality. It is not clear whether the sub-optimal dependencies on problem parameters represent an artifact of our analysis or a limitation of our method.

Our sample complexity guarantee on the estimation of A is a simple corollary of Theorem 1.

Corollary 2 *Under the same assumptions as in Theorem 1, as long as* T *satisfies*

$$T \geq \frac{cd^2}{\lambda_{\min}(\Sigma_\infty)^2\delta}\left[\sigma_\epsilon^4 + \frac{\tau(A_\star,\gamma)^3}{1-\gamma^2}\|\Sigma_\infty\|^2\left(d^2 + 4\log\left(\frac{2T}{\delta}\right)^2\right)\right],$$

the covariance method outputs with probability at least $1 - \delta$ *an estimate* $\widehat{A}$ *that satisfies:*

$$\|\widehat{A} - A_\star\|_F \leq c\frac{1 + \|A_\star\|}{\lambda_{\min}(\Sigma_\infty)}\sqrt{\frac{cd^2}{T\delta}\left[\sigma_\epsilon^4 + \frac{\tau(A_\star,\gamma)^3}{1-\gamma^2}\|\Sigma_\infty\|^2\left(d^2 + 4\log\left(\frac{2T}{\delta}\right)^2\right)\right]}. \tag{7}$$

Therefore, we have shown that it is possible to estimate $A_\star$ at a rate $\mathcal{O}(T^{-1/2})$ without the need of a good initialization as in the case of EM. This corollary inherits all the limitations of Theorem 1. The dependence on the inverse of $\lambda_{\min}(\Sigma_\infty)$ may seem sub-optimal, but it does capture an important aspect of the problem. When Σ_∞ is smaller in a PSD sense it means that the states β_t have smaller norms, which implies that the signal-to-noise ration in the observations $y_t = x_t^\top \beta_t + \epsilon_t$ is lower. In particular, if $\lambda_{\min}(\Sigma_\infty) = 0$, the hidden states β_t lie in a subspace of $\mathbb{R}^d$ of dimension smaller than d, in which case $A_\star$ cannot be fully recovered.

3. A meta-analysis of linear regression with heavy-tailed and dependent noise

In this section we offer a meta-analysis for linear regression that encompasses both the estimation of Σ_∞ and $A_\star\Sigma_\infty$. Suppose we wish to estimate θ from data $\{(\phi_t, z_t)\}_{t=1}^T$, with $z_t = \phi_t^\top \theta + \zeta_t$, where ϕ_t are given features and ζ_t is noise. However, we do not assume ζ_t and ϕ_t are independent nor do we assume the noise terms ζ_t are independent across time. Moreover, we allow ζ_t to be heavy tailed. In this case, standard analyses of linear regression do not apply. To surmount these challenges and prove the following lemma we rely on Chebyshev's inequality.

Lemma 3 *Suppose $\phi_t, \theta \in \mathbb{R}^k$ and that $\sum_{t=1}^T \phi_t \phi_t^\top$ is invertible. Also, suppose $\mathbb{E}[\zeta_t|\phi_{1:T}] = 0$, $\mathbb{E}[\zeta_t^2|\phi_{1:T}] \leq f_1(\phi_{1:T})$ and $\mathbb{E}[\zeta_t\zeta_{t+h}|\phi_{1:t}]] \leq \gamma^h f_2(\phi_{1:T})$ for some $\gamma \in (0,1)$ and functions f_1, f_2. Then, with probability $1-\delta$ the OLS estimate $\widehat{\theta}$ satisfies*

$$\|\widehat{\theta} - \theta\| \leq \sqrt{\frac{k}{\delta}\left(f_1(\phi_{1:T}) + \frac{f_2(\phi_{1:T})}{1-\gamma}\right)\left(\sum_{t=1}^T \phi_t \phi_t^\top\right)^{-1}}.$$

Proof The OLS estimate $\widehat{\theta}$ is given by $\left(\sum_{t=0}^{T-1} \phi_t \phi_t^\top\right)^{-1}\left(\sum_{t=0}^{T-1} \phi_t z_t\right)$. Since $z_t = \phi_t^\top \theta + \zeta_t$, the estimation error can be written as

$$e := \widehat{\theta} - \theta = \left(\sum_{t=0}^{T-1} \phi_t \phi_t^\top\right)^{-1}\left(\sum_{t=0}^{T-1} \phi_t \zeta_t\right).$$

It is convenient to use Chebyshev's inequality to upper bound $\|e\|$ because it only requires an upper bound on the conditional variance of the noise, conditioned on the covariates ϕ_t. Let $\Sigma_{e|x} := \mathbb{E}\left[ee^\top|\phi_{1:T}\right]$. Then, Chebyshev's inequality applied to e yields

$$\mathbb{P}\left(\sqrt{e^\top \Sigma_{e|x}^{-1} e} > z\right) \leq \frac{k}{z^2}. \tag{8}$$

Therefore, we wish to upper bound $\Sigma_{e|x}$ with high probability. Note that

$$\Sigma_{e|x} = \left(\sum_{t=0}^{T-1} \phi_t \phi_t^\top\right)^{-1}\left(\sum_{i,j=0}^{T-1} \mathbb{E}\left[\zeta_i\zeta_j|\phi_{1:T}\right]\phi_i\phi_j^\top\right)\left(\sum_{t=0}^{T-1} \phi_t \phi_t^\top\right)^{-1} \tag{9}$$

To upper bound $\mathbb{E}\left[\zeta_i\zeta_j|\phi_{1:T}\right]\phi_i\phi_j^\top$ note that for any vectors u and v we have $uv^\top + vu^\top \preceq uu^\top + vv^\top$. Then, using our assumptions on $\mathbb{E}\left[\zeta_i\zeta_j|\phi_{1:T}\right]$ yields

$$\Sigma_{e|x} \preceq \left(f_1(\phi_{1:T}) + \frac{f_2(\phi_{1:T})}{1-\gamma}\right)\left(\sum_{t=1}^T \phi_t \phi_t^\top\right)^{-1}. \tag{10}$$

The conclusion follows by putting together (8) and (10). ■

4. Proof sketch for Theorem 1

The proofs for both (6a) and (6b) rely on Lemma 3, where ϕ_t represents the vectorization of $x_t x_t^\top$ and $x_{t+1}x_t^\top$ respectively. The noise terms are ζ_t and ξ_t defined in Section 2.

The assumptions of Theorem 1 imply that all the conditions required by Lemma 3 are satisfied. We are left to prove that $\lambda_{\min}\left(\sum_{t=1}^{T} \phi_t \phi_t^\top\right) \geq cT$ and to upper bound the conditional expectations of $\zeta_t \zeta_{t+k}$ and $\xi_t \xi_{t+k}$. Since the vectors ϕ_t are sub-exponential one can use standard techniques to prove a lower bound on the minimum eigenvalue Tropp (2015); Wainwright (2019).

To upper bound the conditional correlations of the noise terms we note that the state β_{t+k} can be decomposed as

$$\beta_{t+k} = A_\star^k \beta_t + \underbrace{\sum_{j=1}^{k} A_\star^{j-1} w_{t+h-j}}_{u} . \tag{11}$$

Then, we recall that for some $\gamma \in (\rho(A_\star), 1)$ there exists a finite $\tau(A_\star, \gamma)$ such that $\|A_\star^k\| \leq \tau(A_\star, \gamma)\gamma^k$ for all k. Then, the term $A_\star^k \beta_t$ decays exponentially with k. The term u is independent of β_t by assumption. It follows that β_{t+k} and β_t are approximately independent and one can exploit this fact to upper bound the terms $f_1(\phi_{0:T})$ and $f_2(\phi_{0:T})$ and complete the proof of Theorem 1. The complete proof can be found at `https://hmania.github.io/dynamic_LR.pdf`.

5. Linear system identification with known states

We ground our discussion of CM and EM by looking at the case when the states β_t are given. In the fully observed case $A_\star$ can be estimated by the ordinary least squares (OLS) estimate

$$\widehat{A} = \left(\sum_{t=0}^{T-1} \beta_{t+1}\beta_t^\top\right)\left(\sum_{t=0}^{T-1} \beta_t \beta_t^\top\right)^{-1} .$$

In the partially observed case we cannot compute $\beta_{t+1}\beta_t^\top$ and $\beta_t\beta_t^\top$ directly. Nonetheless, EM and CM get a handle on these quantities in different ways. For example, EM uses the available data and an initial guess of the unknown parameters $A_\star$ to compute conditional expectations of $\beta_{t+1}\beta_t^\top$ and $\beta_t\beta_t^\top$. Then, EM uses these conditional expectations to update it's estimate of $A_\star$, repeating these steps until convergence. When the dynamics are stable the averages $\frac{1}{T}\sum_{t=0}^{T-1} \beta_t\beta_t^\top$ and $\frac{1}{T}\sum_{t=0}^{T-1} \beta_{t+1}\beta_t^\top$ converge to Σ_∞ and $A_\star\Sigma_\infty$ respectively. Despite not having access to the hidden states, CM estimates $A_\star\Sigma_\infty$ and Σ_∞.

Recent results have elucidated the sample complexity of estimating $A_\star$ when the states are given. Generally, theoretical results show that the sample complexity of estimating a Markov chain increases with the mixing time of the chain. These analyses exploit the fact that data points that occur with sufficient time delay are approximately independent. Then, they claim that the effective number of data points is equal to the total number divided by the mixing time. However, in the case of linear dynamical systems this line of reasoning does not capture the true relationship between sample complexity and mixing time. Simchowitz et al. (2018) showed that given a trajectory produced by the dynamics $\beta_{t+1} = A_\star \beta_t + w_t$ with $\rho(A_\star) \leq 1$ one can use the ordinary least squares estimator $\widehat{A} = \operatorname{argmin}_A \sum_{t=0}^{T-1} \|A\beta_t - \beta_{t+1}\|^2$ to obtain an estimate $\widehat{A}$ that, with probability at least $1-\delta$, has the following guarantee on the estimation error (informally):

$$\|\widehat{A} - A_\star\| \leq c\sqrt{\frac{d\log(T/\delta)}{T\lambda_{\min}(\sum_{t=0}^{T} A_\star (A_\star^\top)^t)}}, \tag{12}$$

where c is a universal constant. This result is counterintuitive because it applies to all marginally unstable systems (i.e., those with $\rho(A_\star) \leq 1$) and, furthermore, it implies that when all of $A_\star$'s eigenvalues have magnitude one, the estimation rate is $\widetilde{\mathcal{O}}\left(\frac{\sqrt{d}}{T}\right)$ instead of the usual $\widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{T}}\right)$. In other words, in some situations, as the mixing time of the system increases, the estimation rate improves. For scalar linear systems, this property has been know since the work of White (1958). Sarkar and Rakhlin (2019) proved such a guarantee for a class of unstable linear systems and Simchowitz et al. (2019) derived a similar guarantee for partially observable systems.

Therefore, when the state is directly observable a less stable system is easier to estimate. It remains an open question to determine whether the same phenomenon occurs for (1). Douc et al. (2011) showed that the maximum likelihood estimate (MLE) is consistent for the estimation of a class of Markov chains that contains stable partially observed linear systems. However, to the best of the authors' knowledge, proving that the MLE is consistent for the estimation of systems that are linear, marginally unstable, and partially observed is an open question.

6. Expectation-Maximization (EM)

The classical expectation-maximization method (EM) that has already been applied to (1) Shumway and Stoffer (1982). While EM performs well for certain instances of problem (1), it has several drawbacks that motivated us to search for a different estimator.

EM is difficult to analyze theoretically and it can converge to a local maximizer of the likelihood when it is not initialized close enough to a global maximizer, leading to a large estimation error Balakrishnan et al. (2017). Furthermore, in our experience a straightforward implementation of EM is numerically unstable when applied to (1) with $x_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, a problem that does not occur with a constant x_t. Finally, EM needs to know something about the distributions of the process and observation noise processes. Classically, the noise processes are assumed Gaussian.

EM approximately maximizes the likelihood of a probability model that depends on unobserved latent variables. In fact, EM was one of the first estimation methods used to tackle problem (1) Shumway and Stoffer (1982). The maximum likelihood estimate A^{MLE}, as the name suggests, is the maximizer of the likelihood $\mathbb{P}_A(\mathbf{x}, \mathbf{y})$. The likelihood is a nonconvex function of A and it is not known how to provably find A^{MLE}. EM is a local alternating maximization algorithm that is guaranteed to converge to a local maximizer of the likelihood.

EM starts with an initial guess of $A_\star$ and alternates between two steps. In the E-step it constructs a lower bound on the likelihood of the data $\{(x_t, y_t)\}_{t=0}^{T-1}$ by computing the conditional covariances $S_t = \mathbb{E}[\beta_t\beta_t^\top | \mathbf{x}, \mathbf{y}, A']$ and $S_{t,t-1} = \mathbb{E}[\beta_{t+1}\beta_t^\top | \mathbf{x}, \mathbf{y}, A']$. To compute S_t and $S_{t,t-1}$ EM uses the Kalman filter to compute the conditional expectations with respect to past data: $\mathbb{E}[\beta_t | x_{0:t}, y_{0:t}, A']$, $\mathbb{E}[\beta_t\beta_t^\top | x_{0:t}, y_{0:t}, A']$, and $\mathbb{E}[\beta_{t+1}\beta_t^\top | x_{0:t}, y_{0:t}, A']$, followed by a backwards pass over the data to smooth these conditional expectations. In the M-step, EM simply updates the current guess of $A_\star$ to $\left(\sum_{t=1}^{T-1} S_{t,t-1}\right)\left(\sum_{t=1}^{T-1} S_{t-1}\right)^{-1}$. The detailed steps are shown by Shumway and Stoffer (2000).

Analyses of EM are difficult even when presented with i.i.d. data. While one could follow the analysis of Balakrishnan et al. (2017) to guarantee that EM converges to A^{MLE} if initialized close enough, we believe this approach would lead to a result that suggests EM performs poorly when $A_\star$ is close to being unstable. However, given what we know from the performance of the maximum likelihood estimate (the OLS estimate) in the fully observed case, it is likely that EM would perform better when $A_\star$ has eigenvalues close to the unit circle. In fact, EM operates by computing the best guess of the hidden states β_t given the observed data. Therefore, it is likely that less stable systems offer an additional benefit to EM. When $A_\star$ has larger eigenvalues the dynamics have longer memory, which allows for a better estimation of the hidden states. For example, if $A_\star$ was the zero matrix, then the dynamics would have no memory and the observations y_{t+1} and y_{t-1} would not hold any useful information for inferring β_t.

In contrast, the performance of CM degrades as $\rho(A_\star)$ approaches one. Nonetheless, our method does not depend on a good initialization, does not require knowledge concerning the noise processes, and admits a simple theoretical guarantee.

7. Related work

Linear dynamical systems have been studied in depth as models for time-series data and we cannot do justice to its rich history or to the broader work on linear system identification. Brockwell and Davis (2009) and Ljung (1987) discuss these topics in detail. We focus on the closest and most recent related works.

Chow (1981) first proposed (1) as a model for time-varying regression, with Carraro (1984) later generalizing their analysis. The method studied by Carraro (1984) resembles in some ways are own. As an intermediate step it estimates the covariances between different innovations of a Kalman filter, but compared to CM it is complicated and no finite-sample guarantees have been provided.

Shumway and Stoffer (1982) developed model (1) from a different perspective and proposed EM as a viable estimator. They were interested in modelling time-series with missing data and in their model x_t is a matrix of zeros and ones that encodes which entries are available and which are missing in a set of time-series. Then, Khan and Dutt (2007) applied EM to fit a slightly generalized model (1) on EEG time series; they assumed the hidden states β_t follow an auto-regressive model.

Lubik and Matthes (2015) noted that macroeconomic time-series often display nonlinear behaviors that can be captured with time-varying linear models. In particular, they studied a time-varying auto-regressive model in which the unknown parameters evolve randomly and proposed a Bayesian inference method. In a similar vein, Isaksson (1987) studied actuated linear dynamics with randomly evolving parameters and proposed an adaptive Kalman filtering approach for identification. Most prior work on the identification of time-varying linear systems considered the case in which multiple experiments can be conducted on the system whose parameters undergo the same variation Liu (1997); Majji et al. (2010).

The recent development of non-asymptotic guarantees in statistics has led to similar results in system identification. In addition to the works previously mentioned in Sections 1.2 and 5, we mention a few other such works. Hardt et al. (2016) showed that a stochastic gradient descent can recover the parameters of certain linear systems in polynomial time. Sarkar et al. (2021) proved a non-asymptotic guarantee for the identification of partially observed linear dynamics of unknown order. Hazan et al. (2017) and Hazan et al. (2018) proposed spectral filtering methods that predict with sublinear regret the next output of unknown partially observed linear systems. When actuation is available Wagenmaker and Jamieson (2020) showed that active learning can be used for a faster identification of linear dynamics. Finally, Tsiamis et al. (2020) analyzed the sample complexity of designing a good Kalman filter based on observed data from an unknown linear systems.

8. Conclusion and open questions

Counterintuitively, we have shown that it is possible to fit a model for time-varying linear regression by combining just two ordinary least squares estimates. In contrast to EM, our method does not require a good initialization nor knowledge regarding the noise processes. Moreover, CM admits a simple theoretical guarantee that shows the estimation error decays at a $\mathcal{O}(1/\sqrt{T})$ rate.

CM has its own drawbacks. Firstly, it is applicable only to stable systems and its performance degrades as $\rho(A_\star)$ approaches one. Based on the statistical rate achievable when the states β_t are directly observable it should be possible to have a method that performs better when $A_\star$ is less stable. Secondly, we believe our guarantee for CM depends suboptimally on the failure probability δ, the dimension d, and the variances σ_ϵ^2 and Σ_w. We leave this issue and several other open questions for future work:

- Recent work in robust statistics have led to new algorithms and guarantees for linear regression with heavy-tailed noise Cherapanamjeri et al. (2020); Depersin (2020). Can one extend those results to our setting? These methods may address the suboptimal dependence of CM on some of the quantities previously mentioned.
- Similarly to the fully observed setting Simchowitz et al. (2018), can one guarantee that the performance of the maximum-likelihood estimator improves as $A_\star$ becomes less stable?
- Can one characterize the statistical rate of EM? Determining how closely to the MLE one needs to initialize EM would be of particular interest.
- What can one say when the hidden states β_t evolve according to more general linear dynamics or even nonlinear dynamics?

References

Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Annals of Statistics*, 45(1):77–120, 2017.

Bassam Bamieh and Laura Giarre. Identification of linear parameter varying models. *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal*, 12(9):841–853, 2002.

Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer Science & Business Media, 2009.

C Carraro. Identification and estimation of time varying models with structural variability of the parameters. In *Dynamic Modelling and Control of National Economies 1983*, pages 423–429. Elsevier, 1984.

Yeshwanth Cherapanamjeri, Samuel B Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 601–609, 2020.

Gregory C Chow. *Econometric analysis by control methods*. John Wiley & Sons, 1981.

Jules Depersin. A spectral algorithm for robust regression with subgaussian rates. *arXiv preprint arXiv:2007.06072*, 2020.

Randal Douc, Eric Moulines, Jimmy Olsson, and Ramon Van Handel. Consistency of the maximum likelihood estimator for general hidden markov models. *the Annals of Statistics*, 39(1):474–513, 2011.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. *arXiv preprint arXiv:1711.00946*, 2017.

Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. *arXiv preprint arXiv:1802.03981*, 2018.

Alf Isaksson. Identification of time varying systems through adaptive kalman filtering. *IFAC Proceedings Volumes*, 20(5):305–310, 1987.

Jer-Nan Juang and Richard S Pappa. An eigensystem realization algorithm for modal parameter identification and model reduction. *Journal of guidance, control, and dynamics*, 8(5):620–627, 1985.

Robert Kalaba and Leigh Tesfatsion. Time-varying linear regression via flexible least squares. *Computers & Mathematics with Applications*, 17(8-9):1215–1245, 1989.

Eugene Katsevich, Alexander Katsevich, and Amit Singer. Covariance matrix estimation for the cryo-em heterogeneity problem. *SIAM journal on imaging sciences*, 8(1):126–185, 2015.

Mohammad Emtiyaz Khan and Deshpande Narayana Dutt. An expectation-maximization algorithm based kalman smoother approach for event-related desynchronization (erd) estimation from eeg. *IEEE transactions on biomedical engineering*, 54(7):1191–1198, 2007.

Kefu Liu. Identification of linear time-varying systems. *Journal of Sound and Vibration*, 206(4): 487–505, 1997.

Lennart Ljung. System identification: theory for the user. *Prentice Hall*, 1987.

Thomas A Lubik and Christian Matthes. Time-varying parameter vector autoregressions: Specification, estimation, and an application. *Economic Quarterly-Federal Reserve Bank of Richmond*, 101(4):323, 2015.

Manoranjan Majji, Jer-Nan Juang, and John L Junkins. Time-varying eigensystem realization algorithm. *Journal of guidance, control, and dynamics*, 33(1):13–28, 2010.

Samet Oymak and Necmiye Ozay. Revisiting ho-kalman based system identification: robustness and finite-sample analysis. *IEEE Transactions on Automatic Control*, 2021.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618. PMLR, 2019.

Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite time lti system identification. *J. Mach. Learn. Res.*, 22:26–1, 2021.

RH Shumway, AS Azari, and Y Pawitan. Modeling mortality fluctuations in los angeles as functions of pollution and weather effects. *Environmental Research*, 45(2):224–241, 1988.

Robert H Shumway and David S Stoffer. An approach to time series smoothing and forecasting using the em algorithm. *Journal of time series analysis*, 3(4):253–264, 1982.

Robert H Shumway and David S Stoffer. *Time series analysis and its applications*, volume 3. Springer, 2000.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pages 2714–2802. PMLR, 2019.

Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Anastasios Tsiamis and George J Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654. IEEE, 2019.

Anastasios Tsiamis, Nikolai Matni, and George Pappas. Sample complexity of kalman filtering for unknown systems. In *Learning for Dynamics and Control*, pages 435–444. PMLR, 2020.

Stephen Tu, Ross Boczar, Andrew Packard, and Benjamin Recht. Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*, 2017.

Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems. In *Conference on Learning Theory*, pages 3487–3582. PMLR, 2020.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

John S White. The limiting distribution of the serial correlation coefficient in the explosive case. *The Annals of Mathematical Statistics*, pages 1188–1197, 1958.

Ting Zhang, Wei Biao Wu, et al. Inference of time-varying regression models. *The Annals of Statistics*, 40(3):1376–1402, 2012.