2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

# Estimation of Models with Limited Data by Leveraging Shared Structure

Maryann Rui Thibaut Horel Munther Dahleh

***Abstract*— Modern data sets, such as those in healthcare and e-commerce, are often derived from many individuals or systems but have insufficient data from each source alone to separately estimate individual, often high-dimensional, model parameters. If there is shared structure among systems however, it may be possible to leverage data from other systems to help estimate individual parameters, which could otherwise be non-identifiable. In this paper, we assume systems share a latent low-dimensional parameter space and propose a method for recovering $d$-dimensional parameters for $N$ different linear systems, when there are only $T < d$ observations per system. To do so, we develop a three-step method which estimates the low-dimensional subspace spanned by the systems' parameters and produces refined parameter estimates within the subspace. We provide finite sample subspace estimation error guarantees for our proposed method. Finally, we experimentally validate our method on simulations with i.i.d. regression data and as well as correlated time series data.**

## I. Introduction

In a variety of fields such as healthcare and e-commerce, it is often desirable to estimate parameters or provide recommendations for individuals based on data. Consider the common situation where we have $N$ different individuals, each with $T$ observations collected in $(X_i, Y_i) \in \mathbb{R}^{T\times d} \times \mathbb{R}^T$. Assume the data is generated as

$$Y_i = X_i \beta_i^{(0)} + w_i,\ i \in [N] \tag{I.1}$$

where $w_i \in \mathbb{R}^T$ is some independent noise vector, and $\beta_i^{(0)}$ are parameters of interest.

Such a linear model is ubiquitous in statistics, and standard least squares regression provides an estimate of $\beta_i^{(0)}$ based on $(X_i, Y_i)$ when $T \geq d$ and $X_i$ is well-conditioned.

Realistically, however, while a data set may contain many individuals, the data available from each individual may be limited, especially compared to the dimension of the parameter space considered. For instance, in the healthcare setting, patient data may be fragmented and stored on different electronic health record systems, so that each record system may have many individuals but imcomplete data from each [1]. This may lead to problems of non-identifiability for individual systems, as in the case where we only have $T < d$ observations of a $d$-dimensional linear model.

If there is a shared structure among individuals, however, it may be possible to leverage information from other individuals who share similar characteristics to overcome the challenge of non-identifiability of individual parameters.

Laboratory for Information & Decision Systems, Massachusetts Institute of Technology. {maryann,thibauth,dahleh}@mit.edu

In this paper, we examine this possibility and propose a method of estimating each system's $d$-dimensional parameter by exploiting data from other systems along with the assumption that the parameters lie in a common $r$-dimensional subspace, where $r < d$. The questions we wish to answer are: can a sufficiently large number $N$ of systems compensate for a small amount $T$ of data per system in the task of estimating all the parameters? If so, how does the sample complexity scale in the parameters $N, T, r$ and $d$ of the problem?

If we simply count the degrees of freedom of the model, we have $r(d-r)+Nr$ parameters to estimate (the common $r$-dimensional subspace of parameters plus individual factor loadings or coefficients). Intuitively, one may expect that $NT \geq r(d-r) + Nr$ parameters are needed to jointly identify all parameters of the system. It is not obvious how to rigorously justify this intuition, nor how to develop and implement an estimation algorithm for this setting.

To begin to tackle this complex and broad-ranging question, we propose an estimation method based on three separate least squares optimizations. The method first computes initial estimates of each system's parameter vector, which may be significantly far from the true value, but which still contain information about the common underlying subspace spanned by the true system parameters. Next, an estimate of this low dimensional subspace is obtained by extracting the top $r$ principal subspace of the first step estimates. From this, we can obtain a refined estimate of each system's individual parameters by solving another least squares problem, this time constrained to be over the estimated subspace. This last step requires $T \geq r$ for parameter identifiability, which can be a considerably easier condition to satisfy than the naive requirement of $T \geq d$, as $d \gg r$ in many real world datasets.

We provide finite sample subspace estimation error guarantees for a variant of our proposed method that takes into account the possible ill-conditioning of the pseudo-inverse-based least squares solution which arises when $T \approx d$. The analysis relies on obtaining concentration bounds for the sample covariance of the first-step estimates, and then proving that subspace estimation on these first-step estimates will obtain the true underlying subspace in expectation.

Finally, we demonstrate our method and variants on simulations with i.i.d. regression data. We also evaluate our method on time series data with correlated regressors, and find that the method is flexible enough to handle this scenario. These results suggest the applicability of the three-step estimation method for more general settings of estimation of related with a common low rank structure.

### A. Related Work

*a) Mixtures of linear regressions:* The problem of estimating parameters from limited observations of different systems that share a common low dimensional structure is related to the problem of mixtures of linear regressions [2], [3] and multitask, or meta-learning [4], [5]. The main difference to our setting is the systems' parameters are assumed to be clustered, rather than coming from a low dimensional subspace.

*b) Low rank matrix regression:* Furthermore, a large body of work studies the related problem of low rank matrix regression, which usually uses a least squares estimator with nuclear norm regularization to estimate a low rank matrix. We can re-express (I.1) to make the comparison with matrix regression explicit. Let $\Theta^{(0)} := \begin{bmatrix}\beta_1^{(0)} \cdots \beta_N^{(0)}\end{bmatrix}$ be the $d \times N$ matrix whose columns are the system parameters. Then by assumption, $\operatorname{rk} \Theta^{(0)} = r$. However, the data generating process for observations $(X_i, y_i)$ is

$$y_i = X_i \Theta^{(0)} e_i + w_i, \ i \in [N].$$

where $e_i$ is the $i$th coordinate vector in $\mathbb{R}^N$. While we are still trying to estimate a low rank matrix $\Theta^{(0)}$, the dimension of this matrix grows with the number of observations—unlike in matrix regression where it is assumed constant—landing us in a different regime for analysis and optimization.

*c) Dictionary Learning:* Finally, the problem of dictionary learning, also known as sparse coding, and matrix factorization, shares similar structure to the problem considered in this paper [6], [7], [8]. However, we only observe system parameters through the lens of a design matrix $X$ whose rows do not fully span the parameter space. Even if our design matrix $X \in \mathbb{R}^{T \times d}$ were the identity (so $T = d$), though, we also do not impose a sparsity assumption on the dictionary coefficients, as is standard in the dictionary learning literature. The differences are further detailed in Section IV-A.

*d) Meta-learning and transfer learning:* After the initial submission of this paper, we became aware of recent related work on this problem. In [9], the authors present a method of moments (MoM) estimator, and a similar estimator is studied under more general assumptions in [10]. We discuss the relationship between these estimators and ours and provide an empirical comparison in Section V-C.

More generally our work connects with the broader literature in machine learning that studies learning across related tasks or data sets, usually referred to by the umbrella terms meta-learning and transfer learning [11], [12].

## II. Preliminaries

### A. General notations

We define for $N \in \mathbb{N}^*$ the set $[N] := \{1, 2, \ldots, N\}$. The inequality $a \lesssim b$ means that there exists a universal constant $C$ such that $a \leq Cb$.

For vectors $x, y \in \mathbb{R}^d$, $\langle x, y\rangle := x^\intercal y = \sum_{i=1}^d x_i y_i$ and $\|x\|_2 := \sqrt{\langle x, x\rangle}$ denote the Euclidean inner product and norm, respectively.

For a matrix $A$, $\operatorname{tr} A$, $A^\intercal$ and $A^\dagger$ denote its trace, transpose and Moore–Penrose pseudoinverse, respectively. The identity matrix in $\mathbb{R}^{d\times d}$ is written $I_d$. For matrices $A, B \in \mathbb{R}^{d\times r}$, $\langle A, B\rangle := \operatorname{tr}(A^\intercal B)$ denotes the Frobenius or trace inner product, $\|A\|_{\mathrm{F}} = \sqrt{\operatorname{tr}(A^\intercal A)}$ is the Frobenius norm of $A$, and $\|A\|_2$ is its spectral norm. Let $\mathcal{O}(d) := \{Q \in \mathbb{R}^{d\times d} \mid Q^\intercal Q = QQ^\intercal = I_d\}$ denote the orthogonal group on $\mathbb{R}^d$ and $\mathrm{St}(r, d) := \{A \in \mathbb{R}^{d\times r} \mid A^\intercal A = I_r\}$ denote the Stiefel manifold of orthonormal $r$-frames in $\mathbb{R}^d$.

### B. Subspaces

For $r \leq d$, $\mathrm{Gr}(r, d)$ denotes the Grassmanian manifold of $r$-dimensional subspaces of $\mathbb{R}^d$ and we write $P_{\mathcal{A}}$ for the orthogonal projection onto a subspace $\mathcal{A} \in \mathrm{Gr}(r, d)$. For $Q$ an orthogonal projection of rank $r$ and $\mathcal{A} \in \mathrm{Gr}(r, d)$, the identities $Q = P_{\operatorname{im} Q}$ and $\mathcal{A} = \operatorname{im} P_{\mathcal{A}}$ show that the map $\mathcal{A} \mapsto P_{\mathcal{A}}$ is a bijection from $\mathrm{Gr}(r, d)$ to the set of orthogonal projections of rank $r$. This allows us to identify the two sets [13, Sec. 1.3.2]:

$$\mathrm{Gr}(r, d) \cong \{P \in \mathbb{R}^{d\times d} \mid P = P^\intercal = P^2 \wedge \operatorname{tr} P = r\}.$$

Note that the choice of an orthonormal basis of a subspace $\mathcal{A} \in \mathrm{Gr}(r, d)$ gives a representation of $\mathcal{A}$ by an element $A \in \mathrm{St}(r, d)$, although the representation is non-unique. For such a matrix $A$ we have $P_{\mathcal{A}} = AA^\intercal$.

For any two subspaces $\mathcal{A}, \mathcal{B} \in \mathrm{Gr}(r, d)$, we can find $r$ pairs of *principal vectors* $(a_i, b_i) \in \mathcal{A} \times \mathcal{B}$ for $i \in [r]$, and *principal angles* $(\theta_1, \ldots, \theta_r) \in [0, \pi/2]^r$ such that $\langle a_i, b_i\rangle = \cos(\theta_i)$, $\langle a_i, a_j\rangle = 0$, and $\langle b_i, b_j\rangle = 0$, for $i, j \in [r], i \neq j$. $\theta_1$ is the smallest angle between any vector in $\mathcal{A}$ and any vector in $\mathcal{B}$, which is achieved by $a_1$ and $b_1$. The remaining principal vectors and angles are defined inductively, by restricting at each step to the orthogonal complement of the span of the previous vectors [14], [15]. We write $\Theta(\mathcal{A}, \mathcal{B})$ for the diagonal matrix whose diagonal entries are the principal angles. It is possible to show that the nonzero eigenvalues of $P_{\mathcal{A}} - P_{\mathcal{B}}$ are the sines of the nonzero principal angles between $\mathcal{A}$ and $\mathcal{B}$, each counted twice. This implies that $\|P_{\mathcal{A}} - P_{\mathcal{B}}\|_2 = \|\sin \Theta(\mathcal{A}, \mathcal{B})\|_2$ and $\|P_{\mathcal{A}} - P_{\mathcal{B}}\|_{\mathrm{F}} = \sqrt{2}\|\sin \Theta(\mathcal{A}, \mathcal{B})\|_{\mathrm{F}}$.

### C. Random variables

Unless otherwise specified, all random variables are defined on the same probability space. We write $X \overset{d}{=} Y$ for identically distributed variables $X$ and $Y$. We say that a random vector $X \in \mathbb{R}^d$ is sub-Gaussian with variance proxy $\sigma^2$, and write $X \in \mathsf{subG}_d(\sigma^2)$, if for all $\alpha \in \mathbb{R}^d$

$$\mathbb{E}[\exp\langle \alpha, X\rangle] \leq \exp\left(\frac{\sigma^2 \|\alpha\|_2^2}{2}\right).$$

## III. Model

We consider $N$ linear systems of dimension $d$, from each of which we have $T$ observations, with $T < d$. Specifically, each system $i \in [N]$ has observations $(X_i, Y_i) \in \mathbb{R}^{T\times d} \times \mathbb{R}^T$ generated according to

$$Y_i = X_i \beta_i^{(0)} + w_i \tag{III.1}$$

where $\beta_i^{(0)} \in \mathbb{R}^d$ is the parameter of system $i \in [N]$. The central assumption of our model is that the system parameters $(\beta_i^{(0)})_{i\in[N]}$ lie in an $r$-dimensional subspace $\mathcal{B}_0$ of $\mathbb{R}^d$. An orthonormal basis $B_0 \in \mathrm{St}(r,d)$ of $\mathcal{B}_0$ constitutes a common dictionary of $r$ atoms shared by all systems and we can write for each $i \in [N]$

$$\beta_i^{(0)} = B_0\phi_i$$

for some $\phi_i \in \mathbb{R}^r$. We are primarily interested in estimating the $r$-dimensional subspace $\mathcal{B}_0$, from which we can easily recover $\phi_i$, and then $\beta_i^{(0)}$ for $i \in [N]$, from the data.

We now describe our distributional assumptions. All the random variables $(X_i)_{i\in[N]}$, $(\phi_i)_{i\in[N]}$ and $(w_i)_{i\in[N]}$ are mutually independent. The matrices $(X_i)_{i\in[N}$ are identically distributed, each having i.i.d. standard normal entries. Coefficients $\phi_i \in \mathbb{R}^r$ and noise $w_i \in \mathbb{R}^T$ are centered isotropic sub-Gaussian random vectors in $\mathsf{subG}(\sigma_\phi^2)$ and $\mathsf{subG}(\sigma_w^2)$, respectively, with covariance matrices $\sigma_\phi^2 I_r$ and $\sigma_w^2 I_T$, respectively.

## IV. Method

We propose an estimation method that follows a general three step approach. First, compute initial estimates of each system $i$'s parameter. Next, find the $r$-dimensional subspace that best explains the initial estimates. In the last step, the individual system estimates are refined by leveraging the subspace learned in the second step.

Due to the linear structure of our observation model, instantiating the above approach naturally results in formulating and solving a linear least squares problem at each of the three steps. The following subsections describe these least squares problems in more details and Algorithm 1 summarizes the computation of their solution.

**Algorithm 1** Three-step parameter and subspace estimator

**Input:** Samples $(X_i, Y_i) \in \mathbb{R}^{T\times d} \times \mathbb{R}^T$, for $i \in [N]$; model rank $r$.

**Output:** Estimated subspace frame $\widehat{B} \in \mathrm{St}(r,d)$, coefficients $\{\beta^{(2)} \in \mathbb{R}^d : i \in [N]\}$.

1: **for** $i \in [N]$ **do**
2: $\quad \beta_i^{(1)} \leftarrow X_i^\dagger Y_i$ ▷ First-step estimate
3: $\quad \bar{\beta}^{(1)} \leftarrow \beta_i^{(1)}/\|\beta_i^{(1)}\|_2$ ▷ Normalization
4: **end for**
5: $USV^\intercal \leftarrow \mathrm{SVD}([\bar{\beta}_1^{(1)} \cdots \bar{\beta}_N^{(1)}])$
6: $\widehat{B} \leftarrow U[:, 1:r]$ ▷ Subspace estimation
7: **for** $i \in [N]$ **do**
8: $\quad \beta_i^{(2)} \leftarrow \widehat{B}(X_i\widehat{B})^\dagger Y_i$ ▷ Coefficient estimation
9: **end for**

### A. Initial Individual Estimates

We first obtain a least squares estimate $\beta_i^{(1)}$ of $\beta_i^{(0)}$ up to the null space of $X_i$, which is nontrivial as we focus on the regime $T < d$. Let $X_i^\dagger$ be the pseudoinverse of $X_i$. Then our initial estimates are

$$\beta_i^{(1)} = X_i^\dagger Y_i,\ i \in [N].$$

To gain a better understanding of our initial estimates $\beta_i^{(1)}$, under our observation model (III.1), we can write

$$\beta_i^{(1)} = X_i^\dagger X_i\beta_i^{(0)} + X_i{}^\dagger w_i = P_{X_i}\beta_i^{(0)} + X_i{}^\dagger w_i$$

where $P_{X_i} := X_i^\dagger X_i$ is the $d \times d$ projection matrix onto the $T$-dimensional row space of $X_i$. Since the distribution of the rows of $X_i$ is rotationally invariant, one can check that $\mathbb{E}[P_{X_i}] = \frac{T}{d} I_d$. Hence, up to a rescaling by $\frac{d}{T}$, $\beta_i^{(1)}$ is an unbiased estimate of $\beta_i^{(0)}$. However, the noise of this estimate

$$\eta_i := \beta_i^{(1)} - \beta_i^{(0)} = -(I_d - P_{X_i})\beta_i^{(0)} + X_i^\dagger w_i \qquad \text{(IV.1)}$$

is not independent of the true parameter $\beta_i^{(0)}$ as it includes its projection onto the null space of $X_i$ that was left unobserved.

*a) Comparison with dictionary learning:* Using (IV.1) wen can write $\beta_i^{(1)} = \beta_i^{(0)} + \eta_i$. Hence, our setting is reminiscent of the dictionary learning, or sparse coding, problem, in which we have $N$ noisy observations where $\beta_i^{(0)} = B_0\phi_i$, for some fixed $d \times r$ matrix $B_0$, and some vector $\phi_i \in \mathbb{R}^r$. Both $B_0$ and $\phi_i$ are unknown and the goal is to learn $B_0$, which in our setting then allows for a straightforward estimation of $\phi_i$ and $\beta_i^{(0)}$.

However, we cannot apply dictionary learning methods straight off the shelf. First, dictionary learning models assume sparsity of the unknown coefficients $\phi_i$, and for most sample complexity results, a degree of sparsity is necessary (i.e., the size of the support of $\phi_i$ is upper bounded) [6], [16], [17]. Meanwhile, we make no restrictive assumptions on the factor loadings $\phi_i$. Second, the additive noise $\eta_i$ in dictionary learning is assumed to be independent of other randomness in the problem, either with standard subgaussian or bounded distributional assumptions [6], [16], [7]. However, as already mentioned, $\eta_i$ in our method is visibly not independent of the parameters $\beta_i^{(0)}$ to be estimated, as it contains the component of $\beta_i^{(0)}$ in the null space of $X_i$.

*b) Normalization:* The normalization step on line (3) ensures that the first-step estimates are all weighted equally in the subspace estimation step. As will become clear in the simulations (Section V-B) this mitigates issues arising due to the pseudo-inverse $X_i^\dagger$ being ill-conditioned when $T$ is close to $d$. For tractability reasons, our theoretical analysis studies a variant of Algorithm 1 in which the normalization step is replaced with a truncation $\bar{\beta}^{(1)} \leftarrow \beta_i^{(1)}\mathbf{1}\{\|X_i^\dagger\|_2 \le s\}$, for some predefined threshold $s$. We also compare this variant to our main estimator in Section V-B.

### B. Subspace recovery

The goal of the second step is to compute an estimate of the $r$-dimensional subspace of $\mathbb{R}^d$ containing the ground truth parameters $(\beta_i^{(0)})_{i\in[N]}$. We do so by finding the $r$-dimensional subspace that best approximates the (normalized) first step estimates $(\bar{\beta}_i^{(1)})_{i\in[N]}$, in the least squares sense. If $P$ denotes the orthogonal projection onto this optimal subspace, the residual error associated with $\beta_i^{(1)}$

is its distance to the subspace, that is $\|\bar{\beta}_i^{(1)} - P\bar{\beta}_i^{(1)}\|_2$. Consequently, the least squares problem at this step is

$$\min_{P\in \mathrm{Gr}(r,d)} \frac{1}{N}\sum_{i\in[N]} \|\bar{\beta}_i^{(1)} - P\bar{\beta}_i^{(1)}\|_2^2. \tag{IV.2}$$

Note that (IV.2) is exactly the problem of finding the space spanned by the first $r$ *principal components* of the first step estimates $(\bar{\beta}_i^{(1)})_{i\in[N]}$. Those are given by the top $r$ left-singular vectors of the matrix $\Theta^{(1)} \in \mathbb{R}^{d\times N}$ whose columns are $(\bar{\beta}_i^{(1)})_{i\in[N]}$.

Another interpretation of this subspace estimate can be obtained by observing that:

$$\begin{aligned}\sum_{i\in[N]} \|\bar{\beta}_i^{(1)} - P\bar{\beta}_i^{(1)}\|_2^2 &= \|\Theta^{(1)} - P\Theta^{(1)}\|_{\mathrm{F}}^2 \\ &= \langle (I_d - P)\Theta^{(1)}, (I_d - P)\Theta^{(1)}\rangle \\ &= \left\langle I_d - P, \Theta^{(1)}\Theta^{(1)\intercal}\right\rangle,\end{aligned}$$

where the last equality uses that $I_d - P$ is also an orthogonal projection. This allows us to rewrite (IV.2)

$$\begin{aligned}&\operatorname*{arg\,min}_{P\in \mathrm{Gr}(r,d)} \frac{1}{N}\sum_{i\in[N]} \|\bar{\beta}_i^{(1)} - P\bar{\beta}_i^{(1)}\|_2^2 \\ &\qquad = \operatorname*{arg\,max}_{P\in \mathrm{Gr}(r,d)} \left\langle P, \frac{\Theta^{(1)}\Theta^{(1)\intercal}}{N}\right\rangle.\end{aligned}$$

The matrix $\frac{1}{N}\Theta^{(1)}\Theta^{(1)\intercal} = \frac{1}{N}\sum_{i\in[N]} \bar{\beta}_i^{(1)}\bar{\beta}_i^{(1)\intercal}$ appearing on the last line is the sample covariance matrix of the first step estimates. Because this matrix is positive semi-definite, it admits a spectral decomposition with non-negative eigenvalues and orthogonal eigenspaces. The top $r$ left-singular vectors of $\Theta^{(1)}$ are equivalently given by an orthonormal collection of eigenvectors associated with the top $r$ eigenvalues of $\Theta^{(1)}\Theta^{(1)\intercal}$ (counted with multiplicity).

### C. Parameter Recovery

In the third stage of our algorithm we obtain revised estimates of $\beta_i^{(2)}$ for each $i \in [N]$ given the estimate $\widehat{B}$:

$$\begin{aligned}\beta_i^{(2)} &\in \operatorname*{arg\,min}_{\beta\in\widehat{B}} \|y_i - X_i\beta\|_2^2 \\ &= \widehat{B}(X_i\widehat{B})^{\dagger} y_i = P_{\widehat{B}}(X_i P_{\widehat{B}})^{\dagger} y_i.\end{aligned}$$

One can obtain this result by solving for $\beta_i^{(2)} = \widehat{B}\widehat{\phi}_i$ where

$$\widehat{\phi}_i \in \operatorname*{arg\,min}_{\phi\in\mathbb{R}^r} \|y_i - X_i\widehat{B}\phi\|_2^2.$$

## V. Results

### A. Sample complexity

Our main theoretical result is an upper-bound on the sample complexity of the estimator described in Algorithm 1. As already mentioned, it is easier to analyze a variant in which line (3) performs a truncation instead of a normalization. Thus, for the remainder of this section $\bar{\beta}_i^{(1)}$ is defined as $\bar{\beta}_i^{(1)} = \beta_i^{(1)}\mathbf{1}\{\|X_i^{\dagger}\|_2 \le s\}$, for some predefined threshold $s$. Equivalently, Algorithm 1 simply drops the first step estimates $\beta_i^{(1)}$ for which $\|X_i^{\dagger}\|_2 > s$, and uses the remaining ones in the subspace estimation step.

**Theorem V.1.** *Let $\hat{\mathcal{B}}$ be the subspace spanned by the columns of the output $\widehat{B}$ in Algorithm 1 with threshold level $s = \Omega\big(1/(\sqrt{d}-\sqrt{T})\big)$. Then for each $0 < \delta < 1$, with probability at least $1-\delta$*

$$\|\sin\Theta(\hat{\mathcal{B}}, \mathcal{B}_0)\|_2 \lesssim \left(1 + \frac{\sigma_w^2/\sigma_\phi^2}{(\sqrt{d}-\sqrt{T-1})^2}\right)\frac{d^2}{T^2}\sqrt{\frac{d}{N}\log\frac{2}{\delta}}$$

The proof of Theorem V.1 is provided in the full version of the paper. We first show that the $r$th principal subspace of the covariance matrix $\Sigma_\beta = \mathbb{E}[\bar{\beta}_i^{(1)}\bar{\beta}_i^{(1)\intercal}]$ is the ground truth subspace $\mathcal{B}_0$ and quantify its spectral gap. By the Davis–Kahan theorem, upper-bounding the principal subspace angle reduces to upper-bounding the spectral norm of $\frac{1}{N}\sum_{i=1}^N \bar{\beta}_i^{(1)}\bar{\beta}_i^{(1)\intercal} - \Sigma_\beta$. This follows from a standard result on the concentration of covariance matrices of sub-Gaussian random variables.

The bound in Theorem V.1 scales as $1/\sqrt{N}$ as expected when learning from independent observations. While the dimension $r$ of $\mathcal{B}_0$ does not explicitly appear in the bound, it would be conventional to take $\sigma_\phi^2 = 1/r$, so that the norm of the ground truth parameters $\beta_i^{(0)}$ concentrate around 1, as is usually assumed in sample complexity bounds. For this choice of $\sigma_\phi^2$, our error bound scales linearly in $r$. Finally recall that by our modeling assumption, we require $T \le d$, so this bound does not allow us to study the asymptotic regime where $T$ grows to infinity. The first term in the bound degrades as $T$ increases and becomes $\Omega(d)$ when $T = d$. This is due to the design matrix $X_i$ becoming ill-conditioned for $T$ close to $d$, a phenomenon we inspect more closely in Section V-B below.

### B. Simulation

We investigate by simulation the performance of the three-step estimator of Algorithm 1 that uses normalized first-step estimates, as well as the thresholding variant of Algorithm 1 analyzed in Theorem V.1, and which uses a truncated first-step estimate for subspace estimation.

According to (I.1), i.i.d. samples $(X_i, Y_i)$, $i \in [N]$ were generated with $w_i \sim \mathcal{N}(0, \sigma_w^2 I_T)$, $\beta^{(0)}$ drawn uniformly from a common $r$-dimensional subspace of $\mathbb{R}^d$ and normalized to unit norm, and $X_i$ with i.i.d. standard normal entries. Each plot shows the average of 30 trials and error bars indicate one standard deviation.

Figure 1 shows the subspace estimation error of the estimate $\widehat{B}$ from (1) `thresh`, Algorithm 1 with truncated first-step estimates, (2) `norm`, Algorithm 1 with normalized first step estimates. Results are shown for $d = 50, r = 5$, in regimes $d > T = 10$, $T = d = 50$, and $d < T = 80$, as the number of systems $N$ is varied.

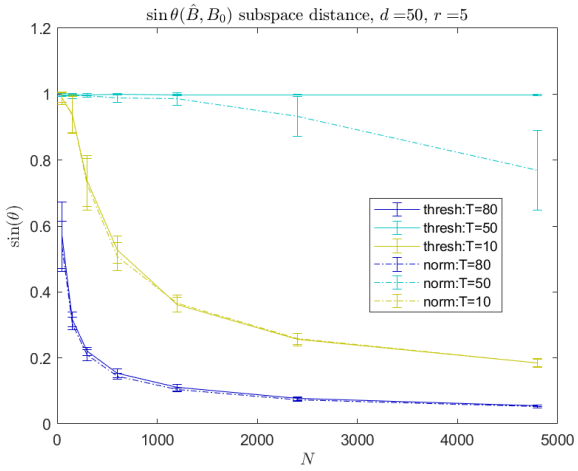


Fig. 1. Subspace estimation error vs $N$ for estimates $\widehat{B}$ from (1) `thresh` and (2) `norm`. Results are shown for $d = 50, r = 5$, in regimes $d > T = 10$, $T = d = 50$, and $d < T = 80$.

Both versions of the three-step estimator do well in both the $T > d$ and $T < d$ regime, though we note suboptimal performance when $T = d$. This arises from the fact that the pseudoinverse of $X_i$ can be ill-conditioned when $T = d$. As described before, we address this issue by normalizing our first step estimates to mitigate the effect of a single sample misdirecting the subspace estimator with an amplified noise term $X_i^\dagger w_i$. For the truncating estimator, we chose an optimal threshold level for each $d$ and $T$ that trades off controlling the effect of possibly ill-conditioned pseudo-inverse-based least squares estimates with losses in effective sample size.

In Figure 2, we show the performance of the refined $\beta_i$ estimates from `thresh` in comparison with two benchmarks: (1) the oracle least squares estimate over $\mathcal{B}_0$ as if it were known, and (2) the naive least squares estimate run separately for each system $i \in [N]$, and which does not share information across systems. The `thresh` estimator is able to leverage information across systems to eventually match the performance of the oracle.

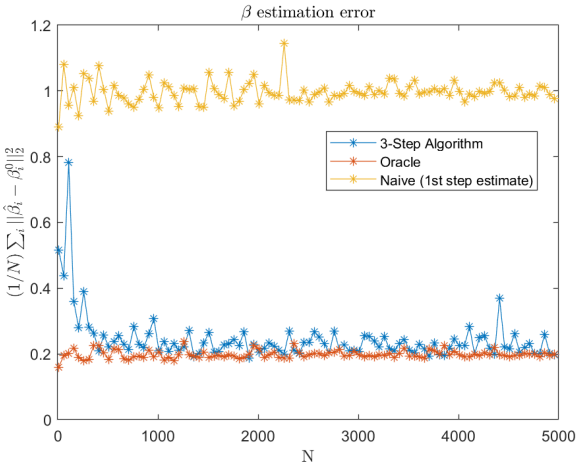


Fig. 2. Comparison of $\beta_i$ estimation error of `thresh` as compared to oracle and naive estimators. Here, $d = 5, r = 1, T = 3$.

### C. Comparison with related work

In [9], the authors present a method of moments estimator, which we refer to as `MoM`, for subspace estimation in the present setting. This estimator can be interpreted under our three-step method as first obtaining a first-step estimate $\beta_i^{(1)} = X_i^\intercal Y_i$ and then estimating the subspace shared by these first-step estimators. Figure 3 and 4 compare the performance of `norm` and of `MoM` in terms of subspace estimation error. Results are shown for the regimes $T < d$, $T = d$, and $T > d$. We see that both estimators perform comparably in the first and third regimes, while `norm` suffers in the $T = d$ regime where $X^\dagger$ may be ill-conditioned. However, the next section on time series data suggests that `norm` may generalize better to settings with non-i.i.d. regressors such as time-series data.

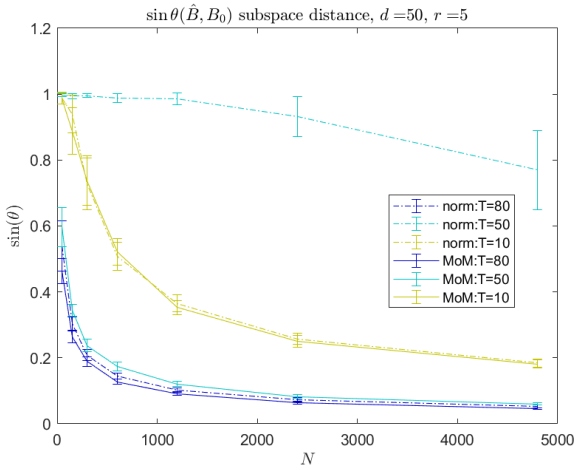


Fig. 3. Subspace estimation error vs. $N$ for estimates $\widehat{B}$ from (1) `norm` and (2) `MoM`, for i.i.d. data. Results are shown for $d = 50, r = 5$, in regimes $d > T = 10$, $T = d = 50$, and $d < T = 80$..

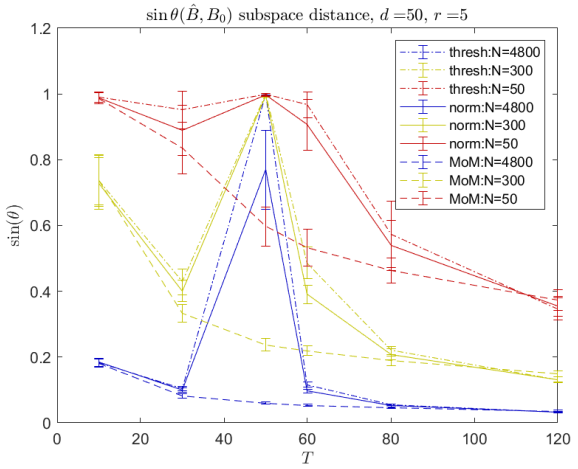


Fig. 4. Subspace estimation error vs $T$ for estimates $\widehat{B}$ from (1) `thresh`, (2) `norm`, and (3) `MoM`, for i.i.d. data. Results are shown for $d = 50, r = 5$, for various values of $N$.

### D. Time series regression

We next evaluate our algorithm and the method of moments estimator, MoM of [9], on time series data. Specifically, consider $T$ observations $(x_{it}, y_{it})_{t\in[T]}$ generated as:

$$x_{i,t+1} = A_i x_{it} + w_{it},\ t \in [T-1],\ x_{i0} \sim \mathcal{N}(0, \sigma_x^2)$$

for each $i \in [N]$, with $x_{it} \in \mathbb{R}^d$, $A_i \in \mathbb{R}^{d\times d}$ and $w_{it} \in \mathbb{R}^d$ a sub-Gaussian random vector in $\mathsf{subG}(\sigma_w^2)$. We assume each dynamics matrix $A_i$ is rank $r \le d$ and can be written in the form

$$A_i = F_i B^\mathsf{T}$$

where the rows of $F_i \in \mathbb{R}^{d\times r}$ are independently distributed and rotationally invariant, and $B \in \mathrm{St}(r, d)$ is an orthonormal $r$-frame for a subspace in $\mathrm{Gr}(r, d)$. $A_i$ is normalized by its operator norm to ensure stability.

Unlike the setting of i.i.d. regression, the covariates $(x_{it})_{t\in[T]}$ are no longer independent of each other and of the collection of noise vectors $(w_{it})_{t\in[T]}$.

We find in our simulations that norm is able to generalize to this setting quite well when $T$ is not close to $d$, as opposed to the MoM, which fails to learn even as $N$ increases. Intuitively, MoM is not robust to the non-isotropy of the regressors $X$, while our least-squares-based first-step estimate is still able to extract useful information in this setting.

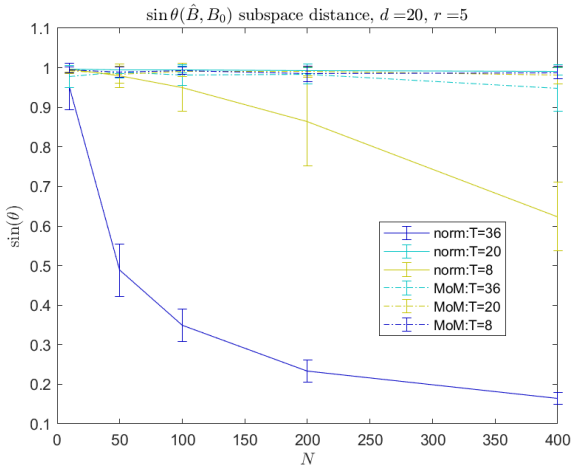


Fig. 5. Subspace estimation error vs. $N$ for estimates $\widehat{B}$ from (1) norm and (2) MoM, for time-series data. Results are shown for $d = 20, r = 5$, in regimes $d > T = 8$, $T = d = 20$, and $d < T = 36$.

## VI. Conclusion

We have shown that when there is shared low-rank structure among systems, we can leverage data from other systems to help estimate individual parameters, even in the regime $r \le T < d$, in which systems would otherwise be non-identifiable from their own data alone. We have presented a method to estimate the common low dimensional subspace as well as the system parameters, by a series of three least squares optimization problems, one of which can be solved simply by singular value decomposition. We then provided finite sample estimation error guarantees of a truncating variant our proposed method. These sample complexity results are not necessarily optimal, and we seek to better understand the trade-offs in the number of systems $N$, and the number of observations per system $T$ in the best achievable estimation error. However, experiments suggest that the three-step estimation procedure may be applied successfully to more general settings such as time-series estimation.

## References


[1] S. B. Dewdney and J. Lachance, "Electronic records, registries, and the development of "big data": crowd-sourcing quality toward knowledge," *Frontiers in oncology*, vol. 6, p. 268, 2017.

[2] S. Faria and G. Soromenho, "Fitting mixtures of linear regressions," *Journal of Statistical Computation and Simulation*, vol. 80, no. 2, pp. 201–225, 2010.

[3] Y. Li and Y. Liang, "Learning mixtures of linear regressions with nearly optimal complexity," in *Conference On Learning Theory*. PMLR, 2018, pp. 1125–1144.

[4] W. Kong, R. Somani, S. Kakade, and S. Oh, "Robust meta-learning for mixed linear regression with small batches," *Advances in neural information processing systems*, vol. 33, pp. 4683–4696, 2020.

[5] W. Kong, R. Somani, Z. Song, S. Kakade, and S. Oh, "Meta-learning for mixed linear regression," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5394–5404.

[6] R. Gribonval, R. Jenatton, and F. Bach, "Sparse and spurious: dictionary learning with noise and outliers," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6298–6319, 2015.

[7] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert, "Sample complexity of dictionary learning and other matrix factorizations," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 3469–3486, 2015.

[8] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 689–696.

[9] N. Tripuraneni, C. Jin, and M. Jordan, "Provable meta-learning of linear representations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 434–10 443.

[10] J. C. Duchi, V. Feldman, L. Hu, and K. Talwar, "Subspace recovery from heterogeneous data with non-isotropic noise," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5854–5866, 2022.

[11] M. Pontil, "The benefit of multitask representation learning," *Machine Learning with Interdependent and Non-identically Distributed Data*, p. 46, 2015.

[12] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei, "Few-shot learning via learning the representation, provably," *arXiv preprint arXiv:2002.09434*, 2020.

[13] Y. Chikuse, *Statistics on Special Manifolds*. Springer, 2003.

[14] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU press, 2013.

[15] K. Ye and L.-H. Lim, "Schubert varieties and distances between subspaces of different dimensions," *SIAM Journal on Matrix Analysis and Applications*, vol. 37, no. 3, pp. 1176–1197, 2016.

[16] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," in *Conference on Learning Theory*. PMLR, 2014, pp. 779–806.

[17] R. Gribonval and K. Schnass, "Dictionary identification—sparse matrix-factorization via $\ell_1$-minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.