

# Optimal Subspace Estimation Using Overidentifying Vectors via Generalized Method of Moments

Jianqing Fan\* and Yiqiao Zhong\*

\*Department of Operations Research and Financial Engineering, Princeton University

## Abstract

Many statistical models seek relationship between variables via subspaces of reduced dimensions. For instance, in factor models, variables are roughly distributed around a low dimensional subspace determined by the loading matrix; in mixed linear regression models, the coefficient vectors for different mixtures form a subspace that captures all regression functions; in multiple index models, the effect of covariates is summarized by the effective dimension reduction space.

Such subspaces are typically unknown, and good estimates are crucial for data visualization, dimension reduction, diagnostics and estimation of unknown parameters. Usually, we can estimate these subspaces by computing moments from data. Often, there are many ways to estimate a subspace, by using moments of different orders, transformed moments, etc. A natural question is: how can we combine all these moment conditions and achieve optimality for subspace estimation?

In this paper, we formulate our problem as estimation of an unknown subspace  $\mathcal{S}$  of dimension  $r$ , given a set of overidentifying vectors  $\{\mathbf{v}_\ell\}_{\ell=1}^m$  (namely  $m \geq r$ ) that satisfy  $\mathbb{E}\mathbf{v}_\ell \in \mathcal{S}$  and have the form

$$\mathbf{v}_\ell = \frac{1}{n} \sum_{i=1}^n \mathbf{f}_\ell(\mathbf{x}_i, y_i),$$

where data are i.i.d. and each function  $\mathbf{f}_\ell$  is known. By exploiting certain covariance information related to  $\mathbf{v}_\ell$ , our estimator of  $\mathcal{S}$  uses an optimal weighting matrix and achieves the smallest asymptotic error, in terms of canonical angles. The analysis is based on the generalized method of moments that is tailored to our problem. Our method is applied to aforementioned models and distributed estimation of heterogeneous datasets, and may be potentially extended to analyze matrix completion, neural nets, among others.

*Keywords:* GMM, ensemble method, aggregation, eigenvectors, factor model, mixture model, index model, distributed estimation.

---

Address: Department of ORFE, Sherrerd Hall, Princeton University, Princeton, NJ 08544, USA, e-mail: [jqfan@princeton.edu](mailto:jqfan@princeton.edu), [yiqiaoz@princeton.edu](mailto:yiqiaoz@princeton.edu). The research was partially supported by NSF grants DMS-1712591 and DMS-1662139 and NIH grant R01-GM072611.

# 1 Introduction

## 1.1 Motivation of subspace estimation

In statistics, many models are used to infer from data as simple relationship between variables as possible. Arguably, it is usually easier to conduct statistical analysis and interpret results if we find a simple relationship. For example, a family of well-studied statistical models is factor models:

$$\mathbf{x}_i = \mathbf{B}\mathbf{z}_i + \boldsymbol{\epsilon}_i, \quad i \in [n] = \{1, 2, \dots, n\}. \quad (1)$$

where  $\mathbf{x}_i \in \mathbb{R}^p$ ,  $\mathbf{z}_i \in \mathbb{R}^r$  and  $r$  is usually (much) smaller than  $p$ . This model is useful since  $\mathbf{x}_i$  is characterized by a smaller number of variables  $\mathbf{z}_i$  (which are called factors) via a linear transformation  $\mathbf{B}$  (which is called the loading matrix), plus unexplained variable or noise  $\boldsymbol{\epsilon}_i$ . In particular, we have

$$\mathbf{x}_i - \boldsymbol{\epsilon}_i = \mathbf{B}\mathbf{z}_i \in \text{span}(\mathbf{B}),$$

where  $\text{span}(\mathbf{B})$  is the linear span of column vectors of  $\mathbf{B}$ . With this model,  $\mathbf{x}_i$  is roughly distributed around a  $r$ -dimensional subspace (if  $\mathbf{B}$  has full column rank), and the coordinates of  $\mathbf{x}_i$  on that subspace are determined by the factors  $\mathbf{z}_i$ . Thus,  $\text{span}(\mathbf{B})$  can be viewed as the intrinsic geometric characteristics of this model. Once this subspace is determined, the degree of freedom is reduced from  $O(pr)$  to  $O(r^2)$ , and then statistical inference on  $\mathbf{B}$  and  $\mathbf{z}_i$  becomes easier. For an overview on dimensionality reduction, see [Li \(2018\)](#).

Another family of models that receives much attention recently is mixture models. Despite having a long history, until recently theoretical analysis about initialization and convergence has been elusive ([Anandkumar et al., 2014](#); [Balakrishnan et al., 2017](#)). Consider a simple mixed linear model:

$$y_i = \sum_{k=1}^K \mathbf{1}\{z_i = k\} (\mathbf{x}_i^T \boldsymbol{\beta}_k + \epsilon_i),$$

where  $\mathbf{x}_i, y_i$  are observed and  $z_i$ , taking values in  $\{1, 2, \dots, K\}$ , is not observed. A special case is  $K = 1$ , in which  $z_i$  is a constant, and the model reduces to the linear regression model. The difficulty for general  $K$  stems from the unobserved latent variable  $z_i$ . Although it is not easy to estimate and analyze each  $\boldsymbol{\beta}_k$ , the subspace  $\text{span}\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K\}$  can be estimated fairly easily using the first and second moments, which is important for subsequent estimation of each  $\boldsymbol{\beta}_k$  ([Yi et al., 2016](#); [Sedghi et al., 2016](#)). More generally, multiple index models assume a semiparametric model

$$y_i = G(\mathbf{x}_i^T \boldsymbol{\beta}_1, \dots, \mathbf{x}_i^T \boldsymbol{\beta}_K, \epsilon_i), \quad (2)$$

where the form of  $G$  is unknown. In all these examples, the subspace

$$\mathcal{S} = \text{span}(\mathbf{B}) \quad \text{or} \quad \mathcal{S} = \text{span}\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K\}$$

plays a pivotal role, since it captures and summarizes the information of one part of the

variables (often covariates) in relation to the other. This motivates us to consider the problem of estimating  $\mathcal{S}$  alone in a general setting, which we call *subspace estimation* in this paper.

The advantage of studying this problem is three-fold. (1) The subspace  $\mathcal{S}$  is intrinsic geometrically, which is invariant to rotation, and therefore there is no identifiability issue for subspace estimation. (2) After obtaining a good estimate of  $\mathcal{S}$ , it is easier for statisticians to visualize data, to reduce data dimensions, to estimate unknown parameters (with a largely reduced degree of freedom), and to study model diagnostics. In this aspect, subspace estimation can be viewed as an intermediate estimation problem. (3) It is less difficult to estimate a subspace than unknown parameters, because instead of finding the maximum likelihood estimator (MLE) or running an EM algorithm, we need only moment conditions from the data, which are much easier to compute.

Relevant to the third point, often we find ourselves in situations where we have many ways to estimate  $\mathcal{S}$ . For example, in factor models, the covariance matrix is usually used to determine  $\text{span}(\mathbf{B})$ ; however, if  $\mathbf{x}_i$  has nonzero means, the non-centered form is also informative since  $\mathbb{E}\mathbf{x}_i \in \text{span}(\mathbf{B})$ . This observation has led to a recent work by [Lettau and Pelger \(2017\)](#). And in multiple index models, it is known that transforming the moments can be helpful ([Li, 1992](#)). These considerations lead naturally to our problem formulation.

## 1.2 Problem formulation

Suppose we have data  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ , where  $y_i$  is the response variable,  $\mathbf{x}_i$  is the vector of covariates (or predictors), and  $n$  is the sample size. A special case is that all  $y_i$  are set to a constant, or equivalently we omit  $y_i$  altogether (a.k.a. unsupervised learning). Assume that  $(\mathbf{x}_i, y_i)$  is i.i.d., with an unknown distribution  $P$ . Our goal is to estimate an unknown linear subspace  $\mathcal{S} = \mathcal{S}(P)$  of  $\mathbb{R}^p$  with dimension  $\dim(\mathcal{S}) = r \geq 1$ , where  $r$  is fixed and unknown, from a set of overidentifying conditions:

$$\mathbf{v}_\ell = \frac{1}{n} \sum_{i=1}^n \mathbf{f}_\ell(\mathbf{x}_i, y_i) \in \mathbb{R}^p, \quad \ell \in [m], \quad (3)$$

$$\text{with } \mathbb{E}\mathbf{v}_\ell \in \mathcal{S}, \quad \forall \ell \in [m], \quad (4)$$

where  $m \geq r$ , and  $\mathbf{f}_1(\mathbf{x}_i, y_i), \dots, \mathbf{f}_m(\mathbf{x}_i, y_i)$  are known functions with finite second moments. Often, the vector  $\mathbf{v}_\ell$  is the empirical moments of  $\mathbf{x}_i$  and  $y_i$ , but our definition here is very general, as  $\mathbf{f}_\ell$  is a generic function. In general, we have more than enough conditions to determine  $\mathcal{S}$ , since  $m \geq r$  and the linear span of  $\mathbb{E}\mathbf{v}_1, \dots, \mathbb{E}\mathbf{v}_m$  is exactly  $\mathcal{S}$  except for degenerate cases. Thus, it is reasonable to expect a good estimator  $\hat{\mathcal{S}}$  from the statistics  $\mathbf{v}_1, \dots, \mathbf{v}_m$ . The question, then, is how to produce an estimator in an *optimal* way.

Note that we do not make assumptions on  $\mathbf{x}_i$  or  $y_i$  directly, other than the i.i.d. assumption. Assumptions will be made, and results will be stated, in terms of  $\mathbf{f}_\ell(\mathbf{x}_i, y_i)$ . Also note that  $\mathbf{v}_\ell$  is not required to have unit  $\ell_2$  norm, since it is implicitly rescaled in our procedure (see Section 2.3).

### 1.3 Combining overidentifying vectors optimally

Let us approach this problem by first assuming  $r$  is known; otherwise, it can be consistently estimated (see Section 4.3). In general, to determine a subspace of dimension  $r$ , we need  $r$  linearly independent vectors. Now given a possibly large pool of  $\mathbf{v}_\ell$ , we have to look for the common and dominant space. A natural way is to consider singular vectors. Let  $\text{svd}_r(\cdot)$ , and respectively  $\text{eigen}_r(\cdot)$ , denote top  $r$  left singular vectors and eigenvectors of a matrix. Suppose we compute

$$\text{svd}_r([\mathbf{v}_1, \dots, \mathbf{v}_m]) \quad \text{or equivalently} \quad \text{eigen}_r\left(\sum_{\ell=1}^m \mathbf{v}_\ell \mathbf{v}_\ell^T\right).$$

Then the resulting  $r$  singular (or eigen-)vectors make use of all given moment conditions. Had these  $\mathbf{v}_\ell$  been i.i.d., this method would be the same as computing the principal components. However, we do not have the luxury to make such assumptions. More often than not, they have different variances, and they may be arbitrarily dependent. To take it into account, we modify our method with a symmetric weighting matrix  $\mathbf{W} = (w_{j\ell}) \in \mathbb{R}^{m \times m}$ , and compute

$$\text{svd}_r([\mathbf{v}_1, \dots, \mathbf{v}_m] \mathbf{W}^{1/2}) \quad \text{equivalently} \quad \text{eigen}_r\left(\sum_{j,\ell=1}^m w_{j\ell} \mathbf{v}_j \mathbf{v}_\ell^T\right) = \text{eigen}_r(\mathbf{V} \mathbf{W} \mathbf{V}^T),$$

where  $\mathbf{W}^{1/2} \in \mathbb{R}^{m \times m}$  is such that its square equals  $\mathbf{W}$ . This method is the same as first applying a linear transformation  $\mathbf{W}^{1/2}$  to  $[\mathbf{v}_1, \dots, \mathbf{v}_m]$  and then implementing the unweighted method. For any fixed  $\mathbf{W}$ , the transformation results in  $m$  new vectors that satisfy (4). Our task, therefore, is to find an optimal  $\mathbf{W}$ .

As we will show, in the sense of large sample asymptotic theory, the optimal choice of  $\mathbf{W}$

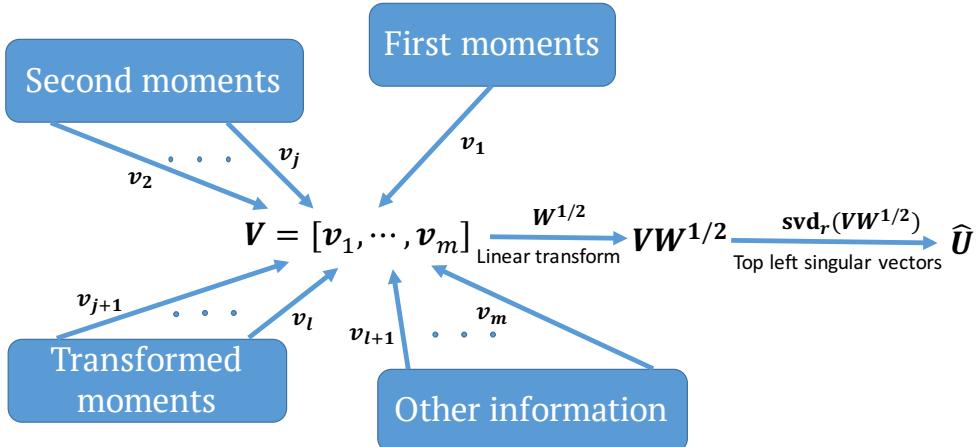


Figure 1: A diagram showing the procedure

is the inverse of certain pseudo-covariance matrix of these vectors (or equivalently  $\mathbf{f}_\ell$ )—see (10). Intuitively, we scale down vectors with large variances, and scale them up otherwise. Our optimality results for such  $\mathbf{W}$  include, for example, the smallest asymptotic expectation of

$$d(\hat{\mathcal{S}}, \mathcal{S})^2 = \|\mathcal{P}_{\hat{\mathcal{S}}} - \mathcal{P}_{\mathcal{S}}\|_F^2,$$

where  $\mathcal{P}_{\hat{\mathcal{S}}}, \mathcal{P}_{\mathcal{S}}$  are projections to a subspace. This criterion is similar to the mean squared error, and can be also expressed in terms of canonical angles—see (23).

Our analysis is based on, but not directly derived from, the generalized method of moments (GMM). Although PCA for subspace estimation has a long history and a well-established theory (Pearson, 1901; Hotelling, 1933; Anderson, 1963), overidentifying vectors are much less studied. General optimality results of GMM are developed in the seminal paper by Hansen (1982), which, however, does not directly apply to our problem. While GMM is widely studied both theoretically and empirically, previous works mostly tackle problems where the MLE of unknown parameters is not available or very difficult to compute. Our paper shows that GMM with a suitable form of weighting matrices produces a simple closed-form estimator, which is useful as an intermediate estimator.

We shall call our proposed method that uses the optimal  $\mathbf{W}$  the *GMM subspace estimator*.

## 1.4 Related works and paper organization

The paper is organized as follows. In Section 2, we propose our estimation procedure whose applications are elucidated in Section 3. In Section 4, we present our theoretical analysis and show optimality of our procedure under nearly minimal assumptions. Numerical simulations and a dataset example are presented in Section 5 and 6 to support our theory. Finally, in Section 7, we discuss possible extensions.

# 2 Subspace GMM estimator

We derive our subspace estimator in a way similar to the classical GMM theory. However, there are several departures: (1) matrix representation of subspaces involves identification of matrices up to rotation, since inherently the parameter space is a Grassmann manifold; (2) a specific blockwise form of the weighting matrix allows simple and fast computation, while enjoying optimality results—see Sections 2.2 and 4.2; (3) for this particular estimation problem, we develop clear methods and analysis for subtle issues, such as singularity of weighting matrices—see Section 4.4.

## 2.1 Estimation via GMM framework

Before formally deriving the GMM estimator, we first address the representation of the subspace  $\mathcal{S}$ . Let  $O(p, r)$  be the set of  $p \times r$  matrices consisting of orthonormal column

vectors. Every  $\mathcal{S}$  is associated with some  $\mathbf{U}^* \in O(p, r)$  whose columns lie in  $\mathcal{S}$ , that is,

$$\mathbf{U}^* = [\mathbf{u}_1^*, \mathbf{u}_2^*, \dots, \mathbf{u}_r^*] \in \mathbb{R}^{p \times r}, \quad \text{with } (\mathbf{U}^*)^T \mathbf{U}^* = \mathbf{I}_r, \text{ and } \mathbf{u}_j^* \in \mathcal{S}, \forall j \in [r]. \quad (5)$$

Such matrix representation  $\mathbf{U}^*$  is unique up to a rotation: a  $p \times r$  matrix satisfies the conditions in (5) if and only if it has the form  $\mathbf{U}^* \mathbf{R}$ , where  $\mathbf{R} \in O(r)$  is an orthogonal matrix of size  $r$ . This is because any two sets of orthonormal bases can be mapped to each other by an orthogonal matrix. It is clear, then, that the projection matrix  $\mathbf{P}_{\mathbf{U}^*}^\perp := \mathbf{I}_p - \mathbf{U}^* (\mathbf{U}^*)^T \in \mathbb{R}^{p \times p}$  is unique.

Therefore, we can represent the set of all possible  $\mathcal{S}$  as  $O(p, r)/O(r)$ , that is, the space  $O(p, r)$  up to rotation. In differential geometry, the space  $O(p, r)$  is called the Stiefel manifold, and  $O(p, r)/O(r)$  is called the Grassmann manifold (Edelman et al., 1998).

Using the representation of  $\mathcal{S}$ , we can rewrite (4) into the following estimating equations, which are usually called the *population moment condition* in the GMM literature:

$$\mathbb{E}(\mathbf{I}_p - \mathbf{U}^* (\mathbf{U}^*)^T) \mathbf{v}_\ell = \mathbf{0}, \quad \forall \ell \in [m].$$

Now it is natural to introduce the GMM estimator as follows. For any matrix  $\mathbf{U}^* \in O(p, r)$  with orthonormal columns, we concatenate all vectors  $(\mathbf{I}_p - \mathbf{U} \mathbf{U}^T) \mathbf{v}_\ell$  into a single vector:

$$\mathbf{g}(\mathbf{U}) := \begin{pmatrix} (\mathbf{I}_p - \mathbf{U} \mathbf{U}^T) \mathbf{v}_1 \\ \vdots \\ (\mathbf{I}_p - \mathbf{U} \mathbf{U}^T) \mathbf{v}_m \end{pmatrix} \in \mathbb{R}^{\bar{m}}, \quad \text{where } \bar{m} := mp. \quad (6)$$

We also denote  $\bar{p} = rp$ . Thus, the function  $\mathbf{g}$  is a nonlinear map from  $\mathbb{R}^{\bar{p}}$  to  $\mathbb{R}^{\bar{m}}$ . For any positive definite matrix  $\mathbf{W} = (w_{k\ell})_{k, \ell \in [m]} \succ \mathbf{0}$  in  $\mathbb{R}^{m \times m}$ , we define the weighting matrix as

$$\overline{\mathbf{W}} = \mathbf{W} \otimes \mathbf{I}_p = \begin{pmatrix} w_{11} \mathbf{I}_p & \cdots & w_{1p} \mathbf{I}_p \\ \vdots & \ddots & \vdots \\ w_{1m} \mathbf{I}_p & \cdots & w_{mm} \mathbf{I}_p \end{pmatrix} \in \mathbb{R}^{\bar{m} \times \bar{m}}, \quad (7)$$

where  $\otimes$  is the Kronecker product. By construction,  $\overline{\mathbf{W}}$  is also a definite positive matrix (see Lemma 2). Here, both  $\mathbf{W}$  and  $\overline{\mathbf{W}}$  can be random. Following the classical GMM approach (Hansen, 1982), we define the GMM estimator  $\widehat{\mathbf{U}}$  as a minimizer of  $Q(\mathbf{U})$ , which is quadratic in  $\mathbf{g}(\mathbf{U})$  and thus quartic (i.e., involving fourth moments) in  $\mathbf{U}$ .

$$\widehat{\mathbf{U}} \in \operatorname{argmin}_{\mathbf{U} \in O(p, r)} Q(\mathbf{U}) \quad \text{where } Q(\mathbf{U}) := [\mathbf{g}(\mathbf{U})]^T \overline{\mathbf{W}} \mathbf{g}(\mathbf{U}). \quad (8)$$

A few remarks are in order. First, as a minimizer of (8),  $\widehat{\mathbf{U}}$  is not uniquely defined. It is clear that  $\widehat{\mathbf{U}}$  is a minimizer of (8) if and only if  $\widehat{\mathbf{U}} \mathbf{R}$  is also a minimizer for any  $\mathbf{R} \in O(r)$ . An alternative way that would circumvent this issue is to define  $Q$  as a function of  $\mathbf{U} \mathbf{U}^T$ , since a minimizer of this function does not depend on the choice of  $\mathbf{R}$ . However, for the

ease of expositions and analysis, we focus on the current form defined in (8). Second, in the minimization problem, the parameter space  $\{\mathbf{U} : \mathbf{U} \in O(p, r)\}$  is a Stiefel manifold in  $\mathbb{R}^{pr}$ , which has dimension  $pr - r(r + 1)/2$ ; whereas, the standard GMM framework usually assumes the parameter space contains some neighborhood (or ball) around  $\mathbf{U}^*$  (Hall, 2005). Third, relevant to the first two remarks, one may consider defining  $\widehat{\mathbf{U}}$  as a minimizer of  $Q$  over the Grassmann manifold, which also resolved the non-uniqueness issue in the first remark. However, we avoid such treatment here due to heavy machinery from differential manifolds.

## 2.2 Computing $\widehat{\mathbf{U}}$ via eigendecomposition

With the block matrix form of  $\overline{\mathbf{W}}$ , we are able to simplify the optimization problem and compute  $\widehat{\mathbf{U}}$  via the standard eigen-decomposition computation.

We observe that although the objective function in (8) has a quartic form, it is equivalent to a quadratic function, and consequently, the optimization problem (8) can be solved very efficiently. This is due to the fact that  $\mathbf{I}_p - \mathbf{U}\mathbf{U}^T$  is a projection matrix, so  $(\mathbf{I}_p - \mathbf{U}\mathbf{U}^T)^2 = \mathbf{I}_p - \mathbf{U}\mathbf{U}^T$ , and

$$\begin{aligned} [\mathbf{g}(\mathbf{U})]^T \overline{\mathbf{W}} \mathbf{g}(\mathbf{U}) &= \sum_{k,\ell=1}^m w_{k\ell} \mathbf{v}_k^T (\mathbf{I}_p - \mathbf{U}\mathbf{U}^T) \mathbf{v}_\ell = \sum_{k,\ell=1}^m w_{k\ell} \mathbf{v}_k^T \mathbf{v}_\ell - \sum_{k,\ell=1}^m w_{k\ell} \text{Tr}(\mathbf{U}^T \mathbf{v}_\ell \mathbf{v}_k^T \mathbf{U}) \\ &= \sum_{k,\ell=1}^m w_{k\ell} \mathbf{v}_k^T \mathbf{v}_\ell - \text{Tr}\left(\mathbf{U}^T \sum_{k,\ell=1}^m w_{k\ell} \mathbf{v}_\ell \mathbf{v}_k^T \mathbf{U}\right), \end{aligned}$$

where  $\text{Tr}(\cdot)$  is the trace of a matrix. Hence, we obtain the following result.

**Proposition 1.** *Solving the optimization problem (8) is equivalent to solving*

$$\max_{\mathbf{U} \in O(p,r)} \text{Tr}(\mathbf{U}^T \mathbf{V} \mathbf{W} \mathbf{V}^T \mathbf{U}), \quad \text{where } \mathbf{V} := [\mathbf{v}_1, \dots, \mathbf{v}_m] \in \mathbb{R}^{p \times m}. \quad (9)$$

The columns of its solution  $\widehat{\mathbf{U}}$  are given by the top  $r$  eigenvectors of  $\mathbf{V} \mathbf{W} \mathbf{V}^T$ .

This proposition provides a way of computing  $\widehat{\mathbf{U}}$  given  $\mathbf{W}$ . Here ‘top eigenvectors’ refer to those eigenvectors with largest eigenvalues. Under a nondegeneracy assumption (Assumptions 1),  $\mathbf{V} \mathbf{W} \mathbf{V}^T$  has a non-vanishing gap between its  $r$ th and  $(r+1)$ -th largest eigenvalues for large  $n$ , and thus  $\widehat{\mathbf{U}}$  is unique up to rotation. (A large eigen-gap also ensures numerical stability.) See details in Section 4.3.

We remark that the above simplification hinges on the block matrix form of  $\overline{\mathbf{W}}$ . For a general  $\overline{m} \times \overline{m}$  weighting matrix  $\overline{\mathbf{W}}$ , there is no simple way to solve (8), which is a genuine quartic function in  $\mathbf{U}$ .

## 2.3 Two-step estimation procedure

With an appropriate choice of the weighting matrix, the GMM produces, in general, asymptotically efficient estimators. This can be usually achieved through a two-step estimation procedure: the first step is obtaining a consistent estimator, which is used to compute an optimal weighting matrix  $\mathbf{W}$ ; and the second step is to solve an optimization problem with the  $\mathbf{W}$  computed from the first step. The first step is often easy, and in particular it is so for our problem, since eigen-decomposition of  $\mathbf{V}\mathbf{W}\mathbf{V}^T$  with any  $\mathbf{W} \succ \mathbf{0}$  leads to a consistent estimator. In the second step, we choose  $\mathbf{W}$  in a way such that it converges in probability (as  $n \rightarrow \infty$ ) to  $(\Sigma^*)^{-1}$ , where  $\Sigma^* = (\Sigma_{j\ell}^*) \in \mathbb{R}^{m \times m}$  is defined as

$$\Sigma_{j\ell}^* = \mathbb{E} [\mathbf{f}_j(\mathbf{x}_i, y_i)^T (\mathbf{I}_p - \mathbf{U}^*(\mathbf{U}^*)^T) \mathbf{f}_\ell(\mathbf{x}_i, y_i)], \quad \forall j, \ell \in [m]. \quad (10)$$

Note that this definition does not depend on  $i$ . In the matrix form, it is the same as

$$\Sigma^* = \mathbb{E} [\mathbf{F}_i^T (\mathbf{I}_p - \mathbf{U}^*(\mathbf{U}^*)^T) \mathbf{F}_i], \quad \text{where } \mathbf{F}_i = [\mathbf{f}_1(\mathbf{x}_i, y_i), \dots, \mathbf{f}_m(\mathbf{x}_i, y_i)] \in \mathbb{R}^{p \times m}.$$

The particular structure of  $\Sigma^*$  is closely related to the covariance matrix of  $\mathbf{f}_1, \dots, \mathbf{f}_m$ , and it comes with optimality guarantees—see Section 4.1 and 4.2. Given an initial consistent  $\widehat{\mathbf{U}}^0$ , we can find a consistent estimator of  $\Sigma^*$ , denoted by  $\widehat{\Sigma}$ , using the natural plug-in estimator, that is, the sample mean of  $\mathbf{f}_j(\mathbf{x}_i, y_i)^T (\mathbf{I}_p - \widehat{\mathbf{U}}^0(\widehat{\mathbf{U}}^0)^T) \mathbf{f}_\ell(\mathbf{x}_i, y_i)$  for all  $j, \ell$ , namely,

$$\widehat{\Sigma} = n^{-1} \sum_{i=1}^n \mathbf{F}_i^T (\mathbf{I}_p - \widehat{\mathbf{U}}^0(\widehat{\mathbf{U}}^0)^T) \mathbf{F}_i. \quad (11)$$

Then, setting  $\mathbf{W}$  to be  $(\widehat{\Sigma})^{-1}$  and solving the eigen-decomposition (1) again, we obtain the final GMM estimator, denoted by  $\widehat{\mathbf{U}}^{\text{GMM}}$ . Our estimation procedure is formally described as follows:

1. Obtain an initial consistent estimator  $\widehat{\mathbf{U}}^0$ . For example, one can choose an initial weighting matrix  $\mathbf{W} = \mathbf{I}_p$  in (9) and compute top  $r$  eigenvectors of  $\mathbf{V}\mathbf{V}^T$ .
2. For each  $j, \ell \in [m]$  with  $j \leq \ell$ , calculate

$$\widehat{\Sigma}_{j\ell} = \frac{1}{n} \sum_{i=1}^n [\mathbf{f}_j(\mathbf{x}_i, y_i)^T (\mathbf{I}_p - \widehat{\mathbf{U}}^0(\widehat{\mathbf{U}}^0)^T) \mathbf{f}_\ell(\mathbf{x}_i, y_i)] \quad (12)$$

and set  $\widehat{\Sigma}_{\ell j} = \widehat{\Sigma}_{j\ell}$ . Form a matrix  $\widehat{\Sigma} = (\widehat{\Sigma}_{j\ell}) \in \mathbb{R}^{m \times m}$ , which is equivalent to computing (11) in matrix form.

3. If  $r = \dim(\mathcal{S})$  is not known in advance, estimate  $r$  as suggested in Section 4.3.
4. Set  $\mathbf{W}$  to be  $(\widehat{\Sigma})^{-1}$  or (13) below, and then compute the top  $r$  eigenvectors of  $\mathbf{V}\mathbf{W}\mathbf{V}^T$ . Obtain the estimator  $\widehat{\mathbf{U}}^{\text{GMM}} \in \mathbb{R}^{p \times r}$  by combining these eigenvectors into a matrix.

We make a fews remarks about the above estimation procedure. First, in principle there are many ways to produce a consistent  $\widehat{\mathbf{U}}^0$ . One may choose a subset of vectors from  $\mathbf{v}_1, \dots, \mathbf{v}_m$ , if it is known that such subset spans the target subspace  $\mathcal{S}$ . Often, standard (or vanilla) estimators serve as good initial estimates. An alternative approach, as often used in the GMM literature, is to conduct iterative GMM, which is to estimate a sequence of  $\widehat{\mathbf{U}}^t$ , where  $\widehat{\mathbf{U}}^t$  is calculated based on  $\widehat{\mathbf{U}}^{t-1}$ .

Second, in cases where  $r$  is not known, one may follow Section 4.3 to estimate  $r$ , and the method only involves the eigenvalues of  $\mathbf{V}\mathbf{W}\mathbf{V}^T$ . Thus, Step 3 does not require additional computational cost.

Third, if  $\widehat{\Sigma}$  is singular (namely, not invertible) or nearly singular, the following variant is preferred over  $(\widehat{\Sigma})^{-1}$ . Given a parameter  $\delta_n \geq 0$ , compute the eigen-decomposition of  $\widehat{\Sigma} = \overline{\mathbf{U}} \text{diag}\{\bar{\lambda}_1, \dots, \bar{\lambda}_m\} \overline{\mathbf{U}}^T$ , and set

$$\mathbf{W} = \overline{\mathbf{U}} \text{diag}\{\psi(\bar{\lambda}_1), \dots, \psi(\bar{\lambda}_m)\} \overline{\mathbf{U}}^T, \quad \text{where } \psi(x) := x^{-1} \mathbf{1}\{x > \delta_n\}. \quad (13)$$

It generalizes the usual matrix inverse: if  $\delta_n = 0$ , then (13) gives the Moore-Penrose pseudoinverse  $(\widehat{\Sigma})^+$ ; and if  $\delta_n > 0$ , then (13) applies a de-noising step with a threshold  $\delta_n$  before taking the Moore-Penrose pseudoinverse. In this regard, (13) can be viewed as computing the pseudoinverse with a hard-thresholding operation. We suggest choosing  $\delta_n$  such that  $\delta_n = o(1)$  and  $\sqrt{n} \delta_n \rightarrow \infty$ . See the analysis in Section 4.4,

As shown in Section 4.4, the presence of redundant vectors is often the cause of singularity of  $\widehat{\Sigma}$ , in which case using the generalized inverse (13) still gives optimality guarantee with a suitable  $\delta_n$ .

## 2.4 Extensions

So far, we have considered combining information that has an average form as required by (4). Typically, this is useful when such  $\mathbf{v}_\ell$  is derived from the method of moments. However, our framework can be used to accommodate the information of  $\mathcal{S}$  which does not have the average form. For example, in the factor model, any single  $\mathbf{x}_i$  satisfies  $\mathbb{E}\mathbf{x}_i = \mathbf{B}\mathbb{E}\mathbf{z}_i \in \mathcal{S}$ , but it does not admit the average form. This kind of individual information can also be incorporated in our GMM framework (9) under certain structure.

In order to apply our asymptotic theory, we assume that we have individual moment information  $\mathbf{m}_i \in \mathcal{S}$  based on the  $i$ th data. For the factor model, we naturally take  $\mathbf{m}_i = \mathbf{x}_i$ . These individual moments are naturally aggregated as the matrix  $\mathbf{M} = n^{-1} \sum_{i=1}^n \mathbf{m}_i \mathbf{m}_i^T$ . Under additional structure that is similar to the factors, we may assume that the eigenspace of  $\mathbb{E}\mathbf{M} - \Sigma_0$  falls in  $\mathcal{S}$  for some given symmetric  $\Sigma_0$ . We can now take the moment conditions

$$\mathbf{v}_\ell = (\mathbf{M} - \Sigma_0) \mathbf{e}_\ell = n^{-1} \sum_{i=1}^n (\mathbf{m}_i \mathbf{m}_i^T \mathbf{e}_\ell - \Sigma_0 \mathbf{e}_\ell),$$

which has the form of averages. Therefore, it falls in our framework. The unweighted

aggregation of these moments in Proposition 1 leads to the principal component analysis of the matrix

$$\sum_{\ell=1}^n \mathbf{v}_\ell \mathbf{v}_\ell^T = (\mathbf{M} - \boldsymbol{\Sigma}_0)^2.$$

We remark that the contribution of the matrix  $\boldsymbol{\Sigma}_0$  is negligible under the pervasive conditions (Bai, 2003; Fan et al., 2013). Hence, one may simply use the eigenspace spanned the matrix  $\mathbf{M}$  to obtain reasonable estimates, while our two-step method allows for more efficient construction.

More generally, suppose that we are given a symmetric matrix  $\mathbf{M} \in \mathbb{R}^{p \times p}$  that is computed from the data such that  $\text{span}(\text{eigen}_r(\mathbf{M}))$  is a consistent estimator of  $\mathcal{S}$ . A natural way to combine  $\mathbf{M}$  with our GMM estimator is to compute

$$\text{eigen}_r(\kappa \mathbf{M} + \mathbf{V} \mathbf{W} \mathbf{V}^T)$$

where  $\kappa$  is a suitable parameter. The hope is that, a good choice of  $\kappa$  may lead to a more efficient estimator.

While it seems difficult to determine an optimal parameter  $\kappa$  in general, we identity cases in Section 4.5, in which introducing  $\kappa \mathbf{M}$  does not bring extra asymptotic efficiency; or in other words, the optimal  $\kappa = 0$ . In practice, one may consider a few choices of values for  $\kappa$ , and use the bootstrap to determine the best  $\kappa$ .

## 3 Examples

We discuss four typical applications to exemplify the general procedure proposed in the previous section.

### 3.1 Factor models

First, let us consider the simple factor model (1), where,  $\mathbf{x}_i, \boldsymbol{\epsilon}_i \in \mathbb{R}^p$  and  $\mathbf{z}_i \in \mathbb{R}^r$  are random vectors, and  $\mathbf{B} \in \mathbb{R}^{p \times r}$  is a fixed and unknown matrix. Only  $\mathbf{x}_i$  is observed, and  $\mathbf{z}_i$  and  $\boldsymbol{\epsilon}_i$  represent, respectively, latent factors and noise. Suppose  $\{\mathbf{z}_i, \boldsymbol{\epsilon}_i\}_{i=1}^n$  are i.i.d., and all vectors are jointly independent. Let  $\boldsymbol{\mu}_{\mathbf{z}} = \mathbb{E}(\mathbf{z}_i)$  and  $\boldsymbol{\Sigma}_{\mathbf{z}} = \text{Cov}(\mathbf{z}_i)$ . We also assume that  $\mathbb{E}\boldsymbol{\epsilon}_i = \mathbf{0}$ , and  $\text{Cov}(\boldsymbol{\epsilon}_i) = \sigma^2 \mathbf{I}_p$ , where  $\sigma$  is known.

This model and its variants are well studied, and have wide applications in econometrics, psychology, etc. Usually,  $r$  is much smaller than  $p$ . Under this model, the  $p$  dimensional vector  $\mathbf{x}_i$  is predominately determined by a linear combination of a small number of ‘factors’  $\mathbf{z}_i$ . The covariance of  $\boldsymbol{\epsilon}_i$  is assumed to be simple in this paper, while more sophisticated and general structures have been considered (Connor and Korajczyk, 1993; Forni et al., 2000; Bai, 2003; Fan et al., 2013; Bai and Ng, 2013; Fan et al., 2016; Forni et al., 2017). Nevertheless, our simple factor model retains essential features, as dimension reduction via principal component analysis is routinely employed in the literature.

Let  $\mathcal{S} = \text{span}(\mathbf{B})$  be the target subspace we want to estimate. It is the left singular subspace of  $\mathbf{B}$ , and is also the subspace spanned by principal component directions. Once we have a good estimate of  $\mathcal{S}$ , it is much easier to estimate  $\mathbf{B}$ , since the degree of freedom reduces from  $pr$  to  $r^2$ . A nice property of  $\mathcal{S}$  is that there is no identifiability issue as is common in factor models, since  $\text{span}(\mathbf{BQ}) = \text{span}(\mathbf{B})$  holds for any invertible matrix  $\mathbf{Q}$ .

Routinely, eigen-decomposition of the empirical covariance matrix of  $\mathbf{x}_i$ , namely  $n^{-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$  where  $\bar{\mathbf{x}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i$ , forms the cornerstone of estimation of  $\mathcal{S}$ . If  $\boldsymbol{\mu}_{\mathbf{z}} \neq \mathbf{0}$ , however, the first moments supply complementary information to the covariance matrix (second moments). To see this, notice

$$\mathbb{E}\mathbf{x}_i = \mathbf{B}\boldsymbol{\mu}_{\mathbf{z}} \in \mathcal{S}.$$

Intuitively, if  $\boldsymbol{\mu}_{\mathbf{z}}$  is very large, then  $\bar{\mathbf{x}}$  is useful to estimate  $\mathcal{S}$ . This motivates combining both the first and second moments. To this end, we set

$$\mathbf{v}_1 = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i, \quad \mathbf{v}_{1+\ell} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \mathbf{e}_\ell - \sigma^2 \mathbf{e}_\ell, \quad \forall \ell \in [p], \quad (14)$$

where  $\{\mathbf{e}_\ell\}$  is the standard basis. Here  $\mathbf{v}_{1+\ell}$  is simply a projection of the matrix  $n^{-1} \sum_i \mathbf{x}_i \mathbf{x}_i^T - \sigma^2 \mathbf{I}_p$  to each basis vector  $\mathbf{e}_\ell$ . We can then follow the procedure outlined in Section 2.3.

Note that, we can replace  $\{\mathbf{e}_\ell\}$  by any set of  $p$  linearly independent vectors, and the same optimality guarantee holds. This is because a change of basis only results in a linear transformation of moments, and our theory ensures optimality among all linear transformations (Section 4.2). Moreover, replacing  $\{\mathbf{e}_\ell\}$  by a set of overcomplete vectors does not give a better estimate, since projections onto overcomplete vectors only provide redundant information (Section 4.4).

We also note that in (14), we assume  $\sigma$  is known, so we can subtract  $\sigma^2 \mathbf{e}_\ell$  to remove the effect from the noise term. This ensures  $\mathbb{E}\mathbf{v}_{1+\ell} \in \mathcal{S}$  and thus our theory is applicable. In the case of an unknown  $\sigma$ , one may consider estimating  $\sigma$  first, or updating estimates iteratively.

Finally, we remark that recently, [Lettau and Pelger \(2017\)](#) also utilize first moments, along with second moments, to achieve improved estimation in factor models. Both theoretical and empirical evidence are shown to justify the use of first moments. In Section 4.5, we will discuss the connection to this work.

## 3.2 Mixture models

Now let us consider mixtures of generalized linear models (GLM). Suppose we have i.i.d. data  $\mathbf{x}_i$  and  $y_i$ , with  $\mathbf{x}_i \in \mathbb{R}^p$ ,  $y_i \in \mathbb{R}$  and  $i \in [n]$ . An unobserved variable  $z_i \in \{1, 2, \dots, K\}$ , independent of  $\mathbf{x}_i$  and  $y_i$ , indicates which model  $\mathbf{x}_i$  and  $y_i$  are generated from. To be precise, each  $z_i$  is a multinomial variable with  $\mathbb{P}(z_i = k) = \pi_k$ , where  $\pi_k$  is a parameter; and

conditioning on  $z_i$  and  $\mathbf{x}_i$ ,  $y_i$  has a density function (or probability mass function):

$$f(y|\mathbf{x}_i, z_i = k; \Theta) = h_k(y; \theta_k(\mathbf{x}_i)), \quad \text{where } \theta_k(\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}_k + \beta_{k0}. \quad (15)$$

Here,  $\boldsymbol{\beta}_k \in \mathbb{R}^p$  and  $\beta_{k0} \in \mathbb{R}$  are the unknown parameters of the  $k$ th GLM model. The parameter space  $\Theta$  is the set of all these parameters. For each  $i \in [n]$ , depending on the value  $z_i$  takes, the response variable  $y_i$  is generated from one of the  $K$  GLMs. Special cases include:

- mixed linear regression

$$y_i = \sum_{k=1}^K \mathbf{1}\{z_i = k\} (\mathbf{x}_i^T \boldsymbol{\beta}_k + \beta_{k0} + \sigma_k \epsilon_i), \quad \text{where } \sigma_k \geq 0; \quad (16)$$

- mixed logistic regression

$$\mathbb{P}(y_i = 1 | \mathbf{x}_i, z_i = k) = \varphi(\mathbf{x}_i^T \boldsymbol{\beta}_k + \beta_{k0}), \quad \text{where } \varphi(t) = \frac{1}{1 + e^{-t}}. \quad (17)$$

Mixtures models and a broader family of latent variable models can be tackled by EM algorithms (Dempster et al., 1977). While local convergence is established under various conditions (Wu, 1983; Xu and Jordan, 1996; Balakrishnan et al., 2017), EM algorithms are usually susceptible to local minima (Jin et al., 2016). In recent years, tensor-based methods, which utilize moments of  $\mathbf{x}_i$  (up to the third order moments), are proved to produce consistent estimators under some conditions on the covariates  $\mathbf{x}_i$  (Anandkumar et al., 2014; Yi et al., 2016). Usually, the covariates are required to be i.i.d. normal, but moderate extensions are possible. A crucial step of these works is to seek a whitening matrix, which is then used to construct tensors that have a special orthogonal structure. The columns of such whitening matrix have the same linear span as  $\text{span}\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K\}$ .

Let  $\mathcal{S} = \text{span}\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K\}$  be our target subspace. Usually  $K$  is much smaller than  $p$ , so a good estimate of  $\mathcal{S}$  is helpful for the estimation of  $\boldsymbol{\beta}_k$ . In the aforementioned papers, typically one can estimate  $\mathcal{S}$  through second moments of  $\mathbf{x}_i$ , assuming  $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{I}_p)$ . For example, in the case of mixed logistic regression, one can use Stein's identity and derive

$$\mathbb{E} [y_i(\mathbf{x}_i \mathbf{x}_i^T - \mathbf{I}_p)] = \sum_{k=1}^K \pi_k \boldsymbol{\beta}_k \boldsymbol{\beta}_k^T \mathbb{E} [\varphi''(\mathbf{x}_i^T \boldsymbol{\beta}_k + \beta_{k0})].$$

Note that each column of  $\mathbb{E} [y_i(\mathbf{x}_i \mathbf{x}_i^T - \mathbf{I}_p)]$  lies in  $\mathcal{S}$ , so it leads to an obvious construction of moments that fit in our framework (3)–(4):

$$\mathbf{v}_\ell = \frac{1}{n} \sum_{i=1}^n y_i (\mathbf{x}_i \mathbf{x}_i^T \mathbf{e}_\ell - \mathbf{e}_\ell), \quad \forall \ell \in [p]. \quad (18)$$

This does not work for mixed linear regression, since  $\mathbb{E}[y_i(\mathbf{x}_i\mathbf{x}_i^T - \mathbf{I}_p)]$  would vanish. A better choice for mixed linear regression is  $\mathbb{E}[y_i^2(\mathbf{x}_i\mathbf{x}_i^T - \mathbf{I}_p)]$ , due to

$$\mathbb{E}[y_i^2(\mathbf{x}_i\mathbf{x}_i^T - \mathbf{I}_p)] = 2 \sum_{k=1}^K \pi_k \boldsymbol{\beta}_k \boldsymbol{\beta}_k^T.$$

This leads to a similar construction as (18), except that  $y_i$  is replaced by  $y_i^2$ .

Apart from these second moments proposed in the literature, one may consider first moments, or moments with  $y_i$  transformed by a nonlinear function. For instance, one may construct

$$\mathbf{v}_\ell = \frac{1}{n} \sum_{i=1}^n h_\ell(y_i) \mathbf{x}_i, \quad \text{or} \quad \mathbf{v}_{\ell,j} = \frac{1}{n} \sum_{i=1}^n h_\ell(y_i) (\mathbf{x}_i \mathbf{x}_i^T \mathbf{e}_j - \mathbf{e}_j), \quad (19)$$

where, for both expressions, one may choose a function  $h_\ell(y_i) = \cos(y_i/t_\ell + \gamma_\ell)$  for any  $t_\ell > 0$  and  $\gamma_\ell \in \mathbb{R}$ . One can also simply choose  $h_\ell(y_i) = y_i$ , leading to the (untransformed) first moments of  $\mathbf{x}_i$ . By Stein's identity, they all satisfy the condition that  $\mathbb{E}\mathbf{v}_\ell$  (or  $\mathbb{E}\mathbf{v}_{\ell,j}$ ) lies in  $\mathcal{S}$ . There are other useful construction in the literature, e.g., Sun et al. (2014).

After constructing these vectors  $\mathbf{v}_\ell$ , we can use our estimation procedure to combine these vectors for optimal asymptotic efficiency, as promised by our theory. In general, the optimality result holds as long as mild regularity conditions (Assumption 1 and 5) are satisfied.

### 3.3 Multiple index models

The mixture model (15) discussed above can be subsumed in multiple index models, which are semiparametric models of the form (2), where  $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K \in \mathbb{R}^p$  are unknown parameters,  $\mathbf{x}_i \in \mathbb{R}^p, y_i \in \mathbb{R}$  are i.i.d. data,  $\epsilon_i$  is unobserved and i.i.d., and the function  $G$  is not known. It is also called the *multi-index model* or *dimension-reduction model*, since usually  $K$  is much smaller than  $p$  and we wish to treat  $\mathbf{x}_i^T \boldsymbol{\beta}_1, \dots, \mathbf{x}_i^T \boldsymbol{\beta}_K$  as new coordinates. A primary interest of this model is  $\text{span}\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K\}$ , denoted again by  $\mathcal{S}$ , since it provides a pathway to dimension reduction, data visualization, estimation of  $g$ , and so on.

Let the regression mean function be

$$\mathbb{E}(y_i|\mathbf{x}_i) = g(\mathbf{x}_i^T \boldsymbol{\beta}_1, \dots, \mathbf{x}_i^T \boldsymbol{\beta}_K).$$

Assume the function  $g$  is twice differentiable, and  $\mathbf{x}_i \sim N(0, \mathbf{I}_p)$ . Well-known estimation methods include sliced inverse regression (Li, 1991), principal Hessian directions (Li, 1992), etc. For example, under our assumptions, the principal Hessian directions (pHd) method seeks to estimate  $\mathbb{E}[y_i(\mathbf{x}_i\mathbf{x}_i^T - \mathbf{I}_p)]$  from the data. (Equivalently, it estimates  $\mathbb{E}[r_i(\mathbf{x}_i\mathbf{x}_i^T - \mathbf{I}_p)]$  where  $r_i = y_i - \beta_0^{LS} - \mathbf{x}_i^T \boldsymbol{\beta}^{LS}$  is the residual after taking out a least square fit). Thus a natural construction is the same as (18). Similar to the previous subsection, we may also

consider utilizing first moments, or applying a nonlinear function  $h$ , which leads to the same form as in (19). These constructions guarantee  $\mathbb{E}\mathbf{v}_\ell \in \mathcal{S}$ , due to Stein's identity.

As pointed out by Li (1992); Cook (1998), a drawback of the pHd is the possibility of vanishing  $\mathbb{E}[y_i(\mathbf{x}_i \mathbf{x}_i^T - \mathbf{I}_p)]$  (or its rank is smaller than  $K$ ). This may occur, for instance, if  $\mathbb{E}[\nabla^2 g(\mathbf{x}_i^T \boldsymbol{\beta}_1, \dots, \mathbf{x}_i^T \boldsymbol{\beta}_K)] = \mathbf{0}$ , which unfortunately includes linear regression. Moreover, any linear trend in  $g$  is missed by pHd, which is an unpleasant feature of pHd. For instance, the direction of  $\boldsymbol{\beta}_2$  is not captured by pHd in the following example:

$$y_i = g_0(\mathbf{x}_i^T \boldsymbol{\beta}_1) + \mathbf{x}_i^T \boldsymbol{\beta}_2 + \epsilon_i.$$

A remedy for pHd is making transformation of  $y_i$  before applying pHd, and some success is reported by (Li, 1992; Sun et al., 2014). In this regard, the vectors constructed through transformation, as suggested in (19), agrees with the aforementioned papers. Besides, our approach can combine transformed moments as before.

### 3.4 Distributed estimation for heterogeneous datasets

As a last example, we consider an estimation problem in a modern setting. Suppose we have  $m$  datasets stored on separate clusters or held by different laboratories/hospitals. Due to communication cost or privacy concerns, we wish to compute statistics locally for each datasets and aggregate these statistics at a central server without accessing the details of these distributed datasets.

Consider the problem of estimating a subspace  $\mathcal{S}$  as before. Let  $n$  be the total sample size; and for ease of exposition, we introduce i.i.d. multinomial variable  $z_i \in \{1, 2, \dots, m\}$  ( $i \in [n]$ ) that indicates which dataset the data unit indexed by  $i$  belongs to. Let  $\rho_\ell := \mathbb{P}(z_i = \ell) \in (0, 1)$  be fixed. Then, the sample size of each dataset is roughly  $\rho_\ell n$ .

Suppose for each  $\ell \in [m]$ , the  $\ell$ th dataset consists of measurements of the form  $\mathbf{f}^{(\ell,k)}(\mathbf{x}_i)$  for all  $i$  such that  $z_i = \ell$ . Here,  $\mathbf{x}_i$  is a random quantity (may or may not observed) associated with, say, a subject with index  $i$  in the  $\ell$ th laboratory with  $k$ th measurement. The total number of measurements for each subject in the  $\ell$ th laboratory is  $K_\ell$ . Each laboratory computes an average of these measurements: for any  $\ell \in [m]$  and  $k \in [K_\ell]$ ,

$$\mathbf{v}^{(\ell,k)} := \frac{1}{|\mathcal{I}_\ell|} \sum_{i \in \mathcal{I}_\ell} \mathbf{f}^{(\ell,k)}(\mathbf{x}_i), \quad \text{where } \mathcal{I}_\ell = \{i : z_i = \ell\}$$

and contributes these vectors to the central server.

These functions  $\mathbf{f}^{(\ell,k)}$  may be different, which reflects different methods of measurements across laboratories. If we assume  $\mathbb{E}\mathbf{v}^{(\ell,k)} \in \mathcal{S}$  and that  $(\mathbf{x}_i; z_i)$  are i.i.d., then we can use our framework to aggregate  $\mathbf{v}^{(\ell,k)}$  across  $\ell$  and  $k$ . To do so, we define  $\mathbf{f}_{\ell,k}(\mathbf{x}_i, z_i) = \mathbf{f}^{(\ell,k)}(\mathbf{x}_i) \mathbf{1}\{z_i = \ell\}$  and rewrite  $\mathbf{v}^{(\ell,k)}$  as

$$\mathbf{v}^{(\ell,k)} = \frac{n}{|\mathcal{I}_\ell|} \cdot \frac{1}{n} \sum_{i=1}^n \mathbf{f}_{\ell,k}(\mathbf{x}_i, z_i).$$

Set  $\mathbf{v}_{\ell,k} := n^{-1}|\mathcal{I}_\ell| \mathbf{v}^{(\ell,k)}$ . It is clear that  $\mathbb{E}\mathbf{v}_{\ell,k} \in \mathcal{S}$ , and our general framework (3)–(4) applies here.

The total number of vectors we have is  $M := K_1 + \dots + K_m$ , and correspondingly  $\Sigma^*$  and  $\mathbf{W}^*$  are of size  $M \times M$ . Note that due to the specific form of  $\mathbf{f}_{\ell,k}(\mathbf{x}_i, z_i)$ ,  $\Sigma^*$  must be a block diagonal matrix, so we only need to estimate each block  $\Sigma_{\ell\ell}^* \in \mathbb{R}^{K_\ell \times K_\ell}$  according to (12), which can be computed locally on each dataset once an initial  $\widehat{\mathbf{U}}_0$  is given. To yield the final estimator, we set  $\mathbf{V}_\ell = [\mathbf{v}_{\ell,1}, \dots, \mathbf{v}_{\ell,K_\ell}] \in \mathbb{R}^{p \times K_\ell}$  and calculate

$$\widehat{\mathbf{U}} = \text{eigen}_r \left( \sum_{\ell=1}^m \mathbf{V}_\ell (\widehat{\Sigma}_{\ell\ell})^{-1} \mathbf{V}_\ell^T \right) \quad (20)$$

where  $\text{eigen}_r(\cdot)$  computes the top  $r$  eigenvectors. Here  $(\widehat{\Sigma}_{\ell\ell})^{-1}$  can be viewed as a local weighting matrix that takes into account the covariance-like information of the  $\ell$ th dataset. A nice property of (20) is that its computation does not require data collection across datasets, and thus may be useful in privacy-sensitive situations.

A special case is  $K_1 = \dots = K_m = 1$ , i.e., all laboratories conduct a single measurement for each of its subjects. In this simple scenario, the weight in (20) simply becomes a scalar.

We remark that in a recent work on distributed estimation for spiked covariance model (Fan et al., 2017), a similar method as (20) is proposed, except there is no weight  $\widehat{\Sigma}_{\ell\ell}^{-1}$  before  $\mathbf{v}_\ell \mathbf{v}_\ell^T$ . While the regimes are different, aggregating  $\mathbf{v}_\ell \mathbf{v}_\ell^T$  seems to be the gist of both methods. Our method, moreover, suggests weighting by  $\widehat{\Sigma}_{\ell\ell}^{-1}$ , which utilizes the variance-like information and thus may be preferable for aggregating heterogeneous datasets.

## 4 Large sample properties

We will establish results about our subspace GMM estimator with an analysis under the classical ‘fixed  $p$  and  $m$ , large  $n$ ’ regime. In this section, to avoid confusion, we explicitly display the dependence on  $n$ , i.e.,  $\widehat{\mathbf{U}}_n = \widehat{\mathbf{U}}$ ,  $\widehat{\Sigma}_n = \widehat{\Sigma}$ ,  $\mathbf{W}_n = \mathbf{W}$ , etc. Proofs can be found in the supplementary materials.

### 4.1 Consistency and asymptotic normality

As the first part of our analysis, we will establish consistency and asymptotic normality, as is often done in the GMM literature. The optimality of asymptotic variance requires a block matrix assumption that could be restrictive in practice. However, if our optimality is gauged not in the original parameter space  $O(p,r)/O(r)$ , but in terms of the canonical angles between two subspaces, then our procedure is superior under fairly mild conditions (Section 4.2).

We embark on our analysis by making a few assumptions. First, we assume that the vectors  $\mathbf{v}_1, \dots, \mathbf{v}_m$  (or equivalently  $\mathbf{f}_1(\mathbf{x}_i, y_i), \dots, \mathbf{f}_m(\mathbf{x}_i, y_i)$ ) contain sufficient information of  $\mathcal{S}$ , in the sense that these vectors, in expectation, span the subspace  $\mathcal{S}$ . Note that

each individual  $\mathbb{E}\mathbf{v}_\ell$  lies in  $\mathcal{S}$ , so an equivalent assumption is stated below in terms of the dimension. In general, this is a mild assumption, since  $m \geq \dim(\mathcal{S}) = r$ .

**Assumption 1** (nondegeneracy). *Suppose  $\dim(\text{span}\{\mathbb{E}\mathbf{v}_1, \dots, \mathbb{E}\mathbf{v}_m\}) = r$ .*

This assumption is a prerequisite for any reasonable estimator of  $\mathcal{S}$ . Indeed, if  $\dim(\mathcal{S}) > \dim(\text{span}\{\mathbb{E}\mathbf{v}_1, \dots, \mathbb{E}\mathbf{v}_m\})$ , even with infinite sample size, it is impossible to uniquely determine  $\mathcal{S}$ . In this regard, Assumption 1 is an identifiability assumption for our problem formulated in (3)–(4).

To state a general consistency result, we consider the following mild condition on the weighting matrix. In particular, in our subspace GMM estimation procedure, as long as  $\Sigma^*$  is invertible, this assumption is satisfied for both the initial estimator (where  $\mathbf{W}_n$  is a constant matrix) and the final GMM estimator (where  $\mathbf{W}_n \xrightarrow{p} (\Sigma^*)^{-1}$ ).

**Assumption 2** (limiting weight matrix). *Suppose  $\mathbf{W}_n$  converges in probability to a non-random positive definite matrix  $\mathbf{W}^* \succ \mathbf{0}$  in  $\mathbb{R}^{m \times m}$ .*

Similar to  $\bar{\mathbf{W}}$  in (7), we define a limiting block matrix  $\bar{\mathbf{W}}^* := \mathbf{W}^* \otimes \mathbf{I}_p \in \mathbb{R}^{\bar{m} \times \bar{m}}$ . It follows from Lemma 2 that  $\bar{\mathbf{W}}^*$  is positive definite under Assumption 2. Since the vectors  $\mathbb{E}\mathbf{v}_1, \dots, \mathbb{E}\mathbf{v}_m$  span the full subspace  $\mathcal{S}$  under Assumption 1, we can show that  $\mathbf{U}^*$  is the only matrix in  $O(p, r)$ , up to rotation, such that  $Q^*(\mathbf{U}) := [\mathbb{E}\mathbf{g}(\mathbf{U})]^T \bar{\mathbf{W}}^* [\mathbb{E}\mathbf{g}(\mathbf{U})]$  is minimized (the minimum is 0). Thus, it is natural to expect  $\hat{\mathbf{U}}_n$ , as a minimizer of  $Q(\mathbf{U})$ , to be close to  $\mathbf{U}^*$  up to rotation when  $n$  is large.

The first theorem is a reassuring consistency result. We consider a generic weighting matrix  $\mathbf{W}^*$  without specifying particular choices.

**Theorem 1.** *Under Assumptions 1 and 2, there exists a sequence of orthogonal matrices  $\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \dots \in O(r)$  such that*

$$\hat{\mathbf{U}}_n \mathbf{R}_n \xrightarrow{p} \mathbf{U}^*, \quad \text{as } n \rightarrow \infty,$$

where  $\xrightarrow{p}$  means convergence in probability.

We remark that if we consider the regime where  $m$  also grows, typically we have consistency if  $m = o(n)$  with additional regularity assumptions on  $\mathbf{W}^*$  (Koenker and Machado, 1999; Donald et al., 2003).

Next, in accordance with our procedure where block weighting matrix is used, we consider the following assumption of block matrix forms for the covariance matrix  $\bar{\mathbf{S}}^*$ .

**Assumption 3** (block-wise covariance). *Suppose the covariance matrix of the concatenated vector*

$$\bar{\mathbf{f}}(\mathbf{x}_i, y_i) := [\mathbf{f}_1(\mathbf{x}_i, y_i); \dots; \mathbf{f}_m(\mathbf{x}_i, y_i)] \in \mathbb{R}^{\bar{m}},$$

denoted by  $\bar{\mathbf{S}}^*$ , has the following block matrix form:  $\bar{\mathbf{S}}^* = \mathbf{S}^* \otimes \mathbf{I}_p$ , where  $\mathbf{S}^* \in \mathbb{R}^{m \times m}$  is positive definite.

Under this assumption, there is a simple connection between  $\mathbf{S}^*$  and  $\boldsymbol{\Sigma}^*$ .

**Lemma 1.** *Under Assumption 3, we have  $\boldsymbol{\Sigma}^* = (p - r)\mathbf{S}^*$ .*

Similar to the classical theory in GMM, we will establish asymptotic normality of our estimator  $\widehat{\mathbf{U}}_n \mathbf{R}_n$  (corrected by a rotation  $\mathbf{R}_n$ ). The asymptotic covariance of our GMM estimator is a function of the weighting matrix  $\mathbf{W}_n$ . Once an explicit expression is obtained, it is then easy to justify, in terms of asymptotic efficiency, the optimality of our estimation procedure.

An important departure from the classical GMM theory is that, our estimator must satisfy the constraint  $\widehat{\mathbf{U}}_n \in O(p, r)$ , which is a manifold of dimension  $pr - r(r + 1)/2$ , and that the estimator is determined up to rotation. For this reason, instead of the difference  $\widehat{\mathbf{U}}_n \mathbf{R}_n - \mathbf{U}^*$ , we measure our estimator through  $\mathbf{P}_{\widehat{\mathbf{U}}_n}^\perp (\widehat{\mathbf{U}}_n \mathbf{R}_n - \mathbf{U}^*)$ , where  $\mathbf{P}_{\widehat{\mathbf{U}}_n}^\perp := \mathbf{I}_p - \widehat{\mathbf{U}}_n \widehat{\mathbf{U}}_n^T \in \mathbb{R}^{p \times p}$ . A desirable property of applying a projection  $\mathbf{P}_{\widehat{\mathbf{U}}_n}^\perp$  is that it does not depend on the choice of  $\mathbf{R}_n$ , as  $\mathbf{P}_{\widehat{\mathbf{U}}_n}^\perp \widehat{\mathbf{U}}_n = \mathbf{0}$  always holds, so we can also write  $\mathbf{P}_{\widehat{\mathbf{U}}_n}^\perp (\widehat{\mathbf{U}}_n \mathbf{R}_n - \mathbf{U}^*) = -\mathbf{P}_{\widehat{\mathbf{U}}_n}^\perp \mathbf{U}^*$ . The following result also serves as an intermediate result towards optimality under  $d(\mathcal{S}, \mathcal{S})$ . Below, the asymptotic variance is a degenerate matrix, but nevertheless, it is the smallest one in terms of the (generalized) inequality  $\succeq$ .

**Theorem 2.** *Define  $\mathbf{P}_{\widehat{\mathbf{U}}_n}^\perp := \mathbf{I}_p - \widehat{\mathbf{U}}_n \widehat{\mathbf{U}}_n^T \in \mathbb{R}^{p \times p}$ ,  $\mathbf{P}_{\mathbf{U}^*}^\perp := \mathbf{I}_p - \mathbf{U}^* (\mathbf{U}^*)^T \in \mathbb{R}^{p \times p}$ , and  $\mathbf{G}^* := [\mathbb{E} \mathbf{v}_1, \dots, \mathbb{E} \mathbf{v}_m]^T \mathbf{U}^* \in \mathbb{R}^{m \times r}$ . Under Assumptions 1 and 2, we have*

$$\begin{aligned} \sqrt{n} \operatorname{Vec}(\mathbf{P}_{\widehat{\mathbf{U}}_n}^\perp \mathbf{U}^*) &\xrightarrow{d} N(\mathbf{0}, (\mathbf{A}^* \otimes \mathbf{P}_{\mathbf{U}^*}^\perp) \bar{\mathbf{S}}^* ((\mathbf{A}^*)^T \otimes \mathbf{P}_{\mathbf{U}^*}^\perp)), \\ \text{where } \mathbf{A}^* &:= [(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^*]^{-1} (\mathbf{G}^*)^T \mathbf{W}^*, \end{aligned} \quad (21)$$

where  $\operatorname{Vec}(\cdot)$  stacks matrix columns into a vector. If we suppose, in addition, Assumption 3, then the choice  $\mathbf{W}^* = (\mathbf{S}^*)^{-1}$  or  $\mathbf{W}^* = (\boldsymbol{\Sigma}^*)^{-1}$  leads to the smallest asymptotic variance, in the following sense

$$(\mathbf{A}^* \otimes \mathbf{P}_{\mathbf{U}^*}^\perp) \bar{\mathbf{S}}^* ((\mathbf{A}^*)^T \otimes \mathbf{P}_{\mathbf{U}^*}^\perp) \succeq [(\mathbf{G}^*)^T (\mathbf{S}^*)^{-1} \mathbf{G}^*]^{-1} \otimes \mathbf{P}_{\mathbf{U}^*}^\perp, \quad \forall \mathbf{W}^* \succ \mathbf{0}. \quad (22)$$

Note that the asymptotic variance is invariant to the scaling of  $\mathbf{W}^*$ ; in particular, the choice  $\mathbf{W}^* = (\mathbf{S}^*)^{-1}$  or  $\mathbf{W}^* = (\boldsymbol{\Sigma}^*)^{-1}$  leads to the same asymptotic variance, in light of Lemma 1. Note also that the right-hand side of (22) is a symmetric matrix, due to basic properties of Kronecker products (see Lemma 2 (iii)). The covariance-like matrix  $\boldsymbol{\Sigma}^*$  is essentially a compact form of  $\bar{\mathbf{S}}^*$ , encoding the covariances through summation over  $p$  coordinates. A natural question is that, without the block-wise assumption, whether the choice  $\mathbf{W}^* = (\boldsymbol{\Sigma}^*)^{-1}$  is optimal for a simpler criterion. The next subsection gives an affirmative answer.

## 4.2 Optimality of estimation procedure

In this subsection, we measure the difference between the estimated subspace  $\widehat{\mathcal{S}} = \text{span}(\widehat{\mathbf{U}}_n)$  and the true  $\mathcal{S} = \text{span}(\mathbf{U}^*)$  using the notion *canonical angles*, which is a generalization of angles between two vectors.

Let  $\sigma_1, \dots, \sigma_r$  be the singular values of  $\widehat{\mathbf{U}}_n^T \mathbf{U}^*$ . Then, the canonical angles  $\theta_1, \dots, \theta_r \in [0, \pi/2]$  between  $\widehat{\mathbf{U}}_n$  and  $\mathbf{U}^*$  are  $\theta_k = \arccos(\sigma_k)$ ,  $k \in [r]$ . Also denote  $\sin \Theta_n(\mathbf{W}_n) = \text{diag}\{\sin \theta_1, \dots, \sin \theta_r\} \in \mathbb{R}^{r \times r}$ , where we show explicitly the dependency on  $\mathbf{W}_n$  and  $\Theta_n$ . Note that rotational invariance of singular values ensures that the canonical angles do not depend on the basis chosen to represent subspaces. The canonical angles are closely related to a common distance between  $\text{span}(\widehat{\mathbf{U}}_n)$  and  $\text{span}(\mathbf{U}^*)$  (Stewart, 1990, 1998):

$$d(\widehat{\mathcal{S}}, \mathcal{S})^2 = \|\widehat{\mathbf{U}}_n \widehat{\mathbf{U}}_n^T - \mathbf{U}^*(\mathbf{U}^*)^T\|_F^2 = 2\|\sin \Theta_n(\mathbf{W}_n)\|_F^2. \quad (23)$$

As briefly discussed in the previous subsection, we may remove the restrictive block-form Assumption 3 if the gauge is a different quantity. To this end, given a weighting matrix  $\mathbf{W}_n \succ \mathbf{0}$ , we gauge  $\widehat{\mathbf{U}} = \widehat{\mathbf{U}}(\mathbf{W}_n)$  through an  $r \times r$  positive semidefinite matrix

$$\Psi_n(\mathbf{W}_n) = (\mathbf{U}^*)^T \mathbf{P}_{\widehat{\mathbf{U}}_n}^\perp \mathbf{U}^* = \mathbf{I}_r - (\mathbf{U}^*)^T \widehat{\mathbf{U}}_n (\widehat{\mathbf{U}}_n)^T \mathbf{U}^*. \quad (24)$$

The eigenvalues of  $\Psi_n(\mathbf{W}_n)$  are determined by the canonical angles between  $\widehat{\mathbf{U}}_n$  and  $\mathbf{U}^*$ . In particular, a useful identity is given below. See the derivation in the proof of Theorem 3.

$$\text{Tr}(\Psi_n(\mathbf{W}_n)) = \|\sin \Theta_n(\mathbf{W}_n)\|_F^2 = \frac{1}{2} d(\widehat{\mathcal{S}}, \mathcal{S})^2. \quad (25)$$

To see how we may relax Assumption 3, we note that  $\Sigma^*$ , as defined in (10), captures covariance-like information of the vectors  $\mathbf{f}_1(\mathbf{x}_i, y_i), \dots, \mathbf{f}_m(\mathbf{x}_i, y_i)$  in an average sense. More precisely, we have the following identity:

$$\Sigma_{j\ell}^* = \text{Tr}[\text{Cov}(\mathbf{P}_{\mathbf{U}^*}^\perp \mathbf{f}_j(\mathbf{x}_i, y_i), \mathbf{P}_{\mathbf{U}^*}^\perp \mathbf{f}_\ell(\mathbf{x}_i, y_i))], \quad \forall j, \ell \in [m]. \quad (26)$$

Here, recall that  $\mathbf{P}_{\mathbf{U}^*}^\perp = \mathbf{I}_r - \mathbf{U}^*(\mathbf{U}^*)^T$  is a deterministic matrix, and that  $\mathbf{P}_{\mathbf{U}^*}^\perp \mathbf{f}_j(\mathbf{x}_i, y_i)$  has zero mean. Since  $\Psi_n(\mathbf{W}_n)$  is determined by summation over  $p$  coordinates, it is reasonable to expect that we only require a condition on  $\Sigma^*$ , rather than the big  $\bar{m} \times \bar{m}$  matrix  $\bar{\mathbf{S}}$ .

This intuition leads to Assumption 4 below, which greatly relaxes Assumption 3. It is possible to further relax this assumption (see Section 4.4), but for now we will content ourselves with it.

**Assumption 4** (invertibility). *The matrix  $\Sigma^*$  defined in (26) is not singular.*

We will use a notion called *asymptotic expectation* to assess and compare the quality of  $\Psi_n(\mathbf{W}_n)$  for different  $\mathbf{W}_n$ . Formally, similar to Shao (2003), we define the asymptotic expectation below.

**Definition 1.** Let  $\{\xi_n\}$  be a sequence of random vectors and  $\{a_n\}$  be a sequence of positive numbers satisfying  $a_n \rightarrow \infty$  or  $a_n \rightarrow a > 0$ . If  $a_n \xi_n \xrightarrow{d} \xi$  and  $\mathbb{E}\|\xi\|_2 < \infty$ , then  $\mathbb{E}\xi/a_n$  is called the asymptotic expectation of  $\xi_n$ , which is denoted by  $\text{AE}(\xi_n)$ .

Note that if  $\mathbb{E}\xi \neq \mathbf{0}$ , then the asymptotic expectation is unique up to a  $(1 + o(1))$  factor. This is a consequence of Prop. 2.3 of Shao (2003). Roughly speaking, the asymptotic expectation is a ‘weaker’ version of the usual expectation, because weak convergence bypasses the issue of potential extreme values of  $\xi_n$ , which are often caused by its singularities. If certain sort of uniform convergence is guaranteed, we will recover the usual expectation, as stated in the next theorem.

**Theorem 3.** Under Assumptions 1, 2 and 4, the asymptotic expectation<sup>1</sup> of  $\Psi_n(\mathbf{W}^*)$  is minimized at  $\mathbf{W}^* = (\Sigma^*)^{-1}$ , that is, for any limiting  $\mathbf{W}^* \succ \mathbf{0}$ ,

$$\text{AE}\left(\Psi_n((\Sigma^*)^{-1})\right) = \frac{1}{n} [(\mathbf{G}^*)^T (\Sigma^*)^{-1} \mathbf{G}^*]^{-1} \preceq \text{AE}\left(\Psi_n(\mathbf{W}^*)\right) = \text{AE}\left(\Psi_n(\mathbf{W}_n)\right).$$

If, in addition,  $\{n\|\Psi_n(\mathbf{W}^*)\|_F\}_n$  and  $\{n\|\Psi_n((\Sigma^*)^{-1})\|_F\}_n$  are both uniformly integrable, then

$$\mathbb{E} \Psi_n((\Sigma^*)^{-1}) \preceq (1 + o(1)) \cdot \mathbb{E} \Psi_n(\mathbf{W}^*).$$

Moreover, the same holds if we replace  $\Psi_n((\Sigma^*)^{-1})$  by  $\Psi_n((\widehat{\Sigma}_n)^{-1})$ .

In the above theorem, the asymptotic expectations of  $\Psi_n((\Sigma^*)^{-1})$  and  $\Psi_n((\widehat{\Sigma}_n)^{-1})$  are essentially the same (up to a  $1 + o(1)$  factor). This is because the asymptotic expectations are determined by the limiting weighting matrix, and  $(\widehat{\Sigma}_n)^{-1}$  implemented in our procedure converges in probability to  $(\Sigma^*)^{-1}$ .

Using (25), we obtain an optimality result for the final estimator  $\widehat{\mathbf{U}}_n^{\text{GMM}} = \widehat{\mathbf{U}}_n(\widehat{\Sigma}_n^{-1})$  produced by our estimation procedure in Section 2.3.

**Corollary 1.** Under Assumptions 1, 2 and 4, the matrix  $\widehat{\Sigma}_n$  computed in Step 2 of our procedure is invertible with probability  $1 - o(1)$ , and for any limiting  $\mathbf{W}^* \succ \mathbf{0}$ ,

$$\text{AE}\left(\left\|\widehat{\mathbf{U}}_n^{\text{GMM}}[\widehat{\mathbf{U}}_n^{\text{GMM}}]^T - \mathbf{U}^*(\mathbf{U}^*)^T\right\|_F^2\right) \leq \text{AE}\left(\left\|\widehat{\mathbf{U}}_n(\mathbf{W}_n)[\widehat{\mathbf{U}}_n(\mathbf{W}_n)]^T - \mathbf{U}^*(\mathbf{U}^*)^T\right\|_F^2\right).$$

Similar to Theorem 3, the asymptotic expectation can be replaced by the usual expectation if uniform integrability is satisfied.

Our results in this subsection rely on Assumption 4, namely, the invertibility of  $\Sigma^*$ . However, if  $m$  is not very small,  $\Sigma^*$  may be singular or nearly singular in practice. Before addressing this issue and justifying the de-noising step suggested in (13), we first focus on a more practical aspect—estimating the dimension of the subspace  $r$  if it is not known in advance—in the next subsection.

---

<sup>1</sup>Statements and inequalities involving asymptotic expectation are up to a  $1 + o(1)$  factor, due to the nature of its definition.

### 4.3 Estimating subspace dimension $r$

In Section 2 and 4.1, we assumed that the subspace dimension  $r$  is known. However, in practice, this is usually unknown. For example, in finite mixture models,  $r$  is the number of mixtures, which typically needs to be estimated from the data. Therefore, in these cases, it is of interest to determine  $r$  prior to obtaining the subspace GMM estimator.

Fortunately, the simple form of (9) allows us to estimate  $r$  consistently using only the eigenvalues of  $\mathbf{V}_n \mathbf{W}_n \mathbf{V}_n^T$ . For a symmetric matrix  $\mathbf{A}$ , let us denote by  $\lambda_j(\mathbf{A})$  the  $j$ th largest eigenvalue of  $\mathbf{A}$ .

**Theorem 4.** Suppose  $r < p$ , and Assumptions 1 and 4 hold. If  $\mathbf{W}_n \xrightarrow{p} \mathbf{W}^* = (\Sigma^*)^{-1}$ , where  $\Sigma^*$  is defined in (10), then  $\sum_{j=r+1}^p \lambda_j(\mathbf{V}_n \mathbf{W}_n \mathbf{V}_n^T) = O_P(n^{-1})$ . Moreover, assuming Assumption 3 in addition, we have

$$n(p-r) \sum_{j=r+1}^p \lambda_j(\mathbf{V}_n \mathbf{W}_n \mathbf{V}_n^T) \xrightarrow{d} \chi_{(p-r)(m-r)}^2.$$

In particular, the choice  $\mathbf{W}_n = (\widehat{\Sigma}_n)^{-1}$  in our procedure satisfies  $\mathbf{W}_n \xrightarrow{p} (\Sigma^*)^{-1}$ .

This result predicts a large eigen-gap between top  $r$  eigenvalues and the remaining eigenvalues of  $\mathbf{V}_n \mathbf{W}_n \mathbf{V}_n^T$  in the large sample regime: as  $n$  grows to  $\infty$ , the sum of the remaining eigenvalues scale with  $1/n$ , whereas, under Assumption 1,  $\lambda_r(\mathbf{V}_n \mathbf{W}_n \mathbf{V}_n^T)$  remains a constant order. This suggests a simple threshold-based method:

$$\widehat{r}_\tau = \operatorname{argmax}_k \{k \mid \lambda_k(\mathbf{V}_n \mathbf{W}_n \mathbf{V}_n^T) > \tau_n\},$$

where  $\tau_n$  is the threshold value with  $\tau_n = o(1)$  and  $n\tau_n \rightarrow \infty$ . In cases where Assumption 3 holds and the chi-squared distribution is a suitable approximation, one can also use the following estimator, which is similar to Li (1991) in spirit.

$$\widehat{r}_\eta = \operatorname{argmin}_k \{k \mid n(p-k) \sum_{j>k} \lambda_j(\mathbf{V}_n \mathbf{W}_n \mathbf{V}_n^T) \leq \eta_n(k)\}. \quad (27)$$

The parameter  $\eta_n(k)$  is related to the quantiles of  $\chi_{(p-k)(m-k)}^2$ . In practice, for example, one may choose  $\eta_n(k)$  to be the 95th quantile of  $\chi_{(p-k)(m-k)}^2$ . The following corollary establishes the consistency of both estimators.

**Corollary 2.** Suppose  $r < p$ , Assumptions 1 and 4 hold, and  $\mathbf{W}_n \xrightarrow{p} (\Sigma^*)^{-1}$ . Then, with  $\tau_n = o(1)$  and  $n\tau_n \rightarrow \infty$ , the estimator  $\widehat{r}_\tau$  is consistent; with  $\eta_n(r) = o(n)$  and  $\eta_n(r) \rightarrow \infty$ , the estimator  $\widehat{r}_\eta$  is consistent.

In the literature, especially in factor models, eigenvalues are the basis of many methods for rank or dimension estimation (e.g., the number of factors in factor models). An early work is the scree test method (Cattell, 1966), which sorts and plots the eigenvalues in descending order. Many recent works on dimension estimation focus on factor models, including Bai

and Ng (2002), in which one minimizes an objective function that is similar to information criterion, and Onatski (2010); Lam et al. (2012); Ahn and Horenstein (2013), in which one uses differences or ratios of eigenvalues to determine the number of factors, etc.

Though being related, these recent works derive sophisticated methods based on different models and typically consider different regimes ( $p$  grow with  $n$ ).

#### 4.4 Redundancy and singular $\Sigma^*$

In this subsection, our goal is to further relax the non-singular Assumption 4, and justifies the use of thresholding in Step 4 of our estimation procedure. Often, we derive the vectors  $\mathbf{v}_\ell$  from different approaches, hoping to improve the estimation of  $\mathcal{S}$ . However, in the process, we may inadvertently introduce redundant vectors that lead to a singular  $\Sigma^*$ . In our simulations, for example, we do observe such singularity issue. Nevertheless, if the singularity of  $\Sigma^*$  is solely caused by redundant vectors, then, as we will show, we can simply resort to the thresholding without losing any optimality guarantee.

Let us first define precisely what we mean by ‘redundancy’. For simplicity, let us write  $\mathbf{f}_\ell := \mathbf{f}_\ell(\mathbf{x}_i, y_i)$ , suppressing the dependency on  $i$  due to i.i.d. data. Recall the notation  $\mathbf{F} = [\mathbf{f}_1, \dots, \mathbf{f}_m] \in \mathbb{R}^{p \times m}$ , where we suppress the subscript  $i$  due to the i.i.d. assumption. Let us partition  $\mathbf{F}$  and  $\Sigma^*$  as follows:

$$\begin{aligned} \mathbf{F} &= [\mathbf{F}_1, \mathbf{F}_2], \quad \text{where } \mathbf{F}_1 \in \mathbb{R}^{p \times m_1}, \mathbf{F}_2 \in \mathbb{R}^{p \times m_2}, \\ \Sigma^* &= \begin{bmatrix} \Sigma_{11}^* & \Sigma_{12}^* \\ \Sigma_{21}^* & \Sigma_{22}^* \end{bmatrix}, \quad \text{where } \Sigma_{ab}^* \in \mathbb{R}^{m_a \times m_b}, a, b \in \{1, 2\}. \end{aligned} \tag{28}$$

By construction, the dimensions must satisfy  $m_1 + m_2 = m$ . We require that such partition separates the non-redundant vectors (columns of  $\mathbf{F}_1$ ) from the redundant vectors (columns of  $\mathbf{F}_2$ ), as will soon be explained. We require  $\mathbf{F}_1$  to be well behaved in the sense that  $\Sigma_{11}^*$  is invertible. Moreover, we let

$$\tilde{\mathbf{F}}_2 := \mathbf{F}_2 - \mathbf{F}_1(\Sigma_{11}^*)^{-1}\Sigma_{12}^* \in \mathbb{R}^{p \times m_2} \tag{29}$$

be the part of  $\mathbf{F}_2$  not explained by  $\mathbf{F}_1$ . This construction ensures  $\mathbf{F}_1$  and  $\tilde{\mathbf{F}}_2$  have zero cross-covariance, i.e.,  $\mathbb{E} \mathbf{F}_1^T \mathbf{F}_2 = \mathbf{0}$ . Also, as seen in Section 1.3, the expectation of any linear combination of vectors  $\mathbf{f}_\ell$  (or  $\mathbf{v}_\ell$ ) always lies in  $\mathcal{S}$ , and thus  $\mathbb{E} \tilde{\mathbf{F}}_2$  satisfies the condition (4). We call  $\mathbf{F}_2$  redundant if  $\mathbb{E} \tilde{\mathbf{F}}_2 = \mathbf{0}$ ; in other words, in expectation,  $\mathbf{F}_2$  is fully explained by  $\mathbf{F}_1$ . This separation of well-behaved vectors and redundant vectors is summarized in the following assumption.

**Assumption 5** (partial invertibility). *Suppose  $0 \leq m_2 < m$ , and that  $\mathbf{F}$  and  $\Sigma^*$  can be partitioned as in (28) such that  $\Sigma_{11}^*$  is invertible, and that  $\tilde{\mathbf{F}}_2$ , as defined in (29), satisfies  $\mathbb{E} \tilde{\mathbf{F}}_2 = \mathbf{0}$ .*

Note that we permit  $m_2 = 0$ , i.e.,  $\Sigma^*$  is non-singular and no partition is needed, which includes Assumption 4 as a special (less general) case. This assumption relaxes the invert-

ibility assumption in that it allows scenarios where some  $\mathbf{f}_\ell$  are fully explained by others, often due to overlapping information practitioners introduce when collecting these  $\mathbf{f}_\ell$ . In particular, this assumption is satisfied if  $\Sigma_{11}^*$  is invertible and  $\mathbf{F}_2$  is a linear transformation of  $\mathbf{F}_1$ .

We also note that our notion of redundancy is related to that in Breusch et al. (1999): in both cases, the redundant part is explained by the other part in a way similar to (29), but our covariance-like matrix  $\Sigma^*$  is tailored to the special structure here.

We will generalize Section 4.3 by showing that the optimal limiting weighting matrix is  $(\Sigma^*)^+$ , i.e., the pseudoinverse of  $\Sigma^*$ , and that the thresholding step (13) guarantees  $\mathbf{W}_n \xrightarrow{p} (\Sigma^*)^+$  for a suitable parameter  $\delta_n$ . The optimality of  $(\Sigma^*)^+$  is established in a slightly larger family of limiting weighting matrices:

$$\mathcal{W} := \{\mathbf{W}^* \in \mathbb{R}^{m \times m} \mid \mathbf{W}^* \succeq \mathbf{0}, (\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^* \succ \mathbf{0}\}.$$

This family  $\mathcal{W}$  includes all positive semidefinite matrices. We can show that  $(\mathbf{G}^*)^T (\Sigma^*)^+ \mathbf{G}^*$  is always positive definite, and thus  $(\Sigma^*)^+ \in \mathcal{W}$ . Moreover, any limiting weighting matrix in  $\mathcal{W}$  leads to a consistent estimator for  $\mathcal{S}$ , which is a generalization of Theorem 1 (see Theorem 7 in the supplementary materials). The next result is a formal optimality statement.

**Theorem 5.** Suppose Assumptions 1 and 5 hold, and  $\mathbf{W}_n \xrightarrow{p} (\Sigma^*)^+$ . Then,  $(\mathbf{G}^*)^T (\Sigma^*)^+ \mathbf{G}^* \succ \mathbf{0}$ , and for any sequence  $\{\mathbf{W}'_n\}$  with  $\mathbf{W}'_n \xrightarrow{p} \mathbf{W}^* \in \mathcal{W}$ ,

$$\text{AE}\left(\Psi_n(\mathbf{W}_n)\right) = \frac{1}{n} [(\mathbf{G}^*)^T (\Sigma^*)^+ \mathbf{G}^*]^{-1} \preceq \text{AE}\left(\Psi_n(\mathbf{W}^*)\right) = \text{AE}\left(\Psi_n(\mathbf{W}'_n)\right). \quad (30)$$

In particular,  $\mathbf{W}_n$  in (13) with  $\delta_n = o(1)$  and  $\sqrt{n} \delta_n \rightarrow \infty$  satisfies  $\mathbf{W}_n \xrightarrow{p} (\Sigma^*)^+$ .

This theorem is a generalization of Theorem 3. Indeed, if  $\Sigma^*$  is invertible, then we recover Theorem 3. Similarly, if we set  $\widehat{\mathbf{U}}_n^{\text{GMM}} = \widehat{\mathbf{U}}_n(\mathbf{W}_n)$  where  $\mathbf{W}_n$  is computed from (13), we obtain a generalized corollary.

**Corollary 3.** Suppose Assumptions 1 and 5 hold, and we compute  $\mathbf{W}_n$  according to (13) with  $\delta_n = o(1)$  and  $\sqrt{n} \delta_n \rightarrow \infty$ . Then, for any sequence  $\{\mathbf{W}'_n\}$  with  $\mathbf{W}'_n \xrightarrow{p} \mathbf{W}^* \in \mathcal{W}$ ,

$$\text{AE}\left(\left\|\widehat{\mathbf{U}}_n^{\text{GMM}}[\widehat{\mathbf{U}}_n^{\text{GMM}}]^T - \mathbf{U}^*(\mathbf{U}^*)^T\right\|_F^2\right) \leq \text{AE}\left(\left\|\widehat{\mathbf{U}}_n(\mathbf{W}'_n)[\widehat{\mathbf{U}}_n(\mathbf{W}'_n)]^T - \mathbf{U}^*(\mathbf{U}^*)^T\right\|_F^2\right). \quad (31)$$

We remark that, in practice, a larger  $\delta_n$  may be used to select more important  $\mathbf{f}_\ell$  (or their linear combinations), apart from eliminating redundant  $\mathbf{f}_\ell$ . This may be useful when  $m$  is large. See more discussion in Section 7.

Note also that asymptotically, adding more moments does not increase errors: if  $\mathbf{V}' \in \mathbb{R}^{p \times m'}$  is a submatrix of  $\mathbf{V}$  with  $m' < m$ , then under similar assumptions, the asymptotic error of the GMM estimator is non-increasing in  $m$ . That is, denoting the left-hand side of

(31) by  $\text{AE}(\text{err}^m)$ , we have  $\text{AE}(\text{err}^m) \leq \text{AE}(\text{err}^{m'})$ . This is because we can constrain  $\mathbf{W}^*$  to have nonzero values only in a  $m' \times m'$  block.

## 4.5 Augmenting regular moments

It is clear that, since the columns of  $\mathbf{V}_n$  are the moments of the data (with or without transformations), the matrix  $\mathbf{V}_n \mathbf{W}_n \mathbf{V}_n^T$  takes into consideration all pairwise products of these moments. However, it is not explicitly clear whether the moments themselves are included. The purpose of this subsection is to show that, implicitly, it also covers the moments themselves.

To elaborate on this point, let us consider the simple factor model we discussed in Section 3.1. Suppose we choose  $\mathbf{V}_n = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{p+1}]$  where  $\mathbf{v}_1$  and  $\mathbf{v}_{1+\ell}$  ( $\ell \in [p]$ ) are vectors given in (14). They consist of the first moments and the second moments of  $\mathbf{x}_i$ . Equivalently, we can write  $\mathbf{V}_n = [\bar{\mathbf{x}}, n^{-1}\mathbf{X}^T\mathbf{X} - \sigma^2\mathbf{I}_p]$ , where  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$ . The GMM estimator finds the top eigenspace of  $\mathbf{V}_n \mathbf{W}_n \mathbf{V}_n^T$ .

The above estimate is predominately based on the pairwise products of the second moment information. A natural question is if the first moment condition, based on  $n^{-1}\mathbf{X}^T\mathbf{X}$ , provides additional information. See Section 2.4 also for discussions. A natural aggregation of the information based on the first and second moment conditions is

$$\text{eigen}_r (\kappa n^{-1}\mathbf{X}^T\mathbf{X} + \mathbf{V}_n \mathbf{W}_n \mathbf{V}_n^T)$$

with a suitably chosen parameter  $\kappa \geq 0$ . However, we shall prove that the most efficient estimator we can hope to obtain is simply the GMM estimator (or equivalently  $\kappa = 0$ ). This means that the information contained in the traditional PCA-based information  $n^{-1}\mathbf{X}^T\mathbf{X}$  is contained in our second moment method  $\text{eigen}_r (\mathbf{V}_n \mathbf{W}_n \mathbf{V}_n^T)$ .

**Theorem 6.** *Assume the same conditions as in Corollary 3. Also suppose that  $\mathbf{M}_n \in \mathbb{R}^{p \times p}$  is a symmetric submatrix of  $\mathbf{V}_n$  satisfying  $\text{rank}(\mathbb{E}\mathbf{M}_n) = r$  and  $\mathbb{E}\mathbf{M}_n \succeq \mathbf{0}$ . For any parameter  $\kappa \geq 0$  and any sequence  $\{\mathbf{W}'_n\}$  with  $\mathbf{W}'_n \xrightarrow{p} \mathbf{W}^* \in \mathcal{W}$ , define*

$$\tilde{\mathbf{U}}_n := \tilde{\mathbf{U}}_n(\kappa, \mathbf{W}'_n) = \text{eigen}_r (\kappa \mathbf{M}_n + \mathbf{V}_n \mathbf{W}'_n \mathbf{V}_n^T),$$

*Then, our GMM estimator  $\hat{\mathbf{U}}_n^{\text{GMM}}$  satisfies*

$$\text{AE} \left( \left\| \hat{\mathbf{U}}_n^{\text{GMM}} [\hat{\mathbf{U}}_n^{\text{GMM}}]^T - \mathbf{U}^* (\mathbf{U}^*)^T \right\|_F^2 \right) \leq \text{AE} \left( \left\| \tilde{\mathbf{U}}_n \tilde{\mathbf{U}}_n^T - \mathbf{U}^* (\mathbf{U}^*)^T \right\|_F^2 \right).$$

Note that, the condition  $\text{rank}(\mathbb{E}\mathbf{M}_n) = r$  essentially requires that  $\mathbf{M}_n$  contains enough information about  $\mathcal{S}$ , which is similar to Assumption 1. Also note that, in general,  $\kappa \geq 0$  is necessary for consistency. The proof idea is that the dominant part of  $\mathbf{M}_n$  (that is, its best rank- $r$  approximation) can be absorbed into  $\mathbf{V}_n \mathbf{W}'_n \mathbf{V}_n^T$ , while its remainder is negligible. See supplementary materials for details.

This result connects our method to [Lettau and Pelger \(2017\)](#), in which eigenvectors of  $n^{-1}\mathbf{XX}^T + \gamma\bar{\mathbf{x}}\bar{\mathbf{x}}^T$  are studied. Our result is more general in the sense that we consider a weighting matrix instead of a single parameter  $\gamma$ .

## 5 Numerical simulations

In this section, we show how our methods may extract information from a pool of commonly used methods, under two kinds of models in Section 3.1–3.2: factor models (Example 1) and mixture models (Example 2). Both examples assume a known  $r$ . In the supplementary materials, we give additional simulation results for the multiple index models (Section 3.3) and consistent dimension estimation (Section 4.3).

In all simulations, the hard-thresholding parameter  $\delta_n$  is fixed at 0.01.

### Example 1: first moments may help

Let us consider the simple factor model (1). We generate each entry of  $\mathbf{B}$  independently from  $N(0, 1)$ , and keep it fixed as the loading matrix throughout the data generating process. We also generate  $\mathbf{z}_i \sim N(\boldsymbol{\mu}_{\mathbf{z}}, \mathbf{I}_r)$  and  $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_p)$  independently, and then compute  $\mathbf{x}_i$ . The left singular vectors of  $\mathbf{B}$  are computed and set to be  $\mathbf{U}^*$ .

We fix  $p = 10$ ,  $r = 2$  and  $\sigma = 2$ , and set  $\boldsymbol{\mu}_{\mathbf{z}} = (\mu, -\mu)^T$ , where  $\mu$  is a parameter. We consider two experiments: (1) fix  $n = 500$ , and let  $\mu$  run through  $\{0, 0.5, 1, 1.5, \dots, 4\}$ ; (2) fix  $\mu = 2$ , and let  $n$  run through  $\{100, 200, 400, 800, 1600, 3200\}$ . The first experiment studies how much information we can extract from the first moments with a varying  $\mu$ , and the second experiment studies the effect of sample sizes on the performance of methods under investigation.

We compare the performance of three estimators: (i) *standard*—the first one is the top  $r$  eigenvectors of the sample covariance of  $\mathbf{y}_i$ , which is the standard method; (ii) *GMM full*—the second one is the two-step estimator proposed in Section 2.3; (iii) *GMM diagonal*—the third one keeps only the diagonal entries of  $\widehat{\Sigma}$  and sets the rest to zero, and elsewhere follows the same procedure as *GMM full*. Both GMM estimators (ii) and (iii) are constructed from the  $p + 1$  vectors in (14).

The performance of estimators is assessed over 100 independent simulations and measured through the estimation error  $\mathbb{E}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T - \mathbf{U}^*(\mathbf{U}^*)^T\|_F$ , where  $\widehat{\mathbf{U}}$  is any of the three estimators. We compute this error by averaging over all 100 simulations, and we also compute the standard errors of the averages. The results are presented in Figure 2. We have also considered other estimation errors, including  $\mathbb{E}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T - \mathbf{U}^*(\mathbf{U}^*)^T\|_F^2$ ,  $\mathbb{E}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T - \mathbf{U}^*(\mathbf{U}^*)^T\|_2$  and  $\mathbb{E}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T - \mathbf{U}^*(\mathbf{U}^*)^T\|_2^2$ . The results for these errors are similar, and figures are omitted here.

On the left plot of Figure 2, as  $\mu$  increases, we can see that the standard method does not benefit from  $\boldsymbol{\mu}_{\mathbf{z}}$ 's deviation from  $\mathbf{0}$ , whereas both of the two GMM methods have decreasing estimation errors. Note that when  $\mu = 0$ , there is no information we can exploit from  $\mathbf{v}_1$

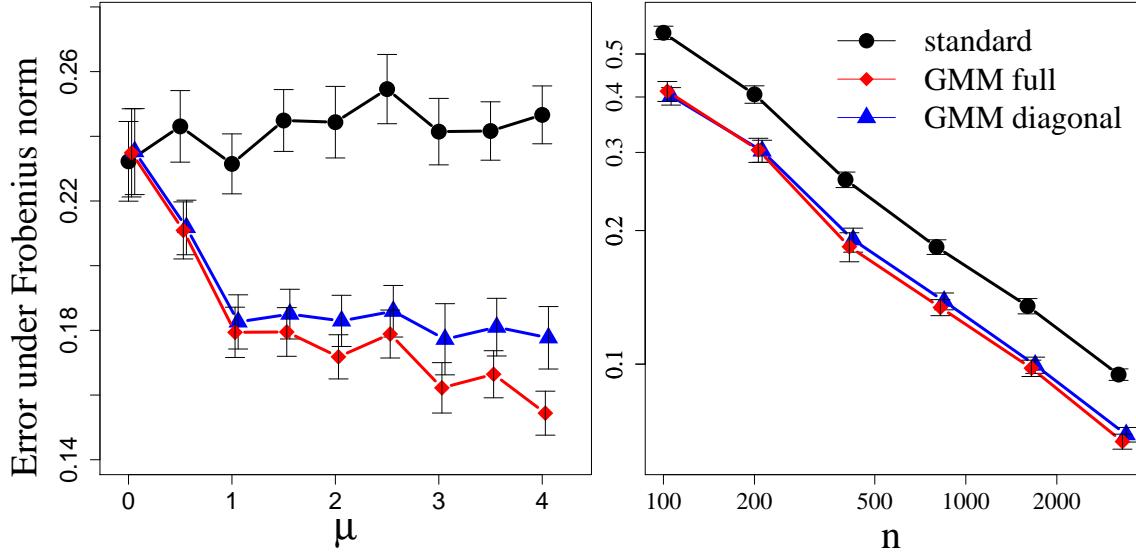


Figure 2: Estimation errors  $\mathbb{E}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T - \mathbf{U}^*(\mathbf{U}^*)^T\|_F$  of three estimators calculated based on averages over 100 simulations. On the markers, bars represent standard errors of the averages. **Left plot:** fix  $n = 500$ . **Right plot** (on the log scale): fix  $\mu = 2$ .

(first moments) since its expectation is zero, so we expect all methods behave similarly. Indeed, the figure supports this prediction at  $\mu = 0$ . However, as  $\mu$  starts increasing, both GMM estimators have superior performance, which suggests that  $\mathbf{v}_1$  indeed contributes to the improvement on the standard method. This contribution grows as  $\mu$  continues to deviate from 0. Thus, the inclusion of first moments through our GMM estimation helps if the factors do not have zero mean.

The right plot of Figure 2 shows the same estimation error against varying  $n$  on the logarithmic scale for all estimators. All three curves have the same alignment, which is consistent with the theory that errors scale with  $n^{-1/2}$ . Moreover, both GMM estimators outperform the standard one regardless of  $n$ , which further demonstrates the benefits of first moments.

## Example 2: advantages of transformations

We now look at another appealing aspect of our GMM methods: using different transformations in the construction of  $\mathbf{v}_\ell$ , we may expose and extract even more information from GMM estimators. Consider the mixed linear regression model (16) with identical  $\sigma_k$ .

We fix  $p = 10$  and  $K = 2$ , and sample each  $\beta_k$  independently and uniformly from a sphere in  $\mathbb{R}^p$  with radius 4. We also generate  $\beta_{k0} \sim N(0, 1)$ . Then, we generate i.i.d. entries of  $\mathbf{x}$  from  $N(0, 1)$ , i.i.d.  $z_i \sim \text{Bernoulli}(1/2)$  and i.i.d.  $\epsilon_i \sim N(0, \sigma^2)$ , where  $\sigma$  is fixed as 1. The left singular vectors of  $[\beta_1, \beta_2]$  are set to  $\mathbf{U}^*$ . We collect  $m = 25$  vectors  $\mathbf{v}_\ell$  as below.

- (a) first moments:  $\mathbf{v}_1 = n^{-1} \sum_{i=1}^n y_i \mathbf{x}_i$ ;

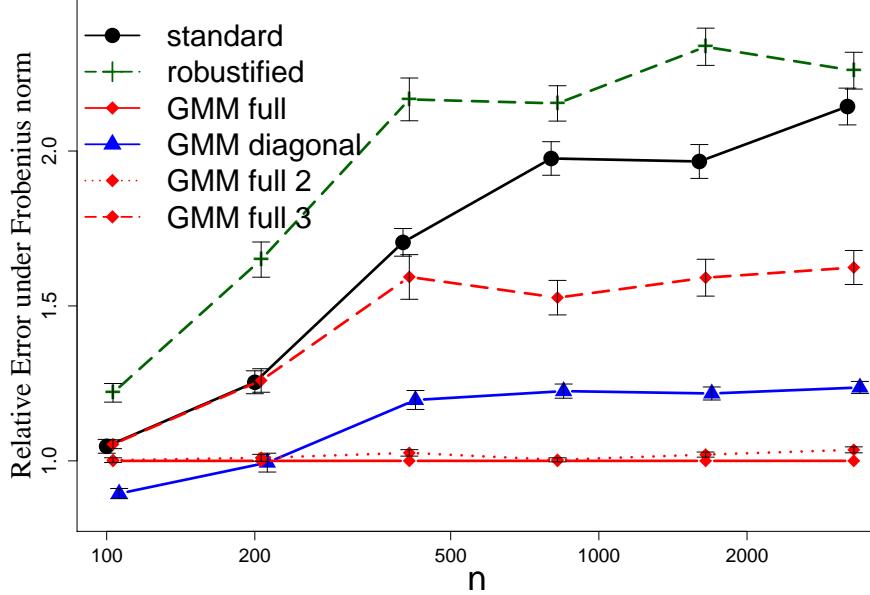


Figure 3: Ratios of estimation errors  $\mathbb{E}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T - \mathbf{U}^*(\mathbf{U}^*)^T\|_F$  calculated based on averages over 100 simulations. Five methods (five curves) are compared against ‘GMM full’ (the horizontal line  $y = 1$ ). On the markers, bars are shown to represent standard errors.

- (b) second moments:  $\mathbf{v}_{1+j} = n^{-1} \sum_{i=1}^n y_i^2 (\mathbf{x}_i \mathbf{x}_i^T \mathbf{e}_j - \mathbf{e}_j)$ ,  $\forall j \in [p]$ ;
- (c) transformed first moments:  $\mathbf{v}_{11+j} = n^{-1} \sum_{i=1}^n \cos(y_i \pi/(2\tau) + (j-1)\pi/4) \mathbf{x}_i$ ,  $\forall j \in [4]$ ;
- (d) transformed second moments:  $\mathbf{v}_{15+j} = n^{-1} \sum_{i=1}^n \tilde{y}_i (\mathbf{x}_i \mathbf{x}_i^T \mathbf{e}_j - \mathbf{e}_j)$ ,  $\forall j \in [p]$ .

In (c),  $\tau$  is a scale parameter chosen to be 0.8 quantile of  $|y_i|$ ; and in (d), the transformed response  $\tilde{y}_i$  is defined by

$$\tilde{y}_i := \text{sign}(y_i) \text{sign}(\mathbf{x}_i^T \tilde{\mathbf{v}}_1), \quad \text{where } \tilde{\mathbf{v}}_1 := n^{-1} \sum_{i=1}^n \text{sign}(y_i) \mathbf{x}_i.$$

In essence, (d) seeks to robustify the second moments by only using the sign of  $y_i$ . The additional  $\text{sign}(\mathbf{x}_i^T \tilde{\mathbf{v}}_1)$  avoids vanishing  $\mathbb{E}\mathbf{v}_{15+j}$ , which is proposed by Sun et al. (2014). Note that (d) does not strictly satisfy the condition (4), because the sign function is not smooth and  $\tilde{\mathbf{v}}_1$  is only asymptotically a linear combination of  $\boldsymbol{\beta}_1$  and  $\boldsymbol{\beta}_2$ . Nevertheless, we find it practically useful for our model here.

Gauged by the same error  $\mathbb{E}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T - \mathbf{U}^*(\mathbf{U}^*)^T\|_F$ , we compare six estimators as listed in Table 1. In Figure 3, we plot the performance of these estimator averaged over 100 simulations. We make a few observations:

---

<i>standard</i> : only using (b)	<i>robustified</i> : only using (d)
<i>GMM full</i> : combining (a)–(d)	<i>GMM diagonal</i> : combining (a)–(d), but $\mathbf{W}$ is diagonal
<i>GMM full 2</i> : combining (a),(b),(d)	<i>GMM full 3</i> : combining (a),(b).

---

Table 1: Comparing six methods: two atomic methods and four GMM-based methods.

- Atomic methods like ‘standard’ or ‘robustified’ have higher errors than GMM methods, suggesting it is better to combine different moments.
- ‘GMM full 3’ has better performance than ‘standard’, which reinforces our conclusion of the previous example that ‘first moments may help’.
- ‘GMM full’ and ‘GMM full 2’ have further improved performance, which indicates that (d) contains useful information. Combining all moments as ‘GMM full’ is the best.

A minor observation is that, when  $n$  is small, ‘GMM diagonal’ may be preferred over ‘GMM full’. It is likely because estimating the full matrix  $\mathbf{W}$  is difficult with a small sample size. To conclude, by constructing  $\mathbf{v}_\ell$  with different transformations, we may exploit and combine different information using our GMM-based methods.

## 6 Real data examples

We apply our procedure to the *ozone* dataset. This dataset has been studied in [Breiman and Friedman \(1985\)](#) and [Li \(1992\)](#), etc. It contains  $n = 330$  days of measurements of ozone concentration and  $p = 8$  meteorological features, where all variables take continuous values. Dataset details can be found in [Breiman and Friedman \(1985\)](#). The goal is to study the relationship between the ozone concentration (response) with meteorological features (predictors). We will use this dataset to study different methods for the multiple index model.

The *abalone* dataset is also studied. Results and details of the dataset are in the supplementary materials. In our experiments, we use the same procedure to produce our GMM estimators as in the simulations. We also fix the parameter  $\delta_n$  to be 0.01 as before.

We obtain the ozone dataset from the R package ‘gclus’ ([Hurley, 2012](#)). First, we run a least squares fit of the ozone concentration  $y$  against all  $p = 8$  standardized covariates, which results in an R-squared 0.69. This serves as a baseline for our subsequence comparisons.

Next, we consider three methods: residual-based pHd, GMM diagonal and GMM full (recall definitions in Sect. 5, Example 3). Before running these methods, we make a linear transformation of the data such that the covariates have identity sample covariance matrix (a.k.a. whitening). Then, we let  $K = 2$ , and run each of these three methods to find 2 orthogonal directions, say,  $\hat{\mathbf{u}}_1$  and  $\hat{\mathbf{u}}_2$ , which allows us to reduce the number of covariates to 8 to 2 (the new covariates have the form  $\mathbf{x}^T \hat{\mathbf{u}}_k$ ). Finally, as in [Li \(1992\)](#), we fit a quadratic regression. We include cross products between variables in the quadratic regression. Using

	least squares	$r$ -based pHd	GMM diagonal	GMM full
$K = 1$	0.69	0.67	0.74	0.74
$K = 2$		0.69	0.76	0.76
$K = 3$		0.72	0.77	0.77

Table 2: R-squared values of the four methods under three settings. A quadratic regression is fit after  $\hat{\mathbf{u}}_k$  is obtained. In the first setting  $K = 1$ , we find a single direction  $\hat{\mathbf{u}}_1$ ; in the second setting  $K = 2$  and the third setting  $K = 3$ , we find  $\hat{\mathbf{u}}_k$  and run a quadratic regression including cross products between the two new covariates  $\mathbf{x}_i^T \hat{\mathbf{u}}_k$ .

the same procedure, we also run these methods with  $K = 1$  and  $K = 3$ . The values of R-squared are summarized in Table 2. The adjusted R-squared values are very similar, which are omitted here. The outperformance of our aggregated GMM methods can be easily seen.

To understand the variability of our results, we sample 100 bootstrap samples and run the same methods. In Figure 7 (see supplementary materials), we report the boxplots from 100 bootstrap samples, which show the R-squared values of quadratic fits with  $K = 2$ .

From Table 2 and Figure 7, we can see that both GMM methods lead to better regression fits than the naive least squares and the pHd method. Moreover, for the pHd method, the outliers of the boxplot suggest that there are failure chances, that is, pHd method does not correctly find good direction  $\hat{\mathbf{u}}$ , whereas, our GMM methods show robust performance in all subsamples. This is consistent with our findings in Example 3 of Section 5.

## 7 Discussions

In this paper, we proposed an estimation framework via GMM to combine information from overidentified vectors and estimate an unknown subspace. This approach is applied to a variety of statistical problems.

A natural question that may be explored is to allow  $p$  or  $m$  to grow with  $n$ . The large  $p$  regime is relevant in the high dimensional literature, in particular, in the presence of sparsity or matrix incoherence structure, e.g. sparse principal component analysis (Zou et al., 2006), matrix completion (Candès and Recht, 2009), etc. The large  $m$  regime is related to moments selection in the GMM literature, which studies selecting informative moments from a large pool of candidate moment conditions (that is, vectors  $\mathbf{v}_\ell$  in our problem).

Also, it is interesting to see whether our method may help modern problems, such as (nonlinear) matrix completion and neural nets. Similar to the pHd, we might consider combining different transformations and activation functions for optimal results. For example, Cohen and Shashua (2016) considered tensor structure for convolutional networks, and Wang et al. (2018) studied data dependent link function (or activation function) for deep neural nets.

## SUPPLEMENTARY MATERIAL

This supplementary document consists of two additional simulation experiments, one additional real data example, proofs, and other details.

### A. Additional simulation and data examples

#### Example 3: a remedy for vanishing moments

In this example, we study multiple index models in Section 3.3. Related to Example 1 and 2, we illustrate how GMM methods are able to avoid the issue of vanishing moments by collecting information from different moments. We focus on three specific forms:

$$y_i = \cos(2\mathbf{x}_i^T \boldsymbol{\beta}_1) - \sin(\mathbf{x}_i^T \boldsymbol{\beta}_2) + 0.5\epsilon_i, \quad (\text{Model A})$$

$$y_i = \cos(2\mathbf{x}_i^T \boldsymbol{\beta}_1) - \mathbf{x}_i^T \boldsymbol{\beta}_2 + 0.5\epsilon_i, \quad (\text{Model B})$$

$$y_i = \cos(2\mathbf{x}_i^T \boldsymbol{\beta}_1) - \cos(\mathbf{x}_i^T \boldsymbol{\beta}_2) + 0.5\epsilon_i. \quad (\text{Model C})$$

Model C has been considered by Li (1992). Here we adopt the same parameter setup as Li (1992). Let  $n = 400$ ,  $p = 10$ , and  $r = 2$  (since there are only two  $\boldsymbol{\beta}$ s). We fix  $\boldsymbol{\beta}_1 = \mathbf{e}_1 = (1, 0, 0, \dots)^T$  and  $\boldsymbol{\beta}_2 = \mathbf{e}_2 = (0, 1, 0, \dots)^T$ , and generate i.i.d.  $\mathbf{x}_i \sim N(0, \mathbf{I}_p)$  and  $\epsilon_i \sim N(0, 1)$ . Set  $\mathbf{U}^* = \text{span}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ .

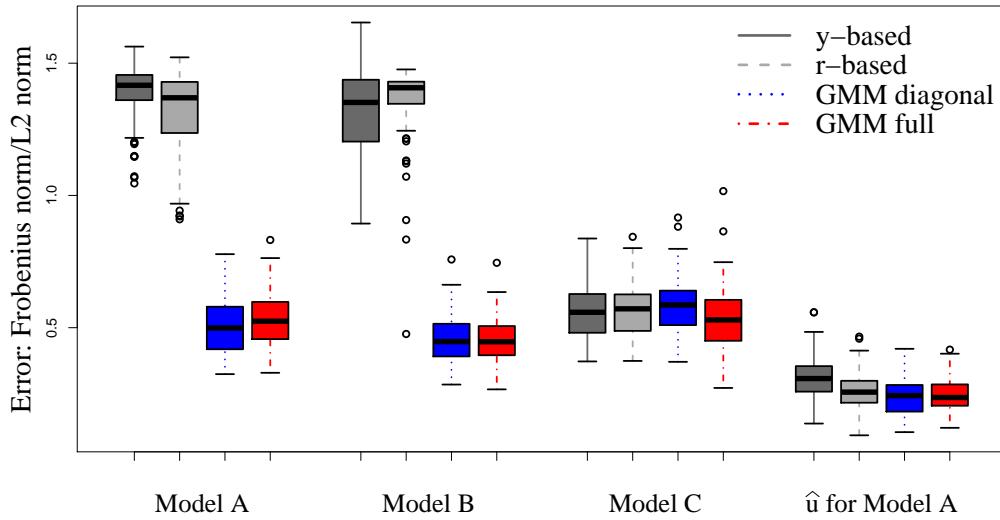


Figure 4: **First three groups:** Distribution of  $\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T - \mathbf{U}^*(\mathbf{U}^*)^T\|_F$  for four methods under three models from 100 simulations. **Last group:** Distribution of  $\|\mathbf{P}_{\mathbf{U}^*}^\perp \widehat{\mathbf{u}}\|_2$  under Model A from 100 simulations.

We compare four methods for these three models:  $y$ -based pHd method,  $r$ -based pHd method, GMM diagonal and GMM full. The first two methods are from [Li \(1992\)](#), and the last two are the same as in Example 1. For the GMM methods, we use first moments and transformed first moments as in Example 2 (see (a) and (c)), as well as the moments constructed from both pHd methods:

$$n^{-1} \sum_{i=1}^n y_i (\mathbf{x}_i \mathbf{x}_i^T \mathbf{e}_j - \mathbf{e}_j), \quad n^{-1} \sum_{i=1}^n r_i (\mathbf{x}_i \mathbf{x}_i^T \mathbf{e}_j - \mathbf{e}_j), \quad \forall j \in [p].$$

As [Li \(1992\)](#), we center the data  $\mathbf{x}_i$  and  $y_i$  first (but this is not essential in our case). We compute the same estimation errors for all methods and all models over 100 simulations. In addition, for the top eigenvector  $\hat{\mathbf{u}}$  produced by all four methods, we compute  $\mathbb{E}\|\mathbf{P}_{\mathbf{U}^*}^\perp \hat{\mathbf{u}}\|_2$ , which is the expected amount of the part of  $\hat{\mathbf{u}}$  unexplained by  $\mathcal{S}$ .

In Figure 4, the first three groups of boxplots show the distribution of  $\|\hat{\mathbf{U}}\hat{\mathbf{U}}^T - \mathbf{U}^*(\mathbf{U}^*)^T\|_F$ , and the last group shows that of  $\|\mathbf{P}_{(\mathbf{U}^*)^\perp} \hat{\mathbf{u}}\|_2$ . Clearly, the GMM methods are much better than the pHd methods for Model A and B, due to the fact that the pHd methods tend to miss linear trend (or more generally odd functions). On model C, all methods perform similarly, since pHd methods capture most useful information from second moments.

From the last group in Figure 4, we also observe that the quality of the first eigenvector  $\hat{\mathbf{u}}$  produced by pHd methods is roughly on a par with the GMM methods. This suggests that, pHd methods can, after all, find a first direction, though they fail to find a second (missing the direction of  $\beta_2$ ). This is where GMM methods are very effective, because they can find directions from various moments and collect them all.

## Estimation of Subspace Dimension

Lastly, we study the estimation of  $r$ . We consider the same factor model as in Example 1. As before, we fix  $p = 10$ ,  $\mu = 2$ ,  $\sigma = 2$ . We consider two cases:  $r = 2$  and  $r = 4$ , with  $\mu_z = (\mu, -\mu)^T$  or  $\mu_z = (\mu, -\mu, \mu, -\mu)^T$ . For both cases, we set  $n = 250r$ . The mechanism for sampling parameters and generating random variables remain the same.

For each case, we make three plots that correspond to different methods (see Figure 5 and 6) from 20 simulations. Let  $\lambda_k := \lambda_k(\mathbf{V}\mathbf{W}\mathbf{V}^T)$  be the  $k$ th largest eigenvalue computed from the GMM methods.

- The left plots show  $\lambda_k$  with different  $k$ .
- The middle plots show  $n(p-k) \sum_{j>k} \lambda_j$  with different  $k$ . The dashed curve with circle markers plots the critical value, namely 95% quantile of  $\chi^2_{(p-k)(m-k)}$ , for different  $k$ . Note that  $k$  starts from 0.
- The right plots show the eigen-ratio  $\lambda_k/\lambda_{k+1}$  with different  $k$ .

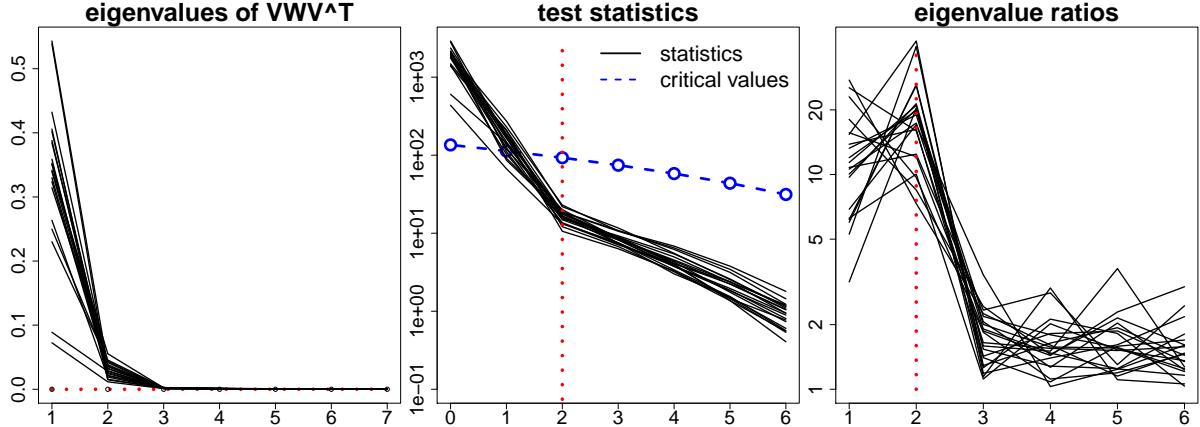


Figure 5:  $r = 2$ . We plot 20 solid curves computed from 20 simulations. The  $y$ -axis is on the log scale on the middle plot and the right plot.

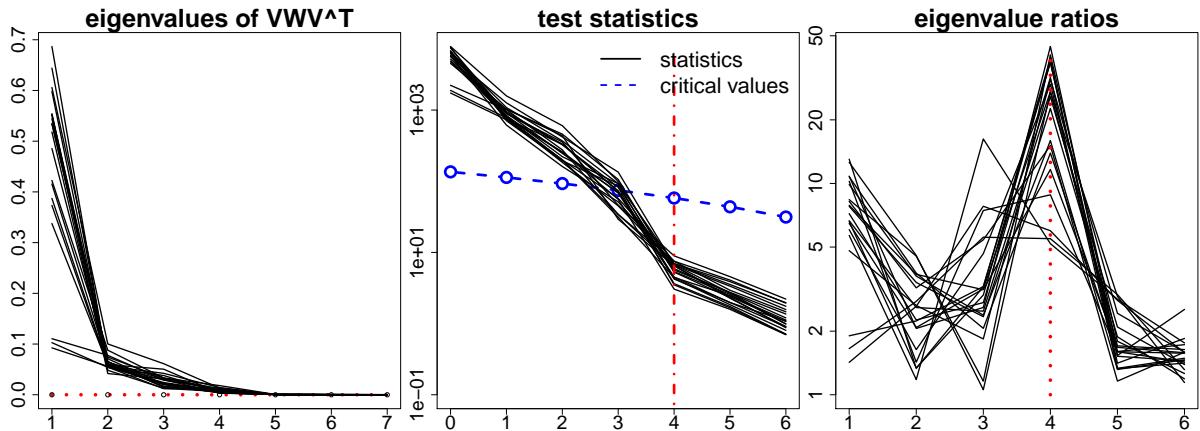


Figure 6:  $r = 4$ . Solid curves are computed from 20 simulations. The  $y$ -axis is on the log scale on the middle plot and the right plot.

Moreover, we add dotted lines on each plots. The horizontal dotted lines on the left plots have zero values on the  $y$ -axis. If some  $\lambda_j$  is close to this line, then the dimension  $r$  should be smaller than  $j$ . Indeed, in both cases, the eigenvalue curves do not visibly touch the dotted lines until  $j = r + 1$ . This leads to good dimension estimator  $\hat{r}_\tau$  with any reasonable parameter  $\tau_n$ .

On the middle and right plots, the vertical dotted lines represent the true dimension  $r$ . The method based on the chi-squared asymptotics (27) mostly likely leads to  $\hat{r} = r$  or  $\hat{r} = r - 1$ , which may underestimate  $r$ . This is probably due to the fact that the chi-squared asymptotics depend on stronger assumptions (see Theorem 4). For both eigen-ratio curves, we see a large spike at the true  $r$  in most simulations, which suggests they are useful in finding  $r$ .

In practice, we can consider combining all these plots to choose  $r$ . More sophisticated

methods, such as those based on information criterion (Bai and Ng, 2002), may be helpful, but they are out of the scope of this paper.

## The abalone dataset

The abalone dataset is a popular one from UCI machine learning repository (Dheeru and Karra Taniskidou, 2017). It contains  $n = 4177$  sets of measurements of abalone, where each set of measurements consists of 7 physical quantities with continuous values, and one discrete variable indicating the sex (male, female or infant). The goal is to predict the rings (or equivalently, the ages) of abalone. For our illustration, we will treat the sex as an unobserved (latent) variable, and study a mixed linear regression model with different methods.

We consider the mixed linear regression model (16) for our second data example. Each data unit has a ‘sex’ variable  $z$ , which can only take three values (male, female and infant). According to the value of this variable, we group our dataset into three sub-datasets, and set  $K = 3$ . We observe strong correlations between the 7 physical variables (the minimum correlation is 0.77), so we decide to only use their principal components. We fix  $p = 5$  and treat the 5 principal components as our covariates.

For each sub-dataset, we run a least squares regression of the ‘rings’ variable  $y$  against the covariates. The resulting coefficient vectors  $\beta_1, \beta_2, \beta_3 \in \mathbb{R}^p$  (without the intercepts) are treated as the true parameters, and our goal is to estimate  $\text{span}\{\beta_1, \beta_2, \beta_3\}$ . Henceforth, we treat  $z$  as a latent variable.

We run and compare three methods: standard (second moments), GMM diagonal and GMM full. The explanations of the three methods are in Table 1. Under the same Frobenius norm as considered in Section 5, the three methods give an error of 1.52, 1.44 and 0.78 respectively. We also implement the same procedure for 100 bootstrap subsamples and make boxplots of the errors. The results are shown in Figure 8.

As shown by the results, our two GMM estimators outperform the standard second moments. Note that this dataset is a challenging one, since the standard method produces an estimation error that exceeds 1. Even so, the GMM estimators have improvements over the standard method, which shows the robustness of our GMM estimators.

## B. Proofs

### Technical lemmas

The next lemma lists a few useful properties of the Kronecker product.

**Lemma 2.** *Suppose  $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$  are matrices with specified dimensions, or with appropriate dimensions such that matrix products are valid. Then, we have the following identities:*

- (i)  $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$ ;
- (ii)  $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$ ; (assuming both  $\mathbf{A}$  and  $\mathbf{B}$  are invertible)

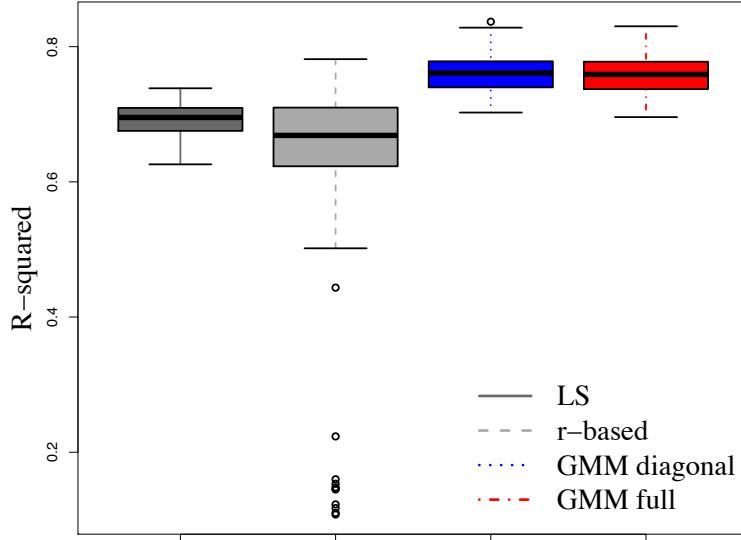


Figure 7: Comparison of the four methods based on 100 bootstrap samples. The R-squared values are calculated after fitting a quadratic regression (including cross products).

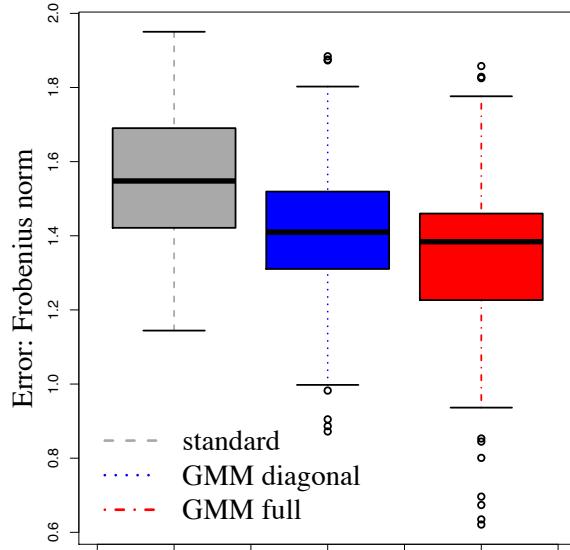


Figure 8: Comparison of the three methods from 100 bootstrap subsamples. The boxplots show the distributions of  $\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T - \mathbf{U}^*(\mathbf{U}^*)^T\|_F$ .

- (iii)  $(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T$ ;
- (iv)  $\text{rank}(\mathbf{A} \otimes \mathbf{B}) = \text{rank}(\mathbf{A}) \text{rank}(\mathbf{B})$ ;
- (v) If  $\mathbf{A}, \mathbf{B}$  are both orthogonal projection matrices, then  $\mathbf{A} \otimes \mathbf{B}$  is also an orthogonal projection;

- (vi)  $\text{Vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\text{Vec}(\mathbf{B})$ ; (recall  $\text{Vec}(\cdot)$  means vectorizing a matrix column-wise)
- (vii) There is a unique permutation matrix  $\mathbf{P}(p, m)$ , depending only on dimension  $p$  and  $m$ , such that  $\text{Vec}(\mathbf{A}^T) = \mathbf{P}(p, m)\text{Vec}(\mathbf{A})$  for any  $\mathbf{A} \in \mathbb{R}^{p \times m}$ ;
- (viii)  $\mathbf{A} \otimes \mathbf{B} = \mathbf{P}(p, m)(\mathbf{B} \otimes \mathbf{A})\mathbf{P}(p, m)^T$  holds for any  $\mathbf{A} \in \mathbb{R}^{p \times p}$ ,  $\mathbf{B} \in \mathbb{R}^{m \times m}$ , where  $\mathbf{P}(p, m)$  is given in (vii);
- (ix)  $\text{Tr}(\mathbf{A}^T \mathbf{B} \mathbf{A}) = (\text{Vec}(\mathbf{A}))^T (\mathbf{I}_p \otimes \mathbf{B}) \text{Vec}(\mathbf{A})$ , where  $p$  is the number of columns of  $\mathbf{A}$ .

The proof of these properties is known: (i)–(iv) are proved in Sect. 4.2 of [Horn and Johnson \(1991\)](#), (vi)–(viii) are proved in Sect. 4.3 of the same book, and (v) is straightforward from the definition. We give a short proof of (ix) below.

**Proof of Lemma 2 (ix).** Let us write  $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_p]$ , where  $\mathbf{a}_j \in \mathbb{R}^p$ . Using this notation, we have  $\text{Tr}(\mathbf{A}^T \mathbf{B} \mathbf{A}) = \sum_j \mathbf{a}_j^T \mathbf{B} \mathbf{a}_j$ . Since  $\mathbf{I}_p \otimes \mathbf{B}$  is a block diagonal matrix, we also have  $(\text{Vec}(\mathbf{A}))^T (\mathbf{I}_p \otimes \mathbf{B}) \text{Vec}(\mathbf{A}) = \sum_j \mathbf{a}_j^T \mathbf{B} \mathbf{a}_j$ , and thus we obtain the desired identity.  $\square$

We will also use an eigenvalue perturbation result from [Kato \(1966\)](#) (see Chap. 2, p. 79, Eq. (2.33)). Here we present a simplified form used in [Li \(1991\)](#) and [Li \(1992\)](#).

**Lemma 3.** Consider the asymptotic expansion

$$\mathbf{T}(\omega) = \mathbf{T} + \omega \mathbf{T}^{(1)} + \omega^2 \mathbf{T}^{(2)} + o(\omega^2),$$

where  $\omega = o(1)$ , and  $\mathbf{T}(\omega), \mathbf{T}, \mathbf{T}^{(1)}, \mathbf{T}^{(2)} \in \mathbb{R}^{p \times p}$  are symmetric matrices. Suppose that  $\mathbf{T}$  has rank  $k$ , where  $k < p$ . Let  $\lambda(\omega)$  be the sum of  $p - k$  eigenvalues of  $\mathbf{T}(\omega)$  with smallest absolute values. Let  $\mathbf{\Pi} \in \mathbb{R}^{p \times p}$  be the projection matrix associate with the null space of  $\mathbf{T}$  so that  $\mathbf{\Pi}\mathbf{T} = \mathbf{T}\mathbf{\Pi} = \mathbf{0}$ . Then,

$$\lambda(\omega) = \omega \lambda^{(1)} + \omega^2 \lambda^{(2)} + o(\omega^2),$$

where  $\lambda^{(1)} = \text{Tr}(\mathbf{T}^{(1)} \mathbf{\Pi})$  and  $\lambda^{(2)} = \text{Tr}(\mathbf{T}^{(2)} \mathbf{\Pi} - \mathbf{T}^{(1)} \mathbf{T}^+ \mathbf{T}^{(1)} \mathbf{\Pi})$ . Here we use  $+$  to denote the Moore-Penrose pseudoinverse.

The next lemma gives a useful elementary property about the Moore-Penrose pseudoinverse. Recall that for any matrix  $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ , the Moore-Penrose pseudoinverse of  $\mathbf{A}$  is denoted by  $\mathbf{A}^+$ .

**Lemma 4.** The matrix  $\mathbf{A}(\mathbf{A}^T \mathbf{A})^+ \mathbf{A}^T$  is an orthogonal projection matrix for any  $\mathbf{A}$ , and its rank is equal to the rank of  $\mathbf{A}$ .

*Proof of Lemma 4.* It can be shown that  $(\mathbf{A}^T \mathbf{A})^+ \mathbf{A}^T = \mathbf{A}^+$  for any  $\mathbf{A}$  (see [Ben-Israel and Greville \(2003\)](#) Chap. 1, p. 49, Ex. 18). Furthermore, it is known that  $\mathbf{A}\mathbf{A}^+$  is an orthogonal projection onto the range of  $\mathbf{A}$  ([Golub and Van Loan, 2013](#), Chap. 5.5.2). Thus, the conclusion follows.  $\square$

## Proofs for Section 4.1

**Proof of Lemma 1.** Denote  $\mathbf{P} = \mathbf{I}_p - \mathbf{U}^*(\mathbf{U}^*)^T$ . It is clear that  $\mathbf{P}$  is a deterministic projection matrix with rank  $p - r$ , so  $\text{Tr}(\mathbf{P}) = p - r$ . For simplicity, let us also denote  $\mathbf{f}_j = \mathbf{f}_j(\mathbf{x}_i, y_i)$  and  $\bar{\mathbf{f}}_j = \mathbb{E}\mathbf{f}_j$ . For any  $j, \ell \in [m]$ , by definition and linearity of  $\text{Tr}(\cdot)$  and  $\mathbb{E}(\cdot)$ ,

$$\begin{aligned}\Sigma_{j\ell}^* &= \text{Tr}(\mathbb{E}[\mathbf{f}_j^T \mathbf{P} \mathbf{f}_\ell]) = \mathbb{E}[\text{Tr}((\mathbf{f}_j - \bar{\mathbf{f}}_j)^T \mathbf{P} (\mathbf{f}_\ell - \bar{\mathbf{f}}_\ell))] \\ &= \mathbb{E}[\text{Tr}(\mathbf{P}(\mathbf{f}_\ell - \bar{\mathbf{f}}_\ell)(\mathbf{f}_j - \bar{\mathbf{f}}_j)^T)] \\ &= \text{Tr}(\mathbb{E}[\mathbf{P}(\mathbf{f}_\ell - \bar{\mathbf{f}}_\ell)(\mathbf{f}_j - \bar{\mathbf{f}}_j)^T]) \\ &= \text{Tr}(\mathbf{P} S_{j\ell} \mathbf{I}_p) \\ &= (p - r) S_{j\ell}.\end{aligned}$$

In the above derivation, we used the fact that  $\mathbf{P}\bar{\mathbf{f}}_j = \mathbf{0}$ , and that  $\mathbb{E}(\mathbf{f}_\ell - \bar{\mathbf{f}}_\ell)(\mathbf{f}_j - \bar{\mathbf{f}}_j)^T$  is the cross-covariance between  $\mathbf{f}_\ell$  and  $\mathbf{f}_j$ , and thus a submatrix of  $\bar{\mathbf{S}}$ .  $\square$

We state and prove a theorem that is more general than Theorem 7.

**Theorem 7.** Suppose Assumption 1 holds and  $\mathbf{W}_n \xrightarrow{p} \mathbf{W}^* \succeq \mathbf{0}$ , where  $\mathbf{W}^* \in \mathbb{R}^{m \times m}$  satisfies  $(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^* \succ \mathbf{0}$ . Then, there exists a sequence of orthogonal matrices  $\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \dots \in O(r)$  such that

$$\widehat{\mathbf{U}}_n \mathbf{R}_n \xrightarrow{p} \mathbf{U}^*, \quad \text{as } n \rightarrow \infty.$$

**Proof of Theorem 7.** Fix any  $\delta > 0$  independent of  $n$ . As stated, we suppress the subscript  $n$ . Let  $N_\delta(\mathbf{U}^*)$  be a neighborhood of  $\mathbf{U}^*$ , up to rotation:

$$\begin{aligned}N_\delta(\mathbf{U}^*) &= \{\mathbf{U} \in O(p, r) : \exists \mathbf{R} \in O(r), \|\mathbf{U} - \mathbf{U}^* \mathbf{R}\|_F < \delta\} \\ &= \cup_{\mathbf{R} \in O(r)} \{\mathbf{U} \in O(p, r) : \|\mathbf{U} - \mathbf{U}^* \mathbf{R}\|_F < \delta\},\end{aligned}$$

which is an open set in  $O(p, r)$ . Recall  $\bar{\mathbf{W}}^* = \mathbf{W}^* \otimes \mathbf{I}_p$  and  $Q^*(\mathbf{U}) = [\mathbb{E}\mathbf{g}(\mathbf{U})]^T \bar{\mathbf{W}}^* [\mathbb{E}\mathbf{g}(\mathbf{U})]$ . Also define  $\varepsilon := \inf_{\mathbf{U} \notin N_\delta(\mathbf{U}^*)} Q^*(\mathbf{U})$ . Note that  $Q^*(\mathbf{U})$  and  $\varepsilon$  do not depend on  $n$ . Also note that  $Q^*(\mathbf{U})$  is a continuous function in  $O(p, r)$ , and that the set  $[N_\delta(\mathbf{U}^*)]^c$ , namely, the complement of  $N_\delta(\mathbf{U}^*)$ , is compact. It follows that the minimum of  $Q^*(\mathbf{U})$  over  $[N_\delta(\mathbf{U}^*)]^c$  can be attained. We claim that  $\varepsilon > 0$ .

To prove this claim, recall  $\mathbf{F} = [\mathbf{f}_1, \dots, \mathbf{f}_m]$  and  $\mathbf{G}^* = (\mathbb{E}\mathbf{F})^T \mathbf{U}^*$ . Since  $\mathbb{E}(\mathbf{I}_p - \mathbf{U}\mathbf{U}^T)\mathbf{v}_\ell = (\mathbf{I}_p - \mathbf{U}\mathbf{U}^T)\mathbb{E}\mathbf{f}_\ell$  and  $\mathbb{E}\mathbf{F} = \mathbf{U}^*(\mathbf{U}^*)^T \mathbb{E}\mathbf{F} = \mathbf{U}^*(\mathbf{G}^*)^T$ , we can rewrite  $Q^*(\mathbf{U})$  as

$$\begin{aligned}Q^*(\mathbf{U}) &= \sum_{j,\ell=1}^m w_{j\ell}^* (\mathbb{E}\mathbf{f}_j)^T (\mathbf{I}_p - \mathbf{U}\mathbf{U}^T) \mathbb{E}\mathbf{f}_\ell = \text{Tr}(\mathbf{W}^* (\mathbb{E}\mathbf{F})^T (\mathbf{I}_p - \mathbf{U}\mathbf{U}^T) \mathbb{E}\mathbf{F}) \\ &= \text{Tr}(\mathbf{W}^* \mathbf{G}^*(\mathbf{U}^*)^T (\mathbf{I}_p - \mathbf{U}\mathbf{U}^T) \mathbf{U}^* (\mathbf{G}^*)^T) \\ &= \text{Tr}((\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^*(\mathbf{U}^*)^T (\mathbf{I}_p - \mathbf{U}\mathbf{U}^T) \mathbf{U}^*).\end{aligned}$$

Note that  $(\mathbf{U}^*)^T(\mathbf{I}_p - \mathbf{U}\mathbf{U}^T)\mathbf{U}^* \succeq \mathbf{0}$ , and by assumption  $(\mathbf{G}^*)^T\mathbf{W}^*\mathbf{G}^* \succ \mathbf{0}$ . Thus  $Q^*(\mathbf{U}) = 0$  if and only if  $(\mathbf{U}^*)^T(\mathbf{I}_p - \mathbf{U}\mathbf{U}^T)\mathbf{U}^* = \mathbf{0}$ . Observe that

$$\text{span}(\mathbf{U}^*) = \text{span}(\mathbf{U}) \Leftrightarrow (\mathbf{I}_p - \mathbf{U}\mathbf{U}^T)\mathbf{U}^* = \mathbf{0} \Leftrightarrow (\mathbf{U}^*)^T(\mathbf{I}_p - \mathbf{U}\mathbf{U}^T)\mathbf{U}^* = \mathbf{0}.$$

Let  $\mathbf{U}_0 \in N_\delta(\mathbf{U}^*)$  be the minimizer of  $Q^*(\mathbf{U})$ . It is clear that  $\text{span}(\mathbf{U}^*) \neq \text{span}(\mathbf{U}_0)$ , so from the above reasoning, we deduce  $\varepsilon = Q^*(\mathbf{U}_0) > 0$ .

Furthermore, since  $\mathbf{I}_p - \mathbf{U}\mathbf{U}^T$  is a projection matrix, for any  $\mathbf{U} \in O(p, r)$ , we have

$$\begin{aligned} \|\mathbf{g}(\mathbf{U}) - \mathbb{E}\mathbf{g}(\mathbf{U})\|_2^2 &= \sum_{\ell=1}^m \|(\mathbf{I}_p - \mathbf{U}\mathbf{U}^T)(\mathbf{v}_\ell - \mathbb{E}\mathbf{v}_\ell)\|_2^2 \leq \sum_{\ell=1}^m \|\mathbf{v}_\ell - \mathbb{E}\mathbf{v}_\ell\|_2^2 \xrightarrow{p} 0, \\ \|\bar{\mathbf{W}} - \bar{\mathbf{W}}^*\|_2 &= \|(\mathbf{W} - \mathbf{W}^*) \otimes \mathbf{I}_p\|_2 \xrightarrow{p} 0. \end{aligned}$$

These also imply  $\|\mathbf{g}(\mathbf{U})\|_2$  and  $\|\bar{\mathbf{W}}\|_2$  are uniformly bounded by a constant, and therefore, uniform convergence:

$$\sup_{\mathbf{U} \in O(p, r)} |Q(\mathbf{U}) - Q^*(\mathbf{U})| \xrightarrow{p} 0.$$

In particular, we have convergence in probability for  $\mathbf{U} = \hat{\mathbf{U}}$  and  $\mathbf{U} = \mathbf{U}^*$ . Using the inequality  $Q(\hat{\mathbf{U}}) \leq Q(\mathbf{U}^*)$  (by definition of  $\hat{\mathbf{U}}$ ), we have, for large  $n$ ,

$$Q^*(\hat{\mathbf{U}}) \leq Q(\hat{\mathbf{U}}) + \varepsilon/3 \leq Q(\mathbf{U}^*) + \varepsilon/3 \leq Q^*(\mathbf{U}^*) + 2\varepsilon/3 < \varepsilon.$$

Thus, we deduce that, for large  $n$ ,  $\hat{\mathbf{U}} \in N_\delta(\mathbf{U}^*)$ . Since  $\delta > 0$  is arbitrary, we conclude that for appropriate choice of  $\mathbf{R} \in O(r)$ , we have  $\hat{\mathbf{U}}\mathbf{R} \xrightarrow{p} \mathbf{U}^*$ .  $\square$

**Proof of Theorem 2.** By definition,  $\hat{\mathbf{U}}$  is a minimizer of (8). First, we establish a useful identity derived from the first-order optimality condition of  $\hat{\mathbf{U}}$ . This is achieved by making use of the connection with the eigenvector formulation in (9).

Since the columns of  $\hat{\mathbf{U}}$  are eigenvectors of  $\mathbf{V}\mathbf{W}\mathbf{V}^T$ , we must have

$$\mathbf{V}\mathbf{W}\mathbf{V}^T\hat{\mathbf{U}} = \hat{\mathbf{U}}\hat{\Lambda}, \quad \text{where } \hat{\Lambda} := \text{diag}\{\hat{\lambda}_1, \dots, \hat{\lambda}_r\}.$$

Here,  $\hat{\lambda}_1, \dots, \hat{\lambda}_r$  are the top  $r$  eigenvalues of  $\mathbf{V}^T\mathbf{W}\mathbf{V}$ . It leads to

$$\mathbf{P}_{\hat{\mathbf{U}}}^\perp \mathbf{V}\mathbf{W}\mathbf{V}^T\hat{\mathbf{U}} = \mathbf{P}_{\hat{\mathbf{U}}}^\perp \hat{\mathbf{U}}\hat{\Lambda} = \mathbf{0}, \tag{32}$$

Define  $\mathbf{G} := [\mathbf{v}_1, \dots, \mathbf{v}_m]^T \hat{\mathbf{U}}\mathbf{R} \in \mathbb{R}^{m \times r}$ . Right multiplying both sides of (32) by  $\mathbf{R}$ , we obtain

$$\mathbf{P}_{\hat{\mathbf{U}}}^\perp \mathbf{V}\mathbf{W}\mathbf{G} = \mathbf{0}. \tag{33}$$

Now we observe  $\mathbf{P}_{\hat{\mathbf{U}}}^\perp (\hat{\mathbf{U}}\mathbf{R} - \mathbf{U}^*) = -\mathbf{P}_{\hat{\mathbf{U}}}^\perp \mathbf{U}^*$  since  $\mathbf{P}_{\hat{\mathbf{U}}}^\perp \hat{\mathbf{U}} = \mathbf{0}$ . Thus, defining  $\mathbf{H} := \mathbf{V}^T\mathbf{U}^* \in$

$\mathbb{R}^{m \times r}$ , we obtain

$$\begin{aligned}\mathbf{P}_{\widehat{\mathbf{U}}}^\perp (\mathbf{P}_{\widehat{\mathbf{U}}}^\perp \mathbf{V} - \mathbf{P}_{\mathbf{U}^*}^\perp \mathbf{V}) &= -\mathbf{P}_{\widehat{\mathbf{U}}}^\perp \left( \widehat{\mathbf{U}} \widehat{\mathbf{U}}^T \mathbf{V} - (\mathbf{U}^*)(\mathbf{U}^*)^T \mathbf{V} \right) = \mathbf{P}_{\widehat{\mathbf{U}}}^\perp (\mathbf{U}^*)(\mathbf{U}^*)^T \mathbf{V} \\ &= -\mathbf{P}_{\widehat{\mathbf{U}}}^\perp (\widehat{\mathbf{U}} \mathbf{R} - \mathbf{U}^*) \mathbf{H}^T.\end{aligned}$$

Using (33), we right multiply the above identity by  $\mathbf{WG}$  and obtain

$$-\mathbf{P}_{\widehat{\mathbf{U}}}^\perp \mathbf{P}_{\mathbf{U}^*}^\perp \mathbf{V} \mathbf{W} \mathbf{G} = -\mathbf{P}_{\widehat{\mathbf{U}}}^\perp (\widehat{\mathbf{U}} \mathbf{R} - \mathbf{U}^*) \mathbf{H}^T \mathbf{W} \mathbf{G}.$$

To derive asymptotic properties for  $\mathbf{P}_{\widehat{\mathbf{U}}}^\perp (\widehat{\mathbf{U}} \mathbf{R} - \mathbf{U}^*)$ , we wish to make inversion of  $\mathbf{H}^T \mathbf{W} \mathbf{G}$  in the above equality. To that end, we make the following observation. Both  $\mathbf{G}$  and  $\mathbf{H}$  converge to the same limit in  $\mathbb{R}^{m \times r}$ , which is  $\mathbf{G}^* = [\mathbb{E} \mathbf{v}_1, \dots, \mathbb{E} \mathbf{v}_m]^T \mathbf{U}^* \in \mathbb{R}^{m \times r}$ , and it has full column rank under Assumption 1. By Assumption 2,  $\mathbf{W} \xrightarrow{p} \mathbf{W}^* \succ \mathbf{0}$ , so  $\mathbf{G}^T \mathbf{W} \mathbf{H}$  converges to  $(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^*$  in probability. These two facts also imply  $(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^* \succ \mathbf{0}$ . Thus, with probability  $1 - o(1)$ ,  $\mathbf{G}^T \mathbf{W} \mathbf{H}$  is invertible, and its limit  $(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^*$  is also invertible. (Note that we have the same conclusion if we directly assumed  $(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^* \succ \mathbf{0}$  instead of the stronger condition  $\mathbf{W}^* \succ \mathbf{0}$ .)

Therefore, with probability  $1 - o(1)$ , we obtain

$$\mathbf{P}_{\widehat{\mathbf{U}}}^\perp (\widehat{\mathbf{U}} \mathbf{R} - \mathbf{U}^*) = \mathbf{P}_{\widehat{\mathbf{U}}}^\perp \mathbf{P}_{\mathbf{U}^*}^\perp \mathbf{V} \mathbf{W} \mathbf{G} (\mathbf{H}^T \mathbf{W} \mathbf{G})^{-1}. \quad (34)$$

Note that  $\text{Vec}(\mathbf{P}_{\mathbf{U}^*}^\perp \mathbf{V}) = \mathbf{g}(\mathbf{U}^*)$ . Using Lemma 2 (vi), we express (34) into the vector form:

$$\text{Vec}(\mathbf{P}_{\widehat{\mathbf{U}}}^\perp (\widehat{\mathbf{U}} \mathbf{R} - \mathbf{U}^*)) = ((\mathbf{G}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{G}^T \mathbf{W} \otimes \mathbf{P}_{\widehat{\mathbf{U}}}^\perp) \mathbf{g}(\mathbf{U}^*)$$

Recall that  $\bar{\mathbf{S}}^* \in \mathbb{R}^{\bar{m} \times \bar{m}}$  is the covariance matrix of  $[\mathbf{f}_1(\mathbf{x}_i, y_i); \dots; \mathbf{f}_m(\mathbf{x}_i, y_i)] \in \mathbb{R}^{\bar{m}}$  (concatenating all  $\mathbf{f}_\ell$  into a vector). Then, by the central limit theorem,

$$\sqrt{n} \mathbf{g}(\mathbf{U}^*) \xrightarrow{d} N(\mathbf{0}, \bar{\mathbf{P}}_{\mathbf{U}^*}^\perp \bar{\mathbf{S}}^* \bar{\mathbf{P}}_{\mathbf{U}^*}^\perp), \quad \text{where } \bar{\mathbf{P}}_{\mathbf{U}^*}^\perp = \mathbf{I}_p \otimes \mathbf{P}_{\mathbf{U}^*}^\perp.$$

Note that the central limit theorem applies despite the covariance matrix of the asymptotic normal distribution being rank deficient. Moreover, since  $(\mathbf{G}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{G}^T \mathbf{W} \xrightarrow{p} [(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^*]^{-1} (\mathbf{G}^*)^T \mathbf{W}^*$  and  $\mathbf{P}_{\widehat{\mathbf{U}}}^\perp \xrightarrow{p} \mathbf{P}_{\mathbf{U}^*}^\perp$ , we have

$$(\mathbf{G}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{G}^T \mathbf{W} \otimes \mathbf{P}_{\widehat{\mathbf{U}}}^\perp \xrightarrow{p} [(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^*]^{-1} (\mathbf{G}^*)^T \mathbf{W}^* \otimes \mathbf{P}_{\mathbf{U}^*}^\perp.$$

Using Slutsky's Theorem, Lemma 2, and the fact that  $(\mathbf{P}_{\mathbf{U}^*}^\perp)^2 = \mathbf{P}_{\mathbf{U}^*}^\perp$ , we arrive at the desired asymptotic normality (21). In particular, when  $\bar{\mathbf{S}}^*$  has the block matrix form  $\bar{\mathbf{S}}^* = \mathbf{S}^* \otimes \mathbf{I}_p$ , where  $\mathbf{S}^* \in \mathbb{R}^{m \times m}$  is invertible, then the asymptotic variance simplifies to  $(\mathbf{A} \mathbf{S}^* \mathbf{A}^T) \otimes \mathbf{P}_{\mathbf{U}^*}^\perp$ . From the standard GMM theory (Hall, 2005), we deduce that the choice  $\mathbf{W}^* = (\mathbf{S}^*)^{-1}$  is the optimal weighting matrix, in the sense that  $\mathbf{A} \mathbf{S}^* \mathbf{A}^T \succeq [(\mathbf{G}^*)^T (\mathbf{S}^*)^{-1} \mathbf{G}^*]^{-1}$  for any  $\mathbf{W}^* \succ \mathbf{0}$ . It follows that (22) is true, and the equality in (22) can be attained at the choice

$\mathbf{W}^* = (\mathbf{S}^*)^{-1}$ . Note that rescaling  $\mathbf{W}^*$  does not change the asymptotic variance in (21). This completes the proof.  $\square$

**Proof of Theorem 3.** In the proof of Theorem 2, we obtained (34) under Assumption 1 and 2. Recall that  $\mathbf{P}_{\widehat{\mathbf{U}}}^\perp(\widehat{\mathbf{U}}\mathbf{R} - \mathbf{U}^*) = -\mathbf{P}_{\widehat{\mathbf{U}}}^\perp\mathbf{U}^*$ , so we have

$$(\mathbf{U}^*)^T \mathbf{P}_{\widehat{\mathbf{U}}}^\perp \mathbf{U}^* = [\mathbf{P}_{\widehat{\mathbf{U}}}^\perp(\widehat{\mathbf{U}}\mathbf{R} - \mathbf{U}^*)]^T \mathbf{P}_{\widehat{\mathbf{U}}}^\perp(\widehat{\mathbf{U}}\mathbf{R} - \mathbf{U}^*).$$

Denote  $\mathbf{A} := (\mathbf{G}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{G}^T \mathbf{W} \in \mathbb{R}^{r \times m}$ . Using (34), we have

$$(\mathbf{U}^*)^T \mathbf{P}_{\widehat{\mathbf{U}}}^\perp \mathbf{U}^* = \mathbf{A} \mathbf{V}^T \mathbf{P}_{\mathbf{U}^*}^\perp \mathbf{P}_{\widehat{\mathbf{U}}}^\perp \mathbf{P}_{\mathbf{U}^*}^\perp \mathbf{V} \mathbf{A}^T. \quad (35)$$

Recall that, in Theorem 2, we have defined  $\mathbf{A}^* := [(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^*]^{-1} (\mathbf{G}^*)^T \mathbf{W}^*$ . Thus,

$$\mathbf{A} \xrightarrow{p} \mathbf{A}^* \quad \text{and} \quad \mathbf{P}_{\widehat{\mathbf{U}}}^\perp \xrightarrow{p} \mathbf{P}_{\mathbf{U}^*}^\perp. \quad (36)$$

Moreover, by the central limit theorem,  $\sqrt{n} \operatorname{Vec}(\mathbf{P}_{\mathbf{U}^*}^\perp \mathbf{V}) \xrightarrow{d} N(\mathbf{0}, \bar{\mathbf{P}}_{\mathbf{U}^*}^\perp \bar{\mathbf{S}}^* \bar{\mathbf{P}}_{\mathbf{U}^*}^\perp)$  where, recall,  $\bar{\mathbf{S}}^*$  is the covariance matrix of the concatenated vector  $\bar{\mathbf{f}}_j$ s. Let  $\boldsymbol{\Xi} = [\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_m] \in \mathbb{R}^{p \times m}$  be a random matrix such that  $\operatorname{Vec}(\boldsymbol{\Xi}) \sim N(\mathbf{0}, \bar{\mathbf{S}}^*)$ . Then,  $\operatorname{Vec}(\mathbf{P}_{\mathbf{U}^*}^\perp \boldsymbol{\Xi}) \sim N(\mathbf{0}, \bar{\mathbf{P}}_{\mathbf{U}^*}^\perp \bar{\mathbf{S}}^* \bar{\mathbf{P}}_{\mathbf{U}^*}^\perp)$ . By the (multivariate) Slutsky's theorem (Van der Vaart, 1998, Thm. 2.7),

$$n(\mathbf{U}^*)^T \mathbf{P}_{\widehat{\mathbf{U}}}^\perp \mathbf{U}^* \xrightarrow{d} \mathbf{A}^* \boldsymbol{\Xi}^T \mathbf{P}_{\mathbf{U}^*}^\perp \boldsymbol{\Xi} (\mathbf{A}^*)^T.$$

We claim that  $\boldsymbol{\Sigma}^* = \mathbb{E}[\boldsymbol{\Xi}^T \mathbf{P}_{\mathbf{U}^*}^\perp \boldsymbol{\Xi}]$ . In fact, for any  $j, \ell \in [m]$ , we have  $\mathbb{E}[\boldsymbol{\xi}_j^T \mathbf{P}_{\mathbf{U}^*}^\perp \boldsymbol{\xi}_\ell] = \operatorname{Tr}(\operatorname{Cov}[\mathbf{P}_{\mathbf{U}^*}^\perp \boldsymbol{\xi}_\ell, \mathbf{P}_{\mathbf{U}^*}^\perp \boldsymbol{\xi}_j])$  and  $\boldsymbol{\Sigma}_{j\ell}^* = \mathbb{E}[\mathbf{f}_j^T \mathbf{P}_{\mathbf{U}^*}^\perp \mathbf{f}_\ell] = \operatorname{Tr}(\operatorname{Cov}[\mathbf{P}_{\mathbf{U}^*}^\perp \mathbf{f}_\ell, \mathbf{P}_{\mathbf{U}^*}^\perp \mathbf{f}_j])$ . By definition,  $\operatorname{Vec}(\boldsymbol{\Xi})$  and  $\operatorname{Vec}(\mathbf{F})$  share the same covariance matrix  $\bar{\mathbf{S}}^*$ , and therefore  $\boldsymbol{\Sigma}_{j\ell} = \mathbb{E}[\boldsymbol{\xi}_j^T \mathbf{P}_{\mathbf{U}^*}^\perp \boldsymbol{\xi}_\ell]$ . This leads to

$$n \operatorname{AE}(\boldsymbol{\Psi}(\mathbf{W})) = n \operatorname{AE}((\mathbf{U}^*)^T \mathbf{P}_{\widehat{\mathbf{U}}}^\perp \mathbf{U}^*) = \mathbf{A}^* \boldsymbol{\Sigma}^* (\mathbf{A}^*)^T. \quad (37)$$

In particular, if  $\mathbf{W}$  is fixed at  $\mathbf{W}^*$  for all  $n$ , then we also have  $n \operatorname{AE}(\boldsymbol{\Psi}(\mathbf{W}^*)) = \mathbf{A}^* \boldsymbol{\Sigma}^* (\mathbf{A}^*)^T$ . With the choice  $\mathbf{W}^* = (\boldsymbol{\Sigma}^*)^{-1}$ , we get  $\operatorname{AE}(\boldsymbol{\Psi}((\boldsymbol{\Sigma}^*)^{-1})) = n^{-1}[(\mathbf{G}^*)^T (\boldsymbol{\Sigma}^*)^{-1} \mathbf{G}^*]^{-1}$ . This is the smallest asymptotic expectation up to a  $(1 + o(1))$  factor, since for any  $\mathbf{W}^* \succ \mathbf{0}$ ,

$$[(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^*]^{-1} (\mathbf{G}^*)^T \mathbf{W}^* \boldsymbol{\Sigma}^* \mathbf{W}^* \mathbf{G}^* [(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^*]^{-1} \succeq [(\mathbf{G}^*)^T (\boldsymbol{\Sigma}^*)^{-1} \mathbf{G}^*]^{-1}. \quad (38)$$

The above inequality is well known (Hall, 2005). This readily implies the weighting matrix  $\mathbf{W}^* = (\boldsymbol{\Sigma}^*)^{-1}$  is optimal. Moreover, assuming  $\{n \|(\mathbf{U}^*)^T \mathbf{P}_{\widehat{\mathbf{U}}}^\perp \mathbf{U}^*\|_F\}$  is uniformly integrable for  $\mathbf{W}^*$ , we obtain

$$\lim_{n \rightarrow \infty} n \mathbb{E}[(\mathbf{U}^*)^T \mathbf{P}_{\widehat{\mathbf{U}}}^\perp \mathbf{U}^*] = n \operatorname{AE}(\boldsymbol{\Psi}(\mathbf{W}^*))$$

by standard results (see, for example, Thm. 3.2.2 and Thm. 5.5.2 in Durrett (2010)). Therefore, assuming uniform integrability, we obtain  $\mathbb{E} \boldsymbol{\Psi}((\boldsymbol{\Sigma}^*)^{-1}) \preceq (1 + o(1)) \mathbb{E} \boldsymbol{\Psi}(\mathbf{W}^*)$ . Finally,

$\widehat{\Sigma} \xrightarrow{p} \Sigma^*$ , and under Assumption 4 we have  $(\widehat{\Sigma})^{-1} \xrightarrow{p} (\Sigma^*)^{-1}$ , so the choice of  $\mathbf{W} = (\widehat{\Sigma})^{-1}$  satisfies Assumption 2, and our proof is complete.  $\square$

**Proof of Corollary 1.** We suppress the dependence on  $\mathbf{W}^*$  whenever there is no confusion. Let the singular value decomposition of  $\widehat{\mathbf{U}}^T \mathbf{U}^*$  be  $\mathbf{U}_0 \Sigma_0 \mathbf{V}_0^T$ , where  $\Sigma_0 = \text{diag}\{\sigma_1, \dots, \sigma_r\}$  is a diagonal matrix with singular values on its diagonal, and  $\mathbf{U}_0, \mathbf{V}_0 \in O(r)$  are orthogonal matrices. Recall the definition of canonical angles  $\theta_1, \dots, \theta_r$  and  $\sin \Theta = \text{diag}\{\sin \theta_1, \dots, \sin \theta_r\}$  in Section 2.3. Using  $\cos \theta_k = \sigma_k$ , we deduce

$$\begin{aligned} \text{Tr}((\mathbf{U}^*)^T \mathbf{P}_{\widehat{\mathbf{U}}}^\perp \mathbf{U}^*) &= \text{Tr}(\mathbf{I}_r - \mathbf{V}_0 \Sigma_0^T \Sigma_0 \mathbf{V}_0^T) = r - \text{Tr}(\Sigma_0^T \Sigma_0) \\ &= r - \sum_{k=1}^r \sigma_k^2 = \sum_{k=1}^r \sin^2 \theta_k \\ &= \|\sin \Theta\|_F^2. \end{aligned}$$

For any  $\mathbf{W}^* \succ \mathbf{0}$ , by (37),  $n \cdot \text{AE}(\Psi(\mathbf{W}^*))$  is the distributional limit of  $n(\mathbf{U}^*)^T \mathbf{P}_{\widehat{\mathbf{U}}}^\perp \mathbf{U}^*$ . Taking the trace and using the above equality, we have  $n \|\sin \Theta(\mathbf{W}^*)\|_F^2 \xrightarrow{d} n \text{Tr}(\text{AE}(\Psi(\mathbf{W}^*)))$ , and thus  $\text{AE}(\|\sin \Theta(\mathbf{W}^*)\|_F^2) = \text{Tr}(\text{AE}(\Psi(\mathbf{W}^*)))$  up to a  $1+o(1)$  factor. Since  $\text{AE}(\Psi((\widehat{\Sigma})^{-1})) \leq \text{AE}(\Psi(\mathbf{W}^*))$  for any  $\mathbf{W}^* \succ \mathbf{0}$  by Theorem 3, we must have

$$\text{Tr}(\text{AE}(\Psi((\widehat{\Sigma})^{-1}))) \leq \text{Tr}(\text{AE}(\Psi(\mathbf{W}^*))), \quad \forall \mathbf{W}^* \succ \mathbf{0}.$$

Expressing this inequality equivalently in term of canonical angles, we have

$$\text{AE}(\|\sin \Theta((\widehat{\Sigma})^{-1})\|_F^2) \leq \text{AE}(\|\sin \Theta(\mathbf{W}^*)\|_F^2), \quad \forall \mathbf{W}^* \succ \mathbf{0}.$$

Finally, using the equivalence (23), we obtain the desired inequality.  $\square$

## Proofs for Section 4.3

**Proof of Theorem 4.** We use the asymptotic result in Lemma 3 to prove the theorem. Recall  $\mathbf{P}_{\mathbf{U}^*}^\perp = \mathbf{I}_p - \mathbf{U}^*(\mathbf{U}^*)^T$ , which we write  $\mathbf{P}$  for shorthand. Also denote  $\mathbf{V}^* = \mathbb{E}\mathbf{V}$ ,  $\omega_n = n^{-1/2}$ . We can rewrite  $\mathbf{V}\mathbf{W}\mathbf{V}^T$  as

$$\begin{aligned} \mathbf{V}\mathbf{W}\mathbf{V}^T &= \mathbf{V}^*\mathbf{W}(\mathbf{V}^*)^T + \omega_n [\sqrt{n}(\mathbf{V} - \mathbf{V}^*)\mathbf{W}(\mathbf{V}^*)^T + \sqrt{n}\mathbf{V}^*\mathbf{W}(\mathbf{V} - \mathbf{V}^*)^T] \\ &\quad + \omega_n^2 n(\mathbf{V} - \mathbf{V}^*)\mathbf{W}(\mathbf{V} - \mathbf{V}^*)^T =: \mathbf{T} + \omega_n \mathbf{T}^{(1)} + \omega_n^2 \mathbf{T}^{(2)}. \end{aligned}$$

Note each column of  $\mathbf{V} - \mathbf{V}^*$  is a sample average of  $\mathbf{f}_\ell(\mathbf{x}_i, y_i) - \mathbb{E}\mathbf{f}_\ell(\mathbf{x}_i, y_i)$ , so  $\sqrt{n}(\mathbf{V} - \mathbf{V}^*)$  is of the order  $O_P(1)$ . Also, since  $\mathbf{W} \xrightarrow{p} \mathbf{W}^* \succ \mathbf{0}$ , we deduce that with probability  $1 - o(1)$ , the rank of  $\mathbf{V}^*\mathbf{W}(\mathbf{V}^*)^T$  is  $r$  due to Assumption 1, and that the projection matrix associated with its null space is exactly  $\mathbf{P}$ . Following a similar argument as in Li (1991) and Li (1992),

by Lemma 3, we deduce<sup>2</sup>

$$\bar{\lambda} := \sum_{j=r+1}^n \lambda_j(\mathbf{V}\mathbf{W}\mathbf{V}^T) = \omega_n \lambda^{(1)} + \omega_n^2 \lambda^{(2)} + o_P(\omega_n^2), \quad (39)$$

where  $\lambda^{(1)} = \text{Tr}(\mathbf{T}^{(1)}\mathbf{P})$  and  $\lambda^{(2)} = \text{Tr}(\mathbf{T}^{(2)}\mathbf{P} - \mathbf{T}^{(1)}\mathbf{T}^+\mathbf{T}^{(1)}\mathbf{P})$ . Note that the column vectors of  $\mathbf{V}^*$  (namely  $\mathbb{E}\mathbf{v}_j$ s) lie in the subspace  $\mathcal{S}$ , so  $(\mathbf{V}^*)^T\mathbf{P} = \mathbf{0}$ . This yields

$$\begin{aligned} \text{Tr}((\mathbf{V} - \mathbf{V}^*)\mathbf{W}(\mathbf{V}^*)^T\mathbf{P}) &= 0, \quad \text{and} \\ \text{Tr}(\mathbf{V}^*\mathbf{W}(\mathbf{V} - \mathbf{V}^*)^T\mathbf{P}) &= \text{Tr}(\mathbf{P}\mathbf{V}^*\mathbf{W}(\mathbf{V} - \mathbf{V}^*)^T) = 0. \end{aligned}$$

Adding the above two equalities, we have  $\text{Tr}(\mathbf{T}^{(1)}\mathbf{P}) = 0$ , which implies  $\lambda^{(1)} = 0$ . Thus, the dominant term in the expansion of  $\bar{\lambda}$  is  $\omega_n^2 \lambda^{(2)}$ . Now we simplify  $\bar{\lambda}$ :

$$\begin{aligned} \bar{\lambda} &= \omega_n^2 \text{Tr}(\mathbf{T}^{(2)}\mathbf{P}) - \omega_n^2 \text{Tr}(\mathbf{T}^{(1)}\mathbf{T}^+\mathbf{T}^{(1)}\mathbf{P}) \\ &= \omega_n^2 \text{Tr}(\mathbf{T}^{(2)}\mathbf{P}) - \omega_n^2 \text{Tr}(\mathbf{P}\mathbf{T}^{(1)}\mathbf{T}^+\mathbf{T}^{(1)}\mathbf{P}) \\ &= \text{Tr}(\mathbf{P}(\mathbf{V} - \mathbf{V}^*)\mathbf{W}(\mathbf{V} - \mathbf{V}^*)^T\mathbf{P}) - \text{Tr}(\mathbf{P}(\mathbf{V} - \mathbf{V}^*)\widetilde{\mathbf{W}}(\mathbf{V} - \mathbf{V}^*)^T\mathbf{P}), \end{aligned}$$

where  $\widetilde{\mathbf{W}} := \mathbf{W}(\mathbf{V}^*)^T\mathbf{T}^+\mathbf{V}^*\mathbf{W}$ . In the second line, we used the identities  $\mathbf{P}^2 = \mathbf{P}$  and  $\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA})$ ; and in the third line, we used the equalities

$$\mathbf{T}^{(1)}\mathbf{P} = \sqrt{n}\mathbf{V}^*\mathbf{W}(\mathbf{V} - \mathbf{V}^*)^T\mathbf{P}, \quad \mathbf{P}\mathbf{T}^{(1)} = \sqrt{n}\mathbf{P}(\mathbf{V} - \mathbf{V}^*)\mathbf{W}(\mathbf{V}^*)^T.$$

Next we denote  $\mathbf{Z} = (\mathbf{V} - \mathbf{V}^*)^T\mathbf{P} = \mathbf{V}^T\mathbf{P} \in \mathbb{R}^{m \times p}$ , and simplify the above expression using Lemma 2 (ix):

$$\bar{\lambda} = \text{Tr}(\mathbf{Z}^T(\mathbf{W} - \widetilde{\mathbf{W}})\mathbf{Z}) = (\text{Vec}(\mathbf{Z}))^T [\mathbf{I}_p \otimes (\mathbf{W} - \widetilde{\mathbf{W}})] \text{Vec}(\mathbf{Z})$$

Note that  $\mathbf{Z}^T = \mathbf{P}\mathbf{V}$  and  $\text{Vec}(\mathbf{Z}^T) = \mathbf{g}(\mathbf{U}^*)$ , so by the central limit theorem,  $\sqrt{n}\text{Vec}(\mathbf{Z}^T) \xrightarrow{d} N(\mathbf{0}, \bar{\Sigma})$  where  $\bar{\Sigma} = \bar{\mathbf{P}}^*\bar{\mathbf{S}}^*\bar{\mathbf{P}}^*$  (recall  $\bar{\mathbf{P}}^* = \mathbf{I}_p \otimes \mathbf{P}$ ). Similarly,  $\sqrt{n}\text{Vec}(\mathbf{Z}) \xrightarrow{d} N(\mathbf{0}, \bar{\Sigma}')$ , where  $\bar{\Sigma}'$  is a row-wise and column-wise permuted version of  $\bar{\Sigma}$  (Lemma 2 (vii)). Also, by Assumption 2, we have  $\mathbf{W} \xrightarrow{p} \mathbf{W}^*$ ; moreover, we claim

$$\widetilde{\mathbf{W}} \xrightarrow{p} \widetilde{\mathbf{W}}^*, \quad \text{where} \quad \widetilde{\mathbf{W}}^* = \mathbf{W}^*(\mathbf{V}^*)^T [\mathbf{V}^*\mathbf{W}^*(\mathbf{V}^*)^T]^+ \mathbf{V}^*\mathbf{W}^*.$$

In fact, although the Moore-Penrose pseudoinverse is not a continuous map in general, we still have  $[\mathbf{V}^*\mathbf{W}(\mathbf{V}^*)^T]^+ \xrightarrow{p} [\mathbf{V}^*\mathbf{W}^*(\mathbf{V}^*)^T]^+$ , because with probability  $1 - o(1)$ , the matrix

---

<sup>2</sup>We think this argument is not rigorous in the cited papers, as Lemma 3 is an asymptotic result for fixed matrices, whereas we substitute  $\mathbf{T}^{(1)}, \mathbf{T}^{(2)}$  by random matrices. This issue, however, can be easily resolved; see the next proof.

$(\mathbf{V}^*)^T \mathbf{W} \mathbf{V}^*$  has exactly rank  $r$ . Thus, by the (multivariate) Slutsky's theorem,

$$n\bar{\lambda} \xrightarrow{d} \boldsymbol{\xi}^T (\bar{\Sigma}')^{1/2} \left[ \mathbf{I}_p \otimes (\mathbf{W}^* - \widetilde{\mathbf{W}}^*) \right] (\bar{\Sigma}')^{1/2} \boldsymbol{\xi}, \quad (40)$$

where  $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{I}_{pm})$ . This implies  $\bar{\lambda} = O_P(1/n)$  and thus proves the first claim of Theorem 4. Moreover, assuming Assumption 3 in addition, we have

$$\bar{\Sigma} = (\mathbf{I}_p \otimes \mathbf{P})(\mathbf{S}^* \otimes \mathbf{I}_p)(\mathbf{I}_p \otimes \mathbf{P}) = \mathbf{S}^* \otimes \mathbf{P},$$

due to Lemma 2 (i). We can proceed to simplify the right-hand side (RHS) of (40).  $\bar{\Sigma}' = \mathbf{P} \otimes \mathbf{S}^*$  by Lemma 2 (vii) and (viii), and thus  $(\bar{\Sigma}')^{1/2} = \mathbf{P} \otimes (\mathbf{S}^*)^{1/2}$  by Lemma 2 (i). Therefore,

$$\text{RHS of (40)} = \boldsymbol{\xi}^T \left[ \mathbf{P} \otimes \left( (\mathbf{S}^*)^{1/2} (\mathbf{W}^* - \widetilde{\mathbf{W}}^*) (\mathbf{S}^*)^{1/2} \right) \right] \boldsymbol{\xi} =: \boldsymbol{\xi}^T \mathbf{M} \boldsymbol{\xi}. \quad (41)$$

By Lemma 1,  $\Sigma^* = (p-r)\mathbf{S}^*$ , so with the choice  $\mathbf{W}^* = (\Sigma^*)^{-1}$ , we derive  $(\mathbf{S}^*)^{1/2} \mathbf{W}^* (\mathbf{S}^*)^{1/2} = \frac{1}{p-r} \mathbf{I}_m$ . We also rewrite  $(\mathbf{S}^*)^{1/2} \widetilde{\mathbf{W}}^* (\mathbf{S}^*)^{1/2}$  as

$$(\mathbf{S}^*)^{1/2} \widetilde{\mathbf{W}}^* (\mathbf{S}^*)^{1/2} = \frac{1}{p-r} \mathbf{K} [\mathbf{K}^T \mathbf{K}]^+ \mathbf{K}^T,$$

where  $\mathbf{K} := (\mathbf{S}^*)^{-1/2} (\mathbf{V}^*)^T \in \mathbb{R}^{m \times p}$ . The matrix  $\mathbf{K}$  has rank  $r$ , since  $(\mathbf{S}^*)^{-1/2}$  has full rank and  $\mathbf{V}^*$  has rank  $r$ . Therefore, by Lemma 4,  $(p-r)(\mathbf{S}^*)^{1/2} \widetilde{\mathbf{W}}^* (\mathbf{S}^*)^{1/2}$  is a projection matrix with rank  $r$ , and consequently,  $(p-r)(\mathbf{S}^*)^{1/2} (\mathbf{W}^* - \widetilde{\mathbf{W}}^*) (\mathbf{S}^*)^{1/2} = \mathbf{I}_m - (\mathbf{S}^*)^{1/2} \widetilde{\mathbf{W}}^* (\mathbf{S}^*)^{1/2}$  is a projection matrix with rank  $m-r$ .

It follows from Lemma 2 (iv) and (v) that  $(p-r)\mathbf{M}$ , where  $\mathbf{M}$  is defined in (41), is a projection matrix with rank  $(p-r)(m-r)$ . Finally, by Thm. 2.7 of Seber and Lee (2003), we conclude that  $(p-r)\boldsymbol{\xi}^T \mathbf{M} \boldsymbol{\xi}$  follows a chi-squared distribution  $\chi_{(p-r)(m-r)}^2$ , which finishes the proof.  $\square$

**Proof of (39).** To derive the above asymptotic result rigorously, we use the notations in Kato (1966) and invoke (3.2), (3.3), (3.4) and (3.6) in Chap. Two of Kato (1966) to bound the residual term in the asymptotic expansion. Under Assumption 1, there is a constant gap between  $\lambda_r(\mathbf{V}^* \mathbf{W}^* (\mathbf{V}^*))$  and  $\lambda_{r+1}(\mathbf{V}^* \mathbf{W}^* (\mathbf{V}^*))$ . Since  $\mathbf{W} \xrightarrow{p} \mathbf{W}^*$ , we deduce that with probability  $1-o(1)$ , the gap between  $\lambda_r(\mathbf{V}^* \mathbf{W} (\mathbf{V}^*))$  and  $\lambda_{r+1}(\mathbf{V}^* \mathbf{W} (\mathbf{V}^*))$  is bounded away from zero by a constant. Thus, choosing any circle with constant radius enclosing  $\lambda=0$ , we have with probability  $1-o(1)$ ,  $\rho$  in (3.4) and  $\max_{\zeta \in \Gamma} \|R(\zeta)\|$  and both bounded by a constant. Since  $\|\mathbf{T}^{(1)}\| = O_P(1)$ ,  $\|\mathbf{T}^{(2)}\| = O_P(2)$  and  $\mathbf{T}^{(n)}$  ( $n \geq 3$ ) are simply zero, (3.2) is satisfied with probability  $1-o(1)$  for  $|x| < r_n$  where  $r_n > 0$  is any vanishing sequence. This implies with probability  $1-o(1)$ ,  $r_0$  in (3.3) satisfies  $r_0 \geq r_n$ . We fix the ' $n$ ' in (3.6) by 2 (not to confuse with our sample size  $n$ ). Then, the upper bound in (3.6) is  $O(n^{-3/2}/r_0^3) = O(n^{-1})$  with the choice, say,  $r_n = n^{-1/10}$ . Therefore, we conclude that the residual term in the second-order expansion of  $\lambda(\omega_n)$  is  $o_P(\omega_n^2)$ .  $\square$

**Proof of Corollary 2.** Let us drop the subscript  $n$  as usual and denote  $\mathbb{E}\mathbf{V}$  by  $\mathbf{V}^*$ . First, we claim that there exists a constant  $c > 0$  such that  $\lambda_r(\mathbf{V}^*(\Sigma^*)^{-1}(\mathbf{V}^*)^T) > c$ . In fact, each column of  $\mathbf{V}^*$  lies in  $\mathcal{S}$ , so we can rewrite  $\mathbf{V}^*$  as  $\mathbf{U}^*(\mathbf{G}^*)^T$  where we recall  $\mathbf{G}^* = (\mathbf{V}^*)^T \mathbf{U}^* \in \mathbb{R}^{m \times r}$ . Under Assumption 1,  $\mathbf{G}^*$  has full (column) rank, so  $\mathbf{G}^* \mathbf{x} \neq \mathbf{0}$  for any  $\mathbf{x} \neq \mathbf{0}$ , which implies invertibility of  $(\mathbf{G}^*)^T(\Sigma^*)^{-1}\mathbf{G}^*$ . Since

$$\mathbf{V}^*(\Sigma^*)^{-1}(\mathbf{V}^*)^T = \mathbf{U}^* ((\mathbf{G}^*)^T(\Sigma^*)^{-1}\mathbf{G}^*) (\mathbf{U}^*)^T,$$

we deduce that top  $r$  eigenvalues of  $\mathbf{V}^*(\Sigma^*)^{-1}(\mathbf{V}^*)^T$  are nonzero, and thus bounded below by some constant  $c > 0$ . Using this claim, together with the fact that  $\mathbf{V}\mathbf{W}\mathbf{V}^T \xrightarrow{p} \mathbf{V}^*(\Sigma^*)^{-1}(\mathbf{V}^*)^T$  and the condition  $\tau_n = o(1)$ , we obtain  $\mathbb{P}(\hat{r}_\tau \geq r) \rightarrow 1$  as  $n \rightarrow \infty$ . On the other hand, Theorem 4 implies that for all  $j > r$ , the eigenvalue  $\lambda_j(\mathbf{V}\mathbf{W}\mathbf{V}^T)$  is  $O_P(n^{-1})$ , so the condition  $n\tau_n \rightarrow \infty$  ensures that for all  $j$ ,  $\lambda_j(\mathbf{V}\mathbf{W}\mathbf{V}^T) \leq \tau_n$  with probability  $1 - o(1)$ . This leads to  $\mathbb{P}(\hat{r}_\tau \leq r) \rightarrow 1$ , and thus completing the proof of the first part of the corollary. The second part follows similarly.  $\square$

## Proofs for Section 4.4

We omit the proof of Corollary 3, which is almost identical to that of Corollary 1.

**Proof of Theorem 5.** Let us break the proof into several steps.

(1) We prove a result parallel to (38). Fix any  $\mathbf{W}^* \in \mathcal{W}$ . We will prove that  $(\mathbf{G}^*)^T(\Sigma^*)^+\mathbf{G}^*$  is invertible, and that

$$[(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^*]^{-1} (\mathbf{G}^*)^T \mathbf{W}^* \Sigma^* \mathbf{W}^* \mathbf{G}^* [(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^*]^{-1} \succeq [(\mathbf{G}^*)^T(\Sigma^*)^+ \mathbf{G}^*]^{-1}. \quad (42)$$

This essentially replaces  $(\Sigma^*)^{-1}$  in (38) by  $(\Sigma^*)^+$ , and is equivalent to

$$(\mathbf{G}^*)^T \mathbf{W}^* \Sigma^* \mathbf{W}^* \mathbf{G}^* \succeq (\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^* [(\mathbf{G}^*)^T(\Sigma^*)^+ \mathbf{G}^*]^{-1} (\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^*.$$

It suffices, therefore, to prove invertibility of  $(\mathbf{G}^*)^T(\Sigma^*)^+ \mathbf{G}^*$  and

$$\Sigma^* \succeq \mathbf{G}^* [(\mathbf{G}^*)^T(\Sigma^*)^+ \mathbf{G}^*]^{-1} (\mathbf{G}^*)^T.$$

The right-hand side above can be simplified under Assumption 5. To do so, we write

$$\mathbf{G}^* = \mathbf{B} \begin{pmatrix} \mathbf{G}_1^* \\ \mathbf{0} \end{pmatrix}, \quad \text{where } \mathbf{B} := \begin{pmatrix} \mathbf{I}_{m_1} & \mathbf{0} \\ \Sigma_{21}^* (\Sigma_{11}^*)^{-1} & \mathbf{I}_{m_2} \end{pmatrix} \in \mathbb{R}^{m \times m}. \quad (43)$$

We also express  $\Sigma^*$  using  $\mathbf{B}$ :

$$\Sigma^* = \mathbf{B} \begin{pmatrix} \Sigma_{11}^* & \mathbf{0} \\ \mathbf{0} & \Sigma_{22}^* - \Sigma_{21}^* (\Sigma_{11}^*)^{-1} \Sigma_{12}^* \end{pmatrix} \mathbf{B}^T.$$

We make two observations: (a) the matrix  $\mathbf{B}$  is invertible; (b)  $\Sigma_{22}^* - \Sigma_{21}^* (\Sigma_{11}^*)^{-1} \Sigma_{12}^* \succeq \mathbf{0}$  due

to  $\Sigma^* \succeq \mathbf{0}$ . This yields

$$(\Sigma^*)^+ = (\mathbf{B}^T)^{-1} \begin{pmatrix} (\Sigma_{11}^*)^{-1} & \mathbf{0} \\ \mathbf{0} & (\Sigma_{22}^* - \Sigma_{21}^*(\Sigma_{11}^*)^{-1}\Sigma_{12}^*)^+ \end{pmatrix} \mathbf{B}^{-1}.$$

This equality can be verified against the definition of Moore-Penrose pseudoinverse. Note that under Assumption 5,  $\Sigma_{11}^*$  is invertible while  $\Sigma_{22}^* - \Sigma_{21}^*(\Sigma_{11}^*)^{-1}\Sigma_{12}^*$  may not. Now

$$\begin{aligned} (\mathbf{G}^*)^T(\Sigma^*)^+\mathbf{G}^* &= ((\mathbf{G}_1^*)^T \mathbf{0}) \begin{pmatrix} (\Sigma_{11}^*)^{-1} & \mathbf{0} \\ \mathbf{0} & (\Sigma_{22}^* - \Sigma_{21}^*(\Sigma_{11}^*)^{-1}\Sigma_{12}^*)^+ \end{pmatrix} \begin{pmatrix} \mathbf{G}_1^* \\ \mathbf{0} \end{pmatrix} \\ &= (\mathbf{G}_1^*)^T(\Sigma_{11}^*)^{-1}\mathbf{G}_1^*. \end{aligned}$$

By Assumption 1,  $\mathbf{G}^*$  has full column rank, and thus  $\mathbf{G}_1^*$  also has, due to (43). It follows that  $(\mathbf{G}_1^*)^T(\Sigma_{11}^*)^{-1}\mathbf{G}_1^*$  is positive definite (thus invertible). This proves the invertibility of  $(\mathbf{G}^*)^T(\Sigma^*)^+\mathbf{G}^*$ . Moreover, to prove (43), it is equivalent to show

$$\mathbf{B} \begin{pmatrix} \Sigma_{11}^* & \mathbf{0} \\ \mathbf{0} & \Sigma_{22}^* - \Sigma_{21}^*(\Sigma_{11}^*)^{-1}\Sigma_{12}^* \end{pmatrix} \mathbf{B}^T \succeq \mathbf{B} \begin{pmatrix} \mathbf{G}_1^* \\ \mathbf{0} \end{pmatrix} [(\mathbf{G}_1^*)^T(\Sigma_{11}^*)^{-1}\mathbf{G}_1^*]^{-1} ((\mathbf{G}_1^*)^T, \mathbf{0}) \mathbf{B}^T,$$

which is also equivalent to show

$$\Sigma_{11}^* \succeq \mathbf{G}_1^* [(\mathbf{G}_1^*)^T(\Sigma_{11}^*)^{-1}\mathbf{G}_1^*]^{-1} (\mathbf{G}_1^*)^T.$$

In the case of nonsingular weighting matrices, a similar inequality is known and is the key to the proof of (38). Our above derivation reduces the singular case to the nonsingular case (note  $\Sigma^*$  may be singular but  $\Sigma_{11}^*$  is nonsingular), so the desired inequality follows.

**(2)** We prove (30). First we observe that (35) still holds—see the sentence in parentheses before (34). This leads to the same expression (35) as in the proof of Theorem 3. Likewise, we still have (36) and (37), because  $(\mathbf{G}^*)^T \mathbf{W}^* \mathbf{G}^*$  is invertible, and  $\widehat{\mathbf{U}}$  (associated with  $\mathbf{W}^*$ ) is consistent up to rotation by Theorem 7.

Thus, for any  $\mathbf{W}^* \in \mathcal{W}$  with  $\mathbf{W} \xrightarrow{p} \mathbf{W}^*$ , the expression for  $\text{AE}(\Psi(\mathbf{W}))$  is given by (37). In particular, if  $\mathbf{W} \xrightarrow{p} (\Sigma^*)^+$ , then we have

$$\begin{aligned} \text{AE}(\Psi(\mathbf{W})) &= \frac{1}{n} \mathbf{A}^* \Sigma^* (\mathbf{A}^*)^T \\ &= \frac{1}{n} [(\mathbf{G}^*)^T(\Sigma^*)^+\mathbf{G}^*]^{-1} (\mathbf{G}^*)^T(\Sigma^*)^+\Sigma^*(\Sigma^*)^+\mathbf{G}^* [(\mathbf{G}^*)^T(\Sigma^*)^+\mathbf{G}^*]^{-1} \\ &= \frac{1}{n} [(\mathbf{G}^*)^T(\Sigma^*)^+\mathbf{G}^*]^{-1}, \end{aligned}$$

where we used the identity  $(\Sigma^*)^+\Sigma^*(\Sigma^*)^+ = (\Sigma^*)^+$  by definition of pseudoinverse. This proves the two equalities in (30). The inequality in (30) is proved in part (1).

**(3)** Now we prove the final claim of the theorem. Let the eigen-decomposition of  $\Sigma^*$  be  $\Sigma^* = \mathbf{U}_0 \Lambda_0 \mathbf{U}_0^T$ , where  $\mathbf{U}_0$  has orthonormal columns, and  $\Lambda_0$  is a diagonal matrix consisting of

all nonzero eigenvalues (the size of  $\Lambda_0$  is possibly smaller than that of  $\Sigma^*$ ). A basic property of the Moore-Penrose pseudoinverse is its representation via the eigen-decomposition:  $(\Sigma^*)^+ = \mathbf{U}_0 \Lambda_0^{-1} \mathbf{U}_0^T$  (Golub and Van Loan, 2013).

By the central limit theorem, we have  $\|\widehat{\Sigma} - \Sigma^*\|_2 = O_P(n^{-1/2})$ . Recall the eigen-decomposition of  $\widehat{\Sigma}$  and (13). Weyl's inequality implies  $|\bar{\lambda}_j - \lambda_j^*| = O_P(n^{-1/2})$ , where  $\lambda_j^*$  is the  $j$ th largest eigenvalue of  $\Sigma^*$ . If  $\lambda_j^* > 0$ , then  $\mathbb{P}(\bar{\lambda}_j > \delta_n) = 1 - o(1)$  due to  $\delta_n = o(1)$ , and thus

$$|\psi(\bar{\lambda}_j) - (\lambda_j^*)^{-1}| = O_P(n^{-1/2}). \quad (44)$$

If  $\lambda_j^* = 0$ , then  $\mathbb{P}(\bar{\lambda}_j \leq \delta_n) = 1 - o(1)$  due to  $\sqrt{n}\delta_n \rightarrow \infty$ , and therefore  $\psi(\bar{\lambda}_j) = 0$  with probability  $1 - o(1)$ . This proves consistency of  $\psi(\bar{\lambda}_1), \dots, \psi(\bar{\lambda}_m)$  in (13).

Next, let  $\mathbf{U}'_0$  be a submatrix of  $\mathbf{U}_0$  such that the columns of  $\mathbf{U}'_0$  correspond to the same eigenvalue of  $\Sigma^*$  (taking into account eigenvalue multiplicity). And denote by  $\bar{\mathbf{U}}'$  the counterpart of  $\bar{\mathbf{U}}$ . by Davis-Kahan's theorem (Davis and Kahan, 1970) and a sin  $\Theta$  formula (Stewart, 1990)

$$\|\bar{\mathbf{U}}'(\bar{\mathbf{U}}')^T - \mathbf{U}'_0(\mathbf{U}'_0)^T\|_2 = O(\|\widehat{\Sigma} - \Sigma^*\|_2) = O_P(n^{-1/2}). \quad (45)$$

We can write  $(\Sigma^*)^+$  as a sum of the form  $\lambda^{-1} \mathbf{U}'_0 (\mathbf{U}'_0)^T$ , where  $\lambda$  is any positive eigenvalue of  $\Sigma^*$  and  $\mathbf{U}'_0$  corresponds to  $\lambda$ . We can do the same for  $\mathbf{W}$ , and combine the bounds in (44) and (45) to deduce  $\mathbf{W} \xrightarrow{p} (\Sigma^*)^+$ . This completes the proof.  $\square$

## Proofs for Section 4.5

**Proof of Theorem 6.** We shall denote by  $\mathbf{M}_n^r$  is the best rank- $r$  approximation of  $n^{-1}\mathbf{X}^T\mathbf{X} - \sigma^2\mathbf{I}_p$  (i.e., in the eigen-decomposition, only keeping top  $r$  terms with largest absolute eigenvalues). We also define

$$\tilde{\mathbf{U}}_n^0 := \tilde{\mathbf{U}}_n^0(\kappa, \mathbf{W}'_n) = \text{eigen}_r(\kappa\mathbf{M}_n^r + \mathbf{V}_n \mathbf{W}'_n \mathbf{V}_n^T).$$

We will prove

$$\begin{aligned} \text{AE} \left( \left\| \widehat{\mathbf{U}}_n^{\text{GMM}} [\widehat{\mathbf{U}}_n^{\text{GMM}}]^T - \mathbf{U}^*(\mathbf{U}^*)^T \right\|_F^2 \right) &\leq \text{AE} \left( \left\| \tilde{\mathbf{U}}_n^0 [\tilde{\mathbf{U}}_n^0]^T - \mathbf{U}^*(\mathbf{U}^*)^T \right\|_F^2 \right) \\ &= \text{AE} \left( \left\| \tilde{\mathbf{U}}_n \tilde{\mathbf{U}}_n^T - \mathbf{U}^*(\mathbf{U}^*)^T \right\|_F^2 \right). \end{aligned} \quad (46)$$

Without loss of generality, we assume that  $\mathbf{M}$  is formed by the first  $p$  columns of  $\mathbf{V}$ . By an elementary property of pseudoinverse, we have  $\mathbf{M}^r = \mathbf{M}(\mathbf{M}^r)^+\mathbf{M}^T$ . This identity can be proved by, for example, using the eigen-decomposition of  $\mathbf{M}$  and rewriting  $\mathbf{M}^r$  and  $(\mathbf{M}^r)^+$  (Golub and Van Loan, 2013, Thm. 2.4.8, Chap. 5.5.2).

(1) First we prove the inequality in (46). Using the above identity of  $\mathbf{M}^r$ , we can absorb

the term  $\kappa\mathbf{M}^r$  into  $\mathbf{V}\mathbf{W}'\mathbf{V}^T$  as follows.

$$\kappa\mathbf{M}^r + \mathbf{V}\mathbf{W}'\mathbf{V}^T = \mathbf{V} \left[ \begin{pmatrix} \kappa(\mathbf{M}^r)^+ & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \mathbf{W}' \right] \mathbf{V}^T =: \mathbf{V}\mathbf{W}^{\text{new}}\mathbf{V}^T$$

We claim that  $(\mathbf{M}^r)^+ \xrightarrow{p} (\mathbb{E}\mathbf{M})^+ = \mathbf{U}^*\Lambda^*(\mathbf{U}^*)^T$ , where  $\Lambda^* \in \mathbb{R}^{r \times r}$  is a diagonal matrix with positive diagonal entries. This can be proved in a way similar to part (3) of the proof of Theorem 5. In fact, since  $\|\mathbf{M} - \mathbb{E}\mathbf{M}\|_2 = O_P(n^{-1/2})$ , by Weyl's inequality, the eigenvalues have convergence  $\lambda_j(\mathbf{M}) \xrightarrow{p} \lambda_j(\mathbb{E}\mathbf{M})$  for all  $j \in [p]$ , where  $\lambda_j(\cdot)$  denotes the  $j$ th largest eigenvalue of a matrix. For  $j \leq r$ , the assumption implies  $\lambda_j(\mathbb{E}\mathbf{M}) > 0$ , so we also have convergence  $[\lambda_j(\mathbf{M})]^{-1} \xrightarrow{p} [\lambda_j(\mathbb{E}\mathbf{M})]^{-1}$ . Moreover, by Davis-Kahan's theorem (Davis and Kahan, 1970), the corresponding eigenspace of  $\lambda_j(\mathbf{M})$  also converges in probability to that of  $\lambda_j(\mathbb{E}\mathbf{M})$ , where  $j \leq r$ . Therefore,  $(\mathbf{M}^r)^+ \xrightarrow{p} (\mathbb{E}\mathbf{M})^+$ , and due to  $\text{rank}(\mathbb{E}\mathbf{M}) = r$  and  $\text{span}(\mathbb{E}\mathbf{M}) = \text{span}(\mathbf{U}^*)$ , we can express  $(\mathbb{E}\mathbf{M})^+$  as  $\mathbf{U}^*\Lambda^*(\mathbf{U}^*)^T$ .

Using this claim, we deduce that  $\mathbf{W}^{\text{new}}$  converges in probability to a semidefinite matrix. Since  $\mathbf{W}' \xrightarrow{p} \mathbf{W}^* \in \mathcal{W}$ , the limit of  $\mathbf{W}^{\text{new}}$  must be also in  $\mathcal{W}$ . Thus, we can treat  $\mathbf{W}^{\text{new}}$  as the new weighting matrix  $\mathbf{W}'$  in Corollary 3, and the inequality in (46) follows from the conclusion of Corollary 3.

**(2)** Denote  $\mathbf{Y} = \kappa\mathbf{M}^r + \mathbf{V}\mathbf{W}'\mathbf{V}^T$  and  $\Delta = \mathbf{M} - \mathbf{M}^r$ . Note that  $\Delta = \sum_{j>r} \lambda_j(\mathbf{M})\mathbf{u}_j(\mathbf{M})[\mathbf{u}_j(\mathbf{M})]^T$ , where  $\mathbf{u}_j(\mathbf{M})$  is the eigenvector of  $\mathbf{M}$  corresponding to  $\lambda_j(\mathbf{M})$ . By the assumption  $\text{rank}(\mathbb{E}\mathbf{M}) = r$  and the analysis in part (1), we have  $|\lambda_j(\mathbf{M})| = O_P(n^{-1/2})$  and  $(\mathbf{U}^*)^T\mathbf{u}_j(\mathbf{M}) \xrightarrow{p} \mathbf{0}$  for  $j > r$ . Applying the consistency result (Theorem 1) to  $\tilde{\mathbf{U}}^0$ , we also obtain  $(\tilde{\mathbf{U}}^0)^T\mathbf{u}_j(\mathbf{M}) \xrightarrow{p} \mathbf{0}$ .

Now, we view  $\Delta$  as a perturbation matrix added to  $\mathbf{Y}$ , and view  $\tilde{\mathbf{U}}$  as a resulting perturbed version of  $\tilde{\mathbf{U}}^0$ . We use (the original version of) Davis-Kahan's theorem (Davis and Kahan, 1970) to obtain

$$\|\tilde{\mathbf{U}}(\tilde{\mathbf{U}})^T - \tilde{\mathbf{U}}^0(\tilde{\mathbf{U}}^0)^T\|_F = O_P\left(\|\Delta\tilde{\mathbf{U}}^0\|_F\right).$$

This form of Davis-Kahan's theorem has appeared in recent works (Yu et al., 2014; Zhong and Boumal, 2018). Using  $|\lambda_j(\mathbf{M})| = O_P(n^{-1/2})$  and  $(\tilde{\mathbf{U}}^0)^T\mathbf{u}_j(\mathbf{M}) \xrightarrow{p} \mathbf{0}$  we established, we obtain

$$\|\tilde{\mathbf{U}}(\tilde{\mathbf{U}})^T - \tilde{\mathbf{U}}^0(\tilde{\mathbf{U}}^0)^T\|_F = o_P(n^{-1/2}). \quad (47)$$

Note that Theorem 5 implies that  $\sqrt{n}\|\tilde{\mathbf{U}}_n^0[\tilde{\mathbf{U}}_n^0]^T - \mathbf{U}^*(\mathbf{U}^*)^T\|_F$  converges in distribution to a non-vanishing limit. Thus, from (47), we deduce that  $\sqrt{n}\|\tilde{\mathbf{U}}_n\tilde{\mathbf{U}}_n^T - \mathbf{U}^*(\mathbf{U}^*)^T\|_F$  must also converge in distribution to the same limit, and therefore, the equality in (46) is true.  $\square$

## References

- AHN, S. C. and HORENSTEIN, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81** 1203–1227.

- ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research* **15** 2773–2832.
- ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics* **34** 122–148.
- BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171.
- BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221.
- BAI, J. and NG, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics* **176** 18–29.
- BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics* **45** 77–120.
- BEN-ISRAEL, A. and GREVILLE, T. N. (2003). *Generalized inverses: theory and applications*, vol. 15. Springer Science & Business Media.
- BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* **80** 580–598.
- BREUSCH, T., QIAN, H., SCHMIDT, P. and WYHOWSKI, D. (1999). Redundancy of moment conditions. *Journal of econometrics* **91** 89–111.
- CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics* **9** 717.
- CATTELL, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research* **1** 245–276.
- COHEN, N. and SHASHUA, A. (2016). Convolutional rectifier networks as generalized tensor decompositions. In *International Conference on Machine Learning*.
- CONNOR, G. and KORAJCZYK, R. A. (1993). A test for the number of factors in an approximate factor model. *the Journal of Finance* **48** 1263–1291.
- COOK, R. D. (1998). Principal hessian directions revisited. *Journal of the American Statistical Association* **93** 84–94.
- DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis* **7** 1–46.
- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* 1–38.
- DHEERU, D. and KARRA TANISKIDOU, E. (2017). UCI machine learning repository.  
URL <http://archive.ics.uci.edu/ml>
- DONALD, S. G., IMBENS, G. W. and NEWHEY, W. K. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics* **117** 55–93.
- DURRETT, R. (2010). *Probability: theory and examples*. Cambridge university press.

- EDELMAN, A., ARIAS, T. A. and SMITH, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications* **20** 303–353.
- FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 603–680.
- FAN, J., LIAO, Y. and WANG, W. (2016). Projected principal component analysis in factor models. *Annals of statistics* **44** 219.
- FAN, J., WANG, D., WANG, K. and ZHU, Z. (2017). Distributed estimation of principal eigenspaces. *arXiv preprint arXiv:1702.06488*.
- FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics* **82** 540–554.
- FORNI, M., HALLIN, M., LIPPI, M. and ZAFFARONI, P. (2017). Dynamic factor models with infinite-dimensional factor space: Asymptotic analysis. *Journal of Econometrics* **199** 74–92.
- GOLUB, G. H. and VAN LOAN, C. F. (2013). *Matrix Computations*, vol. 3. JHU Press.
- HALL, A. R. (2005). *Generalized method of moments*. Oxford University Press.
- HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* 1029–1054.
- HORN, R. A. and JOHNSON, C. R. (1991). Topics in matrix analysis, 1991. *Cambridge University Press, Cambridge* **37** 39.
- HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* **24** 417.
- HURLEY, C. (2012). gclus: Clustering graphics.
- JIN, C., ZHANG, Y., BALAKRISHNAN, S., WAINWRIGHT, M. J. and JORDAN, M. I. (2016). Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems*.
- KATO, T. (1966). *Perturbation theory for linear operators*. New York.
- KOENKER, R. and MACHADO, J. A. (1999). Gmm inference when the number of moment conditions is large. *Journal of Econometrics* **93** 327–344.
- LAM, C., YAO, Q. ET AL. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics* **40** 694–726.
- LETTAU, M. and PELGER, M. (2017). Estimating latent asset-pricing factors. Tech. rep.
- LI, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R*. CRC Press.
- LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86** 316–327.
- LI, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association* **87** 1025–1039.

- ONATSKI, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics* **92** 1004–1016.
- PEARSON, K. (1901). On lines and planes of closest fit to systems of point in space. *Philosophical Magazine* **2** 559–572.
- SEBER, G. A. and LEE, A. J. (2003). *Linear regression analysis*, vol. 936. John Wiley & Sons.
- SEDGHI, H., JANZAMIN, M. and ANANDKUMAR, A. (2016). Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*.
- SHAO, J. (2003). *Mathematical Statistics*. Springer-Verlag New York.
- STEWART, G. W. (1990). Matrix perturbation theory.
- STEWART, G. W. (1998). Perturbation theory for the singular value decomposition. Tech. rep.
- SUN, Y., IOANNIDIS, S. and MONTANARI, A. (2014). Learning mixtures of linear classifiers. In *ICML*.
- VAN DER VAART, A. W. (1998). *Asymptotic statistics*, vol. 3. Cambridge university press.
- WANG, B., LUO, X., LI, Z., ZHU, W., SHI, Z. and OSHER, S. J. (2018). Deep learning with data dependent implicit activation function. *arXiv preprint arXiv:1802.00168* .
- WU, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics* 95–103.
- XU, L. and JORDAN, M. I. (1996). On convergence properties of the em algorithm for gaussian mixtures. *Neural computation* **8** 129–151.
- YI, X., CARAMANIS, C. and SANGHAVI, S. (2016). Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749* .
- YU, Y., WANG, T. and SAMWORTH, R. J. (2014). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323.
- ZHONG, Y. and BOUMAL, N. (2018). Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization* **28** 989–1016.
- ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics* **15** 265–286.