Bernoulli **28**(2), 2022, 997–1020
https://doi.org/10.3150/21-BEJ1374

Inference in latent factor regression with clusterable features

XIN BING[1,*], FLORENTINA BUNEA[1,†] and MARTEN WEGKAMP[1,2]

[1]*Department of Statistics and Data Science, Cornell University, Ithaca, New York, USA.*
E-mail: [*]*xb43@cornell.edu;* [†]*fb238@cornell.edu*
[2]*Department of Mathematics, Cornell University, Ithaca, New York, USA. E-mail: mhw73@cornell.edu*

Regression models, in which the observed features $X \in \mathbb{R}^p$ and the response $Y \in \mathbb{R}$ depend, jointly, on a lower dimensional, unobserved, latent vector $Z \in \mathbb{R}^K$, with $K \ll p$, are popular in a large array of applications, and mainly used for predicting a response from correlated features. In contrast, methodology and theory for inference on the regression coefficient $\beta \in \mathbb{R}^K$ relating Y to Z are scarce, since typically the un-observable factor Z is hard to interpret. Furthermore, the determination of the asymptotic variance of an estimator of β is a long-standing problem, with solutions known only in a few particular cases.

To address some of these outstanding questions, we develop inferential tools for β in a class of factor regression models in which the observed features are signed mixtures of the latent factors. The model specifications are both practically desirable, in a large array of applications, render interpretability to the components of Z, and are sufficient for parameter identifiability.

Without assuming that the number of latent factors K or the structure of the mixture is known in advance, we construct computationally efficient estimators of β, along with estimators of other important model parameters. We benchmark the rate of convergence of β by first establishing its ℓ_2-norm minimax lower bound, and show that our proposed estimator $\widehat{\beta}$ is minimax-rate adaptive. Our main contribution is the provision of a unified analysis of the component-wise Gaussian asymptotic distribution of $\widehat{\beta}$ and, especially, the derivation of a closed form expression of its asymptotic variance, together with consistent variance estimators. The resulting inferential tools can be used when both K and p are independent of the sample size n, and also when both, or either, p and K vary with n, while allowing for $p > n$. This complements the only asymptotic normality results obtained for a particular case of the model under consideration, in the regime $K = O(1)$ and $p \to \infty$, but without a variance estimate.

As an application, we provide, within our model specifications, a statistical platform for inference in regression on latent cluster centers, thereby increasing the scope of our theoretical results.

We benchmark the newly developed methodology on a recently collected data set for the study of the effectiveness of a new SIV vaccine. Our analysis enables the determination of the top latent antibody-centric mechanisms associated with the vaccine response.

Keywords: High dimensional regression; latent factor model; identification; uniform inference; minimax estimation; pure variables; post clustering inference/regression; adaptive estimation

1. Introduction

Latent factor models have been used successfully for several decades for modeling data with embedded low dimensional structures. In particular, they provide a natural framework for regression problems in which the covariate vector $X \in \mathbb{R}^p$ and the response $Y \in \mathbb{R}$ are *jointly* low dimensional. In a latent factor regression model, this is formalized by assuming that there exists a random vector $Z \in \mathbb{R}^K$, for some *unknown* $K < p$, that is connected to the observed pair $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ via the model

$$Y = Z^\top \beta + \varepsilon \tag{1}$$

$$X = AZ + W. \tag{2}$$

The dimension K, matrix $A \in \mathbb{R}^{p\times K}$ and vector $\beta \in \mathbb{R}^K$ are unknown. The random vectors Z and W and random variable ε are independent, with zero means, $\mathbb{E}[Z] = \mathbf{0}$, $\mathbb{E}[W] = \mathbf{0}$ and $\mathbb{E}[\varepsilon] = 0$, and covariance matrices $\Sigma_Z := \mathrm{Cov}(Z)$ and $\Gamma := \mathrm{Cov}(W)$, and variance $\sigma^2 := \mathbb{E}[\varepsilon^2]$, respectively.

Factor regression models, and their many variants [8–10,16,17,21,25,28–30,32,35,42–44] have been introduced to motivate and analyze prediction schemes for $Y \in \mathbb{R}$ from $X \in \mathbb{R}^p$, when p is very large and the components of X are highly correlated. Parameter identifiability is not required for prediction purposes, as unique predictors can still be constructed when the assignment matrix A and the covariance matrix Σ_Z of Z are identifiable only up to orthogonal transformations.

Substantially less work has been devoted to inference in factor models, the problem treated in this work. Classical factor analysis for an observable vector $U \in \mathbb{R}^d$ postulates the existence of factors $Z \in \mathbb{R}^K$ such that $U = BZ + E$, for some $d \times K$ factor loading matrix. Factor regression models are an instance of factor models, where one emphasizes the different roles, response and covariates, respectively, of the observable variables $U = (X, Y)$, and B consists in the matrix A augmented by the vector β.

This paper proposes and analyses computationally efficient estimators for inference on the regression coefficient β in identifiable and interpretable factor regression models, an under-explored problem.

1.1. A framework for regression on interpretable latent factors

We begin by summarizing the model parameters, the nature of the data, as well as the relation between parameter dimensions and sample size. Throughout this work we assume that we have access to an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$, and that (X, Y) have mean zero and satisfy (1) and (2).

We allow for $p > n$, while $K < p$. In this work, we consider the case of non-sparse β, and $K < \sqrt{n}$, but allow K to grow with the sample size n. The complementary cases of $K > \sqrt{n}$ and β sparse will be studied in a follow-up work.

Our central interest is on valid inference for β, which first requires establishing its identifiability. Restrictions on generic factor models of the type $U = BZ + E$ under which the model parameters are identifiable can be traced back to [37]. A very detailed exposition of possible identifiability restrictions was first collected in the seminal work of [4]. They have been revisited in several works, for instance, [6,18,38,50]. Of those restrictions on B, some are of purely mathematical convenience [4], Sections 5 and 6, whereas, as considered in this work, others are practically interpretable, [3,4].

We focus on a class of identifiable factor regression models in which the observed covariates X are signed mixtures of the latent vector's components Z_k, $1 \le k \le K$, with unknown K. A latent Z_k can be interpreted as the representative of one of the mixtures. Inference on β is thus inference for the mixture representatives. The nature of a representative Z_k is the nature of those few observed X_j's that are connected only to that Z_k, justifying their name, pure variables, indexed by I_k, with their totality indexed by $I := \cup_{k=1}^K I_k \subseteq \{1, \ldots, p\}$. In Section 2.1 we formalize this model class, and show that it is identifiable.

Versions of this factor model class are routinely used in educational and psychological testing, where the latent variables are viewed as aptitudes or psychological states [4,18,45]. The X-variables are test results, with some tests specifically designed to measure only one single aptitude Z_k, for each aptitude, whereas others test mixtures of aptitudes. By experimental design, I and K *are known* in this classical literature.

Another important application of this factor regression model, in which both I and K *are unknown*, is to the analysis of biological data sets with hidden signatures. The data sets that we discuss in Section 6 have, by design, what we termed pure variables. Furthermore, because of the inherent biological

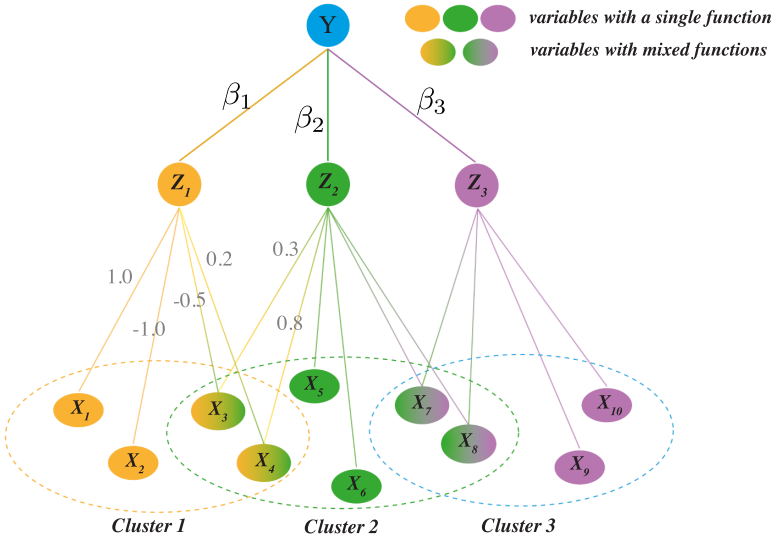

Figure 1. An illustrative example of Essential Regression.

redundancy built into the multi-omic screens that generate the components of X, one expects at least two of the X-variables effectively measuring the same biological signature, and only that one. For instance, there can be two or more paralogous genes with one very specific function, or two or more different subsets of immune cells carrying out the same specific niche immunological function. Such signatures are known to exist, but cannot be measured directly, and correspond to the components of the latent vector Z. Whereas some of these functions are known, it is one of the purposes of the analysis to discover new ones, as well as new X-variables solely associated with them. Thus, neither I nor K can be treated as known, nor can K be treated as fixed, since the number of functions K can grow as p grows, which can in turn grow with n.

Our first contribution is to propose this flexible and, in many applications, more realistic, framework for estimation and inference on β in factor models with pure variables, in which the index set I and the number of factors K are not known and have to be estimated from the data. All the results of this paper are derived in this context, which is the first point of departure from previous results for inference in factor models derived for models with I and K known, for instance in [3,4,6,50].

To emphasize the specific usage of factor models for regression and inference on mixture representatives, under the model specifications formally given in Section 2.1, we refer to it as *Essential Regression*.

Figure 1 below gives an instance of Essential Regression. A response Y depends on three latent factors (Z_1, Z_2, Z_3), which in turn are connected to $(X_1, \ldots, X_{10})$. The measured variables X_1 and X_2 have only (100%) function Z_1. The $\pm$ 1 edge weights indicate that this function activates X_1 and inhibits X_2. Variable X_3 has mixed functions, 50% is devoted to function Z_1, and the sign indicates that Z_1 is an inhibitor, while 30% is devoted to function Z_2, an activator. The fact that the weights, in absolute value, do not sum up to 1 increases the model flexibility, by allowing free association between X_3 and other functions that are not explained by this model.

A similar, data-driven, figure is presented in Section 6, in which we show that the Essential Regression model fits the data collected in a new SIV-study (SIV is the non-human primate equivalent of HIV), and offers insights into immunological signatures driving the vaccine response. This example

illustrates a scientifically-desirable way of modeling a response Y directly at the function (Z) level, when the observed X-variables have either single or mixed functions.

1.2. Our contributions

Minimax adaptive estimation of β. In Section 3, we construct a computationally tractable estimator of β. As part of the estimation procedure, the following unknown quantities are also estimated from the data, under the Essential Regression model: A, I, K, $\Sigma_Z := \mathrm{Cov}(Z)$, $\sigma^2 := \mathbb{E}[\varepsilon^2]$ and $\Gamma := \mathrm{Cov}(W)$. To benchmark the quality of our estimator we derive first, in Theorem 2 of Section 4.1 the minimax optimal rate of estimating β in ℓ_2-norm in an Essential Regression model. We show in Theorem 3 of Section 4.3 that the proposed estimator $\widehat{\beta}$ is minimax rate adaptive, up to logarithmic factors in n and p. Our result uses the fact that only the estimation of the $|I| \times K$ sub-matrix $A_{I\bullet}$, instead of the entire $p \times K$ matrix A, is involved in the construction of $\widehat{\beta}$. In Section 4.4 we introduce and discuss various competitors, including an estimator of β that utilizes estimators of the full matrix A. These estimators are natural consequences of rewriting the identity for $\beta = \Sigma_Z^{-1}\mathrm{Cov}(Z, Y)$, see (4), (5), (6), (24) and (26). We give insight into why these estimators are less efficient than the estimator proposed and studied, and confirm this in our simulation study in Appendix I.

To estimate I and K we use the method proposed in [13], as it guarantees that we can consistently estimate K, without imposing any restrictions on our target for inference, β. Furthermore, this method also guarantees that $I \subseteq \widehat{I} \subseteq I \cup J_1$, where J_1 is an index set of what we term quasi-pure variables, defined formally in Section 4.2. As the name suggests, a quasi-pure variable is a measured X-variable that is very strongly associated with only one Z_k, while having very small, but non-zero, association with other latent factors. A signal strength assumption on the entries of A would render $J_1 = \varnothing$, which would simplify the analysis of $\widehat{\beta}$ considerably.

To maintain a flexible modeling framework, the proofs of all our results, rate optimality and asymptotic distribution, allow for the presence of quasi-pure variables, $J_1 \neq \varnothing$, while controlling their relative number via Assumptions 3 and 3′. The price to pay for considering a more realistic scenario is an increase in the technical difficulty of the proofs of Theorems 3 and 4 and Proposition 5, for instance in Lemmas 7, 10, 17, 27, 28, 29.

Inference for β: Component-wise limiting distribution, asymptotic variance and its estimates. Within various classes of identifiable factor models, and in the classical set-up K and p fixed, [3,4] proposed MLE-based estimators of the rows of identifiable loading factors B, in a generic factor model $U = BZ + E$. They pointed out that the asymptotic covariance matrix of the Gaussian limit of their estimators has a very involved expression, and left its derivation open.

In the regime K fixed and $p \to \infty$, [6] offered a solution to this problem, two decades later. They derived the asymptotic distribution, including the expression of the limiting covariance, of MLE-inspired estimators of the rows of B, under various identifiability restrictions on B, including a version of the conditions given in Section 4.5 below, corresponding to I and K known. Their proof uses a linearization argument, and requires $p \to \infty$ to establish that the corresponding remainder term converges in probability to zero. The practical implementation of the estimator involves an EM-type algorithm that is very sensitive to initialization, and becomes problematic in high dimensions. The estimation of the limiting covariance is not considered in their work.

Computationally feasible estimators of the rows of B, and in particular of the entries of β, with closed form, estimable, asymptotic variances continue to be lacking in the classical regime K, p fixed, and also when both dimensions are allowed to grow. Furthermore, no results of this type have been established when K and I are unknown.

	K and I both known			K and I both unknown		
	K, p fixed	K fixed, $p \to \infty$	K, $p \to \infty$	K, p fixed	K fixed, $p \to \infty$	K, $p \to \infty$
Existing results	[4], MLE estimator, no closed form of the asymptotic variance.	[6], MLE-inspired estimator, computationally involved; closed form asymptotic variance Q_k when $\lambda_K \asymp p$.	NA	NA	NA	NA
Theorem 4, Section 4.5	Computationally tractable $\widehat{\beta}_k$ and asymptotic variance V_k					
	✓	V_k reduces to Q_k when $\lambda_K \to \infty$ and $\lambda_K \gtrsim (p/\sqrt{n}) \log(p \vee n)$.	✓	✓	✓	✓

Table 1. Asymptotically normal estimators of β_k in a class of factor models: existing and new results (Theorem 4)

As our main contribution, we address these open questions in this work, via a unifying analysis, by studying the component-wise distribution of estimators of β_k, for $1 \leq k \leq K$

Theorem 4 of Section 4.5 shows that the computationally tractable estimator proposed in Section 3 is asymptotically normal, with consistently estimable variance, under all scenarios of interest. A consistent estimator of this variance is given in Section 4.5 and its consistency is proved in Proposition 5. Table 1 below offers a snap shot of our asymptotic normality results, relative to existing results.

The quantity $\lambda_K := \lambda_K(A\Sigma_Z A^\top)$ quantifies the size of the signal in $X = AZ + W$. Theoretical analyses under the regime $p > n$ are performed under a conservative signal strength assumption, $\lambda_K \asymp p$, in the existing literature on factor models. See, for instance, [5,9,23,25–27]. This includes results pertaining to inference on β of [6], which are most closely related to our work.

In Section 4.5 we prove that our proposed estimators of β attain a Gaussian limit under a considerably relaxed condition, $\lambda_K \gtrsim p/\sqrt{n}$ (up to logarithmic $\log(p \vee n)$ factors), within a framework in which K can grow as fast as $O(\sqrt{n}/\log(p \vee n))$. A technical discussion of this condition is provided in Section 4.2, and in Remark 3 of Section 4.5.

Table 1 offers a complete picture, to the best of our knowledge, of the existing literature on inference for β, under the modelling framework considered in this work. For completeness, we summarize in Remark 2 of Section 4.3, other approaches proposed in the literature for the selection of K, in other identifiable factor models. We summarize them, and state sufficient conditions for their consistency in Table 2.

For clarity of presentation, we give below the expressions of the limiting variances in the particular case when $\mathrm{Cov}(W) = \tau^2 \mathbf{I_p}$. By letting $\sigma^2 = \mathbb{E}[\varepsilon^2]$, $\mathrm{Cov}(W) = \tau^2 \mathbf{I_p}$, $\Theta := A\Sigma_Z$ and $\Theta^+ = (\Theta^\top \Theta)^{-1}\Theta^\top$, the asymptotic variance derived in [6], in the regime $K = O(1)$ and $\lambda_K \asymp p \to \infty$, is

$$Q_k = \left(\sigma^2 + \tau^2 \|\beta\|_2^2\right) \left[\Sigma_Z^{-1}\right]_{kk}.$$

The assumption that $\lambda_K \asymp p$ is made in [6] indirectly, as a consequence of their assumption (A) (Σ_Z is positive definite and K is fixed) and of their Assumption (C) ($\|A_{i\bullet}\|_2 \leq C$, $c \leq \Gamma_{ii} \leq C$ and $p^{-1}A^\top \Gamma^{-1} A$ converges to some positive definite matrix as $p \to \infty$), see page 438 of [6].

The asymptotic variance of our proposed estimator, valid for all the regimes represented in Table 1 above, has the formula derived in Theorem 4 of Section 4.5,

$$V_k = \left(\sigma^2 + \frac{\tau^2}{m}\|\beta\|_2^2\right)\left[\left[\Sigma_Z^{-1}\right]_{kk} + \tau^2 \mathbf{e}_k^\top \left(\Theta^\top \Theta\right)^{-1} \mathbf{e}_k\right] + \frac{\tau^4}{m(m-1)} \sum_{a=1}^{K} \beta_a^2 \sum_{i \in I_a} \left[\mathbf{e}_k^\top \Theta^+ \mathbf{e}_i\right]^2,$$

where $\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$ is the canonical basis of $\mathbb{R}^K$.

To contrast the two asymptotic variance expressions, we consider the common regime $K = O(1)$ and $p \to \infty$, in which case we note that the signal strength requirement under which our Theorem 4 is established reduces to $\lambda_K \gtrsim (p/\sqrt{n}) \log(p \vee n)$. Theorem 4 shows that, in this case, if we further assume that $\lambda_K \to \infty$, the asymptotic variance of our estimator reduces to

$$V_k = \left(\sigma^2 + \frac{\tau^2}{m}\|\beta\|_2^2\right)\Omega_{kk},$$

and thus $\lim_{n\to\infty} Q_k / V_k \geq 1$. The two asymptotic variances coincide when $m = 1$, which is the minimum identifiability requirement for a factor model with pure variables in which the pure variable set I is known. Hence we recover, in this regime, the asymptotic variance derived in [6], while at the same time relaxing their signal strength conditions required for this derivation.

As noted above, and as we prove formally in Theorem 4, the expression of the asymptotic variance V_k is valid in all the regimes presented in this table. In particular, in the classical regime in which K and p do not vary with n, its derivation requires only $\lambda_K \gtrsim n^{-1/2}$.

In order to provide a unifying analysis, valid for both fixed and growing dimensions, we use the classical Lyapunov CLT for triangular arrays. The verification of the third moment condition of this theorem requires the lengthy, technical, derivations in Lemmas 16 and 19. Finally, although the expression of the asymptotic variance V_k is involved, it can be estimated consistently for each $1 \leq k \leq K$, by a computationally efficient estimator. This result is given in Proposition 5 of Section 4.5 and its proof, which requires considerable attention, is presented in Appendix F, followed by a list of the many technical lemmas used in this proof, Lemmas 21–29.

An application to regression on latent cluster centers. The identifiable factor model $X = AZ + W$ satisfying Assumption 1 in Section 2.1 below can be used to define, uniquely, overlapping clusters of the coordinates of X. The clusters are centered around the components of the latent vector Z, and X-variables in cluster k have indices in the set $G_k := \{j \in [p] : |A_{jk}| > 0\}$, for $1 \leq k \leq K$. A procedure for estimating consistently K and the corresponding clusters has been developed recently in [13]. With this interpretation, the Essential Regression framework can be employed for inference on the latent cluster centers. We show in Section 5 that although it may be tempting to replace the components of Z by weighted averages of variables within a cluster, and subsequently regress Y onto them, this procedure would not estimate β in (1). However, we further show that this can be immediately corrected by regressing on the best linear predictor of Z from these weighted averages. With this correction, we obtain exactly the estimator of β constructed in Section 3, and the inferential tools developed in Section 4.5 can be used for inference in regression on unobserved, latent, cluster centers. In the context of the applications to biological data sets mentioned in Section 1.1, this will be inference at the biological signature level, as illustrated in Section 6.

1.3. Organization of the paper

The rest of the paper is organized as follows.

Section 2 gives a set of modeling assumptions under which the model given by (1) and (2) is identifiable. Section 2.1 introduces and discusses these assumptions, including parameter interpretability. Section 2.2 shows that our central parameter, β, along with other important parameters, is identifiable.

Section 3 introduces our proposed estimator $\widehat{\beta}$ of β. Section 4.4 discusses other natural estimators, and explains why they should be expected to have inferior theoretical and practical perfomance relative to $\widehat{\beta}$. The numerical performance of these alternate estimators is presented in Appendix I.

The performance of estimators of β in factor regression models satisfying Assumption 1 is benchmarked in Section 4.1. Theorem 2 provides the minimax lower bound for estimating β in this class of models, with respect to the ℓ_2 loss.

Section 4.3 shows that the estimator $\widehat{\beta}$ proposed in Section 3 is ℓ_2-norm consistent, and minimax-rate adaptive, under assumptions collected in Section 4.2.

Section 4.5 is devoted to the component-wise asymptotic normality of $\widehat{\beta}$ and to the estimation of the asymptotic variance, as well as to a comparison with existing literature.

Section 5 presents an application of the framework, methodology and theory developed in previous sections to regression on latent cluster centers, when the clusters are allowed to overlap.

Section 6 shows how our methodology can be used to make inference on unobserved, latent, immunological modules, using a data set collected during a study on the effectiveness of a new SIV-type vaccine.

All proofs are deferred to the supplement [15]. Appendix B gives the proofs of Proposition 1 and Theorem 2, on identification and minimax lower bounds, respectively. Appendix C provides necessary preliminary results, and could be skipped at first reading. The proof of Theorem 3 concerning the convergence rates of $\widehat{\beta}$ is given in Appendix D. The proof of Theorem 4 on the asymptotic normality of $\widehat{\beta}$ is given in Appendix E, while Proposition 5 on consistent estimation of the asymptotic variance V_k is proved in Appendix F.

1.4. Notation

For any positive integer q, we let $[q] = \{1, 2, \ldots, q\}$. For two numbers a and b, we write $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. For a set S, we use $|S|$ to denote its cardinality. We use $\mathcal{H}_d$ to denote the set of all $d \times d$ signed permutation matrices and $\mathcal{S}^{d-1}$ to represent the space of the unit vectors in $\mathbb{R}^d$. We denote by $\mathbf{I}_d$ the $d \times d$ identity matrix, by $\mathbf{1}_d$ the d-dimensional vector with entries equal to 1 and by $\{\mathbf{e}_j\}_{1 \le j \le d}$ the canonical basis in $\mathbb{R}^d$. For a generic vector v, we let $\|v\|_q = \left(\sum_i |v_i|^q\right)^{1/q}$ denote its ℓ_q norm for $1 \le q < \infty$. We also write $\|v\|_\infty = \max_i |v_i|$ and $\|v\|_0 = |\operatorname{supp}(v)|$. Let Q be any matrix. We use $\|Q\|_{\mathrm{op}} = \sup_{v \in \mathcal{S}^{d-1}} \|Qv\|$ and $\|Q\|_\infty = \max_{i,j} |Q_{ij}|$ for its operator norm and element-wise maximum norm, respectively. For a symmetric matrix $Q \in \mathbb{R}^{d \times d}$, we denote by $\lambda_k(Q)$ its kth largest eigenvalue for $k \in [d]$. For a positive semi-definite symmetric matrix, we will frequently use the fact that $\lambda_1(Q) = \|Q\|_{\mathrm{op}}$. For an arbitrary real valued matrix M, we let $\sigma_K(M)$ denote its kth singular value (in decreasing order).

For any two sequences a_n and b_n, $a_n \lesssim b_n$ stands for there exists constant $C > 0$ such that $a_n \le C b_n$. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We also use $a_n = o(b_n)$ to denote $a_n / b_n \to 0$ as $n \to \infty$.

2. Modeling assumptions and identifiability

2.1. Modeling assumptions

We begin by formalizing and explaining the set of model identifiability assumptions that will be used in this work.

Assumption 1.

(A0) $\|A_{j\bullet}\|_1 \le 1$ for all $j \in [p]$.
(A1) For every $k \in [K]$, there exists at least two $j \ne \ell \in [p]$, such that $|A_{j\bullet}| = |A_{\ell\bullet}| = \mathbf{e}_k$.

(A2) $\Sigma_Z := \mathrm{Cov}(Z)$ is positive definite. There exists a constant $\nu > 0$ such that

$$\min_{1\le a<b\le K}\Big([\Sigma_Z]_{aa} \wedge [\Sigma_Z]_{bb} - |[\Sigma_Z]_{ab}|\Big) > \nu.$$

In *(A1)*, the absolute value is taken entry-wise and we use $\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$ to denote the canonical basis in $\mathbb{R}^K$. For future reference, we denote the index set corresponding to pure variables as

$$I = \bigcup_{k=1}^{K} I_k, \qquad I_k = \{i \in [p] : |A_{i\bullet}| = \mathbf{e}_k\}. \tag{3}$$

Its complement set is called the non-pure variable set $J := [p] \setminus I$.

Assumption 1 guarantees that A and Σ_Z are identifiable, up to signed permutations [13], Theorem 2. We refer to the second assumption *(A1)* as *the pure variable assumption*. It states that every Z_k, $1 \le k \le K$, must have at least two components of X, the *pure variables*, solely associated with it, up to additive noise with possibly different variance levels. An in-depth comparison with the rich literature on factor models of type (2) and a detailed explanation of assumptions *(A0)–(A2)* can be found in [13], and thus we only offer a brief set of comments here.

If $A\Sigma_Z A^T$ is identifiable, we show in Corollary 1 in Appendix A that A and Σ_Z are identifiable up to signed permutations under Assumption 1, but when *(A1)* is relaxed to

(A1') For each $k \in [K]$, there exists at least one index $i \in [p]$ such that $A_{i\cdot} = \mathbf{e}_k$.

However, the existing conditions under which $A\Sigma_Z A^T$ can be identified from the decomposition $\Sigma = A\Sigma_Z A^T + \Gamma$ can be incompatible with factor models with pure variables, for instance the incoherence condition in [24], or can be very stringent growth conditions on the eigenvalues of $A\Sigma_Z A^T$. For an instance of the latter we refer to [7,27] and also to Table 2. These difficulties can be bypassed under *(A1)* of Assumption 1, using the approach taken in [13], which does not rely on separating out $A\Sigma_Z A^T$ from Σ in the first step. In Assumption 1, *(A0)*, we set the scale to 1 to aid the interpretation of matrix A as a cluster membership matrix, and thus view the model $X = AZ + W$ as a latent clustering model, following [13]. The equality between the weights of two pure variables, $|A_{j\cdot}| = |A_{\ell\cdot}| = \mathbf{e}_k$ has been relaxed in a recent work, [14], but under a slightly different scaling condition than *(A0)*, and we do not pursue that approach here.

Furthermore, we mention, for completeness, that a more rigid form of assumption *(A1')*, specifically

(A1") For each $k \in [K]$, there exists a **known** *index $i \in [p]$ such that $A_{i\cdot} = \mathbf{e}_k$,*

has had a long history, as it is one of the few "user-interpretable" parametrizations of A that eliminates the rotation ambiguity of the latent factors. In psychology, the "pure" variables induced by the parametrization are called factorially simple items [39]. A similar condition on A can be traced back to the econometrics literature, and an early reference is [36], further discussed in [4], who called it "zero elements (of A) in specified positions". We refer to [4,36,45] for more examples in psychology, sociology, etc. This parametrization is also called the errors-in-variable parametrization and has wide applications in structural equation models, see, [33,34]. The more recent review paper [50], and the references therein, provide a nice overview of many other concrete applications that support interest in factor models under a parametrization of this type. We provide another example below.

In the context of Assumption 1, we interpret the entries in A as (signed) mixture weights. Under model (1), each X_j is a signed mixture of $Z_1, \ldots, Z_K$, according to these weights. This assumption, which is sufficient for identifiability, is also a a desirable modelling assumption.

As an illustration, assume that $X \in \mathbb{R}^p$ contains gene-level measurements, and that $Z \in \mathbb{R}^K$ corresponds to their biological functions. Then, *(A0)* enables, in this example, to associate a gene with multiple biological functions, in different proportions per function. The inequality sign in *(A0)* further allows some genes not to be associated with any of the functions captured by this model, thereby increasing the robustness of the model. The second requirement, *(A1)*, simply says that the measured X_j and X_ℓ have the same biological function Z_k, and only that function. We considered signed mixtures to increase the flexibility of the model. In this example, signs correspond to the nature of the function. For instance, if gene X_j activates a signaling pathway, and Z_k has positive sign, then Z_k has a function associated with the activation of this pathway, whereas a negative sign indicates a function associated with its inhibition.

Assumption *(A2)* allows us to depart from the widely used, and restrictive, assumption of independence among the latent factors. We require the variability of the factors to be strictly larger than that between factors. This implies the minimal desideratum that the factors be different.

We first discuss the identifiability of β in Section 2.2. Then in Section 3, we propose our estimator of β which uses its identifiability constructively.

2.2. Identifiability of β: A constructive approach

Under model (1), we have $Y = Z^\top \beta + \varepsilon$, and thus the coefficient β satisfies

$$\beta = [\mathrm{Cov}(Z)]^{-1}\mathrm{Cov}(Z, Y) = \Sigma_Z^{-1}\mathrm{Cov}(Z, Y). \tag{4}$$

Since model (2) and Assumption 1 imply $\mathrm{Cov}(Z, Y) = (A^\top A)^{-1}A^\top \mathrm{Cov}(X, Y)$, we have

$$\beta = \Sigma_Z^{-1}(A^\top A)^{-1}A^\top \mathrm{Cov}(X, Y) \tag{5}$$

$$= (\Theta^\top \Theta)^{-1}\Theta^\top \mathrm{Cov}(X, Y) \tag{6}$$

with $\Theta = A\Sigma_Z$. Therefore, β is uniquely defined whenever Θ is unique. By partitioning the $p \times K$ matrix A as $A_{I\bullet} \in \mathbb{R}^{|I|\times K}$ and $A_{J\bullet} \in \mathbb{R}^{|J|\times K}$ corresponding to I and J, respectively, model (2) and Assumption 1 imply the following decomposition of Σ,

$$\Sigma = \begin{bmatrix} \Sigma_{II} & \Sigma_{IJ} \\ \Sigma_{JI} & \Sigma_{JJ} \end{bmatrix} = \begin{bmatrix} A_{I\bullet}\Sigma_Z A_{I\bullet}^\top & A_{I\bullet}\Sigma_Z A_{J\bullet}^\top \\ A_{J\bullet}\Sigma_Z A_{I\bullet}^\top & A_{J\bullet}\Sigma_Z A_{J\bullet}^\top \end{bmatrix} + \begin{bmatrix} \Gamma_{II} & \\ & \Gamma_{JJ} \end{bmatrix}.$$

In particular, we have $\Sigma_{II} = A_{I\bullet}\Sigma_Z A_{I\bullet}^\top + \Gamma_{II}$ and

$$\Theta = A\Sigma_Z = (\Sigma_{\bullet I} - \Gamma_{\bullet I})\, A_{I\bullet}^\top (A_{I\bullet}^\top A_{I\bullet})^{-1}. \tag{7}$$

The uniqueness of Θ is thus implied by that of $A_{I\bullet}$ and Σ_Z. Theorem 1 in [13] shows that, under Assumption 1, the matrices $A_{I\bullet}$ and Σ_Z can be uniquely determined, up to a signed permutation matrix P, from $\Sigma := \mathrm{Cov}(X)$. As a result, Θ can also be recovered from (7) up to $P^\top$, hence β is identifiable from (6) up to $P^\top$. We remark that the permutation matrix P will not affect either inference or prediction. Indeed, writing $\widetilde{A} = AP$, $\widetilde{Z} = P^\top Z$ and $\widetilde{\beta} = P^\top \beta$, one still has $Y = \widetilde{Z}^\top \widetilde{\beta} + \varepsilon$ and $X = \widetilde{A}\widetilde{Z} + W$. We summarize the identifiability of β in the proposition below. Its proof can be found in Appendix B.1.

Proposition 1. *Under models* (1)–(2) *and Assumption* 1, *the quantities* Σ *and* $Cov(X, Y)$ *define* β *uniquely, via* (6) *and* (7), *up to a signed permutation matrix.*

Our estimator, given in the next section, is based on the representations (6) and (7), followed by appropriate plug-in estimators.

3. Estimation of β

We assume that the data consists of n independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ that satisfy model (1) and (2). We write $\mathbf{X} := (X_1, \ldots, X_n)^\top$ for the observed $n \times p$ data matrix and $\mathbf{y} := (Y_1, \ldots, Y_n)^\top$ for the observed response vector. Let $\widehat{\Sigma} = n^{-1} \sum_{i=1}^n X_i X_i^\top$ denote the sample covariance matrix. Motivated by equations (6) and (7), we consider the plug-in estimator of β via the following steps:

(1) Obtain estimates $\widehat{K}$ and $\{\widehat{I}_1, \ldots, \widehat{I}_{\widehat{K}}\}$ from $\widehat{\Sigma}$ with tuning parameter δ by using Algorithm 1 in [13]. For the reader's convenience, we state the procedure in Algorithm 1 below.

(2) Next, for each $a \in [\widehat{K}]$ and $b \in [\widehat{K}] \setminus \{a\}$, we compute

$$[\widehat{\Sigma}_Z]_{aa} = \frac{1}{|\widehat{I}_a|(|\widehat{I}_a| - 1)} \sum_{i \neq j \in \widehat{I}_a} |\widehat{\Sigma}_{ij}|, \quad [\widehat{\Sigma}_Z]_{ab} = \frac{1}{|\widehat{I}_a||\widehat{I}_b|} \sum_{i \in \widehat{I}_a, j \in \widehat{I}_b} \widehat{A}_{ia} \widehat{A}_{ib} \widehat{\Sigma}_{ij}, \tag{8}$$

to form the estimator $\widehat{\Sigma}_Z$ of Σ_Z. Furthermore, the estimation of $A_{I\bullet}$ follows the procedure in [13]. For each $k \in [\widehat{K}]$ and the estimated pure variable set $\widehat{I}_k$,

$$\text{Pick an element } i \in \widehat{I}_k \text{ at random, and set } \widehat{A}_{i\bullet} = \mathbf{e}_k; \tag{9}$$

$$\text{For the remaining } j \in \widehat{I}_k \setminus \{i\}, \text{ set } \widehat{A}_{j\bullet} = \operatorname{sign}(\widehat{\Sigma}_{ij}) \cdot \mathbf{e}_k. \tag{10}$$

(3) Estimate $\Gamma_{\bullet I}$ by $\widehat{\Gamma}_{\bullet \widehat{I}}$ with

$$\widehat{\Gamma}_{ii} = \widehat{\Sigma}_{ii} - \widehat{A}_{i\bullet}^\top \widehat{\Sigma}_Z \widehat{A}_{i\bullet}, \quad \forall\, i \in \widehat{I}, \qquad \widehat{\Gamma}_{ji} = 0, \quad \forall\, j \neq i. \tag{11}$$

(4) Compute

$$\widehat{\Theta} = \left(\widehat{\Sigma}_{\bullet \widehat{I}} - \widehat{\Gamma}_{\bullet \widehat{I}}\right) \widehat{A}_{\widehat{I}\bullet} \left(\widehat{A}_{\widehat{I}\bullet}^\top \widehat{A}_{\widehat{I}\bullet}\right)^{-1}. \tag{12}$$

Provided that $\widehat{\Theta}^\top \widehat{\Theta}$ is non-singular, estimate β by

$$\widehat{\beta} = \left(\widehat{\Theta}^\top \widehat{\Theta}\right)^{-1} \widehat{\Theta}^\top \frac{1}{n} \mathbf{X}^\top \mathbf{y}. \tag{13}$$

The above procedure requires a single tuning parameter δ, and that $K < n$. The theoretical order of δ is given in (17) of Section 4.2 under the sub-Gaussian distributional assumptions. A practical data-driven procedure of selecting δ is stated in Appendix H. We show in Section 5 that $\widehat{\beta}$ coincides with the ordinary least squares estimator that minimizes $\|\mathbf{y} - \widehat{\mathbf{Z}}\beta\|_2^2$ over β, based on an appropriately constructed predictor $\widehat{\mathbf{Z}}$ of the latent data matrix $\mathbf{Z} := (Z_1, \ldots, Z_n)^\top$. We prove in Theorem 3 of Section 4.3 that $\widehat{\Theta}^\top \widehat{\Theta}$ is non-singular with high probability. In practice, in case that $\widehat{\Theta}^\top \widehat{\Theta}$ is singular or ill-conditioned, we propose to invert $\widehat{\Theta}^\top \widehat{\Theta} + t \cdot \mathbf{I}_{\widehat{K}}$ instead, for any small $t > 0$.

In Section 4.4 we discuss other possible estimators based on the alternative representations of β given in (4), (5) and (6).

Algorithm 1 Estimate the partition of the pure variables $\mathcal{I}$ by $\widehat{\mathcal{I}}$

procedure PUREVAR($\widehat{\Sigma}$, δ)
  $\widehat{\mathcal{I}} \leftarrow \varnothing$.
  for all $i \in [p]$ **do**
    $\widehat{I}^{(i)} \leftarrow \left\{l \in [p] \setminus \{i\} : \max_{j \in [p]\setminus\{i\}} |\widehat{\Sigma}_{ij}| \le |\widehat{\Sigma}_{il}| + 2\delta\right\}$
    $Pure(i) \leftarrow True$.
    for all $j \in \widehat{I}^{(i)}$ **do**
      if $\big||\widehat{\Sigma}_{ij}| - \max_{k \in [p]\setminus\{j\}} |\widehat{\Sigma}_{jk}|\big| > 2\delta$ **then**
        $Pure(i) \leftarrow False$,
        break
    if $Pure(i)$ **then**
      $\widehat{I}^{(i)} \leftarrow \widehat{I}^{(i)} \cup \{i\}$
      $\widehat{\mathcal{I}} \leftarrow$ MERGE($\widehat{I}^{(i)}$, $\widehat{\mathcal{I}}$)
  return $\widehat{\mathcal{I}}$ and $\widehat{K}$ as the number of sets in $\widehat{\mathcal{I}}$

function MERGE($\widehat{I}^{(i)}$, $\widehat{\mathcal{I}}$)
  for all $G \in \widehat{\mathcal{I}}$ **do** ▷ $\widehat{\mathcal{I}}$ is a collection of sets
    if $G \cap \widehat{I}^{(i)} \neq \varnothing$ **then**
      $G \leftarrow G \cap \widehat{I}^{(i)}$ ▷ Replace $G \in \widehat{\mathcal{I}}$ by $G \cap \widehat{I}^{(i)}$
      return $\widehat{\mathcal{I}}$
  $\widehat{I}^{(i)} \in \widehat{\mathcal{I}}$ ▷ add $\widehat{I}^{(i)}$ in $\widehat{\mathcal{I}}$
  return $\widehat{\mathcal{I}}$

4. Statistical guarantees

4.1. Minimax lower bounds for estimators of β in essential regression

To benchmark our estimator of β, we derive the minimax optimal rate of $\|\widehat{\beta} - \beta\|_2$ over the parameter space $(\beta, \Sigma_Z, A) \in \mathcal{S}(R, m)$ with

$$\mathcal{S}(R,m) := \Big\{(\beta, \Sigma_Z, A) : \|\beta\|_2 \le R,\ C_{\min} \le \lambda_{\min}(\Sigma_Z) \le \lambda_{\max}(\Sigma_Z) \le C_{\max},\ A \text{ satisfies Assumption 1 with } \min_k |I_k| = m\Big\},$$

where I_k is defined in (3). For the purpose of the minimax result, it suffices to consider the joint distribution of (X, Y) as

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N_{p+1}\left(0, \begin{bmatrix} A\Sigma_Z A^\top + \tau^2 \mathbf{I}_p & A\Sigma_Z\beta \\ \beta^\top \Sigma_Z A^\top & \beta^\top \Sigma_Z \beta + \sigma^2 \end{bmatrix}\right) \tag{14}$$

for $(\beta, \Sigma_Z, A) \in \mathcal{S}(R, m)$ and some positive constants σ^2 and τ^2.

Theorem 2. *Let $K \le \bar{c}(R^2 \vee m)n$ for some positive constant $\bar{c}$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. random variables from the normal distribution in (14). Then, there exist constants $c' > 0$, $c'' \in (0, 1]$ depending only on $\bar{c}$, $C_{\max}$, $C_{\min}$, σ^2 and τ^2, such that*

$$\inf_{\widehat{\beta}} \sup_{(\beta, \Sigma_Z, A) \in \mathcal{S}(R,m)} \mathbb{P}\left\{\|\widehat{\beta} - \beta\|_2 \ge c'\left(1 \vee \frac{R}{\sqrt{m}}\right) \cdot \sqrt{\frac{K}{n}}\right\} \ge c''. \tag{15}$$

The inf is taken over all estimators $\widehat{\beta}$ of β.

Proof. The proof is deferred to Appendix B.2. It uses the classical technique in [46] for proving minimax lower bounds. After carefully constructing a set of "hypotheses" of β, we observe that the Kullback-Leibler (KL) divergence between joint distributions of (X, Y) parametrized by two hypotheses of β can be calculated from the log ratio of corresponding conditional densities of $Y|X$. This observation greatly simplifies the proof. □

The factor $\sqrt{K/n}$ in (15) is the standard minimax rate of estimation in linear regression with observed Z and sub-Gaussian errors. The factor multiplying it can be viewed as the price to pay for not observing Z. It quantifies the trade-off between not observing Z, with strength $\|\beta\|_2$, and the number of times, given by $m = \min_k |I_k|$, each component of Z is partially observed, up to additive error. The ratio $\|\beta\|_2/\sqrt{m}$ indicates that, under the Essential Regression framework, the fact that Z is not observed can be alleviated by the existence of pure variables, and the quality of estimation is expected to increase as m increases. Theorem 2 above shows that, from the point of view of estimating β consistently in Essential Regression, the number of factors K can grow with n, a scenario not treated in the more classical factor regression literature. It also reveals that, as in the classical regression set-up, although K can grow with n in Essential Regression, consistent estimation of unstructured β cannot be guaranteed when $K > n$. This will be treated in follow-up work.

To the best of our knowledge, the minimax lower bound established above is a new result in the factor regression model literature and it is interesting to place our results in a broader, related, context. For this, note that under the Essential Regression framework, *if I and $A_{I\bullet}$ were known*, the pure variable assumption implies

$$Y = Z^\top \beta + \varepsilon, \qquad \bar{X}_I = Z + \bar{W}_I \tag{16}$$

with $\bar{X}_I := (A_{I\bullet}^\top A_{I\bullet})^{-1} A_{I\bullet}^\top X_I$ and $\bar{W}_I := (A_{I\bullet}^\top A_{I\bullet})^{-1} A_{I\bullet}^\top W_I$. Model (16) becomes an instance of an errors in variables model where the covariance structure of the error term $\bar{W}_I$ is diagonal. The minimax optimal lower bound for estimating β in such models has been derived recently in [11], under sparsity assumptions on β. In the particular case of non-sparse β, which we treat here, their lower bound agrees with that given by our Theorem 2, although their bound is derived over a larger class, and can only be compared with (15) when I is known. The closest result to that of Theorem 2 is the minimax lower bound on rows of A bounded in ℓ_1 norm, and has been derived in [13]. We complement this here, in the latent factor regression context, by providing a minimax lower bound on β with a ℓ_2 norm allowed to increase with n.

4.2. Assumptions

In this section we collect the assumptions under which we evaluate the performance of our proposed estimator $\widehat{\beta}$.

We first make the following distributional specifications for ε, W and Z defined in model (1):

Assumption 2. Let $\gamma_\varepsilon, \gamma_w, \gamma_z$ and B_z be positive finite constants. Assume ε is γ_ε-sub-Gaussian[1] and W has independent γ_w-sub-Gaussian entries. Further assume $\|\Sigma_Z\|_\infty \le B_z$ and the random vector $\Sigma_Z^{-1/2} Z$ is γ_z-sub-Gaussian[2].

[1] A mean zero random variable x is called γ-sub-Gaussian if $\mathbb{E}[\exp(tx)] \le \exp(t^2\gamma^2/2)$ for all $t \in \mathbb{R}$.
[2] A mean zero random vector x is called γ-sub-Gaussian if $v^\top x$ is γ-sub-Gaussian for any $\|v\|_2 = 1$.

The quality of our estimator $\widehat{\beta}$ given by (13) depends on how well we estimate K, I, its partition $\{I_k\}_{1\le k\le K}$, as well as Θ. Our goal is to estimate K consistently, under minimal assumptions. However, consistent estimation of the partition requires a stronger set of assumptions that we would like to avoid. We introduce and discuss below a set of assumptions under which the partition is recovered sufficiently well for the purpose of inference on $\widehat{\beta}$.

Assumption 2 implies that X_j is γ_x-sub-Gaussian with $\gamma_x = (\gamma_z\sqrt{B_z} + \gamma_w)$, as shown in Lemma 4 in Appendix C.3, and it is well known (see, for instance, [12], Lemma 1) that in this case,

$$\mathbb{P}\left\{\max_{1\le j<\ell\le p} |\widehat{\Sigma}_{j\ell} - \Sigma_{j\ell}| \le \delta\right\} \ge 1 - (p\vee n)^{-c'} \tag{17}$$

with $\delta = c\sqrt{\log(p\vee n)/n}$, for some constant $c' > 0$ and $c = c(\gamma_x) > 0$ sufficiently large.

Under Assumptions 1 and 2, and when $\log p \le c''n$ for some constant $c'' > 0$, [13] provides an algorithm for estimating K and I, and prove in their Theorem 3 the following:

(1) $\widehat{K} = K$;
(2) $I_k \subseteq \widehat{I}_{\pi(k)} \subseteq I_k \cup J_1^k$, for all $k \in [K]$,

where $\pi : [K] \to [K]$ is a permutation and $J_1^k := \{j \in J : |A_{jk}| \ge 1 - 4\delta/\nu\}$ with constant ν defined in Assumption 1 of Section 2.1 above.

Since we do not impose any separation condition between the pure variable rows $A_{I\bullet}$ and the remaining rows in $A_{J\bullet}$, the sets $\{J_1^k\}_{k=1}^K$ are typically not empty, and as formalized in (2) above, we cannot expect to recover I perfectly in the presence of *quasi-pure* variables with indices belonging to the set $J_1 := \bigcup_{k=1}^K J_1^k$. Indeed, when $\log p = o(n)$, for any $j \in J_1^k$ we have $|A_{jk}| \approx 1$ and $A_{jk'} \approx 0$ for any $k' \ne k$, so variables corresponding to J_1^k are very close to the pure variables in I_k, and possibly indistinguishable from one another, in finite samples.

While allowing for $J_1 \ne \varnothing$ increases the flexibility of the model, it also poses significant technical difficulties in the analysis of the asymptotic distribution of $\widehat{\beta}_k$, for each k, evidenced in the proofs of Theorem 3, Theorem 4 and Proposition 5. Nevertheless, the limiting distribution can still be derived when $|J_1|$ is small relative to $|I|$. The influence of the misclassified X-variables with entries in J_1 becomes negligible in both the finite sample rate analysis of $\widehat{\beta}$ provided in Section 4.3 and the asymptotic analysis of Section 4.5 under the condition given below. We introduce

$$\bar{\rho}^2 = \sum_{k=1}^K \left(\frac{|J_1^k|}{|I_k| + |J_1^k|}\right)^2 \tag{18}$$

to quantify the influence of quasi-pure variables on the quality of our estimation. Theorem 3 shows that optimal estimation of β is possible in the presence of quasi-pure variables as long as their number is negligible relative to the number of pure variables in the same group, in that the following assumption holds.

Assumption 3. 3 The overall proportion $\bar{\rho}^2$ satisfies $m\bar{\rho}^2 = O(1)$ with $m := \min_{k\in[K]} |I_k|$.

We briefly discuss this assumption below. Let $s \le K$ be the number of factors that have quasi-pure variables, that is,

$$s = |S|, \quad S = \{1 \le k \le K : |J_1^k| \ge 1\}.$$

1. Note that, by (18), we have $\bar{\rho}^2 \leq s$, and therefore $m\bar{\rho}^2 \leq ms$. Thus, if both s (and in particular K, when $s = K$) and m remain bounded, as $n \to \infty$, Assumption 3 holds. In other words, in factor models with a possibly large, but fixed, number of factors K, such that one of the factors has very few X-variables solely associated with it, the assumption holds.
2. Otherwise, if $ms \to \infty$, we note that if we assume that

$$|J_1^k| \leq c(|J_1^k| + |I_k|)/\sqrt{ms}, \tag{19}$$

holds for all $k \in S$ where $c > 0$ is some universal constant, then $\bar{\rho}^2 \leq c^2/m$, and Assumption 3 holds. To gain insight into (19), note that it also implies that, for each cluster $k \in S$, $|J_1^k| = o(|I_k|)$, and therefore $|J_1| = o(|I|)$. Thus, in factor models with a growing number of factors and a growing number of pure variables per factor, (19) prevents the number of quasi-pure variables to grow faster than the number of pure variables.
3. In support of the above discussion, we also offer a calculation of $\bar{\rho}^2$ in a particular case. Assume that each I_k has the same size m, and each J_1^k has the same size m'. Then $\bar{\rho}^2 = s(m'/(m+m'))^2$, and we can verify Assumption 3 explicitly in terms of m, m' and s. Consider
   a) $|I_k| = m$ with $m \geq 1$, m fixed, and $|J_1^k| = m' \geq 1$, for all $k \in S$. Then $\bar{\rho}^2 = s(m'/(m+m'))^2$, and Assumption 3 is met if and only if s remains bounded.
   b) $|I_k| = m$ and $|J_1^k| = m' = O(m^\alpha)$, for all $k \in S$, with $0 \leq \alpha \leq 1$ and $m \to \infty$. A simple calculation shows that Assumption 3 is met only if $s = O(m^{1-2\alpha})$. In particular, $\alpha < 1/2$ allows $s \to \infty$; $\alpha = 1/2$ requires s to be bounded, and the case $\alpha > 1/2$ forces $s = 0$ (all the stated limits are in terms of $n \to \infty$).

Another quantity that needs to be controlled is the covariance matrix Σ_Z. It plays the same role as the Gram matrix in classical linear regression with random design.

Assumption 4. The smallest eigenvalue $\lambda_{\min}(\Sigma_Z) > C_{\min}$ for some constant $C_{\min}$ bounded away from 0.

Assumptions 2–4 allow for a cleaner presentation of our results. We can trace explicitly the dependency of the estimation rate for β on $\bar{\rho}$ and $C_{\min}$ in the proofs. An important feature of this framework is that under Assumptions 1–4 and $K = O(n/\log(p \vee n))$, the matrix $\widehat{\Theta}^\top \widehat{\Theta}$ can be inverted, with probability tending to 1.

We require one more condition, which measures the strength of the signal $A\Sigma_Z A^\top$ retained by the low rank approximation of $\Sigma = A\Sigma_Z A^\top + \Gamma$.

Assumption 5. The Kth eigenvalue $\lambda_K := \lambda_K(A\Sigma_Z A^\top)$ of the signal $A\Sigma_Z A^\top$ satisfies

$$\lambda_K \geq c\, p\sqrt{\frac{\log(p \vee n)}{n}} \tag{20}$$

for some sufficiently small constant $c > 0$. This implies $K \lesssim \sqrt{n/\log(p \vee n)}$.

The implication $K \lesssim \sqrt{n/\log(p \vee n)}$ follows immediately from (20) and the bound $\lambda_K \leq B_z(p/K)$ in Lemma 14.

We recall that the quantity $\lambda_K := \lambda_K(A\Sigma_Z A^\top)$ quantifies the size of the signal in $X = AZ + W$. It is a key quantity in the well studied problem of signal recovery from a $n \times p$ matrix of noisy observations $\mathbf{X}$, with rows corresponding to n i.i.d. copies of X. The signal can be recovered from $n^{-1}\mathbf{X}^\top\mathbf{X}$ as soon as λ_K is above the noise level [20,22,31,48,49]. The noise level is quantified by the largest eigen-value

$\lambda_1(n^{-1}\mathbf{W}^\top\mathbf{W})$, based on the $n \times p$ data matrix $\mathbf{W}$ with rows corresponding to n i.i.d. copies of W. Standard random matrix theory shows that $\lambda_1(n^{-1}\mathbf{W}^\top\mathbf{W})$ concentrates with overwhelming probability around its mean, which is of order $(n+p)/n$, see [47]. Therefore, one needs at least $\lambda_K \gtrsim (n+p)/n$ to distinguish the signal from the noise. For the more specific task of optimal estimation of β, we require Assumption 5 (see Lemma 13 in Appendix D). The investigation of its optimality is beyond the scope of this paper, and will be studied in a follow up work. However, we emphasize that, as mentioned in the Introduction, the consistent estimation of rows of the factor loadings in factor models, especially when $p > n$, has only been established under the stricter condition $\lambda_K \asymp p$ [5,6,26,27]. The intuition behind this more restrictive assumption is as follows. If Σ_Z is positive definite, with finite eigenvalues, and the rows of A are p i.i.d. draws of a K-dimensional sub-Gaussian random vector, $p > K$, then reasoning as above and using the results in [47], $\lambda_K \asymp p$, with high probability. However, for a generic, deterministic, A, it would be an assumption, that we show can be considerably relaxed to Assumption 5. More details are provided in Remark 3 of Section 4.5.

4.3. Consistency of $\widehat{\beta}$ in ℓ_2-norm: Rates of convergence and optimality

The following theorem states the convergence rate of $\min_{P\in\mathcal{H}_K} \|\widehat{\beta} - P\beta\|_2$.

Theorem 3. *Suppose Assumptions 1–4 hold and assume $K\log(p\vee n) \le cn$ for some sufficiently small constant $c > 0$. Then, with probability greater than $1-(p\vee n)^{-c'}$ for some constant $c' > 0$, $\widehat{K} = K$, the matrix $\widehat{\Theta}^\top\widehat{\Theta}$ is non-singular and the estimator $\widehat{\beta}$ given by (13) satisfies*:

$$\min_{P\in\mathcal{H}_K} \left\|\widehat{\beta} - P\beta\right\|_2 \lesssim \left(1 \vee \frac{\|\beta\|_2}{\sqrt{m}}\right)\sqrt{\frac{K\log(p\vee n)}{n}}\left(1 \vee \frac{p}{\lambda_K}\sqrt{\frac{\log(p\vee n)}{n}}\right) \tag{21}$$

If additionally Assumption 5 holds, then with the same probability, $\widehat{\beta}$ given by (13) satisfies

$$\min_{P\in\mathcal{H}_K} \left\|\widehat{\beta} - P\beta\right\|_2 \lesssim \left(1 \vee \frac{\|\beta\|_2}{\sqrt{m}}\right)\sqrt{\frac{K\log(p\vee n)}{n}}. \tag{22}$$

Proof. The proof is given in Appendix D. □

Remark 1. The estimator $\widehat{\beta}$ achieves the minimax rate in Theorem 2 up to a logarithmic $\log(p\vee n)$ term. Inspection of the proof, when K grows with n, shows that the $\log(p\vee n)$ terms appearing in the condition $K\log(p\vee n) \le cn$ and in the upper bound (22) can be improved to $\log K$, but in that case the probability tail in Theorem 3 will change to $1 - K^{-c}$. This additional $\log K$ is the price to pay for not observing Z. On the other hand, comparing (22) with the convergence rate of the oracle least squares estimator (OLS) *when Z is observable*, the extra factor $\|\beta\|_2/\sqrt{m}$ is due to the error term W in $X = AZ + W$. This extra factor becomes negligible when the coefficients of β are uniformly bounded ($\|\beta\|_\infty \lesssim 1$) and the number of latent factors cannot grow much faster than the cardinality of the smallest subgroup of pure variables ($K \lesssim m$). In the worst case scenario the rate of $\widehat{\beta}$ is slower than the aforementioned OLS by a factor of the order of $\|\beta\|_2$.

Remark 2. The selection of K, the number of latent factors, has been thoroughly studied, in general factor models. For completeness, we provide Table 2 summarizing the existing approaches of selecting K, as well as the conditions under which the resulting estimates are consistent. As the table shows, consistent estimation of K via the existing methods is proved under the assumption that there exists

$\widehat{K} = K$ w.h.p.	K is fixed	K grows with n
Existing literature: [2,7,41]	$\lambda_1 \asymp \lambda_K \asymp p$, $\|\Gamma\|_{\infty,1} = O(1)$	NA
Proposed method	Assumption 1, Γ is diagonal with bounded entries, $\log p = o(n)$	

Table 2. Selection of the number of factors K: methods and sufficient conditions for consistency. We write $\lambda_1, \ldots, \lambda_K$ for the top K eigenvalues of $A\Sigma_Z A^T$

a large gap between the eigenvalues of $A\Sigma_Z A^T$ and Γ, respectively. Although in the inference results derived in this manuscript we also need $\lambda_K = \lambda_K(A\Sigma_Z A^T)$ to satisfy Assumption 3, that is,

$$\frac{\lambda_K}{p} \gtrsim \frac{\log(p \vee n)}{\sqrt{n}},$$

this is *not* used to guarantee the correct selection of K (this can be readily seen from the discussion in the middle of page 15), hence it is much milder than the conditions in the existing literature [2, 7,41], as seen from the table. We do, however, also establish that K can be consistently estimated by our procedure. Leveraging the particular class of factor models treated in this work, we show that our method is consistent, for both growing K and fixed K, but under a set of conditions that is very different than those previously considered (please see Table 2). In particular, we do not require λ_1 and λ_K to be of equal order p, but we do consider only diagonal error structures Γ.

4.4. Other possible estimators

We discuss other natural estimators that could be considered in this model, based on equivalent representations of β. Each of these representations offer a valid basis for estimating β via plug-in estimation of the unknown quantities. However, we recommend the estimator $\widehat{\beta}$ in (13) above, as it has several advantages over these other candidates, both theoretically and numerically.

(a) Recall that β can be uniquely defined via identity (5) as long as A and Σ_Z are unique up to signed permutations. The expression (5) suggests the following estimator

$$\widetilde{\beta}^{(A)} = \widehat{\Sigma}_Z^{-1} \left(\widehat{A}^\top \widehat{A}\right)^{-1} \widehat{A}^\top \frac{1}{n} \mathbf{X}^\top \mathbf{y} \tag{23}$$

based on some estimates $\widehat{A}$ of A and $\widehat{\Sigma}_Z$ of Σ_Z. For instance, one can estimate A and Σ_Z by the procedure given in (8), (9), (10) and (33), as these estimates have optimal convergence rates [13].

(b) Building on identity (5), using $\Sigma - \Gamma = A\Sigma_Z A^\top$ and writing $B = A(A^\top A)^{-1}$, we can show that

$$\beta = \left[B^\top(\Sigma - \Gamma)B\right]^{-1} B^\top \mathrm{Cov}(X, Y). \tag{24}$$

If A and Σ_Z are unique up to signed permutations, then so is $\Gamma = \Sigma - A\Sigma_Z A^\top$. Equipped with $\widehat{A}$, one can further estimate Γ via (30) and estimate B by $\widehat{B} = \widehat{A}(\widehat{A}^\top \widehat{A})^{-1}$. Then expression (24) provides another way of estimating β via

$$\widehat{\beta}^{(A)} = \left[\widehat{B}^\top(\widehat{\Sigma} - \widehat{\Gamma})\widehat{B}\right]^{-1} \widehat{B}^\top \frac{1}{n} \mathbf{X}^\top \mathbf{y}. \tag{25}$$

The two estimates are the same, $\widetilde{\beta}^{(A)} = \widehat{\beta}^{(A)}$, if $\widehat{\Gamma} = \widehat{\Sigma} - \widehat{A}\widehat{\Sigma}_Z\widehat{A}^\top$ in (25). Since $\widehat{\beta}^{(A)}$ uses the diagonal structure of Γ, it is expected to have better performance than $\widetilde{\beta}_{(A)}$. However, both (23) and (25) require

separate estimation of A and Σ_Z by $\widehat{A}$ and $\widehat{\Sigma}_Z$, respectively. In contrast, our estimator $\widehat{\beta}$ in (13) estimates $\Theta = A\Sigma_Z$ *as a whole*, leading to better rate performance as evidenced in the simulation section. Furthermore, as we mentioned in Section 3, the proposed $\widehat{\beta}$ has a simple interpretation, as a ordinary least squares estimator, relative to appropriately constructed predictors of Z, which makes it more appealing in practice. We give the details in Section 5.

(c) Recall that $X_I = A_{I\bullet}Z + W_I$ and $A_{I\bullet}$ has full rank under model (2) and Assumption 1. By using (4) and $\mathrm{Cov}(Z, Y) = (A_{I\bullet}^\top A_{I\bullet})^{-1} A_{I\bullet}^\top \mathrm{Cov}(X_I, Y)$, we find the identity

$$\beta = \Sigma_Z^{-1}(A_{I\bullet}^\top A_{I\bullet})^{-1} A_{I\bullet}^\top \mathrm{Cov}(X_I, Y) = \left(\Sigma_Z A_{I\bullet}^\top A_{I\bullet}\Sigma_Z\right)^{-1} \Sigma_Z A_{I\bullet}^\top \mathrm{Cov}(X_I, Y). \tag{26}$$

This expression of β relies only on $A_{I\bullet}\Sigma_Z$ rather than the full matrix $\Theta = A\Sigma_Z$ as in (5). This is yet a different way of estimating β and identity (26) suggests the following estimator of β

$$\widehat{\beta}^{(I)} = \widehat{\Sigma}_Z^{-1}\left(\widehat{A}_{\widehat{I}\bullet}^\top \widehat{A}_{\widehat{I}\bullet}\right)^{-1} \widehat{A}_{\widehat{I}\bullet}^\top \frac{1}{n}\mathbf{X}^\top \mathbf{y}. \tag{27}$$

In Appendix G, we state, without its lengthy proof due to space restrictions, that $\widehat{\beta}^{(I)}$ has the same convergence rate as (22) under Assumptions 1–4. However, we still recommend $\widehat{\beta}$ over $\widehat{\beta}^{(I)}$ as $\widehat{\beta}$ has better numerical performance and has a smaller asymptotic variance (see Theorem 4 and Theorem 31 in the Appendix). We also verify these points in the simulation study presented in Appendix I.

4.5. Component-wise asymptotic normality of $\widehat{\beta}$

In this section, for ease of presentation, we assume that the signed permutation matrix P is identity and we consider $\Gamma = \tau^2 \mathbf{I}_p$ and $|I_k| = m$ for all $k \in [K]$, but our proof holds for the general case and the corresponding explicit, general, expression of the asymptotic variance is given in display (76) of the Appendix.

The component-wise asymptotic normality of $\widehat{\beta}$ is proved under the challenging, but realistic, scenario in which some of the non-pure variables are very close to the pure variables, justifying their name, quasi-pure variables, introduced in Section 4.2 above. Allowing for this situation is similar to relaxing the signal strength conditions used in the literature on support recovery. In our context, they would correspond to requiring that the pure and non-pure variables are well separated, in that

$$\min_{j\in J} \min_{P\in\mathcal{H}_K} \left\| A_{j\bullet} - P\mathbf{e}_1 \right\|_1 \geq c\sqrt{\log(p \vee n)/n}$$

for some universal constant $c > 0$, *an assumption that we do not make*. We note that such assumption would be equivalent to requiring $\bar{\rho} = 0$, for $\bar{\rho}$ defined in (18).

In Section 4.2 we established the convergence rate of $\widehat{\beta}$ when $\bar{\rho} \neq 0$, but satisfies Assumption 3. For the asymptotic normality result, we can still allow for $\bar{\rho} \neq 0$, but require it to be of the smaller size stated in Assumption 3′.

Assumption 3′. The overall proportion $\bar{\rho}$ satisfies $\bar{\rho}^2 \log(p \vee n) = o(1/m)$ as $n \to \infty$.

Assumption 5′. The Kth eigenvalue $\lambda_K = \lambda_K(A\Sigma_Z A^\top)$ of $A\Sigma_Z A^\top$ satisfies

$$\lambda_K \Big/ \frac{p\log(p \vee n)}{\sqrt{n}} \to \infty, \text{ as } n \to \infty.$$

Theorem 4. *Under Assumptions 1, 2, 3′, 4 and 5′, assume $\gamma_w/\tau = O(1)$ and $\gamma_\varepsilon/\sigma = O(1)$. Then, with probability tending to one, $\widehat{K} = K$, $\widehat{\Theta}^\top\widehat{\Theta}$ is non-singular and for any $1 \le k \le K$,*

$$\sqrt{n/V_k}(\widehat{\beta}_k - \beta_k) \xrightarrow{d} N(0,1), \qquad as \quad n \to \infty,$$

where, with $\Theta^+ = (\Theta^\top\Theta)^{-1}\Theta^\top$,

$$V_k = \left(\sigma^2 + \frac{\tau^2}{m}\|\beta\|_2^2\right)\left[\Omega_{kk} + \tau^2 \mathbf{e}_k^\top\left(\Theta^\top\Theta\right)^{-1}\mathbf{e}_k\right] + \frac{\tau^4}{m(m-1)}\sum_{a=1}^{K}\beta_a^2\sum_{i\in I_a}\left[\mathbf{e}_k^\top\Theta^+\mathbf{e}_i\right]^2. \tag{28}$$

Furthermore, if additionally $\lambda_K/\tau^2 \to \infty$ holds, then

$$V_k = \left(\sigma^2 + \frac{\tau^2}{m}\|\beta\|_2^2\right)\Omega_{kk}. \tag{29}$$

Proof. The proof is deferred to Appendix E, and we offer insights into its main steps below. □

Outline of the proof of Theorem 4. There are four main steps in the proof. We briefly explain them below and highlight the difficulties in each step.

In the classical approach of establishing the asymptotic normality for $\widehat{\beta}_k - \beta_k$, a crucial step is to decompose the expression of $\widehat{\beta}_k - \beta_k$ as a sum of independent mean-zero random variables, that serves as the leading term, plus a remainder term of smaller order. The asymptotic variance of the main term determines the asymptotic variance of $\widehat{\beta}_k - \beta_k$.

Under our setting (1) and (2), the first step of proving Theorem 4 is to establish such a decomposition on the event that the dimensions of $\widehat{\beta}$ and β are equal. This event holds with overwhelming probability tending to one. In display (75) of the proof, we show that indeed, on this event,

$$\sqrt{n}(\widehat{\beta}_k - \beta_k) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_{ik} + \sqrt{n}([\mathrm{Rem}_1]_k + [\mathrm{Rem}_2]_k),$$

with Rem_1 and Rem_2 defined in display (72). Each summand ξ_{ik} is in turn a sum of four terms that are bi-linear combinations of ε, Z and W. The ξ_{ik} are independent and form a triangular array since K and p may grow in n. Interestingly, if $X = AZ$ and $W = 0$, ξ_{ik} reduces to $\mathbf{e}_k^\top\Sigma_Z^{-1}\mathbf{Z}_{i\bullet}\varepsilon$, the usual error term in the analysis of the ordinary least squares estimator based on observed Z. For the two remainder terms, we note that Rem_1 depends on the error of estimating $\Theta^+ := (\Theta^\top\Theta)^{-1}\Theta^\top$, while Rem_2 is induced by the existence of quasi-pure variables indexed by J_1.

The second step of the proof is to calculate the first two moments of ξ_{ik} via Lemmas 15 and 18, which is relatively straightforward algebra.

In the third step, we apply Lyapunov's central limit theorem to $\sum_{i=1}^n \xi_{ik}$ for triangular arrays. Verification of the Lyapunov condition requires calculation of the third moments of ξ_{ik}. We rely on Rosenthal's inequality and a careful analysis to accomplish this in Lemma 16.

The final, fourth step is to show that both $[\mathrm{Rem}_1]_k$ and $[\mathrm{Rem}_2]_k$ are negligible as $n \to \infty$. This requires a fair amount of work.

To control the remainder term Rem_1, a key step is to provide upper bound for the quantity $(\widehat{\Theta} - \Theta)^\top\widehat{\Theta}(\widehat{\Theta}^\top\widehat{\Theta})^{-1}$ and even establishing the existence of $(\widehat{\Theta}^\top\widehat{\Theta})^{-1}$ requires a delicate analysis. Lemmas 10 and 13 are devoted to this goal. In order to ensure that $[\mathrm{Rem}_1]_k/\sqrt{V_k} = o_p(1)$, we need the signal strength condition in Assumption 5′, which is slightly stronger relative to Assumption 5. Assumption

5′ implies $K\log(p\vee n)=o(\sqrt{n})$, needed for analyzing the estimator of $(\Theta^\top\Theta)^{-1}$, as we recall that we allow for $\Theta^\top\Theta$ to be general, in particular we do not impose any sparsity assumption on it.

The remainder term $[\mathrm{Rem}_2]_k$ is a complicated function of random quantities that involve sums or maxima over the quasi-pure variable index set J_1. If no such variable exists ($\bar\rho=0$), then $\mathrm{Rem}_2=0$, and the proof ends. However, since we allow for $\bar\rho\neq 0$, it turns out to be challenging to show that $\sqrt{n/V_k}[\mathrm{Rem}_2]_k$ still vanishes asymptotically under Assumption 3′. This is done in Lemmas 17 and 20.

In practice, estimation of V_k is required to construct valid confidence intervals for individual coordinates of β. We propose a simple plug-in estimator $\widehat{V}_k$ by substituting σ^2, τ_i^2, $|I_k|$, β and Ω by their estimates. Specifically, we use $\widehat{\Omega}=\widehat{\Sigma}_Z^{-1}$ and estimate σ^2 and τ_i^2 by

$$\widehat{\tau}_i^2=\widehat{\Sigma}_{ii}-\widehat{A}_{i\bullet}^\top\widehat{\Sigma}_Z\widehat{A}_{i\bullet},\quad \text{for all } i\in[p]; \tag{30}$$

$$\widehat{\sigma}^2=\frac{1}{n}\mathbf{y}^\top\mathbf{y}-2\widehat{\beta}^\top\widehat{h}+\widehat{\beta}^\top\widehat{\Sigma}_Z\widehat{\beta} \tag{31}$$

with

$$\widehat{h}=\frac{1}{n}\left(\widehat{A}_{\widehat{I}\bullet}^\top\widehat{A}_{\widehat{I}\bullet}\right)^{-1}\widehat{A}_{\widehat{I}\bullet}^\top\mathbf{X}_{\bullet\widehat{I}}^\top\mathbf{y}. \tag{32}$$

If either $\widehat{\tau}_i^2$ or $\widehat{\sigma}^2$ is negative, we set it to 0. We estimate $A_{I\bullet}$ according to (9) – (10) and estimate $A_{J\bullet}$ by using the Dantzig-type estimator $\widehat{A}_D$ proposed in [13] given by

$$\widehat{A}_{j\bullet}=\arg\min_{\beta^j\in\mathbb{R}^K}\left\{\|\beta^j\|_1:\ \left\|\widehat{\Sigma}_Z\beta^j-(\widehat{A}_{\widehat{I}\bullet}^\top\widehat{A}_{\widehat{I}\bullet})^{-1}\widehat{A}_{\widehat{I}\bullet}^\top\widehat{\Sigma}_{\widehat{I}j}\right\|_\infty\le c\sqrt{\log(p\vee n)/n}\right\} \tag{33}$$

for any $j\in\widehat{J}$, with some constant $c>0$. The estimator $\widehat{A}$ enjoys the optimal convergence rate of $\max_{j\in[p]}\|\widehat{A}_{j\bullet}-A_{j\bullet}\|_q$ for any $1\le q\le\infty$ [13], Theorem 5. Finally, Θ is estimated by (12).

The next proposition shows that the plug-in estimator $\widehat{V}_k$ consistently estimates the asymptotic variance V_k of $\widehat{\beta}_k$.

Proposition 5. *Under the same conditions of Theorem 4, we have with probability tending to one,* $\widehat{K}=K$ *and*

$$\left|\widehat{V}_k^{1/2}/V_k^{1/2}-1\right|=o_p(1).$$

Consequently, we have with probability tending to one, $\widehat{K}=K$ *and*

$$\sqrt{n/\widehat{V}_k}(\widehat{\beta}_k-\beta_k)\xrightarrow{d}N(0,1),\qquad \text{as } n\to\infty,\quad k\in[K].$$

Proof. The proof of the consistency of $\widehat{V}_k$ is given in Appendix F and the rest of the proof follows from Theorem 4 and an application of the Slutsky's theorem. □

Remark 3. If we treat model (1) and (2) as an augmented factor model, the vector β simply corresponds to a particular row of the augmented matrix $\widetilde{A}=[A^\top,\beta^\top]^\top$. As mentioned in the Introduction, and to the best of our knowledge, the only asymptotic normality results, with explicit asymptotic variance, have been derived in [6], when I and K *are known*, K is a fixed constant, and $p\to\infty$. In this

framework, [6] show

$$\sqrt{n/Q_k}(\widetilde{\beta}_k - \beta_k) \xrightarrow{d} N(0,1), \qquad \text{with} \qquad Q_k = \left(\sigma^2 + \tau^2 \|\beta\|_2^2\right) \Omega_{kk}, \tag{34}$$

as $n \to \infty$, for $1 \le k \le K$, for an MLE-type estimator $\widetilde{\beta}$ of β.

We discuss below the relative computational and theoretical enhancements offered by Theorem 4 above, first within their framework, and then beyond it.

At the computational level, [6] propose to estimate β via an alternating EM-algorithm, but offer no guarantees that this estimator coincides with the MLE-type estimator $\widetilde{\beta}$ that is theoretically studied. In contrast, the estimator $\widehat{\beta}$ constructed in Section 3 is also the estimator analyzed theoretically in this work.

At the theoretical level, the estimator $\widetilde{\beta}$ of [6] is analyzed under

$$p \lesssim \lambda_K(A\Sigma_Z A^\top) \le \lambda_1(A\Sigma_Z A^\top) \lesssim p, \qquad c \le \lambda_K(\Sigma_Z) \le \lambda_1(\Sigma_Z) \le C \tag{35}$$

for some constants $0 < c < C < \infty$, among other technical assumptions. Our estimator $\widehat{\beta}$ is analyzed under considerably weaker assumptions. For instance, our condition on $\lambda_K(A\Sigma_Z A^\top)$ in Assumption 5 considerably relaxes the above condition (35), we don't require $\lambda_1(A\Sigma_Z A^\top) \asymp \lambda_K(A\Sigma_Z A^\top)$, but we allow the condition number of $A\Sigma_Z A^\top$ to grow as fast as $n^{1/2}/\log(p \vee n)$. The latter follows from Assumption 5′ in conjunction with the fact that $\lambda_1(A\Sigma_Z A^\top) \lesssim p$.

Furthermore, when (35) holds, $\|\beta\|_2$ is bounded and $p \to \infty$, as assumed in [6], display (29) shows that the asymptotic variance V_k of our estimator $\widehat{\beta}_k$ reduces to Q_k given in (34) above, when $m = 1$ and I is *known*, as considered in [6].

In summary, Theorem 4 holds uniformly over $\beta \in \mathbb{R}^K$, and for both fixed and growing dimensions K and p, and whether I is known or not. While unifying these cases requires a much more complicated analysis, the reward is that our asymptotic analysis, and in particular the asymptotic variance V_k can be derived simultaneously for all cases of interest. Furthermore, we provide a consistent estimator of V_k, which to the best of our knowledge has not been considered elsewhere.

5. Essential regression as regression on latent cluster centers

The decomposition (2) in our model formulation can be used as a model for possibly overlapping clustering. For this, we interpret A as an allocation matrix that assigns the X-variables to possibly overlapping groups G_k corresponding to the components of Z via

$$G_k = \{j \in [p] : A_{jk} \neq 0\}.$$

This approach was first proposed in [13], and their algorithm, called LOVE, was shown to estimate clusters consistently. With this interpretation, the quantity (at the population level)

$$\bar{X} := (A^\top A)^{-1} A^\top X,$$

can be viewed as weighted cluster averages of all variables. As discussed in the Introduction, Essential Regression provides a framework within which we can analyze when the commonly used cluster averages can be used for downstream analysis, with statistical guarantees. For inference on β, we remark that, at the population level,

$$\beta = \arg\min_\alpha \mathbb{E}\left[\left(Y - \alpha^\top \widetilde{Z}\right)^2\right] \neq \arg\min_\alpha \mathbb{E}\left[\left(Y - \alpha^\top \bar{X}\right)^2\right] \tag{36}$$

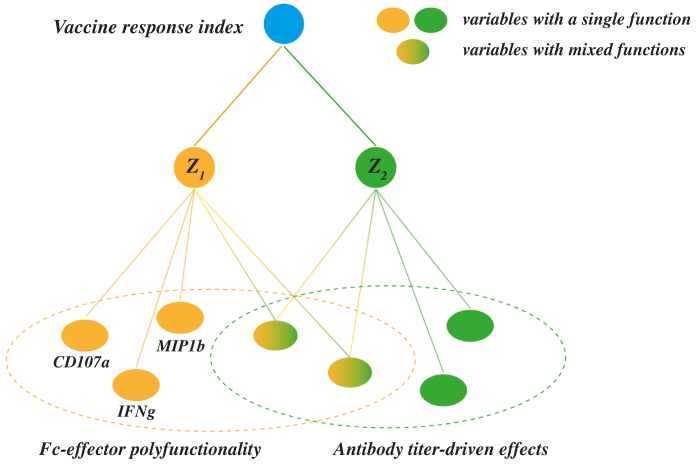

Figure 2. Two representative clusters with their pure variables. (The overlapping variables between these two clusters are more than the plot shows.)

where $\widetilde{Z}$ is the best linear predictor (BLP) of Z from $\bar{X}$, given by

$$\widetilde{Z} = \mathrm{Cov}(Z, \bar{X})[\mathrm{Cov}(\bar{X})]^{-1}\bar{X} = \Theta^\top\Theta\left(\Theta^\top\Sigma\Theta\right)^{-1}\Theta^\top X. \tag{37}$$

Display (36) suggests that estimation of β should be based on the BLP of Z rather than the weighted cluster averages. This is indeed true. We let $\widehat{\mathbf{Z}} = \mathbf{X}\widehat{\Theta}(\widehat{\Theta}^\top\widehat{\Sigma}\widehat{\Theta})^{-1}\widehat{\Theta}^\top\widehat{\Theta}$, which is well defined provided that $(\widehat{\Theta}^\top\widehat{\Sigma}\widehat{\Theta})^{-1}$ exists. The latter is met with high probability under Assumption 5. Consider the least squares estimator $\widetilde{\beta}$ corresponding to regressing $\mathbf{y}$ onto $\mathbf{Z}$. Then

$$\widetilde{\beta} = \left(\widehat{\mathbf{Z}}^\top\widehat{\mathbf{Z}}\right)^{-1}\widehat{\mathbf{Z}}^\top\mathbf{y} = \left(\widehat{\Theta}^\top\widehat{\Theta}\right)^{-1}\left(\widehat{\Theta}^\top\widehat{\Sigma}\widehat{\Theta}\right)\left(\widehat{\Theta}^\top\widehat{\Sigma}\widehat{\Theta}\right)^{-1}\widehat{\Theta}^\top\mathbf{X}^\top\mathbf{y} = \widehat{\beta},$$

by using $\widehat{\Sigma} = n^{-1}\mathbf{X}^\top\mathbf{X}$. This natural approach yields exactly the same estimator $\widehat{\beta}$ we introduced in Section 3. We can thus view $\widehat{\beta}$ as a post-clustering estimator, and the results of Theorems 3 and 4 as pertaining to post-clustering inference, at the factor level.

6. Analysis of SIV-vaccine induced humoral immune responses

We tested Esential Regression on a high-dimensional dataset of vaccine-induced humoral immune responses, from a recently published study that demonstrated multiple antibody-centric mechanisms of vaccine-induced protection against SIV [1], the non-human primate equivalent of HIV. The dataset comprised $p = 191$ antibody functional and biophysical properties, including Fc effector functions, glycosylation profiles and binding to Fc receptors. The properties were measured for $n = 60$ non-human primates (NHPs). For each NHP, the level of protection offered by the vaccine (number of intra-rectal SIV challenges after which the NHP got infected or whether the NHP remained uninfected after the maximum number of challenges for the study, 12, normalized by the total number of challenges) was used as the outcome $Y \in [0, 1]$ we regressed to.

One goal of the study was to determine the un-observed humoral signatures associated with the level of protection offered by the vaccine Y, and suggests therefore a latent factor regression framework.

In particular, the Essential Regression model is ideally suited for this data set, in light of prior biological knowledge on the measured X-variables: some of the measured antibody properties work in tandem with several other properties (mixed-function variables), while others are part of individual immunological signatures (pure/single-function variables) [19,40]. For this data set we used the algorithm developed in [13] to obtain an estimator $\widehat{K} = 10$ of the number of factors.

We used the asymptotic normality of $\widehat{\beta}$, established in Section 4.5 above, to determine the strength of association between Y and the biologically interpretable immunological signatures. This task is difficult to accomplish, with theoretical guarantees, outside a latent factor regression framework. Common existing approaches include standard regularized regression at the observed bio-marker level, followed by an ad-hoc re-creation of clusters and cluster centers [1]. Although subsequent regression of Y onto cluster centers, appropriately defined, can be easily performed, theoretical justifications of such procedure is lacking. In contrast, Essential Regression coupled with Theorem 4 and Proposition 5 provides a principled way of regressing directly onto the latent cluster centers. Figure 2 depicts the top two biological functions associated with the level of protection offered by the vaccine, under FDR-control. The estimated coefficients are $\widehat{\beta}_1 = 0.104$ with asymptotic 90% confidence interval $[0, 0.21]$, and $\widehat{\beta}_2 = 0.105$ with 90% asymptotic confidence interval $[0.02, 0.19]$, corresponding respectively to Z_1 on the left of Figure 2, and to Z_2, on the right. On the basis of the pure and mixed variables in the two associated clusters, Z_1 and Z_2 can be broadly defined as Polyfunctionality involving multiple Fc effector functions and Enhanced IgG titers and FcR2A binding, respectively. These findings are in excellent alignment with biological expectations, providing strong support for the applicability of the methods and theory developed in this work even in data sets of modest sample size.

Acknowledgements

We thank the two anonymous referees for their many insightful and helpful suggestions. We are grateful to Jishnu Das for help with the interpretation of our data analysis results.

Funding

Bunea and Wegkamp are supported in part by NSF grants DMS-1712709 and DMS-2015195.

Bing is supported in part by NSF grant DMS-1407600.

Supplementary Material

Supplement to "Inference in latent factor regression with clusterable features" (DOI: 10.3150/21-BEJ1374SUPP; .pdf). The supplementary document includes all the proofs, the data-driven selection of the tuning parameter, simulations and auxiliary results.

References

[1] Ackerman, M.E., Das, J., Pittala, S., Broge, T., Linde, C., Suscovich, T.J., Brown, E.P., Bradley, T., Natarajan, H., Lin, S., Sassic, J.K., O'Keefe, S., Mehta, N., Goodman, D., Sips, M., Weiner, J.A., Tomaras, G.D., Haynes, B.F., Lauffenburger, D.A., Bailey-Kellogg, C., Roederer, M. and Alter, G. (2018). Route of immunization defines multiple mechanisms of vaccine-mediated protection against SIV. *Nat. Med.* **24** 1590–1598. https://doi.org/10.1038/s41591-018-0161-0

[2] Ahn, S.C. and Horenstein, A.R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81** 1203–1227. MR3064065 https://doi.org/10.3982/ECTA8968
[3] Anderson, T.W. and Amemiya, Y. (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *Ann. Statist.* **16** 759–771. MR0947576 https://doi.org/10.1214/aos/1176350834
[4] Anderson, T.W. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. V* 111–150. Berkeley and Los Angeles, CA: Univ. California Press. MR0084943
[5] Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171. MR1956857 https://doi.org/10.1111/1468-0262.00392
[6] Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40** 436–465. MR3014313 https://doi.org/10.1214/11-AOS966
[7] Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259 https://doi.org/10.1111/1468-0262.00273
[8] Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* **74** 1133–1150. MR2238213 https://doi.org/10.1111/j.1468-0262.2006.00696.x
[9] Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *J. Econometrics* **146** 304–317. MR2465175 https://doi.org/10.1016/j.jeconom.2008.08.010
[10] Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. *J. Amer. Statist. Assoc.* **101** 119–137. MR2252436 https://doi.org/10.1198/016214505000000628
[11] Belloni, A., Rosenbaum, M. and Tsybakov, A.B. (2017). Linear and conic programming estimators in high dimensional errors-in-variables models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 939–956. MR3641415 https://doi.org/10.1111/rssb.12196
[12] Bien, J., Bunea, F. and Xiao, L. (2016). Convex banding of the covariance matrix. *J. Amer. Statist. Assoc.* **111** 834–845. MR3538709 https://doi.org/10.1080/01621459.2015.1058265
[13] Bing, X., Bunea, F., Ning, Y. and Wegkamp, M. (2020). Adaptive estimation in structured factor models with applications to overlapping clustering. *Ann. Statist.* **48** 2055–2081. MR4134786 https://doi.org/10.1214/19-AOS1877
[14] Bing, X., Bunea, F. and Wegkamp, M. (2020). Detecting approximate replicate components of a high-dimensional random vector with latent structure. arXiv:2010.02288.
[15] Bing, X., Bunea, F. and Wegkamp, M. (2022). Supplement to "Inference in latent factor regression with clusterable features." https://doi.org/10.3150/21-BEJ1374SUPP
[16] Blei, D.M. and McAuliffe, J.D. (2007). Supervised topic models. In *Proceedings of the* 20*th International Conference on Neural Information Processing Systems*. *NIPS'*07 121–128. USA: Curran Associates Inc.
[17] Boivin, J. and Ng, S. (2006). Are more data always better for factor analysis? *J. Econometrics* **132** 169–194. MR2271395 https://doi.org/10.1016/j.jeconom.2005.01.027
[18] Bollen, K.A. (1989). *Structural Equations with Latent Variables*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. New York: Wiley. MR0996025 https://doi.org/10.1002/9781118619179
[19] Bournazos, S. and Ravetch, J.V. (2017). Fcγ receptor function and the design of vaccination strategies. *Immunity* **47** 224–233. https://doi.org/10.1016/j.immuni.2017.07.009
[20] Bunea, F., She, Y. and Wegkamp, M.H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39** 1282–1309. MR2816355 https://doi.org/10.1214/11-AOS876
[21] Bunea, F., Strimas-Mackey, S. and Wegkamp, M. (2020). Interpolating Predictors in High-Dimensional Factor Regression. arXiv:2002.02525.
[22] Bunea, F. and Xiao, L. (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. *Bernoulli* **21** 1200–1230. MR3338661 https://doi.org/10.3150/14-BEJ602
[23] Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** 1281–1304. MR0736050 https://doi.org/10.2307/1912275
[24] Chandrasekaran, V., Sanghavi, S., Parrilo, P.A. and Willsky, A.S. (2009). Sparse and low-rank matrix decompositions. *IFAC Proc. Vol.* **42** 1493–1498. 15th IFAC Symposium on System Identification. https://doi.org/10.3182/20090706-3-FR-2004.00249

[25] Connor, G. and Korajczyk, R. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *J. Financ. Econ.* **15** 373–394.
[26] Fan, J., Liao, Y. and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39** 3320–3356. MR3012410 https://doi.org/10.1214/11-AOS944
[27] Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. MR3091653 https://doi.org/10.1111/rssb.12016
[28] Fan, J., Xue, L. and Yao, J. (2017). Sufficient forecasting using factor models. *J. Econometrics* **201** 292–306. MR3717565 https://doi.org/10.1016/j.jeconom.2017.08.009
[29] Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Rev. Econ. Stat.* **82** 540–554.
[30] Forni, M. and Reichlin, L. (1996). Dynamic common factors in large cross-sections. *Empir. Econ.* **21** 27–42.
[31] Giraud, C. (2015). *Introduction to High-Dimensional Statistics*. *Monographs on Statistics and Applied Probability* **139**. Boca Raton, FL: CRC Press. MR3307991
[32] Hahn, P.R., Carvalho, C.M. and Mukherjee, S. (2013). Partial factor modeling: Predictor-dependent shrinkage for linear regression. *J. Amer. Statist. Assoc.* **108** 999–1008. MR3174679 https://doi.org/10.1080/01621459.2013.779843
[33] Jöreskog, K.G. (1970). A general method for analysis of covariance structures. *Biometrika* **57** 239–251. MR0269024 https://doi.org/10.2307/2334833
[34] Jöreskog, K.G. (1970). A general method for estimating a linear structural equation system. *ETS Research Bulletin Series* **1970** i–41. https://doi.org/10.1002/j.2333-8504.1970.tb00783.x
[35] Kelly, B. and Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *J. Econometrics* **186** 294–316. MR3343788 https://doi.org/10.1016/j.jeconom.2015.02.011
[36] Koopmans, T.C. and Reiersøl, O. (1950). The identification of structural characteristics. *Ann. Math. Stat.* **21** 165–181. MR0039967 https://doi.org/10.1214/aoms/1177729837
[37] Lawley, D.N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proc. R. Soc. Edinb.* **60** 64–82. MR0002754
[38] Lawley, D.N. and Maxwell, A.E. (1971). *Factor Analysis as a Statistical Method*, 2nd ed. New York: American Elsevier Publishing Co., Inc. MR0343471
[39] McDonald, R.P. (1999). *Test Theory*: *A Unified Treatment*. London: Taylor & Francis.
[40] Nimmerjahn, F. and Ravetch, J.V. (2007). Fcγ receptors as regulators of immunity. *Advances in Immunology* **96** 179–204. Academic Press. https://doi.org/10.1016/S0065-2776(07)96005-8
[41] Onatski, A. (2009). Testing hypotheses about the numbers of factors in large factor models. *Econometrica* **77** 1447–1479. MR2561070 https://doi.org/10.3982/ECTA6964
[42] Reiß, M. and Wahl, M. (2016). Non-asymptotic upper bounds for the reconstruction error of PCA.
[43] Stock, J.H. and Watson, M.W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179. MR1951271 https://doi.org/10.1198/016214502388618960
[44] Stock, J.H. and Watson, M.W. (2002). Macroeconomic forecasting using diffusion indexes. *J. Bus. Econom. Statist.* **20** 147–162. MR1963257 https://doi.org/10.1198/073500102317351921
[45] Thurstone, L.L. (1931). Multiple factor analysis. *Psychol. Rev.* **38** 406–427.
[46] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. MR2724359 https://doi.org/10.1007/b13794
[47] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge: Cambridge Univ. Press. MR2963170
[48] Wainwright, M.J. (2019). *High-Dimensional Statistics*: *A Non-asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge: Cambridge Univ. Press. MR3967104 https://doi.org/10.1017/9781108627771
[49] Wegkamp, M. and Zhao, Y. (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli* **22** 1184–1226. MR3449812 https://doi.org/10.3150/14-BEJ690
[50] Yalcin, I. and Amemiya, Y. (2001). Nonlinear factor analysis as a statistical method. *Statist. Sci.* **16** 275–294. MR1874155 https://doi.org/10.1214/ss/1009213729

Received April 2020 and revised March 2021