Representation Learning in Linear Factor Models*

José Luis Montiel Olea† and Amilcar Velez‡

August 21th, 2020

Abstract

Modern data sources such as videos, images, or text typically require some form of manual preprocessing prior to their use as input in statistical models. A promise of representation learning—an active area of research in machine learning—is that algorithms will, one day, learn to extract the most useful information from these data sets, thus replacing manual feature engineering. This paper uses a simple Gaussian Linear Factor Model to analyze some recent theoretical developments in the representation learning literature. We start by looking for *sufficient representations*, which are defined as functions of the covariates that, upon conditioning, render the outcomes and covariates independent. The conditional mean of the outcome given covariates, the conditional mean of the latent factors given covariates, and the weighted least squares estimator of the latent factors are sufficient representations. They are also *asymptotically invariant*; which means that the dependence of the representations on the error term in the covariates' regression model vanishes as its dimension grows to infinity. We also use a decision-theoretic approach to understand the extent to which these representations are useful for solving a downstream task. We show that in the Gaussian Linear Factor model the conditional mean of the outcome given covariates can be used to solve *any* task efficiently, not only prediction. We discuss generalizations of our results to a model where a scalar outcome variable, conditional on the factors, has a distribution in the exponential family parameterized by a possibly deep neural network transformation of the factors.

Keywords: Linear Factor Models, Machine Learning, Representation Learning, Neural Networks, Downstream Tasks.

*We would like to thank David Blei for a series of exciting lectures that incited our interest in this topic. We would also like to thank Elliot Oblander for very helpful comments and suggestions. All errors are our own.

†Columbia University, Department of Economics.

‡Northwestern University, Department of Economics.

1 Introduction

Representation Learning—extracting useful information from data to train an algorithm—is an active area of research in machine learning; see Bengio et al. (2013) for a highly cited review. An important promise in this literature is the construction of algorithms that are less dependent on feature engineering and specific domain knowledge—thus reducing the costs of preprocessing data prior to its use as input.

This paper studies *representations* in the context of a Gaussian linear factor model, where a scalar response variable, y_i, and vector-valued covariates, $x_i \in \mathbb{R}^k$, are assumed to be both linear functions of normally distributed errors and latent factors of lower dimension ($z_i \in \mathbb{R}^d$, $d < k$). The motivation for the exercise is to provide a primer on representation learning for econometricians, by connecting the recent information-theoretic definitions of representations and its properties (in particular, the recent framework of Achille and Soatto (2018)) with standard statistical terminology and results.

Factor models (Lawley and Maxwell, 1962, 1973) provide a natural laboratory for exploring representation learning, as the unobserved factors are, in some sense, a systematic and useful lower-dimensional representation of the observed data. The main results in the paper are as follows.

Following the literature, we define a representation z_i^* to be a possibly stochastic function of x_i satisfying $z_i^* \perp y_i | x_i$. Say the representation is *sufficient* if conditioning on it renders the response variable and the covariates independent; i.e., $y_i \perp x_i | z_i^*$. The paper starts by showing that the *conditional mean of y_i given x_i*, the *conditional mean of z_i given x_i*, and any orthogonal rotation of the usual *weighted least squares estimator of z_i*—treating the factor loadings as known and using only the factor model for the covariates x_i—are sufficient representations. These representations are all (non-stochastic) linear transformations of the covariates. They also achieve dimensionality reduction, as they have dimension strictly smaller than k.

Call the error term in the factor model for x_i a *nuisance*. A representation is said to be *invariant* if it is independent of the nuisance. The representations discussed above are, unfortunately, not invariant. In fact, there is no invariant representation in the Gaussian linear factor model. It is herein shown, however, that—as the dimension of the covariates grows large—all of the representations above are *asymptotically invariant* under standard conditions; e.g., the regularity assumptions in Bai and Ng (2006). Asymptotic invariance here means that the *mutual information* between the nuisance and the representation converges to zero as $k \to \infty$.

The definition of asymptotic invariance—and the lack of an invariant representation for a finite k—motivates the search for representations that minimize the mutual information between the nuisance and the representation. Achille and Soatto (2018) refer to such representations as *maximally insensitive* to the nuisance. This paper shows that the conditional mean of y_i given x_i is maximally insensitive among the class of non-stochastic linear sufficient representations. Thus, from the perspective of sufficiency and invariance, learning a good representation in the Gaussian linear model is quite simple. If k is fixed, the conditional mean of y_i given x_i is sufficient and maximally insensitive among sufficient linear representations. If k is large, the weighted least squares estimator of the factors (whose construction does not require information on the distribution of the factors z_i or the outcome) is sufficient and asymptotically invariant.

The representation learning literature has also emphasized the need of constructing representations that are useful for "downstream" tasks. The hope is that one particular representation of the covariates turns out to be useful for different purposes. Prediction and classification problems are two examples of downstream tasks. It is worth mentioning that separating the analysis of the features from the analysis of the outcomes is quite common in the analysis of text data, where, for example, one can use vector embeddings to represent words or sentences, prior to using a text for prediction or classification.

This paper formalizes the notion of a downstream task using a decision-theoretic perspective. We posit an arbitrary loss function (e.g., quadratic loss) involving the outcome variable and an action that is allowed to be contingent on the value of the covariates. We then study the extent to which a representation is useful (or not) for solving a particular task. We formalize this analysis by comparing the smallest expected loss (risk) that would be achieved using all the covariates versus the smallest expected loss that would be achieved by using only the representation.

We show that in the Gaussian factor model the mean of $y_i|x_i$ is—under conditions that we shall spell out clearly—a representation that is useful for *solving any task*. This is an unexpected result, as the conditional mean is typically only optimal for prediction problems under squared loss. We obtain this result by showing that in the Gaussian factor model, the conditional mean of y_i given x_i contains all the information necessary to recover the conditional distribution of $y_i|x_i$. Because the full conditional distribution is encoded in the representation, any task can be solved optimally.

Of course, factor models used in applied work are more complicated than the simple Gaussian linear factor model. Therefore, it is important to understand which of the representations discussed in the paper would still be useful in a more general set-up. To answer this question, we consider a mild departure from the full Gaussian model, by allowing the outcome variable to be a more complicated nonlinear function of the factors, but maintaining the Gaussian linear factor structure for the covariates. To do this, we assume that $y_i|x_i, z_i, \theta$ has a distribution in the exponential family with parameters of the form $\Omega_\theta(z_i)$, where $\Omega_\theta(\cdot)$ denotes a neural network. Importantly, we chose the model for the covariates to remain a Gaussian Linear Factor model. The main assumption here is that—conditional on the factors—the outcome and the covariates are independent.

Our suggested framework relates to several models that have appeared in the literature. First, the nonparametric regression model based on deep neural networks used in Schmidt-Hieber (2020). The difference in our framework is that the regression model is defined in terms of the unobserved factors and it is augmented with a linear factor model for the covariates. Second, the exponential Principal Component Analysis of Collins et al. (2002) that restricts $\Omega_\theta(\cdot)$ to be a linear function of the factors. Third, the Deep Latent Gaussian model of Rezende et al. (2014) where, compared to their general model, we work with only one layer of Gaussian latent variables. Fourth, the Deep Latent Variables models of Mattei and Frellsen (2018), but where we restrict $y_i|z_i$ to have a distribution in the exponential family, as opposed to any arbitrary distribution.

Because the model for the covariates remains linear, the weighted least squares estimator for the factors is still an asymptotically invariant representation. Thus, our interest focuses in understanding the extent to which such a representation can help a decision maker in solving a downstream task. Our main result is that—as k grows large and if we treat the model's parameters as known— this representation can be used to evaluate the expected loss of any action. The key insight is that the expected loss can be computed using the exponential family distribution but assuming that the unobserved factors are actually equal to their estimated value.

The rest of this paper is organized as follows. Section 2 presents the model and main results. Section 3 provides a decision-theoretic definition of a task and shows that the mean of $y_i|x_i$ solves any task. Section 4 discusses the extensions of our main results.

2 Model and Main Results

There is scalar outcome variable y_i, a vector of k covariates x_i, and a vector of d latent features z_i ($d < k$). Consider the linear factor model

$$y_i = \alpha' z_i + u_i, \tag{1}$$

$$x_i = \beta' z_i + v_i, \tag{2}$$

where

$$\begin{pmatrix} u_i \\ v_i \\ z_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & 0 & 0 \\ 0 & \Sigma_v & 0 \\ 0 & 0 & \mathbb{I}_d \end{pmatrix} \right). \tag{3}$$

It is further assumed that Σ_v is diagonal with strictly positive entries, and that $\beta\Sigma_v^{-1}\beta'$ has rank d. The model above parameterizes the joint distribution of (y_i, x_i) by $\theta \equiv (\alpha, \beta, \sigma_u^2, \Sigma_v)$. Equations (1)-(2) can be viewed as a restricted version of the diffusion index forecasting model of Stock and Watson (2002), analyzed in detail in Bai and Ng (2006).

2.1 Sufficient, Invariant Representations

The following definitions of representations are based on Achille and Soatto (2018), but properly adjusted to account for the parametric nature of the factor model.

Definition 1 (Sufficient Representation). We say that z_i^* is a representation of x_i at θ if z_i^* is a function of x_i—possibly stochastic—and

$$\mathbb{P}_\theta(z_i^* | y_i, x_i) = \mathbb{P}_\theta(z_i^* | x_i). \tag{4}$$

The representation is said to be sufficient at θ if the condition

$$y_i \perp x_i | z_i^* \tag{5}$$

holds under $\mathbb{P}_\theta$.

Equation (4) allows for a large class of random variables to serve as representations of x_i. For instance, any function of the form $a + b'x_i + c_i$ where c_i is random vector independent of (u_i, v_i, z_i) is a representation. Not all of these representations, however, are sufficient or satisfy additional properties like the one below.

Definition 2 (Nuisance and Invariance). A random variable n_i is a nuisance at θ if

$$x_i \not\perp n_i \text{ and } y_i \perp n_i$$

under $\mathbb{P}_\theta$. A representation z_i^* is said to be invariant to the nuisance if the *mutual information*

$$I_\theta(z_i^*, n_i) \equiv \mathrm{KL}\left(\mathbb{P}_\theta(z_i^*, n_i) || \mathbb{P}_\theta(z_i^*) \otimes \mathbb{P}_\theta(n_i)\right) \tag{6}$$

equals zero. A representation is *maximally insensitive* to the nuisance n_i—in a class of representations $\mathcal{C}$—if it minimizes (6) among the representations in $\mathcal{C}$. Finally, a representation is said to be *asymptotically invariant* under a sequence of parameters $\{\theta_k\}$—indexed by the dimension of the covariates—if $I_\theta(z_i^*, n_i) \to 0$ as $k \to \infty$.

The definition of nuisance is quite general, and in principle refers to any random variable n_i that affects x_i, but that is independent of y_i. Throughout the rest of the paper we focus on v_i (the error term in the factor model for the covariates x_i) as the nuisance of interest.

2.2 Representations in the Gaussian Linear Factor Model

Consider the following (non-stochastic) linear representations of x_i.

$$\mathbb{E}_\theta[y_i|x_i],\ \mathbb{E}_\theta[z_i|x_i],\ z_i^* \equiv (\beta\Sigma_v^{-1}\beta')^{-1}\beta\Sigma_v^{-1}x_i. \tag{7}$$

The first representation is the conditional mean of y_i given x_i (assuming the the parameter θ is known). The second one is the conditional mean of the factor z_i given x_i, also assuming θ is known.[1] Finally, z_i^* is the weighted least squares estimator of z_i based on equation (2) and assuming β is known (see Anderson (2003), Section 14.7, Equation 1, p. 592).

These representations are infeasible, since the model's parameters are unknown. Let Q denote an arbitrary orthogonal matrix of dimension d.

Proposition 1. In the model given by (1)-(2), $\mathbb{E}_\theta[y_i|x_i]$, $\mathbb{E}_\theta[z_i|x_i]$, and Qz_i^* are sufficient representations of x_i at θ. The mutual information between these representations and the nuisance v_i satisfies

$$I_\theta(\mathbb{E}_\theta[z_i|x_i], v_i) = I_\theta(Qz_i^*; v_i) \geq I_\theta(\mathbb{E}_\theta[y_i|x_i]; v_i) > 0,$$

for any fixed k, where the first inequality is strict only if $d > 1$. These sufficient representations are asymptotically invariant to the nuisance v_i under any sequence of parameters for which $\det(\mathbb{I}_d + (\beta_k\Sigma_{v,k}^{-1}\beta_k')^{-1}) \leq 1 + o(k)$ as $k \to \infty$.

The proposition above establishes a number of results, all of which follow from calculations based on the multivariate normal model.

First, while it is immediate to recognize $\mathbb{E}_\theta[y_i|x_i]$, $\mathbb{E}_\theta[z_i|x_i]$, and Qz_i^* as a representations, it is less evident that such representations are sufficient.

[1] In the Gaussian factor model, both conditional means are linear functions of the covariates.

Consider the case of the weighted least squares estimator of the factors. If Qz_i^* provided a noiseless measure of the factors z_i, sufficiency would be verified by definition (as conditional on the factors, y_i and x_i are independent). The representation Qz_i^* however measures z_i with error, as:

$$Qz_i^* = Qz_i + Q(\beta\Sigma_v^{-1}\beta')^{-1}\beta\Sigma_v^{-1}v_i. \tag{8}$$

The proof of Proposition 1 in the Appendix A.1, verifies that conditioning on Qz_i^* makes y_i and x_i independent. The derivation crucially exploits the Gaussian nature of the factor model.

Second, Proposition 1 also provides a comparison of the representations in terms of mutual information—which is an information-theoretic measure of dependence—with the nuisance v_i. Equation (8) already shows that Qz_i^* and v_i are not independent; and the mutual information formula in Proposition 1 further quantifies the dependence[2]. The relation above shows that the mutual information between Qz_i^* and v_i will equal the mutual information between $\mathbb{E}_\theta[z_i|x_i]$ and v_i. We think this is a somewhat surprising result, as both Qz_i^* and $\mathbb{E}_\theta[z_i|x_i]$ (both of which have dimension d) are typically viewed as legitimate estimators of the latent factors that differ only in their motivation (one of them Frequentist, and the other one Bayesian).

The most interesting message from the comparison of mutual information, however, is that the representation $\mathbb{E}_\theta[y_i|x_i]$ dominates the other in terms of mutual information. It is already a bit surprising that $\mathbb{E}_\theta[y_i|x_i]$ is a sufficient representation (because this conditional mean cannot be viewed as an estimator of the underlying factors). It is even more remarkable that such representation is better in terms of invariance to the nuisance.

Third, Proposition 1 also shows that none of the representations above are invariant to the nuisance. Proposition 1 shows that the mutual information between the representations and v_i converges to zero as the dimension of the covariates goes to infinity. One possible intuition is that, as k grows large, the measurement error in (8) vanishes. The result then follows from the independence between v_i and z_i. To formalize this result we needed to impose some restrictions on the way in which the parameters of the factor model change as k increases. One common assumption in the literature—see Assumption B in Bai and Ng (2006)—is that the factor loadings have a well-defined limit when scaled by the number of covariates; namely

$$k^{-1}\beta_k\Sigma_{v,k}^{-1}\beta_k' \to \Sigma_\beta,$$

where Σ_β is an invertible $d \times d$ matrix. This assumption, which shall be used later, implies

[2] In Appendix A.5, Lemma 2 provides a tractable and close form expression for mutual information.

that

$$\det(\mathbb{I}_d + (\beta_k \Sigma_{v,k}^{-1} \beta_k')^{-1}) \to 1.$$

2.3 Maximally Insensitive Representations

The representation $\mathbb{E}_\theta[y_i|x_i]$ is already appealing because of its sufficient and it has the lowest possible dimension. Also, as $k \to \infty$ this representation is asymptotically invariant. The only limitation is that it is not invariant to the nuisance x_i for a fixed k. Is it possible to find a better representation? The following proposition shows this is not possible, with some qualifications.

Proposition 2: In the model given by (1)-(2), the representation $\mathbb{E}_\theta[y_i|x_i]$ is maximally insensitive to the nuisance v_i among the class of all non-stochastic, linear, and sufficient representations.

The proof of Proposition 2 is constructive, for details see Appendix A.2. The main idea is that for any non-stochastic, linear, sufficient representation of dimension $p \geq 1$, we can find a representation of the same dimension and with the same mutual information with respect to the nuisance, but that explicitly contains $\mathbb{E}_\theta[y_i|x_i]$ as one of its entries. Intuitively, this implies that any non-stochastic, linear, sufficient representation is—in a sense—capturing other features of the covariates that are not $\mathbb{E}_\theta[y_i|x_i]$. As a consequence of the chain rule of conditional mutual information, we can show that the mutual information with respect to nuisance v_i of $\mathbb{E}_\theta[y_i|x_i]$ has to be equal or smaller.

An implication of our result is that in the case of non-stochastic, linear, sufficient representation of dimension one, all of them are proportional to $\mathbb{E}_\theta[y_i|x_i]$ and thus have the same mutual information with respect to v_i. This means that all non-stochastic, linear, sufficient, representations of dimension 1 are maximally insensitive to the nuisance v_i.

A representation that is maximally insensitive to the nuisance v_i in the class of sufficient representation is useful for two reasons. First, sufficient representations and the covariates x_i have the same mutual information with the outcome variable y_i. Secondly, the nuisance v_i affects only the covariates but not the outcome variable, thus a maximally insensitive representation is minimizing the effect of the nuisance in the representation.

3 Downstream Tasks

Intuitively, a good representation ought to be useful in *downstream* tasks, such as prediction. It is therefore important to explore the extent to which the representations discussed in the previous section are useful for solving decision problems that involve (y_i, x_i) such as prediction. In this section, we provide a decision-theoretic definition of a task and show that in the model (1)-(2) the conditional mean of y_i given x_i solves any task efficiently, in a sense we make precise. More generally, in Section 4 we provide an algorithm of how the weighted least-squares estimator of the factors can be used to *asymptotically* solve any task when $k \to \infty$.

Preliminaries: Let $\mathbb{P}_\theta$ denote a joint distribution over $(y_i, x_i) \in \mathcal{Y} \times \mathcal{X}$. Let $\mathcal{A}$ denote some action space. We define a loss function in the usual way: $\mathcal{L} : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$.[3] We refer to any (measurable) function $a : \mathcal{X} \to \mathcal{A}$ as an algorithm. The expected loss of an algorithm $a(\cdot)$ at θ is referred to as the *risk* of $a(\cdot)$ at θ. This is, we define the risk function $R(\cdot, \cdot)$ as

$$R(a(\cdot), \theta) \equiv \mathbb{E}_\theta[\mathcal{L}(y, a(x))]. \tag{9}$$

A *downstream task* (or simply a *task*) is a tuple:

$$\mathcal{T} \equiv (\mathcal{L}, \mathcal{A}, \mathbb{P}_\theta). \tag{10}$$

An algorithm $a(\cdot)$ is *optimal* for task $\mathcal{T}$ at θ if

$$R(a(\cdot), \theta) \leq R(a'(\cdot), \theta), \tag{11}$$

for any other algorithm $a'(\cdot)$.

Definition 3: A representation z^* *solves task* $\mathcal{T}$ if there is an optimal algorithm a^* that only depends on x only through the representation.

That is, a representation z^* solves a task $\mathcal{T}$ if we can find an algorithm $a(\cdot)$ that only uses z^* as input and that has smaller or equal risk than any other algorithm. We further say that a representation z^* solves task $\mathcal{T}$ *efficiently* at θ if there is no other representation of lower dimension that solves the task $\mathcal{T}$ at θ.

[3]Examples of loss functions are quadratic loss, $\mathcal{L}(y, a) = (y - a)^2$, or the check function, $\mathcal{L}(y, a) = y(\tau - \mathbf{1}\{y < 0\})$.

Proposition 3: Suppose $\mathbb{V}_\theta(y)$ and $\mathbb{E}_\theta((\mathbb{E}_\theta[y_i|x_i])^2)$ are known. Then, in the Gaussian Linear Factor Model model given by (1)-(2) the representation $\mathbb{E}_\theta[y_i|x_i]$ solves any task $\mathcal{T}$ efficiently at θ.

The result above is, to some degree, unexpected. It is well-known that $\mathbb{E}_\theta[y_i|x_i]$ is the optimal predictor under quadratic loss. However, the result in Proposition 3 says that for *any* loss, it is possible to dispense with the covariates, retain the representation $\mathbb{E}_\theta[y_i|x_i]$ and still achieve the smallest possible risk at θ.

The idea behind the proof is quite simple and the details are presented in Appendix A.3. In the Gaussian linear factor model the conditional distribution of $y_i|x_i$ is characterized by its first two moments. Our representation is the first moment, and the variance of $y_i|x_i$ is:

$$\mathbb{V}(y_i \mid x_i) = \mathbb{V}(y_i) - \mathbb{E}_\theta[\mathbb{E}_\theta[y_i|x_i]^2].$$

This means that the one-dimensional representation $\mathbb{E}_\theta[y_i|x_i]$ has all the information about the conditional distribution of $y_i|x_i$ (under the assumptions we have made).

4 Extensions

The main results of this paper have been derived under strong assumptions (a Gaussian linear factor model for covariates and response variable). This section discusses a generalization of our main results by allowing a different model for the outcome variable. In addition, we propose an algorithm to asymptotically solve a downstream task using an asymptotically invariant representation.

4.1 A more general model for the outcome variable

Just as before, suppose there is scalar outcome variable y_i with support $\mathcal{Y}$, a vector of k covariates x_i, and a vector of d latent features z_i $(d < k)$. Consider the model

$$\begin{aligned} y_i \mid x_i, z_i, \alpha, \sigma_u &\sim f(y_i \mid z_i, \alpha, \sigma_u), && (12) \\ x_i &= \beta' z_i + v_i, && (13) \end{aligned}$$

where $f(y_i|z_i, \alpha, \sigma_u)$ denotes a density of the form,

$$f(y_i \mid z_i, \alpha, \sigma_u) \equiv h(y, \sigma_u) \exp([\Omega_\alpha(z_i) y_i - \Psi(\Omega_\alpha(z_i))]/a(\sigma_u)). \quad (14)$$

In our notation $h(\cdot, \phi)$ is a real-valued function parametrized by σ_u defined on $\mathcal{Y}$, $a(\cdot)$ is a positive function of σ_u, and $\Psi(\cdot)$ is a smooth function (usually referred to as the log-partition function) defined on all the real line. The density in (14) is a slight modification of the Generalized Linear Models described in McCullagh and Nelder (1989, Equation 2.4) where $\Omega(z_i)$ now plays a role analogous to the natural parameter of the exponential family.[4] Throughout this section we impose the following assumption:

Assumption 1 : $\Omega_\alpha : \mathbb{R}^d \to \mathbb{R}$ is a L_α-Lipschitz function,

$$|\Omega_\alpha(z_1) - \Omega_\alpha(z_2)| \leq L_\alpha |z_1 - z_2|.$$

We maintain the assumption

$$\begin{pmatrix} v_i \\ z_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_v & 0 \\ 0 & \mathbb{I}_d \end{pmatrix} \right) \text{ and } y_i \perp x_i \mid z_i, \tag{15}$$

where Σ_v is diagonal with strictly positive entries, and $\beta \Sigma_v^{-1} \beta'$ has rank d. Once again, the model above parameterizes the joint distribution of (y_i, x_i) by $\theta \equiv (\alpha, \beta, \sigma_u^2, \Sigma_v)$. Throughout this section, we shall also assume θ is known.

We now discuss the relation of (12)-(13) to related models that have appeared in the literature.

1. *Nonlinear Regression model with Neural Networks:* Schmidt-Hieber (2020) has recently analyzed a model of the form

$$y_i = \Omega_\alpha(z_i) + \epsilon_i, \quad \epsilon_i \perp z_i, \epsilon_i \sim \mathcal{N}(0, \sigma_u^2),$$

   where z_i is observed and $\Omega_\alpha(z_i)$ is a deep neural network. Schmidt-Hieber (2020) assumes (y_i, z_i) are observed. In contrast, we assume that z_i is a latent factor, that $\epsilon_i \perp (x_i, z_i)$, and that there is a linear factor model for x_i.

2. *Exponential Family PCA:* If we assume that $\Omega_\alpha(z_i) = \alpha' z_i$ then our model becomes the exponential family principal component analysis model in Collins et al. (2002). Our model assumes that if the latent factors were known, the covariates x_i would have no effect on the distribution of y_i. If we maintain the linear factor model in (13), then the

[4] Normal, Logistic, and Poisson models can be captured with conditional densities of the form (14). See Table 2.1 p. 29 of McCullagh and Nelder (1989)

only use of the covariates is their ability to estimate z_i.

3. *Deep Latent Gaussian Model:* The model (12)-(13) can be described as a particular case of the generative model described in the highly cited work of Rezende et al. (2014). Compared to their general model, we assume there is only one hidden layer of latent variables.

4. *Deep Latent Variable Model:* The outcome model (12) is also a special case of the model in Mattei and Frellsen (2018) for two reasons. First, our outcome variable is scalar. Second, our model uses an exponential family density.

4.2 Computing Expected Loss using the Representation

Characterizing sufficient and maximally insensitive representations in this model is more challenging. However, there is a sense in which the weighted least squares estimator of the factors, z_i^*, is still a useful representation. As we mentioned in Proposition 1, this representation is asymptotically invariant to the nuisance in the factor model for the covariates.

We would like to argue that the representation can be used to simplify the computation of the expected loss of a particular action. To see this, note that for any loss function $\mathcal{L}(y, a)$ the optimal algorithm prescribes the action that minimizes $\mathbb{E}_\theta[\mathcal{L}(y, a)|x]$. The conditional density of Y given X is a mixture distribution:

$$\begin{aligned} f_\theta(y|x) &= \int f(y|z, x, \alpha, \sigma_u) dF_\theta(z|x), \\ &= \int f(y|z, \alpha, \sigma_u) dF_\theta(z|x). \end{aligned}$$

We can show that, as $k \to \infty$, the posterior distribution of $z|x$, $F_\theta(z|x)$, concentrates around the weighted least squares estimator of the factor, z^*, which is a linear function of x. Thus,

$$f_\theta(y|x) \approx f(y|z = z^*, \alpha, \sigma_u).$$

In this case, the best action can be found by computing expected loss according to a model in which y_i has a distribution as in (14) but evaluated at z_i^*. This suggests that we can use a representation z_i^* to compare different actions when solving downstream tasks, provided the dimension of x_i is large. This can be done defining an auxiliary outcome variable y_i^*,

$$y_i^* \mid z_i^*, \alpha, \sigma_u \quad \sim \quad f(y_i^* \mid z_i^*, \alpha, \sigma_u), \tag{16}$$

where this auxiliary outcome variable formalize the discussion described above and not depend on latent factors, z_i, but observable representations, z_i^*.

We claim that, under some regularity assumptions, we can evaluate the performance of different actions in the downstream task by using (16) as $k \to \infty$.

We first restrict the set of downstream tasks that we are interested in, by restricting the loss functions that we are working with:

Assumption 2 : The loss function $\mathcal{L}(\cdot, a) : \mathcal{Y} \to [0, +\infty)$ is dominated by a quadratic polynomial,

$$\mathcal{L}(y, a) \leq c_1 + c_2\, y^2,$$

where $c_1, c_2 > 0$ are constants that could be functions of a.

This assumption allows for quadratic, check, and 0-1 loss.[5] Thus, we are interested in tasks like prediction, quantile estimation, and classification.

We further require some control on the moments of $y|x$. Because of (14), all the moments of $y|z$ will exist. The distribution of $y|x$, however, is a mixture distribution of $y|z$ and z. Consequently, we need to be able to integrate over the moments of $y|z$. We achieve this by requiring that the tails of $y|z$ have polynomial decline and is function of the parameter $\Omega_\alpha(z)$:

Assumption 3 : The exponential family verify a regularity condition,

$$\mathbb{P}_\theta[|y| \geq t \mid\ z, \alpha, \sigma_u] \leq t^{-4}(c_3 + c_4 \exp(c_5|\Omega_\alpha(z)|)),$$

for any z and $t > 0$, where c_3, c_4, c_5 are non-negative constants.[6]

Proposition 4: Suppose Assumptions 1-3 hold. Consider evaluating the expected loss of an action a given some value of the covariates x. Suppose that as $k \to \infty$ the parameters of the model and the covariates satisfy

$$\beta\Sigma_v^{-1}\beta'/k \to \underbrace{\Sigma_\beta}_{d\times d} \text{ and } \beta\Sigma_v^{-1}x/k \to \underbrace{\mu_\beta}_{d\times 1},$$

[5] The quadratic function, $(y-a)^2 \leq 2y^2 + 2a^2$, and the check function, $y(a - 1_{y<0}) \leq 0.5 \max\{a, 1 - a\}y^2 + 0.5 \max\{a, 1-a\}$, satisfies Assumption 2.

[6] This assumption is satisfied for Normal, Logistic and Poisson models, for example.

where Σ_β is invertible. Then,

$$\underbrace{\int \mathcal{L}(y,a) f(y \mid x_k, \alpha, \sigma_u) dy}_{\mathbb{E}_\theta[\mathcal{L}(y,a)|x]} - \underbrace{\int \mathcal{L}(y^*,a) f(y^* \mid z_i^*(x_k), \alpha, \sigma_u) dy^*}_{\text{expected loss for the auxiliary model}} \to 0 \text{ as } k \to \infty. \tag{17}$$

The key insight of this proposition is that the expected loss can be computed using the exponential family distribution but assuming that the unobserved factors are actually equal to their estimated value, which are given by the representation. The main idea is that, by Assumption 2, is sufficient to prove (17) for a quadratic loss function. To conclude the proof, we use the proposition assumptions to verify that p.d.f converges point-wise and Assumption 1 and 3 to guarantee that we can applied a variation of Dominated Convergence Theorem. Details are presented in Appendix A.4.

The proposition above has been derived for a fixed action a and known parameters θ. However, it suggests a strategy for solving downstream tasks when the dimension of x_i is large.

Consider the following approach:

1. Estimate β from the linear factor model for x_i using principal components.

2. Compute the feasible version of z_i^*, denoted $\widehat{z}_i^* \equiv (\widehat{\beta}\widehat{\Sigma}_v\widehat{\beta}')^{-1}\widehat{\beta}\widehat{\Sigma}_v x_i$.

3. Treat $\widehat{z}_i^*$ as z_i and estimate the parameter α and σ_u in the exponential family model.

4. Pick the action that minimizes expected loss according to

$$y_i^* \mid \hat{z}_i^*, \hat{\alpha}, \hat{\sigma}_u \;\sim\; f(y_i^* \mid \hat{z}_i^*, \hat{\alpha}, \hat{\sigma}_u), \tag{18}$$

   In the case of prediction, predict using $\Psi'(\Omega_{\hat{\alpha}}(\hat{z}_i^*))$

These four steps seem to generalize the forecasting algorithm of Stock and Watson (2002) and the 'unsupervised pre-training' strategy described in Chapter 15 of Goodfellow et al. (2016). We think that it is possible to use standard results in the asymptotic analysis of factor models to formalize the validity of this strategy, provided we make high-level assumptions about our ability to consistently estimate the parameters α and σ_u of the function $\Omega_\alpha(\cdot)$) (which could be a neural network). The derivation of these results would need to consider asymptotics where both the N (the number of training examples) and k (the dimension of the covariate vector) diverge to infinity.

A Proofs of Main Results

A.1 Proof of Proposition 1

The proof of this proposition has three parts as was discussed in the main text. First, we will prove that $\mathbb{E}_\theta[y_i \mid x_i]$, $\mathbb{E}_\theta[z_i \mid x_i]$ and Qz_i^* are sufficient representations. Second, we will compute the mutual information with respect to the nuisance v_i. And third, we will prove that these representation are asymptotically invariant.

Part 1: Algebra on multivariate normal distribution shows

$$\mathbb{E}_\theta[y_i|x_i] = \alpha'\beta\Sigma_x^{-1}x_i \text{ and } \mathbb{E}_\theta[z_i|x_i] = \beta\Sigma_x^{-1}x_i,$$

where $\Sigma_x \equiv \Sigma_v + \beta'\beta$. Define by $A_1 \equiv \alpha'\beta\Sigma_x^{-1}$, $A_2 \equiv \beta\Sigma_x^{-1}$ and $A_3 \equiv (\beta\Sigma_v^{-1}\beta')^{-1}\beta\Sigma_v^{-1}$. This means that we can write the three representations as deterministic linear representations of x:

$$\mathbb{E}_\theta[y_i \mid x_i] = A_1 x, \quad \mathbb{E}_\theta[z_i \mid x_i] = A_2 x \quad \text{and } z_i^* = A_3 x$$

By Lemma 1 in Appendix A.5, we conclude that these three representation are sufficient representations since we can verify that inverse matrix of $A_j\Sigma_x A_j'$ exists and

$$\Sigma_x A_j'(A_j\Sigma_x A_j')^{-1}A_j\beta'\alpha = \beta'\alpha,$$

for $j = 1, 2, 3$.

Part 2: By Lemma 2 in Appendix A.5, we knows that for any $\hat{z}_i \equiv Ax_i$ such that the inverse of matrix $(A\Sigma_x A')^{-1}$ and $A\beta'\beta A'$ are well-defined, then the mutual information between $\hat{z}_i$ and v_i is

$$I_\theta(\hat{z}_i; v) = \quad \frac{1}{2}\ln\Big(\frac{\det(A\Sigma_x A')}{\det(A\beta'\beta A')}\Big).$$

By part 1, we know that the representations in this proposition are deterministic and linear. Also we can verify that $A_j\beta'\beta A_j'$ has inverse for $j = 1, 2, 3$. Then, algebra shows

$$I_\theta(\mathbb{E}_\theta[y_i \mid x_i]; v_i) = \frac{1}{2}\ln\left(\frac{\alpha'(\mathbb{I}_d - (\mathbb{I}_d + \Psi)^{-1})\alpha}{\alpha'(\mathbb{I}_d - (\mathbb{I}_d + \Psi)^{-1})^2\alpha}\right),$$
$$I_\theta(\mathbb{E}_\theta[z_i \mid x_i]; v_i) = \frac{1}{2}\ln\left(\frac{1}{\det(\mathbb{I}_d - (\mathbb{I}_d + \Psi)^{-1})}\right),$$
$$I_\theta(z_i^*; v_i) = \frac{1}{2}\ln\left(\frac{\det(\mathbb{I}_d + \Psi)}{\det(\Psi)}\right),$$

where $\Psi = \beta\Sigma_v^{-1}\beta'$.

To conclude the comparison of the representations in terms of mutual information with the nuisance v_i, observes that $I(\mathbb{E}_\theta[z_i \mid x_i]; v_i) = I_\theta(\hat{z}_i; v_i)$ is equivalent to prove

$$\frac{\det(\mathbb{I}_d + \Psi)}{\det(\Psi)} = \frac{1}{\det(\mathbb{I}_d - (\mathbb{I}_d + \Psi)^{-1})},$$

which is true by algebra manipulation.

To prove $I(z_i^*; v_i) \geq I_\theta(\mathbb{E}_\theta[y_i \,|\, x_i]; v_i)$, denote by $\lambda_1 \leq ... \leq \lambda_d$ the eigenvalues of $\mathbb{I}_d - (\mathbb{I}_d + \Psi)^{-1}$ and by $w_1, ..., w_d$ the associated eigenvectors. An important observation is that all these eigenvalues are lower than one and we can use them compute $I(z_i^*; v_i)$ and $I(\mathbb{E}_\theta[y_i \mid x_i]; v_i)$. In particular, we have

$$\frac{1}{\det(\mathbb{I}_d - (\mathbb{I}_d + \Psi)^{-1})} = \frac{1}{\lambda_1 ... \lambda_d},$$

and if we write $\alpha = \sum_{m=1}^d a_m w_m$ using the eigenvectors w_i, we have

$$\frac{\alpha'(\mathbb{I}_d - (\mathbb{I}_d + \Psi)^{-1})\alpha}{\alpha'(\mathbb{I}_d - (\mathbb{I}_d + \Psi)^{-1})^2\alpha} = \frac{\sum_{m=1}^d a_m^2 \lambda_m}{\sum_{m=1}^d a_m^2 \lambda_m^2}$$

This implies that $I(z_2^*; v) \geq I_\theta(z_3^*; v)$ since λ's are lower than one, where equality only holds if $d = 1$.

Part 3: By part 2, it will be sufficient to prove that

$$\lim_{k \to \infty} I_\theta(z_i^*; v_i) = 0,$$

to guarantee that the three representations are asymptotically invariant. By part 2, we have

$$I_\theta(z_i^*; v_i) = \frac{1}{2} \ln \left(\frac{\det(\mathbb{I}_d + \Psi)}{\det(\Psi)} \right) = \frac{1}{2} \ln \left(\det(\mathbb{I}_d + \Psi^{-1}) \right),$$

and by assumption $\det(\mathbb{I}_d + \Psi^{-1}) \to 1$ as $k \to \infty$. This concludes our proof.

A.2 Proof of Proposition 2

Case 1: $p > 1$. Suppose $\hat{z}_i = Ax_i$ is a deterministic linear sufficient representation of dimension p, where $A \in \mathbb{R}^{p \times k}$ and $p < k$. We want to prove

$$I_\theta(\hat{z}_i; v_i) \geq I_\theta(E_\theta[y_i | x_i]; v_i)$$

where $E_\theta[y_i|x_i] = \alpha'\beta\Sigma_x^{-1}x$ is the conditional mean of y_i given x_i and $\Sigma_x \equiv \Sigma_v + \beta'\beta$. By Proposition 1, we know that $E_\theta[y_i|x_i]$ is also a deterministic linear sufficient representation. Define $A_3 \equiv \alpha'\beta\Sigma_x^{-1}$.

By Lemma 1 in Appendix A.5, we know that

$$\Sigma_x A'(A\Sigma_x A')^{-1}A\beta'\alpha = \beta'\alpha$$

and

$$\Sigma_x A_3'(A_3\Sigma_x A_3')^{-1}A_3\beta'\alpha = \beta'\alpha.$$

These two equations imply

$$A' \underbrace{(A\Sigma_x A')^{-1}A\beta'\alpha}_{p\times 1} = A_3' \underbrace{(A_3\Sigma_x A_3')^{-1}A_3\beta'\alpha}_{1\times 1},$$

and this is equivalent to

$$\underbrace{Q_0}_{1\times p}\underbrace{A}_{p\times k} = \underbrace{A_3}_{1\times k}, \tag{19}$$

where

$$Q_0 \equiv \frac{(A\Sigma_x A')^{-1}A\beta'\alpha}{(A_3\Sigma_x A_3')^{-1}A_3\beta'\alpha}.$$

Thus, we can construct a $(p-1)\times 1$ matrix B such that

$$Q \equiv \begin{pmatrix} Q_0 \\ B \end{pmatrix}$$

is an invertible matrix. Define the new representation of dimension p

$$\tilde{z}_i \equiv \underbrace{Q}_{p\times p}\underbrace{Ax_i}_{p\times 1}.$$

The new representation is a linear transformation of $\hat{z}_i$. Equation (19) implies

$$\tilde{z}_i = \begin{pmatrix} Q_0 A x_i \\ \underbrace{BAx_i}_{(p-1)\times 1} \end{pmatrix} = \begin{pmatrix} E_\theta[y_i|x_i] \\ BAx_i \end{pmatrix}.$$

Thus, the first entry of the new representation is the conditional mean of y_i given x_i. By

Lemma 2 in Appendix A.5, we have

$$I_\theta(\tilde{z}_i; v_i) = \quad \frac{1}{2}\ln\Big(\frac{\det(QA\Sigma_x A'Q')}{\det(QA\beta'\beta A'Q')}\Big).$$

Thus, algebra shows that

$$\begin{aligned}
I_\theta(\tilde{z}_i; v_i) = &\quad \frac{1}{2}\ln\Big(\frac{\det(Q)\det(A\Sigma_x A')\det(Q')}{\det(Q)\det(A\beta'\beta A')\det(Q')}\Big),\\
&\text{(as } \det(MN) = \det(M)\det(N))\\
= &\quad \frac{1}{2}\ln\Big(\frac{\det(A\Sigma_x A')}{\det(A\beta'\beta A')}\Big),\\
= &\quad I_\theta(\hat{z}_i; v_i).
\end{aligned}$$

Thus, we have shown that the mutual information between $\tilde{z}$ and the nuisance v is the same as the mutual information between $\hat{z}_i$ and v_i. Note that $\hat{z}_i$ was an arbitrary sufficient representation, and we obtained $\tilde{z}_i$ from $\hat{z}_i$ by transforming the latter to have the conditional mean of y given x in the first coordinate.

Now, we will prove that $I(\tilde{z}_i; v_i) \geq I_\theta(E_\theta[y_i|x_i]; v_i)$. Since $\tilde{z}_i' = [E_\theta[y_i|x_i], x_i'A'B']'$, by chain rule on conditional mutual information we have

$$I_\theta(\tilde{z}; v_i) = I_\theta(E_\theta[y_i|x_i], BAx; v_i) = I_\theta(E_\theta[y_i|x_i]; v_i) + \underbrace{I(BAx; v \mid E_\theta[y_i|x_i])}_{\geq 0} \geq I_\theta(E_\theta[y_i|x_i]; v_i).$$

Then, we conclude the conditional mean of y_i given x_i is maximally insensitive to v_i (among all linear deterministic representations); i.e.,

$$I_\theta(\hat{z}_i; v_i) = I_\theta(\tilde{z}_i; v_i) \geq I_\theta(E_\theta[y_i|x_i]; v_i).$$

Case 2: $p = 1$**.** By Lemma 1 in Appendix A.5, we have

$$\Sigma_x A' \underbrace{(A\Sigma_x A')^{-1} A\beta'\alpha}_{1\times 1} = \beta'\alpha$$

This implies

$$\hat{z}_i = Ax_i = \gamma\alpha'\beta\Sigma_x^{-1}x_i = \gamma E_\theta[y_i|x_i]$$

where $\gamma = (A\Sigma_x A')^{-1}A\beta'\alpha \in \mathbb{R} - \{0\}$. It follows that $I(\hat{z}_i, v_i) = I_\theta(E_\theta[y_i|x_i], v_i)$. Thus, deterministic linear sufficient representation of dimension one are also maximally invariance.

A.3 Proof of Proposition 3

The proof of this proposition has three main observations. First, posterior distribution $y_i|x_i$ is a Gaussian distribution characterized by its two moments (mean and variance), under our model (1)-(2). Second, the assumptions of this proposition implies that we known the variances of $y_i|x_i$,

$$\mathbb{V}(y_i \mid x_i) = \mathbb{V}(y_i) - \mathbb{E}_\theta[\mathbb{E}_\theta[y_i|x_i]^2].$$

Finally, the posterior distribution $y_i|x_i$ is parametrized by the posterior mean, which is $\mathbb{E}_\theta[y_i|x_i]$. These three observations implies that we can solve task $\mathcal{T}$ using only the representation $\mathbb{E}_\theta[y_i|x_i]$, which also has dimension one. This conclude the proof of this proposition.

A.4 Proof of Proposition 4

The conditional distribution of the outcome variable to the covariates, $y_i \mid x_{i,k} \sim f(y_i \mid x_{i,k})$, is expressed as

$$f(y \mid x) \equiv \int f(y \mid x, z)\phi(z \mid \mu_k(x), \Sigma_k(x))\, dz,$$

where $\mu_k(x) \equiv \beta\Sigma_x^{-1}x$ and $\Sigma_k(x) \equiv \mathbb{I}_d - \beta\Sigma_x^{-1}\beta'$ are the posterior mean and variances. Since $y_i \perp x_i \mid z_i$, we can write $f(y \mid z, \alpha, \sigma_u)$ instead of $f(y \mid x, z)$. This give us

$$f(y \mid x) = \int f(y \mid z, \alpha, \sigma_u)\phi(z \mid \mu_k(x), \Sigma_k(x))\, dz.$$

We break the proof of in two main parts. The first part proves that

$$\int \mathcal{L}(y, a) \int f(y \mid z, \alpha, \sigma_u)\phi(z \mid \mu_k(x), \Sigma_k(x))\, dzdy \tag{20}$$

converges to

$$\int \mathcal{L}(y, a) f(y \mid z_0, \alpha, \sigma_u) dy, \tag{21}$$

as $k \to \infty$, and where $z_0 \equiv \Sigma_\beta^{-1}\mu_\beta$. In the second part, we prove that

$$\int \mathcal{L}(y, a) f(y \mid z_i^*(x_{i,k}), \alpha, \sigma_u) dy \tag{22}$$

is also converging to equation (21). These two main parts implies (17).

Proof of Part 1 : In equation (20) all the terms in the integrals are positive. By Tonelli's Theorem we can change the order of the integrals. This implies that equation (20) is equal to

$$\int\int \mathcal{L}(y,a) f(y \mid z, \alpha, \sigma_u)\phi(z \mid \mu_k(x), \Sigma_k(x))\, dzdy. \tag{23}$$

Step 1: Replace $z = \mu_k(x) + \Sigma_k^{1/2}(x)w$ in equation (23) to obtain

$$\int\int \mathcal{L}(y,a) f(y \mid \mu_k(w), \alpha, \sigma_u)\phi(w \mid 0, \mathbb{I}_d)\, dwdy, \tag{24}$$

where $\mu_k(w) \equiv \mu_k(x) + \Sigma_k^{1/2}(x)w$. Equation (21) can be written as

$$\int\int \mathcal{L}(y,a) f(y \mid z_0, \alpha, \sigma_u)\phi(w \mid 0, \mathbb{I}_d)\, dwdy. \tag{25}$$

Algebra shows that $\mu_k(x) = z_i^*(x_{i,k}) + O(k^{-1})$ and $\Sigma_k(x) = (\beta\Sigma_v^{-1}\beta'/k)O(k^{-1})$. By assumptions of the proposition, we have $z_i^* \to \Sigma_\beta^{-1}\mu_\beta = z_0$ as $k \to \infty$. This implies that for a given w and y, we have

$$\mu_k(w) = \mu_k(x) + \Sigma_k^{1/2}(x)w \to z_0 \text{ as } k \to \infty.$$

Thus, we can expected that equation (24) converge to (25) since

$$f(y_i \mid z_i, \alpha, \sigma_u) = h(y, \sigma_u)\exp([\Omega_\alpha(z_i)y_i - \Psi(\Omega_\alpha(z_i))]/a(\sigma_u))$$

is continuous on z_i. This follows by the continuity of $\Omega_\alpha(z_i)$ and $\Psi(\cdot)$, which holds under Assumption 1 and definition of $f(\cdot|z, \alpha, \sigma_u)$.

Step 2: By Assumption 2, equation (24) is bounded by

$$\int\int (c_1 + c_2y^2) f(y \mid \mu_k(w), \alpha, \sigma_u)\phi(w \mid 0, \mathbb{I}_d)\, dwdy, \tag{26}$$

and, in a similar way, equation (25) is bounded by

$$\int\int (c_1 + c_2y^2) f(y \mid z_0, \alpha, \sigma_u)\phi(w \mid 0, \mathbb{I}_d)\, dwdy. \tag{27}$$

By Exercise 12, p. 133 in Dudley (2002), it will be sufficient to prove that (26) and (27) are well-defined, and that equation (26) converges to (27). To do this, we can ignore the constants. Thus, we want to prove that

$$\mathbb{E}_\theta[y_k^2] \to \mathbb{E}_\theta[y_0^2] \text{ as } k \to \infty, \tag{28}$$

where

$$y_k \sim f(y \mid \mu_k(w), \alpha, \sigma_u)\phi(w \mid 0, \mathbb{I}_d)$$

and

$$y_0 \sim f(y \mid z_0, \alpha, \sigma_u)\phi(w \mid 0, \mathbb{I}_d)$$

Since the p.d.f. of y_k converges to y_0 point-wise, it follows that y_k converges weakly to y_0. By the Continuous Mapping Theorem, it follows that y_k^2 converges weakly to y_0^2. By Theorem 3.5, p.31 in Billingsley (1999), we only need to prove that $\{y_k^2\}_k$ is uniformly integrable to conclude (28).

Step 3: We will prove that $\sup \mathbb{E}_\theta[|y_k|^3] < +\infty$, which implies that $\{y_k^2\}_k$ is uniformly integrable. For details see equation (3.18), p.31, in Billingsley (1999). Algebra shows

$$\begin{aligned}
\mathbb{E}_\theta[|y_k|^3] &= \mathbb{E}_\theta[|y_k|^3 1_{\{|y_k|>1\}}] + \mathbb{E}_\theta[|y_k|^3 1_{\{|y_k|\leq 1\}}] \\
&= \int_1^\infty \mathbb{P}_\theta[|y_k|^3 > t]dt + \mathbb{P}_\theta[|y_k| > 1] + \mathbb{E}_\theta[|y_k|^3 1_{\{|y_k|\leq 1\}}] \\
&\leq \int_1^\infty \mathbb{P}_\theta[|y_k| > t^{1/3}]dt + 2 \\
&= \int_1^\infty \int \mathbb{P}_\theta[|y_k| > t^{1/3} \mid \mu_k(w), \alpha, \sigma_u]\phi(w \mid 0, \mathbb{I}_d)dwdt + 2.
\end{aligned}$$

Since all the terms are positive, we can apply Tonelli's Theorem and change the order of the integrals. This implies

$$\mathbb{E}_\theta[|y_k|^3] \leq \int \int_1^\infty \mathbb{P}_\theta[|y_k| > t^{1/3} \mid \mu_k(w), \alpha, \sigma_u]dt\phi(w \mid 0, \mathbb{I}_d)dw + 2,$$

and by Assumption 3, this is lower than

$$\int \int_1^\infty t^{-4/3}\big(c_3 + c_4 \exp(c_5|\Omega_\alpha(\mu_k(w))|)\big)dt\, \phi(w \mid 0, \mathbb{I}_d)dw + 2.$$

Algebra shows that expression above is equal to

$$\int 3\big(c_3 + c_4 \exp(c_5|\Omega_\alpha(\mu_k(w))|)\big)\, \phi(w \mid 0, \mathbb{I}_d)dw + 2,$$

where $\exp(c_5|\Omega_\alpha(\mu_k(w))|)$ can be written as

$$\exp(c_5|\Omega_\alpha(\mu_k(w)) - \Omega_\alpha(z_0) + \Omega_\alpha(z_0)|),$$

which is lower than

$$\exp(c_5|\Omega_\alpha(\mu_k(w)) - \Omega_\alpha(z_0)| + |\Omega_\alpha(z_0)|).$$

By Assumption 1, the previous expression is lower than

$$\exp(c_5 K_\alpha |\mu_k(w) - z_0| + c_5|\Omega_\alpha(z_0)|),$$

where $\mu_k(w) - z_0 = \mu_k(x) - z_0 + \Sigma_k^{1/2}(x)w$. This implies that

$$\exp(c_5|\Omega_\alpha(\mu_k(w))|) \leq C_k \exp(c_5 K_\alpha |\Sigma_k^{1/2}(x)w|),$$

where $C_k \equiv \exp(c_5 K_\alpha |\mu_k(x) - z_0| + c_5|\Omega_\alpha(z_0)|)$.

All this algebra implies,

$$\mathbb{E}_\theta[|y_k|^3] \leq \int 3(c_3 + c_4 C_k \exp(c_5 K_\alpha |\Sigma_k^{1/2}(x)w|))\phi(w \mid 0, \mathbb{I}_d)\, dw + 2, \tag{29}$$

which can be bounded using the Moment Generation Function of the Normal distribution. To see this, define by $\Gamma_k \equiv ||\Sigma_k^{1/2}(x)||$ the matrix norm. This implies

$$|\Sigma_k^{1/2}(x)w| \leq \Gamma_k |w| \leq \Gamma_k \Sigma_{j=1}^d |w_j|,$$

where the second inequality comes from triangle inequality or algebra. Using this, we have that (29) is lower than

$$\int 3(c_3 + c_4 C_k \exp(c_5 K_\alpha \Gamma_k \Sigma_{j=1}^d |w_j|))\phi(w \mid 0, \mathbb{I}_d)\, dw + 2.$$

By definition, C_k converges to $\exp(c_5|\Omega_\alpha(z_0)|)$, thus is uniformly bounded. Then, it will be sufficient to prove that

$$\int \exp(c_5 K_\alpha \Gamma_k \Sigma_{j=1}^d |w_j|)\phi(w \mid 0, \mathbb{I}_d)\, dw$$

is uniformly bounded. To see that, observe that this expression can be written as

$$\prod_{j=1}^d \int \exp(c_5 K_\alpha \Gamma_k |w|)\phi(w \mid 0, 1)\, dw,$$

which is lower that

$$\prod_{j=1}^{d} \int (\exp(-c_5 K_\alpha \Gamma_k w) + \exp(c_5 K_\alpha \Gamma_k w))\, \phi(w \mid 0, 1)\, dw.$$

Define by $M_\phi(t) \equiv \int \exp(tw)\phi(w \,|\, 0, 1)\, dw$ the Moment Generation Function. Then, we have

$$\mathbb{E}_\theta[|y_k|^3] \leq 3c_3 + 3c_4 C_k \big\{ M_\phi(-c_5 K_\alpha \Gamma_k) + M_\phi(c_5 K_\alpha \Gamma_k) \big\}^d + 2. \tag{30}$$

By continuity, we know that $\Gamma_k \to 0$ as $k \to \infty$. This implies that equation (30) is uniformly bounded. This complete the proof of uniformly integrability.

Proof of Part 2 : In a similar way as we did for part 1 in step 2, it will be sufficient to prove that

$$\int y^2 f(y \mid z_i^*(x_{i,k}), \alpha, \sigma_u) dy \to \int y^2 f(y \mid z_0, \alpha, \sigma_u) dy.$$

To conclude this, as we did for part 1 in step 3, it will be sufficient to prove that

$$\int |y|^3 f(y \mid z_i^*(x_{i,k}), \alpha, \sigma_u) dy \tag{31}$$

is uniformly bounded. By Assumption 3, and following step 3 above, this expression is lower than

$$3c_3 + 3c_4 \exp(c_5 |\Omega_\alpha(z_i^*(x_{i,k}))|) + 2,$$

which converges to

$$3c_3 + 3c_4 \exp(c_5 |\Omega_\alpha(z_0)|) + 2.$$

This proves that (31) is uniformly bounded. This complete the proof.

A.5 Technical Lemmas

In this section, we present two technical lemmas to study the deterministic linear representations and its relations with sufficiency concept and to compute mutual information with the nuisance v_i. The derivation of these results use basic algebraic manipulation based on the multivariate normal model.

Lemma 1: Let $\hat{z}_i$ be a deterministic linear representation of x_i,

$$\hat{z}_i \equiv \underbrace{A}_{p \times k} \underbrace{x_i}_{k \times 1}.$$

Suppose the inverse of $\mathbb{E}_\theta[\hat{z}_i\hat{z}_i']$ exists. Then, $\hat{z}_i$ is a sufficient representation of x_i at θ if and only if A solves the Sufficient Representation Equation (SRE):

$$\Sigma_x A'(A\Sigma_x A')^{-1}A\beta'\alpha = \beta'\alpha, \tag{32}$$

where $\Sigma_x \equiv \underbrace{\Sigma_v}_{k\times k} + \beta'\beta$.

Proof. There are two parts:

Part I: Suppose A solves SRE. We will prove that $\hat{z}_i = Ax_i$ is a sufficient representation of x_i, i.e. $y_i \perp x_i \mid \hat{z}_i$. First observe that

$$\begin{pmatrix} x_i \\ y_i \\ \hat{z}_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_x & \beta'\alpha & \Sigma_x A' \\ \alpha'\beta & \Sigma_y & \alpha'\beta A' \\ A\Sigma_x & A\beta'\alpha & A\Sigma_x A' \end{pmatrix}\right).$$

where $\Sigma_x = \Sigma_v + \beta'\beta$, $\Sigma_y = \sigma_u^2 + \alpha'\alpha$ and $\mathbb{E}_\theta[\hat{z}_i\hat{z}_i'] = A\Sigma_x A'$. Since the vector $[x_i\ y_i'\ \hat{z}_i']'$ is Gaussian, it follows that

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \mid \hat{z}_i \sim \mathcal{N}\left(\overline{\mu}, \overline{\Sigma}\right).$$

where $\overline{\mu} = \Sigma_{12}\Sigma_2^{-1}\hat{z}_i$ and $\overline{\Sigma} = \Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{21}$. Here, $\Sigma_2 = A\Sigma_x A'$ has an inverse matrix by assumption and

$$\Sigma_1 = \begin{pmatrix} \Sigma_x & \beta'\alpha \\ \alpha'\beta & \Sigma_y \end{pmatrix}, \quad \text{and} \quad \Sigma_{12} = \begin{pmatrix} \Sigma_x A' \\ \alpha'\beta A' \end{pmatrix} = \Sigma_{21}'$$

Define

$$\Sigma_{12}\Sigma_2^{-1}\Sigma_{21} = \begin{pmatrix} \overline{\Sigma}_1 & \overline{\Sigma}_{12} \\ \overline{\Sigma}_{21} & \overline{\Sigma}_2 \end{pmatrix}$$

Algebra shows

$$\begin{aligned}
\overline{\Sigma}_1 &= \Sigma_v A'\Sigma_2^{-1}A\Sigma_x + \beta'\beta A'\Sigma_2^{-1}A\Sigma_x \\
\overline{\Sigma}_{12} &= \Sigma_x A'\Sigma_2^{-1}A\beta'\alpha \\
\overline{\Sigma}_{21} &= \alpha'\beta A'\Sigma_2^{-1}A\Sigma_x \\
\overline{\Sigma}_2 &= \alpha'\beta A'\Sigma_2^{-1}A\beta'\alpha
\end{aligned}$$

Since A solve SRE and $\Sigma_2 = A\Sigma_x A'$, it follows that $\overline{\Sigma}_{12} = \beta'\alpha$. This implies that correlation between $x_i \,|\, \hat{z}_i$ and $y_i \,|\, \hat{z}_i$ is zero, which proves that $y_i \perp x_i \,|\, \hat{z}_i$ since $(y_i x_i')' \,|\, \hat{z}_i$ is Gaussian.

Part II: Suppose that $\hat{z}_i = Ax_i$ is a sufficient representation of x_i. This implies $y_i \perp x_i \mid \hat{z}_i$, in particular correlation between $x_i \,|\, \hat{z}_i$ and $y_i \,|\, \hat{z}_i$ is zero. This implies that $\overline{\Sigma}_{12} = \beta'\alpha$. Since $\Sigma_2 = A\Sigma_x A'$ we have

$$\Sigma_x A'(A\Sigma_x A')^{-1} A\beta'\alpha = \beta'\alpha$$

which is the Sufficient Representation Equation, then A solves SRE. $\square$

Lemma 2 : Suppose $\hat{z}_i = Ax_i$ is a deterministic linear representation of dimension p and v_i is the noise in the factor model for the covariates x_i. Assume in addition that the inverse of $\mathbb{E}_\theta[\hat{z}_i \hat{z}_i']$ and $A\beta'\beta A'$ exists, in particular that $p < k$. Then, the mutual information between $\hat{z}_i$ and v_i is

$$I_\theta(\hat{z}_i; v) = \quad \frac{1}{2} \ln \left(\frac{\det(A\Sigma_x A')}{\det(A\beta'\beta A')} \right) > 0,$$

where $\Sigma_x \equiv \Sigma_v + \beta'\beta$.

Proof. Since $x_i = \beta' z_i + v_i$, where $z_i \perp v_i$, and $\hat{z}_i = Ax$, then

$$\begin{pmatrix} \hat{z}_i \\ v \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} A\Sigma_x A' & A\Sigma_v \\ \Sigma_v A' & \Sigma_v \end{pmatrix} \right).$$

To compute the mutual information between $\hat{z}_i = Ax_i$ and v_i, we need to calculate the Kullback-Leibler divergence between the multivariate normal distribution defined above and the following multivariate normal distribution (assuming no correlation between $\hat{z}_i$ and v_i):

$$\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} A\Sigma_x A' & 0 \\ 0 & \Sigma_v \end{pmatrix} \right).$$

By assumption, the inverse of both $\mathbb{E}_\theta[\hat{z}\hat{z}'] = A\Sigma_x A'$ and Σ_v exists. By Proposition 1 in Contreras-Reyes and Arellano-Valle (2012), the Kullback-Leibler divergence between these two multivariate normal distributions is

$$\frac{1}{2} \left\{ \ln \left(\frac{\det(\Omega_2)}{\det(\Omega_1)} \right) \right\}.$$

where

$$\Omega_1 = \begin{pmatrix} A\Sigma_x A' & A\Sigma_v \\ \Sigma_v A' & \Sigma_v \end{pmatrix} \quad \text{and} \quad \Omega_2 = \begin{pmatrix} A\Sigma_x A' & 0 \\ 0 & \Sigma_v \end{pmatrix}$$

Since the inverse of both $A\Sigma_x A'$ and Σ_v exists by assumption, Theorem 2 in Silvester (2000) implies that

$$\begin{aligned}
\det(\Omega_1) &= \det(\Sigma_v)\det(A\Sigma_x A' - A\Sigma_v A') \\
&= \det(\Sigma_v)\det(A\beta'\beta A') \\
&\quad (\text{since } \Sigma_x = \Sigma_v + \beta'\beta) \\
\det(\Omega_2) &= \det(\Sigma_v)\det(A\Sigma_x A')
\end{aligned}$$

It follows that

$$\begin{aligned}
I_\theta(\hat{z}_i; v) =& \frac{1}{2}\left\{\ln\left(\frac{\det(\Omega_2)}{\det(\Omega_1)}\right)\right\} \\
=& \frac{1}{2}\left\{\ln\left(\frac{\det(\Sigma_v)\det(A\Sigma_x A')}{\det(\Sigma_v)\det(A\beta'\beta A')}\right)\right\} \\
=& \frac{1}{2}\left\{\ln\left(\frac{\det(A\Sigma_x A')}{\det(A\beta'\beta A')}\right)\right\}
\end{aligned}$$

which is the close form expression of this lemma.

To conclude that mutual information between $\hat{z}_i$ and v_i is positive, let us use the following the general fact. Mutual information of two random variables is zero if and only if these random variables are independent. Since $\hat{z}_i = g(\beta' z_i + v_i)$ and v_i both have in common v_i, it follows that they are not independent. This implies $I(\hat{z}_i, v_i) > 0$. □

References

Achille, A. and S. Soatto (2018): "Emergence of invariance and disentanglement in deep representations," *The Journal of Machine Learning Research*, 19, 1947–1980.

Anderson, T. W. (2003): *An Introduction to Multivariate Statistical Analysis*, Series in Probability and Mathematical Statistics, Wiley-Interscience, third edition ed.

Bai, J. and S. Ng (2006): "Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions," *Econometrica*, 74, 1133–1150.

Bengio, Y., A. Courville, and P. Vincent (2013): "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, 35, 1798–1828.

Billingsley, P. (1999): *Probability and Measure*, John Wiley and Sons, second ed.

Collins, M., S. Dasgupta, and R. E. Schapire (2002): "A generalization of principal components analysis to the exponential family," in *Advances in neural information processing systems*, 617–624.

Contreras-Reyes, J. E. and R. B. Arellano-Valle (2012): "Kullback–Leibler divergence measure for multivariate skew-normal distributions," *Entropy*, 14, 1606–1626.

Dudley, R. M. (2002): *Real Analysis and Probability*, Cambridge Studies in Advanced Mathematics, Cambridge University Press, 2 ed.

Goodfellow, I., Y. Bengio, and A. Courville (2016): *Deep Learning*, MIT Press.

Lawley, D. N. and A. E. Maxwell (1962): "Factor analysis as a statistical method," *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12, 209–229.

——— (1973): "Regression and factor analysis," *Biometrika*, 60, 331–338.

Mattei, P.-A. and J. Frellsen (2018): "Leveraging the exact likelihood of deep latent variable models," in *Advances in Neural Information Processing Systems*, 3855–3866.

McCullagh, P. and J. Nelder (1989): *Generalized Linear Models*, Chapman & Hall.

Rezende, D. J., S. Mohamed, and D. Wierstra (2014): "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:1401.4082*.

SCHMIDT-HIEBER, J. (2020): "Nonparametric regression using deep neural networks with ReLU activation function," *Annals of Statistics*.

SILVESTER, J. R. (2000): "Determinants of block matrices," *The Mathematical Gazette*, 84, 460–467.

STOCK, J. H. AND M. W. WATSON (2002): "Forecasting using principal components from a large number of predictors," *Journal of the American statistical association*, 97, 1167–1179.