Sparse Bayesian Factor Models with Mass-Nonlocal Factor Scores

Yingjie Huang[*,†], Dafne Zorzetto [*,†] and Roberta De Vito[†,‡]

Abstract. Bayesian factor models are widely used for dimensionality reduction and pattern discovery in high-dimensional datasets across diverse fields. These models typically focus on imposing priors on factor loading to induce sparsity and improve interpretability. However, factor scores, which play a critical role in individual-level associations with factors, have received less attention and are assumed to follow a standard normal distribution. This assumption oversimplifies the heterogeneity often observed in real-world applications. We propose the sparse Bayesian Factor model with MAss-Nonlocal factor scores (BFMAN), a novel framework that addresses these limitations by introducing a mass-nonlocal prior on factor scores. This prior allows for both exact zeros and flexible, nonlocal behavior, capturing individual-level sparsity and heterogeneity. The sparsity in the score matrix enables a robust and novel approach to determine the optimal number of factors. Model parameters are estimated via a fast and efficient Gibbs sampler. Extensive simulations demonstrate that BFMAN outperforms standard Bayesian factor models in factor recovery, sparsity detection, score estimation, and selection of the optimal number of factors. We apply BFMAN to the Hispanic Community Health Study/Study of Latinos, identifying meaningful dietary patterns and their associations with cardiovascular disease, showcasing the model's ability to uncover insights into complex nutritional data.

Keywords: Factors selection, nutritional data, pMOM distribution, spike and non-local prior.

1 Introduction

Bayesian factor models play a central role in numerous disciplines, including social sciences [11], genomics [35], nutrition [17], and more broadly in high-dimensional applications [31, 9]. These models are particularly advantageous for large-scale data, providing a structured approach to dimensionality reduction, improving interpretability, and facilitating deeper understanding of the underlying data structure [5]. When dealing with high-dimensional datasets, incorporating sparsity or penalization techniques becomes critical for two primary reasons: first, to ensure interpretability and achieve meaningful insights into the data, and second, to guarantee that the covariance matrix is estimable [25].

Methodological developments have traditionally focused on imposing priors on the factor loading matrix, using approaches such as shrinkage priors [6, 24], sparsity priors

---

[*]Equally contributing co-first authors.

[†]Data Science Institute, Brown University, yingjie_huang@brown.edu; dafne_zorzetto@brown.edu

[‡]Department of Biostatistics, Brow University, roberta_devito@brown.edu

[8, 18], spike-and-slab [30, 4], and non-local mass priors [2]. However, little attention has been given to the factor score matrix, typically assumed to follow a standard multivariate normal distribution, implying independence between factors [34, 15]. While some flexible approaches have been proposed using non-diagonal covariance structures for the factor scores, they often increase model complexity without addressing individual-level heterogeneity. [26].

Factor scores play a critical role, as they quantify the score that each individual expresses on the corresponding factors, making them particularly relevant in various applications [14, 10]. In applications such as nutritional epidemiology, factor analysis is often used to estimate both dietary patterns (i.e., factor loadings) and factor scores, to estimate the association between these patterns and disease outcome [10]. Therefore, a more refined estimation and appropriate prior specification for factor scores are essential to accurately model the relationship between each factor and health outcomes. For instance, in diet-related studies, individuals may exhibit substantial heterogeneity in adherence to dietary patterns—some may strictly follow a given pattern, while others do not at all. Standard Gaussian assumptions fail to capture this heterogeneity and lack the flexibility to induce sparsity in individual-level scores.

To address these limitations, we introduce the Sparse Bayesian Factor Model with Mass-Nonlocal Factor Scores (BFMAN), a novel approach that assumes a mass-nonlocal prior directly on the latent factor scores. This framework introduces a more flexible posterior distribution for factor scores, characterizing the heterogeneity in subject-level associations with latent factors. The mass-nonlocal prior with a non-null probability allows for exact zero in the score matrix, and a non-local slab prior that do not overlap the spike yielding a sparse, heterogeneous structure that reflects real-world variation in individual behavior.

Our model incorporates three key features enabled by this sparse prior on factor scores. First, sparsity in the score enhances interpretability by linking each latent factor to a small subset of individuals. Second, when the sparsity assumption holds, it improves estimation accuracy and model efficiency. Third, inducing sparsity enables a novel, principled approach to inferring the number of latent factors. While existing approaches typically focus on the sparsity or shrinkage in the factor loading matrix [6, 8] or proportion of variance explained [14], our method takes a novel approach by leveraging the level of sparsity in the factor score matrix to infer the optimal number of factors. This unique perspective allows for more precise identification of factors and better captures the structural complexity of the data. To ensure computational scalability, we develop a fast and efficient Gibbs sampler for posterior inference, publicly available at: y1jHuang/nonloc_sparse_bayes.

We conduct extensive simulation studies to evaluate the performance of BFMAN. The results demonstrate that our method consistently outperforms existing methods in factor recovery, sparsity detection, score estimation, and accuracy in selecting the number of latent factors. Moreover, by modeling sparsity at the level of individual scores, BFMAN provides a more nuanced and realistic characterization of latent behavior, making the model particularly well-suited for complex, high-dimensional applications.

To further showcase the utility of our approach, we apply BFMAN to the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) [27], a multi-center epidemiologic study designed to investigate critical components impacting the health of Hispanic/Latino populations [33]. A key aim of the study is the association of diet in cardiovascular disease risk factors, including diabetes, hypertension, and high cholesterol [12]. Using our method, we uncover interpretable dietary patterns and their associations with these three risk factors, providing novel understanding of the diet-disease relationship.

The paper is organized as follows. Section 2 introduces the BFMAN framework, the proposed mass-nonlocal prior for the factor score, and the new procedure for selecting the optimal number of factors. Section 3 presents extensive simulation studies comparing BFMAN to standard methods. Section 4 applies the BFMAN model to the HCHS/SOL data. Finally, Section 5 includes a discussion of our findings and their implications.

2   Bayesian mass-nonlocal factor analysis

2.1   Model and prior specification

Let $\mathbf{Y} \in \mathbb{R}^{n \times p}$ be the observed data matrix where n is the number of observations and p is the number of variables. The latent factor model for each observation $i \in \{1, \cdots, n\}$, is given by:

$$\mathbf{y}_i = \Lambda \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\Lambda \in \mathbb{R}^{p \times k}$ is the factor loading matrix, $\boldsymbol{\eta}_i \in \mathbb{R}^k$ is the latent factor score vector for the i-th observation, where k indicates the number of factors, and $\boldsymbol{\epsilon}_i \sim \mathcal{N}_p(0, \Sigma)$ is the idiosyncratic error matrix, with $\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_p^2)$.

Traditional factor models assume $\boldsymbol{\eta}_i \sim \mathcal{N}(0, I_k)$, which may not capture the sparsity and heterogeneity in the factor scores often observed in practice [34, 15]. To address this, we propose a mixture prior on the factor score η_{ih}, for each observation $i \in \{1, \ldots, n\}$ and factor $h \in \{1, \ldots, k\}$, that includes a Dirac distribution with mass in zero and a slab component given by a product moment (pMOM) prior [19]:

$$\{\eta_{ih} | \theta_h, \phi_h\} \sim (1 - \theta_h)\delta_0(\eta_{ih}) + \theta_h \mathrm{pMOM}(\eta_{ih} \mid \psi), \tag{2}$$

where $\delta_0(\cdot)$ is a Dirac measure with mass at zero, and $\mathrm{pMOM}(\cdot)$ has probability density:

$$p(\eta_{ih} \mid \psi) = \frac{1}{\sqrt{2\pi\psi^3}} \exp\left(-\frac{\eta_{ih}^2}{2\psi}\right) \eta_{ih}^2,$$

with scale parameter $\psi > 0$. This choice ensures a flexible distribution that avoids overlap with the spike in zero while preserving tails similar to a normal distribution. Figure 1 illustrates the shape of the pMOM density across different values of ψ, highlighting its non-locality and zero-avoiding property.
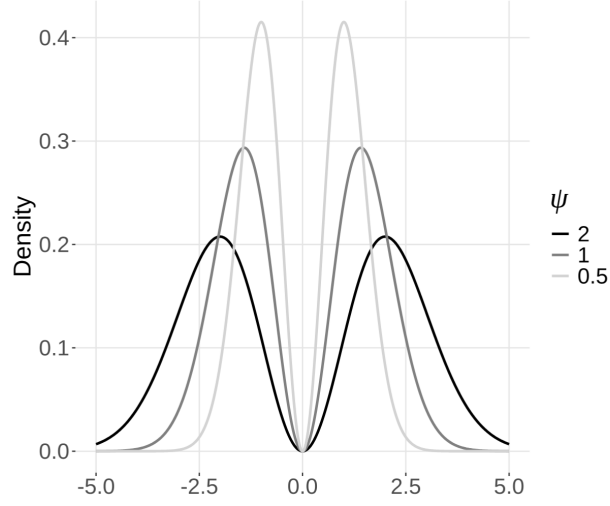
Figure 1: pMOM densities for different dispersion parameters ψ. This distribution defines the slab component of our mass-nonlocal prior.

The prior elicitation for the hyperparameters in the distributions in (1) and (2), respectively, the variance of the idiosyncratic error and the weights of the mixture distribution, is defined as follows:

$$
\begin{aligned}
\sigma_j^{-2} &\sim \mathrm{Ga}(a_\sigma, b_\sigma) \quad \forall j \in \{1, \cdots, p\}, \\
\theta_h &\sim \mathrm{Beta}(a_\theta, b_\theta) \quad \forall h \geq 1.
\end{aligned}
\tag{3}
$$

The formulation of the mass-nonlocal prior (2) allows us to introduce a latent variable Z_{ih} for each observation $i \in \{1, \ldots, n\}$ and factor $h \in \{1, \ldots, k\}$, denoting whether η_{ih} is drawn from the spike or the non-local slab with the following distribution:

$$
Z_{ih} \sim \mathrm{Bern}(\theta_h),
$$

where θ_h is the probability for the corresponding factor score η_{ij} to follow a pMOM probability distribution, with prior (3), such that

$$
\{\eta_{ih}|Z_{ih} = 1, \psi\} \sim \mathrm{pMOM}(\psi) \quad \text{and} \quad \{\eta_{ih}|Z_{ih} = 0\} \sim \delta_0.
$$

For the factor loading matrix Λ, we adopt the multiplicative gamma process shrinkage (MGPS) prior [6]:

$$
\begin{aligned}
\lambda_{jh}|\phi_{jh}, \tau_h &\sim \mathcal{N}(0, \phi_{jh}^{-1}\tau_h^{-1}), \\
\phi_{jh} &\sim \mathrm{Ga}(\nu/2, \nu/2),
\end{aligned}
$$

$$\tau_h = \prod_{l=1}^{h} \delta_l, \quad \delta_1 \sim \text{Ga}(a_1, 1), \quad \delta_l \sim \text{Ga}(a_2, 1), \quad l \geq 2, \tag{4}$$

where $\{\tau_h\}_{h \geq 1}$ increases with h. This prior induces increasing shrinkage on higher-indexed columns of Λ. Alternative priors such as the cumulative shrinkage prior [24] or generalized MGPS [32] could also be used.

We adopt the recommended hyperparameter settings from Bhattacharya and Dunson [6] and Durante [16], ensuring stability and efficiency in posterior inference.

2.2 Posterior Computation

We develop an efficient Gibbs sampler for posterior computation, levering the conjugate prior with the exception of the pMOM distribution, which requires a Metropolis-Hasting step.

Following the steps in the algorithm 2, in each iteration $r = 1, \ldots, R$, we use the observed data $\mathbf{y}$ to update the parameters and random variables. Let $\boldsymbol{\lambda}_j$ denote the j-th row of the factor loading matrix Λ, for $j \in \{1, \ldots, p\}$, and $\boldsymbol{\eta}_i$ the i-th row of the latent score matrix $\boldsymbol{\eta}$, for $i \in \{1, \ldots, n\}$. We indicate with $\mathbf{y}^{(j)} = (y_{1j}, \cdots, y_{nj})^{\mathsf{T}}$ the j variable across all individuals. Let $D_j^{-1} = \text{diag}(\phi_{j1}\tau_1, \cdots, \phi_{jk}\tau_k)$ denote the diagonal prior precision matrix for MGPS prior (4), and $\{\psi_j\}_{j \in \{1,\ldots,k\}}$ the pMOM parameters.

Then the steps for posterior sampling are as follows:

1. The loading matrix entries $\{\boldsymbol{\lambda}_j\}_{j \in \{1,\ldots,p\}}$ are sampled from the following posterior distribution:

$$f(\boldsymbol{\lambda}_j | \tau, \Lambda, D, \sigma_y) \sim \mathcal{N}_k\Big\{ \left(D_j^{-1} + \sigma_j^{-2}\boldsymbol{\eta}^{\mathsf{T}}\boldsymbol{\eta}\right)^{-1} \boldsymbol{\eta}^{\mathsf{T}}\sigma_j^{-2}y^{(j)}, \left(D_j^{-1} + \sigma_j^{-2}\boldsymbol{\eta}^{\mathsf{T}}\boldsymbol{\eta}\right)^{-1} \Big\}.$$

2. The the MGPS prior introduces two parameter. First, the local shrinkage parameter ϕ_{jh}, with the following poster distribution, for $j \in \{1, \ldots, p\}$ and $h \in \{1, \ldots, k\}$:

$$f(\phi_{jh} | \nu, \Lambda, \tau) \sim \text{Ga}\left(\frac{\nu + 1}{2}, \frac{\nu + \tau_h \lambda_{jh}^2}{2}\right).$$

3. Second, the global shrinkage parameter δ_h, with posterior distribution defined as follows:

$$f(\delta_h | a, \tau, \phi, \Lambda) \sim \text{Ga}\Big\{a_h + \frac{p}{2}(k - h + 1), 1 + \frac{1}{2}\sum_{l=1}^{k} \tau_l^{(h)} \sum_{j=1}^{p} \phi_{jl}\lambda_{jl}^2\Big\},$$

where $\tau_l^{(h)} = \prod_{t=1, t \neq h}^{l} \delta_t$ for $h = 1, \cdots, k$.

4. The factor score η_{ih}, conditional to the latent variable Z_{ih}, are sampled from:

$$f(\eta_{ih}|Z_{ih}) = \begin{cases} \pi(\eta_{ih}|c,d) & \text{if } Z_{ih} = 1, \\ 0 & \text{if } Z_{ih} = 0; \end{cases}$$

where $\pi(\eta_{ih}|-)$ indicates the following distribution:

$$\pi(\eta_{ih}|c,d) \propto \exp\left\{ -c\left(\eta_{ih} - \frac{d}{c}\right)^2 \right\}\eta_{ih}^2;$$

with $c = \dfrac{1}{2\psi} + \displaystyle\sum_{j=1}^{p} \dfrac{1}{2\sigma_j^2}\lambda_{jh}^2$ and $d = \displaystyle\sum_{j=1}^{p} \dfrac{1}{2\sigma_j^2}\lambda_{jh}\left(y_{ij} - \sum_{l\neq h}^{k} \lambda_{jl}\eta_{il}\right).$

Due to the non-conjugacy of $\pi(\eta_{ih}|c,d)$, we embedded a Metroplis-Hastings algorithm, which is implemented as follows:

Algorithm 1: Metroplis-Hastings Algorithm

Input: Probability density $\pi(\eta_{ih}|c,d)$, initial state η_{ih}^0
Output: Posterior samples from $\pi(\eta_{ih}|c,d)$
for $m = 1$ **to** M **do**
$\quad$ Generate a random candidate $\eta_{ih}^* \sim \mathcal{N}(\mu = \eta_{ih}^{m-1}, \sigma)$;
$\quad$ Calculate acceptance probability $r = \exp\left(\log \pi(\eta_{ih}^*|c,d) - \log \pi(\eta_{ih}^{m-1}|c,d)\right)$;
$\quad$ Accept or reject:
$\quad$ $\alpha = \min(1, r)$,
$\quad$ $Z = \text{Bern}(\alpha)$,
$\quad$ $\eta_{ih}^m = Z\eta_{ih}^* + (1-Z)\eta_{ih}^{m-1}$.
end

---

5. Sample latent variable Z_{ih}, for $i \in \{1, \dots, n\}$ and $h \in \{1, \dots, k\}$, from a Bernulli distribution such that

$$\begin{aligned} \Pr(Z_{ih} = 0|-) &= \frac{f(Z_{ih}=0)f(\mathbf{y}_i|Z_{ih}=0,-)}{f(Z_{ih}=1)f(\mathbf{y}_i|Z_{ih}=1,-) + f(Z_{ih}=0)f(\mathbf{y}_i|Z_{ih}=0,-)} \\ &= \frac{f(Z_{ih}=0)}{f(Z_{ih}=0) + f(Z_{ih}=1)T}, \end{aligned}$$

where $T = K\sqrt{2\pi}H^{-\frac{1}{2}}(H^{-1}+M^2)$, $H = \frac{1}{\psi}+\boldsymbol{\lambda}_h^\intercal\Sigma^{-1}\boldsymbol{\lambda}_h$, $M = \frac{1}{H}(\mathbf{y}_i - \Lambda_{(-h)}\boldsymbol{\eta}_{i(-h)})\Sigma^{-1}\boldsymbol{\lambda}_h$ and $K = 2\pi\psi^{-\frac{3}{2}}\exp\{\frac{1}{2}HM^2\}$. The $\Lambda_{(-h)}$ represents $p \times (k-1)$ matrix with hth column dropped, and $\boldsymbol{\eta}_{i(-h)}$ denotes $k-1$ vector with η_{ih} entry deleted.

6. Sample the probability parameter θ_h, for each factor $h \in \{1, \dots, k\}$, from:

$$f(\theta_h|\mathbf{Z}_h, a_1, b_1) = \text{Beta}\left(\sum_{i}^{n} Z_{ih} + a_1, n - \sum_{i}^{n} Z_{ih} + b_1\right),$$

where a_1 and b_1 are the hyperparameters.

7. The residual variance σ_j^2, for $j \in \{1, \ldots, p\}$, is drawn from the posterior distribution:

$$f(\sigma_j^{-2}|a_\sigma, b_\sigma, y, \Lambda, \boldsymbol{\eta}) = \text{Ga}\Big\{a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^{n}\big(y_ij - \boldsymbol{\lambda}_j^\intercal\boldsymbol{\eta}_i\big)^2\Big\}.$$

Algorithm 2: Posterior computation.

Input: Outcome matrix $\mathbf{Y}$
Output: Posterior distribution of each parameter
for $r = 1$ **to** R **do**
    Sample factor loading $\boldsymbol{\lambda}_j$ for $j \in \{1, \ldots, p\}$;
    Sample ϕ_{jh}, for $j \in \{1, \ldots, p\}$ and $h \in \{1, \ldots, k\}$;
    Sample δ_h for $h \in \{1, \ldots, k\}$;
    Sample factor score η_{ih} given the latent variable Z_{ih}, for $i \in \{1, \ldots, n\}$
     $h \in \{1, \ldots, k\}$;
    Sample the latent variable Z_{ih}, for $i \in \{1, \ldots, n\}$ and $h \in \{1, \ldots, k\}$;
    Sample the $\theta_h|\mathbf{Z}_h$, for $h \in \{1, \ldots, k\}$;
    Sample the residual variance σ_j^2 for $j \in \{1, \ldots, p\}$.
end

2.3 Model Identification

Latent factor models are non-identifiable due to their invariance under orthogonal transformations. Specifically, for any orthogonal matrix $Q \in \mathbb{R}^{k \times k}$, the latent structure can be equivalently expressed as $\Lambda^* = Q\Lambda$ and $\boldsymbol{\eta}^* = \boldsymbol{\eta}Q^\top$. As a result, the model can be rewritten as: $\mathbf{y}_i = \Lambda^*\boldsymbol{\eta}_i^* + \boldsymbol{\epsilon}_i$, highlighting the rotational ambiguity in the factorization of the latent space.

To address this non-identifiability issue, several approaches have been proposed in the literature. Standard solutions include imposing structural constraints on the loading matrix Λ. For instance, Lopes and West [25] enforce a lower-triangular structure with strictly positive diagonal entries to ensure uniqueness. Classical rotation methods such as the varimax criterion [21], aim to improve interpretability by maximizing the variance of squared loadings post-rotation. More recent approaches, including the parameter expansion framework of Ročková and George [30], extend this idea by proposing EM-based optimization schemes that seek sparse, rotated loading matrices. Their approach mitigates the risk of local optima by expanding the parameter space, enabling greater flexibility in the estimation process. This approach was then followed by Avalos-Pacheco et al. [2], addressing identifiability solely through adding sparsity in the factor loadings via a non-local mass prior. This is further corroborated by the recent paper of Frühwirth-Schnatter et al. [18], the generalized lower-triangular (GLT) decomposition. The decomposition introduces a structure of Λ that satisfies the following condition: for each column $h \in \{1, \cdots, k\}$, let l_h denote the row index of its leading non-zero entry. Then the indices $l_1, l_2, \cdots, l_k$ must be in ascending order, i.e. $l_1 < l_2 < \cdots, l_k$, while the leading entries $\Lambda_{l_h,h} > 0$.

While previous work has primarily addressed non-identifiability by constraining the loading matrix, our contribution lies in a different and novel direction by enforcing identifiability on the factor score matrix $\boldsymbol{\eta}$, instead. This represents a key innovation of our model, as, to the best of our knowledge, no existing work has considered identifiability from the perspective of the factor score matrix. Specifically, we extend the GLT decomposition [18] to the factor score matrix, $\boldsymbol{\eta}$. By applying the GLT constraints to the scores, we simultaneously achieve identifiability and parsimony with fewer assumptions in the latent space, both Λ and $\boldsymbol{\eta}$. Under this structure, the only possible rotation in the score is the identity matrix, eliminating rotational ambiguity in a natural and interpretable way [18].

Furthermore, the sparsity induced by the GLT structure not only resolves identifiability, but also facilitate inference on the number of factors.

2.4   Choosing number of factors

Determining the optimal number of factors is a critical challenge in factor analysis. The objective is to retain a small number of factors that captures the underlying structure of the data without introducing redundancy. Traditional methods for factor number selection typically fall into two broad categories: threshold-based methods and model selection via information criteria.

Threshold-based methods, such as Kaiser's criterion [21] or scree plots [10], rely on thresholds, such as retaining factors with corresponding eigenvalues greater than one, based on Principal Component Analysis (PCA). Although computationally efficient, these approaches are sensitive to the specific structure and variability of the dataset, often resulting in inconsistent or unstable estimates.

Information-theoretic criteria, including the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) [22, 28], offer a more principled alternative by comparing models with different numbers of factors [3]. However, these methods are computationally intensive, especially in high-dimensional settings, as each model configuration must be fit and evaluated independently.

Positioned between heuristic methods and information-theoretic criteria, Bhattacharya and Dunson [6] introduces an adaptive shrinkage approach that starts with a large number of factors and iteratively prunes redundant ones by shrinking the columns of the loading matrix. This strategy still relies on thresholding decisions—defining when a column is "close enough" to zero to be removed—and is sensitive to the tuning of shrinkage hyperparameters [16].

To address these challenges, we introduce a novel method for estimating the number of factors by focusing on sparsity and identifiability in the factor score matrix, not the loadings. Our approach shifts the focus to the factor score matrix $\boldsymbol{\eta}$, leveraging the sparsity-inducing mass-nonlocal prior introduced in (2), motivated by variable selection framework.

Under this formulation, begin with a conservative upper bound $K = 5\log(p)$ [6] following Bhattacharya and Dunson [6], and successively use posterior inference to de-

Table 1: Scenario-specific parameters used to generate the simulation experiments.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
n	100	100	30	3000
p	20	20	60	60
k	3	3	5	6
$\{\theta_h\}_{h=1}^k$	$0.4\ \forall h$	$\{0.8, 0.6, 0.4\}$	$\{0.9, 0.8, 0.7, 0.6, 0.5\}$	$\{0.8, 0.7, 0.6, 0.5, 0.4, 0.3\}$

termine the important factors while discarding the irrelevant ones. Specifically, the posterior estimation of the factor score $\boldsymbol{\eta}$ provides insights into which entries can be considered effectively zero and which deviate significantly from zero. Columns where a high proportion (e.g., $\geq 80\%$) of entries are exactly zero are considered not important and discarded. This procedure is guided by the posterior distribution over the latent indicators Z_{ih}, which directly reflect whether an individual expresses a given factor.

Our proposed sparsity-inducing prior for the score matrix $\boldsymbol{\eta}$ allows us to automatically estimate the number of relevant factors, without relying on arbitrary thresholds for the loadings or model comparisons. This strategy shifts the identifiability constraint from the loading matrix to the score matrix, reducing the need for strong structural assumptions and providing a clear and interpretable mechanism for factor selection. Moreover, it results in a computationally efficient and flexible method that avoids overestimation and adapts naturally to the structure of the data.

3   Simulation study

We conduct extensive simulation experiments to evaluate the performance of our BF-MAN in recovering the sparse structure of the factor score matrix $\boldsymbol{\eta}$ and the factor loading matrix Λ. A particular focus is placed on the model's ability to correctly identify and impute the zeroes entries in $\boldsymbol{\eta}$. To benchmark the performance, we compare our method with the MGPS factor model by Bhattacharya and Dunson [6].

We construct four distinct simulation scenarios with varying levels of complexity in the data generation process, focusing on different sparsity schemes in the factor scores. The data generation process differs across scenarios in terms of sample size n, number of observed variables p, number of factors k, and the sparsity of the factors score induced by $\{\theta_h\}_{h=1}^k$. Table 1 summarizes the scenario-specific parameters, while Table 2 describes the data-generating process used across all scenarios.

Specifically, Scenario 1 represents a setup with a small n and p. Scenario 2 retains the same dimensions as Scenario 1 but introduces heterogeneity in the factor scores sparsity, with increasing probabilities of zeros across factors columns. Scenarios 3 reflects a high-dimensional setting where the number of variables exceeds the sample size, i.e., $p >> n$. Finally, Scenario 4 mimics our real data nutritional application analyzed in Section 4. Each scenario is replicated 50 times.

To evaluate the ability to recover the true latent structure—factor loading Λ and factor score $\boldsymbol{\eta}$—we compute the RV coefficient [29]. The RV coefficient compares the

Table 2: Data generating mechanism across the scenarios, for $i \in \{1, \ldots, n\}$ and $h \in \{1, \ldots, k\}$.

$$
\begin{aligned}
Z_{ih} &\sim \text{Bern}(\theta_h), \\
\eta_{ih} &\sim (1 - Z_{ih})\delta_0 + Z_{ih}\text{pMOM}(\psi = 0.5), \\
\boldsymbol{\lambda}_h &\sim \mathcal{N}_p(0, \mathbb{I}), \\
\boldsymbol{\epsilon}_i &\sim \mathcal{N}_p(0, \Sigma) \text{ with } \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2), \\
\sigma_j^2 &\sim \text{Unif}(0, 1) \; \forall j \in \{1, \ldots, p\}, \\
\mathbf{Y}_i &= \Lambda \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i.
\end{aligned}
$$

estimated and true structures, returning a value between from 0 (no similarity) to 1 (higher similarity).

Figure 2 reports the RV results across all the scenarios. The BFMAN model consistently achieves high RV coefficients, demonstrating excellent recovery of the latent structure. In Scenarios 1 and 2, the RV values exceed 0.95 for both $\Lambda\Lambda^T$ and $\boldsymbol{\eta}\boldsymbol{\eta}^T$, indicating near-perfect recovery and overperforming MGPS model. Even in the more challenging scenarios, which closely mimic real-world data complexities, our model consistently outperforms the MGPS model, while the RV index values remain close to 1. In Scenario 3, where $p >> n$, the RV coefficients are respectively 0.85 for $\boldsymbol{\eta}\boldsymbol{\eta}^T$ and 0.8 for $\Lambda\Lambda^T$. In Scenario 4, with a large sample size and a high-dimensional multivariate variable, both matrices achieve an RV index greater than 0.9. These results highlight the superior ability of the proposed model in accurately recovering the underlying data structure across different levels of sparsity and dimensionality.

As indicated in the previous section, the key strength of our model is its ability to identify the sparsity of the factor score matrix and to exploit it to determine the number of factors. Therefore, Figure 3 illustrates the distribution of the estimated probabilities, $\hat{\theta}$, of non-zero entries in $\boldsymbol{\eta}$ for each factor across the 50 replicates. The estimates match closely the true simlated value (in red), falling within the interquartile range, demonstrating accurate recovery of the sparsity structure. Our method not only accurately estimates the proportion of nonzero entries in the factor score matrix but also correctly assigns $\theta \approx 0$ to the additional factors $k^* = K - k$ that are used to estimated the model but are not part of the data-generating process, where K represents the upper bound used in model estimation.

In contrast, the MGPS model exhibits a tendency to overestimate the number of factors, as illustrated in Figure 4. This suggests that MGPS may require further tuning of its penalty parameters to better control shrinkage and avoid selecting spurious factors—especially in settings with small sample sizes and low dimensions, where overestimation is more likely.

These findings further corroborate the superior performance of our model in accurately selecting the true number of latent factors while preserving interpretability and sparsity. Furthermore, our factor selection procedure—based on discarding columns in $\boldsymbol{\eta}$ that are entirely or mostly zero—shows crucial advantages. In Scenarios 1 and 2, all true factors are correctly retained across all replicates. In Scenario 3, the overall
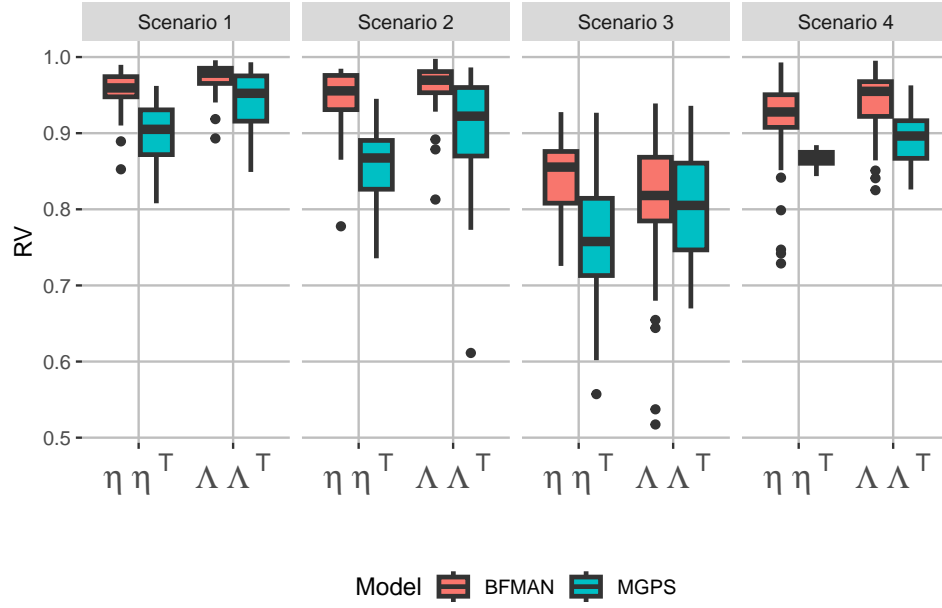
Figure 2: Results comparison: RV coefficient for $\boldsymbol{\eta}\boldsymbol{\eta}^T$ and $\Lambda\Lambda^T$ estimated with our BFMAN model (in red) and MGPS model (in blue) across the four simulated scenarios.
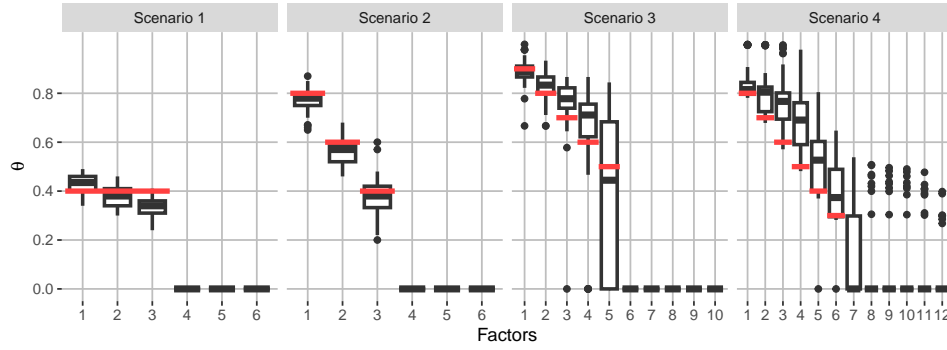

Figure 3: Results for BFMAN model. Distribution, over the 50 replicates, of the estimated probability of nonzero entries in the factor score matrix θ for each factor and for each of the four simulated scenario. The red line indicated the true value. For the factors where the red line is not reported, the true value is zero, indicating the absence of that factor in the data generating process.

identification remained accurate; however, our model occasionally underestimated the number of factors, particularly for Factor 5, which had the highest level of sparsity.

Although Scenario 4 slightly overestimates the number of factors in a few replicates, MGPS consistently shows a much greater overestimation, as shown in Figure 4). In all four scenarios, the MGPS model tends to estimate nearly twice the true number of simulated factors. These results further corroborate the performance of our model in accurately selecting the true number of latent factors while preserving interpretability and sparsity.
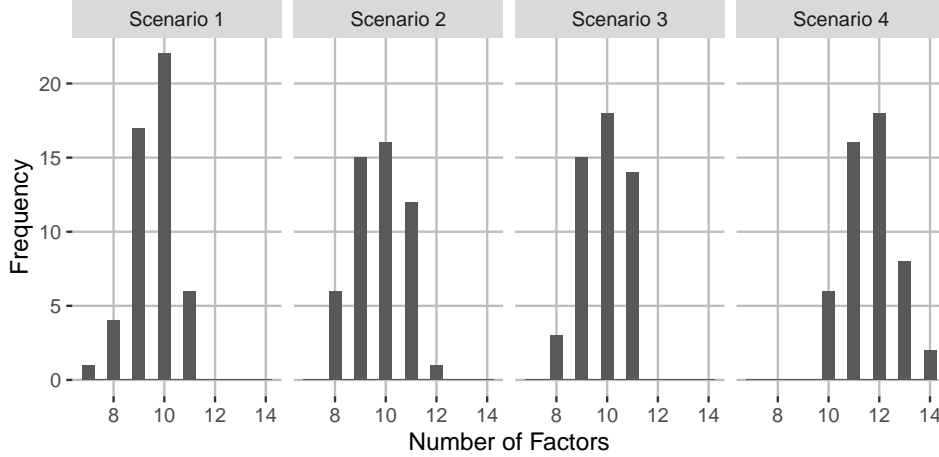

Figure 4: Results for MGPS model. Distribution, over the 50 replicates, of the probability of nonzero entries in the factor score matrix θ estimated for each factor and for each of the four simulated scenario.

4   Nutritional data and cardiovascular diseases

In this section, we apply our proposed model to the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), a large-scale, multi-site cohort designed to investigate the relationship between diet and cardiovascular risk factors in Hispanic/Latino population. The study includes 14,002 adults aged 18-74 years from four U.S. cities (Bronx, Chicago, Miami, and San Diego), recruited using a stratified two-stage probability sampling design as detailed in LaVange et al. [23].

From the original dataset, we exclude individuals who are on relevant medication therapy, have missing data, and/or present unreliable dietary questionnaires (e.g., extreme energy intakes, negative values for nutrient or food intake, or poor quality reported by interviewers) as described in De Vito and Avalos-Pacheco [13]. The resulting dataset includes 2,273 subjects and 53 nutrients. All nutrient values are log-transformed and standardize prior to analysis.

We first estimate the latent dimensionality using the strategy outlined in Section 2.4 starting with $K = 5\log(p)$, i.e., $K = 12$. Then we discard factors in which at least 85%

of the entries in the score matrix are zeros, yielding a final model with 6 factors. Then we rerun the factor analysis setting $k = 6$ to obtain the factor loading and the score matrix.

We then proceed to interpret the estimated factor loading matrix, reported in Figure 5. Following nutritional literature, we name each factor based on important loadings, i.e.$\lambda_{ih} \geq 0.3$. The first factor, namely *plant-based products*, is characterized by high loadings on insoluble and soluble dietary fiber, magnesium, natural folate, and phytic acid. The second factor, labeled *animal and vegeterian food*, reflects a complex, nutrient-dense pattern that incorporates a wide array of nutrients from both plant and animal sources. It includes various proteins, essential fatty acids (such as linoleic, linolenic, LCSFA, and LCMFA), cholesterol, trans fats, a wide range of minerals (including calcium, iron, zinc, and magnesium), and several vitamins (particularly the B-complex and vitamin E). The third factor, namely the *seafood* pattern, is defined by high factor loadings of omega-3 fatty acids, such as eicosapentaenoic acid (EPA), docosapentaenoic acid (DPA), and docosahexaenoic acid (DHA). The fourth factor, labeled *dairy products*, shows significant contributions from short- and medium-chain saturated fatty acids (SCSFA and MCSFA), calcium, and retinol. The fifth factor, representing *animal products*,is driven by high loadings on animal protein, vitamin B12, and vitamin D. Finally, the sixth factor, named *antioxidant products* factor, includes lutein and zeaxanthin, beta carotene, alpha-carotene, and vitamin C, that highlight the antioxidant content of the diet.

Then, we proceed on estimating their association with key cardiovascular risk factors: diabetes, high cholesterol, and hypertension. We fit a Bayesian logistic regression for each outcome, including confounders such as energy, physical activity, depressive symptoms (CESD score), ethnicity, gender, employment, years as US residency, marital status, income, education, alcohol and tobacco use.

The results reported in Figure 6 show that the plant-based product pattern is inversely associated with the risk of diabetes and high cholesterol, aligning with previous evidence on the protective effects of vegetarian diets against cardiometabolic diseases [20]. The animal and vegetarian pattern has a double trend: it is positively associated with hypertension but inversely associated with diabetes. This factor includes both beneficial components such as fiber, linoleic acid, and plant proteins, and potentially adverse components like cholesterol, animal protein, and saturated fats, contributing to these mixed associations. Finally, the pattern of seafood consumption is inversely associated with the risk of high cholesterol, supporting previous evidence that seafood consumption is protective against cardiovascular risk factors [1].

This application illustrates the practical advantages of our method in an epidemiological setting. By inducing sparsity in the factor score matrix, our model not only automatically estimates the number of meaningful latent dietary patterns but also enhances interpretability, allowing for a clearer identification of associations between diet and disease. These features make our approach particularly well suited to uncovering actionable insights in complex, high-dimensional health data.
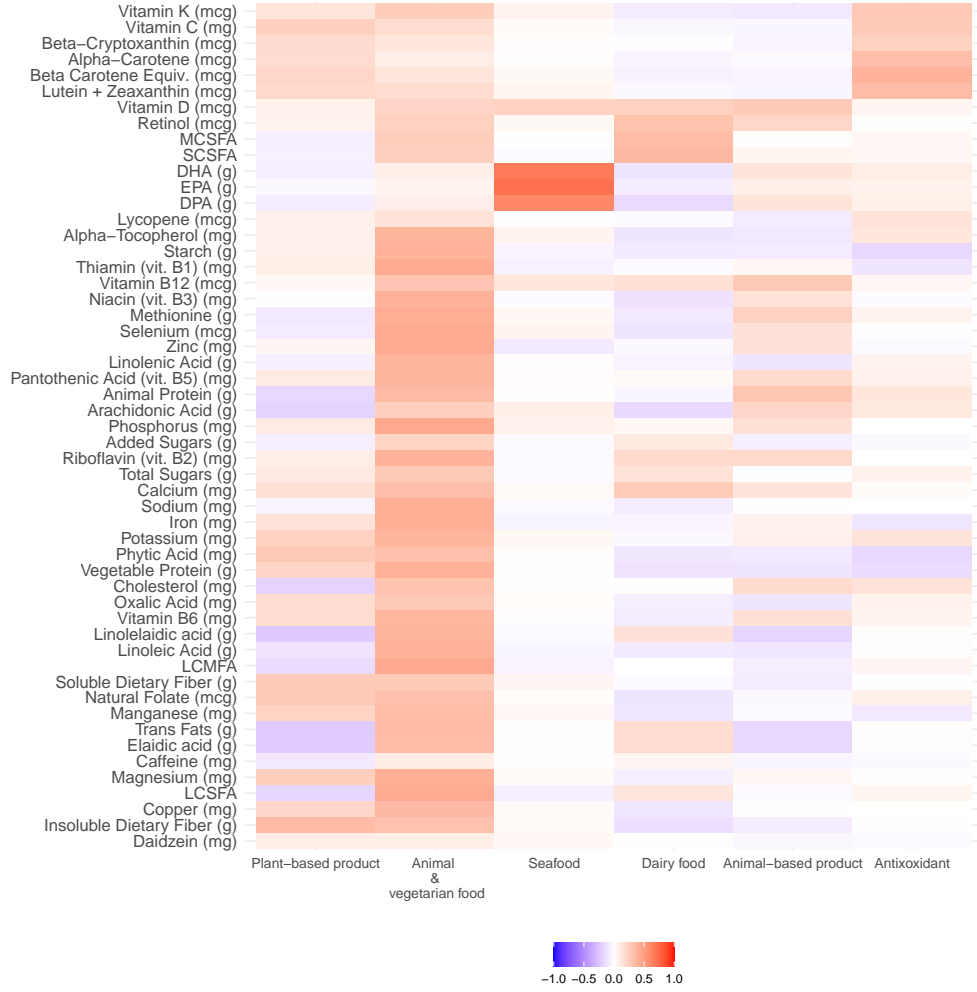
Figure 5: Heatmap of the factor loadings in the HCHS/SOL estimated with BFMAN.

5   Discussion

In this paper, we introduce a novel Bayesian factor model—the BFMAN—that shifts the focus from the commonly studied factor loadings to the factor scores. By incorporating a non-local mass prior on the factor scores, our BFMAN effectively captures individual-level heterogeneity and enforces sparsity in a principled manner. This leads to a richer and more realistic representation of how each subject contributes to latent structures, as demonstrated in our nutritional epidemiology application, where individual adherence to specific dietary patterns varied substantially.

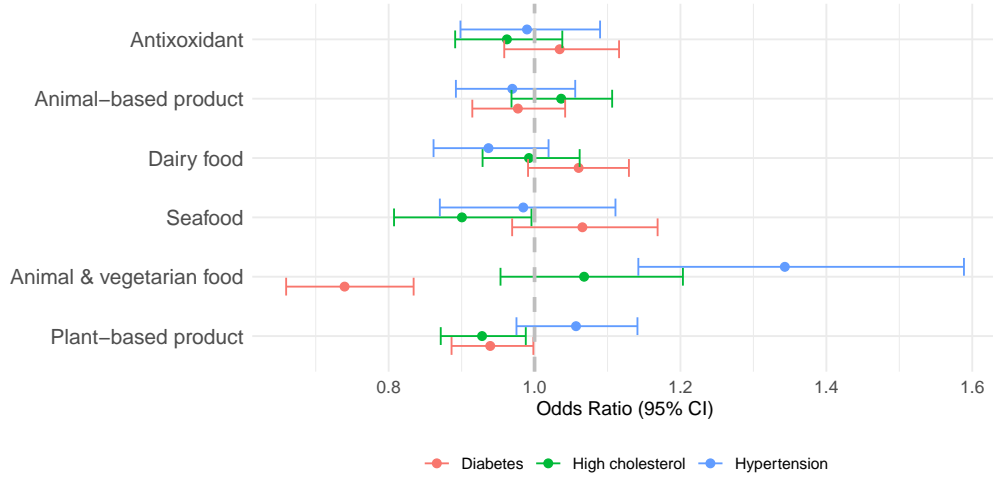Moreover, the sparsity plays a key role in (i) the methodological aspect of defin-

Figure 6: Odds ratio and their corresponding 95% credible intervals for each dietary pattern for the three CVD risk factors: diabetes, hypertension, and high cholesterol.

ing a robust and novel approach to determining the optimal number of factors, and (ii) the real-world application, allowing for a clearer interpretation of factor scores and highlighting which individuals meaningfully engage with certain latent patterns. In the nutritional setting, zero entries in the factor score matrix help identify individuals who do not follow particular dietary behaviors, thereby reducing noise and improving interpretability.

Our simulation results show that BFMAN consistently outperforms the widely used MGPS model. Across all scenarios, BFMAN achieve higher RV coefficients—indicating superior recovery of both the factor score and loading matrices—and provide more accurate estimation of the true number of latent factors. These findings reinforce the model's reliability and support its application in complex, high-dimensional settings.

In the real-world analysis of the HCHS/SOL study, BFMAN identified six interpretable dietary patterns, including plant-based foods, animal products, seafood, dairy products, antioxidants, and a nutrient-dense mixed pattern. The model revealed that only individuals with elevated consumption of processed foods showed a significantly increased probability of developing hypercholesterolemia. By leveraging sparsity in the factor scores, we were able to determine not only the most influential dietary patterns, but also the individuals who truly adhered to them, providing a clearer link between diet and health outcomes.

Our results underscore the central role of factor scores in both methodological innovation and real-world interpretation. While much of the existing literature has focused on imposing structure on the loadings, our work highlights how priors on the scores can yield powerful advantages. A related contribution by Bortolato and Canale [7] introduces adaptive shrinkage priors on factor scores for multi-study settings, further validating

the relevance of this direction. While different prior formulations may be suited to different applications, we believe the mass-nonlocal prior introduced here provides a flexible and interpretable foundation for modeling sparsity and heterogeneity in latent factor models.

Several extensions and generalization can be applied to the model. These include adapting BFMAN to dynamic or longitudinal settings, incorporating structured covariates into the prior on scores, and exploring alternative prior formulations for specific domains. More broadly, we hope this work inspires renewed attention on the modeling of factor scores, which hold rich and underutilized potential for inference and discovery across scientific disciplines.

Funding

References

[1] Aadland, E. K., Lavigne, C., Graff, I. E., Eng, Ø., Paquette, M., Holthe, A., Mellgren, G., Jacques, H., and Liaset, B. (2015). "Lean-seafood intake reduces cardiovascular lipid risk factors in healthy subjects: results from a randomized controlled trial with a crossover design." *The American journal of clinical nutrition*, 102(3): 582–592. 13

[2] Avalos-Pacheco, A., Rossell, D., and Savage, R. S. (2022). "Heterogeneous large datasets integration using Bayesian factor regression." *Bayesian analysis*, 17(1): 33–66. 2, 7

[3] Bai, J. and Ng, S. (2002). "Determining the number of factors in approximate factor models." *Econometrica*, 70(1): 191–221. 8

[4] Bai, R., Ročková, V., and George, E. I. (2021). "Spike-and-slab meets LASSO: A review of the spike-and-slab LASSO." *Handbook of Bayesian variable selection*, 81–108. 2

[5] Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). "Bayesian factor regression models in the "large p, small n" paradigm." *Bayesian Statistics*, 7: 733–742. 1

[6] Bhattacharya, A. and Dunson, D. B. (2011). "Sparse Bayesian infinite factor models." *Biometrika*, 98(2): 291–306. 1, 2, 4, 5, 8, 9

[7] Bortolato, E. and Canale, A. (2024). "Adaptive partition Factor Analysis." *arXiv preprint arXiv:2410.18939*. 15

[8] Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). "High-dimensional sparse factor modeling: applications in gene expression genomics." *Journal of the American Statistical Association*, 103(484): 1438–1456. 2

[9] Casa, A., O'Callaghan, T. F., and Murphy, T. B. (2022). "Parsimonious Bayesian factor analysis for modelling latent structures in spectroscopy data." *The Annals of Applied Statistics*, 16(4): 2417–2436. 1

[10] Castelló, A., Ascunce, N., Salas-Trejo, D., Vidal, C., Sanchez-Contador, C., Santamarina, C., Pedraz-Pingarron, C., Moreno, M. P., Perez-Gomez, B., Lope, V., et al. (2016). "Association between Western and Mediterranean dietary patterns and mammographic density." *Obstetrics & Gynecology*, 128(3): 574–581. 2, 8

[11] Chen, X., Rubin, K. H., and Sun, Y. (1992). "Social reputation and peer relationships in Chinese and Canadian children: A cross-cultural study." *Child development*, 63(6): 1336–1343. 1

[12] Daviglus, M. L., Pirzada, A., and Talavera, G. A. (2014). "Cardiovascular disease risk factors in the Hispanic/Latino population: lessons from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL)." *Progress in cardiovascular diseases*, 57(3): 230–236. 3

[13] De Vito, R. and Avalos-Pacheco, A. (2023). "Multi-study factor regression model: an application in nutritional epidemiology." *arXiv preprint arXiv:2304.13077*. 12

[14] De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2021). "Bayesian multistudy factor analysis for high-throughput biological data." *The annals of applied statistics*, 15(4): 1723–1741. 2

[15] DiStefano, C., Zhu, M., and Mindrila, D. (2019). "Understanding and using factor scores: Considerations for the applied researcher." *Practical assessment, research, and evaluation*, 14(1): 20. 2, 3

[16] Durante, D. (2017). "A note on the multiplicative gamma process." *Statistics & Probability Letters*, 122: 198–204. 5, 8

[17] Edefonti, V. et al. (2012). "Nutrient-based dietary patterns and the risk of head and neck cancer: a pooled analysis in the International Head and Neck Cancer Epidemiology consortium." *Annals of Oncology*, 23(7): 1869–1880. 1

[18] Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2024). "Sparse Bayesian factor analysis when the number of factors is unknown." *Bayesian Analysis*, 1(1): 1–31. 2, 7, 8

[19] Johnson, V. E. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(2): 143–170. 3

[20] Kahleova, H., Levin, S., and Barnard, N. D. (2018). "Vegetarian dietary patterns and cardiovascular disease." *Progress in cardiovascular diseases*, 61(1): 54–61. 13

[21] Kaiser, H. F. (1960). "The application of electronic computers to factor analysis." *Educational and psychological measurement*, 20(1): 141–151. 7, 8

[22] Kp, B. (1998). "Model selection and multimodel inference." *A practical information-theoretic approach*. 8

[23] LaVange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., Liu, K., Giachello, A., Lee, D. J., Ryan, J., et al. (2010). "Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos." *Annals of epidemiology*, 20(8): 642–649. 12

[24] Legramanti, S., Durante, D., and Dunson, D. B. (2020). "Bayesian cumulative shrinkage for infinite factorizations." *Biometrika*, 107(3): 745–752. 1, 5

[25] Lopes, H. F. and West, M. (2004). "Bayesian model assessment in factor analysis." *Statistica Sinica*, 14: 41–67. 1, 7

[26] McDonald, R. P. and Burr, E. (1967). "A comparison of four methods of constructing factor scores." *Psychometrika*, 32(4): 381–401. 2

[27] National Heart, Lung, and Blood Institute and others (2009). "Hispanic Community Health Study/Study of Latinos (HCHS/SOL)." *Bethesda, MD: NHLBI*. 3

[28] Preacher, K. J. and Merkle, E. C. (2012). "The problem of model selection uncertainty in structural equation modeling." *Psychological methods*, 17(1): 1. 8

[29] Robert, P. and Escoufier, Y. (1976). "A unifying tool for linear multivariate statistical methods: the RV-coefficient." *Journal of the Royal Statistical Society Series C: Applied Statistics*, 25(3): 257–265. 9

[30] Ročková, V. and George, E. I. (2016). "Fast Bayesian factor analysis via automatic rotations to sparsity." *Journal of the American Statistical Association*, 111(516): 1608–1622. 2, 7

[31] Roy, A., Lavine, I., Herring, A. H., and Dunson, D. B. (2021). "Perturbed factor analysis: Accounting for group differences in exposure profiles." *The annals of applied statistics*, 15(3): 1386. 1

[32] Schiavon, L., Canale, A., and Dunson, D. B. (2022). "Generalized infinite factorization models." *Biometrika*, 109(3): 817–835. 5

[33] Sorlie, P. D. et al. (2010). "Design and implementation of the Hispanic community health study/study of Latinos." *Annals of epidemiology*, 20(8): 629–641. 3

[34] Tucker, L. R. (1971). "Relations of factor score estimates to their use." *Psychometrika*, 36(4): 427–436. 2, 3

[35] Wang, X. V., Verhaak, R. G., Purdom, E., Spellman, P. T., and Speed, T. P. (2011). "Unifying gene expression measures from multiple platforms using factor analysis." *PloS One*, 6(3). 1