Escaping The Curse of Dimensionality in Bayesian Model-Based Clustering

Noirrit Kiran Chandra   NOIRRITCHANDRA@GMAIL.COM
Department of Mathematical Sciences
The University of Texas at Dallas
Richardson, TX, USA

Antonio Canale   ANTONIO.CANALE@UNIPD.IT
Department of Statistical Sciences
University of Padova
Padova, Italy

David B. Dunson   DUNSON@DUKE.EDU
Department of Statistical Science
Durham, NC, USA

Editor: Shakir Mohamed

Abstract

Bayesian mixture models are widely used for clustering of high-dimensional data with appropriate uncertainty quantification. However, as the dimension of the observations increases, posterior inference often tends to favor too many or too few clusters. This article explains this behavior by studying the random partition posterior in a non-standard setting with a fixed sample size and increasing data dimensionality. We provide conditions under which the finite sample posterior tends to either assign every observation to a different cluster or all observations to the same cluster as the dimension grows. Interestingly, the conditions do not depend on the choice of clustering prior, as long as all possible partitions of observations into clusters have positive prior probabilities, and hold irrespective of the true data-generating model. We then propose a class of latent mixtures for Bayesian clustering (Lamb) on a set of low-dimensional latent variables inducing a partition on the observed data. The model is amenable to scalable posterior inference and we show that it can avoid the pitfalls of high-dimensionality under mild assumptions. The proposed approach is shown to have good performance in simulation studies and an application to inferring cell types based on scRNAseq.

Keywords: Big data; Clustering; Dirichlet process; Exchangeable partition probability function; High dimensional; Latent variables; Mixture model.

1. Introduction

High-dimensional data $y_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}$ for $i = 1, \ldots, n$, with $p \gg n$, have become commonplace, and there is routinely interest in clustering observations $\{1, \ldots, n\}$ into groups. As an illustrative application, we consider single-cell RNA sequencing (scRNASeq) data; clustering of the cells based on their high-dimensional gene expression profiles produces potential cell types and provides information on heterogeneous cell populations of potential utility in disentangling carcinogenic processes. RNAseq data is an exemplary setting in which p is massive and clustering is crucial due to interest in inferring cell types. Although there are a variety of alternatives in the literature (see

Kiselev et al., 2019, for a review), we are particularly motivated to consider a Bayesian approach due to the potential for propagating uncertainty in inferring cell types. Additionally hierarchical Bayes models allow for borrowing of information in a principled manner in complicated scenarios.

Bayesian clustering is typically based on mixture models of the form:

$$y_i \overset{\text{iid}}{\sim} f, \quad f(y) = \sum_{h=1}^{k} \pi_h \mathcal{K}(y; \theta_h), \tag{1}$$

where $f(\cdot)$ is the marginal density of the data, k is the number of components, $\pi = (\pi_1, \ldots, \pi_k)^{\text{T}}$ are probability weights, $\mathcal{K}(y; \theta_h)$ is the density of the data within component h, and the number of clusters in data $y_1, \ldots, y_n$ corresponds to the number of occupied components $k_n \leq k$. When p is large and $y_i \in \mathbb{R}^p$, a typical approach chooses $\mathcal{K}(y; \theta_h)$ as a multivariate Gaussian density with a constrained and parsimonious covariance (see Bouveyron and Brunet-Saumard, 2014, for a review). Examples include matrices that are diagonal (Banfield and Raftery, 1993), block diagonal (Galimberti and Soffritti, 2013) or have a factor analytic representation (Ghahramani et al., 1996).

To avoid sensitivity to a pre-specified k, one can place a prior on k to induce a mixture of finite mixture model (Miller and Harrison, 2018; Frühwirth-Schnatter et al., 2021). Alternatively, a Bayesian nonparametric approach lets $k = \infty$, which allows k_n to increase without a bound as n increases. Under a Dirichlet process (Ferguson, 1973) k_n increases at a log rate in n, while for a Pitman-Yor process (Pitman and Yor, 1997) the rate is a power law.

Notably, Bayesian approaches can be used to intrinsically regularize the model complexity, as discussed by Jefferys and Berger (1992) exploiting the idea of a 'Bayesian Ockham razor'. While in many circumstances relying on the Bayesian Ockham razor is sufficient to choose the appropriate compromise between extremes, e.g. too many or too few clusters, in what follows we will argue that this is not the case in high-dimensional clustering. Indeed, when p is very large, the posterior distribution of k_n can concentrate on large values (Celeux et al., 2018); often the posterior mode of k_n is even equal to n so that each subject is assigned to its own singleton cluster. Consider, for example, the right panel of Figure S.1 in the supplementary materials, which displays the distribution of the mean number of clusters in 100 replicates of a simple simulation example where we generate samples of size $n = 10$ from a $p = 20$ variate normal distribution with mean zero and identity covariance. The boxplot, obtained running a standard Dirichlet process mixture, clearly shows how k_n is concentrated near n even for this moderate value of p. Celeux et al. (2018) conjectured that this aberrant behavior is mainly due to slow mixing of Markov chain Monte Carlo samplers. Frühwirth-Schnatter (2006) combat this problem with a specific prior elicitation criterion; this can be successful for $p \approx 100$, but calibration of hyperparameters is a delicate issue and scaling to $p > 1,000$ is problematic. Alternatively, one may attempt to cluster in lower dimensions via variable selection in clustering (Tadesse et al., 2005; Kim et al., 2006) or by introducing both global and variable-specific clustering indices for each subject, so that only certain variables inform global cluster allocation (Dunson, 2009).

However, we find these approaches complicated and to not address the fundamental question of what is causing the poor performance of Bayesian clustering for large p. To fill this gap, we provide theory showing that, as $p \to \infty$ with n fixed, the posterior can assign probability one to a trivial clustering - either with $k_n = 1$ and all subjects in one cluster or with $k_n = n$ and every subject in a different cluster. We further show that the conditions under which these degenerate limiting behaviors occur are satisfied for seemingly standard priors and multivariate Gaussian kernels. In a related

result for classification, Bickel and Levina (2004) showed that when p increases at a faster rate than n, the Fisher's linear discriminant rule is equivalent to randomly assigning future observations to the existing classes.

Our result has no relationship with the literature studying the posterior behavior of k_n as $n \to \infty$ for nonparametric Bayes procedures (Miller and Harrison, 2014; Cai et al., 2020; Ascolani et al., 2022). Indeed, our result holds for finite n regardless of the true data generating model, and has fundamentally different implications—in particular, that one needs to be extremely careful in specifying the kernel $\mathcal{K}(y; \theta)$ and prior for θ in the large p context. Otherwise, the true posterior can produce degenerate clustering results that have nothing to do with true structure in the data.

A key question is whether it is possible to define models that can circumvent this pitfall? We show that the answer is yes if clustering is conducted on the level of low-dimensional latent variables η_i underlying y_i. When the dimension of η_i is small relative to p, y_i provides abundant information about the lower-dimensional η_i even in low signal-to-noise settings in which each individual y_{ij} contributes very little information on its own. Hence, the curse of dimensionality can be turned into a blessing. This motivates a novel notion of a Bayesian oracle for clustering. The oracle has knowledge of the latent η_is and defines a Bayesian mixture model for clustering based on the η_is; the resulting oracle clustering posterior is thus free of the curse of dimensionality. We propose a particular latent mixture model structure, which can be shown to satisfy this oracle property and additionally leads to straightforward computation.

The article is organized as follows. Section 2 gives details on the limiting behavior of usual clustering methods based on (1). Section 3 introduces our mixture model on the latent variable level with prior specifications and posterior computation strategies. In Section 4, we introduce a Bayesian oracle clustering rule and show that our model achieves this oracle property as the dimension grows to infinity. Section 5 shows simulation studies illustrating how our proposed model learns the latent space with increasing dimensions and compares our method with some popular clustering methods. Section 6 considers an application to scRNASeq data, and Section 7 discusses the results. Proofs of the main results are included in the Appendix while additional technical results, simulation studies, and MCMC convergence diagnostics are reported in the supplementary materials.

2. Limiting Behavior of High-Dimensional Bayesian Clustering

Under a general Bayesian framework, model (1) becomes

$$y_i \sim f, \quad f(y) = \sum_{h \geq 1} \pi_h \mathcal{K}(y; \theta_h), \quad \theta_h \overset{\text{iid}}{\sim} P_0, \quad \{\pi_h\} \sim Q_0, \tag{2}$$

where $\{\pi_h\} \sim Q_0$ denotes a suitable prior for the mixture weights. Examples include stick-breaking (Sethuraman, 1994) constructions or a k-dimensional Dirichlet distribution with the dimension k given a prior following a mixture of finite mixtures (MFMs) approach.

Let $c_i \in \{1, \ldots, \infty\}$ denote the cluster label for subject i (for $i = 1, \ldots, n$), with $k_n = \#\{c_1, \ldots, c_n\}$ denoting the number of clusters represented in the sample. Conditionally on $c_i = h$, we can write $y_i \mid c_i = h \sim \mathcal{K}(y_i; \theta_h)$. Assume that n_j is the size of the jth cluster with $\sum_{j=1}^{k} n_j = n$. The posterior probability of observing the partition Ψ induced by the clusters $c_1, \ldots, c_n$ conditionally on the data $\mathcal{Y} = \{y_1, \ldots, y_n\}$ is

$$\Pi(\Psi \mid \mathcal{Y}) = \frac{\Pi(\Psi) \times \prod_{h \geq 1} \int \prod_{i:c_i=h} \mathcal{K}(y_i; \theta) \, dP_0(\theta)}{\sum_{\Psi' \in \mathscr{P}} \Pi(\Psi') \times \prod_{h \geq 1} \int \prod_{i:c_i=h} \mathcal{K}(y_i; \theta) \, dP_0(\theta)}, \tag{3}$$

where $\mathscr{P}$ is the space of all possible partitions of n data points into clusters. The numerator of (3) is the product of the prior probability of Ψ multiplied by a product of the marginal likelihoods of the observations within each cluster. The denominator is a normalizing constant consisting of an enormous sum over $\mathscr{P}$. Assuming exchangeability, the prior probability of any partition of n subjects into k_n groups depends only on $n_1, \ldots, n_{k_n}$ and k_n through an exchangeable partition probability function (EPPF). The latter is available in closed form for popular choices of Q_0, including the Dirichlet process, Pitman-Yor process and certain MFMs.

The posterior (3) forms the basis for Bayesian inferences on clusterings in the data, while providing a characterization of uncertainty. We are particularly interested in how this posterior behaves in the case in which $y_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}$ are high-dimensional so that p is very large. To study this behavior theoretically, we consider the limiting case as $p \to \infty$ while keeping n fixed. This setting is quite appropriate in our motivating applications to genomics, as there is essentially no limit to the number of variables one can measure on each study subject, while the number of study subjects is often small to moderate.

In such settings with enormous p and modest n, we would like the true posterior distribution in (3) to provide a realistic characterization of clusters in the data. However, this is commonly not the case and as p increases the posterior distribution can have one of two trivial degenerate limits. In particular, depending on the choice of kernel density $\mathcal{K}(\cdot; \theta)$ and the base measure P_0 for the θ_h's, the posterior assigns probability one to either the $k_n = 1$ clustering that places all subjects in the same cluster or the $k_n = n$ clustering that places all subjects in different clusters. We derive sufficient conditions behind such aberrant behaviors as formalized in the following theorem.

Theorem 1 *Let $y_1, \ldots, y_n$ denote p-variate random vectors with joint probability measure $\mathbb{P}_0^p$. Let Ψ denote the partition induced by the cluster labels $c_1, \ldots, c_n$, and let $c_1', \ldots, c_n'$ denote a new set of cluster labels obtained from $c_1, \ldots, c_n$ by merging an arbitrary pair of clusters, with Ψ' the related partition. Assume $Q_0(\pi_h > 0$ for all $h = 1, \ldots, n) > 0$. If*

$$\limsup_{p \to \infty} \frac{\prod_{h \geq 1} \int \prod_{i:c_i=h} \mathcal{K}(y_i; \theta) \, \mathrm{d}P_0(\theta)}{\prod_{h \geq 1} \int \prod_{i:c_i'=h} \mathcal{K}(y_i; \theta) \, \mathrm{d}P_0(\theta)} = 0$$

in $\mathbb{P}_0^p$-probability, then $\lim_{p \to \infty} \Pi(c_1 = \cdots = c_n \mid \mathcal{Y}) = 1$ in $\mathbb{P}_0^p$-probability. Else if

$$\liminf_{p \to \infty} \frac{\prod_{h \geq 1} \int \prod_{i:c_i=h} \mathcal{K}(y_i; \theta) \, \mathrm{d}P_0(\theta)}{\prod_{h \geq 1} \int \prod_{i:c_i'=h} \mathcal{K}(y_i; \theta) \, \mathrm{d}P_0(\theta)} = \infty$$

in $\mathbb{P}_0^p$-probability, then $\lim_{p \to \infty} \Pi(c_1 \neq \cdots \neq c_n \mid \mathcal{Y}) = 1$ in $\mathbb{P}_0^p$-probability.

The condition on Q_0 is equivalent to saying k_n has positive prior mass on $1, \ldots, n$, which is extremely mild and holds for essentially any prior in the literature, including the Dirichlet process, Pitman-Yor process and suitable MFMs that do not pre-specify $k < n$. Changing the condition to $Q_0(\pi_h > 0$ for all $h = 1, \ldots, k) > 0$ with $k < n$, i.e. using a finite mixture model, leads to similar results. Specifically, if the first condition in Theorem 1 holds, then also for finite mixtures we will have a single occupied cluster comprising all samples. If the opposite condition holds, instead, then all of the k mixture components will be occupied. Both results are trivial modifications of the proof of Theorem 1.

Theorem 1 has disturbing implications in terms of the behavior of posterior distributions for Bayesian clustering in large p settings. Notably, the theorem is stated for very general kernel density $\mathcal{K}$ and base measure P_0, and the behavior is controlled by the induced marginal likelihoods obtained in integrating out the kernel parameter θ with respect to P_0. Clearly it is the joint effect of $\mathcal{K}$ and P_0 that leads to the two limiting results and thus it is not immediate to convert the statement of the theorem to simple conditions on $\mathcal{K}$ and P_0. However, as we will discuss in detail, we can argue that these conditions are related to the two extreme situations of complex over-parametrized models having insufficiently informative priors and simpler models equipped with more informative priors. To be more precise, consider the important and widely used special case corresponding to a location-scale mixture of multivariate Gaussian kernels:

$$y_i \overset{\text{iid}}{\sim} f, \quad f(y) = \sum_{h \geq 1} \pi_h \mathrm{N}_p(y; \mu_h, \Sigma_h), \quad (\mu_h, \Sigma_h) \overset{\text{iid}}{\sim} P_0, \tag{4}$$

where $\mathrm{N}_p(\mu, \Sigma)$ denotes the p-dimensional multivariate normal density with mean μ and covariance matrix Σ. We give two practical examples of Theorem 1 in Corollary 2 and 3. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the smallest and largest eigenvalues of a positive definite matrix A and $Y = [y_1, \ldots, y_n]^{\mathrm{T}}$ be the complete $n \times p$ data matrix. Assume, for the true data generating distribution on the data $\mathcal{Y}$,

(A0) $\liminf_{p \to \infty} \lambda_{\min}(YY^{\mathrm{T}})/p > 0$ in $\mathbb{P}_0^p$-probability and $\|y_i\|^2 \leq Kp$ for some $K > 0$ in $\mathbb{P}_0^p$-probability.

Condition (A0) is extremely mild ensuring that the data are non-atomic and is satisfied for any continuous distribution with finite second order moments. Letting $\mathrm{IW}(\nu, \Lambda)$ denote an inverse-Wishart distribution with degrees of freedom ν and scale matrix Λ, we have the following:

Corollary 2 *Assume that the model* (4) *is used to cluster $\mathcal{Y}$ with $\Sigma_h \overset{\text{iid}}{\sim} \mathrm{IW}(\nu_0, \Lambda_0)$ and $\mu_h \mid \Sigma_h \overset{\text{ind}}{\sim} \mathrm{N}_p(\mu_0, \kappa_0^{-1}\Sigma_h)$, with $\|\mu_0\|^2 = O(p)$, $\kappa_0 = O(1)$, $\nu_0 = p + c$ for some fixed constant $c \geq 0$, $\|\Lambda_0\|_2 = O(1)$ and $\|\Lambda_0\|_2/\lambda_{\min}(\Lambda_0) = O(1)$. Under (A0) on the data $\mathcal{Y}$, $\Pi(c_1 = \cdots = c_n \mid \mathcal{Y}) \to 1$ in $\mathbb{P}_0^p$-probability.*

Corollary 3 *Assume that the model* (4) *is used to cluster $\mathcal{Y}$ with $\Sigma_h = \Sigma$ across all clusters, and let $\Sigma \sim \mathrm{IW}(\nu_0, \Lambda_0)$ and $\mu_h \mid \Sigma \overset{\text{iid}}{\sim} \mathrm{N}_p(\mu_0, \kappa_0^{-1}\Sigma)$, with $\|\mu_0\|^2 = O(p)$, $\kappa_0 = O(1)$, $\nu_0 > p - 1$ such that $\lim_{p \to \infty} \nu_0/p > 1$, and $\|\Lambda_0\|_2 = O(1)$ with $\|\Lambda_0\|_2/\lambda_{\min}(\Lambda_0) = O(1)$. Under (A0) on the data $\mathcal{Y}$, $\Pi(c_1 \neq \cdots \neq c_n \mid \mathcal{Y}) \to 1$ in $\mathbb{P}_0^p$-probability.*

Bayesian model-based clustering routinely uses these setups for the kernel parameters and priors (Fruhwirth-Schnatter et al., 2019). The conditions on μ_0 and κ_0 ensure that the Euclidean norm of the prior mean grows with p in the same order as the data $\{y_i\}$, and the conditions on the scale matrix Λ_0 imply that the second moments of the location components are a priori bounded away from 0 while being finite; similar assumptions appear in Yao et al. (2022) in a study on high-dimensional Gaussian location mixture models. In terms of the degrees of freedom parameter ν_0, in Corollary 2 the ratio ν_0/p is 1 in the limit inducing a heavy tailed prior predictive distribution, whereas in Corollary 3 a thinner tailed prior predictive is induced. Corollaries 2 and 3 show that, for mixtures of Gaussians, we can obtain directly opposite aberrant limiting behavior of the posterior depending on the kernel and prior for the kernel parameters but not on the clustering prior Q_0.

Corollary 2 considers the case in which we allow flexible cluster-specific means and dispersion matrices, under typical conjugate multivariate normal IW priors. This case can be viewed as a complex over-parametrized model as p increases and to combat this complexity the Bayesian Ockham razor (Jefferys and Berger, 1992) automatically assigns probability one to grouping all n individuals into the same cluster effectively simplifying the model. At the other extreme, covered by Corollary 3, we assume an under-parametrized relatively simplistic model structure in which all the mixture components have a common covariance. In this case, due perhaps to the relatively concentrated prior predictive distribution, there is not enough penalty for introducing new clusters, and all individuals are assigned to their own singleton cluster. These results hold regardless of the true data-generating model, and in particular the true clustering structure.

These theoretical results demonstrate that in high dimensions it is crucial to choose a good compromise between parsimony and flexibility in Bayesian model-based clustering. Otherwise, the true posterior distribution of clusterings in the data can have effectively no relationship whatsoever with true clustering structure in the data. Although we focus on the limiting case as $p \to \infty$, we conjecture that this behavior can 'kick in' quickly as p increases, based on intuition built through our proofs and through comprehensive simulation experiments.

3. Latent Factor Mixture

To overcome the problems discussed in Section 2, we propose a general class of latent factor mixture models defined as

$$y_i \sim f(y_i; \eta_i, \psi), \quad \eta_i \sim \sum_{h=1}^{\infty} \pi_h \mathcal{K}(\eta_i; \theta_h), \tag{5}$$

where $\eta_i = (\eta_{i1}, \ldots, \eta_{id})^{\mathrm{T}}$ are d-dimensional latent variables, $d < n$ is fixed and not growing with p, $f(\cdot; \eta_i, \psi)$ is the density of the observed data conditional on the latent variables and measurement parameters ψ and $\mathcal{K}(\cdot; \theta)$ is a d-dimensional kernel density.

Under (5), the high dimensional data being collected are assumed to provide error-prone measurements of an unobserved lower-dimensional set of latent variables η_i on subject i. As a canonical example, we focus on a linear Gaussian measurement model with a mixture of Gaussians for the latent factors:

$$y_i \sim \mathrm{N}_p(\Lambda \eta_i, \Sigma), \quad \eta_i \sim \sum_{h=1}^{\infty} \pi_h \mathrm{N}_d(\mu_h, \Delta_h), \quad \{\pi_h\} \sim Q_0, \tag{6}$$

where $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ is a $p \times p$ diagonal matrix, and Λ is a $p \times d$ matrix of factor loadings. The key idea is to incorporate all the cluster-specific parameters at the latent data level instead of the observed data level to favor parsimony. The latent variables are supported on a lower-dimensional hyperplane, and we map from this hyperplane to the observed data level through multiplication by a factor loadings matrix and then adding Gaussian noise. We could further simplify the model by assuming $\Sigma = \sigma^2 I_p$ instead of Σ diagonal; we find it appealing to allow the different y_{ij}s to have varying measurement error variances and hence focus mainly on the unconstrained diagonal case. We refer to model (6) as a LAtent Mixture for Bayesian (Lamb) clustering. The model is highly flexible at the latent variable level, allowing differences across clusters in the mean through μ_h and the shape, size, and orientation through Δ_h.

With different motivations, Galimberti et al. (2009); Baek et al. (2010); Montanari and Viroli (2010) proposed similar latent factor mixture models as (6) albeit with additional constraints.

Moreover, they fixed the number of clusters, used EM algorithms for model fitting and assessed goodness-of-fit via information criteria.

The proposed Lamb model has fundamentally different implications from the popular mixture of factor analyzers of Ghahramani et al. (1996), which defines a mixture of multivariate Gaussians at the p-dimensional observed data level having cluster-specific means and covariance matrices, with the dimension of the covariances reduced via a factor model. In contrast, we are effectively learning a common affine space within which we can define a simple location-scale mixture of Gaussians. Our approach not only massively reduces the effective number of parameters for large p, but also provides a successful compromise between the two extreme cases of Section 2.

3.1 Prior Specifications

In order to accommodate very high-dimensional data, with $p \gg n$, it is important to reduce the effective number of parameters in the $p \times d$ loadings matrix Λ. There is a rich literature on sparse factor modeling using a variety of shrinkage or sparsity priors for Λ; for example, refer to Bhattacharya and Dunson (2011) and the references therein. Although a wide variety of shrinkage priors for Λ are appropriate, we focus on a Dirichlet-Laplace prior (Bhattacharya et al., 2015), as it is convenient both computationally and theoretically. On a p-dimensional vector θ, the Dirichlet-Laplace prior with parameter a, denoted by $\mathrm{DL}(a)$, can be specified in the following hierarchical manner

$$\theta_j \mid \phi, \tau \overset{\text{ind}}{\sim} \mathrm{N}(0, \psi_j \phi_j^2 \tau^2), \quad \psi_j \overset{\text{iid}}{\sim} \mathrm{Exp}(1/2), \quad \phi \sim \mathrm{Dir}(a, \ldots, a), \quad \tau \sim \mathrm{Ga}(pa, 1/2), \quad (7)$$

where θ_j is the j-th element of θ, ϕ is a vector of the same length as θ, $\mathrm{Exp}(a)$ is an exponential distribution with mean $1/a$, $\mathrm{Dir}(a_1, \ldots, a_p)$ is the p-dimensional Dirichlet distribution, and $\mathrm{Ga}(a, b)$ is the gamma distribution with mean a/b and variance a/b^2. To impose shrinkage uniformly on its elements a priori, we let $\mathrm{vec}(\Lambda) \sim \mathrm{DL}(a)$ where $\mathrm{vec}(\Lambda)$ denotes the vectorization of Λ. We then choose inverse-gamma priors for the residual variances, $\sigma_j^{-2} \overset{\text{iid}}{\sim} \mathrm{Ga}(a_\sigma, b_\sigma)$.

For the prior Q_0 on the cluster weights $\{\pi_h\}$, for convenience in computation, we use a stick-breaking prior (Ishwaran and James, 2001) derived from a Dirichlet process, which has concentration parameter α impacting the induced prior on the number of clusters. To allow greater data adaptivity, we choose a $\mathrm{Ga}(a_\alpha, b_\alpha)$ prior for α. We assign the cluster-specific means and covariances $\{\mu_h, \Delta_h\}$ independent multivariate normal inverse-Wishart priors with location μ_0, precision parameter κ_0, inverse scale matrix Δ_0 and degrees of freedom ν_0. Our hierarchical Bayesian model for the η_is can be equivalently represented as

$$\eta_i \mid \mu_i, \Delta_i \overset{\text{ind}}{\sim} \mathrm{N}_d(\mu_i, \Delta_i), \quad \mu_i, \Delta_i \mid G \overset{\text{iid}}{\sim} G, \quad G \sim \mathrm{DP}(\alpha, G_0), \quad \alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha), \quad (8)$$

where $G_0 = \mathrm{NIW}(\mu_0, \Delta_0, \kappa_0, \nu_0)$. The gamma prior on the concentration parameter α is commonly adopted in many applications motivated by Escobar and West (1995). The role of this hyperprior and the elicitation of its hyperparameters has been carefully studied by Frühwirth-Schnatter and Malsiner-Walli (2019), and Ascolani et al. (2022) recently showed the prior to have a crucial impact on consistency in estimating the number of clusters.

In practice, the latent variable dimension d is unknown. Potentially we could put a prior on d and implement a reversible-jump type (Richardson and Green, 1997) Markov chain Monte Carlo (MCMC) algorithm, which may lead to inefficient and expensive computation. Instead we adopt a principal component analysis (PCA) based empirical Bayes type approach (Bai and Ng, 2008) to set d to a large value learned from the data and let the prior shrink the extra columns on Λ. We use the

augmented implicitly restarted Lanczos bidiagonalization algorithm (Baglama and Reichel, 2005) to obtain approximate singular values and eigenvectors, and choose the smallest $\widehat{d}$ explaining at least 95% of the variability in the data. This strategy substantially simplifies the computation. The left and right singular values are used to initialize the Λ and η_i's in our MCMC implementation. We initialize our cluster membership indicators using k-means.

For all the simulation experiments of the next section and the application, we choose $\mu_0 = 0$ and $\Delta_0 = \xi^2 I_d$ for a scalar $\xi^2 > 0$. To specify weakly informative priors, we set $\xi^2 = 20$, $\kappa_0 = 0.001$, $\nu_0 = \widehat{d} + 50$, $a_\alpha = b_\alpha = 0.1$ as the hyper-parameters of the DP mixture prior; $a_\sigma = 1$, $b_\sigma = 0.3$ as the hyper-parameters of the prior on the residual variances. We set $a = 0.5$ as the Dirichlet-Laplace parameter following the recommendation of Bhattacharya et al. (2015).

3.2 Posterior Sampling

For posterior computation we use a Gibbs sampler defined by the following steps.

Step 1 Letting λ_j^{T} denote the jth row of Λ, $\eta = [\eta_1, \ldots, \eta_n]^{\mathrm{T}}$, $D_j = \tau^2 \mathrm{diag}(\psi_{j1}\phi_{j1}^2, \ldots, \psi_{jd}\phi_{jd}^2)$ and $y^{(j)} = (y_{1j}, \ldots, y_{nj})^{\mathrm{T}}$, for $j = 1, \ldots, p$ sample

$$(\lambda_j \mid -) \sim \mathrm{N}_d \left\{ (D_j^{-1} + \sigma_j^{-2}\eta^{\mathrm{T}}\eta)^{-1}\eta^{\mathrm{T}}\sigma_j^{-2}y^{(j)}, \ (D_j^{-1} + \sigma_j^{-2}\eta^{\mathrm{T}}\eta)^{-1} \right\}.$$

Step 2 Update the Δ_h's from the inverse-Wishart distributions $\mathrm{IW}\left(\widehat{\psi}_h, \widehat{\nu}_h\right)$ where

$$\bar{\eta}_h = \frac{1}{n_h} \sum_{i:c_i=h} \eta_i, \quad \widehat{\nu}_h = \nu_0 + n_h,$$

$$\widehat{\psi}_h = \xi^2 I_d + \sum_{i:c_i=h}(\eta_i - \bar{\eta}_h)(\eta_i - \bar{\eta}_h)^{\mathrm{T}} + \frac{\kappa_0 n_h}{\kappa_0 + n_h}\bar{\eta}_h\bar{\eta}_h^{\mathrm{T}}.$$

Due to conjugacy, the location parameters μ_h's can be integrated out of the model.

Step 3 Sample the latent factors, for $i = 1, \ldots, n$, from

$$(\eta_i \mid -) \sim \mathrm{N}_d \left\{ \Omega_h \rho_h, \Omega_h + \Omega_h(\widehat{\kappa}_{h,-i}\Delta_h)^{-1}\Omega_h \right\},$$

where $n_{h,-i} = \sum_{j \neq i} \mathbb{1}(c_j = h)$, $\widehat{\kappa}_{h,-i} = \kappa_0 + n_{h,-i}$, $\bar{\eta}_{h,-i} = \frac{1}{n_{h,-i}} \sum_{j:c_j=h, j \neq i} \eta_i$, $\widehat{\mu}_{h,-i} = \frac{n_{h,-i}\bar{\eta}_{h,-i}}{n_{h,-i}+\kappa_0}$, $\rho_h = \Lambda^{\mathrm{T}}\Sigma^{-1}Y_i + \Delta_h^{-1}\widehat{\mu}_{h,-i}$ and $\Omega_h^{-1} = \Lambda^{\mathrm{T}}\Sigma^{-1}\Lambda + \Delta_h^{-1}$.

Step 4 Sample the cluster indicator variables $c_1, \ldots, c_n$ with probabilities

$$\Pi(c_i = h \mid -) \propto \begin{cases} n_{h,-i} \int \mathrm{N}_d(\eta_i; \mu_h, \Delta_h)\mathrm{d}\Pi(\mu_h, \Delta_h \mid c_{-i}, \eta_{-i}) \text{ for } h \in c_{-i}, \\ \alpha \int \mathrm{N}_d(\eta_i; \mu_h, \Delta_h)\mathrm{d}\Pi(\mu_h, \Delta_h) \text{ for } h \notin c_{-i}. \end{cases} \tag{9}$$

where $\eta_{-i} = \{\eta_j : j \neq i\}$ and $c_{-i} = \{c_j : j \neq i\}$. Due to conjugacy the above integrals are analytically available.

Step 5 Let r be the number of unique c_i's. Following West (1992), first generate $\varphi \sim \mathrm{Beta}(\alpha + 1, n)$, evaluate $\pi/(1 - \pi) = (a_\alpha + r - 1)/\{n(b_\alpha - \log\varphi)\}$ and generate

$$\alpha \mid \varphi, r \sim \begin{cases} \mathrm{Ga}(\alpha + r, b_\alpha - \log\varphi) \text{ with probability } \pi, \\ \mathrm{Ga}(\alpha + r - 1, b_\alpha - \log\varphi) \text{ with probability } 1 - \pi. \end{cases}$$

Step 6 For $j = 1, \ldots, p$ sample σ_j^2 from $\mathrm{Ga} \left\{ a_\sigma + n/2, b_\sigma + \sum_{i=1}^{n} (y_{ij} - \lambda_j^{\mathrm{T}} \eta_i)^2 / 2 \right\}$.

Step 7 Update the hyper-parameters of the Dirichlet-Laplace prior through:

(i) For $j = 1, \ldots, p$ and $h = 1, \ldots d$ sample $\widetilde{\psi}_{jh}$ independently from an inverse-Gaussian $\mathrm{iG}(\tau \phi_{jh} / |\lambda_{jh}|, 1)$ distribution and set $\psi_{jh} = 1/\widetilde{\psi}_{jh}$.

(ii) Sample the full conditional posterior distribution of τ from a generalized inverse Gaussian $\mathrm{giG}\{dp(1-a), 1, 2\sum_{j,h} |\lambda_{jh}|/\phi_{jh}\}$ distribution.

(iii) To sample $\phi \mid \Lambda$, draw T_{jh} independently with $T_{jh} \sim \mathrm{giG}(a-1, 1, 2|\lambda_{jh}|)$ and set $\phi_{jh} = T_{jh}/T$ with $T = \sum_{jh} T_{jh}$.

This simple Gibbs sampler sometimes gets stuck in local modes; a key bottleneck is the exploration Step 4. Therefore, we adopt the split-merge MCMC procedure proposed by Jain and Neal (2004); the authors note that the Gibbs sampler is useful in moving singleton samples between clusters while the split-merge algorithm makes major changes. Hence, we randomly switch between Gibbs and split-merge updates. The split-merge algorithm makes smart proposals by performing restricted Gibbs scans of the same form as in (9).

From the posterior samples of c_i's, we compute summaries following Wade and Ghahramani (2018). Our point estimate is the partition visited by the MCMC sampler that minimizes the posterior expectation of the Binder loss (Binder, 1978) exploiting the posterior similarity matrix obtained from the different sampled partitions.

The sampling algorithm can be easily modified for other priors on Λ having a conditionally Gaussian representation, with Step 7 modified accordingly. For example, we could use horseshoe (Carvalho et al., 2009), increasing shrinkage priors (Bhattacharya and Dunson, 2011; Legramanti et al., 2020; Schiavon et al., 2021), or the fast factor analysis prior (Ročková and George, 2016). Similarly, alternative priors for $\{\pi_h\}$, such as Pitman and Yor (1997) or Miller and Harrison (2018), can be adopted with minor modifications in Steps 4 and 5.

4. Properties of the Latent Mixture for Bayesian Clustering Method

4.1 Bayes Oracle Clustering Rule

We first define a Bayes oracle clustering rule where the observed data follow the distribution in model (5), that is, the high dimensional y_i's provide error-prone measurements on unobserved lower-dimensional latent variables η_i's on subject i, and we assume the oracle has knowledge of the exact values of the latent variables $\{\eta_{0i}\}$, where η_{0i}'s are d-dimensional latent vectors. Given this knowledge, the oracle can define any Bayesian mixture model to induce a posterior clustering of the data, which is not affected by the high-dimensionality of the problem. This leads to the distribution over the space of partitions in the following definition.

Definition 4 *Let $\eta_0 = \{\eta_{01}, \ldots, \eta_{0n}\}$ be the true values of the unobserved latent variables corresponding to each data point. The following mixture model is assumed to cluster η_0*

$$\eta_{0i} \sim \sum_{h=1}^{\infty} \pi_h \mathcal{K}(\eta_{0i}; \theta_h), \quad \{\pi_h\} \sim Q_0, \quad \theta_h \stackrel{\mathrm{iid}}{\sim} G_0.$$

Then the oracle probability of clustering is defined as

$$\Pi(\Psi \mid \eta_0) = \frac{\Pi(\Psi) \times \int \prod_{h \geq 1} \prod_{i:c_i=h} \mathcal{K}(\eta_{0i}; \theta_h) \, \mathrm{d}G_0(\theta_h)}{\sum_{\Psi' \in \mathscr{P}} \Pi(\Psi') \times \int \prod_{h \geq 1} \prod_{i:c'_i=h} \mathcal{K}(\eta_{0i}; \theta_h) \, \mathrm{d}G_0(\theta_h)}. \tag{10}$$

Probability (10) expresses the oracles' uncertainty in clustering if the clustering model could have been applied on the true latent factors. This is a gold standard in being free of the curse of dimensionality through using the oracles' knowledge of the true latent variables, but we make no claims about the relationship between the oracle posterior and any 'true' clustering. Under the framework of Section 3, the high-dimensional measurements on each subject provide information on these latent variables, with the clustering done on the latent variable level. Ideally, we would get closer to the oracle partition probability under the proposed method as p increases, turning the curse of dimensionality into a blessing. We show that this is indeed the case in Section 4.3.

To this end, we assume the oracle uses a location mixture of Gaussians with a common covariance matrix. We assume the following mixture distribution on η_{0i}'s, independent non-informative Jeffreys prior for the common covariance and arbitrary prior Q_0 on the mixture probabilities:

$$\eta_i \overset{\text{iid}}{\sim} \sum_{h=1}^{\infty} \pi_h \mathrm{N}_d(\mu_h, \Delta), \quad \mu_h \mid \Delta \overset{\text{iid}}{\sim} \mathrm{N}_d(0, \kappa_0^{-1}\Delta), \quad \Delta \propto |\Delta|^{-\frac{d+1}{2}}, \quad \{\pi_h\} \sim Q_0. \tag{11}$$

For $d < n$, the oracle rule is well defined for the Jeffreys prior on Δ. Note that the marginal Jeffrey's prior is free of any hyperparameter.

4.2 Assumptions on Data and Prior Specifications

In this section, we show that the posterior probability on the space of partitions induced by the proposed model converges to the oracle probability as $p \to \infty$ in expectation under appropriate conditions on the data generating process and the prior. We assume that the residual error variances σ_j^2's are the same having true common value σ_0^2 for all $j = 1, \ldots, p$. Our result is based on the following assumptions on $\mathbb{P}_0^p$, the true data-generating distribution of $y_1, \ldots, y_n$:

(C1) $y_i \overset{\text{ind}}{\sim} \mathrm{N}_p(\Lambda_0 \eta_{0i}, \sigma_0^2 I_p)$, for each $i = 1, \ldots, n$;

(C2) $\lim_{p \to \infty} \left\| \frac{1}{p} \Lambda_0^{\mathrm{T}} \Lambda_0 - M \right\|_2 = 0$ where M is a $d \times d$ positive-definite matrix;

(C3) $\sigma_L^2 < \sigma_0^2 < \sigma_U^2$ where σ_L^2 and σ_U^2 are known constants;

(C4) $\|\eta_{0i}\| = O(1)$ for each $i = 1, \ldots, n$.

Condition (C1) corresponds to the conditional likelihood of y_i given η_i being correctly specified and the data containing increasing information on the latent factors as p increases. This increasing information assumption is extremely mild; indeed, each individual y_{ij} can be very noisy and provide minimal information about η_i and there will still be a build up of information across $j = 1, \ldots, p$ as long as the additional variables are not completely uncorrelated with the target latent factors. In fact, we have a build up of information even when a proportion of the factor loadings are exactly zero, the factor loadings are very small relative to the residual variance, and the residuals are heavy-tailed. We illustrate this empirically with a simple simulation study in Section S.4.1 of the supplementary

materials. Condition (C2) ensures that Λ_0 is not *ill-conditioned* and its spectral norm does not increase too fast with respect to p since the highest and lowest eigenvalues of $\Lambda_0^{\mathrm{T}}\Lambda_0$ grow in $O(p)$. Related but much stronger conditions appear in the factor modeling (Fan et al., 2008, 2011) and massive covariance estimation literature (Pati et al., 2014). We allow the columns of Λ_0 to be non-orthogonal with varying average squared values which is expected in high-dimensional studies. Condition (C3) bounds the variance of the observed y_is and (C4) is a weak assumption ensuring that the latent variables do not depend on n or p. Additionally, we assume that the latent dimension d is known.

Although we use a stick-breaking prior on the mixture probabilities $\{\pi_h\}$ in Section 3.1, we derive our results for an arbitrary prior Q_0 for wider applicability. We assume the inverse-gamma prior on residual variance σ^2 to be restricted to the compact set $[\sigma_L^2, \sigma_U^2]$.

4.3 Main Results

In Lemma 5 we derive sufficient conditions for the posterior probability on the space of partitions to converge to the oracle probability for $p \to \infty$.

Lemma 5 *Let $\eta = [\eta_1, \dots, \eta_n]^{\mathrm{T}}$, $\zeta^{(p)} = \left[\zeta_1^{(p)}, \dots, \zeta_n^{(p)}\right]^{\mathrm{T}} = (\sqrt{p \log p})^{-1}(\Lambda^{\mathrm{T}}\Lambda)^{1/2}\eta$ and, for any $\delta > 0$, $B_{p,\delta} = \bigcap_{i=1}^{n}\{\Lambda, \eta_i : (\sqrt{p \log p})^{-1}\|\Lambda\eta_i - \Lambda_0\eta_{0i}\| < \delta\}$. Assume for any $\delta > 0$*

$$\Pi(\bar{B}_{p,\delta} \mid \mathcal{Y}) \to 0 \quad \mathbb{P}_0^p\text{-a.s.} \tag{12}$$

where $\bar{B}_{p,\delta}$ is the complement of $B_{p,\delta}$. Let $E(\cdot \mid \mathcal{Y})$ denote expectation with respect to the posterior distribution of the parameters given data $\mathcal{Y}$ and $\Pi(\Psi \mid \zeta^{(p)})$ be the conditional probability of partition Ψ with η_0 replaced by $\zeta^{(p)}$ in (10). Then, $\lim_{p \to \infty} E\left\{\Pi(\Psi \mid \zeta^{(p)}) \mid \mathcal{Y}\right\} = \Pi(\Psi \mid \eta_0)$.

In the following theorem, we show that condition (12) holds for Lamb and hence we avoid the large p pitfall. The proof is in the supplementary materials.

Theorem 6 *Let $B_{p,\delta}$ be as defined in Lemma 5 and $\bar{B}_{p,\delta}$ be its complement set. Then, under (C1)-(C4) and model (6), $\Pi(\bar{B}_{p,\delta} \mid \mathcal{Y}) \to 0$ $\mathbb{P}_0^p$-a.s. for any $\delta > 0$.*

Theorem 6 implies that our model learns the latent factors more accurately with increasing p. In addition to the proof of Theorem 6, this result is further illustrated empirically via a simple simulation experiment reported in Section S.4.2 of the supplementary materials.

The oracle has a slightly simpler model specification than (8) assuming common covariances across components. This simplification is done to make the associated theory more tractable, but the simplified location mixture case is rich enough to provide a nice test case for assessing how the proposed approach can escape the curse of dimensionality.

As conditions (C1)-(C4) imply (A0), the clustering models in Corollaries 2 and 3 would still lead to the two extreme partitions. The Lamb model, in learning the low-dimensional latent space with increasing dimensions, escapes these pitfalls.

5. Simulation Study

We perform a simulation study to analyze the performance of Lamb in clustering high dimensional data. The sampler introduced in Section 3.2 is available from the GitHub page of the first author.
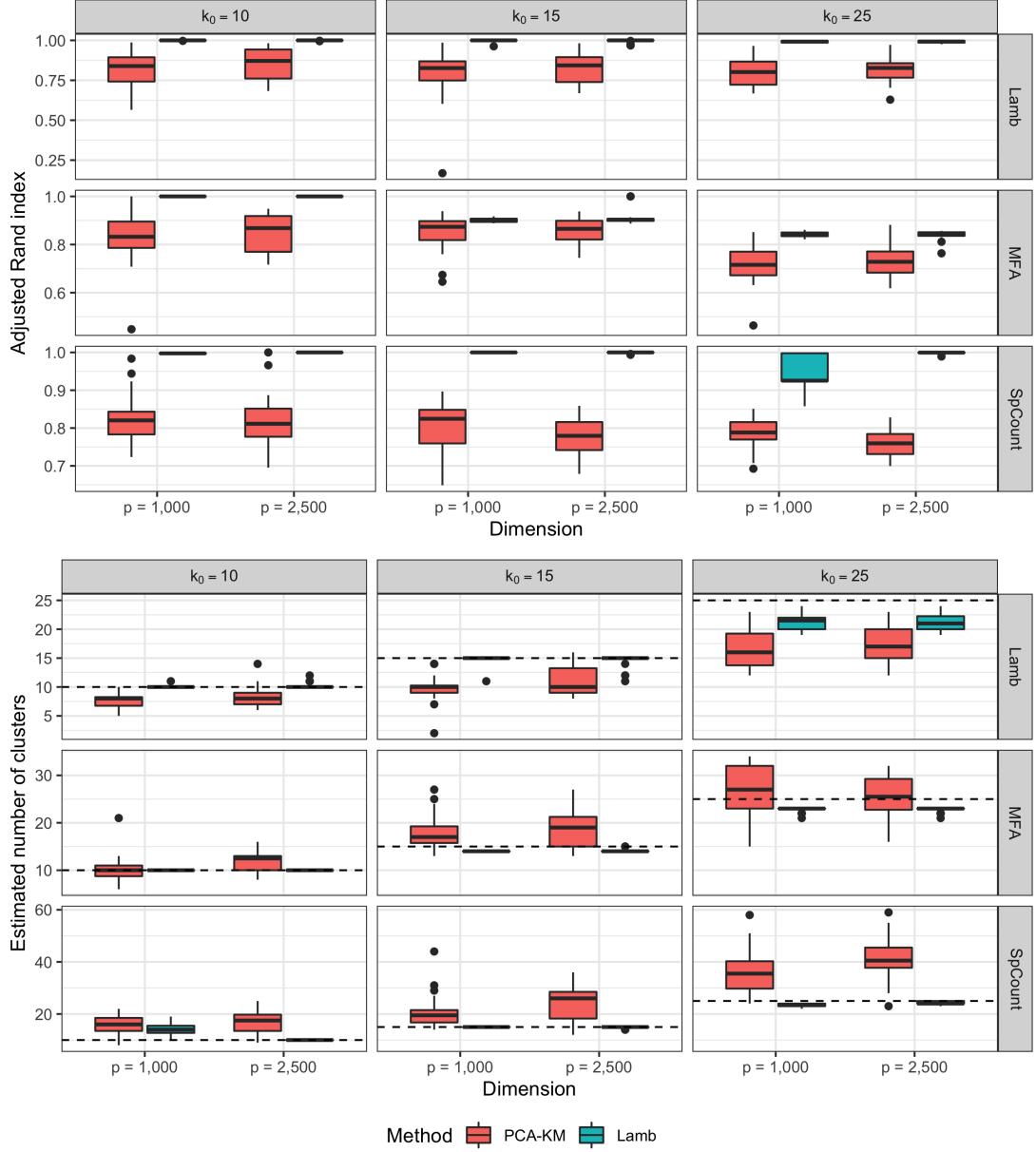
Figure 1: Comparison between our proposed Lamb and the two-stage PCA-KM approach: Distributions of the adjusted Rand indices (upper plot) and estimated number of clusters (lower plot) in 20 replicated experiments. Horizontal dashed lines denote the true number of clusters. The simulation scenarios, reported in each row, are labeled as Lamb for the model of Section 3, MFA for mixture of factor analyzers and SpCount for the log transformed zero inflated sparse Poisson counts.

We compare with a Dirichlet process mixture of Gaussian model with diagonal covariance matrix implemented in R package BNPmix (Corradin et al., 2021), a nonparametric mixture of infinite factor analyzers implemented in R package IMIFA (Murphy et al., 2019), and a pragmatic two-stage approach (PCA-KM) that performs an approximate sparse principal component analysis of

the high dimensional data to reduce dimensionality from p to $\widehat{d}$—with $\widehat{d}$ the minimum number of components explaining at least 95% of the variability as discussed in Section 3.1—and then applies k-means on the principal components, with k chosen by maximizing the average silhouette width (Rousseeuw, 1987). This same approach is used to choose $\widehat{d}$ in implementing Lamb.

For the high-dimensional simulation settings we considered, both the mixture of Gaussians and the mixture of factor analyzers showed high instability, including software crashing for memory issues, lack of convergence, and extremely long running times. For these reasons we report a comparison with PCA-KM approach only. To test the accuracy of the estimated clustering relative to the true clustering, we compute the adjusted Rand index (Rand, 1971).

We generated data under: [1] Lamb, [2] mixture of sparse factor analyzers (MFA), and [3] mixture of log transformed zero inflated sparse Poisson counts (SpCount) [1]-[2] have latent dimension 20, while for [3] the data are discrete and highly non-Gaussian within clusters mimicking the data of Section 6. Details are provided in Section S.3 of the supplementary materials.

We vary true number of clusters $k_0 \in \{10, 15, 25\}$, with the first $\lfloor 2k_0/3 \rceil$ 'main' clusters having the same probability and the remaining ones having together the same probability of a single main cluster. For example if $k_0 = 25$, we set 16 main clusters with probability $1/17$ each and 9 minor clusters of equal weights, whose total probability sums to $1/17$. This is a highly challenging case, as many methods struggle unless there are a small number of close to equal weight clusters that are well separated. The dimension p varies in $p = \{1,000,\ 2,500\}$ while the sample size n is $n = 2,000$. Data visualization plots using McInnes et al. (2018) are in Section S.4.4 of the supplementary materials. For each configuration, we perform 20 independent replications. We run our sampler for 6,000 iterations discarding the first 1,000 as burn in and taking one draw every five to reduce autocorrelation. Prior elicitation follows the default specification of Section 3.1. On average, 6,000 iterations under these settings took between 40 and 50 minutes on a iMac with 4.2 GHz Quad-Core Intel Core i7 processor and 32GB DDR4 RAM.

Figure 1 reports the distribution of the 20 replicates of the adjusted Rand index and mean estimated number of clusters. Our proposed Lamb is uniformly superior in each scenario obtaining high adjusted Rand indices, accurate clustering results, and less variability across replicates. In the MFA scenario, Lamb yields relatively lower Rand index for $k_0 = 25$. This is not unusual due to model misspecification and the large number of clusters.

The Lamb results do not vary much across the simulation replicates because the oracle posterior is quite concentrated at the true clustering. Since the dimension p is in the thousands, the asymptotic results derived in Section 4 kicked in resulting in narrow posterior credible intervals. To understand the performance of our proposed method in smaller sample sizes, we include additional simulation results with $n = 500$ in Section S.4.3 of the supplementary materials.

Furthermore, Section S.4.1 in the supplementary materials reports two simple simulation experiments showing that the degenerate clustering behavior discussed in Section 2 is evident even in moderate dimensions of $p = 20$.

6. Application to ScRNASeq Cell Line Dataset

In this section, we analyze the GSE81861 cell line dataset (Li et al., 2017) to illustrate the proposed method. The dataset profiles 630 cells from 7 cell lines using the Fluidigm based single cell RNA-seq protocol (See et al., 2018). The dataset includes 83 A549 cells, 65 H1437 cells, 55 HCT116 cells, 23 IMR90 cells, 96 K562 cells, 134 GM12878 cells, 174 H1 cells and 57,241 genes. The cell

types are known and hence the data provide a useful benchmark to assess performance in clustering high-dimensional data.

Following standard practice in single cell data analysis, we apply data pre-processing. Cells with low read counts are discarded, as we lack reliable gene expression measurements for these cells, and data are normalized following Lun et al. (2016). We remove non-informative genes using M3Drop (Andrews and Hemberg, 2018). After this pre-processing phase, we obtain a final dataset with $n = 531$ cells and $p = 7,666$ genes.

Applying our empirical Bayes approach, we estimate the latent dimension as $\widehat{d} = 19$. We implement Lamb using our default prior, collecting $10,000$ iterations after a burn-in of $5,000$ and keeping one draw in five. MCMC converge diagnostics are provided in Section S.5 in the supplementary materials. As comparison, we apply the two stage procedure of the previous section and the popular Seurat (Butler et al., 2018) pipeline which performs quality control, normalization, and selects informative genes that exhibit high variation across the cells.

Graphical representations of the different clustering results are shown in Figure 2 via UMAP projections (McInnes et al., 2018). Our proposed Lamb, the two stage approach, and Seurat achieve adjusted Rand indices of 0.977, 0.734 and 0.805 when compared to the true cluster-configuration and yield 12, 10, and 8 clusters, respectively. Seurat is reasonably accurate but splits the H1 cell-type into two clusters, while the two-stage approach is dramatically worse.
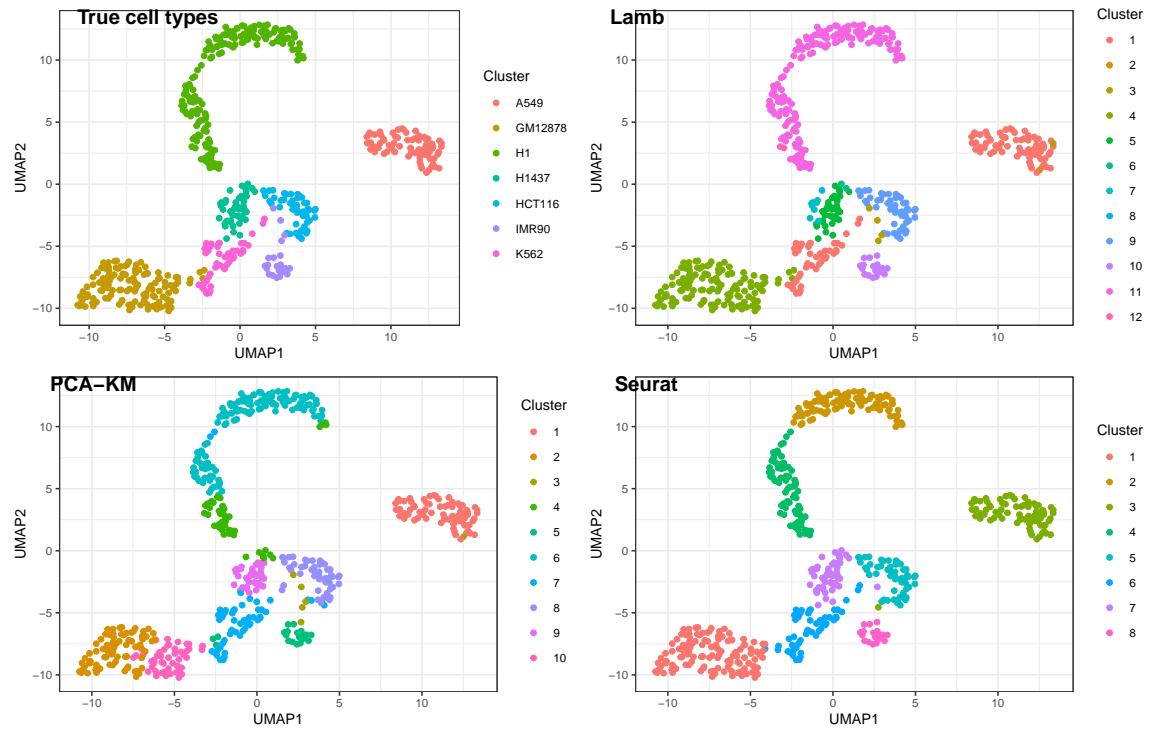

Figure 2: UMAP plots of the cell line dataset: Clusterings corresponding to the true cell-types, Lamb estimate, PCA-KM estimate and Seurat estimate are plotted in clockwise manner. Different panels use different color legends.
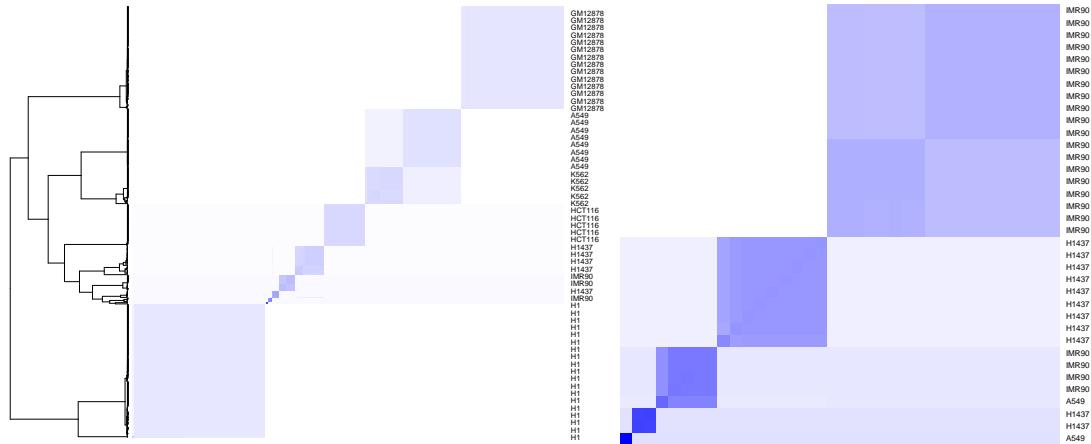
Figure 3: Posterior similarity matrix obtained from the Markov chain Monte Carlo samples of the Lamb method: Left panel reports the similarity matrix for the full cell line dataset along with the dendrogram obtained using complete linkage; row names report the true cluster names; right panel zooms the center of the left panel.

An appealing aspect of our approach is posterior uncertainty quantification. The 95% credible interval for the adjusted Rand index is $[0.900, 0.985]$ and the posterior probability of having between 11 and 13 clusters is 0.98. This suggests that the posterior distribution is highly concentrated, which is consistent with our simulations. The posterior similarity matrix reported in the first panel of Figure 3—also reporting the related dendrogram obtained by using complete linkage—clearly shows that the majority of the observations have a high posterior probability of being assigned to a specific cluster and negligible probability of being assigned to artifactual clusters. Figure 3 also shows micro clusters leading to over-estimation of the number of cell types. Two cells of cluster A549 are put in singleton clusters. Similarly cluster IMR90 is divided into two clusters of size 4 and 19 with negligible posterior probability of being merged. Finally cluster H1437 is split into four clusters with the main one comprising 35 of 47 observations and the smallest one comprising just one observation. Such micro-clusters have negligible impact for practical inference since Lamb does recover the original clustering configurations for most cell-types as reflected by the high adjusted Rand index with the true cell-types. Single-cell experiments are subject to high technical noise (Brennecke et al., 2013) which is not possible to completely remove in pre-processing steps. Such noise can potentially induce differences between cells that may not have any biological significance, for example, the cells in IMR90 (split into the clusters 3 and 10, see the top panel of Figure 2 for details) exhibit a substantial amount of variability although they are biologically of the same type.

7. Discussion

Part of the appeal of Bayesian methods is the intrinsic penalty for model complexity or 'Bayesian Ockham razor' (Jefferys and Berger, 1992), which comes through integrating the likelihood over the prior in obtaining the marginal likelihood. If one adds unnecessary parameters, then the likeli-

hood is integrated over a larger region, which tends to reduce the marginal likelihood. In clustering problems, one relies on the Bayesian Ockham razor to choose the appropriate compromise between the two extremes of too many clusters and over-fitting and too few clusters and under-fitting. Often in low-dimensional problems, this razor is effective and one obtains a posterior providing a reasonable representation of uncertainty in clustering data into groups of relatively similar observations. However, a key contribution of this article is showing that this is fundamentally not the case in high-dimensional problems, and one can obtain nonsensical results using seemingly reasonable priors.

Perhaps our most interesting result is the degenerate behavior in the $p \to \infty$ case for the true posterior on clusterings, regardless of the true data generating model. This negative result provided motivation for our latent factor mixture model, which addresses the large p pitfall by clustering on the latent variable level. Using a low rank factorization with appropriate shrinkage priors, the method can also handle realistic high-dimensional problems. Another interesting theoretical result is our notion of a Bayesian oracle for clustering; to our knowledge, there is not a similar concept in the literature. We show that our proposed Lamb attains the oracle with increasing dimensions.

Several interesting projects stem from the proposed work, which is a first step towards addressing pitfalls of Bayesian approaches to high-dimensional clustering. One important thread is designing faster MCMC algorithms for massive sample size exploiting parallel and distributed computing; for example, running MCMC for different subsets of the variables in parallel and combining the results. Some recent works in the literature discuss related approaches (Ni et al., 2020; Song et al., 2020) but without considering the pitfalls that arise in high-dimensional data clustering. Another thread is to develop fast approximate inference algorithms that avoid MCMC, such as variational Bayes. In addition, it is of substantial interest to generalize the proposed approach to handle more complex data structures; for example, involving data that are not real-valued vectors and allowing for kernel misspecification (Miller and Dunson, 2019). In our settings d and n are fixed and not growing with p. The study of situations in which p, d and n jointly increase—at some rate—would be a very interesting theoretical extension of our results.

Supplementary Materials

Proofs of additional theoretical results, simulation studies and MCMC convergence diagnostics are provided in the supplementary materials.

Acknowledgments

Appendix

Proofs of Section 2

Proof [Theorem 1] Consider the ratio of posterior probabilities:

$$\frac{\Pi(\Psi \mid \mathcal{Y})}{\Pi(\Psi' \mid \mathcal{Y})}. \tag{A.1}$$

If this ratio converges to zero for all $c_1, \ldots, c_n$ in $\mathbb{P}_0^p$-probability as $p \to \infty$, then any partition nested into another partition is more likely *a posteriori* implying $\Pi(c_1 = \cdots = c_n \mid \mathcal{Y}) = 1$ in $\mathbb{P}_0^p$-probability so that all subjects are grouped in the same cluster with probability one. Conversely if the ratio converges to $+\infty$, then $\Pi(c_1 \neq \cdots \neq c_n \mid \mathcal{Y}) = 1$ in $\mathbb{P}_0^p$-probability and each subject is assigned to their own cluster with probability one.

Without loss of generality, assume that $c_1, \ldots, c_n$ define k_n clusters of sizes $n_1, \ldots, n_{k_n}$ and that $c_i' = c_i$ for $c_i \in \{1, \ldots, k_n - 2\}$ and $c_i' = k_n - 1$ for $c_i \in \{k_n - 1, k_n\}$, with $n_1', \ldots, n_{k_n'-1}'$ the cluster sizes under the partition induced by the c_i'. In general, ratio (A.1) can be expressed as

$$\frac{\Pi(\Psi)}{\Pi(\Psi')} \times \frac{\prod_{h=1}^{k_n} \int \prod_{i:c_i=h} \mathcal{K}(y_i; \theta) \, \mathrm{d}P_0(\theta)}{\prod_{h=1}^{k_n-1} \int \prod_{i:c_i'=h} \mathcal{K}(y_i; \theta) \, \mathrm{d}P_0(\theta)}. \tag{A.2}$$

The left hand side of (A.2) can be expressed as the ratio between the EPPFs. Since by assumption there is a positive prior probability for any partition in $\mathscr{P}$, this ratio is finite and does not depend on p or the data generating distribution. Thus, by induction and under the assumptions on the right factor of (A.2) we conclude the proof. ∎

Proof [Corollary 2] Define $c_1, \ldots, c_n$ and $c_1', \ldots, c_n'$ consistently with the proof of Theorem 1. Then, consider the ratio of the marginal likelihoods

$$\frac{\prod_{h=1}^{k_n} \int \prod_{i:c_i=h} \mathrm{N}_p(y_i; \mu_h, \Sigma_h) \, \mathrm{N}_p(\mu_h; \mu_0, \kappa_0^{-1}\Sigma_h) IW(\Sigma_h; \nu_0, \Lambda_0) \, \mathrm{d}(\mu_h, \Sigma_h)}{\prod_{h=1}^{k_n-1} \int \prod_{i:c_i'=h} \mathrm{N}_p(y_i; \mu_h, \Sigma_h) \, \mathrm{N}_p(\mu_h; \mu_0, \kappa_0^{-1}\Sigma_h) IW(\Sigma_h; \nu_0, \Lambda_0) \, \mathrm{d}(\mu_h, \Sigma_h)}. \tag{A.3}$$

The numerator of (A.3) is

$$\prod_{h=1}^{k_n} \left\{ \frac{1}{\pi^{n_h p/2}} \frac{\Gamma_p(\frac{\nu_0+n_h}{2})}{\Gamma_p(\frac{\nu_0}{2})} \left(\frac{\kappa_0}{\kappa_0+n_h}\right)^{\frac{p}{2}} \frac{|\Lambda_0|^{\frac{\nu_0}{2}}}{\left|\Lambda_0 + S_h^{\Psi} + \frac{n_h\kappa_0}{\kappa_0+n_h}(\bar{y}_h^{\Psi} - \mu_0)(\bar{y}_h^{\Psi} - \mu_0)^{\mathrm{T}}\right|^{\frac{\nu_0+n_h}{2}}} \right\},$$

with $\bar{y}_h^{\Psi} = n_h^{-1} \sum_{i:c_i=h} y_i$, $S_h^{\Psi} = \sum_{i:c_i=h} \left(y_i - \bar{y}_h^{\Psi}\right)\left(y_i - \bar{y}_h^{\Psi}\right)^{\mathrm{T}}$, and $\Gamma_p(\cdot)$ being the multivariate gamma function. Obtaining a corresponding expression for the denominator, the ratio (A.3)

becomes

$$
\frac{\Gamma_p\left(\frac{\nu_0+n_{k_n-1}}{2}\right)\Gamma_p\left(\frac{\nu_0+n_{k_n}}{2}\right)}{\Gamma_p\left(\frac{\nu_0+n'_{k_n-1}}{2}\right)\Gamma_p\left(\frac{\nu_0}{2}\right)} \times \left\{\frac{\kappa_0(\kappa_0+n'_{k_n-1})}{(\kappa_0+n_{k_n-1})(\kappa_0+n_{k_n})}\right\}^{p/2}
$$

$$
\times \frac{|\Lambda_0|^{\frac{\nu_0}{2}}\left|\Lambda_0+S_{k_n-1}^{\Psi'}+\frac{n'_{k_n-1}\kappa_0}{\kappa_0+n'_{k_n-1}}(\bar{y}_{k_n-1}^{\Psi'}-\mu_0)(\bar{y}_{k_n-1}^{\Psi'}-\mu_0)^{\mathrm{T}}\right|^{\frac{\nu_0+n'_{k_n-1}}{2}}}{\prod_{h=k_n-1}^{k_n}\left|\Lambda_0+S_h^{\Psi}+\frac{n_h\kappa_0}{\kappa_0+n_h}(\bar{y}_h^{\Psi}-\mu_0)(\bar{y}_h^{\Psi}-\mu_0)^{\mathrm{T}}\right|^{\frac{\nu_0+n_h}{2}}}. \quad \text{(A.4)}
$$

We first study the limit of the first factor of (A.4). From Lemma S.7, we have

$$
\lim_{p\to\infty}\frac{1}{p}\left\{\log\frac{\Gamma_p\left(\frac{\nu_0+n_{k_n}}{2}\right)}{\Gamma_p\left(\frac{\nu_0+n'_{k_n-1}}{2}\right)}+\log\frac{\Gamma_p\left(\frac{\nu_0+n_{k_n-1}}{2}\right)}{\Gamma_p\left(\frac{\nu_0}{2}\right)}\right\}=0.
$$

We now study the limit of the remaining part of (A.4). Note that, if we replace each observation y_i with $\widetilde{y}_i=\Lambda_0^{-1/2}(y_i-\mu_0)$, assumption (A0) is still valid for $\widetilde{y}_i$'s. Moreover, $|\Lambda_0|$ terms get canceled out from (A.4). Hence, without loss of generality we can assume $\mu_0=0$ and $\Lambda_0=I_p$. Without loss of generality we can also assume that $y_{1+\sum_{j=1}^{h-1}n_j},\ldots,y_{\sum_{j=1}^{h}n_j}$ are in cluster h. We define

$$
Y_{(h)}^{\Psi}=\left[y_{1+\sum_{j=1}^{h-1}n_j},\ldots,y_{\sum_{j=1}^{h}n_j}\right]^{\mathrm{T}},
$$

to be the sub-data matrix corresponding to the h-th cluster in partition Ψ. Exploiting lower rank factorization results on matrix determinants, we have

$$
\left|I_p+S_h^{\Psi}+\frac{n_h\kappa_0}{n_h+\kappa_0}\bar{y}_h^{\Psi}\bar{y}_h^{\Psi\mathrm{T}}\right|=\left|I_p+Y_{(h)}^{\Psi\mathrm{T}}Y_{(h)}^{\Psi}-\frac{1}{n_h+\kappa_0}Y_{(h)}^{\Psi\mathrm{T}}\mathbf{1}_{n_h}\mathbf{1}_{n_h}^{\mathrm{T}}Y_{(h)}^{\Psi}\right|
$$

$$
=\left|1-\frac{1}{n_h+\kappa_0}\mathbf{1}_{n_h}^{\mathrm{T}}Y_{(h)}^{\Psi}\left\{I_p+Y_{(h)}^{\Psi\mathrm{T}}Y_{(h)}^{\Psi}\right\}^{-1}Y_{(h)}^{\Psi\mathrm{T}}\mathbf{1}_{n_h}\right|\left|I_p+Y_{(h)}^{\Psi\mathrm{T}}Y_{(h)}^{\Psi}\right|,
$$

where the symbol $|A|$ or $|a|$ is to be interpreted as the determinant of the matrix A or the absolute value of the scalar a, respectively. Then, the second factor of (A.4) simplifies to

$$
\frac{\left\{\left|1-\frac{1}{n'_{k_n-1}+\kappa_0}\mathbf{1}_{n'_{k_n-1}}^{\mathrm{T}}Y_{(k_n-1)}^{\Psi'}\left(I_p+Y_{(k_n-1)}^{\Psi'\mathrm{T}}Y_{(k_n-1)}^{\Psi'}\right)^{-1}Y_{(k_n-1)}^{\Psi'\mathrm{T}}\mathbf{1}_{n'_{k_n-1}}\right|\left|I_p+Y_{(k_n-1)}^{\Psi'\mathrm{T}}Y_{(k_n-1)}^{\Psi'}\right|\right\}^{\frac{\nu_0+n'_{k_n-1}}{2}}}{\prod_{h=k_n-1}^{k_n}\left\{\left|1-\frac{1}{n_h+\kappa_0}\mathbf{1}_{n_h}^{\mathrm{T}}Y_{(h)}^{\Psi}\left(I_p+Y_{(h)}^{\Psi\mathrm{T}}Y_{(h)}^{\Psi}\right)^{-1}Y_{(h)}^{\Psi\mathrm{T}}\mathbf{1}_{n_h}\right|\left|I_p+Y_{(h)}^{\Psi\mathrm{T}}Y_{(h)}^{\Psi}\right|\right\}^{\frac{\nu_0+n_h}{2}}}.
$$

Using Lemma S.9, $\lim_{p\to\infty}\left\|Y_{(h)}\left\{I_p+Y_{(h)}^{\mathrm{T}}Y_{(h)}\right\}^{-1}Y_{(h)}^{\mathrm{T}}-I_{n_h}\right\|_2=0$ in $\mathbb{P}_0^p$-probability and

$$
\lim_{p\to\infty}\left|1-\frac{1}{n_h+\kappa_0}\mathbf{1}_{n_h}^{\mathrm{T}}Y_{(h)}\left\{I_p+Y_{(h)}^{\mathrm{T}}Y_{(h)}\right\}^{-1}Y_{(h)}^{\mathrm{T}}\mathbf{1}_{n_h}\right|=\frac{\kappa_0}{\kappa_0+n_h}\quad\text{in }\mathbb{P}_0^p\text{-probability.}\quad\text{(A.5)}
$$

18

Taking the logarithm of the second and third factor of (A.4) and rearranging it using the previous result

$$\log \frac{\left| 1 - \frac{1}{n'_{k_n-1}+\kappa_0} \mathbf{1}^{\mathrm{T}}_{n'_{k_n-1}} Y^{\Psi'}_{(k_n-1)} \left\{ I_p + Y^{\Psi'\mathrm{T}}_{(k_n-1)} Y^{\Psi'}_{(k_n-1)} \right\}^{-1} Y^{\Psi'\mathrm{T}}_{(k_n-1)} \mathbf{1}_{n'_{k_n-1}} \right|^{\frac{\nu_0+n'_{k_n-1}}{2}}}{\prod_{h=k_n-1}^{k_n} \left| 1 - \frac{1}{n_h+\kappa_0} \mathbf{1}^{\mathrm{T}}_{n_h} Y^{\Psi}_{(h)} \left\{ I_p + Y^{\Psi\mathrm{T}}_{(h)} Y^{\Psi}_{(h)} \right\}^{-1} Y^{\Psi\mathrm{T}}_{(h)} \mathbf{1}_{n_h} \right|^{\frac{\nu_0+n_h}{2}}}$$

$$+ \log \left\{ \frac{\kappa_0(\kappa_0 + n'_{k_n-1})}{(\kappa_0 + n_{k_n-1})(\kappa_0 + n_{k_n})} \right\}^{p/2} + \log \frac{\left| I_p + Y^{\Psi'\mathrm{T}}_{(k_n-1)} Y^{\Psi'}_{(k_n-1)} \right|^{\frac{\nu_0+n'_{k_n-1}}{2}}}{\prod_{h=k_n-1}^{k_n} \left| I_p + Y^{\Psi\mathrm{T}}_{(h)} Y^{\Psi}_{(h)} \right|^{\frac{\nu_0+n_h}{2}}}. \quad \text{(A.6)}$$

Since $\nu_0 = p + c$, in conjunction with (A.5) we have sum of the limits of the first and second terms in (A.6) is 0 in $\mathbb{P}^p_0$-probability. We finally study the last summand of (A.5) and particularly

$$\lim_{p \to \infty} \frac{1}{p} \log \frac{\left| I_p + Y^{\Psi'\mathrm{T}}_{(k_n-1)} Y^{\Psi'}_{(k_n-1)} \right|^{\frac{\nu_0+n'_{k_n-1}}{2}}}{\prod_{h=k_n-1}^{k_n} \left| I_p + Y^{\Psi\mathrm{T}}_{(h)} Y^{\Psi}_{(h)} \right|^{\frac{\nu_0+n_h}{2}}}. \quad \text{(A.7)}$$

Since $\left| I_p + Y^{\Psi\mathrm{T}}_{(h)} Y^{\Psi}_{(h)} \right| = \left| I_{n_h} + Y^{\Psi}_{(h)} Y^{\Psi\mathrm{T}}_{(h)} \right|$ for any partition Ψ, following the same arguments of (S.3) from Lemma S.10 in the supplementary materials, we have

$$\left| I_p + Y^{\Psi'\mathrm{T}}_{(k_n-1)} Y^{\Psi'}_{(k_n-1)} \right| = \prod_{h=k_n-1}^{k_n} \left| I_p + Y^{\Psi\mathrm{T}}_{(h)} Y^{\Psi}_{(h)} \right|^{\frac{\nu_0+n_h}{2}} \times \left| I_{n_{k_n-1}} - ZZ^{\mathrm{T}} \right|$$

where $Z = \{ I_{n_{k_n-1}} + Y^{\Psi}_{(k_n-1)} Y^{\Psi\mathrm{T}}_{(k_n-1)} \}^{-1/2} Y^{\Psi}_{(k_n-1)} Y^{\Psi\mathrm{T}}_{(k_n)} \{ I_{n_{k_n}} + Y^{\Psi}_{(k_n)} Y^{\Psi\mathrm{T}}_{(k_n)} \}^{-1/2}$, and (A.7) reduces to

$$\frac{n_{k_n}}{2p} \log \left| I_{n_{k_n-1}} + Y^{\Psi}_{(k_n-1)} Y^{\Psi\mathrm{T}}_{(k_n-1)} \right| + \frac{n_{k_n-1}}{2p} \log \left| I_{n_{k_n}} + Y^{\Psi}_{(k_n)} Y^{\Psi\mathrm{T}}_{(k_n)} \right| + \frac{\nu_0 + n'_{k_n-1}}{2p} \log \left| I_{n_{k_n-1}} - ZZ^{\mathrm{T}} \right|.$$

From (A0), it can be deduced that the limits of the first two terms in the last expression are 0 in $\mathbb{P}^p_0$-probability as $p \to \infty$. Invoking Lemma S.10 and the fact that $\nu_0 = p + c$, we have

$$\limsup_{p \to \infty} \frac{\nu_0 + n'_{k_n-1}}{2p} \log \left| I_{n_{k_n-1}} - ZZ^{\mathrm{T}} \right| < 0,$$

and henceforth (A.7) is negative. This leads to

$$\limsup_{p \to \infty} \log \frac{\prod_{h=1}^{k_n} \int \prod_{i:c_i=h} \mathrm{N}_p(y_i; \mu_h, \Sigma_h) \, \mathrm{N}_p(\mu_h; \mu_0, \kappa_0^{-1}\Sigma_h) IW(\Sigma_h; \nu_0, \Lambda_0) \, \mathrm{d}(\mu_h, \Sigma_h)}{\prod_{h=1}^{k_n-1} \int \prod_{i:c'_i=h} \mathrm{N}_p(y_i; \mu_h, \Sigma_h) \, \mathrm{N}_p(\mu_h; \mu_0, \kappa_0^{-1}\Sigma_h) IW(\Sigma_h; \nu_0, \Lambda_0) \, \mathrm{d}(\mu_h, \Sigma_h)} = 0,$$

and hence for $p \to \infty$ all the data points are cluster together in $\mathbb{P}^p_0$-probability thanks to Theorem 1. ∎

Proof [Corollary 3] Define $c_1, \ldots, c_n$ and $c_1', \ldots, c_n'$ consistently with the proof of Theorem 1. Then, consider the ratio of the marginal likelihoods

$$\frac{\int \prod_{h=1}^{k_n} \int \prod_{i:c_i=h} \mathrm{N}_p\left(y_i; \mu_h, \Sigma\right) \mathrm{N}_p(\mu_h; \mu_0, \kappa_0^{-1}\Sigma) \mathrm{d}\mu_h IW(\Sigma; \nu_0, \Lambda_0) \, \mathrm{d}\Sigma}{\int \prod_{h=1}^{k_n-1} \int \prod_{i:c_i'=h} \mathrm{N}_p\left(y_i; \mu_h, \Sigma\right) \mathrm{N}_p(\mu_h; \mu_0, \kappa_0^{-1}\Sigma) \mathrm{d}\mu_h IW(\Sigma; \nu_0, \Lambda_0) \, \mathrm{d}\Sigma}. \tag{A.8}$$

The numerator of (A.8) is

$$\prod_{h=1}^{k_n} \left(\frac{\kappa_0}{n_h + \kappa_0}\right)^{\frac{p}{2}} \left|\Lambda_0 + \sum_{h=1}^{k_n} \left\{S_h^\Psi + \frac{n_h\kappa_0}{n_h + \kappa_0}(\bar{y}_h^\Psi - \mu_0)(\bar{y}_h^\Psi - \mu_0)^{\mathrm{T}}\right\}\right|^{-\frac{\nu_0+n}{2}} \pi^{-\frac{np}{2}} \frac{\Gamma_p(\frac{\nu_0+n}{2})}{\Gamma_p(\nu_0/2)} |\Lambda_0|^{\frac{\nu_0}{2}},$$

where $\bar{y}_h^\Psi = \frac{1}{n_h}\sum_{i:c_i=h} y_i$ and $S_h^\Psi = \sum_{i:c_i=h}\left(y_i - \bar{y}_h^\Psi\right)\left(y_i - \bar{y}_h^\Psi\right)^{\mathrm{T}}$. Hence, obtaining a corresponding expression for the denominator, ratio (A.8) becomes

$$\left\{\frac{\kappa_0(\kappa_0 + n_{k_n-1}')}{(\kappa_0 + n_{k_n-1})(\kappa_0 + n_{k_n})}\right\}^{\frac{p}{2}} \left[\frac{\left|\Lambda_0 + \sum_{h=1}^{k_n-1}\left\{S_h^{\Psi'} + \frac{n_h'\kappa_0}{n_h'+\kappa_0}(\bar{y}_h^{\Psi'} - \mu_0)(\bar{y}_h^{\Psi'} - \mu_0)^{\mathrm{T}}\right\}\right|}{\left|\Lambda_0 + \sum_{h=1}^{k_n}\left\{S_h^\Psi + \frac{n_h\kappa_0}{n_h+\kappa_0}(\bar{y}_h^\Psi - \mu_0)(\bar{y}_h^\Psi - \mu_0)^{\mathrm{T}}\right\}\right|}\right]^{\frac{\nu_0+n}{2}}. \tag{A.9}$$

First note that for $n_{k_n}, n_{k_n-1} \geq 1$

$$\frac{\kappa_0(\kappa_0 + n_{k_n-1}')}{(\kappa_0 + n_{k_n-1})(\kappa_0 + n_{k_n})} < 1. \tag{A.10}$$

Similar to Corollary 2, we can assume without loss of generality μ_0 to be a p-dimensional vector of zero and $\Lambda_0 = I_p$. Note that,

$$\sum_{h=1}^{k_n}\left(S_h^\Psi + \frac{n_h\kappa_0}{n_h + \kappa_0}\bar{y}_h^\Psi \bar{y}_h^{\Psi\mathrm{T}}\right) = \sum_{i=1}^{n} y_iy_i^{\mathrm{T}} - \sum_{h=1}^{k} \frac{n_h^2}{n_h + \kappa_0}\bar{y}_h^\Psi \bar{y}_h^{\Psi\mathrm{T}}$$

$$= \sum_{i=1}^{n} y_iy_i^{\mathrm{T}} - \sum_{h=1}^{k} \frac{1}{n_h + \kappa_0}\left(\sum_{i:c_i=h} y_i\right)\left(\sum_{i:c_i=h} y_i\right)^{\mathrm{T}}. \tag{A.11}$$

Also, without loss of generality we can assume $\left\{y_{1+\sum_{j=1}^{h-1} n_j}, \ldots, y_{\sum_{j=1}^{h} n_j}\right\}$ are in cluster h of Ψ similarly to Corollary 2. Then $\sum_{h=1}^{k_n}\left(S_h^\Psi + \frac{n_h\kappa_0}{n_h+\kappa_0}\bar{y}_h^\Psi \bar{y}_h^{\Psi\mathrm{T}}\right) = Y^{\mathrm{T}}\left(I_n - \mathcal{J}_n^\Psi\right)Y$, where $\mathcal{J}_n^\Psi = \mathrm{diag}\left(\frac{J_{n_1}}{n_1+\kappa_0}, \ldots, \frac{J_{n_{k_n}}}{n_{k_n}+\kappa_0}\right)$ is an $n \times n$ order block diagonal matrix and J_r is the $r \times r$ order square matrix with all elements being 1. Clearly, $\mathcal{J}_n^\Psi$ is a positive semi-definite matrix of rank k_n. Henceforth, exploiting the lower rank factorization structure, each determinant in (A.9) can be simplified as

$$\left|I_p + \sum_{h=1}^{k_n}\left(S_h^\Psi + \frac{n_h\kappa_0}{n_h + \kappa_0}\bar{y}_h^\Psi \bar{y}_h^{\Psi\mathrm{T}}\right)\right| = \left|I_p + Y^{\mathrm{T}}\left(I_n - \mathcal{J}_n^\Psi\right)Y\right|$$

$$= \left|I_p + Y^{\mathrm{T}}Y\right|\left|I_n - \mathcal{J}_n^{\Psi 1/2}Y(I_p + Y^{\mathrm{T}}Y)^{-1}Y^{\mathrm{T}}\mathcal{J}_n^{\Psi 1/2}\right|.$$

Hence (A.9) reduces to

$$
\left\{ \frac{\kappa_0(\kappa_0 + n'_{k_n-1})}{(\kappa_0 + n_{k_n-1})(\kappa_0 + n_{k_n})} \right\}^{p/2} \left\{ \frac{\left| I_n - \mathcal{J}_n^{\Psi'1/2} Y(I_p + Y^{\mathrm{T}}Y)^{-1} Y^{\mathrm{T}} \mathcal{J}_n^{\Psi'1/2} \right|}{\left| I_n - \mathcal{J}_n^{\Psi1/2} Y(I_p + Y^{\mathrm{T}}Y)^{-1} Y^{\mathrm{T}} \mathcal{J}_n^{\Psi1/2} \right|} \right\}^{-\frac{\nu_0+n}{2}} . \tag{A.12}
$$

From Lemma S.8 in the supplementary materials, $\lim_{p\to\infty} \left\| Y(I_p + Y^{\mathrm{T}}Y)^{-1}Y^{\mathrm{T}} - I_n \right\|_2 = 0$ in $\mathbb{P}_0^p$-probability. Therefore, from the construction of $\mathcal{J}_n^{\Psi}$,

$$
\lim_{p\to\infty} \left| I_n - \mathcal{J}_n^{\Psi1/2} Y(I_p + Y^{\mathrm{T}}Y)^{-1} Y^{\mathrm{T}} \mathcal{J}_n^{\Psi1/2} \right| = \left| I_n - \mathcal{J}_n^{\Psi} \right| = \prod_{h=1}^{k_n} \left| I_{n_h} - \frac{1}{n_h + \kappa_0} J_{n_h} \right|, \tag{A.13}
$$

in $\mathbb{P}_0^p$-probability. Notably $J_r = 1_r 1_r^{\mathrm{T}}$ where 1_r is the r-dimensional vector of ones, implying that $\left| I_r - \frac{1}{n_r+\kappa_0} J_r \right| = \frac{\kappa_0}{n_r+\kappa_0}$ for any positive integer r. Substituting this in (A.13), we have

$$
\lim_{p\to\infty} \left| I_n - \mathcal{J}_n^{\Psi1/2} Y(I_p + Y^{\mathrm{T}}Y)^{-1} Y^{\mathrm{T}} \mathcal{J}_n^{\Psi1/2} \right| = \prod_{h=1}^{k_n} \frac{\kappa_0}{n_h + \kappa_0}, \quad \text{in } \mathbb{P}_0^p\text{-probability}
$$

and therefore,

$$
\lim_{p\to\infty} \frac{\left| I_n - \mathcal{J}_n^{\Psi'1/2} Y(I_p + Y^{\mathrm{T}}Y)^{-1} Y^{\mathrm{T}} \mathcal{J}_n^{\Psi'1/2} \right|}{\left| I_n - \mathcal{J}_n^{\Psi1/2} Y(I_p + Y^{\mathrm{T}}Y)^{-1} Y^{\mathrm{T}} \mathcal{J}_n^{\Psi1/2} \right|} = \frac{(\kappa_0 + n_{k_n-1})(\kappa_0 + n_{k_n})}{\kappa_0(\kappa_0 + n'_{k_n-1})} \quad \text{in } \mathbb{P}_0^p\text{-probability.}
$$

Thus if we take the log of (A.12) multiplied by p^{-1} and study its limit we have

$$
\liminf_{p\to\infty} \left\{ \frac{1}{2} \log \frac{\kappa_0(\kappa_0 + n'_{k_n-1})}{(\kappa_0 + n_{k_n-1})(\kappa_0 + n_{k_n})} + \frac{n+\nu_0}{2p} \log \frac{\left| I_n - \mathcal{J}_n^{\Psi'1/2} Y(I_p + Y^{\mathrm{T}}Y)^{-1} Y^{\mathrm{T}} \mathcal{J}_n^{\Psi'1/2} \right|}{\left| I_n - \mathcal{J}_n^{\Psi1/2} Y(I_p + Y^{\mathrm{T}}Y)^{-1} Y^{\mathrm{T}} \mathcal{J}_n^{\Psi1/2} \right|} \right\}
$$

$$
= \frac{1}{2} \log \frac{\kappa_0(\kappa_0 + n'_{k_n-1})}{(\kappa_0 + n_{k_n-1})(\kappa_0 + n_{k_n})} \times \left(1 - \limsup_{p\to\infty} \frac{n+\nu_0}{p} \right) > 0.
$$

Since n is fixed with p, the above limit follows from (A.10) and the assumption on ν_0. Thus we have

$$
\liminf_{p\to\infty} \frac{\int \prod_{h=1}^{k_n} \int \prod_{i:c_i=h} \mathrm{N}_p(y_i; \mu_h, \Sigma) \mathrm{N}_p(\mu_h; \mu_0, \kappa_0^{-1}\Sigma) \mathrm{d}\mu_h IW(\Sigma; \nu_0, \Lambda_0) \mathrm{d}\Sigma}{\int \prod_{h=1}^{k_n-1} \int \prod_{i:c'_i=h} \mathrm{N}_p(y_i; \mu_h, \Sigma) \mathrm{N}_p(\mu_h; \mu_0, \kappa_0^{-1}\Sigma) \mathrm{d}\mu_h IW(\Sigma; \nu_0, \Lambda_0) \mathrm{d}\Sigma} = \infty,
$$

and hence for $p \to \infty$ each data point is clustered separately in $\mathbb{P}_0^p$-probability thanks to Theorem 1. ∎

Proofs of Section 4

Proof [Lemma 5] Let $\zeta_0^{(p)} = (\sqrt{p \log p})^{-1} (\Lambda^{\mathrm{T}}\Lambda)^{-1/2} \Lambda^{\mathrm{T}} \Lambda_0 \eta_0$, then $\Pi(\Psi \mid \eta_0) = \Pi(\Psi \mid \zeta_0^{(p)})$. Then,

$$\frac{1}{\sqrt{p \log p}} \left\| \zeta_i^{(p)} - \zeta_{0i}^{(p)} \right\| \leq \left\| (\Lambda^{\mathrm{T}}\Lambda)^{-1/2} \Lambda \right\|_2 \times \frac{1}{\sqrt{p \log p}} \|\Lambda \eta_i - \Lambda_0 \eta_{0i}\| \leq \frac{1}{\sqrt{p \log p}} \|\Lambda \eta_i - \Lambda_0 \eta_{0i}\|. \tag{A.14}$$

From (11) we see that the numerator in the right hand side of (10) can be simplified as

$$C \times \Pi(\Psi) \times \prod_{h=1}^{k_n} \left(\frac{\kappa_0}{n_h + \kappa_0} \right)^{\frac{d}{2}} \times \left| \sum_{h=1}^{k_n} \left\{ S_{\eta_0}^h + \frac{n_h}{n_h + 1} \bar{\eta}_0^h \bar{\eta}_0^{h\mathrm{T}} \right\} \right|^{-\frac{n}{2}}, \tag{A.15}$$

where $n_h = \sum_{i=1}^n I(c_i = h)$, $\bar{\eta}_0^h = \frac{1}{n_h} \sum_{i:c_i=h} \eta_{0i}$, $S_{\eta_0}^h = \sum_{i:c_i=h} (\eta_{0i} - \bar{\eta}_0^h)(\eta_{0i} - \bar{\eta}_0^h)^{\mathrm{T}}$ and C is a positive quantity constant across all $\Psi' \in \mathscr{P}$. Hence it is clear that $\Pi(\Psi \mid \eta)$ is a continuous function of η. Since the function is bounded (being a probability function), the continuity is also uniform. Also note that, for the particular choice of Gaussian kernel and base measure in (11), the oracle partition probability (10) is unchanged if η is multiplied by a full-rank square matrix and therefore $\Pi(\Psi \mid \zeta_0^{(p)}) = \Pi(\Psi \mid \eta_0)$. Therefore, for any $\epsilon > 0$ there exists $\delta > 0$ such that $\left\| \zeta_0^{(p)} - \zeta^{(p)} \right\| < \delta$ implies that $\left| \Pi(\Psi \mid \zeta^{(p)}) - \Pi(\Psi \mid \zeta_0^{(p)}) \right| = \left| \Pi(\Psi \mid \zeta^{(p)}) - \Pi(\Psi \mid \eta_0) \right| < \epsilon$. Again,

$$E\left\{ \left| \Pi(\Psi \mid \zeta^{(p)}) - \Pi(\Psi \mid \eta_0) \right| \,\middle|\, \mathcal{Y} \right\} = E\left\{ \left| \Pi(\Psi \mid \zeta^{(p)}) - \Pi(\Psi \mid \zeta_0^{(p)}) \right| \,\middle|\, B_{p,\delta}, \mathcal{Y} \right\} \Pi(B_{p,\delta} \mid \mathcal{Y})$$
$$+ E\left\{ \left| \Pi(\Psi \mid \zeta^{(p)}) - \Pi(\Psi \mid \eta_0) \right| \,\middle|\, \bar{B}_{p,\delta}, \mathcal{Y} \right\} \Pi(\bar{B}_{p,\delta} \mid \mathcal{Y}). \tag{A.16}$$

Due to continuity, δ can be chosen sufficiently small such that the term inside the first expectation in the right hand side of (A.16) is smaller than arbitrarily small $\epsilon > 0$. Now for any $\delta > 0$, the second term in the right hand side of (A.16) goes to 0 as $\Pi(\bar{B}_{p,\delta} \mid \mathcal{Y}) \to 0$ as $p \to \infty$ by assumption. Therefore, for arbitrarily small $\epsilon > 0$, $E\left\{ \left| \Pi(\Psi \mid \zeta^{(p)}) - \Pi(\Psi \mid \eta_0) \right| \,\middle|\, \mathcal{Y} \right\} < \epsilon$ for large enough p. Hence the proof. ∎

References

Tallulah S Andrews and Martin Hemberg. M3Drop: Dropout-based feature selection for scRNASeq. *Bioinformatics*, 35:2865–2867, 12 2018.

F Ascolani, A Lijoi, G Rebaudo, and G Zanella. Clustering consistency with Dirichlet process mixtures. *Biometrika*, 2022. To appear.

Jangsun Baek, Geoffrey J. McLachlan, and Lloyd K. Flack. Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1298–1309, 2010.

James Baglama and Lothar Reichel. Augmented implicitly restarted Lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing*, 27:19–42, 2005.

Jushan Bai and Serena Ng. Large dimensional factor analysis. *Foundations and Trends in Econometrics*, 3:89–163, 2008.

Jeffrey D. Banfield and Adrian E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.

Anirban Bhattacharya and David B Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98:291–306, 2011.

Anirban Bhattacharya, Debdeep Pati, Natesh S. Pillai, and David B. Dunson. Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110:1479–1490, 2015.

Peter J. Bickel and Elizaveta Levina. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10:989–1010, 2004.

D. A. Binder. Bayesian cluster analysis. *Biometrika*, 65:31–38, 1978.

Charles Bouveyron and Camille Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78, 2014.

Philip Brennecke, Simon Anders, et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10:1093–1095, 2013.

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36:411–420, 2018.

Diana Cai, Trevor Campbell, and Tamara Broderick. Finite mixture models are typically inconsistent for the number of components. *arXiv:2007.04470*, 2020.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. Handling sparsity via the horseshoe. *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, 5: 73–80, 2009.

Gilles Celeux, Kaniav Kamary, Gertraud Malsiner-Walli, Jean-Michel Marin, and Christian P. Robert. Computational solutions for Bayesian inference in mixture models. In Sylvia Frühwirth-Schnatter, Gilles Celeux, and Christian P. Robert, editors, *Handbook of Mixture Analysis*, chapter 5, pages 77–100. CRC Press, Boca Raton, FL, 2018.

Riccardo Corradin, Antonio Canale, and Bernardo Nipoti. BNPmix: an R package for Bayesian nonparametric modelling via Pitman-Yor mixtures. *Journal of Statistical Software*, in press, 2021.

David B. Dunson. Nonparametric Bayes local partition models for random effects. *Biometrika*, 96: 249–262, 2009.

Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147:186–197, 2008.

Jianqing Fan, Yuan Liao, and Martina Mincheva. High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39:3320–3356, 2011.

Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973.

Sylvia Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer Science & Business Media, 2006.

Sylvia Frühwirth-Schnatter and Gertraud Malsiner-Walli. From here to infinity: Sparse finite versus Dirichlet process mixtures in model-based clustering. *Advances in data analysis and classification*, 13:33–64, 2019.

Sylvia Fruhwirth-Schnatter, Gilles Celeux, and Christian P Robert. *Handbook of mixture analysis*. CRC press, 2019.

Sylvia Frühwirth-Schnatter, Gertraud Malsiner-Walli, and Bettina Grün. Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis*, 16:1279–1307, 2021.

Giuliano Galimberti and Gabriele Soffritti. Using conditional independence for parsimonious model-based Gaussian clustering. *Statistics and Computing*, 23:625–638, 2013.

Giuliano Galimberti, Angela Montanari, and Cinzia Viroli. Penalized factor mixture analysis for variable selection in clustered data. *Computational Statistics & Data Analysis*, 53:4301–4310, 2009.

J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian Statistics*, 4:641–649, 1992.

Zoubin Ghahramani, Geoffrey E Hinton, et al. The EM algorithm for mixtures of factor analyzers. Technical report, CRG-TR-96-1, University of Toronto, 1996.

Subhashis Ghosal and Aad Van Der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173, 2001.

Sonia Jain and Radford M Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13:158–182, 2004.

William H. Jefferys and James O. Berger. Ockham's razor and Bayesian analysis. *American Scientist*, 80:64–72, 1992.

Sinae Kim, Mahlet G. Tadesse, and Marina Vannucci. Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93:877–893, 2006.

Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20:273–282, 2019.

Sirio Legramanti, Daniele Durante, and David B Dunson. Bayesian cumulative shrinkage for infinite factorizations. *Biometrika*, 107:745–752, 2020.

Huipeng Li, Elise T Courtois, Debarka Sengupta, Yuliana Tan, Kok Hao Chen, Jolene Jie Lin Goh, Say Li Kong, Clarinda Chua, Lim Kiat Hon, Wah Siew Tan, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature Genetics*, 49:708, 2017.

Aaron TL Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17:75, 2016.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software*, 3:861, 2018.

Jeffrey W. Miller and David B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114:1113–1125, 2019.

Jeffrey W Miller and Matthew T Harrison. Inconsistency of Pitman-Yor process mixtures for the number of components. *The Journal of Machine Learning Research*, 15:3333–3370, 2014.

Jeffrey W. Miller and Matthew T. Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113:340–356, 2018.

Angela Montanari and Cinzia Viroli. Heteroscedastic factor mixture analysis. *Statistical Modelling*, 10:441–460, 2010.

Keefe Murphy, Cinzia Viroli, and Isobel Claire Gormley. IMIFA*: Infinite Mixtures of Infinite Factor Analysers and Related Models*, 2019. R package version 2.1.1.

Yang Ni, Peter Müller, Maurice Diesendruck, Sinead Williamson, Yitan Zhu, and Yuan Ji. Scalable Bayesian nonparametric clustering and classification. *Journal of Computational and Graphical Statistics*, 29:53–65, 2020.

Debdeep Pati, Anirban Bhattacharya, Natesh S. Pillai, and David Dunson. Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *The Annals of Statistics*, 42: 1102–1130, 2014.

Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900, 1997.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6:7–11, 2006.

William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.

Sylvia. Richardson and Peter J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59:731–792, 1997.

Veronika Ročková and Edward I George. Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111:1608–1622, 2016.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

Lorenzo Schiavon, Antonio Canale, and David Dunson. Generalized infinite factorization models. *Biometrica, in press*, 2021.

Peter See, Josephine Lum, Jinmiao Chen, and Florent Ginhoux. A single-cell sequencing guide for immunologists. *Frontiers in Immunology*, 9:2425, 2018.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

Hanyu Song, Yingjian Wang, and David B Dunson. Distributed Bayesian clustering using finite mixture of mixtures. *arXiv preprint arXiv:2003.13936*, 2020.

Mahlet G. Tadesse, Naijun Sha, and Marina Vannucci. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100:602–617, 2005.

Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press, 2012.

Sara Wade and Zoubin Ghahramani. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13:559–626, 06 2018.

Mike West. *Hyperparameter estimation in Dirichlet process mixture models*. Duke University ISDS Discussion Paper# 92-A03, 1992.

Dapeng Yao, Fangzheng Xie, and Yanxun Xu. Bayesian sparse Gaussian mixture model in high dimensions. *arXiv preprint arXiv:2207.10301*, 2022.

Supplementary Materials for
Escaping the Curse of Dimensionality in Bayesian Model-based Clustering

The supplementary materials present proofs of some theoretical results, simulation studies and MCMC convergence diagnostics.

S.1. Additional Theoretical Results

In the supplementary materials, we denote by $\|x\|$ the Euclidean norm of a vector x and by $\|X\|_2$ the spectral norm of a matrix X. The smallest and largest eigenvalues of the matrix $(X^{\mathrm{T}}X)^{\frac{1}{2}}$ are denoted by $s_{\min}(X)$ and $s_{\max}(X)$, respectively. For a positive-definite matrix X, $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ denote the smallest and largest eigenvalues, respectively.

Lemma S.7 *Let $\Gamma_p(\cdot)$ be the multivariate gamma function, $\nu_0 = p + c$ for some constant $c \geq 0$, and ℓ and m be (not varying with p) non-negative integers. Then, $\lim_{p \to \infty} \frac{1}{p} \log\{\Gamma_p(\frac{\nu_0+\ell}{2})/\Gamma_p(\frac{\nu_0+m}{2})\} = 0$.*

Proof Without loss of generality assuming $\ell > m$, we have

$$\frac{\Gamma_p(\frac{\nu_0+\ell}{2})}{\Gamma_p(\frac{\nu_0+m}{2})} = \prod_{j=1}^{p} \frac{\Gamma\left(\frac{\nu_0+\ell-j+1}{2}\right)}{\Gamma\left(\frac{\nu_0+m-j+1}{2}\right)} = \frac{\prod_{j=m+1}^{\ell}\Gamma\left(\frac{\nu_0+j}{2}\right)}{\prod_{j=m}^{\ell}\Gamma\left(\frac{\nu_0+j-p}{2}\right)}. \tag{S.1}$$

Note that the denominator term in the extreme right hand of (S.1) does not depend on p as $\nu_0 - p$ is constant from assumption. Applying Stirling's approximation on the numerator we get

$$\frac{\Gamma_p(\frac{\nu_0+\ell}{2})}{\Gamma_p(\frac{\nu_0+m}{2})} = \frac{1}{\prod_{j=m}^{\ell}\Gamma\left(\frac{\nu_0+j-p}{2}\right)} \times \prod_{j=m+1}^{\ell}\left\{\sqrt{2\pi\frac{\nu_0+j-1}{2}}\left(\frac{\nu_0+j-1}{2e}\right)^{\frac{\nu_0+j-1}{2}}E_j\right\}$$

$$= \frac{1}{\prod_{j=m}^{\ell}\Gamma\left(\frac{\nu_0+j-p}{2}\right)} \times \prod_{j=m+1}^{\ell}\left\{\sqrt{2\pi e}\left(\frac{\nu_0}{2e}\right)^{\frac{\nu_0+j+1}{2}}\left(1+\frac{j-1}{\nu_0}\right)^{\frac{\nu_0+j+1}{2}}E_j\right\},$$

where $E_j = O(\log p)$ arising from the Stirling's approximation formulae. Using the result $\lim_{x \to \infty}(1 + c/x)^x = e^c$, it can be seen that

$$\lim_{p \to \infty} \prod_{j=m+1}^{\ell}\left\{\sqrt{2\pi e}\left(\frac{\nu_0}{2e}\right)^{\frac{\nu_0+j+1}{2}}\left(1+\frac{j-1}{\nu_0}\right)^{\frac{\nu_0+j+1}{2}}\right\}$$

$$= (2\pi e)^{\frac{\ell-m}{2}} \times \left(\frac{\nu_0}{2e}\right)^{\frac{1}{4}(\ell-m)(2\nu_0+m+\ell+3)} \times e^{\frac{1}{2}(\ell-m)(\ell+m-1)},$$

which is a finite quantity. Hence the proof. ∎

Lemma S.8 *For any $n \times p$ order matrix Y satisfying (A0),* $\lim_{p \to \infty} \left\| Y(I_p + Y^{\mathrm{T}}Y)^{-1}Y^{\mathrm{T}} - I_n \right\|_2 = 0$ *in $\mathbb{P}_0^p$-probability.*

Proof Letting $Y = UDV$, the singular value decomposition of Y, we have $Y(I_p + Y^{\mathrm{T}}Y)^{-1}Y^{\mathrm{T}} = U \operatorname{diag} \left(\frac{d_1^2}{1+d_1^2}, \ldots, \frac{d_n^2}{1+d_n^2} \right) U^{\mathrm{T}}$ where $d_1, \ldots, d_n$ are the singular values of Y in descending order. From (A0) we have $\liminf_{p \to \infty} \frac{1}{p} d_i^2 > 0$, which further implies that $\liminf_{p \to \infty} \frac{d_i}{1+d_i} \to 1$ for all $i = 1, \ldots, n$. As $\frac{d_i}{1+d_i} \leq 1$, $\lim_{p \to \infty} \left\| Y(I_p + Y^{\mathrm{T}}Y)^{-1}Y^{\mathrm{T}} - I_n \right\|_2 = 0$ in $\mathbb{P}_0^p$-probability. ∎

Lemma S.9 *Let $\widetilde{Y}$ be an $\widetilde{n} \times p$ order matrix, formed by arbitrarily selecting $\widetilde{n}$ rows from $Y = [y_1, \ldots, y_n]^{\mathrm{T}}$ where $1 \leq \widetilde{n} \leq n$. If Y satisfies (A0), then* $\lim_{p \to \infty} \left\| \widetilde{Y}(I_p + \widetilde{Y}^{\mathrm{T}}\widetilde{Y})^{-1}\widetilde{Y}^{\mathrm{T}} - I_{\widetilde{n}} \right\|_2 = 0$ *in $\mathbb{P}_0^p$-probability.*

Proof Letting $Y = UDV$ the singular value decomposition of Y, we have $\widetilde{Y} = \widetilde{U}DV$ where $\widetilde{U}$ is formed by the corresponding rows of Y which were used to form $\widetilde{Y}$. Using Pati et al. (2014, Lemma 1.1(iii) from the Suppplementary section), we have $s_{\min}(\widetilde{Y}^{\mathrm{T}}) \geq s_{\min}(V^{\mathrm{T}})s_{\min}(D^{\mathrm{T}})s_{\min}(\widetilde{U}^{\mathrm{T}}) = s_{\min}(Y^{\mathrm{T}})$. Since $s_{\min}(\widetilde{U}^{\mathrm{T}}) = s_{\min}(U^{\mathrm{T}}) = 1$, $s_{\min}(\widetilde{Y}^{\mathrm{T}}) \geq s_{\min}(Y^{\mathrm{T}})$. Therefore, $\widetilde{Y}$ also satisfies (A0) if we substitute $Y = \widetilde{Y}$. Consequently applying Lemma S.8, we conclude the proof. ∎

Lemma S.10 *Let Y be an $n \times p$ order matrix satisfying (A0). Let $Y_i = [y_{j_{i,1}}, \ldots, y_{j_{i,n_i}}]^{\mathrm{T}}$, $i = 1, 2$ be an arbitrary partiton of the data-matrix into two sub-matrices such that $n_1 + n_2 = n$. Then $\limsup_{p \to \infty} s_{\max}(Z) < 1$ where $Z = (I_{n_1} + Y_1 Y_1^{\mathrm{T}})^{-1/2} Y_1 Y_2^{\mathrm{T}} (I_{n_2} + Y_2 Y_2^{\mathrm{T}})^{-1/2}$ in $\mathbb{P}_0^p$-probability.*

Proof From (A0) we have $\left\| YY^{\mathrm{T}} \right\|_2 = O(p)$ and $\liminf \lambda_{\min}(YY^{\mathrm{T}})/p > 0$, which implies that

$$0 < \liminf_{p \to \infty} \left| (I_n + YY^{\mathrm{T}})/p \right| \leq \limsup_{p \to \infty} \left| (I_n + YY^{\mathrm{T}})/p \right| = O(1) \text{ in } \mathbb{P}_0^p\text{-probability.} \quad \text{(S.2)}$$

Following the proof of Lemma S.9, we see that Y_i also satisfies (A0), and therefore (S.2) also holds if Y is replaced with Y_i for $i = 1, 2$. Noting that $I_n + YY^{\mathrm{T}} = \begin{bmatrix} I_{n_1} + Y_1 Y_1^{\mathrm{T}} & Y_1 Y_2^{\mathrm{T}} \\ Y_2 Y_1^{\mathrm{T}} & I_{n_2} + Y_2 Y_2^{\mathrm{T}} \end{bmatrix}$ and using matrix factorization results, we have

$$\left| \frac{1}{p}(I_n + YY^{\mathrm{T}}) \right| = \left| \frac{1}{p}(I_{n_1} + Y_1 Y_1^{\mathrm{T}}) \right| \left| \frac{1}{p}(I_{n_2} + Y_2 Y_2^{\mathrm{T}}) \right| \left| I_{n_1} - ZZ^{\mathrm{T}} \right|. \quad \text{(S.3)}$$

Again in $\mathbb{P}_0^p$-probability,

$$\limsup_{p \to \infty} s_{\max}^2(Z) \leq \limsup_{p \to \infty} \left\| Y_1^{\mathrm{T}}(I_{n_1} + Y_1 Y_1^{\mathrm{T}})^{-1}Y_1 \right\|_2 \left\| Y_2^{\mathrm{T}}(I_{n_2} + Y_2 Y_2^{\mathrm{T}})^{-1}Y_2 \right\|_2 \leq 1. \quad \text{(S.4)}$$

For (S.2) to hold, all the terms in the RHS of (S.3) must be bounded away from 0. As $\left| I_{n_1} - ZZ^{\mathrm{T}} \right| = \prod_{j=1}^{n_1} \left| 1 - s_j^2(Z) \right|$, the inequality on (S.4) must be strict. Thus, we conclude the proof. ∎

28

Lemma S.11 *For prior* (7) $\lim_{p\to\infty}\frac{1}{p}\lambda_{\min}(\Lambda^{\mathrm{T}}\Lambda) = \lim_{p\to\infty}\frac{1}{p}\lambda_{\max}(\Lambda^{\mathrm{T}}\Lambda) = v_1$ *for some* $v_1 > 0$ Π-*a.s.*

Proof From Bhattacharya et al. (2015, Eqn (9) and Section 2.4) we get that $\Lambda^{\mathrm{T}}\Lambda = \tau^2 T^{\mathrm{T}}T$ where the (i,j)-th element of T is $t_{ij} = e_{ij}\phi_{ij}$ with $e_{ij} \overset{iid}{\sim} \mathrm{DE}(1)$ where $\mathrm{DE}(b)$ is the double exponential distribution with median 0 and variance $2b^2$. Additionally $\phi \sim \mathrm{Dir}(a,\ldots,a)$ and $\tau \sim \mathrm{Ga}(pda, 1/2)$.

Now by the strong law of large numbers $\left\| \frac{1}{p}\tilde{T}^{\mathrm{T}}\tilde{T} - v_1 I_d \right\|_F \to 0$ as $p \to \infty$ where $\|\cdot\|_F$ is the Frobenius norm of a matrix and $v_1 = \mathrm{Var}(e_{ij}\gamma_{ij})$. Hence, for any $i = 1, \ldots, p \lim_{p\to\infty}\lambda_i(\tilde{T}^{\mathrm{T}}\tilde{T})p^{-1} = v_1$. Also $\lim_{p\to\infty}\tau/(pd) = E(\tau_{ij})$ which implies that $\lim_{p\to\infty}(\tau/\Gamma)^2 = 1$ Π-a.s. Hence the proof. ∎

S.2. Proof of Theorem 6 and Associated Results

To prove Theorem 6, we consider an adaptation of Theorem 6.39 in Ghosal and Van Der Vaart (2017) where instead of having an increasing sample size, we assume an increasing data dimension with fixed sample size. This notion is consistent with the idea that more and more variables are measured on each study subject. We introduce the following notation. Let $\vartheta = (\Lambda, \eta, \sigma)$ with $\eta = [\eta_1, \ldots, \eta_n]^{\mathrm{T}}$ and $\vartheta \in \Theta_p$. Let $\mathbb{P}_\vartheta^p$ and $\mathbb{P}_0^p$ be the joint distributions of the data $y_1, \ldots, y_n$ given ϑ and ϑ_0, respectively, with $\vartheta_0 = (\Lambda_0, \eta_0, \sigma_0)$. We also denote the expectation of a function g with respect to $\mathbb{P}_0^p$ and $\mathbb{P}_\vartheta^p$ by $\mathbb{P}_0^p g$ and $\mathbb{P}_\vartheta^p g$ respectively. Let p_0^p and p_ϑ^p be the densities of $\mathbb{P}_0^p$ and $\mathbb{P}_\vartheta^p$ with respect to the Lebesgue measure. Finally, define the Kullback-Leibler (KL) divergence and the r-th order positive KL-variation between p_0^p and p_ϑ^p, respectively, as $KL(\mathbb{P}_0^p, \mathbb{P}_\vartheta^p) = \int \log \frac{p_0^p}{p_\vartheta^p} \mathrm{d}\mathbb{P}_0^p$ and $V_r^+(\mathbb{P}_0^p, \mathbb{P}_\vartheta^p) = \int \left\{ \left(\log \frac{p_0^p}{p_\vartheta^p} - KL \right)^+ \right\}^r \mathrm{d}\mathbb{P}_0^p$, where f^+ denotes the positive part of a function f.

Theorem S.12 *If for some* $r \geq 2$, $c > 0$ *there exist measurable sets* $B_p \subset \Theta_p$ *with* $\liminf \Pi(B_p) > 0$,

(I) $\sup_{\vartheta \in B_p} \frac{1}{p} KL(\mathbb{P}_0^p, \mathbb{P}_\vartheta^p) \leq c$ *and* $\sup_{\vartheta \in B_p} \frac{1}{p^r} V_r^+(\mathbb{P}_0^p, \mathbb{P}_\vartheta^p) \to 0$.

(II) *For sets* $\tilde{\Theta}_p \subset \Theta_p$ *there exists a sequence of test functions* ϕ_p *such that* $\phi_p \to 0$ $\mathbb{P}_0^p$-*a.s. and* $\int_{\tilde{\Theta}_p} \mathbb{P}_\vartheta^p(1 - \phi_p)\mathrm{d}\Pi(\vartheta) \leq e^{-Cp}$ *for some* $C > 0$.

(III) *Letting* $A_p = \left\{ \vartheta \in \Theta_p : \frac{1}{p}\int \left(\log \frac{p_0^p}{p_\vartheta^p} - KL(\mathbb{P}_0^p, \mathbb{P}_\vartheta^p) \right)\mathrm{d}\tilde{\Pi}_p(\vartheta) < \tilde{\epsilon} \right\}$, *with* $\tilde{\Pi}_p$ *the renormalized restriction of* Π *to set* B_p, *for any* $\tilde{\epsilon} > 0$, $\mathbb{1}(\bar{A}_p) \to 0$ $\mathbb{P}_0^p$-*a.s.*

Then $\Pi(\tilde{\Theta}_p \mid \mathcal{Y}) \to 0$ $\mathbb{P}_0^p$-*a.s.*

Condition (I) ensures that the assumed model is not too far from the true data-generating model. Condition (II) controls the variability of the log-likelihood around its mean. In the Lamb model, the number of parameters grows with p and hence the assumption on V_r^+ is instrumental. The conditions on ϕ_p ensure the existence of a sequence of consistent test functions for $H_0 : \mathbb{P} = \mathbb{P}_0^p$ in which type-II error diminishes to 0 exponentially fast in the critical region. Condition (III) is

a technical condition required to bound the numerator of $\Pi(\tilde{\Theta}_p \mid \mathcal{Y})$. The proof of this theorem follows along the lines of the proof of Theorem 6.39 of Ghosal and Van Der Vaart (2017).

Theorem S.12 is a general result stating sufficient conditions for posterior consistency as $p \to \infty$. We use this theorem to prove Theorem 6.

Proof [Theorem 6] We verify the conditions (I)-(III) from Theorem S.12. Theorems S.13 and S.14 jointly imply that for the Lamb model there exist a sequence of sets B_p such that conditions (I) and (III) are satisfied for any $c > 0$. Theorem S.15 ensures the existence of a sequence of test functions satisfying (II), and finally Theorem S.16 proves (III). Hence the proof. ∎

Theorem S.13 *For any $\epsilon > 0$ define $B_p^\epsilon = \left\{ \Theta : p^{-1} KL(\mathbb{P}_0^p, \mathbb{P}_\vartheta^p) \leq \epsilon \right\}$. Then, under the settings of Section 4, $\liminf \Pi(B_p^\epsilon) > 0$.*

Proof Let P_0 and P be p-variate multivariate normal distributions with $P = \mathrm{N}_p(\mu, \Sigma)$ and $P_0 = \mathrm{N}_p(\mu_0, \Sigma_0)$. Then their Kullback-Leibler divergence is $KL(P_0, P) = \frac{1}{2} \{ \log \frac{|\Sigma|}{|\Sigma_0|} + \mathrm{tr}(\Sigma^{-1}\Sigma_0) + (\mu - \mu_0)^{\mathrm{T}} \Sigma^{-1} (\mu - \mu_0) - p \}$, which, under the settings of Section 4, simplifies to

$$KL(\mathbb{P}_0^p, \mathbb{P}_\vartheta^p) = \frac{1}{2} \left\{ np \log \frac{\sigma^2}{\sigma_0^2} + np \left(\frac{\sigma_0^2}{\sigma^2} - 1 \right) + \frac{1}{\sigma^2} \sum_{i=1}^n \|\mu_i - \mu_{0i}\|^2 \right\}, \qquad (S.5)$$

where $\mu_i = \Lambda \eta_i$ and $\mu_{0i} = \Lambda_0 \eta_{0i}$. Now,

$$\Pi \left\{ p^{-1} KL(\mathbb{P}_0^p, \mathbb{P}_\vartheta^p) < \epsilon \right\} = \Pi \left\{ n \log \frac{\sigma^2}{\sigma_0^2} + n \left(\frac{\sigma_0^2}{\sigma^2} - 1 \right) + \frac{1}{p\sigma^2} \sum_{i=1}^n \|\mu_i - \mu_{0i}\|^2 < \epsilon \right\}$$

$$\geq \Pi \left\{ \log \frac{\sigma^2}{\sigma_0^2} + \left(\frac{\sigma_0^2}{\sigma^2} - 1 \right) \leq \frac{\epsilon}{2n}, \frac{1}{\sigma^2} \sum_{i=1}^n \|\mu_i - \mu_{0i}\|^2 < \frac{p\epsilon}{2} \right\}.$$

Note that for any $x > 0$, $\log x \leq x - 1$ and therefore $\log \frac{\sigma^2}{\sigma_0^2} + \left(\frac{\sigma_0^2}{\sigma^2} - 1 \right) \leq \left(\frac{\sigma_0}{\sigma} - \frac{\sigma}{\sigma_0} \right)^2$ implying that

$$\Pi \left\{ p^{-1} KL(\mathbb{P}_0^p, \mathbb{P}_\vartheta^p) < \epsilon \right\} \geq \Pi \left\{ \left(\frac{\sigma_0}{\sigma} - \frac{\sigma}{\sigma_0} \right)^2 \leq \frac{\epsilon}{2n}, \frac{1}{\sigma^2} \sum_{i=1}^n \|\mu_i - \mu_{0i}\|^2 < \frac{p\epsilon}{2} \right\}$$

$$\geq \Pi \left\{ \left(\frac{\sigma_0}{\sigma} - \frac{\sigma}{\sigma_0} \right)^2 \leq \frac{\epsilon}{2n} \right\} \Pi \left(\sum_{i=1}^n \|\mu_i - \mu_{0i}\|^2 < \sigma_L \frac{p\epsilon}{2} \right),$$

where the second inequality holds thanks to condition (C3). The first factor above is positive under our proposed prior on σ. Now consider the second factor and note that for each $i = 1, \ldots, n$, $\|\mu_i - \mu_{0i}\|^2 = \left\| \Lambda(\eta_i - (\Lambda^{\mathrm{T}}\Lambda)^{-1}\Lambda^{\mathrm{T}}\Lambda_0\eta_{0i}) \right\|^2 + \eta_{0i}^{\mathrm{T}}(\Lambda_0^{\mathrm{T}}\Lambda_0 - \Lambda_0^{\mathrm{T}}\Lambda(\Lambda^{\mathrm{T}}\Lambda)^{-1}\Lambda^{\mathrm{T}}\Lambda_0)\eta_{0i}$. By the triangle inequality

$$\frac{1}{p} \left\| \Lambda_0^{\mathrm{T}}\Lambda_0 - \Lambda_0^{\mathrm{T}}\Lambda(\Lambda^{\mathrm{T}}\Lambda)^{-1}\Lambda^{\mathrm{T}}\Lambda_0 \right\|_2 \leq \left\| \frac{1}{p}\Lambda_0^{\mathrm{T}}\Lambda_0 - M \right\|_2 + \left\| M - \frac{1}{p}\Lambda_0^{\mathrm{T}}\Lambda(\Lambda^{\mathrm{T}}\Lambda)^{-1}\Lambda^{\mathrm{T}}\Lambda_0 \right\|_2. \quad (S.6)$$

The first term on the right hand side of (S.6) goes to 0 as $p \to \infty$ by (C2). Let us define the matrix $B = (\Lambda^{\mathrm{T}}\Lambda)^{-1/2}\Lambda^{\mathrm{T}}$, with $\|B\|_2 = 1$ and $\tilde{\Lambda}_0 = \Lambda_0 M^{-1/2}$ where $M^{1/2}$ is the Cholesky factor of M. From Vershynin (2012, Theorem 5.39) it follows that for any $0 < \epsilon < 1$ and large enough p, $1 - \epsilon \leq \left\| \frac{1}{\sqrt{p}}\tilde{\Lambda}_0 \right\|_2 \leq 1 + \epsilon$. Again, from Lemma 1.1 of the Supplement section of Pati et al. (2014) we have that $1 - \epsilon \leq \left\| \frac{1}{\sqrt{p}}B\tilde{\Lambda}_0 \right\|_2 \leq 1 + \epsilon$, $\quad 1 - \epsilon \leq \frac{1}{\sqrt{p}}s_{\min}(B\tilde{\Lambda}_0) \leq 1 + \epsilon$. Therefore $\lim_{p\to\infty} \lambda_i(\tilde{\Lambda}_0^{\mathrm{T}}\Lambda(\Lambda^{\mathrm{T}}\Lambda)^{-1}\Lambda^{\mathrm{T}}\tilde{\Lambda}_0)p^{-1} = 1$ for all $i = 1, \ldots, d$. Now $\left\| M - \frac{1}{p}\Lambda_0^{\mathrm{T}}\Lambda(\Lambda^{\mathrm{T}}\Lambda)^{-1}\Lambda^{\mathrm{T}}\Lambda_0 \right\|_2 = \|M\|_2 \left\| I_d - \frac{1}{p}\tilde{\Lambda}_0^{\mathrm{T}}\Lambda(\Lambda^{\mathrm{T}}\Lambda)^{-1}\Lambda^{\mathrm{T}}\tilde{\Lambda}_0 \right\|_2$ and therefore the second term on the right hand side of (S.6) goes to 0 as $p \to \infty$. Subsequently we have $\lim_{p\to\infty} \frac{1}{p}\eta_{0i}^{\mathrm{T}}(\Lambda_0^{\mathrm{T}}\Lambda_0 - \Lambda_0^{\mathrm{T}}\Lambda(\Lambda^{\mathrm{T}}\Lambda)^{-1}\Lambda^{\mathrm{T}}\Lambda_0)\eta_{0i} = 0$ for all $i = 1, \ldots, n$. Now (C2) and Lemma S.11 jointly imply that $\left\| (\Lambda^{\mathrm{T}}\Lambda)^{-1}\Lambda^{\mathrm{T}}\Lambda_0 \right\|_2 = O(1)$ Π-a.s. Therefore, for standard normal priors on the latent variables, $\liminf_{p\to\infty} \Pi(\sum_{i=1}^n \|\mu_i - \mu_{0i}\|^2 < \sigma_L p\epsilon) > 0$. From the permanence of KL-property of mixture priors (Ghosal and Van Der Vaart, 2017, Proposition 6.28) we can conclude that the right hand side is also positive. ∎

Theorem S.14 *On the set B_p^ϵ defined in Theorem S.13, we have $V_r^+(\mathbb{P}_0^p, \mathbb{P}_\vartheta^p) = o(p^r)$ for $r = 2$.*

Proof For $r = 2$, $V_r^+(\mathbb{P}_0^p, \mathbb{P}_\vartheta^p) \leq \int \log^2 \frac{p_0^p}{p_\vartheta^p} d\mathbb{P}_0^p - \left\{ \int \log \frac{p_0^p}{p_\vartheta^p} d\mathbb{P}_0^p \right\}^2$. Now conditionally on $\vartheta \in \vartheta$, the observations $y_1, \ldots, y_n$ are independent. Therefore,

$$V_r^+(\mathbb{P}_0^p, \mathbb{P}_\vartheta^p) \leq \sum_{j=1}^n \left[\int \left\{ \log \frac{p_{0j}(y_j)}{p_{\vartheta_j}(y_j)} \right\}^2 p_{0j}(y_j)\mathrm{d}y_j - \left\{ \int \log \frac{p_{0j}(y_j)}{p_{\vartheta_j}(y_j)} p_{0j}(y_j)\mathrm{d}y_j \right\}^2 \right] \quad (\text{S.7})$$

where $p_{0j}(y_j) = \prod_{i=1}^p \mathrm{N}(y_{ji}; \mu_{0ji}, \sigma_0^2)$ and $p_{\vartheta_j}(y_j) = \prod_{i=1}^p \mathrm{N}(y_{ji}; \mu_{ji}, \sigma^2)$ with $\mu_{0j} =$ and $\mu_j = \Lambda\eta_j$. We first show the result for a particular term inside the summation of (S.7). Since $\|\eta_{0i}\| = O(1)$ and n is fixed, the result will readily follow afterwards. For simplicity, we drop the suffix j from the terms of (S.7) henceforth. Consider,

$$\left\{ \log \frac{p_0(y_i)}{p_\vartheta(y_i)} \right\}^2 = \left[\log \frac{\sigma}{\sigma_0} - \frac{1}{2}\left\{ \left(\frac{y_i - \mu_{0i}}{\sigma_0} \right)^2 - \left(\frac{y_i - \mu_i}{\sigma} \right)^2 \right\} \right]^2$$

$$= \frac{1}{4}\left\{ \left(\frac{y_i - \mu_{0i}}{\sigma_0} \right)^2 - \left(\frac{y_i - \mu_i}{\sigma} \right)^2 \right\}^2 + \log^2 \frac{\sigma}{\sigma_0} - \left\{ \left(\frac{y_i - \mu_{0i}}{\sigma_0} \right)^2 - \left(\frac{y_i - \mu_i}{\sigma} \right)^2 \right\} \log \frac{\sigma}{\sigma_0}.$$

Note that,

$$\left\{ \left(\frac{y_i - \mu_{0i}}{\sigma_0} \right)^2 - \left(\frac{y_i - \mu_i}{\sigma} \right)^2 \right\}^2 = \left\{ z_i^2\left(1 - \frac{\sigma_0^2}{\sigma^2} \right) - 2z_i(\mu_{0i} - \mu_i)\frac{\sigma_0}{\sigma} + \left(\frac{\mu_i - \mu_{0i}}{\sigma} \right)^2 \right\}^2$$

$$= z_i^4\left(1 - \frac{\sigma_0^2}{\sigma^2} \right)^2 + 4z_i^2\sigma_0^2\left(\frac{\mu_{0i} - \mu_i}{\sigma} \right)^2 + \left(\frac{\mu_i - \mu_{0i}}{\sigma} \right)^4 - 2z_i^3\left(1 - \frac{\sigma_0^2}{\sigma^2} \right)\frac{\sigma_0}{\sigma}(\mu_{0i} - \mu_i)$$

$$- 2z_i\sigma_0\left(\frac{\mu_{0i} - \mu_i}{\sigma} \right)^3 + 2z_i^2\left(\frac{\mu_{0i} - \mu_i}{\sigma} \right)^2\left(1 - \frac{\sigma_0^2}{\sigma^2} \right)$$

31

where $z_i = (y_i - \mu_{0i})/\sigma_0$ and $z_i \overset{\text{iid}}{\sim} N(0, 1)$. Therefore,

$$E_{y_i}\left\{\left(\frac{y_i - \mu_{0i}}{\sigma_0}\right)^2 - \left(\frac{y_i - \mu_i}{\sigma}\right)^2\right\} = \left(1 - \frac{\sigma_0^2}{\sigma^2}\right) + \left(\frac{\mu_i - \mu_{0i}}{\sigma}\right)^2 \text{ and}$$

$$E_{y_i}\left\{\left(\frac{y_i - \mu_{0i}}{\sigma_0}\right)^2 - \left(\frac{y_i - \mu_i}{\sigma}\right)^2\right\}^2 = 3\left(1 - \frac{\sigma_0^2}{\sigma^2}\right)^2 + 4\sigma_0^2\left(\frac{\mu_{0i} - \mu_i}{\sigma}\right)^2 + \left(\frac{\mu_i - \mu_{0i}}{\sigma}\right)^4$$
$$+ 2\left(\frac{\mu_{0i} - \mu_i}{\sigma}\right)^2\left(1 - \frac{\sigma_0^2}{\sigma^2}\right).$$

Hence,

$$\int\left\{\log\frac{p_0(y_i)}{p_\vartheta(y_i)}\right\}^2 p_0(y_i)\mathrm{d}y_i = \left(\frac{\mu_{0i} - \mu_i}{\sigma}\right)^2 \times \left\{\sigma_0^2 + \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right) - \log\frac{\sigma}{\sigma_0}\right\}$$
$$- \log\frac{\sigma}{\sigma_0}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right) + \frac{1}{4}\left(\frac{\mu_{0i} - \mu_i}{\sigma}\right)^4 + \frac{3}{4}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right)^2 + \log^2\frac{\sigma}{\sigma_0}$$

$$\left\{\int\log\frac{p_0(y_i)}{p_\vartheta(y_i)}p_0(y_i)\mathrm{d}y_i\right\}^2 = \left\{\log\frac{\sigma}{\sigma_0} + \frac{\sigma_0^2 + (\mu_{0i} - \mu_i)^2}{2\sigma^2} - \frac{1}{2}\right\}^2 = \log^2\frac{\sigma}{\sigma_0} + \frac{1}{4}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right)^2$$
$$+ \frac{1}{4}\left(\frac{\mu_{0i} - \mu_i}{\sigma}\right)^4 + \left(\frac{\mu_{0i} - \mu_i}{\sigma}\right)^2 \times \left\{\log\frac{\sigma}{\sigma_0} - \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right)\right\} - \log\frac{\sigma}{\sigma_0}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right),$$

leading to

$$V_r^+(\mathbb{P}_0^p, \mathbb{P}_\vartheta^p) \le \sum_{i=1}^p\left[\int\left\{\log\frac{p_0(y_i)}{p_\vartheta(y_i)}\right\}^2 p_0(y_i)\mathrm{d}y_i - \left\{\int\log\frac{p_0(y_i)}{p_\vartheta(y_i)}p_0(y_i)\mathrm{d}y_i\right\}^2\right]$$
$$= \frac{p}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right)^2 + \left\{\sigma_0^2 - 2\log\frac{\sigma}{\sigma_0} + \left(1 - \frac{\sigma_0^2}{\sigma^2}\right)\right\} \times \sum_{i=1}^p\left(\frac{\mu_{0i} - \mu_i}{\sigma}\right)^2. \quad \text{(S.8)}$$

Note that

$$\sum_{i=1}^p(\mu_{0i} - \mu_i)^2 = \sum_{i=1}^p\left(\lambda_{0i}^{\mathrm{T}}\eta_0 - \lambda_i^{\mathrm{T}}\eta\right)^2 = \eta_0^{\mathrm{T}}\Lambda_0^{\mathrm{T}}\Lambda_0\eta_0 + \eta^{\mathrm{T}}\Lambda^{\mathrm{T}}\Lambda\eta - 2\eta_0^{\mathrm{T}}\Lambda_0^{\mathrm{T}}\Lambda\eta. \quad \text{(S.9)}$$

Now $\eta_0^{\mathrm{T}}\Lambda_0^{\mathrm{T}}\Lambda_0\eta_0 \le \|\Lambda_0\|_2^2\|\eta_0\|^2$ and therefore, by conditions (C2) and (C4), $\eta_0^{\mathrm{T}}\Lambda_0^{\mathrm{T}}\Lambda_0\eta_0 = O(p)$. Also from Lemma S.11, $\frac{1}{p}\|\Lambda\|_2^2 \le c$ for large enough p and some $c > 0$ and therefore $\eta^{\mathrm{T}}\Lambda^{\mathrm{T}}\Lambda\eta \le \|\Lambda\|_2^2\|\eta\|^2 = \|\eta\|^2 O(p)$. From the proof of Theorem S.13 we can see that in the set B_p^ϵ, $\|\eta\|$ is bounded. We have shown that the highest powers in (S.9) and thus in (S.8) are almost surely bounded by p for large enough p. Hence the proof. ∎

32

Theorem S.15 *Let us define the test function* $\phi_p = \mathbb{1}\left\{\left|\frac{1}{\sqrt{np}\sigma_0}\|\sum_{i=1}^n (y_i - \Lambda_0\eta_{0i})\| - 1\right| > \tau\right\}$ *to test the following hypothesis* $H_0 : y_1, \ldots, y_n \sim \mathbb{P}_0^p$ *versus* $H_1 : H_0$ *is false where* τ *is a positive real number. Define the set* $\tilde{\Theta}_p = \bar{B}_{p,\delta}$. *Then there exists a constant* $C > 0$ *such that* $\phi_p \to 0$ $\mathbb{P}_0^p$-*a.s. and* $\int_{\tilde{\Theta}_p} \mathbb{P}_\vartheta^p (1 - \phi_p) \mathrm{d}\Pi(\vartheta) \le e^{-Cp}$.

Proof Let us define $\mu_i = \Lambda\eta_i$ and $\mu_{0i} = \Lambda\eta_{0i}$. Then under H_0, $\frac{1}{\sqrt{n}\sigma_0}\sum_{i=1}^n (y_i - \Lambda\eta_{0i}) \sim \mathrm{N}_p(0, I_p)$ and therefore $\frac{1}{\sqrt{np}\sigma_0}\sum_{i=1}^n (y_i - \Lambda\eta_{0i}) \stackrel{d}{=} \omega/\sqrt{p}$ where $\omega \sim \mathrm{N}_p(0, I_p)$. Then from Rudelson and Vershynin (2013, Theorem 2.1) for some $c > 0$ and any $\tau_{np} > 0$ $\mathbb{P}_0^p\phi_p = \Pr\left(\left|\frac{1}{\sqrt{p}}\|\omega\| - 1\right| > \tau_{np}\right) \le 2\exp\left(-pc\tau_{np}^2\right)$. Since $\sum_{p=1}^\infty \mathbb{P}_0^p\phi_p < \infty$, by Borel-Cantelli lemma $\phi_p \to 0$ $\mathbb{P}_0^p$-a.s.

Notably when H_0 is not true i.e. under $\mathbb{P}_\vartheta^p$, $Y_i \stackrel{d}{=} \sigma\varphi_i + \Lambda\eta_i$ where $\varphi_i \stackrel{\mathrm{iid}}{\sim} \mathrm{N}_p(0, I_p)$ for some $\vartheta \neq (\Lambda_0, \eta_0, \sigma_0)$ and therefore under $\mathbb{P}_\vartheta^p$

$$\mathbb{P}_\vartheta^p(1 - \phi_p) \le \Pr\left\{\frac{1}{\sqrt{pn}\sigma_0}\left\|\sum_{i=1}^n (\sigma\varphi_i + \Lambda\eta_i - \Lambda_0\eta_{0i})\right\| < 1 + \tau_{np}\right\}$$

$$\le \Pr\left\{\frac{1}{\sqrt{pn}\sigma_0}\sum_{i=1}^n \|\Lambda\eta_i - \Lambda_0\eta_{0i}\| - 1 - \tau_{np} - \frac{\sigma}{\sigma_0} \le \frac{\sigma}{\sigma_0}\left(\frac{1}{\sqrt{np}}\sum_{i=1}^n \|\varphi_i\| - 1\right)\right\}. \quad \text{(S.10)}$$

Notably for $\vartheta \in \tilde{\Theta}_p$, $\frac{1}{\sqrt{pn}\sigma_0}\sum_{i=1}^n \|\Lambda\eta_i - \Lambda_0\eta_{0i}\|$ is unbounded above for increasing p and $\frac{\sigma}{\sigma_0}$ is bounded thanks to (C3). Letting $C_p = \frac{1}{\sqrt{pn}\sigma_0}\sum_{i=1}^n \|\Lambda\eta_i - \Lambda_0\eta_{0i}\| - 1 - \tau_{np} - \frac{\sigma}{\sigma_0}$ we have $\liminf_{p\to\infty} C_p > 0$. Therefore, from Rudelson and Vershynin (2013, Theorem 2.1), we have for $\vartheta \in \tilde{\Theta}_p$, $\mathbb{P}_\vartheta^p(1 - \phi_p) \le 2\exp\left(-pncC_p^2\right)$. Hence the proof. ∎

Theorem S.16 *Let* $\tilde{\Pi}_p$ *be the renormalized restriction of* Π *to the set* B_p^ϵ *defined in Theorem S.13. Then* $\mathbb{1}\{\bar{A}_p\} \to 0$ $\mathbb{P}_0^p$-*a.s.*

Proof If we can show that $\sum_{p=1}^\infty \mathbb{P}_0^p(\bar{A}_p) < \infty$, then by Borel-Cantelli lemma $\mathbb{P}_0^p[\limsup \bar{A}_p] = 0$ and henceforth $\mathbb{1}\{\bar{A}_p\} \to 0$ $\mathbb{P}_0^p$-a.s. Now

$$\mathbb{P}_0^p(\bar{A}_p) = \mathbb{P}_0^p\left[\frac{1}{p}\int \sum_{i=1}^n \left\{\frac{1}{\sigma^2}\|y_i - \mu_i\|^2 - \frac{1}{\sigma_0^2}\|y_i - \mu_{0i}\|^2 - \frac{1}{\sigma^2}\|\mu_i - \mu_{0i}\|^2 - p\left(\frac{\sigma_0^2}{\sigma^2} - 1\right)\right\} \mathrm{d}\tilde{\Pi}_p > 2\epsilon\right].$$

Notably under $\mathbb{P}_0^p$, $Y_i \stackrel{d}{=} \sigma_0\varphi_i + \mu_{0i}$ where $\varphi_i \stackrel{\mathrm{iid}}{\sim} \mathrm{N}_p(0, I_p)$. Therefore

$$\mathbb{P}_0^p(\bar{A}_p) = \Pr\left[\frac{1}{p}\int \sum_{i=1}^n \left\{\left(\frac{\sigma_0^2}{\sigma^2} - 1\right)(\|\varphi_i\|^2 - p) + 2\frac{\sigma_0}{\sigma^2}\varphi_i^{\mathrm{T}}(\mu_i - \mu_{0i})\right\} \mathrm{d}\tilde{\Pi}_p > 2\tilde{\epsilon}\right]$$

$$\le \Pr\left[\frac{1}{p}\sum_{i=1}^n (\|\varphi_i\|^2 - p)\int\left(\frac{\sigma_0^2}{\sigma^2} - 1\right)\mathrm{d}\tilde{\Pi}_p > \tilde{\epsilon}\right] +$$

$$\Pr\left[\frac{2}{p}\int \sum_{i=1}^n \left\{\frac{\sigma_0}{\sigma^2}\varphi_i^{\mathrm{T}}(\mu_i - \mu_{0i})\right\} \mathrm{d}\tilde{\Pi}_p > \tilde{\epsilon}\right]. \quad \text{(S.11)}$$

33

Let us consider the first term of (S.11). Notably

$$\Pr\left[\frac{1}{p}\sum_{i=1}^{n}(\|\varphi_i\|^2 - p)\int\left(\frac{\sigma_0^2}{\sigma^2} - 1\right)d\tilde{\Pi}_p > \tilde{\epsilon}\right] \leq \Pr\left[\frac{1}{p}\left|\sum_{i=1}^{n}(\|\varphi_i\|^2 - p)\right|\int\left|\frac{\sigma_0^2}{\sigma^2} - 1\right|d\tilde{\Pi}_p > \tilde{\epsilon}\right].$$
(S.12)

From (C3) we have that σ lies in a compact interval. Hence the integral in the right hand side of (S.12) is bounded above by some positive constant, say $C_{\sigma,1}$. Therefore,

$$\Pr\left[\frac{1}{p}\sum_{i=1}^{n}(\|\varphi_i\|^2 - p)\int\left(\frac{\sigma_0^2}{\sigma^2} - 1\right)d\tilde{\Pi}_p > \tilde{\epsilon}\right] \leq \Pr\left[\frac{1}{p}\left|\sum_{i=1}^{n}(\|\varphi_i\|^2 - p)\right| > \frac{\tilde{\epsilon}}{C_{\sigma,1}}\right] \leq 2e^{-pC_{\sigma,2}}$$

for some positive constant $C_{\sigma,2} > 0$. The second inequality in the above equation follows from Rudelson and Vershynin (2013, Theorem 2.1). Clearly

$$\sum_{p=1}^{\infty}\Pr\left[\frac{1}{p}\sum_{i=1}^{n}(\|\varphi_i\|^2 - p)\int\left(\frac{\sigma_0^2}{\sigma^2} - 1\right)d\tilde{\Pi}_p > \tilde{\epsilon}\right] < \infty.$$
(S.13)

Now we consider the second term of (S.11). As $\varphi_i = (\varphi_{i1}, \ldots, \varphi_{ip})^{\mathrm{T}}$ (similarly μ_i and μ_{0i} are also p-dimensional vectors) we can write

$$\Pr\left[\frac{2}{p}\int\sum_{i=1}^{n}\left\{\frac{\sigma_0}{\sigma^2}\varphi_i^{\mathrm{T}}(\mu_i - \mu_{0i})\right\}d\tilde{\Pi}_p > \tilde{\epsilon}\right] = \Pr\left[\frac{2}{p}\sum_{i=1}^{n}\sum_{j=1}^{p}\varphi_{ij}\int\left\{\frac{\sigma_0}{\sigma^2}(\mu_{ij} - \mu_{0ij})\right\}d\tilde{\Pi}_p > \tilde{\epsilon}\right]$$

$$\leq \exp\left[-\frac{p^2\tilde{\epsilon}^2}{4\sigma_0^2\sum_{i=1}^{n}\sum_{j=1}^{p}E_{\tilde{\Pi}_p}^2\left\{\frac{1}{\sigma^2}(\mu_{ij} - \mu_{0ij})\right\}}\right],$$

where $E_{\tilde{\Pi}_p}$ denotes the expectation with respect to the probability measure $\tilde{\Pi}_p$. The above inequality follows from sub-Gaussian concentration bounds. Now

$$\sum_{i=1}^{n}\sum_{j=1}^{p}E_{\tilde{\Pi}_p}^2\left\{\frac{1}{\sigma^2}(\mu_{ij} - \mu_{0ij})\right\} \leq \sum_{i=1}^{n}E_{\tilde{\Pi}_p}\frac{1}{\sigma^4}\|\mu_i - \mu_{0i}\|^2 \text{ (by Jensen's inequality)}$$

$$= E_{\tilde{\Pi}_p}\frac{1}{\sigma^4}\sum_{i=1}^{n}\times E_{\tilde{\Pi}_p}\|\mu_i - \mu_{0i}\|^2.$$
(S.14)

Since we consider independent priors on σ, Λ and η_i, (S.14) follows from its preceding step. Note that on the set B_p^{ϵ}

$$n\log\frac{\sigma^2}{\sigma_0^2} + n\left(\frac{\sigma_0^2}{\sigma^2} - 1\right) + \frac{1}{p\sigma^2}\sum_{i=1}^{n}\|\mu_i - \mu_{0i}\|^2 < 2\epsilon.$$
(S.15)

From the inequality $\log x < x - 1$ we see that $n\log\frac{\sigma^2}{\sigma_0^2} + n\left(\frac{\sigma_0^2}{\sigma^2} - 1\right) > 0$. Therefore for $\vartheta \in B_p^{\epsilon}$, in conjunction of (S.15) and (C3) we have $\frac{1}{p}\sum_{i=1}^{n}\|\mu_i - \mu_{0i}\|^2 < 2\epsilon\sigma_U^2 \Rightarrow \frac{1}{p}\sum_{i=1}^{n}E_{\tilde{\Pi}_p}\|\mu_i - \mu_{0i}\|^2 <$

$2\epsilon\sigma_U^2$. Also thanks to (C3) $E_{\tilde{\Pi}_p}\frac{1}{\sigma^4}$ is bounded above. Hence the term in (S.14) is bounded above and consequently

$$\sum_{p=1}^{\infty}\Pr\left[\frac{2}{p}\int\sum_{i=1}^{n}\left\{\frac{\sigma_0}{\sigma^2}\varphi_i^{\mathrm{T}}(\mu_i-\mu_{0i})\right\}\mathrm{d}\tilde{\Pi}_p > \tilde{\epsilon}\right] < \infty. \tag{S.16}$$

Combining (S.13) and (S.16) we conclude that $\sum_{p=1}^{\infty}\mathbb{P}_0^p(\bar{A}_p) < \infty$. Hence the proof. ∎

S.3. Details on Simulation Studies

In this section, we discuss the data-generation strategies in three simulation scenarios: [1] Lamb, [2] mixture of sparse factor analyzers (MFA), and [3] mixture of log transformed zero inflated Poisson counts (SpCount) considered in Section 5 of the main manuscript. The observed p-dimensional data are $y_1, \ldots, y_n$, k_0 is the number mixture components in the simulation truth and $\pi_1, \ldots, \pi_{k_0}$ are the mixture probabilities attached to each cluster such that $\sum_{h=1}^{k_0}\pi_h = 1$.

Lamb We let the observed data
$$y_i = \Lambda\eta_i + \epsilon_i, \quad \eta_i \overset{\text{iid}}{\sim} \sum_{h=1}^{k_0}\pi_h\mathrm{N}_d(\mu_h, \Delta_h), \quad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}_p(0, \Sigma),$$
where Λ is a $p \times d$ order sparse matrix with many entries equal to zero, $\mu_h \in \mathbb{R}^d$, Δ_h is a $d \times d$ positive definite matrix, for all $h = 1, \ldots, k_0$ and Σ is a $p \times p$ order diagonal matrix with positive entries.

Mixture of sparse factor analyzers (MFA) We let the observed data
$$y_i \overset{\text{iid}}{\sim} \sum_{h=1}^{k_0}\pi_h\mathrm{N}_p(\mu_h, \Lambda_h\Lambda_h^{\mathrm{T}} + \Sigma_h),$$
where Λ_h is a $p \times d$ order sparse matrix with many entries equal to zero, Σ_h is a $p \times p$ diagonal matrix with positive entries and $\mu_h \in \mathbb{R}^p$, for all $h = 1, \ldots, k_0$.

Mixture of log transformed zero inflated sparse Poisson counts (SpCount) Let $\{\ell_1, \ldots, \ell_p\}$ be a random permutation of $1, \ldots, p$, $r = \lfloor p/k_0 \rfloor$ and define the set $S_h = \{\ell_{(h-1)\times r+1}, \ldots, \ell_{h\times r}\}$ for all $h = 1, \ldots, k_0$. Thus $\{S_1, \ldots, S_{k_0}\}$ can be regarded as a random partition of $\{1, \ldots, p\}$ where each partition has r elements. Additionally fix k_0 positive constants $\lambda_1, \ldots, \lambda_{k_0}$, and let

$$w_{ij} \mid c_i = h \overset{\text{iid}}{\sim} \begin{cases} \mathrm{Pois}(\lambda_h) + \mathrm{N}(0,1) & \text{for all } j \in S_h, \\ 0 & \text{with probability 1 for all } j \notin S_h, \end{cases}$$

$$\Pr(c_i = h) = \pi_h \text{ for all } h = 1, \ldots, k_0.$$

where $\mathrm{Pois}(\lambda)$ is the Poisson distribution with mean λ and set $y_{ij} = \log(w_{ij}+1)$ for all $j = 1, \ldots, p$ and $i = 1, \ldots, n$. Thus the observed data y_i's are highly non-Gaussian within each cluster.

S.4. Additional Simulation Studies

S.4.1 Illustration of the Degenerate Clustering Behaviour

To show the degenerate clustering behavior discussed in Section 2 we performed two simple simulation experiments under the settings of Corollaries 2 and 3.

In the first experiment, we generate data from a five-component mixture model. Specifically, we assumed five well-separated Gaussians with equal proportions. The location vector for the h-th component is $\theta_h \mathbf{1}_p$ with $\mathbf{1}_p$ a p-dimensional vector of ones, $\theta_h \in \mathbb{R}$ and the values of θ_h ranging from -10 to +10. Each mixture component has identity covariance matrix. We fix $n = 10$ and $p = 20$. The left panel of Figure S.1 displays the distribution of the posterior median number of clusters in 100 replicates for a standard DP location mixture with hyperparameter specification satisfying Corollary 2 and proposed Lamb. For the DPM, we use the implementation in the `BNPmix` package (Corradin et al., 2021). Despite coming from a five-component mixture model, the data are grouped into a single cluster for most of the simulation replicates under the DPM specification, consistent with the limiting behavior described by Corollary 2.

In the second experiment, we assume a single p-variate normal distribution with mean zero, and identity covariance. As before, we fix $n = 10$ and $p = 20$. The results obtained assuming a DP mixture with the hyperparameter specification satisfying Corollary 3 and the proposed Lamb are reported in the right panel of Figure S.1. These results clearly show that the limiting behavior described by Corollary 3 is evident already for the moderate $p = 20$. Notably, the proposed Lamb avoids these pitfalls and is associated to a median number of clusters that is centered around the true values.
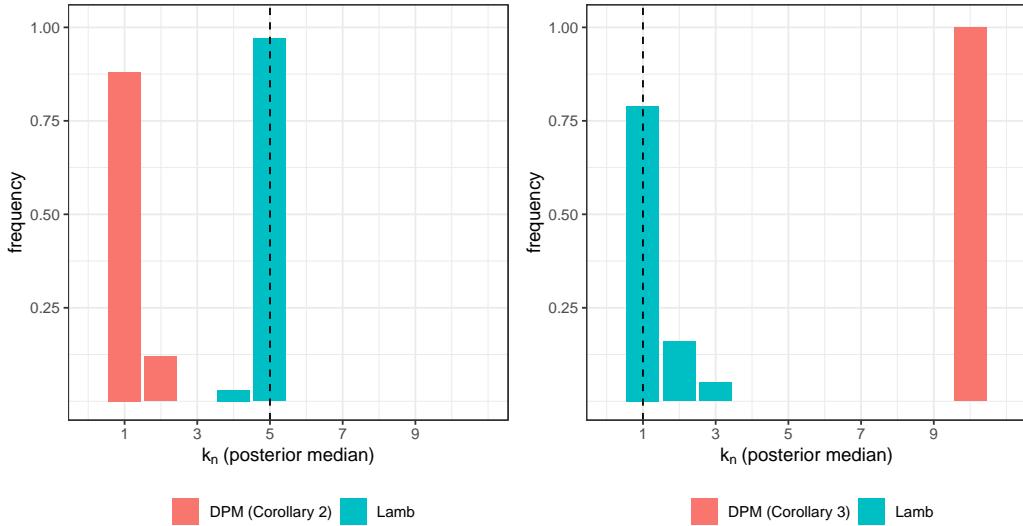
Figure S.1: Empirical distribution of the posterior median number of clusters in 100 replicates under the first (left) and second (right) simulation experiment. DPM hyperparameter specification satisfies Corollary 2 (left) and Corollary 3 (right). Vertical dashed lines represent the true number of clusters.

S.4.2 Recovering the Latent Space

To empirically illustrate the robustness of assumptions (C1) and (C2) used to prove the theory of Section 4, we perform a simple simulation study. These conditions ensure that the data contain increasing information on the latent factors as p increases. Increasing p means that we observe

additional y_{ij} variables for each subject. Each of these variables can have very small correlation with the latent factor η_{0i} and there will still be a build-up of information.

To see this, we generate random y_i for $i = 1, \ldots, 4$ and $p \in \{20, 200, 1000\}$. Data are generated as $y_i = \Lambda_0 \eta_{0i} + \epsilon_i$ where the factor loadings Λ_0's are generated according to

$$\lambda_{0jh} \overset{\text{iid}}{\sim} \pi \delta_{\{0\}} + (1 - \pi)\delta_{\{0.5\}},$$

where $\delta_{\{a\}}$ denotes a Dirac's delta mass at value a. The true latent factors are simulated as $\eta_{0i,j} \sim \mathrm{N}(i + j - 1, 0.05^2)$ where $\eta_{0i} = (\eta_{0i1}, \ldots, \eta_{0id})^{\mathrm{T}}$. We consider two error distributions ensuring low signal-to-noise ratio, and specifically $\epsilon_{ij} \sim \mathrm{N}(0, 25)$ and $\epsilon_{ij} \sim t_3$ where t_3 denotes a central t distribution with 3 degrees of freedom. We set $\pi = 0.2$ and the latent dimension $d = 2$.

To examine the level of recovery, for the m-th MCMC iteration, we regress the true factors with their current value in the m-th iteration. Specifically we stack all η_{0ij} across $i = 1, \ldots, n$ in a vector and use it as response variable, while using as predictor the vector containing all $\eta_{ij}^{(m)}$ of the m-th iteration. We do this for each iteration after the burn-in.
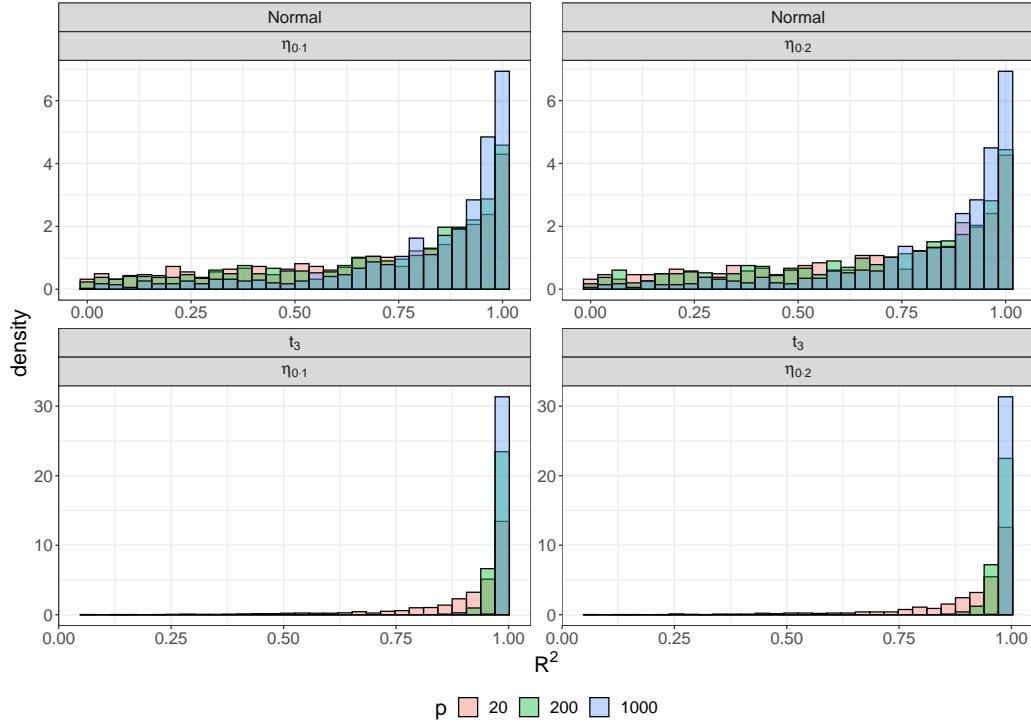
Figure S.2: Posterior distributions of the coefficient of determination R^2 of the linear regressions of the true latent factors on the associated posterior samples for two error distributions ($\mathrm{N}(0, 25)$, in the upper quadrants and t_3, in the bottom quadrants). The different dimensions p are denoted by the different colors.

Clearly, the latent factors are non-identifiable due to the well known rotational ambiguity and thus they can be learned only up to some non-singular matrix multiplication. Hence, to quantify

the accuracy in recovering the latent space, we consider the coefficient of determination R^2 of each fitted regression which is invariant of such identifiability issues. Figure S.2 reports the results. For both error distributions under consideration, as p grows the posterior distributions of the coefficients of determination concentrate near one implying that with more variables we improve on the learning of the latent space even with low signal-to-noise ratios.

S.4.3 Small Sample Studies

In this section, we do additional simulation studies. We consider the same setups considered in Section 5 of the main paper but take the sample size $n = 500$. The true number of clusters is fixed to $k_0 \in \{10, 15\}$. The results depicted in Figure S.3 are overall consistent with those reported in Section 5.
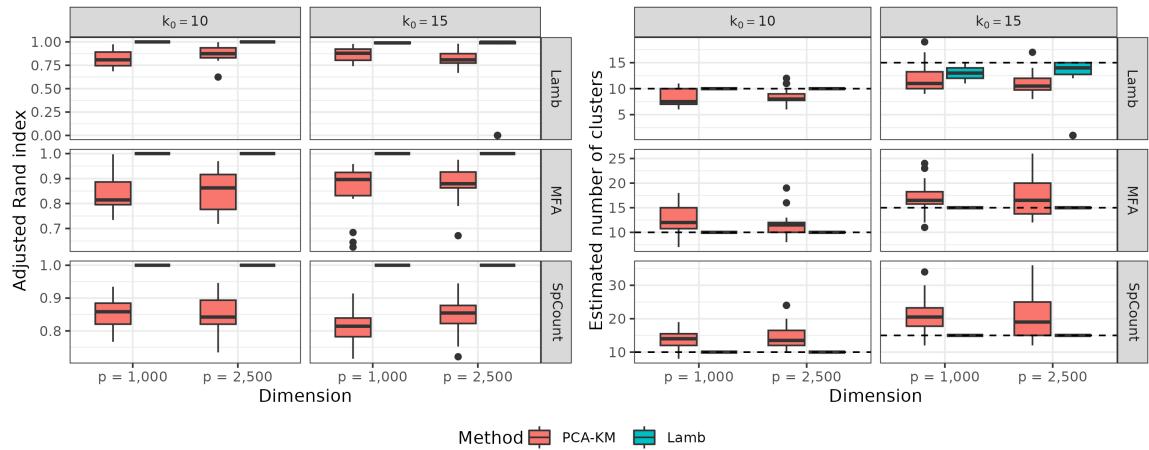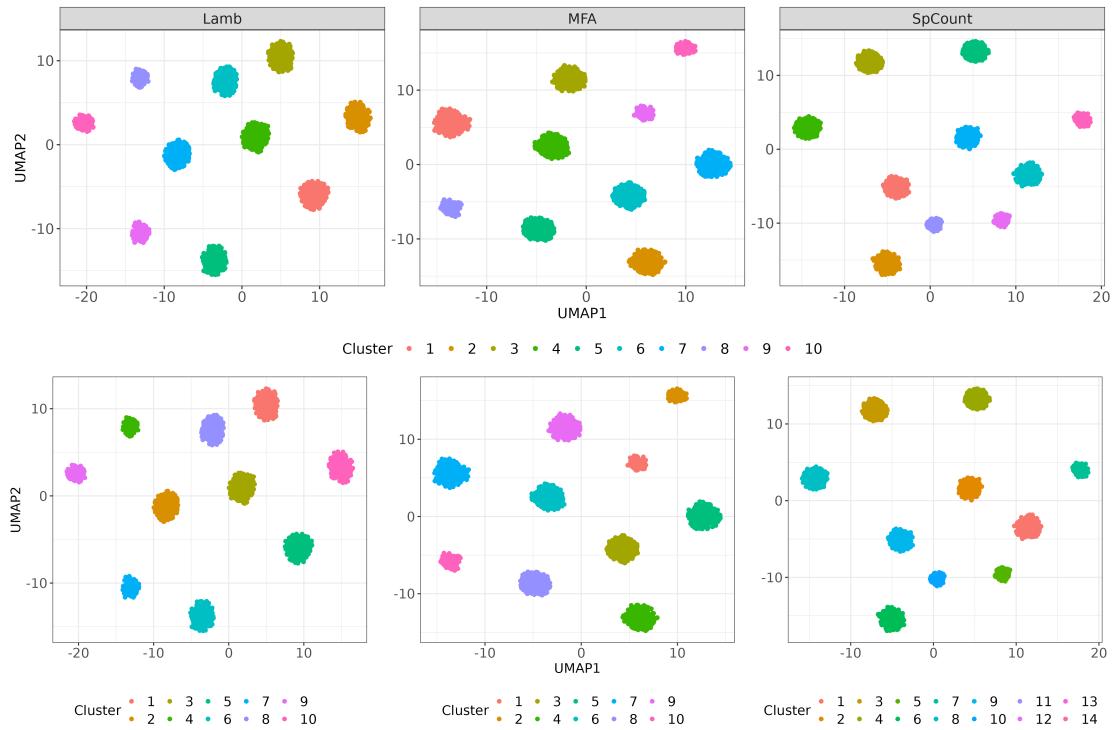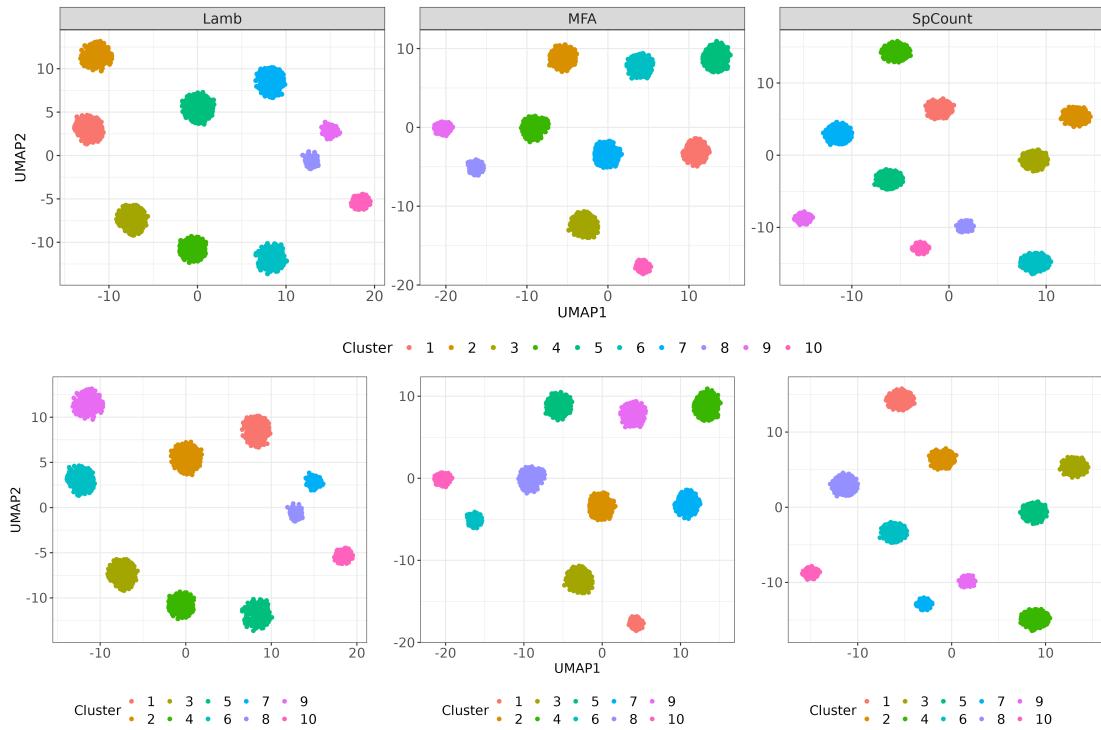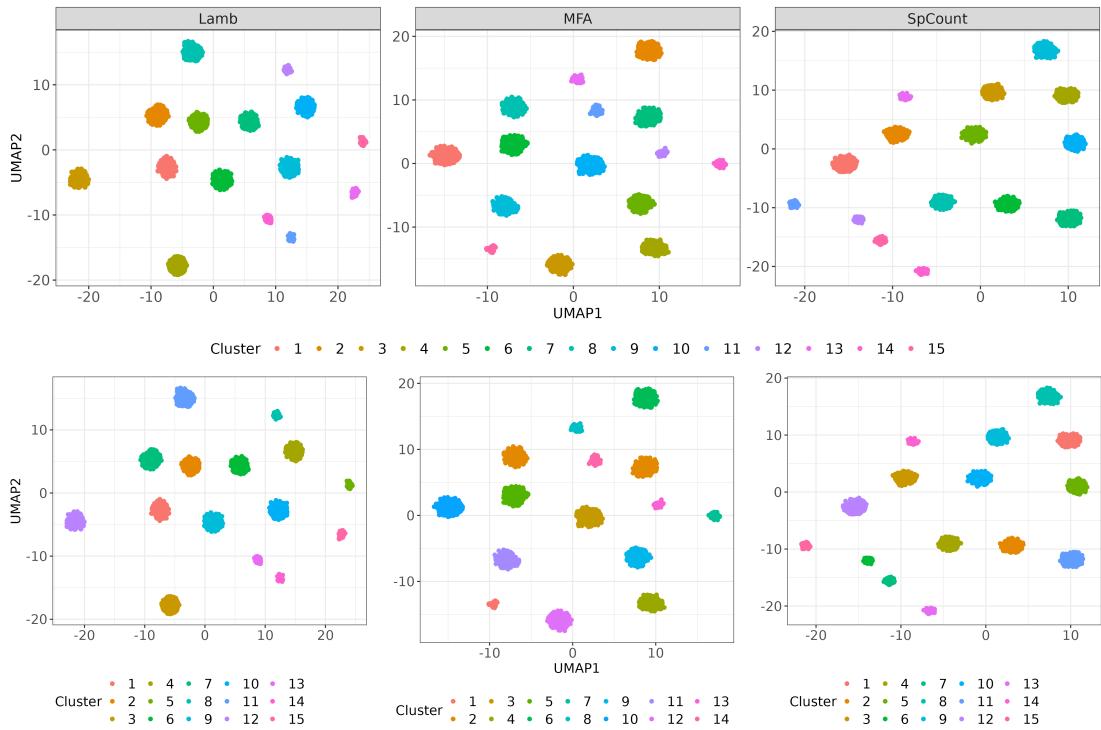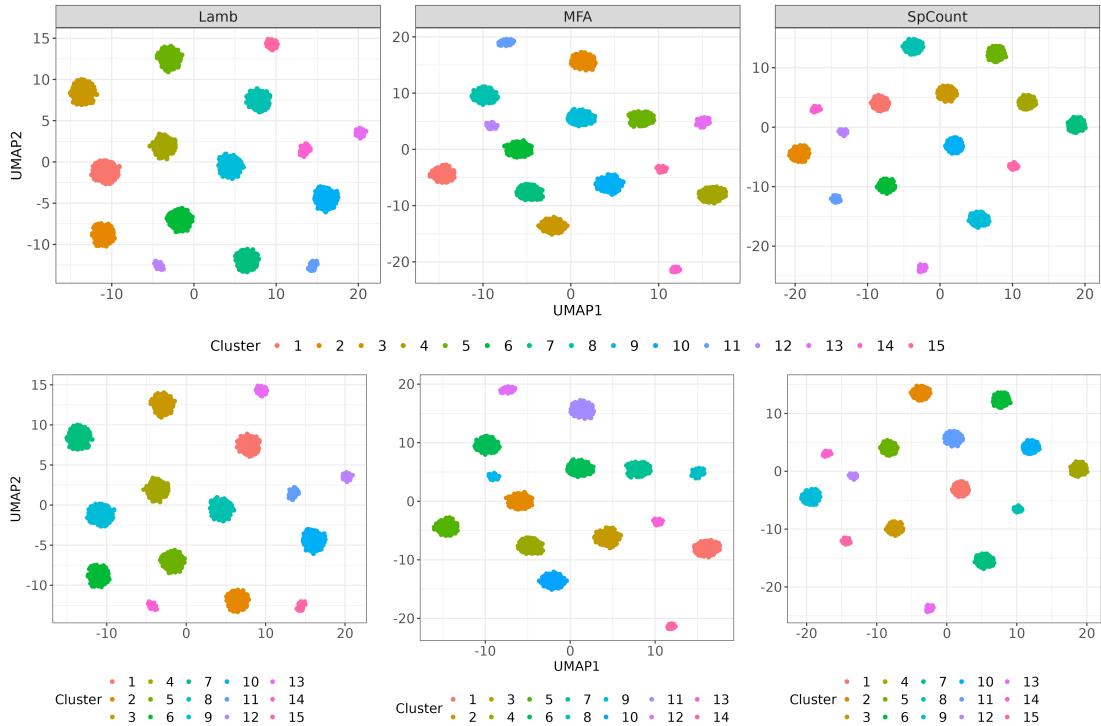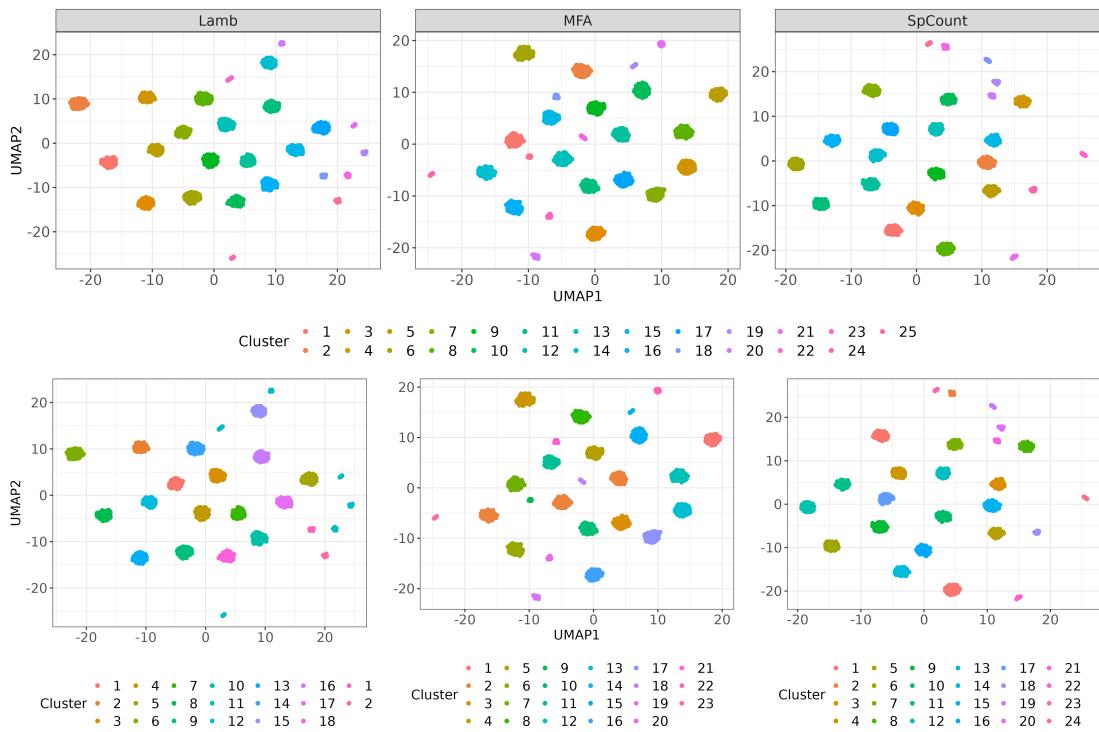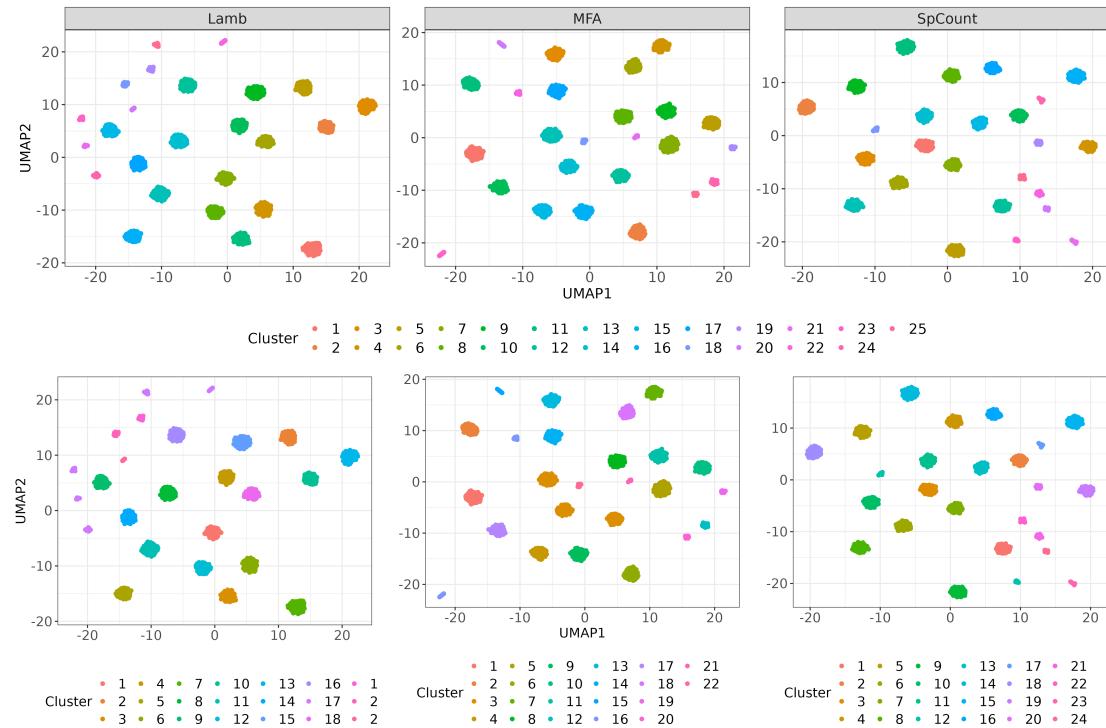
Figure S.3: Comparison between our proposed Lamb and the two-stage PCA-KM approach: Distributions of the adjusted Rand indices (left plot) and estimated number of clusters (right plot) in 20 replicated experiments. Horizontal dashed lines denote the true number of clusters. The true simulation scenarios, reported in each row, are labeled as Lamb for the model of Section 3, MFA for mixture of factor analyzers and SpCount for the $\log$ transformed zero inflated sparse Poisson counts.

S.4.4 Figures Associated to Section 5

Figures S.4-S.9 report the UMAP (McInnes et al., 2018) plots of the simulated datasets of Section 5, corresponding to the replicate with median adjusted Rand index (Rand, 1971). In each figure, the upper and lower panels show the true clustering and the estimated clustering obtained by the Lamb model, respectively. Each figure's caption specifies the true number of clusters (k_0) and the dimension (p).

Figure S.4: $k_0 = 10$, $p = 1,000$.


Figure S.5: $k_0 = 10$, $p = 2,500$

Figure S.6: $k_0 = 15$, $p = 1,000$


Figure S.7: $k_0 = 15$, $p = 2,500$

Figure S.8: $k_0 = 25$, $p = 1,000$

Figure S.9: $k_0 = 25$, $p = 2,500$

S.5. MCMC Convergence Diagnostics in the Cell Line Application

In this section, we provide convergence diagnostics of the MCMC sampler discussed in Section 3.2. Note that most of the variables that are sampled are latent objects and not identifiable. Hence we compute the $\log$-likelihood of $y_{1:n} \mid \Lambda, \eta, \Sigma$ across the MCMC samples. On these $\log$-likelihoods, we show traceplots and Geweke convergence diagnostics (Geweke, 1992) as implemented in the `coda R` package (Plummer et al., 2006). The results are shown in Figure S.10 and they indicate evidence towards good mixing.
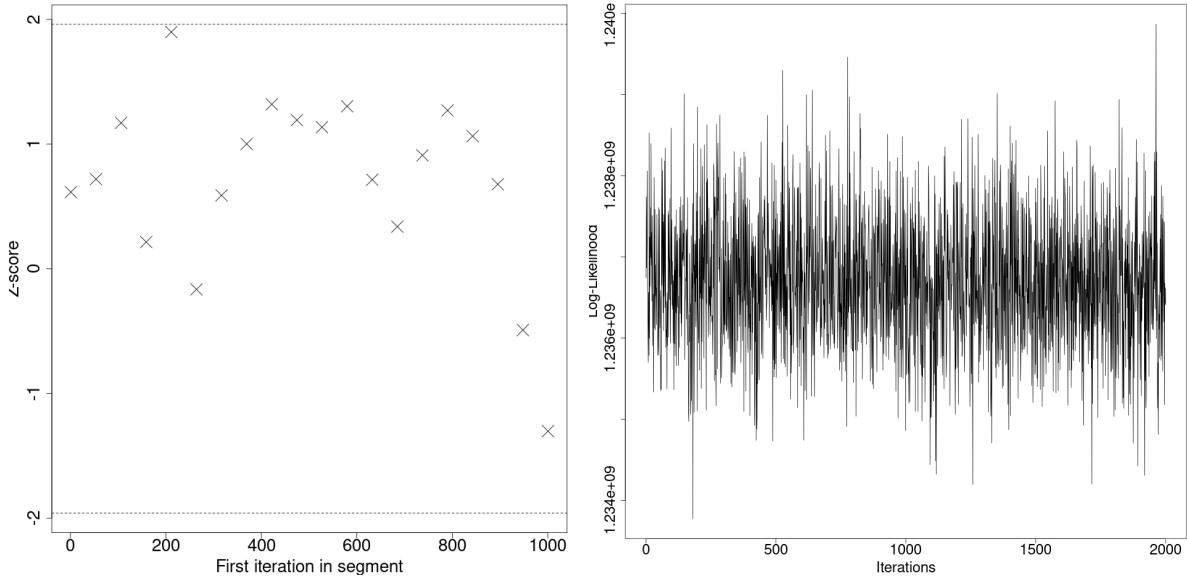
Figure S.10: MCMC Convergence Diagnostics in the Cell Line Application: The joint $\log$-likelihoods of the $y_{1:n} \mid \Lambda, \eta, \Sigma$ are first calculated across the MCMC iterations. The Geweke convergence diagnostic on the $\log$-likelihoods is shown in the left panel and their traceplot in the right panel.