Electronic Journal of Statistics
Vol. 14 (2020) 3869–3913
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1756

Anchored Bayesian Gaussian mixture models

Deborah Kunkel

School of Mathematical and Statistical Sciences
Clemson University, Clemson, SC, USA
e-mail: dekunke@clemson.edu

and

Mario Peruggia

Department of Statistics
The Ohio State University, Columbus, OH, USA
e-mail: peruggia@stat.osu.edu

Abstract: Finite mixtures are a flexible modeling tool for irregularly shaped densities and samples from heterogeneous populations. When modeling with mixtures using an exchangeable prior on the component features, the component labels are arbitrary and are indistinguishable in posterior analysis. This makes it impossible to attribute any meaningful interpretation to the marginal posterior distributions of the component features. We propose a model in which a small number of observations are assumed to arise from some of the labeled component densities. The resulting model is not exchangeable, allowing inference on the component features without post-processing. Our method assigns meaning to the component labels at the modeling stage and can be justified as a data-dependent informative prior on the labelings. We show that our method produces interpretable results, often (but not always) similar to those resulting from relabeling algorithms, with the added benefit that the marginal inferences originate directly from a well specified probability model rather than a *post hoc* manipulation. We provide asymptotic results leading to practical guidelines for model selection that are motivated by maximizing prior information about the class labels and demonstrate our method on real and simulated data.

Keywords and phrases: Label switching, data-dependent prior, identifiability, EM algorithm.

Received October 2019.

1. Introduction

Finite mixture models are flexible tools that are often applied to data from heterogeneous populations or from distributions with irregularly-shaped densities. In these context, they can produce useful approximations to the unknown density functions in both univariate and multivariate settings [Frühwirth-Schnatter, 2006, Marin and Robert, 2014, Rossi, 2014]. Results concerning the accuracy and consistency of the approximations (as the number of components increases at an appropriate rate) have been established both in the frequentist and in the Bayesian settings [Roeder and Wasserman, 1997, Genovese and Wasserman,

2000, Norets and Pelenis, 2012]. In many situations, if the true density is well-behaved in the tails, satisfactory approximations can be obtained using a small or moderate number of mixture components.

When mixture distributions are used to model heterogeneous populations, the mixture components are thought to represent clusters of similar units. Such analyses are found in areas such as medicine, the social sciences, and genetics, where identifying subgroups of similar individuals may help to generate hypotheses for future research. When subgroup identification is an important goal, as it is in this paper, the parameters of the component distributions provide population-level information about the features of groups, which can elucidate the overarching patterns of heterogeneity within a population. Therefore, accurate estimates of component-specific parameters, with attendant measures of uncertainty, become a vital element of inference. Further, the mixture model allows estimation of a probabilistic clustering structure from the data.

Adopting standard notation [e.g., Frühwirth-Schnatter, 2006], we represent the likelihood for a k-component finite mixture model for a response $\boldsymbol{y} = (y_1, \ldots, y_n)$ as

$$f(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\eta}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \eta_j p(y_i|\gamma_j), \tag{1}$$

where $p(\cdot|\gamma_j)$ denotes the jth component density. The model is parameterized by $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$, the vector of mixture proportions whose elements sum to one, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)$, the vector of features of the component densities. It is often helpful to write the model (1) hierarchically, using latent variables $\boldsymbol{S} = \{S_1, \ldots, S_n\}$, $S_i \in \{1, \ldots, k\}$, $i = 1, \ldots, n$, to indicate component membership. The resulting likelihood is

$$f(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} p(y_i|\gamma_{s_i}), \text{ where } P(S_i = j|\boldsymbol{\eta}) = \eta_j, \quad j = 1, \ldots, k, \ i = 1, \ldots, n. \tag{2}$$

If the population comprises well-understood groups, it is appropriate to incorporate information regarding the groups' relative locations and scales into the prior on $(\boldsymbol{\gamma}, \boldsymbol{\eta})$. Often, however, little is known about these groups ahead of time, and it is natural to assume prior exchangeability of component features. Let q index the $k!$ possible permutations of the integers $1, \ldots, k$ and let $\rho_q(\cdot)$ relabel its argument according to the qth permutation. We will use the same symbol $\rho_q(\cdot)$ to denote the action of the qth permutation on a given index between 1 and k and on the elements of a vector argument. For example, if the qth permutation of $(1, 2, 3, 4)$ is $(1, 3, 4, 2)$, then $\rho_q(3) = 4$ and $\rho_q(a_1, a_2, a_3, a_4) = (a_1, a_3, a_4, a_2)$.

An exchangeable prior with density π satisfies $\pi(\boldsymbol{\gamma}, \boldsymbol{\eta}) = \pi(\rho_q(\boldsymbol{\gamma}, \boldsymbol{\eta}))$, $q = 1, \ldots, k!$. This specification produces a posterior distribution that inherits the same label invariance; that is, letting $p_E(\cdot|\boldsymbol{y})$ denote the posterior density of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ under the exchangeable model,

$$p_E(\boldsymbol{\gamma}, \boldsymbol{\eta}|\boldsymbol{y}) = p_E(\rho_q(\boldsymbol{\gamma}, \boldsymbol{\eta})|\boldsymbol{y}), \quad q = 1, \ldots, k!. \tag{3}$$

The posterior density is symmetric with respect to the $k!$ labelings of the components, often producing $k!$ modal regions in the parameter space. The marginal distributions of the component-specific parameters are identical. When Markov

Chain Monte Carlo methods are used to sample from the posterior distribution of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$, a well-mixing chain will jump from one possible labeling to another, a phenomenon referred to as "label-switching."

The posterior symmetry and accompanying label-switching in no way hinders the model's predictive performance, nor does it preclude meaningful inference on objects that do not depend on the component labels. When label-switching occurs, however, ergodic averages cannot be used for inference on the component-specific features, making it impossible to use labeled parameters to learn about distinctions among the mixture components. Much work has been devoted to either preventing or reversing label-switching by placing prior constraints on the parameter space or by post-processing posterior samples in a way that allows only one possible labeling of the mixture components. These approaches, particularly the post-processing approach, are popular in practice.

Prior identifiability constraints create a non-exchangeable prior by requiring $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ to lie in some sub-region of the parameter space that is compatible with only one possible labeling. For example, one could require that $\eta_1 < \ldots < \eta_k$ with probability 1, or establish similar ordering constraints among the component feature parameters. The limitations of these approaches are addressed in detail by, among others, Celeux, Hurn and Robert [2000] and Jasra, Holmes and Stephens [2005]. They are often considered too informative in their strict restrictions of the parameter space and may not effectively isolate a single modal region of the posterior density. It is not always obvious, a priori, what choice of constraint is appropriate for a problem.

Relabeling algorithms, such as those presented by Stephens [2000], Celeux, Hurn and Robert [2000], Marin, Mengersen and Robert [2005], Papastamoulis and Iliopoulos [2010], Bardenet et al. [2012], Rodriguez and Walker [2014], and Li and Fan [2016], tend to be preferred. These algorithms specify a loss function and find the labeling that minimizes the loss function for each posterior sample of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ and, if sampled, $\boldsymbol{s}$. Upon convergence, each unique value of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ is restricted to only one possible labeling. For this reason, Jasra, Holmes and Stephens [2005] have described this strategy as a way of automatically applying an identifiability constraint. Relabeling algorithms often appear to perform "better" than prior constraints, in that they produce relabeled posterior samples that have unimodal and well-separated marginal densities. In contrast to methods based on prior constraints, however, it is not straightforward to obtain expressions for the joint or marginal distributions of the elements of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ corresponding to a relabeling method. The constrained region of the parameter space is the solution to the iterative minimization of the chosen loss function, and, as such, cannot be described concisely as a component of the probability model. Because its constraints are not the result of a clearly defined prior specification, it is difficult to evaluate rigorously the underlying structure that the relabeling algorithm imposes upon a problem. It is not obvious whether this approach can be justified as a basis for making inferential claims about the posterior distribution of the component-specific parameters.

We introduce a modification to the standard finite mixture model, the *anchor* model, in which a small number of observations are assumed to be drawn from

known component densities. This breaks the model's label invariance in a data-dependent manner while avoiding the strong, subjective restrictions imposed by prior identifiability constraints. The anchor model provides an appropriate basis upon which to interpret the component-specific feature parameters by assigning meaning a priori to their labels. The proposed modeling framework requires a modest amount of pre-processing to identify the pre-classified observations but avoids the computational burden of post-processing.

The rest of the paper is organized as follows. In Section 2 we describe our proposed anchored mixture model and its basic properties. We also examine the implications of its asymptotic behavior on the identifiability of the component-specific feature parameters. In Section 3 we outline practical strategies for model specification. In Sections 4 and 5 we present data analysis examples that make use of our proposed methodology. In Section 6 we present concluding remarks and discuss directions for possible future developments. Supplemental materials are presented in Appendices A–F.

2. Anchor models

The idea of assuming known labels for some observations has been considered by Chung, Loken and Schafer [2004], who present this as a way of specifying an informative prior, and, more recently, by Egidi et al. [2018], who propose a post-processing strategy that assigns labels to observations with zero *posterior* probability of being allocated to the same mixture component. Related approaches that disallow specific allocations of the observations to the various mixture components have been suggested as a means of guaranteeing propriety of the posterior distribution if improper priors are specified [Diebolt and Robert, 1994, Wasserman, 2000]. We build on these ideas by formalizing this strategy as a modeling procedure that requires no post-processing of an MCMC sample. A careful assignment of a small number of observations to specific components yields a well-defined mixture model whose components can accurately reflect homogeneous subgroups in the population. We define the anchor model and describe several of its basic properties in the following sections.

2.1. Definition of an anchor model

The anchor model modifies the finite mixture likelihood by selecting a small number of observations whose component labels are assumed to be known. These observations will be called *anchor points*. The resulting model can be fully described using k index sets A_j, $j = 1, \ldots, k$, where A_j contains the indices of those observations in the data set that are to be "anchored" to the jth component and $A = \{A_1, \ldots, A_k\}$ is the set of indices of all anchor points. The likelihood in (1) is replaced by

$$f_A(y_i|\boldsymbol{\gamma}, \boldsymbol{\eta}) = \begin{cases} \sum_{j=1}^{k} \eta_j p(y_i|\gamma_j), & i \notin A, \\ p(y_i|\gamma_j), & i \in A_j, \quad j = 1, \ldots, k. \end{cases} \tag{4}$$

We use m_j to denote the number of points anchored to the jth component and $m = \sum_{j=1}^{k} m_j$ to denote the total number of anchor points. Some of the A_j may be empty and the number of components that contain one or more anchor points is denoted by $k_0 \leq k$. The anchor model can equivalently be represented using latent allocations: if i is the index of an anchored observation, $P(S_i = j) = 1$ for one prespecified component j. This restricts the support of $\boldsymbol{S}$ so that a subset of the possible allocation vectors has prior probability of 0. The hierarchical representation of the probability density for y_i under anchor model A is $f(y_i|S_i = s_i, \boldsymbol{\gamma}) = p(y_i|\gamma_{s_i})$, where

$$P_A(S_i = j|\boldsymbol{\eta}) = \begin{cases} \eta_j, & i \notin A, \\ 1, & i \in A_j, \\ 0, & i \in A_{j'}, \quad \text{with } j' \in \{1, \ldots, k\} \setminus \{j\}, \end{cases} \tag{5}$$

Since an observation can be anchored to at most one component, we require $A_j \cap A_h = \emptyset$ for $j = 1, \ldots, k_0 - 1$ and $h = j + 1, \ldots, k_0$. To impose a unique labeling on each anchor model, we will further require that $A_j \neq \emptyset$ for $j = 1, \ldots, k_0$, (if any components have no anchor points, they will be labeled $k_0 + 1, \ldots, k$) and that $\min_i(A_1) < \min_i(A_2) < \ldots < \min_i(A_{k_0})$. We will denote the values of the anchor points by $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$, where $\boldsymbol{x}_j = \{y_i : i \in A_j\}$.

For the remainder of this paper, we consider the anchored Gaussian mixture model, in which $p(\cdot|\boldsymbol{\gamma}_j)$ is a p-variate Normal distribution with density denoted by $\phi_p(\cdot; \boldsymbol{\gamma}_j)$ and $\boldsymbol{\gamma}_j = (\theta_j, \Sigma_j)$, where θ_j and Σ_j are the mean vector and covariance matrix of the jth component density, respectively. We assume that $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ are a priori independent and that their distributions satisfy two conditions:

C.1: The prior on the mixture proportions, $\eta_1, \ldots, \eta_k$, is a Dirichlet distribution with concentration parameter $\alpha \mathbf{1}_k$, where $\mathbf{1_k}$ is a vector of k ones and $\alpha > 0$.

C.2: The prior on $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)$ has the form $\prod_{j=1}^{k} \pi(\boldsymbol{\gamma}_j|\xi)$, for some continuous density π with (possibly unknown) hyperparameters ξ, and π is positive on an open subset of the parameter space.
For ease of notation, we will suppress ξ from notation and refer to the prior density of $\boldsymbol{\gamma}_j$ using $\pi(\boldsymbol{\gamma}_j)$.

2.2. Basic properties

In this section, we discuss some features of an anchor model which may be readily understood via the latent allocation representation in (5). The notation $\mathcal{S}$ will denote the set of all k^n possible allocation vectors of length n, the sample space of the latent variable $\boldsymbol{S}$ under the exchangeable model. Each allocation vector separates the data into k or fewer groups of observations and we will refer to each unique grouping as a "partition" of the data. All allocation vectors that are equal up to a relabeling of the component labels induce the same partition of the data; e.g., we will say that the allocations $(1, 2, 2, 2, 3)$ and $(2, 1, 1, 1, 3)$ induce the same partition.

Let $\mathcal{S}^A$ be the subset of allocations that have nonzero probability under $A = \{A_1, \ldots, A_k\}$, an anchor model with m anchor points. Then $\mathcal{S}^A$ contains k^{n-m} elements and A assigns probability zero to every allocation that satisfies $s_i = s_{i'}$, for some $i \in A_j$ and $i' \in A_{j'}$, $j \neq j'$. Consequently, all partitions which assign i and i' to the same group have probability zero. This is a key difference between anchor models and relabeling methods that also restrict the set of allocations such as those of Papastamoulis and Iliopoulos [2010] and Rodriguez and Walker [2014]. The relabeling in those references creates a restricted set of allocations that includes exactly one labeling for each partition but does not eliminate any partitions.

Under mild conditions, the anchor model admits no labeling ambiguity, thus eliminating the symmetry of the exchangeable model. The following proposition is a direct consequence of the definition of the anchor models and is proved in the Appendix.

Proposition 1. *The following two statements hold under conditions C.1 and C.2.*

1. *An anchor model $A = \{A_1, \ldots, A_k\}$ imposes a unique labeling on each partition that has nonzero probability if and only if $A_1, \ldots, A_{k-1}$ are nonempty; that is, $k_0 \geq k - 1$.*
2. *For any $j \leq k_0$, $j \neq j'$, the marginal posterior density of $\boldsymbol{\gamma}_j$ is distinct from the marginal posterior density of $\boldsymbol{\gamma}_{j'}$ with probability 1.*

To get an intuition of why the second statement of Proposition 1 holds, notice that the observations anchored to component j contribute to the updating of the distribution of $\boldsymbol{\gamma}_j$ with probability 1, but never contribute to the updating of $\boldsymbol{\gamma}_{j'}$. Any anchor model satisfying the conditions in Proposition 1 produces distinct posterior distributions for the component-specific features. The next two sections describe properties that aid in evaluating which anchor models are most effective at separating distinct groups in the sample.

2.3. Model evidence

One key advantage of the anchor model is that each set of anchor points results in a unique, well-defined probability model, making it possible to compare different anchor models using standard model selection criteria. The goodness-of-fit of an anchor model A may be evaluated using the model marginal likelihood, defined as $m_A(\boldsymbol{y}) = \int f_A(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\eta})\pi(\boldsymbol{\gamma})\pi(\boldsymbol{\eta})d(\boldsymbol{\gamma}, \boldsymbol{\eta})$. This expression can be expressed in terms of the latent allocations as

$$m_A(\boldsymbol{y}) = \sum_{\boldsymbol{s} \in \mathcal{S}^A} m(\boldsymbol{y}|\boldsymbol{s})p_A(\boldsymbol{s}), \tag{6}$$

where we define $m(\boldsymbol{y}|\boldsymbol{s}) = \int f(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{s})\pi(\boldsymbol{\gamma})d\boldsymbol{\gamma}$ and $p_A(\boldsymbol{s}) = \int p_A(\boldsymbol{s}|\boldsymbol{\eta})\pi(\boldsymbol{\eta})d\boldsymbol{\eta}$. Based on Equation (6), the goodness of fit of an anchor model A will be determined by a weighted average of the values $m(\boldsymbol{y}|\boldsymbol{s})$ over all allocations in $\mathcal{S}^A$. The terms $m(\boldsymbol{y}|\boldsymbol{s})$ describe the fit of the model conditional on the partition induced by $\boldsymbol{s}$: broadly, well-fitting anchor models will be those for which the elements of $\mathcal{S}^A$ induce partitions that are supported by the data.

A closed-form expression for $m(\boldsymbol{y}|\boldsymbol{s})$ is available for some models, which can provide heuristic, generalizable insight into which points should be anchored. For example, consider a univariate location mixture model with σ^2 known, so that the component-specific parameter γ_j is simply the mean of the jth Gaussian component. Under the Gaussian prior $\pi(\boldsymbol{\gamma}) = \prod_{j=1}^{k} \phi_1(\gamma_j; \mu, \tau^2)$ with μ and τ^2 known, the conditional marginal likelihood satisfies the condition

$$m(\boldsymbol{y}|\boldsymbol{s}) \propto \exp\left(\frac{-1}{2}\sum_{j=1}^{k}\left(\sum_{i:s_i=j}\frac{(y_i-\bar{y}_j[\mathbf{s}])^2}{\sigma^2}+\frac{\mu^2\tau^{-2}(1+\sigma^2)-n_j\,(\bar{y}_j[\mathbf{s}]-\mu)^2}{(n_j\tau^2+\sigma^2)}\right)\right)\prod_{j=1}^{k}\sqrt{\frac{n_j\tau^2+\sigma^2}{\sigma^2}}, \quad (7)$$

where $n_j = \sum_{i=1}^{n} I(s_i = j)$, and $\bar{y}_j[\boldsymbol{s}] = n_j^{-1}\sum_{i:s_i=j} y_i$.

From Equation (7) we see that, when τ^2 is large relative to σ^2, the magnitude of $m(\boldsymbol{y}|\boldsymbol{s})$ is determined primarily by the term $\sum_{j=1}^{k}\sum_{i:s_i=j}(y_i-\bar{y}_j[\mathbf{s}])^2$, the within-group sum of squares, for the partition induced by $\boldsymbol{s}$. This observation suggests a heuristic notion: well-fitting anchor models will be those for which $\mathcal{S}^A$ contains many allocations that produce well-separated groups in the data. The marginal likelihood on its own is impractical for model selection because the large cardinality of $\mathcal{S}^A$ makes exact computation of the expression in (6) impossible even for moderate values of n and/or k. Consideration of this expression, nonetheless, suggests that, in specifying anchor models, we should promote separation among the mixture components. In Section 3, we propose a computationally feasible method for specifying anchor models that encourage separation and will tend to fit well.

2.4. Anchoring as an informative prior on $\boldsymbol{\gamma}, \boldsymbol{\eta}$

Replacing the exchangeable model with the anchor model (4) can be viewed as creating a data-dependent, non-exchangeable prior on the component-specific parameters.

An anchor model with anchor points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{k_0}$ produces a posterior density of $\boldsymbol{\gamma}, \boldsymbol{\eta}$ that satisfies

$$\begin{aligned} p_A(\boldsymbol{\gamma}, \boldsymbol{\eta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\eta})\left(\prod_{j=1}^{k_0}\pi(\boldsymbol{\gamma}_j)\phi_p(\boldsymbol{x}_j;\boldsymbol{\gamma}_j)\right)\left(\prod_{j=k_0+1}^{k}\pi(\boldsymbol{\gamma}_j)\right)\left(\prod_{i\notin A}\sum_{j=1}^{k}\eta_j\phi_p(y_i;\boldsymbol{\gamma}_j)\right) \\ &= \pi(\boldsymbol{\eta})\left(\prod_{j=1}^{k_0}C_j p(\boldsymbol{\gamma}_j|\boldsymbol{x}_j)\right)\left(\prod_{j=k_0+1}^{k}\pi(\boldsymbol{\gamma}_j)\right)\left(\prod_{i\notin A}\sum_{j=1}^{k}\eta_j\phi_p(y_i;\boldsymbol{\gamma}_j)\right) \quad (8) \end{aligned}$$

where $C_j = \int \pi(\boldsymbol{w})\phi_p(\boldsymbol{x}_j;\boldsymbol{w})d\boldsymbol{w}$ and $p(\boldsymbol{\gamma}_j|\boldsymbol{x}_j)$ denotes the posterior density that results from updating the distribution of $\boldsymbol{\gamma}_j$ with the anchor points $\boldsymbol{x}_j$. Because C_j does not depend on $\boldsymbol{\gamma}_j$, the following proposition holds.

Proposition 2. *The anchor model described in this section produces the same posterior distribution on* $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ *as a model whose likelihood is a Gaussian mixture on the* $n - m$ *unanchored observations and whose prior is equal to*

$$\pi(\boldsymbol{\eta}) \prod_{j=1}^{k_0} p(\boldsymbol{\gamma}_j | \boldsymbol{x}_j) \prod_{j=k_0+1}^{k} \pi(\boldsymbol{\gamma}_j),$$

where $p(\boldsymbol{\gamma}_j | \boldsymbol{x}_j)$ *is the posterior density of* $\boldsymbol{\gamma}_j$ *given the anchor points* $\boldsymbol{x}_j$.

Proposition 2 provides a basis for asymptotic results to be discussed in the next section.

2.5. Large sample properties and quasi-consistency

In this section we establish the asymptotic properties of an anchor model and define a derived notion of *quasi-consistency* that enables us to quantify probabilistically the degree of component-specific parameter identifiability attained by an anchor model.

2.5.1. Limiting behavior of the posterior distribution

To characterize the limiting behavior of an anchor model, we rely on a result from Cooley and MacEachern [1999] which describes the large sample properties of the posterior distribution of the model parameters $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ in a mixture model in the setting where prior information, possibly from pre-labeled samples, is available. Applying results of Berk [1966], they derived statements that, assuming appropriate regularity conditions, hold with probability one with respect to the product measure $F_{\boldsymbol{\gamma}_0, \boldsymbol{\eta}_0}$ on the space of sample paths of the data-generating process with true model parameters $(\boldsymbol{\gamma}_0, \boldsymbol{\eta}_0)$.

In addition to C.1 and C.2 on page 3873, the subsequent results require an additional assumption on the prior density $\pi(\boldsymbol{\gamma})$. The additional assumption asks that

C.3: $\pi(\rho_q(\boldsymbol{\gamma}_0)) > 0, \quad$ for some $q \in \{1, \ldots, k!\}$,

that is, the prior density is positive at some relabeling of the true parameter value. Define also $\Gamma_0 = \{\rho_q(\boldsymbol{\gamma}_0, \boldsymbol{\eta}_0),\ q = 1, \ldots, k!\}$, the set containing all possible labelings of $(\boldsymbol{\gamma}_0, \boldsymbol{\eta}_0)$. Under C.1-C.3, the results in Cooley and MacEachern [1999] give the following. Let U denote an arbitrary open neighborhood of Γ_0. Then,

$$\lim_{n \to \infty} \Pi(U \,|\, y_1, \ldots, y_n) = 1, \ \ a.s. - F_{\boldsymbol{\gamma}_0, \boldsymbol{\eta}_0}, \tag{9}$$

where $\Pi(\cdot \,|\, y_1, \ldots, y_n)$ is the posterior probability measure on the parameter space, given a random sample $y_1, \ldots, y_n$ of size n. In addition, let $N_\epsilon(\boldsymbol{\gamma}, \boldsymbol{\eta})$ denote an open ball of radius $\epsilon > 0$ centered at $(\boldsymbol{\gamma}, \boldsymbol{\eta})$. Consider the qth relabeling

$\rho_q(\boldsymbol{\gamma}_0, \boldsymbol{\eta}_0)$ of the true model parameters, and assume that ϵ is small enough for $\bigcap_{h=1}^{k!} N_\epsilon(\rho_h(\boldsymbol{\gamma}_0, \boldsymbol{\eta}_0)) = \emptyset$ to hold. Then,

$$\lim_{n\to\infty} \Pi(N_\epsilon\left(\rho_q(\boldsymbol{\gamma}_0, \boldsymbol{\eta}_0)\right) \mid y_1, \ldots, y_n) = \frac{\pi(\rho_q(\boldsymbol{\gamma_0}, \boldsymbol{\eta}_0))}{\sum_{h=1}^{k!} \pi(\rho_h(\boldsymbol{\gamma}_0, \boldsymbol{\eta}_0))}, \quad a.s. - F_{\boldsymbol{\gamma}_0, \boldsymbol{\eta}_0}, \tag{10}$$

where π is the prior density of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$. Combined, results (9) and (10) indicate that the posterior distribution concentrates on Γ_0 and, in the limit, the relative posterior mass given to the qth element of Γ_0 is determined solely by its prior density.

It is natural to interpret the limiting values in (10) as defining a discrete asymptotic probability distribution on the $k!$ elements of Γ_0, where the probabilities are determined solely by the prior. Under the exchangeable model, $\pi(\cdot)$ is equal for all elements of Γ_0, and we obtain a discrete *uniform* distribution on its $k!$ elements: no matter how much additional data accumulate, all relabelings of the true parameter remain equally likely.

For an anchor model, we can use the data-dependent prior given in Proposition 2 to determine the limiting values in (10). The probability of the qth relabeling of $(\boldsymbol{\gamma}_0, \boldsymbol{\eta}_0)$ is in fact determined by the anchor points: it is proportional to $\prod_{j=1}^{k_0} p(\boldsymbol{\gamma}_{0\rho_q(j)} | \boldsymbol{x_j})$. (See the proof of Proposition 3. In particular, the expression does not depend on $\boldsymbol{\eta}_0$ because the anchor points provide no information about the mixture proportions.) For this reason, we will use $P_{\boldsymbol{x}}(\boldsymbol{\gamma}_0) = \{p_q,\ q = 1, \ldots, k!\}$ to denote the asymptotic distribution on Γ_0 induced by a set of anchor points $\boldsymbol{x}$, or $P_{\boldsymbol{x}}$ when there is no ambiguity. It is straightforward to derive expressions for the elements of $P_{\boldsymbol{x}}$, which depend only on the Gaussian densities of the anchor points at $\boldsymbol{\gamma}_0$, as stated in the ensuing proposition which is proved in Appendix A.

Proposition 3. *The qth element of* $P_{\boldsymbol{x}}(\boldsymbol{\gamma}_0)$, p_q, *is equal to*

$$\prod_{j=1}^{k_0} \phi_p\left(\boldsymbol{x}_j; \boldsymbol{\gamma}_{0\rho_q(j)}\right) \Big/ \sum_{h=1}^{k!} \prod_{j=1}^{k_0} \phi_p\left(\boldsymbol{x}_j; \boldsymbol{\gamma}_{0\rho_h(j)}\right). \tag{11}$$

2.5.2. Quantifying parameter identifiability

The result in (9) states that, as the sample size goes to infinity, the posterior mass concentrates on arbitrarily small neighborhoods of the $k!$ relabelings of the true value $(\boldsymbol{\gamma}_0, \boldsymbol{\eta}_0)$. Thus, for both the exchangeable model and the anchored model, the asymptotic distributions concentrate on the elements of Γ_0. The influence of the anchor points does not disappear in large samples, however; they determine, through $P_{\boldsymbol{x}}$, the relative posterior mass given to the $k!$ modal regions that are treated symmetrically under the exchangeable model. If the distribution $P_{\boldsymbol{x}}$ assigns high probability to one relabeling, then for large samples, the density $p_A(\boldsymbol{\gamma}, \boldsymbol{\eta} | \boldsymbol{y})$ will approximate, up to a constant factor, the density $p_E(\boldsymbol{\gamma}, \boldsymbol{\eta} | \boldsymbol{y})$ around one of the modal regions and will tend to be flat elsewhere.

The anchor points play a crucial role in disambiguating between class labels and determining the degree with which the anchor model isolates one posterior mode. Anchor models for which $P_{\boldsymbol{x}}$ places high probability on only one relabeling of $(\boldsymbol{\gamma}_0, \boldsymbol{\eta}_0)$ should be preferred. This motivates the following definition.

Definition 1. *Let* $P_{\boldsymbol{x}}(\boldsymbol{\gamma}_0) = \{p_q,\ q = 1, \ldots, k!\}$ *be the limiting probability defined above for an anchor model with anchor points* $\boldsymbol{x}$*. We say that the anchor model is* α quasi-consistent *if* $\max_{q \in \{1,\ldots,k!\}} p_q = \alpha$.

The ideal value of α is one; this is not attainable in practice because all probabilities in (30) are positive, but, for a good anchor model, α will be near one and we can report α as an objective measure of the quality of an anchor model. Note that $\alpha = \alpha(\boldsymbol{\gamma}_0)$ depends on the true, unknown, component-specific parameters $\boldsymbol{\gamma}_0$. In practice, the latter will be replaced by an estimate $\hat{\boldsymbol{\gamma}}_0$ (typically a maximum a posteriori or MAP estimate under the exchangeable model), and we would report the value $\hat{\alpha} = \hat{\alpha}(\hat{\boldsymbol{\gamma}}_0)$.

A related measure of how effectively the anchor points can resolve, asymptotically, the labeling ambiguity is given by the entropy of $P_{\boldsymbol{x}}$, $En(P_{\boldsymbol{x}}, \boldsymbol{\gamma}_0) = -\sum_{q=1}^{k!} p_q \log(p_q)$, where we define $\log(0) = 0$ in the expression. Lower entropy values are preferred. As in the calculation of $\widehat{\alpha}$ above, we can obtain a plug-in estimate of the entropy by substituting an estimate $\widehat{\boldsymbol{\gamma}}_0$ of $\boldsymbol{\gamma}_0$ into the expression. In Section 4, we also demonstrate a modeling strategy in which we select anchor points to minimize the entropy.

The following proposition gives two interesting results in which it can be shown analytically that certain choices of anchor points minimize $En(P_{\boldsymbol{x}}, \boldsymbol{\gamma}_0)$ in a univariate mixture in which $\boldsymbol{\gamma}_j = (\theta_j, \sigma_j^2)$, the mean and variance of the jth Gaussian component.

Proposition 4. *Suppose that* $k = 2$ *and that* $m_j = m$ *observations (with* $1 \leq m \leq n/2$*) are to be anchored to component* j, $j = 1, 2$*. The following results hold:*

1. *(Location mixture.) If* $\sigma_1^2 = \sigma_2^2 = \sigma^2$ *and* $\theta_1 < \theta_2$*, then the optimal anchoring sets* $\boldsymbol{x}_1 = (y_{(1)}, \ldots, y_{(m)})$, $\boldsymbol{x}_2 = (y_{(n-m+1)}, y_{(n)})$*, where* $y_{(l)}$ *denotes the* l*th order statistic.*
2. *(Scale mixture.) If* $\theta_1 = \theta_2 = \theta$ *and* $\sigma_1^2 < \sigma_2^2$*, then the optimal anchoring sets* $\boldsymbol{x}_1$ *equal to the points that minimize* $\sum_{i=1}^{m}(y_i - \theta)^2$ *and* $\boldsymbol{x}_2$ *equal to the points that maximize* $\sum_{i=1}^{m}(y_i - \theta)^2$*.*

Proposition 4 is proved in Appendix A and provides some intuition regarding which anchor models most effectively alleviate the labeling ambiguity: the minimum-entropy anchor model arises from choosing points that are as dissimilar as possible in location and/or in scale.

3. Model specification

We now address two fundamental issues: *how many* and *which* points to anchor.

3.1. Choosing the number of anchor points

Strengthening assumption C.1 by requiring equal mixture proportions (and, as a consequence, equal probabilities on all allocation vectors), the following proposition states that the marginal likelihood of an anchor model can always be increased by introducing additional anchor points. The proposition is proved in Appendix A.

Proposition 5. *Assume that* $\eta_j = 1/k$, $j = 1, \ldots, k$. *Let* $A_*^1, \ldots, A_*^n$ *be a sequence of anchor models where* A_*^m *has the highest marginal likelihood among all anchor models with* m *anchor points. The marginal likelihoods of the models satisfy* $m_{A_*^1}(\boldsymbol{y}) \leq \ldots \leq m_{A_*^n}(\boldsymbol{y})$.

This result indicates that based on goodness-of-fit alone, it is best to specify a larger number of optimally-anchored points. However, increasing the number of anchor points strengthens the degree of prior information built into the model and may increase bias in finite samples. Intuition suggests that limiting the number of anchor points might be desirable to ensure satisfactory out-of-sample predictive performance. To assess the trade-off between goodness of fit and out-of-sample predictive performance, we conducted two small simulation studies. The details of the studies and their results are presented in Appendix B. In simulation 1, we drew data from a two-component location mixture and fit the optimal anchor model with varied values of m. In simulation 2, we drew data from several univariate location-scale mixtures and fit, for varied values of m, anchor models whose anchor points are chosen using the EM strategy to be described in Section 3.2.

In agreement with our intuition, the simulation findings show that, in cases of mixture components that are not well-separated, the out-of-sample predictive ability of the model suffers when too many anchor points are chosen. It is difficult to select a large number anchor points that accurately represent the true densities and they introduce bias in parameter estimation. Further, we found that little benefit accrues from anchoring more than one or two points to each component if the components are well-separated. These results support the recommendation to anchor either one or two points to each component. One anchor point per component will provide unique labeling for all components, as given in Proposition 1. If instead each component has two anchor points, improper priors may be used for the component-specific features of the Gaussian mixture components, a specification that is impossible under the exchangeable model. The next section proposes a method for selecting which points to anchor.

3.2. Anchored EM algorithm for selecting anchor points

We propose selecting anchor points by formulating the optimal anchor model as a solution to a modified Bayesian expectation-maximization (EM) algorithm for maximum a posteriori estimation. The method proceeds by computing a lower

bound on the log posterior density of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ and iteratively updating the parameter values and an anchored posterior distribution on the latent allocations to maximize this lower bound. Intuitively, a good anchor model should concentrate its posterior mass in the vicinity of one of the modal regions of the exchangeable model. Thus, we select as the optimal anchor points those that produce the best approximation (as measured by the lower bound) to the exchangeable posterior density $p_E(\boldsymbol{\gamma}, \boldsymbol{\eta}|\boldsymbol{y})$ near one of the symmetric local modes.

The method draws on the formulation of Neal and Hinton [1998] of the EM algorithm, which makes use of the following lower bound on the log posterior density of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$:

$$\log\left(p_E(\boldsymbol{\gamma}, \boldsymbol{\eta}|\boldsymbol{y})\right) \geq \sum_{\boldsymbol{s}\in\mathcal{S}} q(\boldsymbol{s}) \log\left(\frac{p(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\eta}|\boldsymbol{y})}{q(\boldsymbol{s})}\right). \tag{12}$$

This bound holds for any distribution q on the latent allocations by Jensen's inequality. The expression on the right-hand side of (12) is a function of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ and q and will be denoted by $F(\boldsymbol{\gamma}, \boldsymbol{\eta}, q)$. Neal and Hinton [1998] show that the EM algorithm may be seen as an iterative maximization of the lower bound, F, with respect to $\boldsymbol{\gamma}$ (M step) and q (E step). At iteration t, conditional on the current parameter value $(\boldsymbol{\gamma}^{t-1}, \boldsymbol{\eta}^{t-1})$, the distribution q_*^t that maximizes $F(\boldsymbol{\gamma}^{t-1}, \boldsymbol{\eta}^{t-1}, \cdot)$ is the posterior distribution on the latent variables. For a Gaussian mixture, q_* has the form $q_*(\boldsymbol{S} = (s_1, \ldots, s_n)|\boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\eta}) = \prod_{i=1}^n \prod_{j=1}^k r_{ij}^{I(s_i=j)}$, where

$$r_{ij} = \frac{\eta_j \phi_p(y_i; \boldsymbol{\gamma}_j)}{\sum_{l=1}^k \eta_l \phi_p(y_i; \boldsymbol{\gamma}_l)}, \quad j = 1, \ldots, k, \quad i = 1, \ldots, n. \tag{13}$$

When q^t is set equal to $q_*^t = q_*(\cdot|\boldsymbol{y}, \boldsymbol{\gamma}^{t-1}, \boldsymbol{\eta}^{t-1})$ in the E step of the algorithm, the inequality in (12) is an equality for $(\boldsymbol{\gamma}, \boldsymbol{\eta}) = (\boldsymbol{\gamma}^{t-1}, \boldsymbol{\eta}^{t-1})$ and the lower bound is equal to the log posterior density at that point. Further, Neal and Hinton [1998] state that the value of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ that maximizes $F(\cdot, q_*^t)$ also maximizes the log posterior density.

Ganchev, Taskar and Gama [2008] have modified this EM formulation for settings where q_* cannot arise as the distribution of $\boldsymbol{S}$ because the model imposes certain restrictions on the latent variables. Because the lower bound on $\log(p_E(\boldsymbol{\gamma}, \boldsymbol{\eta}|\boldsymbol{y}))$ holds for any valid probability distribution q, the E-step may be modified so that q^t is chosen to maximize $F(\boldsymbol{\gamma}^{t-1}, \boldsymbol{\eta}^{t-1}, \cdot)$, subject to the problem-specific constraints. It is straightforward to verify that the lower bound on $\log(p_E(\boldsymbol{\gamma}, \boldsymbol{\eta}|\boldsymbol{y}))$ satisfies

$$F(\boldsymbol{\gamma}, \boldsymbol{\eta}, q) = \log(p_E(\boldsymbol{\gamma}, \boldsymbol{\eta}|\boldsymbol{y})) - D_{KL}(q||q_*), \tag{14}$$

where $D_{KL}(q||q_*) = \sum_{\boldsymbol{S}} q(\boldsymbol{s}) \log\left(q(\boldsymbol{s})/q_*(\boldsymbol{s})\right)$ is the Kullback-Leibler (KL) divergence of q_* from q, and q_* is the optimal posterior distribution given in (13). For a given $(\boldsymbol{\gamma}, \boldsymbol{\eta})$, the lower bound F will be largest when q is as close as possible to q_*, in terms of KL divergence.

An anchor model imposes constraints on the distribution of $\boldsymbol{S}$ and fits neatly into this framework. In fact, Neal and Hinton [1998] and Lücke [2016] have suggested using such EM modifications in closely related clustering problems based on Gaussian mixtures. For an anchor model, the posterior distribution of $\boldsymbol{S}$ satisfies $q^t(\boldsymbol{S}|\boldsymbol{y}, \boldsymbol{\gamma}^{t-1}, \boldsymbol{\eta}^{t-1}) = \prod_{i=1}^n q^t(S_i = s_i|\boldsymbol{y}, \boldsymbol{\gamma}^{t-1}, \boldsymbol{\eta}^{t-1})$, where q^t is constrained to satisfy

$$q^t(S_i = j|\boldsymbol{y}, \boldsymbol{\gamma}^{t-1}, \boldsymbol{\eta}^{t-1}) = \begin{cases} \widetilde{r}_{ij}^t, & i \notin A, \\ 1, & i \in A_j, \\ 0, & i \in A_{j'}, \quad j' \neq j. \end{cases} \tag{15}$$

Here, the sets A_j have cardinality $|A_j| = m_j$ for $j = 1 \dots k_0$, and the $\widetilde{r}_{ij}^t$ are probabilities such that $\sum_{j=1}^k \widetilde{r}_{ij}^t = 1$ for all $i \notin A$. Subject to these requirements, the form of the optimal anchor model corresponding to a constrained distribution q that minimizes the KL divergence appearing in (14) is described in the following proposition, which is proved in Appendix A.

Proposition 6. *Let q be the posterior distribution of the allocations under an anchor model, subject to the restrictions in (15). The KL divergence of q_* from q, evaluated at a fixed value of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$, is minimized when the sets A_j are chosen to maximize $\sum_{j=1}^{k_0} \sum_{i \in A_j} r_{ij}$ and when $\widetilde{r}_{ij} = r_{ij}$ for all $i \notin A$.*

The modified EM algorithm will, in the E step, hold $(\boldsymbol{\gamma}^{t-1}, \boldsymbol{\eta}^{t-1})$ constant and update q^t to correspond to a valid anchor model with the optimal anchor points identified in Proposition 6. This amounts to including i in A_j if r_{ij} is among the m_j largest allocation probabilities to component j, except in the case when i satisfies this condition for some other $j' \neq j$. Details on selecting the anchor points is this case are given in Appendix C. (In our experience, this case rarely occurs in real data applications if m is small relative to the sample size.) In the subsequent M step, $(\boldsymbol{\gamma}^t, \boldsymbol{\eta}^t)$ is updated to maximize the lower bound, holding q^t fixed at its current value. As in the standard EM algorithm, the M step can be accomplished by maximizing $E(\log(p(\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{s}, \boldsymbol{y}))$, where the expectation is taken with respect to q^t. This maximization is computationally tractable because $E(\log(p(\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{s}, \boldsymbol{y}))$ can be expressed as a summation of $k \times n$ addenda, by an argument analogous to the one used in the proof of Proposition 6. For the models considered in the examples of Sections 4 and 5 the maximizer can be derived in closed form. The steps of this "anchored EM Algorithm" are detailed in Appendix C.

As discussed by Ganchev, Taskar and Gama [2008] for related approaches, the anchored EM algorithm maximizes a penalized version of the log posterior density, where the penalty is given by the KL divergence of the distribution q_* corresponding to the exchangeable model from the distribution q corresponding to the anchored model. Each EM iteration updates both the parameters and the distribution q in order to increase the lower bound on the log posterior density, yielding an optimal approximation to a local mode of the exchangeable posterior distribution.

3.3. Other strategies for model specification

We propose the anchored EM method as one potential default method for automatically selecting anchor points. The anchor modeling framework admits alternative strategies based on various modeling considerations and, if available, prior information. For example, the large-sample properties outlined in Section 2.5 motivate choosing anchor points that minimize the entropy of the asymptotic distribution on the relabelings, a strategy which we demonstrate in Section 4. In a recent case study, Kunkel and Peruggia [2019] develop a method that uses diagnostic information from a base model to choose anchor points for a mixture of regressions model. Existing classification and clustering tools can also be used to identify anchor points as representative points from the k components. Particularly applicable are a type of semi-supervised algorithms that propose strategies for introducing artificial labels to points in classification problems. For example, the Yarowsky "self-training" algorithm [Yarowsky, 1995, Abney, 2004] and variants thereof use allocation probabilities to assign artificial labels iteratively until a stopping criterion is reached.

4. Univariate examples

In this section we present several examples of the anchor modeling approach applied to univariate data sets.

4.1. Sampling

In the following examples, we specify a conditionally conjugate prior and use a Gibbs sampler to obtain samples from the posterior distribution. Sampling is a well-known challenge in the exchangeable mixture model due to the $k!$ symmetric regions of the posterior [see, e.g., Celeux, Hurn and Robert, 2000, Jasra, Holmes and Stephens, 2005, Geweke, 2007]. A sampler which fully explores the posterior parameter space is, in fact, one in which perfect label switching is observed; each symmetric region must be visited with equal frequency and ergodic averages of the component-specific parameters should be equal. Label switching, then, is a requisite for convergence of the sampling algorithm, but it is difficult to ensure this behavior, particularly in a model with well-separated components. Celeux, Hurn and Robert [2000] recommend replacing the Gibbs sampler with a simulated tempering scheme that promotes swift movement across modal regions of the parameter space. Other strategies leverage the symmetry of the posterior to improve mixing: Frühwirth-Schnatter [2001] presents a sampler that, for each sampled value of the vector of component-specific parameters, proposes a random permutation of the component labels that is accepted with probability one. A similar approach of Geweke [2007] augments each sampled value with every possible relabeling of the value.

We adopted the random permutation strategy recommended by Frühwirth-Schnatter [2001]. The resulting algorithm is a modification of a standard Gibbs

algorithm for sampling draws from the posterior distribution of $(\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{s})$. A standard Gibbs sampler would alternate between updates of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ sampled from the conditional distribution of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ given $(\boldsymbol{s}, \boldsymbol{y})$ and updates of $\boldsymbol{s}$ sampled from the conditional distribution of $\boldsymbol{s}$ given $(\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{y})$. The modified sampler includes an additional step after the update of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ conditional on $(\boldsymbol{s}, \boldsymbol{y})$. Precisely, at iteration t, a value $(\tilde{\boldsymbol{\gamma}}^t, \tilde{\boldsymbol{\eta}}^t)$ is generated conditional on $(\boldsymbol{s}^{t-1}, \boldsymbol{y})$. Then, a random permutation ρ_q of the component labels is selected according to the probabilities described in the following paragraph and $(\boldsymbol{\gamma}^t, \boldsymbol{\eta}^t)$ is set equal to $\rho_q(\tilde{\boldsymbol{\gamma}}^t, \tilde{\boldsymbol{\eta}}^t)$ before proceeding to the update of $\boldsymbol{s}^{t-1}$. The intermediate sample $(\tilde{\boldsymbol{\gamma}}^t, \tilde{\boldsymbol{\eta}}^t)$ is discarded and only $(\boldsymbol{\gamma}^t, \boldsymbol{\eta}^t)$ is retained as part of the chain.

Under the exchangeable model, each random permutation has equal probability of being sampled. Under the anchor model, the asymmetry of the posterior density requires the permutations to be accepted with probabilities proportional to the values of the posterior density $p_A(\rho_q(\tilde{\boldsymbol{\gamma}}^t, \tilde{\boldsymbol{\eta}}^t)|\boldsymbol{y})$ at the various relabelings. Simple algebraic manipulations show that, at iteration t, permutation q is selected with probability w_{tq}, for $q = 1, \ldots, k!$, where w_{tq} is equal to

$$w_{tq} = \frac{\prod_{j=1}^{k} \prod_{i \in A_j} \phi(y_i; \tilde{\gamma}^t_{\rho_q(j)})}{\sum_{h=1}^{k!} \prod_{j=1}^{k} \prod_{i \in A_j} \phi(y_i; \tilde{\gamma}^t_{\rho_h(j)})}, \quad q = 1, \ldots, k!. \tag{16}$$

We prove in Appendix D that the accepted permuted draw $(\boldsymbol{\gamma}^t, \boldsymbol{\eta}^t)$ is in fact a draw from the distribution of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ conditional on $(\boldsymbol{s}^{t-1}, \boldsymbol{y})$. Therefore this procedure results in an algorithm whose invariant distribution is the same as that of the standard algorithm.

In each of the following examples, we used a Gibbs sampler with the random permutation step. We ran 50 chains initialized at different random starting points, discarding the first 1,000 iterations of each chain as burn-in and thinning the chains every 100-th draw to obtain a total of 15,000 retained posterior samples. We estimated standard errors of the computed estimates using the overlapping batch means estimator implemented in the R package `mcmcse` [Flegal et al., 2020].

4.2. Galaxies

Our first example demonstrates the anchoring method using the galaxies data set from Roeder [1990], now a benchmarking staple of the mixture literature. Each of the 82 measurements gives the velocity of a galaxy sampled from the Corona Borealis region of space. Previous analyses have indicated that between three and seven Gaussian components are appropriate for these data; we chose to set $k = 6$, the value given the highest posterior probability in the well-known analysis of Richardson and Green [1997] and used in subsequent analyses by Stephens [2000] and Rodriguez and Walker [2014]. We specify the following priors and hyperparameters following the recommendations of Richardson and Green [1997]: the parameters $\boldsymbol{\gamma}_j = (\theta_j, \sigma_j^2)$ are a priori mutually independent.

The θ_j are normal with mean μ and variance $1/\kappa$, the precisions, σ_j^{-2}, have a Gamma(a_0, b_0) distribution, where the Gamma distribution is parameterized to have a mean of b_0/a_0, and $\boldsymbol{\eta}$ has a Dirichlet$(\mathbf{1}_k)$ distribution, where $\mathbf{1}_k$ is a vector of k ones. The hyperparameters were set as $\mu = 21.7255$, the midpoint of the data, $\sqrt{\kappa} = 1/52$, and $a_0 = 2$. We specified a Gamma$(0.2, 0.016)$ distribution for b_0.

We used two methods to specify anchor models with one anchor point per component. The first model used anchor points selected by the anchored EM method described in Section 3.2. We ran the anchored EM algorithm from 50 random starting positions and selected the anchor points that corresponded to the highest value of the lower bound in (14).

The second set of anchor points were selected to minimize the estimated entropy of the asymptotic probability distribution on the class relabelings as given by (30). To do this, we treated $En(P_{\boldsymbol{x}}, \widehat{\boldsymbol{\gamma}}_0)$ as a continuous function of $\boldsymbol{x}$, where $\widehat{\boldsymbol{\gamma}}_0$ is a maximum a posteriori estimate of $\boldsymbol{\gamma}_0$ under the exchangeable model, and found a value $\boldsymbol{x}^*$ that minimizes this continuous function. The value $\boldsymbol{x}^*$ was identified using the `optim` function in R with the BFGS method and the tolerance parameter set equal to 10^{-10}. We then chose the anchor points to be the observations closest to $\boldsymbol{x}^*$.

We used a Gibbs sampler to fit the two anchor models and the exchangeable model using the random permutation strategy described in Section 4.1. We applied two popular relabeling algorithms to the samples from the exchangeable model: the KL method [Stephens, 2000] and data-based (DB) relabeling [Rodriguez and Walker, 2014]. Both methods are implemented in the R `label.switching` package [Papastamoulis, 2016]. Lastly, we applied an ordering constraint that required $\theta_1 < \ldots < \theta_6$ with probability 1.

The left panel of Figure 1 shows the locations of the minimum-entropy anchor points and the anchored EM points. The EM anchor points are close to the minimum-entropy anchor points for most components except for component 5, where the minimum-entropy point falls in a cluster of high velocity near that of component 6. Using the maximum a posteriori estimate of $\boldsymbol{\gamma}_0$, we calculated the estimated coefficients of quasi-consistency of Definition 1, $\widehat{\alpha}$, to be greater than 0.9999 for both of the anchor models, indicating a high degree of asymptotic identifiability of the component labels.

The middle and right panels of Figure 1 display Monte Carlo estimates of the scaled component densities. The scaled component density for j is $\eta_j \phi_p(\cdot, \gamma_j)$; as noted by [Rodriguez and Walker, 2014], accurate posterior inference on this quantity is tantamount to inference on the posterior classification probability to component j. For components 1, 3, 4, and 6, all methods produce unimodal densities that are similar in location. The estimated density of component 2 has an irregular shape for all methods: the density is skewed under the anchored models and KL relabeling and it is bimodal under the ordering constraint. The location of component 5 differs between the two anchor models, with the minimum-entropy model placing the component in the right-hand cluster in the data. The relabeling methods estimate a more symmetric density with a mode shifted towards the left.

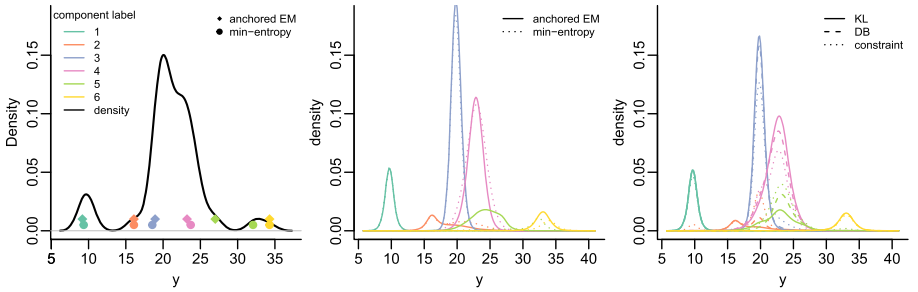

FIG 1. *Left: kernel density estimate of the galaxies data with the EM and minimum-entropy anchor points. Middle: estimated scaled component densities under the two anchor models. Right: estimated scaled component densities under the two relabeling methods.*

TABLE 1
Posterior means (standard errors) of the component-specific parameters for the galaxies data.

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	time(s)
anchored EM	9.713 (0.0022)	16.798 (0.0129)	19.845 (0.0018)	22.803 (0.0038)	25.408 (0.0130)	33.018 (0.0058)	11.91
min-entropy	9.713 (0.0022)	17.289 (0.0214)	19.815 (0.0021)	22.927 (0.0036)	31.512 (0.0181)	33.672 (0.0148)	58.62
KL	9.711 (0.0024)	18.522 (0.0887)	19.847 (0.0169)	22.625 (0.0057)	23.419 (0.0802)	32.795 (0.0223)	183.46
DB	9.711 (0.0024)	18.177 (0.0853)	20.049 (0.0205)	22.222 (0.0130)	23.963 (0.0817)	32.797 (0.0195)	9.04
constraint	8.099 (0.0529)	16.437 (0.0264)	19.878 (0.0118)	22.248 (0.0119)	25.631 (0.0288)	34.625 (0.0553)	N/A
	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	
anchored EM	0.685 (0.0018)	1.104 (0.0058)	0.756 (0.0013)	1.110 (0.0024)	1.289 (0.0050)	1.097 (0.0038)	
min-entropy	0.682 (0.0018)	1.253 (0.0083)	0.755 (0.0017)	1.490 (0.0031)	1.214 (0.0076)	1.134 (0.0069)	
KL	0.708 (0.0037)	1.131 (0.0048)	0.794 (0.0031)	1.610 (0.0020)	1.315 (0.0045)	1.171 (0.0031)	
DB	0.709 (0.0020)	1.080 (0.0055)	0.825 (0.0030)	1.741 (0.0065)	1.197 (0.0048)	1.178 (0.0051)	
constraint	0.749 (0.0031)	0.987 (0.0047)	1.078 (0.0058)	1.369 (0.0056)	1.359 (0.0048)	1.185 (0.0056)	
	η_1	η_2	η_3	η_4	η_5	η_6	
anchored EM	0.090 (0.0002)	0.055 (0.0005)	0.374 (0.0007)	0.330 (0.0009)	0.105 (0.0008)	0.046 (0.0002)	
min-entropy	0.091 (0.0002)	0.078 (0.0011)	0.348 (0.0007)	0.415 (0.0011)	0.038 (0.0003)	0.029 (0.0002)	
KL	0.090 (0.0002)	0.041 (0.0003)	0.323 (0.0009)	0.404 (0.0011)	0.097 (0.0007)	0.045 (0.0002)	
DB	0.090 (0.0002)	0.052 (0.0005)	0.312 (0.0009)	0.376 (0.0013)	0.124 (0.0010)	0.046 (0.0002)	
constraint	0.082 (0.0003)	0.103 (0.0010)	0.294 (0.0013)	0.301 (0.0013)	0.178 (0.0014)	0.042 (0.0002)	

The posterior means of the component-specific parameters are given in Table 1. The anchored models and relabeling methods produce similar estimates for θ_1, θ_3, and θ_4. For component 4, however, the estimated standard deviations lie somewhat between those of the anchor models and relabeling methods, with the estimated values of σ_4 higher under the relabeling methods than the EM anchor model. Both anchor models estimate a greater degree of separation among the means of components 2, 3, 4, and 5 than either of the relabeling methods, which reflects the influence of the separation among the anchor points. Lastly, the parameter estimates for components 5 and 6 under the minimum-entropy model reflect the proximity of the anchor points assigned to these components: both components describe the distinct modal region between 30 and 35, whereas component 5 overlaps with components in the middle region under the other methods. The estimated $\boldsymbol{\eta}$ values are fairly consistent across methods, except for a very low estimate of η_5 under the minimum-entropy model and a tendency for the ordering constraint to produce estimates closer to $1/k$ than the other methods.

TABLE 2
Model parameters used in the simulations.

	Model 1	Model 2	Model 3
$\boldsymbol{\theta}$	$(0, 0)$	$(-3, -1, 1, 3)$	$(19, 19, 23, 29, 33)$
$\boldsymbol{\sigma}$	$(1.5, 0.5)$	$(1, 1, 1, 1)$	$(2.236, 1, 1, 0.707, 1.414)$
$\boldsymbol{\eta}$	$(0.35, 0.65)$	$(0.25, 0.25, 0.25, 0.25)$	$(0.2, 0.2, 0.25, 0.2, 0.15)$

Table 1 also reports the CPU times for implementing each method using an Intel Core i7-8700 processor. For the anchor models, these times are the preprocessing times required to select anchor points. For the relabeling methods, the reported times are for post-processing of the posterior samples and are thus dependent on the number of posterior samples retained for the analysis. None of the reported times includes the time required to sample from the posterior distributions.

4.3. Simulated data

We next fit anchor models to data generated from three univariate Gaussian mixtures with the parameters given in Table 2.

The density functions of models 1, 2, and 3 are shown in the left panels of Figures 2, 3, and 4, respectively. Model 1 is a scale mixture whose two components have identical locations. Models 2 and 3 have been studied previously by [Papastamoulis and Iliopoulos, 2010] (model 3) and [Rodriguez and Walker, 2014] (models 2 and 3) to assess performance of relabeling algorithms. We drew samples of size $n = 200$ from Models 1 and 2 and $n = 600$ from model 3. Following the approach of Rodriguez and Walker [2014], we used "perfect samples," evenly-spaced quantiles of the mixture distribution as evaluated using the R package nor1mix [Maechler, 2019], to eliminate sampling variability.

For each model, we fit three anchor models using the anchored EM method and the minimum-entropy method described in Section 4.2. The third anchor model is an "oracle" model, in which the anchor points are selected to be the observations closest to predetermined quantiles of each true component density. Such information about the true densities would not be available in practice, but we present these results as an illustration of the model's performance when anchor points represent known features of the true mixture components. For the anchor models, we report the value of α evaluated at the true $\boldsymbol{\gamma}_0$. Finally, we fit the exchangeable model and applied the KL and DB relabeling methods and prior ordering constraints.

The following priors and hyperparameters were specified for each example, adhering to the recommendations of Richardson and Green [1997]: θ_j has a Normal distribution with mean μ and variance $1/\kappa$, σ_j^{-2} has a Gamma(a_0, b_0) distribution, $j = 1, \ldots, k$, and $\boldsymbol{\eta}$ has a Dirichlet$(\mathbf{1}_k)$ distribution. Letting $R = y_{(n)} - y_{(1)}$, where $y_{(h)}$ denotes the hth order statistic, the hyperparameters were set as: $\mu = \bar{y}$, $\kappa = 1/R^2$, $a_0 = 2$, $b_0 \sim$ Gamma(g_0, h_0), $g_0 = 0.2$, and $h_0 = 10/R^2$.

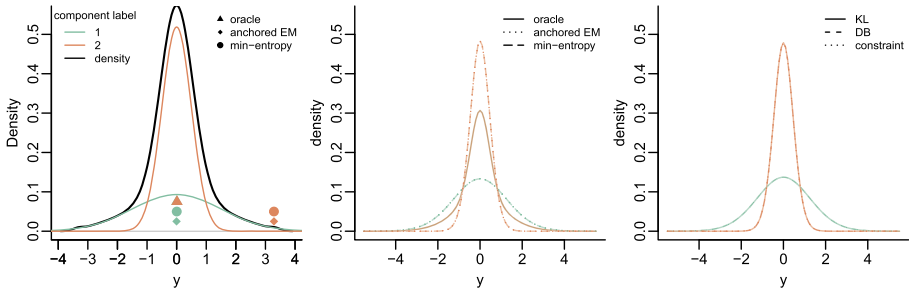

FIG 2. *Left panel: mixture density, anchor points, and true scaled component densities for model 1. Middle panel: estimated scaled component densities for the anchor models. Right panel: estimated scaled component densities for the relabeling methods.*

TABLE 3
Posterior means (standard errors) of the component-specific parameters for model 1.

	θ_1	θ_2	error	time (s)
true	0	0	N/A	
oracle anchors	0.0010 (0.0010)	−0.0010 (0.0010)	0.0020	N/A
anchored EM	−0.0027 (0.0014)	−0.0009 (0.0005)	0.0036	1.03
min-entropy	−0.0003 (0.0014)	−0.0006 (0.0005)	0.0009	0.50
KL	0.0007 (0.0013)	0.0005 (0.0005)	0.0012	584.25
DB	0.0010 (0.0013)	0.0002 (0.0006)	0.0012	574.00
constraint	0.0011 (0.0013)	0.0001 (0.0006)	0.0012	N/A
	σ_1	σ_2	error	
true	1.5	0.5	N/A	
oracle anchors	0.867 (0.0035)	0.866 (0.0035)	0.9988	
anchored EM	1.320 (0.0013)	0.467 (0.0006)	0.2125	
min-entropy	1.320 (0.0013)	0.468 (0.0006)	0.2122	
KL	1.309 (0.0013)	0.464 (0.0006)	0.2272	
DB	1.309 (0.0013)	0.464 (0.0006)	0.2272	
constraint	1.309 (0.0013)	0.464 (0.0006)	0.2272	
	η_1	η_2	error	
true	0.35	0.65	N/A	
oracle anchors	0.501 (0.0009)	0.499 (0.0009)	0.302	
anchored EM	0.430 (0.0009)	0.570 (0.0009)	0.1594	
min-entropy	0.428 (0.0009)	0.572 (0.0009)	0.1568	
KL	0.438 (0.0009)	0.562 (0.0009)	0.1766	
DB	0.439 (0.0009)	0.561 (0.0009)	0.1772	
constraint	0.439 (0.0009)	0.561 (0.0009)	0.1771	

The following results use $m_j = 1$, $j = 1, \ldots, k$ (one anchor point per component) for the three anchor models. The oracle anchor points are the observations closest to the median of each component. Appendix E presents results from using two anchor points per component.

4.3.1. Model 1

Model 1 is a scale mixture whose two components both have means of zero. The oracle anchors are nearly identical points at the median of the distribution. The anchored EM and minimum-entropy methods select the same anchor points in this example: the maximum observation is anchored to component 1 while the observation closest to the sample mean is anchored to component 2. These two anchor models have α values of 1.000, while the oracle anchor model has an α value of 0.500. Figure 2 shows the anchor points and the estimated scaled component densities, and Table 3 gives the estimated posterior means of the component-specific parameters and the total absolute errors, calculated as $\sum_{j=1}^{k} |a_j - \widehat{a}_j|$ for a parameter a_j having posterior mean equal to $\widehat{a}_j$. For this example, in the model with prior ordering constraints, the component standard deviations were assumed to satisfy the condition $\sigma_1 > \sigma_2$. The estimates are very similar across the methods for these data, with the exception of the oracle anchor model. The oracle model performs poorly, producing identical parameter estimates for both components, because the anchor points provide no information about the scale difference between the components. None of the methods accurately capture the difference between η_1 and η_2, with the anchored EM and minimum-entropy models coming the closest of the six methods.

4.3.2. Model 2

The component densities of model 2 overlap substantially, with equal variances and evenly-spaced means. In the model with prior ordering constraints, the component means were assumed to satisfy the condition $\theta_1 < \ldots < \theta_4$. The true model and estimated anchor points are shown in the left panel of Figure 3, with α values of 0.947, 0.972, and 0.996 for the oracle, anchored EM, and minimum-entropy models, respectively. The EM and minimum-entropy method both select the minimum and maximum observations to be anchored to components 1 and 4, respectively. For components 2 and 3, the anchored EM points fall near the true component means of -1 and 1, while the minimum-entropy points fall closer to -2 and 2, respectively. The effect of these differences is seen in the middle panel of Figure 3. The estimated scaled densities for components 2 and 3 are approximately symmetric under the EM anchor model. Under the minimum entropy model, however, these densities are skewed, with excess mass near the anchor points, and the posterior means of θ_2 and θ_3, shown in Table 4, are poor estimates of the true component means. Both of the relabeling methods produce estimated component means that are biased towards zero, most substantially for components 2 and 3, and severely overestimate η_2 while underestimating η_3. The oracle and the EM anchor models produce relatively accurate estimates of all model parameters.

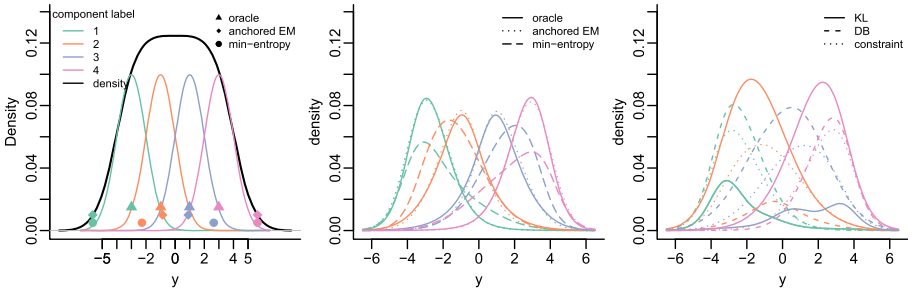

FIG 3. *Left panel: mixture density, anchor points, and true scaled component densities for model 2. Middle panel: estimated scaled component densities for the anchor models. Right panel: estimated scaled component densities for the relabeling methods.*

TABLE 4
Posterior means (standard errors) of the component-specific parameters for model 2.

	θ_1	θ_2	θ_3	θ_4	error	time (s)
true	−3	−1	1	3		
oracle anchors	−2.819 (0.0049)	−0.960 (0.0072)	0.956 (0.0072)	2.818 (0.0049)	0.448	N/A
anchored EM	−2.983 (0.0069)	−0.960 (0.0073)	0.997 (0.0073)	2.997 (0.0068)	0.063	24.87
min-entropy	−2.967 (0.0116)	−1.653 (0.0095)	1.965 (0.0088)	2.864 (0.0134)	1.786	10.37
KL	−1.773 (0.0344)	−1.529 (0.0084)	1.194 (0.0339)	2.074 (0.0069)	2.876	7031.77
DB	−2.556 (0.0142)	−0.250 (0.0384)	0.260 (0.0143)	2.513 (0.0196)	2.421	169.71
constraint	−3.483 (0.0231)	−1.038 (0.0121)	1.034 (0.0124)	3.453 (0.0231)	1.008	N/A
	σ_1	σ_2	σ_3	σ_4	error	
true	1	1	1	1		
oracle anchors	1.024 (0.0026)	1.067 (0.0041)	1.063 (0.0039)	1.026 (0.0025)	0.1805	
anchored EM	1.074 (0.0025)	0.984 (0.0031)	0.984 (0.0032)	1.069 (0.0025)	0.1752	
min-entropy	1.121 (0.0040)	1.089 (0.0036)	1.042 (0.0034)	1.168 (0.0043)	0.4192	
KL	0.749 (0.0067)	1.449 (0.0023)	0.715 (0.0069)	1.304 (0.0031)	1.2890	
DB	1.077 (0.0033)	0.698 (0.0028)	1.455 (0.0051)	0.987 (0.0032)	0.8471	
constraint	0.984 (0.0038)	1.124 (0.0047)	1.122 (0.0047)	0.988 (0.0038)	0.2733	
	η_1	η_2	η_3	η_4	error	
true	0.25	0.25	0.25	0.25		
oracle anchors	0.247 (0.0011)	0.252 (0.0013)	0.251 (0.0013)	0.250 (0.0010)	0.0059	
anchored EM	0.255 (0.0012)	0.247 (0.0012)	0.247 (0.0012)	0.251 (0.0012)	0.0124	
min-entropy	0.222 (0.0016)	0.283 (0.0016)	0.263 (0.0014)	0.232 (0.0018)	0.0923	
KL	0.100 (0.0007)	0.439 (0.0016)	0.082 (0.0006)	0.379 (0.0015)	0.6357	
DB	0.275 (0.0015)	0.078 (0.0007)	0.418 (0.0022)	0.230 (0.0014)	0.3844	
constraint	0.224 (0.0017)	0.275 (0.0020)	0.274 (0.0019)	0.226 (0.0017)	0.0992	

4.3.3. Model 3

Model 3 is a five-component mixture in which the components have varying locations, scales, and weights. The true scaled component densities and anchor points are shown in Figure 4 and the posterior parameter estimates are given in Table 5. The α values for the oracle, anchored EM, and minimum-entropy models are 0.500, 1.000, and 0.978, respectively. The anchored EM points are offset from the true component means but still tend to fall in regions to which their component's true density assigns sizable mass. This model produces accurate estimates of $\theta_2 - \theta_5$, but incorrectly estimates a substantial location shift between components 1 and 2. Its estimates of the component standard devia-

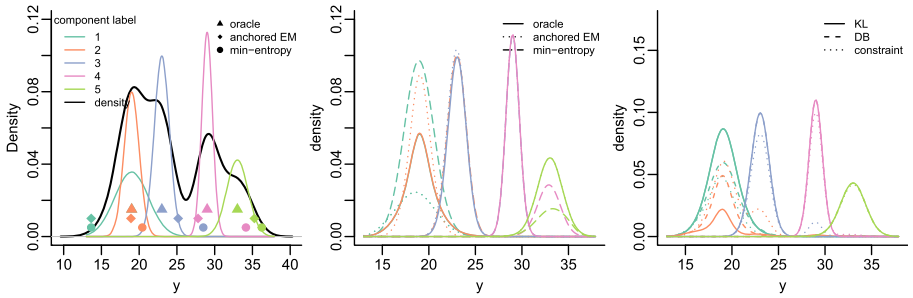

FIG 4. *Left panel: mixture density, anchor points, and true scaled component densities for model 3. Middle panel: estimated scaled component densities for the anchor models. Right panel: estimated scaled component densities for the relabeling methods.*

TABLE 5
Posterior means (standard errors) of the component-specific parameters for model 3.

	θ_1	θ_2	θ_3	θ_4	θ_5	error	time (s)
true	19	19	23	29	33		
oracle anchors	18.955 (0.0049)	18.958 (0.0049)	23.030 (0.0014)	29.006 (0.0006)	32.990 (0.0016)	0.135	N/A
anchored EM	17.273 (0.0290)	19.049 (0.0019)	22.975 (0.0014)	29.006 (0.0006)	32.991 (0.0016)	1.816	16.60
min-entropy	18.948 (0.0014)	23.011 (0.0010)	29.006 (0.0006)	33.031 (0.0056)	33.838 (0.0100)	14.94	10.56
KL	19.080 (0.0028)	20.037 (0.0876)	23.029 (0.0016)	29.006 (0.0007)	32.993 (0.0018)	1.159	1106.45
DB	20.376 (0.0698)	18.732 (0.0417)	23.035 (0.0018)	29.007 (0.0007)	32.995 (0.0018)	1.691	287.54
constraint	17.808 (0.0288)	20.061 (0.0299)	23.847 (0.0360)	29.262 (0.0122)	33.167 (0.0146)	3.529	N/A
	σ_1	σ_2	σ_3	σ_4	σ_5	error	
true	2.236	1	1	0.7071	1.414		
oracle anchors	1.371 (0.0048)	1.380 (0.0048)	0.973 (0.0010)	0.719 (0.0005)	1.364 (0.0012)	1.334	
anchored EM	1.490 (0.0081)	1.073 (0.0042)	1.011 (0.0009)	0.719 (0.0005)	1.363 (0.0012)	0.892	
min-entropy	1.606 (0.0010)	1.011 (0.0007)	0.718 (0.0005)	1.067 (0.0025)	1.134 (0.0037)	1.563	
KL	1.582 (0.0047)	0.976 (0.0051)	0.974 (0.0011)	0.716 (0.0005)	1.358 (0.0013)	0.769	
DB	1.538 (0.0080)	1.020 (0.0053)	0.974 (0.0011)	0.716 (0.0005)	1.358 (0.0013)	0.810	
constraint	1.325 (0.0050)	1.265 (0.0061)	0.934 (0.0022)	0.745 (0.0018)	1.337 (0.0019)	1.356	
	η_1	η_2	η_3	η_4	η_5	error	
true	0.2	0.2	0.25	0.2	0.15		
oracle anchors	0.201 (0.0008)	0.204 (0.0008)	0.245 (0.0003)	0.201 (0.0001)	0.149 (0.0001)	0.0127	
anchored EM	0.132 (0.0021)	0.255 (0.0019)	0.264 (0.0003)	0.201 (0.0001)	0.149 (0.0001)	0.1387	
min-entropy	0.392 (0.0002)	0.255 (0.0002)	0.200 (0.0002)	0.088 (0.0004)	0.064 (0.0004)	0.4954	
KL	0.335 (0.0010)	0.073 (0.0008)	0.246 (0.0005)	0.198 (0.0002)	0.147 (0.0002)	0.2691	
DB	0.257 (0.0015)	0.151 (0.0018)	0.246 (0.0005)	0.198 (0.0002)	0.147 (0.0002)	0.1146	
constraint	0.191 (0.0023)	0.252 (0.0012)	0.228 (0.0010)	0.186 (0.0007)	0.143 (0.0003)	0.1034	

tions are close to their true values, except that of σ_1, which is underestimated as it is by all other methods. The KL and DB relabeling methods accurately estimate the means and standard deviations of all components, and the DB method also produces accurate estimates of $\boldsymbol{\eta}$. Under the prior ordering constraint $\theta_1 < \ldots < \theta_5$, the estimated scaled densities of components 2 and 3 are multimodal and the estimates of $\boldsymbol{\theta}$ are comparatively inaccurate, reflecting the inadequacy of this constraint to describe overlapping components.

The minimum entropy model performs poorly for these data: the anchor points are located at the periphery of plausible regions under their respective component densities, with no points near regions of high density around the mean of component 3. As a result, the mass at this peak is split between components 2 and 3 and the estimated scaled component densities for components 2 and 3 appear bimodal. A similar phenomenon produces bimodality in compo-

nent 4. Table 5 shows that this model produces the least accurate estimates of all of the methods, especially for the component means.

In each of these univariate examples, the anchored EM algorithm has selected anchor points that result in comparatively accurate estimates of the component-specific parameters. In terms of absolute relative error, it outperforms the minimum-entropy anchor model in all cases and tends to have comparable or superior performance to the relabeling methods. Interestingly, the anchored EM model occasionally performs better than the oracle model, due to the anchor points' ability to provide information about both the locations and the relative scales of the components. The estimates produced by the minimum entropy model are less accurate because this method tends to select points in low-density areas where adjacent component begin to overlap. Although these points maximize the model's estimated asymptotic identifiability, their influence introduces bias in finite samples. It is evident that anchor points must fall in areas with non-negligible density of the components to which they belong.

5. A multivariate example: fall detection data

We now apply the anchored modeling framework to a data set called SisFall [Sucerquia, López and Vargas-Bonilla, 2017], one of a growing body of fall data sets that are being used to develop systems that detect falls automatically using wearable devices, cameras, and/or microphones. Experimental data are obtained from volunteer subjects who simulate falls and various activities of daily living (ADLs) and analyzed with the goal of characterizing the distinguishing features of falls compared to ADLs and detecting falls with high accuracy.

Common practices in analyzing these types of data include thresholding [Bourke and Lyons, 2008], in which lower- or upper-thresholds for one variable are set, and a fall is determined to have occurred if the variable exceeds the threshold during a trial. More recent analyses have used supervised classification algorithms on extracted features of the data [Albert et al., 2012, Casilari, Santoyo-Ramón and Cano-García, 2017]. Our approach uses a finite Gaussian mixture model to cluster activities into similar subgroups and to provide a characterization of the features of each group. Analyzing these data in a mixture framework makes it possible to identify groups of experimental activities that share similar features and to describe, with an accompanying appraisal of uncertainty, the typical features of each group. Using this model for classification can provide further insight about which types of ADLs are difficult to distinguish from falls.

The subjects of the SisFall experiments performed 15 types of falls and 19 types of ADLs, repeating 5 trials for most of the activities, while wearing two accelerometers and one gyroscope. We analyzed the data recorded by one of the two accelerometers worn by one subject ("Subject 9", a 24-year-old male) in the SisFall data set. A time series of three-dimensional acceleration vectors (x_t, y_t, z_t) is available for each of the 154 trials. Following common practice in

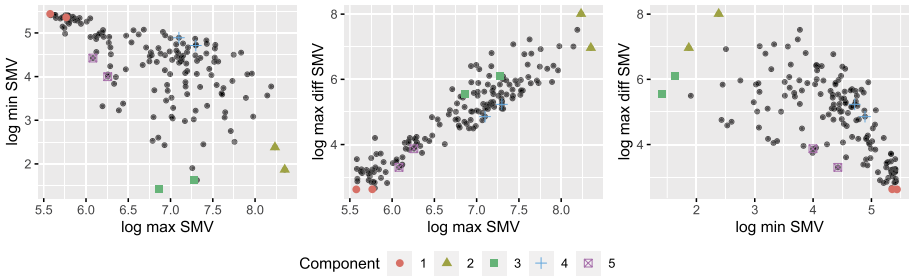

FIG 5. *The data and selected anchor points for the SisFall data example.*

the fall detection literature, we summarized the acceleration at each time point t via the Signal Magnitude Vector (SMV), defined as $SMV_t = \sqrt{x_t^2 + y_t^2 + z_t^2}$. We further summarized the SMV series for each trial using the logarithm of three extracted features arranged in a three-dimensional vector. These features, previously used by Casilari, Santoyo-Ramón and Cano-García [2017] in analyzing several similar fall data sets, are: $\log(\max_t SMV_t)$, $\log(\min_t SMV_t)$, and $\log(\max_t |SMV_t - SMV_{t-1}|)$. Ultimately, the resulting data set contained 154 three-dimensional vectors of extracted log-features.

We fit a multivariate Gaussian mixture model with $k = 5$ components. We selected the number of components based on the Bayesian information criterion (BIC) evaluated at MAP estimates of the exchangeable model parameters. We specified a $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}_j/\kappa)$ prior on $\boldsymbol{\theta}_j$ with $\boldsymbol{\mu} = \bar{\boldsymbol{Y}}$, the sample mean vector of the data, and $\kappa = 0.01$. We specified a Wishart $(\nu, \boldsymbol{A})$ distribution on $\boldsymbol{\Sigma}_j^{-1}$ with $\nu = 6$ degrees of freedom and prior scale $\boldsymbol{A} = 5\boldsymbol{I}_3$, where $\boldsymbol{I}_p$ denotes the $p \times p$ identity matrix. Finally, we specified a Dirichlet$(\mathbf{1}_3)$ prior for $\boldsymbol{\eta}$. We used the anchored EM algorithm to select two anchor points per component. The data and selected anchor points are shown in Figure 5. The coefficient of quasi-consistency was estimated as $\widehat{\alpha} > 0.9999$. Qualitatively, the selected anchor points identify well-separated sites on the periphery of the data cloud, as we would expect in a location problem by generalizing the intuition provided by Proposition 4. The high value of $\widehat{\alpha}$ indicates that we can expect our anchor points to produce high posterior concentration on the true parameter values in large samples. We fit the model using a Gibbs sampler with 200 parallel chains. We thinned the chains after $5,000$ burn-in iterations to obtain $M = 50,000$ samples from the posterior distribution.

Posterior density estimates of $\boldsymbol{\theta}$ are shown in Figure 6. Table 6 lists the average posterior allocation probabilities for selected activities, where each trial's probability of allocation to component j is the relative frequency that $s_i^m = j$, calculated from the Monte Carlo posterior samples of $\boldsymbol{s}^m$, $m = 1, \ldots, M$. The table gives these probabilities averaged over the repeated trials for each activity. The legend of Figure 6 also displays the proportion of falls among the observations classified to each component, if each trial is classified to its most probable component.

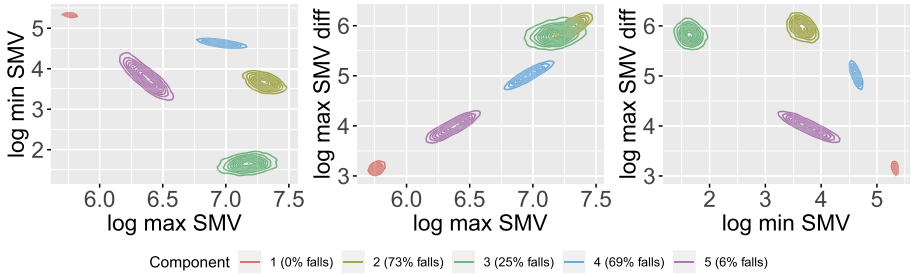

FIG 6. *2D density estimates of the marginal posterior distributions of* $\boldsymbol{\theta}_j$, $j = 1, \ldots, k = 5$, *for the SisFall data set.*

TABLE 6
Posterior allocation probabilities for selected activities in the SisFall data set.

Activity	Component				
	1	2	3	4	5
D03 Jogging slowly	0.000	0.000	1.000	0.000	0.000
D06 Walking upstairs and downstairs quickly	0.000	0.796	0.188	0.000	0.016
D07 Slowly sit in a half height chair, wait a moment, and up slowly	0.981	0.000	0.000	0.006	0.013
D09 Slowly sit in a low height chair, wait a moment, and up slowly	0.245	0.050	0.000	0.575	0.131
D10 Quickly sit in a low height chair, wait a moment, and up quickly	0.001	0.042	0.001	0.148	0.808
D11 Sitting a moment, trying to get up, and collapse into a chair	0.000	0.541	0.003	0.454	0.002
D18 Stumble while walking	0.000	0.957	0.003	0.040	0.000
D19 Gently jump without falling (trying to reach a high object)	0.000	0.308	0.033	0.000	0.658
F02 Fall backward while walking caused by a slip	0.000	0.302	0.001	0.697	0.000
F04 Fall forward while walking caused by a trip	0.000	0.938	0.001	0.061	0.000
F09 Lateral fall when trying to get up	0.000	0.074	0.000	0.926	0.000
F10 Fall forward when trying to sit down	0.000	0.609	0.001	0.390	0.000

Component 1, whose mean is located in a far corner of the posterior parameter space, is characterized by low values of maximum SMV and high values of minimum SMV throughout the trial. It is unsurprising that no activities classified to this component are falls because falls are expected to be associated with large changes in acceleration. Component 5 describes activities with slightly higher values of maximum difference in SMV, and, unlike component 1, is estimated to contain a small number of falls. Table 6 indicates that quick vertical movements, such as quickly sitting and standing (D10), are likely to be classified to this component.

Components 2 and 4 both exhibit high values of maximum SMV and maximum difference in SMV. The majority of activities classified to these components are falls, with a few ADLs such as moving up and down stairs (D06), or trying to get up but collapsing into a chair (D11). The forward falls tend to be classified into component 2, whose SMV values are higher, while falls in other directions are classified into component 4. Component 3, like component 2, describes activities with very high values of maximum SMV and maximum difference, but unlike component 2 its average value of minimum SMV is low. This component contains only 25% falls, suggesting that minimum SMV is a feature that is able to distinguish ADLs from falls.

Table 6 indicates that certain types of ADLs, such as sitting slowly (D07), are unlikely to be confused with falls as indicated by their high probability of allocation to component 1. The small maximum change in (log) acceleration associated with this component, is a feature that is likely to be highly predictive of certain ADLs. Other ADLs, such as going upstairs quickly (D06), share the high-acceleration features that many falls exhibit. The similarities between ADLs that involve fast movement and forward falls suggest that additional measurements, perhaps some including a directional component, may aid in better distinguishing falls in these difficult cases.

6. Discussion

The proposed anchored Bayesian mixture model offers a model-based resolution to label-switching that eliminates prior and posterior exchangeability without imposing highly restrictive identifiability constraints. In Section 2.5 we introduced a notion of quasi-consistency which guarantees that, in large samples, a well-specified anchor model places high posterior probability on one relabeling of the true component-specific parameters. Our anchored EM strategy for selecting optimal anchor points requires several pre-processing steps, but eliminates the need for post-processing of MCMC samples, often resulting in computational savings. A carefully-specified anchor model will produce component-specific parameter estimates that reflect homogeneous subgroups in the population and arise directly from the specified model.

The examples presented in this paper have demonstrated that one or two anchor points per component are often sufficient to eliminate posterior multimodality and produce accurate parameter estimates. The optimal number of anchor points may, in some situations, be higher, and may depend on the method used to select the points. Future work will investigate this question in a variety of univariate and multivariate settings.

Non-Gaussian components The anchor model methodology is readily applicable to non-Gaussian component distributions. The model properties that we presented in this paper do not, in general, rely on Gaussian components. Anchoring always eliminates the model's posterior exchangeability and a minimal number of anchor points will typically produce distinct distributions for the γ_j, as outlined in Proposition 2, when the component distributions are continuous in $\boldsymbol{y}$ and $\boldsymbol{\gamma}$. Proposition 5, which stated that the anchor model's fit improves with the addition of more anchor points, is also true for non-Gaussian mixtures under fairly general conditions. The asymptotic result in Section 2.5 does depend on several regularity conditions on the component likelihoods and priors, which may not hold for all choices of component distributions. The anchored EM algorithm of Section 3 can be implemented in models from other families. Our own applied work [Kunkel and Peruggia, 2019] has demonstrated the use of this strategy in a Gaussian mixture of regressions model. Different component distributions will motivate new approaches to specifying anchor points and these are interesting directions of future research.

Other extensions Future work will explore ways to quantify the sensitivity of analyses to changes in the specification of the anchor points and the number of anchor points. An interesting feature of anchor models is that they admit the possibility of using improper priors for the component-specific parameters. Sensitivity to proper priors is a well-known issue in mixture modeling [Richardson and Green, 1997, Frühwirth-Schnatter, 2006] but non-informative priors are difficult to derive because they tend to be improper, leading to improper posteriors [Grazian and Robert, 2018]. An anchor model can restrict the space of latent allocations to prevent empty components, similarly to the methods proposed by Diebolt and Robert [1994] and Wasserman [2000], and this possibility allows investigation of the model's behavior under a wider class of priors.

From an applied modeling perspective, we plan to extend the anchored mixture methodology to the case of hierarchical mixture models fit to grouped data collected on many experimental units, as in the case, for example, of the entire SisFall data set. Assuming a mixture model with a fixed number of components for the data collected on each of the experimental units, with component specific parameters tied together in a hierarchical structure, several challenging modeling questions will need an answer. Decisions will have to be made concerning the number of components needed to describe the data for each experimental unit. A simple approach would employ the same number (possibly random) of components for each subject. A more refined approach would allow for varying numbers of components across units. With specific regard to the anchored methodology, we plan to investigate different approaches to the specification of the anchor points. These include selecting anchor points using independent fits to the data for each unit and strategies that account for existing dependencies in the data. We will also consider approaches where only a subset of the units will have anchored observations.

Software

An R package implementing the algorithms described in the paper is available at `https://github.com/kunkeldeborah/anchormix`.

Appendix A: Proofs of the propositions

This section presents proofs of Propositions 1 and 3–6. When necessary, references to expressions in the main manuscript will be preceded by M, so that, for example, (M2) refers to Equation (2) in the main manuscript.

Proposition 1. *The following two statements hold under conditions C.1 and C.2.*

1. *An anchor model* $A = \{A_1, \ldots, A_k\}$ *imposes a unique labeling on each partition that has nonzero probability if and only if* $A_1, \ldots, A_{k-1}$ *are nonempty; that is,* $k_0 \geq k - 1$.

2. *For any $j \le k_0$, $j \neq j'$, the marginal posterior density of $\boldsymbol{\gamma}_j$ is distinct from the marginal posterior density of $\boldsymbol{\gamma}_{j'}$ with probability 1.*

Proof of part a. Suppose A_{k-1} and A_k are empty so that at least two components have no points anchored to them. Choose an allocation $\boldsymbol{s}^*$ from $\mathcal{S}^A$ such that $\boldsymbol{s}^*$ has at least one element equal to $k-1$, so that $\boldsymbol{s}^*$ has induced a partition of the data under which one group is labeled $k-1$. The set $\mathcal{S}^A$ also contains the allocation obtained by permuting the label k and $k-1$ in $\boldsymbol{s}^*$, which induces the same partition but a different labeling. In contrast, if $A_1, \ldots, A_{k-1}$ each contain at least one point, any allocation from $\mathcal{S}^A$ induces a partition that cannot be relabeled without relabeling an anchor point.

Proof of part b. Assume that component j has at least one anchor point; that is, $A_j \neq \emptyset$, and that $\boldsymbol{x}_h$ denotes the anchor points for component h. Use $\boldsymbol{\gamma}_{-(h)}$ to denote $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{h-1}, \boldsymbol{\gamma}_{h+1}, \ldots, \boldsymbol{\gamma}_k)$ and Γ to denote the parameter space of $\boldsymbol{\gamma}_h$. The posterior density of $\boldsymbol{\gamma}_j$, denoted by $p_j(\boldsymbol{\gamma}_j|\boldsymbol{y})$, satisfies

$$p_j(\boldsymbol{\gamma}_j|\boldsymbol{y}) \propto \int_{\boldsymbol{\gamma}_{-(j)}} \int_{\boldsymbol{\eta}} p(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k, \boldsymbol{\eta}, \boldsymbol{y}) d\boldsymbol{\eta} d\boldsymbol{\gamma}_{-(j)} \tag{17}$$

$$= \int_{\boldsymbol{\gamma}_{-(j)}} \int_{\boldsymbol{\eta}} \pi(\boldsymbol{\eta}) \prod_{h=1}^{k_0} \phi_p(\boldsymbol{x}_h; \boldsymbol{\gamma}_h) \prod_{h=1}^{k} \pi(\boldsymbol{\gamma}_h) \prod_{i \notin A} \sum_{h=1}^{k} \eta_h \phi_p(y_i; \boldsymbol{\gamma}_h) d\boldsymbol{\eta} d\boldsymbol{\gamma}_{-(j)}. \tag{18}$$

Case 1: $A_{j'} \neq \emptyset$**.** Define $c(\boldsymbol{\gamma}_j, \boldsymbol{\gamma}_{j'})$ to be equal to the following function:

$$c(\boldsymbol{\gamma}_j, \boldsymbol{\gamma}_{j'}) = \int_{\boldsymbol{\gamma}_{-(j,j')}} \prod_{h \neq j,j'}^{k_0} \phi_p(\boldsymbol{x}_h; \boldsymbol{\gamma}_h) \prod_{h \neq j,j'}^{k} \pi(\boldsymbol{\gamma}_h) \int_{\boldsymbol{\eta}} \pi(\boldsymbol{\eta}) \prod_{i \notin A} \sum_{h=1}^{k} \eta_h \phi_p(y_i; \boldsymbol{\gamma}_h) d\boldsymbol{\eta} d\boldsymbol{\gamma}_{-(j,j')}, \tag{19}$$

noting that the label-invariance of the mixture likelihood gives

$$c(\boldsymbol{u}, \boldsymbol{w}) = c(\boldsymbol{w}, \boldsymbol{u}). \tag{20}$$

The expression (18) can be written as

$$p_j(\boldsymbol{\gamma}_j|\boldsymbol{y}) \propto \pi(\boldsymbol{\gamma}_j) \phi_p(\boldsymbol{x}_j; \boldsymbol{\gamma}_j) \int_{\boldsymbol{w}} \pi(\boldsymbol{w}) \phi_p(\boldsymbol{x}_{j'}; \boldsymbol{w}) c(\boldsymbol{w}, \boldsymbol{\gamma}_j) d\boldsymbol{w}. \tag{21}$$

Further, define

$$c_2(\boldsymbol{\gamma}_j, \boldsymbol{x}_{j'}) = \int_{\boldsymbol{u}} \pi(\boldsymbol{u}) \phi_p(\boldsymbol{x}_{j'}; \boldsymbol{u}) c(\boldsymbol{\gamma}_j, \boldsymbol{u}) d\boldsymbol{u}, \tag{22}$$

yielding

$$p_j(\boldsymbol{\gamma}_j|\boldsymbol{y}) \propto \pi(\boldsymbol{\gamma}_j) \phi_p(\boldsymbol{x}_j; \boldsymbol{\gamma}_j) c_2(\boldsymbol{\gamma}_j, \boldsymbol{x}_{j'}). \tag{23}$$

For component j', we have

$$p_{j'}(\boldsymbol{\gamma}_{j'}|\boldsymbol{y}) \propto \pi(\boldsymbol{\gamma}_{j'})\phi_p(\boldsymbol{x}_{j'};\boldsymbol{\gamma}_{j'}) \int_{\boldsymbol{w}} \pi(\boldsymbol{w})\phi_p(\boldsymbol{x}_j;\boldsymbol{w})c(\boldsymbol{w},\boldsymbol{\gamma}_{j'})d\boldsymbol{w} \tag{24}$$

$$= \pi(\boldsymbol{\gamma}_{j'})\phi_p(\boldsymbol{x}_{j'};\boldsymbol{\gamma}_{j'})c_2(\boldsymbol{\gamma}_{j'},\boldsymbol{x}_j), \tag{25}$$

where the equality follows from (20). Assumption C.2 ensures that there is some open set $W \subset \Gamma$ such that $\pi(\boldsymbol{w}) > 0$ for all $\boldsymbol{w} \in W$. Then for $\boldsymbol{w} \in W$, we have the posterior density of γ_j satisfying

$$p_j(\boldsymbol{w}|\boldsymbol{y}) \propto \pi(\boldsymbol{w})\phi_p(\boldsymbol{x}_j;\boldsymbol{w})c_2(\boldsymbol{w},\boldsymbol{x}_{j'}), \tag{26}$$

while the posterior density of $\boldsymbol{\gamma}_{j'}$ satisfies

$$p_{j'}(\boldsymbol{w}|\boldsymbol{y}) \propto \pi(\boldsymbol{w})\phi_p(\boldsymbol{x}_{j'};\boldsymbol{w})c_2(\boldsymbol{w},\boldsymbol{x}_j). \tag{27}$$

With probability 1, $\boldsymbol{x}_j \neq \boldsymbol{x}_{j'}$, and so $p_j(\boldsymbol{w}|\boldsymbol{y}) \neq p_{j'}(\boldsymbol{w}|\boldsymbol{y})$ for all $\boldsymbol{w} \in W$.

Case 2 If $A_{j'} = \emptyset$, then

$$p_j(\boldsymbol{\gamma}_j|\boldsymbol{y}) \propto \pi(\boldsymbol{\gamma}_j)\phi_p(\boldsymbol{x}_j;\boldsymbol{\gamma}_j) \int_{\boldsymbol{w}} \pi(\boldsymbol{w})c(\boldsymbol{w},\boldsymbol{\gamma}_j)d\boldsymbol{w} \tag{28}$$

while

$$p_{j'}(\boldsymbol{\gamma}_{j'}|\boldsymbol{y}) \propto \pi(\boldsymbol{\gamma}_{j'})c_2(\boldsymbol{\gamma}_{j'},\boldsymbol{x}_j). \tag{29}$$

For $\boldsymbol{w} \in W$, is it clear that $p_j(\boldsymbol{w}|\boldsymbol{y}) \neq p_{j'}(\boldsymbol{w}|\boldsymbol{y})$.

Proposition 3. *The qth element of* $P_{\boldsymbol{x}}(\boldsymbol{\gamma}_0)$, p_q, *is equal to*

$$\prod_{j=1}^{k_0} \phi_p\left(\boldsymbol{x}_j;\boldsymbol{\gamma}_{0\rho_q(j)}\right) \Bigg/ \sum_{h=1}^{k!}\prod_{j=1}^{k_0} \phi_p\left(\boldsymbol{x}_j;\boldsymbol{\gamma}_{0\rho_h(j)}\right). \tag{30}$$

Proof. The data dependent prior of $(\boldsymbol{\gamma},\boldsymbol{\eta})$ under model A is proportional to

$$\pi(\boldsymbol{\eta})\prod_{j=1}^{k_0} p(\boldsymbol{\gamma}_j|\boldsymbol{x}_j)\prod_{j=k_0+1}^{k}\pi(\boldsymbol{\gamma}_j).$$

The probability of the qth class label under the anchor model is

$$p_q = \frac{\pi(\rho_q(\boldsymbol{\eta}_0))\prod_{j=1}^{k_0} p(\boldsymbol{\gamma}_{0\rho_q(j)}|\boldsymbol{x}_{\rho_q(j)})\prod_{j=k_0+1}^{k}\pi(\boldsymbol{\gamma}_{0\rho_q(j)})}{\sum_{h=1}^{k!}\pi(\rho_h(\boldsymbol{\eta}_0))\prod_{j=1}^{k_0} p(\boldsymbol{\gamma}_{0\rho_h(j)}|\boldsymbol{x}_j)\prod_{j=k_0+1}^{k}\pi(\boldsymbol{\gamma}_{0\rho_h(j)})} \tag{31}$$

$$= \frac{\prod_{j=1}^{k_0} p(\boldsymbol{\gamma}_{0j}|\boldsymbol{x}_j)\prod_{j=k_0+1}^{k}\pi(\boldsymbol{\gamma}_{0\rho_q(j)})}{\sum_{h=1}^{k!}\prod_{j=1}^{k_0} p(\boldsymbol{\gamma}_{0\rho_h(j)}|\boldsymbol{x}_j)\prod_{j=k_0+1}^{k}\pi(\boldsymbol{\gamma}_{0\rho_h(j)})}, \tag{32}$$

where the terms $\pi(\rho_q(\boldsymbol{\eta}_0))$ are invariant to label permutation and are factored out in (32). For any q, we have

$$\prod_{j=1}^{k_0} p(\boldsymbol{\gamma}_{0\rho_q(j)})|\boldsymbol{x}_j) = \prod_{j=1}^{k_0} \pi(\boldsymbol{\gamma}_{0\rho_q(j)})\phi_p(\boldsymbol{x}_j;\boldsymbol{\gamma}_{0\rho_q(j)})C_j^{-1}. \tag{33}$$

As in Section 2.4 of the manuscript, C_j is defined to be

$$C_j = \int \pi(\boldsymbol{w})\phi_p(\boldsymbol{x}_j;\boldsymbol{w})d\boldsymbol{w}$$

and does not depend on the permutation of the label of $\boldsymbol{\gamma}_j$. Further, $\prod_{j=1}^{k}\pi(\boldsymbol{\gamma}_{\rho_q(j)})$ is equal to $\prod_{j=1}^{k}\pi(\boldsymbol{\gamma}_{\rho_{q'}(j)})$ for any q and q'. The only term in (32) that depends on the qth permutation is $\prod_{j=1}^{k_0}\phi_p(\boldsymbol{x}_j;\boldsymbol{\gamma}_{\rho_q(j)})$. Thus the qth element of $P_{\boldsymbol{x}}(\boldsymbol{\gamma}_0)$ is equal to

$$p_q = \frac{\prod_{j=1}^{k_0}\phi_p(\boldsymbol{x}_j;\boldsymbol{\gamma}_{0\rho_q(j)})}{\sum_{h=1}^{k!}\prod_{j=1}^{k_0}\phi_p(\boldsymbol{x}_j;\boldsymbol{\gamma}_{0\rho_h(j)})},$$

for $q = 1,\ldots,k!$.

Proposition 4. *Suppose that $p = 1$, $k = 2$, and that $m_j = m$ observations (with $1 \le m \le n/2$) are to be anchored to component j, $j = 1, 2$. The following results hold:*

1. *If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and $\theta_1 < \theta_2$, then the optimal anchoring sets $\boldsymbol{x}_1 = (y_{(1)},\ldots,y_{(m)})$ and $\boldsymbol{x}_2 = (y_{(n-m+1)},y_{(n)})$, where $y_{(l)}$ denotes the lth order statistic.*
2. *If $\theta_1 = \theta_2 = \theta$ and $\sigma_1^2 < \sigma_2^2$, then the optimal anchoring sets $\boldsymbol{x}_1$ equal to the points that minimize $\sum_{i=1}^{m}(y_i-\theta)^2$ and $\boldsymbol{x}_2$ equal to the points that maximize $\sum_{i=1}^{m}(y_i-\theta)^2$.*

Proof. When $k = 2$, $P_{\boldsymbol{x}}$ has only two elements and maximizing p_1 will minimize its entropy. Thus, it is sufficient to maximize the ratio p_1/p_2. In the location problem (case 1), p_1/p_2 equals

$$\frac{\phi(\boldsymbol{x}_1;\theta_1,\sigma^2)\phi(\boldsymbol{x}_2;\theta_2,\sigma^2)}{\phi(\boldsymbol{x}_1;\theta_2,\sigma^2)\phi(\boldsymbol{x}_2;\theta_1,\sigma^2)} = \exp\left(\frac{m}{\sigma^2}(\theta_2-\theta_1)(\bar{x}_2-\bar{x}_1)\right),$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Because we assume $\theta_1 < \theta_2$, this expression is increasing in $(\bar{x}_2 - \bar{x}_1)$. In case 2, p_1/p_2 is equal to

$$\frac{\phi(\boldsymbol{x}_1;\theta,\sigma_1^2)\phi(\boldsymbol{x}_2;\theta,\sigma_2^2)}{\phi(\boldsymbol{x}_1;\theta,\sigma_2^2)\phi(\boldsymbol{x}_2;\theta,\sigma_1^2)} = \exp\left(\left(\frac{1}{2\sigma_1^2}-\frac{1}{2\sigma_2^2}\right)\left(\sum_{i=1}^{m}(x_{2i}-\theta)^2 - \sum_{i=1}^{m}(x_{1i}-\theta)^2\right)\right).$$

Because $\sigma_1^2 < \sigma_2^2$, the ratio is increasing in $\sum_{i=1}^{m}(x_{2i}-\theta)^2$ and decreasing in $\sum_{i=1}^{m}(x_{1i}-\theta)^2$. This concludes the proof.

The proof of Proposition 5 below relies on the result stated in the following lemma.

Lemma 1. *For any anchor model A^m, $m \geq 1$, it is possible to form a sequence of anchor models $A^m, \ldots, A^n$ satisfying $A^m \subset A^{m+1} \subset \ldots A^n$ such that $m_{A^m}(\boldsymbol{y}) \leq \ldots \leq m_{A^n}(\boldsymbol{y})$.*

Proof. Given an anchor model A^m with m anchor points and some $i \notin A^m$, let $A^{m+1}_{i,j}$ denote the anchor model obtained from A^m by adding the additional anchor point with index i to component j. We will show that for some j, the marginal likelihood for $A^{m+1}_{i,j}$ is greater than or equal to that of A^m.

Let $\mathcal{S}^{A^m}$ denote the restricted set of allocation vectors for the anchor model A^m. It is possible to write $\mathcal{S}^{A^m}$ as $\cup_{j=1}^{k} \mathcal{S}^{A^{m+1}_{i,j}}$ for any i. The marginal likelihood for A^m can be written as

$$\begin{aligned} m_{A^m}(\boldsymbol{y}) &= \sum_{\mathcal{S}^{A^m}} m(\boldsymbol{y}|\boldsymbol{s}) p_{A^m}(\boldsymbol{s}) \\ &= \frac{1}{k^{n-m}} \sum_{\mathcal{S}^{A^m}} m(\boldsymbol{y}|\boldsymbol{s}) \qquad (34) \\ &= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{k^{n-m-1}} \sum_{\mathcal{S}^{A^{m+1}_{i,j}}} m(\boldsymbol{y}|\boldsymbol{s}). \qquad (35) \end{aligned}$$

Expression (34) follows from the fact that $p_{A^m}(\boldsymbol{s}) = k^{-(n-m)}$ for all $\boldsymbol{s} \in \mathcal{S}^{A^m}$ under the assumption that $\eta_j = k^{-1}$, $j = 1, \ldots, k$. Expression (35) shows that the marginal likelihood for A^m is the average of the marginal likelihoods for $A^{m+1}_{i,j}$, $j = 1, \ldots, k$. Thus,

$$\begin{aligned} m_{A^m}(\boldsymbol{y}) &= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{k^{n-m-1}} \sum_{\mathcal{S}^{A^{m+1}_{i,j}}} m(\boldsymbol{y}|\boldsymbol{s}) \\ &\leq \max_j \frac{1}{k^{n-m-1}} \sum_{\mathcal{S}^{A^{m+1}_{i,j}}} m(\boldsymbol{y}|\boldsymbol{s}) \\ &= m_{A^m_{i,j^*}}(\boldsymbol{y}), \qquad (36) \end{aligned}$$

where j^* is the component to which observation i can be anchored to maximize the marginal likelihood. Thus $m_{A^m}(\boldsymbol{y}) \leq m_{A^{m+1}_{i,j^*}}(\boldsymbol{y})$ for any $m < n$, proving Lemma 1.

Proposition 5. *Assume that $\eta_j = 1/k$, $j = 1, \ldots, k$. Let $A^1_*, \ldots, A^n_*$ be a sequence of anchor models where A^m_* has the highest marginal likelihood among all anchor models with m anchor points. The marginal likelihoods of the models satisfy $m_{A^1_*}(\boldsymbol{y}) \leq \ldots \leq m_{A^n_*}(\boldsymbol{y})$.*

Proof. Fix m and let i be the index of an observation not anchored in A_*^m. Then,

$$m_{A_*^m}(\boldsymbol{y}) \leq m_{A_{i,j^*}^{m+1}}(\boldsymbol{y}) \leq m_{A_*^{m+1}}(\boldsymbol{y}),$$

where the first inequality follows from Lemma 1 and the second from the definition of A_*^{m+1}.

Proposition 6. *Let q be the posterior distribution of the allocations under an anchor model, subject to the restrictions in (M15). The KL divergence of q_* from q, evaluated at a fixed value of $\boldsymbol{\gamma}$, is minimized when the sets A_j are chosen to maximize $\sum_{j=1}^{k_0}\sum_{i\in A_j} r_{ij}$ and when $\widetilde{r}_{ij} = r_{ij}$ for all $i \notin A$.*

Proof. For any distributions q and q_*, and defining $x\log(x) = 0$, the KL-divergence of q_* from q is equal to

$$\begin{aligned}
D_{KL}(q||q_*) &= \sum_{\boldsymbol{s}} q(\boldsymbol{s}) \log\left(\frac{q(\boldsymbol{s})}{q_*(\boldsymbol{s})}\right) \\
&= \sum_{\boldsymbol{s}} \left[\prod_{l=1}^{n} q(s_l)\right] \log\left(\frac{\prod_{i=1}^{n} q(s_i)}{\prod_{i=1}^{n} q_*(s_i)}\right) \\
&= \sum_{\boldsymbol{s}} \left[\prod_{l=1}^{n} q(s_l)\right] \sum_{i=1}^{n}\left(\log\left(\frac{q(s_i)}{q_*(s_i)}\right)\right) \\
&= \sum_{s_1=1}^{k}\cdots\sum_{s_n=1}^{k} \left[\prod_{l=1}^{n} q(s_l)\right] \sum_{i=1}^{n}\left(\log\left(\frac{q(s_i)}{q_*(s_i)}\right)\right) \\
&= \sum_{s_1=1}^{k}\cdots\sum_{s_{n-1}=1}^{k} \left[\prod_{l=1}^{n-1} q(s_l)\right] \sum_{s_n=1}^{k} q(s_n) \sum_{i=1}^{n}\left(\log\left(\frac{q(s_i)}{q_*(s_i)}\right)\right) \\
&= \sum_{s_1=1}^{k}\cdots\sum_{s_{n-1}=1}^{k} \left[\prod_{l=1}^{n-1} q(s_l)\right] \\
&\qquad \left[\sum_{s_n=1}^{k} q(s_n) \sum_{i=1}^{n-1}\left(\log\left(\frac{q(s_i)}{q_*(s_i)}\right)\right) + \sum_{s_n=1}^{k} q(s_n)\left(\log\left(\frac{q(s_n)}{q_*(s_n)}\right)\right)\right] \\
&= \sum_{s_1=1}^{k}\cdots\sum_{s_{n-1}=1}^{k} \left[\prod_{l=1}^{n-1} q(s_l)\right] \\
&\qquad \left[\sum_{i=1}^{n-1}\left(\log\left(\frac{q(s_i)}{q_*(s_i)}\right)\right) + \sum_{s_n=1}^{k} q(s_n)\left(\log\left(\frac{q(s_n)}{q_*(s_n)}\right)\right)\right] \\
&\vdots \\
&= \sum_{i=1}^{n}\sum_{s_i=1}^{k} q(s_i)\log\left(\frac{q(s_i)}{q_*(s_i)}\right).
\end{aligned}$$

Substituting the definition of q_* in (M13) and the restrictions on q from (M15) yields

$$\begin{aligned} D_{KL}(q||q_*) &= \sum_{i=1}^{n}\sum_{j=1}^{k} q(s_i=j)\log\left(\frac{q(s_i=j)}{r_{ij}}\right) \\ &= \sum_{i\in A_j}\sum_{j=1}^{k} I(s_i=j)\log\left(\frac{I(s_i=j)}{r_{ij}}\right) + \sum_{i\notin A}\sum_{j=1}^{k}\widetilde{r}_{ij}\log\left(\frac{\widetilde{r}_{ij}}{r_{ij}}\right) \\ &= -\sum_{j=1}^{k}\sum_{i\in A_j}\log\left(r_{ij}\right) + \sum_{j=1}^{k}\sum_{i\notin A}\widetilde{r}_{ij}\log\left(\frac{\widetilde{r}_{ij}}{r_{ij}}\right). \end{aligned} \tag{37}$$

The second term in (37) is non-negative and can be made equal to zero by setting $\widetilde{r}_{ij} = r_{ij}$ for $i \notin A$. The first term decreases toward zero as $r_{ij} \to 1$ for $i \in A$. Subject to the restriction that $A_j \cap A_{j'} = \emptyset$ and $|A_j| = m_j$, the q that minimizes $D_{KL}(q||q_*)$ is

$$q(S_i=j) = \begin{cases} r_{ij}, & i \notin A, \\ 1, & i \in A_j, \\ 0, & i \in A_{j'}, \quad j' \neq j, \end{cases}$$

where the A_j are selected to maximize $\sum_{j=1}^{k}\sum_{i\in A_j} r_{ij}$. The sets A_j will contain the m_j observations with the highest value(s) of r_{ij}, for $j = 1, \ldots, k_0$, if there are no observations that fit this criterion for more than one value of j.

Appendix B: Simulation study

Here, we describe in detail the simulation studies referenced in Section 3.1 of the main manuscript, evaluating how the number of anchor points affects out-of-sample predictions from the anchor model.

B.1. Simulation 1

In the first simulation study, we simulated data from a mixture with $k = 2$ mixture components to assess the relationship between goodness of fit and predictive performance of the anchor model. For each data set we fit a univariate location model assuming $\sigma_1 = \sigma_2 = 1$ and $\eta_1 = \eta_2 = 0.5$, using independent $N(0, 25)$ prior distributions on the component means θ_1 and θ_2. We generated data from a Gaussian mixture with means $\theta_1 = 0$, $\theta_2 = \delta$ and standard deviations $\sigma_1 = 1$ and $\sigma_2 = \sigma$. We considered several values of δ to assess the effect of separation among the mixture components and several values of σ to assess the effect of model misspecification. For each combination of δ and σ, we generated 1,000 small data sets, $\boldsymbol{y}_{j,(\delta,\sigma)}$, $j = 1, \ldots, 1,000$, of size $n = 10$. To each of these data sets, we fit nine anchor models $A^2_{j,(\delta,\sigma)}, \ldots, A^{10}_{j,(\delta,\sigma)}$, having $m = 2, \ldots, 10$

anchor points such that $A^m_{j,(\delta,\sigma)}$ has the highest marginal likelihood among anchor models with m anchor points, subject to the additional requirement that at least one anchor point be assigned to each of the two components. To block out uninteresting sources of variation, we used a common master batch of 1,000 data sets sampled from a model with standard normal mixture components and obtained the 1,000 data sets for each (δ, σ) pair through appropriate rescaling and translation of the observations.

We assessed the out-of-sample predictive performance of each model using its expected log pointwise predictive density (ELPPD) [Gelman, Hwang and Vehtari, 2014] as follows. For a given simulated data set, $\boldsymbol{y}_{j,(\delta,\sigma)}$, we generated $\widetilde{N} = 1,000$ replicate data sets, $\widetilde{\boldsymbol{y}}^1_{j,(\delta,\sigma)}, \ldots, \widetilde{\boldsymbol{y}}^{\widetilde{N}}_{j,(\delta,\sigma)}$, from the same distribution as that of $\boldsymbol{y}_{j,(\delta,\sigma)}$. Again, this was done using a common master batch of 1,000 standardized data sets. For each anchor model $A^m_{j,(\delta,\sigma)}$, we generated $T = 3,000$ Monte Carlo samples of the parameters $\boldsymbol{\gamma}^1, \ldots, \boldsymbol{\gamma}^T$ from the posterior distribution of $\boldsymbol{\gamma}$ conditional on $\boldsymbol{y}_{j,(\delta,\sigma)}$.

We then estimated the ELPPD for that anchor model fit to $\boldsymbol{y}_{j,(\delta,\sigma)}$ by the quantity $1/\widetilde{N} \sum_{i=1}^{\widetilde{N}} \left\{ \sum_{k=1}^{10} \left[\log \left(T^{-1} \sum_{t=1}^{T} f(\widetilde{y}^{i,k}_{j,(\delta,\sigma)} | \boldsymbol{\gamma}^t) \right) \right] \right\}$, which provides a Monte Carlo estimate of the expected log predictive density, $p(\tilde{\boldsymbol{y}} | \boldsymbol{y}_{j,(\delta,\sigma)})$, of a new sample, $\tilde{\boldsymbol{y}}$, and will be large when the model has strong predictive performance.

Figure A7 shows boxplots of the 1,000 estimated ELPPD values for each of the (δ, σ) simulation settings, with $\delta = 0.25, 1.75, 2.75$ and $\sigma_2 = 0.10, 1$, and each value for the number of anchor points, m, between 2 and 9. For the settings with the largest value of $\delta = 2.75$ in which the mixture components are well-separated, there is only a slight increase in predictive performance as the number of anchored points increases. For the smallest value of $\delta = 0.25$, however, the predictive performance deteriorates when the number of anchor points is large: the ELPPD values appear to have both lower medians and higher variability for models with more anchor points. For the settings in which the data were generated with $\sigma \neq 1$, this pattern of deterioration as m grows is more apparent.

B.2. Simulation 2

The second simulation sought to investigate the effect of increasing the number of anchor points when using the anchored EM algorithm. We simulated data from models 1, 2, and 3, whose parameters are given in Table M2 and analyses of which are presented in Section M4.3 with the same prior and hyperparameter specification. From each model we generated 250 data sets of size $n = 100$ and fit eight different anchor models, requiring $m_j = m$, $j = 1, \ldots, k$, using the values of $m = 1, 2, 3, 5, 7, 9, 10, 12$. The anchor points were chosen using the anchored EM method as well as two oracle methods. The first set of oracle anchor points (random oracle points) were selected randomly from, for component j, the observations generated from component j. The second

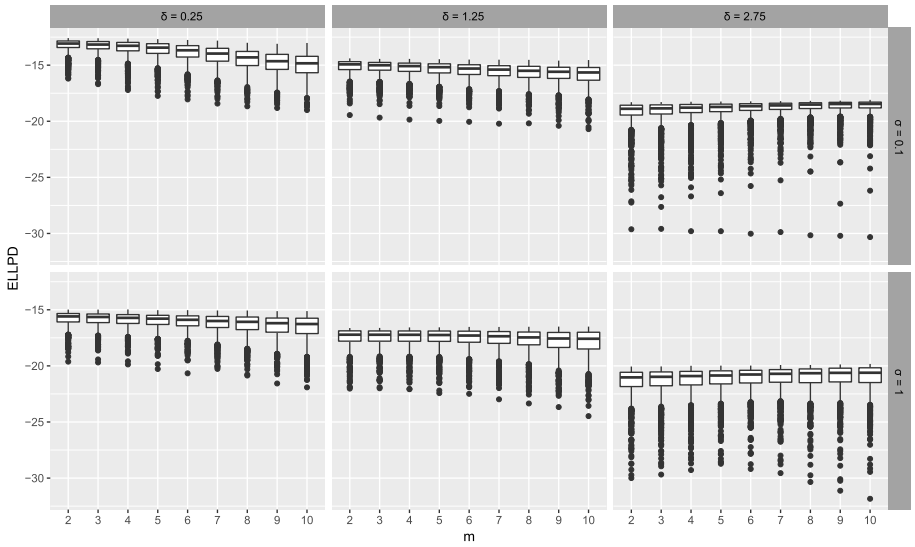

FIG A7. *Boxplots of the values of the ELPPD of each simulated data set for anchor models with* $m = 2, \ldots, 10$. *The panels show results for selected experimental conditions with* $\delta = 0.25, 1.75, 2.75$ *and* $\sigma = 0.1, 1$.

set of oracle anchor points (quantile oracle points) were chosen to be evenly-spaced percentiles from the true component densities, beginning at the 5th percentile and ending at the 95th. The model with $m = 1$ uses only the 5th percentile.

The estimated ELPPD was calculated for each data set to assess the model's predictive performance. Figure A8 shows boxplots of the ELPPD values for models 1, 2, and 3 respectively.

The left-hand panels display the values under anchor models specified using anchored EM, as would occur in practice. For model 1, the two-component scale mixture, the median predictive performance improves slightly as m increases, but for model 2, the four-component location mixture, the performance deteriorates noticeably as the number of anchor points increases. For model 3 there is also a slight tendency for the ELPPD to decrease with higher values of m. In all models there is more variability in the ELPPD when m is large. This pattern suggest that in models with substantial overlap among components, using many points selected by anchored EM leads to models that may introduce bias. In all models there is little apparent difference in the average ELPPD between models with $m = 1$ and $m = 2$.

The random and quantile oracle methods demonstrate how predictive performance changes with anchor points that represent true features of the distributions. In the random oracle models, shown in the middle panels of Figure A8 the anchor points are guaranteed to be a random sample from the component to which they are anchored. In models 2 and 3, there is little difference in the

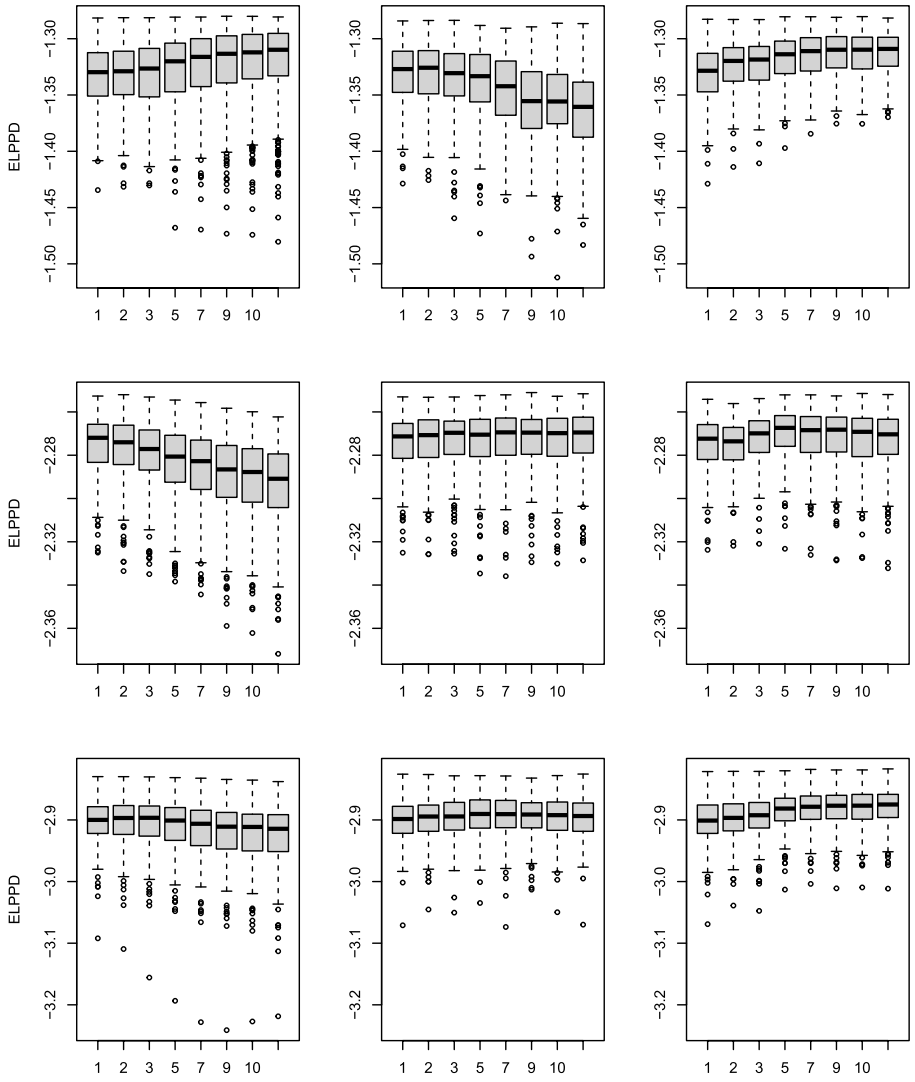

FIG A8. *Estimated ELPPD values under the anchored EM (left), random oracle (center), and quantile oracle (right) anchor models. Rows 1, 2 and 3 give results for models 1, 2, and 3, respectively.*

model performance as m increases. In model 1, larger numbers of points lead to poorer predictive performance. The estimated ELPPD values under the quantile oracle models, are shown in the right panels of Figure A8. As m increases, these points will become increasingly representative of the features of the true component densities. It is unsurprising that, for these methods, there is a slight increase in predictive performance as m increases.

Appendix C: Anchored EM algorithm

In this section we list the pseudo-code of the Anchored EM algorithm described in Section 3.2 of the main manuscript.

Algorithm 1 Anchored EM Algorithm

0: Initialize $\boldsymbol{\gamma}^0$ and select $\epsilon > 0$. Set $t = 1$ and $\delta > \epsilon$.
while $\delta \geq \epsilon$ **do**
E step: Calculate the unconstrained posterior probabilities r_{ij}^t at the current value of $\boldsymbol{\gamma}^{t-1}$. Initialize $A^t = \emptyset$.
Find A^t to maximize $\sum_{j=1}^{k_0} \sum_{i \in A_j} r_{ij}^t$.

Set $\widetilde{r}_{ij}^t = 1$ and $\widetilde{r}_{il}^t = 0, l \neq j$ for $i \in A_j^t$, $j = 1, \ldots, k_0$.
Set $\widetilde{r}_{ij}^t = r_{ij}^t$ for $i \notin A^t, j = 1, \ldots, k$. Define q^t as in $M14$ using $\widetilde{r}_{ij}^t$.
M step: Set $\boldsymbol{\gamma}^t = \arg\max_{\boldsymbol{\gamma}} F(\boldsymbol{\gamma}, q^t)$.
$\delta = F(\boldsymbol{\gamma}^t, q^t) - F(\boldsymbol{\gamma}^{t-1}, q^{t-1})$; t++.
end while

An exact solution to the maximization of A^t above may be framed as a solution to a transportation problem where k_0 sources must be connected to n destinations and the cost of connecting source j to destination i is $-r_{ij}^t$. The solution can be expressed as an $n \times k_0$ matrix B of 0's and 1's, where $B_{ij} = 1$ if $i \in A_j$, subject to the constraints $\sum_{j=1}^{k_0} B_{ij} \leq 1$ and $\sum_{i=1}^{n} B_{ij} = m_j$. The function `lp.transport` in the R package `lpSolve` [Berkelaar, 2015] can be used to perform this step of the algorithm. Alternatively, an approximate solution may be found using, for example, Vogel's approximation method and variants thereof [Mathirajan and Meenakshi, 2004, Juman and Hoque, 2015] or using the heuristic method described in the pseudo-code below.

while $|A^t| < m$ **do**
Find $i', j' = \max_{i,j: i \notin A^t, |A_j^t| < m_j} r_{ij}^t$. Set $i' \in A_{j'}$.
Set $\widetilde{r}_{i'j'}^t = 1$ and $\widetilde{r}_{i'l}^t = 0, l \neq j'$.
end while

Appendix D: Random permutation sampler

As noted in Section 4.1, to prove that the random permutation sampler is a Gibbs sampler that generates draws from the target posterior distribution we need to show that the accepted randomly permuted draw $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ is in fact a draw from the distribution of $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ given $(\boldsymbol{s}, \boldsymbol{y})$. Let C be a measurable subset and let the symbol $p_T(\cdot|\cdot)$ denote generically a conditional density under the target posterior for the given model (either the exchangeable or the anchor model). We note first that the absolute values of the Jacobian determinants for the $k!$ permutations and their inverses are all equal to one and can therefore be ignored in the following derivation. Then we have

$$
\begin{aligned}
&P((\boldsymbol{\gamma},\boldsymbol{\eta}) \in C|\boldsymbol{s},\boldsymbol{y}) \\
&= \sum_{q=1}^{k!} \int\limits_{\rho_q^{-1}(C)} \frac{p_T(\rho_q(\rho_q^{-1}(\boldsymbol{\gamma},\boldsymbol{\eta}))|\boldsymbol{y})}{\sum_{h=1}^{k!} p_T(\rho_h(\rho_q^{-1}(\boldsymbol{\gamma},\boldsymbol{\eta}))|\boldsymbol{y})} p_T(\rho_q^{-1}(\boldsymbol{\gamma},\boldsymbol{\eta})|\boldsymbol{s},\boldsymbol{y})\, d(\rho_q^{-1}(\boldsymbol{\gamma},\boldsymbol{\eta})) \\
&= \sum_{q=1}^{k!} \int\limits_{C} \frac{p_T(\rho_q(\boldsymbol{\gamma},\boldsymbol{\eta})|\boldsymbol{y})}{\sum_{h=1}^{k!} p_T(\rho_h(\boldsymbol{\gamma},\boldsymbol{\eta})|\boldsymbol{y})} p_T(\boldsymbol{\gamma},\boldsymbol{\eta}|\boldsymbol{s},\boldsymbol{y})\, d(\boldsymbol{\gamma},\boldsymbol{\eta}) \\
&= \int\limits_{C} \frac{\sum_{q=1}^{k!} p_T(\rho_q(\boldsymbol{\gamma},\boldsymbol{\eta})|\boldsymbol{y})}{\sum_{h=1}^{k!} p_T(\rho_h(\boldsymbol{\gamma},\boldsymbol{\eta})|\boldsymbol{y})} p_T(\boldsymbol{\gamma},\boldsymbol{\eta}|\boldsymbol{s},\boldsymbol{y})\, d(\boldsymbol{\gamma},\boldsymbol{\eta}) \\
&= \int\limits_{C} p_T(\boldsymbol{\gamma},\boldsymbol{\eta}|\boldsymbol{s},\boldsymbol{y})\, d(\boldsymbol{\gamma},\boldsymbol{\eta}) \\
&= P_T((\boldsymbol{\gamma},\boldsymbol{\eta}) \in C|\boldsymbol{s},\boldsymbol{y}),
\end{aligned}
$$

proving the claim. Note that for the exchangeable model all relabelings will have the same probability $1/k!$ of being selected and the algorithm reduces to the original algorithm proposed in Frühwirth-Schnatter [2001].

Appendix E: Univariate examples with $m_j = 2$

In this section, we revisit the univariate examples of Section (M4) using the anchored EM method to select two anchor points per component. We also use oracle anchor points chosen as the 5th and 95th percentiles of the true component distributions. The minimum-entropy method is not demonstrated for this case.

Figure A9 shows the scaled component densities under model 1 for the anchor models and relabeling methods. Table A7 gives posterior means of the model parameters for all methods, including the anchor models with $m_j = 1$ as presented in the main manuscript. Figure A10 and Table A8 show the scaled component densities and posterior means of the model parameters, respectively, for all methods for model 2. The results for model 3 are shown in Figure A11 and Table A9.

In model 1, the anchor models with two anchor points provide parameter estimates with higher accuracy than the ones with one anchor point. This improvement is, however, not seen for all parameters in model 2 and model 3. The anchored EM model with $m_j = 1$ provides more accurate estimates of both $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ for model 2, although $m_j = 2$ produces better estimates of $\boldsymbol{\sigma}$. For model 3, the anchored EM model is more accurate with $m_j = 1$ than with $m_j = 2$.

Appendix F: Analysis of the SisFall data

This section provides additional details on the data analysis example presented in Section 5 of the main manuscript.

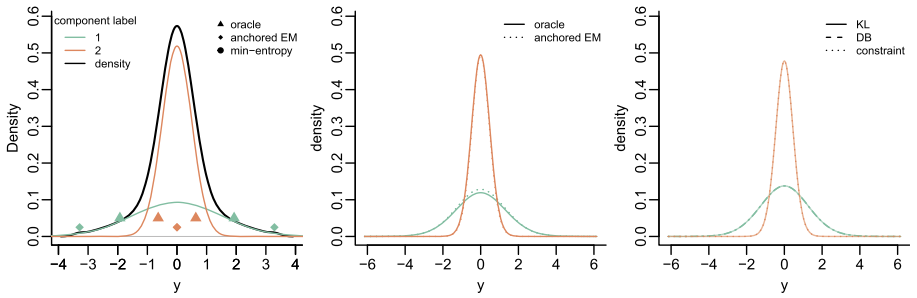

FIG A9. *Left panel: anchor points and true scaled component densities for model 1. Middle panel: estimated scaled component densities for the anchor models. Right panel: estimated scaled component densities for the relabeling methods and prior constraints.*

TABLE A7
Posterior means of the component-specific parameters for model 1.

	θ_1	θ_2	error	time (s)
true	0	0		
oracle anchors ($m_j = 2$)	0.0004 (0.0014)	−0.0008 (0.0005)	0.0012	N/A
anchored EM ($m_j = 2$)	−0.0004 (0.0014)	−0.0007 (0.0005)	0.0011	1.03
oracle anchors ($m_j = 1$)	0.0010 (0.0010)	−0.0010 (0.0010)	0.0020	N/A
anchored EM ($m_j = 1$)	−0.0027 (0.0014)	−0.0009 (0.0005)	0.0036	1.03
min-entropy($m_j = 1$)	−0.0003 (0.0014)	−0.0006 (0.0005)	0.0009	0.50
KL	0.0005 (0.0013)	−0.0008 (0.0005)	0.0013	1102.78
DB	0.0005 (0.0013)	−0.0008 (0.0005)	0.0013	547.59
constraint	0.0005 (0.0013)	−0.0008 (0.0005)	0.0013	N/A
	σ_1	σ_2	error	
true	1.5	0.5		
oracle anchors ($m_j = 2$)	1.356 (0.0013)	0.485 (0.0005)	0.1593	
anchored EM ($m_j = 2$)	1.333 (0.0013)	0.470 (0.0006)	0.1973	
oracle anchors ($m_j = 1$)	0.867 (0.0035)	0.866 (0.0035)	0.9988	
anchored EM ($m_j = 1$)	1.320 (0.0013)	0.467 (0.0006)	0.2125	
min-entropy ($m_j = 1$)	1.320 (0.0013)	0.468 (0.0006)	0.2122	
KL	1.307 (0.0013)	0.464 (0.0005)	0.2293	
DB	1.307 (0.0013)	0.464 (0.0006)	0.2293	
constraint	1.733 (0.0013)	0.220 (0.0005)	0.5129	
	η_1	η_2	error	
true	0.35	0.65		
oracle anchors ($m_j = 2$)	0.395 (0.0008)	0.605 (0.0008)	0.0907	
anchored EM ($m_j = 2$)	0.418 (0.0008)	0.582 (0.0008)	0.1363	
oracle anchors ($m_j = 1$)	0.501 (0.0009)	0.499 (0.0009)	0.3020	
anchored EM ($m_j = 1$)	0.430 (0.0009)	0.570 (0.0009)	0.1594	
min-entropy ($m_j = 1$)	0.428 (0.0009)	0.572 (0.0009)	0.1568	
KL	0.439 (0.0009)	0.561 (0.0009)	0.1781	
DB	0.439 (0.0009)	0.561 (0.0009)	0.1782	
constraint	0.439 (0.0009)	0.561 (0.0009)	0.1781	

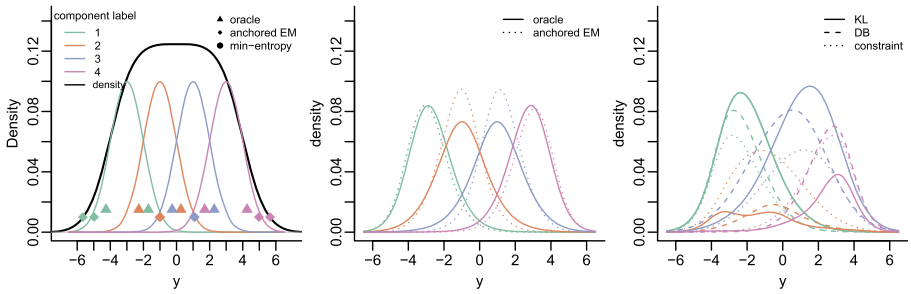

FIG A10. *Left panel: anchor points and true scaled component densities for model 2. Middle panel: estimated scaled component densities for the anchor models. Right panel: estimated scaled component densities for the relabeling methods and prior constraints.*

TABLE A8
Posterior means of the component-specific parameters for model 2.

	θ_1	θ_2	θ_3	θ_4	error	time (s)
true	−3	−1	1	3		
oracle anchors ($m_j = 2$)	−2.802 (0.0037)	−0.978 (0.0059)	0.979 (0.0057)	2.800 (0.0038)	0.442	N/A
anchored EM ($m_j = 2$)	−3.104 (0.0051)	−1.061 (0.0050)	1.138 (0.0050)	3.140 (0.0051)	0.443	21.39
oracle anchors ($m_j = 1$)	−2.819 (0.0049)	−0.960 (0.0072)	0.956 (0.0072)	2.818 (0.0049)	0.448	N/A
anchored EM ($m_j = 1$)	−2.983 (0.0069)	−0.960 (0.0073)	0.997 (0.0073)	2.997 (0.0068)	0.063	24.87
min-entropy ($m_j = 1$)	−2.967 (0.0116)	−1.653 (0.0095)	1.965 (0.0088)	2.864 (0.0134)	1.786	10.37
KL	−2.208 (0.0078)	−0.942 (0.0376)	1.288 (0.0090)	1.864 (0.0318)	2.274	6134.31
DB	−2.536 (0.0141)	−0.259 (0.0393)	0.278 (0.0136)	2.520 (0.0199)	2.408	159.00
constraint	−3.477 (0.0230)	−1.039 (0.0129)	1.027 (0.0127)	3.492 (0.0225)	1.035	N/A
	σ_1	σ_2	σ_3	σ_4	error	
true	1	1	1	1		
oracle anchors ($m_j = 2$)	1.109 (0.0020)	1.257 (0.0030)	1.259 (0.0031)	1.108 (0.0020)	0.7333	
anchored EM ($m_j = 2$)	1.011 (0.0020)	0.954 (0.0027)	0.945 (0.0026)	1.001 (0.0020)	0.1135	
oracle anchors ($m_j = 1$)	1.024 (0.0026)	1.067 (0.0041)	1.063 (0.0039)	1.026 (0.0025)	0.1805	
anchored EM ($m_j = 1$)	1.074 (0.0025)	0.984 (0.0031)	0.984 (0.0032)	1.069 (0.0025)	0.1752	
min-entropy ($m_j = 1$)	1.121 (0.0040)	1.089 (0.0036)	1.042 (0.0034)	1.168 (0.0043)	0.4192	
KL	1.243 (0.0037)	0.699 (0.0069)	1.497 (0.0023)	0.766 (0.0068)	1.270	
DB	1.071 (0.0033)	0.687 (0.0028)	1.476 (0.0050)	0.970 (0.0031)	0.8893	
constraint	0.981 (0.0038)	1.123 (0.0048)	1.122 (0.0048)	0.979 (0.0038)	0.2848	
	η_1	η_2	η_3	η_4	error	
true	0.25	0.25	0.25	0.25		
oracle anchors ($m_j = 2$)	0.245 (0.0007)	0.254 (0.0009)	0.255 (0.0009)	0.246 (0.0007)	0.0181	
anchored EM ($m_j = 2$)	0.236 (0.0010)	0.269 (0.0010)	0.266 (0.0010)	0.229 (0.0009)	0.0702	
oracle anchors ($m_j = 1$)	0.247 (0.0011)	0.252 (0.0013)	0.251 (0.0013)	0.250 (0.0010)	0.0059	
anchored EM ($m_j = 1$)	0.255 (0.0012)	0.247 (0.0012)	0.247 (0.0012)	0.251 (0.0012)	0.0124	
min-entropy ($m_j = 1$)	0.222 (0.0016)	0.283 (0.0016)	0.263 (0.0014)	0.232 (0.0018)	0.0923	
KL	0.353 (0.0015)	0.075 (0.0006)	0.459 (0.0017)	0.114 (0.0008)	0.6230	
DB	0.273 (0.0015)	0.077 (0.0007)	0.427 (0.0022)	0.223 (0.0014)	0.4010	
constraint	0.224 (0.0017)	0.276 (0.0020)	0.277 (0.0019)	0.224 (0.0017)	0.1047	

Data The full SisFall data set collected by [Sucerquia, López and Vargas-Bonilla, 2017], with additional details on the experimental procedure, is available at http://sistemic.udea.edu.co/en/research/projects/english-falls/.

Additional tables Table A10 gives posterior means of $\boldsymbol{\theta}_j$ for each mixture component. Table A11 gives the full table of posterior allocation probabilities for each of the activities considered in the analysis. Activities beginning with "D" are ADLs and activities beginning with "F" are falls.

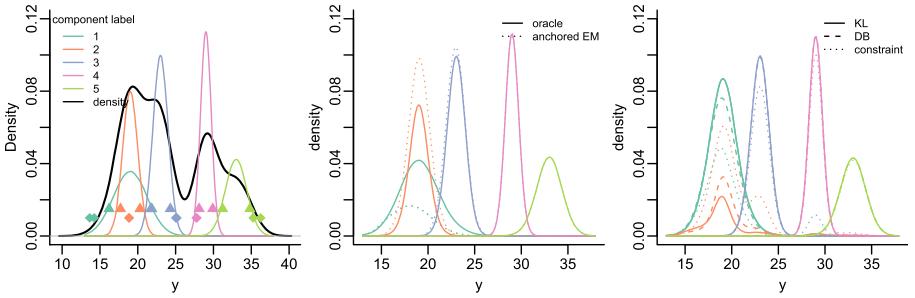

FIG A11. *Left panel: anchor points and true scaled component densities for model 3. Middle panel: estimated scaled component densities for the anchor models. Right panel: estimated scaled component densities for the relabeling methods and prior constraints.*

TABLE A9
Posterior means of the component-specific parameters for Model 3.

	θ_1	θ_2	θ_3	θ_4	θ_5	error	time (s)
true	19	19	23	29	33		
oracle anchors ($m_j = 2$)	18.990 (0.0053)	19.012 (0.0020)	23.023 (0.0011)	29.007 (0.0006)	32.993 (0.0016)	0.059	N/A
anchored EM ($m_j = 2$)	16.822 (0.0224)	19.044 (0.0015)	22.954 (0.0012)	29.006 (0.0006)	32.990 (0.0016)	2.284	16.11
oracle anchors ($m_j = 1$)	18.955 (0.0049)	18.958 (0.0049)	23.030 (0.0014)	29.006 (0.0006)	32.990 (0.0016)	0.135	N/A
anchored EM ($m_j = 1$)	17.273 (0.0290)	19.049 (0.0019)	22.975 (0.0014)	29.006 (0.0006)	32.991 (0.0016)	1.816	16.60
min-entropy ($m_j = 1$)	18.948 (0.0014)	23.011 (0.0010)	29.006 (0.0006)	33.031 (0.0056)	33.838 (0.0100)	14.94	10.56
KL	19.086 (0.0026)	20.087 (0.0868)	23.028 (0.0016)	29.005 (0.0007)	32.992 (0.0018)	1.215	986.31
DB	18.799 (0.0062)	20.335 (0.0831)	23.065 (0.0044)	29.007 (0.0007)	32.994 (0.0018)	1.613	285.12
constraint	17.848 (0.0255)	20.062 (0.0287)	23.856 (0.0351)	29.266 (0.0117)	33.167 (0.0150)	3.504	N/A
	σ_1	σ_2	σ_3	σ_4	σ_5	error	
true	2.236	1	1	0.7071	1.414		
oracle anchors ($m_j = 2$)	2.059 (0.0028)	1.018 (0.0020)	0.973 (0.0008)	0.717 (0.0005)	1.363 (0.0012)	0.284	
anchored EM ($m_j = 2$)	1.377 (0.0067)	1.125 (0.0031)	1.023 (0.0009)	0.719 (0.0005)	1.364 (0.0012)	1.070	
oracle anchors ($m_j = 1$)	1.371 (0.0048)	1.380 (0.0048)	0.973 (0.0010)	0.719 (0.0005)	1.364 (0.0012)	1.334	
anchored EM ($m_j = 1$)	1.490 (0.0081)	1.073 (0.0042)	1.011 (0.0009)	0.719 (0.0005)	1.363 (0.0012)	0.892	
min-entropy ($m_j = 1$)	1.606 (0.0010)	1.011 (0.0007)	0.718 (0.0005)	1.067 (0.0025)	1.134 (0.0037)	1.563	
KL	1.581 (0.0023)	0.986 (0.0062)	0.973 (0.0013)	0.716 (0.0011)	1.358 (0.0007)	0.760	
DB	1.579 (0.0036)	0.989 (0.0052)	0.971 (0.0012)	0.716 (0.0006)	1.358 (0.0013)	0.761	
constraint	1.322 (0.0049)	1.273 (0.0060)	0.934 (0.0024)	0.746 (0.0018)	1.338 (0.0019)	1.368	
	η_1	η_2	η_3	η_4	η_5	error	
true	0.2	0.2	0.25	0.2	0.15		
oracle anchors ($m_j = 2$)	0.219 (0.0006)	0.187 (NA)	0.244 (0.0003)	0.201 (0.0001)	0.149 (0.0001)	0.0405	
anchored EM ($m_j = 2$)	0.094 (0.0015)	0.286 (0.0013)	0.269 (0.0003)	0.201 (0.0001)	0.149 (0.0001)	0.2134	
oracle anchors ($m_j = 1$)	0.201 (0.0008)	0.204 (0.0008)	0.245 (0.0003)	0.201 (0.0001)	0.149 (0.0001)	0.0127	
anchored EM ($m_j = 1$)	0.132 (0.0021)	0.255 (0.0019)	0.264 (0.0003)	0.201 (0.0001)	0.149 (0.0001)	0.1387	
min-entropy ($m_j = 1$)	0.392 (0.0002)	0.255 (0.0002)	0.200 (0.0002)	0.088 (0.0004)	0.064 (0.0004)	0.4954	
KL	0.334 (0.0011)	0.074 (0.0008)	0.246 (0.0005)	0.198 (0.0002)	0.147 (0.0002)	0.2688	
DB	0.303 (0.0016)	0.106 (0.0013)	0.245 (0.0005)	0.198 (0.0002)	0.147 (0.0002)	0.2063	
constraint	0.189 (0.0023)	0.253 (0.0012)	0.228 (0.0010)	0.186 (0.0007)	0.143 (0.0003)	0.1060	

TABLE A10
Posterior means of $\boldsymbol{\theta}_j$ for each component from the SisFall data.

Component	1	2	3	4	5
$\log(\max_t SMV_t)$	5.769	7.304	7.169	6.972	6.366
$\log(\min_t SMV_t)$	5.323	3.661	1.686	4.626	3.795
$\log(\max_t \lvert SMV_t - SMV_{t-1}\rvert)$	3.162	5.963	5.802	5.021	3.979

TABLE A11
Posterior allocation probabilities for selected activities in the SisFall data set.

Activity	Component				
	1	2	3	4	5
D01 Walking slowly	0.099	0.008	0.000	0.490	0.403
D02 Walking quickly	0.000	0.086	0.000	0.066	0.848
D03 Jogging slowly	0.000	0.000	1.000	0.000	0.000
D04 Jogging quickly	0.000	0.011	0.989	0.000	0.000
D05 Walking upstairs and downstairs slowly	0.059	0.027	0.000	0.887	0.027
D06 Walking upstairs and downstairs quickly	0.000	0.796	0.188	0.000	0.016
D07 Slowly sit in a half height chair, wait a moment, and up slowly	0.981	0.000	0.000	0.006	0.013
D08 Quickly sit in a half height chair, wait a moment, and up quickly	0.000	0.019	0.001	0.024	0.956
D09 Slowly sit in a low height chair, wait a moment, and up slowly	0.245	0.050	0.000	0.575	0.131
D10 Quickly sit in a low height chair, wait a moment, and up quickly	0.001	0.042	0.001	0.148	0.808
D11 Sitting a moment, trying to get up, and collapse into a chair	0.000	0.541	0.003	0.454	0.002
D12 Sitting a moment, lying slowly, wait a moment, and sit again	0.938	0.020	0.000	0.035	0.007
D13 Sitting a moment, lying quickly, wait a moment, and sit again	0.173	0.110	0.000	0.695	0.022
D14 Being on one's back change to lateral position, wait a moment, and change to one's back	0.942	0.001	0.000	0.038	0.019
D15 Standing, slowly bending at knees, and getting up	0.978	0.000	0.000	0.010	0.012
D16 Standing, slowly bending without bending knees, and getting up	0.986	0.000	0.000	0.008	0.005
D17 Standing, get into a car, remain seated and get out of the car	0.797	0.003	0.000	0.156	0.045
D18 Stumble while walking	0.000	0.957	0.003	0.040	0.000
D19 Gently jump without falling (trying to reach a high object)	0.000	0.308	0.033	0.000	0.658
F01 Fall forward while walking caused by a slip	0.000	0.915	0.001	0.084	0.000
F02 Fall backward while walking caused by a slip	0.000	0.302	0.001	0.697	0.000
F03 Lateral fall while walking caused by a slip	0.000	0.250	0.000	0.750	0.000
F04 Fall forward while walking caused by a trip	0.000	0.938	0.001	0.061	0.000
F05 Fall forward while jogging caused by a trip	0.000	0.799	0.201	0.000	0.000
F06 Vertical fall while walking caused by fainting	0.000	0.058	0.000	0.942	0.000
F07 Fall while walking, with use of hands in a table to dampen fall, caused by fainting	0.000	0.211	0.001	0.788	0.000
F08 Fall forward when trying to get up	0.001	0.586	0.001	0.412	0.000
F09 Lateral fall when trying to get up	0.000	0.074	0.000	0.926	0.000
F10 Fall forward when trying to sit down	0.000	0.609	0.001	0.390	0.000
F11 Fall backward when trying to sit down	0.000	0.230	0.002	0.760	0.008
F12 Lateral fall when trying to sit down	0.000	0.239	0.001	0.759	0.000
F13 Fall forward while sitting, caused by fainting or falling asleep	0.014	0.726	0.001	0.124	0.135
F14 Fall backward while sitting, caused by fainting or falling asleep	0.001	0.527	0.003	0.467	0.002
F15 Lateral fall while sitting, caused by fainting or falling asleep	0.000	0.439	0.001	0.561	0.000

Acknowledgments

This paper is based on results presented in the first author's Ph.D. dissertation [Kunkel, 2018]. The authors' work was partially supported by the National Science Foundation under Grant No. SES-1424481 and No. SES-1921523. The authors thank Steven MacEachern, Bettina Grün, and two anonymous reviewers for their helpful comments.

References

[1] ABNEY, S. (2004). Understanding the Yarowsky Algorithm. *Computational Linguistics* **30** 365–395. MR2087949

[2] ALBERT, M. V., KORDING, K., HERRMANN, M. and JAYARAMAN, A. (2012). Fall Classification by Machine Learning Using Mobile Phones. *PLOS ONE* **7** 1-6.

[3] BARDENET, R., CAPPE, O., FORT, G. and KEGL, B. (2012). Adaptive Metropolis with Online Relabeling. In *Proceedings of the Fif-*

teenth International Conference on Artificial Intelligence and Statistics (N. D. Lawrence and M. Girolami, eds.). *Proceedings of Machine Learning Research* **22** 91–99. PMLR, La Palma, Canary Islands.

[4] Berk, R. H. (1966). Limiting Behavior of Posterior Distributions when the Model is Incorrect. *Annals of Mathematical Statistics* **37** 51–58. MR0189176

[5] Berkelaar, M. (2015). lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs R package version 5.6.13.

[6] Bourke, A. K. and Lyons, G. M. (2008). A Threshold-Based Fall-Detection Algorithm Using a Bi-axial Gyroscope Sensor. *Medical Engineering & Physics* **30** 84–90.

[7] Casilari, E., Santoyo-Ramón, J.-A. and Cano-García, J.-M. (2017). Analysis of Public Datasets for Wearable Fall Detection Systems. *Sensors* **17**.

[8] Celeux, G., Hurn, M. and Robert, C. P. (2000). Computational and Inferential Difficulties with Mixture Posterior Distributions. *Journal of the American Statistical Association* **95** 957-970. MR1804450

[9] Chung, H., Loken, E. and Schafer, J. L. (2004). Difficulties in Drawing Inferences With Finite-Mixture Models. *The American Statistician* **58** 152-158. MR2109393

[10] Cooley, C. A. and MacEachern, S. N. (1999). Prior Elicitation in the Classification Problem. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **27** 299-313. MR1704403

[11] Diebolt, J. and Robert, C. P. (1994). Estimation of Finite Mixture Distributions through Bayesian Sampling. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **56** 363-375. MR1281940

[12] Egidi, L., Pappadà, R., Pauli, F. and Torelli, N. (2018). Relabelling in Bayesian Mixture Models by Pivotal Units. *Statistics and Computing* **28** 957–969. MR3766053

[13] Flegal, J. M., Hughes, J., Vats, D. and Dai, N. (2020). mcmcse: Monte Carlo Standard Errors for MCMC, Riverside, CA, Denver, CO, Coventry, UK, and Minneapolis, MN R package version 1.4-1.

[14] Frühwirth-Schnatter, S. (2001). Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models. *Journal of the American Statistical Association* **96** 194-209. MR1952732

[15] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer. MR2265601

[16] Ganchev, K., Taskar, B. and Gama, J. (2008). Expectation Maximization and Posterior Constraints. In *Advances in Neural Information Processing Systems 20* (J. C. Platt, D. Koller, Y. Singer and S. T. Roweis, eds.) 569–576. Curran Associates, Inc. MR1047248

[17] Gelman, A., Hwang, J. and Vehtari, A. (2014). Understanding Predictive Information Criteria for Bayesian Models. *Statistics and Computing* **24** 997–1016. MR3253850

[18] Genovese, C. R. and Wasserman, L. (2000). Rates of Convergence for the Gaussian Mixture Sieve. *The Annals of Statistics* **28** 1105–1127.

MR1810921

[19] Geweke, J. (2007). Interpretation and Inference in Mixture Models: Simple MCMC Works. *Computational Statistics & Data Analysis* **51** 3529–3550. MR2367818

[20] Grazian, C. and Robert, C. P. (2018). Jeffreys Priors for Mixture Estimation: Properties and Alternatives. *Computational Statistics & Data Analysis* **121** 149–163. MR3759204

[21] Jasra, A., Holmes, C. C. and Stephens, D. A. (2005). Markov chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science* **20** 50–67. MR2182987

[22] Juman, Z. A. M. S. and Hoque, M. A. (2015). An Efficient Heuristic to Obtain a Better Initial Feasible Solution to the Transportation Problem. *Applied Soft Computing* **34** 813–826.

[23] Kunkel, D. (2018). Anchored Bayesian Gaussian Mixture Models, PhD thesis, The Ohio State University.

[24] Kunkel, D. and Peruggia, M. (2019). Statistical inference with anchored Bayesian mixture of regressions models: A case study analysis of allometric data. *arXiv preprint* arXiv:1905.04389.

[25] Li, H. and Fan, X. (2016). A Pivotal Allocation-Based Algorithm for Solving the Label-Switching Problem in Bayesian Mixture Models. *Journal of Computational and Graphical Statistics* **25** 266-283. MR3474047

[26] Lücke, J. (2016). Truncated Variational Expectation Maximization. *arXiv preprint* arXiv:1610.03113.

[27] Maechler, M. (2019). nor1mix: Normal aka Gaussian (1-d) Mixture Models R package version 1.3-0.

[28] Marin, J.-M., Mengersen, K. and Robert, C. P. (2005). Bayesian Modelling and Inference on Mixtures of Distributions. In *Bayesian Thinking Modeling and Computation*, (D. K. Dey and C. R. Rao, eds.). *Handbook of Statistics* **25** 459–507. Elsevier. MR2490536

[29] Marin, J.-M. and Robert, C. P. (2014). *Bayesian Essentials with R.* Springer. MR3136532

[30] Mathirajan, M. and Meenakshi, B. (2004). Experimental Analysis of Some Variants of Vogel's Approximation Method. *Asia-Pacific Journal of Operational Research* **21** 447-462.

[31] Neal, R. M. and Hinton, G. E. (1998). *A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants.* In *Learning in Graphical Models* 355–368. Springer Netherlands, Dordrecht.

[32] Norets, A. and Pelenis, J. (2012). Bayesian Modeling of Joint and Conditional Distributions. *Journal of Econometrics* **168** 332-346. MR2923772

[33] Papastamoulis, P. (2016). label.switching: An R Package for Dealing with the Label Switching Problem in MCMC Outputs. *Journal of Statistical Software, Code Snippets* **69** 1–24.

[34] Papastamoulis, P. and Iliopoulos, G. (2010). An Artificial Allocations Based Solution to the Label Switching Problem in Bayesian Analysis of Mixtures of Distributions. *Journal of Computational and Graphical Statistics* **19** 313-331. MR2758306

[35] Richardson, S. and Green, P. J. (1997). 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. MR1483213

[36] Rodriguez, C. E. and Walker, S. G. (2014). Label Switching in Bayesian Mixture Models. *Journal of Computational and Graphical Statistics* **23** 25-45. MR3173759

[37] Roeder, K. (1990). Density Estimation with Confidence Sets Exemplified by Superclusters and Voids in the Galaxies. *Journal of the American Statistical Association* **85** 617-624.

[38] Roeder, K. and Wasserman, L. (1997). Practical Bayesian Density Estimation Using Mixtures of Normals. *Journal of the American Statistical Association* **92** 894-902. MR1482121

[39] Rossi, P. E. (2014). *Bayesian Non- and Semi-parametric Methods and Applications*. Princeton University Press. MR3288097

[40] Stephens, M. (2000). Dealing with Label Switching in Mixture Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **62** 795-809. MR1796293

[41] Sucerquia, A., López, J. D. and Vargas-Bonilla, J. F. (2017). SisFall: A Fall and Movement Dataset. *Sensors* **17**.

[42] Wasserman, L. (2000). Asymptotic Inference for Mixture Models by Using Data-Dependent Priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62** 159–180. MR1747402

[43] Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics. ACL '95* 189–196. Association for Computational Linguistics, Stroudsburg, PA, USA.