The Annals of Statistics
2021, Vol. 49, No. 6, 3043–3069
https://doi.org/10.1214/21-AOS2072

ESTIMATING THE NUMBER OF COMPONENTS IN FINITE MIXTURE MODELS VIA THE GROUP-SORT-FUSE PROCEDURE

BY TUDOR MANOLE[1] AND ABBAS KHALILI[2]

[1]*Department of Statistics and Data Science, Carnegie Mellon University, tmanole@andrew.cmu.edu*
[2]*Department of Mathematics and Statistics, McGill University, abbas.khalili@mcgill.ca*

Estimation of the number of components (or order) of a finite mixture model is a long standing and challenging problem in statistics. We propose the Group-Sort-Fuse (GSF) procedure—a new penalized likelihood approach for simultaneous estimation of the order and mixing measure in multidimensional finite mixture models. Unlike methods which fit and compare mixtures with varying orders using criteria involving model complexity, our approach directly penalizes a continuous function of the model parameters. More specifically, given a conservative upper bound on the order, the GSF groups and sorts mixture component parameters to fuse those which are redundant. For a wide range of finite mixture models, we show that the GSF is consistent in estimating the true mixture order and achieves the $n^{-1/2}$ convergence rate for parameter estimation up to polylogarithmic factors. The GSF is implemented for several univariate and multivariate mixture models in the R package `GroupSortFuse`. Its finite sample performance is supported by a thorough simulation study, and its application is illustrated on two real data examples.

1. Introduction. Mixture models are a flexible tool for modelling data from a population consisting of multiple hidden homogeneous subpopulations. Applications in economics (Bosch-Domènech et al. (2010)), machine learning (Goodfellow, Bengio and Courville (2016)), genetics (Bechtel et al. (1993)) and other life sciences (Thompson, Smith and Boyle (1998), Morris, Richmond and Grimshaw (1996)) frequently employ mixture distributions. A comprehensive review of statistical inference and applications of finite mixture models can be found in the book by McLachlan and Peel (2000).

Given integers $N, d \geq 1$, let $\mathcal{F} = \{f(\mathbf{y}; \boldsymbol{\theta}) : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top \in \Theta \subseteq \mathbb{R}^d, \mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^N\}$ be a parametric family of density functions with respect to a σ-finite measure ν, with a compact parameter space Θ. The density function of a finite mixture model with respect to $\mathcal{F}$ is given by

$$p_G(\mathbf{y}) = \int_\Theta f(\mathbf{y}; \boldsymbol{\theta}) \, \mathrm{d}G(\boldsymbol{\theta}) = \sum_{j=1}^K \pi_j f(\mathbf{y}; \boldsymbol{\theta}_j), \tag{1.1}$$

where

$$G = \sum_{j=1}^K \pi_j \delta_{\boldsymbol{\theta}_j} \tag{1.2}$$

is the mixing measure with $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jd})^\top \in \Theta$, $j = 1, \ldots, K$, and the mixing probabilities $0 \leq \pi_j \leq 1$ satisfy $\sum_{j=1}^K \pi_j = 1$. Here, $\delta_{\boldsymbol{\theta}}$ denotes the Dirac measure placing mass at $\boldsymbol{\theta} \in \Theta$. The $\boldsymbol{\theta}_j$ are said to be atoms of G, and K is called the *order* of the model.

Received October 2019; revised December 2020.
MSC2020 subject classifications. Primary 62F10, 62F12; secondary 62H12.
Key words and phrases. Finite mixture models, maximum penalized likelihood estimation, Wasserstein distance, strong identifiability.

Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a random sample from a finite mixture model (1.1) with true mixing measure $G_0 = \sum_{j=1}^{K_0} \pi_{0j} \delta_{\boldsymbol{\theta}_{0j}}$. The true order K_0 is defined as the smallest number of atoms of G_0 for which the component densities $f(\cdot; \boldsymbol{\theta}_{0j})$ are different, and the mixing proportions π_{0j} are nonzero. This paper is concerned with parametric estimation of K_0.

In practice, the order of a finite mixture model may not be known. An assessment of the order is important even if it is not the main object of study. Indeed, a mixture model whose order is less than the true number of underlying subpopulations provides a poor fit, while a model with too large of an order, which is said to be overfitted, may be overly complex and hence uninformative. From a theoretical standpoint, estimation of overfitted finite mixture models leads to a deterioration in rates of convergence of standard parametric estimators. Indeed, given a consistent estimator G_n of G_0 with $K > K_0$ atoms, the parametric $n^{-1/2}$ convergence rate is generally not achievable. Under the so-called second-order strong identifiability condition, Chen (1995) and Ho and Nguyen (2016a) showed that the optimal pointwise rate of convergence in estimating G_0 is bounded below by $n^{-1/4}$ with respect to an appropriate Wasserstein metric. In particular, this rate is achieved by the maximum likelihood estimator up to a polylogarithmic factor. Minimax rates of convergence have also been established by Heinrich and Kahn (2018), under stronger regularity conditions on the parametric family $\mathcal{F}$. Remarkably, these rates deteriorate as the upper bound K increases. This behaviour has also been noticed for pointwise estimation rates in mixtures which do not satisfy the second-order strong identifiability assumption—see, for instance, Chen and Chen (2003) and Ho and Nguyen (2016b). These results warn against fitting finite mixture models with an incorrectly specified order. In addition to poor convergence rates, the consistency of G_n does not guarantee the consistent estimation of the mixing probabilities and atoms of the true mixing measure, though they are of greater interest in most applications.

The aforementioned challenges have resulted in the development of many methods for estimating the order of a finite mixture model. It is difficult to provide a comprehensive list of the research on this problem, and thus we give a selective overview. One class of methods involves hypothesis testing on the order using likelihood-based procedures (Dacunha-Castelle and Gassiat (1999), McLachlan (1987), Liu and Shao (2003)), and the EM-test (Chen and Li (2009), Li and Chen (2010)). These tests typically assume knowledge of a candidate order; when such a candidate is unavailable, estimation methods can be employed. Minimum distance-based methods for estimating K_0 have been considered by Chen and Kalbfleisch (1996), James, Priebe and Marchette (2001), Woo and Sriram (2006), Heinrich and Kahn (2018), and Ho, Nguyen and Ritov (2020). The most common parametric methods involve the use of an information criterion, whereby a penalized likelihood function is evaluated for a sequence of candidate models. Examples include Akaike's Information Criterion (AIC; Akaike (1974)) and the Bayesian Information Criterion (BIC; Schwarz (1978)). The latter is arguably the most frequently used method for mixture order estimation (Keribin (2000), Leroux (1992), McLachlan and Peel (2000)), though it was not originally developed for nonregular models. This led to the development of information criteria such as the Integrated Completed Likelihood (ICL; Biernacki, Celeux and Govaert (2000)), and the Singular BIC (sBIC; Drton and Plummer (2017)). Bayesian approaches include the method of Mixtures of Finite Mixtures, whereby a prior is placed on the number of components (Miller and Harrison (2018), Nobile (1994), Richardson and Green (1997), Stephens (2000)), and model selection procedures based on Dirichlet Process mixtures, such as those of Ishwaran, James and Sun (2001) and the Merge–Truncate–Merge method of Guha, Ho and Nguyen (2019). Motivated by regularization techniques in regression, Chen and Khalili (2008) proposed a penalized likelihood method for order estimation in finite mixture models with a one-dimensional parameter space Θ, where the regularization is applied to the difference between sorted atoms of the overfitted mixture model. Hung et al. (2013) adapted this method to estimation of

the number of states in Gaussian Hidden Markov models, which was also limited to one-dimensional parameters for different states. Despite its model selection consistency and good finite sample performance, the extension of this method to multidimensional mixtures has not been addressed. In this paper, we take on this task and propose a far-reaching generalization called the Group-Sort-Fuse (GSF) procedure.

The GSF postulates an overfitted finite mixture model with a large tentative order $K > K_0$. The true order K_0 and the mixing measure G_0 are simultaneously estimated by merging redundant mixture components, by applying two penalties to the log-likelihood function of the model. The first of these penalties groups the estimated atoms, while the second penalty shrinks the distances between those which are in high proximity. The latter is achieved by applying a sparsity-inducing regularization function to consecutive distances between these atoms, sorted using a so-called *cluster ordering* (Definition 2). Unlike most existing methods, this form of regularization, which uses continuous functions of the model parameters as penalties, circumvents the fitting of mixture models of all orders $1, 2, \ldots, K$. In our simulations we noticed that using EM-type algorithms (Dempster, Laird and Rubin (1977)), the GSF is less sensitive to the choice of starting values than methods which involve maximizing likelihoods of mixture models with different orders. By increasing the amount of regularization, the GSF produces a series of fitted mixture models with decreasing orders, as shown in Figure 1 for a simulated dataset. This qualitative representation, inspired by coefficient plots in penalized regression (Friedman, Hastie and Tibshirani (2008)), can also provide insight on the mixture order and parameter estimates for purposes of exploratory data analysis.

The main contributions of this paper are summarized as follows. For a wide range of second-order strongly identifiable parametric families, the GSF is shown to consistently estimate the true order K_0, and achieves the $n^{-1/2}$ rate of convergence in parameter estimation up to polylogarithmic factors. To achieve this result, the sparsity-inducing penalties used in the GSF must satisfy conditions which are nonstandard in the regularization literature. We also derived, for the first time, sufficient conditions for the strong identifiability of multinomial mixture models. Thorough simulation studies based on multivariate location-Gaussian and multinomial mixture models show that the GSF performs well in practice. The method

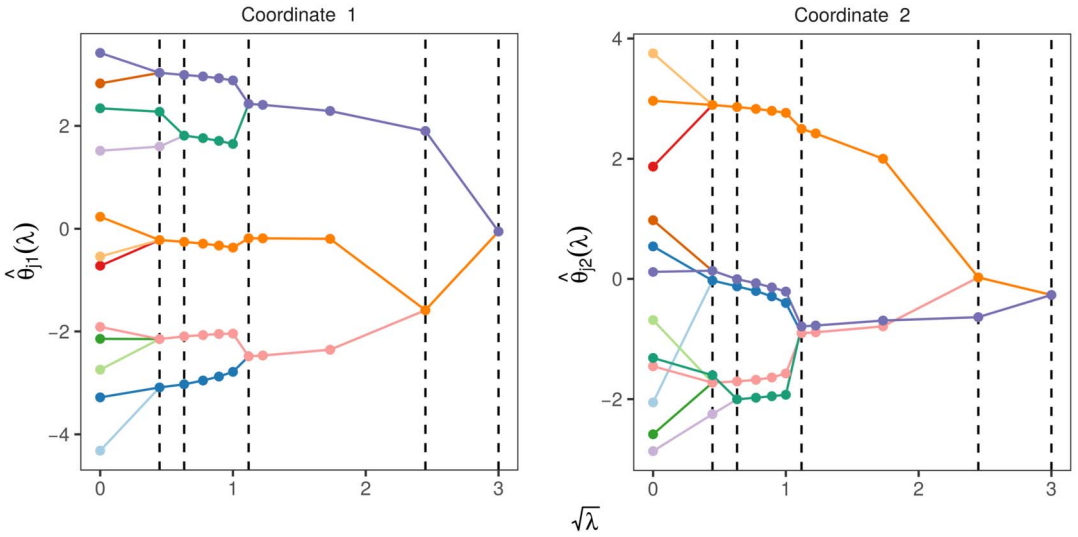

FIG. 1. *Regularization plots based on simulated data from a location-Gaussian mixture with* $K_0 = 5$, $d = 2$. *The fitted atoms* $\widehat{\boldsymbol{\theta}}_j(\lambda) = (\widehat{\theta}_{j1}(\lambda), \widehat{\theta}_{j2}(\lambda))^\top$, $j = 1, \ldots, K = 12$, *are plotted against a regularization parameter* λ. *Across coordinates, each estimated atom is identified by a unique color.*

is implemented for several univariate and multivariate mixture models in the R package `GroupSortFuse`.[1]

The rest of this paper is organized as follows. We describe the GSF method, and compare it to a naive alternative in Section 2. Asymptotic properties of the method are studied in Section 3. Our simulation results and two real data examples are respectively presented in Sections 4 and 5, and Supplement E.6 (Manole and Khalili (2021)). We close with some discussions in Section 6. Proofs, numerical implementation, and additional simulation results are given in Supplements A–F.

Notation. Throughout the paper, $|A|$ denotes the cardinality of a set A, and for any integer $K \geq 1$, $A^K = A \times \cdots \times A$ denotes the K-fold Cartesian product of A with itself. S_K denotes the set of permutations on K elements $\{1, 2, \ldots, K\}$. Given a vector $\mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, we denote its ℓ_p-norm by $\|\mathbf{x}\|_p = (\sum_{j=1}^d |x_j|^p)^{1/p}$, for all $1 \leq p < \infty$. In the case of the Euclidean norm $\|\cdot\|_2$, we omit the subscript and write $\|\cdot\|$. The diameter of a set $A \subseteq \mathbb{R}^d$ is denoted $\mathrm{diam}(A) = \sup\{\|x - y\| : x, y \in A\}$. Given two sequences of real numbers $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we write $a_n \lesssim b_n$ to indicate that there exists a constant $C > 0$ such that $a_n \leq C b_n$ for all $n \geq 1$. We write $a_n \asymp b_n$ if $a_n \lesssim b_n \lesssim a_n$. For any $a, b \in \mathbb{R}$, we write $a \wedge b = \min\{a, b\}$, $a \vee b = \max\{a, b\}$, and $a_+ = a \vee 0$. Finally, we let $\mathcal{G}_K = \{G : G = \sum_{j=1}^K \pi_j \delta_{\boldsymbol{\theta}_j}, \boldsymbol{\theta}_j \in \Theta, \pi_j \geq 0, \sum_{j=1}^K \pi_j = 1\}$ be the class of mixing measures with at most K components.

Figures. All the numerical and algorithmic details of the illustrative figures throughout this paper are given in Section 4 and Supplement D.

2. The Group-Sort-Fuse (GSF) method. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a random sample arising from p_{G_0}, where $G_0 \in \mathcal{G}_{K_0}$ is the true mixing measure with unknown order K_0. Assume an upper bound K on K_0 is known—further discussion on the choice of K is given in Section 3.3. The log-likelihood function of a mixing measure G with $K > K_0$ atoms is said to be overfitted, and is defined by

$$l_n(G) = \sum_{i=1}^n \log p_G(\mathbf{Y}_i). \tag{2.1}$$

The overfitted maximum likelihood estimator (MLE) of G is given by

$$\bar{G}_n = \sum_{j=1}^K \bar{\pi}_j \delta_{\bar{\boldsymbol{\theta}}_j} = \operatorname*{argmax}_{G \in \mathcal{G}_K} l_n(G). \tag{2.2}$$

As discussed in the Introduction, though the overfitted MLE is consistent in estimating G_0 under suitable metrics, it suffers from slow rates of convergence, and there may exist atoms of $\bar{G}_n$ whose corresponding mixing probabilities vanish, and do not converge to any atoms of G_0. Furthermore, from a model selection standpoint, $\bar{G}_n$ typically has order greater than K_0. In practice, $\bar{G}_n$ therefore overfits the data in the following two ways which we will refer to below: (a) certain fitted mixing probabilities $\bar{\pi}_j$ may be near-zero, and (b) some of the estimated atoms $\bar{\boldsymbol{\theta}}_j$ may be in high proximity to each other. In this section, we propose a penalized maximum likelihood approach which circumvents both types of overfitting, thus leading to a consistent estimator of K_0.

Overfitting (a) can readily be addressed by imposing a lower bound on the mixing probabilities, as was considered by Hathaway (1986). This lower bound, however, could be particularly challenging to specify in overfitted mixture models. An alternative approach is to

[1] https://github.com/tmanole/GroupSortFuse

penalize against near-zero mixing probabilities (Chen and Kalbfleisch (1996)). Thus, we begin by considering the following preliminary penalized log-likelihood function

$$l_n(G) - \varphi(\pi_1, \ldots, \pi_K), \quad G \in \mathcal{G}_K, \tag{2.3}$$

where $\varphi \equiv \varphi_n$ is a nonnegative penalty function such that $\inf_{n\geq 1} \varphi_n(\pi_1, \ldots, \pi_K) \to \infty$ as $\min_{1\leq j\leq K} \pi_j \to 0$. We further require that φ is invariant to relabeling of its arguments, that is, $\varphi(\pi_1, \ldots, \pi_K) = \varphi(\pi_{\tau(1)}, \ldots, \pi_{\tau(K)})$, for any permutation $\tau \in S_K$. Examples of φ are given at the end of this section. The presence of this penalty ensures that the maximizer of (2.3) has mixing probabilities which stay bounded away from zero. Consequently, as shown in Theorem 1 below, this preliminary estimator is consistent in estimating the atoms of G_0, unlike the overfitted MLE in (2.2). It does not, however, consistently estimate the order K_0 of G_0, as it does not address overfitting (b).

Our approach is to introduce a second penalty which has the effect of merging fitted atoms that are in high proximity. We achieve this by applying a sparsity-inducing penalty r_{λ_n} to the distances between appropriately chosen pairs of atoms of the overfitted mixture model with order K. It is worth noting that one could naively apply r_{λ_n} to all $\binom{K}{2}$ pairwise atom distances. However, our simulations shown toward the end of this section suggest that such an exhaustive form of penalization increases the sensitivity of the estimator to the upper bound K. Instead, given a carefully chosen sorting of the atoms in $\mathbb{R}^d$, our method merely penalizes their $K-1$ consecutive distances. This results in the double penalized log-likelihood $L_n(G)$ in (2.6), which we now describe using the following definitions.

DEFINITION 1. Let $\mathbf{t}_1, \ldots, \mathbf{t}_K \in \Theta \subseteq \mathbb{R}^d$, and let $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_H\}$ be a partition of $\{\mathbf{t}_1, \ldots, \mathbf{t}_K\}$, for some integer $1 \leq H \leq K$. Suppose

$$\max_{\mathbf{t}_i, \mathbf{t}_j \in \mathcal{C}_h} \|\mathbf{t}_i - \mathbf{t}_j\| < \min_{\substack{\mathbf{t}_i \in \mathcal{C}_h \\ \mathbf{t}_l \notin \mathcal{C}_h}} \|\mathbf{t}_i - \mathbf{t}_l\|, \quad h = 1, \ldots, H. \tag{2.4}$$

Then, each set $\mathcal{C}_h$ is said to be an atom cluster, and $\mathcal{P}$ is said to be a cluster partition.

According to Definition 1, a partition is said to be a cluster partition if the within-cluster distances between atoms are always smaller than the between-cluster distances. The penalization in (2.3) (asymptotically) induces a cluster partition $\{\mathcal{C}_1, \ldots, \mathcal{C}_{K_0}\}$ of the estimated atoms. Heuristically, the estimated atoms falling within each atom cluster $\mathcal{C}_h$ approximate some true atom $\boldsymbol{\theta}_{0j}$, and the goal of the GSF is to merge these estimates, as illustrated in Figure 2. To do so, the GSF hinges on the notion of *cluster ordering*—a generalization of the natural ordering on the real line, which we now define.

DEFINITION 2. Let $\mathbf{t} = (\mathbf{t}_1, \ldots, \mathbf{t}_K) \in \Theta^K$. A cluster ordering is a permutation $\alpha_{\mathbf{t}} \in S_K$ such that the following two properties hold:

(i) *Symmetry.* For any permutation $\tau \in S_K$, if $\mathbf{t}' = (\mathbf{t}_{\tau(1)}, \ldots, \mathbf{t}_{\tau(K)})$, then $\alpha_{\mathbf{t}'} = \alpha_{\mathbf{t}}$.

(ii) *Atom Ordering.* For any integer $1 \leq H \leq K$ and for any cluster partition $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_H\}$ of $\{\mathbf{t}_1, \ldots, \mathbf{t}_K\}$, $\alpha_{\mathbf{t}}^{-1}(\{j : \mathbf{t}_j \in \mathcal{C}_h\})$ is a set of consecutive integers for all $h = 1, \ldots, H$.

If $t_1, \ldots, t_K \in \Theta \subseteq \mathbb{R}$ and $\mathbf{t} = (t_1, \ldots, t_K)$, then the permutation $\alpha_{\mathbf{t}} \in S_K$ which induces the natural ordering $t_{\alpha_{\mathbf{t}}(1)} \leq \cdots \leq t_{\alpha_{\mathbf{t}}(K)}$ is a cluster ordering. When $\Theta \subseteq \mathbb{R}^d$, property (ii) is satisfied for any permutation $\alpha_{\mathbf{t}} \in S_K$ such that

$$\alpha_{\mathbf{t}}(k) = \operatorname*{argmin}_{\substack{1\leq j\leq K \\ j\notin\{\alpha_{\mathbf{t}}(i):1\leq i\leq k-1\}}} \|\mathbf{t}_j - \mathbf{t}_{\alpha_{\mathbf{t}}(k-1)}\|, \quad k = 2, \ldots, K. \tag{2.5}$$

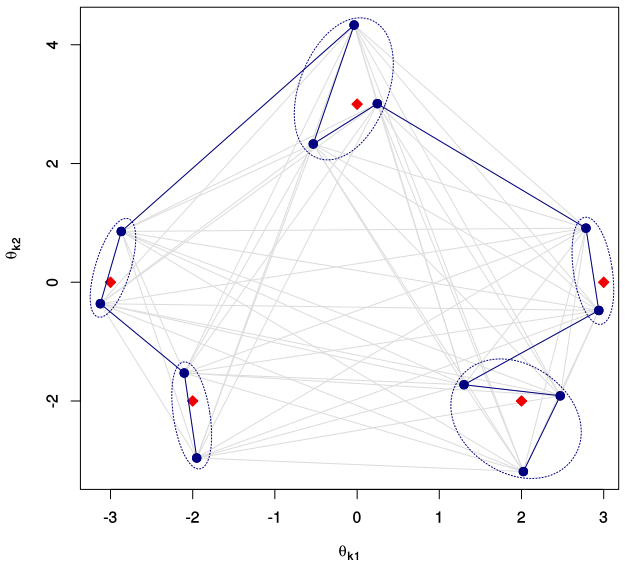

FIG. 2. *Illustration of a cluster partition $\mathcal{P}$ and a cluster ordering $\alpha_{\tilde{\boldsymbol{\theta}}}$ with $K = 12$, based on the simulated sample used in Figure 1, with true atoms $\boldsymbol{\theta}_{01}, \ldots, \boldsymbol{\theta}_{05}$ denoted by lozenges (◆), and atoms $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{12})$, obtained by maximizing the penalized log-likelihood (2.3), denoted by disks (●). The ellipses (…) represent a choice of $\mathcal{P}$ with $K_0 = 5$ atom clusters. The blue line (—) represents a cluster ordering $\alpha_{\tilde{\boldsymbol{\theta}}}$, in the sense that $\alpha_{\tilde{\boldsymbol{\theta}}}(1)$ is the index of the bottommost point, $\alpha_{\tilde{\boldsymbol{\theta}}}(2)$ is the index of the following point on the line, etc. The grey lines (—) represent all the pairwise distances penalized by the naive method defined in Figure 3.*

$\alpha_{\mathbf{t}}$ further satisfies property (i) provided $\alpha_{\mathbf{t}}(1)$ is invariant to relabeling of the components of $\mathbf{t}$. Any such choice of $\alpha_{\mathbf{t}}$ is therefore a cluster ordering in $\mathbb{R}^d$, and an example is shown in Figure 2 based on a simulated sample.

Given a mixing measure $G = \sum_{j=1}^{K} \pi_j \delta_{\boldsymbol{\theta}_j}$ with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, let $\alpha_{\boldsymbol{\theta}}$ be a cluster ordering. For ease of notation, in what follows we write $\alpha \equiv \alpha_{\boldsymbol{\theta}}$. Let $\boldsymbol{\eta}_j = \boldsymbol{\theta}_{\alpha(j+1)} - \boldsymbol{\theta}_{\alpha(j)}$, for all $j = 1, \ldots, K-1$. We define the penalized log-likelihood function

$$L_n(G) = l_n(G) - \varphi(\pi_1, \ldots, \pi_K) - n \sum_{j=1}^{K-1} r_{\lambda_n}\big(\|\boldsymbol{\eta}_j\|; \omega_j\big), \tag{2.6}$$

where the penalty $r_{\lambda_n}(\eta; \omega)$ is a nonsmooth function at $\eta = 0$ for all $\omega > 0$, satisfying conditions (P1)–(P3) discussed in Section 3. In particular, $\lambda_n \geq 0$ is a regularization parameter, and $\omega_j \equiv \omega_j(G) > 0$ are possibly random weights as defined in Section 3. Property (i) in Definition 2, and the invariance of φ to relabelling of its arguments, guarantee that $L_n(G)$ is well-defined in the sense that it does not change upon relabeling the atoms of G. Finally, the Maximum Penalized Likelihood Estimator (MPLE) of G is given by

$$\widehat{G}_n = \sum_{j=1}^{K} \widehat{\pi}_j \delta_{\widehat{\boldsymbol{\theta}}_j} = \operatorname*{argmax}_{G \in \mathcal{G}_K} L_n(G). \tag{2.7}$$

To summarize, the penalty φ ensures the asymptotic existence of a cluster partition $\{\mathcal{C}_1, \ldots, \mathcal{C}_{K_0}\}$ of $\{\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_K\}$. Heuristically, the estimated atoms in each $\mathcal{C}_h$ approximate one of the atoms of G_0, and the goal of the GSF is to merge their values to be equal. To achieve this, Property (ii) of Definition 2 implies that any cluster ordering α is among the permutations in S_K which maximize the number of indices j such that $\boldsymbol{\theta}_{\alpha(j)}, \boldsymbol{\theta}_{\alpha(j+1)} \in \mathcal{C}_h$, and minimize the number of indices l such that $\boldsymbol{\theta}_{\alpha(l)} \in \mathcal{C}_h$ and $\boldsymbol{\theta}_{\alpha(l+1)} \notin \mathcal{C}_h$, for all $h = 1, \ldots, K_0$. Thus our choice of α maximizes the number of penalty terms $r_{\lambda_n}(\|\boldsymbol{\eta}_j\|; \omega_j)$ acting on distances between atoms of the same atom cluster $\mathcal{C}_h$. The nondifferentiability of r_{λ_n} at zero

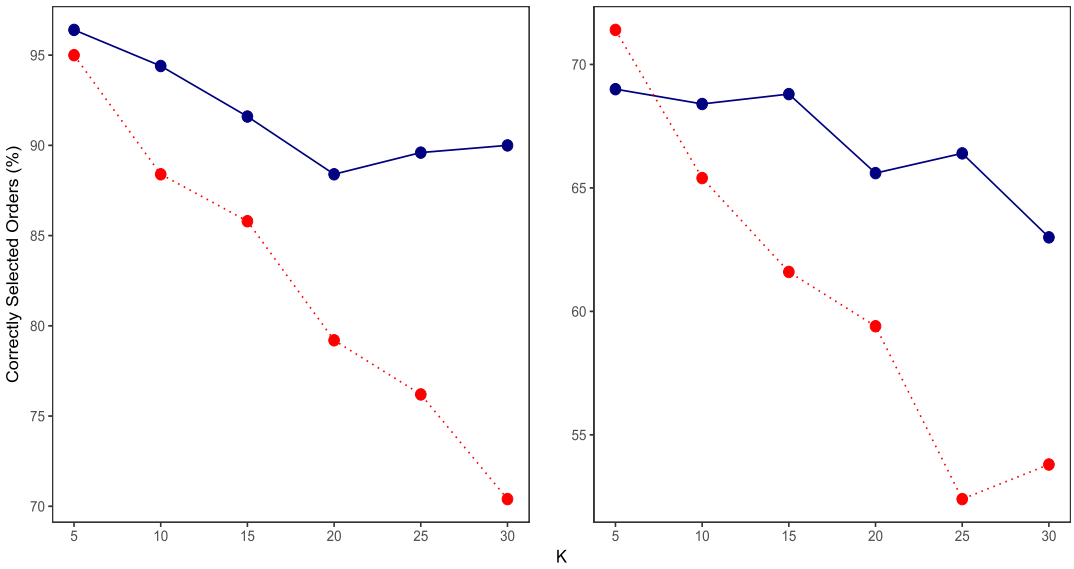

FIG. 3. *A comparison of the GSF* (—), *and the naive alternative* (...) *given by* $\operatorname{argmax}_{G\in\mathcal{G}_K}\{l_n(G) - \varphi(\pi_1,\ldots,\pi_K) - n\sum_{j\neq k} r_{\lambda_n}(\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_k\|;\omega_{jk})\}$. *The results are based on* 500 *simulated samples of size* $n = 200$ *from the bivariate Gaussian mixture Models F.*1 (*left*, $K_0 = 2$) *and F.*2 (*right*, $K_0 = 3$) *given in Supplement F. Each point represents the percentage of times that a method with varying upper bounds* K *correctly estimated* K_0.

ensures that, asymptotically, $\widehat{\boldsymbol{\eta}}_j = \mathbf{0}$ or equivalently $\widehat{\boldsymbol{\theta}}_{\alpha(j)} = \widehat{\boldsymbol{\theta}}_{\alpha(j+1)}$ for certain indices j, and thus the effective order of $\widehat{G}_n$ becomes strictly less than the postulated upper bound K. This is how the GSF simultaneously estimates both the mixture order and the mixing measure. The choice of the tuning parameter λ_n determines the size of the penalty r_{λ_n} and thus the estimated mixture order. In Section 3, under certain regularity conditions, we prove the existence of a sequence λ_n for which $\widehat{G}_n$ has order K_0 with probability tending to one, and in Section 4 we discuss data-driven choices of λ_n. Figure 3 compares the sensitivity with respect to K of the GSF and a naive alternative that applies the penalty r_{λ_n} to all $\binom{K}{2}$ pairwise atom distances.

Examples of the penalties φ *and* r_{λ_n}. We now discuss some examples of penalty functions φ and r_{λ_n}. The functions $\varphi(\pi_1,\ldots,\pi_K) \propto -\sum_{j=1}^K \log\pi_j$ and $\varphi(\pi_1,\ldots,\pi_K) \propto \sum_{j=1}^K \pi_j^{-\iota}$ (for some $\iota > 0$) were used by Chen and Kalbfleisch (1996) in the context of distance-based methods for mixture order estimation. As seen in Supplement D.1, the former is computationally convenient for EM-type algorithms, and we use it in all demonstrative examples throughout this paper. Li, Chen and Marriott (2009) also discuss the function $\varphi(\pi_1,\ldots,\pi_K) \propto -\min_{1\leq j\leq K}\log\pi_j$ in the context of hypothesis testing for the mixture order, which is more severe (up to a constant) than the former two penalties.

Regarding r_{λ_n}, satisfying conditions (P1)–(P3) in Section 3, we consider the following three penalties. For convenience, the first two penalties are written in terms of their first derivatives with respect to η.

(i) The Smoothly Clipped Absolute Deviation (SCAD; Fan and Li (2001)),

$$r'_{\lambda_n}(\eta;\omega) \equiv r'_{\lambda_n}(\eta) = \lambda_n I\{|\eta| \leq \lambda_n\} + \frac{(a\lambda_n - |\eta|)_+}{a-1} I\{|\eta| > \lambda_n\}, \quad a > 2.$$

(ii) The Minimax Concave Penalty (MCP; Zhang (2010)),

$$r'_{\lambda_n}(\eta;\omega) \equiv r'_{\lambda_n}(\eta) = \left(\lambda_n - \frac{|\eta|}{a}\right)_+, \quad a > 1.$$

(iii) The Adaptive Lasso (ALasso; Zou (2006)),

$$r_{\lambda_n}(\eta;\omega) = \lambda_n w|\eta|.$$

The Lasso penalty $r_{\lambda_n}(\eta;\omega) = \lambda_n|\eta|$ does not satisfy all the conditions (P1)–(P3), and is further discussed in Section 3.

3. Asymptotic study. In this section, we study asymptotic properties of the GSF, beginning with preliminaries. We also introduce more notation in the sequence that it will be needed. Throughout this section, except where otherwise stated, we fix $K \geq K_0$.

3.1. *Preliminaries.* Inspired by Nguyen (2013), we analyze the convergence of mixing measures in $\mathcal{G}_K$ using the Wasserstein distance. Recall that the Wasserstein distance of order $r \geq 1$ between two mixing measures $G = \sum_{j=1}^{K} \pi_j \delta_{\boldsymbol{\theta}_j}$ and $G' = \sum_{k=1}^{K'} \pi'_k \delta_{\boldsymbol{\theta}'_k}$ is given by

$$W_r(G, G') = \left(\inf_{\mathbf{q} \in \mathcal{Q}(\boldsymbol{\pi}, \boldsymbol{\pi}')} \sum_{j=1}^{K} \sum_{k=1}^{K'} q_{jk} \|\boldsymbol{\theta}_j - \boldsymbol{\theta}'_k\|^r \right)^{\frac{1}{r}}, \tag{3.1}$$

where $\mathcal{Q}(\boldsymbol{\pi}, \boldsymbol{\pi}')$ denotes the set of joint probability distributions $\mathbf{q} = \{q_{jk} : 1 \leq j \leq K, 1 \leq k \leq K'\}$ supported on $\{1, \ldots, K\} \times \{1, \ldots, K'\}$, such that $\sum_{j=1}^{K} q_{jk} = \pi'_k$ and $\sum_{k=1}^{K'} q_{jk} = \pi_j$. We note that the ℓ_2-norm of the underlying parameter space Θ is embedded into the definition of W_r. The distance between two mixing measures is thus largely controlled by that of their atoms. The definition of W_r also bypasses the nonidentifiability issues arising from mixture label switching. These considerations make the Wasserstein distance a natural metric for the space $\mathcal{G}_K$.

A condition which arises in likelihood-based asymptotic theory of finite mixture models with unknown order, called strong identifiability (in the second-order), is defined as follows.

DEFINITION 3 (Strong Identifiability; Chen (1995), Ho and Nguyen (2016a)). The family $\mathcal{F}$ is said to be strongly identifiable (in the second-order) if $f(\mathbf{y};\boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$ for all $\mathbf{y} \in \mathcal{Y}$, and the following assumption holds for all integers $K \geq 1$.

(SI) Given distinct $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K \in \Theta$, if we have $\zeta_j \in \mathbb{R}$, $\boldsymbol{\beta}_j, \boldsymbol{\gamma}_j \in \mathbb{R}^d$, $j = 1, \ldots, K$, such that

$$\operatorname*{ess\,sup}_{\mathbf{y} \in \mathcal{Y}} \left| \sum_{j=1}^{K} \left\{ \zeta_j f(\mathbf{y};\boldsymbol{\theta}_j) + \boldsymbol{\beta}_j^{\top} \frac{\partial f(\mathbf{y};\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}} + \boldsymbol{\gamma}_j^{\top} \frac{\partial^2 f(\mathbf{y};\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\top}} \boldsymbol{\gamma}_j \right\} \right| = 0$$

then $\zeta_j = 0$, $\boldsymbol{\beta}_j = \boldsymbol{\gamma}_j = \mathbf{0} \in \mathbb{R}^d$, for all $j = 1, \ldots, K$.

For strongly identifiable mixture models, the likelihood ratio statistic with respect to the overfitted MLE $\bar{G}_n$ is stochastically bounded (Dacunha-Castelle and Gassiat (1999)). In addition, under condition (SI), upper bounds relating the Wasserstein distance between a mixing measure G and G_0 to the Hellinger distance between the corresponding densities p_G and p_{G_0} have been established by Ho and Nguyen (2016a). In particular, there exist $\delta_0, c_0 > 0$ depending on the true mixing measure G_0 such that for any $G \in \mathcal{G}_K$ satisfying $W_2(G, G_0) < \delta_0$,

$$h(p_G, p_{G_0}) \geq c_0 W_2^2(G, G_0), \tag{3.2}$$

where h denotes the Hellinger distance,

$$h(p_G, p_{G_0}) = \left(\frac{1}{2} \int (\sqrt{p_G} - \sqrt{p_{G_0}})^2 \, d\nu \right)^{\frac{1}{2}}.$$

Specific statements and discussion of these results are given in Supplement B, and are used throughout the proofs of our Theorems 1–3. Further discussion of condition (SI) is given in Section 3.3. We also require regularity conditions (A1)–(A4) on the family $\mathcal{F}$, condition (C) on the cluster ordering $\alpha_{\mathbf{t}}$, and condition (F) on the penalty φ, which we state below.

Define the family of mixture densities

$$\mathcal{P}_K = \left\{ p_G(\mathbf{y}) = \int_{\boldsymbol{\Theta}} f(\mathbf{y};\boldsymbol{\theta})\, dG(\boldsymbol{\theta}) : G \in \mathcal{G}_K \right\}. \tag{3.3}$$

Let $p_0 = p_{G_0}$ be the density of the true finite mixture model with its corresponding probability distribution P_0. Furthermore, define the empirical process

$$\nu_n(G) = \sqrt{n} \int_{\{p_0>0\}} \frac{1}{2} \log\left\{ \frac{p_G + p_0}{2p_0} \right\} d(P_n - P_0), \quad G \in \mathcal{G}_K, \tag{3.4}$$

where $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{Y}_i}$ denotes the empirical measure.

For any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top \in \Theta$, $\mathbf{y} \in \mathcal{Y}$, and $G \in \mathcal{G}_K$, let

$$U(\mathbf{y};\boldsymbol{\theta}, G) = \frac{1}{p_G(\mathbf{y})} f(\mathbf{y};\boldsymbol{\theta}), \tag{3.5}$$

$$U_{\kappa_1\ldots\kappa_M}(\mathbf{y};\boldsymbol{\theta}, G) = \frac{1}{p_G(\mathbf{y})} \frac{\partial^M f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_{\kappa_1}\cdots\partial\theta_{\kappa_M}} \tag{3.6}$$

for all $\kappa_1, \ldots, \kappa_M = 1, \ldots, d$, and any integer $M \geq 1$.

The regularity conditions are given as follows.

(A1) *Uniform Law of Large Numbers.* We have

$$\sup_{G\in\mathcal{G}_K} \frac{1}{\sqrt{n}} \left|\nu_n(G)\right| \xrightarrow{\text{a.s.}} 0, \quad \text{as } n \to \infty.$$

(A2) *Uniform Lipschitz Condition.* The kernel density f is uniformly Lipschitz up to the second order (Ho and Nguyen (2016a)). That is, there exist $C, \delta > 0$ such that for any $\boldsymbol{\gamma} \in \mathbb{R}^d$, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, and $\mathbf{y} \in \mathcal{Y}$,

$$\left| \boldsymbol{\gamma}^\top \left(\frac{\partial^2 f(\mathbf{y};\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} - \frac{\partial^2 f(\mathbf{y};\boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \right) \boldsymbol{\gamma} \right| \leq C \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_1^\delta \|\boldsymbol{\gamma}\|_2^2.$$

(A3) *Smoothness.* There exists $h_1 \in L^1(\nu)$ such that $|\log f(\mathbf{y};\boldsymbol{\theta})| \leq h_1(\mathbf{y})$ ν-almost everywhere. Moreover, the kernel density $f(\mathbf{y};\boldsymbol{\theta})$ possesses partial derivatives up to order 5 with respect to $\boldsymbol{\theta}$. For all $M \leq 5$, all $\kappa_1, \ldots, \kappa_M$, and any atom $\boldsymbol{\theta}_0$ of G_0,

$$U_{\kappa_1\ldots\kappa_M}(\cdot;\boldsymbol{\theta}_0, G_0) \in L^3(P_0).$$

There also exists $h_2 \in L^3(P_0)$ and $\epsilon > 0$ such that for all $\mathbf{y} \in \mathcal{Y}$,

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\epsilon} \left|U_{\kappa_1\ldots\kappa_5}(\mathbf{y};\boldsymbol{\theta}, G_0)\right| \leq h_2(\mathbf{y}).$$

(A4) *Uniform Boundedness.* There exist $\epsilon_1, \epsilon_2 > 0$, and $q_1, q_2 \in L^2(P_0)$ such that for all $\mathbf{y} \in \mathcal{Y}$, $|U(\mathbf{y};\boldsymbol{\theta}, G)| \leq q_1(\mathbf{y})$, and for every $\kappa_1 = 1, \ldots, d$, $|U_{\kappa_1}(\mathbf{y};\boldsymbol{\theta}, G)| \leq q_2(\mathbf{y})$, uniformly for all G such that $W_2(G, G_0) < \epsilon_1$, and for all $\boldsymbol{\theta} \in \Theta$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{0k}\| < \epsilon_2$, for some $k \in \{1, \ldots, K_0\}$.

(A1) is a standard condition required to establish consistency of nonparametric maximum likelihood estimators. A sufficient condition for (A1) to hold is that the kernel density $f(\mathbf{y};\boldsymbol{\theta})$ is continuous with respect to $\boldsymbol{\theta}$ for ν-almost every $\mathbf{y}$ (see Example 4.2.4 of van de Geer

(2000)). Under condition (A2) and the Strong Identifiability condition (SI) in Definition 3, local upper bounds relating the Wasserstein distance over $\mathcal{G}_K$ to the Hellinger distance over $\mathcal{P}_K$ in (3.3) have been established by Ho and Nguyen (2016a)—see Theorem B.2 of Supplement B. Under conditions (A3) and (SI), Dacunha-Castelle and Gassiat (1999) showed that the likelihood ratio statistic for overfitted mixtures is stochastically bounded—see Theorem B.1 of Supplement B. Condition (A4) is used to perform an order assessment for a score-type quantity in the proof of the order selection consistency of the GSF (Theorem 3).

We further assume that the cluster ordering $\alpha_{\mathbf{t}}$ satisfies the following continuity-type condition.

(C) Let $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}, \ldots, \boldsymbol{\theta}_{0K_0})$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) \in \Theta^K$. Suppose there exists a cluster partition $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{K_0}\}$ of $\boldsymbol{\theta}$ of size K_0. Let $\tau \in S_{K_0}$ be the permutation such that $(\boldsymbol{\theta}_{\alpha_{\boldsymbol{\theta}}(1)}, \ldots, \boldsymbol{\theta}_{\alpha_{\boldsymbol{\theta}}(K)}) = (\mathcal{C}_{\tau(1)}, \ldots, \mathcal{C}_{\tau(K_0)})$, as implied by the definition of cluster ordering. Then, there exists $\delta > 0$ such that, if for all $k = 1, \ldots, K_0$ and $\boldsymbol{\theta}_j \in \mathcal{C}_k$, we have $\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_{0k}\| < \delta$, then $\tau = \alpha_{\boldsymbol{\theta}_0}$.

An illustration of condition (C) is provided in Figure 4. It is easy to verify that the example of cluster ordering in (2.5) satisfies (C) whenever the minimizers therein are unique. Finally, we assume that the penalty $\varphi \equiv \varphi_n$ satisfies the following condition:

(F) $\varphi_n = a_n \phi$, where $0 < a_n = o(n)$, $a_n \not\to 0$, and $\phi : \bigcup_{j=1}^K (0, 1]^j \to \mathbb{R}_+$ is Lipschitz on any compact subset of $(0, 1]^j$, $1 \le j \le K$. Also, for all $\pi_1, \ldots, \pi_K \in (0, 1]$ and $\rho_k \ge \pi_k$, $1 \le k \le K_0 \le K$, $\phi(\pi_1, \ldots, \pi_K) \ge \phi(\rho_1, \ldots, \rho_{K_0})$, and $\phi(\pi_1, \ldots, \pi_K) \to \infty$ as $\min_j \pi_j \to 0$.

Condition (F) holds for all examples of functions φ stated in Section 2. When $r_{\lambda_n}(\eta; \omega)$ is constant with respect to η away from zero, as is the case for the SCAD and MCP, condition (P2) below implies that a_n is constant with respect to n. For technical purposes, we require a_n to diverge when r_{λ_n} is the ALasso penalty, ensuring that φ_n and $n r_{\lambda_n}$ are of comparable order. In practice, however, we notice that the GSF is hardly sensitive to the choice of a_n.

Given $G = \sum_{j=1}^K \pi_j \delta_{\boldsymbol{\theta}_j} \in \mathcal{G}_K$, we now define a choice of the weights $\omega_j \equiv \omega_j(G)$ for the penalty function r_{λ_n} in (2.6), which are random and depend on G. It should be noted that the choice of these weights is relevant for the ALasso penalty but not for the SCAD and MCP.

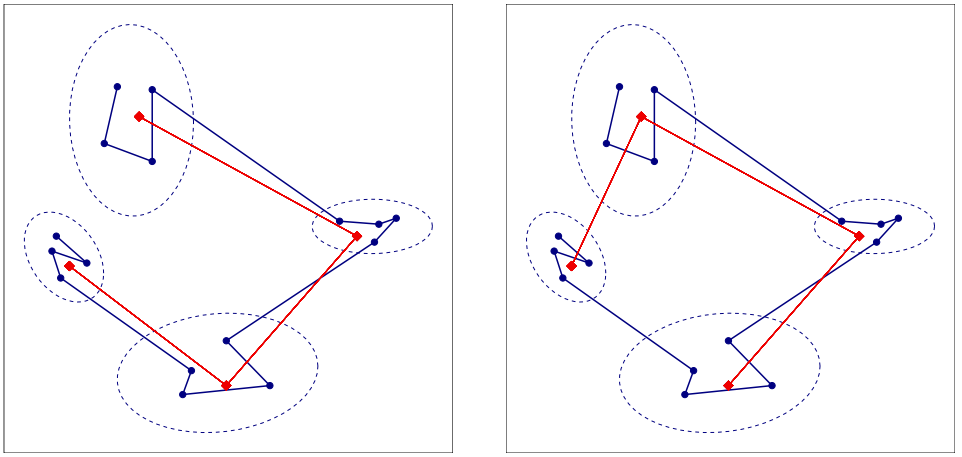

FIG. 4. *Illustration of condition* (C). *The points of* $\boldsymbol{\theta}$ *are depicted in blue* (●) *and the points of* $\boldsymbol{\theta}_0$ *are depicted in red* (◆). *The blue solid lines* (—) *denote the permutation* $\alpha_{\boldsymbol{\theta}}$, *while the red solid lines* (—) *denote the permutation* $\alpha_{\boldsymbol{\theta}_0}$. *The ellipses* (- - -) *represent a choice of cluster partition of* $\boldsymbol{\theta}$. *The choice of cluster ordering in the left plot satisfies condition* (C), *while that of the right plot does not.*

Define the estimator

$$\widetilde{G}_n = \sum_{j=1}^{K} \tilde{\pi}_j \delta_{\tilde{\boldsymbol{\theta}}_j} = \operatorname*{argmax}_{G \in \mathcal{G}_K} \{ l_n(G) - \phi(\pi_1, \ldots, \pi_K) \}, \tag{3.7}$$

and let $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_K)$. Define $\tilde{\boldsymbol{\eta}}_j = \tilde{\boldsymbol{\theta}}_{\tilde{\alpha}(j+1)} - \tilde{\boldsymbol{\theta}}_{\tilde{\alpha}(j)}$, for all $j = 1, \ldots, K-1$, where $\tilde{\alpha} \equiv \alpha_{\tilde{\boldsymbol{\theta}}}$, and recall that $\boldsymbol{\eta}_j = \boldsymbol{\theta}_{\alpha(j+1)} - \boldsymbol{\theta}_{\alpha(j)}$, where $\alpha \equiv \alpha_{\boldsymbol{\theta}}$. Let $u, v \in S_{K-1}$ be the permutations such that

$$\|\boldsymbol{\eta}_{u(1)}\| \geq \cdots \geq \|\boldsymbol{\eta}_{u(K-1)}\|, \qquad \|\tilde{\boldsymbol{\eta}}_{v(1)}\| \geq \cdots \geq \|\tilde{\boldsymbol{\eta}}_{v(K-1)}\|,$$

and set $\psi = v \circ u^{-1}$. Inspired by Zou (2006), for some $\beta > 1$, we then define

$$\omega_j = \|\tilde{\boldsymbol{\eta}}_{\psi(j)}\|^{-\beta}, \quad j = 1, \ldots, K-1. \tag{3.8}$$

Finally, we define the Voronoi diagram of the atoms $\{\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_K\}$ of $\widehat{G}_n$ in (2.7) by $\{\widehat{\mathcal{V}}_k : 1 \leq k \leq K_0\}$, where for all $k = 1, \ldots, K_0$, the sets

$$\widehat{\mathcal{V}}_k = \{\widehat{\boldsymbol{\theta}}_j : \|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_{0k}\| < \|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_{0l}\|, \forall l \neq k, 1 \leq j \leq K\}, \tag{3.9}$$

are called Voronoi cells with corresponding index sets $\widehat{\mathcal{I}}_k = \{1 \leq j \leq K : \widehat{\boldsymbol{\theta}}_j \in \widehat{\mathcal{V}}_k\}$.

3.2. *Main results*. We are now ready to state our main results. Theorem 1 below shows that $\{\widehat{\mathcal{V}}_k : 1 \leq k \leq K_0\}$ asymptotically forms a cluster partition of $\{\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_K\}$. This result, together with the rate of convergence established in Theorem 2, leads to the consistency of the GSF in estimating K_0, as stated in Theorem 3.

THEOREM 1. *Assume conditions* (SI), (A1)–(A2) *and* (F) *hold, and let the penalty function* r_{λ_n} *satisfy the following condition*:

(P1) $r_{\lambda_n}(\eta; \omega) \geq 0$ *is a nondecreasing function of* $\eta \in \mathbb{R}_+$ *which satisfies* $r_{\lambda_n}(0; \omega) = 0$ *and* $\lim_{n \to \infty} r_{\lambda_n}(\eta; \omega) = 0$, *for all* $\eta, \omega \in \mathbb{R}_+$. *Furthermore, for any fixed compact sets* $I_1, I_2 \subseteq (0, \infty)$, $r_{\lambda_n}(\cdot; \omega)$ *is convex over* I_1 *for large* n, *and* $\operatorname{diam}(n r_{\lambda_n}(I_1; I_2)) = O(a_n)$.

Then, as $n \to \infty$:

(i) $W_r(\widehat{G}_n, G_0) \to 0$, *almost surely, for all* $r \geq 1$.

Assume further that condition (A3) *holds. Then*:

(ii) $\phi(\widehat{\pi}_1, \ldots, \widehat{\pi}_K) = O_p(1)$. *In particular, for every* $k = 1, \ldots, K_0$, $\sum_{j \in \widehat{\mathcal{I}}_k} \widehat{\pi}_j = \pi_{0k} + o_p(1)$.

(iii) *For every* $1 \leq l \leq K$, *there exists a unique* $1 \leq k \leq K_0$, *such that* $\|\widehat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_{0k}\| = o_p(1)$, *thus* $\{\widehat{\mathcal{V}}_k : 1 \leq k \leq K_0\}$ *is a cluster partition of* $\{\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_K\}$, *with probability tending to one*.

Theorem 1(i) establishes the consistency of $\widehat{G}_n$ under the Wasserstein distance—a property shared by the overfitted MLE $\bar{G}_n$ (Ho and Nguyen (2016a)). This is due to the fact that, by conditions (F) and (P1), the log-likelihood function is the dominant term in L_n, in (2.6). Theorem 1(ii) implies that the estimated mixing proportions $\widehat{\pi}_j$ are stochastically bounded away from 0, which then results in Theorem 1(iii) showing that every atom of $\widehat{G}_n$ is consistent in estimating an atom of G_0. A straightforward investigation of the proof shows that this property also holds for $\widetilde{G}_n$ in (3.7), but not for the overfitted MLE $\bar{G}_n$, which may have a subset of atoms whose limit points are not among those of G_0.

When $K > K_0$, the result of Theorem 1 does not imply the consistency of $\widehat{G}_n$ in estimating K_0. The latter is achieved if the number of distinct elements of each Voronoi cell $\widehat{\mathcal{V}}_k$ is equal to one with probability tending to one, which is shown in Theorem 3 below. To

establish this result, we require an upper bound on the rate of convergence of $\widehat{G}_n$ under the Wasserstein distance. We obtain this bound by studying the rate of convergence of the density $p_{\widehat{G}_n}$ to p_0, with respect to the Hellinger distance, and appeal to inequality (3.2). van de Geer (2000) (see also Wong and Shen (1995)) established convergence rates for nonparametric maximum likelihood estimators under the Hellinger distance in terms of the bracket entropy integral

$$\mathcal{J}_B\big(\gamma, \bar{\mathcal{P}}_K^{\frac{1}{2}}(\gamma), \nu\big) = \int_0^{\gamma} \sqrt{H_B\big(u, \bar{\mathcal{P}}_K^{\frac{1}{2}}(u), \nu\big)}\, du, \quad \gamma > 0,$$

where $H_B(u, \bar{\mathcal{P}}_K^{\frac{1}{2}}(u), \nu)$ denotes the u-bracket entropy with respect to the $L^2(\nu)$ metric of the density family

$$\bar{\mathcal{P}}_K^{\frac{1}{2}}(u) = \left\{ \sqrt{\frac{p_G + p_0}{2}} : G \in \mathcal{G}_K, h\left(\frac{p_G + p_0}{2}, p_0\right) \le u \right\}, \quad u > 0.$$

In our work, however, the main difficulty in bounding $h(p_{\widehat{G}_n}, p_0)$ is the presence of the penalty r_{λ_n}. The following Theorem shows that, as $n \to \infty$, if the growth rate of r_{λ_n} away from zero, as a function of η, is carefully controlled, then $p_{\widehat{G}_n}$ achieves the same rate of convergence as the MLE $p_{\bar{G}_n}$.

THEOREM 2. *Assume the same conditions as Theorem 1, and that the cluster ordering $\alpha_{\mathbf{t}}$ satisfies condition* (C). *For a universal constant $J > 0$, assume there exists a sequence of real numbers $\gamma_n \gtrsim (\log n/n)^{1/2}$ such that for all $\gamma \ge \gamma_n$,*

$$\mathcal{J}_B\big(\gamma, \bar{\mathcal{P}}_K^{\frac{1}{2}}(\gamma), \nu\big) \le J\sqrt{n}\gamma^2. \tag{3.10}$$

Furthermore, assume r_{λ_n} satisfies the following condition:

(P2) *The restriction of r_{λ_n} to any compact subset of $\{(\eta, \omega) \subseteq \mathbb{R}^2 : \eta, \omega > 0\}$ is Lipschitz continuous in both η and ω, with Lipschitz constant $\ell_n = O(\gamma_n^{3/2}/\log n)$, and $a_n \asymp n\ell_n \vee 1$.*

Then, $h(p_{\widehat{G}_n}, p_0) = O_p(\gamma_n)$.

Gaussian mixture models are known to satisfy condition (3.10) for $\gamma_n \asymp (\log n/n)^{\frac{1}{2}}$, under certain boundedness assumptions on Θ (Ghosal and van der Vaart (2001), Genovese and Wasserman (2000)). Lemma 3.2.1 of Ho (2017) shows that (3.10) also holds for this choice of γ_n for many of the strongly identifiable density families which we discuss below. For these density families, $p_{\widehat{G}_n}$ achieves the parametric rate of convergence up to polylogarithmic factors.

Let $\widehat{K}_n$ be the order of $\widehat{G}_n$, namely the number of distinct components $\widehat{\boldsymbol{\theta}}_j$ of $\widehat{G}_n$ with nonzero mixing proportions. We now prove the consistency of $\widehat{K}_n$ in estimating K_0.

THEOREM 3. *Assume the same conditions as Theorem 2, and assume that the family $\mathcal{F}$ satisfies condition* (A4). *Suppose further that the penalty r_{λ_n} satisfies the following condition*:

(P3) *$r_{\lambda_n}(\cdot; \omega)$ is differentiable for all $\omega > 0$, and*

$$\lim_{n\to\infty} \inf\left\{ \gamma_n^{-1} \frac{\partial r_{\lambda_n}(\eta; \omega)}{\partial \eta} : 0 < \eta \le \gamma_n^{\frac{1}{2}} \log n, \omega \ge \big(\gamma_n^{\frac{\beta}{2}} \log n\big)^{-1} \right\} = \infty,$$

where γ_n is the sequence defined in Theorem 2, and $\beta > 1$ is the constant in (3.8).

Then, as $n \to \infty$:

(i) $\mathbb{P}(\widehat{K}_n = K_0) \to 1$. *In particular,* $\mathbb{P}(\bigcap_{k=1}^{K_0} \{|\widehat{\mathcal{V}}_k| = 1\}) \to 1$.
(ii) $W_1(\widehat{G}_n, G_0) = O_p(\gamma_n)$.

Condition (P3) ensures that as $n \to \infty$, r_{λ_n} grows sufficiently fast in a vanishing neighborhood of $\eta = 0$ to prevent any mixing measure of order greater than K_0 from maximizing L_n. In addition to being model selection consistent, Theorem 3 shows that for most strongly identifiable parametric families $\mathcal{F}$, $\widehat{G}_n$ is a $(\log n/n)^{1/2}$-consistent estimator of G_0. Thus, $\widehat{G}_n$ improves on the $(\log n/n)^{1/4}$ rate of convergence of the overfitted MLE $\bar{G}_n$. This fact combined with Theorem 1(iii) implies that the fitted atoms $\widehat{\boldsymbol{\theta}}_j$ are also $(\log n/n)^{1/2}$-consistent in estimating the true atoms $\boldsymbol{\theta}_{0k}$, up to relabeling.

3.3. *Remarks.* We now discuss several aspects of the GSF in regards to the (SI) condition, penalty r_λ, upper bound K, and its relation to existing approaches in Bayesian mixture modeling.

(I) *The Strong Identifiability (SI) Condition.* A wide range of univariate parametric families are known to be strongly identifiable, including most exponential families (Chen (1995), Chen, Chen and Kalbfleisch (2004)), and circular distributions (Holzmann, Munk and Stratmann (2004)). Strongly identifiable families with multidimensional parameter space include multivariate Gaussian distributions in location or scale, certain classes of Student t-distributions, as well as von Mises, Weibull, logistic and generalized Gumbel distributions (Ho and Nguyen (2016a)). In this paper, we also consider finite mixture of multinomial distributions. To establish conditions under which this family satisfies condition (SI), we begin with the following result.

PROPOSITION 1. *Consider the binomial family with known number of trials* $M \geq 1$,

$$\mathcal{F} = \left\{ f(y;\theta) = \binom{M}{y} \theta^y (1-\theta)^{M-y} : \theta \in (0,1),\, y \in \{0,\ldots,M\} \right\}. \tag{3.11}$$

Given any integer $r \geq 1$, *the condition* $(r+1)K - 1 \leq M$ *is necessary and sufficient for* $\mathcal{F}$ *to be strongly identifiable in the* r*th order (Heinrich and Kahn (2018)). That is, for any* K *distinct points* $\theta_1, \ldots, \theta_K \in (0,1)$, *and* $\beta_{jl} \in \mathbb{R}$, $j = 1, \ldots, K$, $l = 0, \ldots, r$, *if*

$$\sup_{y \in \{0,\ldots,M\}} \left| \sum_{j=1}^{K} \sum_{l=0}^{r} \beta_{jl} \frac{\partial^l f(y;\theta_j)}{\partial \theta^l} \right| = 0,$$

then $\beta_{jl} = 0$ *for every* $j = 1, \ldots, K$ *and* $l = 0, \ldots, r$.

The inequality $(r+1)K - 1 \leq M$ is comparable to the classical identifiability result of Teicher (1963), which states that binomial mixture models are identifiable with respect to their mixing measure if and only if $2K - 1 \leq M$. Using Proposition 1, we can readily establish the following result.

COROLLARY 1. *A sufficient condition for the multinomial family*

$$\mathcal{F} = \left\{ \binom{M}{y_1, \ldots, y_d} \prod_{j=1}^{d} \theta_j^{y_j} : \theta_j \in (0,1), 0 \leq y_j \leq M, \sum_{j}^{d} \theta_j = 1, \sum_{j}^{d} y_j = M \right\} \tag{3.12}$$

with known number of trials $M \geq 1$, *to satisfy condition* (SI) *is* $3K - 1 \leq M$.

(II) *The Penalty Function* r_{λ_n}. Condition (P1) is standard and is satisfied by most well-known regularization functions, including the Lasso, ALasso, SCAD and MCP, as long as $\lambda_n \to 0$, for large enough a_n, as $n \to \infty$. Conditions (P2) and (P3) are satisfied by SCAD and MCP when $\lambda_n \asymp \gamma_n^{\frac{1}{2}} \log n$. When $\gamma_n \asymp (\log n/n)^{1/2}$, it follows that λ_n decays slower than the $n^{-1/4}$ rate, contrasting the typical rate $\lambda_n \asymp n^{-1/2}$ encountered in variable selection problems for parametric regression (see, for instance, Fan and Li (2001)).

We now consider the ALasso with the weights ω_j in (3.8), which are similar to those proposed by Zou (2006) in the context of variable selection in regression. Condition (P2) implies $\lambda_n \gamma_n^{-\frac{3}{2}} \log n \to 0$, while condition (P3) implies $\lambda_n \gamma_n^{-\frac{\beta+2}{2}} \to \infty$, where β is the parameter in the weights. Thus, both conditions (P2) and (P3) are satisfied by the ALasso with the weights in (3.8) only when $\beta > 1$ and by choosing $\lambda_n \asymp \gamma_n^{3/2}/\log n$. In particular, the value $\beta = 1$ is invalid. When $\gamma_n \asymp (\log n/n)^{1/2}$, it follows that $\lambda_n \asymp n^{-3/4}(\log n)^{-1/4}$ which decays much faster than the sequence λ_n required for the SCAD and MCP discussed above. This discrepancy can be anticipated from the fact the weights ω_j corresponding to nearby atoms of $\widetilde{G}_n$ diverge. It is worth noting that the typical tuning parameter for the ALasso in parametric regression is required to satisfy $\sqrt{n}\lambda_n \to 0$ and $n^{\frac{1+\beta}{2}}\lambda_n \to \infty$, for any $\beta > 0$.

Finally, we note that the Lasso penalty $r_{\lambda_n}(\eta; \omega) = \lambda_n|\eta|$ cannot simultaneously satisfy conditions (P2) and (P3), since they would require opposing choices of λ_n. Furthermore, for this penalty, when $\Theta \subseteq \mathbb{R}$ and α is the natural ordering on the real line, that is $\theta_{\alpha(1)} \leq \cdots \leq \theta_{\alpha(K)}$, we obtain the telescoping sum

$$\lambda_n \sum_{j=1}^{K-1} |\eta_j| = \lambda_n \sum_{j=1}^{K-1} (\theta_{\alpha(j+1)} - \theta_{\alpha(j)}) = \lambda_n(\theta_{\alpha(K)} - \theta_{\alpha(1)})$$

which fails to penalize the vast majority of the overfitted components.

(III) *Choice of the Upper Bound* K. By Theorem 3, as long as the upper bound on the mixture order satisfies $K \geq K_0$, the GSF provides a consistent estimator of K_0. The following result shows the behaviour of the GSF for a misspecified bound $K < K_0$.

PROPOSITION 2. *Assume that the family* $\mathcal{F}$ *satisfies condition* (A3), *and that the mixture family* $\{p_G : G \in \mathcal{G}_K\}$ *is identifiable*, *Then*, *for any* $K < K_0$, *as* $n \to \infty$, *the GSF order estimator* $\widehat{K}_n$ *satisfies*: $\mathbb{P}(\widehat{K}_n = K) \to 1$.

Guided by the above result, if the GSF chooses the prespecified upper bound K as the estimated order, the bound is likely misspecified and larger values should also be examined. This provides a natural heuristic for choosing an upper bound K for the GSF in practice, which we further elaborate upon in Section 4.2 of the simulation study.

(IV) *Connections between the GSF and Existing Bayesian Approaches.* When $\varphi(\pi_1, \ldots, \pi_K) = (1-\gamma)\sum_{j=1}^K \log \pi_j$, for some $\gamma > 1$, the estimator $\widetilde{G}_n$ in (3.7) can be viewed as the posterior mode of the overfitted Bayesian mixture model

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K \overset{\text{i.i.d.}}{\sim} H, \tag{3.13}$$

$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\gamma, \ldots, \gamma), \qquad \mathbf{Y}_i | G = \sum_{j=1}^K \pi_j \delta_{\boldsymbol{\theta}_j} \overset{\text{i.i.d.}}{\sim} p_G, \quad i = 1, \ldots, n, \tag{3.14}$$

where H is a uniform prior on the (compact) set $\Theta \subseteq \mathbb{R}^d$. Under this setting, Rousseau and Mengersen (2011) showed that when $\gamma < d/2$, the posterior distribution of G has the effect of asymptotically emptying out redundant components of the overfitted mixture model, such that the posterior expectation of the mixing probabilities of the $(K - K_0 + 1)$ extra components

decay at the rate $n^{-1/2}$, up to polylogarithmic factors. On the other hand, if $\gamma > d/2$, two or more of the posterior atoms with nonnegligible mixing probabilities will have the tendency to approach each other. The authors discuss that the former case results in more stable behaviour of the posterior distribution. In contrast, under our setting with the choice $\gamma > 1$, Theorem 1(i) implies that all the mixing probabilities of $\widetilde{G}_n$ are bounded away from zero with probability tending to one. This behaviour matches their above setting $\gamma > d/2$, though with a generally different cutoff for γ. We argue that the GSF does not suffer from the instability described by Rousseau and Mengersen (2011) in this setting, as it proposes a simple procedure for merging nearby atoms using the second penalty r_λ in (2.6), hinging upon the notion of cluster ordering. From a Bayesian standpoint, this penalty can be viewed as replacing the i.i.d. prior H in (3.13) by the following exchangeable and non-i.i.d. prior

$$(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) \sim p_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) \propto \prod_{j=1}^{K-1} \exp\{-r_\lambda(\|\boldsymbol{\theta}_{\alpha_{\boldsymbol{\theta}}(j+1)} - \boldsymbol{\theta}_{\alpha_{\boldsymbol{\theta}}(j)}\|; \omega_j)\} \tag{3.15}$$

up to rescaling of r_λ, which places high-probability mass on nearly-overlapping atoms. On the other hand, Petralia, Rao and Dunson (2012), Xie and Xu (2020) replace H by so-called repulsive priors, which favour diverse atoms, and are typically used with $\gamma < d/2$. For example, Petralia, Rao and Dunson (2012) study the prior

$$(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) \sim p_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) \propto \prod_{j<k}^{K} \exp\{-\tau \|\boldsymbol{\theta}_j - \boldsymbol{\theta}_k\|^{-1}\}, \quad \tau > 0. \tag{3.16}$$

In contrast to the GSF, the choice $\gamma < d/2$ ensures vanishing posterior mixing probabilities corresponding to redundant components, which is further encouraged by the repulsive prior (3.16). Without a post-processing step which thresholds these mixing probabilities, however, this methods do not yield consistent order selection. It turns out that by further placing a prior on K, order consistency can be obtained (Miller and Harrison (2018), Nobile (1994)).

A distinct line of work in nonparametric Bayesian mixture modeling places a prior, such as a Dirichlet process, directly on the mixing measure G. Though the resulting posterior typically has infinitely-many atoms, consistent estimators of $K_0 < \infty$ can be obtained using post-processing techniques, such as the Merge–Truncate–Merge (MTM) method of Guha, Ho and Nguyen (2019). Both the GSF and MTM aim at reducing the overfitted mixture order by merging nearby atoms. Unlike the GSF, however, the Dirichlet process mixture's posterior may have vanishing mixing probabilities, hence a single merging stage of its atoms is insufficient to obtain an asymptotically correct order. The MTM thus also truncates such redundant components, and performs a second merging of their mixing probabilities to recover a proper mixing measure. Both the truncation and merging stages use hard-thresholding rules. We compare the two methods in our simulation study, Section 4.3.

4. Simulation study. We conduct a simulation study to assess the finite-sample performance of the GSF. We develop a modification of the EM algorithm to obtain an approximate solution to the optimization problem in (2.7). The main ingredients are the Local Linear Approximation algorithm of Zou and Li (2008) for nonconcave penalized likelihood models, and the proximal gradient method (Nesterov (2004)). Details of our numerical solution are given in Supplement D.1. The algorithm is implemented in our R package `GroupSortFuse`.

In the GSF, the tuning parameter λ regulates the order of the fitted model. Figure 1 (see also Figure 3 in Supplement E.7) shows the evolution of the parameter estimates $\widehat{\boldsymbol{\theta}}_j(\lambda)$ for a simulated dataset, over a grid of λ-values. These qualitative representations can provide insight about the order of the mixture model, for purposes of exploratory data analysis. For instance, as seen in the figures, when small values of λ lead to a significant reduction in the

postulated order K, a tighter bound on K_0 can often be obtained. In applications where a specific choice of λ is required, common techniques include v-fold Cross Validation and the BIC, applied directly to the MPLE for varying values of λ (Zhang, Li and Tsai (2010)). In our simulation, we use the BIC due to its low computational burden.

Default choices of penalties, tuning parameters, and cluster ordering. Throughout all simulations and real data analyses in this paper, including those contained in Figures 1–3, the following choices were used by default unless otherwise specified. We used the penalty $\varphi(\pi_1, \ldots, \pi_K) = (1-\gamma)\sum_{j=1}^{K} \log \pi_j$, with the constant $1-\gamma \approx -\log 20$ following the suggestion of Chen and Kalbfleisch (1996). The penalty r_λ is taken to be the SCAD by default, though we also consider simulations below which employ the MCP and ALasso penalties. For the ALasso, the weights ω_j are specified as in (3.8). The tuning parameter λ is selected using the BIC as described above. The cluster ordering $\alpha_{\boldsymbol{\theta}}$ is chosen as in (2.5). We recall that this choice does not constrain $\alpha_{\boldsymbol{\theta}}(1)$—in our simulations, we chose this value using a heuristic which ensures that $\alpha_{\boldsymbol{\theta}}$ reduces to the natural ordering on $\mathbb{R}$ in the case $d = 1$. Further numerical details are given in Supplement D.2.

4.1. *Parameter settings and order selection results.* Our simulations are based on multinomial and multivariate location-Gaussian mixture models. We compare the GSF under the SCAD (GSF-SCAD), MCP (GSF-MCP) and ALasso (GSF-ALasso) penalties to the AIC, BIC, and ICL (Biernacki, Celeux and Govaert (2000)), as implemented in the R packages `mixtools` (Benaglia et al. (2009)) and `mclust` (Fraley and Raftery (1999)). ICL performed similarly to the BIC in our multinomial simulations, but generally underperformed in our Gaussian simulations. Therefore, below we only discuss the performance of AIC and BIC.

We report the proportion of times that each method selected the correct order K_0, out of 500 replications, based on the models described below. For each simulation, we also report detailed tables in Supplement E with the number of times each method incorrectly selected orders other than K_0. We fix the upper bound $K = 12$ throughout this section. For this choice, the effective number of parameters of the mixture models hereafter is less than the smallest sample sizes considered.

Multinomial mixture models. The density function of multinomial mixture model of order K is given by

$$p_G(\mathbf{y}) = \sum_{j=1}^{K} \pi_j \binom{M}{y_1, \ldots, y_d} \prod_{l=1}^{d} \theta_{jl}^{y_l} \tag{4.1}$$

with $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jd})^\top \in (0,1)^d$, $\mathbf{y} = (y_1, \ldots, y_d)^\top \in \{1, \ldots, M\}^d$, where $\sum_{l=1}^{d} \theta_{jl} = 1$, $\sum_{l}^{d} y_l = M$. We consider 7 models with true orders $K_0 = 2, 3, \ldots, 8$, dimensions $d = 3, 4, 5$, and $M = 35, 50$ to satisfy the strong identifiability condition $3K - 1 \leq M$ described in Corollary 1. The parameter settings are given in Table 1. The results for $M = 50$ are reported in Figure 5 below. Those for $M = 35$ are similar, and are relegated to Supplement E.1. The simulation results are based on the sample sizes $n = 100, 200, 400$.

Under Model 1, all five methods selected the correct order most often, and exhibited similar performance across all the sample sizes—the results are reported in Table 1 of Supplement E.1. The results for Models 2–7 with orders $K_0 = 2, 3, 4, 5$, are plotted by percentage of correctly selected orders in Figure 5. Under Model 2, the correct order is selected most frequently by the BIC and GSF-ALasso, for all the sample sizes. Under Models 3 and 4, the GSF with all three penalties, in particular the GSF-ALasso, outperforms AIC and BIC. Under

TABLE 1
Parameter settings for the multinomial mixture Models 1–7

Model	1	2	3
$\pi_1, \boldsymbol{\theta}_1$	0.2, (0.2, 0.2, 0.2, 0.2, 0.2)	$\frac{1}{3}$, (0.2, 0.2, 0.2, 0.2, 0.2)	0.25, (0.2, 0.2, 0.6)
$\pi_2, \boldsymbol{\theta}_2$	0.8, (0.1, 0.3, 0.2, 0.1, 0.3)	$\frac{1}{3}$, (0.1, 0.3, 0.2, 0.1, 0.3)	0.25, (0.2, 0.6, 0.2)
$\pi_3, \boldsymbol{\theta}_3$		$\frac{1}{3}$, (0.3, 0.1, 0.2, 0.3, 0.1)	0.25, (0.6, 0.2, 0.2)
$\pi_4, \boldsymbol{\theta}_4$			0.25, (0.45, 0.1, 0.45)

Model	4	5	6	7
$\pi_1, \boldsymbol{\theta}_1$	0.2, (0.2, 0.2, 0.6)	$\frac{1}{6}$, (0.2, 0.2, 0.6)	$\frac{1}{7}$, (0.2, 0.2, 0.6)	0.125, (0.2, 0.2, 0.2, 0.4)
$\pi_2, \boldsymbol{\theta}_2$	0.2, (0.6, 0.2, 0.2)	$\frac{1}{6}$, (0.2, 0.6, 0.2)	$\frac{1}{7}$, (0.2, 0.6, , 2)	0.125, (0.2, 0.2, 0.4, 0.2)
$\pi_3, \boldsymbol{\theta}_3$	0.2, (0.45, 0.1, 0.45)	$\frac{1}{6}$, (0.6, 0.2, 0.2)	$\frac{1}{7}$, (0.6, 0.2, 0.2)	0.125, (0.2, 0.4, 0.2, 0.2)
$\pi_4, \boldsymbol{\theta}_4$	0.2, (0.2, 0.7, 0.1)	$\frac{1}{6}$, (0.45, 0.1, 0.45)	$\frac{1}{7}$, (0.45, 0.1, 0.45)	0.125, (0.4, 0.2, 0.2, 0.2)
$\pi_5, \boldsymbol{\theta}_5$	0.2, (0.1, 0.7, 0.2)	$\frac{1}{6}$, (0.2, 0.7, 0.1)	$\frac{1}{7}$, (0.1, 0.7, 0.2)	0.125, (0.1, 0.3, 0.1, 0.5)
$\pi_6, \boldsymbol{\theta}_6$		$\frac{1}{6}$, (0.1, 0.7, 0.2)	$\frac{1}{7}$, (0.7, 0.2, 0.1)	0.125, (0.1, 0.3, 0.5, 0.1)
$\pi_7, \boldsymbol{\theta}_7$			$\frac{1}{7}$, (0.1, 0.2, 0.7)	0.125, (0.1, 0.5, 0.3, 0.1)
$\pi_8, \boldsymbol{\theta}_8$				0.125, (0.5, 0.1, 0.3, 0.1)

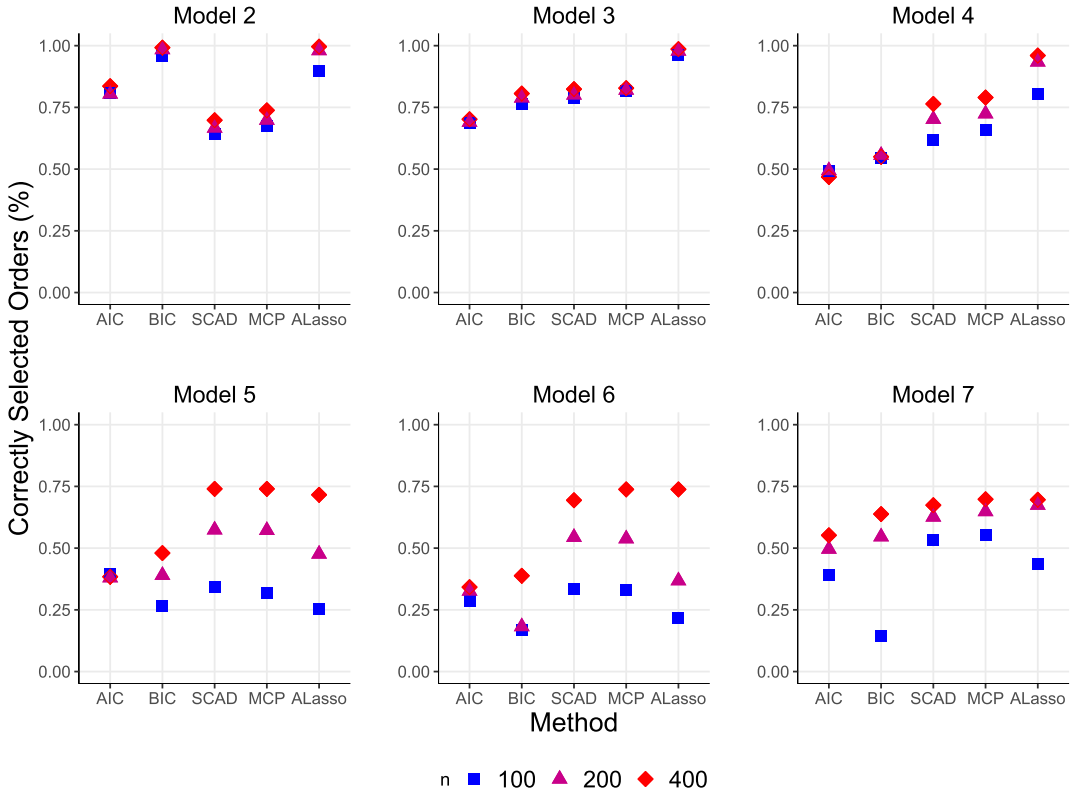

FIG. 5. *Percentage of correctly selected orders for multinomial mixture Models* 2–7.

Models 5–7, all methods selected the correct order for $n = 100$ fewer than 55% of the time. For $n = 200$, the GSF-SCAD and GSF-MCP select the correct number of components more than 55% of the time, unlike AIC and BIC. All three GSF penalties continue to outperform the other methods when $n = 400$.

Multivariate location-Gaussian mixtures with unknown covariance matrix. The density function of a multivariate Gaussian mixture model in mean, of order K, is given by

$$p_G(\mathbf{y}) = \sum_{j=1}^{K} \pi_j \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}_j)\right\},$$

where $\boldsymbol{\mu}_j \in \mathbb{R}^d$, $j = 1, \ldots, K$, and $\boldsymbol{\Sigma} = \{\sigma_{ij} : i, j = 1, \ldots, d\}$ is a positive definite $d \times d$ covariance matrix. We consider the 10 mixture models in Table 2 with true orders $K_0 = 2, 3, 4, 5$, and with dimension $d = 2, 4, 6, 8$. For each model, we consider both an identity and nonidentity covariance matrix $\boldsymbol{\Sigma}$, which is estimated as an unknown parameter. The simulation results are based on the sample sizes $n = 200, 400, 600, 800$.

The results for Models 1.a, 1.b, 3.a, 3.b, 4.a, 4.b are plotted by percentage of correctly selected orders in Figure 6 below. Detailed results for the more challenging Models 2.a, 2.b, 5.a and 5.b are reported by percentage of selected orders between $1, \ldots, K(= 12)$ in Tables 15 and 18 of Supplement E.2.

In Figure 6, under Models 1.a and 1.b with $d = 2$, all the methods selected the correct number of components most frequently for $n = 400, 600, 800$; however, the performance of all methods deteriorates in Model 1.b with nonidentity covariance matrix when $n = 200$. Under Model 3.a with $d = 4$, all methods perform similarly for $n = 400, 600, 800$, but the

TABLE 2
Parameter settings for the multivariate Gaussian mixture models

Model	σ_{ij}	$\pi_1, \boldsymbol{\mu}_1$	$\pi_2, \boldsymbol{\mu}_2$	$\pi_3, \boldsymbol{\mu}_3$	$\pi_4, \boldsymbol{\mu}_4$	$\pi_5, \boldsymbol{\mu}_5$
1.a	$I(i = j)$	$0.5, (0, 0)^\top$	$0.5, (2, 2)^\top$			
1.b	$(0.5)^{\lvert i-j \rvert}$	$0.5, (0, 0)^\top$	$0.5, (2, 2)^\top$			
2.a	$I(i = j)$	$0.25, (0, 0)^\top$	$0.25, (2, 2)^\top$	$0.25, (4, 4)^\top$	$0.25, (6, 6)^\top$	
2.b	$(0.5)^{\lvert i-j \rvert}$	$0.25, (0, 0)^\top$	$0.25, (2, 2)^\top$	$0.25, (4, 4)^\top$	$0.25, (6, 6)^\top$	
3.a	$I(i = j)$	$\frac{1}{3}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$	$\frac{1}{3}, \begin{pmatrix} 2.5 \\ 1.5 \\ 2 \\ 1.5 \end{pmatrix}$	$\frac{1}{3}, \begin{pmatrix} 1.5 \\ 3 \\ 2.75 \\ 2 \end{pmatrix}$		
3.b	$(0.5)^{\lvert i-j \rvert}$					
4.a	$I(i = j)$	$\frac{1}{5}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$	$\frac{1}{5}, \begin{pmatrix} -1.5 \\ 2.25 \\ -1 \\ 0 \\ 0.5 \\ 0.75 \end{pmatrix}$	$\frac{1}{5}, \begin{pmatrix} 0.25 \\ 1.5 \\ 0.75 \\ 0.25 \\ -0.5 \\ -1 \end{pmatrix}$	$\frac{1}{5}, \begin{pmatrix} -0.25 \\ 0.5 \\ -2.5 \\ 1.25 \\ 0.75 \\ 1.5 \end{pmatrix}$	$\frac{1}{5}, \begin{pmatrix} -1 \\ -1.5 \\ -0.25 \\ 1.75 \\ -0.5 \\ 2 \end{pmatrix}$
4.b	$(0.5)^{\lvert i-j \rvert}$					
5.a	$I(i = j)$	$\frac{1}{5}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$	$\frac{1}{5}, \begin{pmatrix} 1 \\ 1.5 \\ 0.75 \\ 2 \\ 1.5 \\ 1.75 \\ 0.5 \\ 2.5 \end{pmatrix}$	$\frac{1}{5}, \begin{pmatrix} 2 \\ 0.75 \\ 1.5 \\ 1 \\ 1.75 \\ 0.5 \\ 2.5 \\ 1.5 \end{pmatrix}$	$\frac{1}{5}, \begin{pmatrix} 1.5 \\ 2 \\ 1 \\ 0.75 \\ 2.5 \\ 1.5 \\ 1.75 \\ 0.5 \end{pmatrix}$	$\frac{1}{5}, \begin{pmatrix} 0.75 \\ 1 \\ 2 \\ 1.5 \\ 0.5 \\ 2.5 \\ 1.5 \\ 1.75 \end{pmatrix}$
5.b	$(0.5)^{\lvert i-j \rvert}$					

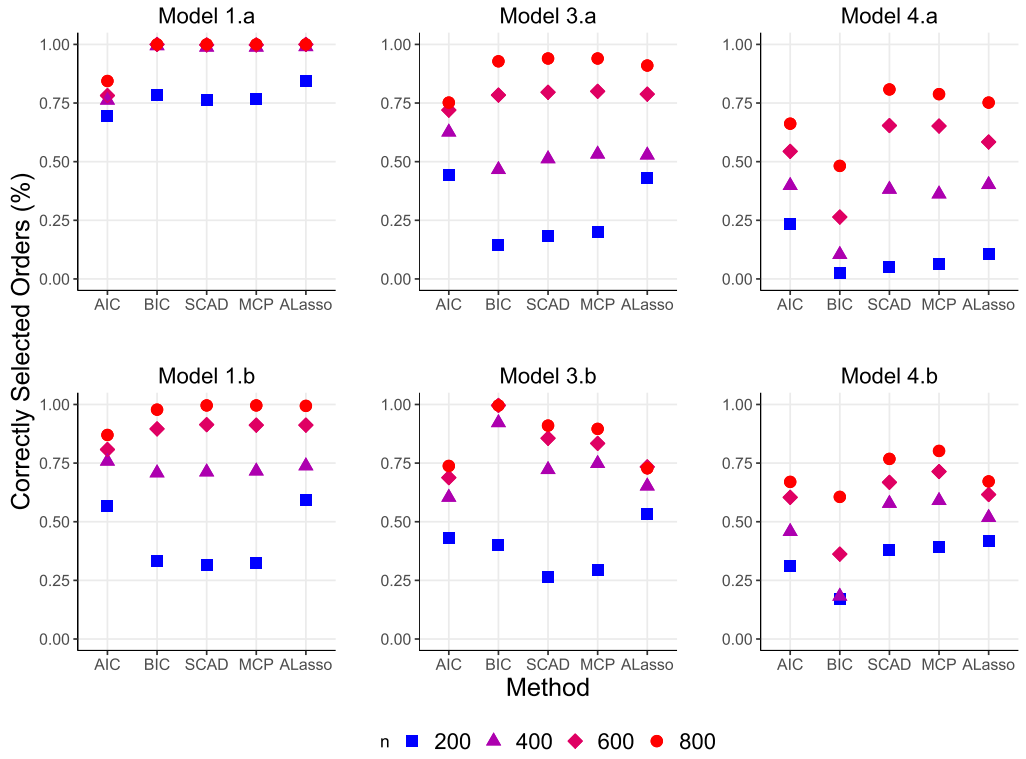

FIG. 6. *Percentage of correctly selected orders for the multivariate Gaussian mixture models.*

GSF-ALasso and the AIC outperformed the other methods for $n = 200$. Under Model 3.b, the BIC outperformed the other methods for $n = 400, 600, 800$, but the GSF-ALasso again performed the best for $n = 200$. In Models 4.a and 4.b with $d = 6$, the GSF with the three penalties outperformed AIC and BIC across all sample sizes.

From Table 15, under Model 2.a with $d = 2$ and identity covariance matrix, the BIC and the GSF with the three penalties underestimate and the AIC overestimates the true order, for sample sizes $n = 200, 400$. The three GSF penalties significantly outperform the AIC and BIC, when $n = 600, 800$. For the more difficult Model 2.b with nonidentity covariance matrix, all methods underestimate across all sample sizes considered, but the AIC selects the correct order most frequently. From Table 18, under Model 5.a, all methods apart from AIC underestimated K_0 for $n = 200, 400, 600$, and the three GSF penalties outperformed the other methods when $n = 800$. Interestingly, the performance of all methods improves for Model 5.b with nonidentity covariance matrix. Though all methods performed well for $n = 400, 600, 800$, the BIC did so the best, while the GSF-ALasso exhibited the best performance when $n = 200$.

In summary, depending on the models and sample sizes considered here, in some cases AIC or BIC exhibit the best performance, while in others the GSF based on at least one of the penalties (ALasso, SCAD, or MCP) outperforms. The universality of information criteria in almost any model selection problem is in part due to their ease of use on the investigator's part, while many other methods require specification of multiple tuning parameters. Though we defined the GSF in its most general form, our empirical investigation suggests that, other than λ and K, its tuning parameters ($\alpha_{\mathbf{t}}$, φ, ω_j, and choices therein) may not need to be tuned beyond their default choices used here. We have shown that off-the-shelf data-driven methods for selecting λ yield reasonable performance. We next discuss the choice of the bound K.

4.2. *Sensitivity analysis for the upper bound K.* In this section, we assess the sensitivity of the GSF with respect to the choice of upper bound K via simulation. Specifically, we show the behaviour of the GSF for a range of K-values which are both misspecified ($K < K_0$) and well-specified ($K \geq K_0$). In the former case, by Proposition 2, the GSF is expected to select the order K, whereas in the latter case, by Theorem 3, the GSF selects the correct K_0 with high probability.

We consider the multinomial Models 3 ($K_0 = 4$) and 5 ($K_0 = 6$) with sample size $n = 400$, and the Gaussian Models 3.a ($K_0 = 3$) and 4.a ($K_0 = 5$) with sample size $n = 600$. The results are based on 80 simulated samples from each model. For each sample, we apply the GSF-SCAD with $K = 2, \ldots, 25$, and then report the most frequently estimated order $\widehat{K}$, as well as the average estimated order over the 80 samples. The results are given in Figure 7. Detailed results are reported by percentage of selected orders with respect to the bounds $K = 2, \ldots, 25$, in Tables 19–22 of Supplement E.3.

For all four models, it can be seen that the GSF estimates the order K most frequently when $K < K_0$. In fact, it does so on every replication for $K = 1, 2$ (resp. $K = 1, 2, 3$) under multinomial Model 3 (resp. Model 5). When $K \geq K_0$, the GSF correctly estimates the order K_0 most frequently for all four models. Although the average selected order is seen to slightly deviate from K_0 as K increases (as was already noted in Figure 3), the overall behaviour of the GSF is remarkably stable with respect to the choice of K. The resulting elbow shape of the solid red lines in Figure 7 is anticipated by Theorem 3 and Proposition 2.

Guided by the above results, in applications where finite mixture models ($K_0 < \infty$) have meaningful interpretations in capturing population heterogeneity, we suggest to examine the

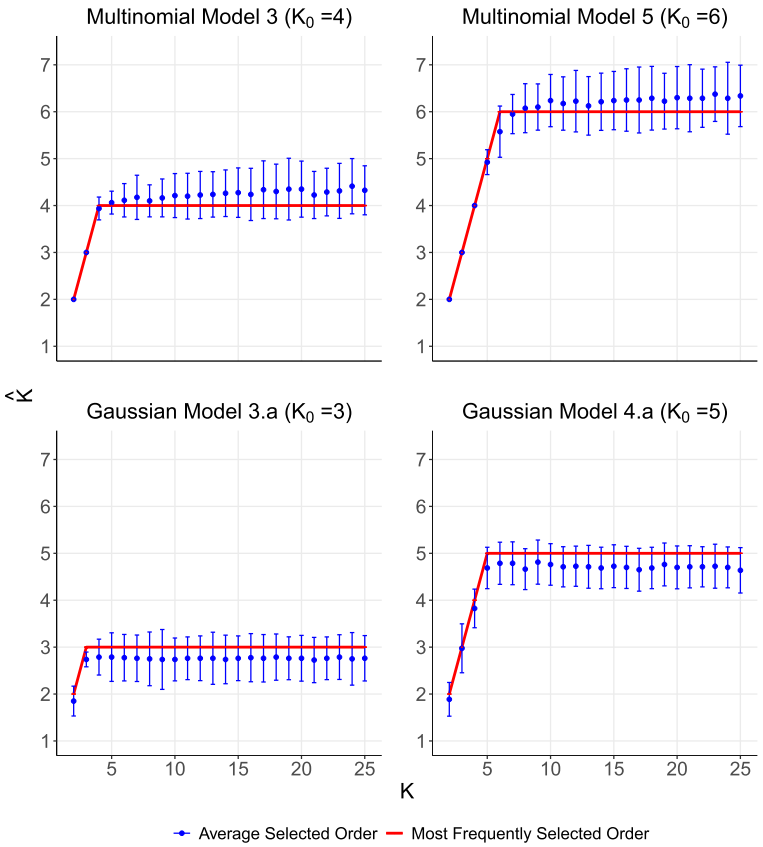

FIG. 7. *Sensitivity analysis of the GSF with respect to the upper bound K. Error bars represent one standard deviation of the fitted order.*

GSF over a range of small to large values of K. This range may be chosen with consideration of the resulting number of mixture parameters, with respect to the sample size n. An elbow-shaped scatter plot of $(K, \widehat{K})$ can shed light on a safe choice of the bound K and the selected order $\widehat{K}$. We illustrate such a strategy through the real data analysis in Section 5.

4.3. *Comparison of merging-based methods.* We now compare the GSF to alternate order selection methods which are also based on merging the components of an overfitted mixture. Our simulations are based on location-Gaussian mixture models, though unlike Section 4.1, we now treat the common covariance $\mathbf{\Sigma}$ as known. In addition to the GSF, and to the AIC/BIC which are included as benchmarks, we consider the following two methods:

– The Merge–Truncate–Merge (MTM) procedure (Guha, Ho and Nguyen (2019)) described in Section 3.3(IV), applied to posterior samples from a Dirichlet Process mixture (DPM).
– A hard-thresholding analogue of the GSF, denote by GSF-Hard, which is obtained by first computing the estimator $\widetilde{G}_n$ in (2.3), and then merging the atoms of $\widetilde{G}_n$ which fall within a sufficiently small distance $\lambda > 0$ of each other (see Algorithm 2 in Supplement D.2 for a precise description). The GSF-Hard thus replaces the penalty r_λ in the GSF with a post-hoc merging rule. By a straightforward simplification of our asymptotic theory, the GSF-Hard estimator satisfies the same properties as $\widehat{G}_n$ in Theorems 1–3.

We fit the MTM procedure using the same algorithm and parameter settings as described in Section 5 of Guha, Ho and Nguyen (2019). The truncation and (second) merging stages of the MTM require a tuning parameter $c > 0$, which plays a similar role as λ in the GSF-Hard. The authors recommend considering various choices of c in practice, though we are not aware of a method for tuning c. We therefore follow them by reporting the performance of the MTM for a range of c-values. For the GSF-Hard, we tune λ using the BIC. Further implementation details are provided in Supplement D.2.

We report the proportion of times that each method selected the correct order under Gaussian Models 1.b and 2.a in Figure 8, based on $n = 50, 100, 200, 400$. More detailed results

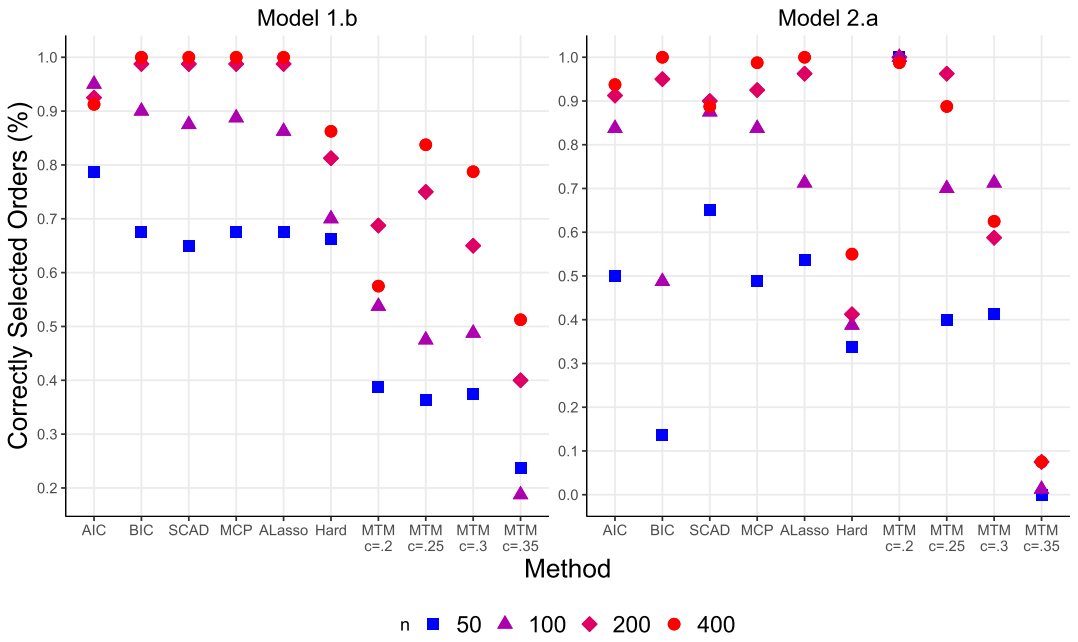

FIG. 8. *Percentage of correctly selected orders for the multivariate Gaussian models with common and known covariance matrix.*

TABLE 3
Average computational time (in seconds) per replication for the multivariate Gaussian models with common and known covariance matrix

	Model 1.b						Model 2.a					
n	AIC/ BIC	GSF-SCAD	GSF-MCP	GSF-ALasso	GSF-Hard	MTM	AIC/ BIC	GSF-SCAD	GSF-MCP	GSF-ALasso	GSF-Hard	MTM
50	23.6	1.30	1.2	5.3	3.8	2830.0	21.1	2.6	1.8	5.3	3.9	2502.6
100	29.8	2.7	2.0	9.9	5.2	7148.2	25.6	6.3	3.8	9.9	5.4	5607.2
200	38.7	6.7	4.5	19.6	6.8	25,428.3	34.9	17.2	8.6	19.6	7.0	21,008.0
400	47.6	12.4	8.5	35.8	7.6	34,911.9	45.8	43.5	16.4	35.8	9.0	20,151.2
600	54.4	24.2	15.3	49.3	8.8	51,131.0	51.3	57.8	21.8	49.3	10.0	37,535.7
800	60.0	32.2	22.8	67.1	9.9	74,185.0	56.7	103.6	39.9	67.1	10.3	57,469.7

can be found in Supplement E.4, including those for $n = 600, 800$. For each sample size, we perform 80 replications due to the computational burden associated with fitting Dirichlet Process mixture models. The MTM results are based on the posterior mode.

The AIC, BIC, and GSF under all three penalties exhibit improved performance under the current setting with fixed $\mathbf{\Sigma}$, compared to that of Section 4.1. The GSF-Hard performs reasonably under Model 1.b but markedly underperforms in Model 2.a. Regarding the MTM, we report the results under four consecutive c-values which were most favourable from a range of 16 candidate values. Under Model 1.b, the MTM under all four c-values estimates K_0 most of the time, under most sample sizes, but underperforms compared to the remaining methods. In contrast, under Model 2.a, there exists a value of c for which the MTM remarkably estimates K_0 on nearly all replications. However, the sensitivity to c is also seen to increase, which can be problematic in the absence of a data-driven tuning procedure. Finally, we recall that the MTM is based on a nonparametric Bayes procedure, while the other methods are parametric and might generally require smaller sample sizes to achieve reasonable accuracy.

We emphasize that MTM and GSF-Hard are both post-hoc procedures for reducing the order of an overfitted mixing measure G_n, which is respectively equal to a sample from the DPM posterior, or to the estimator $\widetilde{G}_n$. This contrasts the GSF, which uses continuous penalties of the parameters to simultaneously perform order selection and mixing measure estimation, and does not vary discretely with the tuning parameter λ. On the other hand, these two post-hoc procedures have the practical advantage of being computationally inexpensive wrappers on top of the well-studied estimators G_n, for which standard implementations are available. To illustrate this point, in Table 3 we report the computational time associated with the results from Figure 8, including also the sample sizes $n = 600, 800$. It can be seen that GSF-Hard is typically computable with an order of magnitude fewer seconds than the GSF under any of the three penalties. The computational times for the MTM are largely dominated by the time required to sample the DPM posterior with the implementation we used—the post-processing procedure itself accounts for a negligible fraction of this time.

5. Real data example. We consider the data analyzed by Mosimann (1962), arising from the study of the Bellas Artes pollen core from the Valley of Mexico, in view of reconstructing surrounding vegetation changes from the past. The data consists of $M = 100$ counts on the frequency of occurrence of $d = 4$ kinds of fossil pollen grains, at $n = 73$ different levels of a pollen core. A simple multinomial model provides a poor fit to this data, due to over-dispersion caused by clumped sampling. Mosimann (1962) modelled this extra variation using a Dirichlet-multinomial distribution, and Morel and Nagaraj (1993) fitted a 3-component multinomial mixture model.

We applied the GSF-SCAD with upper bounds $K = 2, \ldots, 25$. For each K, we fitted the GSF based on five different initial values for the modified EM algorithm, and selected the model with optimal tuning parameter value. For $K = 2$, the estimated order was 2 and for $K \geq 3$, the most frequently selected order was $\widehat{K} = 3$. Given the similarity of the sample size and dimension with those considered in the simulations, below we report the fitted model corresponding to the upper bound $K = 12$.

The models obtained by the GSF with the three penalties are similar—for instance, the fitted model obtained by the GSF-SCAD is

$$0.15\,\mathrm{Mult}(\widehat{\boldsymbol{\theta}}_1) + 0.25\,\mathrm{Mult}(\widehat{\boldsymbol{\theta}}_2) + 0.60\,\mathrm{Mult}(\widehat{\boldsymbol{\theta}}_3),$$

where $\mathrm{Mult}(\boldsymbol{\theta})$ denotes the multinomial distribution with 100 trials and probabilities $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_1 = (0.94, 0.01, 0.03, 0.02)^\top$, $\widehat{\boldsymbol{\theta}}_2 = (0.77, 0.02, 0.15, 0.06)^\top$ and $\widehat{\boldsymbol{\theta}}_3 = (0.87, 0.01, 0.09, 0.03)^\top$. The log-likelihood value for this estimate is -499.87. The coefficient plots produced by the tuning parameter selector for GSF-SCAD are shown in Figure 9. Interestingly, the fitted order equals 3, for all $\lambda > 0.9$ in the range considered, coinciding with the final selected order, and with the aforementioned sensitivity analysis on K.

We also ran the AIC, BIC and ICL on this data. The AIC selected six components, while the BIC and ICL selected three components. The fitted model under the latter two methods is

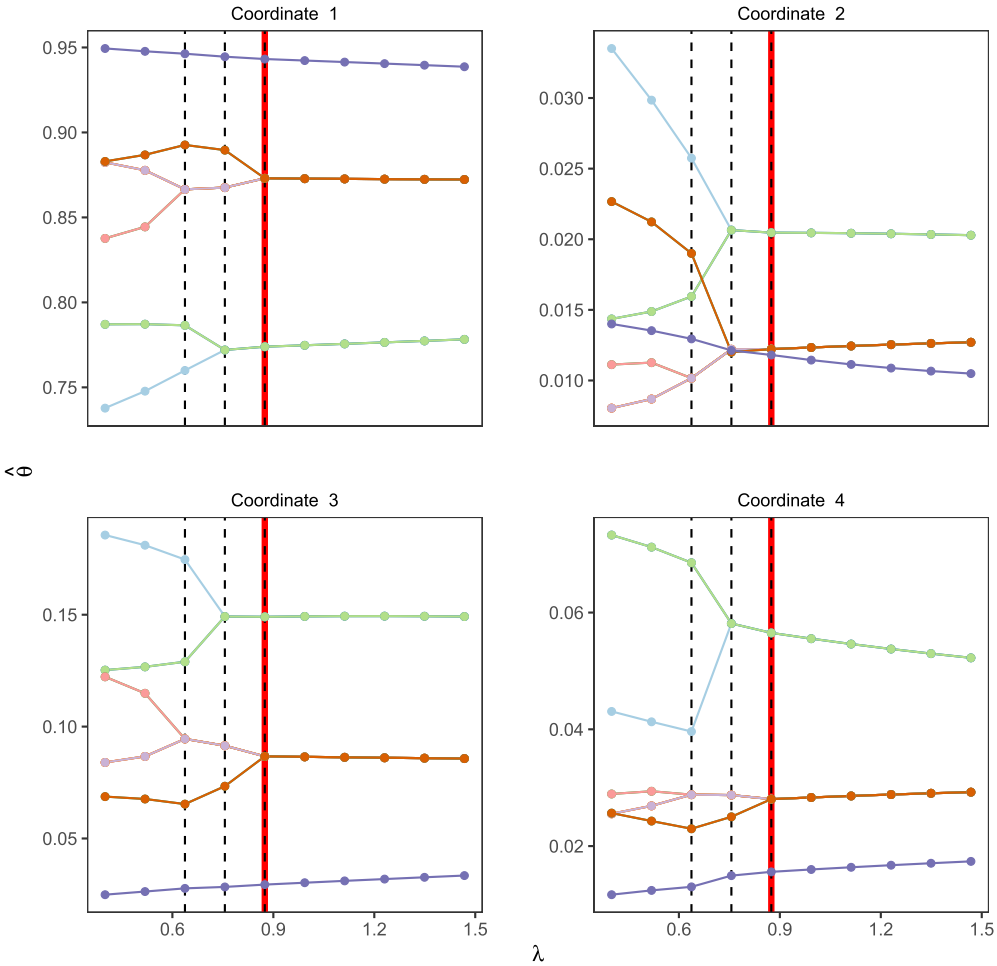

FIG. 9. *Coefficient plots for the GSF-SCAD on the pollen data. The vertical red lines indicate the selected tuning parameter.*

given by

$$0.17\,\mathrm{Mult}(\widehat{\boldsymbol{\theta}}_1) + 0.22\,\mathrm{Mult}(\widehat{\boldsymbol{\theta}}_2) + 0.61\,\mathrm{Mult}(\widehat{\boldsymbol{\theta}}_3),$$

where $\widehat{\boldsymbol{\theta}}_1 = (0.95, 0.02, 0.03, 0.01)^\top$, $\widehat{\boldsymbol{\theta}}_2 = (0.77, 0.02, 0.15, 0.07)^\top$ and $\widehat{\boldsymbol{\theta}}_3 = (0.87, 0.01, 0.09, 0.03)^\top$, with entries rounded to the nearest hundredths. The log-likelihood value for this estimate is -496.39.

6. Conclusion and discussion. In this paper, we developed the Group-Sort-Fuse (GSF) method for estimating the order of finite mixture models with a multidimensional parameter space. By starting with a conservative upper bound K on the mixture order, the GSF estimates the true order by applying two penalties to the overfitted log-likelihood, which group and fuse redundant mixture components. Under certain regularity conditions, the GSF is consistent in estimating the true order and it further provides a $\sqrt{n}$-consistent estimator for the true mixing measure (up to polylogarithmic factors). We examined its finite sample performance via thorough simulations, and illustrated its application to two real datasets, one of which is relegated to Supplement E.6.

We suggested the use of off-the-shelf methods, such as v-fold cross validation or the BIC, for selecting the tuning parameter λ_n involved in the penalty r_{λ_n}. Properties of such choices with respect to our theoretical guidelines, or alternative methods specialized to the GSF, require further investigation.

The methodology developed in this paper may be applicable to mixtures which satisfy weaker notions of strong identifiability (Ho and Nguyen (2016b)). Extending our proof techniques to such models is, however, nontrivial. In particular, bounding the log-likelihood ratio statistic for the overfitted MLE $\bar{G}_n$ (Dacunha-Castelle and Gassiat (1999)), and the penalized log-likelihood ratio for the MPLE $\widehat{G}_n$, would require new insights in the absence of (second-order) strong identifiability. Empirically, we illustrated in Section 4.1 the promising finite sample performance of the GSF under location-Gaussian mixtures with an unknown but common covariance matrix, which themselves violate condition (SI).

We have shown that the GSF achieves a near-parametric rate of convergence under the Wasserstein distance, but this rate only holds pointwise in the true mixing measure G_0. Our work leaves open the behaviour of the GSF when the true mixing measure is permitted to vary with the sample size n—indeed, the minimax risk is known to scale at a rate markedly slower than parametric (Heinrich and Kahn (2018), Wu and Yang (2020)).

We established in Proposition 2 the asymptotic behaviour of the GSF when the upper bound K is underspecified. However, our work provides no guarantees when other aspects of the mixture model $\mathcal{P}_K = \{p_G : G \in \mathcal{G}_K\}$ are misspecified, such as the kernel density family $\mathcal{F}$. We note that the recent work of Guha, Ho and Nguyen (2019) establishes the asymptotic behaviour of various Bayesian procedures under such misspecification, in terms of a suitable Kullback–Leibler projection of the true mixture distribution. While we expect the GSF to obey similar asymptotics, we are not aware of a general theory for maximum likelihood estimation under misspecification in nonconvex models such as $\mathcal{P}_K$. We leave a careful investigation of such properties to future work.

We believe that the framework developed in this paper paves the way to a new class of methods for order selection problems in other latent-variable models, such as mixture of regressions and Markov-switching autoregressive models (Frühwirth-Schnatter (2006)). Results of the type developed by Dacunha-Castelle and Gassiat (1999) in understanding large sample behaviour of likelihood ratio statistics for these models, and the recent work of Ho, Yang and Jordan (2019) in characterizing rates of convergence for parameter estimation in over-specified Gaussian mixtures of experts, may provide first steps toward such extensions. We also mention applications of the GSF procedure to nonmodel-based clustering methods,

such as the K-means algorithm. While the notion of order, or true number of clusters, is generally elusive in the absence of a model, extensions of the GSF may provide a natural heuristic for choosing the number of clusters in such methods.

Acknowledgements. We would like to thank the Editor, an Associate Editor, and two referees for their insightful comments and suggestions which significantly improved the quality of this paper. We thank Jiahua Chen for discussions related to the proof of Proposition 1, Russell Steele for bringing to our attention the multinomial dataset analyzed in Section 5, and Aritra Guha for sharing an implementation of the Merge–Truncate–Merge procedure. We also thank Sivaraman Balakrishnan and Larry Wasserman for useful discussions.

Funding. Tudor Manole was supported by the Natural Sciences and Engineering Research Council of Canada and also by the Fonds de recherche du Québec–Nature et technologies. Abbas Khalili was supported by the Natural Sciences and Engineering Research Council of Canada through Discovery Grant (NSERC RGPIN-2015-03805 and NSERC RGPIN-2020-05011), and the CRM StatLab.

SUPPLEMENTARY MATERIAL

Supplement to "Estimating the number of components in finite mixture models via the Group-Sort-Fuse procedure" (DOI: 10.1214/21-AOS2072SUPP; .pdf). Supplementary information containing proofs, numerical implementation, and additional simulation results.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723. MR0423716 https://doi.org/10.1109/tac.1974.1100705

BECHTEL, Y. C., BONAITI-PELLIE, C., POISSON, N., MAGNETTE, J. and BECHTEL, P. R. (1993). A population and family study of N-acetyltransferase using caffeine urinary metabolites. *Clin. Pharmacol. Ther.* **54** 134–141.

BENAGLIA, T., CHAUVEAU, D., HUNTER, D. and YOUNG, D. (2009). mixtools: An R package for analyzing finite mixture models. *J. Stat. Softw.* **32** 1–29.

BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 719–725.

BOSCH-DOMÈNECH, A., MONTALVO, J. G., NAGEL, R. and SATORRA, A. (2010). A finite mixture analysis of beauty-contest data using generalized beta distributions. *Exp. Econm.* **13** 461–475.

CHEN, J. H. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** 221–233. MR1331665 https://doi.org/10.1214/aos/1176324464

CHEN, H. and CHEN, J. (2003). Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statist. Sinica* **13** 351–365. MR1977730

CHEN, H., CHEN, J. and KALBFLEISCH, J. D. (2004). Testing for a finite mixture model with two components. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 95–115. MR2035761 https://doi.org/10.1111/j.1467-9868.2004.00434.x

CHEN, J. and KALBFLEISCH, J. D. (1996). Penalized minimum-distance estimates in finite mixture models. *Canad. J. Statist.* **24** 167–175. MR1406173 https://doi.org/10.2307/3315623

CHEN, J. and KHALILI, A. (2008). Order selection in finite mixture models with a nonsmooth penalty. *J. Amer. Statist. Assoc.* **103** 1674–1683. MR2722574 https://doi.org/10.1198/016214508000001075

CHEN, J. and LI, P. (2009). Hypothesis test for normal mixture models: The EM approach. *Ann. Statist.* **37** 2523–2542. MR2543701 https://doi.org/10.1214/08-AOS651

DACUNHA-CASTELLE, D. and GASSIAT, E. (1999). Testing the order of a model using locally conic parametrization: Population mixtures and stationary ARMA processes. *Ann. Statist.* **27** 1178–1209. MR1740115 https://doi.org/10.1214/aos/1017938921

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DRTON, M. and PLUMMER, M. (2017). A Bayesian information criterion for singular models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 323–380. MR3611750 https://doi.org/10.1111/rssb.12187

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

FRALEY, C. and RAFTERY, A. E. (1999). MCLUST: Software for model-based cluster analysis. *J. Classification* **16** 297–306.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York.

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. *Springer Series in Statistics*. Springer, New York. MR2265601

GENOVESE, C. R. and WASSERMAN, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28** 1105–1127. MR1810921 https://doi.org/10.1214/aos/1015956709

GHOSAL, S. and VAN DER VAART, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263. MR1873329 https://doi.org/10.1214/aos/1013203453

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3617773

GUHA, A., HO, N. and NGUYEN, X. (2019). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli* Available at arXiv:1901.05078.

HATHAWAY, R. J. (1986). A constrained EM algorithm for univariate normal mixtures. *J. Stat. Comput. Simul.* **23** 211–230.

HEINRICH, P. and KAHN, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *Ann. Statist.* **46** 2844–2870. MR3851757 https://doi.org/10.1214/17-AOS1641

HO, N. P. M. (2017). *Parameter Estimation and Multilevel Clustering with Mixture and Hierarchical Models*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—University of Michigan. MR3809858

HO, N. and NGUYEN, X. (2016a). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electron. J. Stat.* **10** 271–307. MR3466183 https://doi.org/10.1214/16-EJS1105

HO, N. and NGUYEN, X. (2016b). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Ann. Statist.* **44** 2726–2755. MR3576559 https://doi.org/10.1214/16-AOS1444

HO, N., NGUYEN, X. and RITOV, Y. (2020). Robust estimation of mixing measures in finite mixture models. *Bernoulli* **26** 828–857. MR4058353 https://doi.org/10.3150/18-BEJ1087

HO, N., YANG, C.-Y. and JORDAN, M. I. (2019). Convergence rates for Gaussian mixtures of experts. arXiv preprint. Available at arXiv:1907.04377.

HOLZMANN, H., MUNK, A. and STRATMANN, B. (2004). Identifiability of finite mixtures—with applications to circular distributions. *Sankhyā* **66** 440–449. MR2108200

HUNG, Y., WANG, Y., ZARNITSYNA, V., ZHU, C. and WU, C. F. J. (2013). Hidden Markov models with applications in cell adhesion experiments. *J. Amer. Statist. Assoc.* **108** 1469–1479. MR3174722 https://doi.org/10.1080/01621459.2013.836973

ISHWARAN, H., JAMES, L. F. and SUN, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Statist. Assoc.* **96** 1316–1332. MR1946579 https://doi.org/10.1198/016214501753382255

JAMES, L. F., PRIEBE, C. E. and MARCHETTE, D. J. (2001). Consistent estimation of mixture complexity. *Ann. Statist.* **29** 1281–1296. MR1873331 https://doi.org/10.1214/aos/1013203454

KERIBIN, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A* **62** 49–66. MR1769735

LEROUX, B. G. (1992). Consistent estimation of a mixing distribution. *Ann. Statist.* **20** 1350–1360. MR1186253 https://doi.org/10.1214/aos/1176348772

LI, P. and CHEN, J. (2010). Testing the order of a finite mixture. *J. Amer. Statist. Assoc.* **105** 1084–1092. MR2752604 https://doi.org/10.1198/jasa.2010.tm09032

LI, P., CHEN, J. and MARRIOTT, P. (2009). Non-finite Fisher information and homogeneity: An EM approach. *Biometrika* **96** 411–426. MR2507152 https://doi.org/10.1093/biomet/asp011

LIU, X. and SHAO, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Ann. Statist.* **31** 807–832. MR1994731 https://doi.org/10.1214/aos/1056562463

MANOLE, T. and KHALILI, A. (2021). Supplement to "Estimating the Number of Components in Finite Mixture Models via the Group-Sort-Fuse Procedure." https://doi.org/10.1214/21-AOS2072SUPP

MCLACHLAN, G. J. (1987). On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **36** 318–324.

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. Wiley Interscience, New York. MR1789474 https://doi.org/10.1002/0471721182

MILLER, J. W. and HARRISON, M. T. (2018). Mixture models with a prior on the number of components. *J. Amer. Statist. Assoc.* **113** 340–356. MR3803469 https://doi.org/10.1080/01621459.2016.1255636

MOREL, J. G. and NAGARAJ, N. K. (1993). A finite mixture distribution for modelling multinomial extra variation. *Biometrika* **80** 363–371. MR1243510 https://doi.org/10.1093/biomet/80.2.363

MORRIS, T. H., RICHMOND, D. R. and GRIMSHAW, S. D. (1996). Orientation of dinosaur bones in riverine environments: Insights into sedimentary dynamics and taphonomy. In *The Continental Jurassic* 521–530. Museum of Northern Arizona, Flagstaff, AZ.

MOSIMANN, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* **49** 65–82. MR0143299 https://doi.org/10.1093/biomet/49.1-2.65

NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization. Applied Optimization* **87**. Kluwer Academic, Boston, MA. MR2142598 https://doi.org/10.1007/978-1-4419-8853-9

NGUYEN, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.* **41** 370–400. MR3059422 https://doi.org/10.1214/12-AOS1065

NOBILE, A. (1994). *Bayesian Analysis of Finite Mixture Distributions*. ProQuest LLC, Pittsburgh, PA. Thesis (Ph.D.)—Carnegie Mellon University. MR2692049

PETRALIA, F., RAO, V. and DUNSON, D. (2012). Repulsive mixtures. In *Advances in Neural Information Processing Systems* (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) **25** 1889–1897. Curran Associates, Red Hook, NY.

RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 731–792. MR1483213 https://doi.org/10.1111/1467-9868.00095

ROUSSEAU, J. and MENGERSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 689–710. MR2867454 https://doi.org/10.1111/j.1467-9868.2011.00781.x

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

STEPHENS, M. (2000). Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.* **28** 40–74. MR1762903 https://doi.org/10.1214/aos/1016120364

TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Stat.* **34** 1265–1269. MR0155376 https://doi.org/10.1214/aoms/1177703862

THOMPSON, T. J., SMITH, P. J. and BOYLE, J. P. (1998). Finite mixture models with concomitant information: Assessing diagnostic criteria for diabetes. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **47** 393–404.

VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge.

WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339–362. MR1332570 https://doi.org/10.1214/aos/1176324524

WOO, M.-J. and SRIRAM, T. N. (2006). Robust estimation of mixture complexity. *J. Amer. Statist. Assoc.* **101** 1475–1486. MR2279473 https://doi.org/10.1198/016214506000000555

WU, Y. and YANG, P. (2020). Optimal estimation of Gaussian mixtures via denoised method of moments. *Ann. Statist.* **48** 1981–2007. MR4134783 https://doi.org/10.1214/19-AOS1873

XIE, F. and XU, Y. (2020). Bayesian repulsive Gaussian mixture model. *J. Amer. Statist. Assoc.* **115** 187–203. MR4078456 https://doi.org/10.1080/01621459.2018.1537918

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHANG, Y., LI, R. and TSAI, C.-L. (2010). Regularization parameter selections via generalized information criterion. *J. Amer. Statist. Assoc.* **105** 312–323. MR2656055 https://doi.org/10.1198/jasa.2009.tm08013

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443 https://doi.org/10.1214/009053607000000802