ON CHOOSING MIXTURE COMPONENTS VIA NON-LOCAL PRIORS

JAIRO FÚQUENE, MARK STEEL, DAVID ROSSELL

Department of Statistics, University of Warwick, UK

Abstract

Choosing the number of components remains a central but elusive challenge in mixture models. Traditional model selection criteria can fail to enforce parsimony or result in poorly separated components of limited practical use. Non-local priors (NLPs) are a family of distributions that encourage parsimony by enforcing a separation between the models under consideration. We formalize NLPs in the context of mixtures and show how they lead to extra parsimony and well-separated components that have non-negligible weight, hence interpretable as distinct subpopulations. We derive tractable expressions and suggest default prior settings aimed at detecting multi-modal densities. We also give a theoretical characterization of the sparsity induced by NLPs and propose easily implementable algorithms to obtain the integrated likelihood and parameter estimates. Although the framework is generic we fully develop the multivariate Normal mixture case based on a novel family of exchangeable moment priors. The proposal is illustrated using simulated and real data sets. Our results show a serious lack of sensitivity of the Bayesian information criterion (BIC) and insufficient parsimony of local prior and shrinkage counterparts to our formulation, which strikes a reasonable balance between power and parsimony.

KEYWORDS: Mixture models, Non-local priors, Model selection, Bayes factor.

1. INTRODUCTION

Mixture models have many applications, *e.g.* in human genetics (Schork et al., 1996), false discovery rate control (Efron, 2008), signal deconvolution (West and Turner, 1994), density estimation (Escobar and West, 1995) and cluster analysis (*e.g.* Fraley and Raftery (2002); Baudry et al. (2010)). An extensive treatment of mixtures is provided in Frühwirth-Schnatter (2006) and Mengersen et al. (2011). In spite of their fundamental role in statistics, their irregular nature (*e.g.* multi-modal unbounded likelihood, non-identifiability) means that choosing the number of components remains an elusive problem both in the Bayesian and frequentist paradigms. As discussed below, formal criteria often lead to too many or too few components, requiring the data analyst to perform some ad-hoc post-processing. Our main contributions are proposing the use of non-local priors (NLPs) to

Corresponding author: David Rossell (rosselldavid@gmail.com).

select the number of components, characterizing the properties of the associated inference and developing computationally tractable expressions and algorithms.

Consider a sample $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$ of independent observations from a finite mixture where $\mathbf{y}_i \in \Re^p$ arises from the density

$$p(\mathbf{y}_i|\boldsymbol{\vartheta}_k, \mathcal{M}_k) = \sum_{j=1}^{k} \eta_j p(\mathbf{y}_i|\boldsymbol{\theta}_j). \tag{1.1}$$

The component densities $p(\mathbf{y}|\boldsymbol{\theta})$ are indexed by a parameter $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\eta} = (\eta_1, ..., \eta_k) \in \mathcal{E}_k$ are the weights, $\mathcal{E}_k$ the unit simplex and $\mathcal{M}_k$ denotes the model with k components. Our main goal is to infer k. For simplicity we assume that there is an upper bound K such that $k \in \{1, \ldots, K\}$, *e.g.* given by subject-matter or practical considerations, but when this is not the case the proposed Bayesian framework remains valid by setting a prior distribution on k with support on the natural numbers.

The parameter $\boldsymbol{\vartheta}_k = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k, \boldsymbol{\eta})$ takes values in $\Theta_k = \Theta^k \times \mathcal{E}_k$. As an example, a multivariate Normal mixture is characterised by $p(\mathbf{y} \mid \boldsymbol{\theta}_j) = N(\mathbf{y} \mid \boldsymbol{\mu}_j, \Sigma_j)$ with $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \Sigma_j)$ where $\boldsymbol{\mu}_j \in \Re^p$ is the mean and Σ_j is the covariance matrix of component j. Throughout, we assume that $\mathbf{y}$ are truly generated by $p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})$ for some $k^* \in \{1, \ldots, K\}$, $\boldsymbol{\vartheta}^*_{k^*} \in \Theta_{k^*}$.

Mixture models suffer from a lack of identifiability that plays a fundamental role both in estimation and model selection. This non-identifiability can be either caused by the invariance of the likelihood (1.1) to relabeling the components or to posing overfitted models that could be equivalently defined with $\mathcal{M}_{k'}$ for $k' < k$, *e.g.* setting $\eta_j = 0$ or $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$ for some $i \neq j$. The former issue is known as label switching and is due to there being $k!$ equivalent ways of rearranging the components giving rise to the same $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$. Although it creates some technical difficulties, label switching by itself does not seriously hamper inference, *e.g.* if $k = k^*$ then the maximum likelihood estimator (MLE) is consistent and asymptotically normal as $n \to \infty$ in the quotient topology (Redner, 1981), and from a Bayesian perspective the integrated likelihood behaves asymptotically as in regular models (Crawford, 1994). Non-identifiability due to overfitting has more serious consequences, *e.g.* estimates for $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ are consistent under mild conditions (Ghosal and der Vaart, 2001) but the MLE and posterior mode of $\boldsymbol{\vartheta}_k$ can behave erratically (Leroux, 1992; Rousseau and Mengersen, 2011). In addition, as we now discuss frequentist and Bayesian tests to assess the adequacy of $\mathcal{M}_k$ can behave insatisfactorily.

The literature on criteria for choosing k is too large to be covered here, instead the reader is referred to Fraley and Raftery (2002); Baudry et al. (2010) and Richardson and Green (1997). We review a few model-based criteria, as these are most closely related to our proposal and can be generically applied to any probability model. From a frequentist perspective the likelihood ratio test statistic between $\mathcal{M}_k$ and $\mathcal{M}_{k+1}$ may diverge as $n \to \infty$ when data truly arise from $\mathcal{M}_k$ (see for example Liu and Chao (2004)), unless restrictions on the parameter space or likelihood penalties are imposed (Gosh and Sen (1985); Chen and Li (2009)). A model-based alternative to testing is to consider selection criteria such as the Bayesian information criterion (BIC). Although the formal

BIC justification as an approximation to the Bayesian evidence (Schwarz, 1978) is not valid for overfitted mixtures (Frühwirth-Schnatter, 2006), it is nevertheless often adopted as a useful criterion to choose k (*e.g.* Fraley and Raftery (2002)). One issue with the BIC is that due to its evaluating $p(\boldsymbol{\vartheta} \mid \mathbf{y}, \mathcal{M}_k)$ at a single MLE it ignores that there are $k!$ maxima, causing a loss of sensitivity to detect truly present components. More importantly, as discussed below, the measure of dimensionality $\dim(\Theta_k)$ used by the BIC is overly stringent for overfitted mixtures (Watanabe, 2013), again decreasing power. These theoretical observations align with the empirical results we present here.

From a Bayesian perspective, model selection is usually based on the posterior probability $P(\mathcal{M}_k|\mathbf{y}) = p(\mathbf{y}|\mathcal{M}_k)P(\mathcal{M}_k)/p(\mathbf{y})$, where $P(\mathcal{M}_k)$ is the prior probability assigned to $\mathcal{M}_k$,

$$p(\mathbf{y}|\mathcal{M}_k) = \int_{\Theta_k} p(\mathbf{y}|\boldsymbol{\vartheta}_k, \mathcal{M}_k)p(\boldsymbol{\vartheta}_k|\mathcal{M}_k)d\boldsymbol{\vartheta}_k \tag{1.2}$$

is the integrated (or marginal) likelihood and $p(\boldsymbol{\vartheta}_k|\mathcal{M}_k)$ a prior distribution under $\mathcal{M}_k$. Equivalently, one may use Bayes factors $B_{k',k}(\mathbf{y}) = p(\mathbf{y}|\mathcal{M}_{k'})/p(\mathbf{y}|\mathcal{M}_k)$ to compare any $\mathcal{M}_{k'}, \mathcal{M}_k$. A common argument for (1.2) is that it automatically penalizes overly complex models, however this parsimony enforcement is not as strong as one would ideally wish. To gain intuition, for regular models with fixed $p_k = \dim(\Theta_k)$ one obtains

$$\log p(\mathbf{y} \mid \mathcal{M}_k) = \log p(\mathbf{y} \mid \hat{\vartheta}_k, \mathcal{M}_k) - \frac{p_k}{2}\log(O_p(n)) + O_p(1) \tag{1.3}$$

as $n \to \infty$ (Dawid, 1999). This implies that $B_{k^*,k}(\mathbf{y})$ grows exponentially as $n \to \infty$ when $\mathcal{M}_{k*} \not\subset \mathcal{M}_k$ but is only $O_p(n^{-(p_k - p_{k^*})/2})$ when $\mathcal{M}_{k*} \subset \mathcal{M}_k$. That is, overfitted models are only penalized at a slow polynomial rate. Key to the current manuscript, Johnson and Rossell (2010) showed that either faster polynomial or quasi-exponential rates are obtained by letting $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ be a NLP (defined below). Expression (1.3) remains valid for many mixtures with $k \leq k^*$ (*e.g.* including Normal mixtures, Crawford (1994)), however this is no longer the case for $k > k^*$. Using algebraic statistics, Watanabe (2009, 2013) gave expressions analogous to (1.3) for such overfitted mixtures, where $p_k/2$ is substituted with a rational number $\lambda \in [p_{k^*}/2, p_k/2]$ called the *real canonical threshold* and the remainder term is $O_p(\log\log n)$ instead of $O_p(1)$. The exact value of λ is complicated but the implication is that p_k in (1.3) imposes an overly stringent penalty when $k > k^*$ and, more importantly, that $B_{k,k^*}(\mathbf{y}) = O_p(n^{-(\lambda - p_{k^*}/2)})$. That is, akin to regular models overfitted mixtures are penalized only at a slow polynomial rate. These results align with those in Chambaz and Rousseau (2008). Denoting the posterior mode by $\tilde{k} = \text{argmax}_{k\in\{1,\ldots,K\}} P(\mathcal{M}_k \mid \mathbf{y})$, these authors found that the frequentist probability $P_{\vartheta^*}(\tilde{k} < k^*) = O(e^{-an})$ but in contrast $P_{\vartheta^*}(\tilde{k} > k^*) = O((\log n)^b/\sqrt{n})$ for some constants $a, b > 0$, again implying that overfitted mixtures are not sufficiently penalized. We emphasize that these results apply to a wide class of priors (including most commonly used ones), but not to the NLPs class proposed in this paper, for which faster rates are attained.

An interesting alternative to considering a series of $k = 1, \ldots, K$ is to set a single large k and induce posterior shrinkage. Rousseau and Mengersen (2011) showed that the prior

$p(\boldsymbol{\eta}_k|\mathcal{M}_k)$ strongly influences posterior inference on $\boldsymbol{\vartheta}_k$ when $k > k^*$. Under $p(\boldsymbol{\eta}_k|\mathcal{M}_k) = \text{Dirichlet}(\boldsymbol{\eta}_k; q_1, ..., q_k)$ with $\max_j q_j < d/2$ where $d = \dim(\Theta)$ the posterior for η_j collapses to 0 for redundant components, whereas if $\min_j q_j > d/2$ then at least two components tend to be identical with non-zero weights. That is, the posterior shrinkage induced by $q_j < d/2$ helps discard spurious components. Gelman et al. (2013) set as default $q_1 = ... = q_k = 1/k$, although Havre et al. (2015) argued that this leads to insufficient shrinkage. Instead, Havre et al. (2015) proposed setting smaller q_j and counting the number of empty components (with no assigned observations) at each Markov Chain Monte Carlo (MCMC) iteration to estimate k^*. Petralia et al. (2012) showed that faster shrinkage may be obtained by considering repulsive priors that assign vanishing density to $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$ for $i \neq j$, a framework extended using determinantal point processes by Affandi et al. (2013) and Xu et al. (2016). A recent approach by Malsiner-Walli et al. (2015) resembling repulsive mixtures is to encourage components merging into groups at a first hierarchical level and to separate groups at the second level. Interestingly, as we shall see, repulsive mixtures can be seen as shrinkage counterparts to our framework.

In spite of their usefulness, shrinkage priors also bear limitations. One issue is that for regular models their induced posterior shrinkage is strong but ultimately inferior to that from assigning positive prior probability to submodels (Bhattacharya et al., 2015), which when set adequately leads to optimal minimax concentration in linear models (Castillo et al., 2015). Based on the $n^{-\frac{1}{2}}$ posterior shrinkage in Rousseau and Mengersen (2011) we hypothesize that a similar result may hold for mixtures, although their irregular nature requires a separate study. On the practical side, inference may be sensitive to q_j or the chosen k, and it may be hard to set a threshold for selecting the non-zero η_j. Finally, shrinkage priors do not lead to posterior model probabilities, whereas here we adhere to formal Bayesian model selection. Building upon Johnson and Rossell (2010, 2012), our key idea is to induce additional parsimony by defining $p(\vartheta_k \mid \mathcal{M}_k)$ to be a NLP.

Definition 1. *Let $\mathcal{M}_k$ be a mixture with k components as in (1.1). A continuous prior density $p(\boldsymbol{\vartheta}_k|\mathcal{M}_k)$ is a NLP iff*

$$\lim_{\vartheta_k \to \mathbf{t}} p(\boldsymbol{\vartheta}_k|\mathcal{M}_k) = 0$$

for any $\mathbf{t} \in \Theta_k$ such that $p(\mathbf{y}|\mathbf{t}, \mathcal{M}_k) = p(\mathbf{y}|\boldsymbol{\vartheta}_{k'}, \mathcal{M}_{k'})$ for some $\boldsymbol{\vartheta}_{k'} \in \Theta_{k'}$, $k' < k$.

Intuitively a NLP under $\mathcal{M}_k$ assigns vanishing density to $\boldsymbol{\vartheta}_k$ that define a mixture with redundant components, *e.g.* $\eta_j = 0$ or $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$ for $i \neq j$ (see Section 2). A local prior (LP) is any $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ not satisfying Definition 1. Beyond their philosophical appeal in establishing a probabilistic separation between the models under consideration, Johnson and Rossell (2010) showed that for asymptotically normal models NLPs penalize models with spurious parameters at a faster rate than that in (1.3), the specific rate depending on the speed at which $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ converges to 0. Johnson and Rossell (2012) found that NLPs are a necessary and sufficient condition to achieve the strong consistency $P(\mathcal{M}_{k^*} \mid \mathbf{y}) \xrightarrow{P} 1$ in certain high-dimensional linear regression with $O(n^a)$ predictors and $a < 1$, whereas Shin et al. (2015) showed a similar result with $O(e^{n^a})$

predictors and certain NLP classes. These authors also observed an improved model selection performance relative to several popular penalized likelihood methods.

Here we investigate theoretical, computational and practical issues to enable the use of NLPs in mixtures. In Section 2 we formulate a general NLP class, show how it leads to stronger parsimony than LPs, propose a particular choice leading to tractable expressions and consider default elicitation for prior parameters. In Section 3 we propose computational schemes for model selection and parameter estimation. In Section 4 we illustrate the performance of BIC, LPs and NLPs in Normal mixtures using synthetic and real examples. Conclusions are presented in Section 5. All proofs are in the Supplementary material.

2. Prior formulation and parsimony properties

A NLP under $\mathcal{M}_k$ assigns vanishing density to any $\boldsymbol{\vartheta}_k$ such that (1.1) is equivalent to a mixture with $k' < k$ components. A necessary condition is to avoid vanishing ($\eta_j = 0$) and overlapping components ($\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$) but for this to also be a sufficient condition one needs to require *generic identifiability*. Definition 2 is adapted from Leroux (1992).

Definition 2. *Let* $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) = \sum_{j=1}^{k} \eta_j p(\mathbf{y} \mid \boldsymbol{\theta}_j)$ *and* $p(\mathbf{y} \mid \widetilde{\boldsymbol{\vartheta}}_{k'}, \mathcal{M}_{k'}) = \sum_{j=1}^{k'} \tilde{\eta}_j p(\mathbf{y} \mid \widetilde{\boldsymbol{\theta}}_j)$ *be two mixtures as in (1.1). Assume that* $\eta_j > 0, \tilde{\eta}_j > 0$ *for all* j *and that* $\boldsymbol{\theta}_j \neq \boldsymbol{\theta}_{j'}$, $\widetilde{\boldsymbol{\theta}}_j \neq \widetilde{\boldsymbol{\theta}}_{j'}$ *for all* $j \neq j'$. *The class* $p(\mathbf{y} \mid \boldsymbol{\theta})$ *defines a generically identifiable mixture if* $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) = p(\mathbf{y} \mid \widetilde{\boldsymbol{\vartheta}}_{k'}, \mathcal{M}_{k'})$ *for almost every* $\mathbf{y}$ *implies that* $k = k'$ *and* $\boldsymbol{\vartheta}_k = \tilde{\boldsymbol{\vartheta}}_{\Psi(k')}$ *for some permutation* $\Psi(k')$ *of the component labels in* $\mathcal{M}_{k'}$.

That is, assuming that all components have non-zero weights and distinct parameters a generically identifiable mixture is uniquely identified by its parameters up to label permutations. Teicher (1963) showed that mixtures of univariate Normal, Exponential and Gamma distributions are generically identifiable. Yakowitz and Spragins (1968) extended the result to several multivariate distributions, including the Normal case. Throughout we assume $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ to be generically identifiable in which case $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ defines a NLP if and only if $\lim\ p(\boldsymbol{\vartheta}_k|\mathcal{M}_k) = 0$ as (i) $\eta_j \to 0$ for any $j = 1, ..., k$ or (ii) $\boldsymbol{\theta}_i \to \boldsymbol{\theta}_j$ for any $i \neq j$. Let $d(\boldsymbol{\vartheta}_k)$ be a continuous penalty function converging to 0 under (i) or (ii), then a general NLP class is defined by

$$p(\boldsymbol{\vartheta}_k|\mathcal{M}_k) = d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k), \tag{2.1}$$

where $p^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k)$ is an arbitrary LP with the only restriction that $p(\boldsymbol{\vartheta}_k|\mathcal{M}_k)$ is proper. For simplicity we consider $p^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k) = p^L(\boldsymbol{\theta}|\mathcal{M}_k)p^L(\boldsymbol{\eta}|\mathcal{M}_k)$ and $d_\vartheta(\boldsymbol{\vartheta}_k) = d_\theta(\boldsymbol{\theta})d_\eta(\boldsymbol{\eta})$, where

$$d_\theta(\boldsymbol{\theta}) = \frac{1}{C_k} \left(\prod_{1 \leq i < j \leq k} d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \right), \tag{2.2}$$

$C_k = \int \left(\prod_{1 \leq i < j \leq k} d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \right) p^L(\boldsymbol{\theta}|\mathcal{M}_k) d\boldsymbol{\theta}$ is the prior normalization constant for $\boldsymbol{\theta}$ and $d_\eta(\boldsymbol{\eta}) \propto \prod_{i=1}^{k} \eta_j^r$ with $r > 0$. In general evaluating C_k can be cumbersome but below we give closed-form expressions for a specific $d_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $p^L(\boldsymbol{\theta} \mid \mathcal{M}_k)$. We set the symmetric

Dirichlet $p(\boldsymbol{\eta} \mid \mathcal{M}_k) = \text{Dir}(\boldsymbol{\eta}|q) \propto d_\eta(\boldsymbol{\eta})\text{Dir}(\boldsymbol{\eta}|q-r)$, where $q > 1$ to satisfy (i) above and $r \in [q-1, q)$. Summarizing, we focus attention on

$$p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = d_\theta(\boldsymbol{\theta}) p^L(\boldsymbol{\theta} \mid \mathcal{M}_k)\text{Dir}(\boldsymbol{\eta} \mid q), \tag{2.3}$$

where $q > 1$ and $d_\theta(\boldsymbol{\theta})$ is as in (2.2). The specific form of $d_\theta(\boldsymbol{\theta})$ depends on the model under consideration. We focus on the case where $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_i)$ where $\boldsymbol{\mu}_i$ is a location parameter and Σ_i is a scale matrix. Adapting earlier proposals for variable selection one may define MOM penalties (Johnson and Rossell, 2010) $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)/g$ where A is a symmetric positive-definite matrix, or alternatively eMOM penalties (Rossell et al., 2013) $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \exp\{-g/(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\}$ where g is a prior dispersion parameter, also adopted by Petralia et al. (2012) in the context of repulsive mixtures. The main difference between MOM and eMOM is that the latter induce a stronger model separation that under regularity conditions gives faster sparsity rates. However, the empirical results in Johnson and Rossell (2010, 2012) and Rossell and Telesca (2016) suggest that by setting g adequately both MOM and eMOM are often equally satisfactory. We now offer some theoretical results for both penalties, but in our implementations we focus on the MOM for the practical reason that C_k has a closed-form, hence avoiding a doubly-intractable problem where one needs to determine both prior and posterior normalizing constants. Note that C_k is guaranteed to be finite for eMOM penalties as $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \leq 1$. We defer discussion of prior elicitation to Section 2.3.

2.1. **Parsimony enforcement.** We show that NLPs induce extra parsimony via the penalty term $d_\vartheta(\boldsymbol{\vartheta}_k)$, which affects specifically overfitted mixtures. We first lay out technical conditions for the result to hold. Recall that k^* is the true number of components and $\boldsymbol{\vartheta}^*_{k^*}$ the true parameter value. Let $p^*_k(\mathbf{y})$ be the density $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ with $\boldsymbol{\vartheta}_k \in \Theta_k$ that minimises Kullback-Leibler (KL) divergence with respect to $p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})$. When $k \leq k^*$ for generically identifiable mixtures $p^*_k(\mathbf{y})$ is defined by a unique parameter denoted $\boldsymbol{\vartheta}^*_k \in \Theta_k$ (up to label permutations), whereas when $k > k^*$ there are multiple minimizers giving $p^*_k(\mathbf{y}) = p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})$. $p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ denotes a LP and $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ a NLP as in (2.1), whereas $P^L(\cdot \mid \mathbf{y}, \mathcal{M}_k)$ and $E^L(\cdot \mid \mathbf{y}, \mathcal{M}_k)$ are the posterior probability and expectation under $p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$.

NLP parsimony conditions

B1 L_1 *consistency.* For all $\epsilon > 0$ as $n \to \infty$

$$P^L\left(\int |p(\mathbf{z} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) - p^*_k(\mathbf{z})|\, d\mathbf{z} > \epsilon \mid \mathbf{y}, \mathcal{M}_k\right) \to 0$$

in probability with respect to $p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})$.

B2 *Continuity.* $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ is a continuous function in $\boldsymbol{\vartheta}_k$.

B3 *Penalty boundedness.* There is a constant c_k such that $d_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}_k) \leq c_k$ for all $\boldsymbol{\vartheta}_k$. Alternatively if $p(\boldsymbol{\vartheta}_k|\mathcal{M}_k)$ is the MOM prior and $k > k^*$ then there exist finite $\epsilon, U > 0$ such that

$$\lim_{n\to\infty} P\left(E^L\left[\exp\left\{\frac{1}{2g}\sum_{j=1}^{k} \boldsymbol{\mu}_j' A^{-1} \boldsymbol{\mu}_j \frac{\epsilon}{1+\epsilon}\right\} \mid \mathbf{y}, \mathcal{M}_k\right] < U\right) = 1.$$

Condition B1 amounts to posterior L_1 consistency of $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ to the data-generating truth when $k \geq k^*$ and to the KL-optimal density when $k < k^*$. This is a milder version of Condition A1 in Rousseau and Mengersen (2011) where rather than fixed ϵ one has $\epsilon = \sqrt{\log n}/\sqrt{n}$. See the discussion therein and Ghosal and der Vaart (2001) for results on finite Normal mixtures, Rousseau (2007) for Beta mixtures and Ghosal and Van Der Vaart (2007) for infinite Normal mixtures. For strictly positive $p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) > 0$ condition B1 is intimately connected to MLE consistency (Ghosal, 2002), proven for fairly general mixtures by Redner (1981) for $k \leq k^*$ and by Leroux (1992) for $k > k^*$. Condition B2 holds when the component-specific $p(\mathbf{y} \mid \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, as in the vast majority of common models. B3 is trivially satisfied when NLPs are defined using bounded penalties (*e.g.* eMOM), whereas for the MOM we require the technical condition that the given posterior exponential moment is bounded in probability when $k > k^*$. To gain intuition, B3 requires that under the posterior distribution $p^L(\boldsymbol{\mu}|\mathcal{M}_k, \mathbf{y})$ none of the elements in $\boldsymbol{\mu}$ diverges to infinity.

Theorem 1 below states that $d_\vartheta(\boldsymbol{\vartheta}_k)$ imposes a complexity penalty concentrating around 0 when $k > k^*$ and a positive constant for $k \leq k^*$. Part (i) applies to any model, Part (ii) only requires B1-B3 and Part (iii) holds under the mild conditions A1-A4 in Rousseau and Mengersen (2011) (Supplementary Section 6), hence the result is not restricted to Normal mixtures.

Theorem 1. Let $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ be a generically identifiable mixture, $p(\mathbf{y}|\mathcal{M}_k)$ and $p^L(\mathbf{y} \mid \mathcal{M}_k)$ the integrated likelihoods under $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ and $p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$. Then

(i) $p(\mathbf{y}|\mathcal{M}_k) = p^L(\mathbf{y}|\mathcal{M}_k) E^L\left(d_\vartheta(\boldsymbol{\vartheta}_k)|\mathbf{y}\right)$, where

$$E^L\left(d_\vartheta(\boldsymbol{\vartheta}_k|\mathbf{y})\right) = \int d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k|\mathbf{y}, \mathcal{M}_k) d\boldsymbol{\vartheta}_k.$$

(ii) If B1-B2 are satisfied then as $n \to \infty$

$$P^L\left(|d_\vartheta(\boldsymbol{\vartheta}_k) - d_k^*| > \epsilon \mid \mathbf{y}, \mathcal{M}_k\right) \to 0$$

where $d_k^* = 0$ for $k > k^*$ and $d_k^* = d_\vartheta(\boldsymbol{\vartheta}_k^*)$ for $k \leq k^*$.

If B3 also holds then $E^L\left(d_\vartheta(\boldsymbol{\vartheta}_k)|\mathbf{y}\right) \xrightarrow{P} d_k^*$.

(iii) Let $k > k^*$ and $p(\boldsymbol{\vartheta}_k|\mathcal{M}_k) \propto d_\theta(\boldsymbol{\theta}) p^L(\boldsymbol{\theta}|\mathcal{M}_k) \mathrm{Dir}(\boldsymbol{\eta}; q)$, where $d_\theta(\boldsymbol{\theta}) \leq c_k$ for some finite c_k and $q > 1$. If A1-A4 in Rousseau and Mengersen (2011) hold for $p^L(\boldsymbol{\theta}|\mathcal{M}_k)$ then for all $\epsilon > 0$ and all $\delta \in (0, \dim(\Theta)/2)$ there exists a finite $\tilde{c}_k > 0$ such that

$$P^L\left(d_\vartheta(\boldsymbol{\vartheta}_k) > \tilde{c}_k n^{-\frac{k-k^*}{2}(q-\delta)+\epsilon}\right) \to 0$$

in probability.

Part (i) extends Theorem 1 in Rossell and Telesca (2016) to mixtures and shows that $p(\mathbf{y}|\mathcal{M}_k)$ differs from $p^L(\mathbf{y}|\mathcal{M}_k)$ by a term $E^L\left(d_\vartheta(\boldsymbol{\vartheta}_k)|\mathbf{y}\right)$ that intuitively should converge to 0 for overfitted models. Part (i) also eases computation as $E^L(d_\vartheta(\boldsymbol{\vartheta}_k)|\mathbf{y})$ can be estimated from standard MCMC output from $p^L(\boldsymbol{\vartheta}_k|\mathbf{y}, \mathcal{M}_k)$, as we exploit in Section 3. Part (ii) confirms that the posterior of $d_\vartheta(\boldsymbol{\vartheta}_k)$ under $p^L(\boldsymbol{\vartheta}_k|\mathbf{y}, \mathcal{M}_k)$ concentrates around 0 for overfitted models and a finite constant otherwise, and that its

expectation also converges. Part (iii) states that for overfitted models this concentration rate is essentially $n^{-(k-k^*)q/2}$, leading to an accelerated sparsity-inducing Bayes factor $B_{k,k^*}(\mathbf{y}) = E^L(O_p(n^{-(k-k^*)q/2}))B^L_{k,k^*}(\mathbf{y})$, where as discussed earlier the LP-based $B^L_{k,k^*}(\mathbf{y}) = O_p(n^{-(\lambda-p_{k^*}/2)})$ for some $\lambda \in [p_{k^*}/2, p_k/2]$ under the conditions in Watanabe (2013). For instance, one may set q such that $(k-k^*)q/2 = \lambda - p_{k^*}/2$ so that $B_{k,k^*}(\mathbf{y})$ converges to 0 at twice the rate for $B^L_{k,k^*}(\mathbf{y})$. As λ is unknown in general a conservative option is to take its upper bound $\lambda = p_k/2$, so that $q = (p_k - p_{k^*})/(k-k^*)$ is set to the number of parameters per component. See Section 2.3 for further discussion on prior elicitation.

2.2. **MOM prior for location parameters.** We propose $d_\theta(\boldsymbol{\theta})$ leading to closed-form C_k in (2.2) in the common case where $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_i)$, $\boldsymbol{\mu}_i$ is a location parameter and Σ_i a positive-definite scale matrix. We define the MOM-Inverse Wishart (MOM-IW) prior

$$p(\boldsymbol{\theta}_k|\mathcal{M}_k) = \frac{1}{C_k} \prod_{1\le i<j\le k} \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A_\Sigma^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{g} \prod_{j=1}^{k} N\left(\boldsymbol{\mu}_j \mid \mathbf{0}, gA_\Sigma\right) \mathrm{IW}(\Sigma_j \mid \nu, S), \tag{2.4}$$

where A_Σ^{-1} is an arbitrary symmetric positive definite matrix and (g, ν, S) are given prior hyperparameters. A trivial choice is $A_\Sigma^{-1} = I$, but it has the inconvenience of not being invariant to changes in scale of $\mathbf{y}$. To address this in our examples we use $A_\Sigma^{-1} = \frac{1}{k}\sum_{j=1}^k \Sigma_j^{-1}$, which is symmetric and positive-definite and is related to the L_2 distance between Normal distributions. In the particular case where $\Sigma_1 = \ldots = \Sigma_k = \Sigma$, a parsimonious model sometimes considered to borrow information across components, clearly $A_\Sigma = \Sigma$. In our model-fitting algorithms and examples we consider both the equal and unequal covariance cases. We remark that in the latter case (2.4) defines a NLP that penalizes $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j$ even when $\Sigma_i \neq \Sigma_j$. We do not view this as problematic, given that in most applications the interest is to identify components with well-separated locations.

To compute C_k we need to deal with a non-trivial expectation of a product of quadratic forms. Corollary 1 gives a recursive formula for C_k building upon Kan (2006), and Corollary 2 provides a simpler expression for the univariate case.

Corollary 1. The normalization constant in (2.4) is

$$C_k = \frac{1}{s!} \sum_{\upsilon_{(1,2)}=0}^{1} \ldots \sum_{\upsilon_{(k-1,k)}=0}^{1} (-1)^{\sum_{i,j}^{s} \upsilon_{(i,j)}} \mathcal{Q}_s(B_\upsilon), \tag{2.5}$$

where $s = \binom{k}{2}$, $\mathcal{Q}_s(B_\upsilon) = s!2^s d_s(B_\upsilon)$, $d_s(B_\upsilon) = \frac{1}{2s}\sum_{i=1}^s tr(B_\upsilon^i) d_{s-i}(B_\upsilon)$, $d_0(B_\upsilon) = 1$ and B_υ is a $pk \times pk$ matrix with element (l,m) given by

$$\begin{cases} b_{ll} = \dfrac{1}{2}(k-1) - \sum_{i<j} \upsilon_{(i,j)}, & l = 1 + p(i-1), \ldots, pi \\ b_{lm} = b_{ml} = -\dfrac{1}{2} + \sum_{i<j} \upsilon_{(i,j)}, & (l,m) = (1+p(i-1), 1+p(j-1)), \ldots, (pi, pj) \end{cases}$$

where $i \neq j$, $i = 1, \ldots, k$, $j = 1, \ldots, k$ and $b_{lm} = 0$ otherwise.

Corollary 2. If $p = 1$ then

$$C_k = \prod_{j=1}^{k} \frac{\Gamma(jt+1)}{\Gamma(t+1)}. \tag{2.6}$$

2.3. **Prior elicitation.** The most critical aspect in a NLP is its induced separation between components, driven by g and q in our formulation (2.3)-(2.4). Below we propose defaults that can be used in the absence of a priori knowledge, and whenever the latter is available we naturally recommend to include it in the prior.

We first discuss g, for concreteness focusing on the Normal $p(\mathbf{y} \mid \boldsymbol{\theta}_j) = N(\mathbf{y} \mid \boldsymbol{\mu}_j, \Sigma_j)$ case, although the ideas remain applicable to other mixtures. In many applications the interest is in finding clearly-separated components that facilitate interpreting the data-generating process in terms of distinct sub-populations, thus we assign low prior probability to $\boldsymbol{\vartheta}_k$ resulting in unimodal $p(\mathbf{y}|\boldsymbol{\vartheta}_k, \mathcal{M}_k)$. The number of modes in Normal mixtures depends on non-trivial combinations of parameter values (Ray and Lindsay, 2005), but fortunately when focusing on a pair of components simpler conditions are available. Specifically, for $k = 2$ and focusing on $\eta_1 = \eta_2 = 0.5$ and $\Sigma_1 = \Sigma_2$ to keep the elicitation simple the mixture is bimodal when $\kappa = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 4$. Thus we set g such that $P(\kappa < 4|g, \mathcal{M}_k) = 0.1$ or 0.05, say, which is trivial given that the prior on κ implied by (2.4) is $p(\kappa|g, \mathcal{M}_k) = \text{Gamma}(\kappa|p/2 + 1, 1/4g)$. For instance, in a univariate Normal mixture $g = 5.68$ gives $P(\kappa < 4 \mid g, \mathcal{M}_2) = 0.05$. The left panel in Figure 1 illustrates the associated prior and for comparison the middle panel shows a Normal prior with $g^L = 11.56$, which also assigns $P^L(\kappa < 4 \mid g, \mathcal{M}_2) = 0.05$. Based on simulation and sensitivity analyses (Supplementary Section 9) we found $P(\kappa < 4 \mid g, \mathcal{M}_k) = 0.05$ to be slightly preferable to 0.1 for balancing parsimony and detection power. We remark that alternative strategies to set g arise from using different measures of separation between components, *e.g.* within/between sums of squares instead of multimodality (see Malsiner-Walli et al. (2015)).

Regarding q, as discussed earlier $q > 1$ is required for (2.3) to define a NLP. One possible option is to set $q = 3$ so that $p(\boldsymbol{\eta} \mid \mathcal{M}_k) \propto \prod_{j=1}^{k} \eta_j^2$ induces a quadratic penalty comparable to the MOM prior (2.4). Alternatively, from the discussion following Proposition 1, setting $q = (p_k - p_{k^*})/(k - k^*)$, the number of additional parameters associated with adding one component, provides a conservative choice aimed at (at least) doubling the Bayes factor sparsity rate of the underlying LP. For instance, for Normal mixtures with common covariance this leads to $q = p + 1$, and under unequal covariances to $q = p + 0.5p(p + 1) + 1$. These are the values we used in our examples with $p = 1$ or $p = 2$ (Section 4), but we remark that for larger p this default choice may lead to an overly informative prior on $\boldsymbol{\eta}$. Based on our experience $q \in [2, 4]$ (Supplementary Section 9) gives fairly robust results and satisfactory sparsity enforcement, thus larger values do not seem warranted. The prior distribution on the remaining parameters, which may be thought of as nuisance parameters, will typically reduce to a standard form for which defaults are already available. As an illustration for Normal mixtures we set $p(\Sigma_1, \ldots, \Sigma_k \mid \mathcal{M}_k) = \prod_{j=1}^{k} \text{IW}(\Sigma_j; \nu, S)$. We follow the recommendation in Hathaway (1985) that eigenvalues of $\Sigma_i \Sigma_j^{-1}$ for any $i \neq j$ should be bounded away from 0 to prevent

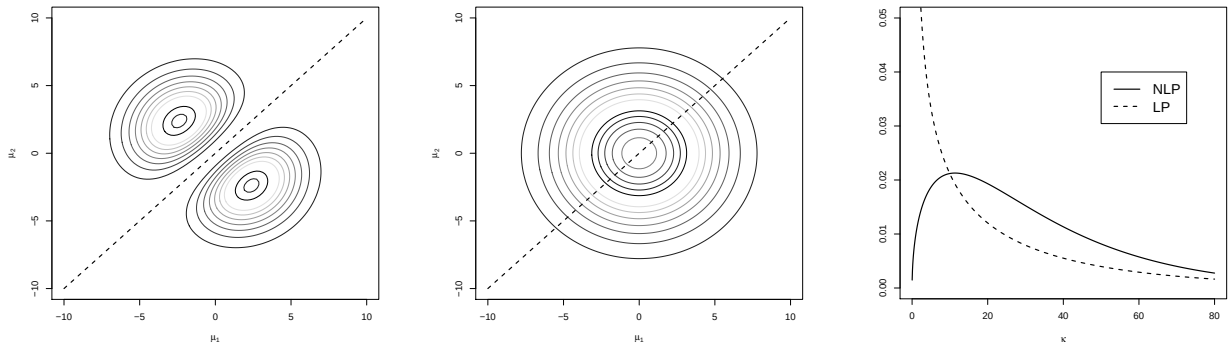


FIGURE 1. Default NLP $p(\mu_1, \mu_2|\sigma^2 = 1, g = 5.68, \mathcal{M}_2)$ (left), Default LP $p^L(\mu_1, \mu_2|\sigma^2 = 1, g^L = 11.56, \mathcal{M}_2)$ (middle) and the implied prior densities on κ (right).

the posterior from becoming unbounded, which is achieved if $\nu \geq 5 + (p-1)$ (Frühwirth-Schnatter (2006), Chapter 6). Throughout we assume that variables in the observed data are standardized to have mean 0 and variance 1 and set a default $S = I(5 + (p-1))^{-1}$, so that $E(\Sigma_j^{-1}) = I$.

3. Computational algorithms

Computation for mixtures can be challenging, and potentially more so when embarking upon a non-standard formulation such as ours. Fortunately, it is possible to estimate the integrated likelihood $p(\mathbf{y} \mid \mathcal{M}_k)$ and posterior means $E(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$ for arbitrary mixtures using direct extensions of existing MCMC algorithms (Section 3.1) and, although our main interest is to infer k, to obtain posterior modes via an Expectation-Maximimation (EM) algorithm (Dempster et al. (1977)), as described in Section 3.2.

3.1. Approximation of $p(\mathbf{y} \mid \mathcal{M}_k)$.

Theorem 1(i) suggests the estimator

$$\hat{p}(\mathbf{y}|\mathcal{M}_k) = \hat{p}^L(\mathbf{y}|\mathcal{M}_k)\frac{1}{T}\sum_{t=1}^{T}\omega(\boldsymbol{\vartheta}_k^{(t)}), \tag{3.1}$$

where $\omega(\boldsymbol{\vartheta}_k) = p(\boldsymbol{\vartheta}_k|\mathcal{M}_k)/p^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k)$ and $p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ is an arbitrary LP conveniently chosen so that MCMC algorithms to sample $\boldsymbol{\vartheta}_k^{(t)} \sim p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k) \propto p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ are readily available. For $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ as in (2.3)-(2.4) we used

$$p^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k) = \mathrm{Dir}(\boldsymbol{\eta} \mid q)\prod_{j=1}^{k} N\left(\boldsymbol{\mu}_j \mid \mathbf{0}, g\Sigma_j\right)\mathrm{IW}(\Sigma_j \mid \nu, S), \tag{3.2}$$

which gives

$$\omega(\boldsymbol{\vartheta}_k) = \frac{1}{C_k} \prod_{1 \leq i < j \leq k} \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A_\Sigma^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{g} \prod_{j=1}^{k} \frac{N(\boldsymbol{\mu}_j \mid \mathbf{0}, g A_\Sigma)}{N(\boldsymbol{\mu}_j \mid \mathbf{0}, g \Sigma_j)}. \tag{3.3}$$

We remark that $\omega(\boldsymbol{\vartheta}_k^{(t)})$ also act as particle weights, *i.e.* $\boldsymbol{\vartheta}_k^{(t)} \sim p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$ weighted by $\omega(\boldsymbol{\vartheta}_k^{(t)})$ give a posterior sample from the target $p(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$. For alternative strategies to sample $\boldsymbol{\vartheta}_k$ under NLPs see Rossell and Telesca (2016); Petralia et al. (2012); Affandi et al. (2013) or Xu et al. (2016).

Since $p^L(\mathbf{y}|\mathcal{M}_k)$ is not available in closed form we resort to computational approximations. One option is to use trans-dimensional Markov chain Monte Carlo as in Richardson and Green (1997). Marin and Robert (2008) argue that this may be hard to calibrate and require a large number of simulations to explore each model adequately, so that when K is small exploring each model separately may be preferable. However, approximating $p^L(\mathbf{y}|\mathcal{M}_k)$ directly can also be challenging due to label switching. Chib (1995) presented an approach based on the Gibbs output and Neal (1999) showed that this estimator fails when the Gibbs sampler does not explore the $k!$ modes. A correction presented in Berkhof et al. (2003) and revisited in Marin and Robert (2008) uses the estimator

$$\hat{p}^L(\mathbf{y}|\mathcal{M}_k) = \frac{p(\mathbf{y}|\hat{\boldsymbol{\vartheta}}_k, \mathcal{M}_k) p^L(\hat{\boldsymbol{\vartheta}}_k|\mathcal{M}_k)}{\hat{p}^L(\hat{\boldsymbol{\vartheta}}_k|\mathbf{y}, \mathcal{M}_k)} \tag{3.4}$$

where $\hat{\boldsymbol{\vartheta}}_k$ is *e.g.* the MLE or posterior mode of $\boldsymbol{\vartheta}_k$. The numerator in (3.4) simply requires evaluating the likelihood and prior at $\hat{\boldsymbol{\vartheta}}_k$. To evaluate the denominator we note that under exchangeable $p^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k)$ the posterior distribution is also exchangeable, thus

$$p^L(\boldsymbol{\vartheta}_k|\mathbf{y}, \mathcal{M}_k) = p^L(\psi(\boldsymbol{\vartheta}_k)|\mathbf{y}, \mathcal{M}_k) = \frac{1}{k!} \sum_{\psi \in \mathfrak{N}(k)} p^L(\psi(\boldsymbol{\vartheta}_k)|\mathbf{y}, \mathcal{M}_k), \tag{3.5}$$

where $\mathfrak{N}(k)$ is the set of $k!$ possible permutations of the set $\{1, ..., k\}$. Using the standard Rao-Blackwell argument in Marin and Robert (2008) and defining the latent indicator z_i where $z_i = j$ if observation i is assigned to component j, we estimate (3.5) by

$$\hat{p}^L(\hat{\boldsymbol{\vartheta}}_k|\mathbf{y}, \mathcal{M}_k) = \frac{1}{Tk!} \sum_{\psi \in \mathfrak{N}(k)} \sum_{t=1}^{T} p^L(\psi(\hat{\boldsymbol{\vartheta}}_k)|\mathbf{y}, \boldsymbol{z}^{(t)}, \mathcal{M}_k), \tag{3.6}$$

where $\mathbf{z}^{(t)} = (z_1^{(t)}, \ldots, z_n^{(t)})$ are posterior samples from $p^L(\mathbf{z}, \boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$, which can also be used to evaluate $\omega(\boldsymbol{\vartheta}_k)$ in (3.1). We remark that the algorithm so far can be applied to any model for which posterior samples are available. Algorithm 1 provides an adaptation for Normal mixtures, which we used in our examples.

Algorithm 1: $p(\mathbf{y}|\mathcal{M}_k)$ for Normal mixtures under prior (2.3)-(2.4).

Initialize $\boldsymbol{\vartheta}_k^{(0)} = (\boldsymbol{\theta}_1^{(0)}, ..., \boldsymbol{\theta}_k^{(0)}, \boldsymbol{\eta}^{(0)})$. **for** $t = 1, ..., T$ **do**

Draw $z_i^{(t)} = j$ with probability:

$$\frac{\eta_k^{(t-1)} N(\mathbf{y}_i; \boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})}{\sum_{j=1}^k \eta_j^{(t-1)} N(\mathbf{y}_i; \boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})}.$$

Let $n_j^{(t)} = \sum_{i=1}^n \mathrm{I}(z_i^{(t)} = j)$ and $\bar{\mathbf{y}}_j^{(t)} = \dfrac{1}{n_j} \sum_{z_i^{(t)}=j} \mathbf{y}_i$ if $n_j^{(t)} > 0$, else $\bar{\mathbf{y}}_j^{(t)} = 0$. Draw

$$\boldsymbol{\eta}^{(t)} \sim \mathrm{Dir}(q + n_1^{(t)}, ..., q + n_k^{(t)}).$$

Let $S_j = S^{-1} + \sum_{z_i=j} (\mathbf{y}_i - \boldsymbol{\mu}_j^{(t-1)})(\mathbf{y}_i - \boldsymbol{\mu}_j^{(t-1)})' + \sum_{j=1}^k \dfrac{n_j/g}{n_j + 1/g} \bar{\mathbf{y}}_j^{(t)} \bar{\mathbf{y}}_j^{\prime(t)}$. Draw

$$\Sigma_j^{(t)} \sim \mathrm{IW}\left(\nu + n_j, S_j\right),$$

Draw

$$\boldsymbol{\mu}_j^{(t)} \sim N\left(\frac{g n_j^{(t)} \bar{\mathbf{y}}_j^{(t)}}{1 + g n_j^{(t)}}, \frac{g}{1 + g n_j^{(t)}} \Sigma_j^{(t)}\right),$$

end

Compute $\hat{p}^L(\mathbf{y}|\mathcal{M}_k)$ as in (3.4) where $\hat{\boldsymbol{\vartheta}}_k$ is the posterior mode, and $\hat{p}(\mathbf{y}|\mathcal{M}_k)$ as in (3.1).

3.2. **Posterior modes.** The EM algorithm and variations thereof are a popular strategy to obtain $\hat{\boldsymbol{\vartheta}}_k = \arg\max_{\boldsymbol{\vartheta}_k} p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) P(\mathcal{M}_k)$, a convenient way to quickly estimate parameters, cluster individuals or set $\hat{\boldsymbol{\vartheta}}_k$ in (3.4) to compute the integrated likelihood. We briefly describe the algorithm, see Supplementary Section 8 for an outline of its derivation. As usual, at iteration t the E-step computes $\bar{z}_{ij}^{(t)} = P(z_i = j|\mathbf{y}_i, \boldsymbol{\vartheta}_j^{(t-1)}) = \eta_j^{(t-1)} p(\mathbf{y}_i|\boldsymbol{\theta}_j^{(t-1)}) / \sum_{j=1}^k \eta_j^{(t-1)} p(\mathbf{y}_i|\boldsymbol{\theta}_j^{(t-1)})$ and is trivial to implement. The M-step requires updating $\boldsymbol{\vartheta}_k^{(t)}$ in a manner that increases the expected complete log-likelihood, which we denote by $\xi(\boldsymbol{\vartheta}_k)$, but under our prior $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ in general this cannot be done in closed-form. A key observation is that if $p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ leads to closed-form updates, the corresponding target $\xi^L(\boldsymbol{\vartheta}_k)$ only differs from $\xi(\boldsymbol{\vartheta}_k)$ by a term $d_\vartheta(\boldsymbol{\vartheta}_k)$, thus one may approximate $\xi(\boldsymbol{\vartheta}_k)$ via a first order Taylor expansion of $d_\vartheta(\boldsymbol{\vartheta}_k)$. Naturally these updates are approximate and need not lead to an increase in $\xi(\boldsymbol{\vartheta}_k)$, but typically they do given that $d_\vartheta(\boldsymbol{\vartheta}_k)$ only features in the prior and thus only has a mild influence for moderately large n. Nevertheless, whenever $\xi(\boldsymbol{\vartheta}_k)$ is not increased we resort to updates given by the gradient algorithm. Algorithm 2 details the algorithm for Normal mixtures (extensions to other models follow similar lines), for simplicity outlining only the closed-form updates (see Supplementary Section 8 for the gradient algorithm). Algorithm 2 is initialized at an arbitrary $\boldsymbol{\vartheta}_k^{(0)}$ (in our implementation the MLE) and stops whenever the increase in $\xi(\boldsymbol{\vartheta}_k)$ is below a tolerance ϵ^* or a maximum number of iterations T is reached. In our examples we set $T = 10000$ and $\epsilon^* = 0.0001$. For ease of notation we define $d_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A_\Sigma^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ evaluated at the current value of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \Sigma_1, \ldots, \Sigma_k$.

Algorithm 2: EM under MOM-IW-Dir priors.

Set $t = 1$. **while** $\zeta > \epsilon^*$ *and* $t < T$ **do**

for $t \geq 1$ *and* $j = 1, ..., k$ **do**

E-step. Let $\bar{z}_{ij}^{(t)} = \dfrac{\eta_j^{(t-1)} N(\mathbf{y}_i; \boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})}{\sum_{j=1}^k \eta_j^{(t-1)} N(\mathbf{y}_i; \boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})}$ and $n_j^{(t)} = \sum_{i=1}^n \bar{z}_{ij}^{(t)}$.

M-step. Let $\bar{\mathbf{y}}_j^{(t)} = \sum_{i=1}^n \bar{z}_{ij}^{(t)} \mathbf{y}_i / n_j^{(t)}$. Update

$$\boldsymbol{\mu}_j^{(t)} = \left(\Sigma_j^{-1(t-1)} n_j^{(t)} + A_{\Sigma^{(t-1)}}^{-1} \left(\frac{1}{g} + \sum_{i \neq j} \frac{2}{d_{ij}} \right) \right)^{-1} \times \left(\Sigma^{-1(t-1)} n_j^{(t)} \bar{\mathbf{y}}_j^{(t)} + A_{\Sigma^{(t-1)}}^{-1} \left(\sum_{i \neq j} \frac{\boldsymbol{\mu}_j^{(t-1)} - (\boldsymbol{\mu}_i^{(t-1)} - \boldsymbol{\mu}_j^{(t-1)})}{d_{ij}} \right) \right),$$

Update $(\nu - p + n_j^{(t)}) \Sigma_j^{(t)} =$

$$S^{-1} + \frac{\boldsymbol{\mu}_j^{(t)} (\boldsymbol{\mu}_j^{(t)})'}{kg} + \sum_{i=1}^n \bar{z}_{ij}^{(t)} (\mathbf{y}_i - \boldsymbol{\mu}_j^{(t)})(\mathbf{y}_i - \boldsymbol{\mu}_j^{(t)})' - \frac{1}{k} \sum_{i \neq j} \frac{2(\boldsymbol{\mu}_j^{(t)} - \boldsymbol{\mu}_k^{(t)})(\boldsymbol{\mu}_j^{(t)} - \boldsymbol{\mu}_k^{(t)})'}{d_{ij}}.$$

Update $\eta_j^{(t)} = \dfrac{n_j^{(t)} + q - 1}{n + k(q-1)}$.

end

Compute $\zeta = |\xi(\boldsymbol{\vartheta}_k^{(t)}) - \xi(\boldsymbol{\vartheta}_k^{(t-1)})|$ and set $t = t + 1$.

end

4. Results

We compare the performance of our MOM-IW-Dir (2.3)-(2.4) under default prior parameters to its LP counterpart Normal-IW-Dir with dispersion g^L set to match the 95% percentile for the separation parameter κ (Section 2.3) and to the BIC. Throughout we set uniform prior model probabilities $P(\mathcal{M}_1) = \ldots = P(\mathcal{M}_K) = 1/K$. We estimated the integrated likelihoods using Algorithm 1 based on 5,000 MCMC draws after a 2,500 burn-in. We focus attention on posterior model probabilities $P(\mathcal{M}_k \mid \mathbf{y})$. To facilitate comparison of the BIC with MOM-IW-Dir and Normal-IW-Dir we transformed the BIC into a criterion taking values in [0,1] using (1.3), *i.e.* the asymptotic relationship between the BIC and the integrated likelihood in regular models. For simplicity we denote this criterion by $P(\mathcal{M}_k \mid \mathbf{y})$ and, although it cannot be interpreted as a bona fide probability, it can still be viewed as measuring the strength of evidence provided by BIC in favour of each model. Section 4.1 presents a simulation study for univariate and bivariate Normal mixtures. Section 4.2 explores a model mispecification case where data come from a Student-t mixture. Sections 4.3 and 4.4 analyze the Old Faithful and Fisher's Iris datasets.

4.1. **Simulation study.** We consider choosing amongst the three competing models

$$\mathcal{M}_1 : N(\mathbf{y}_i; \boldsymbol{\mu}, \Sigma),$$
$$\mathcal{M}_2 : \eta_1 N(\mathbf{y}_i; \boldsymbol{\mu}_1, \Sigma) + (1-\eta_1) N(\mathbf{y}_i; \boldsymbol{\mu}_2, \Sigma)$$
$$\mathcal{M}_3 : \eta_1 N(\mathbf{y}_i; \boldsymbol{\mu}_1, \Sigma) + \eta_2 N(\mathbf{y}_i; \boldsymbol{\mu}_2, \Sigma) + (1-\eta_1-\eta_2) N(\mathbf{y}_i; \boldsymbol{\mu}_3, \Sigma),$$

where independence is assumed across $i = 1, \ldots, n$. We simulated 100 datasets under each of the 8 data-generating truths depicted in Figure 2 for univariate (Cases 1-4) and bivariate outcomes (Cases 5-8). Case 1 corresponds to $k^* = 1$ components, Cases 2-3 to $k^* = 2$ moderately and strongly-separated components respectively, Case 4 to $k^* = 3$ with two strongly overlapping components and a third component with smaller weight, and Cases 5-8 are analogous cases for the bivariate outcome.

Figure 3 shows the average posterior probability assigned to the data-generating model $P(\mathcal{M}_{k^*} \mid \mathbf{y})$. Supplementary Figures 7-9 show the corresponding posterior expected number of components and, to assess sensitivity to prior specification, for alternative choices of g giving $P(\kappa < 4 \mid \mathcal{M}_k) = 0.1$ instead of 0.05 and various q. Overall a similar behavior is observed in the univariate and bivariate cases. The BIC favoured sparse solutions but lacked sensitivity to detect some truly present components, the LP exhibited better sensitivity but did not enforce sparsity sufficiently, and the NLP achieved a balance between the two. For instance, the BIC (correctly) strongly supported $k = 1$ in Cases 1 and 5 and $k = 2$ in Cases 3 and 7, but it lacked sensitivity to detect moderately-separated components in Cases 2, 4, 6 and 8, especially in Case 6 where $P(\mathcal{M}_{k^*} \mid \mathbf{y})$ was essentially 0 even with $n = 1,000$ observations. Here LPs showed a much improved sensitivity, although assigning relatively large probability to models with spurious components in the remaining four cases. In contrast, NLPs supported $\mathcal{M}_{k^*}$ to an extent comparable to BIC in Cases 1, 3 and 7 but in Cases 2, 4 (for $q = 4$) and 6 showed even higher sensitivity than LPs. This last observation illustrates that although NLPs may have slightly reduced power to detect poorly-separated components, they may in fact increase power for moderately-separated components due to assigning larger $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ consistent with that degree of separation. Supplementary Figures 7-9 show similar results under alternative prior settings, although $P(\kappa < 4 \mid \mathcal{M}_k) = 0.05$ was found to lead to slightly improved parsimony relative to $P(\kappa < 4 \mid \mathcal{M}_k) = 0.10$.

4.2. **Inference under a misspecified model.** In practice the data-generating density may present non-negligible departures from the assumed class. An important case we investigate here is the presence of heavy tails, which under an assumed Normal mixture likelihood may affect both the chosen number of components and the parameter estimates. We generated $n = 600$ observations from $k^* = 3$ bivariate Student-t components with 4 degrees of freedom, means $\boldsymbol{\mu}_1 = (-1, 1)'$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_3 = (6, 6)'$ and a common scale matrix with elements $\sigma_{11} = \sigma_{22} = 2$ and $\sigma_{12} = \sigma_{21} = -1$. We considered up to 5 components with either homogeneous $\Sigma_1 = \ldots = \Sigma_k$ or heterogeneous covariance matrices, giving a total of 11 models.

Supplementary Table 1 shows $P(\mathcal{M}_k \mid \mathbf{y})$ given by the BIC approximation and the Normal-IW-Dir and MOM-IW-Dir priors. BIC assigned overwhelming evidence in favour

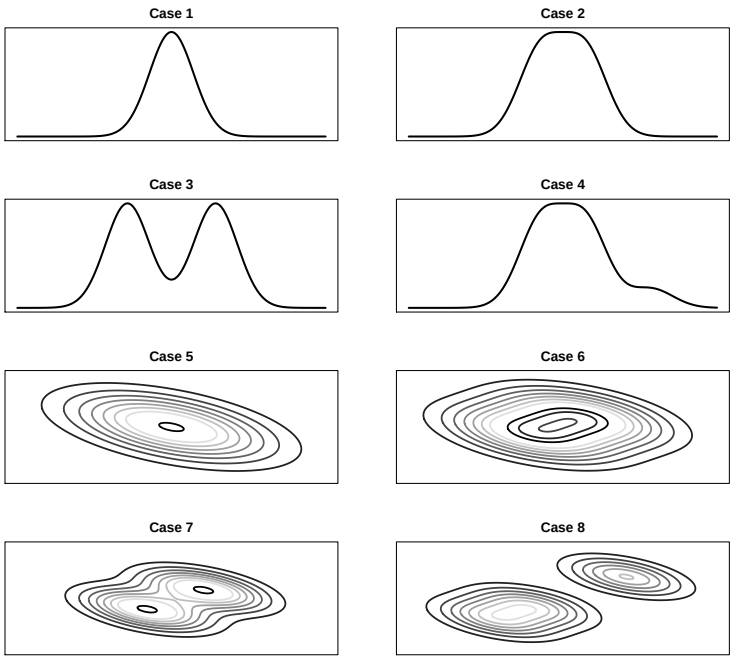


FIGURE 2. Data-generating densities in simulation study. Case 1: $k^* = 1$, $\boldsymbol{\mu}_1 = 0$; Case 2: $k^* = 2$, $\boldsymbol{\mu}_1 = -1, \boldsymbol{\mu}_2 = 1$, $\eta_1 = \eta_2 = 0.5$; Case 3: $k^* = 2$, $\boldsymbol{\mu}_1 = -2, \boldsymbol{\mu}_2 = 2$, $\eta_1 = \eta_2 = 0.5$; Case 4: $k^* = 3$, $\boldsymbol{\mu}_1 = -1, \boldsymbol{\mu}_2 = 1, \boldsymbol{\mu}_3 = 4$, $\eta_1 = 0.45, \eta_2 = 0.45$, $\eta_3 = 0.1$; Case 5: $k^* = 1$, $\boldsymbol{\mu} = (0, 0)'$; Case 6: $k^* = 2$, $\boldsymbol{\mu}_1 = (-0.4, -0.6)'$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$; Case 7: $k^* = 2$, $\boldsymbol{\mu}_1 = (-0.65, -0.85)'$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$; Case 8: $k^* = 3$, $\boldsymbol{\mu}_1 = (-0.65, -0.85)'$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_3 = (3, 3)'$. We set $\Sigma = 1$ in Cases 1-4, $\sigma^2_{11} = \sigma^2_{22} = 1$ and $\sigma^2_{12} = \sigma^2_{21} = -0.5$ in Cases 5-8.

of $k = 4$ components with heterogeneous covariances, whereas the LP placed most posterior probability on $k \in \{5, 6\}$ components with common covariances. In contrast, our NLP assigned posterior probability 1 (up to rounding) to $k = 3$ with equal covariances. To provide further insight Figure 4 shows the component contours for the modal model of each method, setting the parameter estimates to the posterior modes, as well as the classification of observations into their most probable component. The NLP solution (lower panel) returned three components with means reflecting the location of the true Student-t components. The BIC (upper panel) approximated the two mildly-separated components with two normals centered roughly at (0,0), analogously to its lack of sensitivity to discern overlapping components observed in Section 4.1, whereas the heavier-than-normal tails of the third Student-t component are captured via two normals with different variance. Three of the components returned by the LP solution (middle panel) capture the location of the three components, but two extra components are added to account for the heavy tails. This example illustrates how by penalizing solutions with poorly-separated

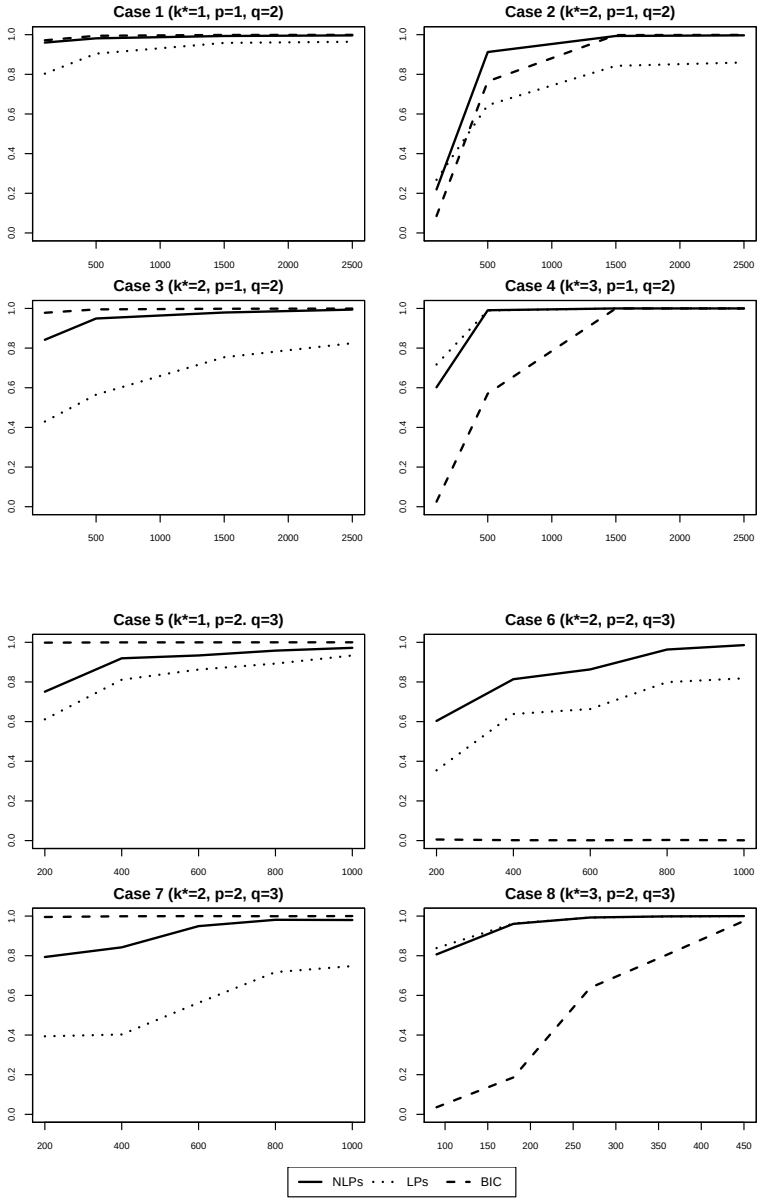


FIGURE 3. Average posterior probability for the data-generating truth model for Cases 1-8.

or low-weight components NLPs may induce a form of robustness to model mispecification, although we remark that this is a finite-sample effect and would eventually vanish as n grows to infinity.

An alternative to conducting formal model selection is to fit a single mixture with a large k and then discard components that are deemed unnecessary. Although this can be computationally convenient we remark that in general the criterion to discard components is case-dependent and, more importantly, that the quality of the inference may suffer. For instance, in this example if one were to set $k = 6$ both BIC and the LP would return

at least four components with an estimated weight $\hat{\eta}_j$ >0.15 and that would thus not be discarded in practice.

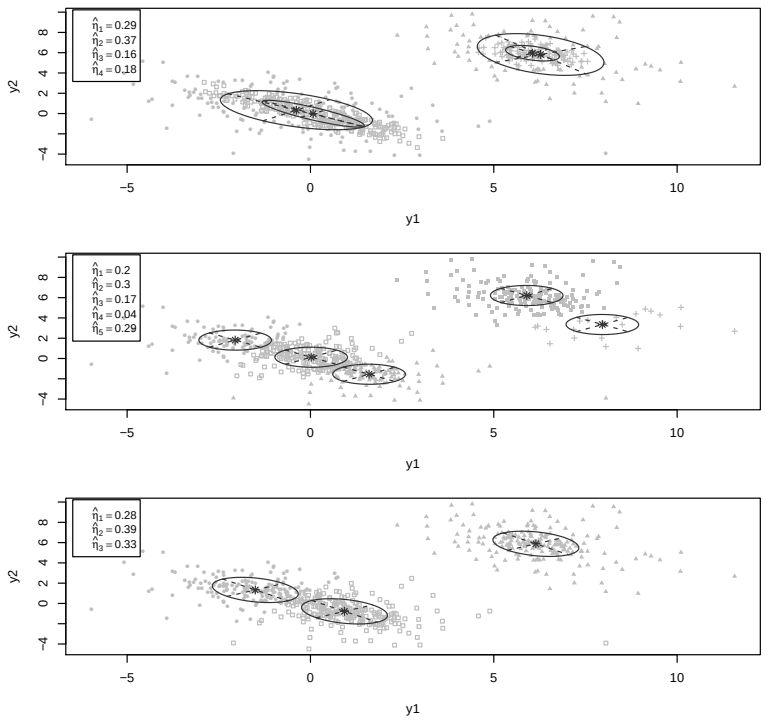


FIGURE 4. Classification and contours in the misspecified model example for BIC (top), Normal-IW-Dir (middle) and MOM-IW-Dir (bottom). Different symbols indicate modal cluster allocations.

4.3. **Old Faithful.** The Old Faithful is the biggest cone-type geyser in the Yellowstone National Park and is one of the most interesting geographical features on Earth. We seek clusters in a dataset with $n = 272$ eruptions recording their duration and the time to the next eruption (dataset faithful in R). We consider up to $K = 5$ normal components either with homoscedastic or heteroscedastic covariance matrices. The posterior model probabilities under BIC and our NLP favoured $k = 3$ components with equal covariances (0.927 and 0.967 respectively, Supplementary Table 2), but the cluster shapes are slightly different (Figure 5). The LP respectively assigned 0.473 and 0.353 posterior probability to $k = 4$ and $k = 5$ components, although in the latter case $\hat{\eta}_5 = 0.004$ is negligible. Relative to the BIC and our NLP, the LP splits the component in the lower-left corner into two.

4.4. **Fisher's Iris data.** Fisher's Iris flower data (Fisher, 1936) contain four variables measuring the dimensions of $n = 150$ plants, which we transformed into principal components to facilitate plotting the results. The plants are known to belong to three species, iris setosa, iris versicolor and iris virginica, each with 50 observations. We compare the

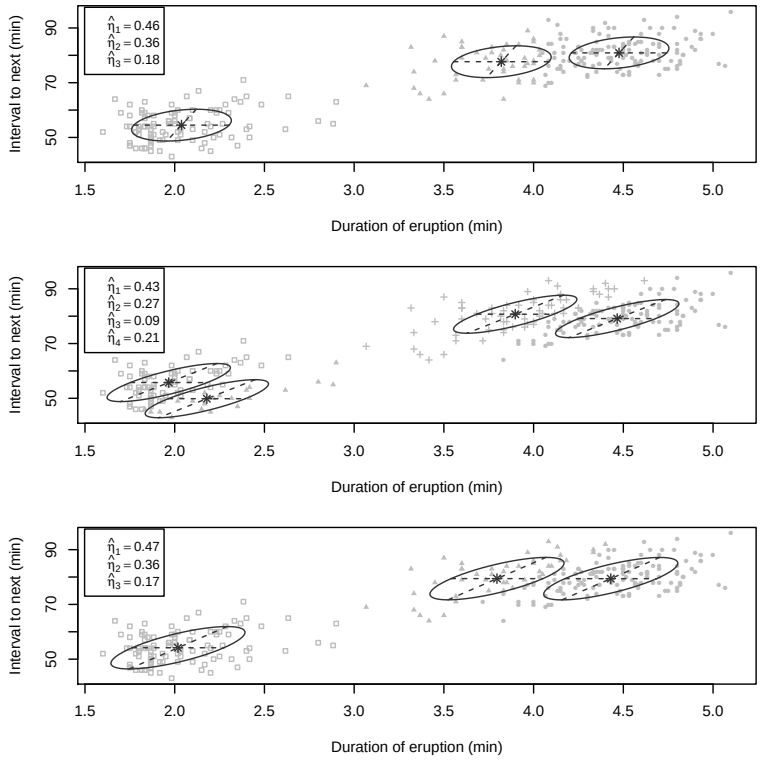


FIGURE 5. Faithful dataset. Classification of observations and contours for the model chosen by BIC (top), Normal-IW-Dir (middle) and MOM-IW-Dir (bottom). Different symbols indicate modal cluster allocations.

ability of the various methods to recover these underlying three species in an unsupervised fashion. We considered up to $k = 6$ normal components with either equal or unequal covariances.

Supplementary Table 3 provides posterior model probabilities. The BIC strongly supported two components with heteroscedastic covariances (posterior probability=0.968). Upon inspection this solution failed to distinguish the versicolor and virginica species, which are merged into a single component, akin to its lack of sensitivity observed in our other examples. Both LP and our NLP supported a three component Normal, albeit the evidence for the former was weaker than for the latter with $P^L(\mathcal{M}_3|\mathbf{y}) = 0.80$ and $P(\mathcal{M}_3|\mathbf{y}) = 1$. Figure 6 shows the contours of the NLP solution for the first two principal components (accounting for 96.0% of the variance in the data), which closely resemble the three species. We remark that a strategy based on fitting a single model with $k = 6$ components and dropping those with low estimated weight result in the addition of multiple spurious components (Supplementary Table 3).

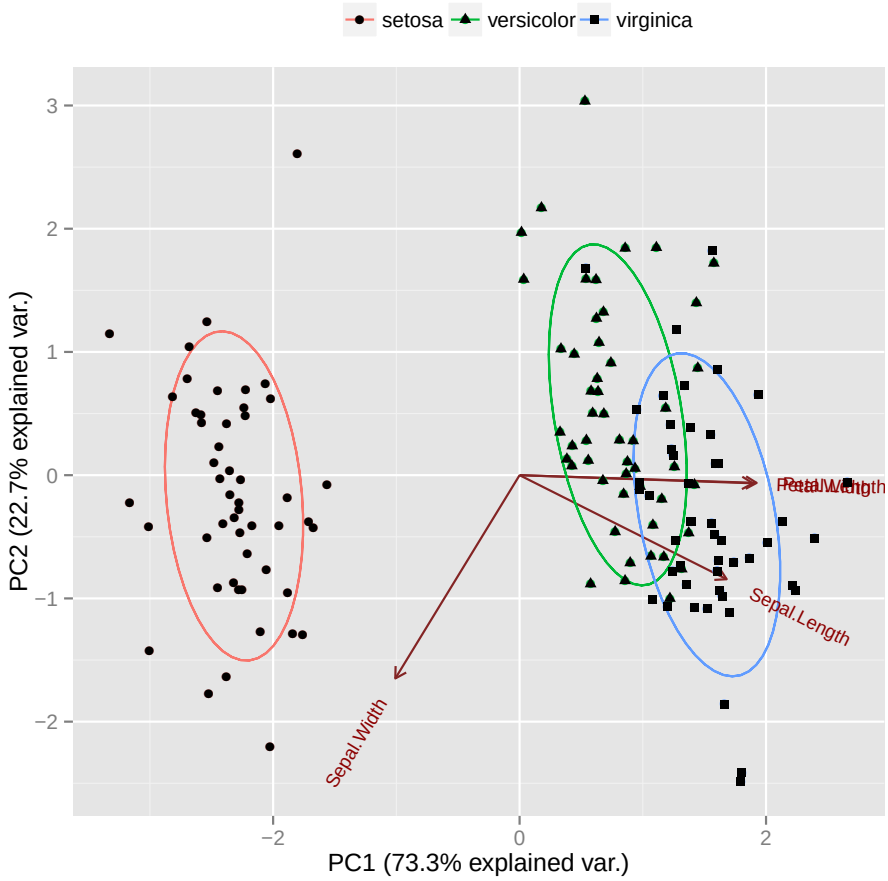


FIGURE 6. Principal components for the Fisher's Iris data-set, classification of observations and contours using EM algorithm under NLPs.

5. CONCLUSIONS

We propose the use of NLPs to select the number of mixture components. The primary goal was to encourage solutions that not only balance parsimony and sensitivity, but that also facilitate interpretation in terms of well-separated underlying subpopulations. From a theoretical standpoint our chosen formulation asymptotically enforces parsimony under the wide class of generically identifiable mixtures, with specific asymptotic rates being available under mild additional regularity conditions. From a computational standpoint we found a closed-form expression for the normalization constant of a prior formulation that is applicable to any location-scale mixture, avoiding a doubly-intractable problem. Further, we provide direct extensions of existing algorithms to obtain the marginal likelihood, posterior samples and posterior modes at negligible additional cost relative to standard formulations, rendering the approach practical.

Our results showed a systematic trend of the BIC to miss necessary components, in some instances even with large sample sizes, and of a LP counterpart to our formulation to add spurious components. This was observed in simulations but also in datasets with known subgroup structure such as Fisher's Iris data. Interestingly, as an alternative to the formal Bayesian model selection framework adopted in our approach we attempted fitting a single model with many components and then dropping those deemed unnecessary. This alternative led to the addition of spurious components in our three examples, *e.g.* in 85%-95% of the MCMC iterations there were no empty components (Supplementary Tables 4-5), which may naively suggest to keep all components in the model. We remark that in our examples we used a uniform prior on the model space, thus Bayesian model selection may achieve further parsimony by reinforcing sparse models a priori.

An intriguing observation to be pursued in follow-up work is that, by penalizing poorly-separated and low-weight components, NLPs showed robustness to model mispecification in an example, suggesting that it may be interesting to combine this prior formulation with robust likelihoods (*e.g.* based on asymmetric or heavy-tailed components). Another interesting venue is to consider extensions to non-parametric infinite mixtures and their connections to determinantal point processes. Overall, our findings suggest that NLPs are a sensible basis to tackle a long-standing open model selection problem in mixture models.

Acknowledgments

David Rossell was partially funded by the NIH grant R01 CA158113-01 and the EPSRC First Grant EP/N011317/1.

References

R. H. Affandi, E. B. Fox, and B. Taskar. Approximate inference in continuous determinantal process. In *Neural information processing systems*, pages 1–9, Lake Tahoe, 2013.

J-P. Baudry, A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo. Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, 19: 332–353, 2010.

J. Berkhof, I. V. Mechelen, and A. Gelman. A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica*, 13:423–442, 2003.

A. Bhattacharya, D. Pati, N.S. Pillai, and D.B. Dunson. Dirichlet-Laplace Priors for Optimal Shrinkage. *Journal of the American Statistical Association*, 110:1479–1490, 2015.

I. Castillo, J. Schmidt-Hieber, and A.W. van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43:1986–2018, 2015.

A. Chambaz and J. Rousseau. Bounds for Bayesian order identification with application to mixtures. *The Annals of Statistics*, 36:928–962, 2008.

J. Chen and P. Li. Hypothesis test for Normal mixture models: The EM approach. *The Annals of Statistics*, 37:2523–2542, 2009.

S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321, 1995.

S.L. Crawford. An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association*, 89:259–267, 1994.

A.P. Dawid. The trouble with Bayes factors. Technical report, University College London, 1999.

A.P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, B*, 39-1:1–38, 1977.

B. Efron. Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, 23-1:1–22, 2008.

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

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

S. Frühwirth-Schnatter. *Finite Mixtures and Markov Switching Models*. Springer, New York, 2006.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis, Third Edition*. Boca Raton: Chapman and Hall/CRC, 2013.

S. Ghosal. A review of consistency and convergence of posterior distribution. In *Division of theoretical statistics and mathematics*, pages 1–10, Indian Statistical Institute, 2002.

S. Ghosal and A. V. der Vaart. Entropies and rates of converegence for maximum likelihood and bayes estimation for mixture of normal densities. *Annals of Statistics*, 29: 1233–1263, 2001.

S. Ghosal and A. Van Der Vaart. Posterior convergence rates of dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35:697–723, 2007.

J. K. Gosh and P. K. Sen. On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results. In *Le Cam, L. M., Olshen, R. A. (Eds.), Proceedings of the Berkeley conference in Honor of Jerzy Neyman and Jack Kiefer*, volume II, pages 789–806, Wadsworth, Monterey, 1985.

R. J. Hathaway. A constrained formulation of maximum-likelihood estimation for Normal mixture distributions. *The annals of Statistics*, 13:795–800, 1985.

Z. V. Havre, N. White, J. Rousseau, and K. Mengersen. Overfitting bayesian mixture models with and unknown number of components. *PLoS ONE*, 10 (7):1–27, 2015.

V. E. Johnson and D. Rossell. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal Royal Statistical Society, B*, 72:143–170, 2010.

V. E. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107:649–655, 2012.

R. Kan. From moments of sums to moments of product. *Journal of Multivariate Analysis*, 99:542–554, 2006.

B. G. Leroux. Consistence estimation of a mixing distribution. *The Annals of Statistics*, 20:1350–1360, 1992.

X. Liu and Y. Z. Chao. Asymptotics for likelihood ratio test in a two-component normal mixture model. *Journal Statistical Planing and Inference*, 123:61–81, 2004.

I-Li Lu and Donald Richards. Random discriminants. *The Annals of Statistics*, 21: 1982–2000, 1993.

G. Malsiner-Walli, S. Frühwirth-Schnatter, and B. Grün. Identifying mixtures of mixtures using bayesian estimation. Technical report, 2015. ArXiv: 1502.06449v2.

J. M. Marin and C. P. Robert. Approximating the marginal likelihood in mixture models. *Bulleting of the Indian Chapter of ISBA*, 1:2–7, 2008.

K. L. Mengersen, C. P. Robert, and D. M. Titterington. *Mixtures: Estimation and Applications*. Wiley, 2011.

R. Neal. Erroneous results in "Marginal Likelihood from the Gibbs Output". *Technical Report*, 1999.

F. Petralia, V. Rao, and D.B. Dunson. Repulsive mixtures. In *Advances in Neural Information Processing Systems*, pages 1889–1897, 2012.

S. Ray and B. Lindsay. The topography of multivariate normal mixtures. *The Annals of Statistics*, 33:2042–2065, 2005.

R. Redner. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics*, 9:225–228, 1981.

S. Richardson and P. J. Green. On Bayesian analysis of mixture models with an unknown number of components. *Journal of the Royal Statical Society*, B-59:731–792, 1997.

D. Rossell and D. Telesca. Non-local priors for high-dimensional estimation. *Journal of the American Statistical Association*, (in press), 2016.

D. Rossell, D. Telesca, and V. E. Johnson. High-dimensional Bayesian classifiers using non-local priors. In *Statistical Models for Data Analysis*, pages 305–314, Springer, 2013.

J. Rousseau. Approximating interval hypotheses: p-values and Bayes factors. In J.M. Bernardo, J. O. Berger, A. P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 8*, pages 417–452. Oxford University Press, 2007.

J. Rousseau and K. Mengersen. Asymptotic behavior of the posterior distribution in over-fitted models. *Journal of the Royal Statistical Society B*, 73:689–710, 2011.

N. J. Schork, D. B. Allison, and B. Thiel. Mixture distribution in human genetics. *Statistical Methods in Medical Research*, 5:155–178, 1996.

G. Schwarz. Estimating the dimension of a model. *The Annals of statistics*, 6:461–464, 1978.

M. Shin, A. Bhattacharya, and V.E. Johnson. Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. Technical report, Texas A&M University, 2015.

H. Teicher. Identifibility of finite mixtures. *The Annals of Mathematical Statistics*, 34: 1265–1269, 1963.

S. Watanabe. *Algebraic geometry and statistical learning theory*, volume 25 of *Cambridge monographs on applied and computational mathematics*. Cambridge University Press, 2009.

S. Watanabe. A widely applicable Bayesian information criteria. *Journal of Machine Learning Research*, 14:867–897, 2013.

M. West and D. A. Turner. Deconvolution of mixtures in analysis of neural synaptic transmission. *The Statistician*, 43:31–43, 1994.

Y. Xu, P. Mueller, and D. Telesca. Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics*, (in press), 2016.

S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematics and Statistics*, 39:209–214, 1968.

Supplementary material

6. Conditions A1-A4 in Rousseau and Mengersen (2011)

For convenience we reproduce verbatim Conditions A1-A4 in Rousseau and Mengersen (2011), adjusted to the notation we used in this paper. Their Condition A5 is trivially satisfied by our $\boldsymbol{\eta} \sim \text{Dir}(q)$ prior, hence is not reproduced here. Recall that we defined $p^*_{k^*}(\mathbf{y}) = p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})$ to be the data-generating truth.

We denote $\Theta^*_k = \{\boldsymbol{\vartheta}_k \in \Theta_k; p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) = p^*_{k^*}(\mathbf{y})\}$ and let $\log(p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k))$ be the log-likelihood calculated at $\boldsymbol{\vartheta}_k$. Denote $F_0(g) = \int p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*}) g(\mathbf{y}) d\mathbf{y}$ where $g(\cdot)$ is a probability density function, denote by $\text{Leb}(A)$ the Lebesgue measure of a set A and let $\nabla p(\mathbf{y}|\boldsymbol{\theta})$ be the vector of derivatives of $p(\mathbf{y}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, and $\nabla^2 p(\mathbf{y}|\boldsymbol{\theta})$ be the second derivatives with respect to $\boldsymbol{\theta}$. Define for $\epsilon \geq 0$

$$\bar{p}(\mathbf{y}|\boldsymbol{\theta}) = \sup_{|\boldsymbol{\theta}^l - \boldsymbol{\theta}| \leq \epsilon} p(\mathbf{y}|\boldsymbol{\theta}^l), \qquad \underline{p}(\mathbf{y}|\boldsymbol{\theta}) = \inf_{|\boldsymbol{\theta}^l - \boldsymbol{\theta}| \leq \epsilon} p(\mathbf{y}|\boldsymbol{\theta}^l)$$

We now introduce some notation that is useful to characterize Θ^*_k, following Rousseau and Mengersen (2011). Let $\mathbf{w} = (w_i)_{i=0}^{k^*}$ with $0 = w_0 < w_1 < ... < w_{k^*} \leq k$ be a partition of $\{1, ..., k\}$. For all $\boldsymbol{\vartheta}_k \in \Theta_k$ such that $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) = p^*_k(\mathbf{y})$ there exists $\mathbf{w}$ as defined above such that, up to a permutation of the labels,

$$\forall i = 1, ..., k^*, \quad \boldsymbol{\theta}_{w_{i-1}+1} = ... = \boldsymbol{\theta}_{w_i} = \boldsymbol{\theta}^*_i, \quad \eta(i) = \sum_{j=w_{i-1}+1}^{w_i} \eta_j = \eta^*_i, \quad \eta_{w_{k^*+1}} = ... = \eta_k = 0.$$

In other words, $I_i = \{w_{i-1} + 1, ..., w_i\}$ represents the cluster of components in $\{1, ..., k\}$ having the same parameter as $\boldsymbol{\theta}^*_i$. Then define the following parameterisation of $\boldsymbol{\vartheta}_k \in \Theta_k$ (up to permutation)

$$\boldsymbol{\iota}_{\boldsymbol{w}} = \left((\boldsymbol{\theta}_j)_{j=1}^{w_{k^*}}, (r_i)_{i=1}^{k^*-1}, (\eta_j)_{j=w_{k^*+1}}^{k} \right) \in \Re^{pw_{k^*} + k^* + k - w_{k^*} - 1}, \quad r_i = \eta(i) - \eta^*_i, \quad i = 1, ..., k^*,$$

and

$$\boldsymbol{\varpi}_{\boldsymbol{w}} = \left((f_j)_{j=1}^{w_{k^*}}, \boldsymbol{\theta}_{w_{k^*}+1}, ..., \boldsymbol{\theta}_k \right), \qquad f_j = \frac{\eta_j}{\eta(i)}, \quad \text{when} \quad j \in I_i = \{w_{i-1} + 1, ..., w_i\},$$

note that for $p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})$

$$\boldsymbol{\iota}^*_{\boldsymbol{w}} = (\boldsymbol{\theta}^*_1, ..., \boldsymbol{\theta}^*_1, \boldsymbol{\theta}^*_2, ..., \boldsymbol{\theta}^*_2, ..., \boldsymbol{\theta}^*_{k^*}, ..., \boldsymbol{\theta}^*_{k^*}, 0...0...0)$$

where $\boldsymbol{\theta}^*_i$ is repeated $w_i - w_{i-1}$ times in the above vector for any $\boldsymbol{\varpi}_{\boldsymbol{w}}$. Then we parameterize $(\boldsymbol{\iota}_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}})$, so that $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) = p(\mathbf{y} \mid (\boldsymbol{\iota}_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}}), \mathcal{M}_k)$ and we denote

$\nabla p(\mathbf{y} \mid (\boldsymbol{\iota}^*_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}}), \mathcal{M}_k)$ and $\nabla^2 p(\mathbf{y} \mid (\boldsymbol{\iota}^*_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}}), \mathcal{M}_k)$ the first and second derivatives of $p(\mathbf{y} \mid (\boldsymbol{\iota}_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}}), \mathcal{M}_k)$ with respect to $\boldsymbol{\iota}_{\boldsymbol{w}}$ and computed at $\boldsymbol{\vartheta}^*_{k^*} = (\boldsymbol{\iota}^*_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}})$. We also denote by $P^L(\cdot \mid \mathbf{y}, \mathcal{M}_k)$ the posterior distribution using a LP.

Conditions

A1 L_1 *consistency.* For all $\epsilon = (\log n)^e/\sqrt{n}$ with $e \geq 0$ as $n \to \infty$

$$P^L\left(\int |p(\mathbf{z} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) - p_k^*(\mathbf{z})|\, d\mathbf{z} > \epsilon \mid \mathbf{y}, \mathcal{M}_k\right) \to 0$$

in probability with respect to $p(\mathbf{y} \mid \boldsymbol{\vartheta}_{k^*}^*, \mathcal{M}_{k^*})$.

A2 *Regularity.* The component density $p(\mathbf{y}|\boldsymbol{\theta})$ indexed by a parameter $\boldsymbol{\theta} \in \Theta$ is three times differentiable and regular in the sense that for all $\boldsymbol{\theta} \in \Theta$ the Fisher information matrix associated with $p(\mathbf{y}|\boldsymbol{\theta})$ is positive definite at $\boldsymbol{\theta}$. Denote $\nabla^3 p(\mathbf{y}|\boldsymbol{\theta})$ the array whose components are

$$\frac{\partial^3 p(\mathbf{y}|\boldsymbol{\theta})}{\partial_{\boldsymbol{\theta}_{i1}} \partial_{\boldsymbol{\theta}_{i2}} \partial_{\boldsymbol{\theta}_{i3}}}$$

For all $i \leq k^*$, there exists $\epsilon > 0$ such that

$$F_0\left(\frac{\bar{p}(\mathbf{y}|\boldsymbol{\theta}_i^*)^3}{\underline{p}(\mathbf{y}|\boldsymbol{\theta}_i^*)^3}\right) < \infty, \quad F_0\left(\frac{\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}^*|\leq\epsilon} |\nabla p(\mathbf{y}|\boldsymbol{\theta})|^3}{\underline{p}(\mathbf{y}|\boldsymbol{\theta}_i^*)^3}\right) < \infty, \quad F_0\left(\frac{|p(\mathbf{y}|\boldsymbol{\theta}_i^*)|^4}{(p(\mathbf{y} \mid \boldsymbol{\vartheta}_{k^*}^*, \mathcal{M}_{k^*}))^4}\right) < \infty,$$

$$F_0\left(\frac{\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}^*|\leq\epsilon} |\nabla^2 p(\mathbf{y}|\boldsymbol{\theta})|^2}{\underline{p}(\mathbf{y}|\boldsymbol{\theta}_i^*)^2}\right) < \infty, \qquad F_0\left(\frac{\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}^*|\leq\epsilon} |\nabla^3 p(\mathbf{y}|\boldsymbol{\theta})|^2}{\underline{p}(\mathbf{y}|\boldsymbol{\theta}_i^*)}\right) < \infty.$$

Assume also that for all $i = 1, ..., k^*$, $\boldsymbol{\theta}_i^* \in \text{int}(\Theta^k)$ the interior of Θ^k.

A3 *Integrability.* There exists $\Theta^{k^*} \subset \Theta^k$ satisfying $\text{Leb}(\Theta^{k^*}) > 0$ and for all $i \leq k^*$

$$d(\boldsymbol{\theta}_i^*, \Theta^{k^*}) = \inf_{\boldsymbol{\theta} \in \Theta^{k^*}} |\boldsymbol{\theta} - \boldsymbol{\theta}_i^*| > 0$$

and such that for all $\boldsymbol{\theta} \in \Theta^{k^*}$,

$$F_0\left(\frac{p(\mathbf{y}|\boldsymbol{\theta})^4}{(p(\mathbf{y} \mid \boldsymbol{\vartheta}_{k^*}^*, \mathcal{M}_{k^*}))^4}\right) < \infty, \qquad F_0\left(\frac{p(\mathbf{y}|\boldsymbol{\theta})^3}{\underline{p}(\mathbf{y}|\boldsymbol{\theta}_i^*)^3}\right) < \infty, \qquad \forall i \leq k^*.$$

A4 *Stronger identifiability.*

For all $\mathbf{w}$ partitions of $\{1, ..., k\}$ as defined above, let $\boldsymbol{\vartheta}_k \in \Theta_k$ and write $\boldsymbol{\vartheta}_k$ as $(\boldsymbol{\iota}_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}})$; then

$$(\boldsymbol{\iota}_{\boldsymbol{w}} - \boldsymbol{\iota}_{\boldsymbol{w}}^*)' \nabla p(\mathbf{y} \mid (\boldsymbol{\iota}_{\boldsymbol{w}}^*, \boldsymbol{\varpi}_{\boldsymbol{w}}), \mathcal{M}_k) + \frac{1}{2}(\boldsymbol{\iota}_{\boldsymbol{w}} - \boldsymbol{\iota}_{\boldsymbol{w}}^*)' \nabla^2 p(\mathbf{y} \mid (\boldsymbol{\iota}_{\boldsymbol{w}}^*, \boldsymbol{\varpi}_{\boldsymbol{w}}), \mathcal{M}_k)(\boldsymbol{\iota}_{\boldsymbol{w}} - \boldsymbol{\iota}_{\boldsymbol{w}}^*) = 0 \Leftrightarrow$$
$$\forall i \leq k^*, r_i = 0 \text{ and } \forall j \in I_i \;\; f_j(\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*) = 0, \;\; \forall i \geq w_{k^*} + 1, \;\; p_i = 0.$$

Assuming also that if $\boldsymbol{\theta} \notin \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k\}$ then for all functions $h_{\boldsymbol{\theta}}$ which are linear combinations of derivatives of $p(\mathbf{y}|\boldsymbol{\theta})$ of order less than or equal to 2 with respect to $\boldsymbol{\theta}$, and all functions h_1 which are also linear combinations of derivatives of the $p(\mathbf{y}|\boldsymbol{\theta}_j)$'s

$j = 1, 2, .., k$ and its derivatives of order less or equal to 2, then $\alpha h_{\boldsymbol{\theta}} + \beta h_1 = 0$ if and only if $\alpha h_{\boldsymbol{\theta}} = \beta h_1 = 0$.

Extension to non compact spaces: If $\subset \Theta^k$ is not compact then we also assume that for all sequences $\boldsymbol{\theta}_n$ converging to a point in $\partial\Theta^k$ the frontier of Θ^k, considered as a subset of $\Re \cup \{-\infty, \infty\}^p$, $p(\mathbf{y}|\boldsymbol{\theta}_n)$ converges pointwise either to a degenerate function or to a proper density $p(\cdot)$ such that $p(\cdot)$ is linearly independent of any null combinations of $p^*(\mathbf{y}|\boldsymbol{\theta}_i)$, $\nabla p^*(\mathbf{y}|\boldsymbol{\theta}_i)$ and $\nabla^2 p^*(\mathbf{y}|\boldsymbol{\theta}_i)$, $i = 1, ..., k^*$.

7. Proofs

Proof of Theorem 1

Part (i). The result is straightforward. Briefly, $p(\mathbf{y}|\mathcal{M}_k) =$

$$\begin{aligned}
&\int d_\vartheta(\boldsymbol{\vartheta}_k) p(\mathbf{y}|\boldsymbol{\vartheta}_k, \mathcal{M}_k) p^L(\boldsymbol{\vartheta} \mid \mathcal{M}_k) d\boldsymbol{\vartheta}_k \\
&= \int d_\vartheta(\boldsymbol{\vartheta}_k) \frac{p(\mathbf{y}|\boldsymbol{\vartheta}_k, \mathcal{M}_k) p^L(\boldsymbol{\vartheta} \mid \mathcal{M}_k)}{p^L(\mathbf{y}|\mathcal{M}_k)} p^L(\mathbf{y}|\mathcal{M}_k) d\boldsymbol{\vartheta}_k \\
&= p^L(\mathbf{y}|\mathcal{M}_k) E^L(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y}),
\end{aligned}$$

as desired.

Part (ii). Posterior concentration. We need to prove that

$$P^L\left(|d_\vartheta(\boldsymbol{\vartheta}_k|\mathbf{y}) - d_k^*| > \epsilon \mid \mathbf{y}, \mathcal{M}_k\right) \to 0$$

where $d_k^* = 0$ for $k > k^*$ and $d_k^* = d_\vartheta(\boldsymbol{\vartheta}_k^*)$ for $k \leq k^*$. Intuitively, the result follows from the fact that by the L_1 posterior concentration assumption B1 the posterior concentrates on the KL-optimal model $p_k^*(\mathbf{y})$, but for generically identifiable mixtures this corresponds to parameter values satisfying $d(\boldsymbol{\vartheta}_k) = 0$ if $k > k^*$ and $d(\boldsymbol{\vartheta}_k) > 0$ if $k \leq k^*$.

More formally, let A_k be the set of $\boldsymbol{\vartheta}_k \in \Theta_k$ defining $p_k^*(\mathbf{y})$, *i.e.* minimizing KL divergence between the data-generating $p(\mathbf{y} \mid \boldsymbol{\vartheta}_{k^*}^*, \mathcal{M}_{k^*})$ and $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$. Consider first the overfitted model case $k > k^*$, then generic identifiability gives that

$$A_k = \{\boldsymbol{\vartheta}_k \in \Theta_k : \eta_j = 0 \text{ for some } j = 1, \ldots, k \text{ or } \boldsymbol{\theta}_i = \boldsymbol{\theta}_j \text{ for some } i \neq j\}.$$

This implies that for all $\boldsymbol{\vartheta}_k \in A_k$ we have that $d_\vartheta(\boldsymbol{\vartheta}_k) = 0$ and also that the L_1 distance

$$l(\boldsymbol{\vartheta}_k) = \int |p_k^*(\mathbf{y}) - p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)| \, d\mathbf{y} = 0.$$

Thus $d_\vartheta(\boldsymbol{\vartheta}_k) > 0 \Rightarrow \boldsymbol{\vartheta}_k \notin A_k \Rightarrow l(\boldsymbol{\vartheta}_k) > 0$. Given that by assumption $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ and $d_\vartheta(\boldsymbol{\vartheta}_k)$ are continuous in $\boldsymbol{\vartheta}_k$, for all $\epsilon' > 0$ there is an $\epsilon > 0$ such that $d_\vartheta(\boldsymbol{\vartheta}_k) > \epsilon'$ implies $l(\boldsymbol{\vartheta}_k) > \epsilon$ and hence that the probability of the former event must be smaller. That is,

$$P^L\left(d_\vartheta(\boldsymbol{\vartheta}_k) > \epsilon' \mid \mathbf{y}, \mathcal{M}_k\right) \leq P^L(l(\boldsymbol{\vartheta}_k) > \epsilon \mid \mathbf{y}, \mathcal{M}_k)$$

and the right hand side converges to 0 in probability for an arbitrary ϵ by Condition B1, proving the result for the case $k > k^*$.

The proof for the $k \leq k^*$ case proceeds analogously. Briefly, when $k \leq k^*$ generic identifiability gives that $A_k = \{\boldsymbol{\vartheta}_k^*\}$ is a singleton with positive weights $\eta_j^* > 0$ for all $j = 1, \ldots, k$ and $\boldsymbol{\theta}_i^* \neq \boldsymbol{\theta}_j^*$ for $i \neq j$. Thus $d_k^* = d_\vartheta(\boldsymbol{\vartheta}_k^*) > 0$. By continuity of $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ and $d_\vartheta(\boldsymbol{\vartheta}_k)$ with respect to $\boldsymbol{\vartheta}_k$ this implies that for all $\epsilon' > 0$ there exists an $\epsilon > 0$ such that $l(\boldsymbol{\vartheta}_k) < \epsilon \Rightarrow |d_\vartheta(\boldsymbol{\vartheta}_k) - d_k^*| > \epsilon'$, and thus that

$$P^L\left(|d_\vartheta(\boldsymbol{\vartheta}_k) - d_k^*| > \epsilon' \mid \mathbf{y}, \mathcal{M}_k\right) \leq P^L\left(l(\boldsymbol{\vartheta}_k) < \epsilon \mid \mathbf{y}, \mathcal{M}_k\right),$$

where the right hand side converges to 1 in probability by Condition B1, proving the result.

Part (ii). Convergence of $E^L(d_\vartheta(\boldsymbol{\vartheta}_k)|\mathbf{y})$

Consider first the case where $d_\vartheta(\boldsymbol{\vartheta}_k) \in [0, c_k]$ is bounded below some finite constant c_k. Then Part (ii) above and Lemma 2 below give that

$$\begin{aligned} E^L\left(d_\vartheta(\boldsymbol{\vartheta}) \mid \mathbf{y}, \mathcal{M}_k\right) &\xrightarrow{P} 0, \text{ for } k > k^* \\ E^L\left(d_\vartheta(\boldsymbol{\vartheta}) \mid \mathbf{y}, \mathcal{M}_k\right) &\xrightarrow{P} d_k^* > 0, \text{ for } k \leq k^* \end{aligned} \tag{7.1}$$

as we wished to prove. Next consider the MOM prior case $d_\vartheta(\boldsymbol{\vartheta}_k) =$

$$d_\eta(\boldsymbol{\eta}) \frac{1}{C_k} \prod_{1 \leq i < j \leq k} \left((\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A_\Sigma^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right)^t,$$

where $d_\eta(\boldsymbol{\eta})$ is bounded by assumption. From Lemma 1

$$\begin{aligned} E^L\left(d_\vartheta(\boldsymbol{\vartheta}) \mid \mathbf{y}, \mathcal{M}_k\right) &= \int \tilde{d}_\theta(\boldsymbol{\theta}) d_\eta(\boldsymbol{\eta}) \frac{p(\mathbf{y}|\boldsymbol{\vartheta}_k, \mathcal{M}_k)\tilde{p}(\boldsymbol{\vartheta}_k|\mathcal{M}_k)}{p^L(\mathbf{y}|\mathcal{M}_k)} \frac{\tilde{p}^L(\mathbf{y}|\mathcal{M}_k)}{\tilde{p}^L(\mathbf{y}|\mathcal{M}_k)} d\boldsymbol{\vartheta}_k \\ &= \frac{\tilde{p}^L(\mathbf{y}|\mathcal{M}_k)}{p^L(\mathbf{y}|\mathcal{M}_k)} \int \tilde{d}_\theta(\boldsymbol{\theta}) d_\eta(\boldsymbol{\eta}) \tilde{p}^L(\boldsymbol{\vartheta}_k|\mathbf{y}, \mathcal{M}_k) d\boldsymbol{\vartheta}_k, \end{aligned} \tag{7.2}$$

where $\tilde{d}_\theta(\boldsymbol{\theta}) d_\eta(\boldsymbol{\eta})$ is bounded and hence by Part (ii) and Lemma 2 the integral in (7.2) converges to 0 in probability when $k > k^*$ and to a non-zero finite constant when $k \leq k^*$. Therefore it suffices to show that $\tilde{p}^L(\mathbf{y}|\mathcal{M}_k)/p^L(\mathbf{y}|\mathcal{M}_k)$ is bounded in probability, as this

would then immediately imply the desired result (7.1). From Lemma 1 $\tilde{p}^L(\mathbf{y}|\mathcal{M}_k) =$

$$\begin{aligned}
&\int p(\mathbf{y}|\boldsymbol{\vartheta}_k, \mathcal{M}_k)\tilde{p}^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k)d\boldsymbol{\vartheta}_k = \\
&\int p(\mathbf{y}|\boldsymbol{\vartheta}_k, \mathcal{M}_k)p^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k)\frac{\tilde{p}^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k)}{p^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k)}d\boldsymbol{\vartheta}_k = \\
&\int p(\mathbf{y}|\boldsymbol{\vartheta}_k, \mathcal{M}_k)p^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k)\prod_{j=1}^{k}\frac{N(\boldsymbol{\mu}_j; \mathbf{0}, (1+\epsilon)g\Sigma_j)}{N(\boldsymbol{\mu}_j; \mathbf{0}, g\Sigma_j)}d\boldsymbol{\vartheta}_k \\
&\int p(\mathbf{y}|\boldsymbol{\vartheta}_k, \mathcal{M}_k)p^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k)\frac{1}{(1+\epsilon)^{kp/2}}\exp\left\{\frac{1}{2g}\sum_{j=1}^{k}\boldsymbol{\mu}_j' A_\Sigma^{-1}\boldsymbol{\mu}_j\frac{\epsilon}{1+\epsilon}\right\}d\boldsymbol{\vartheta}_k \\
&= \frac{p^L(\mathbf{y} \mid \mathcal{M}_k)}{(1+\epsilon)^{kp/2}}E^L\left(\exp\left\{\frac{1}{2g}\sum_{j=1}^{k}\boldsymbol{\mu}_j' A_\Sigma^{-1}\boldsymbol{\mu}_j\frac{\epsilon}{1+\epsilon}\right\} \mid \mathbf{y}, \mathcal{M}_k\right) \\
&\geq \frac{p^L(\mathbf{y}|\mathcal{M}_k)}{(1+\epsilon)^{kp/2}}, \qquad (7.3)
\end{aligned}$$

thus $\tilde{p}^L(\mathbf{y}|\mathcal{M}_k)/p^L(\mathbf{y}|\mathcal{M}_k) \geq \frac{1}{(1+\epsilon)^{kp/2}}$. From (7.2) this implies that when $k \leq k^*$ we obtain $E^L\left(d_\vartheta(\boldsymbol{\vartheta}) \mid \mathbf{y}, \mathcal{M}_k\right) \xrightarrow{P} d_k^* > 0$. Further, by Condition B3 the $E^L()$ term in (7.3) is bounded above in probability when $k > k^*$, implying that $E^L\left(d_\vartheta(\boldsymbol{\vartheta}) \mid \mathbf{y}, \mathcal{M}_k\right) \xrightarrow{P} 0$. □

Part (iii).

By assumption $p(\boldsymbol{\eta}|\mathcal{M}_k) = \text{Dir}(\boldsymbol{\eta}; q) \propto d_\eta(\boldsymbol{\eta})\text{Dir}(\boldsymbol{\eta}; q-r)$, where $d_\eta(\boldsymbol{\eta}) = \prod_{j=1}^{k}\eta_j^r$ and $q > 1$, $q - r < 1$. Consider the particular choice $q - r < \dim(\Theta)/2$ and without loss of generality let $k^* + 1, \ldots, k$ be the labels for the spurious components. Theorem 1 in Rousseau and Mengersen (2011) showed that under the assumed A1-A4 and a further condition A5 trivially satisfied by $p^L(\boldsymbol{\eta}|\mathcal{M}_k) = \text{Dir}(\boldsymbol{\eta}; q-r)$ the corresponding posterior distribution of the spurious weights concentrates around 0, specifically

$$P^L\left(\sum_{j=k^*+1}^{k}\eta_j > n^{-\frac{1}{2}+\tilde{\epsilon}}|\mathbf{y}, \mathcal{M}_k\right) \to 0 \qquad (7.4)$$

in probability for all $\tilde{\epsilon} > 0$ as $n \to \infty$. Now, the fact that the geometric mean is smaller than the arithmetic mean gives that

$$(k - k^*)\left(\prod_{j=k^*+1}^{k}\eta_j\right)^{\frac{1}{k-k^*}} \leq \sum_{j=k^*+1}^{k}\eta_j,$$

and thus

$$P^L\left(\sum_{j=k^*+1}^{k}\eta_j > n^{-\frac{1}{2}+\tilde{\epsilon}}|\mathbf{y},\mathcal{M}_k\right)\geq$$

$$P^L\left((k-k^*)\left(\prod_{j=k^*+1}^{k}\eta_j\right)^{\frac{1}{k-k^*}} > n^{-\frac{1}{2}+\tilde{\epsilon}}|\mathbf{y},\mathcal{M}_k\right)=$$

$$P^L\left(\prod_{j=k^*+1}^{k}\eta_j^r > \frac{1}{(k-k^*)^r}n^{-\frac{r(k-k^*)}{2}+\epsilon}|\mathbf{y},\mathcal{M}_k\right), \tag{7.5}$$

where $\epsilon = r(k-k^*)\tilde{\epsilon}$ is a constant. Thus (7.4) implies that (7.5) also converges to 0 in probability. Finally, given that by assumption $d_\vartheta(\boldsymbol{\vartheta}) = d_\theta(\boldsymbol{\theta})d_\eta(\boldsymbol{\eta}) \leq c_k\prod_{j=k^*+1}^{k}\eta_j^r$ we obtain

$$P^L\left(d_\vartheta(\boldsymbol{\vartheta}) > n^{-\frac{r(k-k^*)}{2}+\epsilon}|\mathbf{y},\mathcal{M}_k\right) \leq P^L\left(\prod_{j=k^*+1}^{k}\eta_j^r > \frac{1}{c_k}n^{-\frac{r(k-k^*)}{2}+\epsilon}|\mathbf{y},\mathcal{M}_k\right), \tag{7.6}$$

where the right hand side converges in probability to 0 given that (7.5) converges to 0 in probability and c_k, k, k^*, r are finite constants. As mentioned earlier this result holds for any $r > 0$ satisfying $q - r < \dim(\Theta)/2$, in particular we may set $q - r = \delta < \dim(\Theta)/2$ (where $\delta > 0$ can be arbitrarily small) so that plugging $r = q - \delta$ into the left hand side of (7.6) gives the desired result. □

Lemma 1. *Let $p(\boldsymbol{\vartheta}_k|\mathcal{M}_k) = d_\theta(\boldsymbol{\theta})p^L(\boldsymbol{\theta}|\mathcal{M}_k)p(\boldsymbol{\eta}|\mathcal{M}_k)$ be the MOM prior in (2.4). Then $p(\boldsymbol{\vartheta}_k|\mathcal{M}_k) = \tilde{d}_\theta(\boldsymbol{\theta})\tilde{p}^L(\boldsymbol{\theta}|\mathcal{M}_k)p(\boldsymbol{\eta}|\mathcal{M}_k)$, where $\tilde{d}_\theta(\boldsymbol{\theta}) \leq c_k$ for some finite c_k,*

$$\tilde{p}^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k) = \prod_{j=1}^{k} N\left(\boldsymbol{\mu}_j|\mathbf{0},(1+\epsilon)gA_\Sigma\right),$$

and $\epsilon \in (0,1)$ is an arbitrary constant.

Proof. The MOM prior has an unbounded penalty

$$d_\theta(\boldsymbol{\theta}) = \frac{1}{C_k}\prod_{1\leq i<j\leq k}\left((\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)'A_\Sigma^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)/g\right)^t,$$

however we may rewrite $d_\theta(\boldsymbol{\theta})p^L(\boldsymbol{\theta}|\mathcal{M}_k)$

$$=d_\theta(\boldsymbol{\theta})\prod_{j=1}^{k} N\left(\boldsymbol{\mu}_j|\mathbf{0},gA_\Sigma\right)\frac{N\left(\boldsymbol{\mu}_j|\mathbf{0},(1+\epsilon)gA_\Sigma\right)}{N\left(\boldsymbol{\mu}_j|\mathbf{0},(1+\epsilon)gA_\Sigma\right)},$$

$$=\tilde{d}_\theta(\boldsymbol{\theta})\prod_{j=1}^{k} N\left(\boldsymbol{\mu}_j|\mathbf{0},(1+\epsilon)gA_\Sigma\right), \tag{7.7}$$

where $\epsilon \in (0,1)$ is an arbitrary constant and $\tilde{d}_\theta(\boldsymbol{\theta}) =$

$$d_\theta(\boldsymbol{\theta}) \prod_{j=1}^{k} \frac{N\left(\boldsymbol{\mu}_j|\mathbf{0}, gA_\Sigma\right)}{N\left(\boldsymbol{\mu}_j|\mathbf{0}, (1+\epsilon)gA_\Sigma\right)} = d_\theta(\boldsymbol{\theta}) \prod_{j=1}^{k} (1+\epsilon)^{1/2} \exp\left\{-\frac{1}{2}\frac{\epsilon \boldsymbol{\mu}_j^{'} A_\Sigma^{-1} \boldsymbol{\mu}_j}{(1+\epsilon)g}\right\}.$$

The fact that $\tilde{d}_\theta(\boldsymbol{\theta})$ is bounded follows from the fact that the product term is a Normal kernel and hence bounded, whereas $d_\theta(\boldsymbol{\theta})$ can only become unbounded when $\boldsymbol{\mu}_j A_\Sigma^{-1} \boldsymbol{\mu}_j \to \infty$ for some j, but this polynomial increase is countered by the exponential decrease in $\exp\left\{-\frac{1}{2}\frac{\epsilon \boldsymbol{\mu}_j^{'} A_\Sigma^{-1} \boldsymbol{\mu}_j}{(1+\epsilon)g}\right\}$. □

Lemma 2. *Let $d_\vartheta(\boldsymbol{\vartheta}_k) \in [0, c_k]$ be a bounded continuous function in $\boldsymbol{\vartheta}_k$, where c_k is a finite constant. Let*

$$g_k(\mathbf{y}) = E^L(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y}, \mathcal{M}_k) = \int d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k|\mathbf{y}, \mathcal{M}_k) d\boldsymbol{\vartheta}_k.$$

If for any $\epsilon > 0$ we have that $P^L(d_\vartheta(\boldsymbol{\vartheta}) > \epsilon|\mathbf{y}, \mathcal{M}_k) \xrightarrow{P} 0$ then $g_k(\mathbf{y}) \xrightarrow{P} 0$. Alternatively, if there exists some $d_k^* > 0$ such that for any $\epsilon > 0$ $P^L(|d_\vartheta(\boldsymbol{\vartheta}_k) - d_k^*| > \epsilon|\mathbf{y}, \mathcal{M}_k) \xrightarrow{P} 1$, then $g_k(\mathbf{y}) \xrightarrow{P} d_k^*$.

Proof. Consider the case $P^L(d_\vartheta(\boldsymbol{\vartheta}) > \epsilon|\mathbf{y}, \mathcal{M}_k) \xrightarrow{P} 0$, then $g_k(\mathbf{y}) =$

$$\begin{aligned}
&\int_{d_\vartheta(\boldsymbol{\vartheta}_k)<\epsilon} d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k|\mathbf{y}, \mathcal{M}_k) d\boldsymbol{\vartheta}_k + \int_{d_\vartheta(\boldsymbol{\vartheta}_k)>\epsilon} d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k|\mathbf{y}, \mathcal{M}_k) d\boldsymbol{\vartheta}_k \\
&\leq \epsilon P^L(d_\vartheta(\boldsymbol{\vartheta}_k) < \epsilon|\mathbf{y}, \mathcal{M}_k) + c_k P^L(d_\vartheta(\boldsymbol{\vartheta}_k) > \epsilon|\mathbf{y}, \mathcal{M}_k) \\
&\leq \epsilon + c_k P^L(d_\vartheta(\boldsymbol{\vartheta}_k) > \epsilon|\mathbf{y}, \mathcal{M}_k) \xrightarrow{P} \epsilon,
\end{aligned}$$

where $\epsilon > 0$ is arbitrarily small. Hence $g_k(\mathbf{y}) \xrightarrow{P} 0$.

Next consider the case $P^L(|d_\vartheta(\boldsymbol{\vartheta}_k) - d_k^*| > \epsilon|\mathbf{y}, \mathcal{M}_k) \xrightarrow{P} 1$. Then

$$\begin{aligned}
g_k(\mathbf{y}) &> \int_{d_\vartheta(\boldsymbol{\vartheta}_k)>d_k^*-\epsilon} d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k|\mathbf{y}) d\boldsymbol{\vartheta}_k \\
&\geq (d_k^* - \epsilon) P^L\left(d_\vartheta(\boldsymbol{\vartheta}_k) > d_k^* - \epsilon|\mathbf{y}, \mathcal{M}_k\right) \xrightarrow{P} d_k^* - \epsilon,
\end{aligned}$$

and analogously $g_k(\mathbf{y}) =$

$$\begin{aligned}
&\int_{d_\vartheta(\boldsymbol{\vartheta}_k)<d_k^*+\epsilon} d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k|\mathbf{y}, \mathcal{M}_k) d\boldsymbol{\vartheta}_k + \int_{d_\vartheta(\boldsymbol{\vartheta}_k)>d_k^*+\epsilon} d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k|\mathbf{y}, \mathcal{M}_k) d\boldsymbol{\vartheta}_k \\
&\leq (d_k^* + \epsilon) + c_k P^L(d_\vartheta(\boldsymbol{\vartheta}_k) > d_k^* + \epsilon|\mathbf{y}, \mathcal{M}_k) \xrightarrow{P} d_k^* + \epsilon,
\end{aligned}$$

for any $\epsilon > 0$ and hence $g_k(\mathbf{y}) \xrightarrow{P} d_k^*$. □

Proof of Corollary 1

In order to compute the normalization, C_k we need to find the expectation:

$$C_k = E\left(\prod_{1\leq i<j\leq k}\left(\frac{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)' A_\Sigma^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}{g}\right)\right).$$

with respect to $(\boldsymbol{\mu}_1,...,\boldsymbol{\mu}_k \sim N(\mathbf{0}, A_\Sigma)$. Moreover consider the Cholesky decomposition $A_\Sigma = \boldsymbol{L}\boldsymbol{L}'$ where $A_\Sigma^{-1} = (\boldsymbol{L}')^{-1}\boldsymbol{L}^{-1}$, by setting $\sqrt{g}\boldsymbol{L}\boldsymbol{\mu}_j^* = \boldsymbol{\mu}_j$ the jacobian of the transformation is the determinant of the block diagonal matrix:

$$|J(\boldsymbol{\mu}_1^*,...,\boldsymbol{\mu}_k^*)| = \left|\begin{pmatrix} \sqrt{g}\boldsymbol{L} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{g}\boldsymbol{L} \end{pmatrix}\right| = g^{k/2}(\det(\boldsymbol{L}))^k,$$

where $(\det(\boldsymbol{L}))^k = (\det(A_\Sigma))^{k/2}$. The normalization constant C_k can be found by using the following expectation

$$C_k = E\left(\prod_{1\leq i<j\leq k}((\boldsymbol{\mu}_i^*-\boldsymbol{\mu}_j^*)'(\boldsymbol{\mu}_i^*-\boldsymbol{\mu}_j^*))\right), \tag{7.8}$$

where $\boldsymbol{\mu}_k^* \sim N_p(\boldsymbol{\mu}_k^*|\mathbf{0}, \boldsymbol{I_p})$.

To obtain the result we apply the adapted Proposition 4 in Kan (2006) to the $p\times k$ vector $\boldsymbol{\mu}^* = (\boldsymbol{\mu}_1^*,...,\boldsymbol{\mu}_k^*)$, where k is the number of components and $\boldsymbol{\mu}_j^* \in \Re^p$ for $j = 1,\ldots,k$, which for convenience we reproduce below as Proposition 1.

Proposition 1. *Suppose* $\boldsymbol{\mu}^* = (\mu_1^*,...,\mu_k^*)' \sim N_k(\mathbf{0}, \boldsymbol{I}_k)$, *for symmetric matrices* $A_{(1,2)},...,A_{(k-1,k)}$, *we have*

$$E\left(\prod_{1\leq i<j\leq k}(\boldsymbol{\mu}^{*'}A_{(i,j)}\boldsymbol{\mu}^*)\right) = \frac{1}{s!}\sum_{\upsilon_{(1,2)}=0}^{1}\ldots\sum_{\upsilon_{(k-1,k)}=0}^{1}(-1)^{\sum_{i,j}^{\binom{k}{2}}\upsilon_{(i,j)}}\mathcal{Q}_s(B_\upsilon), \tag{7.9}$$

where $s = \binom{k}{2}$, $B_\upsilon = (\frac{1}{2}-\upsilon_{(1,2)})A_{(1,2)}+,...,+(\frac{1}{2}-\upsilon_{(k-1,k)})A_{(k-1,k)}$ *and* $\mathcal{Q}_s(B_\upsilon)$ *is given by the recursive equation:* $\mathcal{Q}_s(B_\upsilon) = s!2^s d_s(B_\upsilon)$ *where* $d_s(B_\upsilon) = \frac{1}{2s}\sum_{i=1}^{s} tr(B_\upsilon^i)d_{s-i}(B_\upsilon)$ *and* $d_0(B_\upsilon) = 1$ *and* $A_{(i,j)}$ *is a* $pk\times pk$ *matrix* (l,m) *element*

$$\begin{cases} a_{ll} = 1, \quad l = 1+p(i-1)...p_i \quad and \quad l = 1+p(j-1)...p_j. \\ a_{lm} = a_{ml} = -1, \quad (l,m) = (1+p(i-1), 1+p(j-1))...(pi,pj). \\ a_{lm} = 0 \quad otherwise. \end{cases}$$

We define now the $A_{(1,2)},...,A_{(k-1,k)}$ matrices with dimensions $pk\times pk$. These matrices can be found using $p*p$ identity matrices in the diagonal blocks corresponding to the i and j components minus the identity matrix in the "cross-blocks" corresponding to (i,j). Finally using the $A_{(i,j)}$ matrices, B_υ can be expressed as a $pk\times pk$ matrix with element

(l,m) as follows

$$\begin{cases} b_{ll} = \frac{1}{2}(k-1) - \sum_{i<j} \upsilon_{(i,j)}, \quad l = 1 + p(i-1)...p_i \quad \text{and} \quad l = 1 + p(j-1)...p_j. \\ b_{lm} = b_{ml} = -\frac{1}{2} + \sum_{i<j} \upsilon_{(i,j)}, \quad (l,m) = (1 + p(i-1), 1 + p(j-1))...(pi, pj). \end{cases}$$

□

Proof of Corollary 2

Using the Corollary 2.2 in Lu and Richards (1993), if $z > -1/n$, then

$$(2\pi)^{-n/2} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \prod_{1\leq i<j\leq n} (x_i - x_j)^{2z} \prod_{j=1}^{n} \exp\{-x_j^2/2\} dx_j = \prod_{j=1}^{n} \frac{\Gamma(jz+1)}{\Gamma(z+1)}, \tag{7.10}$$

and using $x_i = (\mu_i - m)/(\sqrt{a_{\sigma^2} g})$ with $i = 1, ..., k$, we have that the normalization constant is

$$C_k = E_{\mu_1,...,\mu_k|a_{\sigma^2}} \left(\prod_{1\leq i<j\leq k} \left(\frac{\mu_i - \mu_j}{\sqrt{a_{\sigma^2} g}} \right)^{2t} \right) = \prod_{j=1}^{k} \frac{\Gamma(jt+1)}{\Gamma(t+1)}. \tag{7.11}$$

□

8. EM algorithm for multivariate Normal mixtures under MOM-Wishart-Dirichlet priors

The complete-data posterior can be written as follows

(8.1)

$$p(\boldsymbol{\vartheta}_k | \mathbf{y}, \boldsymbol{z}, \mathcal{M}_k) = \prod_{j=1}^{k} \prod_{i=1}^{n} (\eta_j N(\mathbf{y}|\boldsymbol{\mu_j}, \boldsymbol{\Sigma}_j))^{z_{ij}} N(\boldsymbol{\mu}_j | \mathbf{0}, g A_\Sigma) \, \text{Wishart}(\Sigma_j^{-1} | \nu, S) \text{Dir}(\boldsymbol{\eta}; q).$$

For the E-step and the t-th iteration we compute the expectation of $\bar{z}_{ij}^{(t)} = p(z_i = j | \mathbf{y}_i, \boldsymbol{\vartheta}_j^{(t-1)})$ the latent cluster allocations given $\boldsymbol{\eta}^{(t-1)}$, $\boldsymbol{\mu}_1^{(t-1)}$,...,$\boldsymbol{\mu}_k^{(t-1)}$ and $\boldsymbol{\Sigma}_1^{(t-1)}, ..., \boldsymbol{\Sigma}_k^{(t-1)}$, as follows:

$$\bar{z}_{ij}^{(t)} = \frac{\eta_j^{(t-1)} N(\mathbf{y}_i; \boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})}{\sum_{j=1}^{k} \eta_j^{(t-1)} N(\mathbf{y}_i; \boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})} \tag{8.2}$$

In the M-step and the t-th iteration we find the maximizers $\boldsymbol{\eta}^{(t)}$, $\boldsymbol{\mu}_1^{(t)}$,...,$\boldsymbol{\mu}_K^{(t)}$ and $\boldsymbol{\Sigma}_1^{(t)}, ..., \boldsymbol{\Sigma}_k^{(t)}$, given the expectations of the missing data, of the following function:

(8.3)

$$\begin{aligned}\log(p(\boldsymbol{\vartheta}_k|\mathbf{y}, \bar{z}_{ij}, \mathcal{M}_k)) = &\sum_{j=1}^{k} n_j \log(\eta_j) + \sum_{j=1}^{k}\sum_{i=1}^{n} \bar{z}_{ij} \log(N(\boldsymbol{y}_i|\boldsymbol{\mu_j}, \boldsymbol{\Sigma}_j)) + \sum_{j=1}^{k} \log(N(\boldsymbol{\mu}_j|\mathbf{0}, gA_\Sigma)) \\ &+ \sum_{1\le i<j\le k} \log((\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A_\Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)) + \sum_{j=1}^{k} \log(\text{Wishart}(\Sigma_j^{-1}|\nu, S)) \\ &+ \log(\text{Dir}(\boldsymbol{\eta}; q)) + \text{Constant},\end{aligned}$$

with $n_j^{(t)} = \sum_{i=1}^{n} \bar{z}_{ij}^{(t)}$. We successively update $\boldsymbol{\eta}^{(t)}$, $\boldsymbol{\mu}_1^{(t)}$,...,$\boldsymbol{\mu}_K^{(t)}$ and $\boldsymbol{\Sigma}_1^{(t)}, ..., \boldsymbol{\Sigma}_k^{(t)}$ in a fashion that guarantees that (8.3) increases at each step. The M-step for $\eta_j^{(t)}$ is computed by using

$$\eta_j^{(t)} = \frac{n_j^{(t)} + q - 1}{n + k(q-1)}, \tag{8.4}$$

which maximizes (8.3) with respect to $\boldsymbol{\eta}$ conditional on the current $\boldsymbol{\mu}_1^{(t-1)}$,...,$\boldsymbol{\mu}_K^{(t-1)}$ and $\boldsymbol{\Sigma}_1^{(t-1)}, ..., \boldsymbol{\Sigma}_k^{(t-1)}$. For $\boldsymbol{\mu}_j^{(t)}$ let

$$\xi(\boldsymbol{\mu}_j^{(t)}) = \sum_{i\neq j} \log(\boldsymbol{C}_{ij}^{(t)'} A_{\Sigma^{(t-1)}}^{-1} \boldsymbol{C}_{ij}^{(t)}) - \frac{1}{2g} \boldsymbol{\mu}_j^{'(t)} A_{\Sigma^{(t-1)}}^{-1} \boldsymbol{\mu}_j^{(t)} - \frac{1}{2} \sum_{i=1}^{n} \bar{z}_{ij}^{(t)} (\boldsymbol{y}_i - \boldsymbol{\mu}_j^{(t)})' A_{\Sigma^{(t-1)}}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_j^{(t)}),$$

be the corresponding target where $\boldsymbol{C}_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$. The first derivative of $\xi(\boldsymbol{\mu}_j^{(t)})$ is

$$\nabla\xi(\boldsymbol{\mu}_j^{(t)}) = -2 \sum_{i\neq j} \frac{A_{\Sigma^{(t-1)}}^{-1} \boldsymbol{C}_{ij}^{(t)}}{\boldsymbol{C}_{ij}^{(t)'} A_{\Sigma^{(t-1)}}^{-1} \boldsymbol{C}_{ij}^{(t)}} - \frac{1}{g} (A_{\Sigma^{(t-1)}}^{-1} \boldsymbol{\mu}_j^{(t)}) - \sum_{i=1}^{n} \bar{z}_{ij}^{(t)} (A_{\Sigma^{(t-1)}}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_j^{(t)})).$$

Because an analytic solution of $\nabla\xi(\boldsymbol{\mu}_j^{(t)}) = \mathbf{0}$ in terms of $\boldsymbol{\mu}_j^{(t)}$ is not feasible we resort to a first order Taylor's approximation for $-2\sum_{i\neq j}(A_{\Sigma^{(t-1)}}^{-1} \boldsymbol{C}_{ij}^{(t)})/(\boldsymbol{C}_{ij}^{(t)'} A_{\Sigma^{(t-1)}}^{-1} \boldsymbol{C}_{ij}^{(t)})$ around $\boldsymbol{\mu}_j^{(t-1)}$ and we now compute the M-step for $\boldsymbol{\mu}_j^*$ as follows:

$$\begin{aligned}\boldsymbol{\mu}_j^* = &\left(\Sigma_j^{-1(t-1)} n_j^{(t)} + A_{\Sigma^{(t-1)}}^{-1} \left(\frac{1}{g} + \sum_{j\neq k} \frac{2}{d_{ij}^{(t-1)}} \right) \right)^{-1} \\ &\times \left(\Sigma^{-1(t-1)} n_j^{(t)} \bar{\mathbf{y}}_j^{(t)} + A_{\Sigma^{(t-1)}}^{-1} \left(\sum_{i\neq j} \frac{\boldsymbol{\mu}_j^{(t-1)} - (\boldsymbol{\mu}_i^{(t-1)} - \boldsymbol{\mu}_j^{(t-1)})}{d_{ij}^{(t-1)}} \right) \right),\end{aligned} \tag{8.5}$$

with $d_{ij}^{(t-1)} = (\boldsymbol{\mu}_i^{(t-1)} - \boldsymbol{\mu}_j^{(t-1)})' A_{\Sigma^{(t-1)}}^{-1} (\boldsymbol{\mu}_i^{(t-1)} - \boldsymbol{\mu}_j^{(t-1)})$. If $\xi(\boldsymbol{\mu}_j^*) > \xi(\boldsymbol{\mu}_j^{(t-1)})$ set $\boldsymbol{\mu}_j^{(t)} = \boldsymbol{\mu}_j^*$, else take the gradient step in supplementary algorithm 3. Let $\xi(\Sigma_j^{(t)})$ be the corresponding

target for $\Sigma_j^{(t)}$. Due to the penalty term $\sum_{i\neq j}\log(\boldsymbol{\mu}_i^{(t)}-\boldsymbol{\mu}_j^{(t)})' A^{-1}_{\Sigma^{(t)}}(\boldsymbol{\mu}_i^{(t)}-\boldsymbol{\mu}_j^{(t)})$ and the term $-\frac{1}{2}\log(|A^{-1}_{\Sigma^{(t)}}|)$ an analytic solution of $\nabla\xi(\Sigma_j^{(t)})$ in terms of $\Sigma_j^{(t)}$ is not feasible. Therefore we use a first order Taylor's approximation around the previous iteration $(t-1)$ for the logarithm of this expression, so that

$$\sum_{i\neq j}\log(\boldsymbol{\mu}_i^{(t)}-\boldsymbol{\mu}_j^{(t)})' A^{-1}_{\Sigma^{(t)}}(\boldsymbol{\mu}_i^{(t)}-\boldsymbol{\mu}_j^{(t)}) - \frac{1}{2}\log(|A^{-1}_{\Sigma^{(t)}}|) \approx$$

$$\sum_{i\neq j}\frac{(\boldsymbol{\mu}_i^{(t)}-\boldsymbol{\mu}_j^{(t)})' A^{-1}_{\Sigma^{(t)}}(\boldsymbol{\mu}_i^{(t)}-\boldsymbol{\mu}_j^{(t)})}{(\boldsymbol{\mu}_i^{(t-1)}-\boldsymbol{\mu}_j^{(t-1)})' A^{-1}_{\Sigma^{(t-1)}}(\boldsymbol{\mu}_i^{(t-1)}-\boldsymbol{\mu}_j^{(t-1)})} - \frac{1}{2}\log(|\Sigma_j^{(t)}|).$$

Note that when a common variance-covariance is considered on all component densities, i.e. $A^{-1}_{\boldsymbol{\Sigma}_{\boldsymbol{K}}^{(t)}} = \boldsymbol{\Sigma}^{-1(t)}$, we only need to use a Taylor's approximation of the penalty term around the previous iteration $(t-1)$. So we compute the M-step for Σ_j^* using $(\nu-p+n_j^{(t)})\Sigma_j^{(t)} =$

$$S^{-1} + \frac{\boldsymbol{\mu}_j^{(t)}(\boldsymbol{\mu}_j^{(t)})'}{kg} + \sum_{i=1}^{n}\bar{z}_{ij}^{(t)}(\mathbf{y}_i-\boldsymbol{\mu}_j^{(t)})(\mathbf{y}_i-\boldsymbol{\mu}_j^{(t)})' - \frac{1}{k}\sum_{i\neq j}\frac{2(\boldsymbol{\mu}_j^{(t)}-\boldsymbol{\mu}_k^{(t)})(\boldsymbol{\mu}_j^{(t)}-\boldsymbol{\mu}_k^{(t)})'}{d_{ij}^{(t-1)}}.$$

If $\xi(\Sigma_j^*) > \xi(\Sigma_j^{(t-1)})$ set $\Sigma_j^{(t)} = \Sigma_j^*$, else take the gradient step conditional to obtain Σ_j^* positive define in supplementary algorithm 3.

Supplementary Algorithm 3: Gradient Ascend algorithm.

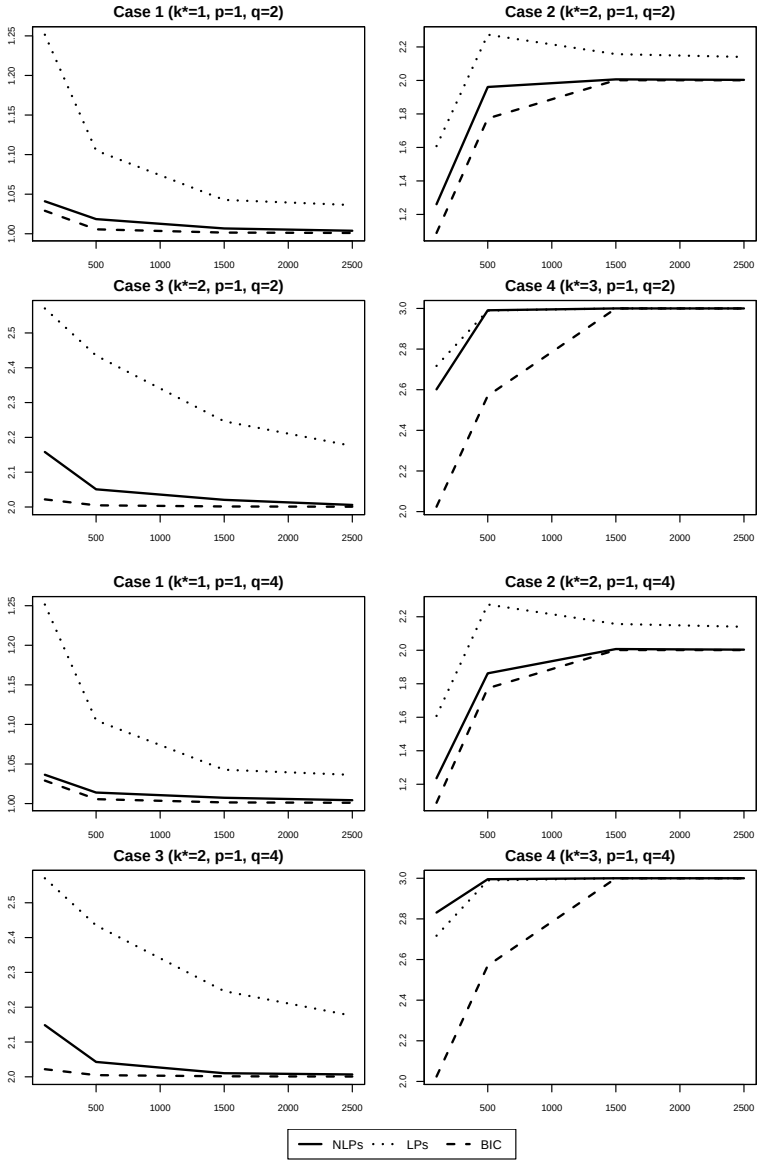
1 Initialization $\boldsymbol{\zeta} = \boldsymbol{\zeta}^*$, $\bar{k} = \sqrt{\frac{\|\boldsymbol{\zeta}^* - \boldsymbol{\zeta}^{(t-1)}\|}{\nabla\xi(\boldsymbol{\zeta}^{(t-1)})}}$ and $h = 0$;
2 **while** $(\xi(\boldsymbol{\zeta}^{(t-1)}) > \xi(\boldsymbol{\zeta}^*))$ **do**
3 $\quad \boldsymbol{\zeta}^* = \boldsymbol{\zeta}^{(t-1)} + \frac{\bar{k}}{2^h}\nabla\xi(\boldsymbol{\zeta}^{(t-1)})$;
4 $\quad h = h+1$
5 **end**
6 $\boldsymbol{\zeta}^{(t)} = \boldsymbol{\zeta}^*$

9. Supplementary results

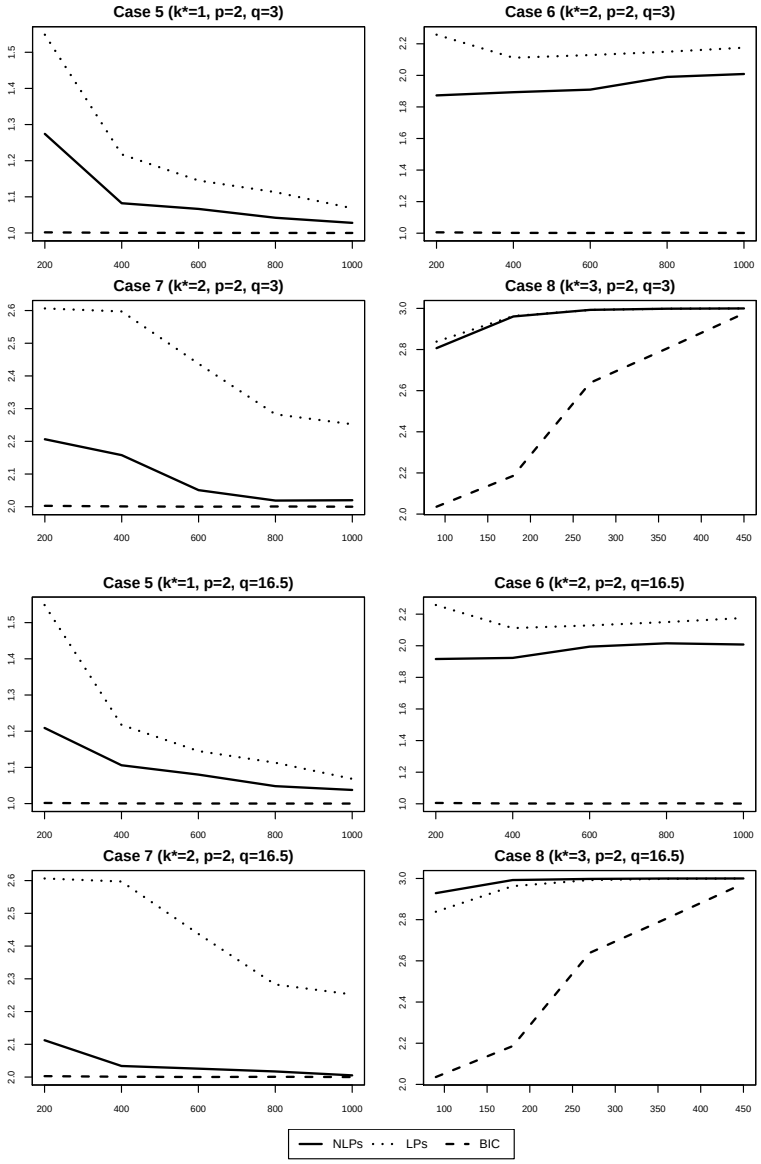
We provide additional results for the simulation study in Section 4.1. Regarding the univariate Normal mixtures in Cases 1-4, the four top panels in Supplementary Figure 7 show the posterior expected number of components given by $E(k \mid \mathbf{y}) = P(\mathcal{M}_1 \mid \mathbf{y}) + 2P(\mathcal{M}_2 \mid \mathbf{y}) + 3P(\mathcal{M}_3 \mid \mathbf{y})$ for $q = 2$ and $P(\kappa < 4) = 0.05$. The four bottom panels show analogous results for $q = 4$ and $P(\kappa < 4) = 0.05$, showing that the findings are fairly robust to mild deviations from our default q.

Regarding the bivariate Normal mixtures in Cases 5-8, the four top panels in Supplementary Figure 8 shows $E(k \mid \mathbf{y})$ for $q = 3$ and $P(\kappa < 4) = 0.05$. The four bottom panels show the same results for $q = 16.5$ (a value recommended in Frühwirth-Schnatter (2006) and Mengersen et al. (2011), Chapter 10) and $P(\kappa < 4) = 0.05$, showing again that the findings are fairly robust to mild deviations from our recommended prior setting.

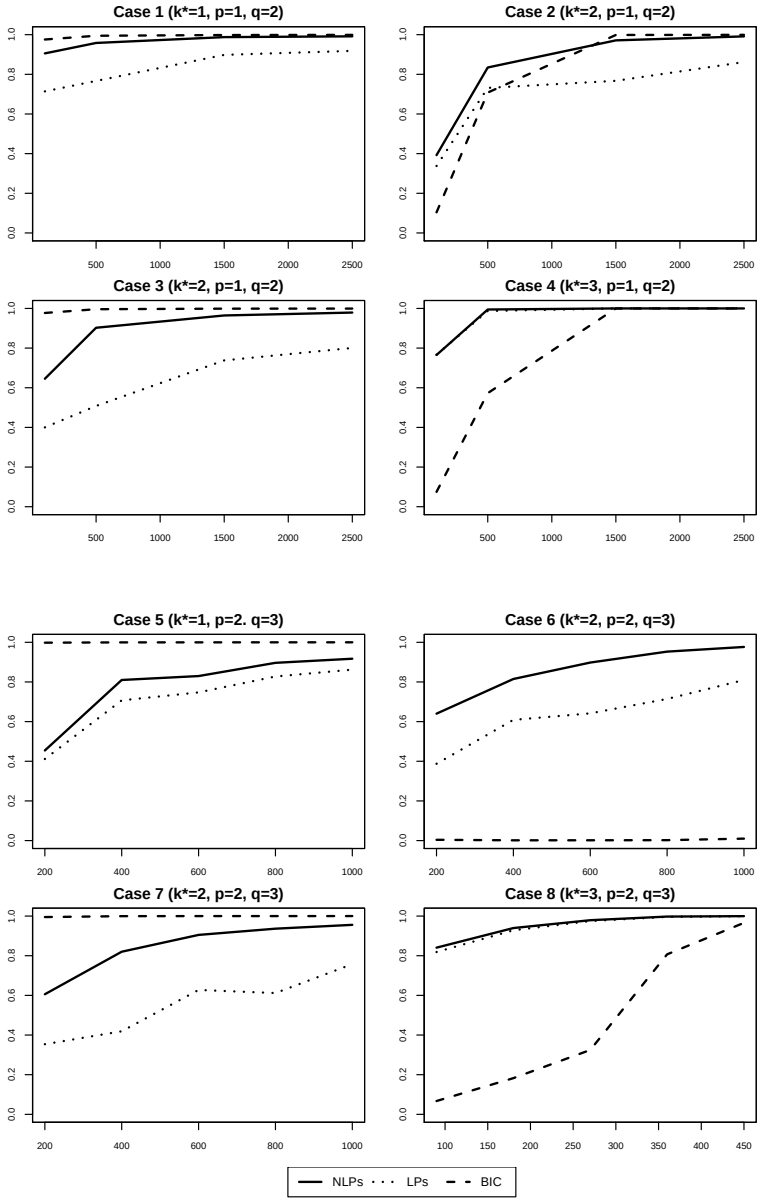
Finally, to assess sensitivity to the prior elicitation of g, Supplementary Figure 9 shows the average posterior probability $P(\mathcal{M}_{k^*} \mid \mathbf{y})$ for Cases 1-8 with $P(\kappa < 4) = 0.1$ and q set as in Figure 3. Although the results are largely similar to those in Figure 3, the benefits in parsimony enforcement are somewhat reduced in some situations (*e.g.* Case 5), indicating that $P(\kappa < 4 \mid g, \mathcal{M}_K) = 0.05$ may be slightly preferable to 0.1 to achieve a better balance between parsimony and detection power.



SUPPLEMENTARY FIGURE 7. Posterior expected model size in simulation study. Sensitivity to q in univariate Cases 1-4.



SUPPLEMENTARY FIGURE 8. Posterior expected model size in simulation study. Sensitivity to q in bivariate Cases 5-8.



SUPPLEMENTARY FIGURE 9. Average $P(\mathcal{M}_{k^*} \mid \mathbf{y})$ in simulation study (Cases 1-8) under $P(\kappa < 4 \mid \mathcal{M}_k) = 0.1$.

Supplementary Table 1 provides more detailed results for the misspecified Normal model (Section 4.2). It indicates the posterior probability of 11 models with $k = 1, ..., 6$ components, for each k, considering either homogeneous ($\Sigma_j = \Sigma$) or heterogeneous ($\Sigma_i \neq \Sigma_j$) covariance matrices. The table also gives the posterior modes for the weights $\boldsymbol{\eta}$ under each model, *i.e.* obtained from $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\theta}}) = \arg\max_{\eta,\theta} p(\boldsymbol{\eta}, \boldsymbol{\theta} \mid \mathbf{y}, \mathcal{M}_k)$. The model with highest posterior is indicated in bold face. Supplementary Table 2 shows analogous results for the Faithful data (Section 4.3) and Supplementary Table 3 for the Iris data (Section 4.4).

As an alternative to formal Bayesian model selection suppose one fits a model with a large number of components ($k = 6$ in our examples) to successively discard those deemed unnecessary. One strategy to discard components is to set a threshold on the estimated $\hat{\boldsymbol{\eta}}$, which results in the addition of spurious components. An alternative illustrated in Supplementary Tables 4-5 is to describe the number m of non-empty components (no allocated observations) at each MCMC iteration when obtaining posterior draws from $p^L(\boldsymbol{z}, \boldsymbol{\vartheta}|\mathbf{y}, \mathcal{M}_6)$ and $p(\boldsymbol{z}, \boldsymbol{\vartheta}|\mathbf{y}, \mathcal{M}_6)$ (respectively). For instance, for the misspecified model roughly 95% of the MCMC iterations had 6 components with some allocated observations, and similarly for other data sets, which naively suggest that at least $k = 6$ components are needed. This is in stark contrast with posterior model probabilities $P(\mathcal{M}_k|\mathbf{y})$ in Supplementary Tables 1-3, which suggest more parsimonious models. This difference is explained by the fact that $P(m \mid \mathbf{y}, \mathcal{M}_6)$ reported in Supplementary Tables 4-5 conditions on the larger model whereas $P(\mathcal{M}_k \mid \mathbf{y})$ is a formal measure of uncertainty for each of the models under consideration conditional on the observed data.

SUPPLEMENTARY TABLE 1. Misspecified model. $P(\mathcal{M}_k \mid \mathbf{y})$ and posterior modes $\hat{\boldsymbol{\eta}}$ for 11 models with $k \in \{1, \ldots, 6\}$ and either homogeneous ($\Sigma_j = \Sigma$) or heterogeneous ($\Sigma_i \neq \Sigma_j$) under BIC, LPs and NLPs

				BIC				
	k	$P(\mathcal{M}_k\|\mathbf{y})$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\eta}_4$	$\hat{\eta}_5$	$\hat{\eta}_6$
	1	0.000						
$\Sigma_j = \Sigma$	2	0.000	0.665	0.335				
	3	0.000	0.277	0.388	0.335			
	4	0.000	0.285	0.380	0.159	0.176		
	5	0.000	0.280	0.245	0.140	0.159	0.176	
	6	0.000	0.259	0.071	0.213	0.122	0.157	0.178
$\Sigma_i \neq \Sigma_j$	2	0.000	0.665	0.335				
	3	0.002	0.293	0.372	0.335			
	4	**0.998**	**0.293**	**0.372**	**0.159**	**0.176**		
	5	0.000	0.354	0.212	0.099	0.168	0.167	
	6	0.000	0.211	0.127	0.208	0.118	0.168	0.168
				LPs				
	1	0.000						
$\Sigma_j = \Sigma$	2	0.000	0.665	0.335				
	3	0.003	0.278	0.387	0.335			
	4	0.062	0.295	0.370	0.139	0.196		
	5	**0.469**	**0.196**	**0.300**	**0.169**	**0.041**	**0.294**	
	6	0.465	0.242	0.087	0.273	0.063	0.041	0.294
$\Sigma_i \neq \Sigma_j$	2	0.000	0.657	0.343				
	3	0.000	0.306	0.357	0.337			
	4	0.000	0.384	0.058	0.231	0.327		
	5	0.000	0.356	0.047	0.255	0.055	0.287	
	6	0.000	0.053	0.227	0.304	0.065	0.063	0.288
				NLPs				
	1	0.000						
$\Sigma_j = \Sigma$	2	0.000	0.665	0.335				
	3	**1.000**	**0.278**	**0.387**	**0.335**			
	4	0.000	0.293	0.372	0.148	0.187		
	5	0.000	0.189	0.306	0.170	0.039	0.296	
	6	0.000	0.271	0.077	0.303	0.014	0.037	0.298
$\Sigma_i \neq \Sigma_j$	2	0.000	0.657	0.343				
	3	0.000	0.306	0.357	0.337			
	4	0.000	0.328	0.314	0.061	0.297		
	5	0.000	0.293	0.134	0.229	0.060	0.284	
	6	0.000	0.192	0.152	0.118	0.203	0.159	0.176

SUPPLEMENTARY TABLE 2. Faithful dataset. $P(\mathcal{M}_k \mid \mathbf{y})$ and posterior modes $\hat{\boldsymbol{\eta}}$ for 11 models with $k \in \{1, \ldots, 6\}$ and either homogeneous ($\Sigma_j = \Sigma$) or heterogeneous ($\Sigma_i \neq \Sigma_j$) under BIC, LPs and NLPs

				BIC				
	k	$P(\mathcal{M}_k\|\mathbf{y})$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\eta}_4$	$\hat{\eta}_5$	$\hat{\eta}_6$
	1	0.000						
$\Sigma_j = \Sigma$	2	0.004	0.641	0.359				
	3	**0.927**	**0.462**	**0.356**	**0.182**			
	4	0.050	0.400	0.346	0.030	0.224		
	5	0.001	0.352	0.345	0.029	0.216	0.058	
	6	0.000	0.367	0.135	0.057	0.121	0.100	0.220
$\Sigma_i \neq \Sigma_j$	2	0.018	0.644	0.356				
	3	0.000	0.080	0.332	0.588			
	4	0.000	0.243	0.334	0.077	0.346		
	5	0.000	0.355	0.338	0.133	0.128	0.046	
	6	0.000	0.255	0.227	0.093	0.171	0.128	0.126
				LPs				
	1	0.000						
$\Sigma_j = \Sigma$	2	0.000	0.641	0.359				
	3	0.132	0.475	0.354	0.171			
	4	**0.473**	**0.429**	**0.267**	**0.089**	**0.215**		
	5	0.353	0.424	0.269	0.087	0.216	0.004	
	6	0.042	0.216	0.176	0.044	0.124	0.299	0.141
$\Sigma_i \neq \Sigma_j$	2	0.000	0.631	0.369				
	3	0.000	0.599	0.262	0.139			
	4	0.000	0.572	0.172	0.151	0.105		
	5	0.000	0.545	0.146	0.107	0.108	0.094	
	6	0.000	0.518	0.204	0.045	0.045	0.087	0.101
				NLPs				
	1	0.000						
$\Sigma_j = \Sigma$	2	0.000	0.641	0.359				
	3	**0.967**	**0.472**	**0.355**	**0.173**			
	4	0.028	0.425	0.263	0.093	0.219		
	5	0.002	0.317	0.216	0.139	0.107	0.221	
	6	0.003	0.222	0.080	0.062	0.111	0.310	0.215
$\Sigma_i \neq \Sigma_j$	2	0.000	0.630	0.370				
	3	0.000	0.600	0.246	0.154			
	4	0.000	0.571	0.214	0.116	0.099		
	5	0.000	0.410	0.205	0.110	0.092	0.183	
	6	0.000	0.250	0.205	0.087	0.130	0.230	0.098

SUPPLEMENTARY TABLE 3. Iris dataset. $P(\mathcal{M}_k \mid \mathbf{y})$ and posterior modes $\hat{\boldsymbol{\eta}}$ for 11 models with $k \in \{1, \ldots, 6\}$ and either homogeneous ($\Sigma_j = \Sigma$) or heterogeneous ($\Sigma_i \neq \Sigma_j$) under BIC, LPs and NLPs

				BIC				
	k	$P(\mathcal{M}_k\|\mathbf{y})$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\eta}_4$	$\hat{\eta}_5$	$\hat{\eta}_6$
	1	0.000						
$\Sigma_j = \Sigma$	2	0.000	0.333	0.667				
	3	0.000	0.334	0.330	0.336			
	4	0.000	0.333	0.102	0.335	0.230		
	5	0.000	0.333	0.176	0.205	0.207	0.079	
	6	0.000	0.284	0.049	0.174	0.205	0.209	0.079
$\Sigma_i \neq \Sigma_j$	**2**	**0.968**	**0.333**	**0.667**				
	3	0.032	0.333	0.300	0.367			
	4	0.000	0.333	0.135	0.216	0.316		
	5	0.000	0.333	0.193	0.227	0.160	0.087	
	6	0.000	0.287	0.046	0.193	0.227	0.160	0.087
				LPs				
	1	0.000						
$\Sigma_j = \Sigma$	2	0.000	0.333	0.667				
	3	**0.809**	**0.333**	**0.336**	**0.331**			
	4	0.029	0.334	0.101	0.336	0.229		
	5	0.132	0.333	0.187	0.178	0.219	0.083	
	6	0.030	0.195	0.138	0.200	0.177	0.211	0.079
$\Sigma_i \neq \Sigma_j$	2	0.000	0.362	0.638				
	3	0.000	0.334	0.305	0.361			
	4	0.000	0.309	0.135	0.230	0.326		
	5	0.000	0.287	0.198	0.201	0.125	0.189	
	6	0.000	0.188	0.145	0.139	0.177	0.124	0.227
				NLPs				
	1	0.000						
$\Sigma_j = \Sigma$	2	0.000	0.333	0.667				
	3	**1.000**	**0.334**	**0.331**	**0.335**			
	4	0.000	0.333	0.101	0.336	0.230		
	5	0.000	0.333	0.182	0.194	0.212	0.079	
	6	0.000	0.232	0.102	0.190	0.186	0.211	0.079
$\Sigma_i \neq \Sigma_j$	2	0.000	0.362	0.638				
	3	0.000	0.333	0.307	0.360			
	4	0.000	0.310	0.129	0.233	0.328		
	5	0.000	0.287	0.210	0.203	0.115	0.185	
	6	0.000	0.186	0.147	0.197	0.188	0.107	0.175

SUPPLEMENTARY TABLE 4. Number of non-empty components m in the MCMC posterior samples under the Normal-IW-Dir prior and $\Sigma_j = \Sigma$. The Misspecified, Faithful and Fisher's Iris data are considered (Section 4).

	$P(m = k\|\mathbf{y}, \mathcal{M}_6)$					
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
Misspecified	0	0	0	0	0.046	0.954
Faithful	0	0	0	0	0.130	0.870
Fisher's Iris	0	0.020	0	0.020	0	0.960

SUPPLEMENTARY TABLE 5. Number of non-empty components m in the MCMC posterior samples under the MOM-IW-Dir prior and $\Sigma_j = \Sigma$. The Misspecified, Faithful and Fisher's Iris data are considered (Section 4).

	$P(m = k\|\mathbf{y}, \mathcal{M}_6)$					
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
Misspecified	0	0	0	0	0.045	0.950
Faithful	0	0	0	0	0.120	0.880
Iris	0	0.002	0	0.002	0	0.996