Finite mixture models are typically inconsistent for the number of components

Diana Cai
Dept. of Computer Science
Princeton University
Princeton, NJ 08544
dcai@cs.princeton.edu

Trevor Campbell
CSAIL
MIT
Cambridge, MA 02139
tdjc@mit.edu

Tamara Broderick
CSAIL
MIT
Cambridge, MA 02139
tbroderick@csail.mit.edu

1 Introduction

A generative model is, of necessity, a vast simplification of the deeply complex real-world phenomena that govern any observed data set. It is only via this simplification that we can arrive at a tractable data analysis and discover meaningful and actionable patterns in data. In this sense, typically any model of a real-world data set is misspecified, and misspecification is unavoidable. But while misspecification in the form of simplification is powerful, it can also be potentially dangerous. In particular, certain kinds of misspecification can lead to fundamentally inaccurate or misleading inferences. For instance, recent work by Miller and Harrison (2013, 2014) serves as a cautionary tale about mixture modeling. In particular, mixture models are often matched with a nonparametric Bayesian prior by practitioners in order to discover the number of clusters in a set of data. Miller and Harrison (2013, 2014) demonstrate that such models are "severely" inconsistent for the number of clusters; that is, the probability of the correct number of clusters being recovered decreases to zero as the amount of data increases. An implication is that finite mixture models would be a more appropriate modeling choice. But empirical work by Miller and Dunson (2015) suggests otherwise. We here aim to demonstrate theoretically that even finite mixture models with an unknown number of clusters generally exhibit severe inconsistency, just as the nonparametric Bayesian models do. We discuss the implications for practical modeling and inference in mixture models.

Mixture models are widely used across the sciences and engineering to discover latent groups in a data set. Typically the number of groups, or clusters, is unknown in advance, and one of the principal inferential goals is estimating and interpreting the number of clusters. For instance, practitioners might wish to find the number of latent genetic populations (Pritchard et al., 2000; Lorenzen et al., 2006; Huelsenbeck and Andolfatto, 2007), gene tissue profiles (Yeung et al., 2001; Medvedovic and Sivaganesan, 2002), cell types (Chan et al., 2008; Prabhakaran et al., 2016), haplotypes (Xing et al., 2006), switching Markov regimes in US dollar exchange rate data (Otranto and Gallo, 2002), gamma-ray burst types (Mukherjee et al., 1998), or segmentation regions in an image (e.g., tissue types in an MRI scan (Banfield and Raftery, 1993)).

Suppose we take a Bayesian approach and compute a posterior distribution over the number of clusters. A natural check on our analysis is to establish that—when the true, generating number of clusters is known—our posterior increasingly concentrates near the truth as the number of data points becomes arbitrarily large. That is, we wish to check for a form of *consistency*. A nonparametric Bayesian prior for mixture models implicitly gives a prior with support on the natural numbers for any number of data points and is often used for learning the number of components in a mixture model. But Miller and Harrison (2013, 2014) demonstrate that the posterior under such a prior concentrates strictly away from the true, generating number of clusters when that number is finite. In fact, the estimate of the number of clusters diverges to infinity as the amount of data grows. A recommended alternative (Green and Richardson, 2001; Miller and Harrison, 2013, 2014, 2016) is to instead consider a prior on the number of clusters that does not vary with the size of the data but still maintains support on all possible positive integer numbers of clusters. We call this the *finite mixture model* in what follows to emphasize that, unlike in the nonparametric Bayesian model, the expected number of clusters in the generative model is fixed and finite across data set sizes. Nobile (1994)

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

has shown that the resulting posterior in this case does concentrate at the true, generating number of clusters. But crucially this result depends on the assumption that the cluster likelihoods, roughly the shapes of the clusters, are perfectly specified.

In practice, though, we can expect that the cluster likelihoods are at least somewhat imperfectly specified since they are necessarily simplifications of real-world phenomena. For instance, while Gaussian mixture models are ubiquitous, data are rarely perfectly Gaussian. Miller and Dunson (2015) provide empirical evidence of undesirable posterior behavior in a finite mixture model with misspecified likelihood. We conjecture that even in the finite mixture model, the posterior number of clusters typically concentrates strictly away from the generating number of clusters when that number is finite. We further conjecture that, in fact, with probability 1, the estimate of the number of clusters diverges to infinity as the amount of data grows, just as in the nonparametric case. After reviewing consistency and inconsistency, we posit results for mixture model consistency and rates of convergence (or divergence) in Section 3. We give empirical evidence for the severe inconsistency problem in finite mixture models in Section 4. We conclude in Section 5 with a discussion of the implications of these results for the practitioner who wishes to discover the number of clusters in a mixture model; we also discuss the role of consistency in practical data analysis.

2 Bayesian mixture models

We consider a Bayesian model consisting of (1) a *prior* distribution Π_0 on a parameter space Ω and (2) conditionally i.i.d. data governed by a *likelihood* distribution $P_\theta(\cdot)$ on an observation space $\mathcal{X}$ and defined for every parameter $\theta \in \Omega$. All spaces are endowed with appropriate σ-algebras, and we assume there exists a measure ν on $\mathcal{X}$ such that $\forall \theta \in \Omega$, $P_\theta \ll \nu$; therefore P_θ has density p_θ with respect to ν. The joint distribution of the parameter $\Theta \in \Omega$ and observations $X_1, \ldots, X_N \in \mathcal{X}$ is defined by

$$\Theta \sim \Pi_0 \qquad X_1, \ldots, X_N \,|\, \Theta \overset{\text{i.i.d.}}{\sim} P_\Theta. \tag{1}$$

The *posterior* distribution Π on Ω describes the practitioner's state of knowledge after observing N observations $X_1, \ldots, X_N$ and is defined by

$$\forall \text{ measurable } A \subseteq \Omega, \quad \Pi(A \,|\, X_1, \ldots, X_N) = \frac{\int_A \prod_{n=1}^N p_\theta(X_n) \mathrm{d}\Pi_0(\theta)}{\int_\Omega \prod_{n=1}^N p_\theta(X_n) \mathrm{d}\Pi_0(\theta)}. \tag{2}$$

We focus on a particular Bayesian model choice: a *mixture model* with an unknown number of components. In this case, the likelihood P_θ is a weighted sum of component distributions F_ξ for $\xi \in \Xi$. The full parameter space here is the union across possible cluster cardinalities $k \in \mathbb{N}$ of: the product space of $\{k\}$, the k-dimensional probability simplex, and k copies of the component parameter space Ξ; i.e., $\Omega = \bigcup_{k=1}^\infty \{k\} \times \Delta^{k-1} \times \Xi^k$. Thus, each parameter $\theta \in \Omega$ can be expressed as $\theta = (k, \pi_1, \ldots, \pi_k, \xi_1, \ldots, \xi_k)$ for some $k \in \mathbb{N}$. We then give P_θ the form

$$\forall \text{ measurable } A \subseteq \mathcal{X}, \quad P_\theta(A) = \sum_{j=1}^k \pi_j F_{\xi_j}(A).$$

Since we have assumed ν dominates $\{P_\theta : \theta \in \Omega\}$, the above equality implies ν also dominates $\{F_\xi : \xi \in \Xi\}$, so we have that there exist densities f_ξ such that $p_\theta = \sum_{j=1}^k \pi_j f_{\xi_j}$.

3 Posterior inconsistency

A desirable property of the posterior distribution Π is that it becomes arbitrarily more "accurate" as we gather more data. In particular, suppose our model is *well-specified* in that the data $X_1, \ldots, X_N$ are truly generated from P_θ for some $\theta \in \Omega$. Then we might expect the posterior to concentrate on neighborhoods of θ. That is, for some large class of measurable subsets $A \subseteq \Omega$, the posterior $\Pi(A \,|\, X_1, \ldots, X_N)$ should converge to $\mathbb{1}(\theta \in A)$ in some sense as $N \to \infty$. Properties of this form are known collectively as *posterior consistency*. There are several popular formulations of posterior consistency (e.g., Doob (1949); LeCam (1953); Freedman (1963); Schwartz (1965)), and we provide

an overview in the appendix. For an in-depth review, see Ghosh and Ramamoorthi (2003, Ch. 4) and Ghosal and van der Vaart (2017, Ch. 6).

The posterior consistency theorems above assume the data are truly generated by P_θ for some $\theta \in \Omega$. In our setting, we are specifically interested in posterior (in)consistency for misspecified models. Posterior consistency for models under likelihood misspecification has been investigated by Berk (1966); Kleijn and van der Vaart (2006); De Blasi and Walker (2013); Ramamoorthi et al. (2015), where the conditions for consistency are much stronger than in the case when the model is well-specified.

Posterior (in)consistency in mixture models. Posterior consistency for density estimation in a wide class of mixture models is well-established (Ghosal et al., 1999; Lijoi et al., 2004). But posterior consistency for the number of components is not as thoroughly characterized. There are several results establishing consistency for well-specified finite mixture models. Nobile (1994) demonstrates that finite mixtures exhibit posterior consistency (see Appendix A.3) assuming the model is well-specified and the class of densities $\{p_\theta : \theta \in \Omega\}$ is identifiable (up to duplicate components and component reordering). Ishwaran et al. (2001) shows that in a well-specified setting, the posterior does not asymptotically *underestimate* the number of components when assuming a stronger identifiability condition (see Definition A.3 and theorem A.4), but this result does not cover its behavior at numbers of components larger than the true number. Rousseau and Mengersen (2011) shows that in a mixture model with well-specified densities but an excess of components, the posterior will concentrate properly by emptying the extra components.

Miller and Harrison (2013, 2014) considers data generated by a finite mixture but modeled with a nonparametric mixture. The authors show that the posterior on the number of components is inconsistent and instead suggest using a finite mixture model for inference, given the posterior consistency results (under correct specification of the likelihood) cited above. Additionally, the authors note that despite achieving consistency, the practitioner should be wary of misspecification of the mixture likelihood. However, in practice, likelihoods are almost always misspecified. Thus, it is important to understand when misspecification is problematic and exactly what problems manifest in practice. We focus on the case where the likelihood model does not contain the generating data distribution and the resulting posterior inconsistency for the number of mixture components. For example, suppose the data are generated from a mixture of Laplace distributions, while the model is chosen to be a Gaussian family; in Section 4, we examine this example empirically. A proof sketch for the following result may be found in Appendix A.4.

Conjecture 3.1. *Suppose the data are generated by a density p that is not in $\{p_\theta : \theta \in \Omega\}$. Then the posterior on the number of components is severely inconsistent, i.e.*

$$\forall k \in \mathbb{N}, \quad \Pi(k \mid X_1, \ldots, X_N) \overset{P_\theta \text{ a.s.}}{\to} 0, \quad N \to \infty. \tag{3}$$

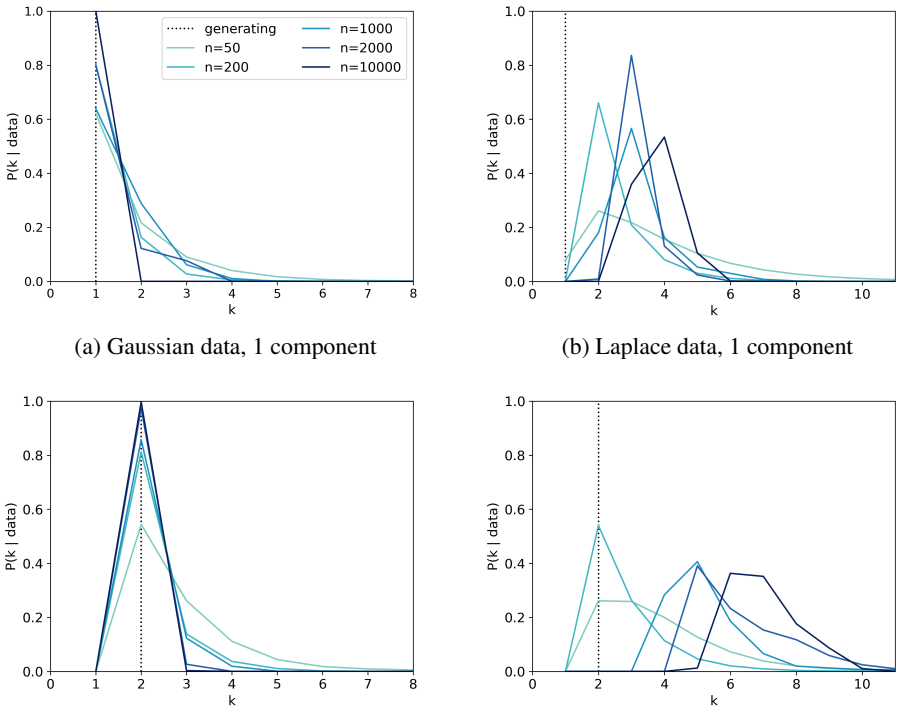
Since mixture models are typically misspecified in practice, this result implies that the posterior for the number of components is typically inconsistent. However, since inconsistency is an asymptotic property, it is useful to quantify when, for a finite sample, a posterior can still give "useful" inferences. For instance, if the data generating density p is close to some member of the model class p_θ, inferences made given data from p and data from p_θ should be similar until a large enough sample size has been obtained to differentiate them. We quantify this difference precisely by viewing the posterior $\Pi(\cdot \mid X_1, \ldots, X_N)$ as a random element in the space of probability measures on Ω, and comparing the distribution over the posterior given data $X_1, \ldots, X_N \sim p$ and data $X_1, \ldots, X_N \sim p_\theta$. Although specified for the general setting, this result applies to the posterior on the number of components in a mixture model. A proof sketch for the following result may be found in Appendix A.5.

Conjecture 3.2. *Suppose $p = (1-\epsilon)p_\theta + \epsilon q$ for some density q, $\epsilon \in [0,1]$, and $\theta \in \Omega$. Then if μ is the distribution over the posterior given data $(X_n)_{n=1}^N \overset{i.i.d.}{\sim} p$ and μ_θ is the same for data $(X_n)_{n=1}^N \overset{i.i.d}{\sim} p_\theta$,*

$$d_{\mathrm{TV}}(\mu, \mu_\theta) \leq 1 - (1-\epsilon)^N. \tag{4}$$

4 Simulations

We examine two simple examples of model misspecification and posterior inconsistency in the number of components. Our experiments and empirical results are very similar to Figure 2 of Miller



(a) Gaussian data, 1 component

(b) Laplace data, 1 component

(c) Gaussian data, 2 components

(d) Laplace data, 2 components

Figure 1: Posterior probability of the number of components k for the Gaussian mixture model, fit to univariate data generated from a Gaussian mixture model, and a Laplace mixture model.

and Dunson (2015), who study the posterior of a skewed Gaussian. Specifically, we study data generated from a 1-component and 2-component Gaussian and Laplace mixture models, where the data size varies from $n = 50, 200, 1000, 2000, 10000$. Each subset of the data is fit to a univariate Gaussian mixture model, with a geometric(0.1) prior on the number of components. Inference for the model is performed using a split-merge Gibbs sampler (Miller and Harrison, 2016)[1]. We ran a total of 30000 iterations per dataset, discarding 5000 burn-in samples. The results of the simulations are in Figure 1. The top row shows the results of the 1-component Gaussian and Laplace models. The posterior on the number of components concentrates around 1 in the case of Gaussian-generated data as the sample size increases (left), where as the posterior on the number of components diverges for the Laplace data (right). The bottom row shows a similar behavior in the 2-component case, where the posterior concentrates around the correct value in the Gaussian case but not the Laplace case.

5 Discussion and future directions

We have posited that the posterior for the number of components is inconsistent for mixture models with a misspecified component family. Misspecification is inevitable in practice; in some cases, it can severely affect the interpretability of results, but in other cases, misspecified models can be useful (Grünwald, 2006). Thus, it is important to understand under what conditions we can expect to make reasonable inferences—and how we can mitigate the effect of model misspecification. In our work, we have outlined a few promising first steps towards understanding misspecification and its effect on the number of components in the model. A number of authors have recently proposed robust Bayesian inference methods to mitigate likelihood misspecification (Grünwald and van Ommen, 2014; Miller and Dunson, 2015; Wang et al., 2017), for the setting when the empirical distribution of the observed data is close in Kullback-Leibler divergence to the empirical distribution of the data sampled from the model. It remains to better understand connections between our results and these methods. For instance, it would be interesting to investigate connections between our Conjecture 3.2 and the asymptotic results of Miller and Dunson (2015) and whether our result might provide insights into setting the parameter in the coarsened posterior.

[1]Code available at https://github.com/jwmi/BayesianMixtures.jl

Acknowledgments

This research was supported in part by ONR grant N00014-17-1-2072, an MIT Lincoln Laboratory Advanced Concepts Committee Award, and a Google Faculty Research Award.

References

J. D. Banfield and A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, pages 803–821, 1993.

A. Barron. Uniformly powerful goodness of fit tests. *The Annals of Statistics*, 17(1):107–124, 1989.

R. H. Berk. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

C. Chan, F. Feng, J. Ottinger, D. Foster, M. West, and T. B. Kepler. Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry Part A*, 73(8):693–701, 2008.

P. De Blasi and S. G. Walker. Bayesian asymptotics with misspecified models. *Statistica Sinica*, pages 169–187, 2013.

J. Doob. Application of the theory of martingales. Colloque international Centre Nat. Rech. Sci., 1949.

D. A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case: I. *The Annals of Mathematical Statistics*, pages 1386–1403, 1963.

S. Ghosal and A. W. v. d. Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics, 2017.

S. Ghosal, J. Ghosh, and R. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158, 1999.

J. Ghosh and R. Ramamoorthi. *Bayesian Nonparametrics*. Springer Series in Statistics, 2003.

P. J. Green and S. Richardson. Modelling heterogeneity with and without the Dirichlet process. *Scandinavian journal of statistics*, 28(2):355–375, 2001.

P. Grünwald and T. v. Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *arXiv preprint arXiv:1412.3730*, 2014.

P. D. Grünwald. Bayesian inconsistency under misspecification. *International Social of Bayesian Analysis Conference*, 2006.

J. P. Huelsenbeck and P. Andolfatto. Inference of population structure under a Dirichlet process model. *Genetics*, 175(4):1787–1802, 2007.

H. Ishwaran, L. F. James, and J. Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96(456):1316–1332, 2001.

B. J. Kleijn and A. W. v. d. Vaart. Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, pages 837–877, 2006.

L. LeCam. *On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates*. University of California: Publ. in statistics. Univ.of California Press, 1953.

A. Lijoi, I. Prünster, and S. Walker. Extending Doob's consistency theorem to nonparametric densities. *Bernoulli*, 10(4):651–663, 2004.

E. D. Lorenzen, P. Arctander, and H. R. Siegismund. Regional genetic structuring and evolutionary history of the impala aepyceros melampus. *Journal of Heredity*, 97(2):119–132, 2006.

M. Medvedovic and S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206, 2002.

J. W. Miller and D. B. Dunson. Robust Bayesian inference via coarsening. *arXiv preprint arXiv:1506.06101*, 2015.

J. W. Miller and M. T. Harrison. A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems*, pages 199–206, 2013.

J. W. Miller and M. T. Harrison. Inconsistency of Pitman-Yor process mixtures for the number of components. *Journal of Machine Learning Research*, 15(1):3333–3370, 2014.

J. W. Miller and M. T. Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 2016.

S. Mukherjee, E. D. Feigelson, G. J. Babu, F. Murtagh, C. Fraley, and A. Raftery. Three types of gamma-ray bursts. *The Astrophysical Journal*, 508(1):314, 1998.

A. Nobile. *Bayesian analysis of finite mixture distributions*. PhD thesis, Carnegie Mellon University, 1994.

E. Otranto and G. M. Gallo. A nonparametric Bayesian approach to detect the number of regimes in Markov switching models. *Econometric Reviews*, 21(4):477–496, 2002.

S. Prabhakaran, E. Azizi, A. Carr, and D. Pe'er. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *International Conference on Machine Learning*, pages 1070–1079, 2016.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

R. Ramamoorthi, K. Sriram, R. Martin, et al. On posterior concentration in misspecified models. *Bayesian Analysis*, 10(4):759–789, 2015.

J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73 (5):689–710, 2011.

L. Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie*, 4:10–26, 1965.

Y. Wang, A. Kucukelbir, and D. M. Blei. Reweighted data for robust probabilistic models. In *International Conference on Machine Learning*, 2017.

E. P. Xing, K.-A. Sohn, M. I. Jordan, and Y.-W. Teh. Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture. In *International Conference on Machine Learning*, pages 1049–1056, 2006.

K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318, 2001.

A Posterior consistency details

A.1 Doob's consistency theorem

There are several popular formulations of posterior consistency. The earliest method, pioneered by Doob (1949), is a *Bayesian* formulation in which consistency is guaranteed to hold at almost every $\theta \in \Omega$ under the prior Π_0 (Theorem A.1).

Theorem A.1 (Doob (1949))**.** *Suppose the mapping $\theta \mapsto P_\theta$ is one-to-one. Then under mild technical conditions (see Schwartz (1965, Theorem 3.2)),*

$$\forall \textit{ measurable } A \subseteq \Omega, \quad \Pi(A \,|\, X_1, \ldots, X_N) \overset{P_\Theta \text{ a.s.}}{\to} 1\,(\Theta \in A) \quad \Pi_0 \textit{ a.e.} \tag{A.1}$$

This result is remarkably elegant: assuming the model is identifiable, it asserts that Bayesian posteriors are almost surely consistent on all neighborhoods of Θ for a very large class of models. Its proof based on the theory of martingales is likewise elegant. However, the parameter θ governing the generation of data must truly be generated via $\Theta \sim \Pi_0$, since the above result only holds outside a set of Π_0-measure 0. Unfortunately, this set of measure 0 can be quite large (in a practical sense) in many applications, and if we take a frequentist perspective in which θ is simply some unknown, fixed value, the above result cannot be applied.

A.2 Conditions for Schwartz's consistency theorem and extensions

If we take a frequentist perspective in which θ is simply some unknown, fixed value, then Theorem A.1 is not applicable. Therefore, Schwartz (1965) developed a second, *frequentist* formulation of posterior consistency in which the data are truly generated by P_θ for some $\theta \in \Omega$, and the goal is to show that the posterior concentrates on neighborhoods V of θ. Roughly, V has to be large enough that it is supported by the prior (Definition A.1) and that there exists a test that distinguishes data $X_1, \ldots, X_N$ generated by θ from that generated by any parameter in V^c (Definition A.2). If these two properties hold, then the posterior is consistent on V (a.s. P_θ) (Theorem A.2). Note that in contrast to Doob's result (Theorem A.1), Theorem A.2 makes no assumption that the parameter is generated by the prior Π_0; however, this comes at the cost of verifying the KL-neighborhood and uniformly exponentially consistent (UEC) test conditions given in Appendix A.

Definition A.1. *A measurable subset $V \subset \Omega$ is said to be a* KL-neighborhood of θ *if for every $\epsilon > 0$ there exists a subset $W \subseteq V$ such that*

$$\Pi_0(W) > 0 \quad \textit{and} \quad \sup_{\zeta \in W} \mathbb{E}_\theta \left[\log \frac{p_\theta(X)}{p_\zeta(X)} \right] < \epsilon. \tag{A.2}$$

Definition A.2. *A* uniformly exponentially consistent (UEC) test *for a measurable subset $V \subseteq \Omega$ at θ is a sequence of measurable subsets $A_n \subseteq \mathcal{X}^n$ and $r > 0$ such that*

$$P_\theta^N(A_N) \geq 1 - e^{-Nr} \quad \textit{and} \quad \sup_{\zeta \in V^c} P_\zeta^N(A_N) \leq e^{-Nr}, \tag{A.3}$$

We now state Schwartz's consistency theorem.

Theorem A.2 (Schwartz (1965, Theorem 6.1))**.** *Suppose the data are generated from P_θ, and $V \subset \Omega$. Then if measurable $V \subset \Omega$ is a KL-neighborhood of θ and there exists a UEC test for V at θ,*

$$\Pi(V^c \,|\, X_1, \ldots, X_N) \to 0, \quad P_\theta \textit{ a.s.} \tag{A.4}$$

Many extensions to Schwartz's consistency theorem have been developed and are important for establishing posterior consistency in nonparametric Bayesian models. We refer to Ghosh and Ramamoorthi (2003, Ch. 4) and Ghosal and van der Vaart (2017, Ch. 6) for details and additional references.

A.3 Posterior consistency in mixture models

Theorem A.3 (Nobile (1994, Theorem 3.2))**.** *If*

$$\Pi_0\left(\{\theta \in \Omega : \exists i \neq j \textit{ with either } \xi_i = \xi_j \textit{ or } \pi_i = \pi_j\}\right) = 0, \tag{A.5}$$

and the class of densities $\{p_\theta : \theta \in \Omega\}$ is identifiable up to component reordering and duplicate components, then

$$\forall k \in \mathbb{N}, \quad \Pi(k \mid X_1, \ldots, X_N) \overset{P_\Theta \text{ a.s.}}{\to} 1(k = K) \quad \Pi_0 \text{ a.e.} \tag{A.6}$$

In the frequentist posterior consistency setting, Ishwaran et al. (2001, proof of Theorem 1) establishes that for an $\mathcal{F}$-identifiable class of densities $\{p_\theta : \theta \in \Omega\}$, if the true density p_θ is in the KL-support of the prior, the posterior weights the true k exponentially more than any $k' < k$. $\mathcal{F}$-identifiability is a stronger condition than identifiability, and is related to the existence of UEC tests (Barron, 1989):

Definition A.3. *A class of densities $\{p_\theta : \theta \in \Omega\}$ is* $\mathcal{F}$-identifiable *if there exists a countable sequence of sets $(A_n)_{n=1}^N$ for which $\theta \neq \theta'$ means $p_\theta(A_n) \neq p_{\theta'}(A_n)$ for some $n \in \mathbb{N}$.*

Definition A.4. *A density $p_{\theta^\star}$ is in the* KL-support *of the prior Π_0 if for all $\epsilon > 0$, $\Pi_0\left(\{p_\theta : \theta \in \Omega, d_{KL}(p_{\theta^\star}, p_\theta) < \epsilon\} > 0\right.$.*

Theorem A.4 (Ishwaran et al. (2001, Proof of Theorem 1))**.** *Suppose the data are generated by a k-component mixture p_θ, where p_θ is in the KL-support of the prior and $\{p_\theta : \theta \in \Omega\}$ is $\mathcal{F}$-identifiable. Then there exists an $\epsilon > 0$ such that for any $k' < k$*

$$\frac{\Pi(k \mid X_1, \ldots, X_N)}{\Pi(k' \mid X_1, \ldots, X_N)} \geq e^{N\epsilon} \text{ as } N \to \infty \quad P_\theta \text{ a.s.} \tag{A.7}$$

A.4 Proof sketch of Conjecture 3.1

Proof Sketch. Recent results on density consistency under model misspecification (Ramamoorthi et al., 2015) show that the posterior will concentrate on a density $p_{\theta^\star}$ in the model family that is "as close as possible" to the true density p in some sense. Since $p \notin \{p_\theta : \theta \in \Omega\}$, the closest density $p_{\theta^\star}$ to p will have an infinite number of components, making the posterior on the number of components approach 0 for any finite number. □

A.5 Proof sketch of Conjecture 3.2

Proof Sketch. Let $\mathcal{M}$ be a σ-algebra of sets of probability measures on Ω, $\mathcal{I} = \{0,1\}^N$, $z = [1, \ldots, 1]$, and $g(I) = \sum_{n=1}^N I_n$ be the number of nonzero entries in $I \in \mathcal{I}$. Further, let f_N map a dataset $(X_n)_{n=1}^N$ to the corresponding posterior on Ω given that dataset. Then

$$d_{\mathrm{TV}}(\ldots) = \frac{1}{2} \sup_{A \in \mathcal{M}} \left| \int_{x \in f_N^{-1}(A)} \prod_{n=1}^N p(x_n) - \prod_{n=1}^N p_\theta(x_n) \mathrm{d}x \right| \tag{A.8}$$

$$= \frac{1}{2} \sup_{A \in \mathcal{M}} \left| \int_{x \in f_N^{-1}(A)} \prod_{n=1}^N ((1-\epsilon)p_\theta(x_n) + \epsilon q(x_n)) - \prod_{n=1}^N p_\theta(x_n) \mathrm{d}x \right| \tag{A.9}$$

$$= \frac{1}{2} \sup_{A \in \mathcal{M}} \left| \int_{x \in f_N^{-1}(A)} \sum_{I \in \mathcal{I}} \prod_{n=1}^N ((1-\epsilon)p_\theta(x_n))^{I_n} (\epsilon q(x_n))^{1-I_n} - \prod_{n=1}^N p_\theta(x_n) \mathrm{d}x \right|. \tag{A.10}$$

Using the triangle inequality to extract the right hand term with the $I = [1, \ldots, 1]$ term,

$$d_{\mathrm{TV}}(\ldots) \leq \frac{1}{2} \sup_{A \in \mathcal{M}} (1 - (1-\epsilon)^N) \int_{x \in f_N^{-1}(A)} \prod_{n=1}^N p_\theta(x_n) \mathrm{d}x + \tag{A.11}$$

$$\sum_{I \in \mathcal{I} \setminus \{z\}} \epsilon^{N - g(I)} (1-\epsilon)^{g(I)} \int_{x \in f_N^{-1}(A)} \prod_{n=1}^N p_\theta(x_n)^{I_n} q(x_n)^{1-I_n} \mathrm{d}x. \tag{A.12}$$

Noting that all integrands above are densities, the integrals are probabilities and so lie in $[0,1]$. Therefore, we can bound these constants by 1 and drop the supremum,

$$d_{\mathrm{TV}}(\ldots) \leq \frac{1}{2} \left((1 - (1-\epsilon)^N) + \sum_{n=1}^N \binom{N}{n} \epsilon^n (1-\epsilon)^{N-n} \right). \tag{A.13}$$

Using the binomial theorem,

$$d_{\mathrm{TV}}\left(\dots\right) \leq 1 - (1-\epsilon)^N. \tag{A.14}$$

□