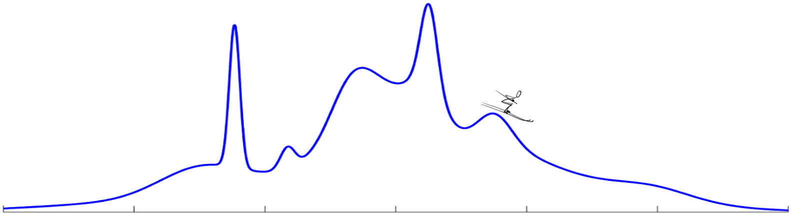
Dirichlet process mixture inconsistency for the number of components

Jeffrey W. Miller
and
Matthew T. Harrison

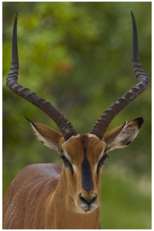
Brown University
Division of Applied Mathematics

NIPS 2013, Lake Tahoe



DPs are often used to infer the number of groups

Population structure



Huelsenbeck & Andolfatto (2007)

Leaché & Fujita (2010)

Richards et al. (2009)

Chen et al. (2009)

Gonzales & Zardoya (2007)

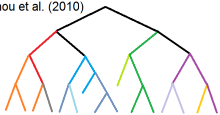
Fogelqvist et al. (2010)

Exchange rate modeling

Otranto & Gallo (2002)

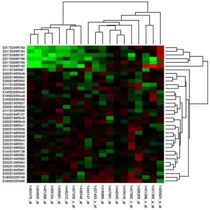
Heterotachy in phylogenetic trees

Lartillot & Philippe (2004)
Zhou et al. (2010)



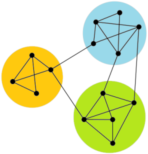
Gene expression profiling

Medvedovic & Sivaganesan (2002)



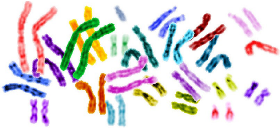
Network communities

Baskerville et al. (2011)



Haplotype inference

Xing et al. (2006)



The DPM is great as a flexible prior on densities . . .

The DPM is great as a flexible prior on densities . . .

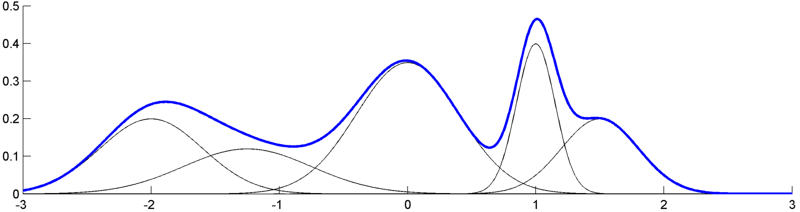
. . . what about for **estimating the number of groups**?

Finite mixture model

$$(\pi_1, \ldots, \pi_k) \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$$

$$\theta_1, \ldots, \theta_k \overset{\text{iid}}{\sim} H$$

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x) = \sum_{i=1}^{k} \pi_i \, p_{\theta_i}(x)$$

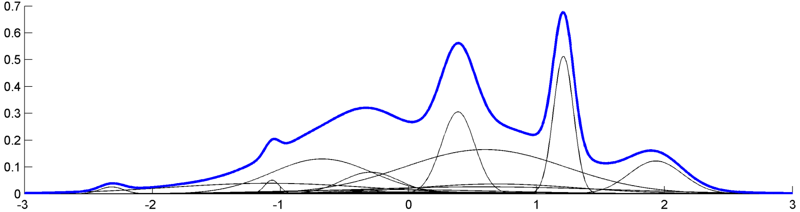


Dirichlet process mixture model

$$(\pi_1, \pi_2, \dots) \sim \text{Stick-breaking process}$$

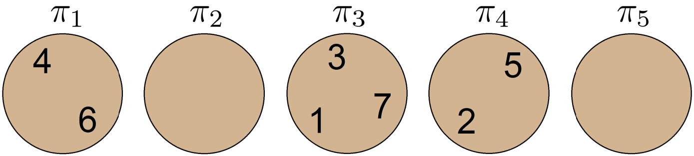
$$\theta_1, \theta_2, \dots \overset{\text{iid}}{\sim} H$$

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} f(x) = \sum_{i=1}^{\infty} \pi_i \, p_{\theta_i}(x)$$



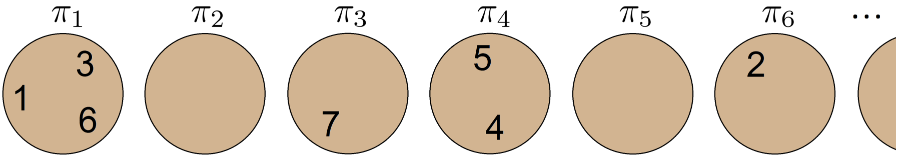
Ferguson (1983), Lo (1984), Sethuraman (1994),
West, Müller, and Escobar (1994), MacEachern (1994)

Finite mixture



5 tables (i.e. components)
3 occupied tables

Dirichlet process mixture



∞ tables (i.e. components)
4 occupied tables

What if we use a DPM on data from finite mixture?

It is known that in many cases the posterior concentrates at the true density f_0,

$$P(\|f - f_0\|_{L_1} < \varepsilon \mid X_{1:n}) \xrightarrow[n\to\infty]{} 1 \;\; \forall \varepsilon > 0,$$

(often at essentially the minimax-optimal rate), for *any* sufficiently regular f_0.
(Contributions by: Ghosal, van der Vaart, Scricciolo, Lijoi, Prünster, Walker, James, Tokdar, Dunson, Bhattacharya, Wu, Ghosh, Ramamoorthi, Ishwaran, and others.)

What if we use a DPM on data from finite mixture?

It is known that in many cases the posterior concentrates at the true density f_0,

$$P(\|f - f_0\|_{L_1} < \varepsilon \mid X_{1:n}) \xrightarrow[n\to\infty]{} 1 \ \ \forall \varepsilon > 0,$$

(often at essentially the minimax-optimal rate), for *any* sufficiently regular f_0.
(Contributions by: Ghosal, van der Vaart, Scricciolo, Lijoi, Prünster, Walker, James, Tokdar, Dunson, Bhattacharya, Wu, Ghosh, Ramamoorthi, Ishwaran, and others.)

In fact, the posterior on the mixing distribution concentrates (in Wasserstein distance) at the true mixing distribution (Nguyen, 2013).

What if we use a DPM on data from finite mixture?

It is known that in many cases the posterior concentrates at the true density f_0,

$$P(\|f - f_0\|_{L_1} < \varepsilon \mid X_{1:n}) \xrightarrow[n\to\infty]{} 1 \ \ \forall \varepsilon > 0,$$

(often at essentially the minimax-optimal rate), for *any* sufficiently regular f_0.
(Contributions by: Ghosal, van der Vaart, Scricciolo, Lijoi, Prünster, Walker, James, Tokdar, Dunson, Bhattacharya, Wu, Ghosh, Ramamoorthi, Ishwaran, and others.)

In fact, the posterior on the mixing distribution concentrates (in Wasserstein distance) at the true mixing distribution (Nguyen, 2013).

Does the posterior on the number of occupied tables concentrate at the true number of components? i.e.

$$P(\#\text{occupied} = k_0 \mid X_{1:n}) \xrightarrow[n\to\infty]{?} 1$$

Outline

1. Empirical evidence
2. Theoretical results
3. Intuition

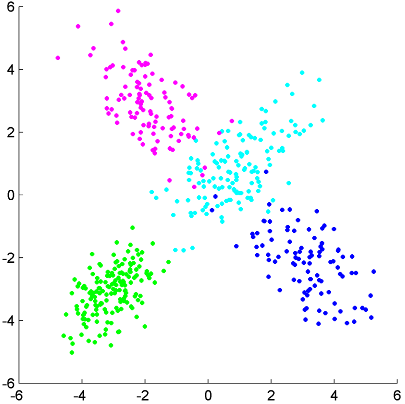
Some interesting experiments

Tiny extra clusters often appear in posterior samples.

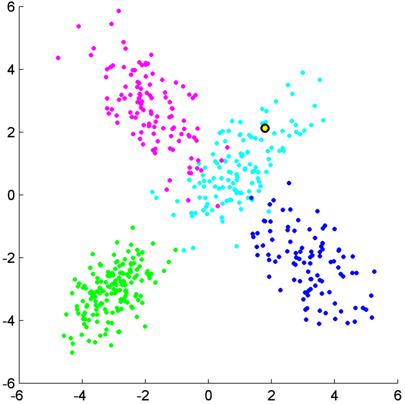
Empirically, this is well-known (e.g. West, Müller, and Escobar, 1994).

Bivariate Gaussian mixture with 4 components

True cluster assignments

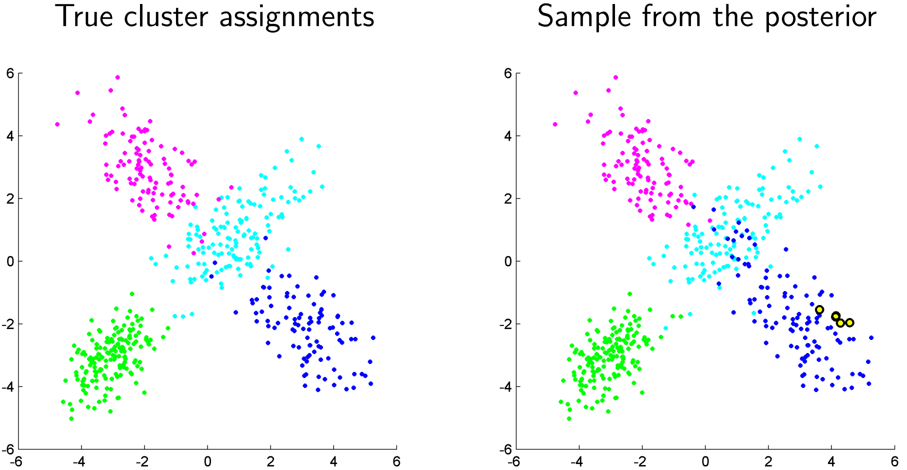


Sample from the posterior



Tiny extra clusters often appear in posterior samples.

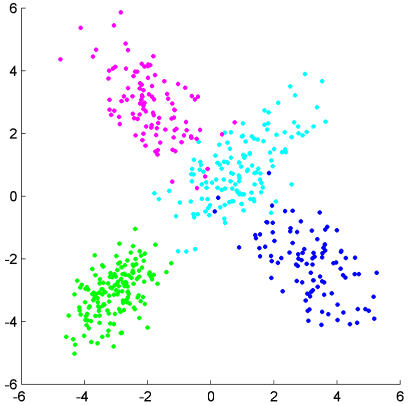
Bivariate Gaussian mixture with 4 components



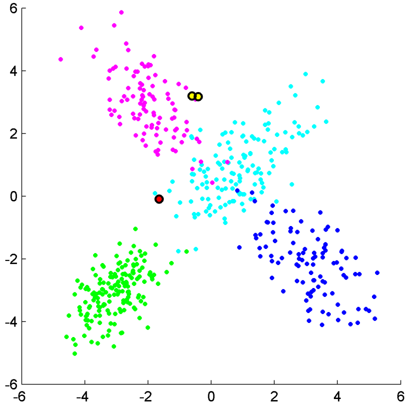
Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

True cluster assignments

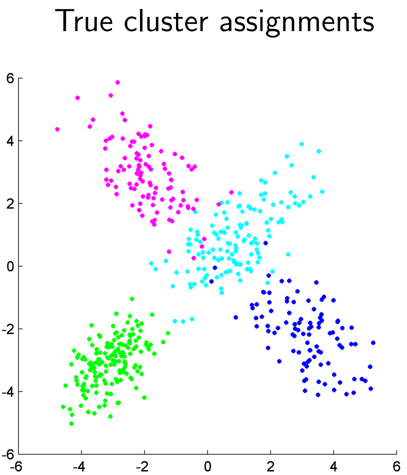


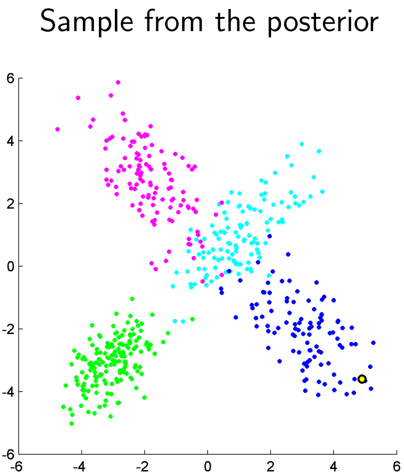
Sample from the posterior



Tiny extra clusters often appear in posterior samples.

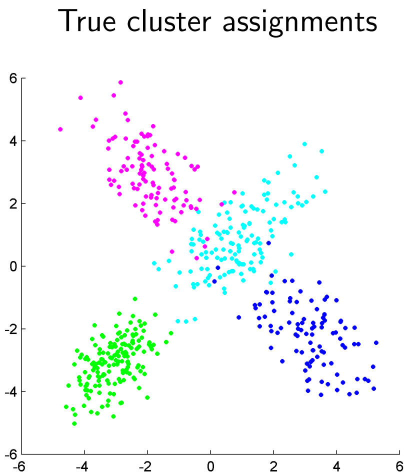
Bivariate Gaussian mixture with 4 components

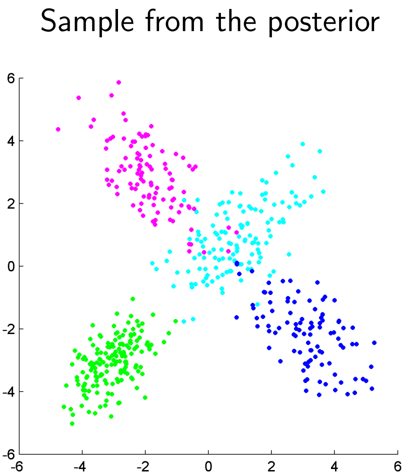




Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

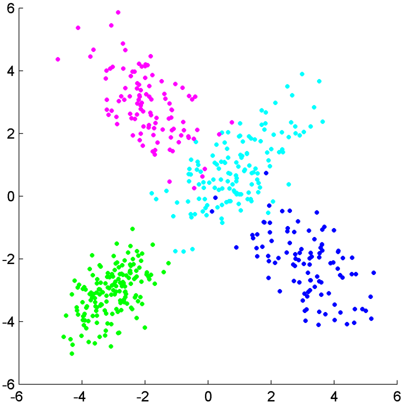




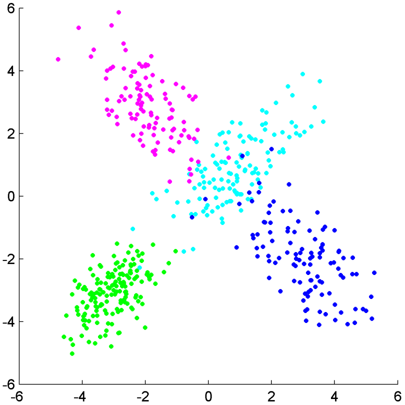
Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

True cluster assignments



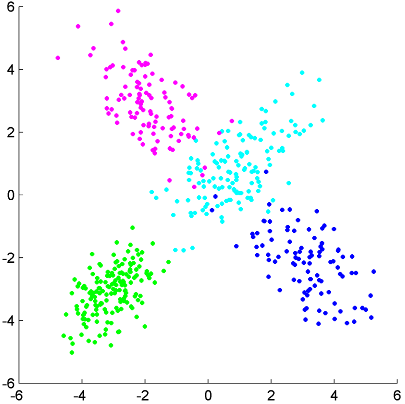
Sample from the posterior



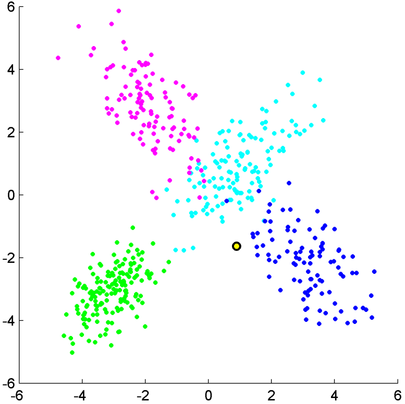
Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

True cluster assignments



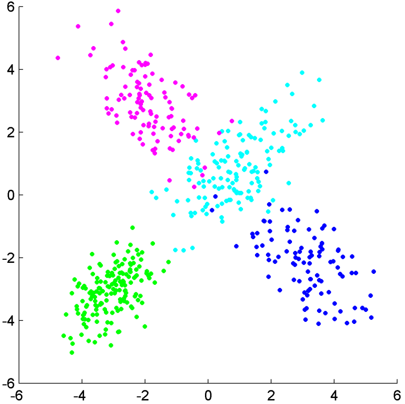
Sample from the posterior



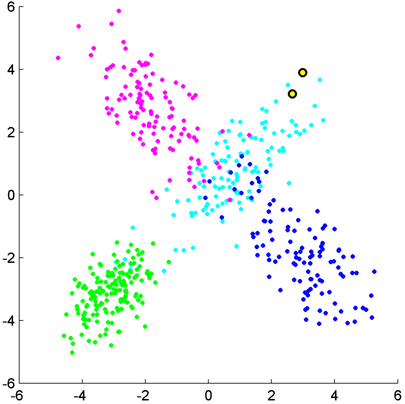
Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

True cluster assignments



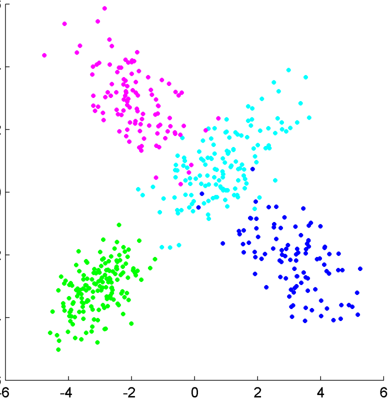
Sample from the posterior



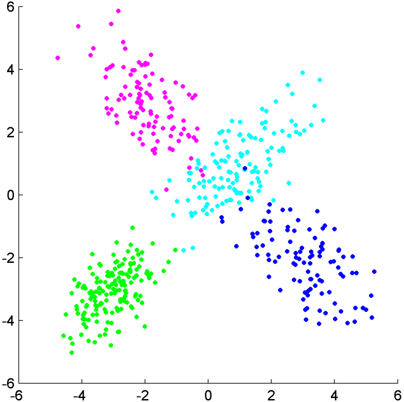
Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

True cluster assignments



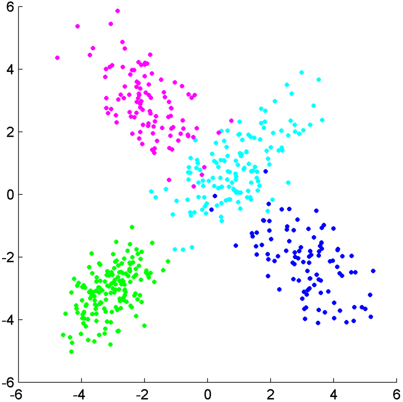
Sample from the posterior



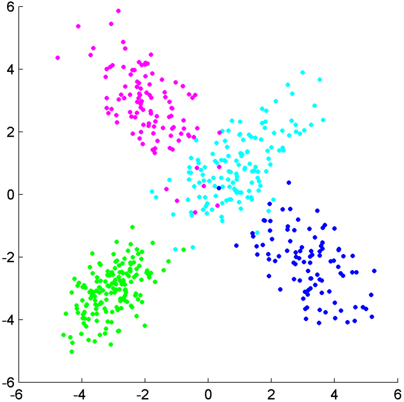
Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

True cluster assignments

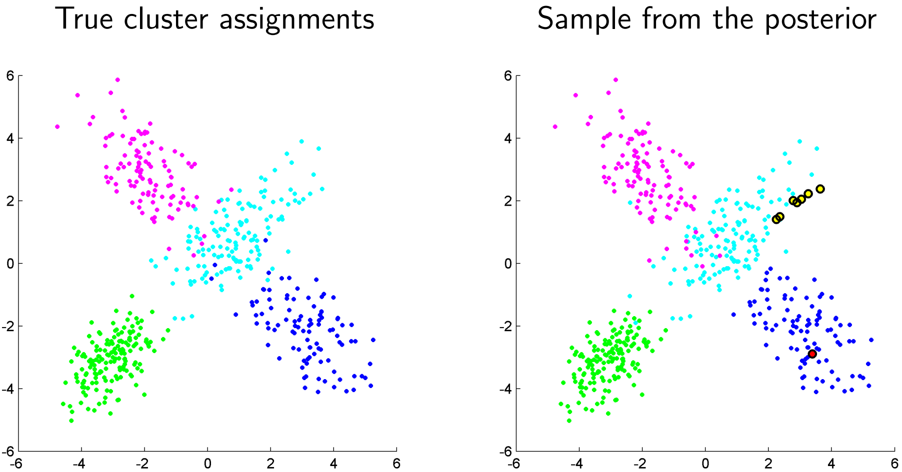


Sample from the posterior



Tiny extra clusters often appear in posterior samples.

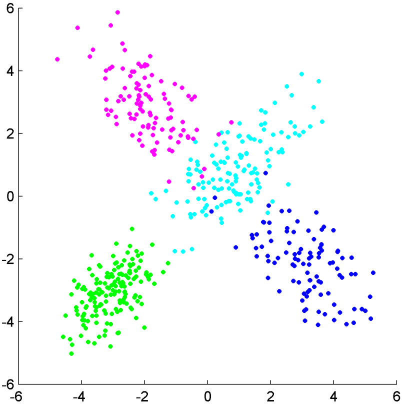
Bivariate Gaussian mixture with 4 components



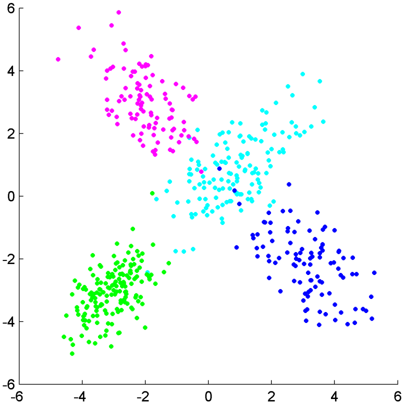
Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

True cluster assignments



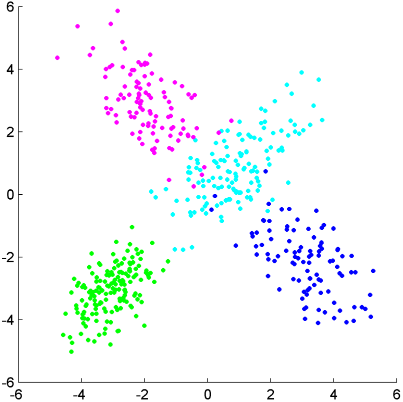
Sample from the posterior



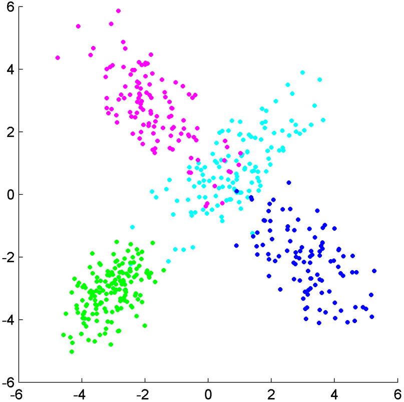
Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

True cluster assignments



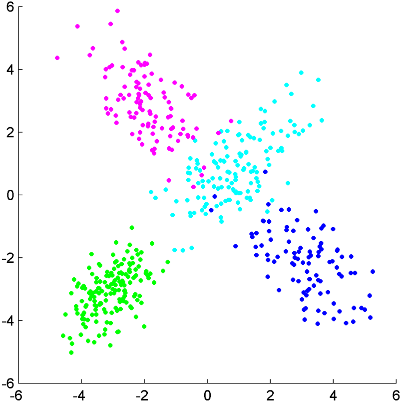
Sample from the posterior



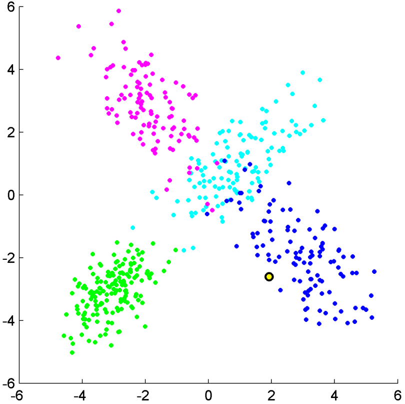
Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

True cluster assignments



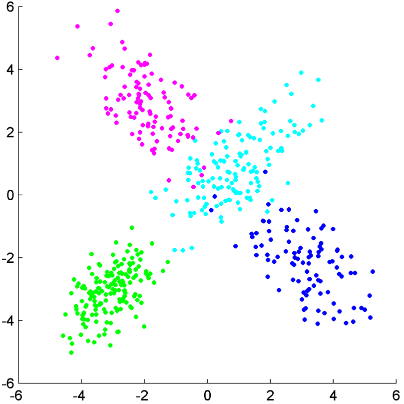
Sample from the posterior



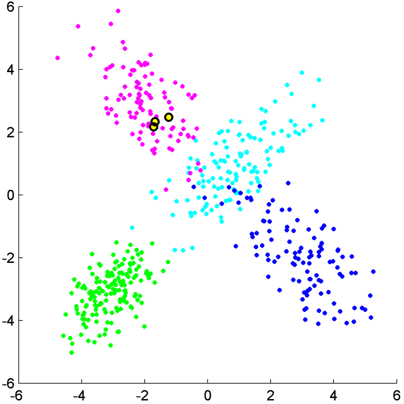
Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

True cluster assignments

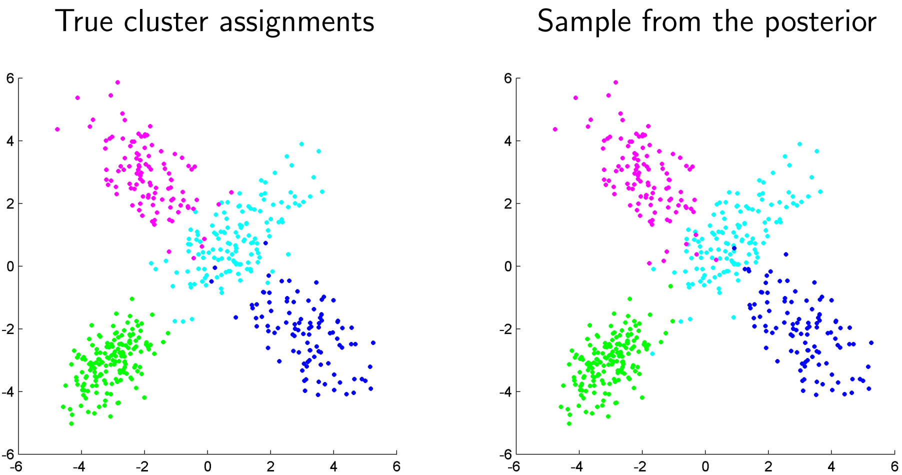


Sample from the posterior



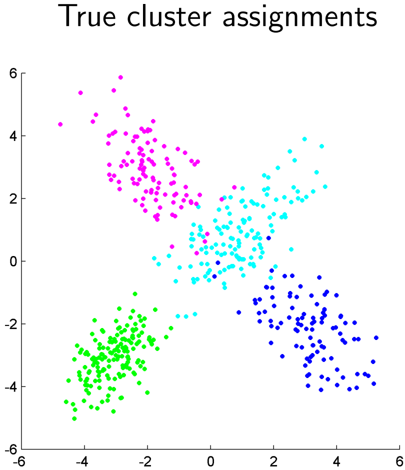
Tiny extra clusters often appear in posterior samples.

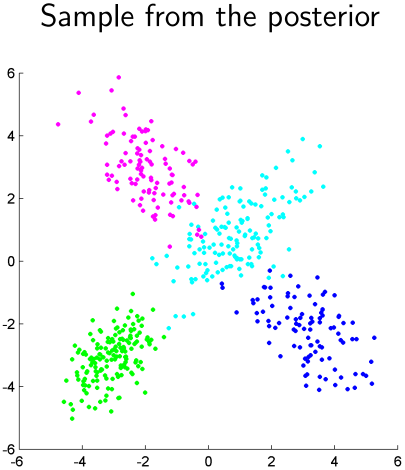
Bivariate Gaussian mixture with 4 components



Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

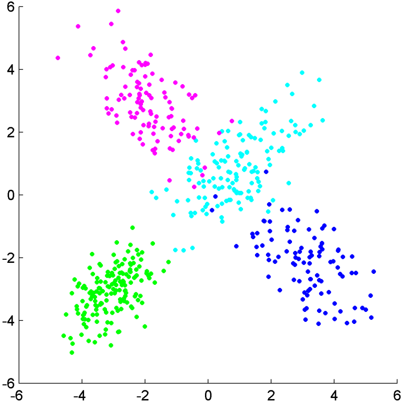




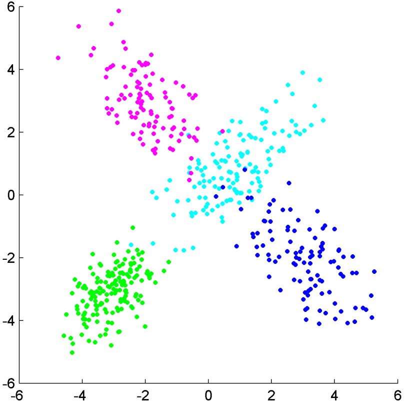
Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

True cluster assignments

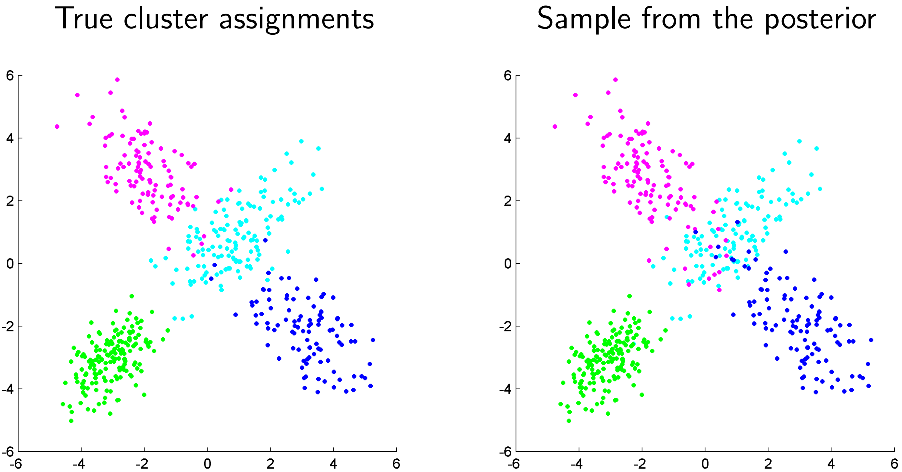


Sample from the posterior



Tiny extra clusters often appear in posterior samples.

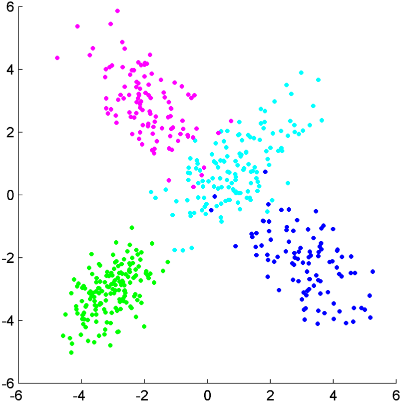
Bivariate Gaussian mixture with 4 components



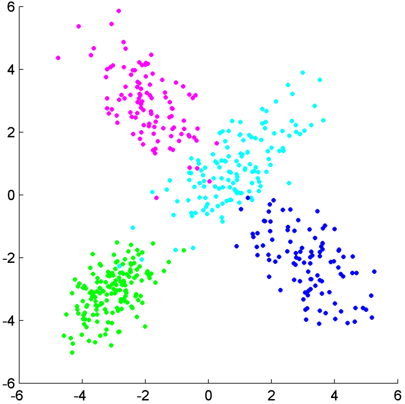
Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

True cluster assignments



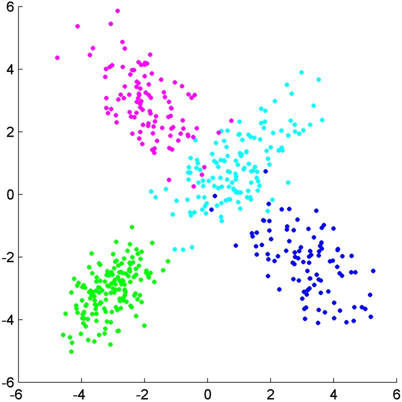
Sample from the posterior



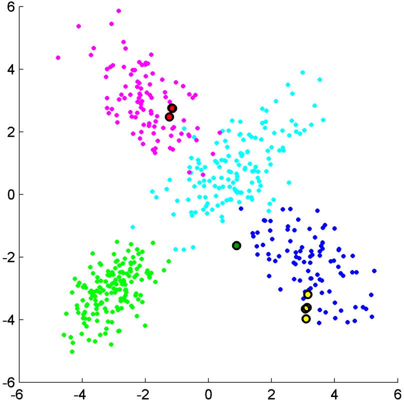
Tiny extra clusters often appear in posterior samples.

Bivariate Gaussian mixture with 4 components

True cluster assignments

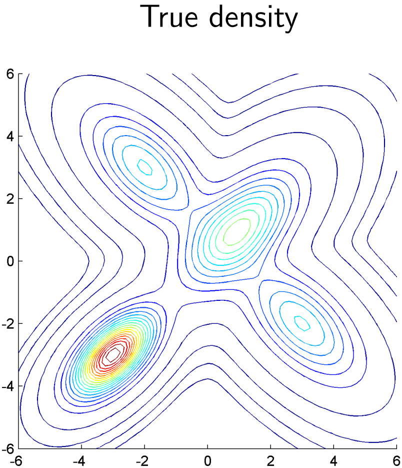


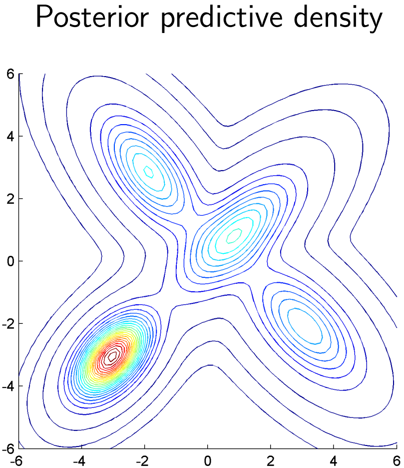
Sample from the posterior



Tiny extra clusters often appear in posterior samples.

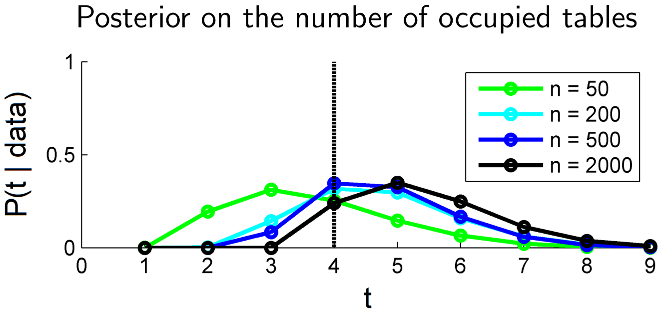
Bivariate Gaussian mixture with 4 components





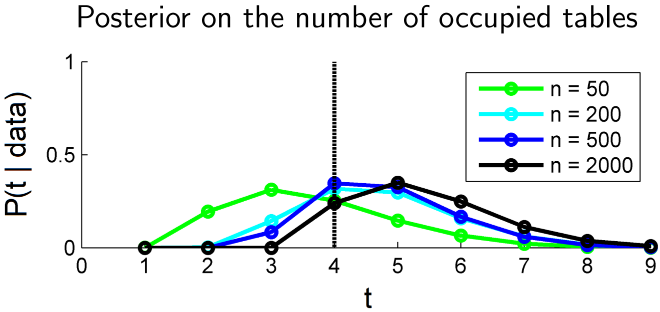
These tiny clusters have negligible impact on density estimates . . .

Bivariate Gaussian mixture with 4 components



. . . but they do affect the posterior on the number of occupied tables.

Bivariate Gaussian mixture with 4 components



. . . but they do affect the posterior on the number of occupied tables.

Will it eventually concentrate at the true value?

Theoretical results

Theorem (M. & Harrison, 2013)

Under mild regularity conditions, if $X_1, X_2, \ldots$ are i.i.d. from a finite mixture with k_0 components, then the DPM posterior on the number of occupied tables T_n satisfies

$$\limsup_{n\to\infty} P(T_n = k_0 \mid X_1, \ldots, X_n) < 1$$

with probability 1.

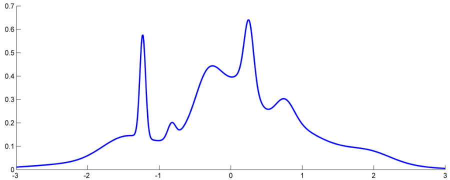
- This implies inconsistency.
- We assume the concentration parameter α is fixed.
- This generalizes to Pitman–Yor process mixtures.
- See Miller & Harrison (2013) arXiv:1309.0024 for details.

This implies inconsistency of Dirichlet process mixtures over:

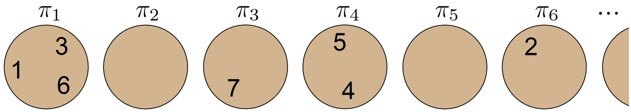
1. a large class of continuous exponential families, including
   - multivariate Gaussian
   - Exponential
   - Gamma
   - Log-Normal
   - Weibull with fixed shape
2. essentially any discrete family, including
   - Poisson
   - Geometric
   - Negative Binomial
   - Binomial
   - Multinomial
   - (and many more)

To be clear: It's fine to use DPMs ...

1. as a flexible prior on densities
   (viewing the latent variables as nuisance parameters)



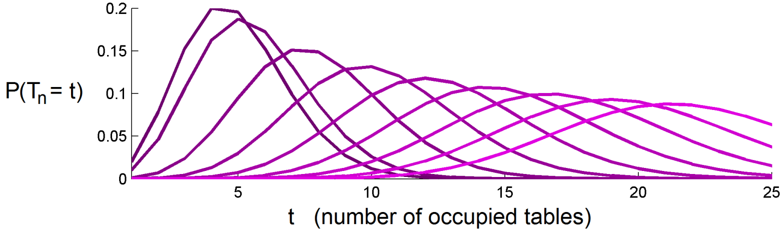
2. or if the data-generating process is well-modeled by a DPM
   (and in particular, is not a finite mixture!)



Intuition

The wrong intuition

It is tempting to think that the prior on the number of occupied tables is the culprit, since it is diverging as $n \to \infty$.



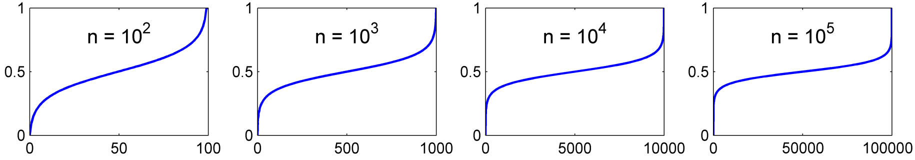
However, this is not the fundamental reason why inconsistency occurs.

The right intuition

Given that there are t occupied tables, the conditional distribution of their sizes $n_1, \ldots, n_t$ is

$$P(n_1, \ldots, n_t \mid T_n = t) \;\propto\; n_1^{-1} \cdots n_t^{-1}\, I(\textstyle\sum n_i = n).$$

CDF of n_1 given $T_n = 2$ occupied tables

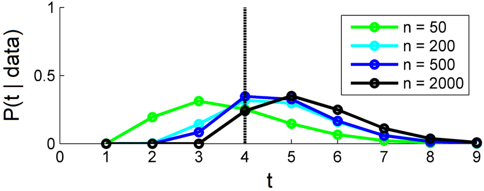


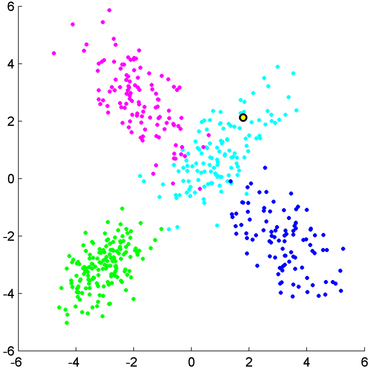
Key observation

As n grows, this becomes concentrated in the "corners". In other words, the DPM really likes to have one or more tables with very few customers.

The DPM really likes to have one or more tables with very few customers.

This explains the tiny extra clusters, since (it turns out) they do not significantly reduce the likelihood.





Solutions?

What if we . . .

- put a prior on the concentration parameter?
- ignore tables with very few customers? (busy waiter strategy)
- put a prior on the number of components?

  This works in principle (Nobile, 1994), but . . .

beware of misspecification.

Summary

The DPM posterior on the number of occupied tables should not be used to estimate the number of components in a finite mixture.

Dirichlet process mixture inconsistency for the number of components

Jeffrey W. Miller

and

Matthew T. Harrison

Brown University

Division of Applied Mathematics

Poster: Fri37

References I

E. B. Baskerville, A. P. Dobson, T. Bedford, S. Allesina, T. M. Anderson, and M. Pascual. Spatial guilds in the Serengeti food web revealed by a Bayesian group model. *PLoS computational biology*, 7(12):e1002321, 2011.

H. Chen, P. L. Morrell, V. E. T. M. Ashworth, M. de la Cruz, and M. T. Clegg. Tracing the geographic origins of major avocado cultivars. *Journal of Heredity*, 100(1):56–65, 2009.

T. Ferguson. Bayesian density estimation by mixtures of normal distributions. In M. H. Rizvi, J. Rustagi, and D. Siegmund, editors, *Recent Advances in Statistics*, pages 287–302. Academic Press, 1983.

J. Fogelqvist, A. Niittyvuopio, J. Ågren, O. Savolainen, and M. Lascoux. Cryptic population genetic structure: the number of inferred clusters depends on sample size. *Molecular ecology resources*, 10(2):314–323, 2010.

S. Ghosal. The Dirichlet process, related priors and posterior asymptotics. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics*, pages 36–83. Cambridge University Press, 2010.

E. G. Gonzalez and R. Zardoya. Relative role of life-history traits and historical factors in shaping genetic population structure of sardines (Sardina pilchardus). *BMC evolutionary biology*, 7(1):197, 2007.

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

N. Lartillot and H. Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109, 2004.

References II

A. D. Leaché and M. K. Fujita. Bayesian species delimitation in West African forest geckos (Hemidactylus fasciatus). *Proceedings of the Royal Society B: Biological Sciences*, 277 (1697):3071–3077, 2010.

A. Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357, 1984.

S. N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3):727–741, 1994.

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 M. T. Harrison. Inconsistency of Pitman–Yor process mixtures for the number of components. *arXiv:1309.0024*, 2013.

X. L. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

A. Nobile. *Bayesian Analysis of Finite Mixture Distributions.* PhD thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, 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.

C. M. Richards, G. M. Volk, A. A. Reilley, A. D. Henk, D. R. Lockwood, P. A. Reeves, and P. L. Forsline. Genetic diversity and population structure in Malus sieversii, a wild progenitor species of domesticated apple. *Tree Genetics & Genomes*, 5(2):339–347, 2009.

References III

C. Scricciolo. Adaptive Bayesian density estimation using Pitman–Yor or normalized inverse-Gaussian process kernel mixtures. *arXiv:1210.8094*, 2012.

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

M. West, P. Müller, and M. Escobar. *Hierarchical priors and mixture models, with application in regression and density estimation*. Institute of Statistics and Decision Sciences, Duke University, 1994.

E. Xing, K. Sohn, M. Jordan, and Y. Teh. Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 1049–1056, 2006.

Y. Zhou, H. Brinkmann, N. Rodrigue, N. Lartillot, and H. Philippe. A Dirichlet process covarion mixture model and its assessments using posterior predictive discrepancy tests. *Molecular biology and evolution*, 27(2):371–384, 2010.