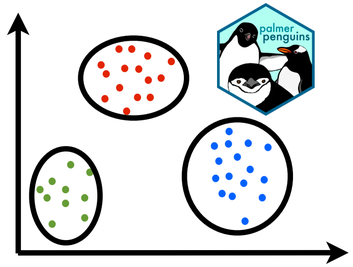
Finite mixture modeling under misspecification

Learning number of subpopulations in data:

- cell types
- gene tissue profiles
- tissue types in MRI scan
- latent genetic subpopulations
- gamma-ray burst types
- communities in a social network



[Chan et al., 2008; Prabhakran et al., 2016; Banfield et al., 1993; Lorenzen et al., 2006]

Finite mixture models are used to discover latent groups.
Finite but unknown # of latent subpopulations.

Model: put a prior on # of components (e.g. Poisson distribution)
Inference: compute posterior on # of components $\Pi(k \mid X_{1:N})$
Proved to be consistent for the # of components [Nobile, 1994]

Real data are not perfectly Gaussian — *misspecification*!

Data analysis folk wisdom: will learn too many components. But people still use finite mixtures. [Miller and Harrison, 2018]

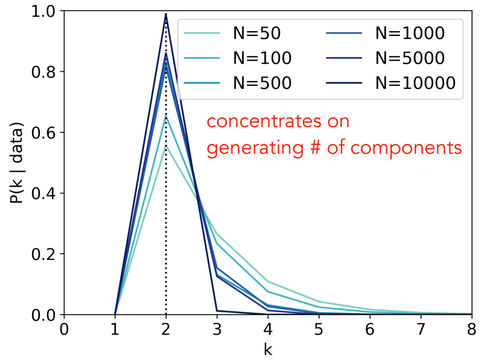
Our work: we prove that under model misspecification, the posterior number of components *diverges* to infinity.

Theorem. Suppose the data are generated from a distribution that cannot be represented as a finite mixture of our model component family.

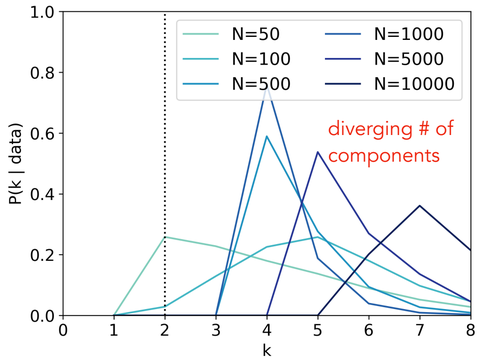
Under regularity conditions on the *prior* and *component family*, then the posterior number of components ***diverges:*** i.e., for any $k \in \mathbb{N}$,

$$\Pi(k \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 0$$

Well-specified model (Gaussian mixture data)

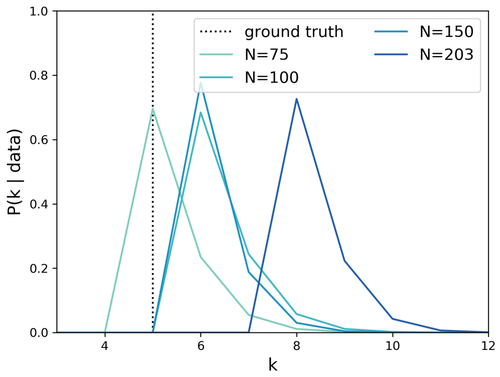


Misspecified model (Laplace mixture data)

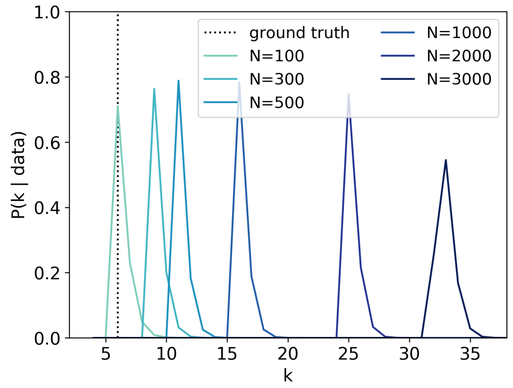


Real data experiments: High-dimensional gene expression data sets fit with Gaussian mixtures.

Cancer gene expression data



Mouse single-cell RNA-seq



The number of components varies substantially within realistic data size ranges.

Paper available at: https://arxiv.org/abs/2007.04470