Nonparametric Deconvolution Models

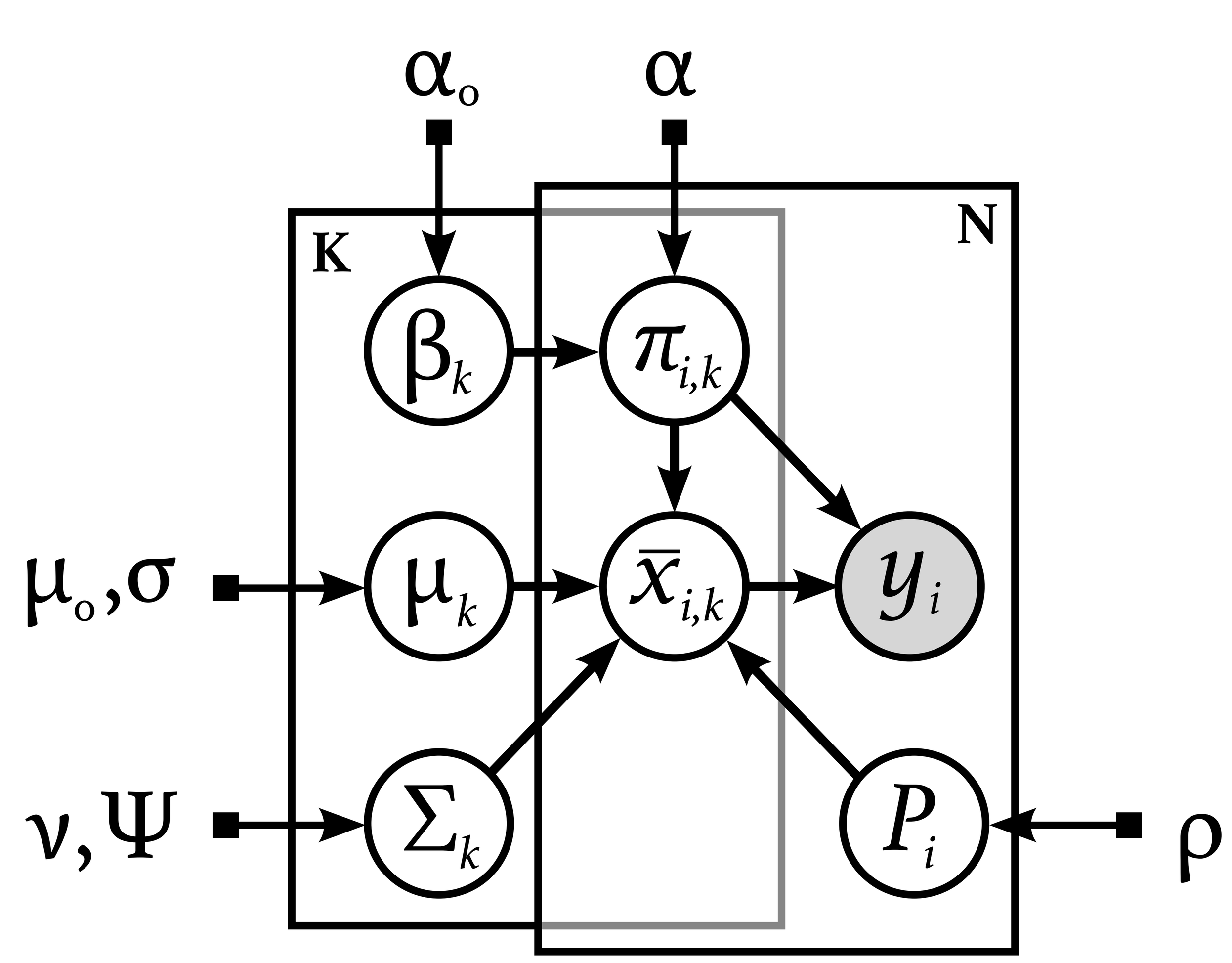
Allison Chaney, Princeton University Advisor: Barbara Engelhardt
In collaboration with Archit Verma and Young-suk Lee

Objective

We consider the problem of modeling collections of convolved data points. Specifically, each observation is composed of particles that originate from diverse factors. The objective of this work is to create a general family of models to learn 1) the features of global factors shared among all observations as well as the number and global proportions of these factors; 2) for each observation, the proportion (or membership) of particles that belong to each factor; and 3) the features of observation-specific (or local) factors for each observation. While the first two objectives are fulfilled by existing models, the final objective is unique to our model. This model framework will allow us to ask scientific questions about observations whose local factors deviate from their corresponding global factors (e.g., anomalous voting behavior or cancerous cells).

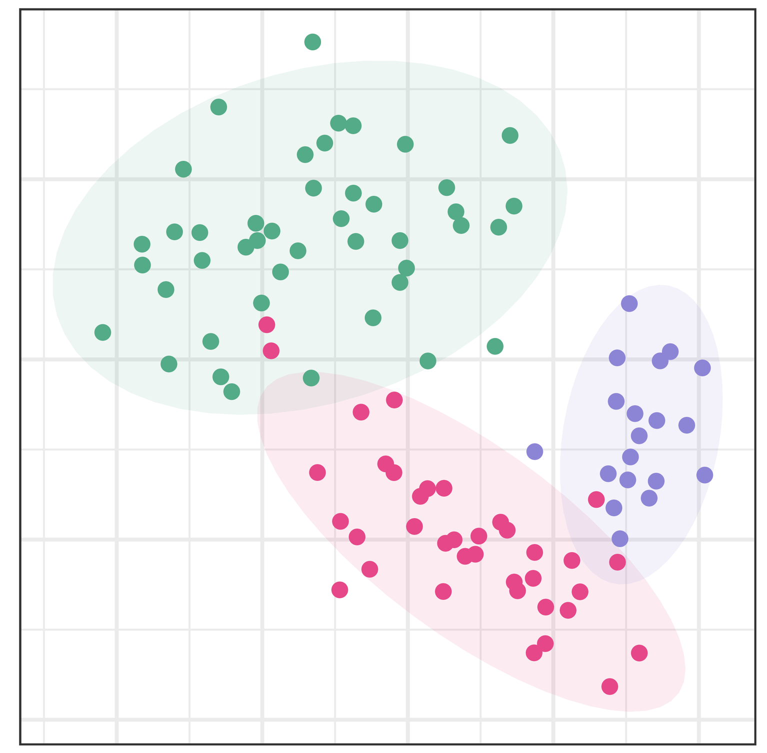
Examples of convolved observations in multiple domains

General	Voting	Bulk RNA-seq	Images
observation y_i	district vote tally	sample	image
feature m	issue or candidate	gene expression level	pixel
particle p	individual voter	one cell	light particle
factor k	voting cohort	cell type	visual pattern

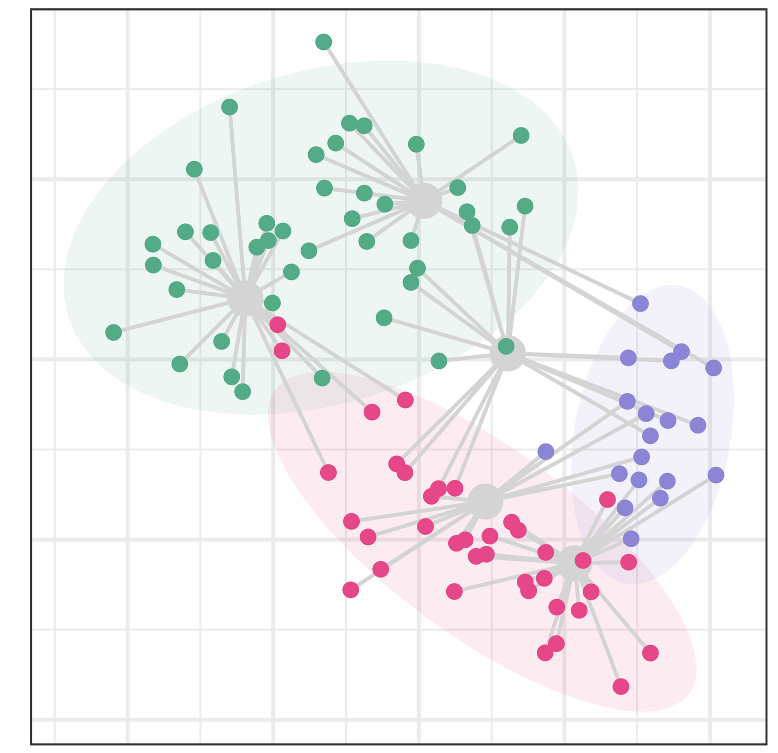


Notation			
K	number of latent factors	P	number of particles in an observation
N	number of observations	$\boldsymbol{y}$	observations
$\boldsymbol{\beta}$	global distribution of factors	$\boldsymbol{\pi}$	local distribution of factors
$\boldsymbol{\mu}$	global factor feature means	$\bar{\boldsymbol{x}}$	local factor features
$\boldsymbol{\Sigma}$	global factor feature covariances		

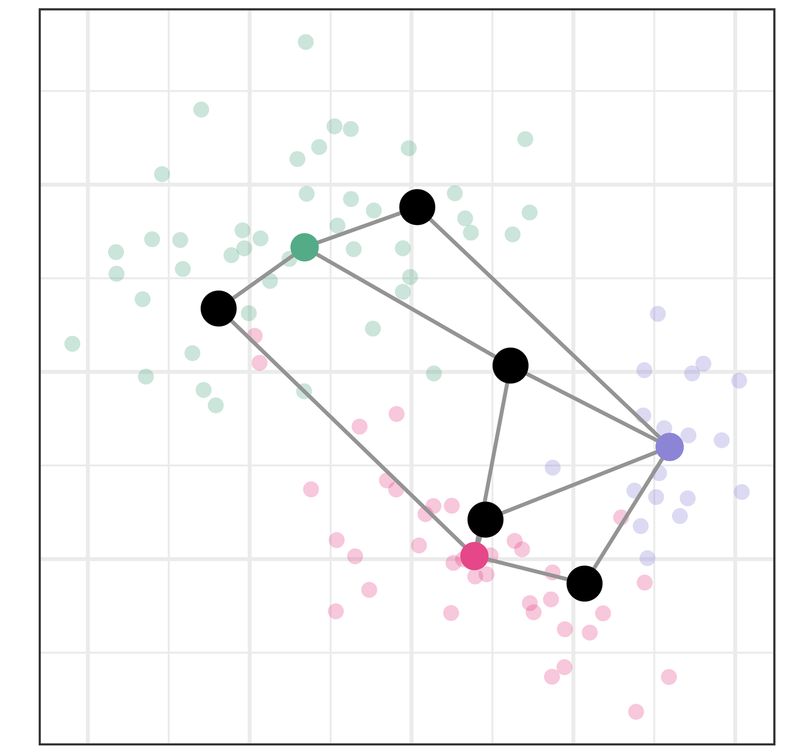
Related Models



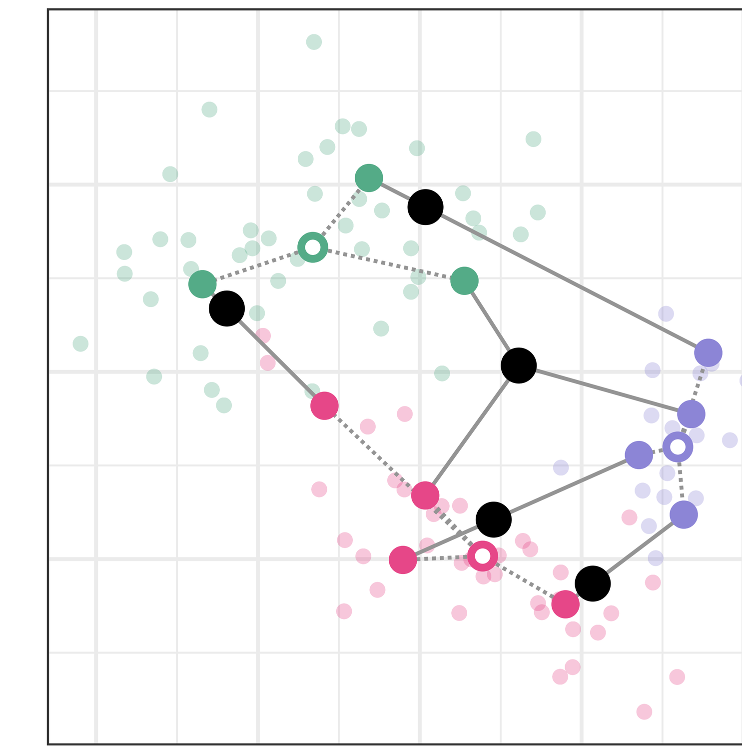
Mixture models assign each observation to one of K clusters, or factors.



Admixture models represent groups of observations, each with its own mixture of K shared factors.

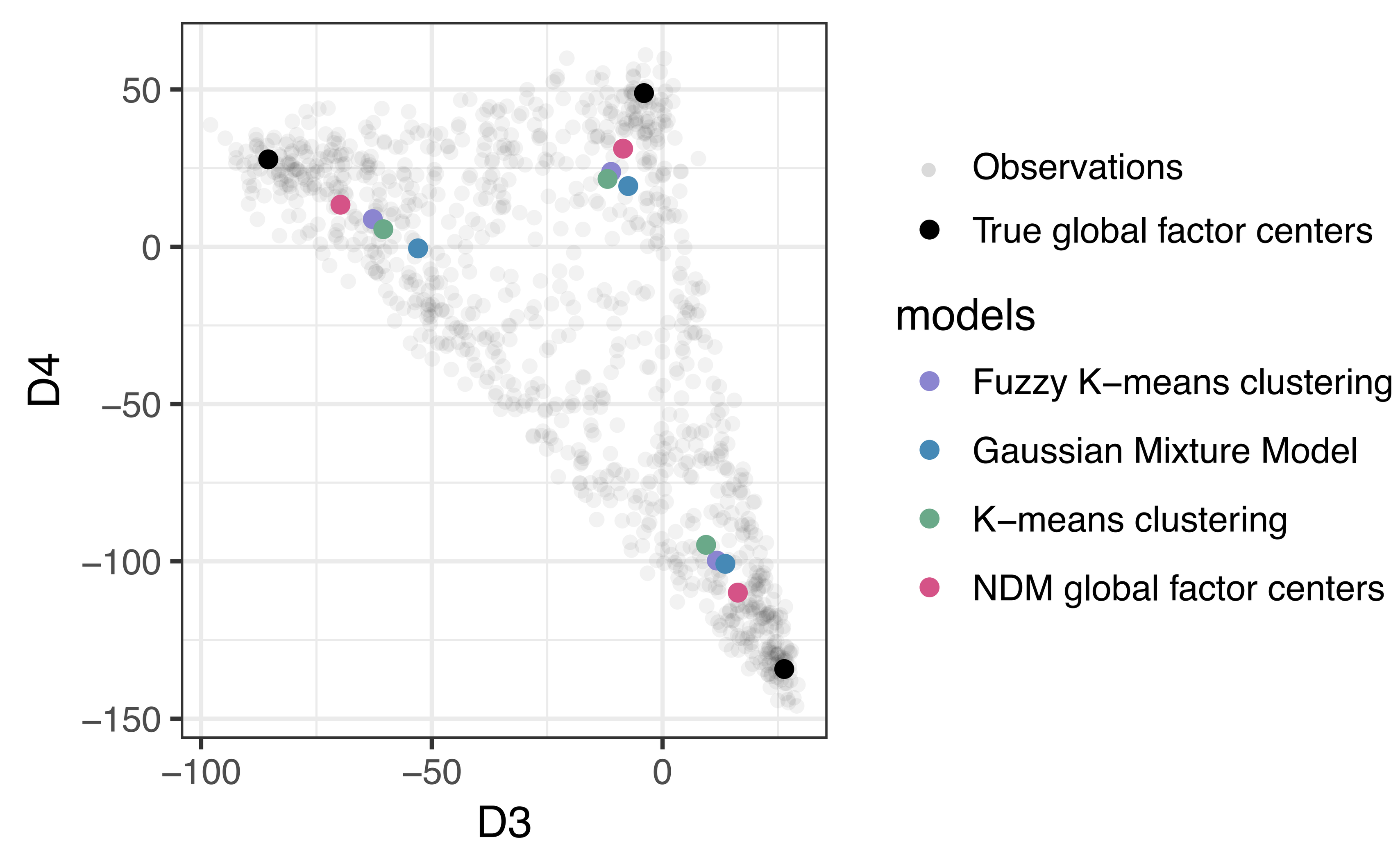


Decomposition models decompose observations into constituent parts by representing observations as a product between group representations and factor features.



Deconvolution models (this work) similarly decompose, or deconvolve, observations into constituent parts, but also capture group-specific (or local) fluctuations in factor features.

Results on Simulated Data



Inference

To support a range of model instances, we rely on black box variational inference (Ranganath, 2014); this allows us to fit a variety of models within the NDM family. We use separate split-merge procedure (Bryant, 2012) to learn the number of latent factors K.

Algorithm Pseudocode

set K to an initial value
initialize variational parameters
repeat until convergence:
  repeat until batch convergence:
    update variational parameters for $\bar{\boldsymbol{x}}, \boldsymbol{\pi}, \boldsymbol{P}, \boldsymbol{\beta}$ using BBVI
    update variational parameters for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ using analytic updates
  split/merge latent factors, defining new K and updating variational parameters accordingly

References

M. Bryant and E. B. Sudderth. Truly nonparametric online variational inference for hierarchical Dirichlet processes. NIPS, 2012.
D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. NIPS, 2001.
R. Ranganath, S. Gerrish, and D. M. Blei. Black Box Variational Inference. AISTATS, 2014.
Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. JASA, 2012.

This research was supported by an appointment to the Intelligence Community Postdoctoral Research Fellowship Program at Princeton University, administered by Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Office of the Director of National Intelligence.