Clustering Initialization

when data points are indistinguishable between clusters

ESTIMATING MIXED MULTINOMIAL CHOICE MODELS

ANDREEA GEORGESCU, MIT ORC

PRESENTED AT MIT ML RETREAT

Advisors: Retsef Levi, Vivek Farias

ESTIMATING PREFERENCES

DATA. Sales data, including assortment offered and product bought

Offered assortment:			Bought:
A	B	C	A

$$\mathbf{x} = \{\mathbf{A}, \mathbf{B}, \mathbf{C}\} | \mathbf{A}$$

TASK. Predict purchase probability for each item in each assortment

ESTIMATING PREFERENCES

DATA. Sales data, including assortment offered and product bought

TASK. Predict purchase probability for each item in each assortment

COMMON MODELING APPROACH. Fit a mixed MNL model

- **MNL MODEL**. Purchase probabilities given by $p(i|S) = \lambda_i / \sum_{j \in S} \lambda_j$
- **MIXED MNL**. There are K customer types, each following a MNL model.

$$p(i|S) = \sum_{k=1:K} \pi_k p^k(i|S)$$

- **ESTIMATION**. EM algorithm

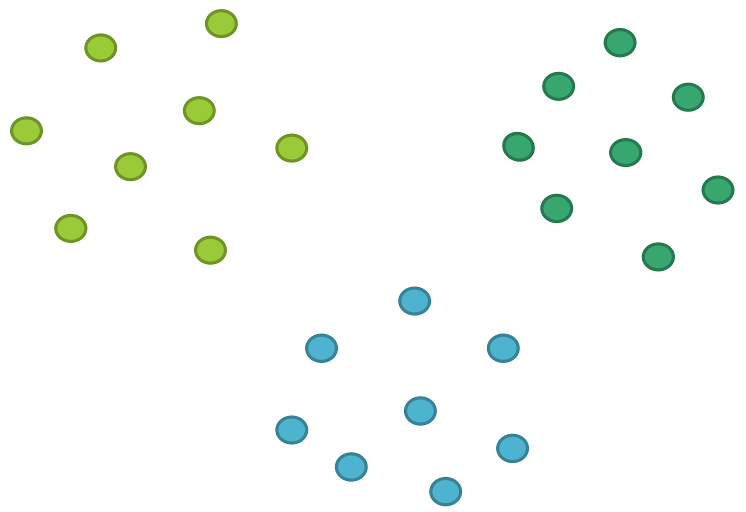
WHY IS THIS INTERESTING?

EM PERFORMANCE NOTORIOUSLY DEPENDENT ON INITIALIZATION

- Convergence with common initialization for mixed MNL is slow.

K-MEANS ~ EM FOR MIXED GAUSSIANS, $\sigma = 1$

- Efficient initialization due to [Arthur, D. 2006]



Arthur, D. and Vassilvitskii, S., 2006. *k-means++: The advantages of careful seeding*.

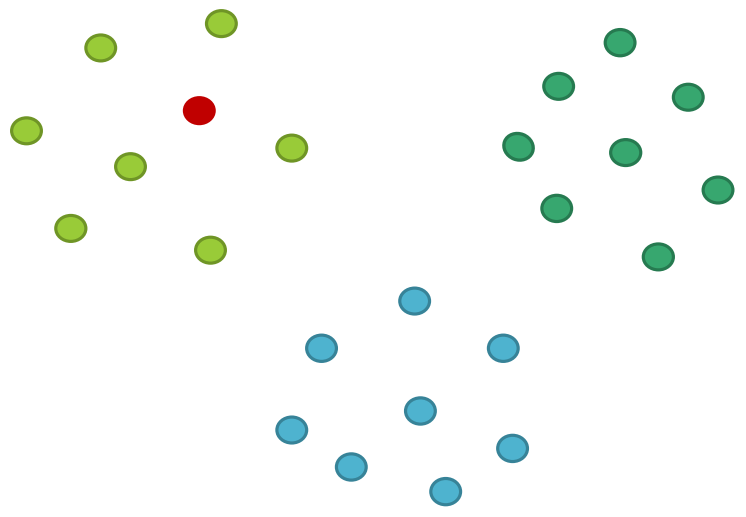
WHY IS THIS INTERESTING?

EM PERFORMANCE NOTORIOUSLY DEPENDENT ON INITIALIZATION

- Convergence with common initialization for mixed MNL is slow.

K-MEANS ~ EM FOR MIXED GAUSSIANS, $\sigma = 1$

- Efficient initialization due to [Arthur, D. 2006]



Arthur, D. and Vassilvitskii, S., 2006. *k-means++: The advantages of careful seeding*.

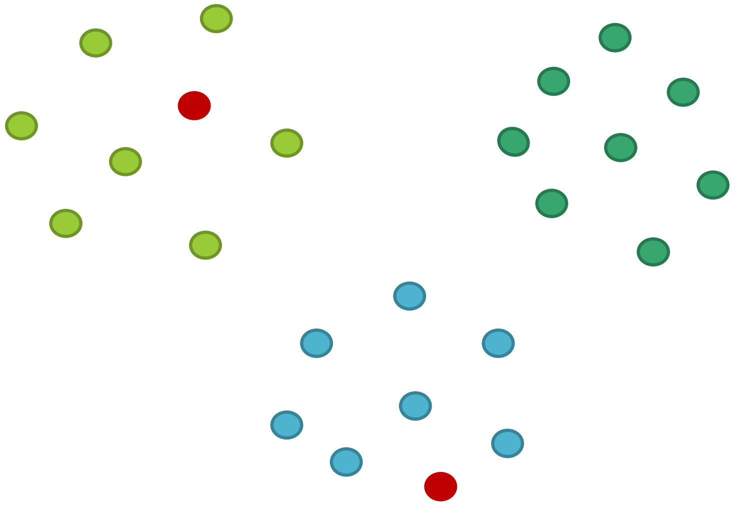
WHY IS THIS INTERESTING?

EM PERFORMANCE NOTORIOUSLY DEPENDENT ON INITIALIZATION

- Convergence with common initialization for mixed MNL is slow.

K-MEANS ~ EM FOR MIXED GAUSSIANS, $\sigma = 1$

- Efficient initialization due to [Arthur, D. 2006]



Arthur, D. and Vassilvitskii, S., 2006. *k-means++: The advantages of careful seeding*.

WHY IS THIS INTERESTING?

EM PERFORMANCE NOTORIOUSLY DEPENDENT ON INITIALIZATION

- Convergence with common initialization for mixed MNL is slow.

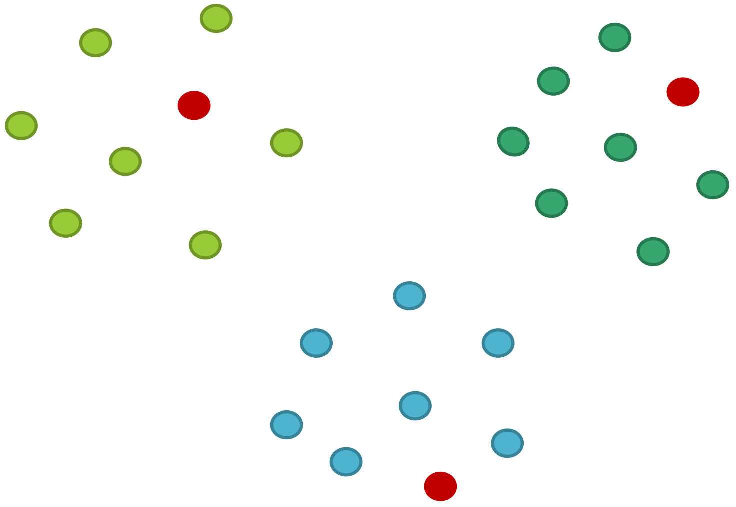
K-MEANS ~ EM FOR MIXED GAUSSIANS, $\sigma = 1$

- Efficient initialization due to [Arthur, D. 2006]

- Good seeding: centroids as different as possible
- What makes seeding possible?

Datapoints in different clusters are distinguishable

(in separate parts of the plane)



Arthur, D. and Vassilvitskii, S., 2006. *k-means++: The advantages of careful seeding*.

NAÏVE EM ++ INITIALIZATION FOR MIXED MNL

True models

MNL Model / Cluster	Product weights		
	0	1	2
A ($\pi = 0.5$)	25	25	50
B ($\pi = 0.5$)	40	20	40

Data

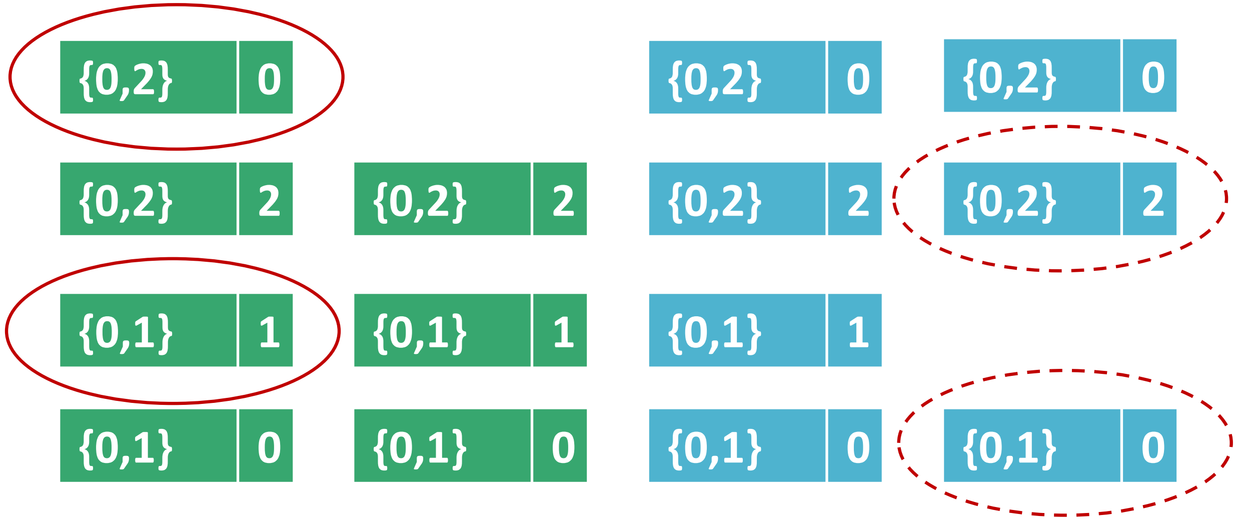
{0,2}	0			{0,2}	0	{0,2}	0
{0,2}	2	{0,2}	2	{0,2}	2	{0,2}	2
{0,1}	1	{0,1}	1	{0,1}	1		
{0,1}	0	{0,1}	0	{0,1}	0	{0,1}	0

NAÏVE EM ++ INITIALIZATION FOR MIXED MNL

True models			
MNL Model / Cluster	Product weights		
	0	1	2
A $(\pi = 0.5)$	25	25	50
B $(\pi = 0.5)$	40	20	40

Choose centroids far apart.

Data



NAÏVE EM ++ INITIALIZATION FOR MIXED MNL

True models			
MNL Model / Cluster	Product weights		
	0	1	2
A $(\pi = 0.5)$	25	25	50
B $(\pi = 0.5)$	40	20	40

Data

{0,2}	0			{0,2}	0	{0,2}	0
{0,2}	2	{0,2}	2	{0,2}	2	{0,2}	2
{0,1}	1	{0,1}	1	{0,1}	1		
{0,1}	0	{0,1}	0	{0,1}	0	{0,1}	0

Choose centroids far apart.

Model estimated:

Product 0 is ~never bought in {0,1,2} as opposed to 32.5% of time.

EM ++ INITIALIZATION FOR MIXED MNL

HOW IS EM INITIALIZED FOR MIXED MNL?

- Split observations into k groups randomly and estimate an MNL on each cluster.
- The k smaller datasets will be very similar statistically, so the k initial MNL models are very close to each other
- **Leads to slow convergence.**

EM ++ INITIALIZATION FOR MIXED MNL

LEVERAGE MNL STRUCTURE TO CONSTRUCT ++ STYLE SEEDING

MNL Signature:

$$r(S) = \left[p(i|S) \big/ p(no\ item\ bought \mid S) \right]_{i \in S}$$

- MNL: r(S) constant in all assortments
- MNL: fully specified by r(S) (on S)
- Mixed MNL: r(S) linear combinations of $r^k(S)$

EM ++ INITIALIZATION FOR MIXED MNL

LEVERAGE MNL STRUCTURE TO CONSTRUCT ++ STYLE SEEDING

MNL Signature:

$$r(S) = \left[{p(i|S)}/{p(no\ item\ bought\ |\ S)} \right]_{i \in S}$$

- MNL: r(S) constant in all assortments
- MNL: fully specified by r(S) (on S)
- Mixed MNL: r(S) linear combinations of $r^k(S)$

- If r(S) is relatively stable on all assortments S observed, MNL is enough.
- If not, then every $r^k(S)$ gives a different MNL, in the space spanned by the mixtures.

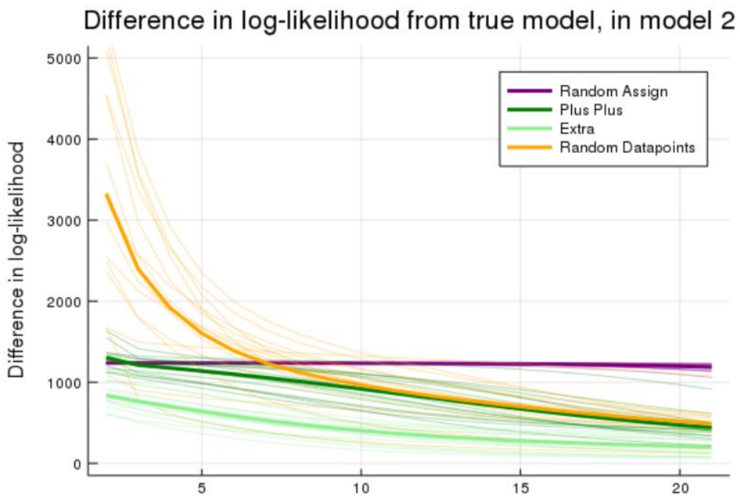
EM ++ INITIALIZATION FOR MIXED MNL

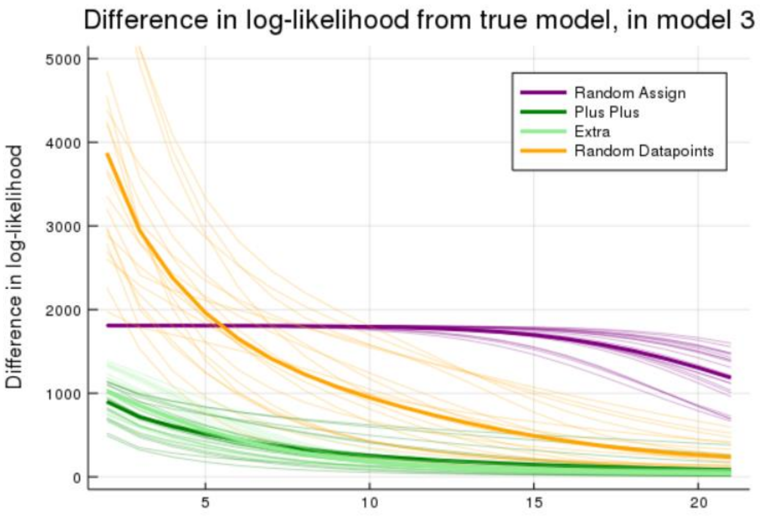
FINAL SEEDING TECHNIQUE

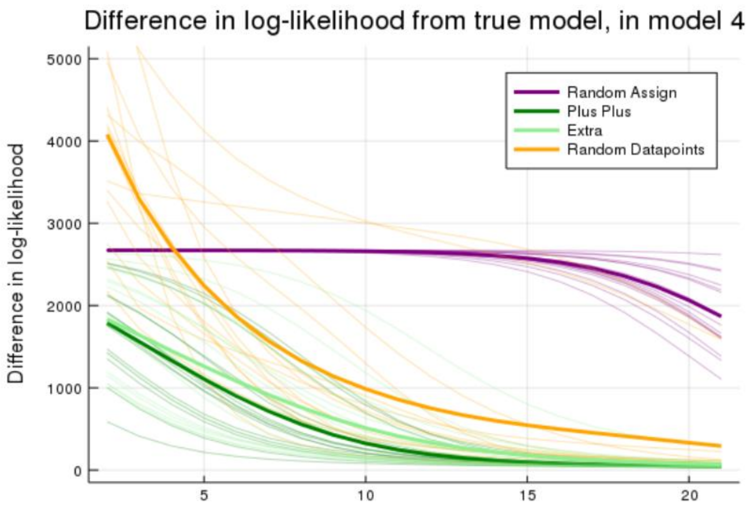
- Compute r(S) for every assortment in the dataset.
- Technique 1: Random
  - Choose K r(S) randomly and initialize EM with MNL models they specify.
- Technique 2: Plus Plus
  - Choose K r(S) that are furthest apart *(how far apart the MNL models induced are)* – inspired by the original k-means ++.

- Results:
  - Both methods significantly outperform classic initialization.
  - Technique 2 is more stable.

RESULTS ON SYNTHETIC DATA







Classic initialization (datapoints are split into K similar clusters)
Plus Plus (Method 2)
Described on last slide.
Random (Method 1)

RESULTS ON SYNTHETIC DATA

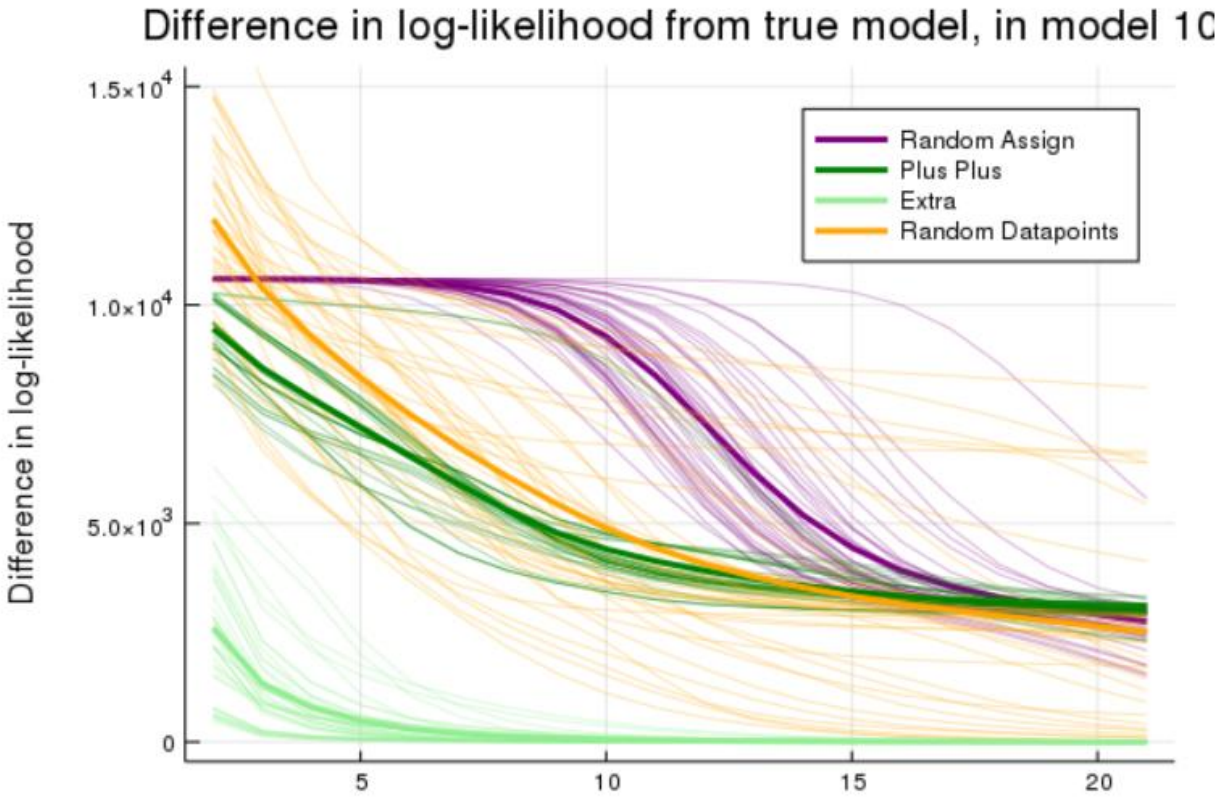
MAIN LIMITATION OF SEEDING

Initial MNL models have information only on some products

LIGHT GREEN METHOD

Ratios by assortment are imputed using the known linear combination and true ratios.

Then ++ seeding applied.



Appendix

EM-algorithm and fuzzy k-means

In fuzzy k-means we are minimizing through block coordinate descent on $q, \lambda, c = (c_1, \dots, c_K)$ the function

$$\min_{q,\lambda,c_1,\dots,c_K} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} q_i(k) \|y_i - c_k\|^2 + q_i(k) \log \frac{q_i(k)}{\lambda(k)} \right) \textit{ subject to } \sum\nolimits_{k\in[K]} q_i(k) = 1, \forall i, \qquad \sum\nolimits_{k\in[K]} \lambda(k) = 1$$

In EM we are minimizing through coordinate descent on $q, \lambda, u = (u_1, \dots, u_K)$ the function

$$\min_{q,\lambda,u_1,\dots,u_K} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} q_i(k)(-\log \mathbb{P}[y_i; u_k]) + q_i(k) \log \frac{q_i(k)}{\lambda_k} \right) \textit{ subject to } \sum\nolimits_{k\in[K]} q_i(k) = 1, \forall i, \qquad \sum\nolimits_{k\in[K]} \lambda_k = 1$$

The two problems are equivalent (and the same for mixed Gaussians)

- $c_1, \dots, c_K$ and $u_1, \dots, u_K$ are the centroids of the clusters, the objects that describe the cluster properties.
- Distance is Euclidean distance in k-means and negative log likelihood in EM.
- q gives the fuzzy cluster partition of data points, and it is regularized through entropy.
- λ is the relative size of the clusters.