Memoized Online Variational Inference for Dirichlet Process Mixture Models

BROWN

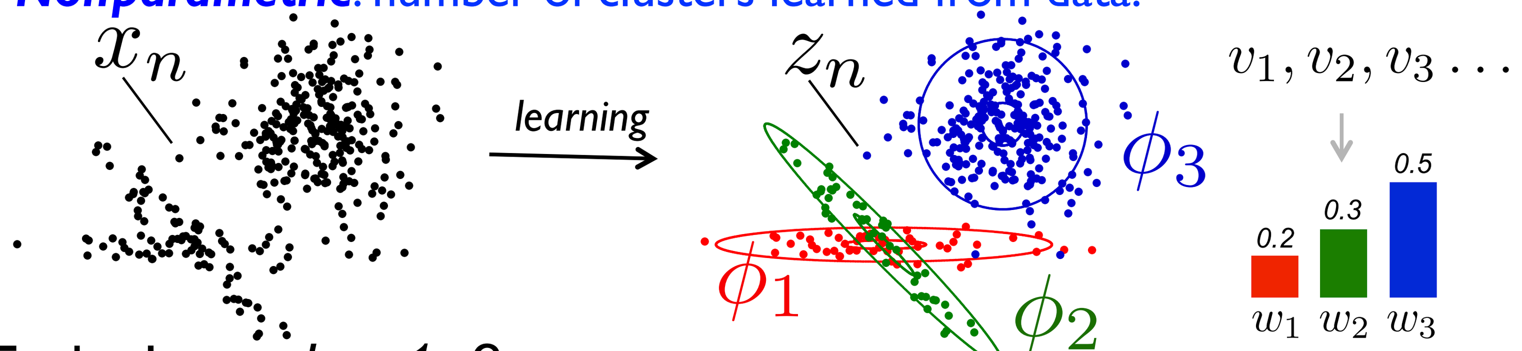
Michael C. Hughes and Erik B. Sudderth

Python code: bitbucket.org/michaelchughes/bnpy/

Dirichlet Process Mixture Model

Assigns data to discrete clusters

Nonparametric: number of clusters learned from data.



Each cluster *k* = 1, 2, … :

- Stick fraction $v_k \sim \text{Beta}(1, \alpha_0)$
- Appearance probability $w_k = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell)$ *stick-breaking*
- Data-generating parameter $\phi_k \sim H(\lambda_0)$

Each data item *n* = 1, 2, … *N*:

- Draw cluster assignment $z_n \sim \text{Discrete}(w_1, w_2, \dots)$
- Draw observed data $x_n \sim \text{F}(\phi_{z_n}) = \exp\left(\phi_{z_n}^T t(x_n) - a(\phi_{z_n})\right)$ *exponential family*

Algorithms **generalize** to any likelihood F, not just Gaussian

Multivariate Gaussian likelihood **F**

$x_n \sim \mathcal{N}(\mu_{z_n}, \Lambda_{z_n}^{-1})$ $\quad t(x_n) = [x_n \quad x_n x_n^T]$

mean, precision matrix — *sufficient statistics*

Summary

Memoized online (MO) variational inference

- No pesky learning rates, insensitive to batch size

New online moves add/remove clusters on-the-fly

- *Birth*: add useful clusters, escape local optima
- *Merge*: remove redundancy, improve speed

MO-BM *(MO with births and merges):*

Scalable, robust exploration of nonparametric posterior.

Start with just K=1 cluster, grow as needed!

Stochastic Online (SO)

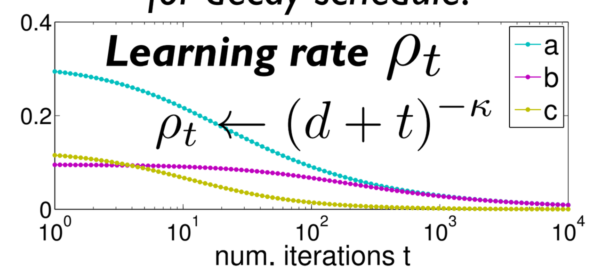
At batch b, perform usual E-step, then [Hoffman et al. JMLR '13]

Update global factors via **noisy** gradient

$\lambda_k^b \leftarrow \lambda_0 + \frac{N}{|\mathcal{B}_b|} s_k^b$ ***M-step*** *amplifies current batch*

$\lambda_k \leftarrow \rho_t \lambda_k^b + (1-\rho_t)\lambda_k$ ***Gradient step*** *natural gradients make updates simple*

Many options (a,b,c,…) for decay schedule.



Finds (local) optima of full-data objective in expectation.

Sensitive *to learning rate schedule and batch size. Careful tuning required.*

Birth Moves

Escape poor solutions by adding useful clusters.

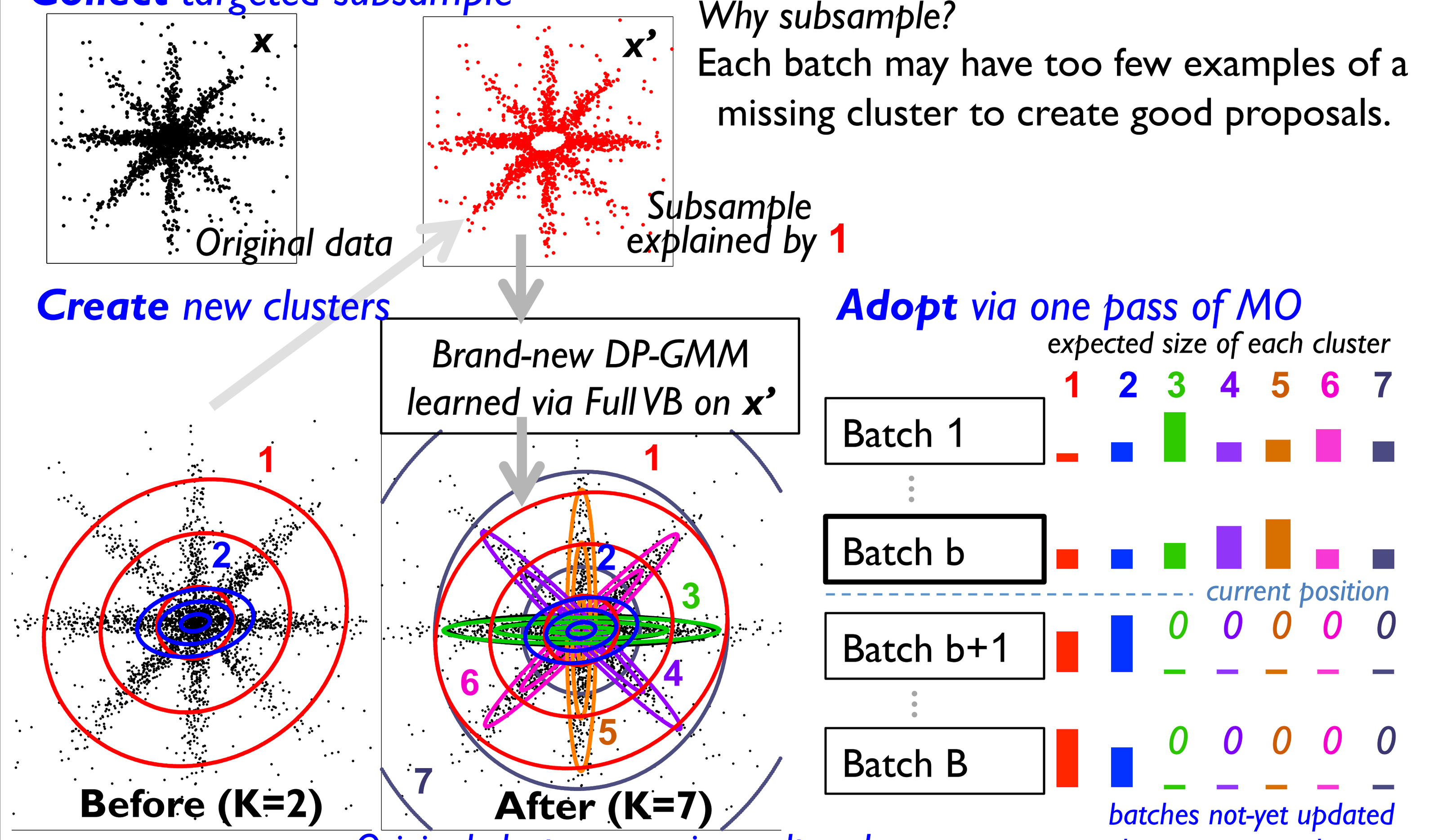
- Each move adds many clusters via fresh analysis of one cluster's data.

Collect *targeted subsample*

Why subsample? Each batch may have too few examples of a missing cluster to create good proposals.

Create *new clusters*

Adopt *via one pass of MO*

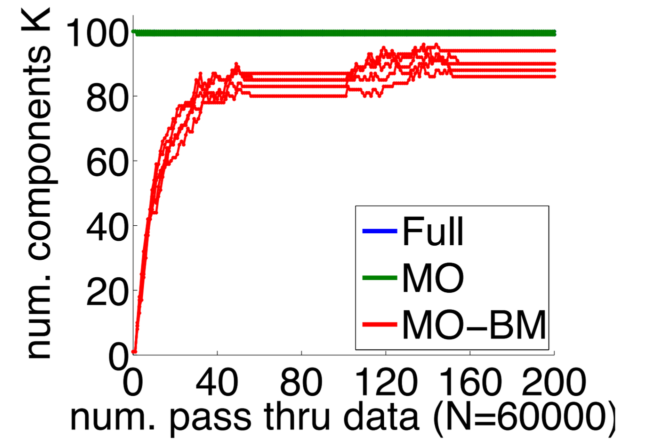


Original clusters remain unaltered.
Expansion possible via ***nested*** *truncation.*

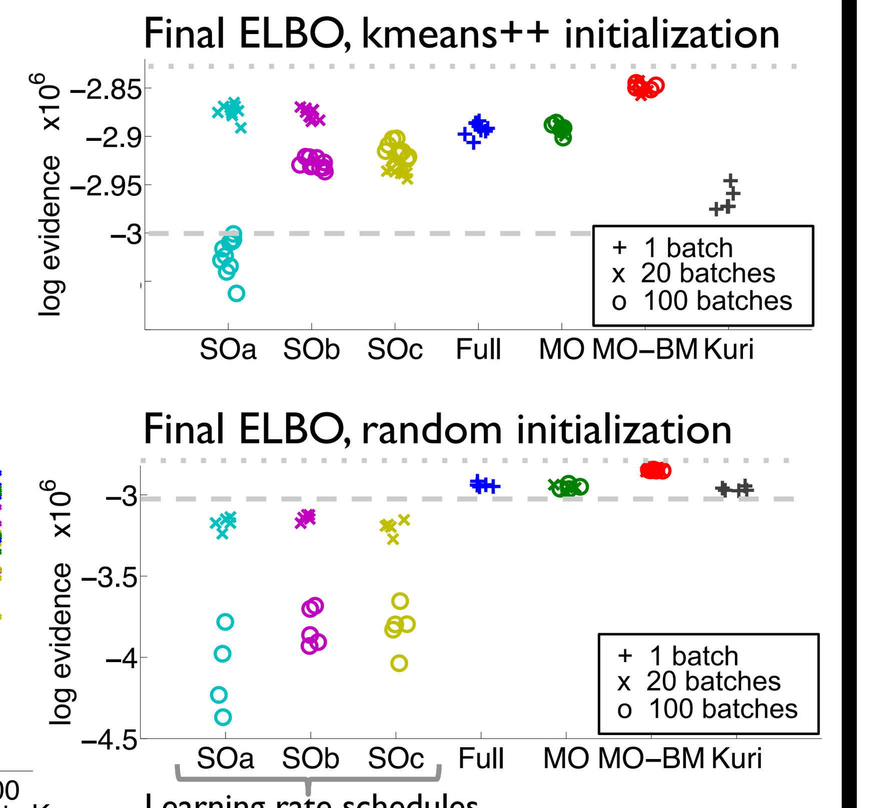
MNIST Handwritten Digits

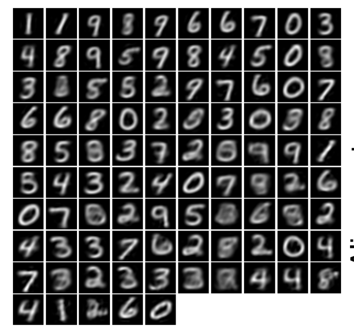
Cluster 60000 images of digits 0-9. PCA projected to 50 dimensions.
10 runs of each algorithm, from 10 fixed sets of initial parameters.

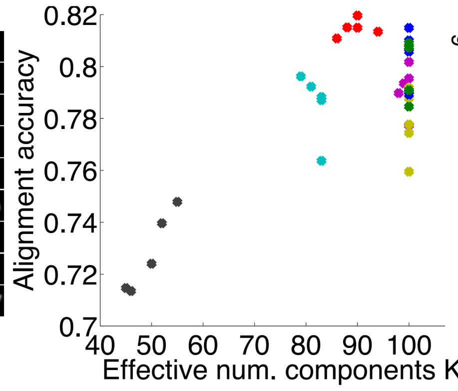
MO-BM started at K=1 discovers >80 useful clusters via births



MO-BM from K=1 reaches better ELBO than smart initializations with K=100







MO-BM estimated clusters have best many-to-one alignment to true digits 0-9

MO reliable, while SO very sensitive to learning rate, # batches, & initialization

Variational Bayes Inference (VB)

Algorithm that finds approximate posterior *q*

- Coordinate ascent optimization, minimizes KL divergence
- Like EM, but learns *distributions* not just point estimates

$p(z, v, \phi | x) \approx \prod_{n=1}^{N} q(z_n) \prod_{k=1}^{\infty} q(v_k) q(\phi_k)$

assume q is ***factorized***

Truncation to K clusters $q(z_n > K) = 0$
is ***nested:*** *allows K to grow/shrink*

Data-specific factors

$q(z_n) = \text{Disc}(r_{n1}, \dots r_{nK})$

r_{nk} *Posterior "responsibility" cluster k has for item n*

$N_k^0 = \sum_{n=1}^{N} r_{nk}$ *Expected size of cluster k*

Global factors

$q(\phi_k) = H(\lambda_k)$

$q(v_k) = \text{Beta}(\alpha_{k1}, \alpha_{k0})$

Updates just simple function of $\{N_k^0\}_{k=1}^K$

Update at each iteration

For data item n = 1, 2, … N:
$r_{n1} \dots r_{nK} \leftarrow \text{Estep}(x_n, \alpha, \lambda)$
$r_{nk} \propto e^{\mathbb{E}_q[\log p(z_n = k|v) + \log p(x_n|z_n = k, \phi)]}$

For cluster k = 1, 2, … K:
$s_k^0 \leftarrow \sum_{n=1}^{N} r_{nk} t(x_n)$ *Expected sufficient stats*
$\lambda_k \leftarrow \lambda_0 + s_k^0$ ***M-step***

Process **entire** dataset between global updates.
Slow to propagate information.

Evidence lower bound (ELBO) objective $\log p(x) \geq \mathcal{L}(q)$

$\mathcal{L}(q) = \sum_{k=1}^{K} \mathbb{E}[\phi_k]^T s_k^0 - N_k^0 \mathbb{E}[a(\phi_k)] + N_k^0 \mathbb{E}[\log w_k] - \sum_{n=1}^{N} r_{nk} \log r_{nk} + \mathcal{L}(q(v), q(\phi))$

linear function of sufficient statistics — *q(z) entropy* — *global factors*

Memoized Online (MO)

New variational algorithm, inspired by [Neal & Hinton '99]

- Analyze huge datasets by dividing into small, fixed batches
- Modest memory required, but still scales to millions of examples
- Several passes through all batches yield quality solutions

Update for each batch b

$r(\mathcal{B}_b) \leftarrow \text{Estep}(x(\mathcal{B}_b), \alpha, \lambda)$

For cluster k = 1, 2, … K:
$s_k^0 \leftarrow s_k^0 - s_k^b$
$s_k^b \leftarrow \sum_{n \in \mathcal{B}_b} r_{nk} t(x_n)$ *Expected sufficient stats*
$s_k^0 \leftarrow s_k^0 + s_k^b$
$\lambda_k \leftarrow \lambda_0 + s_k^0$ ***M-step***

Data	Batch Summaries
$x(\mathcal{B}_1)$	$s_1^1 \quad s_2^1 \quad \cdots \quad s_K^1$
$x(\mathcal{B}_2)$	$s_1^2 \quad s_2^2 \quad \cdots \quad s_K^2$
$x(\mathcal{B}_b)$	$\vdots \quad \vdots \quad \vdots$
$x(\mathcal{B}_B)$	$s_1^B \quad s_2^B \quad \cdots \quad s_K^B$

Global factors updated at ***every*** *batch.*

Global Summary

$s_1^0 \quad s_2^0 \quad \cdots \quad s_K^0$

$s_k^0 = s_k^1 + s_k^2 + \dots s_k^B$
Global summaries are ***additive***

ELBO objective:

Exact *full-dataset objective via cached entropy at each batch b:*

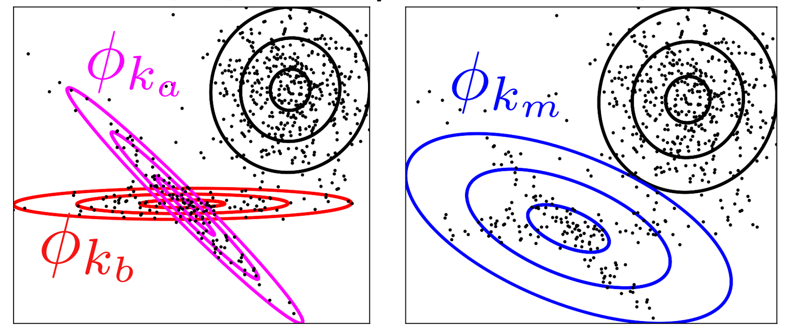
$H_k^b = -\sum_{n \in \mathcal{B}_b} r_{nk} \log r_{nk}$ *q(z) entropy*

Aggregate across batches $H_k^0 = H_k^1 + H_k^2 + \dots H_k^B$
allows runtime ***independent*** *of dataset size N*

Merge Moves

Merge two clusters into one. Simpler models & faster learning.

- ***Online*** *proposal, requires no batch processing.*
- *Run many proposals after each pass.*



New cluster takes over all responsibility for any data assigned to old clusters.
$r_{nk_m} \leftarrow r_{nk_a} + r_{nk_b}$

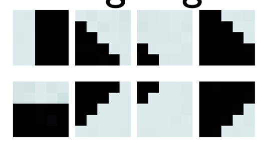
Direct construction of global summaries:
$s_{k_m}^0 \leftarrow s_{k_a}^0 + s_{k_b}^0$ *additivity*

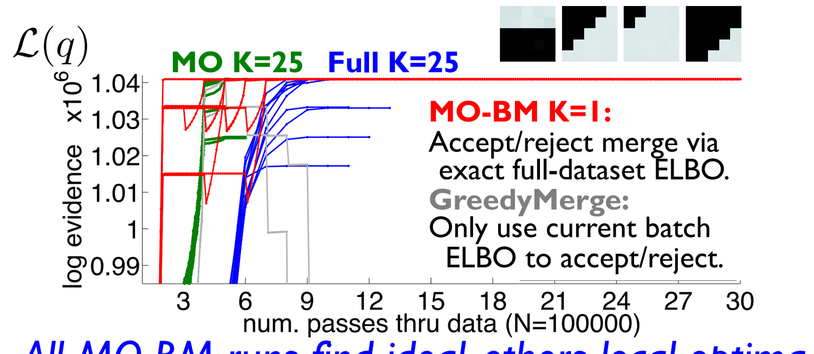
Accept/reject decision via **exact,** full-dataset ELBO comparison

accept if $\mathcal{L}(q_{merge}) > \mathcal{L}(q)$ Requires cached entropy H_{k_a, k_b}^b for all pairs.

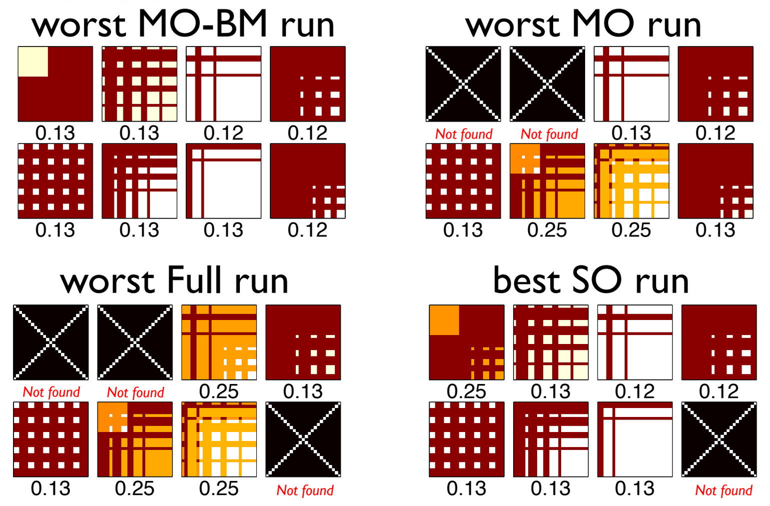
Toy Data

5x5 image patches, with strong edges
K=8 true clusters





All MO-BM runs find ideal, others local optima.
Exact ELBO essential to successful merges.

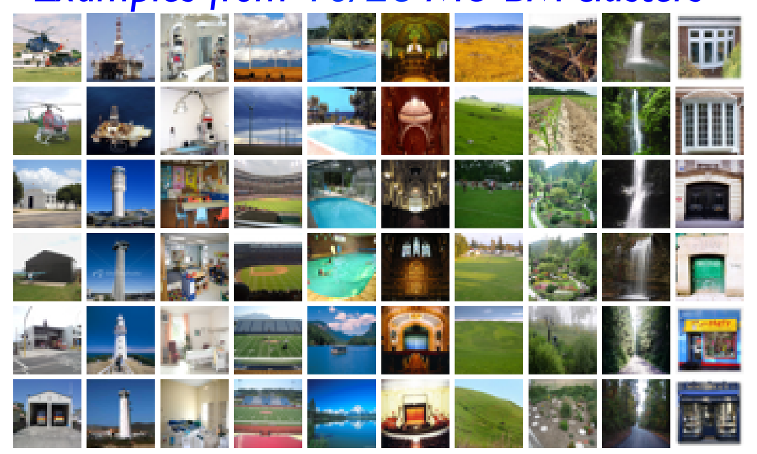


Covariances of estimated clusters

SUN Scene Categories

Cluster 108,754 tiny images (32x32 pixels). PCA projected to 50 dims.

Examples from 10/28 MO-BM clusters



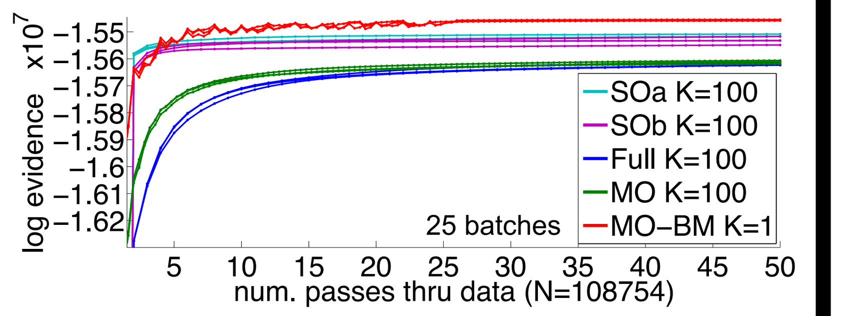
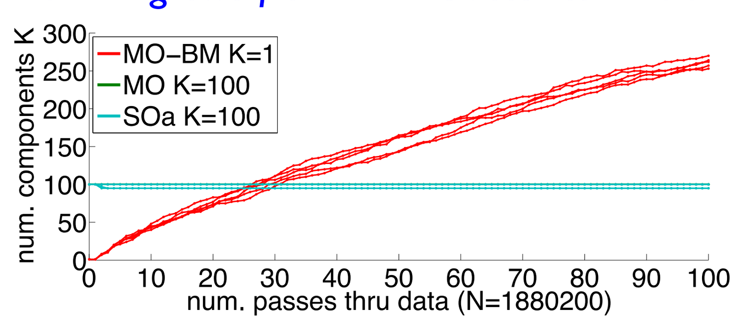


Image Patches

Cluster 1.88 million 8x8 patches from Berkeley Segmentation.

MO-BM grows from K=1 cluster to >250



MO-BM final ELBO better than competitors

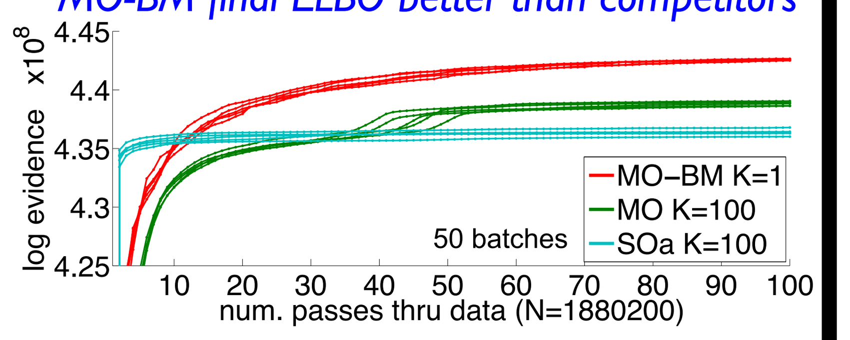


Image denoising. *Expected patch log likelihood* [Zoran & Weiss ICCV '11]
MO-BM final PSNR within 0.05 dB of best published GMM.