Variational Inference for DPGMM with Coresets

Zalán Borsos, Olivier Bachem, Andreas Krause

ETH zürich — Learning & Adaptive Systems

Overview

Goal: perform approximate **posterior inference** for finite and infinite Gaussian mixture models under **time** and **memory constraints**

Method: variational inference with **coresets**

Contributions: a novel coreset construction algorithm for posterior inference for **BGMM** and **DPGMM**

Why Coresets?

Coresets are weighted subsets of the data, with strong performance guarantees for a *specific* problem.

$$\mathrm{cost}(\mathbf{C}, \mathbf{Q}) = \sum_{(\gamma, \mathbf{x}) \in \mathbf{C}} \gamma f(\mathbf{x}, \mathbf{Q})$$

coreset query weights

$$\underbrace{|\mathrm{cost}(\mathbf{X}, \mathbf{Q}) - \mathrm{cost}(\mathbf{C}, \mathbf{Q})| \leq \varepsilon \mathrm{cost}(\mathbf{X}, \mathbf{Q})}_{\text{strong coreset [3, 6]}} + \underbrace{\varepsilon\Delta}_{\text{lightweight coreset [1]}}, \quad \forall \mathbf{Q}$$

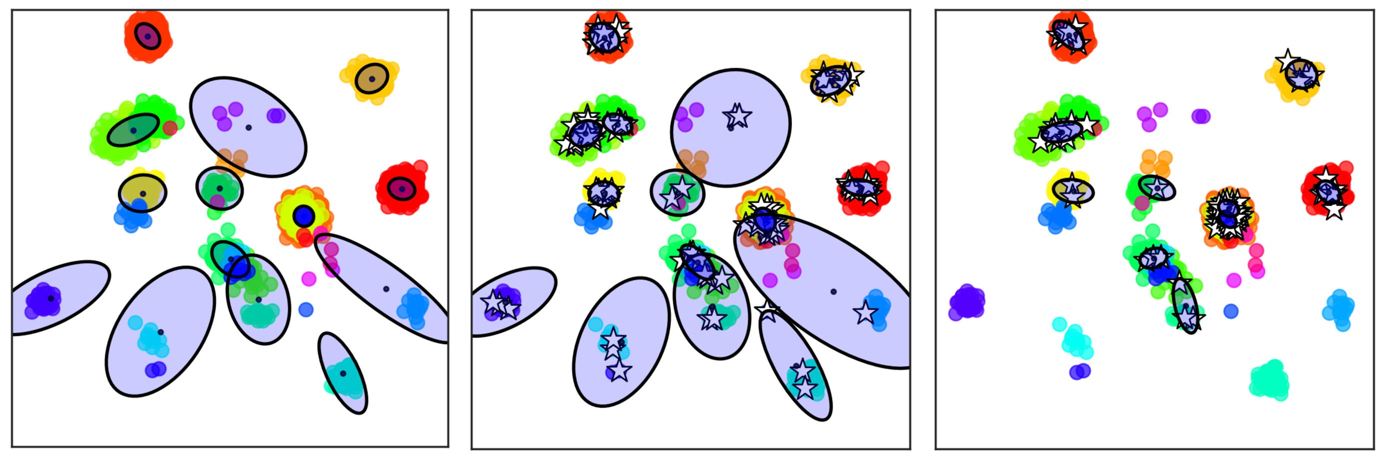


Fig. 1. Posterior means and covs of DPGMM on the full data, coreset and uniform subsample of size 2%

	easy to sample	N/M speedup	theoretical guarantees	captures large components	captures small components
coreset	✓	✓	✓	✓	✓
uniform	✓	✓	✗	✓	✗

Coresets for VI in BGMM and DPGMM

Algorithm 1 Coreset for GMM

Input: $\mathbf{X}$ data set, M summary size

$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$

for $\boldsymbol{x} \in \mathbf{X}$ **do**

  $q(\boldsymbol{x}) = \frac{1}{2N} + \frac{||\boldsymbol{x}-\boldsymbol{\mu}||^2}{2\sum_{n=1}^{N}||\boldsymbol{x}_n-\boldsymbol{\mu}||^2}$

end for

$\mathbf{C} \leftarrow$ sample M points with probability $q(\boldsymbol{x})$ from $\mathbf{X}$ and assign weights $\gamma_{\boldsymbol{x}} = \frac{1}{M \cdot q(\boldsymbol{x})}$

return Coreset $\mathbf{C}$

- coresets constructed for the log-likelihood can be used for posterior inference [4]
- the size of previous coresets for GMM log-likelihood [5] depends on $\kappa(\Sigma)$
- integration over GMM parameter space
  $$\boldsymbol{\theta} = [(w_1, \boldsymbol{\mu}_1, \boldsymbol{\Lambda}_1), ..., (w_T, \boldsymbol{\mu}_T, \boldsymbol{\Lambda}_T)]$$
  is problematic

Theorem. For $M \in \Omega\left(\frac{D^4 T^4 + \log\frac{1}{\delta}}{\varepsilon^2}\right)$ Alg 1 return $\mathbf{C}$ s.t. w.p. $1-\delta$:

$$|\phi(\mathbf{X}|\boldsymbol{\theta}) - \phi(\mathbf{C}|\boldsymbol{\theta})| \leq \varepsilon\phi(\mathbf{X}|\boldsymbol{\theta}) + \varepsilon\sum_{t=1}^{T}\mathrm{Tr}(\boldsymbol{\Lambda}_t)\sum_{n=1}^{N}||\boldsymbol{x}_n - \boldsymbol{\mu}||^2 \quad (1)$$

where $\phi(\mathbf{X}|\boldsymbol{\theta}) = -\mathcal{L}(\mathbf{X}|\boldsymbol{\theta}) + n \cdot \ln\sum_{t=1}^{T}\frac{w_t}{\sqrt{|2\pi\boldsymbol{\Lambda}_t^{-1}|}}$

Corollary. Similar approximation guarantee holds for the ELBO.

✓the dependence on $\kappa(\Sigma)$ in the error guarantee and not on the coreset size
→ suitable both for ML and posterior inference in (B)GMMs.

truncated VI for DPMM [2]: truncated q at T

↓

handle infinite mixtures

$\mathbb{E}_{\sim q}[(1)]$ ⟶ Alg 1 works for DPGMM too

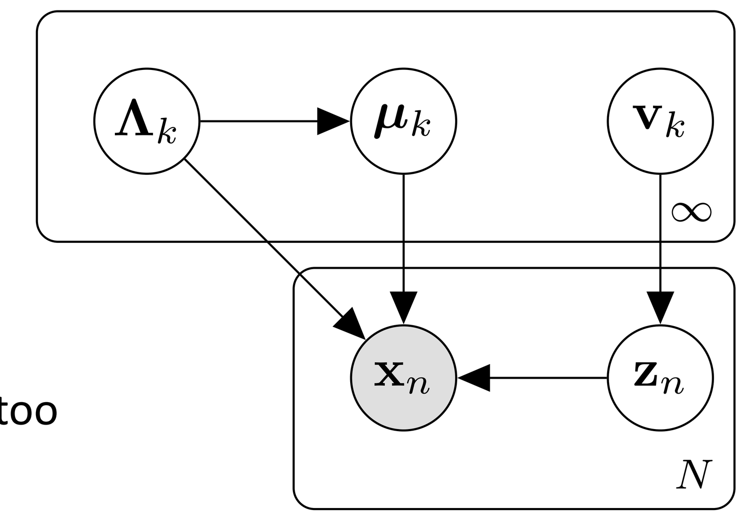
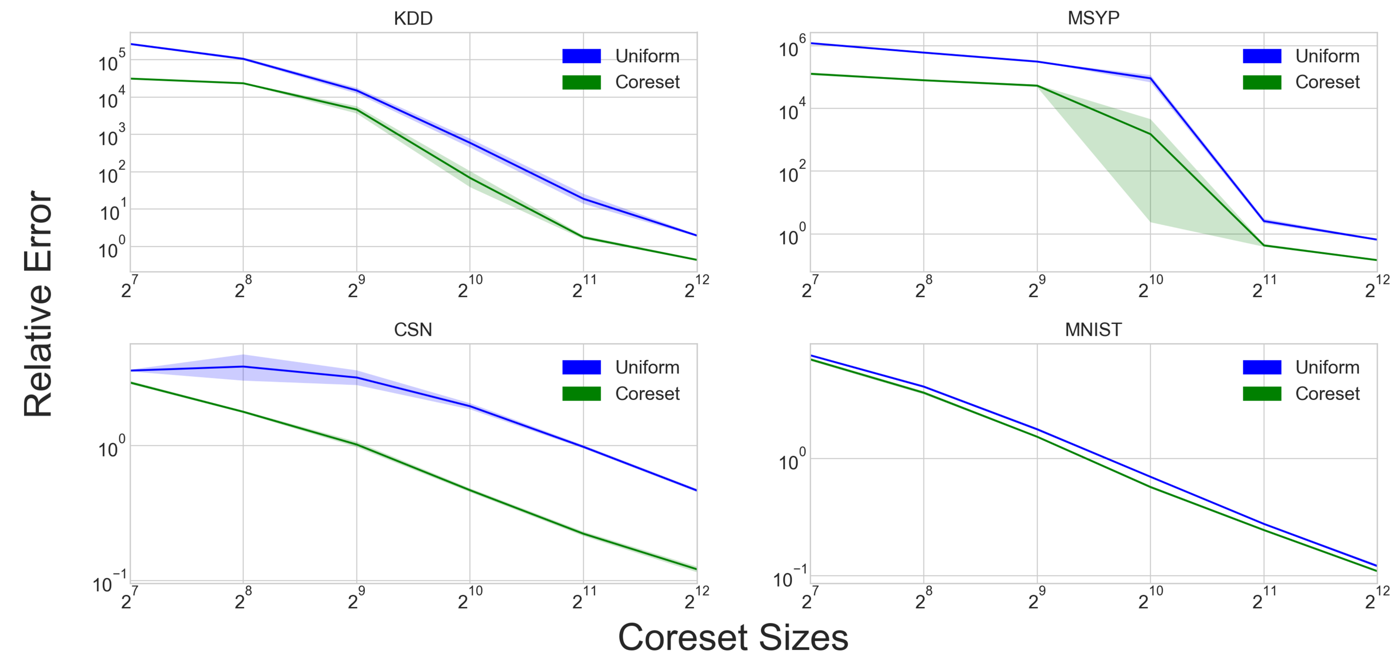


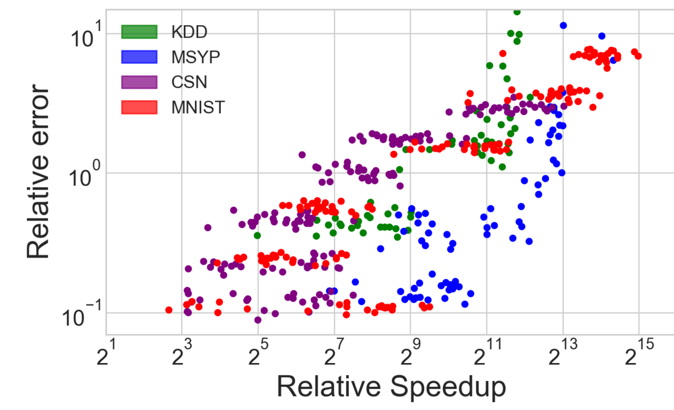
Fig. 2. Graphical representation of DPGMM

Alg 1 can be used with **weighted**: ✓ CAVI ✓ SVI ✓ ADVI ✓ BBVI

repeat
  ...
  for $n = 1$ **to** N **do**
    for $t = 1$ **to** T **do**
      $\phi_{nt} = \exp\left(\mathbb{E}[\ln v_t] + \sum_{i=1}^{t-1}\mathbb{E}[\ln(1-v_i)] + \mathbb{E}[|\boldsymbol{\Lambda}_t|] - \frac{D}{2\beta_t} - \frac{\nu_t}{2}(\mathbf{x}_n - \mathbf{m}_t)^\intercal \mathbf{W}_t(\mathbf{x}_n - \mathbf{m}_t)\right)$
    end for
    renormalize ϕ_n
    $\phi_n = \gamma_n \phi_n$ ▷ weight modification
  end for
until convergence

Experiments and Discussion





↑ Fig. 3. Relative approximation error of the held-out log likelihood with coresets and uniform sampling under DPGMM. Optimization via weighted CAVI.

← Fig. 4. Speedup-accuracy tradeoff, with coreset construction time included

- same coreset construction as for K-Means [1] works for BGMM and DPGMM
- coresets help if sampling distribution q has low entropy ← the data is not evenly spread out
- coresets offer a N/M reduction in runtime and memory
- but VI converges with fewer iterations on the coreset

References

1. Bachem, O., Lucic, M., and Krause, A. (2017). Scalable and distributed clustering via lightweight coresets. arXiv preprint arXiv:1702.08248.
2. Blei, D. M., Jordan, M. I., et al. (2006). Variational inference for dirichlet process mixtures. Bayesian analysis, 1(1):121–144.
3. Feldman, D., Schmidt, M., and Sohler, C. (2013). Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1434–1453. SIAM.
4. Huggins, J., Campbell, T., and Broderick, T. (2016). Coresets for scalable bayesian logistic regression. In Advances In Neural Information Processing Systems, pages 4080–4088
5. Lucic, M., Faulkner, M., Krause, A., and Feldman, D. (2017). Training Mixture Models at Scale via Coresets. ArXiv e-prints.
6. Lucic, M., Bachem, O., and Krause, A. (2016). Strong coresets for hard and soft bregman clustering with applications to exponential family mixtures. In International Conference on Artificial Intelligence and Statistics.