Collapsed Variational Dirichlet Process Mixture Models*

Kenichi Kurihara
Dept. of Computer Science
Tokyo Institute of Technology, Japan
kurihara@mi.cs.titech.ac.jp

Max Welling
Dept. of Computer Science
UC Irvine, USA
welling@ics.uci.edu

Yee Whye Teh
Dept. of Computer Science
National University of Singapore
tehyw@comp.nus.edu.sg

Discussion led by
Youssef Barhomi

Motivation

- Gibbs sampling is not efficient
- Sampling requires careful monitoring of the convergence of the Markov chain

Variational Bayesian methods

Motivation

Variational methods:

- A good approximation of the DP
- Deterministic
- Handles “modern” datasets faster than Gibbs Sampling

TSB and FSD

TSB: Truncated Stick Breaking process with standard variational bayesian model

FSD: Finite Symmetric Dirichlet representation with standard variational bayesian model

1 - TSB

TSB: Truncated Stick Breaking process with standard variational bayesian model

$$v_i \sim \mathcal{B}(v_i; 1, \alpha) \qquad i = 1, ..., T-1 \qquad (1)$$

$$v_T = 1 \qquad (2)$$

$$\pi_i = v_i \prod_{j<i} (1 - v_j) \qquad i = 1, ..., T \qquad (3)$$

$$\pi_i = 0 \qquad i > T \qquad (4)$$

$$P(X, \mathbf{z}, \mathbf{v}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n | \eta_{z_n}) \, p(z_n | \boldsymbol{\pi}(\mathbf{v})) \right] \left[\prod_{i=1}^{T} p(\eta_i) \mathcal{B}(v_i; 1, \alpha) \right] \qquad (5)$$

X are data points, ***z*** are the assignments, ***v*** are the stick breaking weighs, and $\boldsymbol{\eta}$ is cluster parameters

II - FSD

$$P(X, \mathbf{z}, \mathbf{v}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\, p(z_n|\boldsymbol{\pi}(\mathbf{v}))\right] \left[\prod_{i=1}^{T} p(\eta_i)\mathcal{B}(v_i; 1, \alpha)\right] \quad (5)$$

+

$\boldsymbol{\pi} \sim \mathcal{D}(\boldsymbol{\pi}; \frac{\alpha}{K}, ..., \frac{\alpha}{K})$ mixture weights following a symmetric dirichelet

+

assume a large number of clusters K

[Ishwaran and Zarepour, 2002]

$$P(X, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\, p(z_n|\boldsymbol{\pi})\right] \left[\prod_{i=1}^{K} p(\eta_i)\right] \mathcal{D}(\boldsymbol{\pi}; \tfrac{\alpha}{K}, ..., \tfrac{\alpha}{K}) \quad (7)$$

Marginalizing the mixture of weights

$$P(X, \mathbf{z}, \mathbf{v}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\, p(z_n|\boldsymbol{\pi}(\mathbf{v}))\right] \left[\prod_{i=1}^{T} p(\eta_i)\mathcal{B}(v_i; 1, \alpha)\right] \quad (5)$$

TSB

$$P(X, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\, p(z_n|\boldsymbol{\pi})\right] \left[\prod_{i=1}^{K} p(\eta_i)\right] \mathcal{D}(\boldsymbol{\pi}; \tfrac{\alpha}{K}, ..., \tfrac{\alpha}{K}) \quad (7)$$

FSD

$$P(X, \mathbf{z}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\right] p(\mathbf{z}) \left[\prod_{i=1}^{\infty} p(\eta_i)\right] \quad (9)$$

Collapsed model

Marginalizing the mixture of weights

$$p_{\text{TSB}}(\mathbf{z}) = \prod_{i<T} \frac{\Gamma(1+N_i)\Gamma(\alpha+N_{>i})}{\Gamma(1+\alpha+N_{\geq i})} \qquad (10)$$

$$p_{\text{FSD}}(\mathbf{z}) = \frac{\Gamma(\alpha)\prod_{k=1}^{K}\Gamma(N_k+\frac{\alpha}{K})}{\Gamma(N+\alpha)\Gamma(\frac{\alpha}{K})^K} \qquad (12)$$

with

$$N_i = \sum_{n=1}^{N} \mathbb{I}(z_n = i) \qquad N_{>i} = \sum_{n=1}^{N} \mathbb{I}(z_n > i) \qquad (11)$$

and $N_{\geq i} = N_i + N_{>i}$. For FSD we find instead,

Lower bound formulation

log marginal likelihood

$$\mathcal{L}(X) \geq \mathbf{B}(X) = \sum_{\mathbf{z}} \int_{\mathrm{d}\boldsymbol{\theta}} Q(\mathbf{z})Q(\boldsymbol{\theta}) \log \frac{P(X, \mathbf{z}, \boldsymbol{\theta})}{Q(\mathbf{z})Q(\boldsymbol{\theta})} \quad (13)$$

lower bound

where $\boldsymbol{\theta}$ is either $\{\boldsymbol{\eta}, \mathbf{v}\}$, $\{\boldsymbol{\eta}, \boldsymbol{\pi}\}$ or $\{\boldsymbol{\eta}\}$

$$Q_{\mathrm{TSB}}(\mathbf{z}, \boldsymbol{\eta}, \mathbf{v}) = \left[\prod_{n}^{N} q(z_n)\right] \left[\prod_{i=1}^{T} q(\eta_i) q(v_i)\right] \quad (14)$$

$$Q_{\mathrm{FSD}}(\mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\pi}) = \left[\prod_{n}^{N} q(z_n)\right] \left[\prod_{k=1}^{K} q(\eta_k)\right] q(\boldsymbol{\pi}) \quad (15)$$

Lower bound formulation

$$\mathbf{B}(X) = \sum_{n=1}^{N}\sum_{z_n}\int_{d\eta_{z_n}} q(z_n)q(\eta_{z_n})\log p(\mathbf{x}_n|\eta_{z_n}) + \sum_{i}\int_{d\eta_i} q(\eta_i)\log\frac{p(\eta_i)}{q(\eta_i)} - \sum_{n=1}^{N}\sum_{z_n} q(z_n)\log q(z_n) + \text{Extra Term}$$

$$\text{Term}_{\text{TSB}} = \sum_{n=1}^{N}\sum_{z_n=1}^{T} q(z_n)\int_{d\mathbf{v}}\left[\prod_{i=1}^{z_n} q(v_i)\right]\log p(z_n|\mathbf{v}) + \sum_{i=1}^{T}\int_{dv_i} q(v_i)\log\frac{p(v_i)}{q(v_i)} \quad (17)$$

$$\text{Term}_{\text{FSD}} = \sum_{n}\sum_{z_n=1}^{K}\int_{d\boldsymbol{\pi}} q(z_n)q(\boldsymbol{\pi})\log p(z_n|\boldsymbol{\pi}) + \int_{d\boldsymbol{\pi}} q(\boldsymbol{\pi})\log\frac{p(\boldsymbol{\pi})}{q(\boldsymbol{\pi})} \quad (18)$$

$$\text{Term}_{\text{CTSB/CFSD}} = \sum_{\mathbf{z}}\left[\prod_{n=1}^{N} q(z_n)\right]\log p(\mathbf{z}) \quad (19)$$

Update equations

$$q(\eta_i) \propto p(\eta_i) \exp\left(\sum_n q(z_n = i) \log p(\mathbf{x}_n|\eta_i)\right)$$

$$q(z_n) \propto \exp\left(\sum_{\mathbf{z}_{\neg n}} \prod_{m \neq n} q(z_m) \log p(z_n|\mathbf{z}_{\neg n})\right) \times \exp\left(\int_{\mathrm{d}\eta_{z_n}} q(\eta_{z_n}) \log p(\mathbf{x}_n|\eta_{z_n})\right)$$

for TSB formulation: $$p(z_n = i|\mathbf{z}_{\neg n}) = \frac{1 + N_i^{\neg n}}{1 + \alpha + N_{\geq i}^{\neg n}} \prod_{j<k} \frac{\alpha + N_{>i}^{\neg n}}{1 + \alpha + N_{\geq i}^{\neg n}}$$

for FSD formulation: $$p(z_n = k|\mathbf{z}_{\neg n}) = \frac{N_k^{\neg n} + \frac{\alpha}{K}}{N^{\neg n} + \alpha}$$

Optimal labels re-ordering

$$v_i \sim \mathcal{B}(v_i; 1, \alpha) \qquad i = 1, ..., T-1 \qquad (1)$$

$$v_T = 1 \qquad (2)$$

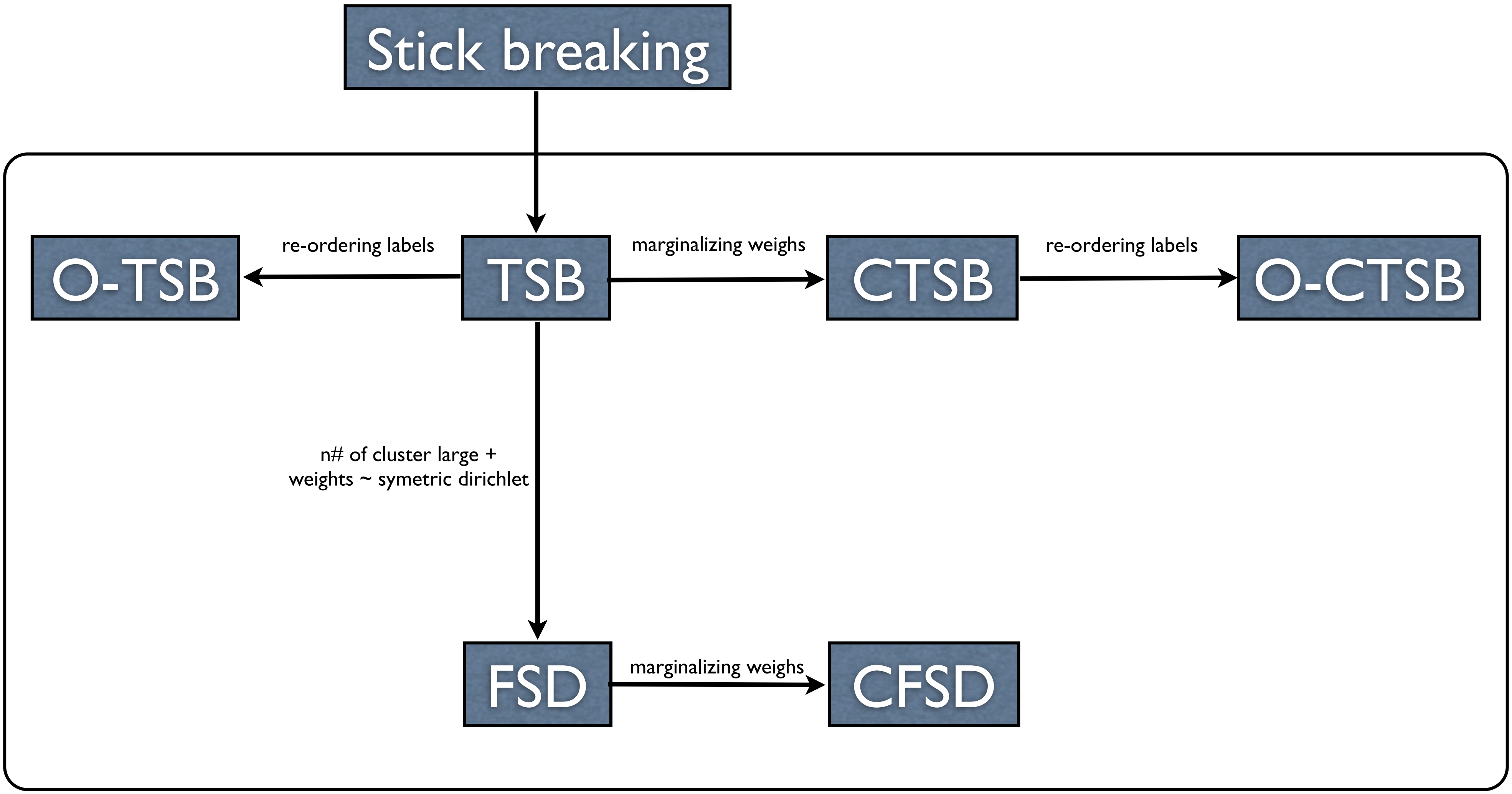
$$\pi_i = v_i \prod_{j<i} (1 - v_j) \qquad i = 1, ..., T \qquad (3)$$

$$\pi_i = 0 \qquad i > T \qquad (4)$$

$$P(X, \mathbf{z}, \mathbf{v}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n | \eta_{z_n}) p(z_n | \boldsymbol{\pi}(\mathbf{v})) \right] \left[\prod_{i=1}^{T} p(\eta_i) \mathcal{B}(v_i; 1, \alpha) \right] \qquad (5)$$

- Permutation of cluster labels change the probability, therefore, an optimal reordering of the labels will maximize that probability

Big picture


Stick breaking
O-TSB
re-ordering labels
TSB
marginalizing weighs
CTSB
re-ordering labels
O-CTSB
n# of cluster large +
weights ~ symetric dirichlet
FSD
marginalizing weighs
CFSD

Experiments

Exp1:

- Synthetic data from a mixture of 10 Gaussians in 16 dimensions with a separation coefficient c = 2
- 30 independently sampled training/testing data, 1000 test datapoints

Exp II:

- MNIST dataset 28*28 images reduced to 50 dimensions with a PCA.
- 30 splits of the data, 5000 training and 10,000 testing.

Exp I

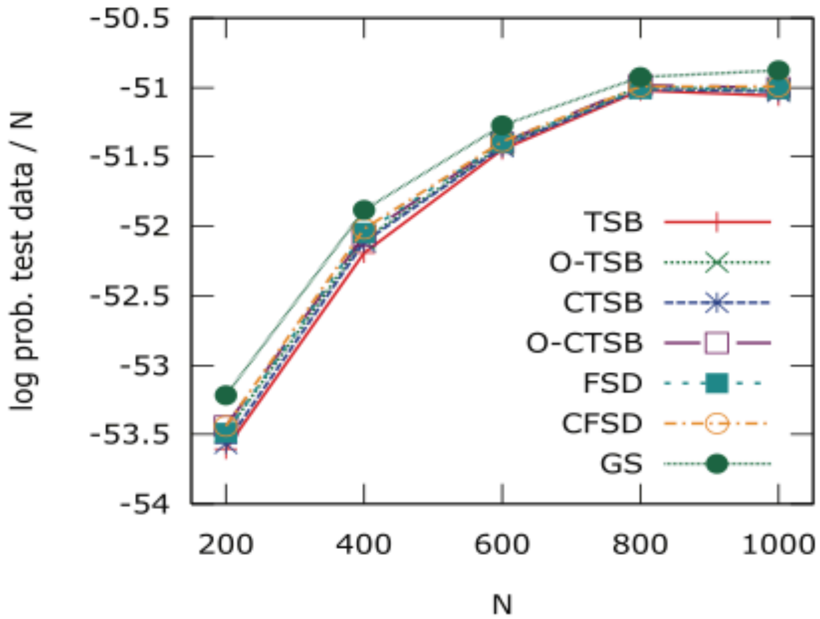


Figure 2: Average log probability per data-point for test data as a function of N.

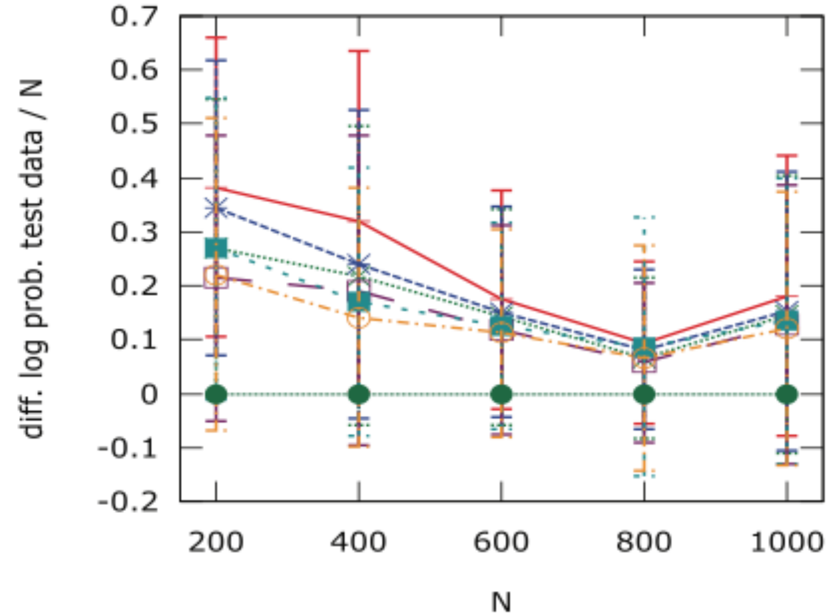


Figure 3: Relative average log probability per data-point for test data as a function of N.

Exp I

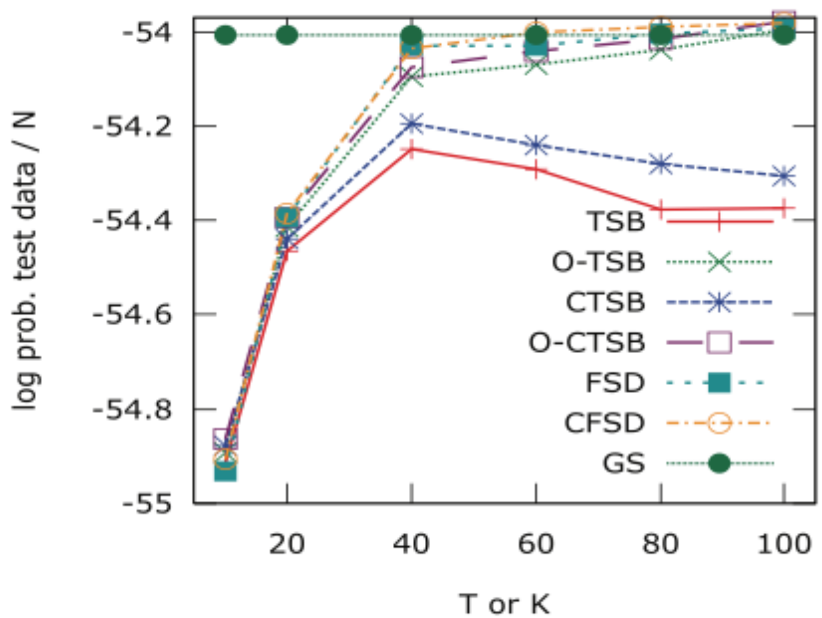


Figure 4: Average log probability per data-point for test data as a function of T (for TSB methods) or K (for FSD methods).

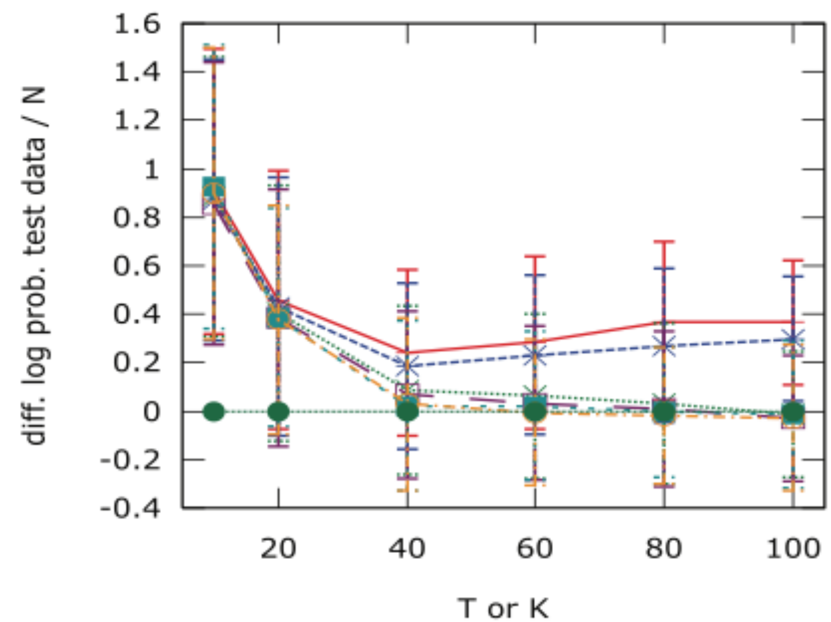


Figure 5: Relative average log probability per data-point for test data as a function of T (for TSB methods) or K (for FSD methods).

Exp II

Conclusion

- There is little difference between TSB and FSD.
- Label re-ordering is important for the stick breaking representation (especially when we have no clue about how many clusters we may have).
- Variational bayesian algorithms are much more efficient computationally than Gibbs sampling, with almost no loss in accuracy.