Tree-structured clustering method

Yinqiao Yan

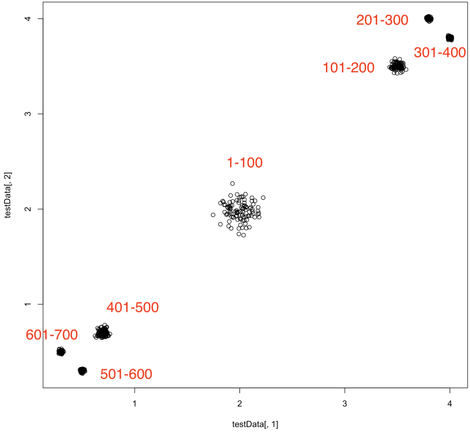
Institute of Statistics and Big Data
Renmin University of China

Introduction

- Cluster data in a tree-structured subclones.
- Adams et al. (2010) proposed a novel nonparametric Bayesian prior named tree-structured stick-breaking prior (TSSB).
- We propose a truncation version of TSSB, referred to as TSSB-DW (**TSSB** with finite **D**epth and **W**idth).

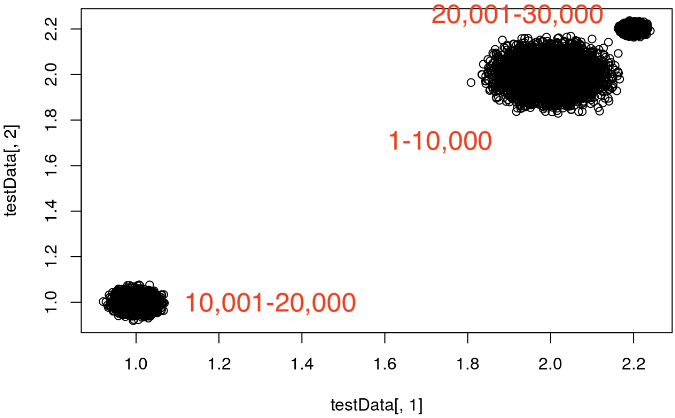
Data Preprocessing

- Simulation study 1: Seven classes normal data.
  Dimension: 700×2



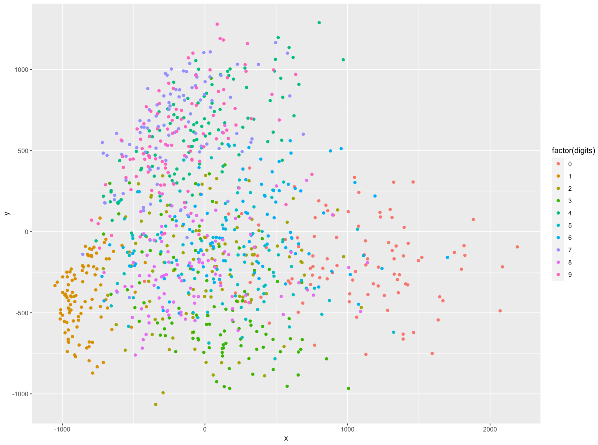
Data Preprocessing

- Simulation study 2: Three classes normal data.
  Dimension: 30,000$\times$500



Data Preprocessing

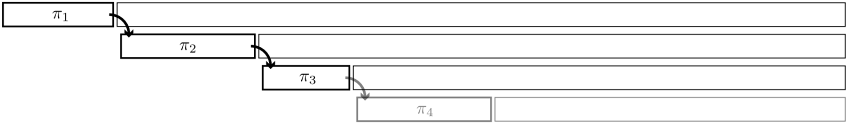
- Real data: MNIST
  Dimension: 60,000×154 (whole) 1,000×154 (mini)



miniData: 100 samples for each digit.

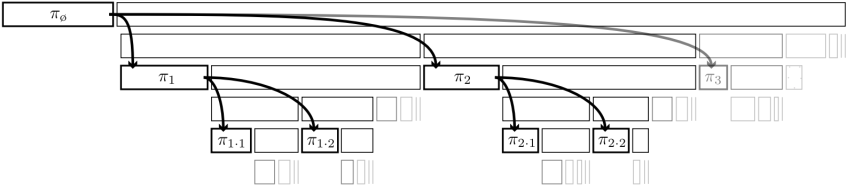
Preliminaries

Dirichlet process (DP)



(a) Dirichlet process stick breaking

Tree-structured stick breaking process (TSSB)



(b) Tree-structured stick breaking

Our model

Motivated by the work of Ishwaran and James (2001), our model:

- based on a truncation version of TSSB.
- Use factored normal likelihood to avoid the high dimensionality problem.
- Parameters:
  - node parameters θ_ε^ℓ and $\sigma_\varepsilon^{2\ell}$, for $\ell = 1, ..., L$.
  - data assignment c_i, for $i = 1, ..., n$.
  - stick length ν-sticks and ψ-sticks, which derive the random weights π_ε.
  - hyper-parameter drift λ^ℓ.
  - stick-breaking hyper-parameters $\alpha_0, \lambda, \gamma$.
  - **Fixed** parameters: $\eta_\mathcal{N}$ and η_Θ
- Search for new tree structure (Yuan et al.2015).
  The authors added another swap-nodes step to propose a new tree structure.

Our model

The complete model is

$$(X_i \mid \theta, \Sigma, c_i = \varepsilon) \overset{\text{ind}}{\sim} \prod_{\ell=1}^{L} N\left(X_i^{\ell} \mid \theta_{\varepsilon}^{\ell}, \eta_{\mathcal{N}}^{|\varepsilon|}\sigma_{\varepsilon}^{2\ell}\right), \quad i = 1, \ldots, n$$

$$c_i \mid \pi \overset{\text{iid}}{\sim} \sum_{\varepsilon} \pi_{\varepsilon}\delta_{\varepsilon}$$

$$\pi \sim \text{TSSB-DW}(\alpha_0, \rho, \gamma)$$

$$\theta_{\varnothing}^{\ell} \sim N(\theta_{\varnothing}^{\ell} \mid \mu_0^{\ell}, \lambda^{\ell}), \quad \ell = 1, \ldots, L$$

$$\theta_{\varepsilon}^{\ell} \mid \theta_{pa(\varepsilon)}^{\ell}, \eta_{\Theta} \overset{\text{iid}}{\sim} N(\theta_{\varepsilon}^{\ell} \mid \theta_{pa(\varepsilon)}, \eta_{\Theta}^{|\varepsilon|}\lambda^{\ell})$$

$$\sigma_{\varepsilon}^{2\ell} \overset{\text{iid}}{\sim} \text{InvGamma}(v_{sig}, s_{sig})$$

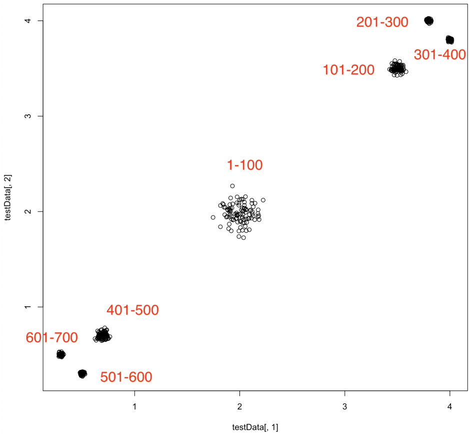
$$\lambda^{\ell} \overset{\text{iid}}{\sim} \text{InvGamma}(v_{dft}, s_{dft})$$

Posterior inference

- Resample data assignments c_i for $i = 1, ..., n$. (multinoulli)
- Resample node parameters θ_ε^ℓ and $\sigma_\varepsilon^{2\ell}$ for $\ell = 1, ..., L$. (normal-invGamma)
- Resample stick length ν-sticks and ψ-sticks, in order to update the random weights π_ε. (like in DP)
- Resample hyper-parameter drift λ^ℓ for $\ell = 1, ..., L$. (normal-invGamma)
- Resample stick-breaking hyper-parameters $\alpha_0, \lambda, \gamma$ by slice sampler.

Simulation Study

Study 1. Seven classes normal data (low dimension).

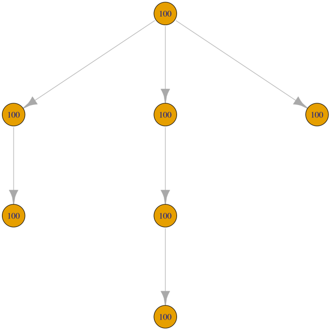


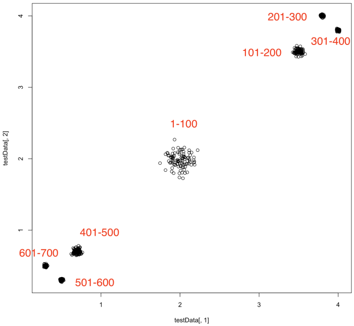
Simulation Study

Update Settings:

- $\eta_{\mathcal{N}} = 1$ and $\eta_{\Theta} = 0.5$
- set.seed(9)
- Update order: (1) NodeParams (2) Assignments (3) SearchTree
- Iter = 200, burnIn = 0
- priorSigmaScale = mean(diag(cov(t(testData))))
- D = 3, W = 3
- "OnlyTree"
- $\lambda^{\ell} \overset{\text{iid}}{\sim} \text{InvGamma}(v_{dft}, s_{dft})$

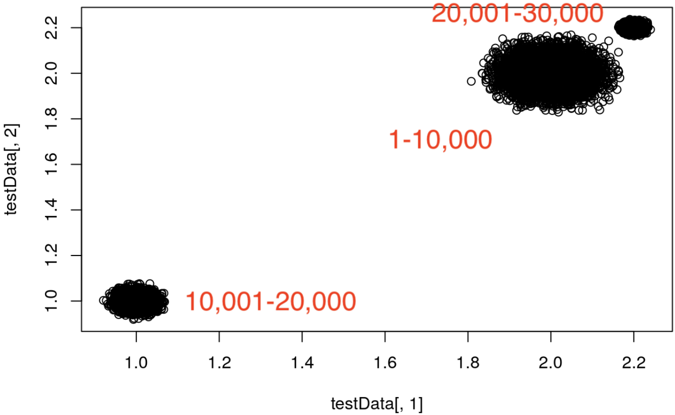
Results





Node 0 contains dataIds 401-500, Node 1 contains dataIds 1-100
Node 11 contains dataIds 501-600, Node 2 contains dataIds 301-400
Node 23 contains dataIds 201-300, Node 233 contains dataIds 101-200
Node 3 contains dataIds 601-700.

Study 2. Three classes normal data (high dimension).

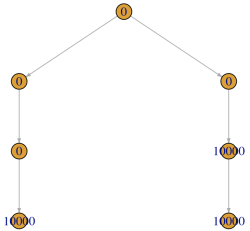


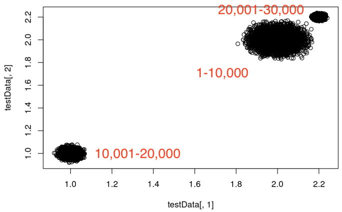
Simulation Study

Update Settings:

- $\eta_{\mathcal{N}} = 1$ and $\eta_{\Theta} = 1$
- set.seed(12)
- Update order: (1) NodeParams (2) Assignments *(WITHOUT searchTree step)*
- Iter = 50, burnIn = 0
- priorSigmaScale = 1e-4
- D = 3, W = 3
- "OnlyTree"
- $\lambda^{\ell} \overset{\text{iid}}{\sim} \text{Unif}(0.01, 1)$

Results

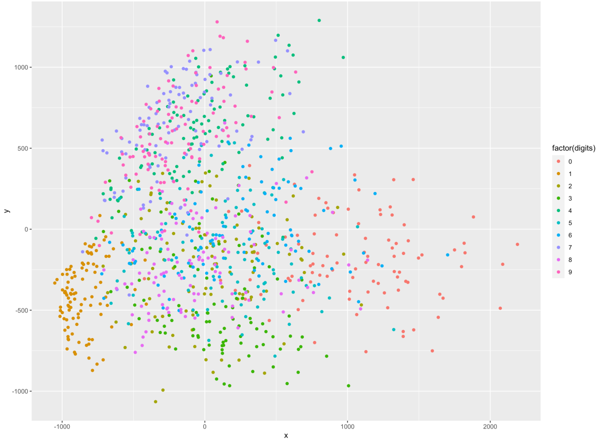




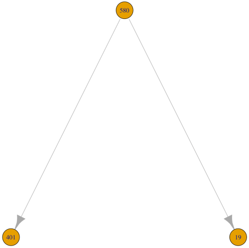
Node 232 contains dataIds 1-10000
Node 33 contains dataIds 10001-20000
Node 331 contains dataIds 20001-30000

MNIST

MNIST: A famous handwritten digits dataset. It includes a training set with 60,000 images and a test set with 10,000 images. Each image has 28x28 pixels. Each pixel in the image matrix is in [0,255].



Results under same settings.



Root: 0 (53), 1 (0), 2 (92), 3 (62), 4 (64), 5 (77), 6 (63), 7 (45), 8 (84), 9 (40)
Child One: 0 (46), 1 (100), 2 (6), 3 (36), 4 (36), 5 (17), 6 (36), 7 (55), 8 (12), 9 (57
Child Two: 0 (1), 1 (0), 2 (2), 3 (2), 4 (0), 5 (6), 6 (1), 7 (0), 8 (4), 9 (3)

Conclusion

- TSSB-DW model works well when the data are separable.
- MCMC approach obtains samples from the joint posterior distribution, so it is possible to derive a tree structure different from the original setting.
- MNIST may not be a good example for tree-structured clustering. Most digital samples overlap and the variances of each class have no obvious difference.

Future Work

- Try more simulation studies.
- Try to solve the inseparable data clustering problem.
- Find a better way to choose the tree structure from all the samples and interpret the result.