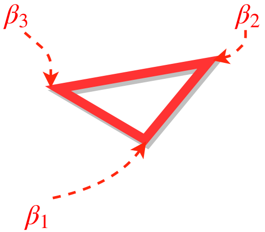
Dirichlet Simplex Nest and Geometric Inference

Mikhail Yurochkin[1,2,*]
Aritra Guha[3,*], Yuekai Sun[3], XuanLong Nguyen[3]

IBM Research[1], MIT-IBM Watson AI Lab[2], University of Michigan[3]
* authors contributed equally

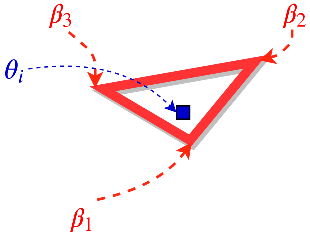
ICML 2019, June 11th

Dirichlet Simplex Nest (DSN)



The parent nest: $\mathscr{B} = \mathrm{Conv}(\beta_1, \ldots, \beta_K)$

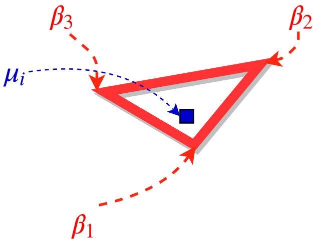
Dirichlet Simplex Nest (DSN)



The parent nest: $\mathscr{B} = \mathrm{Conv}(\beta_1, \ldots, \beta_K)$

Admixture weights: $\theta_i \sim \mathrm{Dir}_K(\alpha) \in \Delta^{K-1}$

Dirichlet Simplex Nest (DSN)

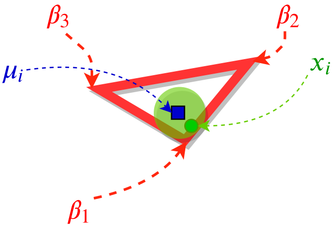


The parent nest: $\mathscr{B} = \mathrm{Conv}(\beta_1, \ldots, \beta_K)$

Admixture weights: $\theta_i \sim \mathrm{Dir}_K(\alpha) \in \Delta^{K-1}$

Offspring i is born: $\mu_i = \sum_{k=1}^{K} \theta_{ik}\beta_k \in \mathbb{R}^D$

Dirichlet Simplex Nest (DSN)



The parent nest: $\mathscr{B} = \mathrm{Conv}(\beta_1, \ldots, \beta_K)$

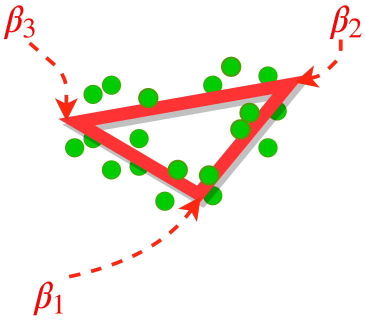
Admixture weights: $\theta_i \sim \mathrm{Dir}_K(\alpha) \in \Delta^{K-1}$

Offspring i is born: $\mu_i = \sum_{k=1}^{K} \theta_{ik}\beta_k \in \mathbb{R}^D$

Offspring i leaves the nest:

$x_i|\mu_i \sim F(\cdot|\mu_i) \in \mathbb{R}^D$ s.t. $\mathbb{E}[x_i|\mu_i] = \mu_i$

Dirichlet Simplex Nest (DSN)



Dirichlet Simplex Nest examples:

- LDA: $x_i|\mu_i \sim \text{Multinomial}(\mu_i)$ (Blei et al. (2003))
- Gaussian: $x_i|\mu_i \sim \mathcal{N}(\mu_i, \Sigma)$ (Schmidt et al. (2009))
- Poisson-Gamma: $x_{i,d}|\mu_{i,d} \overset{\text{iid}}{\sim} \text{Pois}(\mu_{i,d})$ (Cemgil (2009))

Centroidal Voronoi tessellation (CVT)

Du et al (1999)

- open set Ω
- distance $d : \Omega \times \Omega \to \mathbb{R}_+$
- density $\rho : \Omega \to \mathbb{R}_+$
- For $\{z_1, \ldots, z_K\} \subset \bar{\Omega}$, the Voronoi region corresponding to z_k is

$$V_k = \{x \in \Omega : d(x, z_k) < d(x, z_l) \text{ for any } l \neq k\}.$$

  - ⋆ $\{V_1, \ldots, V_K\}$ is a Voronoi tessallation of Ω
  - ⋆ z_k's are the generators of this Voronoi tessellation
- $\{V_1, \ldots, V_K\}$ is a CVT iff the z_k's satisfy

$$z_k = \frac{\int_{V_k} x\rho(x)\mathrm{d}x}{\int_{V_k} \rho(x)\mathrm{d}x}.$$

Variational characterization of CVT's

- define the cost function

$$J(z_1,\ldots,z_K) = \sum_{k=1}^K \int_{V_k} d(x,z_k)^2 \rho(x)\mathrm{d}x,$$

  where the V_k's are Voronoi regions corresponding to the z_k's,
- If $(z_1,\ldots,z_K) \in \operatorname{argmin} J$, then the corresponding VT is a CVT.
- estimate $(z_1,\ldots,z_K)$ by minimizing empirical counterpart of J:

$$\frac{1}{n}\sum_{k=1}^K \sum_{i=1}^n d(x_i,z_k)^2 1\{x_i \in V_k\}, \quad x_i \overset{\text{iid}}{\sim} \rho$$

- If d is Euclidean distance, this is the K-means cost function.

CVT of equilateral simplices

- Ω is a scaled, rotated, translated copy of Δ^{K-1}
- equipped with distance $d(x_1, x_2) = \|x_1 - x_2\|_2$
- ρ is $\mathrm{Dir}_K(\alpha 1_K)$ density "transported" to Ω
- **fact:** The CVT generators z_k's are on the line segments between centroid μ of the simplex and its extreme points.
- **idea:** estimate extreme points by estimating μ and z_k's and searching along ray from $\hat{\mu}$ to $\hat{z}_k$ (Geometric Dirichlet Means (GDM), Y. and Nguyen (2016))

Toy example: Gaussian DSN with $K = D = 3$

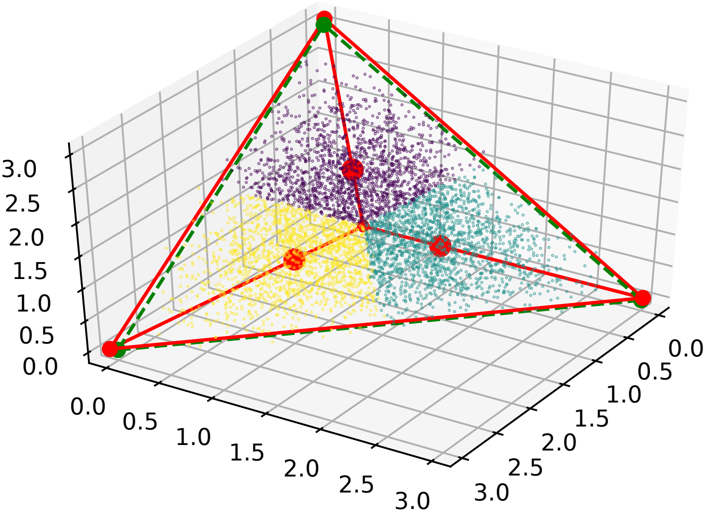


Figure: GDM with equilateral $\mathscr{B}$, $\alpha = 2.5$

Toy example: Gaussian DSN with $K = D = 3$

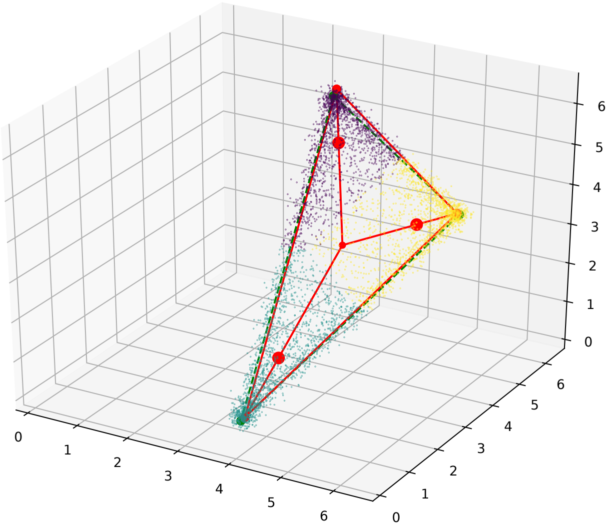


Figure: GDM with general $\mathscr{B}$, $\alpha = 0.3$

Toy example: Gaussian DSN with $K = D = 3$

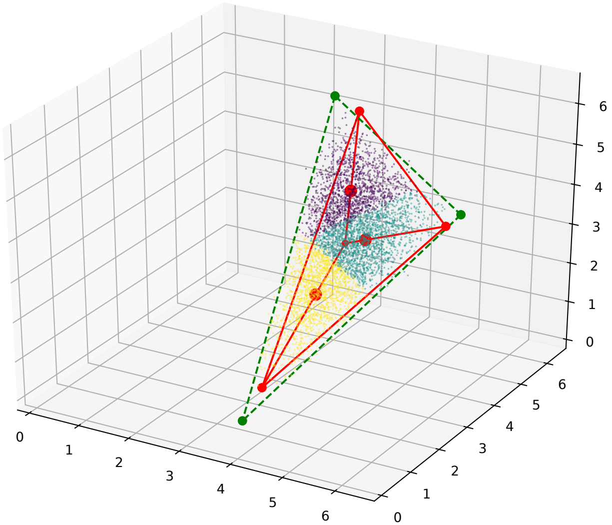


Figure: GDM with general $\mathscr{B}$, $\alpha = 2.5$

Toy example: Gaussian DSN with $K = D = 3$

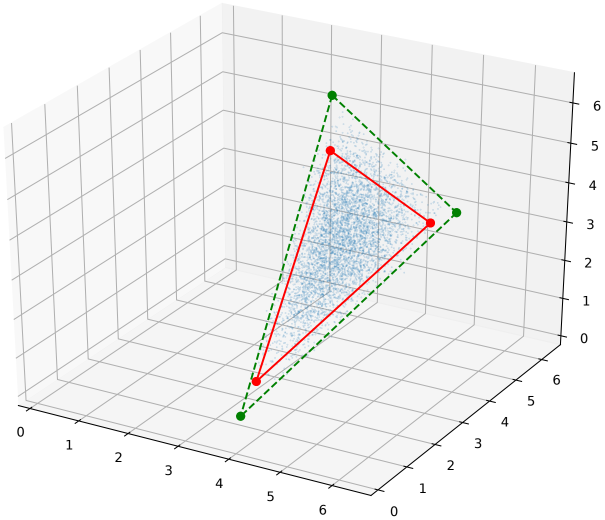


Figure: Xray (Kumar et al., 2013) - separability is not satisfied

Toy example: Gaussian DSN with $K = D = 3$

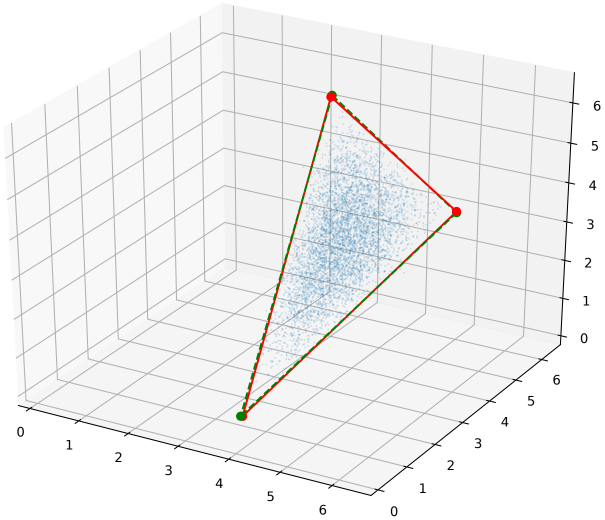


Figure: Stan-HMC (Carpenter et al., 2017) - too slow (10min)

Estimating extreme points of non-equilateral simplices

- **idea:** transform non-equilateral simplex to equilateral simplex
- map from equilateral to non-equilateral simplex is the linear map B^T
- CVT centroids in $\|\cdot\|_{(BB^T)^\dagger}$-norm are images of CVT generators of Δ^{K-1} under B^T (for any concentration parameter $\alpha > 0$)
- K-means clustering the x_i's in the $\|\cdot\|_{(BB^T)^\dagger}$-norm is equivalent to K-means clustering the (left) singular vectors of $\bar{X} = X - 1_n\hat{\mu}^T$

Voronoi Latent Admixture (VLAD)

inputs: $x_1, \ldots, x_n \in \mathbb{R}^D$, extension size $\gamma > 0$

outputs: $\hat{\beta}_1, \ldots, \hat{\beta}_K \in \mathbb{R}^D$

1a. $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$
1b. $\bar{x}_i \leftarrow x_i - \hat{\mu}$

2. compute top K singular factors of centered data matrix: $\bar{X} \approx U \Lambda V^T$

3. $(\hat{w}_1, \ldots, \hat{w}_K) \leftarrow K$-means$(u_1, \ldots, u_n)$, $u_i \in \mathbb{R}^K$ is the i-th row of U

4. $\hat{z}_k \leftarrow \hat{\mu} + V \Lambda \hat{w}_k$

5. $\hat{\beta}_k \leftarrow \hat{\mu} + \gamma(\hat{z}_k - \hat{\mu})$

Toy example: Gaussian DSN with $K = D = 3$

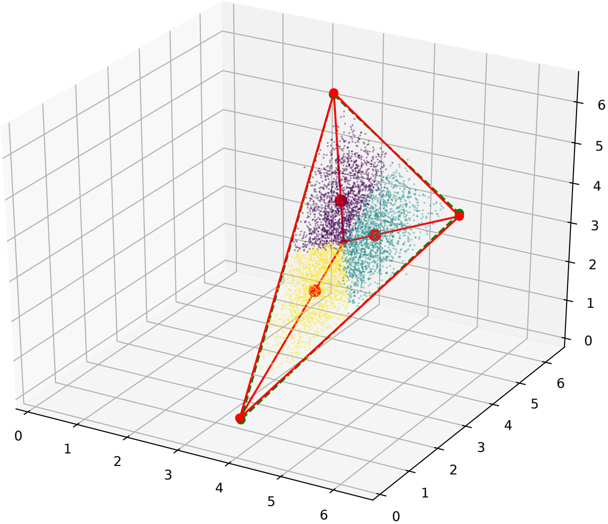


Figure: Voronoi Latent Admixture (VLAD), under 1s

Choosing extension size γ

- **fact:** exact $\gamma = \frac{\|\beta_k - \mu\|_2}{\|z_k - \mu\|_2}$ depends only on the concentration parameter α (and K). It does not depend on the geometry of the simplex
- choosing extension size γ is equivalent to estimating concentration parameter α
- tabulate $\gamma(\alpha)$ for all $\alpha > 0$ in a particular simplex (eg. Δ^{K-1}) by Monte Carlo
- estimate α by a method of moments approach:

$$\hat{\alpha} = \text{argmin}\{\|\frac{\hat{B}(\gamma(\alpha))(I_K - \frac{1}{K}1_K 1_K^T)\hat{B}(\gamma(\alpha))^T}{K(K\alpha+1)} - \hat{\Sigma}\|\} : \alpha > 0\}$$

  - ⋆ $\hat{B}(\gamma) = 1_K\hat{\mu} + \gamma(\alpha)(\hat{Z} - 1_K\hat{\mu}^T)$
  - ⋆ $\frac{I_K - \frac{1}{K}1_K 1_K^T}{K(K\alpha+1)}$ is the covariance matrix of a $\text{Dir}_K(\alpha 1_K)$ random vector
  - ⋆ $\hat{\Sigma}$ is the sample covariance matrix of the x_i's

Estimation error of VLAD

- **Pollard's condition:** the noiseless CVT cost function

$$\sum_{k=1}^{K} \int_{V_k} \text{Dirichlet}(\theta; \alpha 1_K) \|B^T\theta - z_k\|_2^2 d\theta$$

  is λ-strongly convex

- quantify the noise level by $\sigma^2 := \sup\{\|\text{Cov}[x_i \mid \theta_i]\|_2 : \theta_i \in \Delta^{K-1}\}$
- there is a permutation π of $[K]$ such that

$$\|(\hat{\beta}_{\pi_{(1)}}, \ldots, \hat{\beta}_{\pi_{(K)}}) - (\beta_1, \ldots, \beta_K)\|_2 \leq O(\frac{\sigma^{1/4}}{\sqrt{\lambda}}) + O_P(\frac{1}{\sqrt{n}})$$

Simulations: varying DSN geometry

- $\alpha = 2$, $D = 500/2000$, $K = 10$, $n = 10k$
- $\beta_k \sim \mathcal{N}(0, K)/\mathrm{Dir}_D(0.1)$ and scaled towards mean by $c_k \sim \mathrm{Unif}(c_{\min}, 1)$
- small $c_{\min}$ $\rightarrow$ more DSN skewness $\rightarrow$ harder problem

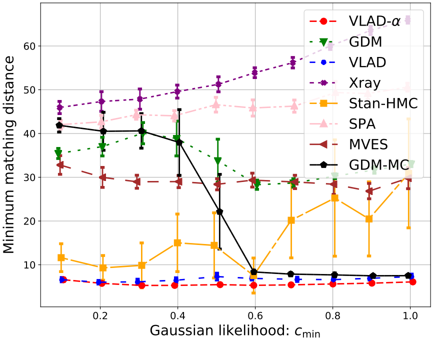


Figure: Gaussian data

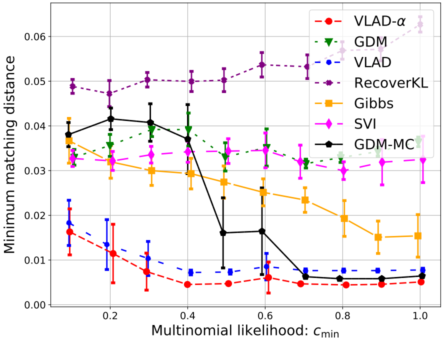


Figure: Multinomial data (LDA)

Simulations: varying concentration parameter α

- varying α, $D = 500/2000$, $K = 10$, $n = 10k$
- $\beta_k \sim \mathcal{N}(0, K)/\mathrm{Dir}_D(0.1)$ and scaled towards mean by $c_k \sim \mathrm{Unif}(0.5, 1)$
- large $\alpha \rightarrow$ non-separable data $\rightarrow$ harder problem

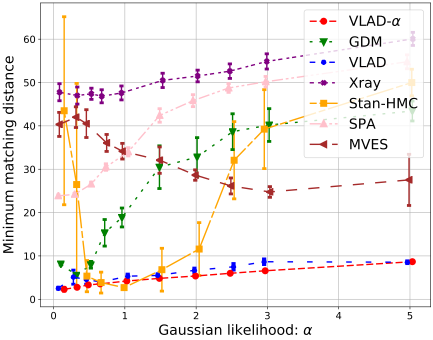


Figure: Gaussian data

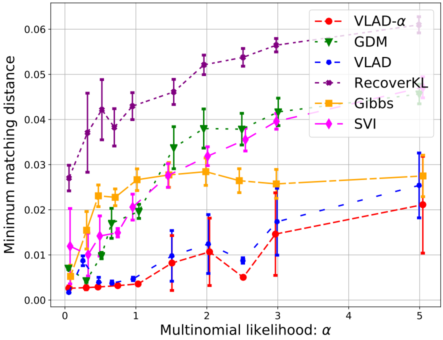


Figure: Multinomial data (LDA)

Factor analysis of stock data

- $D = 55$ stocks
- $n = 3k$ days
- 400 data points left out for validation
- VLAD factors
  - ⋆ growth component related to banks (e.g., Bank of America, Wells Fargo)
  - ⋆ Stocks of fuel companies (Valero Energy, Chevron) are inversely related to defense contractors (Boeing, Raytheon)

	Residual norm	Volume	Runtime
VLAD	0.300	**0.14**	**1 sec**
GDM	0.294	1499	**1 sec**
HMC	0.299	1.95	10 min
MVES	**0.287**	5.39×10^9	3 min
SPA	0.392	3.31×10^7	**1sec**

Topic discovery in New York Times corpus

- $D = 5320$ words in vocabulary
- $n \approx 100k$ articles
- $25k$ articles left out for perplexity evaluation
- VLAD topics
  - ⋆ ballot al-gore votes election recount florida voter court vote counties count county board hand bush george-bush official republican counted lawyer
  - ⋆ cup sugar minutes cream butter teaspoon tablespoon chocolate egg pan add bowl mixture flour milk oven baking water serving cake

	Perplexity	Coherence	Runtime
VLAD	1767	0.86	**6 min**
GDM	1777	0.88	30 min
Gibbs	**1520**	0.80	5.3 hrs
RecoverKL	2365	0.70	17 min
SVI	1669	0.81	40min

Open problem: asymmetric concentration parameter α

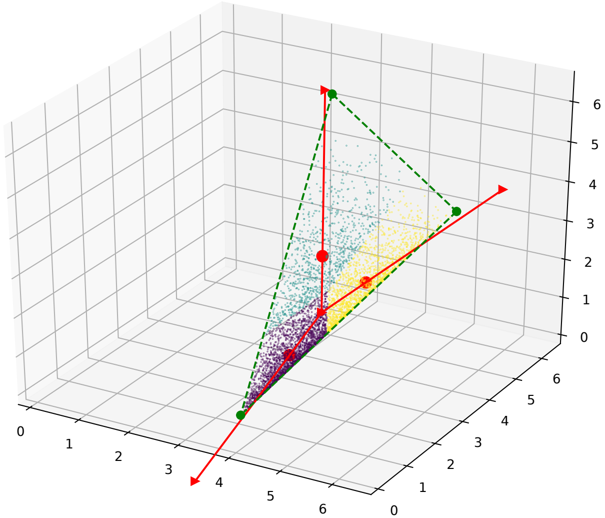


Figure: VLAD, $\alpha = (0.5, 1.5, 2.5)$

THANK YOU!

Please come to poster #230

Questions?