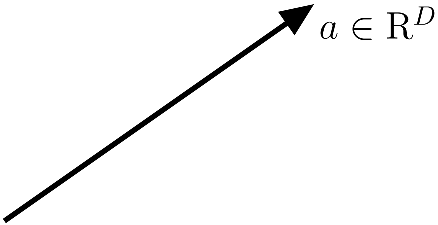
A Dirichlet Process Mixture Model for Spherical Data

Julian Straub
Jason Chang, Oren Freifeld, John W. Fisher III
Massachusetts Institute of Technology

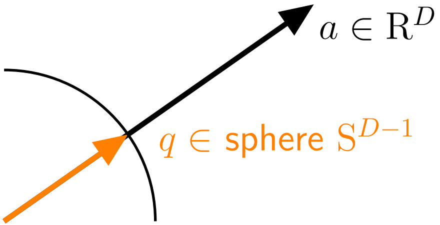
September 22, 2017

CSAIL

Data on the Sphere = Directional Data

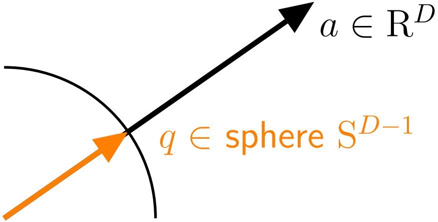


Data on the Sphere = Directional Data



$$a = q||a||_2$$

Data on the Sphere = Directional Data



$$a = q||a||_2$$

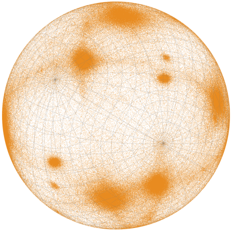
Interested in:

directional data = data on the sphere S^{D-1}
$\Rightarrow$ all information contained in direction q

Native Directional Data

Examples of directional data ($||a||_2 = 1$):

- surface normals [Furukawa 2009, Holz 2011, Straub 2014]

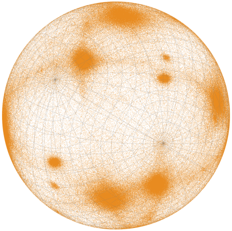


[Straub 2014]

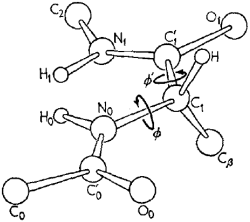
Native Directional Data

Examples of directional data ($||a||_2 = 1$):

- surface normals [Furukawa 2009, Holz 2011, Straub 2014]
- protein backbone configurations [Ramachandran 1965, Lennox 2009, Ting 2010]



[Straub 2014]

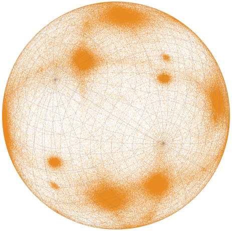


[Ramachandran 1965]

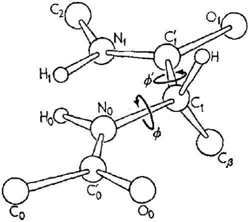
Native Directional Data

Examples of directional data ($||a||_2 = 1$):

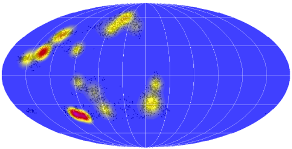
- surface normals [Furukawa 2009, Holz 2011, Straub 2014]
- protein backbone configurations [Ramachandran 1965, Lennox 2009, Ting 2010]
- beam directions in external radiation therapy [Bangert 2010]



[Straub 2014]



[Ramachandran 1965]

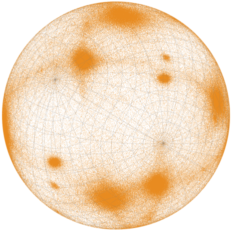


[Bangert 2010]

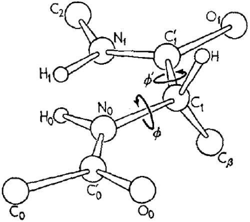
Native Directional Data

Examples of directional data ($||a||_2 = 1$):

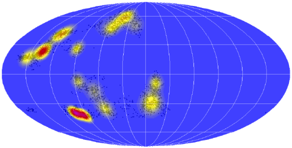
- surface normals [Furukawa 2009, Holz 2011, Straub 2014]
- protein backbone configurations [Ramachandran 1965, Lennox 2009, Ting 2010]
- beam directions in external radiation therapy [Bangert 2010]
- quaternions representing 3D rotations [Choe 2006]



[Straub 2014]



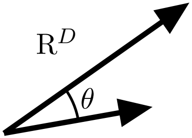
[Ramachandran 1965]



[Bangert 2010]

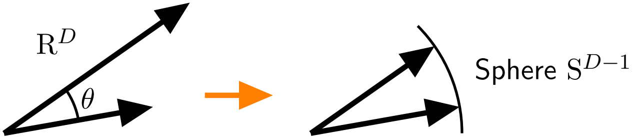
Euclidean Data Treated as Directional Data

have Euclidean data and use angular deviation for comparison
(cosine distance $\cos(\theta) = \frac{a^T}{||a||_2} \frac{b}{||b||_2}$)



Euclidean Data Treated as Directional Data

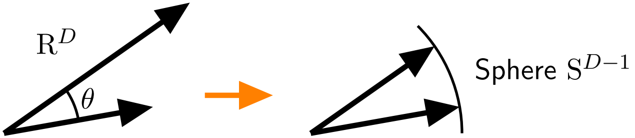
have Euclidean data and use angular deviation for comparison
(cosine distance $\cos(\theta) = \frac{a^T}{||a||_2} \frac{b}{||b||_2}$)



$\Rightarrow$ all information is in direction $\Rightarrow$ directional data

Euclidean Data Treated as Directional Data

have Euclidean data and use angular deviation for comparison
(cosine distance $\cos(\theta) = \frac{a^T}{||a||_2} \frac{b}{||b||_2}$)

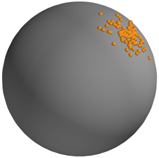


$\Rightarrow$ all information is in direction $\Rightarrow$ directional data

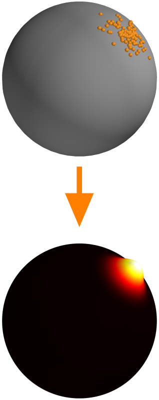
Representative examples:

- semantic word vectors (word2vec) [Mikolov 2013]
- word frequency counts [Dhillon 2001, Strehl 2000]
- gene expression data [Banerjee 2005]

Distributions on the Sphere

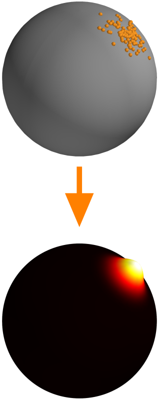


Distributions on the Sphere



von-Mises-Fisher (vMF)
Kent, Bingham

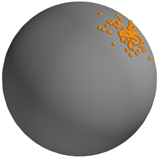
Distributions on the Sphere



+ no linearization
− difficult normalizers
− posterior sampling

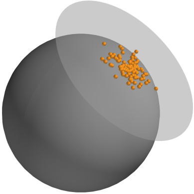
von-Mises-Fisher (vMF)
Kent, Bingham

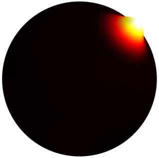
Distributions on the Sphere



Linearization using

Riemannian Geometry

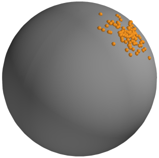




+ no linearization
− difficult normalizers
− posterior sampling

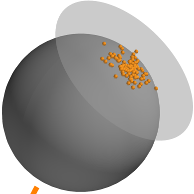
von-Mises-Fisher (vMF)
Kent, Bingham

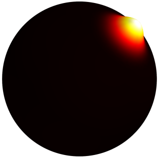
Distributions on the Sphere



Linearization using

Riemannian Geometry

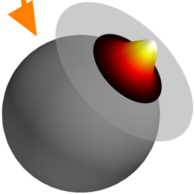




\+ no linearization

− difficult normalizers

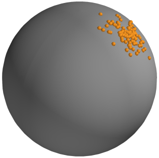
− posterior sampling



von-Mises-Fisher (vMF)
Kent, Bingham

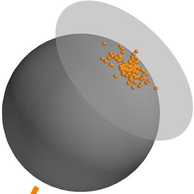
Tangent space Gaussian (TG)

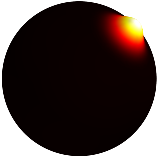
Distributions on the Sphere



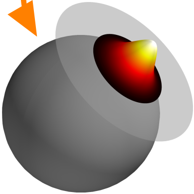
Linearization using

Riemannian Geometry





+ no linearization
– difficult normalizers
– posterior sampling



– linearization errors
+ established inference
+ generalization to other manifolds

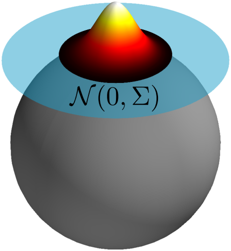
von-Mises-Fisher (vMF)
Kent, Bingham

Tangent space Gaussian (TG)

Tangent Space Gaussian via Riemannian Geometry

Tangent space Gaussian (TG)

(see Pennec 1999)



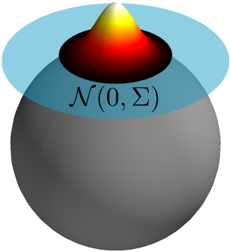
mean $\mu \in \mathrm{S}^{D-1}$

covariance Σ in tangent space

Tangent Space Gaussian via Riemannian Geometry

Tangent space Gaussian (TG)

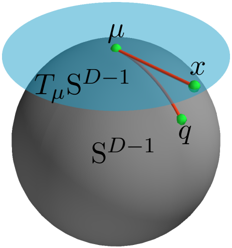
(see Pennec 1999)



mean $\mu \in \mathrm{S}^{D-1}$

covariance Σ in tangent space

Riemannian geometry



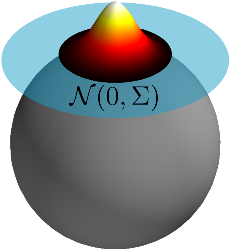
$x = \mathrm{Log}_\mu(q)$ logarithm map

$q = \mathrm{Exp}_\mu(x)$ exponential map

Tangent Space Gaussian via Riemannian Geometry

Tangent space Gaussian (TG)

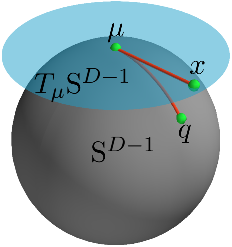
(see Pennec 1999)



mean $\mu \in \mathrm{S}^{D-1}$

covariance Σ in tangent space

Riemannian geometry

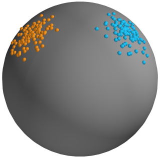


$x = \mathrm{Log}_\mu(q)$ logarithm map

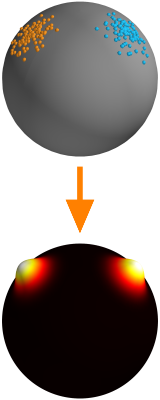
$q = \mathrm{Exp}_\mu(x)$ exponential map

$$q \sim \mathcal{N}\left(\mathrm{Log}_\mu(q); 0, \Sigma\right)$$

Mixture Models on the Sphere

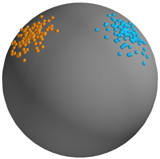


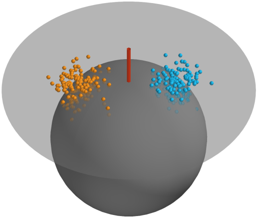
Mixture Models on the Sphere



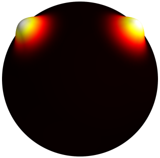
[Peel 2001, Bangert 2010]

Mixture Models on the Sphere



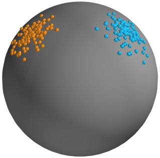


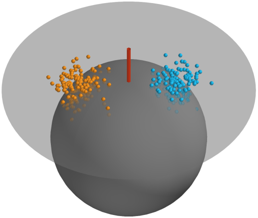
[Goddard 1998, Choe 2006]



[Peel 2001, Bangert 2010]

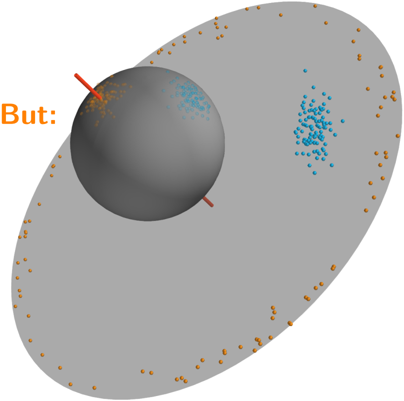
Mixture Models on the Sphere

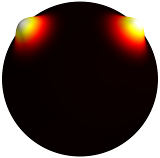




[Goddard 1998, Choe 2006]

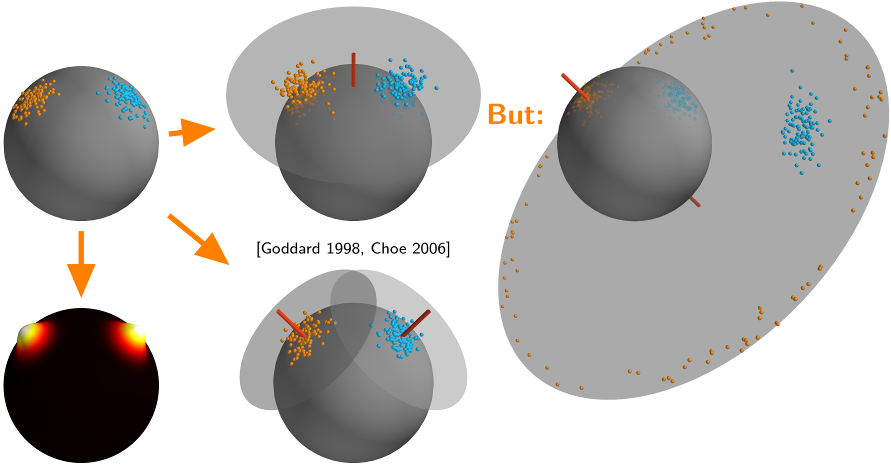
But:





[Peel 2001, Bangert 2010]

Mixture Models on the Sphere



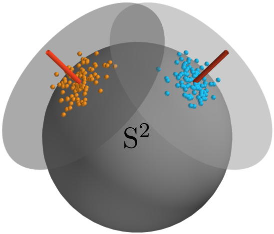
[Peel 2001, Bangert 2010]

[Feiten 2013, Simo-Serra 2014]

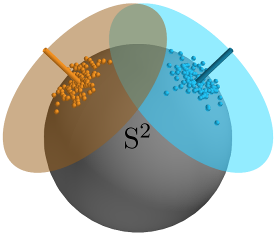
Dirichlet Process Tangent Gaussian Mixture Model

Goals:

- flexibel mixture model for directional data
- anisotropic component distributions
- efficient manifold-respecting inference



Dirichlet Process Tangent Gaussian Mixture Model



Goals:

- flexibel mixture model for directional data
- anisotropic component distributions
- efficient manifold-respecting inference

Approach and Contributions:

- Bayesian nonparametric Dirichlet process mixture model for directional data

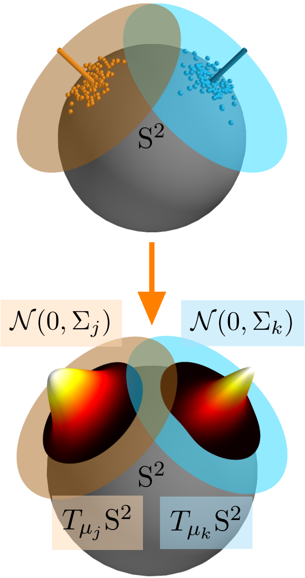
Dirichlet Process Tangent Gaussian Mixture Model

Goals:

- flexibel mixture model for directional data
- anisotropic component distributions
- efficient manifold-respecting inference

Approach and Contributions:

- Bayesian nonparametric Dirichlet process mixture model for directional data
- Gaussian distributions in cluster-dependent tangent spaces



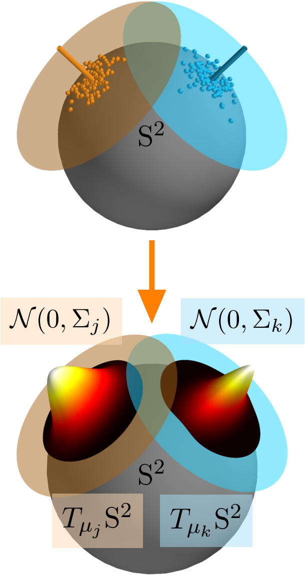
Dirichlet Process Tangent Gaussian Mixture Model

Goals:

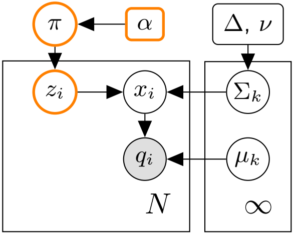
- flexibel mixture model for directional data
- anisotropic component distributions
- efficient manifold-respecting inference

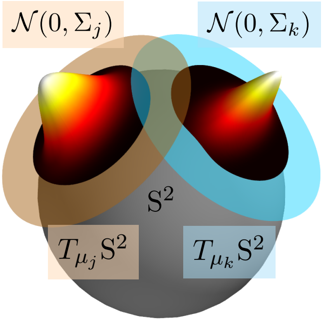
Approach and Contributions:

- Bayesian nonparametric Dirichlet process mixture model for directional data
- Gaussian distributions in cluster-dependent tangent spaces
- manifold-aware sub-cluster split/merge MCMC sampler



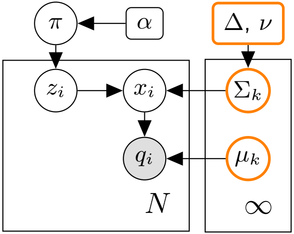
DP-TGMM Generative Model

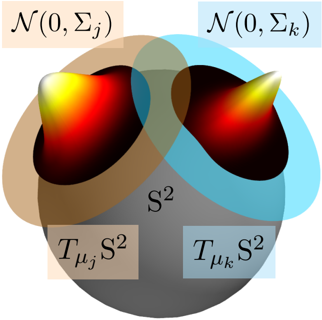




$$\pi \sim \mathsf{GEM}(1, \alpha) \qquad z_i \sim \mathsf{Cat}(\pi)$$

DP-TGMM Generative Model

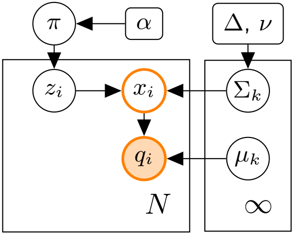


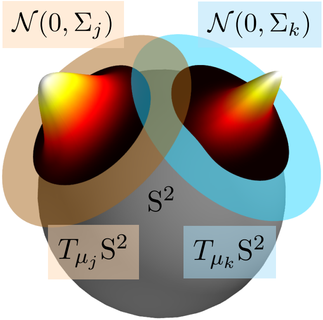


$$\pi \sim \mathsf{GEM}(1, \alpha) \qquad z_i \sim \mathsf{Cat}(\pi)$$

$$\mu_k \sim \mathsf{Unif}(\mathrm{S}^{D-1}) \qquad \Sigma_k \sim \mathsf{IW}(\Delta, \nu)$$

DP-TGMM Generative Model





$$\pi \sim \mathsf{GEM}(1, \alpha) \qquad z_i \sim \mathsf{Cat}(\pi)$$

$$\mu_k \sim \mathsf{Unif}(\mathrm{S}^{D-1}) \qquad \Sigma_k \sim \mathsf{IW}(\Delta, \nu)$$

$$x_i \sim \mathcal{N}(0, \Sigma_{z_i}) \qquad q_i = \mathrm{Exp}_{\mu_{z_i}}(x_i)$$

DP-TGMM Inference

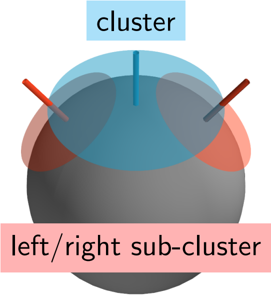
restricted Gibbs sampler with Metropolis-Hastings split/merge proposals

z $\leftarrow 1\,, K \leftarrow 1$
for $t \in \{1, \ldots, T\}$ **do**
    sample parameters
    sample labels
    propose splits of all clusters
    propose merges between all pairs of clusters
end for

DP-TGMM Inference

restricted Gibbs sampler with Metropolis-Hastings split/merge proposals

1: $\mathbf{z} \leftarrow 1\,, K \leftarrow 1$
2: **for** $t \in \{1, \ldots, T\}$ **do**
3: sample parameters
4: sample labels
5: propose splits of all clusters
6: propose merges between all pairs of clusters
7: **end for**

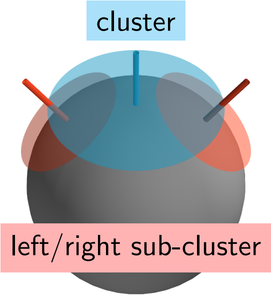


- MH split/merge proposals use inferred sub-clusters [Chang 2014]

DP-TGMM Inference

restricted Gibbs sampler with Metropolis-Hastings split/merge proposals

$\mathbf{z} \leftarrow 1\,, K \leftarrow 1$
for $t \in \{1, \ldots, T\}$ **do**
    sample parameters
    sample labels
    propose splits of all clusters
    propose merges between all pairs of clusters
end for

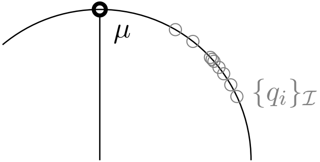


- MH split/merge proposals use inferred sub-clusters [Chang 2014]
- efficient manifold-aware inference

Efficient Manifold-aware Inference

Problem: inference requires frequent computation of sufficient statistics of $\{q_i\}_{\mathcal{I}}$ in $T_\mu \mathrm{S}^{D-1}$, where μ changes each iteration.

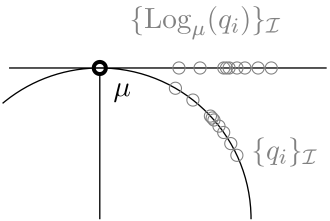
$$S_\mu = \sum_{i \in \mathcal{I}} \mathrm{Log}_\mu(q_i) \mathrm{Log}_\mu(q_i)^T$$



Efficient Manifold-aware Inference

Problem: inference requires frequent computation of sufficient statistics of $\{q_i\}_{\mathcal{I}}$ in $T_\mu \mathrm{S}^{D-1}$, where μ changes each iteration.

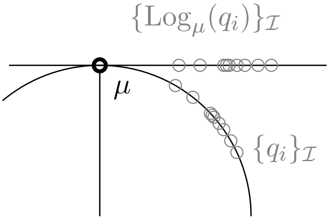
$$S_\mu = \sum\nolimits_{i \in \mathcal{I}} \mathrm{Log}_\mu(q_i) \mathrm{Log}_\mu(q_i)^T$$



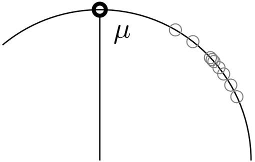
Efficient Manifold-aware Inference

Problem: inference requires frequent computation of sufficient statistics of $\{q_i\}_\mathcal{I}$ in $T_\mu \mathrm{S}^{D-1}$, where μ changes each iteration.

$$S_\mu = \sum_{i \in \mathcal{I}} \mathrm{Log}_\mu(q_i) \mathrm{Log}_\mu(q_i)^T$$



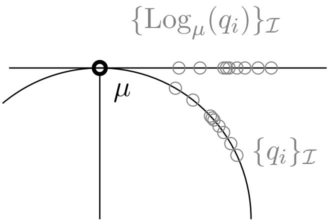
Solution: approximate S_μ using Karcher mean $\tilde{q}$ of data [Karcher 1977]



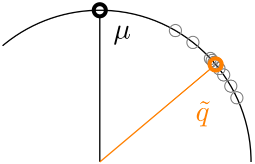
Efficient Manifold-aware Inference

Problem: inference requires frequent computation of sufficient statistics of $\{q_i\}_\mathcal{I}$ in $T_\mu \mathrm{S}^{D-1}$, where μ changes each iteration.

$$S_\mu = \sum_{i \in \mathcal{I}} \mathrm{Log}_\mu(q_i) \mathrm{Log}_\mu(q_i)^T$$



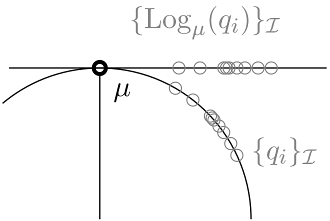
Solution: approximate S_μ using Karcher mean $\tilde{q}$ of data [Karcher 1977]



Efficient Manifold-aware Inference

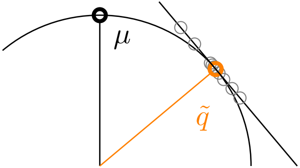
Problem: inference requires frequent computation of sufficient statistics of $\{q_i\}_\mathcal{I}$ in $T_\mu \mathrm{S}^{D-1}$, where μ changes each iteration.

$$S_\mu = \sum\nolimits_{i \in \mathcal{I}} \mathrm{Log}_\mu(q_i) \mathrm{Log}_\mu(q_i)^T$$



Solution: approximate S_μ using Karcher mean $\tilde{q}$ of data [Karcher 1977]

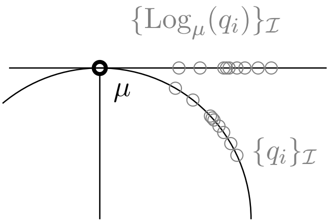
$$S_\mu \approx S_{\tilde{q}}$$



Efficient Manifold-aware Inference

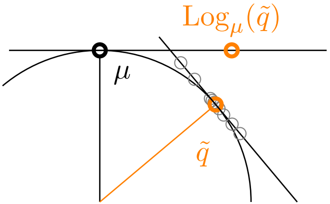
Problem: inference requires frequent computation of sufficient statistics of $\{q_i\}_\mathcal{I}$ in $T_\mu \mathrm{S}^{D-1}$, where μ changes each iteration.

$$S_\mu = \sum\nolimits_{i \in \mathcal{I}} \mathrm{Log}_\mu(q_i) \mathrm{Log}_\mu(q_i)^T$$



Solution: approximate S_μ using Karcher mean $\tilde{q}$ of data [Karcher 1977]

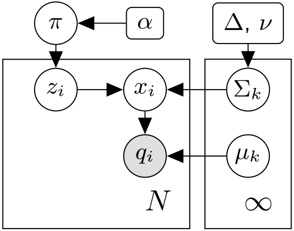
$$S_\mu \approx S_{\tilde{q}} + N \mathrm{Log}_\mu(\tilde{q}) \mathrm{Log}_\mu(\tilde{q})^T$$

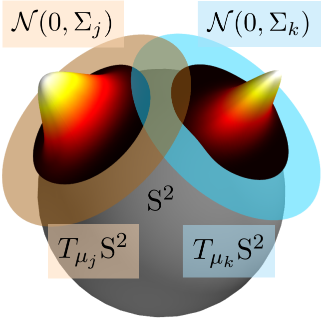


DP-TGMM Recap

Approach and Contributions:

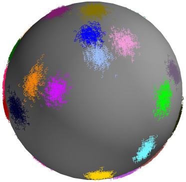
- BNP Dirichlet process mixture model for directional data
- Gaussian distributions in cluster-dependent tangent spaces
- manifold-aware sub-cluster split/merge MCMC sampler





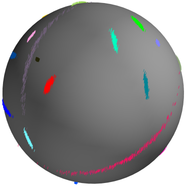
Synthetic Data – Number of Clusters

isotropic

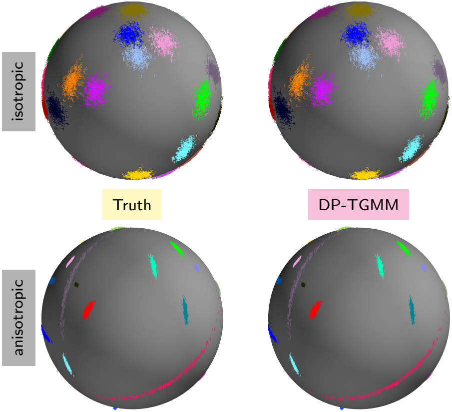


Truth

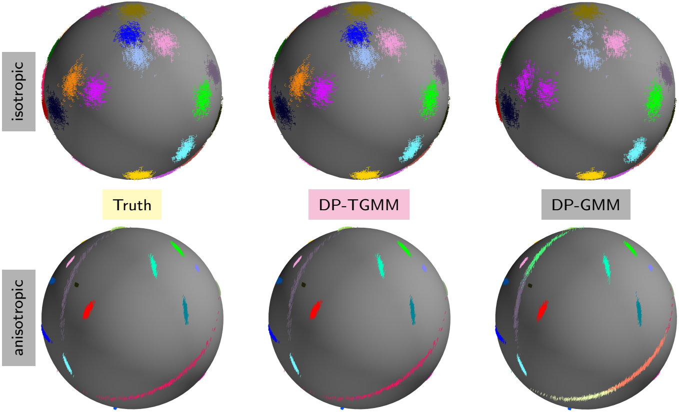
anisotropic



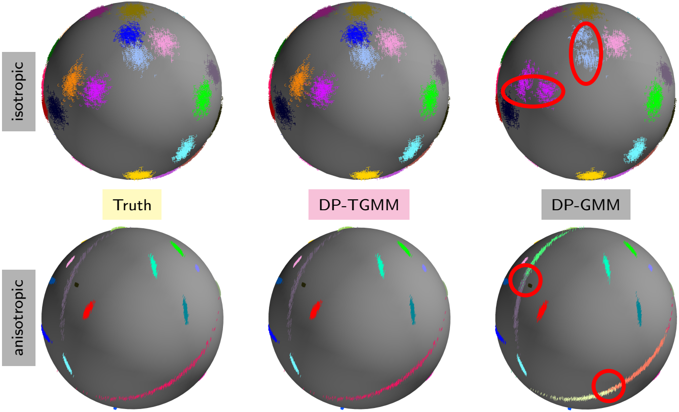
Synthetic Data – Number of Clusters



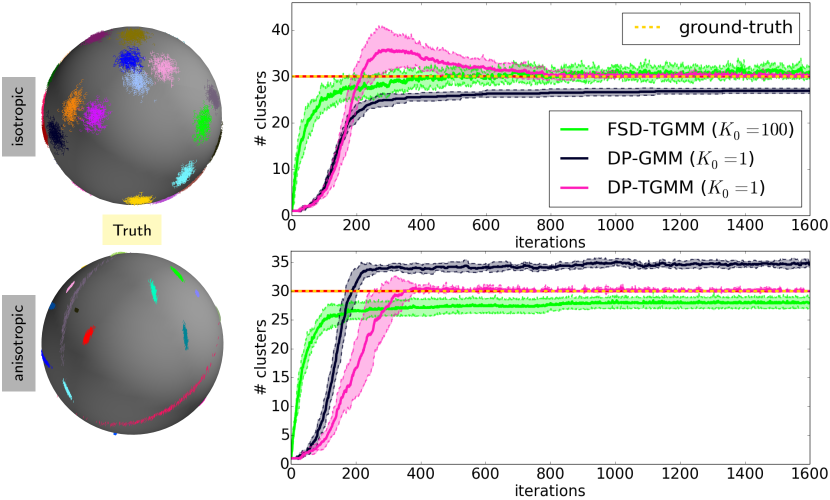
Synthetic Data – Number of Clusters



Synthetic Data – Number of Clusters



Synthetic Data – Number of Clusters



Semantic Word Vectors (word2vec)

word $\overset{\text{word2vec}}{\Longrightarrow}$ embedding space (here 20D) [Mikolov et al. 2013]

Semantic Word Vectors (word2vec)

word $\overset{\text{word2vec}}{\Longrightarrow}$ embedding space (here 20D) [Mikolov et al. 2013]

semantically similar words have small angular deviation

Semantic Word Vectors (word2vec)

word $\overset{\text{word2vec}}{\Longrightarrow}$ embedding space (here 20D) [Mikolov et al. 2013]

semantically similar words have small angular deviation

"finance"	**"music"**	**"religion "**	**"leisure"**	**"government"**	**"food"**
funding	symphonic	orthodoxy	malls	parliamentary	tomatoes
prospective	operatic	orthodox	hotels	enacted	edible
loans	soloists	evangelical	dining	parliament	fruit
financing	orchestral	christians	nightlife	delegation	meats
funds	music	primacy	outdoor	unanimously	meat
contracts	waltz	preaching	upscale	granting	vegetables
compensation	trios	doctrines	shopping	mandate	juice
regulations	lute	rabbis	restaurants	constitutional	baked
assets	soloist	clergy	taverns	citizenship	corn
investors	flute	catholicism	shops	committee	tasting

Semantic Word Vectors (word2vec) – Covariances

condition number $\kappa = \frac{\max \operatorname{eig}(\Sigma)}{\min \operatorname{eig}(\Sigma)}$ for each inferred cluster

$\kappa > 1 \Rightarrow$ anisotropic covariance

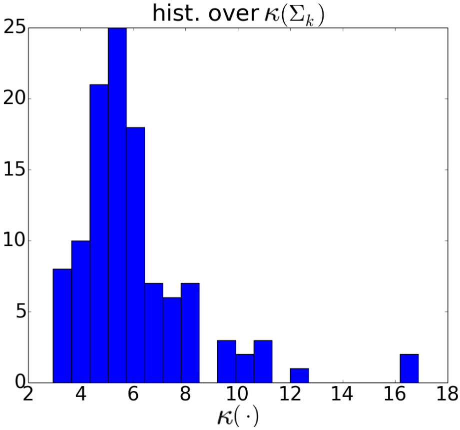
$\kappa = 1 \Rightarrow$ isotropic covariance

Semantic Word Vectors (word2vec) – Covariances

condition number $\kappa = \frac{\max \mathrm{eig}(\Sigma)}{\min \mathrm{eig}(\Sigma)}$ for each inferred cluster

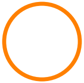
$\kappa > 1 \Rightarrow$ anisotropic covariance

$\kappa = 1 \Rightarrow$ isotropic covariance

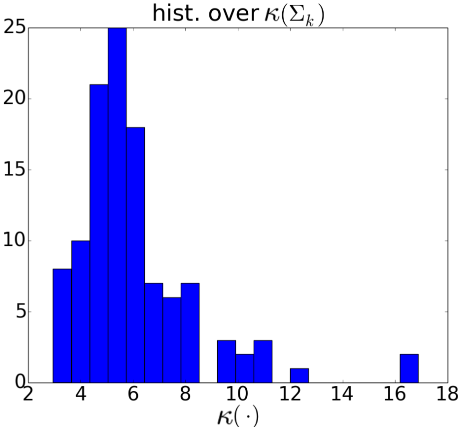


condition number $\kappa = \frac{\max \operatorname{eig}(\Sigma)}{\min \operatorname{eig}(\Sigma)}$ for each inferred cluster

$\kappa > 1 \Rightarrow$ anisotropic covariance



$\kappa = 1 \Rightarrow$ isotropic covariance

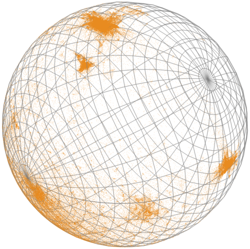


$\Rightarrow$ inferred clusters are anisotropic

Directional Segmentation

directional scene segmentation = clustering of scene's surface normals

image of scene

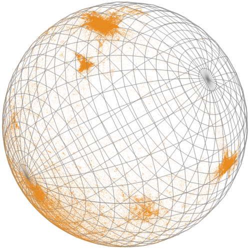


surface normals

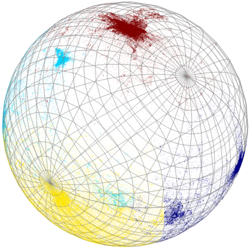
Directional Segmentation

directional scene segmentation = clustering of scene's surface normals

image of scene



surface normals

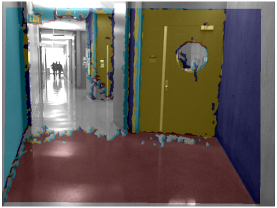


surface normal clustering

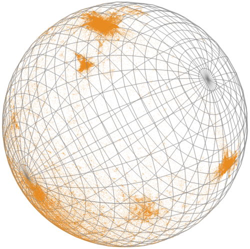
Directional Segmentation

directional scene segmentation = clustering of scene's surface normals

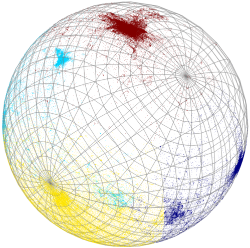
image of scene



directional segmentation



surface normals

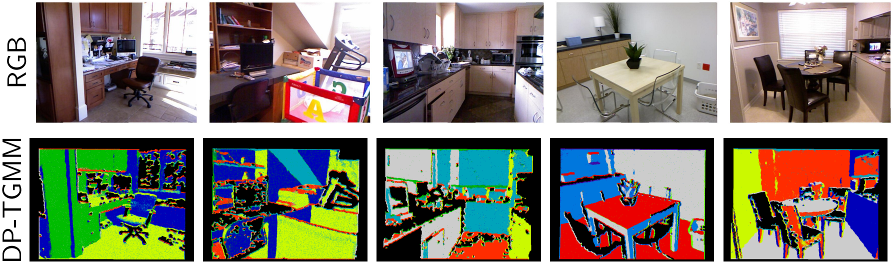


surface normal clustering

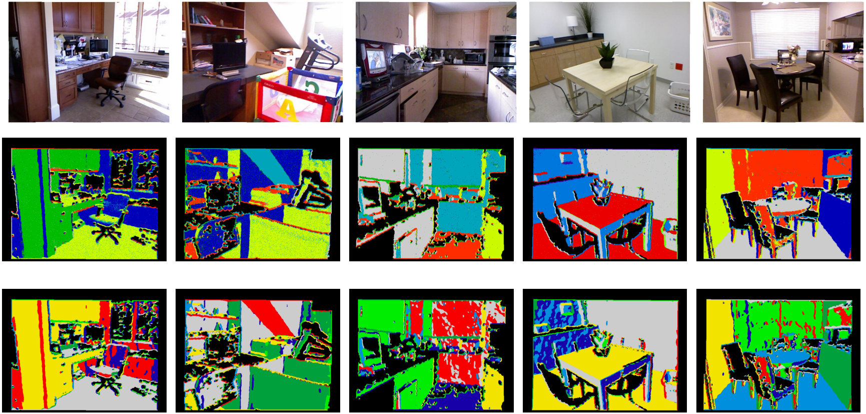
Directional Segmentation of NYU RGB-D Dataset

RGB

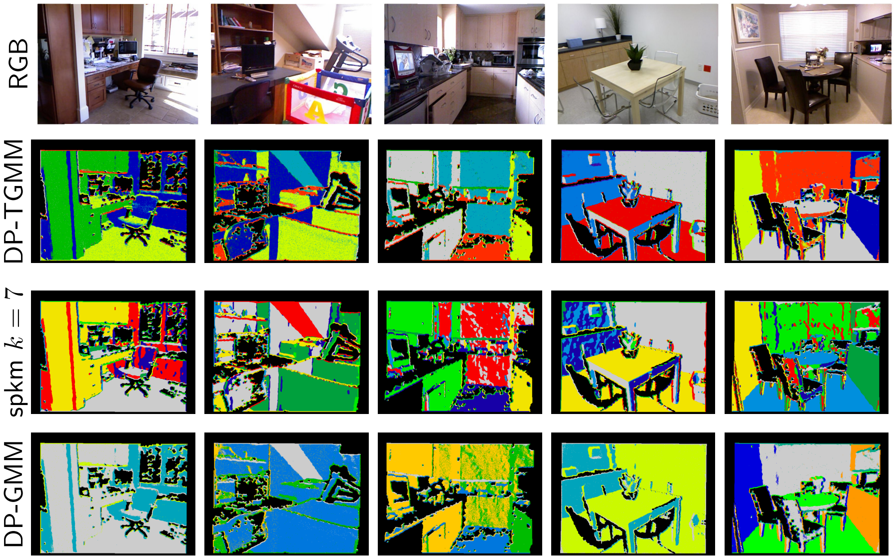
Directional Segmentation of NYU RGB-D Dataset



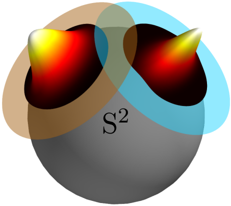
Directional Segmentation of NYU RGB-D Dataset

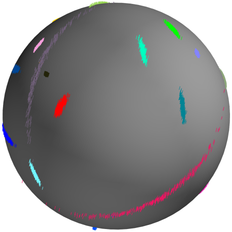


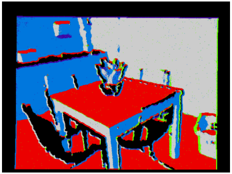
Directional Segmentation of NYU RGB-D Dataset



Conclusion

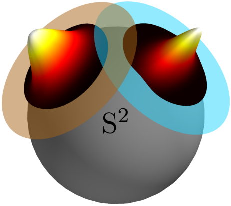


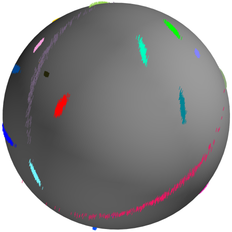


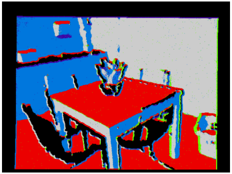


- Bayesian nonparametric mixture model for directional data
- anisotropic component distributions on the sphere
- efficient manifold-aware MCMC inference

Conclusion





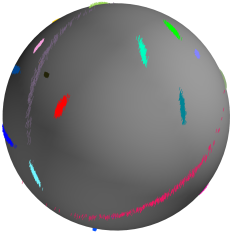


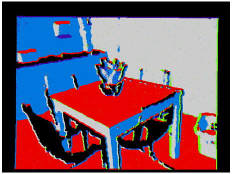
- Bayesian nonparametric mixture model for directional data
- anisotropic component distributions on the sphere
- efficient manifold-aware MCMC inference

Future work:

- generalization to Riemannian manifolds
- hierarchical models

Conclusion





- Bayesian nonparametric mixture model for directional data
- anisotropic component distributions on the sphere
- efficient manifold-aware MCMC inference

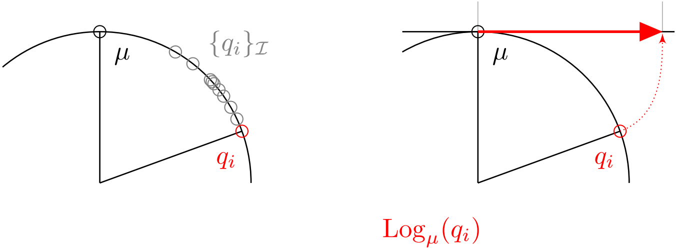
Future work:

- generalization to Riemannian manifolds
- hierarchical models

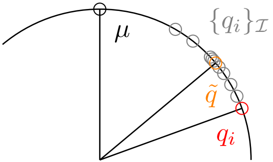
links to paper, slides and code are at http://people.csail.mit.edu/jstraub/

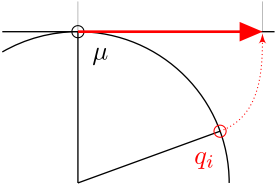
Efficient Manifold-aware Inference

Problem: inference requires frequent mapping $\{q_i\}_{\mathcal{I}}$ into $T_\mu \mathrm{S}^{D-1}$, where μ changes each iteration.



Problem: inference requires frequent mapping $\{q_i\}_{\mathcal{I}}$ into $T_\mu \mathrm{S}^{D-1}$, where μ changes each iteration.
Solution: approximate $\mathrm{Log}_\mu(q)$ using Karcher mean [Karcher 1977] $\tilde{q}$ of data

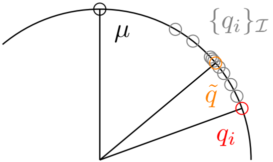


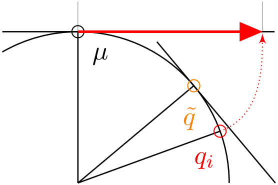


$$\tilde{q} = \arg\min_{p \in \mathrm{S}^{D-1}} \sum_{i \in \mathcal{I}} \arccos^2(p^T q_i) \qquad \mathrm{Log}_\mu(q_i)$$

Efficient Manifold-aware Inference

Problem: inference requires frequent mapping $\{q_i\}_{\mathcal{I}}$ into $T_\mu \mathrm{S}^{D-1}$, where μ changes each iteration.
Solution: approximate $\mathrm{Log}_\mu(q)$ using Karcher mean [Karcher 1977] $\tilde{q}$ of data

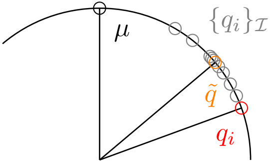




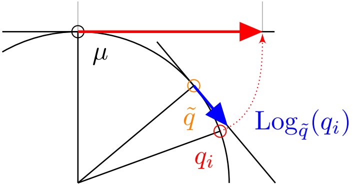
$$\tilde{q} = \arg\min_{p \in \mathrm{S}^{D-1}} \sum_{i \in \mathcal{I}} \arccos^2(p^T q_i) \qquad \mathrm{Log}_\mu(q_i)$$

Efficient Manifold-aware Inference

Problem: inference requires frequent mapping $\{q_i\}_\mathcal{I}$ into $T_\mu \mathrm{S}^{D-1}$, where μ changes each iteration.
Solution: approximate $\mathrm{Log}_\mu(q)$ using Karcher mean [Karcher 1977] $\tilde{q}$ of data



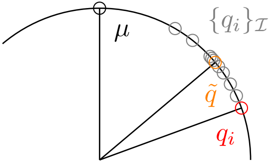
$$\tilde{q} = \underset{p \in \mathrm{S}^{D-1}}{\arg\min} \sum_{i \in \mathcal{I}} \arccos^2(p^T q_i)$$

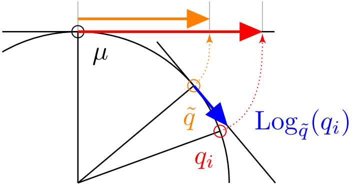


$\mathrm{Log}_\mu(q_i)$

Efficient Manifold-aware Inference

Problem: inference requires frequent mapping $\{q_i\}_{\mathcal{I}}$ into $T_\mu \mathrm{S}^{D-1}$, where μ changes each iteration.
Solution: approximate $\mathrm{Log}_\mu(q)$ using Karcher mean [Karcher 1977] $\tilde{q}$ of data

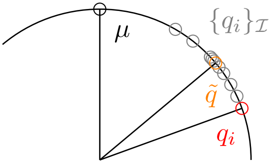


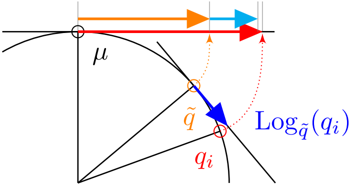


$$\tilde{q} = \underset{p \in \mathrm{S}^{D-1}}{\arg\min} \sum_{i \in \mathcal{I}} \arccos^2(p^T q_i) \qquad \mathrm{Log}_\mu(q_i) \approx \mathrm{Log}_\mu(\tilde{q})$$

Efficient Manifold-aware Inference

Problem: inference requires frequent mapping $\{q_i\}_\mathcal{I}$ into $T_\mu \mathrm{S}^{D-1}$, where μ changes each iteration.
Solution: approximate $\mathrm{Log}_\mu(q)$ using Karcher mean [Karcher 1977] $\tilde{q}$ of data





$$\tilde{q} = \underset{p \in \mathrm{S}^{D-1}}{\arg\min} \sum_{i \in \mathcal{I}} \arccos^2(p^T q_i)$$

$$\mathrm{Log}_\mu(q_i) \approx \mathrm{Log}_\mu(\tilde{q}) + R_{\tilde{q}}^{\mu} \mathrm{Log}_{\tilde{q}}(q_i)$$