Speaker clustering via the mean-shift algorithm

Themos Stafylakis

National Technical University of Athens,

Institute for Language & Speech Processing,

Athens, Greece

Outline of the presentation

- Non-parametric density estimation
- Baseline mean shift
- Why it requires adaptation
- Bayesian setting (seek the modes of the posterior)
- Distances – Divergences
- Proposed Kernels
- Exponential family basics
- Derivation of the proposed algorithm
- Experiments & future work

Basics about the mean-shift algorithm

What's that?

- An elegant non-parametric approach to clustering
- # clusters are not required to be known a priori
- Also known as mode seeking algorithm
- Alternative to hierarchical clustering, spectral clustering, etc.

Current applications

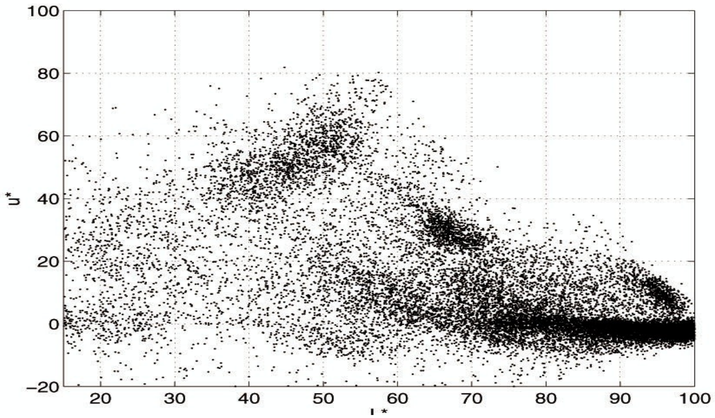
- Image segmentation
- Discontinuity preserving filtering
- Boundary detection
- Object tracking (2D & 3D)

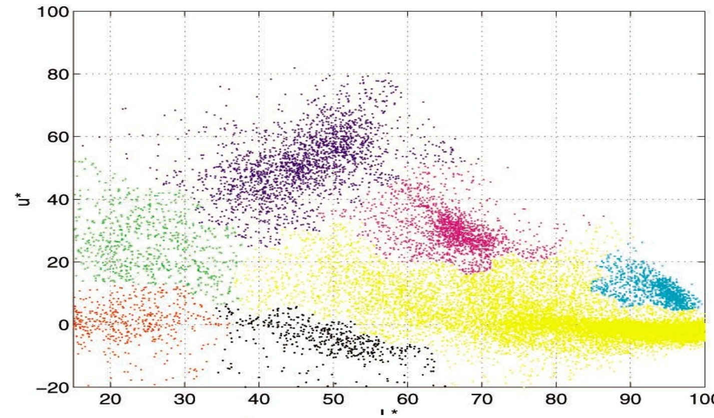
Main references

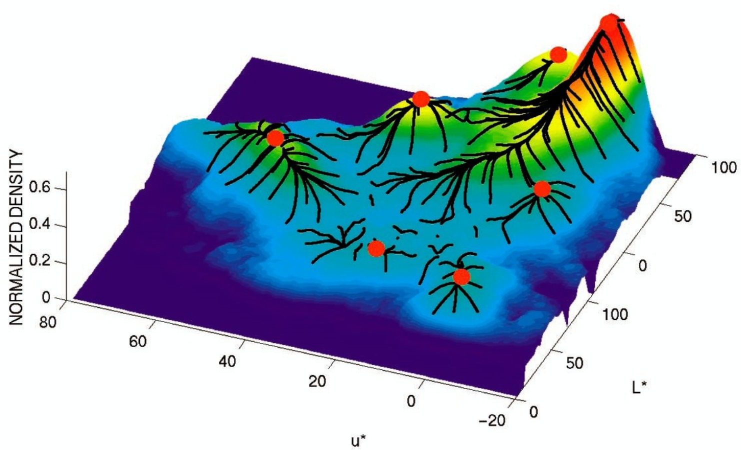
All of D. Comaniciu & P. Meer, expecially

“Mean Shift: a robust approach towards feature space analysis”, IEEE-PAMI, May 2002

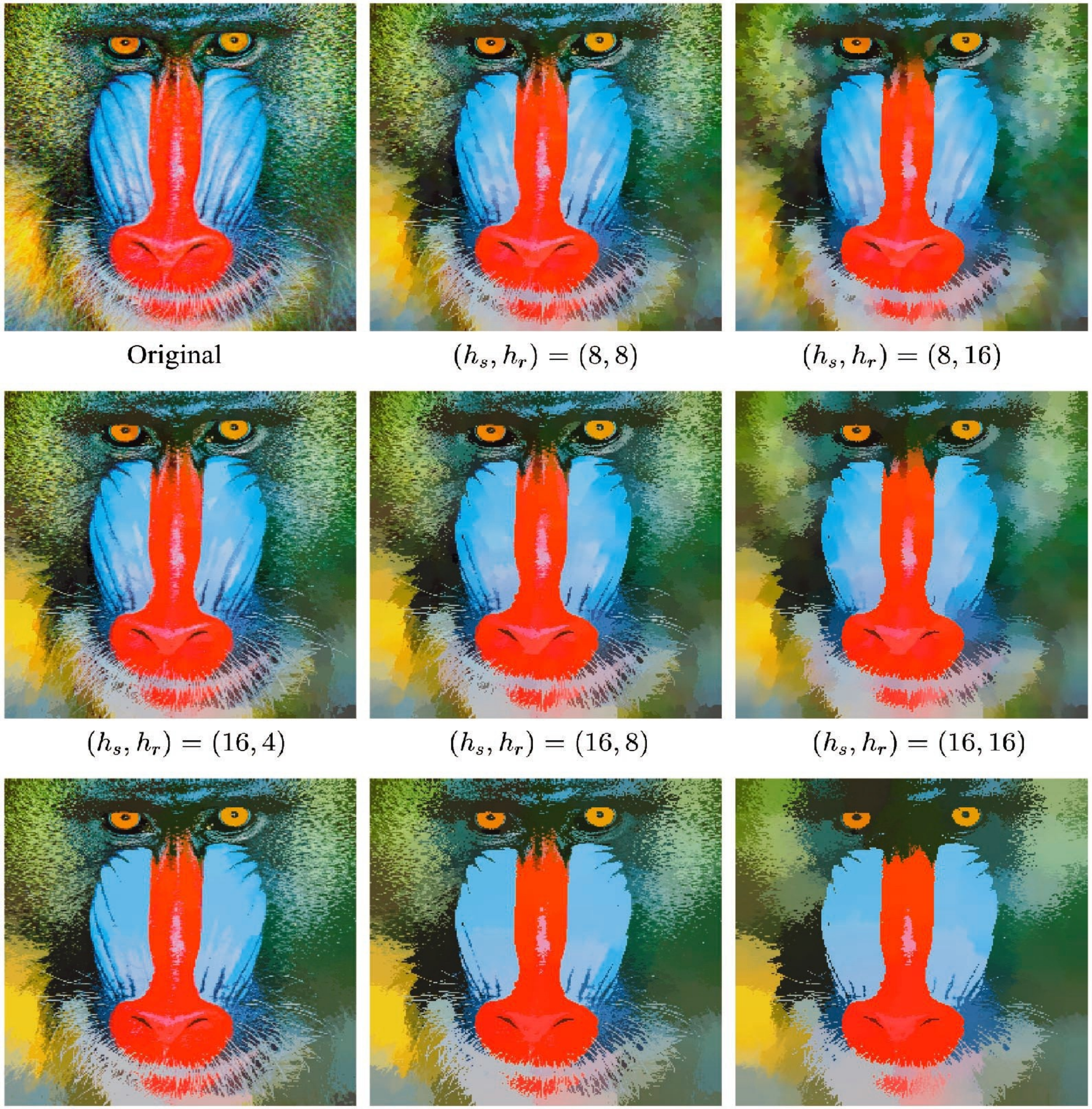
An example from image segmentation







An example from discontinuity preserving filtering



Original

$(h_s, h_r) = (8, 8)$

$(h_s, h_r) = (8, 16)$

$(h_s, h_r) = (16, 4)$

$(h_s, h_r) = (16, 8)$

$(h_s, h_r) = (16, 16)$

$(h_s, h_r) = (32, 4)$

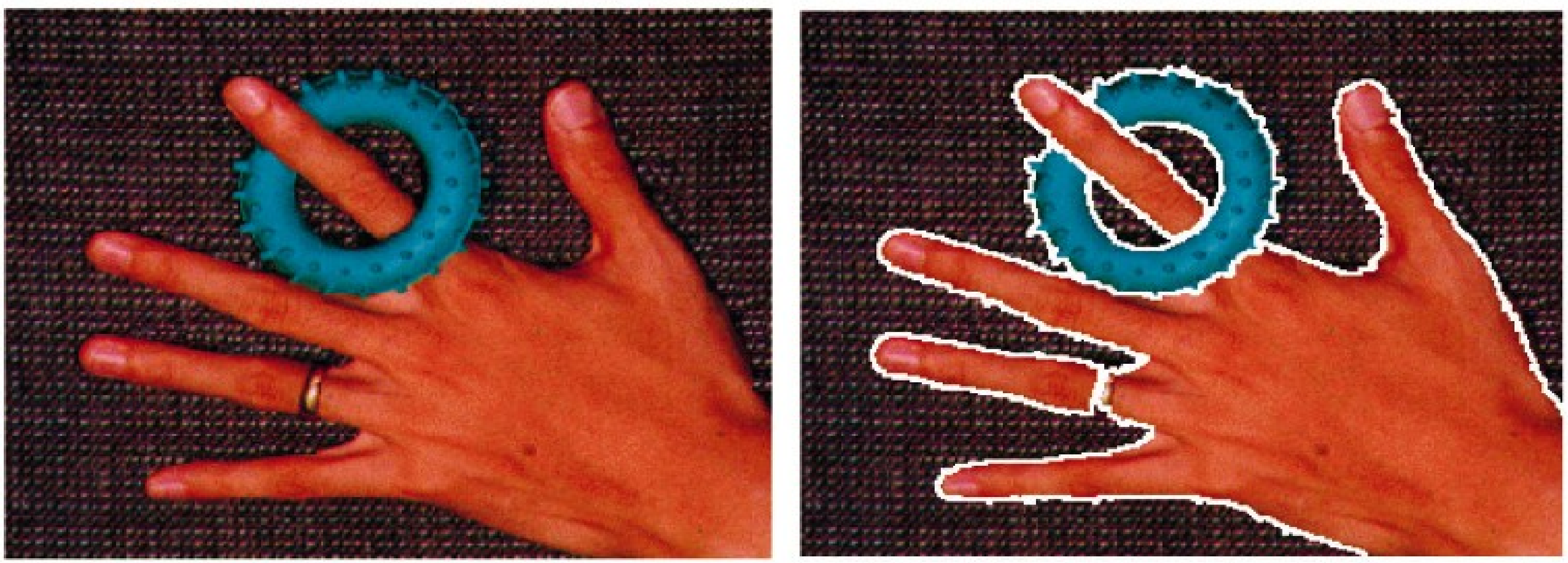
$(h_s, h_r) = (32, 8)$

$(h_s, h_r) = (32, 16)$

(Comaniciu & Meer, IEEE - PAMI, '02)

Examples from boundary detection

(a) (b) (c)



(a) (b)

(Comaniciu & Meer, IEEE - PAMI, '02)

Contribution of the proposed method

Limitations of the mean-shift algorithm:

- The original mean-shift acts on the space of observations (RGB, LUV, etc.)
- Several clustering tasks require probabilistic parametric models as entities
- Example: speaker clustering, i.e. given N utterances merge those being from the same speaker
- Note: We always assume that the #clusters is unknown
- Task: Adapt the mean-shift to act on the space of parametric models.

The proposed method:

- is based on the exponential family (Normal, Poisson, Gamma, Beta, multinomial, categorical, etc.)
- uses a Bayesian statistical setting (basically conjugate-exponential models)
- Can be explained completely using the theory of Information Geometry (Amari, Rodriguez, Snoossi, a.o.)

The original mean-shift algorithm (I)

Standard non-parametric density estimation

We have some data X, generated from an unknown pdf $f(\mathbf{x})$

$$\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{n}, \mathbf{x}^{(i)} \in \Re^d$$

We estimate the pdf using Parzen windows

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}^{(i)}), \quad K_{\mathbf{H}}(\mathbf{x}) = c_d|\mathbf{H}|^{-1/2}K(\mathbf{H}^{-1/2}\mathbf{x})$$

Assume only radically symmetric kernels, i.e. $K(\mathbf{x}) = c_{k,d}k(||\mathbf{x}||^2)$.

The normal (Gaussian) kernel

$$K_N(\mathbf{x}) = (2\pi)^{d/2}\exp\left(-\frac{1}{2}||\mathbf{x}||^2\right)$$

The estimated pdf is as follows

$$\hat{f}_{h,K}(\mathbf{x}) = \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{x}-\mathbf{x}^{(i)}}{h}\right\|^2\right)$$

where h is the bandwidth of the kernel.

The original mean-shift algorithm (II)

Parametric vs. non-parametric

- Non-parametric models allow #parameters grow linearly with #data points, n,
- Make very few assumptions about the data generating process.
- The parameters are the data points themselves + the bandwidth.
- The bandwidth can be variable, depending of the density of the region.

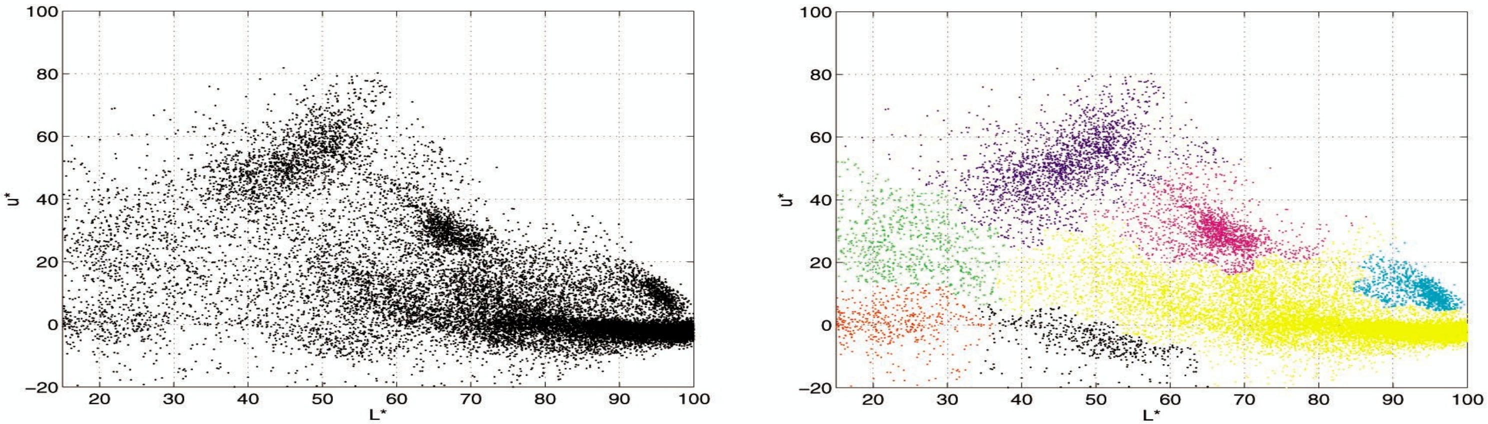


Fig.1: Real world clusters exhibit arbitrary shapes

Basic problem with non-parametric modeling

- You rarely have enough data to estimate the pdf robustly.
- #observations required grows exponentially with the dimensionality.
- You don't obtain compact representation of the models.

The original mean-shift algorithm (III)

But do actually we need to estimate the underlying pdf robustly?

Assume a standard clustering task...

Target:

a point-estimate about the cluster assignments for each observation.

Requirements:

a) a method to detect the modes of the pdf.

b) a method to assign each observation to the appropriate mode.

That's all we need – that's what the mean-shift does!

a) It uses direclty the gradient of the pdf to estimate the modes.

b) It provides a method to assign the data to the modes.

Let's see how it works...

The original mean-shift algorithm (IV)

Recall the expression of the estimated pdf

$$\hat{f}_{h,K}(\mathbf{x}) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} k\left(\left\| \frac{\mathbf{x} - \mathbf{x}^{(i)}}{h} \right\|^2 \right)$$

Differentiate it w.r.t. $\mathbf{x}$, and set to zero (i.e. mode seeking)

$$\nabla \hat{f}_{h,K}(\mathbf{x}) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (\mathbf{x} - \mathbf{x}^{(i)}) k'\left(\left\| \frac{\mathbf{x} - \mathbf{x}^{(i)}}{h} \right\|^2 \right)$$

Define the differential kernel profile $g(x) = -k'(x)$ the gradient yields

$$\hat{\nabla} f_{h,K}(\mathbf{x}) = \frac{2c_{k,d}}{h^2 c_{g,d}} \hat{f}_{h,G}(\mathbf{x}) \mathbf{m}_{h,G}(\mathbf{x}),$$

where

$$\hat{f}_{h,G}(\mathbf{x}) = \frac{c_{g,d}}{nh^d} \sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x}^{(i)} - \mathbf{x}}{h} \right\|^2 \right)$$

proportional to the density estimated using the kernel with profile $g(x)$

and the mean-shift vector

$$\mathbf{m}_{h,G}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}^{(i)} g\left(\left\| \frac{\mathbf{x}^{(i)} - \mathbf{x}}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x}^{(i)} - \mathbf{x}}{h} \right\|^2 \right)} - \mathbf{x}$$

which vanishes iff a mode (or a saddle point) has been detected!

The original mean-shift algorithm (V)

The actual algorithm

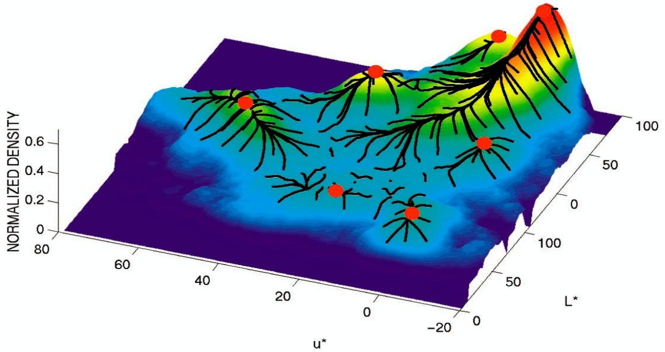
We need to find where the mean-shift vector vanishes.

Recall that $\mathbf{m}_{h,G}(\mathbf{x}) = \dfrac{\sum_{i=1}^{n} \mathbf{x}^{(i)} g\left(\left\|\frac{\mathbf{x}^{(i)}-\mathbf{x}}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}^{(i)}-\mathbf{x}}{h}\right\|^2\right)} - \mathbf{x}$

For each observation $i = 1, 2 \ldots, n$ set $t = 0, \mathbf{x}_t \leftarrow \mathbf{x}^{(i)}$

1. calculate $\mathbf{m}_{h,G}(\mathbf{x}_t)$
2. set $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \mathbf{m}_{h,G}(\mathbf{x}_t)$
3. if $\|\mathbf{x}_{t+1} - \mathbf{x}_t\| < \epsilon$ goto 4; else $t \leftarrow t + 1$ and goto 1.
4. store $\tilde{\mathbf{x}}^{(i)} = \mathbf{x}_{t+1}$.

The matrix $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}^{(1)}, \tilde{\mathbf{x}}^{(2)}, \ldots, \tilde{\mathbf{x}}^{(n)}]$ contains the convergent points.



Each iteration independent of the other. Ideal for cluster computing

Motivation for the proposed method

How to adapt this idea to operate on the space of distributions?

- We start having N distribution of the same family and order (one for each segment).
- We then define the kernel, i.e. its shape and its distance.
- The pdf can be regarded as the posterior of θ, given the complete data (X,Z).
- Task: find the modes of the posterior, assign each θ to the correct mode.
- Note the correspondence: Kernel function & posterior distribution of θ.

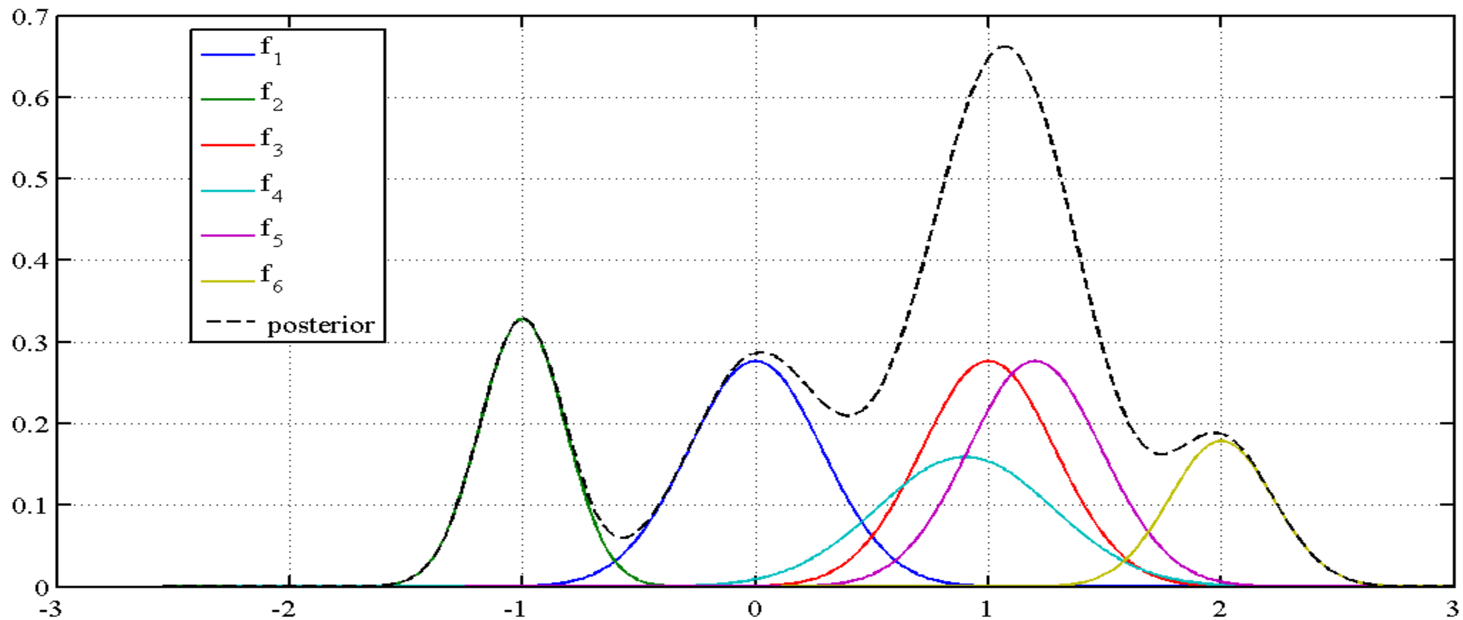
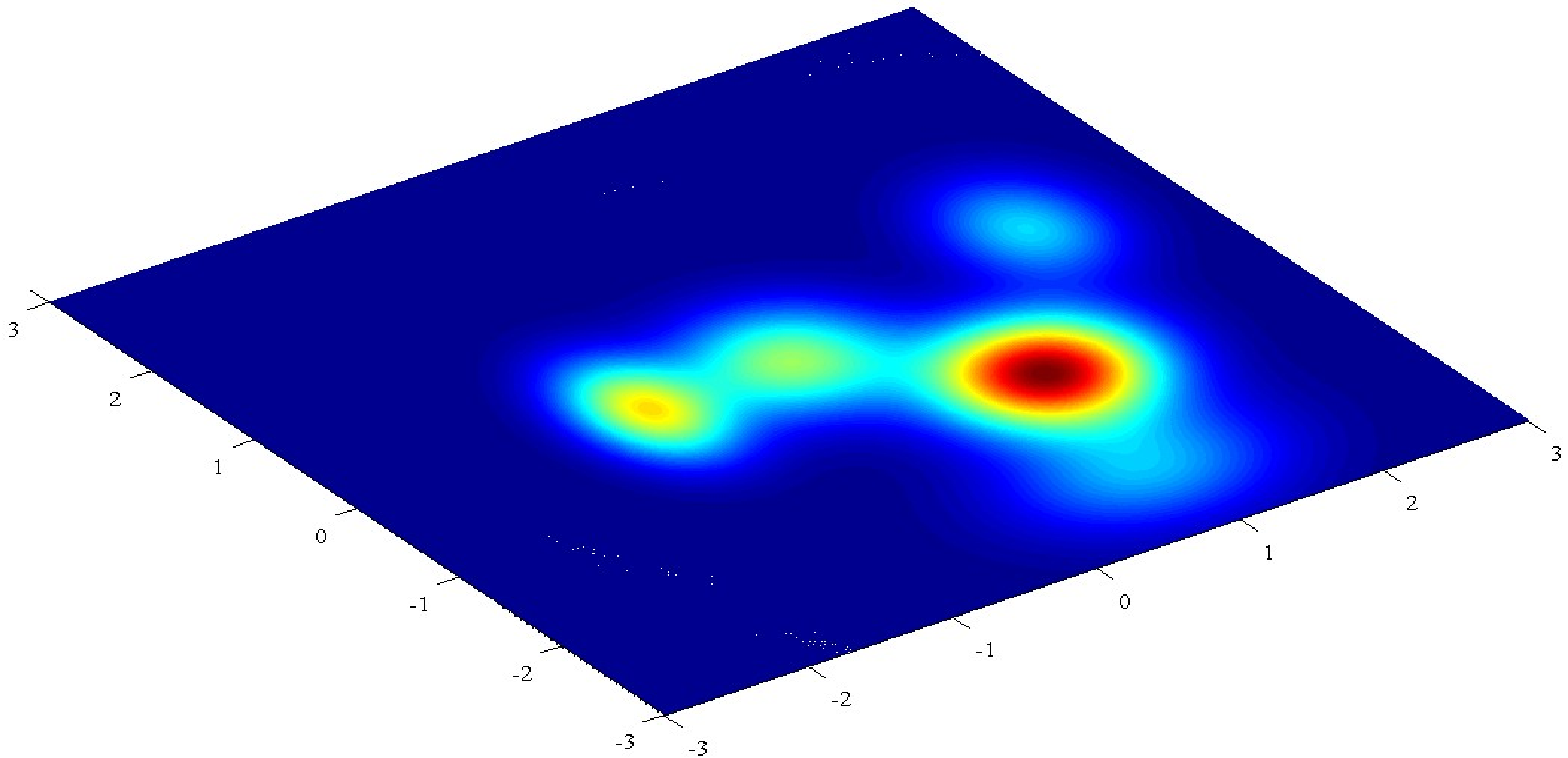


Fig.2: 6 initial segments that form 4 clusters, assuming normal kernel

Motivation for the proposed method

Same example in 2-D parameter space



- Starting from each point, find its closest maximum in the parameter space.
- Note: It can be it self, i.e. attracted by it self!
- Consider the alteration as dealing with a higher level in the Bayesian Hierarchy.

Kernels on the space of distributions (I)

Step 1: Define an appropriate measure of deviation

$$D_\delta(p,q) = \begin{cases} \frac{\int p dx}{1-\delta} + \frac{\int q dx}{\delta} - \frac{\int p^\delta q^{1-\delta} dx}{\delta(1-\delta)}, \text{ if } \delta \in (0,1) \\ \int p \log(\frac{p}{q}) dx, \text{ if } \delta = 1 \\ \int q \log(\frac{q}{p}) dx, \text{ if } \delta = 0 \end{cases}$$

- For $\delta = 0$ or 1: Kullback-Leibler divergence
- For $\delta = 1/2$: Twice the Hellinger (squared) distance, the only symmetric deviation

$$D_{1/2}(p,q) = 2 \int (\sqrt{p} - \sqrt{q})^2 \, dx$$

We may also symmetrize the KL divergence by using the summation, twice the minimum, twice the harmonic mean, etc.

Note, for all δ:

$$D(p(\mathbf{x};\theta) || p(\mathbf{x};\theta + \delta\theta)) \approx \frac{1}{2} \delta\theta^T G(\theta) \delta\theta$$

where $G(\theta)$ the Fisher Information Matrix.

Kernels on the space of distributions (II)

Step 2: Define the shape of the kernel

$$K_{\delta,\alpha}(p_\theta, p) \propto \begin{cases} \sqrt{G(\theta)}[1+\lambda D_\delta(p_\theta,p)]^{\frac{-1}{1-\alpha}}, \text{ if } 0 \le \alpha < 1 \\ \sqrt{G(\theta)} e^{-\lambda D_\delta(p_\theta,p)}, \ \alpha = 1 \end{cases}$$

The is a very rich geometry underlying this family!

See Information Geometry (Amari, Kass, Rodriguez, Snoossi, a.o.)

- R. Kass, "*The Geometry of asymptotic inference*"
- S.-I. Amari, "*Differential Geometry of Curved Exponential Families-Curvatures and Information Loss*"
- C. Rodriguez, "*A geometric theory of Ignorance*"
- H. Snoossi, "*Bayesian Information Geometry. Application to Prior Selection on Statistical Manifolds*"

Kernels on the space of distributions (III)

Bayesian rationale and derivation of the family of kernels

Consider the cost function

$$\mathcal{J}_{\delta,\alpha}(\Pi) = \gamma_e \int \Pi(\theta) D_\delta(p_\theta, p_0) d\theta + \gamma_u D_\alpha(\Pi(\theta), \sqrt{G(\theta)})$$

The family is generated by minimizing the cost function w.r.t. $\Pi(\theta)$ using calculus of variations (Rodriguez, Snoossi, a.o.)

γ_e : how confident you are about the location p_0

γ_u : how close should be to the uninformative (Jeffreys) prior, $\lambda \leftarrow \frac{\gamma_e}{\gamma_u}$

δ : the type of deviation between the Likelihood functions (observation space)

α : the type of deviation between $\Pi(\theta)$ and the Jeffreys prior (parameter space)

- (δ,α) = (1,1): the entropic prior, Normal-Wishart (for Gaussian likelihoods),
- (δ,α) = (0,1): the usual conjugate prior, Normal-Inverse Wishart,
- If Euclidean geometry in θ, $\alpha<1$: *t*-distribution, $\alpha=1$: Gaussian distribution
- γ_e/γ_u goes to zero: the Jeffreys prior
- γ_e/γ_u goes to infinity: a single probability mass at p_0

Kernels on the space of distributions (IV)

Step 3: Differentiate the posterior of θ to obtain the mean-shift vector

$$\pi(p_\theta|X,Z) \propto \sqrt{|G(\theta)|}\sum_{k=1}^{K}\frac{n_k}{n}\exp\left(-\lambda_k D_\delta(p_\theta||p_{\theta_k})\right)$$

By differentiating it w.r.t. θ and setting it to zero you obtain the mean-shift vector.

Analytic solution? Yes, if the likelihood belongs to the exponential family of distributions:

$$p(\mathbf{x};\boldsymbol{\theta}) = h(\mathbf{x})\exp(\boldsymbol{\theta}\cdot\mathbf{t}(\mathbf{x}) - \psi(\boldsymbol{\theta}))$$

$$\psi(\boldsymbol{\theta}) = \log\int_{\mathcal{X}}\exp(\boldsymbol{\theta}\cdot\mathbf{t}(\mathbf{x}))h(\mathbf{x})d\mathbf{x}$$

$\psi(\boldsymbol{\theta})$: the log-partition function (convex in θ),

$\boldsymbol{\theta}$: the natural parameters, $\boldsymbol{\theta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$

$\mathbf{t}(\mathrm{x})$: the sufficient statistics of x, $\mathbf{t}(x) = (x, x^2)$

$h(\mathrm{x})$: the dominant measure, constant for Gaussian likelihoods.

The exponential family has many appealing properties....

Kernels on the space of distributions (V)

Fundamental properties of the exponential family

Due to the convexity of $\psi(\theta)$ w.r.t. θ we may define the expectation parameters

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) = (\mu, \sigma^2 + \mu^2)$$
$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \int_{\mathcal{X}} \mathrm{t}(\mathrm{x})p(\mathrm{x};\boldsymbol{\theta})h(\mathrm{x})d\mathrm{x}$$

Second order derivatives

$$G(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\boldsymbol{\eta}$$
$$G(\boldsymbol{\theta}) = \begin{pmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 4\mu^2\sigma^2 + 2\sigma^4 \end{pmatrix}$$

Fisher Information: Lower bound of variance when estimating η based on a sample of a single observation (Cramer-Rao bound)

Hence, log-likelihood of θ given $\mathrm{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{n}$

$$\mathcal{L}(\boldsymbol{\theta}; X) = n(\boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta}))$$

For unitary sample size (negative Shannon entropy)

Legendre Transforms:

$$\phi(\boldsymbol{\eta}) = \max_{\boldsymbol{\theta}}\{\boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta})\} \qquad \psi(\boldsymbol{\theta}) = \max_{\boldsymbol{\eta}}\{\boldsymbol{\theta} \cdot \boldsymbol{\eta} - \phi(\boldsymbol{\eta})\}$$

Kernels on the space of distributions (VI)

Q: Why do we need all this theory?

A: To be able to differentiate the kernel and avoid heuristics!

Derivatives of the Kullback Leibler Divergence

$$D_1(p^k||p^l) = (\boldsymbol{\theta}^k - \boldsymbol{\theta}^l) \cdot \boldsymbol{\eta}^k - (\psi(\boldsymbol{\theta}^k) - \psi(\boldsymbol{\theta}^l))$$

$$D_0(p^k||p^l) = (\boldsymbol{\eta}^k - \boldsymbol{\eta}^l) \cdot \boldsymbol{\theta}^k - (\phi(\boldsymbol{\eta}^k) - \phi(\boldsymbol{\eta}^l))$$

We obtain

$$\nabla_{\boldsymbol{\theta}^k} D_1(p^k||p^l) = G(\boldsymbol{\theta}^k)(\boldsymbol{\theta}^k - \boldsymbol{\theta}^l) \approx (\boldsymbol{\eta}^k - \boldsymbol{\eta}^l)$$

$$\nabla_{\boldsymbol{\theta}^k} D_0(p^k||p^l) = \boldsymbol{\eta}^k - \boldsymbol{\eta}^l$$

Note: To obtain a gradient algorithm you need the natural gradient!

Definition: $\tilde{\nabla}_{\theta^k} = G(\theta^k)^{-1} \nabla_{\theta^k}$

Differentiate: $\tilde{\nabla}_{\boldsymbol{\theta}^k} D_1(p^k||p^l) = \boldsymbol{\theta}^k - \boldsymbol{\theta}^l$

$$\tilde{\nabla}_{\boldsymbol{\theta}^k} D_0(p^k||p^l) = G(\boldsymbol{\eta}^k)(\boldsymbol{\eta}^k - \boldsymbol{\eta}^l) \approx \boldsymbol{\theta}^k - \boldsymbol{\theta}^l$$

The latter approximation holds since:

$$G(\boldsymbol{\eta}^k)(\boldsymbol{\eta}^k - \boldsymbol{\eta}^l) = (\nabla_{\boldsymbol{\eta}} \boldsymbol{\theta}(\boldsymbol{\eta}))\big|_{\boldsymbol{\eta}=\boldsymbol{\eta}^k} (\boldsymbol{\eta}^k - \boldsymbol{\eta}^l) \approx \boldsymbol{\theta}^k - \boldsymbol{\theta}^l$$

The modified mean-shift algorithm

Assuming (δ,α) = (0,1) (i.e. normal-inverse Wishart) the estimated posterior is

$$\pi(p_\theta|X,Z) \propto \sqrt{|G(\theta)|} \sum_{k=1}^{K} \frac{n_k}{n} \exp\left(-\lambda_k D_\delta(p_\theta||p_{\theta_k})\right)$$

Note: The normalizing constant is unnecessary (we are applying mode seeking)

Set the derivative of the posterior *w.r.t.* $\boldsymbol{\theta}$ to zero to obtain

$$\boldsymbol{\eta}_{t+1} \leftarrow \frac{\sum_{k=1}^{K} \frac{n_k}{n} \exp\left(-\lambda_k D_\delta(p_t||p_{\theta_k})\right)\left(\lambda_k \boldsymbol{\eta}^k + \frac{1}{2}|G(\boldsymbol{\theta})|^{-1}\nabla_{\boldsymbol{\theta}}|G(\boldsymbol{\theta})|\right)}{\sum_{k=1}^{K} \frac{n_k}{n}\lambda_k \exp\left(-\lambda_k D_\delta(p_t||p_{\theta_k})\right)}$$

Approximation for sufficienlty large sample sizes

$$\boldsymbol{\eta}_{t+1} \leftarrow \frac{\sum_{k=1}^{K} \frac{n_k}{n}\lambda_k \exp\left(-\lambda_k D_\delta(p_t||p_{\theta_k})\right)\boldsymbol{\eta}^k}{\sum_{k=1}^{K} \frac{n_k}{n}\lambda_k \exp\left(-\lambda_k D_\delta(p_t||p_{\theta_k})\right)}$$

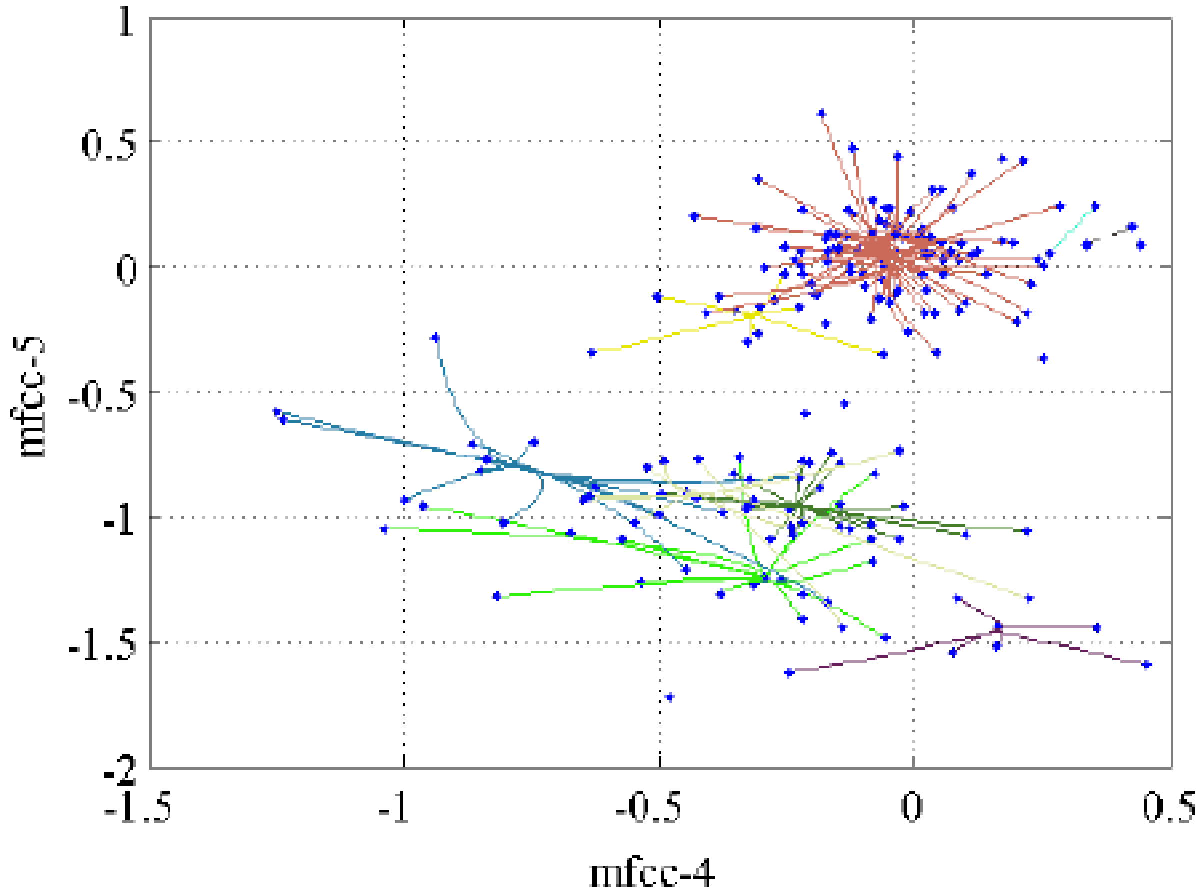
i.e. the usual weighted average in the $\boldsymbol{\eta}$-parametrization.

Assuming (δ,α) = (1,1) i.e. normal-Wishart prior we obtain (by differentiating w.r.t. $\boldsymbol{\theta}$)

$$\boldsymbol{\theta}_{t+1} \leftarrow \frac{\sum_{k=1}^{K} \frac{n_k}{n}\lambda_k \exp\left(-\lambda_k D_\delta(p_t||p_{\theta_k})\right)\boldsymbol{\theta}^k}{\sum_{k=1}^{K} \frac{n_k}{n}\lambda_k \exp\left(-\lambda_k D_\delta(p_t||p_{\theta_k})\right)}$$

i.e. the usual weighted average in the $\boldsymbol{\theta}$-parametrization.

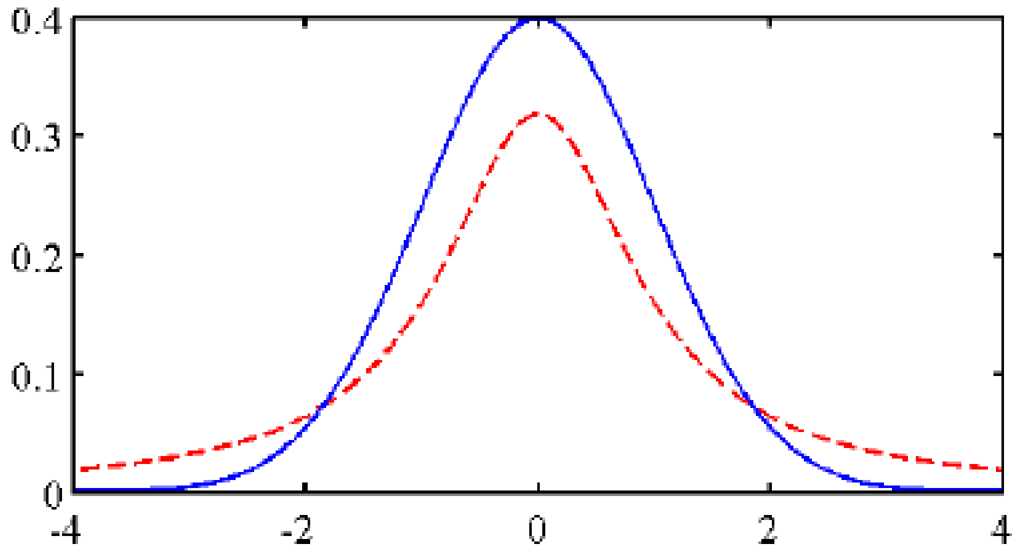
An illustrative example from BN



- 191 segments merged into 16 clusters
- Blue dots are their initial position
- 6 clusters are singletones

Avoid fast transitions between speaker

- Decrease the divergence between segments that are close enough
- Multiply the kernels by a pdf having heavy tails
- Cauchy and Laplacian work well (prefer the Cauchy)
- Consider the posterior as being a function of time (as seen by each segment)



Cauchy vs. Gaussian

Note the heavy tails of the cauchy distribution

Some experiments on speaker clustering

Benchmark Test

- ESTER Speaker Clustering Dataset – 32 Brodcasts from the French Radio
- ESTER-DEV set (14 BN shows, ~7h total duration)
- ESTER-TEST set (18 BN shows, ~9h total duration)

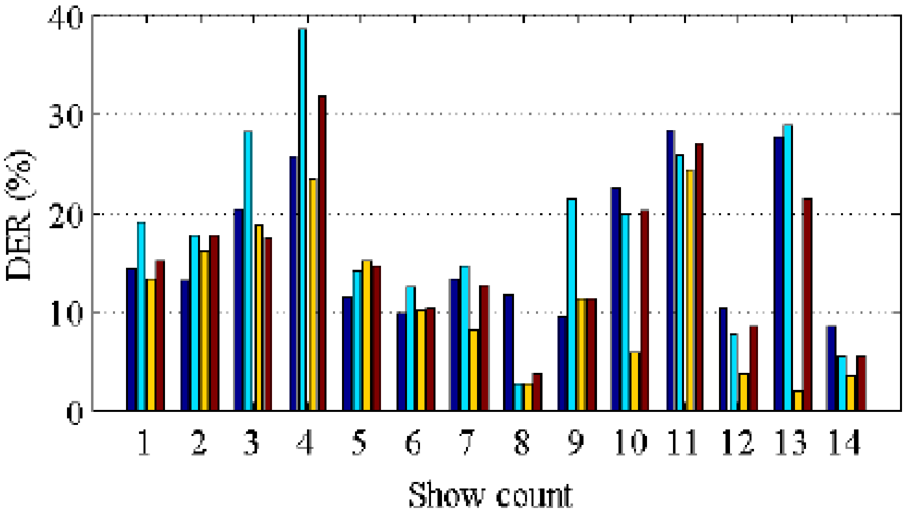
Method used

- Preprocessing (i.e. MFCC extraction)
- Speaker turn detection (oversegmentation)
- No Viterbi alignment

For comparison

- Hierarchical Clustering using ΔBIC (LIUM open-source software)
- Scoring Metric (Hamming distance between ground truth & estimated state sequence)

Experimental Results on ESTER



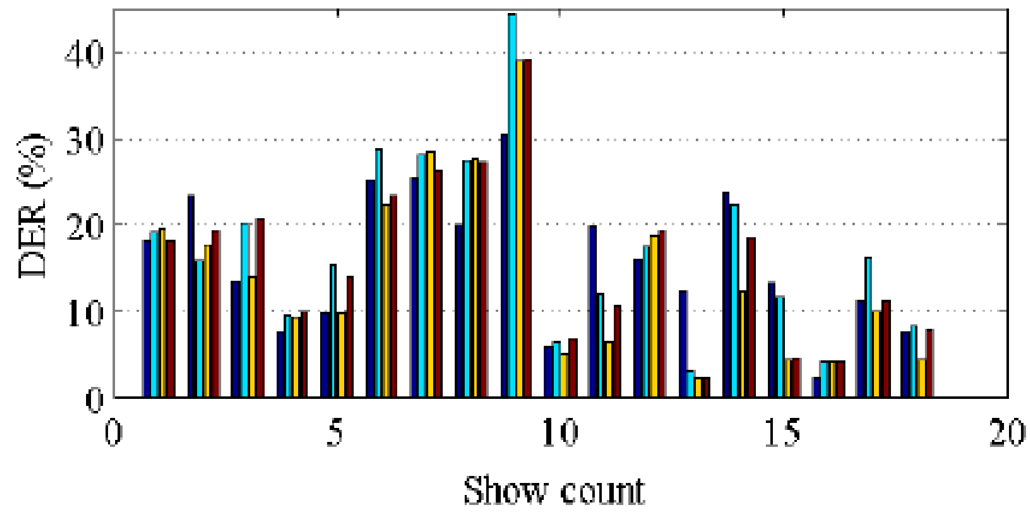


Table 1: Overall Speaker Diarization Error Rate (%) on ESTER

	ESTER-DEV	ESTER-TEST
HC Local-BIC	15.76	16.28
MS summed KLD	18.78	17.77
MS Harmonic mean KLD	14.88	15.49
MS asymmetric KLD	16.55	16.83
False Alarm Rate	0.3	0.6
Missed Speech Rate	0.9	1.2

Note: DER with Hierarchical Clustering using KL-Divergence >30%

Conclusions

What we proposed...

- We proposed an adaptation of the mean-shift algorithm to the parameter space.
- We showed how to deal with a higher level in the Bayesian hierarchy.
- We derived a rich family of kernels, including both shape and distance.
- We showed that for the exponential family, no heuristic is involved.

When they are relevant...

- All these approaches lead to point-estimates.
- Use them when a point-estimate is sufficient.
- Avoid them when not dealing with real-time apps.
- Avoid them in large scale problems, or when combining information streams

More complex models than Gaussians?

- GMMs: They belong to the exponential family only if the complete data likelihood is considered. Use UBM. Get the memberships from the last E-step.
- I-vectors: Express the uncertainty in estimating them by a Gaussian, a t-distribution and it may work.

Thanx for your attention!

Apologies for the maths!

For any question, suggestion, collaboration,

themosst@ilsp.athena-innovation.gr

Q&A