Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms

Frank Nielsen
Frank.Nielsen@acm.org
5793b870

Sony Computer Science Laboratories, Inc.

April 2013

Why histogram clustering?

Task: Classify documents into categories:
Bag-of-Word (BoW) modeling paradigm [3, 6].

- Define a word dictionary, and
- Represent each document by a *word count* histogram.

Centroid-based k-means clustering [1]:

- Cluster document histograms to learn categories,
- Build visual vocabularies by quantizing image features: Compressed Histogram of Gradient descriptors [4].

$\rightarrow$ *histogram centroids*

$w_h = \sum_{i=1}^{d} h^i$: cumulative sum of bin values
$\tilde{}$: normalization operator

Why Jeffreys divergence?

Distance between two frequency histograms $\tilde{p}$ and $\tilde{q}$:
Kullback-Leibler divergence or *relative entropy*.

$$\begin{aligned}
\mathrm{KL}(\tilde{p}:\tilde{q}) &= H^{\times}(\tilde{p}:\tilde{q}) - H(\tilde{p}), \\
H^{\times}(\tilde{p}:\tilde{q}) &= \sum_{i=1}^{d} \tilde{p}^i \log \frac{1}{\tilde{q}^i}, \mathrm{cross-entropy} \\
H(\tilde{p}) &= H^{\times}(\tilde{p}:\tilde{p}) = \sum_{i=1}^{d} \tilde{p}^i \log \frac{1}{\tilde{p}^i}, \mathrm{Shannon\ entropy}.
\end{aligned}$$

$\rightarrow$ expected extra number of bits per datum that must be transmitted when using the "wrong" distribution $\tilde{q}$ instead of the true distribution $\tilde{p}$.
$\tilde{p}$ is hidden by nature (and hypothesized), $\tilde{q}$ is estimated.

Why Jeffreys divergence?

When clustering histograms, all histograms play the *same role* $\rightarrow$ Jeffreys [8] divergence:

$$\begin{aligned} J(p,q) &= \mathrm{KL}(p:q) + \mathrm{KL}(q:p), \\ J(p,q) &= \sum_{i=1}^{d} (p^i - q^i) \log \frac{p^i}{q^i} = J(q,p). \end{aligned}$$

$\rightarrow$ symmetrizes the KL divergence.
(also called J-divergence or symmetrical Kullback-Leibler divergence, etc.)

Jeffreys centroids: frequency and positive centroids

A set $\mathcal{H} = \{h_1, ..., h_n\}$ of *weighted histograms*.

$$c = \arg\min_x \sum_{j=1}^{n} \pi_j J(h_j, x),$$

$\pi_j > 0$'s histogram positive weights: $\sum_{j=1}^{n} \pi_j = 1$.

- Jeffreys positive centroid c:

$$c = \arg\min_{x \in \mathbb{R}_+^d} \sum_{j=1}^{n} \pi_j J(h_j, x),$$

- Jeffreys frequency centroid $\tilde{c}$:

$$\tilde{c} = \arg\min_{x \in \Delta_d} \sum_{j=1}^{n} \pi_j J(\tilde{h}_j, x),$$

Δ_d: Probability $(d-1)$-dimensional simplex.

Prior work

- Histogram clustering wrt. χ^2 distance [10]
- Histogram clustering wrt. Bhattacharyya distance [11, 13]
- Histogram clustering wrt. Kullback-Leibler distance as Bregman k-means clustering [1]
- Jeffreys frequency centroid [16] (Newton numerical optimization)
- Jeffreys frequency centroid as equivalent symmetrized Bregmen centroid [14]
- Mixed Bregman clustering [15]
- Smooth family of KL symmetrized centroids including Jensen-Shannon centroids and Jeffreys centroids in *limit* case [12]

Jeffreys positive centroid

$$c = \arg\min_{x \in \mathbb{R}^d_+} J(\mathcal{H}, x) = \arg\min_{x \in \mathbb{R}^d_+} \sum_{j=1}^{n} \pi_j J(h_j, x).$$

Theorem (Theorem 1)

The Jeffreys positive centroid $c = (c^1, ..., c^d)$ *of a set* $\{h_1, ..., h_n\}$ *of* n *weighted positive histograms with* d *bins can be calculated component-wise* exactly *using the Lambert* W *analytic function:*

$$c^i = \frac{a^i}{W(\frac{a^i}{g^i} e)},$$

where $a^i = \sum_{j=1}^{n} \pi_j h_j^i$ *denotes the coordinate-wise arithmetic weighted means and* $g^i = \prod_{j=1}^{n} (h_j^i)^{\pi_j}$ *the coordinate-wise geometric weighted means.*

Lambert analytic function [2] $W(x)e^{W(x)} = x$ for $x \geq 0$.

Jeffreys positive centroid (proof)

$$\min_x \quad \sum_{j=1}^{n} \pi_j J(h_j, x)$$

$$\min_x \quad \sum_{j=1}^{n} \pi_j \sum_{i=1}^{d} (h_j^i - x^i)(\log h_j^i - \log x^i)$$

$$\equiv \min_x \quad \sum_{i=1}^{d} \sum_{j=1}^{n} \pi_j (x^i \log x^i - x^i \log h_j^i - h_j^i \log x^i)$$

$$\sum_{i=1}^{d} x^i \log x^i - x^i \log \underbrace{\prod_{j=1}^{n} (h_j^i)^{\pi_j}}_{g} - \underbrace{\sum_{j=1}^{n} \pi_j h_j^i}_{a} \log x^i$$

$$\min_x \sum_{i=1}^{d} x^i \log \frac{x^i}{g} - a \log x^i$$

Jeffreys positive centroid (proof)

Coordinate-wise minimize:

$$\min_x x \log \frac{x}{g} - a \log x$$

Setting the derivative to zero, we solve:

$$\log \frac{x}{g} + 1 - \frac{a}{x} = 0$$

and get

$$x = \frac{a}{W(\frac{a}{g}e)}$$

Jeffreys frequency centroid: A guaranteed approximation

$$\tilde{c} = \arg\min_{x \in \Delta_d} \sum_{j=1}^{n} \pi_j J(\tilde{h}_j, x),$$

Relaxing x from probability simplex Δ_d to $\mathbb{R}_+^d$, we get

$$\tilde{c}' = \frac{c}{w_c}, c^i = \frac{a^i}{W(\frac{a^i}{g^i} e)}, w_c = \sum_i c^i$$

Lemma (Lemma 1)

The cumulative sum w_c of the bin values of the Jeffreys positive centroid c of a set of frequency histograms is less or equal to one:

$$0 < w_c \leq 1.$$

Proof of Lemma 1

From Theorem 1:

$$w_c = \sum_{i=1}^{d} c^i = \sum_{i=1}^{d} \frac{a^i}{W(\frac{a^i}{g^i} e)}.$$

Arithmetic-geometric mean inequality: $a^i \geq g^i$
Therefore $W(\frac{a^i}{g^i} e) \geq 1$ and $c^i \leq a^i$. Thus

$$w_c = \sum_{i=1}^{d} c^i \leq \sum_{i=1}^{d} a^i = 1$$

Lemma 2

Lemma (Lemma 2)

For any histogram x and frequency histogram $\tilde{h}$, we have $J(x, \tilde{h}) = J(\tilde{x}, \tilde{h}) + (w_x - 1)(\mathrm{KL}(\tilde{x} : \tilde{h}) + \log w_x)$, where w_x denotes the normalization factor ($w_x = \sum_{i=1}^{d} x^i$).

$$J(x, \tilde{H}) = J(\tilde{x}, \tilde{H}) + (w_x - 1)(\mathrm{KL}(\tilde{x} : \tilde{H}) + \log w_x),$$

where $J(x, \tilde{H}) = \sum_{j=1}^{n} \pi_j J(x, \tilde{h}_j)$ and
$\mathrm{KL}(\tilde{x} : \tilde{H}) = \sum_{j=1}^{n} \pi_j \mathrm{KL}(\tilde{x}, \tilde{h}_j)$ (with $\sum_{j=1}^{n} \pi_j = 1$).

Proof of Lemma 2

$$x^i = w_x \tilde{x}^i$$

$$J(x, \tilde{h}) = \sum_{i=1}^{d} (w_x \tilde{x}^i - \tilde{h}^i) \log \frac{w_x \tilde{x}^i}{\tilde{h}^i}$$

$$\begin{aligned}
J(x, \tilde{h}) &= \sum_{i=1}^{d} (w_x \tilde{x}^i \log \frac{\tilde{x}^i}{\tilde{h}^i} + w_x \tilde{x}^i \log w_x + \tilde{h}^i \log \frac{\tilde{h}^i}{\tilde{x}^i} - \tilde{h}^i \log w_x) \\
&= (w_x - 1) \log w_x + J(\tilde{x}, \tilde{h}) + (w_x - 1) \sum_{i=1}^{d} \tilde{x}^i \log \frac{\tilde{x}^i}{\tilde{h}^i} \\
&= J(\tilde{x}, \tilde{h}) + (w_x - 1)(\mathrm{KL}(\tilde{x} : \tilde{h}) + \log w_x)
\end{aligned}$$

since $\sum_{i=1}^{d} \tilde{h}^i = \sum_{i=1}^{d} \tilde{x}^i = 1$.

Guaranteed approximation of $\tilde{c}$

Theorem (Theorem 2)

Let $\tilde{c}$ denote the Jeffreys frequency centroid and $\tilde{c}' = \frac{c}{w_c}$ the normalized Jeffreys positive centroid. Then the approximation factor $\alpha_{\tilde{c}'} = \frac{J(\tilde{c}', \tilde{H})}{J(\tilde{c}, \tilde{H})}$ is such that $1 \leq \alpha_{\tilde{c}'} \leq \frac{1}{w_c}$ (with $w_c \leq 1$).

Proof of Theorem 2

$$J(c, \tilde{H}) \leq J(\tilde{c}, \tilde{H}) \leq J(\tilde{c}', \tilde{H})$$

From Lemma 2, since
$J(\tilde{c}', \tilde{H}) = J(c, \tilde{H}) + (1 - w_c)(\mathrm{KL}(\tilde{c}', \tilde{H}) + \log w_c))$ and
$J(c, \tilde{H}) \leq J(\tilde{c}, \tilde{H})$

$$1 \leq \alpha_{\tilde{c}'} \leq 1 + \frac{(1 - w_c)(\mathrm{KL}(\tilde{c}', \tilde{H}) + \log w_c)}{J(\tilde{c}, \tilde{H})}$$

$$\mathrm{KL}(\tilde{c}' : \tilde{H}) = \frac{1}{w_c}\mathrm{KL}(c, \tilde{H}) - \log w_c$$

$$\alpha_{\tilde{c}'} \leq 1 + \frac{(1 - w_c)\mathrm{KL}(c, \tilde{H})}{w_c J(\tilde{c}, \tilde{H})}$$

Since $J(\tilde{c}, \tilde{H}) \geq J(c, \tilde{H})$ and $\mathrm{KL}(c, \tilde{H}) \leq J(c, \tilde{H})$, we get
$\alpha_{\tilde{c}'} \leq \frac{1}{w_c}$.
When $w_c = 1$ the bound is tight.

In practice...

c in closed-form $\rightarrow$ compute w_c, $\mathrm{KL}(c, \tilde{H})$, $J(c, \tilde{H})$.
Bound the approximation factor $\alpha_{\tilde{c}'}$ as:

$$\alpha_{\tilde{c}'} \leq 1 + \left(\frac{1}{w_c} - 1 \right) \frac{\mathrm{KL}(c, \tilde{H})}{J(c, \tilde{H})} \leq \frac{1}{w_c}$$

Fine approximation

From [16, 14], minimization of Jeffreys frequency centroid equivalent to:

$$\tilde{c} = \arg\min_{\tilde{x} \in \Delta_d} \mathrm{KL}(\tilde{a} : \tilde{x}) + \mathrm{KL}(\tilde{x} : \tilde{g})$$

Lagrangian function enforcing $\sum_i c^i = 1$:

$$\log \frac{\tilde{c}^i}{\tilde{g}^i} + 1 - \frac{\tilde{a}^i}{\tilde{c}^i} + \lambda = 0$$

$$\tilde{c}^i = \frac{\tilde{a}^i}{W\left(\frac{\tilde{a}^i e^{\lambda+1}}{\tilde{g}^i}\right)}$$

$\lambda = -\mathrm{KL}(\tilde{c} : \tilde{g}) \leq 0$

Fine approximation: Bisection search

$$c^i \leq 1 \Rightarrow c^i = \frac{\tilde{a}^i}{W\left(\frac{\tilde{a}^i e^{\lambda+1}}{\tilde{g}^i}\right)} \leq 1$$

$$\lambda \geq \log(e^{\tilde{a}^i}\tilde{g}^i) - 1 \forall i, \quad \lambda \in [\max_i \log(e^{\tilde{a}^i}\tilde{g}^i) - 1, 0]$$

$$s(\lambda) = \sum_i c^i(\lambda) = \sum_{i=1}^{d} \frac{\tilde{a}^i}{W\left(\frac{\tilde{a}^i e^{\lambda+1}}{\tilde{g}^i}\right)}$$

Function s: monotonously decreasing with $s(0) \leq 1$.
$\rightarrow$ Bisection search for $s(\lambda^*) \simeq 1$ for arbitrary precision.

Experiments: Caltech-256

Caltech-256 [7]: 30607 images labeled into 256 categories (256 Jeffreys centroids).
Arbitrary floating-point precision: `http://www.apfloat.org/`

$$\tilde{c}'' = \frac{\tilde{a} + \tilde{g}}{2}$$

	α_c(optimal positive)	$\alpha_{\tilde{c}'}$(n$'$lized approx.)	$w_c \leq 1$(n$'$lizing coeff.t)	$\alpha_{\tilde{c}''}$(Veldhuis' approx.)
avg	0.9648680345638155	**1.000220508096425**5	0.9338228644308926	1.065590178484613
min	0.906414219584823	**1.000000507952880**9	0.8342819488534723	1.002770738209519**5**
max	0.995639922067858**5**	**1.000003148954177**2	0.9931975105809021	1.358229667539775**4**

Experiments: Synthetic data-sets

Random binary histograms

$$\alpha = \frac{J(\tilde{c}')}{J(\tilde{c})} \geq 1$$

Performance:
$\bar{\alpha} \sim 1.0000009, \alpha_{\max} \sim 1.00181506, \alpha_{\min} = 1.000000.$

Express better worst-case upper bound performance?

Summary and conclusion

- Jeffreys positive centroid c in closed-form
- normalized Jeffreys positive centroid $\tilde{c}'$ within approximation factor $\frac{1}{w_c}$
- Bisection search for arbitrary fine approximation of $\tilde{c}$.

$\rightarrow$ Variational Jeffreys k-means clustering

Other Kullback-Leibler symmetrizations:

- Jensen-Shannon divergence [9]
- Chernoff divergence [5]
- Family of symmetrized centroids including Jensen-Shannon and Jeffreys centroids [12]

Thank you!

http://www.informationgeometry.org

```
@Article{JeffreysCentroid-2013,
  author = {Frank Nielsen},
  title = {Jeffreys centroids: {A} closed-form expression for positive histograms
                   and a guaranteed tight approximation for frequency histograms},
  journal = {IEEE Signal Processing Letters (SPL)},
  year = {2013}
}
```

Bibliographic references I

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh.

Clustering with Bregman divergences.

Journal of Machine Learning Research, 6:1705–1749, 2005.

D. A. Barry, P. J. Culligan-Hensley, and S. J. Barry.

Real values of the W-function.

ACM Trans. Math. Softw., 21(2):161–171, June 1995.

Brigitte Bigi.

Using Kullback-Leibler distance for text categorization.

In *Proceedings of the 25th European conference on IR research (ECIR)*, ECIR'03, pages 305–319, Berlin, Heidelberg, 2003. Springer-Verlag.

Vijay Chandrasekhar, Gabriel Takacs, David M. Chen, Sam S. Tsai, Yuriy A. Reznik, Radek Grzeszczuk, and Bernd Girod.

Compressed histogram of gradients: A low-bitrate descriptor.

International Journal of Computer Vision, 96(3):384–399, 2012.

Herman Chernoff.

A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations.

Annals of Mathematical Statistics, 23:493–507, 1952.

Bibliographic references II

G. Csurka, C. Bray, C. Dance, and L. Fan.

Visual categorization with bags of keypoints.

Workshop on Statistical Learning in Computer Vision (ECCV), pages 1–22, 2004.

G. Griffin, A. Holub, and P. Perona.

Caltech-256 object category dataset.

Technical Report 7694, California Institute of Technology, 2007.

Harold Jeffreys.

An invariant form for the prior probability in estimation problems.

Proceedings of the Royal Society of London, 186(1007):453–461, March 1946.

Jianhua Lin.

Divergence measures based on the Shannon entropy.

IEEE Transactions on Information Theory, 37:145–151, 1991.

Huan Liu and Rudy Setiono.

Chi2: Feature selection and discretization of numeric attributes.

In *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence (TAI)*, pages 88–, Washington, DC, USA, 1995. IEEE Computer Society.

Max Mignotte.

Segmentation by fusion of histogram-based k-means clusters in different color spaces.

IEEE Transactions on Image Processing (TIP), 17(5):780–787, 2008.

Bibliographic references III

Frank Nielsen.

A family of statistical symmetric divergences based on Jensen's inequality.

CoRR, abs/1009.4004, 2010.

Frank Nielsen and Sylvain Boltz.

The Burbea-Rao and Bhattacharyya centroids.

IEEE Transactions on Information Theory, 57(8):5455–5466, August 2011.

Frank Nielsen and Richard Nock.

Sided and symmetrized Bregman centroids.

IEEE Transactions on Information Theory, 55(6):2048–2059, June 2009.

Richard Nock, Panu Luosto, and Jyrki Kivinen.

Mixed Bregman clustering with approximation guarantees.

In *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases*, pages 154–169, Berlin, Heidelberg, 2008. Springer-Verlag.

Raymond N. J. Veldhuis.

The centroid of the symmetrical Kullback-Leibler distance.

IEEE signal processing letters, 9(3):96–99, March 2002.