Hypothesis testing, information divergence and computational geometry

Frank Nielsen
Frank.Nielsen@acm.org
www.informationgeometry.org

Sony Computer Science Laboratories, Inc.

August 2013, GSI, Paris, FR

The Multiple Hypothesis Testing (MHT) problem

Given a rv. X with n hypothesis $H_1 : X \sim P_1$, ..., $H_n : X \sim P_n$, decide for a IID sample $x_1, ..., x_m \sim X$ which hypothesis holds true?

$$P^m_{\text{correct}} = 1 - P^m_{\text{error}}$$

Asymptotic regime:

$$\alpha = -\frac{1}{m} \log P^m_e, \quad m \to \infty$$

Bayesian hypothesis testing (preliminaries)

prior probabilities: $w_i = \Pr(X \sim P_i) > 0$ (with $\sum_{i=1}^{n} w_i = 1$)
conditional probabilities: $\Pr(X = x | X \sim P_i)$.

$$\begin{aligned} \Pr(X = x) &= \sum_{i=1}^{n} \Pr(X \sim P_i) \Pr(X = x | X \sim P_i) \\ &= \sum_{i=1}^{n} w_i \Pr(X | P_i) \end{aligned}$$

Let $c_{i,j}$ = cost of deciding H_i when in fact H_j is true.
Matrix $[c_{ij}]$= cost *design* matrix
Let $p_{i,j}(u)$ = probability of making this decision using rule u.

Bayesian detector

Minimize the *expected cost*:

$$E_X[c(r(x))], \quad c(r(x)) = \sum_i \left(w_i \sum_{j \neq i} c_{i,j} p_{i,j}(r(x)) \right)$$

Special case: Probability of error P_e obtained for $c_{i,i} = 0$ and $c_{i,j} = 1$ for $i \neq j$:

$$\boxed{P_e = E_X \left[\sum_i \left(w_i \sum_{j \neq i} p_{i,j}(r(x)) \right) \right]}$$

The maximum *a posteriori* probability (MAP) rule considers classifying x:

$$\boxed{\mathrm{MAP}(x) = \mathrm{argmax}_{i \in \{1,\ldots,n\}} \; w_i p_i(x)}$$

where $p_i(x) = \Pr(X = x | X \sim P_i)$ are the conditional probabilities.

$\rightarrow$ MAP Bayesian detector minimizes P_e over all rules [8]

Probability of error and divergences

Without loss of generality, consider equal priors ($w_1 = w_2 = \frac{1}{2}$):

$$P_e = \int_{x \in \mathcal{X}} p(x) \min(\Pr(H_1|x), \Pr(H_2|x)) \mathrm{d}\nu(x)$$

($P_e > 0$ as soon as $\mathrm{supp} p_1 \cap \mathrm{supp} p_2 \neq \emptyset$)

From Bayes' rule $\Pr(H_i|X = x) = \frac{\Pr(H_i)\Pr(X=x|H_i)}{\Pr(X=x)} = w_i p_i(x)/p(x)$

$$\boxed{P_e = \frac{1}{2} \int_{x \in \mathcal{X}} \min(p_1(x), p_2(x)) \mathrm{d}\nu(x)}$$

Rewrite or bound P_e using tricks of the trade:

Trick 1. $\forall a, b \in \mathbb{R},\ \min(a, b) = \frac{a+b}{2} - \frac{|a-b|}{2}$,

Trick 2. $\forall a, b > 0,\ \min(a, b) \leq \min_{\alpha \in (0,1)} a^\alpha b^{1-\alpha}$,

Probability of error and total variation

$$
\begin{aligned}
P_e &= \frac{1}{2}\int_{x\in\mathcal{X}} \left(\frac{p_1(x)+p_2(x)}{2} - \frac{|p_1(x)-p_2(x)|}{2}\right) \mathrm{d}\nu(x), \\
&= \frac{1}{2}\left(1 - \frac{1}{2}\int_{x\in\mathcal{X}} |p_1(x)-p_2(x)|\mathrm{d}\nu(x)\right)
\end{aligned}
$$

$$
\boxed{P_e = \frac{1}{2}(1 - \mathrm{TV}(P_1, P_2))}
$$

total variation metric distance:

$$
\boxed{\mathrm{TV}(P,Q) = \frac{1}{2}\int_{x\in\mathcal{X}} |p(x)-q(x)|\mathrm{d}\nu(x)}
$$

$\rightarrow$ Difficult to compute when handling multivariate distributions.

Bounding the Probability of error P_e

$\min(a,b) \leq \min_{\alpha\in(0,1)} a^\alpha b^{1-\alpha}$ for $a,b>0$, *upper bound* P_e:

$$\begin{aligned} P_e &= \frac{1}{2}\int_{x\in\mathcal{X}} \min(p_1(x),p_2(x))\mathrm{d}\nu(x) \\ &\leq \frac{1}{2}\min_{\alpha\in(0,1)}\int_{x\in\mathcal{X}} p_1^\alpha(x)p_2^{1-\alpha}(x)\mathrm{d}\nu(x). \end{aligned}$$

$$C(P_1,P_2) = -\log \min_{\alpha\in(0,1)}\int_{x\in\mathcal{X}} p_1^\alpha(x)p_2^{1-\alpha}(x)\mathrm{d}\nu(x) \geq 0,$$

Best error exponent α^* [7]:

$$P_e \leq w_1^{\alpha^*} w_2^{1-\alpha^*} e^{-C(P_1,P_2)} \leq e^{-C(P_1,P_2)}$$

Bounding technique can be extended using any quasi-arithmetic α-means [13, 9]...

Computational information geometry

Exponential family manifold [4]:

$$\mathcal{M} = \{p_\theta \mid p_\theta(x) = \exp(t(x)^\top \theta - F(\theta))\}$$

Dually flat manifolds [1] enjoy dual affine connections [1]: $(\mathcal{M}, \nabla^2 F(\theta), \nabla^{(e)}, \nabla^{(m)})$.

$$\eta = \nabla F(\theta), \quad \theta = \nabla F^*(\eta)$$

Canonical divergence from Young inequality:

$$A(\theta_1, \eta_2) = F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2 \geq 0$$

$$F(\theta) + F^*(\eta) = \theta^\top \eta$$

MAP decision rule and additive Bregman Voronoi diagrams

$$\boxed{\mathrm{KL}(p_{\theta_1} : p_{\theta_2}) = B(\theta_2 : \theta_1) = A(\theta_2 : \eta_1) = A^*(\eta_1 : \theta_2) = B^*(\eta_1 : \eta_2)}$$

Canonical divergence (mixed primal/dual coordinates):

$$A(\theta_2 : \eta_1) = F(\theta_2) + F^*(\eta_1) - \theta_2^\top \eta_1 \geq 0$$

Bregman divergence (uni-coordinates, primal or dual):

$$B(\theta_2 : \theta_1) = F(\theta_2) - F(\theta_1) - (\theta_2 - \theta_1)^\top \nabla F(\theta_1)$$

$$\boxed{\log p_i(x) = -B^*(t(x) : \eta_i) + F^*(t(x)) + k(x), \quad \eta_i = \nabla F(\theta_i) = \eta(P_{\theta_i})}$$

Optimal MAP decision rule:

$$\begin{aligned} \mathrm{MAP}(x) &= \mathrm{argmax}_{i \in \{1,\ldots,n\}} w_i p_i(x) \\ &= \mathrm{argmax}_{i \in \{1,\ldots,n\}} - B^*(t(x) : \eta_i) + \log w_i, \\ &= \mathrm{argmin}_{i \in \{1,\ldots,n\}} B^*(t(x) : \eta_i) - \log w_i \end{aligned}$$

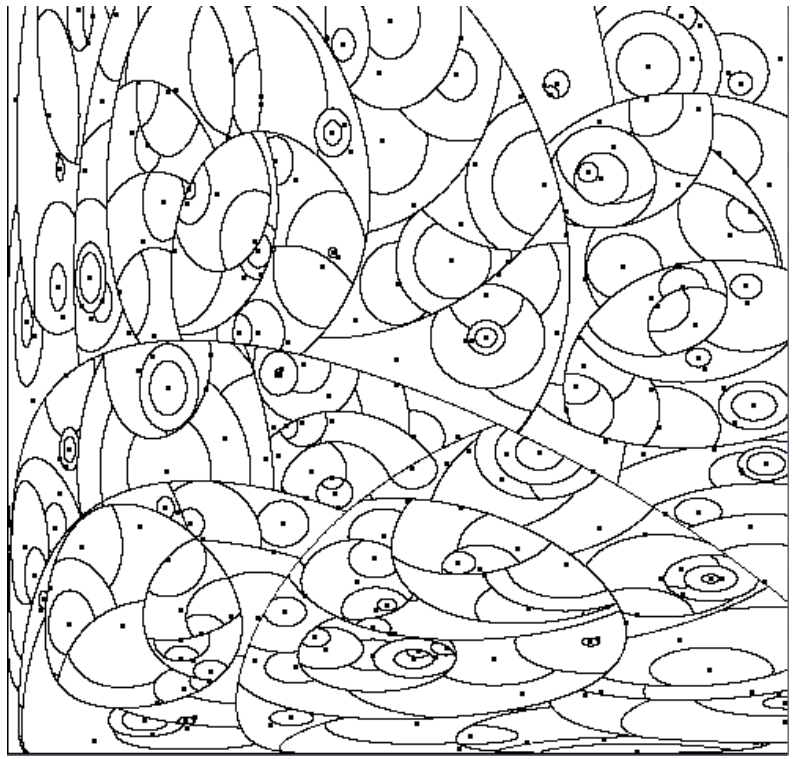
$\rightarrow$ *nearest neighbor classifier* [2, 10, 15, 16]

MAP & nearest neighbor classifier

Bregman Voronoi diagrams (with additive weights) are affine diagrams [2].

$$\mathrm{argmin}_{i\in\{1,\ldots,n\}} B^*(t(x) : \eta_i) - \log w_i$$

- point location in arrangement [3] (small dims),
- Divergence-based search trees [16],
- GPU brute force [6].



Geometry of the best error exponent: binary hypothesis

On the exponential family manifold, Chernoff α-coefficient [5]:

$$c_\alpha(P_{\theta_1} : P_{\theta_2}) = \int p_{\theta_1}^\alpha(x) p_{\theta_2}^{1-\alpha}(x) \mathrm{d}\mu(x) = \exp(-J_F^{(\alpha)}(\theta_1 : \theta_2)),$$

Skew Jensen divergence [14] on the natural parameters:

$$J_F^{(\alpha)}(\theta_1 : \theta_2) = \alpha F(\theta_1) + (1-\alpha) F(\theta_2) - F(\theta_{12}^{(\alpha)}),$$

Chernoff information = Bregman divergence for exponential families:

$$\boxed{C(P_{\theta_1} : P_{\theta_2}) = B(\theta_1 : \theta_{12}^{(\alpha^*)}) = B(\theta_2 : \theta_{12}^{(\alpha^*)})}$$

Geometry of the best error exponent: binary hypothesis

Chernoff distribution P^* [12]:

$$\boxed{P^* = P_{\theta^*_{12}} = G_e(P_1, P_2) \cap \mathrm{Bi}_m(P_1, P_2)}$$

e-geodesic:

$$G_e(P_1, P_2) = \{E_{12}^{(\lambda)} \mid \theta(E_{12}^{(\lambda)}) = (1-\lambda)\theta_1 + \lambda\theta_2, \lambda \in [0,1]\},$$

m-bisector:

$$\mathrm{Bi}_m(P_1, P_2) : \{P \mid F(\theta_1) - F(\theta_2) + \eta(P)^\top \Delta\theta = 0\},$$

Optimal natural parameter of P^*:

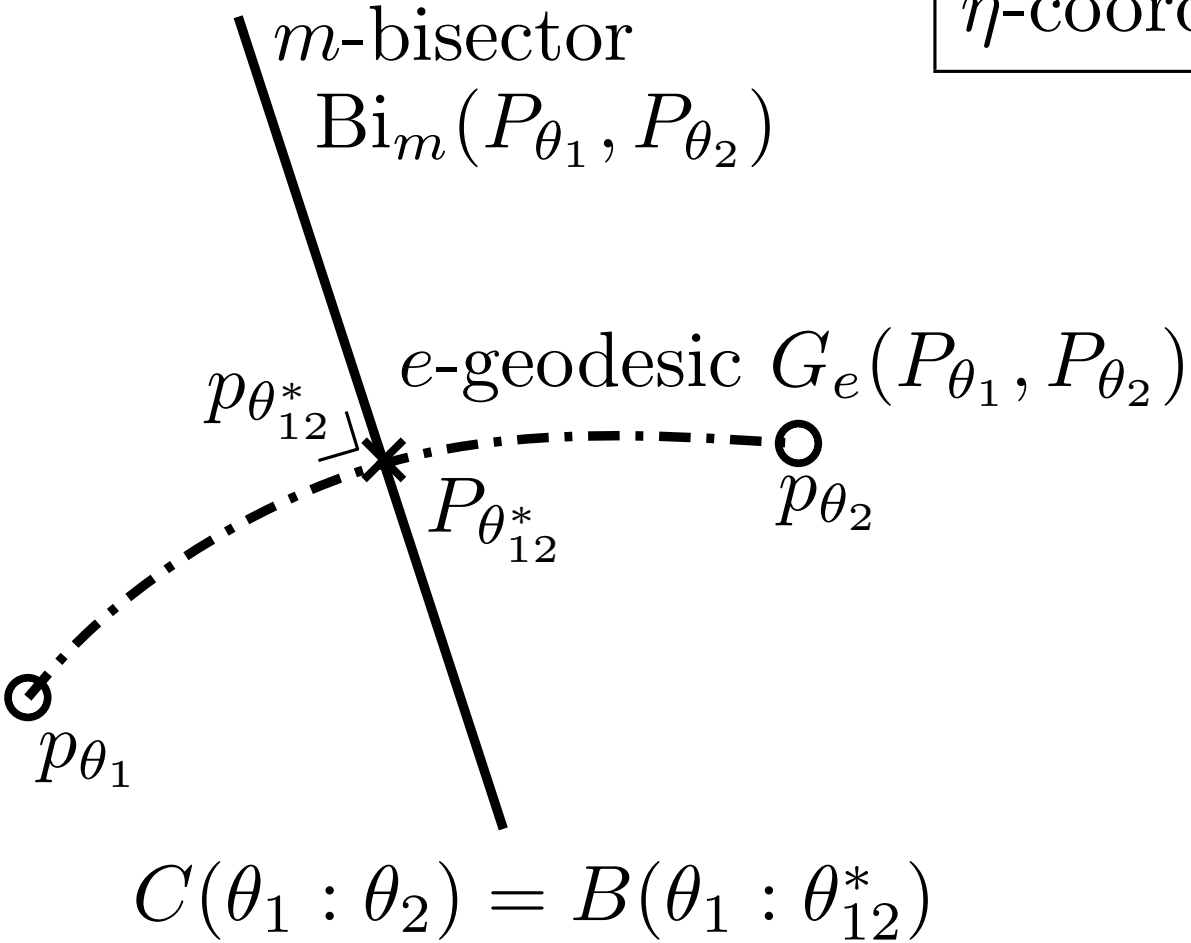
$$\theta^* = \theta_{12}^{(\alpha^*)} = \mathrm{argmin}_{\theta\in\Theta} B(\theta_1 : \theta) = \mathrm{argmin}_{\theta\in\Theta} B(\theta_2 : \theta).$$

$\rightarrow$ closed-form for order-1 family, or efficient bisection search.

Geometry of the best error exponent: binary hypothesis

$$P^* = P_{\theta^*_{12}} = G_e(P_1, P_2) \cap \mathrm{Bi}_m(P_1, P_2)$$

η-coordinate system



$$C(\theta_1 : \theta_2) = B(\theta_1 : \theta^*_{12})$$

Geometry of the best error exponent: multiple hypothesis

n-ary MHT [8] from *minimum pairwise Chernoff distance*:

$$\boxed{C(P_1,...,P_n) = \min_{i,j\neq i} C(P_i,P_j)}$$

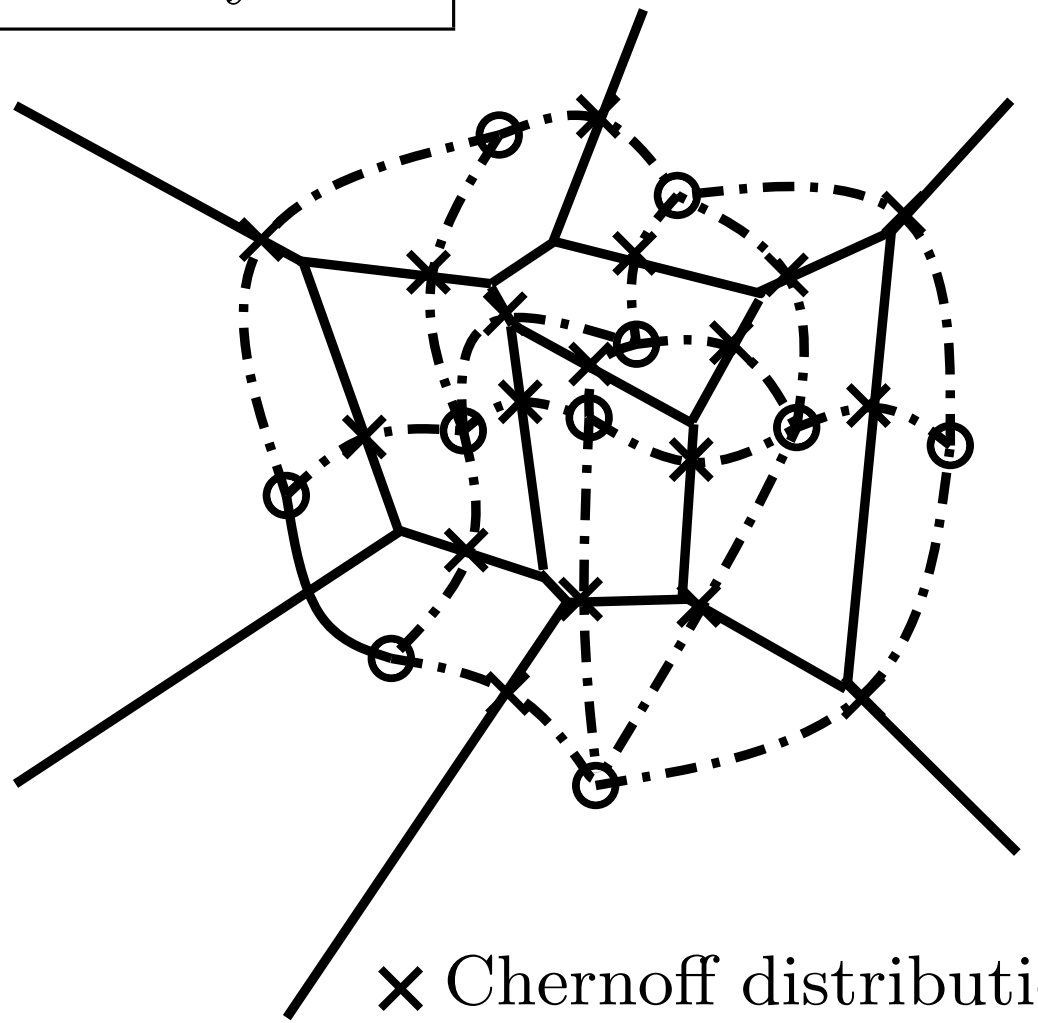
$$P_e^m \leq e^{-mC(P_{i^*},P_{j^*})}, \quad (i^*,j^*) = \mathrm{argmin}_{i,j\neq i} C(P_i,P_j)$$

Compute for each pair of natural neighbors [3] P_{θ_i} and P_{θ_j}, the Chernoff distance $C(P_{\theta_i},P_{\theta_j})$, and choose the pair with minimal distance. (Proof by contradiction using Bregman Pythagoras theorem.)

$\rightarrow$ Closest Bregman pair problem (Chernoff distance fails triangle inequality).

Hypothesis testing: Illustration

η-coordinate system



× Chernoff distribution between natural neighbours

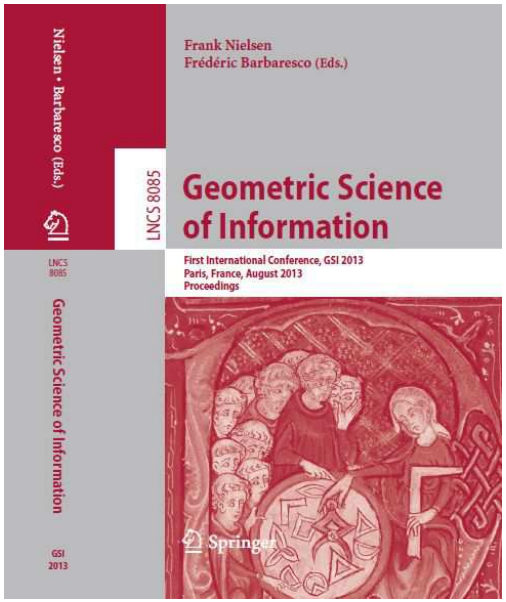
Summary

Bayesian multiple hypothesis testing...
... from the viewpoint of computational geometry.

- probability of error & best MAP Bayesian rule
- total variation & P_e, upper-bounded by the Chernoff distance.
- Exponential family manifolds:
  - MAP rule = NN classifier (additive Bregman Voronoi diagram)
  - best error exponent from intersection geodesic/bisector for binary hypothesis,
  - best error exponent from closest Bregman pair for multiple hypothesis.

Thank you

28th-30th August, Paris.




```
@incollection{HTIGCG-GSI-2013,
year={2013},
booktitle={Geometric Science of Information},
volume={8085},
series={Lecture Notes in Computer Science},
editor={Frank Nielsen and Fr\'ed\'eric Barbaresco},
title={Hypothesis testing, information divergence and computational geometry},
publisher={Springer Berlin Heidelberg},
author={Nielsen, Frank},
pages={241-248}
}
```

Bibliographic references I

Shun-ichi Amari and Hiroshi Nagaoka.

Methods of Information Geometry.

Oxford University Press, 2000.

Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock.

Bregman Voronoi diagrams.

Discrete & Computational Geometry, 44(2):281–307, 2010.

Jean-Daniel Boissonnat and Mariette Yvinec.

Algorithmic Geometry.

Cambridge University Press, New York, NY, USA, 1998.

Lawrence D. Brown.

Fundamentals of statistical exponential families: with applications in statistical decision theory.

Institute of Mathematical Statistics, Hayworth, CA, USA, 1986.

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.

Vincent Garcia, Eric Debreuve, Frank Nielsen, and Michel Barlaud.

k-nearest neighbor search: Fast GPU-based implementations and application to high-dimensional feature matching.

In *IEEE International Conference on Image Processing (ICIP)*, pages 3757–3760, 2010.

Bibliographic references II

Martin E. Hellman and Josef Raviv.

Probability of error, equivocation and the Chernoff bound.

IEEE Transactions on Information Theory, 16:368–372, 1970.

C. C. Leang and D. H. Johnson.

On the asymptotics of M-hypothesis Bayesian detection.

IEEE Transactions on Information Theory, 43(1):280–282, January 1997.

Frank Nielsen.

Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means.

submitted, 2012.

Frank Nielsen.

k-MLE: A fast algorithm for learning statistical mixture models.

In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2012.

preliminary, technical report on arXiv.

Frank Nielsen.

Hypothesis testing, information divergence and computational geometry.

In Frank Nielsen and Frédéric Barbaresco, editors, *Geometric Science of Information*, volume 8085 of *Lecture Notes in Computer Science*, pages 241–248. Springer Berlin Heidelberg, 2013.

Bibliographic references III

Frank Nielsen.

An information-geometric characterization of Chernoff information.

IEEE Signal Processing Letters (SPL), 20(3):269–272, March 2013.

Frank Nielsen.

Pattern learning and recognition on statistical manifolds: An information-geometric review.

In Edwin Hancock and Marcello Pelillo, editors, *Similarity-Based Pattern Recognition*, volume 7953 of *Lecture Notes in Computer Science*, pages 1–25. Springer Berlin Heidelberg, 2013.

Frank Nielsen and Sylvain Boltz.

The Burbea-Rao and Bhattacharyya centroids.

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

Frank Nielsen, Paolo Piro, and Michel Barlaud.

Bregman vantage point trees for efficient nearest neighbor queries.

In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo (ICME)*, pages 878–881, 2009.

Paolo Piro, Frank Nielsen, and Michel Barlaud.

Tailored Bregman ball trees for effective nearest neighbors.

In *European Workshop on Computational Geometry (EuroCG)*, LORIA, Nancy, France, March 2009. IEEE.