The entropic centers of multivariate normal distributions

Frank Nielsen* Richard Nock†

Abstract

In this paper, we seek for a single best representative of a set of statistical multivariate normal distributions. To define the "best" center, we consider either minimizing the average or the maximum relative entropy of the center to the given set of normal distributions. Since the relative entropy is an asymmetric divergence, this yields the notion of left- and right-sided, and symmetrized entropic centroids and circumcenters along with their respective information radii. We show how to instantiate and implement for this special case of multivariate normals very recent work that tackled the broader case of finding centers of point sets with respect to Bregman divergences.

1 Information-theoretic centers

Consider a set of n multivariate normal distributions $\mathcal{D} = \{N(\mu_1, \Sigma_1), ..., N(\mu_n, \Sigma_n)\}$ with $\mu_i \in \mathbb{R}^d$ denoting the mean vector and Σ_i the $d \times d$ symmetric positive semi-definite variance-covariance matrix (i.e., $x^T \Sigma_i x \geq 0$ for all $x \in \mathbb{R}^d$). The probability density function $\Pr(X = x) = p(x; \mu, \Sigma)$ of a normal random variable $X \sim N(\mu, \Sigma)$ is given as: $p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det\Sigma}} \exp\left(-\frac{(x-\mu)^T \Sigma^{-1}(x-\mu)}{2}\right)$.

A normal distribution $N(\mu, \Sigma)$ can thus be uniquely characterized by a parameter point $\tilde{\Lambda} = (\mu, \Sigma)$ in dimension $D = d + \frac{d(d+1)}{2} = \frac{d(d+3)}{2}$ by stacking the d mean coordinates of μ with the $\frac{d(d+1)}{2}$ matrix coefficients of Σ. We use the tilde notation $\tilde{}$ to emphasize on the *mixed-type* nature vector/matrix of the parameter. For example, bivariate parametric distributions are represented by 5D points lying in the parameter space $\mathcal{X} = \mathbb{R}^2 \times \mathrm{Cone}(\mathbb{R}^{2\times 2})$, where $\mathrm{Cone}(\mathbb{R}^{2\times 2})$ denotes the convex cone of positive semi-definite matrices. Given a set of n d-variate distributions $\mathcal{D}$ handled as D-dimensional point set $\mathcal{S} = \{\tilde{\Lambda}_1, ..., \tilde{\Lambda}_n\}$ with $\tilde{\Lambda}_i = (\mu_i, \Sigma_i)$, we seek to to define a proper center. Ignoring for a while the fact that $\mathcal{S}$ is a point set lying in a parameter space $\mathcal{X}$, we may consider the two usual centers in Euclidean geometry $\mathbb{E}^D$: (1) The *centroid* that is commonly called and defined as the center of mass $\bar{\tilde{\Lambda}} = \frac{1}{n}\sum_{i=1}^n \tilde{\Lambda}_i$ of $\mathcal{S}$, and (2) The *circumcenter* $\tilde{\Lambda}^*$ that defines the smallest radius enclosing ball of $\mathcal{S}$.

Both the centroid and the circumcenter are appropriate centers for simplifying the point set down to its best single representative. In other words, these centers solve the 1-clustering task for the following respective minimization criteria:

(1) The *centroid* c^+ is found as the unique minimizer of the minimum average for the *squared* Euclidean distance:

$$c^+ = \arg\min_{x\in\mathbb{R}^D} \sum_{i=1}^n \frac{1}{n}||x\tilde{\Lambda}_i||^2.$$

(2) The *circumcenter* C^* is defined as the center that minimizes the radius of enclosing balls:

$$C^* = \arg\min_{x\in\mathbb{R}^D} \max_{i=1}^n ||x\tilde{\Lambda}_i||.$$

While these minimization problems look quite similar at first glance, they bear in fact very different mathematical properties. Although it could be tempting to consider "as is" these Euclidean centers for the parameter space Λ, this may not yield meaningful centers properly characterizing well the normal sets. The reason is that the Euclidean distance (or its squared distance) does not make sense[1] for normals. Indeed, consider two univariate normals $X_1 \sim N(\mu, \sigma_1^2)$ and $X_2 \sim N(\mu, \sigma_2^2)$ centered on the *same* mean μ. Applying the Euclidean distance on the parameter point $\lambda_1 = (\mu, \sigma_1^2)$ and $\lambda_2 = (\mu, \sigma_2^2)$, we get a bigger distance $\sqrt{(\sigma_2^2 - \sigma_1^2)^2}$ for σ_1 deviating much from σ_2, a clearly wrong notion of statistical distance. An appropriate distance between statistical distributions is the *Kullback-Leibler divergence* (KL) better known as *relative entropy*. The KL divergence is an *asymmetric* measure of dissimilarity of probability distributions defined as $\mathrm{KL}(p||q) = \int_x p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x$. The KL divergence relates the Shannon entropy $H(p) = -\int p(x)\log p(x)\mathrm{d}x$ with the cross-entropy $H(p;q) = \int p(x)\log\frac{1}{q(x)}\mathrm{d}x$ as follows: $\mathrm{KL}(p||q) = -H(p) - \int p\log q\mathrm{d}x = H(p;q) - H(p)$. For multivariate normal distributions, the closed-form formula for the en-

*Sony Computer Science Laboratories Inc. (FRL) and École Polytechnique (LIX), Frank.Nielsen@acm.org. Web: http://www.sonycsl.co.jp/person/nielsen/

†CEREGMIA, University of Antilles-Guyane, Richard.Nock@martinique.univ-ag.fr

[1] Except iff. all covariance matrices are the same.

tropy and relative entropy are obtained after carrying out fastidious integral computations as $H(p(x;\mu,\Sigma)) = \frac{d}{2} + \frac{1}{2}\log(2\pi)^d \det\Sigma$, (independent of μ) and $\mathrm{KL}(p(x;\mu_i,\Sigma_i)||p(x;\mu_j,\Sigma_j)) = \frac{1}{2}(\mu_i-\mu_j)^T\Sigma_j^{-1}(\mu_i-\mu_j) + \frac{1}{2}\log\det(\Sigma_i^{-1}\Sigma_j) + \frac{1}{2}\mathrm{tr}\left(\Sigma_j^{-1}\Sigma_i\right) - \frac{d}{2}$ (*), where $\mathrm{tr}(\Sigma) = \sum_{i=1}^d \Sigma_{i,i}$ denotes the matrix trace (i.e. the sum of the diagonal elements $\Sigma_{i,i}$). Observe that the relative entropy of normals with identity covariance matrices collapses to the *squared* Euclidean distance.

2 Relative entropy of exponential families

It turns out that the normal distributions belong to the (full regular) exponential families [2, 5] in statistics. Besides normal distributions, exponential families include many familiar discrete or continuous distributions such as Poisson, Bernoulli, Beta, Gamma but do not fully cover the spectrum of usual distributions either (e.g., uniform, Lévy SαS or Cauchy distributions). Exponential families admit the following canonical decomposition of their probability measures: $p(x;\lambda) = p(x;\theta) = \exp\{< \theta, t(x) > -F(\theta) + C(x)\}$, where $\theta \in \mathbb{R}^D$ are the *natural parameters* associated with the *sufficient statistics* $t(x)$ ($t : \mathbb{R}^d \mapsto \mathbb{R}^D$). The real-valued *log normalizer* function $F(\theta)$ is a strictly convex and differentiable function that specifies uniquely the exponential family, and the function $C(x)$ is the base *counting* or *Lebesgue* measure. Once this canonical decomposition is figured out, we can apply the key *equivalence theorem* [2, 5] Kullback-Leibler of distributions of the *same exponential* family$\Longleftrightarrow$Bregman divergence for the log normalizer F: $\mathrm{KL}(p(x;\mu_i,\Sigma_i)||p(x;\mu_j,\Sigma_j)) = D_F(\theta_j||\theta_i)$, to get without integral computations the closed-form formula (notice that parameter order swaps). The Bregman divergence [2, 5] D_F is defined as the tail of a Taylor expansion for a strictly convex and differentiable function F as $D_F(\theta_j||\theta_i) = F(\theta_j) - F(\theta_i) - < \theta_j - \theta_i, \nabla F(\theta_i) >$, where $< \cdot,\cdot >$ denote the vector inner product ($< p, q >= p^T q = \sum_{i=1}^d p_i q_i$) and ∇F is the gradient operator. For multivariate normals, we get the *mixed-type* natural parameters $\tilde{\Theta} = (\theta, \Theta) = (\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1})$, $F(\tilde{\Theta}) = \frac{1}{4}\mathrm{tr}(\Theta^{-1}\theta\theta^T) - \frac{1}{2}\log\det\Theta + \frac{d}{2}\log 2\pi$ and the one-to-one mapping from the source $\tilde{\Lambda} = (\mu, \Sigma)$ to natural parameters $\tilde{\Theta}$: $\tilde{\Lambda} = \begin{pmatrix} \lambda = \mu \\ \Lambda = \Sigma \end{pmatrix} \Longleftrightarrow \tilde{\Theta} = \begin{pmatrix} \theta = \Sigma^{-1}\mu \\ \Theta = \frac{1}{2}\Sigma^{-1} \end{pmatrix}$. The inner product $< \tilde{\Theta}_p, \tilde{\Theta}_q >$ in the corresponding Bregman divergence D_F is a *composite* inner product obtained as the sum of two inner products for vectors and matrices: $< \tilde{\Theta}_p, \tilde{\Theta}_q >=< \Theta_p, \Theta_q > + < \theta_p, \theta_q >$. For matrices, the inner product $< \Theta_p, \Theta_q >$ is defined by the trace of the matrix product $\Theta_p\Theta_q^T$: $< \Theta_p, \Theta_q >= \mathrm{Tr}(\Theta_p\Theta_q^T)$. One can check by hand that $\mathrm{KL}(p(x;\mu_i,\Sigma_i)||p(x;\mu_j,\Sigma_j)) = D_F(\tilde{\Theta}_j||\tilde{\Theta}_i)$ yields formula (*) by elementary calculus, bypassing complex integral computations.

3 Legendre transformation and duality

We refer to [2, 5] for detailed explanations that we quickly summarize here: Any Bregman generator function F admits a *dual* Bregman generator function $G = F^*$ via the Legendre transformation $G(y) = \sup_{x\in\mathcal{X}}\{< y, x > -F(x)\}$. The supremum is reached at the *unique* point where the gradient of $G(x) =< y, x > -F(x)$ vanishes, that is when $y = \nabla F(x)$. Writing $\mathcal{X}'_F$ for the *gradient space* $\{x' = \nabla F(x) | x \in \mathcal{X}\}$, the convex conjugate $G = F^*$ of F is the function defined by $F^*(x') =< x, x' > -F(x)$. It follows from Legendre transformation that *any* Bregman divergence D_F admits a *dual* Bregman divergence D_{F^*} related to D_F as follows: $D_F(p||q) = F(p) + F^*(\nabla F(q)) - < p, \nabla F(q) >= F(p) + F^*(q') - < p, q' >= D_{F^*}(q'||p')$. Yoshizawa and Tanabe [10] carried out that non-trivial Legendre transformation for multivariate normals. The strictly convex and differentiable dual Bregman generator function F^* (ie., potential function in information geometry) is $F^*(\tilde{H}) = -\frac{1}{2}\log(1 + \eta^T H^{-1}\eta) - \frac{1}{2}\log\det(-H) - \frac{d}{2}\log(2\pi e)$. The $\tilde{H} \Leftrightarrow \tilde{\Theta}$ coordinate transformations obtained from the Legendre transformation are given by $\tilde{H} = \nabla_{\tilde{\Theta}} F(\tilde{\Theta}) = \begin{pmatrix} \frac{1}{2}\Theta^{-1}\theta \\ -\frac{1}{2}\Theta^{-1} - \frac{1}{4}(\Theta^{-1}\theta)(\Theta^{-1}\theta)^T \end{pmatrix}$, and $\tilde{\Theta} = \nabla_{\tilde{H}} F^*(\tilde{H}) = \begin{pmatrix} -(H + \eta\eta^T)^{-1}\eta \\ -\frac{1}{2}(H + \eta\eta^T)^{-1} \end{pmatrix}$. This yields the dual *expectation* coordinate systems arising from the canonical decomposition:

$$\tilde{H} = \begin{pmatrix} \eta = \mu \\ H = -(\Sigma + \mu\mu^T) \end{pmatrix} \Longleftrightarrow \tilde{\Lambda} = \begin{pmatrix} \lambda = \mu \\ \Lambda = \Sigma \end{pmatrix}.$$

These formula simplify when we restrict ourselves to diagonal-only covariance matrices Σ_i, spherical Gaussians $\Sigma_i = \sigma_i I$, or univariate normals $\mathcal{N}(\mu_i, \sigma_i^2)$. The expectation parameter $\tilde{H}$ plays an important role for infering the source parameters $\tilde{\Lambda}$ from a sequence of identically and independently distributed observations $x_1, ..., x_v$. Indeed, the maximum likelihood estimator (MLE) of exponential families is $\hat{\tilde{H}} = \frac{1}{v}\sum_{i=1}^v t(x_i)$, where $t(x_i)$ is the sufficient statistics evaluated at x_i. This yields a simple procedure to in-

Figure 1: Right-sided (red), left-sided (blue) and symmetrized centroids (green) for 2D normals.

fer from raw data $x_1, ..., x_v \in \mathbb{R}^d$ the multivariate normal parameters $\hat{\Lambda} \Leftrightarrow \hat{\tilde{H}} \in \mathbb{R}^D$ by taking the centroid on the sufficient statistics for $t(x) = \tilde{x} = (x, -\frac{1}{2}xx^T)$. (This MLE is biased.) Gaussian distribution modeling abound in practice as explicited for three applications in [3].

4 Entropic centroids of multivariate normals

The *entropic centroids* e^+ of normals are defined similarly to the Euclidean geometry case by considering minimizing the *average distance*: the information radius r^+. Because the relative entropy is asymmetric, we consider three entropic centroids defined as the following unique minimizers: $e_L^+ = \arg\min_{\tilde{\Lambda}} \sum_{i=1}^n \frac{1}{n}\mathrm{KL}(p(x;\tilde{\Lambda})||p(x;\tilde{\Lambda}_i)), e_R^+ = \arg\min_{\tilde{\Lambda}} \sum_{i=1}^n \frac{1}{n}\mathrm{KL}(p(x;\tilde{\Lambda}_i||p(x;\tilde{\Lambda})),$ and $e^* = \arg\min_{\tilde{\Lambda}} \sum_{i=1}^n \frac{1}{2n}\mathrm{KL}(p(x;\tilde{\Lambda}_i||p(x;\tilde{\Lambda})) + \mathrm{KL}(p(x;\tilde{\Lambda})||p(x;\tilde{\Lambda}_i)).$ The latter symmetrical KL divergence is also called J-divergence and plays an important role in signal processing [4]. Using the equivalence theorem $\mathrm{KL} \leftrightarrow D_F$, it follows that the minimizers match up to source↔natural parameter conversions the following *Bregman centroids* c^+ for the log normalizer F by the swapping argument order: $e_L^+ \leftrightarrow c_R^+ = \arg\min_{\tilde{\Theta}} \sum_{i=1}^n \frac{1}{n} D_F(\tilde{\Theta}_i||\tilde{\Theta})$, $e_R^+ \leftrightarrow c_L^+ = \arg\min_{\tilde{\Theta}} \sum_{i=1}^n \frac{1}{n} D_F(\tilde{\Theta}||\tilde{\Theta}_i)$, and $e^+ \leftrightarrow c^+ = \arg\min_{\tilde{\Theta}} \sum_{i=1}^n \frac{1}{n} \frac{D_F(\tilde{\Theta}_i||\tilde{\Theta}) + D_F(\tilde{\Theta}||\tilde{\Theta}_i)}{2}$. It has been shown in [7] that the sided Bregman centroids admit *closed-form* formulas that are generalized means[2]: c_R^+ is simply the center of mass $c_R^+ = \bar{\tilde{\Theta}}$ (independent of F, a mean for the identity function) and $c_L^+ = \nabla F^{-1}(\sum_{i=1}^n \nabla F(\tilde{\Theta}_i))$, a ∇F-mean. The information radius [7] r^+ coincides for the sided centroid, and is expressed as a Burbea-Rao divergence (i.e., a generalized F-Jensen reminder): $r^+(\mathcal{S}) = \frac{1}{n}\sum_{i=1}^n F(\tilde{\Theta}_i) - F(\bar{\tilde{\Theta}}) \geq 0$. These results extend to barycenters as well, and allows one to perform interpolation and model merging on statistical normal manifolds [9]. (Note that centroids are robust to outliers.) Further, the *symmetrized entropic centroid* e^+ does not admit closed-form solution but is characterized geometrically exactly as the intersection of the geodesic linking the left-sided and right-sided Bregman centroids (say, c_L and c_R respectively) with the mixed-type bisector: $M_F(c_R^F, c_L^F) = \{x \in \mathcal{X} \mid D_F(c_R^F||x) = D_F(x||c_L^F)\}$. This yields an efficient approximation algorithm by walking dichotomically on the geodesic (wrt. the relative entropy) linking the two sided Bregman (c_L^+ and c_R^+) or equivalently entropic centroids (e_L^+ and e_R^+). Figure 1 depicts the sided and symmetrized KL entropic centroids derived from Bregman centroids. The geodesic walk algorithm [7] simplifies and generalizes a former complex and time consuming *ad-hoc method* [4], and allows one to extend the k-means algorithm [2] to hard sided and symmetrized entropic clustering of normals [3]. The Bregman loss function of sided k-means monotonously decreases. (k-means is a Bregman k-means [2] in disguise for the generator $F(x) = x^2$.)

[2] A f-mean is defined as $f^{-1}(\frac{1}{n}\sum_{i=1}^n f(x_i))$.

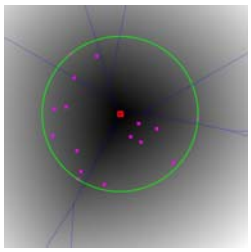

Figure 2: The circumcenter of the smallest enclosing disk is lying on the furthest Voronoi diagram.

5 Entropic circumcenters of normals

The MinMax optimization problem differs from the MinAvg optimization in the sense that it further optimizes the KL radius r^+ to its smallest possible value r^* but becomes sensitive to outliers. The MinMax optimization problem is *not* differentiable on the furthest Voronoi diagram [5], as depicted in Figure 2. We similarly define the sided and symmetrized entropic circumcenters: $E_L^* \leftrightarrow C_R^* = \arg\min_{\tilde{\Theta}} \max_{i=1}^n D_F(\tilde{\Theta}_i||\tilde{\Theta})$, $E_R^* \leftrightarrow C_L^* = \arg\min_{\tilde{\Theta}} \max_{i=1}^n D_F\tilde{\Theta}||(\tilde{\Theta}_i)$, and $E^* \leftrightarrow C^* = \arg\min_{\tilde{\Theta}} \max_{i=1}^n \frac{D_F(\tilde{\Theta}_i||\tilde{\Theta}) + D_F(\tilde{\Theta}||\tilde{\Theta}_i)}{2}$. We showed that Welzl's MiniBall algorithm extends

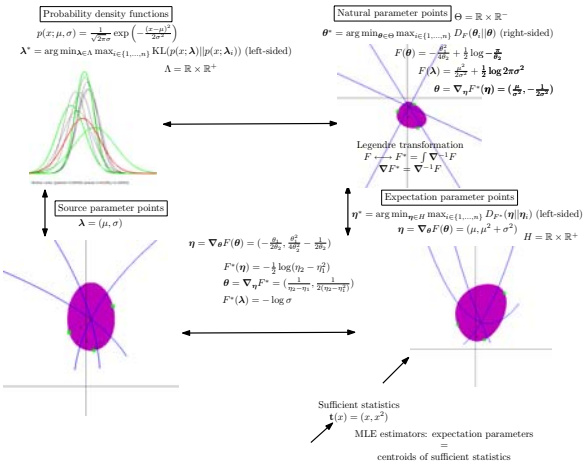

Figure 3: The left KL circumcenter of a set of 1D normals. (Zoom in pdf please for explanations.)

to arbitary Bregman divergences [8] allowing us to compute exactly the sided entropic circumcenters E_R^* and E_L^* on the plane. The basis computations relies on using the fact that the *right-sided Bregman Voronoi bisector* [5] is a straight line. (In small dimensions, computing these basis proceed by dimension reduction by constructing the Bregman generator restricted to affine subspaces.) Note that the circumcenter of the entropic ball passing through exactly three points may not exist as this circumcenter obtained as the common intersection point of three linear bisectors may potentially fall out of domain $\mathcal{X}$. However, this never happens for the recursive generalization of Welzl's algorithm [8]. As dimension increases, it is not possible to compute in practice the exact circumcenter as Welzl's algorithm exhibits the *curse of dimensionality*: an exponential time dependence with the dimension. We considered in [6] a generalization of the approximation of the smallest enclosing ball based on the notion of *core-sets* working in very large dimensions ($d \sim 10000$). As mentioned above, the computation of the smallest enclosing entropic balls rely on the property that right-type Bregman bisector are hyperplanes [5], and therefore the right-type Bregman Voronoi is an affine diagram that can be computed equivalently using a power diagram [5]. This allows us to define *entropic Voronoi diagrams* for multivariate normals with corresponding dual regular/geodesic entropic Delaunay triangulations.

6 Concluding remarks

We have concisely presented in view of our results on information-theoretic Bregman centers [7, 8, 6] the entropic centers of statistical multivariate normal distributions, i.e. the Kullback-Leibler entropic centroids and circumcenters. We have described the Legendre transformation for the mixed-type vector/matrix log normalizer of that exponential family and reported on our implementation. Although these results extend theoretically to arbitrary exponential families, the potential non-existence of closed-form solutions for the Legendre transformation or its gradient (e.g., Beta or Gamma distributions) may require to tabulate these functions in practice. We can reinterpret these entropic centers under the auspices of information geometry [1] by considering the dually flat Riemannian manifolds where Bregman divergences arise naturally as the canonical divergences [10].

References

[1] S.-I. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000. ISBN-10:0821805312.

[2] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[3] J. V. Davis and I. S. Dhillon. Differential entropic clustering of multivariate gaussians. In *Neural Information Processing Systems*, pages 337–344, 2006.

[4] T. A. Myrvoll and F. K. Soong. On divergence-based clustering of normal distributions and its application to HMM adaptation. In *EuroSpeech*, volume 2, pages 1517–1520, 2003.

[5] F. Nielsen, J.-D. Boissonnat, and R. Nock. Bregman Voronoi diagrams: Properties, algorithms and applications, 2007. arXiv.org:0709.2196 (Extend SODA'07).

[6] F. Nielsen and R. Nock. On approximating the smallest enclosing Bregman balls. In *Proc. 22nd Symposium on Computational Geometry*, pages 485–486, 2006.

[7] F. Nielsen and R. Nock. On the centroids of symmetrized Bregman divergences. 2007. arXiv.org:0711.3242.

[8] F. Nielsen and R. Nock. On the smallest enclosing information disk. *Information Processing Letters*, 105(3):93–97, 2008.

[9] B. Pelletier. Informative barycentres in statistics. *Annals of the Institute of Statistical Mathematics*, 57(4):767–780, 2005.

[10] S. Yoshizawa and K. Tanabe. Dual differential geometry associated with Kullback-Leibler information on the Gaussian distributions and its 2-parameter deformations. *SUT Journal of Mathematics*, 35(1):113–137, 1999.