A tutorial on Information Geometry

April 8, 2025

1 Introduction

Professor Shun-ichi Amari, the founder of modern information geometry, stated that Information Geometry is a method of exploring the world of information (e.g. statistical models) by means of modern geometry [1]. If so, a natural question arises:

- Why are we using a geometric approach to study Statistical models?

Geometry allows us to study the invariance of a problem in a coordinate-free framework (here an example would be the invariance of the distance between the probability distribution) and get a geometric structure (e.g a statistical manifold). Futhermore, it helps us reason intuitively about problems.

1.1 Manifold of Statistical Model

Let us consider a family of probability distribution S = $\{p(x, \theta)\}$ of a statistical model. Now, let us define a mapping $\phi : S \to \mathbb{R}^n$ by $\phi[p(x, \theta)] = \theta$. This function,$\phi$, plays a role of a coordinate function, and the vector θ is used as the coordinate for $p(x, \theta)$, and a differential structure is introduced in S by this coordinate function. This S is a differential statistical manifold.
To give the reader a visual representation of a statistical manifold, we show an example:

Let us consider X to be a discrete random variable taking values on a finite set $X = \{0, 1, \cdots, n\}$. The probability distribution defined by a vector $p(x)$ is specified by $n + 1$ probabilities

$$p_i = Prob\{x = i\}, i = 0, 1, \cdots, n \tag{1}$$

where $p(x)$ is a probability vector

$$\mathbf{p} = (p_0, p_1, \cdots, p_n) \tag{2}$$

we also have a constraint

$$\sum_{i=0}^{n} p_i = 1, \quad p_i > 0 \tag{3}$$

Thus we can visualize this discrete distribution as an n-dimensional simplex, which is a manifold. Every point in the simplex satisfies the constraint

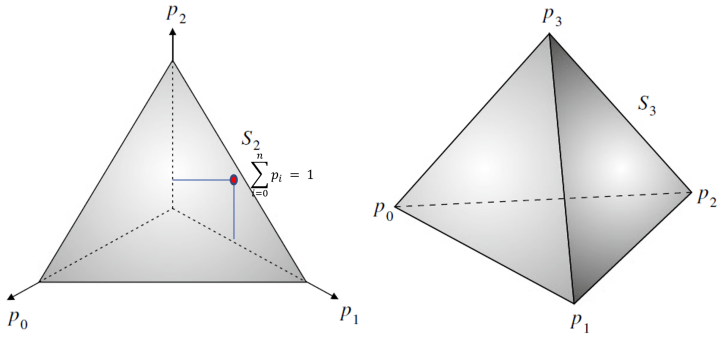

Figure 1: Manifold S_2 and S_3 of discrete probability distribution[2]

Remark : One of the motivations for using the modern geometric perspective to study statistical problems is to use these tools to measure the goodness of fit of data to a model using the concept of divergence (dissimilarity or loosely we can call it distances). For example, we want to know how close our cost or objective function in an estimation problem or likelihood in statistics matches the observed data. We also might want to measure discrepancy in models [3]. Thus, the divergence function helps us define a metric (Riemannian metric) on the manifold and allow us to study these notions of similarity (fitness) or dissimilarity of probability distribution(models). In order to understand the metric better let us formally define these divergence functions.

2 Divergence between two points

Let us consider P and Q in a manifold M, where the coordinates are ξ_P and ξ_Q, a divergence function, $D[P:Q]$ is a function that satisfies certain criteria.

Defintion1.1 $D[P:Q]$ is called divergence when it satisfies certain criteria [2]:

1. $D[P:Q] \geq 0$
2. $D[P:Q] = 0$, when and only when $P = Q$
3. When P and Q are sufficiently close, by denoting their coordinates by ξ_P and $\xi_Q = \xi_P + d\xi$, the Taylor series expansion of D is written as

$$D[\xi_P : \xi_P + d\xi] = \frac{1}{2} \sum g_{ij}(\xi_P) d\xi_i d\xi_j + O(|d\xi|^3), \tag{4}$$

and matrix $G = g_{ij}$ is a positive deinite, depending on ξ_P

Remark : The first two conditions are very easy to interpret and we might interpret them as distances, however, divergence does not necessarily mean distance because they do not satisfy the triangle inequality and also is not symmetric $D[P : Q] \neq D[Q : P]$. In contrast, the third statement might be confusing. The third statement shows us, how the divergence D provides, manifold M with a Riemannian structure. A manifold is Riemannian when a positive definite matrix $G(\xi)$ is defined on M.

2.1 Example of Divergence

First, let us look at some familiar divergences some of us might have seen in other courses or while performing research.

2.1.1 Euclidean Divergence

The coordinate system is an orthonormal Cartesian coordinate system and divergence is half of the square of the Euclidean distance.

$$D[P : Q] = \frac{1}{2}\sum(\xi_{P_i} - \xi_{Q_i})^2 \tag{5}$$

Here the Riemannian metric is the identity

2.1.2 Kullback Liebler Divergence

This divergence is very popular in the literature and has many applications in machine learning, optimization, and information theory. The divergence is written as:

$$D_{KL}[p(x) : q((x)] = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{6}$$

We will go into detail about the KL-Divergence in the later section and see how it is derived.

In order to define KL divergence, we need to introduce Bregmann Divergence, which is a generalized divergence for dually flat manifolds. I will explain these dual flat nature later in the report. For now, let us consider a strictly convex function, defined over the coordinate ξ. A function $\psi(\xi)$ is convex when this inequality is satisfied.

$$\psi\{\lambda\xi_1 + (1-\lambda)\xi_2\} \leq \lambda\psi(\xi_1) + (1-\lambda)\psi(\xi_2) \tag{7}$$

where, $0 \leq \lambda \leq 1$.

2.2 Bregman Divergence

If we can define a convex function in a manifold, with coordinate ξ as shown in Fig. 2. We can draw a tangent hyperplane touching it at a point ξ_0. The tangent hyperplane can be represented as :

$$z = \psi(\xi_0) + \nabla\psi(\xi_0).(\xi - \xi_0) \tag{8}$$

The graph of ψ is always above the hyperplane. Thus, We want to figure out how diverged(far away) is the approximation(tangent plane) from the true function. We can quantify it using the relation below.

$$D_\psi[\xi : \xi_0] = \psi(\xi) - \psi(\xi_0) - \nabla\psi(\xi_0) \cdot (\xi - \xi_0) \tag{9}$$

This function D_ψ satisfies the criteria of divergence. This is called a Bregann divergence derived from the convex function ψ.

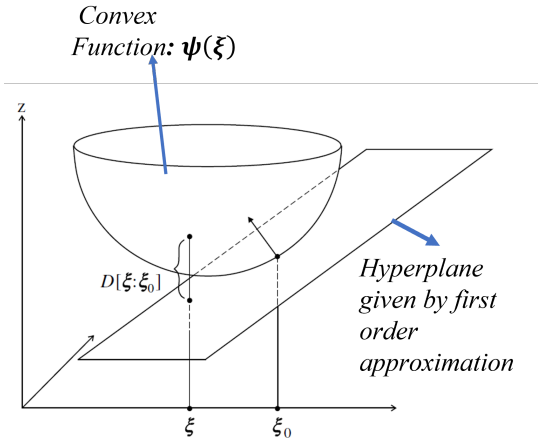

Figure 2: Bregman Divergence $D_\psi[\xi : \xi_0]$ of a convex function $\psi(\xi)$[2]

Remark : Why is the Bregmann divergence important in information geometry?
The exponential family is a powerful statistical model, which includes a family of Probability distribution such as discrete probability distribution , Gaussian , Multinomial , Gamma distributions. These exponential families are associated with a convex function known as the cumulant generating function or free energy. Let us define a exponential family of probability density function.

$$p(x, \theta) = exp\{\theta x - \psi(\theta)\} \tag{10}$$

and

$$\int p(x, \theta) d\mu(x) = 1 \tag{11}$$

and ψ or the cumulant generating function (normalizing factor) is convex and can be derived using Equation (10,11)written as

$$\psi(\theta) = log \int exp(\theta \cdot x) d\mu(x) \tag{12}$$

Now, since we have a convex function associated with the exponential family, we can use Bregman divergence for determining the similarity between probability density functions.

Furthermore, let us introduce an important transformation technique called Legendre Transformation. The importance of Legendre transformation in information geometry lies in the fact that it allows one to switch between different parametrizations of a statistical manifold, and to relate different notions of divergence and entropy to it. We will see the definition first and how these claims hold below:

3 Legendre Transformation

The Legendre transformation is used to transform a function into another function of a different variable that contains the same information as the original function. An example in mechanics is the conversion of a Lagrangian into a Hamiltonian using the Legendre Transformation[4].

A graphical interpretation will make more sense.

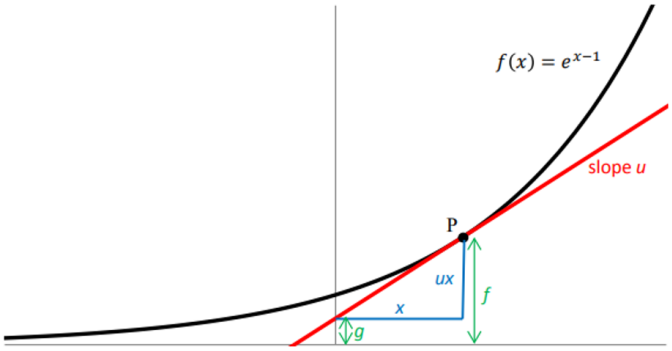

Figure 3: A convex function (e^{x-1}) plotted in (x, y) coordinate; A tangent line approximates the function at each point in the curve [4]

In Figure 3, we consider a convex function $f(x) = e^{(x-1)}$.We can approximate the function at point P using the tangent line (first-order Taylor series expansion). Now let's think about the definition of Legendre transformation, we want to find another function with different variables that preserve the information of $f(x)$, i.e, we want another coordinate $\xi^\star = (x^\star, y^\star)$. Since a normal vector at each point is unique, we can specify a point of M by using a normal

vector. Here in the Figure 3, we see that we can use slope and y-intercept $\xi^\star = (\mu, g(u))$, to define the same curve.

An example would be: The slope becomes:

$$u = \frac{df}{dx} = e^{x-1} \longrightarrow x = 1 + \ln u \tag{13}$$

Similarly, the Intercept can be written as:

$$f(x) = xu + g(u) \qquad g(u) = f(x) - xu = u - (1 + lnu)u = -ulnu \tag{14}$$

If we plot this function $g(u)$, we also get

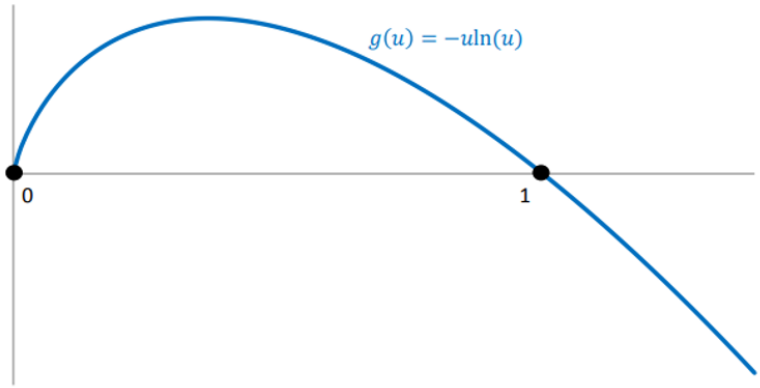

Figure 4: Function $g(u)$ derived using Legendre Transformation [4]

If we write this graphical explanation in a generalized form we can derive the following. Let us define a convex function $\psi(\xi)$:

$$\xi^\star = \nabla\psi(\xi) \tag{15}$$

The legendre transformation $\psi^\star$ is a dual of ψ and satisfies

$$\psi^\star(\xi^\star) = \max_{\xi'}\{\xi^{'}.\xi^\star = \psi(\xi^{'})\} \tag{16}$$

You can compare this to the definition we came up with using the graphical interpretation $g(u) = f(x) - xu$, we can see they are exactly the same.

Remark : We see that the Legendre transformation establishes a duality between objects in dual spaces. This is because if you take the derivative of $\psi^\star(\xi^\star)$ with respect to $(\xi^\star)$, we get our original coordinate ξ. This will help us define the dual expression of Bergman divergence later on.

4 Dual Connections

Let us consider a manifold M and a tangent space T_ξ at ξ. If we wanted to compare the two basis vectors of two tangent spaces at ξ and $\xi + d\xi$. For example, an interesting question is, How different are they? We cannot compare them directly because they are in different tangent spaces. In our class, we used the affine connections, Christoffel symbols, in order to define a linear correspondence between these two tangent spaces. And we also showed that the Levi-Civita Connection unifies the picture by relating the Riemannian metric (G_{ij}) and the affine connection Γ_{ij}. Thus, there are two ways to define straightness or flatness, one using the tangent vector along the curve pointing in the same direction and the other using the shortest path connecting two points.

An interesting property of Affine connections allows us to define a metric in the manifold using two affine connections. For example, if there exists an affine connection $\Gamma_{ij}(\theta)$, where the geodesic (a generalization of a straight line) can be defined. Correspondingly, $\Gamma^{\star}_{ij}(\theta)$ defines a dual affine connection where another geodesic called the dual geodesic can be defined. This is important for us because the two tangent vectors defined from these dual and primal coordinate systems will help us define orthogonality in a manifold. A manifold is dually flat when two affine coordinates are connected by the Legendre Transformation. The primal and dual geodesic are not the same measure because the Coordinate transformation that connects them might be nonlinear.

5 Bregman Divergence and Exponential Family

An exponential family induces a convex function as shown in Equation 10. The Bregman divergence from $p(x, \theta)$ to $p(x, \theta^{'})$ for the exponential family is defined as:

$$D_{\psi}[\theta^{'} : \theta] = \psi(\theta^{'}) - \psi(\theta) - \theta^{*} \cdot (\theta^{'} - \theta) \tag{17}$$

where $\theta^{\star}$, is the Legendre transformation of the coordinate θ

$$\theta^{\star} = \nabla\psi(\theta) \tag{18}$$

The dual convex of the cumulant generating function (12) is given by the Legendre transformation(16) which can be written as:

$$\psi^{\star}(\theta^{\star}) = \int p(X, \theta) \log p(x, \theta) dx \tag{19}$$

The above equation is also a convex function (we call it a dual convex function $\psi^{\star}$ of ψ) and is given by negative entropy.
If we take the divergence from $p(x, \theta)$ to $p(x, \theta^{'})$, Equation(17) will be reduced to the following equation:

$$D_{\psi^\star}[\theta^{'} : \theta] = \int p(x,\theta) \log \frac{p(x,\theta)}{p(x,\theta^{'})} d\mu(x) = D_{KL}[\theta : \theta'] \tag{20}$$

$$\tag{21}$$

The above Equation 20 is called the KL Divergence.

Remark : The KL divergence is important because it can be used to quantify the information loss or gain when approximating one distribution by another. For example, it can be used to compare the performance of different models or classifiers, to measure the distance between clusters, to estimate the entropy of a random variable, or to perform Bayesian inference.

Now, let's introduce some theorems that are very useful in estimation problems.

6 Generalized Pythagorean Theorem

In Euclidean space, a right-angled triagle can be defined when two basis vectors are independent to each other (or two straight lines are orthogonal to each other) as shown in the left of Figure 5. If such a triangle exists, the Pythagoras theorem can be defined. For a non-euclidean dually-flat manifold, we can also define the Pythagorean theorem.

Remark : Here is the intuition, two curves $\theta_1(t)$ and $\theta_2(t)$, intersect orhtogonally when their tangent vectors are orhtogonal, i.e.

$$< \dot{\theta}_1(t), \dot{\theta}_2(t) > \quad = g_{ij}\theta_1^i(t)\theta_2^j(t) = 0 \tag{22}$$

where g_{ij} defines the Riemannian metric(*The Riemannian metric is induced because of the inner product of the basis vectors*).

A generalized Pythagorean theorem only holds in a dually flat manifold, M. The notion of straightness in a dually flat manifold is defined using geodesics and a triangle PQR is orthogonal when the dual geodesic connecting P and Q is orthogonal to geodesic connecting Q and R as shown in Figure 5.

Theorem 1. *When triangle PQR is orthogonal such that a dual geodesic connecting P and Q is orthogonal to the geodesic connecting Q and R , the following relation holds:*

$$D_\psi(R : P) = D_\psi(Q : P) + D_\psi(R : Q) \tag{23}$$

Since the divergence function is asymmetric as mentioned in Section 2, a dual for the Pythagorean theorem also arises. You can refer to the book by Amari [2].

The Pythagorean theorem helps us define the Projection theorem, which is a very important tool in information geometry. First, let us define projection theorem and why it is important.

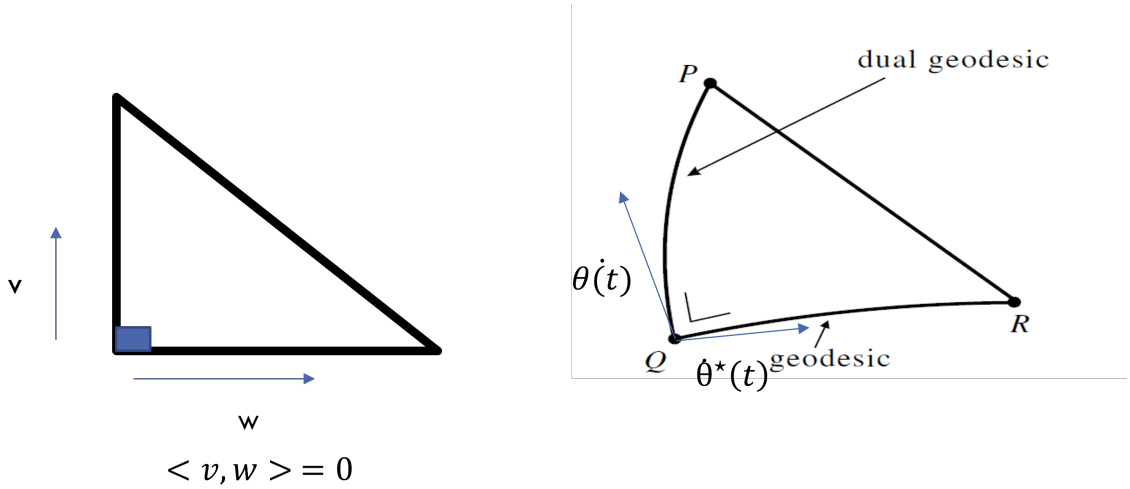

Figure 5: left: Pythagorean Theorem in Euclidean space ; right: Generalized Pythagorean Theorem [2]

7 Projection Theorem

Let's say we have a dually flat manifold M and S is a submanifold of that dually flat manifold as shown in Figure 6. A natural question arises: if there is a point in the manifold P in M, what is the point in S that is closest to P (minimizes the divergence)? This question arises when we try to solve various approximation problems. Before pointing out the power of the projection theorem, we need to define some properties necessary to state the theorem. A geodesic projection is a generalization of projection in a dually flat manifold and is defined as follows

Definition : $\hat{P}_s$ is the geodesic projection of P to S when the geodesic connecting P and $\hat{P}_s$ and dual geodesic projection is defined using the dual geodesic [2].

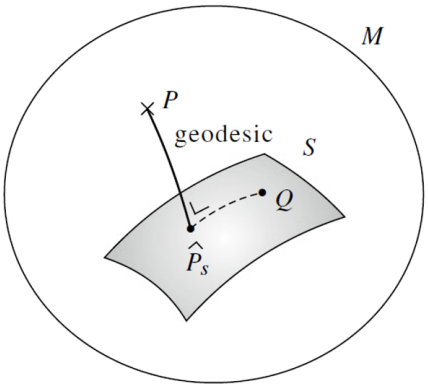

Figure 6: Projection of point $P \in M$ to S [2]

Remark : This definition is not clear, what does it mean for a geodesic to be orthogonal to submanifold S. A geodesic is a curve, let's call it $\theta(t)$. This curve θ is orthogonal to S when its tangent vector $\dot{\theta}(t)$ is orthogonal to any tangent vector of S at the intersection.

Now we can define the projection theorem:

Theorem 2. *(Projection Theorem) Given $P \in M$ and a smooth submanifold $S \subset M$, the point $\hat{P}_s^\star$ that is the dual geodesic projection of P to S is the point in S that mninimizes the divergence*

$$D_\psi[P : R], R \in S \tag{24}$$

Remark: The proof for this statement is illustrated in Amari's book [2]. A brief way to summarize the proof (refer to the Figure(5,and 6) for visual insights) would be the dual geodesic projection $\hat{P}_s^\star$ on S and $Q \in S$ are connected using an infinitesimal line element, $\hat{P}_s^\star QR$ form an orthogonal triangle and hence we can use the Pythagorean theorem. The theorem shows dependence on$\hat{P}_s^\star$ while computing $D_\psi[P : R]$. Thus the dual projection determining the minimization of the primal divergence makes sense.

Note: A similar, dual-primal alternate relationship is seen in KL-Divergence. KL Divergence is derived from the dual convex of the primal convex function defining the exponential family.

These two theorems are very useful in many applications. I would like to highlight some, however, I will only illustrate the EM algorithm for the brevity of the report.

8 Applications

Some of the applications of Pythagorean Theorem and Projection theorem are as follows:

- Whenever one needs to choose a model from a manifold M that is the least biased, one would choose the one that maximizes the entropy. The geodesic projection (also called e-projection) on a statistical manifold gives the distribution that maximizes the entropy. Refer to Amari [2] for detail.
- Mutual information is a measure of the discrepancy of probability from its independence. Let's say we have a non-independent distribution $p(x, y)$ and we want to find an independent distribution is close to the distribution $p(x, y)$. It is given by a dual-geodesic projection of p(x,y) to a submanifold of independent distribution M_I

Besides, these applications I want to illustrate a powerful algorithm called the EM algorithm, that has been used in many applications like data clustering, image recognition, estimation of Gaussian mixture models, Hidden Markov Models, etc.

8.1 Expectation Maximization(EM) Algorithm

Rather than going into too much technicality, I will try my best to explain it using some visual connections for the ease of the reader.

Let us say we have a statistical model $M = \{p(x, \xi)\}$, where x is divided into two parts $x = (y, h)$, where y is observed and h is called a hidden variable.M is called the model manifold. Now we need to estimate the parameters (ξ) of the model just by using the observed data. Since we can observe y, we find the empirical distribution which is given by the equation:

$$\bar{q}(y) = \frac{1}{N} \sum y_i, \tag{25}$$

We want a candidate of joint distribution $q(y, h)$, that estimates the statistical model, $p(x, \xi)$, well. This joint distribution can be written as

$$\bar{q}(y, h) = \bar{q}(y)q(h|y) \tag{26}$$

$\bar{q}(y)$ is known, however $q(h|y)$ is unknown. We want a conditional distribution that is compatible with the empirical distribution $\bar{q}(y)$. However, we don't know which one to choose, so lets us consider all the probability distributions, i.e. arbitrary set of distributions $q(h|y$, defined in a submanifold. Thus the candidate joint probability becomes a submanifold (it is no longer a point in the manifold, since the $q(h|y)$ is arbitrary). We call this submanifold data manifold and it is denoted by $D = \{\bar{q}(y, h)\}$ as shown in Figure 7

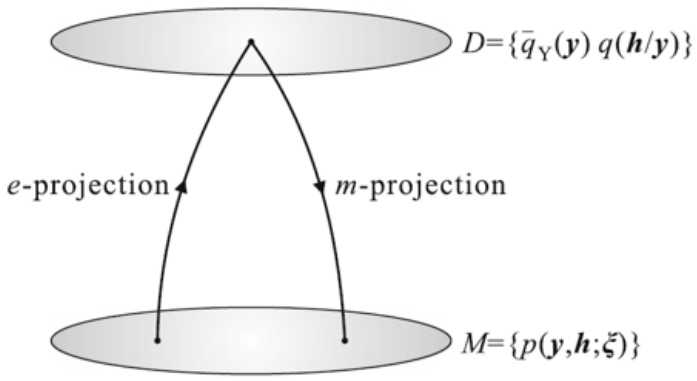

Figure 7: Alternating projection, e-projection is the geodesic projection and m-projection is dual geodesic projection [2]

Remark : The consequence of having a hidden variable is that, our estimator is not a single point anymore, but a set of points that creates a submanifold called the Data Manifold D. In this scenario, we can't just use the projection theorem that minimizes the divergence between a point and a manifold. Here, we need to perform an iterative alternate minimization algorithm called an EM algorithm as shown in Figure 8. First, we initialize a parameter randomly in in the model manifold, and we perform this alternating dual and primal geodesic projection algorithm until these geodesic converge, leading us to a locally optimal estimate of the model manifold.

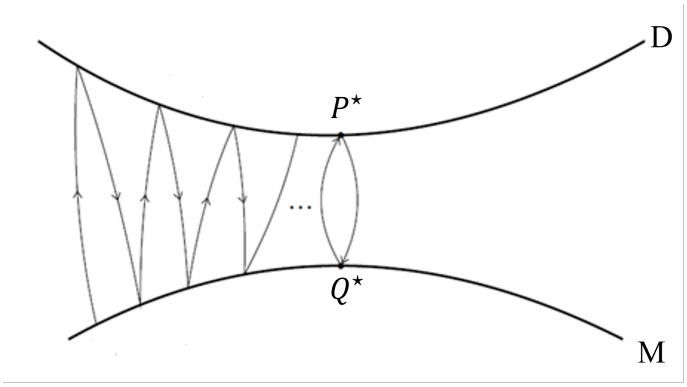

Figure 8: The procedure for the iterative dual geodesic projection (EM algorithm) [2]

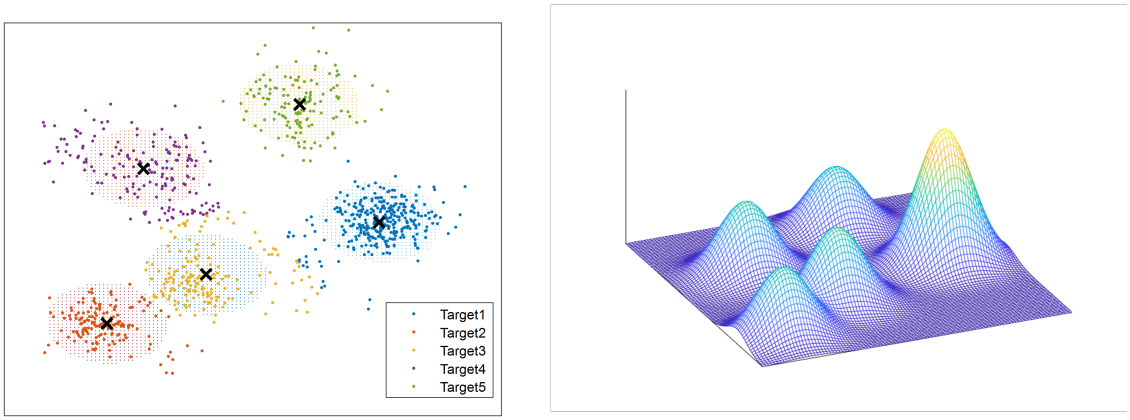

Figure 9: Estimating a Gaussian Mixture Model using EM algorithm, which also resulted in a cluster

8.2 Simulations

These are some of the results I obtained after using the EM algorithm for multiple datasets. Figure 9 is an example of approximating a Gaussian Mixture Model of a given dataset used for a monitoring application. Figure 10 is a clustering example using the medical dataset from UCI.

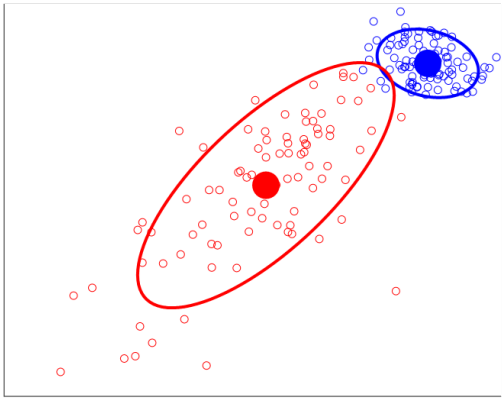

Figure 10: Em algorithm for clustering the UCI Epidemiology Dataset, One group is healthy another group has Iron deficiency Anemia

References

[1] S.-i. Amari, *Differential-geometrical methods in statistics*, vol. 28. Springer Science & Business Media, 2012.

[2] S.-i. Amari, *Information geometry and its applications*, vol. 194. Springer, 2016.

[3] F. Nielsen, "An elementary introduction to information geometry," *Entropy*, vol. 22, no. 10, p. 1100, 2020.

[4] V. Hirvonen, "Legendre transformations for dummies: Intuition examples," 2020.