Information geometry and prior selection

Hichem Snoussi* and Ali Mohammad-Djafari*

*Laboratoire des Signaux et Systèmes (L2S),
Supélec, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France

Abstract. In this contribution, we study the problem of prior selection arising in Bayesian inference. There is an extensive literature on the construction of non informative priors and the subject seems far from a definite solution [1]. Here we revisit this subject with differential geometry tools and propose to construct the prior in a Bayesian decision theoretic framework. We show how the construction of a prior by projection is the best way to take into account the restriction to a particular family of parametric models. For instance, we apply this procedure to the curved parametric families where the ignorance is directly expressed by the relative geometry of the restricted model in the wider model containing it.

INTRODUCTION

Experimental science can be modeled as a learning machine mapping the inputs $\boldsymbol{x}$ to the outputs $\boldsymbol{y}$ (see figure 1). The complexity of the physical mechanism underlying the mapping inputs/outputs or the lack of information make the prediction of the outputs given the inputs (forward model) or the estimation of the inputs given the outputs (inverse problem) a difficult task. When a parametric forward model $p(\boldsymbol{y} \,|\, \boldsymbol{x}, \boldsymbol{\theta})$ is assumed to be available from the knowledge of the system, one can use the classical ML or when a prior model $p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})$ is assumed to be available too, the classical Bayesian methods can be used to obtain the joint *a posteriori* $p(\boldsymbol{x}, \boldsymbol{\theta} \,|\, \boldsymbol{y})$ and then both $p(\boldsymbol{x} \,|\, \boldsymbol{y})$ and $p(\boldsymbol{\theta} \,|\, \boldsymbol{y})$ from which we can make any inference about $\boldsymbol{x}$ and $\boldsymbol{\theta}$. But in many practical situations the question of modeling $p(\boldsymbol{y} \,|\, \boldsymbol{x})$ and $p(\boldsymbol{x})$ is still open and to validate a model, one uses what is called the training data $\boldsymbol{z} = (\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1..N}$. Then the role of statistical learning become trying to find a joint distribution $p(\boldsymbol{z})$ belonging in general to the whole set of probability distributions and to exploit the maximum of relevant information to provide some desired predictions. In this paper, we suppose that we are given some training data $\boldsymbol{x}_{1..N}$ and $\boldsymbol{y}_{1..N}$ and some information about the mapping which consists in a model $\mathcal{Q} = \{P(\boldsymbol{z})\}$ of probability distributions, parametric ($\mathcal{Q} = \{P(\boldsymbol{z} \,|\, \boldsymbol{\theta})\}$) or non parametric. Our objective is to construct a learning rule τ mapping the set $\mathcal{Z}$ of training data $\boldsymbol{z} = (\boldsymbol{x}_{1..N}, \boldsymbol{y}_{1..N})$ to a probability distribution $p \in \mathcal{Q}$ or to a probability distribution in the whole set of probabilities $p \in \mathcal{P}$:

$$\begin{array}{rccc} \tau: & \mathcal{Z} & \longrightarrow & \mathcal{Q} \\ & \boldsymbol{z} & \mapsto & q=\tau(\boldsymbol{z}) \end{array}$$

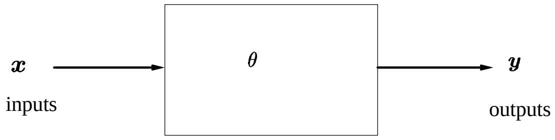

Figure 1. Learning machine model of experimental science

The Bayesian statistical learning leads to a solution depending on the prior distribution of the unknown distribution p. In the parametric case, this is equivalent to the prior $\Pi(\boldsymbol{\theta})$ on the parameter $\boldsymbol{\theta}$. Finding a general expression for $\Pi(\boldsymbol{\theta})$ and how this expression reflects the relationship between a restricted model and the closer set of ignorance containing it are the main objectives of this paper. We show the prior expression depends on the chosen geometry (subjective choice) of the set of probability measures. We show that the entropic prior [Rodriguez 1991, [2]] and the conjugate prior of exponential families are special cases related to special geometries.

In section I, we review briefly some concepts of Bayesian geometrical statistical learning and the role of differential geometry. In section II, we develop the basics of prior selection in a Bayesian decision perspective and we discuss the effect of model restriction both from non parametric to parametric modelization and from parametric family to a curved family. In section III, we study the particular case of δ-flat families where previous results have explicit formula. In section IV, we come across the case of δ-flat families mixture. In section V, we apply these results to a couple of learning examples, the mixture of multivariate Gaussian classification and blind source separation. We end with a conclusion and indicate some future scopes.

I. STATISTICAL GEOMETRIC LEARNING

I.1/ Mass and Geometry

The statistical learning consists in constructing a learning rule τ which maps the training measured data $\boldsymbol{z}$ to a probability distribution $q=\tau(\boldsymbol{z})\in\mathcal{Q}\subset\mathcal{P}=\{p\mid\int p=1\}$ (the predictive distribution). The subset $\mathcal{Q}$ is in general a parametric model and it is called the computational model. Therefore, our target space is the space of distributions and it is fundamental to provide this space with, at least in this work, two attributes which are the mass (a scalar field) and a geometry. The mass is defined by an *a priori* distribution $\Pi(p)$ on the space $\mathcal{P}$ before collecting the data $\boldsymbol{z}$ and modified according to Bayesian rule after observing the data to give the *a posteriori* distribution (see figure 2):

$$P(p\mid\boldsymbol{z})\propto P(\boldsymbol{z}\mid p)\,\Pi(p)$$

where $P(\boldsymbol{z} \mid p)$ is $p(\boldsymbol{z})$ the likelihood of the probability p to generate the data $\boldsymbol{z}$.

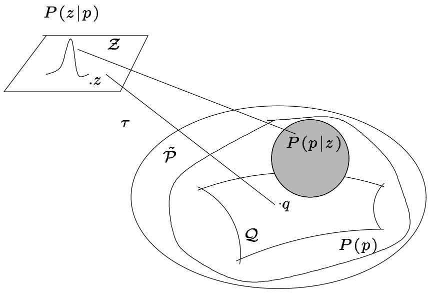

Figure 2. *a posteriori* mass proportional to the product of the *a priori* mass and the likelihood function.

The geometry can be defined by the δ-divergence D_δ:

$$D_\delta(p,q) = \frac{\int p}{1-\delta} + \frac{\int q}{\delta} - \frac{\int p^\delta q^{1-\delta}}{\delta(1-\delta)}$$

which is an invariant measure under reparametrization of the restricted parametric model $\mathcal{Q}$. It is shown [Amari 1985, [3]] that, in the parametric manifold $\mathcal{Q}$, the δ-divergence induces a dualistic structure $(g, \nabla^\delta, \nabla^{1-\delta})$, where g is the Fisher metric, ∇^δ the δ connection with Christoffel symbols $\Gamma^\delta_{ij,k}$ and $\nabla^* = \nabla^{1-\delta}$ its dual connection:

$$\begin{cases} g_{ij} & = & E_{\boldsymbol{\theta}}\left[\partial_i l(\boldsymbol{\theta})\, \partial_j l(\boldsymbol{\theta})\right] \\ \Gamma^\delta_{ij,k} & = & E_{\boldsymbol{\theta}}\left[\left(\partial_i \partial_j l(\boldsymbol{\theta}) + \delta \partial_i l(\boldsymbol{\theta}) \partial_j l(\boldsymbol{\theta})\right) \partial_k l(\boldsymbol{\theta})\right] \end{cases}$$

The parametric manifold $\mathcal{Q}$ is δ-flat if and only if there exists a parameterization $[\theta_i]$ such that the Christoffel symbols vanish: $\Gamma^\delta_{ij,k}(\boldsymbol{\theta}) = 0$. The coordinates $[\theta_i]$ are called the affine coordinates. If for a different coordinate system $[\theta^{'}_i]$, the connection coefficients are null then the two coordinate systems $[\theta_i]$ and $[\theta^{'}_i]$ are related by an affine transformation, i.e there exists a $(n \times n)$ matrix $\boldsymbol{A}$ and a vector $\boldsymbol{b}$ such that $\boldsymbol{\theta}^{'} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{b}$.

All the above definitions can be extended to non parametric families by replacing the partial derivatives with the Fréchet derivatives. Embedding the model $\mathcal{Q}$ in the whole space of finite measures $\tilde{\mathcal{P}}$ [Zhu et al. 1995, [4, 5]] not only the space of probability distributions $\mathcal{P}$, many results can be proven easily for the main reason that $\tilde{\mathcal{P}}$ is δ-flat and δ-convex $\forall$ δ in $[0,1]$. However, $\mathcal{P}$ is δ-flat for only $\delta = \{0,1\}$ and δ-convex for $\delta = 1$. For notation convenience, we use the δ-coordinates $\overset{\delta}{l}$ of a point $p \in \tilde{\mathcal{P}}$ defined as:

$$\overset{\delta}{l}(p) = p^\delta / \delta$$

A curve linking 2 points a and b is a function $\gamma : [0,1] \longrightarrow \tilde{\mathcal{P}}$, such that $\gamma(0) = a$ and $\gamma(1) = b$. A curve is a δ-geodesic in the δ-geometry if it is a straight line in the δ-coordinates.

I.2/ Bayesian learning

The loss quantity of a decision rule τ with a fixed δ-geometry can be measured by the δ-divergence $D_\delta(p, \tau(\boldsymbol{z}))$ between the true probability p and the decision $\tau(\boldsymbol{z})$. This divergence is first averaged with respect to all possible measured data $\boldsymbol{z}$ and then with respect to the unknown true probability p which gives the generalization error $E(\tau)$:

$$E_\delta(\tau) = \int_p P(p) \int_{\boldsymbol{z}} P(\boldsymbol{z} \,|\, p) D_\delta(p, \tau(\boldsymbol{z}))$$

Therefore, the optimal rule τ_δ is the minimizer of the generalization error:

$$\tau_\delta = \arg\min_\tau \{E_\delta(\tau)\}$$

The coherence of Bayesian learning is shown in [Zhu et al. 1995, [4, 5]] and means that the optimal estimator τ_δ can be computed pointwise as a function of $\boldsymbol{z}$ and we don't need a general expression of the optimal estimator τ_δ:

$$\hat{p}(\boldsymbol{z}) = \tau_\delta(\boldsymbol{z}) = \arg\min_q \int_p P(p \,|\, \boldsymbol{z}) D_\delta(p, q) \qquad (1)$$

By variational calculation, the solution of (1) is straightforward and gives:

$$\hat{p}^\delta = \int p^\delta P(p \,|\, z)$$

The above solution is exactly the gravity center of the set $\tilde{\mathcal{P}}$ with mass $P(p \,|\, \boldsymbol{z})$, the *a posteriori* distribution of p and the δ-geometry induced by the δ-divergence D_δ. Here we have the analogy with the static mechanics and the importance of the geometry defined on the space of distributions. The whole space of finite measures $\tilde{\mathcal{P}}$ is δ-convex and thus, independently on the *a posteriori* distribution $P(p \,|\, \boldsymbol{z})$ the solution $\hat{p}$ belongs to $\tilde{\mathcal{P}}$ $\forall\ \delta \in [0, 1]$.

I.3/ Restricted Model

In practical situations, we restrict the space of decisions to a subset $\mathcal{Q} \in \tilde{\mathcal{P}}$. $\mathcal{Q}$ is in general a parametric manifold that we suppose to be a differentiable manifold. Thus $\mathcal{Q}$ is parametrized with a coordinate system $[\theta_i]_{i=1}^n$ where n is the dimension of the manifold. $\mathcal{Q}$ is also called the computational model and we prefer this appellation because the main reason of the restriction is to design and manipulate the points p with their coordinates which belong to an open subset of $\mathbb{R}^n$. However, the computational model $\mathcal{Q}$ is not disconnected from non parametric manipulations and we will show that both *a priori* and final decisions can be located outside the model $\mathcal{Q}$.

Let's compare now the non parametric learning with the parametric learning when we are constrained to a parametric model $\mathcal{Q}$:

1. **Non parametric modeling:** The optimal estimate is the minimizer of the generalization error where the true unknown point p is allowed to belong to the whole

space $\tilde{\mathcal{P}}$ and the minimizer q is constrained to $\mathcal{Q}$:

$$\hat{q}(\boldsymbol{z}) = \tau_\delta(\boldsymbol{z}) = \arg\min_{q \in \mathcal{Q}} \int_{p \in \tilde{\mathcal{P}}} P(p \,|\, \boldsymbol{z}) D_\delta(p, q) \tag{2}$$

Thus the solution is the δ-projection of the barycentre $\hat{p}$ of $(\tilde{\mathcal{P}}, P(p \,|\, \boldsymbol{z}), D_\delta)$ onto the model $\mathcal{Q}$ (see figure 3).

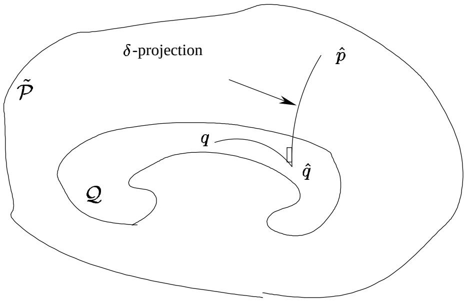

Figure 3. Projection of the non parametric solution onto the computational model

2. **Parametric modeling:** The optimal estimate is the minimizer of the same cost function as in the non parametric case but the true unknown point p is also constrained to be in $\mathcal{Q}$:

$$\hat{q}(\boldsymbol{z}) = \tau_\delta(\boldsymbol{z}) = \arg\min_{q \in \mathcal{Q}} \int_{p \in \mathcal{Q}} P(p \,|\, \boldsymbol{z}) D_\delta(p, q) = \arg\min_{q \in \mathcal{Q}} \int_{\boldsymbol{\theta}} P(\boldsymbol{\theta} \,|\, \boldsymbol{z}) D_\delta(p_\theta, q) d\boldsymbol{\theta} \tag{3}$$

The solution is the δ-projection of the barycentre $\hat{p}$ of $(\mathcal{Q}, P(\boldsymbol{\theta} \,|\, \boldsymbol{z}), D_\delta)$ onto the model $\mathcal{Q}$ (see figure 4).

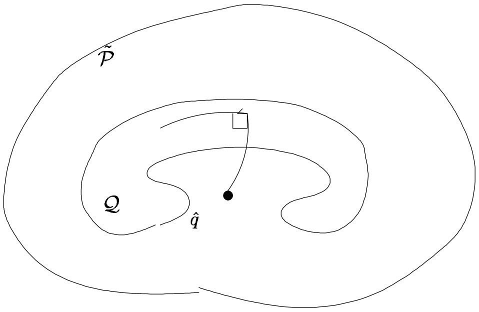

Figure 4. Projection of the barycentre solution onto the parametric model

The interpretation of the parametric modeling as a non parametric one and the effect of such restriction can be done in two ways:

1. The cost function to be minimized in equation (3) is the same as the cost function in (2) when p is allowed to belong to the whole set $\tilde{\mathcal{P}}$ and the *a posteriori* $P(p \,|\, \boldsymbol{z})$ is zero outside the model $\mathcal{Q}$. This is the case when the prior $P(p)$ has $\mathcal{Q}$ as its support. However this interpretation implies that the best solution $\hat{p}$ which is the barycentre of $\mathcal{Q}$ can be located outside the model $\mathcal{Q}$ and thus has *a priori* a zero probability !
2. The second interpretation is to say that the cost function to be minimized in equation (3) is the same as the cost function in (2) when the *a posteriori* $P(\boldsymbol{\theta} \,|\, \boldsymbol{z})$ is the projected mass of the *a posteriori* $P(p \,|\, \boldsymbol{z})$ onto the model $\mathcal{Q}$. We note here the role of the geometry defined on the space $\mathcal{P}$ and the relative geometric shape of the manifold. For instance, the ignorance is directly related to the geometry of the model $\mathcal{Q}$. The projected *a posteriori* or *a priori* can be computed by:

   $$f^{\perp}(q) \propto \int_{p \in \mathcal{S}_q} f(p)$$

   where $f(p)$ designs the *a priori* or the *a posteriori* distribution and $\mathcal{S}_q = \{p \in \tilde{\mathcal{P}} \mid p^{\perp} = q\}$ the set of points p whose the δ-projection is the q in $\mathcal{Q}$.
   The manipulation of these concepts in the general case is very abstract. However, in section IV, we present the explicit computations in the case of restricted autoparallel parametric submanifold $\mathcal{Q}_1 \in \mathcal{Q}$ of δ-flat families.

II. PRIOR SELECTION

The present section is the main contribution of this paper. We address here the problem of prior selection in a Bayesian decision framework. By prior selection, we mean how to construct a prior $P(p)$ respecting the following rule: Exploit the prior knowledge without adding irrelevant information. We note that this represents a trade off between some desirable behaviour and uniformity of the prior. We want to insist here, that the prior selection must be performed before collecting the data $\boldsymbol{z}$, otherwise the coherence of the Bayesian rule is broken down.

In a decision framework, the desirable behaviour can be stated as follows: Before collecting the training data, provide a reference distribution p_0 as a decision. The reference distribution can be provided by an expert or by our previous experience. Now, we have the inverse problem of the statistical learning. Before, the *a posteriori* distribution (mass) is fixed and we have to find the optimal decision (barycentre). Now, the optimal decision p_0 (barycentre) is fixed and we have to find the optimal repartition $\Pi(p)$ according to the uniformity constraint. In order to have the usual notions of integration and derivation, we assume that our objective is to find the prior on the parametric model $\mathcal{Q} = \{q_\theta \mid \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^n\}$.

The cost function can be constructed as a weighted sum of the generalization error of the reference prior and the divergence of the prior from the Jeffreys prior (The square root of the determinant of the Fisher information [6]) representing the uniformity. It is worth noting that we are considering two different spaces: the space $\tilde{\mathcal{P}}$ of finite measures and the space of prior distributions on the finite measures. Since we have two distinct spaces, we can choose two different geometries on each space. For example, if we consider the δ-geometry on the space $\tilde{\mathcal{P}}$ and the 1-geometry on the space of priors, we have the following cost function:

$$J(\Pi) = \gamma_e \int \Pi(\boldsymbol{\theta}) D_\delta(p_\theta, p_0) d\boldsymbol{\theta} + \gamma_u \int \Pi(\boldsymbol{\theta}) \log \Pi(\boldsymbol{\theta}) / \sqrt{g(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{4}$$

where γ_e is the confidence degree in the reference distribution p_0 and γ_u the uniformity degree. Considered independently, these two coefficients are not significant. However, their ratio is relevant in the following. The cost (4) can be rewritten as:

$$\begin{cases} J(\Pi) = \gamma_e E(\tau_0) + \gamma_u \int \Pi(\boldsymbol{\theta}) \log \Pi(\boldsymbol{\theta}) / \sqrt{g(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ \frac{\partial \tau_0}{\partial \boldsymbol{z}} = 0 \end{cases}$$

where $E(\tau_0)$ is the generalisation error of a fixed learning rule τ_0. By variational calculation, we obtain the solution of the minimization of the function (4):

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_0)} \sqrt{g(\boldsymbol{\theta})} \tag{5}$$

We note that if $\delta = 1$ then the cost function (4) is the kullback-Leibler divergence between the joint distributions of data and parameters as considered in [Rodriguez 1991, [2]] and if $\delta = 0$ we obtain the conjugate prior for exponential families (see examples in section VI). When the value of the ratio γ_e/γ_u goes to 0, we obtain the Jeffreys prior and when this ratio goes to ∞ we obtain the Dirac concentrated on p_0.

The model restriction to the parametric manifold $\mathcal{Q}$ is essentially for computational reasons. However, the reference distribution is a prior decision and does not depend on a post processing after collecting the data. Therefore, the reference distribution p_0 can be located in the whole space of probability measures. We can also have either a discrete set of N reference distributions $(p_0^i)_{i=1}^N$ weighted by $(\gamma_e^i)_{i=1}^N$ or a continuous set of reference distributions (a region or the whole set of probability distributions) with a probability measure $P(p_0)$ corresponding to the weights $(\gamma_e^i)_{i=1}^N$ in the discrete case. We show in the following that the prior solution Π has the same form as (5).

1. $p_0 \notin \mathcal{Q}$: When the reference distribution p_0 is located outside the model $\mathcal{Q}$, the δ-divergence $D_\delta(p_\theta, p_0)$ in the expression (4) can be decomposed according to the generalized Pythagore relation [Amari et al. 2000 [7]]:

   $$D_\delta(p_\theta, p_0) = D_\delta(p_\theta, p_0^\perp) + D_\delta(p_0^\perp, p_0)$$

   where $p_0^\perp$ is the $1-\delta$-projection of p_0 onto $\mathcal{Q}$ (see figure 5).

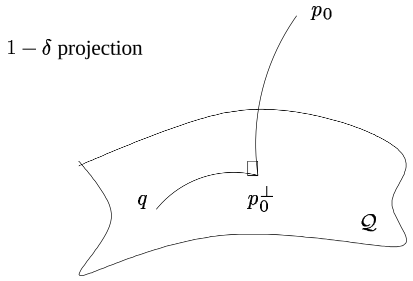

Figure 5. The equivalent of the non parametric reference distribution is its $1-\delta$ projection onto the parametric model $\mathcal{Q}$.

Giving the prior solution:

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_0^\perp)} \sqrt{g(\boldsymbol{\theta})}$$

2. When we have N reference distributions $\{(p_1,\gamma_1),...,(p_N,\gamma_N)\}$, the cost function (4) becomes:

$$J_N(\Pi) = \sum_{i=1}^{N} \gamma_i \int \Pi(\boldsymbol{\theta}) D_\delta(p_\theta, p_i) d\boldsymbol{\theta} + \gamma_u \int \Pi(\boldsymbol{\theta}) \log \Pi(\boldsymbol{\theta}) / \sqrt{g(\boldsymbol{\theta})} d\boldsymbol{\theta} \qquad (6)$$

If we define the $1-\delta$-barycentre p_G of the system $\{(p_1,\gamma_1),...,(p_N,\gamma_N)\}$ as

$$\overset{1-\delta}{l}(p_G) = \sum_{i=1}^{N} \gamma_i \overset{1-\delta}{l}(p_i) / \sum_{i=1}^{N} \gamma_i$$

and the $p_G^\perp$ the $1-\delta$ projection of p_G onto $\mathcal{Q}$, the solution Π of the minimization of (6) is:

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\sum \gamma_i}{\gamma_u} D_\delta(p_\theta, p_G^\perp)} \sqrt{g(\boldsymbol{\theta})}$$

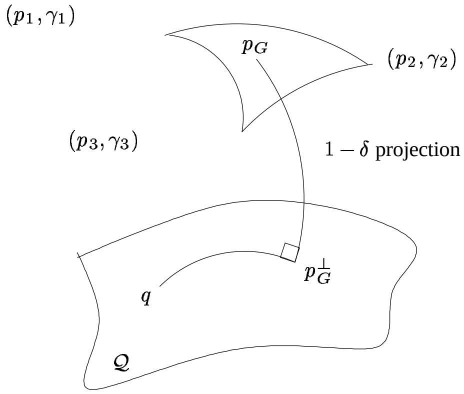

Figure 6. The equivalent reference distribution is the $1-\delta$ projection of the $1-\delta$ barycentre of the N references distributions.

3. When we have a continuous set $\mathcal{P}_r \subseteq \tilde{\mathcal{P}}$ of reference distributions with a mass distribution $P_r(p_0)$, the cost function is transformed to:

$$J_c(\Pi) = \int_{p_0 \in \mathcal{P}_r} P_r(p_0) \int \Pi(\boldsymbol{\theta}) D_\delta(p_\theta, p_0) d\boldsymbol{\theta} + \gamma_u \int \Pi(\boldsymbol{\theta}) \log \Pi(\boldsymbol{\theta}) / \sqrt{g(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{7}$$

In the same way, we define the $1-\delta$ barycentre p_G of $(\mathcal{P}_r, P_r)$ as:

$$\overset{1-\delta}{l}(p_G) = \int_{\mathcal{P}_r} P_r(p_0) \overset{1-\delta}{l}(p_0) / \int_{\mathcal{P}_r} P_r(p_0)$$

and the $p_G^\perp$ the $1-\delta$ projection of p_G onto $\mathcal{Q}$, the solution Π of the minimization of (7) is:

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\int P(p_0)}{\gamma_u} D_\delta(p_\theta, p_G^\perp)} \sqrt{g(\boldsymbol{\theta})}$$

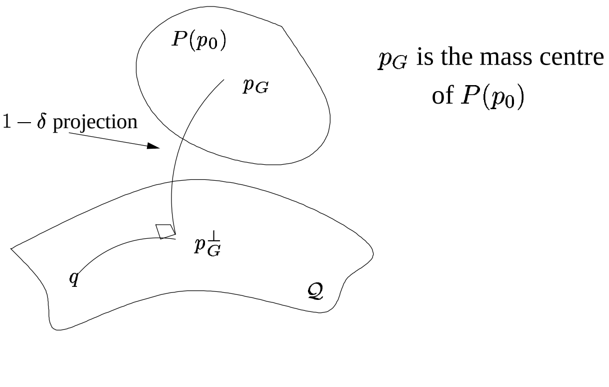

Figure 7. The equivalent reference distribution of a continum reference region is the $1-\delta$ projection of the $1-\delta$ expectation reference.

The above results show that whatever the choice of the reference distribution is, the resulting prior has the same form with a certain (non arbitrary) reference prior belonging to the model $\mathcal{Q}$. The existence of many reference distributions (or even a continuous set) indicates implicitly the existence of hyperparameter and the resulting solution shows that this hyperparameter is integrated and at the same time optimized if the *a priori* average (the barycentre) is considered as an optimization operation.

III. δ-FLAT FAMILIES

In this section we study the particular case of δ-flat families. $\mathcal{Q}$ is a δ flat manifold if and only if there exists a coordinate system $[\theta_i]$ such that the connection coefficients $\Gamma_\delta(\boldsymbol{\theta})$ are null. We call $[\theta_i]$ an affine coordinate system. It is known that δ-flatness is equivalent to $1-\delta$ flatness. Therefore, there exist dual affine coordinates $[\eta_i]$ such that $\Gamma_{1-\delta}(\boldsymbol{\eta}) = 0$. One of the many properties of δ-flat families is that we can express, in a simple way, the δ-divergence D_δ as a function of the coordinates $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ and thus any decision can be computed while manipulating the real coordinates. It is shown in [Amari 1985, [3]] that the dual affine coordinates $[\theta_i]$ and $[\eta_i]$ are related by Legendre transformations and the canonical divergence is:

$$D_\delta(p,q) = \psi(p) + \phi(q) - \theta_i(p)\eta_i(q)$$

where ψ and ϕ are the dual potentials such that:

$$\begin{cases} \frac{\partial \eta_j}{\partial \theta_i} = g_{ij} & \frac{\partial \theta_i}{\partial \eta_j} = g_{ij}^{-1} \\ \\ \partial_i \psi = \eta_i & \partial_i \phi = \theta_i \end{cases}$$

For example, the exponential families are 1-flat with the canonical parameters as 1-affine coordinates, the mixture family is 0-flat with the mixture coefficients as 0-affine coordinates, $\tilde{\mathcal{P}} = \{p, \int p < \infty\}$ is δ flat for all $\delta \in [0,1]$.

δ optimal estimates in δ flat families

As indicated in section II, the δ optimal estimate is the δ projection of $\int_\theta p^\delta P(\boldsymbol{\theta} \,|\, \boldsymbol{z})$ which is the minimizer of the functional $\int_\theta P(\boldsymbol{\theta} \,|\, \boldsymbol{z}) D_\delta(p_\theta, q)$. We see that, in general, the divergence as a function of the parameters $[\theta_i]$ has not a simple expression. However, with δ-flat manifolds, we obtain an explicit solution. Noting that:

$$\partial_i D_\delta(p_\theta, q) = D_\delta(p_\theta, (\partial_i)_q) = \theta_i(q) - \theta_i(p)$$

the solution is:

$$\hat{q} = q(\hat{\boldsymbol{\theta}}), \;\; \hat{\boldsymbol{\theta}} = \int \boldsymbol{\theta} P(\boldsymbol{\theta} \,|\, \boldsymbol{z}) d\boldsymbol{\theta} = E_{\theta \,|\, \boldsymbol{z}}[\boldsymbol{\theta}]$$

This means that the δ optimal estimate is the *a posteriori* expectation of the δ affine coordinates. Since the only degree of freedom of the affine coordinates is the affine transformation, this estimate is invariant under affine reparameterization.

Noting also that:

$$\partial_i D_{1-\delta}(p,q) = D_{1-\delta}(p,(\partial_i)_q) = \eta_i(q) - \eta_i(p)$$

Then the *a posteriori* expectation of the $1-\delta$ affine coordinates is the $1-\delta$ optimal estimate.

Prior selection with δ flat families

The δ prior Π has the following general expression:

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta,p_0)} \sqrt{g(\boldsymbol{\theta})}$$

where $p_0 \in \mathcal{Q}$ is the equivalent reference distribution in the manifold $\mathcal{Q}$. When we assume that $\mathcal{Q}$ is δ flat with affine coordinates $[\theta_i]$ and dual affine coordinates $[\eta_i]$, the expression of the prior becomes:

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u}\left(\psi(\boldsymbol{\theta}) - \theta_i \eta_i^0\right)} \sqrt{g(\theta)}$$

where $[\theta_i^0]$ and $[\eta_i^0]$ are the affine coordinates of p_0.

Therefore, we have an explicit analytic expression of the prior.

In the Euclidean case, that is when the connection ∇ is equal to its dual connection ∇^*, which is equivalent to equality of the affine coordinates $[\theta_i] = [\eta_i]$, the δ prior distribution is Gaussian with mean $\boldsymbol{\theta}_0$ and precision $2\,\frac{\gamma_e}{\gamma_u}$:

$$\Pi(\theta) \propto e^{-\frac{\gamma_e}{\gamma_u}||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2}$$

We detail here the notion of prior projection in the particular case of ∇^*-autoparallel submanifolds $\mathcal{Q}_a \subset \mathcal{Q}$. $\mathcal{Q}_a$ is $(1-\delta)$-autoparallel in $\mathcal{Q}$ if and only if, at every point $p \in \mathcal{Q}_a$, the covariant derivative $\nabla^*_{\partial_a}\partial_b$ remains in the tangent space $\mathcal{T}_p$ of the submanifold $\mathcal{Q}_a$ at the point p. A simple characterization in flat manifolds is that the $(1-\delta)$-affine coordinates $[u_i]$ of $\mathcal{Q}_a$ form an affine subspace of the coordinates $[\eta_i]$. We can show that by a suitable affine reparametrization of $\mathcal{Q}$, the submanifold $\mathcal{Q}_a$ is defined as:

$$\begin{cases} \mathcal{Q}_a = \{p_\eta \in \mathcal{Q}\,|\, \boldsymbol{\eta}_I = \boldsymbol{\eta}_I^0 \text{ is fixed }\} \\ \\ I \subset \{1..n\} \end{cases}$$

where $n - |I|$ is the dimension of $\mathcal{Q}_a$. If we consider the space $\mathcal{Q}_a^c$ such the complementary dual affine coordinates $\boldsymbol{\theta}_{II} = \boldsymbol{\theta}_{II}^0$ are fixed ($II = \{1..n\} - I$), then the tangent spaces $\mathcal{T}_p$ and $\mathcal{T}_p^c$ at the point $p(\boldsymbol{\eta}_I^0, \boldsymbol{\theta}_{II}^0)$ are orthogonal. Consequently, the projected prior from $\mathcal{Q}$ onto $\mathcal{Q}_a$ is simply:

$$\Pi^\perp(p) = \int_{q\in\mathcal{Q}_a^c} \Pi(q) = \int_{\theta_I} \Pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{II}) d\boldsymbol{\theta}_I$$

Hence, we see that the projected prior onto a ∇^*-autoparallel manifold is the marginalization in the δ affine coordinates and not in with respect to the $\boldsymbol{\eta}_I$ coordinates as it seems intuitive at a first look. This is essential due to the dual affine structure of the space $\tilde{\mathcal{P}}$.

IV. MIXTURE OF δ-FLAT FAMILIES AND SINGULARITIES

The mixture of distributions has attracted a great attention in that it gives a wider exploration of the probability distributions space based on a simple parametric manifold. For instance, by the mixture of Gaussians (which belongs to a 0-flat family) we can approach any probability distribution in total variation norm. In this section, we study the general case of the mixture of δ flat families. The space can be defined as:

$$\begin{cases} \mathcal{Q} = \{p_\theta \,|\, p_\theta = \sum_{j=1}^{k} w_j p_j(.\,;\boldsymbol{\theta}^j)\} \\ \\ p_j \in \mathcal{Q}_j, \;\; \mathcal{Q}_j \text{ is } \delta \text{ flat} \end{cases}$$

where the manifolds $\mathcal{Q}_j$ are either distinct or not.

The mixture distribution can be viewed as an incomplete model where the weighted sum is considered as a marginalization over the hidden variable z representing the label of the mixture. Thus $p_\theta = \sum_z p(z)p(x\,|\,z,\boldsymbol{\theta}_z)$ and the weights $p(z)$ are the parameters of a mixture family. We consider now the statistical learning problem within the mixture family. A mixture of δ flat families is not, in general, δ flat. Therefore the δ optimal estimates have no more a simple expression. However, with data augmentation procedure we can construct iterative algorithms computing the solution. Here, we focus on the computation of the δ prior of the mixture density.

The δ prior has the following expression:

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_0)} \sqrt{g(\boldsymbol{\theta})} \tag{8}$$

The mixture (marginalization) form of the distribution p_θ leads to a complex expression of the δ divergence and the determinant of the Fisher information. However, the computation of these expressions in the complete data distribution space [Rodriguez 2001, [8]] is feasible and gives explicit formula. By complete data $\boldsymbol{y}$, we mean the union of the observed data $\boldsymbol{x}$ and the hidden data $\boldsymbol{z}$. Therefore, the divergence will be considered between complete data distributions:

$$D_\delta(p^c, p_0^c) = \frac{\int p^c}{1-\delta} + \frac{\int p_0^c}{\delta} - \frac{\int (p^c)^\delta (p_0^c)^{1-\delta}}{\delta(1-\delta)}$$

where p^c is the complete likelihood $p(x,\, z\,|\,\boldsymbol{\theta})$ and $\boldsymbol{\theta}$ includes the parameters of the conditionals $p(x\,|\,z, \theta_z)$ and the discrete probabilities $p(z)$.

The additivity property of the δ-divergence is not conserved unless δ is equal to 0 or 1 [Amari1985, [3]]:

$$D_\delta(p_1p_2, q_1q_2) = D_\delta(p_1, q_1) + D_\delta(p_2, q_2) -$$

$$\delta\,(1-\delta) D_\delta(p_1, q_1) D_\delta(p_2, q_2)$$

Consequently, in the special case of $\delta \in \{0, 1\}$, we have the following simple formula:

$$\begin{cases} D_0(p, p_0) = \sum_{j=1}^{k} w_j^0 \left[D_0(p_j, p_j^0) + \log \frac{w_j^0}{w_j} \right] \\ \\ D_1(p, p_0) = \sum_{j=1}^{k} w_j \left[D_1(p_j, p_j^0) + \log \frac{w_j}{w_j^0} \right] \end{cases}$$

Singularities with mixture families

It is known that in learning the parameters of Gaussian mixture densities [Snoussi 2001] the maximum likelihood fails because of the degeneracy of the likelihood function to infinity when certain variances go to zero or certain covariance matrices approach the boundary of singularity. In [Snoussi 2001, [9]], there is an analysis of the occurrence of this situation in the multivariate Gaussian mixture case. In this section, we give a general condition leading to this problem of degeneracy occurring in the learning within the mixture of δ flat families.

Let $\mathcal{Q}$ a δ flat manifold and $[\theta_i]$ the natural affine coordinates and $[\eta_i]$ the dual affine coordinates. The two coordinate systems are related by Legendre transformation [Amari 1985, [3]]:

$$\begin{cases} \frac{\partial \eta_j}{\partial \theta_i} = g_{ij} \quad \frac{\partial \theta_i}{\partial \eta_j} = g_{ij}^{-1} \\ \\ \partial_i \psi = \eta_i \quad \partial_i \phi = \theta_i \end{cases}$$

where $(g_{ij})_{i=1..n}^{j=1..n}$ is the Fisher matrix and ψ and ϕ are the dual potentials.

It is clear from the expression of the variable transformation between the two affine coordinates that a singularity of the Fisher information matrix g leads to non differentiability in the transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. A singularity of g means that the determinant of this matrix is zero. Therefore, it is interesting to study the behaviour of the dual divergence at the boundary of singularity and we will show in an example that the dual divergences may have different behaviour as the distribution p approaches the boundary of singularity.

To illustrate such behaviour, we take a Gaussian family $\{\mathcal{N}(\mu, \sigma^2) \,|\, \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+\}$ which is a 2-dimensional statistical manifold 0-flat. The 0-affine coordinates are $\boldsymbol{\theta}$ and the 1-affine coordinates are $\boldsymbol{\eta}$ given by the following expressions:

$$\begin{cases} \theta_1 = \frac{\mu}{\sigma^2}, \quad \theta_2 = \frac{-1}{2\sigma^2} \\ \\ \eta_1 = \mu, \quad \eta_2 = \mu^2 + \sigma^2 \end{cases} \qquad (9)$$

The corresponding Fisher information are:

$$|g(\theta)| \propto \sigma^6, \quad |g(\eta)| \propto 1/\sigma^6 \tag{10}$$

The canonical divergence has the following expression:

$$D_\delta(p_1,p_2) = D_{1-\delta}(p_2,p_1) = \psi(p_1) + \phi(p_2) - \theta_i(p_1)\,\eta_i(p_2) \tag{11}$$

where ψ and ϕ are the potentials given by:

$$\psi = \frac{\mu^2}{2\sigma^2} + \log\sqrt{2\pi}\sigma, \quad \phi = \frac{-1}{2} - \log\sqrt{2\pi}\sigma \tag{12}$$

We see that the degeneracy occurs when the variance σ goes to zero. A detailed study of how this degeneracy occurs in the Gaussian mixture case is in [Snoussi 2001, [9]] and is reviewed in the example of the next section. Here we focus on the difference of behaviour of the two canonical divergences D_0 and D_1.

The expression of the δ prior is:

$$\Pi_\delta \propto e^{-D_\delta(p_\theta,p_0)}\sqrt{g(\theta)}$$

Following the complete data procedure:

$$\begin{cases} \Pi_0 \propto e^{-\frac{\gamma_e}{\gamma_u}\sum w_{i0}\{D_0(p_\theta^i,p_0^i)+\log\frac{w_{i0}}{w_i}\}}\sqrt{g(\boldsymbol{\theta},\boldsymbol{w})} \\ \Pi_1 \propto e^{-\frac{\gamma_e}{\gamma_u}\sum w_i\{D_1(p_\theta^i,p_0^i)+\frac{w_i}{w_{i0}}\}}\sqrt{g(\boldsymbol{\theta},\boldsymbol{w})} \end{cases}$$

The resulting prior is factorized and separated into independent priors on the components of the Gaussian mixture. Combining expressions of (9), (10), (11) and (12) we note the following comparison of the 0 and 1 priors through their dependences on the variance σ_j:

$$\begin{array}{c|c} \delta = 0 & \delta = 1 \\ \downarrow & \downarrow \\ p \longrightarrow \partial\mathcal{Q} & p \longrightarrow \partial\mathcal{Q} \\ \Pi_0 \text{ is } O(\sigma_j^\alpha e^{-k_0/\sigma_j^2}) & \Pi_1 \text{ is } O(\sigma_j^{2w_j\frac{\gamma_j}{\gamma_u}}) \\ \downarrow & \downarrow \\ \text{Exponential} & \text{Polynomial} \end{array}$$

where α, k_0 are constant.

We note that:

- For $\delta = 0$, the prior decreases to 0 when p approaches the boundary of singularity $\partial\mathcal{Q}$ with an **exponential** term leading to an inverse Gamma prior for the variance.
- For $\delta = 1$, the prior decreases to 0 when p approaches the boundary of singularity $\partial\mathcal{Q}$ with a **polynomial** term leading to a Gamma prior for the variance. We note the presence of the parameter w_i in the power term.

This kind of behaviour pushes us to use the 0 prior in that it is able to eliminate the degeneracy of the likelihood function.

V. EXAMPLES

In this section we develop the δ prior in 2 learning problems: Multivariate Gaussian mixture and blind source separation and segmentation.

V.1/ Multivariate Gaussian mixture

The multivariate Gaussian mixture distribution of $\boldsymbol{x} \in \mathbb{R}^n$ is:

$$p(\boldsymbol{x}_i) = \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{x}_i\,; \boldsymbol{m}_k, \boldsymbol{R}_k) \tag{13}$$

where w_k, $\boldsymbol{m}_k$ and $\boldsymbol{R}_k$ are the weight, mean and covariance of the cluster k. This can be interpreted as an incomplete data problem where the missing data are the labels $(z_i)_{i=1..T}$ of the clusters. Therefore, the mixture (13) is considered as a marginalization over z:

$$p(\boldsymbol{x}_i) = \sum_{z_i} p(z_i)\, \mathcal{N}(\boldsymbol{x}_i \,|\, z_i, \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ is the set of the unknown means and covariances. Our objective is the prediction of the future observations given the trained data $\boldsymbol{x}_i$, $i = 1..T$. The whole parameter characterizing the statistical model is $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{w})$. We consider now the derivation of the δ prior for $\delta \in \{0\,,1\}$ and compare the two resulting priors.

The δ prior has the following form:

$$\Pi_\delta(\boldsymbol{\eta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_{\boldsymbol{\eta}}, p_0)} \sqrt{g(\boldsymbol{\eta})}$$

Therefore, we have to compute the D_δ divergence and the Fisher information matrix. As noted in the previous section and following [Rodriguez 2001, [8]], the computation is considered in the complete data space $(\mathcal{X} \times \mathcal{Z})^T$ of observations $\boldsymbol{x}_i$ and labels z_i, T is the number of observations. In fact, we mean the number of virtual observations as the construction of the prior precedes the real observations. We have:

$$\begin{cases} D_0(\boldsymbol{\eta} : \boldsymbol{\eta}^0) = \underset{\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta}_0}{E} \left[\log \frac{p(\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta}_0)}{p(\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta})} \right] \\ D_1(\boldsymbol{\eta} : \boldsymbol{\eta}^0) = \underset{\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta}}{E} \left[\log \frac{p(\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta})}{p(\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta}_0)} \right] \\ g_{ij}(\boldsymbol{\eta}) = -\underset{\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta}}{E} \left[\frac{\partial^2}{\partial_i \partial_j} \log p(\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta}) \right] \end{cases}$$

By classifying the labels $\boldsymbol{z}_{1..T}$ and using the sequential Bayes rule between $\boldsymbol{x}_{1..T}$ and $\boldsymbol{z}_{1..T}$, the δ divergences become:

$$\begin{cases} D_0(\boldsymbol{\eta} : \boldsymbol{\eta}^0) = T \sum_{i=1}^{k} w_i^0 \left(D_0(\mathcal{N}_i : \mathcal{N}_i^0) + \log \frac{w_i^0}{w_i} \right) \\ \\ D_1(\boldsymbol{\eta} : \boldsymbol{\eta}^0) = T \sum_{i=1}^{k} w_i \left(D_1(\mathcal{N}_i : \mathcal{N}_i^0) + \log \frac{w_i}{w_i^0} \right) \end{cases}$$

where $D_0(\mathcal{N}_i : \mathcal{N}_i^0) = D_1(\mathcal{N}_i^0 : \mathcal{N}_i)$ is the 0 divergence between two multivariate Gaussians:

$$\begin{cases} D_0(\mathcal{N}_i \,\|\, \mathcal{N}_i^0) = \frac{1}{2} \left(\log |\boldsymbol{R}_i \boldsymbol{R}_{i0}^{-1}| + \mathrm{Tr} \left(\boldsymbol{R}_{i0} \boldsymbol{R}_i^{-1} \right) - n + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i0})^* \boldsymbol{R}_i^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i0}) \right) \\ \\ D_1(\mathcal{N}_i \,\|\, \mathcal{N}_i^0) = D_0(\mathcal{N}_i^0 \,\|\, \mathcal{N}_i) \end{cases}$$

The Fisher matrix is block diagonal with K diagonal blocks corresponding to the components of the mixture. Each block g_i with size $(n + n^2 + 1)$ has also a diagonal form (n is the dimension of the vector $\boldsymbol{x}_t$):

$$g = \begin{bmatrix} [g_1] & & \\ & \ddots & \\ & & [g_K] \end{bmatrix} \quad , \quad g_i = \begin{bmatrix} w_i \, g_{\mathcal{N}}(\boldsymbol{m}_i, \boldsymbol{R}_i) & [0] \\ \\ [0] & 1/w_i \end{bmatrix}$$

where $g_{\mathcal{N}}$ is the Fisher matrix of the multivariate Gaussian and has the following expression:

$$g_{\mathcal{N}}(\boldsymbol{m}, \boldsymbol{R}) = \begin{bmatrix} \boldsymbol{R}^{-1} & [0] \\ \\ [0] & -\frac{1}{2} \frac{\partial \boldsymbol{R}^{-1}}{\partial \boldsymbol{R}} \end{bmatrix}$$

whose determinant is:

$$|g_{\mathcal{N}}(\boldsymbol{m}, \boldsymbol{R})| = |\boldsymbol{R}|^{-(n+2)}$$

Thus, the determinant of the block g_i is:

$$|g_i(w_i, \boldsymbol{m}_i, \boldsymbol{R}_i)| = \left(\frac{1}{2} \right)^{n^2} w_i^{(n^2+n-1)} \, |\boldsymbol{R}_i|^{-(n+2)} \tag{14}$$

The additional form of the $\{0, 1\}$ divergences (implying the multiplicative form of their exponentials) and the multiplicative form of the determinant of the Fisher matrix (due to its block diagonal form) lead to an independent priors of the components $\boldsymbol{\eta}_i = (w_i, \boldsymbol{m}_i, \boldsymbol{R}_i)$: $\Pi(\boldsymbol{\eta}) = \prod_{k=1}^{K} \Pi(\boldsymbol{\eta}_i)$. The two values of $\delta = \{0, 1\}$ lead to two different priors Π_δ:

- $\delta = 0$:

$$\begin{aligned} \Pi_0(\boldsymbol{\eta}_i) \quad & \propto \quad \exp \left[-\frac{\gamma_e}{\gamma_u} \left(w_i^0 D_0(\mathcal{N}_i : \mathcal{N}_i^0) + w_i^0 \log \frac{w_i^0}{w_i} \right) \right] \sqrt{|g_i(\boldsymbol{\eta}_i)|} \\ & \propto \quad \mathcal{N} \left(\boldsymbol{m}_i ; \boldsymbol{m}_0, \frac{\boldsymbol{R}_i}{\alpha w_i^0} \right) \mathcal{W}_n \left(\boldsymbol{R}_i^{-1} ; \nu_0, \boldsymbol{R}_0^{-1} \right) w_i^{\beta_0} \end{aligned} \tag{15}$$

with,

$$\alpha = \tfrac{\gamma_e}{\gamma_u}, \quad \nu_0 = \alpha w_i^0, \quad \beta_0 = \alpha w_i^0 + \tfrac{n^2+n-1}{2}$$

$\mathcal{W}_n$ is the wishart distribution of an $n \times n$ matrix:

$$\mathcal{W}_n(\boldsymbol{R}; \nu, \boldsymbol{\Sigma}) \propto |\boldsymbol{R}|^{\frac{\nu-(n+1)}{2}} \exp\left[-\frac{\nu}{2}\mathrm{Tr}\left(\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\right)\right]$$

The 0-prior is Normal Inverse Wishart for the mean and covariance $(\boldsymbol{m}_i, \boldsymbol{R}_i)$ and Dirichlet for the weight w_i, that is the **conjugate** prior.

- $\delta = 1$:

$$\begin{aligned} \Pi_1(\boldsymbol{\eta}_i) &\propto \exp\left[-\tfrac{\gamma_e}{\gamma_u}\left(w_i D_1(\mathcal{N}_i : \mathcal{N}_i^0) + w_i \log \tfrac{w_i}{w_i^0}\right)\right] \sqrt{|g_i(\boldsymbol{\eta}_i)|} \\ &\propto \mathcal{N}\left(\boldsymbol{m}_i ; \boldsymbol{m}_0, \tfrac{\boldsymbol{R}_i}{\alpha w_i}\right) \mathcal{W}_n\left(\boldsymbol{R}_i ; \alpha w_i - 1, \tfrac{\alpha w_i - 1}{\alpha w_i}\boldsymbol{R}_0\right) \qquad (16) \\ & \quad w_i^{\frac{n^2+n-1}{2} - (1+\frac{n}{2})\alpha w_i} \, (w_i^0)^{\alpha w_i} \, \Gamma_n\left(\tfrac{\alpha w_i - 1}{2}\right) \end{aligned}$$

where Γ_n is the generalized Gamma function of dimension n ([6] page 427):

$$\Gamma_n(b) = \left[\Gamma(\frac{1}{2})\right]^{\frac{1}{2}n(n-1)} \prod_{i=1}^{n} \Gamma(b + \frac{i-n}{2}), \;\; b > \frac{n-1}{2}$$

The 1-prior Π_1 (16) is the generalized entropic prior [Rodriguez 2001, [8]] to the multivariate case. We see that the prior Π_1 is a **Wishart** function of the covariance matrices $\boldsymbol{R}_i$ and the prior Π_0 is an **inverse Wishart** function of the covariances. This leads to a difference of the behaviour of these functions on the boundary of singularity (the set of singular matrices).

V.2/ Source separation

The second example deals with the source separation problem. The observations $\boldsymbol{x}_{1..T}$ are T samples of m-vectors. At each time t, the vector data $\boldsymbol{x}_t$ is supposed to be a noisy instantaneous mixture of an observed n-vector source $\boldsymbol{s}_t$ with unknown mixing coefficients forming the mixing matrix $\boldsymbol{A}$. This is simply modeled by the following equation:

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{n}_t, \; t = 1..T$$

where given the data $\boldsymbol{x}_{1..T}$, our objective is the recovering of the original sources $\boldsymbol{s}_{1..T}$ and the unknown matrix $\boldsymbol{A}$. The Bayesian approach taken to solve this inverse problem [Knuth 1998 [10], Djafari 1999 [11], Snoussi 2002, [12]] needs also the estimation of the noise covariance matrix $\boldsymbol{R}_n$ and the learning of the statistical parameters of the original sources $\boldsymbol{s}_{1..T}$. In the following, we suppose that the sources are statistically independent and that each source is modeled by a mixture of univariate Gaussians, so that we have to learn each set of source j parameters $\boldsymbol{\eta}^j$ which contains the weights, means and variances composing the mixture j:

$$\begin{cases} \boldsymbol{\eta}^j = \left(\boldsymbol{\eta}^j_i\right)_{i=1..K_j} \\ \\ \boldsymbol{\eta}^j_i = (w^j_i, m^j_i, \sigma^j_i) \end{cases}$$

The index j indicates the source j and i indicates the Gaussian component i of the distribution of the source j. Therefore we don't have a multidimensional Gaussian mixture but instead independent unidimensional Gaussian mixtures.

In the following, our parameter of interest is $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{R}_n, \boldsymbol{\eta})$: the mixing matrix $\boldsymbol{A}$, the noise covariance $\boldsymbol{R}_n$ and $\boldsymbol{\eta}$ contains all the parameters of the sources model. Our objective is the computation of the δ priors for $\delta \in \{0, 1\}$. We have an incomplete data problem with two hierarchies of hidden variables, the sources $\boldsymbol{s}_{1..T}$ and the labels $\boldsymbol{z}_{1..T}$ so that the complete data are $(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T})$. We begin by the computation of the Fisher information matrix which is common to the both geometries.

a/ Fisher information matrix

The Fisher matrix $\mathcal{F}(\boldsymbol{\theta})$ is defined as:

$$\mathcal{F}_{ij}(\boldsymbol{\theta}) = -\mathop{E}_{\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T}} \left[\frac{\partial^2}{\partial_i \partial_j} \log p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta}) \right]$$

The factorization of the joint distribution $p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta})$ as:

$$p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta}) = p(\boldsymbol{x}_{1..T} \,|\, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T}, \theta)\, p(\boldsymbol{s}_{1..T} \,|\, \boldsymbol{z}_{1..T}, \theta)\, p(\boldsymbol{z}_{1..T} \,|\, \theta)$$

and the corresponding expectations as

$$\mathop{E}_{\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T}} [.] = \mathop{E}_{\boldsymbol{z}_{1..T}} [.] \mathop{E}_{\boldsymbol{s}_{1..T} \,|\, \boldsymbol{z}_{1..T}} [.] \mathop{E}_{\boldsymbol{x}_{1..T} \,|\, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T}} [.]$$

and taking into account the conditional independencies $((\boldsymbol{x}_{1..T} \,|\, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T}) \Leftrightarrow (\boldsymbol{x}_{1..T} \,|\, \boldsymbol{s}_{1..T})$ and $(\boldsymbol{s}_{1..T} \,|\, \boldsymbol{z}_{1..T}) \Leftrightarrow \prod \boldsymbol{s}^j_{1..T} \,|\, \boldsymbol{z}^j_{1..T})$, the Fisher information matrix will have a block diagonal structure as follows:

$$g(\boldsymbol{\theta}) = \begin{bmatrix} g(\boldsymbol{A}, \boldsymbol{R}_n) & \dots & & [0] \\ \vdots & g(\boldsymbol{\eta}^1) & & \\ & & \ddots & \\ & & & \\ [0] & & \dots & g(\boldsymbol{\eta}^n) \end{bmatrix}$$

a.1/ $(\boldsymbol{A}, \boldsymbol{R}_n)$-block

The Fisher information matrix of $(\boldsymbol{A}, \boldsymbol{R}_n)$ is:

$$\mathcal{F}_{ij}(\boldsymbol{A}, \boldsymbol{R}_n) = -\mathop{E}_{\boldsymbol{s}} \mathop{E}_{\boldsymbol{x} \,|\, \boldsymbol{s}} \left[\frac{\partial^2}{\partial_i \partial_j} \log p(\boldsymbol{x}_{1..T} \,|\, \boldsymbol{s}_{1..T}, \boldsymbol{A}, \boldsymbol{R}_n) \right]$$

which is very similar to the Fisher information matrix of the mean and covariance of a multivariate Gaussian distribution. The obtained expression is

$$g(\boldsymbol{A}, \boldsymbol{R}_n) = \begin{bmatrix} \left(\underset{\boldsymbol{s}_{1..T}}{E} \boldsymbol{R}_{ss}\right) \otimes \boldsymbol{R}_n^{-1} & [0] \\ & \\ [0] & -\frac{1}{2}\frac{\partial \boldsymbol{R}_n^{-1}}{\partial \boldsymbol{R}_n} \end{bmatrix}$$

where $\boldsymbol{R}_{ss} = \frac{1}{T}\sum \boldsymbol{s}_t \boldsymbol{s}_t^*$ and $\otimes$ is the Kronecker product.

We note the block diagonality of the $(\boldsymbol{A}, \boldsymbol{R}_n)$-Fisher matrix. The term corresponding to the mixing matrix $\boldsymbol{A}$ is the signal to noise ratio as can be expected. Thus, the amount of information about the mixing matrix is proportional to the signal to noise ratio. The induced volume of $(\boldsymbol{A}, \boldsymbol{R}_n)$ is then:

$$|g(\boldsymbol{A}, \boldsymbol{R}_n)|^{1/2}\, d\boldsymbol{A}\, d\boldsymbol{R}_n = \frac{|\underset{\boldsymbol{\eta}}{E} \boldsymbol{R}_{ss}|^{m/2}}{|\boldsymbol{R}_n|^{\frac{m+n+1}{2}}}\, d\boldsymbol{A}\, d\boldsymbol{R}_n$$

a.$2/$ $(\boldsymbol{\eta}^j)$**-block**

Each $g(\boldsymbol{\eta}^j)$ is the Fisher information of a one-dimensional Gaussian distribution. Therefore, it is obtained by setting $n = 1$ in the expression (14) of the previous section:

$$\left|g(\boldsymbol{\eta}^j)\right|^{1/2} d\boldsymbol{\eta}^j = \left\{ \prod_{i=1}^{K_j} \frac{w_i^{1/2}}{v_i^{3/2}} \right\} d\boldsymbol{\eta}^j$$

b$/$ δ**-Divergence (**$\delta = 0, 1$**)**

The δ-divergence between two parameters $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{R}_n, \boldsymbol{\eta})$ and $\boldsymbol{\theta}^0 = (\boldsymbol{A}^0, \boldsymbol{R}_n^0, \boldsymbol{\eta}^0)$ for the complete data likelihood $p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta})$ is:

$$\begin{cases} D_0(\boldsymbol{\theta} : \boldsymbol{\theta}^0) = \underset{x,s,z|\theta^0}{E} \log \frac{p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta}^0)}{p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta})} \\ \\ D_1(\boldsymbol{\theta} : \boldsymbol{\theta}^0) = \underset{x,s,z|\theta}{E} \log \frac{p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta})}{p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta}^0)} \end{cases}$$

Similar developments of the above equation as in the computation of the Fisher matrix based on the conditional independencies, lead to an affine form of the divergence, which is a sum of the expected divergence between the $(\boldsymbol{A}, \boldsymbol{R}_n)$ parameters and the divergence between the sources parameters $\boldsymbol{\eta}$:

$$\begin{cases} D_0(\boldsymbol{\theta} : \boldsymbol{\theta}^0) = \underset{s|\eta^0}{E}\, \underset{|s}{D_0}(\boldsymbol{A}, \boldsymbol{R}_n \,:\, \boldsymbol{A}^0, \boldsymbol{R}_n^0) + D_0(\boldsymbol{\eta} \,:\, \boldsymbol{\eta}^0) \\ \\ D_1(\boldsymbol{\theta} : \boldsymbol{\theta}^0) = \underset{s|\eta}{E}\, \underset{|s}{D_1}(\boldsymbol{A}, \boldsymbol{R}_n \,:\, \boldsymbol{A}^0, \boldsymbol{R}_n^0) + D_1(\boldsymbol{\eta} \,:\, \boldsymbol{\eta}^0) \end{cases}$$

where $\underset{|s}{D_\delta}$ means the divergence between the distributions $p(\boldsymbol{x}_{1..T} \,|\, \boldsymbol{A}, \boldsymbol{R}_n, \boldsymbol{s}_{1..T})$ and $p(\boldsymbol{x}_{1..T} \,|\, \boldsymbol{A}^0, \boldsymbol{R}_n^0, \boldsymbol{s}_{1..T})$ keeping the sources $\boldsymbol{s}_{1..T}$ fixed.

The δ-divergence between $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_0$ is the sum of the δ-divergences between each source parameter $\boldsymbol{\eta}^j$ and $\boldsymbol{\eta}_0^j$ due to the *a priori* independence between the sources. Then, the divergence between $\boldsymbol{\eta}^j$ and $\boldsymbol{\eta}_0^j$ is obtained as a particular case ($n = 1$) of the general expression derived in the multivariate case. Therefore we have the same form of the prior as in equations (15) and (16).

The expressions of the averaged divergences between the $(\boldsymbol{A}, \boldsymbol{R}_n)$ parameters are:

$$\left\{ \begin{array}{rl} \underset{s|\eta^0}{E} \underset{|s}{D_0}(\boldsymbol{A}, \boldsymbol{R}_n : \boldsymbol{A}_0, \boldsymbol{R}_{n0}) = & \frac{1}{2}\left(\log\left|\boldsymbol{R}_n \boldsymbol{R}_{n0}^{-1}\right| + \mathrm{Tr}\,(\boldsymbol{R}_n^{-1} \boldsymbol{R}_{n0})\right. \\ & \left. +\mathrm{Tr}\left(\boldsymbol{R}_n^{-1}(\boldsymbol{A} - \boldsymbol{A}_0) \underset{s|\eta^0}{E}[\boldsymbol{R}_{ss}](\boldsymbol{A} - \boldsymbol{A}_0)^*\right)\right) \\ \underset{s|\eta}{E} \underset{|s}{D_1}(\boldsymbol{A}, \boldsymbol{R}_n : \boldsymbol{A}_0, \boldsymbol{R}_{n0}) = & \frac{1}{2}\left(\log|\boldsymbol{R}_{n0} \boldsymbol{R}_n^{-1}| + \mathrm{Tr}\left(\boldsymbol{R}_{n0}^{-1} \boldsymbol{R}_n\right)\right. \\ & \left. +\mathrm{Tr}\left(\boldsymbol{R}_{n0}^{-1}(\boldsymbol{A} - \boldsymbol{A}_0) \underset{s|\eta}{E}[\boldsymbol{R}_{ss}](\boldsymbol{A} - \boldsymbol{A}_0)^*\right)\right) \end{array} \right.$$

leading to the following δ priors on $(\boldsymbol{A}, \boldsymbol{R}_n)$:

$$\left\{ \begin{array}{lcl} \Pi_0(\boldsymbol{A}, \boldsymbol{R}_n^{-1}) & \propto & \mathcal{N}\left(\boldsymbol{A}; \boldsymbol{A}_0, \frac{1}{\alpha}{\boldsymbol{R}_{ss}^0}^{-1} \otimes \boldsymbol{R}_n\right) \mathcal{W}_{im}\left(\boldsymbol{R}_n^{-1}; \alpha, {\boldsymbol{R}_n^0}^{-1}\right) |\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]|^{\frac{m}{2}} \\ \Pi_1(\boldsymbol{A}, \boldsymbol{R}_n) & \propto & \mathcal{N}\left(\boldsymbol{A}; \boldsymbol{A}_0, \frac{1}{\alpha}\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]^{-1} \otimes \boldsymbol{R}_n^0\right) \mathcal{W}_{im}\left(\boldsymbol{R}_n; \alpha - n, \frac{\alpha - n}{\alpha}\boldsymbol{R}_n^0\right) \end{array} \right.$$

Therefore, the 0-prior is a normal inverse Wishart prior (conjugate prior). The mixing matrix and the noise covariance are not *a priori* independent. In fact, the covariance matrix of $\boldsymbol{A}$ is the noise to signal ratio $\frac{1}{\alpha}{\boldsymbol{R}_{ss}^0}^{-1} \otimes \boldsymbol{R}_n$. We note a multiplicative term which is a power of the determinant of the *a priori* expectation of the source covariance $\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]$. This term can be injected in the prior $p(\boldsymbol{\eta})$ and thus the $(\boldsymbol{A}, \boldsymbol{R}_n)$ parameters and the $\boldsymbol{\eta}$ parameters are *a priori* independent.

The 1-prior (entropic prior) is normal Wishart. The mixing matrix and the noise covariance are *a priori* independent since the noise to signal ratio $\frac{1}{\alpha}\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]^{-1} \otimes \boldsymbol{R}_n^0$ depend on the reference parameter $\boldsymbol{R}_n^0$. However, we have in counterpart the dependence of $\boldsymbol{A}$ and $\boldsymbol{\eta}$ through the term $\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]^{-1}$ present in the covariance matrix of $\boldsymbol{A}$. In practice, we prefer to replace the expected covariance $\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]$, in the two priors, by its reference value $\boldsymbol{R}_{ss}^0$.

We note that the precision matrix for the mixing matrix $\boldsymbol{A}$ ($\alpha \boldsymbol{R}_{ss}^0 \otimes \boldsymbol{R}_n^{-1}$ for Π_0 and $\alpha \underset{s|\eta}{E}[\boldsymbol{R}_{ss}] \otimes {\boldsymbol{R}_n^0}^{-1}$ for Π_1) is the product of the confidence term $\alpha = \frac{\gamma_e}{\gamma_u}$ in the reference parameters and the signal to noise ratio. Therefore, the resulting precision of the reference matrix $\boldsymbol{A}_0$ is not only our *a priori* coefficient γ_e but the product of this coefficient and the signal to noise ratio.

VI. CONCLUSION AND DISCUSSION

In this paper, we have shown the importance of providing a geometry (a measure of distinguishibility) to the space of distributions. A different geometry will give a different learning rule mapping the training data to the space of predictive distributions. The prior selection procedure established in a statistical decision framework needs to be taken in a specified geometry. We have tried to elucidate the interaction between the parametric and non parametric modeling. The notion of "projected mass" gives to the restricted parametric modelization a non parametric sense and shows the role of the relative geometry of the parametric model in the whole space of distributions. The same investigations are considered in the interaction between a curved family and the whole parametric model containing it. Exact expressions are shown in a simple case of auto-parallel families and we are working on the more abstract space of distributions.

REFERENCES

1. R. E. Kass and L. Wasserman, “Formal rules for selecting prior distributions: A review and annotated bibliography”, Technical report no. 583, Department of Statistics, Carnegie Mellon University, 1994.
2. C. Rodríguez, “Entropic priors”, *Tech. rep. Electronic form* `http:omega.albany.edu:8008/entpriors.ps`, (1991).
3. S. Amari, *Differential-Geometrical Methods in Statistics*, Volume 28 of Springer Lecture Notes in Statistics, Springer-Verlag, New York, 1985.
4. H. Zhu and R. Rohwer, “Bayesian invariant measurements of generalisation”, in *Neural Proc. Lett.*, 1995, vol. 2 (6), pp. 28–31.
5. H. Zhu and R. Rohwer, “Bayesian invariant measurements of generalisation for continuous distributions”, Technical report, NCRG/4352, ftp://cs.aston.ac.uk/neural/zhuh/continuous.ps.z, Aston University, 1995.
6. G. E. P. Box and G. C. Tiao, *Bayesian inference in statistical analysis*, Addison-Wesley publishing, 1972.
7. S. Amari and H. Nagaoka, *Methods of Information Geometry*, vol. Volume 191 of Translations of Mathematical Monographs, AMS, OXFORD, University Press, 2000.
8. C. Rodríguez, “Entropic priors for discrete probabilistic networks and for mixtures of Gaussians models”, in *Bayesian Inference and Maximum Entropy Methods*, R. L. FRY, Ed. MaxEnt Workshops, August 2001, pp. 410–432, Amer. Inst. Physics.
9. H. Snoussi and A. Mohammad-Djafari, “Penalized maximum likelihood for multivariate gaussian mixture”, in *Bayesian Inference and Maximum Entropy Methods*, R. L. Fry, Ed. MaxEnt Workshops, August 2002, pp. 36–46, Amer. Inst. Physics.
10. K. Knuth, “A Bayesian approach to source separation”, in *Proceedings of Independent Component Analysis Workshop*, 1999, pp. 283–288.
11. A. Mohammad-Djafari, “A Bayesian approach to source separation”, in *Bayesian Inference and Maximum Entropy Methods*, J. R. G. Erikson and C. Smith, Eds., Boise, IH, July 1999, MaxEnt Workshops, Amer. Inst. Physics.
12. H. Snoussi and A. Mohammad-Djafari, “MCMC Joint Separation and Segmentation of Hidden Markov Fields”, in *Neural Networks for Signal Processing XII*. IEEE workshop, September 2002, pp. 485–494.