The geometry of prior selection

Hichem Snoussi*

to appear in NeuroComputing, 2005

Abstract

This contribution is devoted to the selection of prior in a Bayesian learning framework. There is an extensive literature on the construction of non informative priors and the subject seems far from a definite solution [1]. We consider this problem in the light of the recent development of information geometric tools [2]. The differential geometric analysis allows the formulation of the prior selection problem in a general manifold valued set of probability distributions. In order to construct the prior distribution, we propose a criteria expressing the trade off between decision error and uniformity constraint. The solution has an explicit expression obtained by variational calculus. In addition, it has two important invariance properties: invariance to the dominant measure of the data space and also invariance to the parametrization of a restricted parametric manifold. 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 autoparallel restricted families. Two practical examples illustrate the proposed construction of prior. The first example deals with the learning of a mixture of multivariate Gaussians in a classification perspective. We show in this learning problem how the penalization of likelihood by the proposed prior eliminates the degeneracy occurring when approaching singularity points. The second example treats the blind source separation problem.

Keywords

Differential geometry, prior selection, Bayesian learning, mixture of Gaussians, blind source separation.

I. Introduction

A learning machine can be described as a system mapping some inputs $\boldsymbol{x}$ to some outputs $\boldsymbol{y}$ (see Figure 1). The inputs $\boldsymbol{x}$ and the outputs $\boldsymbol{y}$ lie in two general sets either euclidian or not. The learning of the machine consists essentially in extracting information from some collected data in order to perform a specific task related to the behavior of the modeled machine. The distinction inputs/outputs is not in general related to the task performed by the learning machine. For instance, filtering consists in estimating the inputs $\boldsymbol{x}$ given the outputs $\boldsymbol{y}$. The inference consists in finding some parameter $\boldsymbol{\theta}$ characterizing the mapping $\boldsymbol{y} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$. The prediction consists in estimating the stochastic behavior of the outputs given some previous recorded data, and so on. 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 (maximum likelihood) to estimate either the parameter $\boldsymbol{\theta}$ or the inputs $\boldsymbol{x}$ given the data (outputs) $\boldsymbol{y}$. 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 $\boldsymbol{\theta}$. The problem of prediction can be stated as follows: given some training data $D = (\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1..N}$, where i is the time index and N is the sample size, our purpose is the estimation of the output probability distribution (prediction). In this work, we focus on the prediction problem. We note that, in this situation, before designing the learning algorithm, one is confronted to two important questions: (i) how to choose the parametric model $p(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\theta})$ (model selection), (ii) how to select a prior distribution on the parameter $\boldsymbol{\theta}$. In words, the first question concerns the selection of an appropriate manifold in the whole set of probability distributions $\mathcal{P}$, on which the learning algorithm will estimate the

Hichem Snoussi is with IRCCyN, Institut de Recherche en Communications et Cybernétiques de Nantes, Ecole Centrale de Nantes, 1, Rue de la Noë, BP 92101, 44321, Nantes, France. Email = snoussi@irccyn.ec-nantes.fr

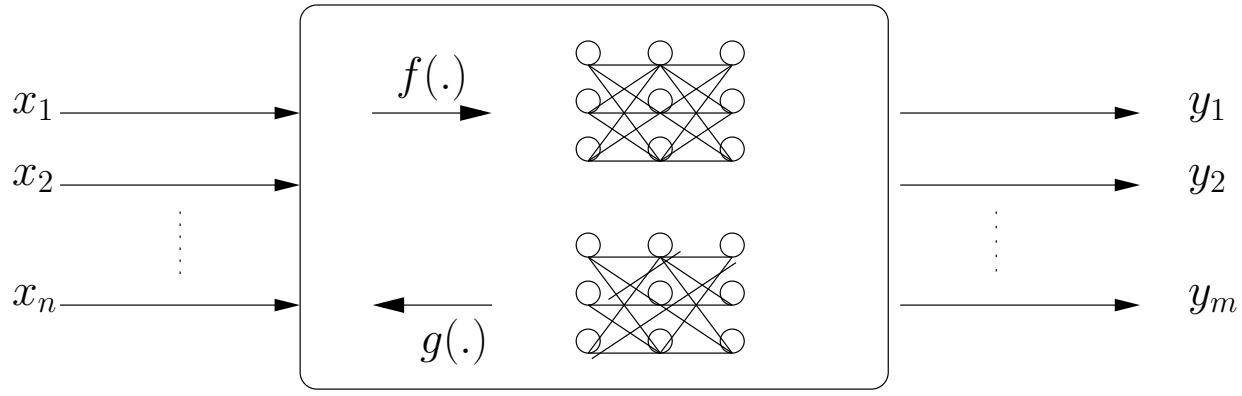


Fig. 1. Learning machine model of experimental science

prediction $p(\boldsymbol{y} \mid \mathcal{D})$. This problem is out of the scope of our paper. In [3,4], one can find a geometric insight of the selection of a model among a finite set of available models. Our contribution rather concerns the second question of the selection of appropriate prior distribution. Assuming given a statistical model (differentiable manifold) either parametric or not, we propose a novel method to construct a prior distribution. This method can be interpreted as an inverse problem of geometric Bayesian learning [5, 6]. In fact, Bayesian learning consists in construction a decision rule (a mapping from the data space to the manifold $\mathcal{Q}$ of predicted distributions, see Figure 2) by in minimizing a cost function (generalization error) given a chosen manifold and a prior distribution on this manifold [5]. However, in the proposed method, we assume a fixed prediction (reference distribution) and we minimize the decision error cost under an uniformity constraint (a measure of ignorance).

In the sequel, we assume 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 $\mathcal{Q}$ of probability distributions, either parametric ($\mathcal{Q} = \{P(\boldsymbol{z} \mid \boldsymbol{\theta})\}$) or non parametric. The statistical manifold $\mathcal{Q}$ is the set of probability distributions on the space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ (see Figure 2). The objective of a learning algorithm is to construct a learning rule τ mapping the set $\mathcal{D}$ of training data $D = (\boldsymbol{x}_{1..N}, \boldsymbol{y}_{1..N})$ to a probability distribution $p \in \mathcal{Q} \subset \mathcal{P}$ ($\mathcal{P}$ is the whole set of probability densities):

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

The Bayesian statistical learning leads to a solution depending on the prior distribution of the unknown distribution p. In the parametric case, where the points p of the manifold $\mathcal{Q}$ are parametrized by a coordinate system $\boldsymbol{\theta}$, 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 ($\mathcal{Q}$) and the closer set of ignorance containing it are the main objectives of this paper. We show that the prior expression depends on the chosen geometry (subjective choice) of the set of probability measures. The entropic prior[1] [7, 8] and the conjugate prior of exponential families are special cases related to special geometries.

In section II, we review briefly some concepts of Bayesian geometrical statistical learning and the role of differential geometry. In section III, 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 IV, we study the particular case of δ-flat families where previous results have explicit formula. In section V, we come across the case of δ-flat families mixture. In section VI, 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.

[1]Some related work about ignorance and prior selection in a geometric framework can be found in *http://omega.albany.edu:8008/ignorance*

II. Statistical geometric learning

A. Mass and Geometry

The statistical learning consists in constructing a learning rule τ which maps the training measured data D to a probability distribution[2] $q = \tau(D) \in \mathcal{Q} \subset \mathcal{P} = \{p \mid \int p = 1\}$ (the predictive distribution). We will discuss the consequences of the restriction to a subset $\mathcal{Q}$ in subsection II-C. Therefore, our target space is a 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 D and modified according to Bayesian rule, after observing the data to give the *a posteriori* distribution (see Figure 2):

$$P(p_0 \mid D) \propto P(D \mid p_0)\, \Pi(p_0), \text{ for all } p_0 \in \mathcal{P} \tag{1}$$

where $P(D \mid p_0)$ is the likelihood of the probability p_0 to generate the data D (the distribution is to be compared to a parameter in the classic maximum likelihood methods). In the sequel, $\boldsymbol{z}$ is the couple $(\boldsymbol{x}, \boldsymbol{y})$ introduced in the introduction I. In the case of i.i.d samples $D = \{\boldsymbol{z}_i\}_{i=1..N}$, the likelihood of the probability p_0 is simply $P(D \,|\, p_0) = \prod_{i=1}^{N} p_0(\boldsymbol{z}_i)$. For the parametric case $\mathcal{Q} = \{p_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \mathbb{R}^n\}$, where $\boldsymbol{\theta}$ is a coordinate system and n is the dimension of the manifold, just replace p_0 in equation (1) by $\boldsymbol{\theta}$ to find the classic Bayesian parametric formulation. We assume that the data D are generated by an unknown distribution p^*. As the number N of data samples becomes large, the *a posteriori* distribution $P(p_0 \mid \mathcal{D})$ is concentrated around the true distribution p^* (consistency), under some weak regular conditions [3].

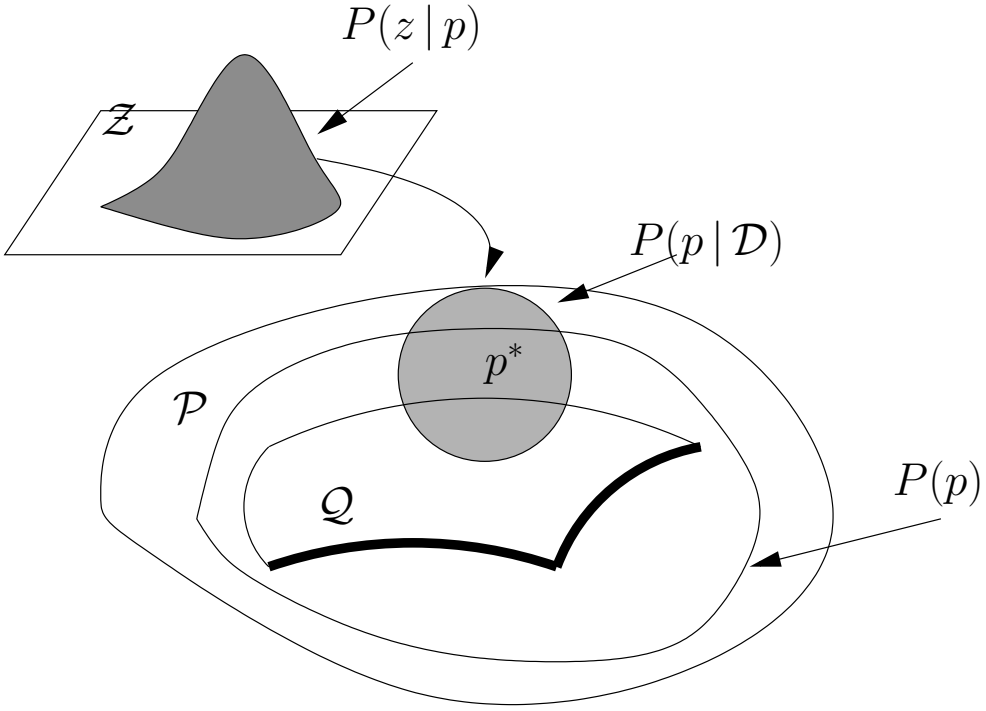


Fig. 2. The *a posteriori* mass is proportional to the product of the *a priori* mass and the likelihood function. As the number of samples N grows, the *a posteriori* distribution $P(p \,|\, \mathcal{D})$ (the dark ball) is more and more concentrated around the true distribution p^*.

[2]In literature, the considered subset $\mathcal{Q}$ is parametric. This restriction to parametric manifold is important for computational reasons, that is why $\mathcal{Q}$ is also called the computational model. However, for the derivation of the main results of this contribution, there is no need to restrict $\mathcal{Q}$ to be parametric.

[3]In section V-A, there is an example illustrating the violation of these conditions and how the construction of prior and the use of Bayesian approach eliminate the singularity problem and ensure the consistency of the MAP solution.

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

$$\begin{cases} D_\delta(p,q) = \frac{\int p}{1-\delta} + \frac{\int q}{\delta} - \frac{\int p^\delta q^{1-\delta}}{\delta(1-\delta)}, \;\; \delta \neq 0,1 \\ \\ D_1(p,q) = D_0(q,p) = \int q - \int p + \int p \log p/q \end{cases} \tag{2}$$

where the integration is defined with respect to a dominant measure. We notice that this definition is parametric free. Therefore, in the case of a parametric restricted manifold $\mathcal{Q}$, this measure is invariant under reparametrization. It is shown in [9] 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 (defining the scalar product in the tangent spaces), ∇^δ the δ connection with Christoffel symbols $\Gamma^\delta_{ij,k}$ and $\nabla^* = \nabla^{1-\delta}$ its dual connection:

$$\begin{cases} g_{ij} &= < \partial_i, \partial_j > = \; E_{\boldsymbol{\theta}} \left[\partial_i l(\boldsymbol{\theta}) \, \partial_j l(\boldsymbol{\theta}) \right] \\ \\ \Gamma^\delta_{ij,k} &= E_{\boldsymbol{\theta}} \left[(\partial_i \partial_j l(\boldsymbol{\theta}) + \delta \, \partial_i l(\boldsymbol{\theta}) \partial_j l(\boldsymbol{\theta})) \, \partial_k l(\boldsymbol{\theta}) \right] \end{cases} \tag{3}$$

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 then 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 $\tilde{\mathcal{P}}$ [5,6] 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 $[0\,,\,1]$, whereas, $\mathcal{P}$ 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:

$$\overset{\delta}{l}(p) = p^\delta/\delta \tag{4}$$

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:

$$\overset{\delta}{l}(t) = (1-t) \; \overset{\delta}{l}(a) + t \; \overset{\delta}{l}(b)$$

B. 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 [5,6] and means that the optimal estimator τ_δ can be computed pointwise as a function of $\boldsymbol{z}$ and we do not need a general expression of the optimal estimator τ_δ:

$$\hat{p}(\boldsymbol{z}) = \tau_\delta(\boldsymbol{z}) = \arg\min_q \int_p P(p \,|\, \boldsymbol{z}) D_\delta(p, q) \tag{5}$$

By variational calculation, the solution of (5) is straightforward and gives:

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

which is exactly the *a posteriori* mean of the δ coordinates. The above result can be considered as the extension of the classic parametric Bayesian inference to the more abstract set of probability distributions. For example, consider the estimation of a parameter $\boldsymbol{\eta}$ from its *a posteriori* distribution $p(\boldsymbol{\eta} \,|\, \boldsymbol{z})$. The δ divergence is to be compared to the quadratic cost $\|\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}\|^2$. The minimization of the expected cost leads to the EAP (*a posteriori* expectation) solution: $\hat{\boldsymbol{\eta}} = \mathrm{E}_p[\boldsymbol{\eta} \,|\, \boldsymbol{z}] = \int \boldsymbol{\eta}\, p(\boldsymbol{\eta} \,|\, \boldsymbol{z}) d\,\boldsymbol{\eta}$.

From a physical point of view, the above solution is exactly the gravity center of the set $\tilde{\mathcal{P}}$ within a mass $P(p \,|\, \boldsymbol{z})$, the *a posteriori* distribution of p and with 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 $\tilde{\mathcal{P}}\ \forall\ \delta \in [0\,,\,1]$.

C. 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 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 $\mathcal{Q}$:

C.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 $\mathcal{Q}$ (the integration is computed over the whole set $\tilde{\mathcal{P}}$ but the minimization is performed on the subset $\mathcal{Q}$):

$$\hat{q}(\boldsymbol{z}) = \tau_\delta(\boldsymbol{z}) = \arg\min_{q \in \mathcal{Q}} \int_{p \in \tilde{\mathcal{P}}} P(p \,|\, \boldsymbol{z}) D_\delta(p, q) \tag{6}$$

Thus the solution $\hat{q}$ 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). A point b is the δ projection of a point a onto the manifold $\mathcal{Q}$ if b minimizes the δ divergence $D_\delta(a, q)$, $\forall q \in \mathcal{Q}$. The projection b can also be characterized by the property that the geodesic line linking a and b is orthogonal to all curves in $\mathcal{Q}$ passing through the point b. For details, refer to [5] where the authors define the point $\hat{p}$ as the ideal δ-estimate and the point $\hat{q}$ as the δ estimate within the model $\mathcal{Q}$.

C.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 $\mathcal{Q}$:

$$\hat{q}(\boldsymbol{z}) = \tau_\delta(\boldsymbol{z}) = \arg\min_{q \in \mathcal{Q}} \int_{p \in \mathcal{Q}} P(p \,|\, \boldsymbol{z}) D_\delta(p, q) = \arg\min_{q \in \mathcal{Q}} \int_{\boldsymbol{\theta}} P(\boldsymbol{\theta} \,|\, \boldsymbol{z}) D_\delta(p_\theta, q) d\,\boldsymbol{\theta} \tag{7}$$

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).

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

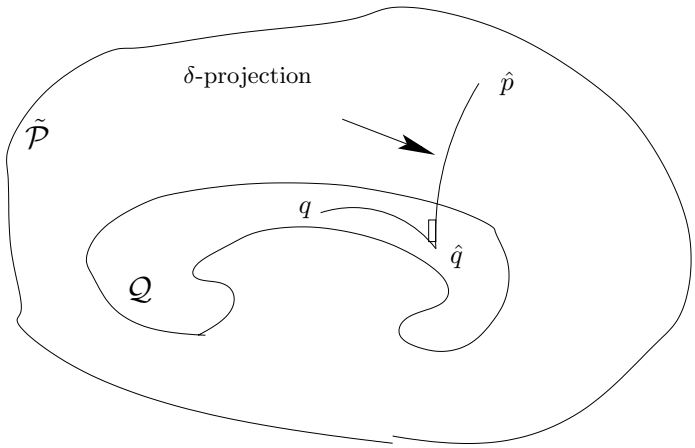


Fig. 3. The δ estimate $\hat{q}$ is the δ projection of the non parametric solution $\hat{p}$ onto the computational model $\mathcal{Q}$.

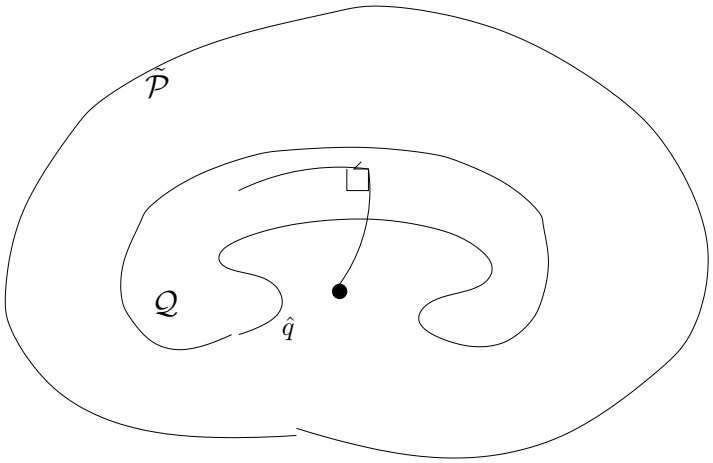


Fig. 4. Projection of the barycentre solution onto the parametric model

1. The cost function to be minimized in equation (7) is the same as the cost function in (6) when p is allowed to belong to the whole set $\tilde{\mathcal{P}}$ and the *a posteriori* $P(p \mid \boldsymbol{z})$ 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 (7) is the same as the cost function in (6) when the *a posteriori* $P(\boldsymbol{\theta} \mid \boldsymbol{z})$ is the projected mass of the *a posteriori* $P(p \mid \boldsymbol{z})$ onto the model $\mathcal{Q}$. This interpretation is more consistent than the first one. In fact, it is more robust with respect to the model deviation. For instance, assume that the data are generated according to a true distribution p^* outside the manifold $\mathcal{Q}$. As the sample size N gets larger, the *a posteriori* distribution is more and more concentrated around the point p^*. The classic *a posteriori* measure of the manifold $\mathcal{Q}$ will converge to 0. Consequently, the inference on the manifold $\mathcal{Q}$ has no meaning. However, when considering the projected *a posteriori* distribution, the measure on the manifold $\mathcal{Q}$ will concentrate around the δ projection of the true distribution p^*. Therefore, the parametric modeling is equivalent to the non parametric modeling in the restricted case.
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 point q in $\mathcal{Q}$ (see Figure 5).

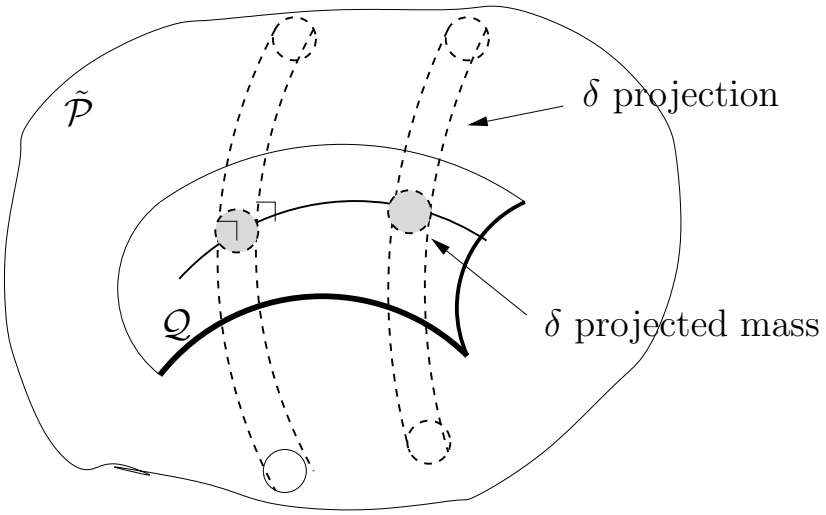


Fig. 5. Projection of the *a priori/ a posteriori* distribution on the manifold $\mathcal{Q}$ leads to an equivalence between the parametric and non parametric modeling.

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.

III. 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 (ignorance) 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 $\mathcal{Q} = \{q_\theta \mid \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^n\}$.

A. Family of (δ, α)-Priors

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 [10]) representing the uniformity. In fact, the Fisher matrix is a bilinear form which is a natural metric of the statistical manifold and it is shown that the square root of its determinant represents an equal prior for all the distributions of the model [3]. It is worth noting that we are considering two different spaces: the space $\tilde{\mathcal{P}}$ of finite measures and the space $\mathcal{G} = \{\Pi, \int \Pi = 1\}$ of prior distributions on the finite measures. Since we have two distinct spaces, we can choose two different geometries on each space. In the sequel, we consider the δ-geometry on the space $\tilde{\mathcal{P}}$ and the α-geometry on the space of priors. For the same reason as for the distributions $p_{\boldsymbol{\theta}}$, we embed the space $\mathcal{G}$ of priors Π in the corresponding space of finite measure priors $\tilde{\mathcal{G}} = \{\Pi, \int \Pi > 0\}$. We have the following family of cost functions parametrized [4] by the couple (δ, α) :

[4]The cost function is also parametrized by the weights γ_e and γ_u

$$J_{\delta,\alpha}(\Pi) = \gamma_e \int \Pi(\boldsymbol{\theta}) D_\delta(p_\theta, p_0) d\,\boldsymbol{\theta} + \gamma_u D_\alpha(\Pi, \sqrt{g}) \tag{8}$$

where D_δ and D_α are the δ divergence and the α divergence defined on the spaces $\tilde{\mathcal{P}}$ and $\tilde{\mathcal{G}}$ respectively, according to equation (2). γ_e is the confidence degree in the reference distribution p_0 (reflecting some *a priori* knowledge) and γ_u the uniformity degree (constraint of ignorance). Considered independently, these two coefficients are not significant. However, their ratio is relevant in the following. The cost (8) can be rewritten as:

$$\begin{cases} J_{\delta,\alpha}(\Pi) = \gamma_e E_\delta(\tau_0) + \gamma_u D_\alpha(\Pi, \sqrt{g}) \\ \\ \frac{\partial \tau_0}{\partial \boldsymbol{z}} = 0 \end{cases}$$

where $E_\delta(\tau_0)$ is the generalization error of a fixed learning rule τ_0. This learning rule is fixed as we have not collected any data:

$$\begin{aligned} E_\delta(\tau_0) &= \int_{\boldsymbol{\theta}} \Pi(\boldsymbol{\theta}) \int_z p(\boldsymbol{z} \,|\, \boldsymbol{\theta}) D_\delta(p_{\boldsymbol{\theta}}, \tau_0(\boldsymbol{z})) d\,\boldsymbol{z}\, d\,\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \Pi(\boldsymbol{\theta}) D_\delta(p_{\boldsymbol{\theta}}, p_0)\, d\,\boldsymbol{\theta} \end{aligned}$$

The cost function represents a balance between a fixed predictive density p_0 (the prior knowledge of the user) and the uniformity constraint reflecting our prior ignorance. Its minimization is the inverse problem of Bayesian statistical learning introduced in the previous section as the predictive density is fixed and the cost function is minimized with respect to the prior density.

Theorem 1: The following (δ, α) measure:

$$\begin{cases} \Pi_{\delta,\alpha}(\boldsymbol{\theta}) \propto \dfrac{\sqrt{g(\boldsymbol{\theta})}}{\left[1 + (1-\alpha)\frac{\gamma_e}{\gamma_u} D_\delta(p_{\boldsymbol{\theta}}, p_0)\right]^{1/(1-\alpha)}}, & \alpha \neq 1 \\ \\ \Pi_\delta(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_0)} \sqrt{g(\boldsymbol{\theta})}, & \alpha = 1 \end{cases} \tag{9}$$

minimizes the cost function $J_{\delta,\alpha}(\Pi)$ over the space $\tilde{\mathcal{G}}$. □

See Appendix VIII-A for the proof of Theorem 1. The minimization of the function $J_{\delta,\alpha}(\Pi)$ relies on variational calculus. In the sequel, we call this measure the (δ, α)-**Prior** . For notational convenience, we refer to the particular case $(\delta, 1)$-Prior as δ-Prior [5].

Remark 1: The obtained (δ, α)-Prior family contains many particular known cases corresponding to particular values of the couple (δ, α) and the ratio γ_e/γ_u. For instance:
- If $(\delta, \alpha) = (1, 1)$ then the cost function (8) is the kullback-Leibler divergence between the joint distributions of data and parameters. The $(1, 1)$-Prior is then the **Entropic prior** considered in [7].
- If $(\delta, \alpha) = (0, 1)$ we obtain the **conjugate** prior for exponential families (see examples in Section VI).
- For the particular case where $\mathcal{Q}$ is a δ Euclidean family (δ-flat + self dual[6]), we obtain the **t-distribution** for $\alpha \neq 1$ and the **Gaussian distribution** for $\alpha = 1$.
- If the ratio γ_e/γ_u goes to 0, we obtain the Jeffreys prior $\sqrt{g(\boldsymbol{\theta})}$.
- If the ratio γ_e/γ_u goes to ∞ we obtain the Dirac concentrated on p_0.

Remark 2: We note that the (δ, α)-Prior (9) can be extended to a coordinate free space $\mathcal{Q}$. If we consider

[5] In the original contribution [11], the author proposed the particular case of $\alpha = 1$ and considered the family of the $(\delta, 1)$-Priors.
[6] A differentiable manifold is self dual if the dual connections are equal: $\nabla^* = \nabla^{1-\delta} = \nabla$.

the prior on the elements p of the non parametric space $\mathcal{Q}$, then we have the following expression:

$$\begin{cases} \Pi_{\delta,\alpha}(p) & \propto & \dfrac{\sqrt{g(p)}}{\left[1 + (1-\alpha)\frac{\gamma_e}{\gamma_u} D_\delta(p, p_0)\right]^{1/(1-\alpha)}}, & \alpha \neq 1,\ p \in \mathcal{Q} \\ \Pi_\delta(p) & \propto & e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p,p_0)} \sqrt{g(p)}, & \alpha = 1,\ p \in \mathcal{Q} \end{cases} \tag{10}$$

where $g(p)$ is a measure of statistical curvature of the space $\mathcal{Q}$.

B. Choice of reference distribution

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 assume in both cases (discrete and continu) that the weights sum to one: $\sum \gamma_e^i = \int P_r(p_0) = 1$. We show in the following that the (δ, α)-Prior has the same expression form as (9) but with additional terms measuring:

- the relative accuracy of the reference distributions, i.e the mean distance from the reference distribution to the manifold $\mathcal{Q}$.
- the dispersion of the reference distributions.

In the following, we give exact definitions of the two above notions (accuracy and dispersion) before introducing the expression of the (δ, α)-Prior.

Definition 1: [**β-Barycentre**] The distribution p_G is the β-barycentre of the discrete set $\{(p_1, \gamma_e^1), ..., (p_N, \gamma_e^N)\}$ if its β-coordinate $\overset{\beta}{l}$ (4) is:

$$\overset{\beta}{l}(p_G) = \sum_{i=1}^{N} \gamma_e^i \overset{\beta}{l}(p_i)$$

Definition 2: [**β-Barycentre**] The distribution p_G is the β-barycentre of the continuous set $(\tilde{\mathcal{P}}, P_r)$ if its β-coordinate $\overset{\beta}{l}$ (4) is:

$$\overset{\beta}{l}(p_G) = \int \overset{\beta}{l}(p_0) P_r(p_0)$$

We introduce the notions of **accuracy** and **dispersion** of a set of reference distributions (either discrete or continuous).

Definition 3: [**β-Accuracy**] The β-accuracy of a set of reference distributions $(\tilde{\mathcal{P}}, P_r)$ (resp. $\{(p_i, \gamma_e^i)\}$) relatively to a manifold $\mathcal{Q}$ is the inverse of the β-divergence between the β-barycentre of $(\tilde{\mathcal{P}}, P_r)$ (resp. $\{(p_i, \gamma_e^i)\}$) and its β-projection on the manifold $\mathcal{Q}$ (see Figure 6):

$$A_\beta = 1/D_\beta(p_G, p_G^\perp) \tag{11}$$

Definition 4: [**β-Dispersion**] The β-dispersion of a set of reference distributions $(\tilde{\mathcal{P}}, P_r)$ (resp. $\{(p_i, \gamma_e^i)\}$) is the average of the β-divergence to the β-barycentre (see Figure 6):

$$V_\beta = \int D_\beta(p_0, p_G) P_r(p_0) \ \ (\text{resp.} \sum \gamma_e^i D_\beta(p_i, p_G)) \tag{12}$$

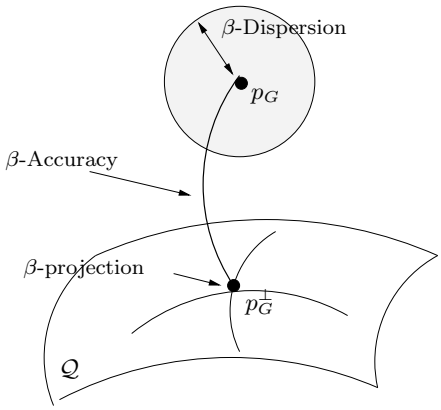


Fig. 6. The continuous set of reference distributions is represented by the filled ball. The point p_G is the β-barycentre and $p_G^\perp$ its β-projection on the manifold $\mathcal{Q}$. The β-accuracy is the inverse of $D_\beta(p_G, p_G^\perp)$. The β-dispersion is the mean (according to the distribution P_r) of the divergence to p_G.

Theorem 2: In the general case where we are given a set of reference distributions (not necessarily included in the manifold $\mathcal{Q}$) with the corresponding probability measure (P_r in the continuous case and $\{\gamma_e^i\}$ in the discrete case) and if $\mathcal{Q}$ is δ-convex[7], the (δ,α)-Prior has the following expression:

$$\begin{cases} \Pi_{\delta,\alpha}(\boldsymbol{\theta}) \propto \dfrac{\sqrt{g(\boldsymbol{\theta})}}{\left[1 + \dfrac{(1-\alpha)\frac{\gamma_e}{\gamma_u}}{1+(1-\alpha)\frac{\gamma_e}{\gamma_u}(1/A_{1-\delta}+V_{1-\delta})} D_\delta(p_{\boldsymbol{\theta}}, p_G^\perp)\right]^{1/(1-\alpha)}}, & \alpha \neq 1 \\ \Pi_\delta(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u}(1/A_{1-\delta}+V_{1-\delta})}\, e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_G^\perp)} \sqrt{g(\boldsymbol{\theta})}, & \alpha = 1 \end{cases} \tag{13}$$

where p_G is the $(1-\delta)$-barycentre of reference distributions, $p_G^\perp$ its $(1-\delta)$-projection on $\mathcal{Q}$, $A_{1-\delta}$ and $V_{1-\delta}$ are the accuracy and the dispersion of reference distributions set. □

Proof: see Appendix VIII-B.

First, we notice that the expression of the (δ,α)-Prior has a similar form as in the original expression (9) (where the reference distribution belongs to the manifold $\mathcal{Q}$ ($p_0 = p_{\boldsymbol{\theta}_0}$)). Second, we notice the additional term $(1-\alpha)\frac{\gamma_e}{\gamma_u}(1/A_{1-\delta} + V_{1-\delta})$ in the denominator of the coefficient weighting the divergence $D_\delta(p_{\boldsymbol{\theta}}, p_G^\perp)$. The presence of this term is intuitive. In fact, it reduces the confidence coefficient $(1-\alpha)\frac{\gamma_e}{\gamma_u}$ in the reference distribution p_0, in particular when the reference distribution is located outside the manifold $\mathcal{Q}$ ($A_{1-\delta} < \infty$) or when there is an uncertainty about p_0 ($V_{1-\delta} > 0$). In words, when the confidence coefficient γ_e/γ_u is very high ($\longrightarrow \infty$), the resulting weighting coefficient converges to $1/(1/A_{1-\delta} + V_{1-\delta})$. Therefore, the (δ,α)-Prior does not converge to a dirac at $p_G^\perp$ and implicitly takes into account the accuracy and the dispersion of the reference set (see Figure 7). The confidence term is bounded as follows:

$$1 \leq \frac{(1-\alpha)\frac{\gamma_e}{\gamma_u}}{1+(1-\alpha)\frac{\gamma_e}{\gamma_u}(1/A_{1-\delta}+V_{1-\delta})} \leq 1/(1/A_{1-\delta}+V_{1-\delta})$$

Example 1: In the particular case of only one reference distribution p_0 located outside the manifold $\mathcal{Q}$ (see Figure 7-a), the barycentre p_G is p_0. The accuracy is the inverse of the $(1-\delta)$-divergence of p_0 to $\mathcal{Q}$ ($1/D_{\delta-1}(p_0, p_0^\perp)$) and the dispersion is null.

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

[7]A manifold is β-convex if all the β-geodesics are contained in $\mathcal{Q}$.

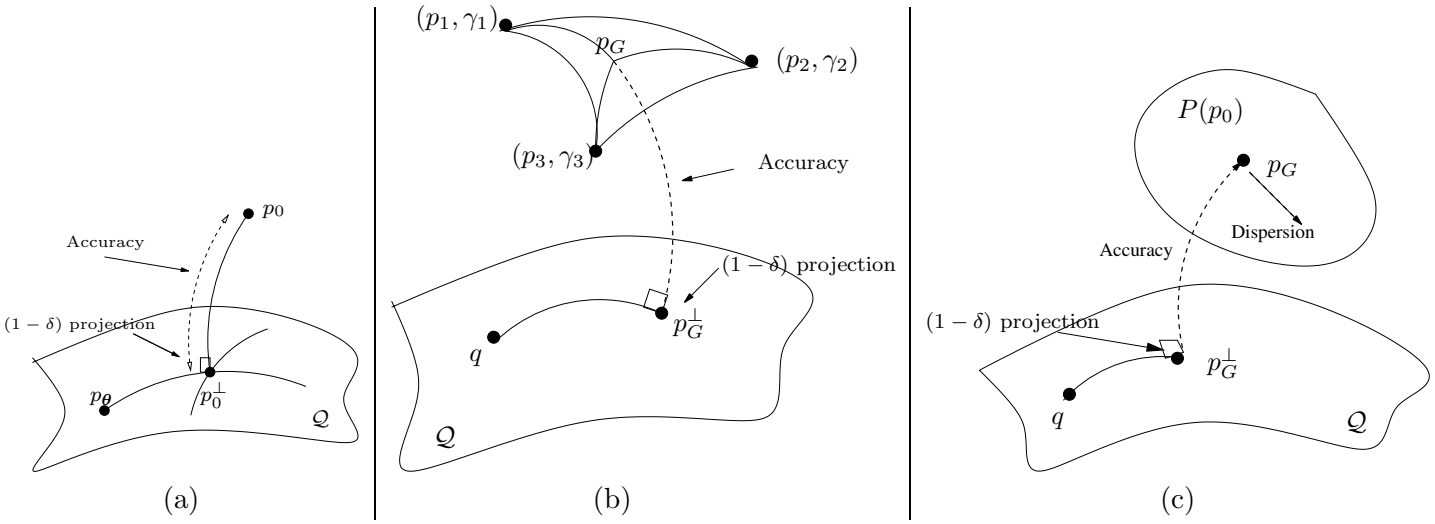


Fig. 7. (a) The equivalent of the reference distribution p_0 located outside $\mathcal{Q}$ is its $1-\delta$ projection, (b) the equivalent reference distribution is the $1-\delta$ projection of the $1-\delta$ barycentre of the N references distributions, (c) The equivalent reference distribution of a continum reference region is the $1-\delta$ projection of the $1-\delta$ barycentre.

reference distributions (or even a continuous set) indicates implicitly the existence of hyperparameters and the resulting solution shows that these hyperparameters are integrated and at the same time optimized if the *a priori* average (the barycentre) is considered as an optimization operation.

IV. 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 [9] 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:

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

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

A. 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:

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

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. This property of invariance is well expected since we are using a parametric free geometric construction of estimates.

In addition, noting 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. We can directly obtain this result by just replacing δ by $(\delta-1)$, since a δ-flat manifold is also $(1-\delta)$-flat. In general, the δ-estimate is different from the $(1-\delta)$-estimate. They are equal in the case of an Euclidean manifold ($\nabla = \nabla^*$).

B. Prior selection with δ-flat families

The (δ,α)-Prior $\Pi_{\delta,\alpha}$ has the following general expression:

$$\begin{cases} \Pi_{\delta,\alpha}(\boldsymbol{\theta}) & \propto & \dfrac{\sqrt{g(\boldsymbol{\theta})}}{[1 + \lambda D_\delta(p_{\boldsymbol{\theta}}, p_0)]^{1/(1-\alpha)}}, & \alpha \neq 1 \\ \Pi_{\delta}(\boldsymbol{\theta}) & \propto & e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_0)} \sqrt{g(\boldsymbol{\theta})}, & \alpha = 1 \end{cases} \tag{14}$$

where λ is a fixed coefficient depending on the confidence ration γ_e/γ_u, the accuracy and the dispersion (see Section III-B). $p_0 \in \mathcal{Q}$ 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:

$$\begin{cases} \Pi_{\delta,\alpha}(\boldsymbol{\theta}) & \propto & \dfrac{\sqrt{g(\boldsymbol{\theta})}}{\left[1 + \lambda(\psi(\boldsymbol{\theta}) - \theta_i\, \eta_i^0)\right]^{1/(1-\alpha)}}, & \alpha \neq 1 \\ \Pi_{\delta}(\boldsymbol{\theta}) & \propto & e^{-\frac{\gamma_e}{\gamma_u}(\psi(\boldsymbol{\theta}) - \theta_i\, \eta_i^0)} \sqrt{g(\boldsymbol{\theta})}, & \alpha = 1 \end{cases} \tag{15}$$

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.

Example 2: In the Euclidean case, that is when the connection ∇ is equal to its dual connection ∇^*, which is equivalent to equality of the affine coordinates $[\theta_i] = [\eta_i]$: (**i**) the (δ,α)-Prior distribution is a **t-distribution** with $\frac{1+\alpha}{1-\alpha}$ degrees of freedom, mean $\boldsymbol{\theta}_0$ and precision λ (**ii**) the δ-Prior ($\alpha = 1$) is **Gaussian** with mean $\boldsymbol{\theta}_0$ and precision $2\,\gamma_e/\gamma_u$:

$$\begin{cases} \Pi_{\delta,\alpha}(\boldsymbol{\theta}) & \propto & \dfrac{\sqrt{g(\boldsymbol{\theta})}}{[1 + \lambda\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2]^{1/(1-\alpha)}}, & \alpha \neq 1 \\ \Pi_{\delta}(\boldsymbol{\theta}) & \propto & e^{-\frac{\gamma_e}{\gamma_u}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2} \sqrt{g(\boldsymbol{\theta})}, & \alpha = 1 \end{cases}$$

C. Projection of Priors

We detail here the notion of prior projection. Our objective is how to determine a prior (or in general a probability mass) on the subspace $\mathcal{Q}_a$ taking into account the prior of the embedding space $\mathcal{Q}$. The essence of the *projection mass* notion is to define a prior on a restricted set by suitably projecting the prior of the embedding space. Then, when working in the restricted space, we do not lose the information about the initial space. This notion is completely different from the common notion of defining the prior on $\mathcal{Q}_a$ by just restricting the prior on $\mathcal{Q}$ (see Figure 8). This idea is very ambitious comparing to our limited understanding of the geometry of the space under hand. For this reason, we will illustrate the computation in the particular case

of ∇^*-autoparallel submanifolds $\mathcal{Q}_a \subset \mathcal{Q}$. The general case needs a more abstract mathematical investigation about how to perform the projection.

$\mathcal{Q}_a$ 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:

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

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 $\mathcal{T}_p^c$ are orthogonal at the point $p(\boldsymbol{\eta}_I^0, \boldsymbol{\theta}_{II}^0)$. Consequently, the projected prior from $\mathcal{Q}$ onto $\mathcal{Q}_a$ is simply:

$$\Pi^{\perp}(p) = \int_{q \in \mathcal{Q}_a^c} \Pi(q) = \int_{\theta_I} \Pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{II}^0) d\,\boldsymbol{\theta}_I$$

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 $\tilde{\mathcal{P}}$. In fact, this wrong intuition is due to our experience with Euclidean spaces. In an Euclidean space, the θ-coordinates are equal to the η-coordinates. Therefore, the projection is obtained by simply marginalizing the coordinates (see Figure 8).

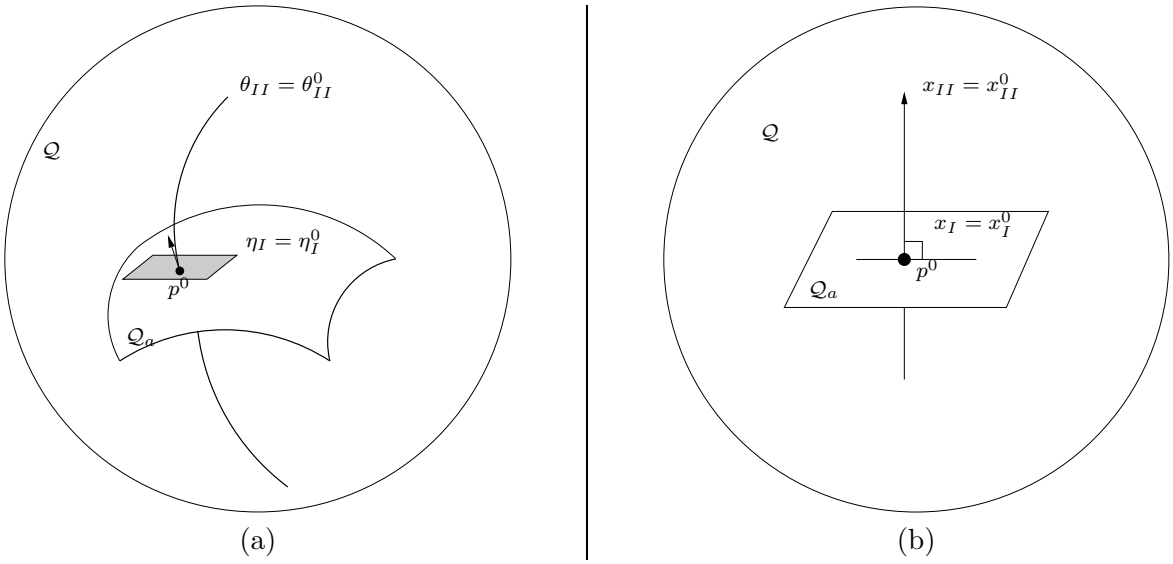


Fig. 8. (a) The orthogonal manifold to $\mathcal{Q}_a$ is the manifold $\mathcal{Q}_a^c$ obtained by fixing the complementary part of the dual coordinates. The projected mass is then the integral along $\mathcal{Q}_a^c$, (b) In the Euclidean case, the dual coordinates are equal. The projected mass is then obtained by marginalizing in the same coordinate system.

V. 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:

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

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. In this section and the following one, we focus on the computation of the particular case of δ-Prior ($\alpha = 1$) of the mixture density.

The δ-Prior has the following expression:

$$\Pi_\delta(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_0)} \sqrt{g(\boldsymbol{\theta})} \tag{16}$$

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 [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 [9]:

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

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

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

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

A. Singularities with mixture families

It is known that in learning the parameters of Gaussian mixture densities [12] 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 [12], 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 [9]:

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

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:

$$\begin{cases} \theta_1 = \frac{\mu}{\sigma^2}, & \theta_2 = \frac{-1}{2\,\sigma^2} \\ \\ \eta_1 = \mu, & \eta_2 = \mu^2 + \sigma^2 \end{cases} \tag{17}$$

The corresponding Fisher information are:

$$|g(\theta)| \propto \sigma^6, \quad |g(\eta)| \propto 1/\sigma^6 \tag{18}$$

The canonical divergence has the following expression:

$$D_\delta(p_1, p_2) = D_{1-\delta}(p_2, p_1) = \psi(p_1) + \phi(p_2) - \theta_i(p_1)\,\eta_i(p_2) \tag{19}$$

where ψ and ϕ are the potentials given by:

$$\psi = \tfrac{\mu^2}{2\,\sigma^2} + \log\sqrt{2\pi}\sigma, \quad \phi = \tfrac{-1}{2} - \log\sqrt{2\pi}\sigma \tag{20}$$

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 [12] and this is recalled 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 (17), (18), (19) and (20) we note the following comparison of the 0 and 1 priors through their dependences on the variance σ_j:

$\delta = 0$	$\delta = 1$
$\downarrow$	$\downarrow$
$p \longrightarrow \partial\mathcal{Q}$	$p \longrightarrow \partial\mathcal{Q}$
Π_0 is $O(\sigma_j^\alpha e^{-k_0/\sigma_j^2})$	Π_1 is $O(\sigma_j^{2\,w_j\frac{\gamma_j}{\gamma_u}})$
$\downarrow$	$\downarrow$
Exponential	Polynomial

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.

VI. Examples

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

A. Multivariate Gaussian mixture

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

$$p(\boldsymbol{x}_i) = \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{x}_i\,;\,\boldsymbol{m}_k, \boldsymbol{R}_k) \tag{21}$$

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 (21) 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 [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:

$$|g_i(w_i, \boldsymbol{m}_i, \boldsymbol{R}_i)| = \left(\frac{1}{2}\right)^{n^2} w_i^{(n^2+n-1)}\, |\boldsymbol{R}_i|^{-(n+2)} \tag{22}$$

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:

- $\delta = 0$:

$$\begin{aligned} \Pi_0(\boldsymbol{\eta}_i) \;&\propto\; \exp\left[-\tfrac{\gamma_e}{\gamma_u}\left(w_i^0 D_0(\mathcal{N}_i : \mathcal{N}_i^0) + w_i^0 \log\tfrac{w_i^0}{w_i}\right)\right]\, \sqrt{|g_i(\boldsymbol{\eta}_i)|} \\ &\propto\; \mathcal{N}\left(\boldsymbol{m}_i\,;\, \boldsymbol{m}_0, \tfrac{\boldsymbol{R}_i}{\alpha\, w_i^0}\right)\, \mathcal{W}_n\left(\boldsymbol{R}_i^{-1}\,;\, \nu_0, \boldsymbol{R}_0^{-1}\right)\, w_i^{\beta_0} \end{aligned} \tag{23}$$

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.

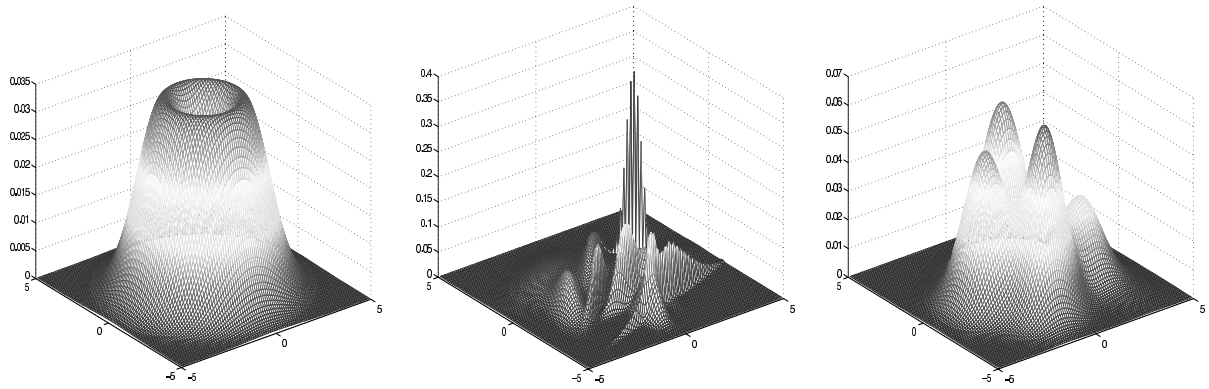
- $\delta = 1$:

$$\begin{aligned} \Pi_1(\boldsymbol{\eta}_i) \;&\propto\; \exp\left[-\tfrac{\gamma_e}{\gamma_u}\left(w_i D_1(\mathcal{N}_i : \mathcal{N}_i^0) + w_i \log\tfrac{w_i}{w_i^0}\right)\right]\, \sqrt{|g_i(\boldsymbol{\eta}_i)|} \\ &\propto\; \mathcal{N}\left(\boldsymbol{m}_i\,;\, \boldsymbol{m}_0, \tfrac{\boldsymbol{R}_i}{\alpha\, w_i}\right)\, \mathcal{W}_n\left(\boldsymbol{R}_i\,;\, \alpha w_i - 1, \tfrac{\alpha w_i - 1}{\alpha w_i}\boldsymbol{R}_0\right) \\ &\quad\; w_i^{\frac{n^2+n-1}{2} - (1+\frac{n}{2})\alpha w_i}\, (w_i^0)^{\alpha w_i}\, \Gamma_n(\tfrac{\alpha w_i - 1}{2}) \end{aligned} \tag{24}$$

where Γ_n is the generalized Gamma function of dimension n ([10] page 427):

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

The 1-prior Π_1 (24) is the generalized entropic prior [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). Figure 9 illustrates the problem of degeneracy and highlights the advantage of penalizing the likelihood by a 0-Prior when learning the parameters of the Gaussian mixture. In this simulation example, we have considered the ML estimation of a mixture of 10 Gaussians of bi-dimensional vectors ($n = 2$). The 10 multivariate Gaussians have the same covariance and the means are located on a circle. The graph on the left of the Figure 9 represents the original distribution which is a mixture of 10 Gaussians. The graph in the middle shows the estimated distribution with the maximum likelihood estimator. We note the degeneracy of the maximum likelihood which diverges to very sharp Gaussians (because of the singularity of the estimated covariances). The graph on the right shows the effect of regularization produced by the penalization of the likelihood by a 0-Prior.



(a) (b) (c)

Fig. 9. (a) Original distribution, (b) estimated distribution with maximum likelihood, given 100 samples, (c) estimated distribution with penalized maximum likelihood, given 100 samples.

B. 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 [13–15] 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:

$$\begin{cases} \boldsymbol{\eta}^j = \left(\boldsymbol{\eta}_i^j\right)_{i=1..K_j} \\ \\ \boldsymbol{\eta}_i^j = (w_i^j, m_i^j, \sigma_i^j) \end{cases}$$

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.

B.1 Fisher information matrix

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

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

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

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

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:

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

$(\boldsymbol{A}, \boldsymbol{R}_n)$-block

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

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

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:

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

($\boldsymbol{\eta}^j$)-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 (22) 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.2 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 (23) and (24).

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

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

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

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

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.

VII. 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. The problem of prior selection is considered as an inverse problem of a geometric statistical decision learning problem. The solving of a variational cost function leads to a family a prior distributions called the (δ, α)-Priors. This family contains many known particular cases of probability distributions such as the exponential family, the student distribution, etc, which correspond to particular geometries.

All the results in this paper can be extended to manifold valued parametric models. Indeed, when in a specific problem, the space of parameters is not Euclidean but rather a manifold, we can apply this work results to construct a prior on the manifold. This can be done by
1. replacing the statistical manifold $\mathcal{Q}$ of probability distributions by the manifold of parameters under hand.
2. choosing a suitable metric and an affine connection on the manifold.

We have also derived the expression of this family in the more general case of a set of reference distributions by introducing the notions of accuracy and dispersion. 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.

VIII. Appendix

A. Proof of Theorem 1

Consider the (δ,α)-cost as a function of the prior Π:

$$J_{\delta,\alpha}(\Pi) = \gamma_e \int \Pi(\boldsymbol{\theta}) D_\delta(p_\theta, p_0) d\,\boldsymbol{\theta} + \gamma_u D_\alpha(\Pi, \sqrt{g})$$

where the β-divergence D_β is defined as:

$$\begin{cases} D_\beta(p,q) = \frac{\int p}{1-\beta} + \frac{\int q}{\beta} - \frac{\int p^\beta q^{1-\beta}}{\beta(1-\beta)}\,, \quad \beta \neq 0,1 \\ \\ D_1(p,q) = \int q - \int p + \int p \log p/q = D_0(q,p) \end{cases}$$

For the first case ($\alpha \neq 0,1$), by variational calculus, we have the following expression of the variation $\Delta\, J_{\delta,\alpha}$:

$$\begin{aligned} \Delta\, J_{\delta,\alpha} &= \gamma_e \int D_\delta(p_\theta, p_0) \Delta\, \Pi\, d\,\boldsymbol{\theta} + \gamma_u \Delta\, D_\alpha(\Pi, \sqrt{g}) \\ &= \gamma_e \int D_\delta(p_\theta, p_0) \Delta\, \Pi\, d\,\boldsymbol{\theta} + \frac{\gamma_u}{1-\alpha} \int (1 - (\frac{\Pi}{\sqrt{g(\boldsymbol{\theta})}})^{\alpha-1}) \Delta\Pi d\,\boldsymbol{\theta} \\ &= \int \Delta\, \Pi \left\{ \gamma_e D_\delta(p_\theta, p_0) + \frac{\gamma_u}{1-\alpha} - \frac{\gamma_u}{1-\alpha} (\frac{\Pi}{\sqrt{g(\boldsymbol{\theta})}})^{\alpha-1} \right\} d\,\boldsymbol{\theta} \end{aligned}$$

Equating $\Delta\, J_{\delta,\alpha}$ to 0 yields the (δ,α)-Prior:

$$\Pi_{\delta,\alpha}(\boldsymbol{\theta}) \propto \frac{\sqrt{g(\boldsymbol{\theta})}}{\left[1 + (1-\alpha)\frac{\gamma_e}{\gamma_u} D_\delta(p_{\boldsymbol{\theta}}, p_0)\right]^{1/(1-\alpha)}}, \quad \alpha \neq 0,1$$

We note that the case $\alpha = 0$ can be obtained simply by replacing α by 0 in the previous equation. We have obtained the same result when considering the 0-divergence in the cost function.

For the case $\alpha = 1$, the variation of $J_{\delta,1}$ is:

$$\begin{aligned} \Delta\, J_{\delta,1} &= \gamma_e \int D_\delta(p_\theta, p_0) \Delta\, \Pi\, d\,\boldsymbol{\theta} + \gamma_u \Delta\, D_1(\Pi, \sqrt{g}) \\ &= \gamma_e \int D_\delta(p_\theta, p_0) \Delta\, \Pi\, d\,\boldsymbol{\theta} + \gamma_u \int \log \Pi / \sqrt{g(\boldsymbol{\theta})} \Delta\Pi d\,\boldsymbol{\theta} \end{aligned}$$

$\Delta\, J_{\delta,1} = 0$ yields the δ-Prior:

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

□

B. Proof of Theorem 2

Before proving the theorem, we recall some important definitions and results (see [2, 5, 9] for details):

Theorem 3: [**Pythagorean relation**] If the β-geodesic connecting p and r is orthogonal to the $(1-\beta)$-geodesic connecting r and q (the geodesics are considered in a δ-flat space), then

$$D_\beta(p,q) = D_\beta(p,r) + D_\beta(r,q)$$

Corollary 1: [β-**Projection**] Let p a point in a dually β-flat space $\mathcal{S}$ and $\mathcal{Q}$ a $(1-\beta)$-autoparallel manifold. Then a necessary and sufficient condition for a point q in $\mathcal{Q}$ to satisfy $D_\beta(p,q) = \min_{r\in\mathcal{Q}} D_\beta(p,r)$ is for the β-geodesic connecting p and q to be orthogonal to $\mathcal{Q}$ at q.

The point q is called the β-projection of p onto $\mathcal{Q}$.

Using the above results, the following decomposition of the divergence is straightforward:

Corollary 2: Let p a point in a δ-convexe $\mathcal{Q}$ (with respect to the whole set $\tilde{\mathcal{P}}$) and let p_0 a point in $\tilde{\mathcal{P}}$, then

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

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

Consider the cost function to be minimized, in the general case of not restricted reference distributions:

$$J_{\delta,\alpha}(\Pi) = \gamma_e \int P_r(p_0) \int \Pi(\boldsymbol{\theta}) D_\delta(p_\theta,p_0) d\boldsymbol{\theta} + \gamma_u D_\alpha(\Pi,\sqrt{g})$$

With the definition of the barycentre (Definition 2 in Section III-B) and the expression of the δ-divergence (2), we have a simple expression of the integral with respect to the reference distribution p_0:

$$\begin{aligned} \int P(p_0) D_\delta(p_\theta,p_0) &= D_\delta(p_\theta,p_G) + \tfrac{1}{\delta}(< \textstyle\int p_0 >_{P_r} - \int p_G) \\ &= D_\delta(p_\theta,p_G) + < D_\delta(p_G,p_0) > \end{aligned} \tag{25}$$

Using Corollary 2 with the point $p_G^\perp$ as the $(1-\delta)$-projection of p_G onto $\mathcal{Q}$, we can decompose the divergence between the points p_θ and p_G as the geodesics are orthogonal (see Figure 10),

$$\begin{aligned} \int P(p_0) D_\delta(p_\theta,p_0) &= D_\delta(p_\theta,p_G^\perp) + D_\delta(p_G^\perp,p_G) + < D_\delta(p_G,p_0) > \\ &= D_\delta(p_\theta,p_G^\perp) + 1/A_{1-\delta} + V_{1-\delta} \end{aligned} \tag{26}$$

where the accuracy $A_{1-\delta}$ and the dispersion $V_{1-\delta}$ are defined according to Definition 3 and 4 respectively.

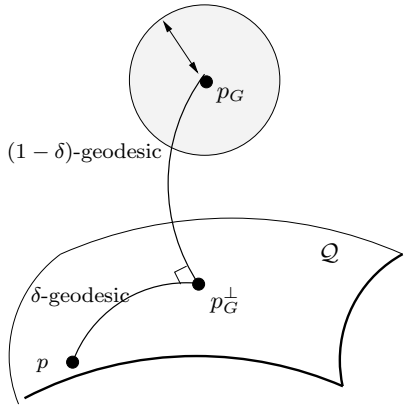


Fig. 10. Pythagorean relation: the δ-geodesic $l(p,p_G^\perp)$ is orthogonal to the $(1-\delta)$-geodesic $l(p_G^\perp,p_G)$.

Then, replacing the expression of the mean divergence (26) in the cost function (25) and minimizing with respect to the prior Π using the same variational arguments as in the proof of Theorem VIII-A, we obtain the

expression of the (δ, α)-Prior:

$$\begin{cases} \Pi_{\delta,\alpha}(\boldsymbol{\theta}) & \propto & \dfrac{\sqrt{g(\boldsymbol{\theta})}}{\left[1 + \dfrac{(1-\alpha)\frac{\gamma_e}{\gamma_u}}{1+(1-\alpha)\frac{\gamma_e}{\gamma_u}(1/A_{1-\delta}+V_{1-\delta})} D_\delta(p_{\boldsymbol{\theta}}, p_G^{\perp})\right]^{1/(1-\alpha)}}, & \alpha \neq 1 \\ \Pi_{\delta}(\boldsymbol{\theta}) & \propto & e^{-\frac{\gamma_e}{\gamma_u}(1/A_{1-\delta}+V_{1-\delta})}\, e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_G^{\perp})}\, \sqrt{g(\boldsymbol{\theta})}, & \alpha = 1 \end{cases}$$

□

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] S. Amari and H. Nagaoka, *Methods of Information Geometry*, vol. 191 of Translations of Mathematical Monographs, AMS, OXFORD, University Press, 2000.

[3] V. Balasubramanian, "A Geometric Formulation of Occam's Razor for Inference of Parametric Distributions", Tech. Rep., Princeton, Preprint PUPT-1588 and http://xyz.lanl.gov/adap-org/9601001, January 1996.

[4] V. Balasubramanian, "Statistical Inference, Occam's Razor and Statistical Mechanics on the Space of Probability Distributions", *cond-mat/9601030 and Neural Computation*, vol. 9, no. 2, February 1997.

[5] H. Zhu and R. Rohwer, "Bayesian invariant measurements of generalisation", in *Neural Proc. Lett.*, 1995, vol. 2 (6), pp. 28–31.

[6] 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.

[7] C. Rodríguez, "Entropic priors", *Tech. rep. Electronic form http: omega.albany.edu:8008/entpriors.ps*, 1991.

[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] S. Amari, *Differential-Geometrical Methods in Statistics*, Volume 28 of Springer Lecture Notes in Statistics, Springer-Verlag, New York, 1985.

[10] G. E. P. Box and G. C. Tiao, *Bayesian inference in statistical analysis*, Addison-Wesley publishing, 1972.

[11] H. Snoussi and A. Mohammad-Djafari, "Information Geometry and Prior Selection", in *Bayesian Inference and Maximum Entropy Methods*, C. Williams, Ed. MaxEnt Workshops, August 2002, pp. 307–327, Amer. Inst. Physics.

[12] 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 2001, pp. 36–46, Amer. Inst. Physics.

[13] K. Knuth, "A Bayesian approach to source separation", in *Proceedings of Independent Component Analysis Workshop*, 1999, pp. 283–288.

[14] 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.

[15] 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.