The direct L2 geometric structure on a manifold of probability densities with applications to Filtering

Damiano Brigo

The direct L^2 geometric structure on a manifold of probability densities with applications to Filtering

Damiano Brigo*
Dept. of Mathematics
King's College, London
damiano.brigo@kcl.ac.uk

November 12, 2011

Abstract

In this paper we introduce a projection method for the space of probability distributions based on the differential geometric approach to statistics. This method is based on a direct L^2 metric as opposed to the usual Hellinger distance and the related Fisher Information metric. We explain how this apparatus can be used for the nonlinear filtering problem, in relationship also to earlier projection methods based on the Fisher metric. Past projection filters focused on the Fisher metric and the exponential families that made the filter correction step exact. In this work we introduce the mixture projection filter, namely the projection filter based on the direct L^2 metric and based on a manifold given by a mixture of pre-assigned densities.

keywords Finite Dimensional Families of Probability Distributions, Exponential Families, Mixture Families, Hellinger distance, Fisher information metric, Direct L2 metric, Kullback Leibler information

AMS Classification codes: 53B25, 53B50, 60G35, 62E17, 62M20, 93E11

1 Introduction

In this paper we consider the (scalar) nonlinear filtering problem in continuous time. For a quick introduction to the filtering problem see Davis and Marcus (1981) [14]. For a more complete treatment see Liptser and Shiryayev (1978) [21] from a mathematical point of view or Jazwinski (1970) [19] for a more applied perspective. For recent results see the collection of papers [13].

The nonlinear filtering problem has an infinite–dimensional solution in general. Constructing of approximate finite-dimensional filters is an important area of research.

When the system has continuous time signal and continuous time observations, the solution of the filtering problem is a Stochastic PDE which can be seen as a generalization of the Fokker–Planck equation expressing the evolution of the density of a diffusion process. This filtering

*I am grateful towards Giuseppe Tinaglia and Alexander Pushnitski for help with geometry and topology. All remaining errors are my own.

equation is called Kushner–Stratonovich equation, and an unnormalized (simpler) version of it is known as the Duncan–Mortensen–Zakai Stochastic Partial Differential Equation. When observations are in discrete time, the filtering problem decomposes into a prediction step, given by the Fokker-Planck equation, and a correction step, given by Bayes formula.

In [11], [7] and [8] the Fisher metric is used to project the Kushner–Stratonovich (or the Fokker–Planck) equation onto an exponential family of probability densities, yielding the new class of approximate filters called *projection filters*. The projection filters are based on the differential geometric approach to statistics, as developed by [2] and [25]. It is also shown that one can choose the family so as to make the prediction step exact. Moreover, it is shown that for exponential families the projection filters coincide with the assumed density filters.

In [9, 10] the Gaussian projection filter is studied in the small-noise setting.

In the present paper we choose a different differential geometric structure based on a direct L^2 metric as opposed to the usual Hellinger distance and the related Fisher Information metric. We explain how this structure can be used to derive a different family of finite dimensional filters that form a good approximation for the solution of the nonlinear filtering problem. This structure is particularly suited to be applied to mixture families of distributions, similarly to how exponential families are well suited to work with the Fisher information metric. In this work we thus introduce the mixture projection filter, namely the projection filter based on the direct L^2 metric and based on a manifold given by a mixture of pre-assigned densities. The prediction step is given by a linear differential equation, whereas the correction step can be made exact by updating the basis functions for the tangent space of the manifold, namely the mixture components, at each observation time.

The exponential projection filter had a clear relationship with the assumed density filters, as documented in [8]. This method has a clear relationship with earlier Galerkin-based approaches to non-linear filtering, see for example [23]. In our opinion however the geometric structure and the exact projection make the method in this paper more rigorous than the usual Galerkin methods. We will explore in detail the relationship between our mixture projection filter based on direct L^2 metric and Galerkin methods in future research, where we will also implement the mixture projection filter equations numerically. We will also investigate the choice of the specific mixture family, starting with gaussian or lognormal mixtures.

2 Statistical manifolds

On the measurable space $(\mathbf{R}^n, \mathcal{B}(\mathbf{R}^n))$ we consider a non–negative and σ–finite measure λ, and we define $\mathcal{M}(\lambda)$ to be the set of all non–negative and finite measures μ which are absolutely continuous w.r.t. λ, and whose density

$$p_\mu = \frac{d\mu}{d\lambda}$$

is positive λ–a.e. For simplicity, we restrict ourselves to the case where λ is the Lebesgue measure on $\mathbf{R}^n$. We also assume that the total measure is normalized to one, so as to represent a probability measure. This in turn implies that p_μ integrates to one.

In the following, we denote by $H(\lambda)$ the set of all the densities of measures contained in $\mathcal{M}(\lambda)$. Notice that, as all the measures in $\mathcal{M}(\lambda)$ are non–negative and finite, we have that if p is a density in $H(\lambda)$ then $p \in L_1(\lambda)$, that is $\sqrt{p} \in L^2(\lambda)$. The above remark implies that the set $\mathcal{R}(\lambda) := \{\sqrt{p} : p \in H(\lambda)\}$ of square roots of densities of $H(\lambda)$ is a subset of $L^2(\lambda)$. Notice that all $\sqrt{p}$ in $\mathcal{R}(\lambda)$ satisfy $\sqrt{p(x)} > 0$, for almost every $x \in \mathbf{R}^n$.

We notice the important point that neither $H(\lambda)$ nor $\mathcal{R}(\lambda)$ are vector subspaces of L_1 or L^2 respectively. Hence, we cannot view them as normed subspaces or topological vector spaces.

We will be able to use the L^2 norm to define a *metric* in $\mathcal{R}$, but we will not be able to view $\mathcal{R}$ as a normed space.

2.1 The Hellinger distance

The above remarks lead to the definition of the following metric in $\mathcal{R}(\lambda)$, see Jacod and Shiryayev [18] or Hanzon [16], $d_{\mathcal{R}}(\sqrt{p_1}, \sqrt{p_2}) := \|\sqrt{p_1} - \sqrt{p_2}\|$,where $\|\cdot\|$ denotes the norm of the Hilbert space $L^2(\lambda)$. This leads to the Hellinger metric on $H(\lambda)$ (or $\mathcal{M}(\lambda)$), obtained by using the bijection between densities (or measures) and square roots of densities : if μ_1 and μ_2 are the measures having densities p_1 and p_2 w.r.t. λ, the Hellinger metric is defined as $d_{\mathcal{M}}(\mu_1, \mu_2) = d_H(p_1, p_2) = d_{\mathcal{R}}(\sqrt{p_1}, \sqrt{p_2})$. It can be shown, see e.g. [16], that the distance $d_{\mathcal{M}}(\mu_1, \mu_2)$ in $\mathcal{M}(\lambda)$ is defined independently of the particular λ we choose as basic measure, as long as both μ_1 and μ_2 are absolutely continuous w.r.t. λ. As one can always find a λ such that both μ_1 and μ_2 are absolutely continuous w.r.t. λ (take for example $\lambda := (\mu_1 + \mu_2)/2$), the distance is well defined on the set of all finite and positive measures on $(\Omega, \mathcal{F})$.

2.2 The L^2 direct distance

There is another possibility for defining a metric in H. We consider the following subset of H:

$$H_2(\lambda) = H(\lambda) \cap L^2(\lambda)$$

i.e. the set of L^2 densities. Notice that here we do not take the square root, but we use the L^2 structure directly on the densities. If we further assume that densities in H are bounded, then

$$H_2(\lambda) = H(\lambda)$$

since bounded positive functions that are in L_1 are also in L^2.

This structure leads to the definition of the following metric in $H_2(\lambda)$: $d_2(p_1, p_2) := \|p_1 - p_2\|$. H_2 with this metric is a metric space but, again, it is not a normed space, since it is not a vector space. We call this metric the direct L^2 distance, since it is taken directly on the densities rather than mapping them to their square roots.

2.3 Neither $(H(\lambda), d_H)$ nor $(H_2(\lambda), d_2)$ are L^2 Hilbert manifolds

Despite being subsets of L^2, neither $(H(\lambda), d_H)$ (or the equivalent $(\mathcal{R}(\lambda), d_{\mathcal{R}})$) nor $(H_2(\lambda), d_2)$ are locally homeomorphic to $L^2(\lambda)$, hence they are not manifolds modeled on $L^2(\lambda)$. Indeed, any open set of $L^2(\lambda)$ contains functions which are negative in a set with positive λ–measure. There is no open set of $L^2(\lambda)$ which contains only positive functions such as the functions of $H_2(\lambda)$ or $\mathcal{R}(\lambda)$.

2.4 Definition of Tangent vectors through the L^2 structure

Consider an open subset M of $L^2(\lambda)$. Let x be a point of M, and let $\gamma : (-\epsilon, \epsilon) \to M$ be a curve on M around x, i.e. a differentiable map between an open neighborhood of $0 \in \mathbf{R}$ and M such that $\gamma(0) = x$. We can define the tangent vector to γ at x as the Fréchet derivative

$D\gamma(0) : (-\epsilon, \epsilon) \rightarrow L^2(\lambda)$, i.e. the linear map defined in $\mathbf{R}$ around 0 and taking values in $L^2(\lambda)$ such that the following limit holds :

$$\lim_{|h|\to 0} \frac{\|\gamma(h) - \gamma(0) - D\gamma(0)\cdot h\|}{|h|} = 0 \ .$$

The map $D\gamma(0)$ approximates linearly the change of γ around x. Let $\mathcal{C}_x(M)$ be the set of all the curves on M around x. If we consider the space

$$L_xM := \{D\gamma(0) \ : \ \gamma \in \mathcal{C}_x(M)\} \ ,$$

of tangent vectors to all the possible curves on M around x, we obtain again the space $L^2(\lambda)$. This is due to the fact that for every $v \in L^2(\lambda)$ we can always consider the straight line $\gamma^v(h) := x + h\,v$. Since M is open, $\gamma^v(h)$ takes values in M for $|h|$ small enough. Of course $D\gamma^v(0) = v$, so that indeed $L_xM = L^2(\lambda)$.

2.5 Finite dimensional submanifold embedded in L^2

The situation becomes different if we consider an *m–dimensional* manifold N that is a subset of L^2 (and, possibly, a subset of $\mathcal{R}$ or H_2 above). As such, it can be endowed with the topology induced by the L^2 norm. Because N is m-dimensional, it is also locally homeomorphic to R^m.

We can consider the induced L^2 structure on N as follows : suppose $x \in N$, and define again

$$L_xN := \{D\gamma(0) \ : \ \gamma \in \mathcal{C}_x(N)\} \ .$$

This is a linear subspace of $L^2(\lambda)$ called the *tangent vector space* at x, which does not coincide with $L^2(\lambda)$ in general (due to the finite dimension of N, this tangent space will be m-dimensional). The set of all tangent vectors at all points x of N is called the *tangent bundle*, and will be denoted by $L\,N$. In our work we shall consider finite dimensional manifolds N embedded in $L^2(\lambda)$, which are contained in $\mathcal{R}(\lambda)$ or H_2 as a set, i.e. $N \subset \mathcal{R}(\lambda) \subset L^2(\lambda)$ or $N \subset H_2(\lambda) \subset L^2(\lambda)$, so that usually $x = \sqrt{p}$ or $x = p$, respectively.

We analyze the two cases separately.

2.6 Finite dimensional manifolds N in $(\mathcal{R}, d_{\mathcal{R}})$

If N is m–dimensional, it is locally homeomorphic to $\mathbf{R}^m$, and it may be described locally by a chart : if $\sqrt{p} \in N$, there exists a pair $(S^{1/2}, \phi)$ with $S^{1/2}$ open neighbourhood of $\sqrt{p}$ in N for the topology induced by $d_{\mathcal{R}}$ and $\phi : S^{1/2} \to \Theta$ homeomorphism of $S^{1/2}$ with the topology induced by $d_{\mathcal{R}}$ onto an open subset Θ of $\mathbf{R}^m$ with the usual topology of $\mathbf{R}^m$. By considering the inverse map i of ϕ,

$$\begin{array}{rcl} i : \Theta & \longrightarrow & S^{1/2} \\ \theta & \longmapsto & \sqrt{p(\cdot,\theta)} \end{array}$$

we can express $S^{1/2}$ as

$$i(\Theta) = \{\sqrt{p(\cdot,\theta)}\,,\, \theta \in \Theta\} = S^{1/2}.$$

We will work only with the single coordinate chart $(S^{1/2}, \phi)$ as it is done in [2]. From the fact that $(S^{1/2}, \phi)$ is a chart, it follows that

$$\{\frac{\partial i(\cdot,\theta)}{\partial \theta_1}, \cdots, \frac{\partial i(\cdot,\theta)}{\partial \theta_m}\}$$

is a set of linearly independent vectors in $L^2(\lambda)$. In such a context, let us see what the vectors of $L_{\sqrt{p(\cdot,\theta)}}S^{1/2}$ are. We can consider a curve in $S^{1/2}$ around $\sqrt{p(\cdot,\theta)}$ to be of the form $\gamma : h \mapsto \sqrt{p(\cdot,\theta(h))}$, where $h \mapsto \theta(h)$ is a curve in Θ around θ. Then, according to the chain rule, we compute the following Fréchet derivative:

$$D\gamma(0) = D\sqrt{p(\cdot,\theta(h))}\Big|_{h=0} = \sum_{k=1}^{m} \frac{\partial\sqrt{p(\cdot,\theta)}}{\partial\theta_k}\,\dot{\theta}_k(0) = \sum_{k=1}^{m} \frac{1}{2\sqrt{p(\cdot,\theta)}}\frac{\partial p(\cdot,\theta)}{\partial\theta_k}\,\dot{\theta}_k(0)\ .$$

We obtain that a basis for the tangent vector space at $\sqrt{p(\cdot,\theta)}$ to the space $S^{1/2}$ of square roots of densities of S is given by :

$$L_{\sqrt{p(\cdot,\theta)}}S^{1/2} = \mathrm{span}\{\frac{1}{2\sqrt{p(\cdot,\theta)}}\frac{\partial p(\cdot,\theta)}{\partial\theta_1},\cdots,\frac{1}{2\sqrt{p(\cdot,\theta)}}\frac{\partial p(\cdot,\theta)}{\partial\theta_m}\}\ . \tag{1}$$

As i is the inverse of a chart, these vectors are actually linearly independent, and they indeed form a basis of the tangent vector space. One has to be careful, because if this were not true, the dimension of the above spanned space could drop.

The inner product of any two basis elements is defined, according to the L^2 inner product

$$\langle\frac{1}{2\sqrt{p(\cdot,\theta)}}\,\frac{\partial p(\cdot,\theta)}{\partial\theta_i},\frac{1}{2\sqrt{p(\cdot,\theta)}}\,\frac{\partial p(\cdot,\theta)}{\partial\theta_j}\rangle = \tfrac{1}{4}\int\frac{1}{p(x,\theta)}\,\frac{\partial p(x,\theta)}{\partial\theta_i}\,\frac{\partial p(x,\theta)}{\partial\theta_j}\,d\lambda(x) = \tfrac{1}{4}\,g_{ij}(\theta)\ . \tag{2}$$

This is, up to the numeric factor $\frac{1}{4}$, the Fisher information metric, see for example [2], [22] and [1]. The matrix $g(\theta) = (g_{ij}(\theta))$ is called the Fisher information matrix.

Next, we introduce the orthogonal projection between any linear subspace V of $L^2(\lambda)$ containing the finite dimensional tangent vector space (1) and the tangent vector space (1) itself. Let us remember that our basis is not orthogonal, so that we have to project according to the following formula:

$$\begin{aligned}\Pi\ : V\ &\longrightarrow\ \mathrm{span}\{w_1,\cdots,w_m\}\\ v\ &\longmapsto\ \sum_{i=1}^{m}[\sum_{j=1}^{m}W^{ij}\,\langle v,w_j\rangle]\,w_i\end{aligned}$$

where $\{w_1,\cdots,w_m\}$ are m linearly independent vectors, $W := (\langle w_i,w_j\rangle)$ is the matrix formed by all the possible inner products of such linearly independent vectors, and (W^{ij}) is the inverse of the matrix W. In our context $\{w_1,\cdots,w_m\}$ are the vectors in (1), and of course W is, up to the numeric factor $\frac{1}{4}$, the Fisher information matrix given by (2). Then we obtain the following projection formula, where $(g^{ij}(\theta))$ is the inverse of the Fisher information matrix $(g_{ij}(\theta))$:

$$\begin{aligned}&\Pi_\theta\ : L^2(\lambda)\supseteq V\longrightarrow\mathrm{span}\{\frac{1}{2\sqrt{p(\cdot,\theta)}}\,\frac{\partial p(\cdot,\theta)}{\partial\theta_1},\cdots,\frac{1}{2\sqrt{p(\cdot,\theta)}}\,\frac{\partial p(\cdot,\theta)}{\partial\theta_m}\}\\ &\Pi_\theta[v] = \sum_{i=1}^{m}[\sum_{j=1}^{m}4g^{ij}(\theta)\,\langle v,\frac{1}{2\sqrt{p(\cdot,\theta)}}\,\frac{\partial p(\cdot,\theta)}{\partial\theta_j}\rangle]\,\frac{1}{2\sqrt{p(\cdot,\theta)}}\,\frac{\partial p(\cdot,\theta)}{\partial\theta_i}\ .\end{aligned} \tag{3}$$

Let us go back to the definition of tangent vectors for our statistical manifold. Amari [2] uses a different representation of tangent vectors to S at p. Without exploring all the assumptions

needed, let us say that Amari defines an isomorphism between the actual tangent space and the vector space

$$\text{span}\{\frac{\partial \log p(\cdot,\theta)}{\partial \theta_1}, \cdots, \frac{\partial \log p(\cdot,\theta)}{\partial \theta_m}\} \ .$$

On this representation of the tangent space, Amari defines a Riemannian metric given by

$$E_{p(\cdot,\theta)}\{\frac{\partial \log p(\cdot,\theta)}{\partial \theta_i} \ \frac{\partial \log p(\cdot,\theta)}{\partial \theta_j}\} \ ,$$

where $E_p\{\cdot\}$ denotes the expectation w.r.t. the probability density p. This is again the Fisher information metric, and indeed this is the most frequent definition of Fisher metric. In fact, it is easy to check that

$$\begin{aligned} E_{p(\cdot,\theta)}\{\frac{\partial \log p(\cdot,\theta)}{\partial \theta_i} \ \frac{\partial \log p(\cdot,\theta)}{\partial \theta_j}\} &= \int \frac{\partial \log p(x,\theta)}{\partial \theta_i} \ \frac{\partial \log p(x,\theta)}{\partial \theta_j} \, p(x,\theta) \, d\lambda(x) \\ &= \int \frac{1}{p(x,\theta)} \ \frac{\partial p(x,\theta)}{\partial \theta_i} \ \frac{\partial p(x,\theta)}{\partial \theta_j} \, d\lambda(x) = g_{ij}(\theta) \ . \end{aligned} \tag{4}$$

From the above relation and from (2) it is clear that, up to the numeric factor $\frac{1}{4}$, the Fisher information metric and the Hellinger metric coincide on the two representations of the tangent space to S at $p(\cdot,\theta)$.

There is another way of measuring how close two densities of S are. Consider the Kullback–Leibler information between two densities p and q of $H(\lambda)$:

$$K(p,q) := \int \log \frac{p(x)}{q(x)} \, p(x) \, d\lambda(x) = E_p\{\log \frac{p}{q}\} \ .$$

This is not a metric, since it is not symmetric and it does not satisfy the triangular inequality. When applied to a finite dimensional manifold such as S, both the Kullback–Leibler information and the Hellinger distance are particular cases of α–divergence, see [2] for the details. One can show that the Fisher metric and the Kullback–Leibler information coincide infinitesimally. Indeed, consider the two densities $p(\cdot,\theta)$ and $p(\cdot,\theta+d\theta)$ of S. By expanding in Taylor series, we obtain

$$\begin{aligned} K(p(\cdot,\theta), p(\cdot,\theta+d\theta)) &= -\sum_{i=1}^m E_{p(\cdot,\theta)}\{\frac{\partial \log p(\cdot,\theta)}{\partial \theta_i}\} \, d\theta_i \\ &\qquad - \sum_{i,j=1}^m E_{p(\cdot,\theta)}\{\frac{\partial^2 \log p(\cdot,\theta)}{\partial \theta_i \partial \theta_j}\} \, d\theta_i \, d\theta_j + O(|d\theta|^3) \\ &= \sum_{i,j=1}^m g_{ij}(\theta) \, d\theta_i \, d\theta_j + O(|d\theta|^3) \ . \end{aligned}$$

The interested reader is referred to [1].

Example 2.1 (The Gaussian family and the Fisher metric with canonical parameters). *We may consider the Fisher metric for the Gaussian family of densities. The Gaussian family may be defined as a particular exponential family, represented with canonical parameters θ, given by*

$$\{p(x,\theta) = \exp(\theta_1 x + \theta_2 x^2 - \psi(\theta)), \ \theta_2 < 0\}$$

where one has easily

$$\psi(\theta) = \tfrac{1}{2}\ln\left(\frac{\pi}{-\theta_2}\right) - \frac{\theta_1^2}{4\theta_2}$$

and the Fisher metric is

$$g(\theta) = \begin{bmatrix} -1/(2\theta_2) & \theta_1/(2\theta_2^2) \\ \theta_1/(2\theta_2^2) & 1/(2\theta_2^2) - \theta_1^2/(2\theta_2^3) \end{bmatrix}$$

The familiar representation of Gaussian densities is in terms of mean and variance, given respectively by

$$\mu = -\theta_1/(2\theta_2), \quad v = \sigma^2 = (1/\theta_2 - \theta_1^2/\theta_2^2)/2$$

The Fisher metric is used ideally to compute the distance between two infinitesimally near points $p(\cdot,\theta)$ and $p(\cdot,\theta+d\theta)$. Informally, we can write

$$d_H(p(\cdot,\theta), p(\cdot,\theta+d\theta)) = (d\theta)^T g(\theta) d\theta$$

Notice that the matrix changes when changing coordinates, whereas the distance must clearly be the same. Hence if we have another set of coordinates η related by diffeomorphism $\eta = \eta(\theta)$ to θ, with inverse $\theta = \theta(\eta)$, then clearly

$$d_H(p(\cdot,\eta), p(\cdot,\eta+d\eta)) = (d\eta)^T \ (\partial_\eta \theta(\eta))^T \ g(\theta(\eta)) \ \partial_\eta \theta(\eta) \ d\eta$$

where $\partial_\eta \theta(\eta)$ is the Jacobian matrix of the transformation. It follows that

$$g(\eta) = (\partial_\eta \theta(\eta))^T \ g(\theta(\eta)) \ \partial_\eta \theta(\eta)$$

Example 2.2 (The Gaussian family and the Fisher metric with expectation parameters). *We may consider the Fisher metric for the Gaussian family of densities in the parameters μ and v. These are related to the so called expectation parameters μ and $v+\mu^2$. With this coordinate system the Fisher metric is much simpler and the matrix is diagonal, resulting in*

$$g(\mu, v) = \frac{1}{v}\begin{bmatrix} 1 & 0 \\ 0 & 1/(2v) \end{bmatrix}$$

This can be derived either by applying the change of coordinates formula, or Eq. 2 directly, with the parameters θ_1, θ_2 replaced by μ, v.

2.7 Finite dimensional manifolds N in (H_2, d_2)

Alternatively, if we use H_2 instead of $\mathcal{R}$ as a set where N is contained, N can still be described locally by a chart : if $p \in N$, there exists a pair (S,ψ) with S open neighbourhood of p in N for the topology induced by d_2 and $\psi : S \to \Theta$ homeomorphism of S with the topology induced by d_2 onto an open subset Θ of $\mathbf{R}^m$ with the usual topology.

By considering the inverse map j of ψ,

$$\begin{array}{rcl} j : \Theta & \longrightarrow & S \\ \theta & \longmapsto & p(\cdot,\theta) \end{array}$$

we can express S as

$$j(\Theta) = \{p(\cdot,\theta)\,,\ \theta \in \Theta\} = S.$$

We will work only with the single coordinate chart (S, ψ). From the fact that (S, ψ) is a chart, it follows that

$$\{\frac{\partial j(\cdot,\theta)}{\partial \theta_1}, \cdots, \frac{\partial j(\cdot,\theta)}{\partial \theta_m}\}$$

is a set of linearly independent vectors in $L^2(\lambda)$. In such a context, let us see what the vectors of $L_{p(\cdot,\theta)}S$ are. We can consider a curve in S around $p(\cdot,\theta)$ to be of the form $\gamma : h \mapsto p(\cdot, \theta(h))$, where $h \mapsto \theta(h)$ is a curve in Θ around θ. Then, according to the chain rule, we compute the following Fréchet derivative:

$$D\gamma(0) = Dp(\cdot,\theta(h))|_{h=0} = \sum_{k=1}^{m} \frac{\partial p(\cdot,\theta)}{\partial \theta_k}\ \dot{\theta}_k(0) = \sum_{k=1}^{m} \frac{\partial p(\cdot,\theta)}{\partial \theta_k}\ \dot{\theta}_k(0)\ .$$

We obtain that a basis for the tangent vector space at $p(\cdot,\theta)$ to the space S is given by :

$$L_{p(\cdot,\theta)}S = \text{span}\{\frac{\partial p(\cdot,\theta)}{\partial \theta_1}, \cdots, \frac{\partial p(\cdot,\theta)}{\partial \theta_m}\}\ . \tag{5}$$

As j is the inverse of a chart, these vectors are actually linearly independent, and they indeed form a basis of the tangent vector space. One has to be careful, because if this were not true, the dimension of the above spanned space could drop.

The inner product of any two basis elements is defined, according to the L^2 inner product

$$\langle \frac{\partial p(\cdot,\theta)}{\partial \theta_i}, \frac{\partial p(\cdot,\theta)}{\partial \theta_j} \rangle = \int \frac{\partial p(x,\theta)}{\partial \theta_i}\ \frac{\partial p(x,\theta)}{\partial \theta_j}\, d\lambda(x) = h_{ij}(\theta)\ . \tag{6}$$

This is different from the Fisher information metric. The matrix $h(\theta) = (h_{ij}(\theta))$ is called the direct L^2 metric.

Next, we introduce the orthogonal projection between any linear subspace V of $L^2(\lambda)$ containing the finite dimensional tangent vector space (5) and the tangent vector space (5) itself.

$$\begin{aligned} &\Pi_\theta\ : L^2(\lambda) \supseteq V \longrightarrow \text{span}\{\frac{\partial p(\cdot,\theta)}{\partial \theta_1}, \cdots, \frac{\partial p(\cdot,\theta)}{\partial \theta_m}\} \\ &\Pi_\theta[v] = \sum_{i=1}^{m}[\sum_{j=1}^{m} h^{ij}(\theta)\ \langle v, \frac{\partial p(\cdot,\theta)}{\partial \theta_j}\rangle]\ \frac{\partial p(\cdot,\theta)}{\partial \theta_i}\ . \end{aligned} \tag{7}$$

Example 2.3 (The Gaussian family and the direct L^2 metric in canonical parameters). *We may consider the L^2 metric for the Gaussian family of densities introduced earlier. The L^2 metric is*

$$h(\theta) = \frac{1}{8}\frac{\sqrt{2}}{\sqrt{-\theta_2 \pi}} \begin{bmatrix} 1 & \frac{\theta_1}{-\theta_2} \\ \frac{\theta_1}{-\theta_2} & \frac{3}{4}\frac{1}{(-\theta_2)} + \frac{{\theta_1}^2}{{\theta_2}^2} \end{bmatrix}$$

and, as expected, it is different from the Fisher metric seen earlier.

Example 2.4 (The Gaussian family and the direct L^2 metric in expectation parameters). *We may consider the L^2 metric for the Gaussian family in the coordinates μ, v. The L^2 metric is*

$$h(\mu, v) = \frac{1}{8v\sqrt{v\pi}} \begin{bmatrix} 1 & 0 \\ 0 & \frac{3}{4v} \end{bmatrix}$$

and, as expected, it is different from the μ, v Fisher metric seen earlier, although it is still a diagonal matrix.

3 Exponential families and Mixture families

Earlier research in [7], [8], [5] and [6] illustrated in detail how the Hellinger distance and the related Fisher information metric are ideal tools when using the projection onto exponential families of densities. This idea was first sketched by Hanzon in [15]. The above references illustrate this by applying the above framework to the infinite dimensional stochastic PDE describing the optimal solution of the nonlinear filtering problem. This generates an approximate filter that is locally the closest filter in Fisher metric to the optimal one. The use of exponential families allows the correction step in the filtering algorithm to become exact, so that only the prediction step is approximated. Furthermore, and independently from the filtering application, exponential families and the Fisher metric are known to interact well. For example, the Fisher metric is obtained by double differentiation of the normalizing exponent in the exponential family and has a straightforward link with the expectation parameters. See for example [4].

The study of the projection filter for exponential families has been carried out in details int he above references, especially [11], [7] and [8].

However, besides exponential families, there is another general framework that is powerful in modeling probability densities, and this is the mixture family. Mixture distributions are ubiquitous in statistics and may account for important stylized features such as skewness, multi-modality and fat tails.

We define a mixture family as follows. Suppose we are given $m+1$ fixed squared integrable probability densities in H_2, say $\underline{q} = [q_1, q_2, \ldots, q_{m+1}]^T$. Suppose we define the following space of probability densities:

$$S^M(\underline{q}) = \{\theta_1 q_1 + \theta_2 q_2 + \cdots + \theta_m q_m + (1 - \theta_1 - \cdots - \theta_m) q_{m+1}, \theta_i \geq 0 \text{ for all } i, \quad \theta_1 + \cdots + \theta_m < 1\}$$

For convenience, define the transformation

$$\hat{\theta}(\theta) := [\theta_1, \theta_2, \ldots, \theta_m, 1 - \theta_1 - \theta_2 - \ldots - \theta_m]^T$$

for all θ. We will often write $\hat{\theta}$ instead of $\hat{\theta}(\theta)$ for brevity. With this definition,

$$S^M(\underline{q}) = \{\hat{\theta}(\theta)^T \underline{q}, \quad \theta_i \geq 0 \text{ for all } i, \quad \theta_1 + \cdots + \theta_m < 1\}$$

While for exponential families the Hellinger distance and the related Fisher metric are ideal, given also the expression (4), for mixture families it is less than ideal. For example, the calculation of the Fisher information matrix $g(\theta)$ becomes cumbersome, and the related projection is quite convoluted. Instead, if we consider the L^2 distance and the related structure, the metric $h(\theta)$ and the related projection become very simple. Indeed, one can immediately check from the definition of h that for the mixture family we have

$$\frac{\partial p(\cdot, \theta)}{\partial \theta_i} = q_i - q_{m+1}$$

and

$$h_{ij}(\theta) = \int (q_i(x) - q_m(x))(q_j(x) - q_m(x)) d\lambda(x) =: h_{ij}$$

i.e., the L^2 metric (and matrix) does not depend on the specific point θ of the manifold. The same holds for the tangent space at $p(\cdot, \theta)$, which is given by

$$L_{p(\cdot,\theta)} S = \text{span}\{q_1 - q_{m+1}, q_2 - q_{m+1}, \cdots, q_m - q_{m+1}\}$$

Also the L^2 projection becomes particularly simple:

$$\begin{aligned} &\Pi_\theta \ : L^2(\lambda) \supseteq V \longrightarrow \mathrm{span}\{q_1 - q_{m+1}, q_2 - q_{m+1}, \cdots, q_m - q_{m+1}\} \\ &\Pi_\theta[v] = \sum_{i=1}^{m} [\sum_{j=1}^{m} h^{ij} \ \langle v, q_j - q_{m+1} \rangle] \ (q_i - q_{m+1}) \ . \end{aligned} \tag{8}$$

It is therefore worthwhile to try and apply the L^2 metric and the related structure to the projection of the infinite dimensional filter onto the mixture family above.

4 The nonlinear filtering problem

In order to present the key geometric ideas without being overwhelmed by technicalities on stochastic PDEs, we consider the filtering problem with continuous time state and discrete time observations.

In this model, the state process is a continuous time stochastic differential equation

$$dX_t = f_t(X_t)\, dt + \sigma_t(X_t)\, dW_t \ ,$$

but only discrete–time observations are available

$$Z_n = h(X_{t_n}) + V_n \ ,$$

at times $0 = t_0 < t_1 < \cdots < t_n < \cdots$ regularly sampled, where $\{V_n \, , \, n \geq 0\}$ is a Gaussian white noise sequence independent of $\{X_t \, , \, t \geq 0\}$.

The nonlinear filtering problem consists in finding the conditional density $p_n(x)$ of the state X_{t_n} given the observations up to time t_n, i.e. such that $P[X_{t_n} \in dx \mid \mathcal{Z}_n] = p_n(x)\, dx$, where $\mathcal{Z}_n := \sigma(Z_0, \cdots, Z_n)$. We define also the prediction conditional density $p_n^-(x)\, dx = P[X_{t_n} \in dx \mid \mathcal{Z}_{n-1}]$. The sequence $\{p_n \, , \, n \geq 0\}$ satisfies a recurrent equation, and the transition from p_{n-1} to p_n is decomposed in two steps, as explained for example in [19].

There is first a prediction step: Between time t_{n-1} and t_n, we solve the Fokker–Planck equation

$$\frac{\partial p_t^n}{\partial t} = \mathcal{L}_t^* \, p_t^n \ , \qquad p_{t_{n-1}}^n = p_{n-1} \ .$$

The solution at final time t_n defines the prediction conditional density $p_n^- = p_{t_n}^n$.

We have then a second step, the correction step:

At time t_n, the newly arrived observation Z_n is combined with the prediction conditional density p_n^- via the Bayes rule

$$p_n(x) = c_n \ \Psi_n(x) \ p_n^-(x) \ , \tag{9}$$

where c_n is a normalizing constant, and $\Psi_n(x)$ denotes the likelihood function for the estimation of X_{t_n} based on the observation Z_n only, i.e.

$$\Psi_n(x) := \exp\Big\{ -\tfrac{1}{2} |Z_n - h(x)|^2 \Big\} \ . \tag{10}$$

5 The mixture projection filter (MPF)

We now introduce the mixture projection filter.

We will now work on the prediction step first, in order to derive the projected version of the Fokker Planck equation, living in the manifold S^M. We adopt the following technique. Take a curve in the mixture family S^M,

$$t \mapsto p(\cdot, \theta(t))$$

and notice that the left hand side of the Fokker Planck equation for this density would read

$$\frac{\partial p(\cdot,\theta(t))}{\partial t} = \sum_{i=1}^{m} \frac{\partial p(\cdot,\theta(t))}{\partial \theta_i} \frac{d}{dt}\theta_i(t) = \sum_{i=1}^{m} (q_i - q_{m+1}) \frac{d}{dt}\theta_i(t)$$

and project the right hand side of the Fokker Planck equation as

$$\Pi_\theta[\mathcal{L}_t^* p(\cdot,\theta)] = \sum_{i=1}^{m} [\sum_{j=1}^{m} h^{ij} \ \langle \mathcal{L}_t^* p(\cdot,\theta), q_j - q_{m+1} \rangle] \ (q_i - q_{m+1}) \ =$$

$$= \sum_{i=1}^{m} [\sum_{j=1}^{m} h^{ij} \ \langle p(\cdot,\theta), \mathcal{L}_t(q_j - q_{m+1}) \rangle] \ (q_i - q_{m+1})$$

where we used integration by parts in the last step. Now equating the two sides we obtain

$$\sum_{i=1}^{m} (q_i - q_{m+1}) \frac{d}{dt}\theta_i(t) = \sum_{i=1}^{m} [\sum_{j=1}^{m} h^{ij} \ \langle p(\cdot,\theta), \mathcal{L}_t(q_j - q_{m+1}) \rangle] \ (q_i - q_{m+1})$$

which yields the ordinary differential equation for the parameters θ of the projected density:

$$\frac{d}{dt}\theta_i(t) = \sum_{j=1}^{m} h^{ij} \ \langle p(\cdot,\theta), \mathcal{L}_t(q_j - q_{m+1}) \rangle$$

Now, by taking into account the structure of $p(\cdot,\theta)$ and the fact that such densities are linear in θ, we see that the above equation is a linear differential equation:

$$\frac{d}{dt}\theta_i(t) = \sum_{j=1}^{m} h^{ij} \ \left[\sum_{k=1}^{m} \theta_k \langle q_k, \mathcal{L}_t(q_j - q_{m+1}) \rangle + (1 - \theta_1 - \cdots - \theta_m) \langle q_{m+1}, \mathcal{L}_t(q_j - q_{m+1}) \rangle \right].$$

If we define, for two vector functions f and g, the matrix $\langle f, g \rangle$ and the vector $\mathcal{L}_t f$ as

$$(\langle f, g \rangle)_{i,j} := \langle f_i, g_j \rangle, \quad (\mathcal{L}_t f)_i := \mathcal{L}_t(f_i)$$

then we can write the above ODE in compact form as

$$\frac{d}{dt}\underline{\theta}(t) = h^{-1} \langle \mathcal{L}_t(\underline{q}_{1:m} - 1_m q_{m+1}), \ \underline{q} \rangle \ \hat{\theta}(\underline{\theta}(t))$$

where $\underline{q}_{1:m}$ is the vector with the first m components of $\underline{q}$, and 1_m is a m-dimensional (column) vector of ones.

In [7] and [8] it is shown that, by carefully choosing the exponential family, the Fisher metric exponential projection filter makes the correction step exact. In the mixture framework under the L^2 metric we are using now, this is harder to achieve unless we are willing to redefine the manifold at every correction step. Let us therefore focus on the correction step first. Suppose we are in $[t_{n-1}, t_n)$ and we obtained a prediction for the density up to t_n^-, whose parameter we call $\underline{\theta}_n^-$. At t_n a new observation Z_n arrives and we update the density. Substituting the prediction $p(\cdot, \underline{\theta}_n^-)$ into formula (9), we observe that the resulting density leaves the original mixture family $S^M(\underline{q})$. The updated density at t_n is

$$c_n \Psi_n(x) p(x, \theta_n^-) = c_n \Psi_n(x) \ \hat{\underline{\theta}}^T \underline{q}$$

and is outside $S^M(\underline{q})$. However, we may keep the update step exact by re-defining the basis functions q as follows.

Suppose that we change basis functions at every discrete date observation step. The first basis function vector is $\underline{q}^0$, then at update time t_1 we will select a new vector of basis functions $\underline{q}^1$, and so on. At every point in time we keep the vector $m+1$ dimensional. Suppose the basis functions in $[t_{n-1}, t_n)$ are $\underline{q}^{n-1}$. We run the prediction step up to t_n^-, getting $\underline{\theta}_n^-$. At time t_n, we define the new basis functions as

$$q_i^n(x) := c_{i,n}\Psi_n(x)q_i^{n-1}(x) \text{ for all } i = 1, \ldots, m+1$$

and where $c_{i,n}$ is the normalizing constant for the density on the right hand side. Every q_i^n is a normalized densities and we can define a mixture of such densities as the new space. In this case, the correction step amounts to set, at t_n:

Correction Step:

$$\text{At } t_n : \ \underline{\theta}_n = \underline{\theta}^n_{t_n}, \ \text{ and the new manifold is } S^M(\underline{q}^n)$$

We may now focus on the prediction step.

Before doing so, it is important to notice that the L^2 metric changes as well when we change the manifold, so that it is safe to index as follows:

$$h^n_{ij} = \int (q_i^n(x) - q_m^n(x))(q_j^n(x) - q_m^n(x))d\lambda(x)$$

Prediction step Between time t_{n-1} and t_n, we solve the ODE's

$$\frac{d}{dt}\underline{\theta}^n(t) = (h^{n-1})^{-1}\langle \mathcal{L}_t(\underline{q}^{n-1}_{1:m} - 1_m q^{n-1}_{m+1}),\ \underline{q}^{n-1}\rangle\ \hat{\theta}(\underline{\theta}^n(t)), \ \ \underline{\theta}^n_{t_{n-1}} := \underline{\theta}_{n-1}\ .$$

The solution at final time t_n defines the prediction parameters $\underline{\theta}_n^- = \theta^n_{t_n}$.

6 Conclusion and Further Research

We introduced a projection method for the space of probability distributions based on the differential geometric approach to statistics. This method makes use of a direct L^2 metric as opposed to the usual Hellinger distance and the related Fisher Information metric. We applied this apparatus to the nonlinear filtering problem. Past projection filters concentrated on the Fisher metric and the exponential families that made the filter correction step exact. Instead, in this work we introduce the mixture projection filter, namely the projection filter based on the direct L^2 metric and based on a manifold given by a mixture of pre-assigned densities. We derived the filter equations and showed how an update on the manifold functions can make the correction step exact. The prediction step is a simple linear ordinary differential equation.

We finally remarked that the exponential projection filter had a clear relationship with the assumed density filters, as documented in [8]. The mixture projection filter introduced here has a clear relationship with earlier Galerkin-based approaches that needs to be explored further. In future work we will also implement the mixture projection filter equations numerically and will investigate the choice of the specific mixture family.

References

[1] J. Aggrawal: Sur l'information de Fisher. In: Thories de l'Information (J. Kamp de Friet, ed.), Springer-Verlag, Berlin–New York 1974, pp. 111-117.

[2] Amari, S. Differential-geometrical methods in statistics, Lecture notes in statistics, Springer-Verlag, Berlin, 1985

[3] Bain, A., and Crisan, D. (2010). Fundamentals of Stochastic Filtering. Springer-Verlag, Heidelberg.

[4] Barndorff-Nielsen, O.E. (1978). Information and Exponential Families. John Wiley and Sons, New York.

[5] Brigo, D. Diffusion Processes, Manifolds of Exponential Densities, and Nonlinear Filtering, In: Ole E. Barndorff-Nielsen and Eva B. Vedel Jensen, editor, Geometry in Present Day Science, World Scientific, 1999

[6] Brigo, D, On SDEs with marginal laws evolving in finite-dimensional exponential families, STAT PROBABIL LETT, 2000, Vol: 49, Pages: 127 – 134

[7] Brigo, D, Hanzon, B, LeGland, F, A differential geometric approach to nonlinear filtering: The projection filter, IEEE T AUTOMAT CONTR, 1998, Vol: 43, Pages: 247 – 252

[8] Brigo, D, Hanzon, B, Le Gland, F, Approximate nonlinear filtering by projection on exponential manifolds of densities, BERNOULLI, 1999, Vol: 5, Pages: 495 – 534

[9] D. Brigo, On the nice behaviour of the Gaussian Projection Filter with small observation noise, *Systems and Control Letters* **26** (1995) 363–370

[10] D. Brigo, New results on the Gaussian projection filter with small observation noise, to appear in *Systems and Control Letters*.

[11] D. Brigo, *Filtering by Projection on the Manifold of Exponential Densities*, PhD Thesis, Free University of Amsterdam, 1996.

[12] Brigo, D., and Pistone, G. (1996). Projecting the Fokker-Planck Equation onto a finite dimensional exponential family. Available at arXiv.org

[13] Crisan, D., and Rozovskii, B. (Eds) (2011). The Oxford Handbook of Nonlinear Filtering, Oxford University Press.

[14] M. H. A. Davis, S. I. Marcus, An introduction to nonlinear filtering, in: M. Hazewinkel, J. C. Willems, Eds., *Stochastic Systems: The Mathematics of Filtering and Identification and Applications* (Reidel, Dordrecht, 1981) 53–75.

[15] Hanzon, B. A differential-geometric approach to approximate nonlinear filtering. In C.T.J. Dodson, Geometrization of Statistical Theory, pages 219 - 223,ULMD Publications, University of Lancaster, 1987.

[16] B. Hanzon, Identifiability, recursive identification and spaces of linear dynamical systems, CWI Tracts 63 and 64, CWI, Amsterdam, 1989

[17] M. Hazewinkel, S.I.Marcus, and H.J. Sussmann, Nonexistence of finite dimensional filters for conditional statistics of the cubic sensor problem, *Systems and Control Letters* **3** (1983) 331–340.

[18] J. Jacod, A. N. Shiryaev, Limit theorems for stochastic processes. Grundlehren der Mathematischen Wissenschaften, vol. 288 (1987), Springer-Verlag, Berlin,

[19] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, New York, 1970.

[20] J. Lévine, Finite dimensional realizations of stochastic PDE's and application to filtering, *Stochastics and Stochastic Reports* **43** (1991) 75–103.

[21] R.S. Liptser, A.N. Shiryayev, *Statistics of Random Processes I, General Theory* (Springer Verlag, Berlin, 1978).

[22] M. Murray and J. Rice - Differential geometry and statistics, Monographs on Statistics and Applied Probability 48, Chapman and Hall, 1993.

[23] Kenney, J., Stirling, W. Nonlinear Filtering of Convex Sets of Probability Distributions. Presented at the 1st International Symposium on Imprecise Probabilities and Their Applications, Ghent, Belgium, 29 June - 2 July 1999

[24] D. Ocone, E. Pardoux, *A Lie algebraic criterion for non-existence of finite dimensionally computable filters*, Lecture notes in mathematics 1390, 197–204 (Springer Verlag, 1989)

[25] Pistone, G., and Sempi, C. (1995). An Infinite Dimensional Geometric Structure On the space of All the Probability Measures Equivalent to a Given one. The Annals of Statistics 23(5), 1995