Maximum likelihood eigenfunctions of the Fokker Planck equation and Hellinger projection

Damiano Brigo  
Dept. of Mathematics  
Imperial College London  
180 Queen’s Gate  
damiano.brigo@imperial.ac.uk

Giovanni Pistone  
de Castro Statistics  
Collegio Carlo Alberto  
Via Real Collegio 30  
10024 Moncalieri, IT

March 15, 2016

Abstract

We apply the $L^2$ based Fisher-Rao vector-field projection by Brigo, Hanzon and LeGland (1999) to finite dimensional approximations of the Fokker Planck equation on exponential families. We show that if the sufficient statistics are chosen among the diffusion eigenfunctions the finite dimensional projection or the equivalent assumed density approximation provide the exact maximum likelihood density. The same result had been derived earlier by Brigo and Pistone (2016) in the infinite-dimensional Orlicz based geometry of Pistone and co-authors as opposed to the $L^2$ structure used here.

keywords Finite Dimensional Families of Probability Distributions, Exponential Families, Mixture Families, Hellinger distance, Fisher information metric, Direct $L^2$ metric, Kullback Leibler information, Eigenfunctions, Fokker Planck equation, Forward Kolmogorov Equation, Maximum Likelihood Estimator.

1 Introduction

We apply the $L^2$ based Hellinger distance and the metric it induces on finite dimensional families, namely the Fisher-Rao metric, to the vector-field projection of measure-valued evolution equations, and of the Fokker Planck equation in particular, onto finite dimensional manifolds. This follows the work initially sketched in Hanzon (1987) [9] and fully developed in Brigo et al (1999)[7] in the context of stochastic filtering and provides a general method to derive finite dimensional approximations of the Fokker Planck equation on exponential families. Using an alternative metric we also derive a second type of vector field projection that works better with mixture families. In the exponential family/Fisher Rao case, we show that if the sufficient statistics of the exponential family on which we project are chosen among the diffusion eigenfunctions then the finite dimensional projection or the equivalent assumed density approximation provide the exact maximum likelihood density. For the assumed density approximation see for example [10] for Gaussian families and [7] for exponential families. For references and a detailed literature review see [8], where this same result is presented in the conclusions of a much longer work and in the context of the richer statistical manifolds geometry of G. Pistone and co-authors [12], based on Orlicz spaces, rather than on the minimal $L^2$ structure we use here.
2 Statistical manifolds

For a full summary see [7]. We consider parametric families of probability densities, \( \{ p(\cdot, \theta), \theta \in \Theta \} \) with \( \Theta \) convex open set in \( \mathbb{R}^n \). The set of square roots of such densities is a subset of \( L^2 \) that we may view as a finite dimensional manifold. In general the \( L^2 \) distance between square roots of densities leads to the Hellinger distance. A curve in such a manifold is given by \( t \mapsto \sqrt{p(\cdot, \theta(t))} \). Differentiating with respect to \( t \) we obtain that all tangent vectors at \( \theta \) are in the space span \( \left\{ \frac{\partial \sqrt{p(\cdot, \theta)}}{\partial \theta_i}, i = 1, \ldots, n \right\} \). We can use the \( L^2 \) inner product \( \langle \cdot, \cdot \rangle \) to introduce an inner product on the tangent space and a metric. Define \( g_{i,j}(\theta)/4 = \left\langle \frac{\partial \sqrt{p(\cdot, \theta)}}{\partial \theta_i}, \frac{\partial \sqrt{p(\cdot, \theta)}}{\partial \theta_j} \right\rangle \). This is, up to the factor 4, the familiar Fisher-Rao information matrix. If we have a \( L^2 \) vector \( v \), we can project it via the orthogonal projection

\[
\Pi_{g}^{\theta}[v] = \sum_{i=1}^{m} \sum_{j=1}^{m} 4g_{ij}(\theta) \left\langle v, \frac{1}{2\sqrt{p(\cdot, \theta)}} \frac{\partial p(\cdot, \theta)}{\partial \theta_j} \right\rangle \frac{1}{2\sqrt{p(\cdot, \theta)}} \frac{\partial p(\cdot, \theta)}{\partial \theta_i}
\]

where upper indices denote the inverse matrix. A different possibility is a geometry that does not use the square root. This would be done most generally using a duality argument involving \( L^1 \) and \( L^\infty \), but here we will assume that all densities are square integrable, so that all \( p' \)’s we deal with are in \( L^2 \). Then we can mimick the above structure but without square roots. The curve is \( t \mapsto p(\cdot, \theta(t)) \), the tangent space is span \( \left\{ \frac{\partial p(\cdot, \theta)}{\partial \theta_i}, i = 1, \ldots, n \right\} \). Define \( \gamma_{i,j}(\theta) = \left\langle \frac{\partial p(\cdot, \theta)}{\partial \theta_i}, \frac{\partial p(\cdot, \theta)}{\partial \theta_j} \right\rangle \). This leads to what we called direct metric projection in [3],

\[
\Pi_{\gamma}^{\theta}[v] = \sum_{i=1}^{m} \sum_{j=1}^{m} \gamma_{ij}(\theta) \left\langle v, \frac{\partial p(\cdot, \theta)}{\partial \theta_j} \right\rangle \frac{\partial p(\cdot, \theta)}{\partial \theta_i}.
\]

There is another way of measuring how close two densities are. Consider the Kullback–Leibler information or relative entropy between two densities \( p \) and \( q \): \( K(p, q) := \int \log \frac{p(x)}{q(x)} p(x) \, dx \). This is not a metric, since it is not symmetric and it does not satisfy the triangular inequality. It is a classic result that the Fisher metric and the Kullback–Leibler information coincide infinitesimally. Indeed, by Taylor expansion it is easy to show that

\[
K(p(\cdot, \theta), p(\cdot, \theta + d\theta)) = \sum_{i,j=1}^{m} g_{ij}(\theta) d\theta_i d\theta_j + O(|d\theta|^3).
\]

3 Vector field projection of Fokker Planck

We first summarize the key results for projection in Hellinger distance \( g \) on exponential families and for projection in direct metric \( \gamma \) on mixture families. For a full account see [7] for the Hellinger case and [3] for the direct metric case. Consider a stochastic differential equation on a probability space \((\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})\) taking values in \( \mathbb{R}^N \),

\[
dX_t = f(X_t, t)dt + \sigma(X_t, t)dW_t, \quad X_0, \quad a(x, t) = \sigma(x, t)\sigma'(x, t)
\]

where the prime index denotes transposition. Let us start from the Fokker Planck equation for the probability density \( p_t(x) = p(x, t) \) of the solution \( X_t \) of our SDE. Examples of possible applications of approximating the Fokker Planck equation and its stochastic PDE extensions are given in [3], and include signal processing, stochastic- (local-) volatility modeling in quantitative
For our SDE the Fokker Planck equation reads
\[ \frac{\partial p(x,t)}{\partial t} = \mathcal{L}_t^\theta p(x,t), \quad \mathcal{L}_t^\theta p = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i} [f_i(\cdot, t) p] + \frac{1}{2} \sum_{i,j=1}^{n} \frac{\partial^2}{\partial x_i \partial x_j} [a_{ij}(\cdot, t) p]. \]

In many cases the solution of this equation is infinite dimensional. We may need finite dimensional approximations. We can project this parabolic PDE according to either the L2 direct metric (inducing the metric $\gamma(\theta)$) or, by deriving the analogous equation for $\sqrt{p_t}$,
\[ \frac{\partial \sqrt{p_t}}{\partial t} = \frac{1}{\sqrt{p_t}} \mathcal{L}_t^\theta p, \]
according to the Hellinger metric (inducing $g(\theta)$). The respective projections
\[ \frac{d}{dt} p(\cdot, \theta(t)) = \Pi_{\theta(t)}^\gamma [\mathcal{L}_t^\theta p(\cdot, \theta)], \quad \frac{d}{dt} \sqrt{p(\cdot, \theta(t))} = \Pi_{\theta(t)}^g \left[ \frac{1}{\sqrt{\sqrt{p(\cdot, \theta)}}} \mathcal{L}_t^\theta p(\cdot, \theta) \right] \]
transform the PDE into a finite dimensional ODE for $\theta$ via the chain rule:
\[ \frac{d}{dt} \rho(\cdot, \theta_t) = \sum_{j=1}^{m} \frac{\partial \rho(\cdot, \theta)}{\partial \theta_j} \delta_j(t), \quad \rho(p) = p \text{ or } \rho(p) = \sqrt{p}. \]

For brevity we write, for suitable functions $\varphi$, $\mathcal{E}_\theta[\varphi] := \int \varphi(x)p(x,\theta)dx$. The projections in direct and Hellinger metric respectively yield, after an integration by parts using the fact that $\mathcal{L}_t^\theta$ is the formal adjoint of $\mathcal{L}$:
\[ \dot{\theta}_i^\gamma = \sum_{j=1}^{m} \gamma^{ij}(\theta) E_{\theta_i} \left( \mathcal{L} \left( \frac{\partial p(\cdot, \theta)}{\partial \theta_j} \right) \right), \quad \dot{\theta}_i^g = \sum_{j=1}^{m} g^{ij}(\theta) E_{\theta_i} \left( \mathcal{L} \left( \frac{\partial \log p(\cdot, \theta)}{\partial \theta_j} \right) \right), \quad \theta_0^0. \]

We now choose specific families to carry out the projection. In particular, we project on the exponential family using the Hellinger distance and on the mixture family using the direct distance. The exponential families we consider are $\text{EF}(c)$, whose generic density is defined as $p(x, \theta) := \exp[\theta^t c(x) - \psi(\theta)]$. The functions $c$ are the sufficient statistics of the family, the parameters $\theta \in \mathbb{R}^n$ are the canonical parameters of the family. This works well with the Hellinger/Fisher Rao choice because the tangent space has a simple structure: square roots do not complicate issues thanks to the exponential structure. Moreover, the Fisher matrix has a simple structure: $\partial^2 \psi(\theta) = g_{ij}(\theta)$. The structure of the projection $\Pi$ is simple for exponential families. Finally, alternative coordinates, expectation parameters, defined via $\eta(\theta) = \mathcal{E}_\theta[c] = \partial_\theta \psi(\theta)$ are available, with the two coordinate systems $\eta$ and $\theta$ being bi-orthogonal.

For the direct metric, we will project on mixture families. We define a simple mixture family as follows. Given $m+1$ fixed squared integrable probability densities $q = [q_1, q_2, \ldots, q_{m+1}]^t$, define $\hat{\theta}(\theta) := [\theta_1, \theta_2, \ldots, \theta_m, 1 - \theta_1 - \theta_2 - \ldots - \theta_m]^t$ for all $\theta \in \mathbb{R}^m$. The simple mixture family (on the simplex) is defined as $p(\cdot, \theta) = \hat{\theta}(\theta)^t q$. If we consider the $L^2 / \gamma(\theta)$ distance, the metric $\gamma(\theta)$ itself and the related projection become very simple and do not depend on the point $\theta$. For example, $\frac{\partial p(\cdot, \theta)}{\partial \theta_i} = q_i - q_{m+1}$. Accordingly, the tangent space at $p(\cdot, \theta)$ does not depend on $\theta$ and is given by span$\{q_1 - q_{m+1}, q_2 - q_{m+1}, \ldots, q_m - q_{m+1}\}$. Also the $L^2$ projection becomes particularly simple and is equivalent to a classic Galerkin method. One has that the first Eq. specializes to
\[ \dot{\theta}_i^\gamma = \sum_{j=1}^{m} \gamma^{ij} \sum_{k=1}^{m+1} \hat{\theta}_k(t) \int (\mathcal{L} (q_j(x) - q_{m+1}(x))) (q_k(x) - q_{m+1}(x)) dx, \quad \theta_0^0. \]
which is a linear equation. See [3] for more details, we do not pursue mixtures further here and we focus on Hellinger distance and exponential families.

The projection of the Fokker Planck equation in Fisher metric has been introduced in [4]. First of all we refer to the the Hellinger projection above, the second Eq. in [4], as to the “vector field projection”. This is because we project the $L^2$ vector field of the Fokker Planck equation for $\sqrt{p}$ instant by instant onto the tangent space of $EF(c)$, thus obtaining a vector field in $EF(c)$. For the second Eq. in [3] to hold we need to ensure that $(\mathcal{L}_t p(\cdot, \theta))/\sqrt{p(\cdot, \theta)}$ is indeed an $L^2$ vector for all $t$ and $\theta$. This holds in turn if the condition

$$\sup_{t \geq 0} E_\theta \left[ \left\| \mathcal{L}_t p(\cdot, \theta) \right\|^2 \right] < \infty \text{ for all } \theta \in \Theta$$

holds, see [7], where it is shown that sufficient conditions guaranteeing this in $EF(c)$ are the following: $f_t$ and its first derivatives, $a_t$ and its first and second derivatives, and $c$ with its first and second derivatives have at most polynomial growth, and densities in $EF(c)$ integrate any polynomial. The paper [7] also discusses conditions under which all vector fields are well defined and the projection is well defined, see Theorem 5.4 for the special case $h = 0$.

Again for brevity, we write $E_\eta[\varphi] := \int \varphi(x)p(x; \eta(\theta))dx$. When projecting on $EF(c)$ the second Eq. [4] specializes into

$$\dot{\theta}_i^t = \sum_{j=1}^m g^{ij}(\theta_t) E_{\theta_t} [\mathcal{L}_i], \quad \theta_0^t \quad \text{or} \quad \eta_t^i = E_\eta [\mathcal{L}_i^c], \quad \eta_0^i$$

where the second equation has been obtained from the first by recalling that $d\eta(\theta) = g(\theta)d\theta$.

A natural question when projecting is how good the projection is locally? Also, one would like to have a measure for how far the projected evolution is, locally, from the original one. We now define a local projection residual as the $L^2$ norm of the Fokker Planck infinite dimensional vector field minus its finite-dimensional orthogonal projection. Define the vector field minus its projection and its norm as

$$\varepsilon_t(\theta) := \frac{\mathcal{L}_t p(\cdot, \theta)}{2\sqrt{p(\cdot, \theta)}} - \Pi^\theta \left[ \frac{\mathcal{L}_t p(\cdot, \theta)}{2\sqrt{p(\cdot, \theta)}} \right], \quad R_t^2 := \|\varepsilon_t(\theta)\|^2$$

The projection residual $R_t$ can be computed jointly with the projected equation evolution (6) to have a local measure of the goodness of the approximation involved in the projection.

Monitoring the projection residual and its peaks can be helpful in tracking the projection method performance, see also [4] for examples of $L^2$-based projection residuals in the more complex case of the Kushner-Stratonovich equations of nonlinear filtering. However, the projection residual only allows for a local approximation error numerical analysis. To illustrate this, assume for a moment that time is discrete $1, 2, 3, \ldots$. To make the point, we are artificially separating projection and propagation and the local and global errors. This is not completely precise but allows us to make an important point on our method. If we start from the manifold with a $p_0 = p(\cdot, \theta^0) \in EF(c)$ (omitting square roots in the notation and with upper indices denoting time), the Fokker Planck vector field driven by $\mathcal{L}_t$ will move us out of the manifold as the $L^2$ vector related to $\mathcal{L}_t p^0$ will not be tangent to $EF(c)$. We then project this $L^2$ vector on the tangent space of $EF(c)$ and follow it, obtaining a new $p(\cdot, \theta^1)$ in the manifold. Now we continue, again the Fokker Planck vector field driven by $\mathcal{L}_t p(\cdot, \theta^1)$ would bring us out of $EF(c)$, and to avoid this we project it onto the tangent space and follow the projected vector, obtaining $p(\cdot, \theta^2)$. The crucial point here is that this second step was done starting from an approximate
point \( p(\cdot, \theta^1) \) rather than from the true Fokker Planck density \( p_1 \). This means that, besides the local projection error measured by \( R_t \), we have a second error coming from the fact that we start the projection from the wrong point. If we leave the global approximation error analysis aside for a minute, the big advantage of the above method is that it does not require us to know the true solution of the Fokker Planck equation to be implemented. Indeed, Equation (6) works perfectly well without knowing the true solution \( p_t \).

4 Entropy point projection & eigenfunctions based ML

Now, to study the global error, we introduce a second projection method. We call this method “point projection”, since here we will project directly the \( L^2 \) densities onto \( \text{EF}(c) \) rather than their evolutions, and in particular we will make no use of tangent spaces. Point projection will require us to know the true solution, so as an approximation method it will be pointless. However, it will help us with the global error analysis, and a modification of the method based on the assumed density approximation will allow us to find an algorithm that does not require the true solution. The point projection method works as follows. Starting again from the manifold with a \( p_0 = p(\cdot, \theta^0) \in \text{EF}(c) \), the Fokker Planck vector field driven by \( \mathcal{L}^c \) will move us out of the manifold; we follow this vector field and reach \( p_1 \). To go back to \( \text{EF}(c) \), we project \( p_1 \) onto the exponential family by minimizing the relative entropy, or Kullback Leibler information of \( p_1 \) with respect to \( \text{EF}(c) \), finding the orthogonal projection of \( p_1 \) on \( \text{EF}(c) \). It is well known that the orthogonal projection in relative entropy is obtained by matching the sufficient statistics expectations of the true density. Namely, the projection is the particular exponential density of \( \text{EF}(c) \) with \( c \)-expectations \( \eta_1 = E_{p_1}[c] \). See for example [5] for a quick proof and an application to filtering in discrete time. We know that \( \text{EF}(c) \), besides \( \theta \), admits another important coordinate system, the expectation parameters \( \eta \). If one defines \( \eta(\theta) = E_{p(\theta)}[c] \) as above, then \( d\eta(\theta) = g(\theta)d\theta \) where \( g \) is the Fisher metric. Thus, we can take the \( \eta^1 \) above coming from the true density \( p_1 \) and look for the exponential density \( p(\cdot; \eta^1) \) sharing these \( c \)-expectations. This will be the closest in relative entropy to the true \( p_1 \) in \( \text{EF}(c) \). We then continue moving forward in time, iterating this algorithm.

The advantage of this method compared to the previous vector field based one is that we find at every time the best possible approximation (“maximum likelihood”) of the true solution in \( \text{EF}(c) \). The disadvantage is that in order to compute the projection at every time, such as for example \( \eta^1 = E_{p_1}[c] \), we need to know the true solution \( p_1 \) at that time. However it turns out that we can somewhat combine the two ideas and analyze the error if we invoke the assumed density approximation. This works as follows. Differentiate both sides of \( \eta_t = E_{p_t}[c] \) to obtain

\[
\dot{\eta}_t = \frac{d}{dt} \int c(x)p_t(x)dx = \int c(x)\frac{\partial p_t(x)}{\partial t}dx = \int c(x)\mathcal{L}_t^* p_t(x)dx = E_{p_t}[\mathcal{L}c]
\]

so that \( \dot{\eta}_t = E_{p_t}[\mathcal{L}c] \). This last equation is not a closed equation, since \( p_t \) in the right hand side is not characterized by \( \eta \). Thus, to be solved this equation should be coupled with the original Fokker Planck for \( p_t \) and we would still be in infinite dimension. However, at this point we can close the equation by invoking the assumed density approximation: we replace \( p_t \) with the exponential density \( p(\cdot; \eta_t) \). We obtain

\[
\frac{d}{dt} \dot{\eta}_t = E_{\dot{\eta}_t}[\mathcal{L}c].
\]

This is now a finite dimensional ODE for the expectation parameters. There is more: this last equation is the same as our earlier vector field based projected equation (6). This result
had been proven for nonlinear filtering in [7], although here we gave a more direct derivation. Intuitively, the result is related to the fact that the Fisher-Rao metric and relative entropy are infinitesimally equivalent, see Eq \( \text{(2)} \).

**Theorem 4.1** Closing the evolution equation for the relative entropy point projection of the Fokker Planck solution onto \( \text{EF}(c) \) by forcing an exponential density on the right hand side is equivalent to the approximation based on the vector field projection in Fisher metric.

We can now attempt an analysis of the error between the best possible projection \( \eta_t \) and the vector field based (or equivalently assumed density approximation based) projection \( \tilde{\eta}_t \). To do this, write \( \epsilon_t := \eta_t - \tilde{\eta}_t \), expressing the difference between the best possible approximation and the vector field projection / assumed density one, in expectation coordinates. Differentiating we see easily that \( \dot{\epsilon}_t = (E_p[c] - E_p[\tilde{\eta}_t][Lc]) \). Now suppose that the \( c \) statistics in \( \text{EF}(c) \) are chosen among the eigenfunctions of the operator \( L \), so that \( Lc = -\Lambda c \), where \( \Lambda \) is a \( n \times n \) diagonal matrix with the eigenvalues corresponding to the chosen eigenfunctions. Substituting, we obtain

\[
\dot{\epsilon}_t = -\Lambda (E_p[c] - E_p[\tilde{\eta}_t][c]) \quad \text{or} \quad \dot{\epsilon}_t = -\Lambda \epsilon_t \Rightarrow \epsilon_t = \exp(-\Lambda t) \epsilon_0
\]

so that if we start from the manifold \((\epsilon_0 = 0)\) the error is always zero, meaning that the vector field projection gives us the best possible Maximum Likelihood (ML) approximation. If we don’t start– from the manifold, ie if \( p_0 \) is outside \( \text{EF}(c) \), then the difference between the vector field approach and the best possible approximation dies out exponentially fast in time provided we have negative eigenvalues for the chosen eigenfunctions. This leads to the main result in this paper:

**Theorem 4.2** (ML for the Fokker Planck Equation & Fisher-Rao projection.) *The vector field projection approach leading to (6) provides also the best possible approximation of the Fokker Planck equation solution in relative entropy in the family \( \text{EF}(c) \), provided that the sufficient statistics \( c \) are chosen among the eigenfunctions of the adjoint operator \( L \) of the original Fokker Planck equation, and provided that \( \text{EF}(c) \) is an exponential family when using such eigenfunctions. In other words, under such conditions the Fisher Rao vector field projected equation (6) provides the exact maximum likelihood density for the solution of the Fokker Planck equation in the related exponential family.*

The choice or availability of suitable eigenfunctions is not always straightforward, except in a few simple cases. See [11] for a discussion on eigenfunctions for the Fokker Planck equation. For example, in the one dimensional case \( N = 1 \) where the diffusion is on a bounded domain \([\ell, r]\) with reflecting boundaries and strictly positive diffusion coefficient \( \sigma \) then the spectrum of the operator \( L \) is discrete, there is a stationary density and eigenfunctions can be expressed with respect to this stationary density, that could be taken as background measure, see [8]. For the case \( N > 1 \) special types of SDEs allow for a specific eigenfunctions/eigenvalue analysis, see again [11]. Further research is needed to explore the eigenfunctions approach in connection with maximum likelihood.

**References**


