Information Geometric Nonlinear Filtering: a Hilbert Space Approach

Nigel Newton (University of Essex)

Information Geometry and its Applications IV,
Liblice, June 2016

In honour of Shun-ichi Amari
on the occasion of his 80th birthday

Overview

- Nonlinear Filtering (recursive Bayesian estimation)
  - The need for a proper state space for posterior distributions
- The infinite-dimensional Hilbert manifold of probability measures, M, (and Banach variants)
- An M-valued Itô stochastic differential equation for the nonlinear filter
- Information geometric properties of the nonlinear filter

Nonlinear Filtering

- Markov "signal" process: $\left(X_t \in \mathbf{X},\ t \in [0,\infty)\right)$
  - $(\mathbf{X}, \mu)$ is a metric space, with reference probability measure μ
  - Eg. $\mathbf{X} = \mathrm{R}^d,\ \ \mu = N(0, I)$
- Partial "observation" process: $\left(Y_t \in \mathrm{R},\ t \in [0,\infty)\right)$

$$Y_t = \int_0^t h(X_s)ds + W_t$$

Brownian Motion, independent of X

Nonlinear Filtering

- Markov "signal" process: $\left(X_t \in \mathbf{X},\ t \in [0,\infty)\right)$
  - $(\mathbf{X}, \mu)$ is a metric space, with reference probability measure μ
  - Eg. $\mathbf{X} = \mathrm{R}^d,\ \ \mu = N(0, I)$
- Partial "observation" process: $\left(Y_t \in \mathrm{R},\ t \in [0,\infty)\right)$

$$Y_t = \int_0^t h(X_s)\,ds + W_t$$

Brownian Motion, independent of X

- Estimate X_t at each time t from its prior distribution P_t and the history of the observation:

$$Y_0^t := (Y_s,\ s \in [0, t])$$

- The linear-Gaussian case yields the *Kalman-Bucy filter*

Nonlinear Filtering

- Regular conditional (posterior) distribution: $\Pi_t : \Omega \to \mathcal{P}(\mathbf{X})$

$$\Pi_t(B) = \mathbf{P}\left(X_t \in B \mid Y_0^t\right)$$

- Π_t is a random probability measure evolving on $\mathcal{P}(\mathbf{X})$. How should we represent it?

Nonlinear Filtering

- Regular conditional (posterior) distribution: $\Pi_t : \Omega \to \mathcal{P}(\mathbf{X})$

$$\Pi_t(B) = \mathbf{P}\left(X_t \in B \mid Y_0^t\right)$$

- Π_t is a random probability measure evolving on $\mathcal{P}(\mathbf{X})$. How should we represent it?
- We could consider the conditional density (w.r.t μ), π_t
  - typical differential equation (Shiriyayev, Wonham, Stratonovich, Kushner):

$$"d\pi_t = \mathcal{A}\pi_t dt + \pi_t(h - \overline{h}_t)(dY_t - \overline{h}_t dt)" \qquad \left(\overline{h}_t := \int h(x)\Pi_t(dx)\right)$$

- Spaces of densities are not necessarily optimal

Mean-Square Errors

- Suppose $\mathbf{E}f(X_t)^2 < \infty$ for some $f : \mathbf{X} \to \mathrm{R}$
- Then $\bar{f}_t := \mathrm{E}_{\Pi_t} f$ minimises the mean-square error

$$\mathbf{E}(f(X_t) - \hat{f}_t)^2 = \mathbf{E}\left(\underbrace{\mathrm{E}_{\Pi_t}(f - \bar{f}_t)^2}_{\text{estimation error}} + \underbrace{(\bar{f}_t - \hat{f}_t)^2}_{\text{approximation error}} \right)$$

Mean-Square Errors

- Suppose $\mathbf{E}f(X_t)^2 < \infty$ for some $f : \mathbf{X} \to \mathrm{R}$
- Then $\bar{f}_t := \mathrm{E}_{\Pi_t} f$ minimises the mean-square error

$$\mathbf{E}(f(X_t) - \hat{f}_t)^2 = \mathbf{E}\left(\mathrm{E}_{\Pi_t}(f - \bar{f}_t)^2 + (\bar{f}_t - \hat{f}_t)^2\right)$$

estimation error + approximation error

- If $\hat{f}_t = \mathrm{E}_{\hat{\Pi}_t} f$ for some $\hat{\Pi}_t : \Omega \to \mathcal{P}(X)$, and $\Pi_t, \hat{\Pi}_t << \mu$ then

$$(\bar{f}_t - \hat{f}_t)^2 \le \mathrm{E}_\mu f^2 \mathrm{E}_\mu (\pi_t - \hat{\pi}_t)^2$$

and so the $L^2(\mu)$ norm on densities may be useful

Mean-Square Errors

- Suppose $\mathbf{E}f(X_t)^2 < \infty$ for some $f : \mathbf{X} \to \mathrm{R}$
- Then $\bar{f}_t := \mathrm{E}_{\Pi_t} f$ minimises the mean-square error

$$\mathbf{E}(f(X_t) - \hat{f}_t)^2 = \mathbf{E}\left(\mathrm{E}_{\Pi_t}(f - \bar{f}_t)^2 + (\bar{f}_t - \hat{f}_t)^2\right)$$

estimation error + approximation error

- If $\hat{f}_t = \mathrm{E}_{\hat{\Pi}_t} f$ for some $\hat{\Pi}_t : \Omega \to \mathcal{P}(X)$, and $\Pi_t, \hat{\Pi}_t << \mu$ then

$$(\bar{f}_t - \hat{f}_t)^2 \le \mathrm{E}_\mu f^2 \mathrm{E}_\mu (\pi_t - \hat{\pi}_t)^2$$

and so the $L^2(\mu)$ norm on densities may be useful

- Not if $f = 1_B$ and $\Pi_t(B)$ is very small (Eg. fault detection)

Mean-Square Errors

- Suppose $\mathbf{E} f(X_t)^2 < \infty$ for some $f : \mathbf{X} \to \mathrm{R}$
- Then $\bar{f}_t := \mathrm{E}_{\Pi_t} f$ minimises the mean-square error

$$\mathbf{E}(f(X_t) - \hat{f}_t)^2 = \mathbf{E}\left(\underbrace{\mathrm{E}_{\Pi_t}(f - \bar{f}_t)^2}_{\text{estimation error}} + \underbrace{(\bar{f}_t - \hat{f}_t)^2}_{\text{approximation error}} \right)$$

- If $\hat{f}_t = \mathrm{E}_{\hat{\Pi}_t} f$ for some $\hat{\Pi}_t : \Omega \to \mathcal{P}(\mathcal{X})$, and $\Pi_t, \hat{\Pi}_t << \mu$ then

$$(\bar{f}_t - \hat{f}_t)^2 \le \mathrm{E}_\mu f^2 \mathrm{E}_\mu (\pi_t - \hat{\pi}_t)^2$$

  and so the $L^2(\mu)$ norm on densities may be useful
- Not if $f = 1_B$ and $\Pi_t(B)$ is very small (Eg. fault detection)
- When topologised in this way, $\mathcal{P}(\mathbf{X})$ has a boundary

Multi-Objective Mean-Square Errors

- Maximising the L^2 error over square-integrable functions

$$\mathcal{M}(\hat{\Pi}_t \mid \Pi_t) := \sup_{f \in L^2(\Pi_t)} \frac{(\bar{f}_t - \hat{f}_t)^2}{\mathrm{E}_{\Pi_t}(f - \bar{f}_t)^2} \qquad \left(\frac{\text{approximation error}}{\text{estimation error}}\right)$$

$$= \sup_{f \in F} \left(\mathrm{E}_{\Pi_t} f(1 - d\hat{\Pi}_t / d\Pi_t)\right)^2$$

$$= \mathrm{E}_{\Pi_t}(1 - d\hat{\Pi}_t / d\Pi_t)^2$$

where $F := \left\{ f \in L^2(\Pi_t) : \ \bar{f}_t = 0, \ \mathrm{E}_{\Pi_t} f^2 = 1 \right\}$

Multi-Objective Mean-Square Errors

- Maximising the L^2 error over square-integrable functions

$$\mathcal{M}(\hat{\Pi}_t \mid \Pi_t) := \sup_{f \in L^2(\Pi_t)} \frac{(\bar{f}_t - \hat{f}_t)^2}{\mathrm{E}_{\Pi_t}(f - \bar{f}_t)^2} \qquad \left(\frac{\text{approximation error}}{\text{estimation error}}\right)$$

$$= \sup_{f \in F} \left(\mathrm{E}_{\Pi_t} f(1 - d\hat{\Pi}_t / d\Pi_t)\right)^2$$

$$= \mathrm{E}_{\Pi_t}(1 - d\hat{\Pi}_t / d\Pi_t)^2$$

where $\quad F := \left\{ f \in L^2(\Pi_t) : \ \bar{f}_t = 0, \ \mathrm{E}_{\Pi_t} f^2 = 1 \right\}$

- In time-recursive approximations, the accuracy of $\hat{\Pi}_t$ is affected by that of $\hat{\Pi}_s$ ($s < t$). This naturally induces multi-objective criteria at time s (nonlinear dynamics).

Geometric Sensitivity

- $\mathcal{M}$ is "geometrically sensitive". (It requires small probabilities to be approximated with greater absolute accuracy than large probabilities)
- When topologised by $\mathcal{M}$, $\mathcal{P}(\mathbf{X})$ does not have a boundary

Geometric Sensitivity

- $\mathcal{M}$ is "geometrically sensitive". (It requires small probabilities to be approximated with greater absolute accuracy than large probabilities)
- When topologised by $\mathcal{M}$, $\mathcal{P}(\mathbf{X})$ does not have a boundary
- This is highly desirable in the context of recursive Bayesian estimation, where conditional probabilities are repeatedly multiplied by the likelihood functions of new observations.

Geometric Sensitivity

- $\mathcal{M}$ is "geometrically sensitive". (It requires small probabilities to be approximated with greater absolute accuracy than large probabilities.)
- When topologised by $\mathcal{M}$, $\mathcal{P}(\mathbf{X})$ does not have a boundary.
- This is highly desirable in the context of recursive Bayesian estimation, where conditional probabilities are repeatedly multiplied by the likelihood functions of new observations.
- $\mathcal{M}$ is Pearson's χ^2 divergence. It belongs to the one-parameter family of α-divergences: $\mathcal{M} = \mathcal{D}_{-3}$

Geometric Sensitivity

- $\mathcal{M}$ is "geometrically sensitive". (It requires small probabilities to be approximated with greater absolute accuracy than large probabilities.)
- When topologised by $\mathcal{M}$, $\mathcal{P}(\mathbf{X})$ does not have a boundary.
- This is highly desirable in the context of recursive Bayesian estimation, where conditional probabilities are repeatedly multiplied by the likelihood functions of new observations.
- $\mathcal{M}$ is Pearson's χ^2 divergence. It belongs to the one-parameter family of α-divergences: $\mathcal{M} = \mathcal{D}_{-3}$
- It is too restrictive to use in practice

α-Divergences

- As $|\alpha|$ becomes larger $\mathcal{D}_\alpha$ becomes increasingly "geometrically sensitive"
- The case $\alpha = 0$ yields the *Hellinger metric*

α-Divergences

- As $|\alpha|$ becomes larger $\mathcal{D}_\alpha$ becomes increasingly "geometrically sensitive"
- The case $\alpha = 0$ yields the *Hellinger metric*
- The case $\alpha = \pm 1$ yields the *KL-Divergence:*

$$\mathcal{D}(P \mid Q) := \mathcal{D}_{-1}(P \mid Q) = \mathrm{E}_Q \frac{dP}{dQ} \log \frac{dP}{dQ}$$

- This is widely used in practice.

α-Divergences

- As $|\alpha|$ becomes larger $\mathcal{D}_\alpha$ becomes increasingly "geometrically sensitive"
- The case $\alpha = 0$ yields the *Hellinger metric*
- The case $\alpha = \pm 1$ yields the *KL-Divergence:*

$$\mathcal{D}(P \,|\, Q) := \mathcal{D}_{-1}(P \,|\, Q) = \mathrm{E}_Q \frac{dP}{dQ} \log \frac{dP}{dQ}$$

- This is widely used in practice.
- Symmetric error criteria may be appropriate, such as

$$\mathcal{D}(\hat{\Pi}_t \,|\, \Pi_t) + \mathcal{D}(\Pi_t \,|\, \hat{\Pi}_t)$$

Connections with Information Theory

- Conditional mutual information (un-averaged):

$$I(X;Y \mid Z) := \mathcal{D}\left(P_{XY|Z} \mid P_{X|Z} \otimes P_{Y|Z}\right)$$

- Additivity property:

$$I(X;(Y,Z)) = I(X;Z) + \mathbf{E}\, I(X;Y \mid Z)$$

Connections with Information Theory

- Conditional mutual information (unaveraged):

$$I(X;Y\,|\,Z) := \mathcal{D}\left(P_{XY|Z}\,|\,P_{X|Z}\otimes P_{Y|Z}\right)$$

- Additivity property:

$$I(X;(Y,Z)) = I(X;Z) + \mathbf{E}\,I(X;Y\,|\,Z)$$

- *Information Supply* to the nonlinear filter:

$$S(t) := I(X;Y_0^t)$$

- The filter continuously *fuses* new observation information

$$S(t) = S(s) + \mathbf{E}\,I(X;Y_s^t\,|\,Y_0^s)$$

Appropriate Metrics on $\mathcal{P}(\mathbf{X})$

- The KL divergence is bilinear in the density and its log (regarded as elements of dual spaces of functions).
- For $P, Q \in \mathcal{P}(\mathbf{X})$ with $P, Q << \mu$

$$\mathcal{D}(P \mid Q) = \left\langle p, \log p \right\rangle - \left\langle p, \log q \right\rangle$$

where p and q are the densities

Appropriate Metrics on $\mathcal{P}(\mathbf{X})$

- The KL divergence is <u>bilinear</u> in the density and its log (regarded as elements of dual spaces of functions).

- For $P, Q \in \mathcal{P}(\mathbf{X})$ with $P, Q << \mu$

$$\mathcal{D}(P \mid Q) = \langle p, \log p \rangle - \langle p, \log q \rangle$$

where p and q are the densities

- So we would like the metric to "control" both p and $\log p$

Maximal Exponential Model
(G. Pistone et al.)

- $\mathcal{E}(\mu) = \left\{ P \in \mathcal{P}(\mathbf{X}) : \ p = \exp(a - K_\mu(a)) \,|\, a \in S_\mu \right\}$
- Model space (exponential Orlicz):

$$B_\mu = \left\{ a : \mathbf{X} \to \mathrm{R} : \mathrm{E}_\mu \, a = 0, \ \mathrm{E}_\mu \cosh(\alpha a) < \infty \text{ for some } \alpha > 0 \right\}$$

Maximal Exponential Model

(G. Pistone et al.)

- $\mathcal{E}(\mu) = \left\{ P \in \mathcal{P}(\mathbf{X}) : \; p = \exp(a - K_\mu(a)) \,|\, a \in S_\mu \right\}$
- Model space (exponential Orlicz):

$$B_\mu = \left\{ a : \mathbf{X} \to \mathrm{R} : \mathrm{E}_\mu\, a = 0,\; \mathrm{E}_\mu \cosh(\alpha a) < \infty \text{ for some } \alpha > 0 \right\}$$

- Global Chart: s_μ: $\mathcal{E}(\mu) \to B_\mu$

$$s_\mu(P) := \log(p) - \mathrm{E}_\mu \log(p)$$

Maximal Exponential Model
(G. Pistone et al.)

- $\mathcal{E}(\mu) = \left\{ P \in \mathcal{P}(\mathbf{X}) :\ p = \exp(a - K_\mu(a)) \,|\, a \in S_\mu \right\}$
- Model space (exponential Orlicz):

$$B_\mu = \left\{ a : \mathbf{X} \to \mathrm{R} : \mathrm{E}_\mu\, a = 0,\ \mathrm{E}_\mu \cosh(\alpha a) < \infty \text{ for some } \alpha > 0 \right\}$$

- Global Chart: s_μ: $\mathcal{E}(\mu) \to B_\mu$

$$s_\mu(P) := \log(p) - \mathrm{E}_\mu \log(p)$$

- Mixture Map: η_μ: $\mathcal{E}(\mu) \to {}^*B_\mu$

$$\eta_\mu(P) := p - 1$$

Injective and of class C^∞, but not homeomorphic

The Hilbert Manifold M

- M is the subset of $\mathcal{P}(\mathbf{X})$ whose members have the following properties:

$$P \sim \mu, \quad \mathrm{E}_\mu\, p^2 < \infty \quad \text{and} \quad \mathrm{E}_\mu \log^2 p < \infty$$

The Hilbert Manifold M

- M is the subset of $\mathcal{P}(\mathbf{X})$ whose members have the following properties:

$$P \sim \mu, \quad \mathrm{E}_\mu\, p^2 < \infty \quad \text{and} \quad \mathrm{E}_\mu \log^2 p < \infty$$

- Model space:

$$H = L_0^2(\mu) = \left\{ a : \mathbf{X} \to \mathrm{R} : \mathrm{E}_\mu\, a = 0,\ \mathrm{E}_\mu\, a^2 < \infty \right\}$$

- Global Chart: $\phi : M \to H$

$$\phi(P) := p - 1 + \log p - \mathrm{E}_\mu \log p$$

The Hilbert Manifold M

- M is the subset of $\mathcal{P}(\mathbf{X})$ whose members have the following properties:

$$P \sim \mu, \quad \mathrm{E}_\mu\, p^2 < \infty \quad \text{and} \quad \mathrm{E}_\mu \log^2 p < \infty$$

- Model space:

$$H = L_0^2(\mu) = \left\{ a : \mathbf{X} \to \mathrm{R} : \mathrm{E}_\mu\, a = 0,\ \mathrm{E}_\mu\, a^2 < \infty \right\}$$

- Global Chart: $\phi : M \to H$

$$\phi(P) := p - 1 + \log p - \mathrm{E}_\mu \log p$$

- Proposition 1: ϕ is a bijection onto H

M as a Generalised Exponential Family

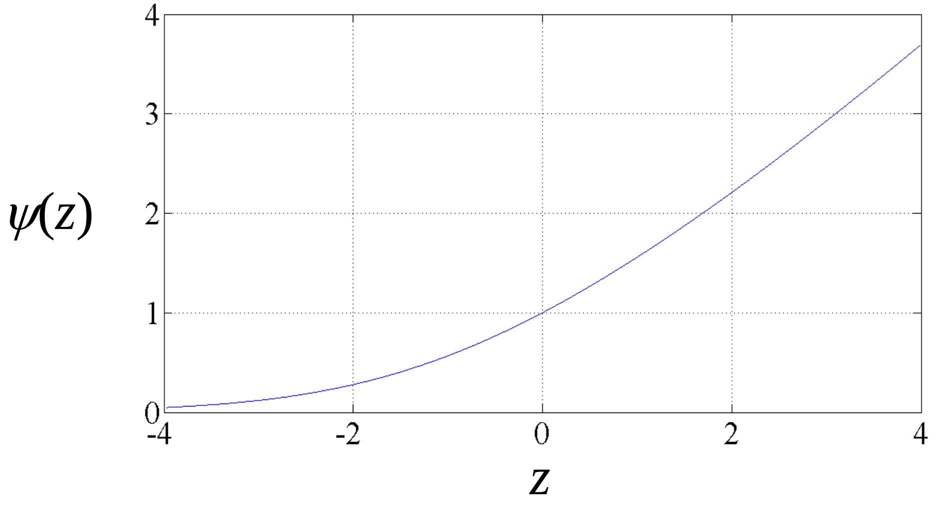
- The exponential function is replaced by the inverse of the function $(0, \infty) \ni y \rightarrow y - 1 + \log y \in \mathrm{R}$:

$$p(x) = \psi(a(x) + Z(a)) \quad \text{where} \quad a = \phi(P)$$

M as a Generalised Exponential Family

- The exponential function is replaced by the inverse of the function $(0, \infty) \ni y \to y - 1 + \log y \in \mathrm{R}$:

$$p(x) = \psi(a(x) + Z(a)) \quad \text{where} \quad a = \phi(P)$$



- <u>Convex</u>, <u>linear growth</u>, <u>bounded derivatives</u> of all orders.

Mixture and Exponential Maps

- The maps $m, e : M \to H$, defined by

$$m(P) = p - 1 \quad \text{and} \quad e(P) = \log p - \mathrm{E}_\mu \log p$$

are <u>injective</u>, but not homeomorphic (like η_μ of $\mathcal{E}(\mu)$)

Mixture and Exponential Maps

- The maps $m, e : M \to H$, defined by

$$m(P) = p - 1 \quad \text{and} \quad e(P) = \log p - \mathrm{E}_\mu \log p$$

are <u>injective</u>, but not homeomorphic (like η_μ of $\mathcal{E}(\mu)$)

- They satisfy:

$$\mathcal{D}(P \mid Q) + \mathcal{D}(Q \mid P) = \left\langle m(P) - m(Q),\, e(P) - e(Q) \right\rangle_H$$

- So that

$$\| m(P) - m(Q) \|_H^2 + \| e(P) - e(Q) \|_H^2 \le \| \phi(P) - \phi(Q) \|_H^2$$

and $$\mathcal{D}(P \mid Q) + \mathcal{D}(Q \mid P) \le \frac{1}{2} \| \phi(P) - \phi(Q) \|_H^2$$

The Tangent Bundle

- Global Chart: $\Phi : TM \to H \times H$

$$\Phi(P,U) \coloneqq (\phi(P), U\phi_P)$$

The Tangent Bundle

- Global Chart: $\Phi : TM \to H \times H$

$$\Phi(P,U) \coloneqq (\phi(P), U\phi_P)$$

- m and e representations:

$$\Phi_m(P,U) \coloneqq (\phi(P), Um_P) \in H \times H, \quad \Phi_e(P,U) \coloneqq (\phi(P), Ue_P) \in H \times H$$

Injective but not homeomorphic

The Tangent Bundle

- Global Chart: $\Phi : TM \to H \times H$

$$\Phi(P,U) := (\phi(P), U\phi_P)$$

- m and e representations:

$$\Phi_m(P,U) := (\phi(P), Um_P) \in H \times H, \quad \Phi_e(P,U) := (\phi(P), Ue_P) \in H \times H$$

Injective but not homeomorphic

- The Fisher metric: for $U, V \in T_P M$

$$\langle U, V \rangle_P := -UV\mathcal{D}_P = \langle Um_P, Ve_P \rangle_H \qquad \text{(Eguchi)}$$

The Tangent Bundle

- Global Chart: $\Phi : TM \to H\times H$

$$\Phi(P,U) := (\phi(P), U\phi_P)$$

- m and e representations:

$$\Phi_m(P,U) := (\phi(P), Um_P) \in H \times H, \qquad \Phi_e(P,U) := (\phi(P), Ue_P) \in H \times H$$

  Injective but not homeomorphic

- The Fisher metric: for $U, V \in T_P M$

$$\langle U, V \rangle_P := -UV\mathcal{D}_P = \langle Um_P, Ve_P \rangle_H \qquad \text{(Eguchi)}$$

- $(T_P M, <\cdot\,,\cdot>)$ is an inner product space with

$$\| U \|_P = \langle Um_P, Ue_P \rangle_H \leq \| U\phi \|_H$$

e and m Parallel Transport

- These are obtained by considering the inclusions:

$$\Phi_m(TM) \subset H \times H \quad \text{and} \quad \Phi_e(TM) \subset H \times H$$

together with the parallel transport on $H \times H$ defined by:

$$T_{a,b}(a,u) = (b,u)$$

e and m Parallel Transport

- These are obtained by considering the inclusions:

$$\Phi_m(TM) \subset H \times H \quad \text{and} \quad \Phi_e(TM) \subset H \times H$$

together with the parallel transport on $H \times H$ defined by:

$$T_{a,b}(a,u) = (b,u)$$

- Like the m parallel transport on the maximal exponential model, they coincide with m parallel transport on the tangent bundle only in special cases.

- α-parallel transports can be defined in the same way on statistical Hilbert bundles.

Submanifolds

Like the maximal exponential model, M admits many useful submanifolds. For example…

- Proposition 2: If $N \subset M$ is a finite-dimensional exponential family, then it is a C^∞-embedded submanifold of M, on which m, e and $\mathcal{D}$ are of class C^∞

Submanifolds

Like the maximal exponential model, M admits many useful submanifolds. For example…

- Proposition 2: If $N \subset M$ is a finite-dimensional exponential family, then it is a C^∞-embedded submanifold of M, on which m, e and $\mathcal{D}$ are of class C^∞

- Example: the non-singular Gaussian measures on R^m form a C^∞-embedded submanifold of $M(\mathrm{R}^m, \mu)$, where

$$\mu(dx) := 2^{-m} \exp(-|x|)\, dx$$

Submanifolds

Like the maximal exponential model, M admits many useful submanifolds. For example…

- Proposition 2: If $N \subset M$ is a finite-dimensional exponential family, then it is a C^∞-embedded submanifold of M, on which m, e and $\mathcal{D}$ are of class C^∞
- Example: the non-singular Gaussian measures on R^m form a C^∞-embedded submanifold of $M(\mathrm{R}^m, \mu)$, where

$$\mu(dx) := 2^{-m} \exp(-|x|)\, dx$$

- Similar results hold for mixture models and α-models
- Subspaces of H also provide natural submanifolds of M

Banach Variants

- The α-divergences are twice differentiable on M.
- Greater regularity can be obtained by the use of stronger topologies on the model space: $L^{\lambda}(\mu)$, for $\lambda > 2$
- This enables the definition of α-covariant derivatives on the statistical bundles mentioned above.
- Details in:

N.J. Newton, Infinite-dimensional statistical manifolds based on a balanced chart, *Bernoulli* 22, 711-731 (2016)

Nonlinear Filtering

- Markov "signal" process: $\left(X_t \in \mathbf{X},\ t \in [0,\infty)\right)$
  - $\left(\mathbf{X}, \mu\right)$ is a metric space, with reference probability measure μ
  - Eg. $\mathbf{X} = \mathrm{R}^d,\ \ \mu = N(0, I)$
- Partial "observation" process: $\left(Y_t \in \mathrm{R},\ t \in [0,\infty)\right)$

$$Y_t = \int_0^t h(X_s)\,ds + W_t$$

- Estimate X_t at each time t from its prior distribution P_t and the history of the observation:

$$Y_0^t := (Y_s,\ s \in [0,t])$$

- Typical equation for the density:

$$d\pi_t = \mathcal{A}\pi_t dt + \pi_t (h - \overline{h}_t) d\overline{W}_t \quad \text{where } d\overline{W}_t := dY_t - \overline{h}_t dt$$

M-Valued Nonlinear Filters

Proposition 3: Under some technical conditions:

1. $\mathbf{P}\left(\Pi_t \in M \text{ for all } t \geq 0\right) = 1$

M-Valued Nonlinear Filters

Proposition 3: Under some technical conditions:

1. $\mathbf{P}\left(\Pi_t \in M \text{ for all } t \geq 0\right) = 1$

2. The coordinate representation $\phi(\Pi)$ satisfies the following (infinite-dimensional) Itô equation

$$d\phi(\Pi_t) = (u_t - \zeta_t)dt + v_t d\overline{W}_t$$

where

$$u_t := \Lambda(1 + \pi_t^{-1})\mathcal{A}\pi_t$$

$$\zeta_t := \Lambda(h - \overline{h}_t)^2 / 2$$

$$v_t := \Lambda(\pi_t + 1)(h - \overline{h}_t)$$

$$\Lambda f = \begin{cases} f - E_\mu f & \text{if } f \in L^2(\mathbf{X}, \mu) \\ 0 & \text{otherwise} \end{cases}$$

Components

- Since H is of countable dimension, it admits a complete orthonormal basis $(\eta_i, i = 1, 2, 3, \dots)$
- So the filter equations can be written in terms of the components:

$$\phi(\Pi_t)^i := \left\langle \phi(\Pi_t), \eta_i \right\rangle_H \quad \text{for} \quad i = 1, 2, 3, \dots$$

Components

- Since H is of countable dimension, it admits a complete orthonormal basis $(\eta_i,\ i = 1, 2, 3, \ldots)$
- So the filter equations can be written in terms of the components:

$$\phi(\Pi_t)^i := \left\langle \phi(\Pi_\mathrm{t}), \eta_i \right\rangle_H \quad \text{for} \quad i = 1, 2, 3, \ldots$$

- The Fisher metric can be expressed in terms of the (η_i)

$$\left\langle U, V \right\rangle_P = G(P)_{i,j} u^i v^j$$

where $G(P)_{i,j} = \left\langle D_i, D_j \right\rangle_P$, $(P, D_i) = \Phi^{-1}(\phi(P), \eta_i)$ and $U = u^i D_i$

Components

- Since H is of countable dimension, it admits a complete orthonormal basis $(\eta_i,\ i = 1, 2, 3, \ldots)$
- So the filter equations can be written in terms of the components:

$$\phi(\Pi_t)^i := \left\langle \phi(\Pi_t), \eta_i \right\rangle_H \quad \text{for} \quad i = 1, 2, 3, \ldots$$

- The Fisher metric can be expressed in terms of the (η_i)

$$\left\langle U, V \right\rangle_P = G(P)_{i,j} u^i v^j$$

where $G(P)_{i,j} = \left\langle D_i, D_j \right\rangle_P$, $(P, D_i) = \Phi^{-1}(\phi(P), \eta_i)$ and $U = u^i D_i$

- The basis can be chosen to suit the problem (wavelets)
- Truncated series could be used in approximations

Quadratic Variation

- Semimartingales on M have well-defined quadratic variation in the Fisher metric; in particular

$$[\Pi]_t := \int_0^t G(\Pi_s)_{i,j}\, d\left[\phi(\Pi)^i, \phi(\Pi)^j\right]_s$$

Quadratic Variation

- Semimartingales on M have well-defined quadratic variation in the Fisher metric; in particular

$$[\Pi]_t := \int_0^t G(\Pi_s)_{i,j}\, d\left[\phi(\Pi)^i, \phi(\Pi)^j\right]_s$$

- Proposition 4: Under the conditions of Proposition 3:

$$I(X; Y_s^t \mid Y_0^s) = \frac{1}{2}\mathbf{E}\left([\Pi]_t - [\Pi]_s \mid Y_0^s\right)$$

Quadratic Variation

- Semimartingales on M have well-defined quadratic variation in the Fisher metric; in particular

$$[\Pi]_t := \int_0^t G(\Pi_s)_{i,j}\, d\left[\phi(\Pi)^i, \phi(\Pi)^j\right]_s$$

- Proposition 4: Under the conditions of Proposition 3:

$$I(X; Y_s^t \mid Y_0^s) = \frac{1}{2}\mathbf{E}\left([\Pi]_t - [\Pi]_s \mid Y_0^s\right)$$

- Results of this type are of interest in *Non-equilibrium Statistical Mechanic*s, where interactions between systems set up “flows of entropy”.

Finite Dimensional Filters

- A number of filters are known to evolve on finite-dimensional exponential manifolds (Kalman-Bucy, Benes…)

Finite Dimensional Filters

- A number of filters are known to evolve on finite-dimensional exponential manifolds (Kalman-Bucy, Benes…)
- Proposition 5: Under some technical conditions, Π is the unique strong solution of the following intrinsic Stratonovich equation on such a manifold:

$$\circ\, d\Pi_t = \left(U_t(\Pi_t) - \frac{1}{2} \nabla^{(-1)}_{V_t} V_t(\Pi_t) \right) dt + V_t(\Pi_t) \circ d\overline{W}_t$$

where $\nabla^{(-1)}$ is Amari's (-1)-covariant derivative, and U and V are suitably regular, time-dependent vector fields.

Projections onto Submanifolds

(Brigo, Pistone, Hanzon, Le Gland, Armstrong…)

1. Choose a suitable C^2-embedded finite-dimensional submanifold $N \subset M$.
2. The tangent space $T_P N$ is complete w.r.t. the Fisher metric.
3. Evaluate $u_t - z_t$ and v_t at points of N. (These are tangent vectors of M.)
4. Project onto $T_P N$ in the Fisher metric to obtain an evolution equation on N.

Projections onto Submanifolds

(Brigo, Pistone, Hanzon, Le Gland, Armstrong…)

1. Choose a suitable C^2-embedded finite-dimensional submanifold $N \subset M$.
2. The tangent space $T_P N$ is complete w.r.t. the Fisher metric.
3. Evaluate $u_t - z_t$ and v_t at points of N. (These are tangent vectors of M.)
4. Project onto $T_P N$ in the Fisher metric to obtain an evolution equation on N.

- The Hilbert manifold is very suited to this purpose
- One could also project in the model space metric

Details in:

1. N.J. Newton, An infinite-dimensional statistical manifold modelled on Hilbert space, *J. Functional Anal.* 263, 1661-1681 (2012).
2. N.J. Newton, Information Geometric Nonlinear Filtering, *Infin. Dimens. Anal. Quantum Probab. Relat. Top.* 18, 1550014 (2015).
3. N.J. Newton, Infinite-dimensional statistical manifolds based on a balanced chart, *Bernoulli* 22, 711-731 (2016)

Related Work

4. J. Armstrong and D. Brigo, Stochastic filtering via L2 projection on mixture manifolds with computer algorithms and numerical examples, arXiv:1303.6236 (2013)
5. D. Brigo, B. Hanzon and F. Le Gland, Approximate nonlinear filtering on exponential manifolds of densities, *Bernoulli* 5, 495-534 (1999).
6. D. Brigo and G. Pistone, Projection-based dimensionality reduction for measure-valued evolution equations in statistical manifolds, arXiv:1601.04189 (2016)
7. A. Cena and G. Pistone, Exponential statistical manifold, *Ann. Inst. Statist. Math*. 59, 27-56 (2007)

Related Work (cont.)

8. P. Gibilisco and G. Pistone, Connections on non-parametric statistical manifolds by Orlicz space geometry, *Infin. Dimens. Anal. Quantum Probab. Relat. Top. 1*, 325-347 (1998)
9. M.R. Grasselli, Dual connections in non-parametric classical information geometry, *Ann. Inst. Statist. Math*. 62, 873-896 (2010)
10. G. Pistone and M.P. Rogantin, The exponential statistical manifold: mean parameters, orthogonality and space transformations, *Bernoulli* 5, 721-760 (1999).
11. G. Pistone and C. Sempi, An infinite-dimensional geometric structure on the space of all probability measures equivalent to a given one, *Ann. Statist*. 23, 1543-1561 (1995).