Using prior knowledge in dynamic settings for multivariate Gaussian processes

Dan Cornford

d.cornford@aston.ac.uk

Aston University, Birmingham, UK

http://wiki.aston.ac.uk/DanCornford

Joint work with: Yuan Shen, Michael Vrettas, Manfred Opper, Remi Barillec and thanks to Ross Bannister.

SLIM, 24 July 2009, Manchester

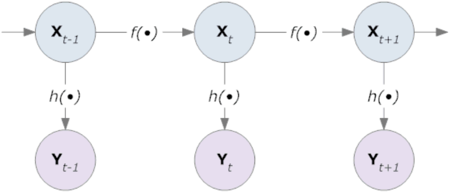
Outline

In this talk I will cover how prior knowledge can be used to help formulate joint structure in multivariate settings. In particular I will address:

- the context I am thinking about this in – data assimilation;
- some older, well known methods – balance and joint structure;
- some more recent, but still well known methods – ensemble (unscented) methods;
- our recent variational Bayesian approach;
- open questions and future directions.

I think that almost all interesting structure in real systems arises through some (unobserved / able?) dynamics, so understanding the dynamics is one way to model joint structure in almost all systems.

The basic setting: dynamical systems



- I'll work in the state space modelling formalism; i.e. treat the state as a latent process.
- A dynamic model in this context is typically a model defined by a set of differential or difference equations.

The main things we will consider

$\mathbf{X} \equiv \mathbf{X}(\mathbf{s}, t)$ – the simulator state
$\mathbf{s}$ – spatial position
t – time

$\mathbf{X}_t = \mathbf{X}(\mathbf{s}, t)$ at time t
$\mathbf{X}_{t+1} = f(\mathbf{X}_t) + \boldsymbol{\eta}_t$ – simulator
$\mathbf{Y}_t = h(\mathbf{X}_t) + \epsilon_t$ – observation

Inference in dynamical systems (data assimilation)

Assume we have a sequence of discrete time observations from $t = t_0$ to $t = t_k$, which I will denote $\mathbf{Y}_{t_0:t_k}$. The corresponding simulator states are given by $\mathbf{X}_{t_0:t_k}$.

In state inference we are interested in $\mathrm{p}(\mathbf{X}_t \mid \mathbf{Y}_{t_0:t_k})$ which is:

- smoothing if $t < t_k$;
- filtering if $t = t_k$;
- prediction if $t > t_k$.

Here I will largely stick with the filtering problem, and focus on the static (state at a fixed time) data assimilation problem of inferring $\mathrm{p}(\mathbf{X}_{t_k} \mid \mathbf{Y}_{t_0:t_k})$, although I will revisit this later.
Note here $\mathbf{X}$ is assumed to be a random variable, which can be induced from many aspects, e.g. initial condition error, $\mathrm{p}(\mathbf{X}_{t_0})$, observation error, ϵ, model error, $\boldsymbol{\eta}$.

Filtering in dynamical systems

Filtering is the most simple algorithm involving a prediction step and an update step:

- Prediction:

$$\mathrm{p}(\mathbf{X}_{t_k} \,|\, \mathbf{Y}_{t_0:t_{k-1}}) = \int \mathrm{p}(\mathbf{X}_{t_k} \,|\, \mathbf{X}_{t_{k-1}}; f)\mathrm{p}(\mathbf{X}_{t_{k-1}} \,|\, \mathbf{Y}_{t_0:t_{k-1}}) d\mathbf{X}_{t_{k-1}} \ .$$

- Update:

$$\mathrm{p}(\mathbf{X}_{t_k} \,|\, \mathbf{Y}_{t_0:t_k}) \propto \mathrm{p}(\mathbf{Y}_{t_k} \,|\, \mathbf{X}_{t_k}; h)\mathrm{p}(\mathbf{X}_{t_k} \,|\, \mathbf{Y}_{t_0:t_{k-1}}) \ .$$

In words this is:

- Prediction: passing a distribution through a (non-linear) function $\mathbf{X}_{t+1} = f(\mathbf{X}_t) + \boldsymbol{\eta}$.
- Update: Bayesian update of a static latent variable model with likelihood derived from $\mathbf{Y}_t = h(\mathbf{X}_t) + \boldsymbol{\epsilon}_t$

What is the simulator, f, and the state, $\mathbf{X}$?

E.g., the model (conservation) equations for the atmosphere are:

$$\begin{aligned}
\frac{D\mathbf{v}}{Dt} &= -\frac{1}{\rho}\nabla p - \nabla\phi - 2\Omega\times\mathbf{v} + F \ , \text{ – momentum}\\
\frac{\partial\rho}{\partial t} &= \nabla\cdot(\rho\mathbf{v}) \ , \text{ – mass}\\
\frac{DT}{Dt} &= \frac{1}{\rho c_p}\frac{Dp}{Dt} + \frac{Q}{c_p} \ , \text{ – energy (2nd LoT)}\\
\frac{\partial\rho q}{\partial t} &= -\nabla\cdot(\rho q\mathbf{v}) + \rho(E-C) \ , \text{ – water vapour}\\
p &= \rho RT \ , \text{ – ideal gas law}
\end{aligned}$$

So $\mathbf{X} = \{\mathbf{v}, T, p, \rho, q\}$ typically, and we discretise PDE to:

an ODE, $d\mathbf{X} = \mathcal{M}(\mathbf{X})dt$, and f represents the (integral) operator that maps the state at time t to time $t+1$.

Aston University
Birmingham

Different characters of ’multivariateness’

I think there are three main cases for the state vector, $\mathbf{X}$:

1. traditional multivariate – $\mathbf{X}$, is composed of different quantities, e.g. Lorenz 3D system;
2. spatio-temporal multivariate – $\mathbf{X} = \mathbf{X}(\mathbf{s}, t)$, a function of space and time, which is typically discretised, e.g. Kuramoto-Shivashinsky system;
3. full multivariate – $\mathbf{X}$ covers both of the above, e.g. primitive equations.

With 1, we need a joint specification, which is not trivial to parametrise, with 2 we can parametrise, for example assuming stationarity and separability, 3 needs a bit of both.

- I’ll start by looking at 3, in the context of dynamic models.

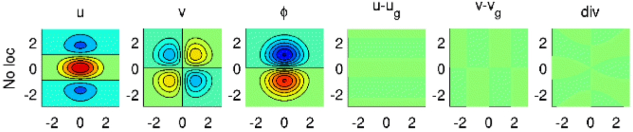
Multiple variables in data assimilation - balance

A (simplification and) scale analysis at a fixed time gives:

$$u \approx -\frac{\partial \Phi}{\partial y} \text{ and } v \approx \frac{\partial \Phi}{\partial x}$$

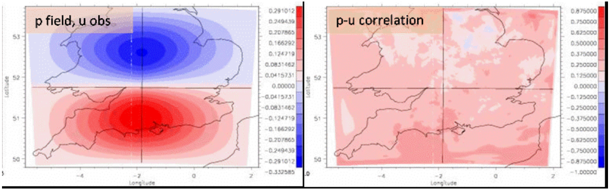
Using this geostrophic balance we can develop consistent multivariate covariances for u, v, Φ e.g.:

$$\begin{aligned} C_{uv}((x_1, y_1)(x_2, y_2)) &= E[u_1 . v_2] = -E\left[\left(\frac{\partial \Phi_1}{\partial y}\right) . \left(\frac{\partial \Phi_2}{\partial x}\right)\right] \\ &= \frac{\partial^2}{\partial y \partial x} E[\Phi_1 . \Phi_2] = \frac{\partial^2}{\partial y \partial x} C_{\Phi\Phi}((x_1, y_1)(x_2, y_2)) \end{aligned}$$



based on a U observation in the centre of domain – from J. D. Kepert

Problems using balances for covariances



from Ross Bannister

There are many problems with using such balances:

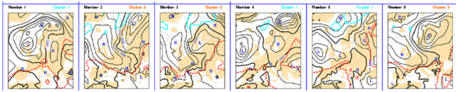
- They are often rather crude approximations.
- They really only operate in static settings; if you want space-time correlations there are very few analytic formulations.
- One must still posit a model for e.g. $C_{\Phi\Phi}((x_1, y_1)(x_2, y_2))$ – this is typically done on the basis of variogram fitting to historical data (the 'NMC method'[1]).

[1] This works on the innovations – the difference between the forecast and reality.

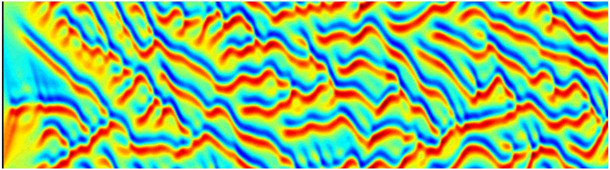
Alternatives - the Ensemble methods

- Many areas have a definition of ensemble: in the physical sciences this means 'a small number of'!
- Simplistically, if I gave you a function $f(\mathbf{X})$, and asked for $\text{Cov}[f(\mathbf{X}), f(\mathbf{X})] = \text{E}[(f(\mathbf{X}) - \mu)(f(\mathbf{X}) - \mu)^{T}], \mu = \text{E}[f(\mathbf{X})]$, evaluated at $\mathbf{X} = \mathbf{X}_t$ and told you nothing else about $f(\mathbf{X})$ ...

  ... you might sample from $\text{p}(\mathbf{X}_t)$ and propagate this through $f(\mathbf{X})$, using the samples to compute the moments.
- All operational ensemble systems use this Monte Carlo motivation, but the members are not typically sampled randomly from $\text{p}(\mathbf{X}_t)$, and typically the number $n < 100$.
- A more principled alternative is the unscented transform, which samples deterministically based on the current estimate of the covariance of $\text{p}(\mathbf{X}_t)$.



The Kuramoto-Shivashinsky system



Consider the univariate system given in differential form:

$$\frac{\partial \mathbf{X}}{\partial t} = -\frac{\partial^2 \mathbf{X}}{\partial s^2} - \frac{\partial^4 \mathbf{X}}{\partial s^4} - 0.5 \left(\frac{\partial \mathbf{X}}{\partial s} \right)^2 .$$

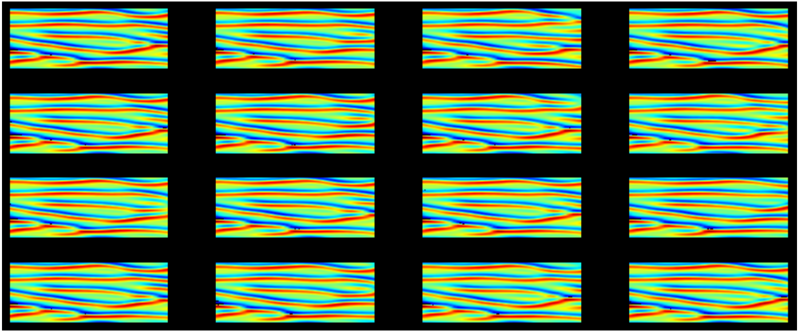
where as before t is time and s is the single spatial dimension.

- This is a PDE, so the solution is over a function space in (s, t) and the solutions are like 'waves', but not readily predictable.
- In practice the system cannot be solved in function space, and is discretised (often in a spectral domain) to produce a set of m coupled ODEs.

How to compute the covariance of $\mathbf{X}(s)$ or $\mathbf{X}(s, t)$?

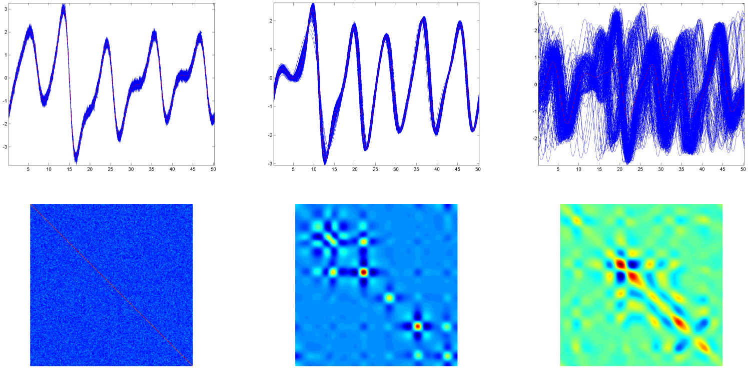
The Kuramoto-Shivashinsky system

The below shows a series of 16 ensemble members from a KS simulation where the initial $p(\mathbf{X}) = N(\mu, \sigma_0^2 I)$.



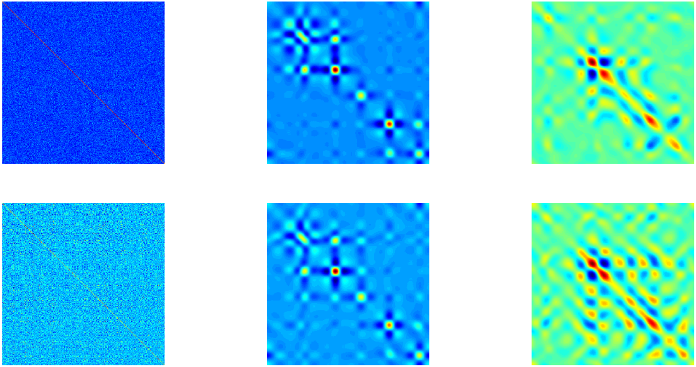
The initial noise being independent is not terribly realistic, but the KS system soon imposes it's dynamics.

The Kuramoto-Shivashinsky system



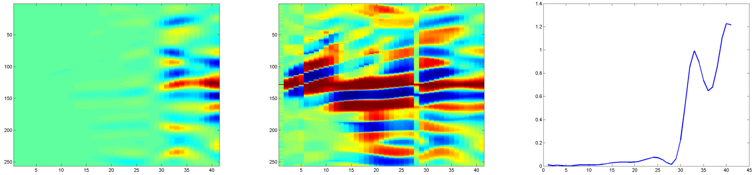
Using 256 ensemble members, it is possible to get good estimates of the mean and covariance at times 0, 10 and 40.

The Kuramoto-Shivashinsky system

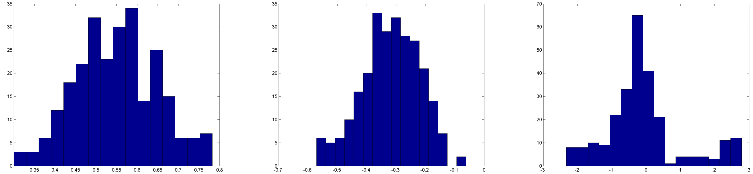


Using 16 ensemble members, finite sample sizes affect the quality of the estimates of the mean and covariance (shown at times 0, 10 and 40).

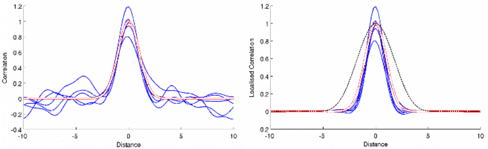
The Kuramoto-Shivashinsky system



We can also explore how the spatial covariance between a single point (this time in the middle of the domain) evolves in time - but beware things are not Gaussian at all times:

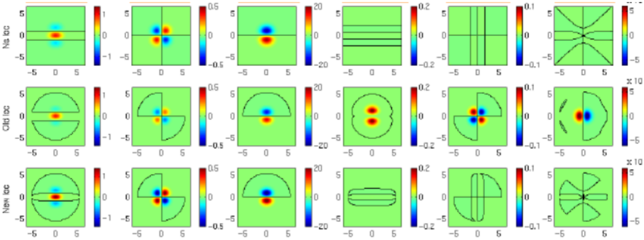


Balance and ensemble methods



from J. D. Keppert

One way to improve covariance estimation when using ensemble methods is to use localisation (this reduces impact of noise and rank deficiency, and is widely used in practice).

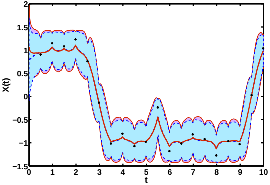


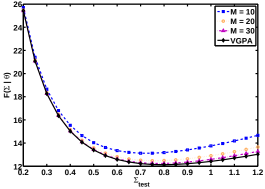
localisation can also exploit balance, if the localising functions obey the balance constraints – from J. D. Keppert

Recap

- Balance constraints can get us so far – but these are static, approximate, and parameters need to be estimated in the underlying covariances!
- Ensemble methods can be used to get time varying, state dependant covariances, and using localisation do a reasonable job.
- In practice ensemble methods are increasingly dominating in the geosciences.
- The alternatives to ensemble methods are the variational approaches, but the existing ones simply seek a MAP solution to the smoothing problem of estimating $\mathrm{p}(\mathbf{X}_{t_0:t_k} \mid \mathbf{Y}_{t_0:t_k})$.
- Next I'll describe briefly our variational approach ...

Approximate inference in diffusion processes





- The time evolution of a diffusion process can be described by a stochastic differential equation, henceforth SDE, (interpreted in the Itō sense):

$$d\mathbf{X}(t) = \mathbf{f}_{\boldsymbol{\theta}}(t, \mathbf{X}(t))dt + \mathbf{\Sigma}^{1/2} d\mathbf{W}(t), \qquad (1)$$

  where $\mathbf{f}_{\boldsymbol{\theta}}(t, \mathbf{X}(t)) \in \Re^D$ is the (usually non-linear) drift function, $\mathbf{\Sigma} = \text{diag}\{\boldsymbol{\sigma}_1^2, \ldots, \boldsymbol{\sigma}_D^2\}$ is the system noise.
- Most geoscience models have this structure with a model error term – the diffusion process is probably too simple an approximation in most cases, but is a start!

Approximate inference in diffusion processes

- In our work, the key idea is to approximate the true (latent) posterior process, $p(\mathbf{X}(t))$ by another one that belongs to a family of tractable ones (e.g. Gaussian processes), $q(\mathbf{X}(t))$.
- We do so by minimizing the $KL[q_t \| p_t]$ divergence, between the approximating process and the true one.
- The Gaussian process assumption implies a linear SDE:

$$d\mathbf{X}(t) = -\mathbf{A}(t)\mathbf{X}(t) + \mathbf{b}(t)dt + \mathbf{\Sigma}^{1/2}d\mathbf{W}(t)$$

where $\mathbf{A}(t) \in \Re^{D \times D}$ and $\mathbf{b}(t) \in \Re^{D}$ define the linear drift.

These time varying functions, $\mathbf{A}(t)$ and $\mathbf{b}(t)$, are the control parameters in the optimisation which minimises the KL.

Approximate inference in diffusion processes

- The time evolution of this Gaussian process can be expressed by a set of ordinary differential equations:

$$\begin{aligned}\dot{\mathbf{m}}(t) &= -\mathbf{A}(t)\mathbf{m}(t) + \mathbf{b}(t)\\ \dot{\mathbf{S}}(t) &= -\mathbf{A}(t)\mathbf{S}(t) - \mathbf{S}(t)\mathbf{A}(t)^\top + \boldsymbol{\Sigma}\end{aligned}$$

- To enforce these constraints the following $\mathcal{L}$agrangian is formulated:

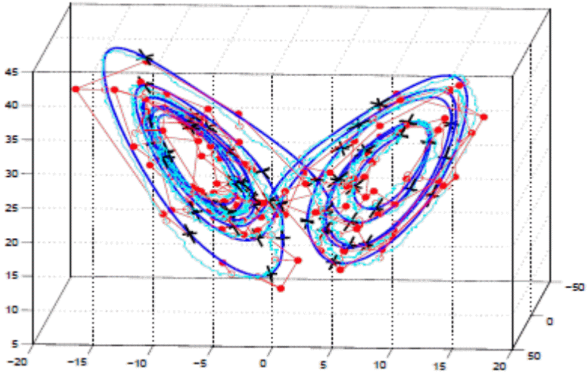
$$\begin{aligned}\mathcal{L} &= \int_{t_0}^{t_f} \Big\{\mathcal{E}(t) - \mathrm{tr}\{\boldsymbol{\Psi}(t)(\dot{\mathbf{S}}(t) + \mathbf{A}(t)\mathbf{S}(t) + \mathbf{S}(t)\mathbf{A}(t)^\top - \boldsymbol{\Sigma})\}\\ &\quad -\boldsymbol{\lambda}(t)^\top(\dot{\mathbf{m}}(t) + \mathbf{A}(t)\mathbf{m}(t) - \mathbf{b}(t))\Big\}dt\end{aligned}$$

where $\mathcal{E}(t) \in \Re$ is the energy term, $\boldsymbol{\lambda}(t) \in \Re^D$ and $\boldsymbol{\Psi}(t) \in \Re^{D\times D}$ are time dependent Lagrange multipliers.

Application to the Lorenz system

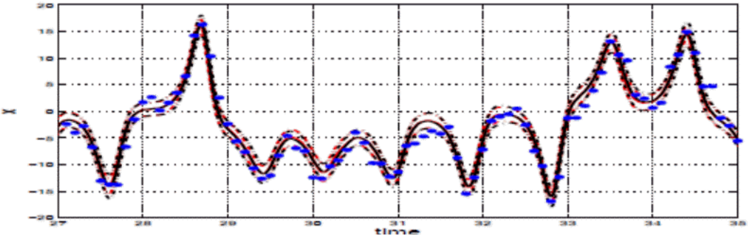
The Lorenz 3D system is given by:

$$\mathcal{M}_{\mathrm{L3D}}(\mathbf{X}_{\sqcup};\boldsymbol{\theta}) = \begin{pmatrix} \sigma(y_t - x_t) \\ \rho x_t - y_t - x_t z_t \\ x_t z_t - \beta z_t \end{pmatrix},$$



Application to the Lorenz system

The $\mathbf{A}(t)$ and $\mathbf{b}(t)$ generate the time varying GP structure, thus we 'learn' the best approximating joint covariance structure directly as part of the inference method.



- This approximation has some nice features: relative speed, no finite sample errors, really does smoothing, provides a bound on the marginal likelihood.
- But the are still problems: KL is 'the wrong way', doesn't scale to high D without further approximations, particularly in $\mathbf{A}(t)$.

Open questions

- Modelling multi-output systems is unsolved – still things to find there I'd guess!
- Should we ever contemplate more than central tendency and dispersion in large systems? Bayes Linear view?
- For really complicated systems how much can we really learn from data, or how much data might we need?
- I now see the real issue is in the discrepancy / model error - can we learn this?
- Real systems (which generate pretty much all observations) have dynamics, and these typically induce the structure – we should always try to exploit this where possible.
- I think this makes me a scientist ...

This work has been funded by EPSRC as part of the Variational Inference for Stochastic Dynamic Environmental Models (VISDEM) project (EP/C005848/1).