Estimating parameters in stochastic systems: A variational Bayesian approach.

Michail D. Vrettas*[,a], Dan Cornford[a], Manfred Opper[b]

[a]Aston University - Non-linearity and Complexity Research Group
Aston Triangle, Birmingham B4 7ET - United Kingdom
[b]Technical University Berlin - Artificial Intelligence Group
Franklinstrave, 28/29, D-10587, Berlin - Germany

Abstract

This work is concerned with approximate inference in dynamical systems, from a variational Bayesian perspective. When modelling real world dynamical systems, stochastic differential equations appear as a natural choice, mainly because of their ability to model the noise of the system by adding a variation of some stochastic process to the deterministic dynamics. Hence, inference in such processes has drawn much attention. Here a new extended framework is derived and present that is based on a local polynomial approximation of a recently proposed variational Bayesian algorithm. The paper begins by showing that the new extension of this variational algorithm can be used for state estimation (smoothing) and converges to the original algorithm. However, the main focus is on estimating the (hyper-) parameters of these systems (i.e. drift parameters and diffusion coefficients). The new approach is validated on a range of different systems which vary in dimensionality and non-linearity. These are the Ornstein-Uhlenbeck process, which its exact likelihood can be computed analytically, the univariate and highly non-linear, stochastic double well and the multivariate chaotic stochastic Lorenz '63 (3D model). As a special case the algorithm is also applied to the 40 dimensional stochastic Lorenz '96 system. In our investigation we compare this new approach with a variety of other well known methods, such as the hybrid Monte Carlo, dual unscented Kalman filter, full weak-constraint 4D-Var algorithm and analyse empirically their asymptotic behaviour as a function of observation density or length of time window increases. In particular we show we are able to estimate parameters in both the drift (deterministic) and diffusion (stochastic) part of the model evolution equations using our new methods.

Key words: Bayesian inference, variational techniques, dynamical systems, stochastic differential equations, parameter estimation

*Corresponding author
Email addresses: vrettasm@aston.ac.uk (Michail D. Vrettas), d.cornford@aston.ac.uk (Dan Cornford),

1. Introduction

Stochastic differential equations (SDEs) (Kloeden and Platen [35]) are a powerful tool in the modelling of real-world dynamical systems (Honerkamp [27]). Most phenomena observed in nature are time dependent and a common characteristic is that they consist of many sub-systems which, quite frequently, have different time scales. Hence in the description of the dynamics of the slow components of a system, the very fast ones can often be treated as *noise*. One strength of SDEs lies in their ability to model these very fast sub-systems as a stochastic process and incorporate a deterministic drift, which usually includes all the available knowledge of the system via physical laws, to formulate a model that best describes the observed system. However, such dynamical systems are usually formed by a very large number of unknown variables (or degrees of freedom) and are only partially/sparsely observed at discrete times, which makes statistical inference necessary if one wants to estimate the complete state vector at arbitrary times.

1.1. Bayesian treatment of SDEs

Inference for such systems is challenging because the *missing paths* between observed values must also be estimated, together with any unknown parameters. A variety of different approaches has been developed to undertake inference in SDEs; for a review see Sorensen [54]. This paper focuses largely on Bayesian approaches which from a methodological point of view can be grouped into three main categories.

The first category attempts to solve the *Kushner-Stratonovich-Pardoux* (*KSP*) equations (Kushner [39]). The *KSP* method, described briefly in Eyink et al. [19], can be applied to give the optimal (in terms of the variance minimising estimator) Bayesian posterior solution to the state inference problem, providing the exact conditional statistics (often expressed in terms of the mean and covariance) given a set of observations and serves as a benchmark for other approximation methods. Initially, the optimal filtering problem was solved by Kushner and Stratonovich [55, 37, 39] and later the optimal smoothing setting was given by an adjoint (backward) algorithm due to Pardoux [49]. Unfortunately, the KSP method is computationally intractable when applied to high dimensional non-linear systems (Kushner [38], Miller et al. [44]), hence a number of approximations have been developed to deal with this issue.

For instance, when the problem is linear the filtering part of the KSP equations (i.e. the forward Kolmogorov equations) boil down to the Kalman-Bucy filter [31], which is the continuous time version of the well known Kalman filter [30]. When dealing with systems that exhibit non-linear behaviour a variety of approximations, based on the original Kalman filter (KF), have been proposed. The first approach is to

`opperm@cs.tu-berlin.de` (Manfred Opper) linearise the model (usually up to first order) around the current state estimate, which through a Taylor expansion, requires the derivation of the Jacobian of the model evolution equations. However, this Jacobian might not always be easy to compute. Moreover the model should be smooth enough in the time-scales of interest, otherwise linearisation errors will grow causing the filter estimates to diverge. This method is known as the extended Kalman filter (EKF) (Maybeck [42]) and was succeeded by a family of methods based on statistical linearisation exploiting the observation that it is sometimes easier to approximate a probability distribution than a non-linear operator.

A widely used method that has produced a large body of literature is the ensemble Kalman filter (EnKF) (Evensen [17]), or when dealing with the smoothing problem the ensemble Kalman smoother (EnKS) (Evensen and van Leeuwen [18]). In DelSole and Yang [8] an ensemble Kalman filter (EnKF) is developed for stochastic dynamical systems and the paper includes an interesting discussion of the issues of parameter estimation in such system which is discussed further later. Recently another sampling strategy has been proposed. Rather than sampling this ensemble of particles randomly from the initial distribution it is preferable to select a *design* (i.e. deterministically chose them), so as to capture specific information (usually the first two moments), about the distribution of interest. A widely used example is the *unscented transform* and the filtering method is thus referred to as the unscented Kalman filter (UnKF), first introduced by Julier et al. [29]. Another popular, non-parametric, approach is the particle filter (Kitagawa [33]), in which the solution of the posterior density (or KSP equations) is approximated by a discrete set of particles with random support [34, 20]. This method can be seen as a generalisation of the ensemble Kalman filter, because it does not make the linear Gaussian assumption when the ensemble is updated in the light of the observations. In other words, if the dynamics of the system are linear then both filters should give the same answer, given a sufficiently large number of particles / ensemble members.

The second category applies Monte Carlo methods to sample from the desired posterior process, focusing on areas (in the state space) of high probability, based on Markov chains (Neal [47]). When the dynamics of the system is deterministic, then the sampling problem is on the space of initial conditions. In contrast, when the dynamics is stochastic the sampling problem is on the space of (infinite dimensional) sample paths. Therefore Markov chain Monte Carlo (MCMC) methods for diffusions are also known as "*path-sampling*" techniques. Although early sampling techniques such as the Gibbs sampler Geman and Geman [21] can be applied to systems, convergence is often very slow due to poor mixing in the Markov chains. In order to achieve better mixing of the chain and faster convergence other more complex and sophisticated techniques were developed. Stuart et al. [56], introduced the *Langevin MCMC* method, which essentially generalises the Langevin equation to sampling in infinite dimensions. A similar approach is the *hybrid Monte Carlo* (HMC) method (see Duane et al. [13]) which was later generalized for path sampling problems by Alexander et al. [1]. Both algorithms need information on the gradient of the target log-posterior distribution and update the entire trajectory (sample path) at each iteration. They combine ideas of molecular dynamics, employing the Hamiltonian of the system (including a kinetic energy term), to produce new configurations which are then accepted or rejected in a probabilistic way using the Metropolis criterion. Further details of this method are given in Section 4.2.

Following the work of Pedersen [50], on *simulated maximum likelihood estimation* (SMLE), Durham and Gallant [14], examine a variety of numerical techniques to refine the performance of the SMLE method by introducing the notion of the *Brownian bridge*, between two consecutive observations, instead of the Euler discretisation scheme that was used in [50]. This lead to various "blocking strategies", for sampling the sub-paths, such as the one proposed by Golightly and Wilkinson [22], as an extension to the previous "modified bridge" [14]. The work of Elerian et al. [15], Eraker [16] and Roberts and Stramer [52] is essentially based on a similar direction, that is augmenting the state with additional data between the measured values, in order to form a complete data likelihood and then facilitate the use of a Gibbs sampler or other sampling techniques (e.g. MCMC). A rather different sampling approach is presented by Beskos et al. [7], where an "*exact sampling*" algorithm (in the sense that there are no discretisation errors), is developed that does not depend on data imputation between the observable values, but rather on a technique called *retrospective sampling* (see Papaspiliopoulos and Roberts [48] for further details). Although this method is very appealing and computationally efficient compared to other sampling methods that depend on fine temporal discretisation to achieve sufficient accuracy, the applicability of the method depends heavily on the *exact algorithm*, as introduced by Beskos et al. [6].

Another alternative methodology, considered in this paper, approximates the posterior process using variational techniques (Jaakkola [28]). A popular treatment, which is operational at the *European Centre for Medium-Range Weather Forecasts* (ECMWF), is the four dimensional variational data assimilation method, also known as "*4D-Var*" (Dimet and Talagrand [12]). This method seeks the most probable trajectory (or the mode), of the approximate posterior smoothing distribution, within a predefined time window. This is found by minimizing a cost function which depends on the measured values and the model dynamics. However, this method does not provide uncertainty estimates around the most probable solution. The "*4D-Var*" method, as adopted by the ECMWF and others, makes the strong assumption that the model is either perfectly known, or that any uncertainties are negligible and hence can be ignored. A generalization of this strong *perfect model* assumption, is to accept that the model is not perfect and should be treated as an approximate solution the real equations governing the system. This leads to a *weak formulation* of *4D-Var* [11, 63]. The theory behind the *weak formulation* was introduced in early 70's by Sasaki [53] - several versions are described in Tremolet [57] and will be discussed later (see also Appendix B).

Another variational technique that seeks the conditional mean and variance of the posterior smoothing distribution is described in [19]. Eyink et al. [19] advocates that the ultimate goal of a data assimilation method is to recover not a specific history that generated the observations, but rather the correct posterior distribution, conditioned upon the observations. To achieve that Eyink et al. [19] apply a *mean field* approximation to the KSP equations. More recently the work of Archambeau et al. [4], suggested a rather different approach, where the true posterior process is approximated by a time-varying linear dynamical system inducing a non-stationary Gaussian process, rather than assuming a fully factorising form to the joint posterior. This linear dynamic approximation assumption implies a fine time discretisation if good accuracy is to be achieved, and globally optimises the approximate posterior process in terms of minimizing the Kullback-Leibler divergence (Kullback and Leibler [36]), between the two probability measures. This method is further reviewed in Section 2.2.

1.2. Motivation & Aim

This paper extends Vrettas et al. [59] and is motivated by inference of the state and (hyper-) parameters in models of real dynamical systems, such as the atmosphere (Kalnay [32]), where only a handful of Kalman filter approaches have been applied to *joint* state and parameter inference (Annan et al. [2]). In this work we develop a local polynomial approximation to extend the variational treatment proposed in Archambeau et al. [5]. The argument behind the use of the polynomial approximation in the variational algorithm is to control the number of free parameters that need to be optimized within the algorithm and constrain the space of functions admitted as solutions, in order to increase the robustness of the original algorithm with respect to different initialisations. In addition, this re-parametrisation of the original variational algorithm helps to improve the accuracy of the estimates, by allowing the application of higher order (and accuracy) integration schemes, such as Runge-Kutta 2nd order methods, when solving the resulting system of ordinary differential equations. The aim of this paper is three-fold: (a) to introduce this new *local polynomial based* extension, (b) to provide evidence that it converges to the original variational Gaussian process approximation algorithm, with less demand on computational resources (e.g. computer memory) and (c) to present an empirical comparison of the proposed extension, estimating both state and system parameters. The comparison is performed by applying well known methods that cover all the aforementioned categories dealing with the Bayesian inference problem for SDEs to a range of increasingly complex systems.

1.3. Paper outline

The paper is structured as follows. Section 2 briefly reviews the variational Gaussian process based algorithm (Archambeau et al. [5]), hereafter VGPA. Only the basic, but essential, information is given so that the reader can understand the rest of the paper. Section 3 presents the new polynomial approximation. Details of the approximation framework are explained thoroughly and mathematical expressions for the general multivariate case are provided. After the experimental set-up is illustrated, the stability and convergence of the new extensions are tested, on both univariate and multivariate systems, in Section 4. Section 5 empirically explores the asymptotic (infill and expanding domains) behaviour of the algorithm with increasing observation numbers, in comparison to other estimation methods. In addition the application of the new method to a system of 40 dimensions (stochastic Lorenz '96) is demonstrated which shows that the proposed method can attain good estimates of the system parameters in this smoothing framework on reasonably high dimensional systems. Conclusions are given in Section 6, with a discussion of the shortcomings and possible future directions.

2. Approximate Bayesian inference

This section reviews the VGPA algorithm first introduced in [4]. This algorithm, for approximate inference in diffusions, was initially proposed for state estimation (smoothing) and later was extended to include also estimation of (hyper-) parameters [5]. In this paper the VGPA provides the *backbone* on which the new extensions are built. Before proceeding to the basic setting a very short overview of partially observed diffusions will be given. This is necessary to provide a precise description of the approach adopted to the treatment of dynamical systems.

2.1. Markov processes and diffusions

A stochastic process can be seen as a collection of random variables indexed by a set, which here is regarded as time (i.e. $\boldsymbol{X} = \{\mathbf{X}_t, t \geq 0\}$). An informal and short introduction to stochastic processes can be found in [43]. A *Markov process* is a stochastic process in which if one wants to make a prediction about the state of the system at a future time 't_{n+1}', the only information necessary is the state of the system at the present time 't_n'. Any knowledge about the past is redundant. This is also known as the *Markov property.*

Diffusion processes are a special class of continuous time Markov processes with continuous sample paths (Kloeden and Platen [35]). The time evolution of a general, D dimensional, diffusion process $\boldsymbol{X} = \{\mathbf{X}_t\}_{t=t_0}^{t_f}$ can be described by a stochastic differential equation (here to be interpreted in the Itō sense):

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

where $\mathbf{X}_t \in \Re^D$ is the D dimensional latent state vector, $\mathbf{f}(t, \mathbf{X}_t; \boldsymbol{\theta}) \in \Re^D$ is the (usually) non-linear drift function, that models the deterministic part of the system, $\boldsymbol{\Sigma}(t, \mathbf{X}_t; \boldsymbol{\theta}) \in \Re^{D\times D}$ is the diffusion or system noise covariance matrix and $d\mathbf{W}_t$ is the differential of a D dimensional Wiener process, $\boldsymbol{W} = \{\mathbf{W}_t, t_0 \leq t \leq t_f\}$, which often models the effect of faster dynamical modes not explicitly represented in the drift function but present in the real system. $\boldsymbol{\theta} \in \Re^m$ is a set of (hyper-) parameters within the drift and diffusion functions.

Often the latent process $\mathbf{X}$ is only partially observed, at a finite set of discrete times $\{t_k\}_{k=1}^K$, subject to error. Hence

$$\mathbf{Y}_k = h_k(\boldsymbol{X}_{t_k}) + \boldsymbol{\epsilon}_k\ , \tag{2}$$

where $\mathbf{Y}_k \in \Re^d$ denotes the k'th observation taken at time t_k, $h_k(\cdot) : \Re^D \to \Re^d$ is the general observation operator and the observation noise $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}) \in \Re^d$, is assumed to be i.i.d. Gaussian white, with covariance matrix $\mathbf{R} \in \Re^{d\times d}$, although this can be generalised. In what follows, the general notation $\mathcal{N}(\mu, \Sigma)$ will denote the normal distribution with mean μ and (co)variance Σ.

2.2. Variational Gaussian approximation of the posterior measure

Equation (1) defines a stochastic system with multiplicative noise (i.e. state dependent). The VGPA framework considers diffusion processes with additive system noise [4, 7], although re-parametrisation makes it possible to map a class of multiplicative noise models into this additive class, as stated in Kloeden and Platen [35]. Consider the following SDE:

$$d\mathbf{X}_t = \mathbf{f}(t, \mathbf{X}_t; \boldsymbol{\theta})\ dt + \boldsymbol{\Sigma}^{1/2}\ d\mathbf{W}_t\ , \tag{3}$$

where for simplicity the covariance matrix $\boldsymbol{\Sigma}$ is assumed diagonal and all the assumptions about the dimensions of the drift and diffusion functions and the Wiener process remain the same as Eq.(1).

In addition, for notational convenience, it is further assumed that the discrete time measurements are "direct observations" of the state variables (i.e. $\boldsymbol{Y}_k = \boldsymbol{X}_{t_k} + \boldsymbol{\epsilon}_k$). This assumption simplifies the presentation of the algorithm and is the most common case in practice. Adding arbitrary observation operators to the equations only affects the system in the observation energy term in (5) and can be readily included if required. In this work the interest is on the *conditional posterior distribution* of the state variables given the observations, thus following the Bayesian paradigm one seeks the posterior measure given as follows:

$$p(\mathbf{X}_{t_0:t_f}|\mathbf{Y}_{1:K}) = \frac{1}{Z} \times \prod_{k=1}^{K} p(\mathbf{Y}_k|\mathbf{X}_{t_k}) \times p(\mathbf{X}_{t_0:t_f}) \ , \tag{4}$$

where K denotes the number of noisy observations, Z is the normalising marginal likelihood (i.e. $Z = p(\mathbf{Y}_{1:K})^1$), the posterior measure is over paths $\boldsymbol{X} = \{\mathbf{X}_t, t_0 \leq t \leq t_f\}$, the prior measure $p(\mathbf{X}_{t_0:t_f})$ is over paths defined by (3) and $p(\mathbf{Y}_k|\mathbf{X}_{t_k})$ is the likelihood for the observation at time t_k from (2).

The VGPA algorithm approximates the true posterior process by another that belongs to a family of tractable ones, in this case the Gaussian processes. This is achieved by minimising the "*variational free energy*", defined as follows (see also Appendix A):

$$\mathcal{F}(q(\mathbf{X}|\boldsymbol{\Sigma}), \boldsymbol{\theta}, \boldsymbol{\Sigma}) = -\left\langle \ln \frac{p(\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\Sigma})}{q(\mathbf{X}|\boldsymbol{\Sigma})} \right\rangle_{q(\mathbf{X}|\boldsymbol{\Sigma})} \ , \tag{5}$$

where p is the true posterior process, q is the approximate posterior process, $\langle . \rangle_{q(\mathbf{X}|\boldsymbol{\Sigma})}$ denotes the expectation with respect to $q(\mathbf{X}|\boldsymbol{\Sigma})$ and time indices have been omitted for simplicity.

The approximation of the true posterior process by a Gaussian process implies that q must be defined using a *linear* SDE. It follows that

$$d\mathbf{X}_t = \mathbf{g}_L(t, \mathbf{X}_t) \ dt + \boldsymbol{\Sigma}^{1/2} \ d\mathbf{W}_t \ , \tag{6}$$

where $\mathbf{g}_L(t, \mathbf{X}_t) = -\mathbf{A}_t\mathbf{X}_t + \mathbf{b}_t$, with $\mathbf{A}_t \in \Re^{D \times D}$ and $\mathbf{b}_t \in \Re^D$ define the linear drift in the approximating process. Both of these variational parameters, $\mathbf{A}_t$ and $\mathbf{b}_t$, are time dependent functions that need to be optimised as part of the estimation procedure. The time dependence of these parameters is a necessity due to the non-stationarity that is introduced in the process by the observations and system equations. Another point worth noting is the diffusion coefficient $\boldsymbol{\Sigma}$, which is chosen to be identical to that of the true process Eq.(3). This is a necessary condition because in the case where these two parameters are not identical then, as shown in [5], the bound on the negative log-marginal likelihood, given by Eq.(5), would not be finite.

[1] For notational brevity $p(\mathbf{Y}_{1:K})$ is shorthand notation for the joint density $p(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_K)$.

The time evolution of this general time varying linear system (6) is determined by two ordinary differential equations (ODEs), one for the marginal means $\mathbf{m}_t$ and one for the marginal covariances $\mathbf{S}_t$. These are given by the following equations (see also Kloeden and Platen [35], Ch. 4):

$$\dot{\mathbf{m}}_t = -\mathbf{A}_t \mathbf{m}_t + \mathbf{b}_t \ , \tag{7}$$

$$\dot{\mathbf{S}}_t = -\mathbf{A}_t \mathbf{S}_t - \mathbf{S}_t \mathbf{A}_t^\top + \boldsymbol{\Sigma} \ , \tag{8}$$

where $\dot{\mathbf{m}}_t$ and $\dot{\mathbf{S}}_t$ denote the time derivatives $\frac{d\mathbf{m}_t}{dt}$ and $\frac{d\mathbf{S}_t}{dt}$ accordingly. Thus we can write

$$q(\mathbf{X}_t) = \mathcal{N}(\mathbf{X}_t; \mathbf{m}_t, \mathbf{S}_t) \ , \tag{9}$$

where $\mathbf{m}_t \in \Re^D$ and $\mathbf{S}_t \in \Re^{D \times D}$.

Equations (7) and (8) are constraints to be satisfied ensuring consistency in the algorithm (Archambeau et al. [4, 5]). One way to enforce these constraints, within a predefined time window $[t_0, t_f]$, is to formulate the following $\mathcal{L}$agrangian functional:

$$\mathcal{L} = \mathcal{F}(q(\mathbf{X}_t), \boldsymbol{\theta}, \boldsymbol{\Sigma}) - \int_{t_0}^{t_f} \left(\boldsymbol{\lambda}_t^\top \underbrace{(\dot{\mathbf{m}}_t + \mathbf{A}_t \mathbf{m}_t - \mathbf{b}_t)}_{\text{ODE for the means}} + \mathrm{tr}\{\boldsymbol{\Psi}_t \underbrace{(\dot{\mathbf{S}}_t + \mathbf{A}_t \mathbf{S}_t + \mathbf{S}_t \mathbf{A}_t^\top - \boldsymbol{\Sigma})}_{\text{ODE for the covariances}}\} \right) dt \ , \tag{10}$$

where $\boldsymbol{\lambda}_t \in \Re^D$, $\boldsymbol{\Psi}_t \in \Re^{D \times D}$ are time dependent Lagrange multipliers, with $\boldsymbol{\Psi}_t$ being symmetric matrix. Given a set of fixed parameters for the diffusion coefficient $\boldsymbol{\Sigma}$ and the drift $\boldsymbol{\theta}$, minimising this quantity (10) and hence the free energy (5), will lead to the optimal approximate posterior process.

3. Local polynomial approximation

This section proposes a new extension to the previously described VGPA algorithm [5], in terms of polynomial approximations. Connections with previous work, on the same subject, will be given first, followed by the general multi-dimensional case, which will be derived and explained in detail.

The linear drift $\mathbf{g}_L(t, \mathbf{X}_t)$ in Eq.(6) is defined in terms of $\mathbf{A}_t$ and $\mathbf{b}_t$. These functions are discretised with a small time discretisation step (e.g. $\delta t = 0.01$), resulting in set of discrete time variables that need to be inferred during the optimisation procedure. In Vrettas et al. [59], these time varying functions were approximated with basis function expansions that cover the whole time domain of inference (i.e. $T = [t_0, t_f]$).

This allowed a reduction in the total number of control variables in the optimisation step, as well as some prior control over the space of functions admitted as solutions. However, the $\mathbf{A}_t$ and $\mathbf{b}_t$ variational parameters are, by construction, *discontinuous* when observations occur. Thus a large number of basis functions was required to capture the *roughness* at observation times.

The solution proposed here is to define the approximation only between observation times such as, $[t_0, t_{k=1}], (t_{k=1}, t_{k=2}], \dots, (t_{k=K}, t_f]$. This way one approximating function can be defined on each sub-interval (without overlap), further reducing the total number of parameters to be optimised.

3.1. Re-parametrisation of the variational parameters

The variational parameters $\mathbf{A}_t$ and $\mathbf{b}_t$ in Archambeau et al. [5] are represented as a set of discrete time variables whose size scales proportionally to the length of the time window of inference, the dimensionality of the data (state vector $\mathbf{X}_t$) and the time discretisation step. In total we need to optimise

$$N_{total} = (D+1) \times D \times |t_f - t_0| \times \delta t^{-1} \ , \tag{11}$$

variables, where D is the system dimension, t_0 and t_f are the initial and final times and δt must be sufficiently small for numerical stability in the system being considered.

By replacing $\mathbf{A}_t$ and $\mathbf{b}_t$ with local polynomials on each sub-interval the following expressions are obtained:

$$\begin{aligned} \tilde{\boldsymbol{A}}_t^j &= \boldsymbol{A}_0^j + \boldsymbol{A}_1^j \times t + \cdots + \boldsymbol{A}_M^j \times t^M \ , \\ \tilde{\boldsymbol{b}}_t^j &= \boldsymbol{b}_0^j + \boldsymbol{b}_1^j \times t + \cdots + \boldsymbol{b}_M^j \times t^M \ , \end{aligned} \tag{12}$$

where $\tilde{\boldsymbol{A}}_t^j$ and $\tilde{\boldsymbol{b}}_t^j$ are the approximating functions defined on the j'th sub-interval, $\boldsymbol{A}_i^j \in \Re^{D \times D}$ and $\boldsymbol{b}_i^j \in \Re^D$ are the i'th order coefficients of the j'th polynomial and $i \in \{0, 1, \dots, M\}$, with M being the order of the polynomial.

It is important to distinguish from the case where the polynomials are fitted between the actual *measurable values* (e.g. cubic splines). Here the polynomials are rather inferred between *observation times*. Note also that the order of the polynomials between $\tilde{\boldsymbol{A}}_t^j$ and $\tilde{\boldsymbol{b}}_t^j$, or even between the j'th polynomial of each approximation, need not to be the same; however in the absence of any additional information about the functions, or lack of any theoretical guidance, an empirical approach is followed that suggest the same order of polynomials, under the condition that they provide enough flexibility to capture the *discontinuity* of the variational parameters at observation times, as shown in Figure 1.

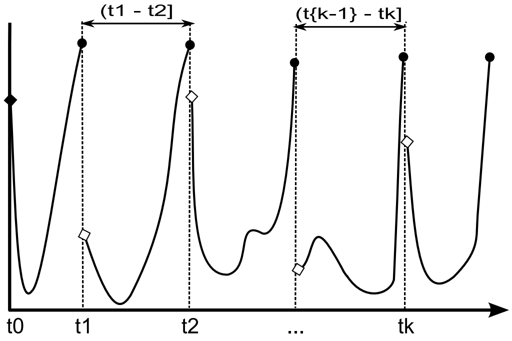


Figure 1: An example of the *local* polynomial approximation, on a univariate system. The vertical dashed lines represent the times the observations occur and each polynomial is defined *locally* between two observation times. The filled diamond and circles indicate closed sets, while the clear diamonds define open sets. Note that only the first polynomial is defined in a closed set from both sides, to avoid overlapping.

The expression for the (approximate) $\mathcal{L}$agrangian for the j'th sub-interval thus becomes:

$$\tilde{\mathcal{L}}^j = \tilde{\mathcal{F}}^j(q(\mathbf{X}_t),\boldsymbol{\theta},\boldsymbol{\Sigma}) - \int_{t\in T^j} \Big(\boldsymbol{\lambda}_t^\top(\dot{\mathbf{m}}_t + \tilde{\boldsymbol{A}}_t^j\mathbf{m}_t - \tilde{\boldsymbol{b}}_t^j) + \mathrm{tr}\{\boldsymbol{\Psi}_t(\dot{\mathbf{S}}_t + \tilde{\boldsymbol{A}}_t^j\mathbf{S}_t + \mathbf{S}_t\tilde{\boldsymbol{A}}_t^{j\top} - \boldsymbol{\Sigma})\}\Big)\,dt\,, \tag{13}$$

where $T^j \subset T$, or $T = \{T^1 \cup \cdots \cup T^j \cup \cdots \cup T^J\}$, with $J \geq 1$, being the total number of disjoint sub-sets.

The expressions for the polynomial approximations, Eq. (12), can be presented more compactly using matrix notation. This simplified presentation is used from this point forward:

$$\begin{aligned}\tilde{\boldsymbol{A}}_t^j &= \boldsymbol{A}^j \times \boldsymbol{p}^j(t)\,,\\ \tilde{\boldsymbol{b}}_t^j &= \boldsymbol{b}^j \times \boldsymbol{p}^j(t)\,.\end{aligned} \tag{14}$$

Schematically these matrix - vector products can be seen as:

$$\tilde{\boldsymbol{A}}_t^j \xleftarrow{\text{reshape to}} \begin{pmatrix} A_1^j(t) \\ A_2^j(t) \\ \vdots \\ A_{D^2}^j(t) \end{pmatrix} = \underbrace{\begin{pmatrix} A_{1,0}^j & A_{1,1}^j & \cdots & A_{1,M}^j \\ A_{2,0}^j & A_{2,1}^j & \cdots & A_{2,M}^j \\ \vdots & \vdots & \ddots & \vdots \\ A_{D^2,0}^j & A_{D^2,1}^j & \cdots & A_{D^2,M}^j \end{pmatrix}}_{\boldsymbol{A}^j} \times \underbrace{\begin{pmatrix} 1 \\ t \\ \vdots \\ t^M \end{pmatrix}}_{\boldsymbol{p}^j(t)}\,.$$

Here $A^j_{r,i}$ represents the r'th (scalar) component of the $\boldsymbol{A}^j_i$ coefficient in the j'th sub-interval. Effectively, we have reshaped the $\boldsymbol{A}^j_i$ weights in column vectors and packed them all together in one matrix of size $D^2 \times (M+1)$. For the $\tilde{\boldsymbol{b}}^j_t$ a similar procedure is followed, which is simpler because the $\boldsymbol{b}^j_i$ coefficients are already vectors, so there is no need to reshape them. Hence we have:

$$\tilde{\boldsymbol{b}}^j_t \leftarrow \begin{pmatrix} b^j_1(t) \\ b^j_2(t) \\ \vdots \\ b^j_D(t) \end{pmatrix} = \underbrace{\begin{pmatrix} b^j_{1,0} & b^j_{1,1} & \cdots & b^j_{1,M} \\ b^j_{2,0} & b^j_{2,1} & \cdots & b^j_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ b^j_{D,0} & b^j_{D,1} & \cdots & b^j_{D,M} \end{pmatrix}}_{\boldsymbol{b}^j} \times \underbrace{\begin{pmatrix} 1 \\ t \\ \vdots \\ t^M \end{pmatrix}}_{\boldsymbol{p}^j(t)} ,$$

where $b^j_{r,i}$ represents the r'th component of the $\boldsymbol{b}^j_i$ coefficient.

Eq. (14) shows that the vectors $\boldsymbol{p}^j(t)$ can be precomputed off-line for all predefined discrete time domains, reducing the computational complexity of estimating the coefficients of the polynomials. $\boldsymbol{p}^j(t)$ is precomputed and stored column-wise in a matrix, as shown on Table 1. Thus the reconstruction of the approximate variational parameters $\tilde{\boldsymbol{A}}^j_t$ and $\tilde{\boldsymbol{b}}^j_t$, for their whole time domain, can be done by a simple matrix - matrix multiplication (e.g. $\tilde{\boldsymbol{A}}^j_t = \boldsymbol{A}^j \times \boldsymbol{\Pi}^j(t)$).

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ t_{k+\delta t} & t_{k+2\delta t} & t_{k+3\delta t} & \cdots & t_{k+1} \\ t^2_{k+\delta t} & t^2_{k+2\delta t} & t^2_{k+3\delta t} & \cdots & t^2_{k+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t^M_{k+\delta t} & t^M_{k+2\delta t} & t^M_{k+3\delta t} & \cdots & t^M_{k+1} \end{pmatrix}$$

Table 1: Example of $\boldsymbol{\Pi}^j(t)$ matrix, defined on $\boldsymbol{T}^j = (t_k, t_{k+1}]$.

The number of coefficients for both variational parameters $\tilde{\boldsymbol{A}}_t$ and $\tilde{\boldsymbol{b}}_t$ is:

$$L_{total} = (D+1) \times D \times (M+1) \times J \ , \tag{15}$$

variables, where D is the system dimension, M is the order of the polynomials and J is the total number of disjoint sub-intervals (i.e. the number of observation times increased by one). Usually, it is anticipated that $L_{total} \ll N_{total}$, thus making the optimisation problem smaller.

The original VGPA algorithm, uses a scaled conjugate gradient (SCG) algorithm (see Nabney [45]), to minimize Eq.(10) with respect to the variational parameters $\mathbf{A}_t$ and $\mathbf{b}_t$. The same procedure is used here computing the gradients of the approximate Lagrangian Eq.(13), with respect to the coefficients $\boldsymbol{A}^j$ and $\boldsymbol{b}^j$, of the re-parametrized variational parameters, for each sub-interval. To further improve computational efficiency and stability a modified Gram-Schmidt orthogonalisation is applied (Golub and van Loan [23]) to the rows of the pre-computed $\boldsymbol{\Pi}^j(t)$ matrices, as shown in Table 1, on each sub-interval separately. In practice this orthogonalisation dramatically reduces the number of iterations required for the algorithm to reach convergence.

4. Numerical simulations on artificial data

This section explores the convergence properties of the new Local Polynomial (hereafter LP) approximation algorithm comparing to the original VGPA framework. The new LP approach is validated on one linear and two non-linear dynamical systems. The experimental set up will be shown first, followed by results for the uni- and multi-variate systems.

4.1. Choice of systems & experimental design

The first system considered is the linear one dimensional Ornstein-Uhlenbeck process (OU). Originating from the physics literature it was proposed as a model for the velocity of a particle undergoing Brownian motion (Uhlenbeck and Ornstein [58]). Here it is understood as a continuous Markov process with dynamics that can be represented by the following SDE:

$$dX_t = -\theta X_t \; dt + \Sigma^{1/2} \; dW_t \; , \tag{16}$$

where $\theta > 0$ is the drift parameter, $\Sigma \in \Re$ is the diffusion coefficient[2] and $W_t \in \Re$ is the univariate Wiener process. In fact this system is one of very few on which exact inference can be performed. The prior process is Gaussian (linear), and given that the initial state is fixed ($X_0 = x0$), the (non-stationary) covariance function for the posterior process is given by:

$$Cov(X_t, X_s) = \frac{\Sigma}{2\theta}(\exp\{-\theta|t-s|\} - \exp\{-\theta(t+s)\}) \; , \tag{17}$$

which can then be used in a Gaussian process regression smoother to compute the exact posterior (Rasmussen and Williams [51]).

[2] To keep the notation consistent we use Σ instead of σ^2, and we represent the scalars with normal fonts while vectors and matrices are represented with bold fonts.

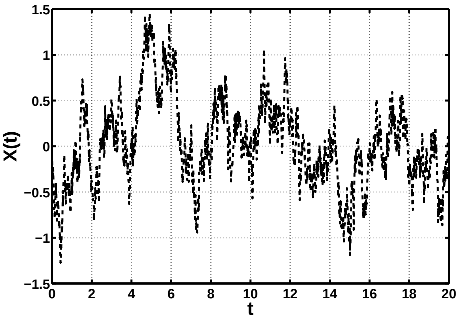
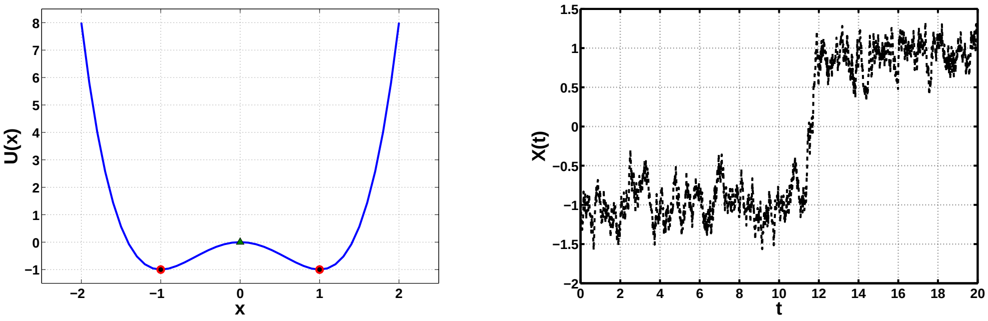


Figure 2: An example of the Ornstein-Uhlenbeck diffusion process that will be used later in the simulations.

Secondly, the non-linear double well model (DW), which is a stochastically forced scalar differential equation with three equilibrium values at $X_t = 0$ and $X_t = \pm\, \theta$ (Miller et al. [44]) is considered. As shown in Fig. 3(a) the position of a particle at 0 is unstable, while stable equilibria are found at $\pm\, \theta$ in the absence of noise. Mathematically, the potential is given by $U(x) = -2x^2 + x^4$. Notice that the drift function in Eq. (18), is simply the derivative: $-\frac{dU(x)}{dx} = 4x(1 - x^2)$, for $\theta = 1$. However, within our setting random forces occur and occasionally drive the particle from one basin to the other (see Fig. 3(b)). This effect is known as “transition” between the two stable states. The SDE that describes the dynamics of this system is the following:

$$dX_t = 4X_t(\theta - X_t^2)\; dt + \Sigma^{1/2}\; dW_t\;, \tag{18}$$

where $\theta > 0$, is the drift parameter which determines the stable points. Although a simple system, the double well has served as a benchmark in a number of references such as [19, 5].



(a) Double well potential

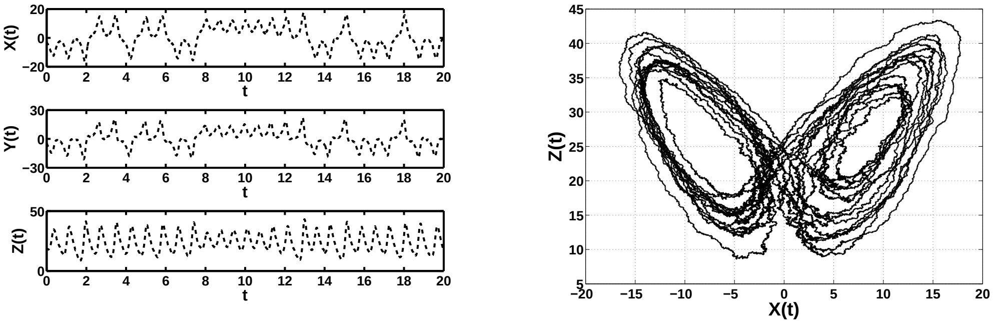
(b) Double well simulation

Figure 3: (a) The double well potential. The circles indicate the stable points (in this example ±1), in the absence of stochastic forcing, while the triangle denotes the unstable point. (b) An example of a DW sample path including a transition. This sample path will be used as the history in the experimental simulations.

The final system is a stochastic version of the three dimensional chaotic Lorenz '63 (L3D), driven by the following SDE:

$$d\mathbf{X}_t = \begin{bmatrix} \sigma(y_t - x_t) \\ \rho x_t - y_t - x_t z_t \\ x_t y_t - \beta z_t \end{bmatrix} dt + \mathbf{\Sigma}^{1/2}\, d\mathbf{W}_t \;, \tag{19}$$

where $\mathbf{X}_t = [x_t\ y_t\ z_t]^\top \in \Re^3$ is the state vector representing all three dimensions, $\boldsymbol{\theta} = [\sigma\ \rho\ \beta]^\top \in \Re^3$, is the drift parameter vector, $\mathbf{\Sigma} \in \Re^{3\times3}$ is a (diagonal) covariance matrix and $\mathbf{W}_t \in \Re^3$ is an uncorrelated multivariate Wiener process. The deterministic version of this model (i.e. without the noisy part of Eq. (19)) was first introduced by Lorenz [40] as a low dimensional analogue for large scale thermal convection in the atmosphere. This multi-dimensional non-linear system produces chaotic behaviour when its drift parameters σ, ρ and β lie within a specific range of values and is used in a large body of literature (see Evensen and van Leeuwen [18], Miller et al. [44] and Hansen and Penland [25]). The choice of the drift values, in this work, are those which produce chaotic behaviour (as shown in Table 2) and are the most commonly used values.



(a) Lorenz 3D simulation

(b) Lorenz 3D in X-Z plane

Figure 4: (a) A typical realisation of the stochastic Lorenz '63 system as time series in each dimension. (b) The same data but in X-Z plane where the effect of the random fluctuations is more clear.

Following a similar strategy to Apte et al. [3], the time discretisation is applied only in the posterior approximation; the inference problem is derived in an infinite dimensional framework (continuous time sample paths), as shown in Section 2.2. The Euler-Maruyama representation of the prior process (3), leads to the following discrete time analogue

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{f}(t, \mathbf{x}_k; \boldsymbol{\theta})\ \delta t + \sqrt{\mathbf{\Sigma}\delta t}\ \boldsymbol{\xi}_k \;, \tag{20}$$

where $\boldsymbol{\xi}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and the positive *infinitesimal* dt in Eq.(3), has now been replaced by a positive *finite*

$System$	t_0	t_f	δt	$\boldsymbol{\theta}$	$\boldsymbol{\Sigma}$	N_{obs}	$\boldsymbol{R}$
OU	0	20	0.01	2	1	2	0.04
DW	0	20	0.01	1	0.8	2	0.04
$L3D$	0	20	0.01	$[10, 28, 2.6667]$	6	10	2

Table 2: Experimental setup that generated the data (trajectories and observations). Initial times (t_0) and final times (t_f) define a fixed time window of inference, whilst δt is the time discretisation step. $\boldsymbol{\theta}$ are the parameters related to the drift function, while $\boldsymbol{\Sigma}$ and $\boldsymbol{R}$ represent the noise (co)variances of the stochastic process and the discrete observations accordingly. In the multivariate system these covariance matrices are diagonal. N_{obs} represents the number of available i.i.d. observations *per time unit* (i.e. observation density), which without loss of generality is measured at equidistant time instants.

number δt. In addition, this expression can be used to provide approximate sample paths (in terms of discretising a stochastic differential equation) from the prior process (Higham [26], Kloeden and Platen [35]). Under this first order approximation we impose a suitably small discretisation step δt to achieve good accuracy.

In the numerical experiments a fixed inference window of twenty time units (i.e. $T = [0, 20]$) was considered for all systems and the time discretisation was set to $\delta t = 0.01$ for numerical stability. For the $L3D$ the deterministic equations were integrated forwards in time for $T_{burn} = 5000$ units, in order to get the initial state vector $\mathbf{X}_0$ on the attractor and then generated the stochastic sample path (Figures 4(a) and 4(b)). Table 2 summarizes the *true* parameter values, that generated the sample paths for the simulations that follow. Note that 20 time units within these systems corresponds to a rather long assimilation window compared with operational systems.

4.2. State estimation results

The presentation of the experimental simulations begins with results for the OU process. Figure 5 shows the results from the LP approximation of the VGPA algorithm, of polynomial order $M = 5$. For this example the measurement density of 2 observations per time unit (hence 40 in the whole time domain $T = [0, 20]$), with $M = 5$ and $J = 41$, produces a set of $L_{total} = 492$ coefficients to be inferred, compared to $N_{total} = 4000$ in the original VGPA framework. This is roughly 12.3% of the size of the original VGPA optimisation problem. For this system, as mentioned earlier, one can use the induced non-stationary covariance kernel function Eq.(17) and compute the exact posterior process. Comparing the results obtained from the LP approximation with the results from a GP regression smoother with the OU kernel the match is excellent, as expected for a linear system, where the approximation is theoretically exact (in the limiting case as $\delta t \rightarrow 0$).

To provide a robust demonstration of the consistency of the results of the LP approximation, with respect to the original discretized VGPA, fifty different realisations of the observation noise, from a single trajectory, were used. The order of the polynomials was increased to explore convergence of the LP to the

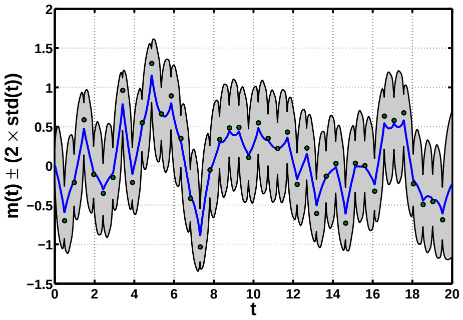
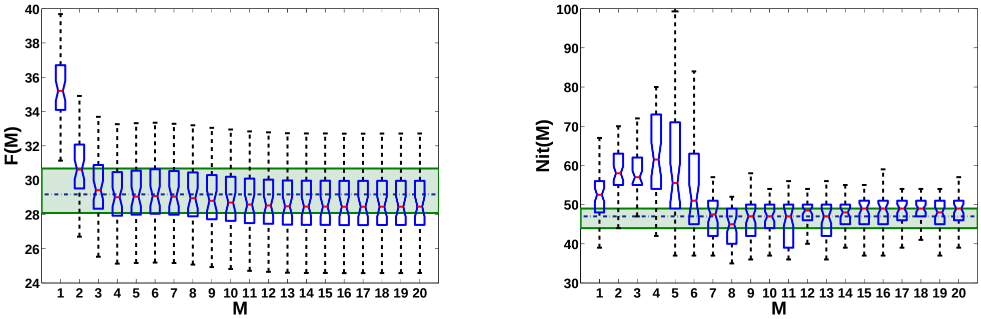


Figure 5: The marginal values of the means (solid blue line) and variances (shaded grey area) obtained by the LP approximation of 5'th order on a single realisation of the OU system. The results from the GP regression, on the same observation set, are visually indistinguishable and are omitted. The circles indicate noisy observations.

original VGPA. Summary statistics from these experiments, on the OU system, concerning the convergence of the free energy obtained from the LP approximation algorithm compared with the one from the original VGPA, are shown in Figure 6(a). Here the median is plotted along with the 25'th and 75'th percentiles in box-plots, while the extended vertical dashed lines indicate the 5'th and 95'th percentiles, from these 50 realisations, when the system has converged to its free energy minimum. For this example, with only second order polynomials (i.e. $M = 2$), the LP algorithm reaches the same free energy values as the original VGPA.



(a) Variational free energy

(b) Number of SCG iterations

Figure 6: (a) The median and the 25'th to 75'th percentiles as box-plots of the variational free energy, from fifty realisations of the observation noise, as a function of the increasing order of polynomials M, keeping the drift and diffusion parameters fixed to their true values. Extended vertical dashed lines indicate the 5'th and 95'th percentiles. The horizontal dashed (blue) line represents the 50'th percentile of the free energy obtained from the original VGPA on the same 50 realisations and the shaded area encloses the 25'th to 75'th percentiles. (b) The summaries from the same experiment concerning the number of iterations both algorithms needed to converge to optimality. Again, the horizontal lines (and shaded area) represent results obtained for the original VGPA, while boxplot results from the LP approximation, as in (a).

Figure 7(a) compares the results obtained from the LP approximation with 5'th order polynomials, on a single realisation of the DW system, to the outcomes of a Hybrid Monte Carlo (HMC) sample from the posterior process, using the true values for the drift and diffusion parameters. The HMC algorithm [13], combines Hamiltonian molecular dynamics with the Metropolis-Hastings accept/reject criterion to propose a new configuration (or a new sample path) of the posterior process (Eq. 4). The algorithm begins with an initial (discrete time) sample path $\mathbf{X}^j = \{x_k^j\}_{k=0}^N$, where $j > 0$ is the step in the iterative process and proposes a new sample path $\mathbf{X}^{j+1} = \{x_k^{j+1}\}_{k=0}^N$. This is done by simulating, forwards in time, a fictitious time deterministic system:

$$\frac{dx_k^j}{d\tau} = p_k \quad \text{and} \quad \frac{dp_k}{d\tau} = -\frac{\partial \mathcal{H}(x_k^j, p_k)}{\partial x_k^j} \ , \tag{21}$$

where $p_k \sim \mathcal{N}(0, 1)$ are the fictitious momentum variables assigned to each state variable x_k, resulting in a finite size random vector $\mathbf{p} = \{p_k\}_{k=0}^N$. These deterministic equations are discretised with a time step $\delta\tau$ and solved with a *leapfrog* integration scheme. The Hamiltonian of the system $\mathcal{H}(\mathbf{X}, \mathbf{p})$ is:

$$\mathcal{H}(\mathbf{X}, \mathbf{p}) = E_{pot} + E_{kin} \ , \tag{22}$$

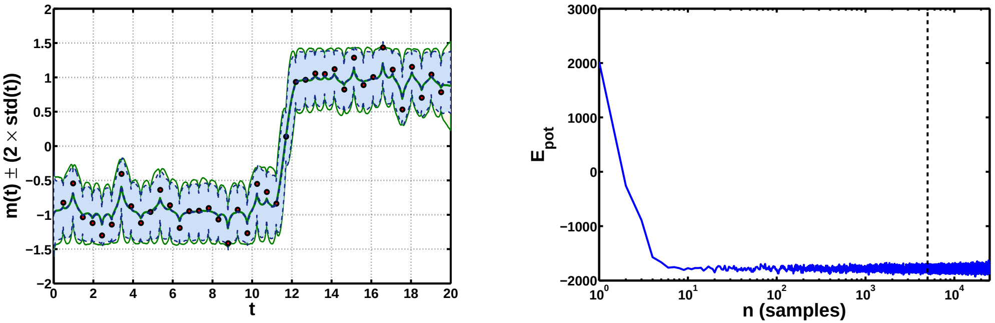
where $E_{pot} = -\ln p(\mathbf{X}_{t_0:t_f}|\mathbf{Y}_{1:K})$ is the potential energy associated with the dynamics of the system (SDE) as well as the observations Eq.(4) and $E_{kin} = \frac{1}{2}\mathbf{p}\mathbf{p}^\top$ is the kinetic energy.

The HMC solution is assumed to provide a *reference solution* to the smoothing problem. The setting for the DW example is $25,000$ iterations of which the first $5,000$ are considered as a *burn-in* period and discarded. Each HMC iteration generates 80 posterior sample paths (or configurations) of the system with artificial time $\delta\tau = 0.01$, of which only the last one is considered as candidate state. In total $2,000,000$ sample paths are generated from which only $20,000$ are sampled uniformly to compute the marginal mean and variance as shown in Figure 7(a). The convergence results of this simulation are shown in Figure 7(b). Even though there exist recently proposed MC sampling algorithms, such as the *generalised HMC* as suggest by Alexander et al. [1] to speed up the convergence of the Markov chain, here a rather classical hybrid Monte Carlo, as was first introduced by Duane et al. [13] is used.

Although the variance of the LP approximation is slightly underestimated, the mean path matches the HMC results and the time of the transition between the two wells is tracked accurately. The variational approximation as shown in Section 2.2 is likely to underestimate the variance of the approximating process, as is often the case when the expectation in the KL divergence is taken with respect to the approximating distribution[3] in Eq.(5). Empirically we have found this to have a relatively minor impact as long as the system is well observed, which keeps the posterior process close to Gaussian. Where the true posterior process

[3] That is $KL[q_t\|p_t]$ instead of computing $KL[p_t\|q_t]$, where p_t is the true posterior while q_t is the approximate one.

is strongly non-Gaussian, and in particular where it is multi-modal there is more significant underestimation, as might be expected.

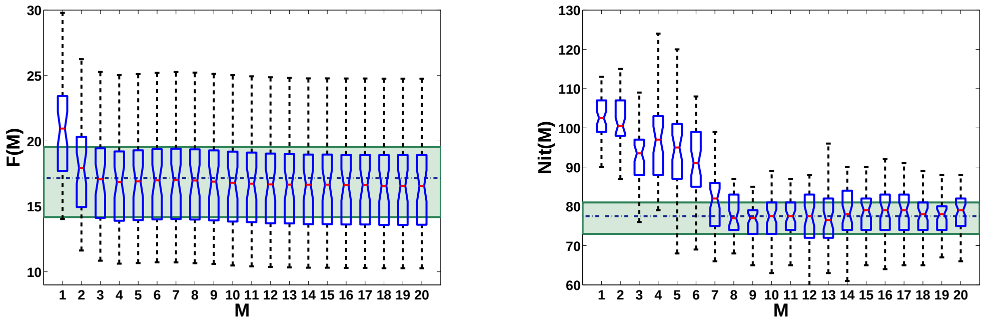


(a) HMC vs LP

(b) Potential energy trace

Figure 7: (a) Comparison of the approximate marginal mean and variance (of a single DW realisation), between the "correct" HMC posterior estimates (solid green lines and light shaded area) and the LP approximation, of 5'th order, (dashed blue lines and dark shaded area). The circles indicate noisy observations. (b) Trace of the potential energy (-x- axis is in log-space), of the Hamiltonian, in the HMC posterior sampling. The vertical dashed line, indicates the end of the burn in period and the beginning of the posterior sampling.

Figures 8(a) and 8(b), present results comparable to Figures6(a) and 6(b), but for the DW system. Again 50 different realisations of the observation noise, from a single trajectory, were generated and both LP approximation and VGPA algorithms were applied, given the true parameter values for the drift and diffusion coefficients. The summaries from these runs show the consistency of the LP approximation, when applied to non-linear systems. The algorithm exhibits stability and slightly outperforms the original VGPA framework, in terms of minimizing the free energy, although this has a very minor impact in terms of solving the ODEs (Eq. 7, 8) to produce the marginal means and variances as shown in Figure 7(a).

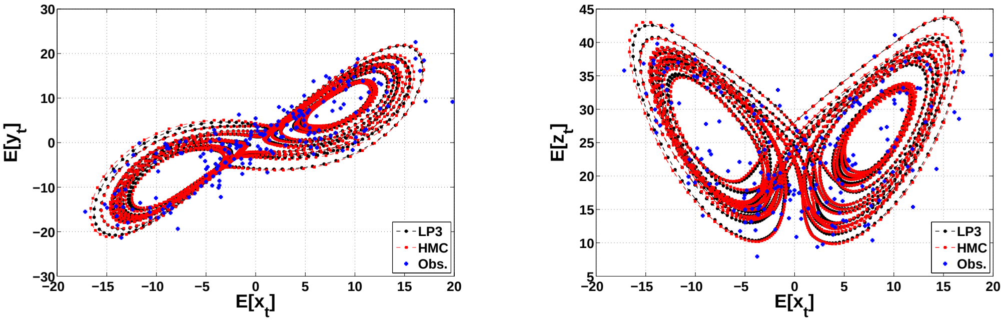


(a) Variational free energy

(b) Number of SCG iterations

Figure 8: (a) Similar to Fig.6(a), but from fifty different realizations of the observation noise of the DW system. (b) Again, similar to Fig.6(b), but for the DW system.

To provide a more complete assessment of how this new LP approximation approach to the VGPA algorithm scales with higher dimensions the same experiments were repeated on a multivariate system, namely the Lorenz '63 (L3D). Figures 9(a) and 9(b), show the approximated mean paths obtained with a 3'rd order LP algorithm, against the posterior mean paths computed using HMC, in XY and XZ planes respectively, from a single realisation of the stochastic L3D shown in Figure 4(a). The observation density for this example was relatively high ($N_{obs} = 10$, per time unit), hence it was possible to set the order of the polynomials to $M = 3$. In this example, unlike the previous case of the DW, the LP approximation slightly overestimates the marginal variance (Figure 10(b)) compared with the estimates obtained by using HMC. However, the same effect is observed when applying also the original VGPA framework, hence this is not an artefact of the polynomial approximation but rather of the variational framework.

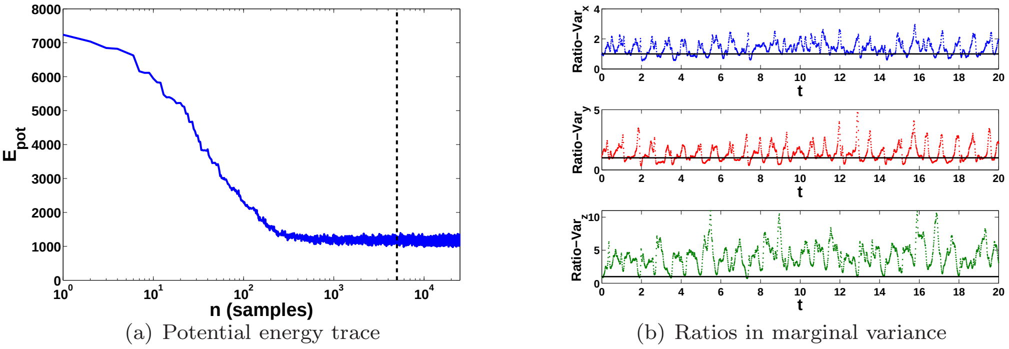


(a) XY plane

(b) XZ plane

Figure 9: The marginal means, obtained from the LP approximation and the HMC sampling in XY (a) and XZ (b) planes respectively, on a single realisation of the L3D (see Fig. 4(b)). In both plots, the dots (black) are the results from the LP approximation (of 3'rd order), while the squares (red) are results from HMC. Crosses (blue) indicate the noisy observations. The $E[\cdot]$ notation in the figures axis represents *expected* value.

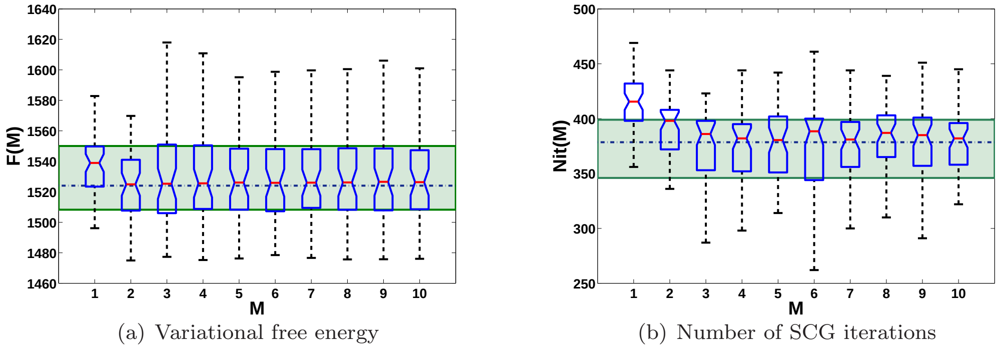
The tuning of the HMC sampling scheme was similar to the one used to obtain the posterior estimates for the DW system, only in this case a smaller artificial time step was necessary to correctly sample the posterior process. In total $25,000$ iterations of the HMC algorithm were used, with the first $5,000$ considered as *burn-in*. Each HMC iteration produced 50 new configurations of the system (posterior sample paths), where only the last one was proposed as a new configuration. The artificial time step was $\delta\tau = 0.004$. Sampling from high dimensional distributions, with the HMC, is not a trivial task. Sampling continuous time *sample paths*, which when discretised result in a large number of random variables that need to be jointly sampled at each iteration is challenging. For the L3D system considered here, we had to sample $N_{rv} = 6003$, random variables at each iteration. The trace of the potential energy of the Hamiltonian (for the L3D example), is presented in Fig. 10(a). Considerable effort was expended to ensure that the HMC sampler converged and gave a sufficiently uncorrelated set of samples.



(a) Potential energy trace (b) Ratios in marginal variance

Figure 10: (a) Trace of the potential energy of the Hamiltonian in the HMC posterior sampling of the L3D example. The vertical dashed line, indicates the end of the burn in period and the beginning of the posterior sampling. Notice also the logarithmic scale on the horizontal axis. (b) The ratios, in each dimension of the L3D, between the LP approximate variance over the one obtained by the HMC sampling (i.e. $\frac{Var_{\mathbf{LP}}}{Var_{\mathbf{HMC}}}$). The overestimation from the LP approximation is apparent in all three dimensions.

The performance of the new polynomial framework seems to scale well for this multivariate system. As shown in Figures 11(a) and 11(b), when comparing the minimisation of the free energy and the number of iterations to reach convergence, the LP approximation is very stable and fully converges to the original VGPA with only $M = 2$ order of polynomial. The experiments were extended up to $M = 20$, and showed similar outcomes although with higher computational cost and are omitted from the plots. The observation density considered (i.e. $N_{obs} = 10$) implies that $M = 9$ is the limit where both algorithms LP and VGPA optimise the same number of parameters. For values of $M > 9$, the LP becomes more demanding in computational resources. However, when tested with $M = 3$, we obtain $L_{total} = 9,648$ whilst $N_{total} = 24,000$ hence achieving a 59.8% reduction in the number of variables to be optimised.



(a) Variational free energy (b) Number of SCG iterations

Figure 11: (a) Box-plots of the free energy attained from 50 realisations of the observation noise (on a single L3D sample path) as a function of the order of polynomials M. The horizontal dashed line (and the solid ones above and below) represent the 25, 50 and 75 percentiles from the VGPA free energy on the same data sets. (b) Presents a similar plot but for the number of iterations in the SCG optimisation routine at which convergence was achieved. In both plots the extreme values (outliers) have been removed for better presentation.

The reduction in the memory requirements of the algorithm does not produce a similar reduction in computational time. Figures 6(b), 8(b) and 11(b) compare the number of iterations of the LP algorithm to reach convergence with the number of iterations from the VGPA. These results are summaries from 50 different realizations (of the observation noise on a single trajectory) of the OU, DW and L3D systems respectively, and show that the original VGPA algorithm, while optimising a larger number of parameters, still converges in slightly fewer iterations.

5. Parameter estimation in stochastic systems

The original VGPA algorithm can be used to estimate unknown model parameters (Archambeau et al. [5]). The new LP algorithm is also able to estimate the (hyper-) parameters of the aforementioned dynamical systems. In this work the focus is on estimating the drift parameters $\boldsymbol{\theta}$ and diffusion coefficients $\mathbf{\Sigma}$, although estimation of the prior distribution over the initial state (i.e. $\mathcal{N}(\mu_0, \tau_0)$) and the noise related to the observations $\mathbf{R}$ can also be included.

The classical approach to parameter estimation, from incomplete data, is the Expectation-Maximization (EM) algorithm, that was first introduced by Dempster et al. [10] and later extended to partially observed diffusions by Dembo and Zeitouni [9]. However, even though the EM algorithm is well studied with a broad range of applications it cannot be applied successfully in the current variational framework, because the approximate posterior distribution q_t, induced by Eq. (6), is restricted to have the same diffusion coefficient $\mathbf{\Sigma}$. Therefore, although an EM approach can be used to estimate the drift parameters $\boldsymbol{\theta}$, the system noise $\mathbf{\Sigma}$ would have to be held constant during the Maximization step. As a result a different approach for estimating the parameters is adopted.

Based on the fact that the *variational free energy*, Eq. (5), provides an upper bound to the negative log-marginal likelihood (details are in Vrettas et al. [60]):

$$\begin{aligned} -\ln p(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{\Sigma}) &= \mathcal{F}(q(\mathbf{X}|\mathbf{\Sigma}), \boldsymbol{\theta}, \mathbf{\Sigma}) - \mathrm{KL}[q(\mathbf{X}|\mathbf{\Sigma}) \| p(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}, \mathbf{\Sigma})] \\ &\leq \mathcal{F}(q(\mathbf{X}|\mathbf{\Sigma}), \boldsymbol{\theta}, \mathbf{\Sigma}) \,, \end{aligned} \tag{23}$$

where $KL[q\|p] \geq 0$, is the Kullback-Leibler divergence between the approximate and correct posteriors and the time dependence has been omitted, two approaches are considered. Initially a discrete approximation to the posterior is constructed, based on a fixed set of possible parameter values. Subsequently gradient based methods are developed to find the approximate "*maximum a posteriori*" (MAP) values of the parameters.

5.1. Discrete approximations to the posterior distribution

As seen from Equation (23), the negative *free energy* can be substituted for the log marginal likelihood and by choosing suitable prior distributions $p_0(\boldsymbol{\theta})$ and $p_0(\boldsymbol{\Sigma})$, with $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ treated as random variables. To illustrate this approach an example, for the drift parameter $\boldsymbol{\theta}$ is given.

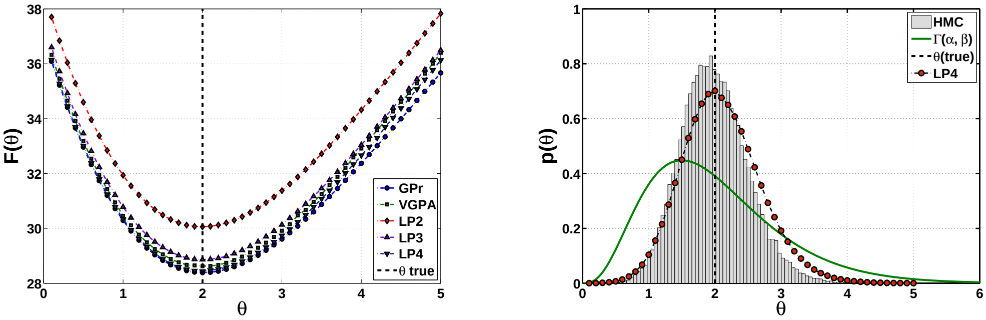
Keeping the diffusion noise $\boldsymbol{\Sigma}$ fixed to its true value, initially select a set of points $\mathcal{D}_{\boldsymbol{\theta}} = \{\boldsymbol{\theta}_i\}_{i=1}^{n_\theta}$ at which to approximate the posterior distribution. Run the variational approximation to convergence with these selected values. This yields a corresponding set of free energy values $\mathcal{D}_{\mathcal{F}} = \{\mathcal{F}(q(\mathbf{X}), \boldsymbol{\theta}_i, \boldsymbol{\Sigma})\}_{i=1}^{n_\theta}$ that can be used to evaluate $\exp\{-\mathcal{F}(q(\mathbf{X}), \boldsymbol{\theta}_i, \boldsymbol{\Sigma})\}$ instead of the true likelihood $p(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Thus

$$p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Sigma}) \propto \left\{ \exp\{-\mathcal{F}(q(\mathbf{X}|\boldsymbol{\Sigma}), \boldsymbol{\theta}_i, \boldsymbol{\Sigma})\} \times p_0(\boldsymbol{\theta}_i) \right\}_{i=1}^{n_\theta}, \tag{24}$$

where $n_\theta \in N$ is the number of discrete points. Similar discrete approximations, to the posterior distribution, can be computed for the system noise $\boldsymbol{\Sigma}$. In the above procedure the parameters that are not approximated are kept fixed (to their true values). In the results that follow Gamma priors are defined for the drift parameters and inverse Gamma for the system noise covariance, i.e. $p_0(\boldsymbol{\theta}) = \mathcal{G}(\alpha, \beta)$ and $p_0(\boldsymbol{\Sigma}) = \mathcal{G}^{-1}(a, b)$. The values of the parameters α, β, a and b, were chosen such that the mean value of the distribution coincides to the true values of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$, but with large variance to reflect our "ignorance" about the true values of the parameters.

Figure 12(a), compares the profile of the approximate marginal likelihood, of the OU drift parameter, obtained with the original variational framework (VGPA) and the local polynomial (LP), on a typical realisation. For this system we also show the "true" marginal likelihood obtained using a Gaussian process regression smoother (with OU kernel function). The LP framework converges to the original VGPA when 4'th order polynomials are employed, which is consistent with the state estimation results in Fig 6(a). The minimum of the profile can be well identified with only 2'nd order polynomials, which suggests that for the drift parameter, in this example, the bound on the true likelihood does not need to be very precise, if a point estimator is sought.

Figure 12(b), shows the results from the LP (of 4'th order) discrete approximation to the posterior distribution of the drift parameter θ using a $\mathcal{G}(4.0, 0.5)$ prior. Here the results are compared with $80,000$ samples from the posterior (presented as a histogram), obtained from four independent Markov chains ($20,000$ samples per chain), using HMC sampling. The same prior distribution (continuous green line) is

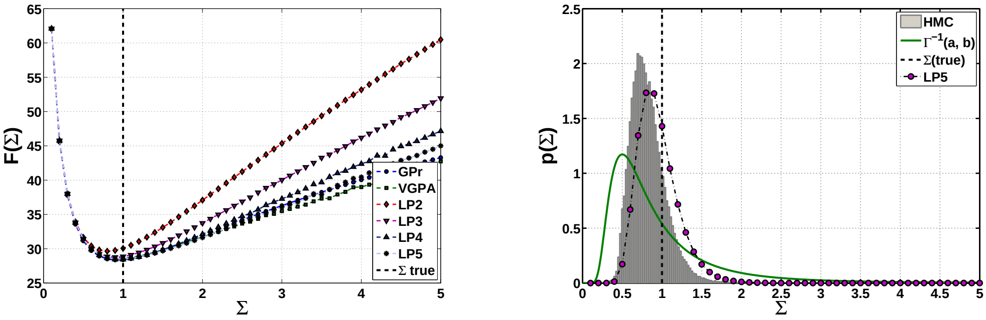


(a) θ marginal profile

(b) Posterior distribution p(θ)

Figure 12: *OU system:* (a) The profile marginal likelihood of the drift parameter θ, keeping the system noise Σ fixed to its true value, obtained by the GP regression (blue circles) with the OU kernel, which gives the exact likelihood, against the original VGPA algorithm (green squares) and the new LP extension with different order of polynomials. (b) The histogram of the posterior samples obtained with the HMC. The continuous green line shows the $\mathcal{G}(4.0, 0.5)$ prior of the (hyper)-parameter θ, while the red circles connected with the dot-dashed line represent the discrete approximation to the posterior distribution obtained by the point estimates of the LP algorithm with 4'th order polynomials. Both the HMC posterior sample histogram and the LP approximation have been normalized, such that the area they define sums to unity. In both figures the vertical dashed line represents the true parameter value that generated the data.

used in both cases and in addition the results are presented such that the areas defined by the histogram and the approximate discrete estimates (red circles), sum to one. Although the results, for both algorithms, are slightly biased the LP algorithm provides a better approximation because for a linear system, such as the OU, the variational Gaussian process yields an optimal approximation while the HMC approximation remains subject to finite sample effects.

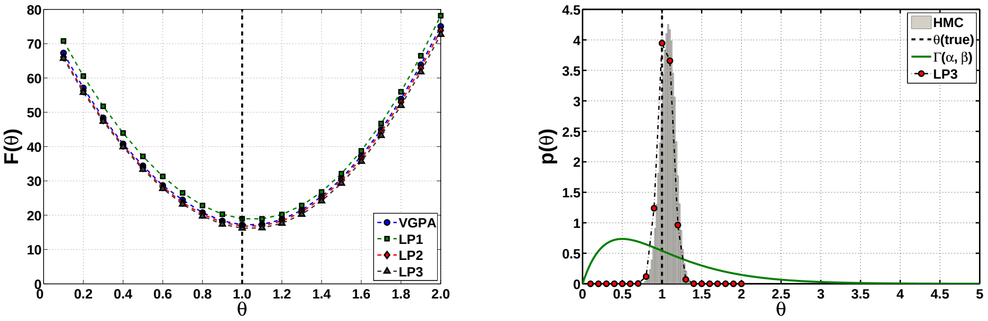


(a) Σ marginal profile

(b) Posterior distribution p(Σ)

Figure 13: *OU system:* (a) Plot similar to Fig. 12(a) only for the system noise Σ and keeping the drift θ fixed to its true value. Again, the results of the GP regression represent the exact marginal likelihood. (b) As Fig. 12(b), only the continuous line now is the $\mathcal{G}^{-1}(3.0, 2.0)$ prior of the (hyper-) parameter Σ.

Figures 13(a) and 13(b), show similar profile and posterior results, but for the OU system noise coefficient Σ. It is apparent that for this parameter the LP method needs higher order of polynomials to match the results from the original VGPA. All methods locate the minimum of the profile at a smaller value than the true one. Furthermore, both methods seem to deviate from the true likelihood (blue circles), as the value of this parameter becomes more distant from the true value that generated the data. The same bias effect can also be seen in Figure 13(b), where the LP method (5'th order) is compared with the HMC posterior sampling. However, MCMC methods for sampling this parameter can be problematic due to the high dependencies between the system noise Σ and the states of the system X_t, which results in slow rates of convergence (Roberts and Stramer [52], Golightly and Wilkinson [22]). Again the same $\mathcal{G}^{-1}(3.0, 2.0)$ prior (continuous green line), was used for both algorithms.



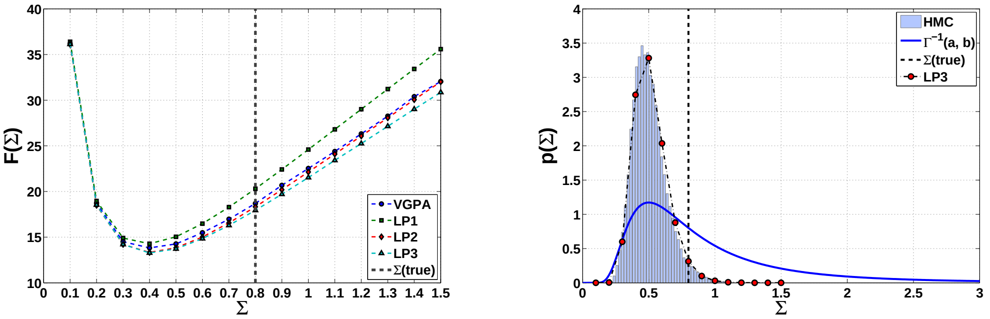
(a) θ marginal profile

(b) Posterior distribution p(θ)

Figure 14: *DW system:* (a) The profile approximate marginal likelihood of the drift parameter θ, keeping the system noise Σ fixed to its true value, obtained by original VGPA algorithm (blue circles) and the new LP extension with different order of polynomials. (b) The histogram of the posterior samples obtained using the HMC. The continuous green line shows the $\mathcal{G}(2.0, 0.5)$ prior of the (hyper-) parameter θ, whilst the red circles connected with the dot-dashed line represent the approximate posterior distribution obtained by the discrete estimates of the LP algorithm with 3'rd order polynomials. Both the HMC posterior sample histogram and the LP point estimates have been normalized, such that the area they define sums to unity.

Likewise, the approximate posterior distributions and profile likelihoods, for a single realisation of the DW system are presented for the drift θ in Figures 14(a) and 14(b) and for the diffusion coefficient Σ in Figs. 15(a) and 15(b). Here there is no method to compute the exact likelihood, hence the only comparison is between the profiles obtained from the VGPA algorithm against those obtained with the LP method. For both parameters θ and Σ, the results are almost identical with 3'rd order polynomials. Both estimates are biased, the drift towards a higher value, while the noise towards a smaller value, but these biases are consistent with those seen in the HMC posterior samples.

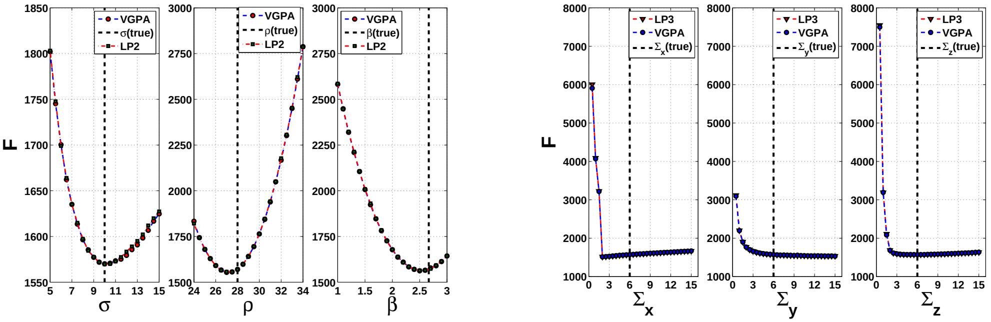
The profiles of the drift parameter vector $\boldsymbol{\theta} = [\sigma \ \rho \ \beta]^\top$ for the $L3D$ system are shown in Fig. 16(a) where the original VGPA algorithm (red circles) is plotted against the LP approximation, with 2'nd order polynomials (green squares). The results are almost indistinguishable and the minimum values are well estimated for all parameters. Figure 16(b), presents similar profiles but for the diagonal elements of the $\boldsymbol{\Sigma}$ matrix (i.e. Σ_x, Σ_y and Σ_z). Both the VGPA and the LP (3'rd order) exhibit identical behaviour; unlike



(a) Σ marginal profile (b) Posterior distribution $\mathrm{p}(\Sigma)$

Figure 15: *DW system:* (a) Plot similar to Fig. 14(a) only for the system noise Σ and keeping the drift θ fixed to its true value. (b) As in Fig. 14(b), only the continuous line now is the $\mathcal{G}^{-1}(3.0, 2.0)$ prior of the (hyper-) parameter Σ. Again the areas that both algorithms define (HMC and LP) have been normalized. In both figures the vertical dashed line represent the true parameter value that generated the data.

the drift parameters the system noise profiles are not as informative. Only the first dimension 'x', shows a clear minimum, although biased towards a smaller value (the true values are indicated with vertical dashed lines). The third dimension 'z', shows a weak minimum, i.e. there is quite flat region around the minimum value and the second dimension 'y', does not posses a minimum within the range of values explored.



(a) $\boldsymbol{\theta}$ marginal profiles (b) $\boldsymbol{\Sigma}$ marginal profiles

Figure 16: *L3D system:* (a) The profile approximate marginal likelihood for all three parameters of the *L3D* drift vector. From left to right the profiles for σ, ρ and β obtained from the original VGPA algorithm (red circles) are compared against those obtained with the LP with 2'nd order polynomials (green squares). (b) As before but for the system noise, on each dimension (Σ_x, Σ_y and Σ_z). Here the LP approximation uses 3'rd order polynomials. The vertical dashed lines indicate the true values of the parameters that generated the datasets.

Figure 17 (upper three panels), presents the posterior estimates of the *L3D* drift vector $\boldsymbol{\theta}$, obtained from the HMC algorithm. The lower three panels present the approximate posterior distributions (discrete estimates) from the LP algorithm. Both algorithms used the same prior distributions ($p_0(\sigma) = \mathcal{G}(20, 0.5)$, $p_0(\rho) = \mathcal{G}(56, 0.5)$ and $p_0(\beta) = \mathcal{G}(6, 0.5)$), nonetheless the comparison between the upper and lower panels is not straightforward, because the approximate posterior distributions obtained with the LP algorithm are conditional, in the sense that the two other drift parameters are kept fixed to their true values, whereas the posterior distributions from the HMC are obtained jointly (i.e. all the drift parameters are sampled simultaneously). The results from the LP method show week biases towards smaller values in all parameters, which is consistent with the HMC results, except the σ parameter (first column) which the LP approximation estimates more accurately.

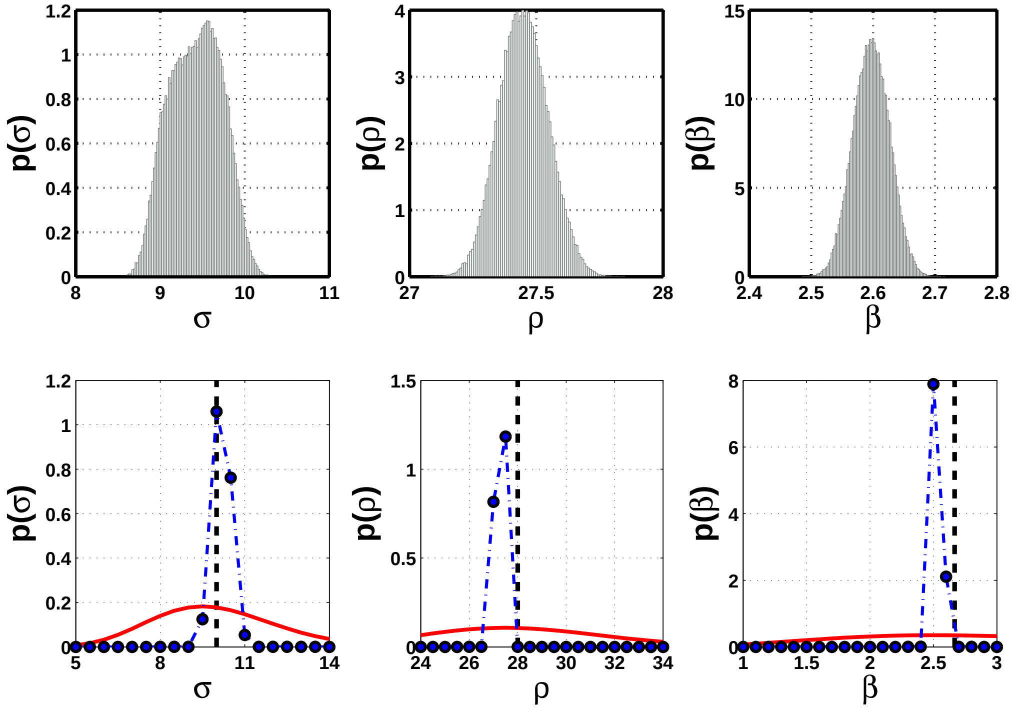


Figure 17: *L3D system:* The upper three panels, starting from left to right, present the joint posterior HMC samples for the drift parameters σ, ρ and β. The lower three panels, following the same order, show the approximate posterior distributions (blue dots connected with the dot-dashed line) obtained from the LP algorithm with 2'nd order polynomials. The continuous lines represent the $\mathcal{G}amma$ prior distributions that were used. Notice that the priors are very broad. In all the above results the system noise is assumed to be known and fixed to its true value.

5.2. Maximum likelihood type-II point estimates

Another approach for estimating the (hyper-) parameters, as suggested in [5], is also based on the bound that the *variational free energy* provides to the marginal likelihood (Eq. 23), but instead of constructing approximate posterior distributions to the (hyper-) parameters, as in the previous section, it employs a conjugate gradient algorithm to provide point estimates. More specifically, the algorithm works in an outer/inner loop optimisation framework, where in the inner loop the variational approximation framework is used to compute the optimal approximate posterior process $q(\mathbf{X}_t)$, given a fixed set of the parameters ($\boldsymbol{\theta}$

and $\boldsymbol{\Sigma}$). Then, in the outer loop, a gradient step is taken to improve the current estimates of the (hyper-) parameters. This procedure, as shown in Table 3, alternates until the gradients of the optimal process (Eq.10), with respect to the $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ are zero ($\nabla_{\boldsymbol{\theta}}\mathcal{L} = 0$ and $\nabla_{\boldsymbol{\Sigma}}\mathcal{L} = 0$), or the estimates cannot improve any further (i.e. the optimal Gaussian process estimated in the inner loop does not change significantly, e.g. $\Delta\mathcal{L} \leq 1.0e - 6$).

```
ML type-II parameter estimation algorithm
initialize{$\boldsymbol{\theta}_0$, $\boldsymbol{\Sigma}_0$, $n = 0$, $N_{max} = 1,000$}        \* initialize the algorithm *\
$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$, $\boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma}_0$                              \* set the initial parameter values *\
$\mathcal{L} \leftarrow$ inner-loop($\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$)                          \* optimal posterior process *\
outer-loop:
  compute{$\nabla_{\boldsymbol{\theta}}\mathcal{L}$, $\nabla_{\boldsymbol{\Sigma}}\mathcal{L}$}                    \* gradients w.r.t. the parameters *\
  if ($\nabla_{\boldsymbol{\theta}}\mathcal{L}^{\top}\nabla_{\boldsymbol{\theta}}\mathcal{L} == 0$ or $\nabla_{\boldsymbol{\Sigma}}\mathcal{L}^{\top}\nabla_{\boldsymbol{\Sigma}}\mathcal{L} == 0$)    \* check if the gradients are zero *\
    return{$\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$}                              \* return the old parameter values *\
  end
  update{$\boldsymbol{\theta}^*$, $\boldsymbol{\Sigma}^*$}                                \* new parameter values *\
  $\mathcal{L}^* \leftarrow$ inner-loop($\boldsymbol{\theta}^*$, $\boldsymbol{\Sigma}^*$)                     \* new cost function value *\
  if {$\Delta\mathcal{L}^*$ & $\Delta\boldsymbol{\theta}^*$ & $\Delta\boldsymbol{\Sigma}^*$} $\leq 1.0e - 6$            \* check for termination *\
    return{$\boldsymbol{\theta}^*$, $\boldsymbol{\Sigma}^*$}                            \* return the new parameter values *\
  end
  $\mathcal{L} \leftarrow \mathcal{L}^*$, $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^*$, $\boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma}^*$                 \* set the old values to the new *\
  n ← n+1                                        \* increase the loop counter by one *\
while($n \leq N_{max}$)                                \* maximum number of iterations *\
return{$\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$}                                  \* if it has not convergence yet *\
```

Table 3: Pseudo-code of the "*maximum a posteriori*" (MAP) estimation algorithm in practice. Every time the parameters are updated the *inner-loop($\boldsymbol{\theta}$,$\boldsymbol{\Sigma}$)* function recomputes the optimal Gaussian process approximation for a given set of fixed parameter values.

The same dual optimisation approach can also be used with the LP approximation framework, without any change in the code, since the re-parametrisation of the variational parameters $\mathbf{A}_t$ and $\mathbf{b}_t$, affects only the smoothing algorithm (inner loop), while leaving the outer loop unaffected. In fact, the new approach is more flexible, because we can adjust the bound of the variational algorithm to the marginal likelihood, by tuning the order of the polynomial approximation. To present a more comprehensive study the new LP approximation framework is compared, in terms of estimating the (hyper-) parameters of the aforementioned dynamical systems with other well known method that cover all the main categories that deal with the Bayesian inference problem.

The first method considered is based on the unscented Kalman filter (UnKF). As discussed in Section 1, this method utilizes a technique known as the "*unscented transformation*", to estimate the states of the dynamical system considered and was primarily introduced, as an alternative to the extended Kalman filter (EKF), to address its linearisation limitations. The UnKF has been extended to model parameter estimation problems [61, 62]. Two approaches were taken: (a) augmenting the state vector with the model parameters and then applying a single filter recursion to estimate both of them *jointly* and (b) using two separate filters one to estimate the system states, given the current estimates for the parameters, and one to estimate the model parameters given the current state estimates. In approach (b) two filters are run in parallel and are known as the *dual filter*. In this work a dual unscented Kalman filter (dual UnKF), similar to the one used by Gove and Hollinger [24] to assimilate net CO_2 exchange between the surface and the atmosphere, is implemented.

The second algorithm considered is based on the four dimensional variational assimilation method. As described earlier, the *4D-Var* method minimizes a cost function that measures the distance of the most probable trajectory from the observations, within a predefined time window of inference. In most operational implementations the model equations are assumed perfect (strong constraint), or that the errors are sufficiently small to be ignored. In this work the model is assumed to be known only approximately, hence allowing for model error to exist in the problem formulation. This formulation is known as "*weak constraint 4D-Var*". Tremolet [57], describes different variations of this algorithm, with the one closer to our approach denoted in his work, as "$4D - Var_x$", where the subscript "x" denotes the control variable in the optimisation procedure. In our implementation since every (discrete in time) system state $\mathbf{x}_k$ is a control variable we also refer to it as "*full weak constraint 4D-Var*".

Although this method is well studied for estimating the states of a system, not much work has been done in estimating model parameters. Navon [46] provides a useful review for parameter estimation, in the context of meteorology and oceanography. In our work a dual approach similar to the LP approximation algorithm is taken. The estimation framework is based on an outer/inner optimisation loop. The inner loop estimates the most probable trajectory, given the current estimates for the drift and diffusion parameters and subsequently the outer loop, conditioning on the most probable trajectory, updates the estimates of the parameters by taking a gradient descent step. The cost function to optimize is given by:

$$\mathcal{J}_{cost} = \mathcal{J}_{x_0} + \mathcal{J}_f + \mathcal{J}_{obs} + \mathcal{J}_{hp} + \mathcal{C} \ , \tag{25}$$

where $\mathcal{J}_{x_0}$, is the contribution of the prior over the initial state $\mathbf{x}_{k=t0}$, $\mathcal{J}_f$ is the influence of the model equations (drift function), $\mathcal{J}_{obs}$ is the contribution of the observations, $\mathcal{J}_{hp}$ comes from the priors over the (hyper-) parameters and $\mathcal{C}$ is a constant value that depends on the system noise coefficient $\mathbf{\Sigma}$ (details of the cost function can be found in Appendix B). In practice, one needs to compute the gradients of the cost function with respect to the control variables (i.e. $\nabla_{\mathbf{x}_{0:N}} \mathcal{J}_{cost}$), for estimating the most probable trajectory (inner loop) and then the gradients of the cost function with respect to the (hyper-) parameters (i.e. $\nabla_{\boldsymbol{\theta}}\mathcal{J}_{cost}$ and $\nabla_{\boldsymbol{\Sigma}}\mathcal{J}_{cost}$), for updating their values in the outer optimization loop.

The following sections present an empirical comparison of the marginal and joint estimation of the drift $\boldsymbol{\theta}$ and diffusion coefficient $\boldsymbol{\Sigma}$, using the UnKF, 4D-Var and LP methodologies in two distinct asymptotic regimes: (a) *infill asymptotics*, where the observations are sampled more and more densely, within a fixed time domain (i.e. $N_{obs} \to \infty$, while $T = [t_0, t_f]$) and (b) *increasing domain asymptotics*, where the observation density remains fixed, whilst the time window of inference increases (i.e. $N_{obs} =$ const. and $T \to \infty$).

5.2.1. Infill asymptotic behaviour, $N_{obs} \to \infty$

Before proceeding a few issues need to be clarified concerning the presentation of the results. As mentioned earlier the variational LP approximation method and the weak constraint 4D-Var based algorithm, provide point estimates of the (hyper-) parameters, in a gradient based optimisation framework. The dual unscented Kalman filter approach provides mean estimates (of the parameters), as a function of time. To make the results of the dual UnKF more comparable with those from the other two methods we treated the collection of the mean estimates as a (filtered) distribution and compute estimates of its moments, such as the mean value (Hansen and Penland [25]). An example of this procedure is shown in Figure 18, where the dual UnKF is applied to estimate the drift parameter of the DW system, on a single data set. As a general rule, we used only the second half of the mean estimate values. We argue that in these controlled experiments[4] there is no need to average over the whole time window because the initial estimated value is wrong by construction. Hence we allow the filter to converge around a value before averaging. The second remark has to do with the quantities that we plot. In order to provide a more general analysis thirty different observation noise realisations were created, for each observation density. The results are presented as summary statistics, illustrated using the 25'th, 50'th (or median value) and the 75'th percentile of the estimated values from each algorithm.

We begin with the conditional[5] drift estimation of the OU and DW systems (see Figures 19(a) and 19(b) accordingly). The results for the OU system show that the LP approximation has a small increasing trend and settles to a higher value, compared with 4D-Var, although this higher value is also seen in the HMC posterior estimates of this parameter (Fig. 12(b)). Also both algorithms narrow the range of estimates, as

[4] Here we imply that we know a priori the true values that generated the data and also we know that the initial value of the estimation process is deliberately wrong but close to the true one.

[5] This term is used to signify that all the other parameters, such as the system and observation noises ($\boldsymbol{\Sigma}$ and $\mathbf{R}$), are assumed known and fixed to their true values.

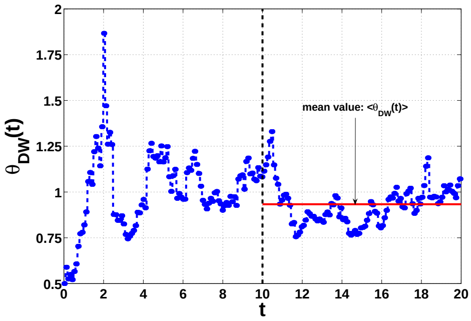
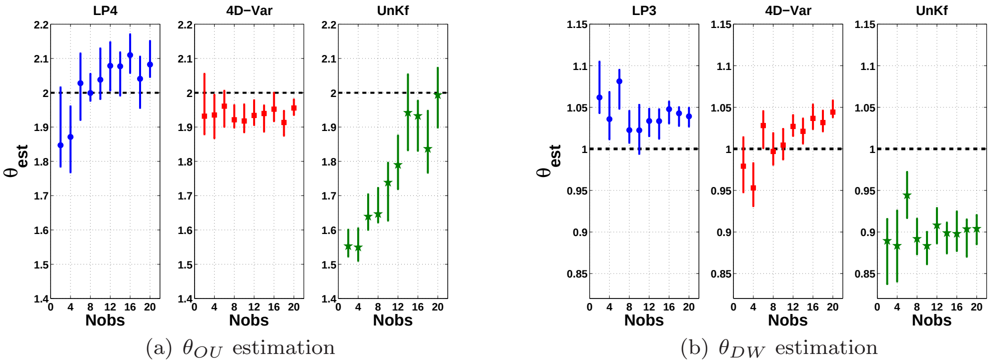


Figure 18: An example of mapping the results from the application of the dual UnKF algorithm applied to a single trajectory, estimating the DW drift parameter, to a point estimate (mean value). The blue circles indicate the ensemble mean estimates as a function of time, while the continuous red line is the mean value of these estimates over the period used for averaging. The vertical dashed line marks the beginning of the time window where the average takes place.

the observation density increases (the error bars are closer to the median value), as one would expect. On the other hand the results from the UnKF based algorithm, show a more steep trend and only when the system is highly observed are the estimates close to the true generating value. Here, as in all the experiments that follow, all three algorithms were initialized with the same value for the parameter(s) that were estimated.



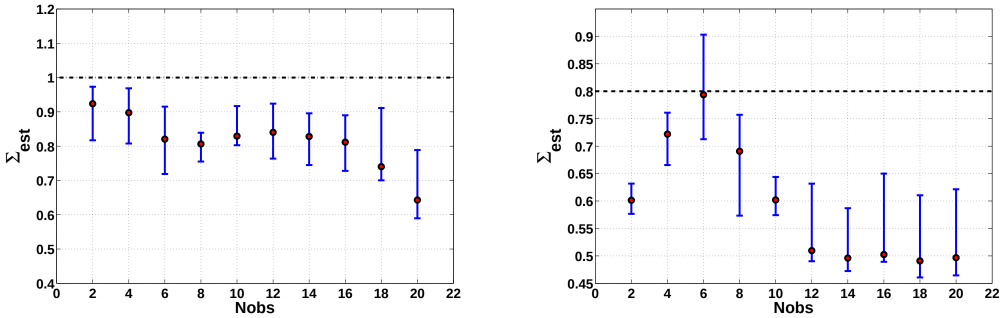
(a) θ_{OU} estimation

(b) θ_{DW} estimation

Figure 19: *Drift (conditional) estimation:* (a) Presents the summary statistics (25'th, 50'th and 75'th percentiles) after estimating the drift parameter θ from thirty different realizations, of the observation noise, on the OU system keeping the system noise coefficient Σ fixed to its true value. The left panel (blue) presents the results from the *LP* algorithm, while the middle (red) and the right (green) the results from the *(full) weak-constrained 4D-Var* and the *dual UnKF* accordingly. In (b) we repeat the same estimation experiment but for thirty different realizations, of the observation noise, of the DW system. All estimation results are presented as functions of increasing observation density.

For the DW system the algorithms were more stable, in the sense that they converge to a stable value and there are no major trends as in the OU case. The results from all methods are biased either towards higher values (LP and 4D-Var), or lower values (UnKF). Once again the LP algorithm bias matches the HMC posterior estimates as shown in Fig. 14(b). Although the results from the dual UnKF seem inferior compared to the other two algorithms, it should be recalled that this is a filter estimation, which means that it “sees” the observations sequentially, only up to the current time and does not take into account the future observations.

Figures 20(a) and 20(b), present the results of estimating the system noise Σ, of the OU and DW systems. It is obvious that the estimation for the OU system is stable, while for the DW the process needs to be well observed (e.g. $N_{obs} \geq 10$), before convergence to a value is seen. Both plots show consistency with the HMC posterior estimates from the previous section. Here we show only the estimates obtained from the LP approximation method. The other methods, although they were applied to the same datasets, they were unable to provide good estimates, hence were omitted. Recently, DelSole and Yang [8], presented an ensemble Kalman filter (EnKF) for providing general maximum likelihood estimates for the state and model parameters, of stochastic dynamical systems. In this paper the authors obtain good estimates of the noise (stochastic) parameters, although in a rather different setting then the one considered here. However, this Kalman filtering approach is unable to estimate simultaneously the drift and diffusion parameters as we present later.



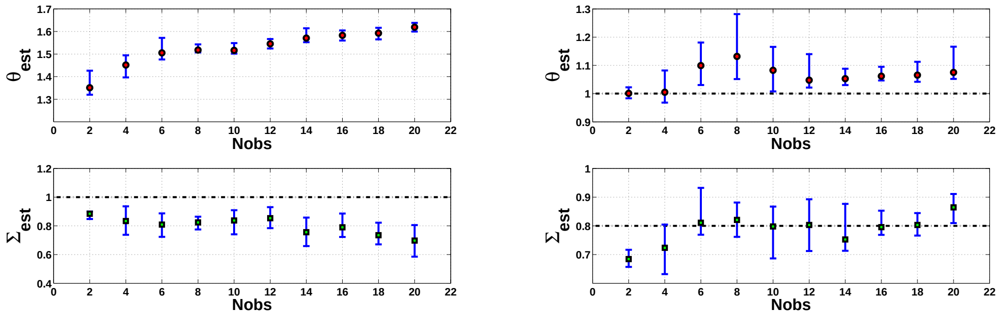
(a) Σ_{OU} estimation

(b) Σ_{DW} estimation

Figure 20: *Noise (conditional) estimation:* (a) shows the conditional estimation of the system noise coefficient Σ, keeping θ to its true value. The plot presents the 50’th percentile (red circles) and the 25’th to 75’th percentiles (blue vertical lines). (b) repeats the same experiment but for the DW system. All results were obtained with the *LP* method (3’rd order) and presented as functions of increasing observation density.

The experiments on the uni-variate systems conclude with the joint estimation of the drift parameter θ and the system noise coefficient Σ. Figures 21(a) and 21(b), summarize the results obtained from the LP approximation method. The drift estimation for the OU system, shows a significant bias to smaller values (compared with the conditional estimation of Fig. 19(a)), where the bias was towards a higher value. These estimates become more confident as the observation density increases (smaller error bars). Meanwhile, the estimation of the OU diffusion noise is consistent with the conditional outcomes. Unlike the OU system, the

DW shows consistent estimation for the drift parameter and a surprising improvement of the system noise estimation. In these plots, in contrast to the conditional ones, we can not refer directly to the posterior HMC estimates, because here the parameters are estimated simultaneously, while the results of the HMC were obtained by fixing the parameters that are not estimated to their true values.



(a) OU system

(b) DW system

Figure 21: *Joint estimation:* In (a) the drift and diffusion coefficient, of the OU system, are estimated jointly. The left upper panel shows the results for θ, while the left lower panel for Σ. The results are summaries (25'th, 50'th and 75'th percentiles) from thirty different observation realizations. (b) shows the same joint estimation but for the DW system. The right upper panel shows the results for θ, while the right lower panel for Σ. All results were obtained with the *LP* method (3'rd order) and presented as functions of increasing observation density.

Next we consider the conditional estimation of the drift vector $\boldsymbol{\theta}$, of the L3D system (Figure 22). It is clear that in this example the 4D-Var estimation method (middle column), performs better and produces more stable and certain results. The LP algorithm when tested with 4 and 6 observations per time unit seems to be under-sampled, hence the state estimation (inner loop of the optimisation procedure), does not actually converge to the optimal posterior process. Therefore, the parameter estimates are also not reliable. When the process is observed more frequently (e.g. $N_{obs} \geq 8$), it produces more stable results. The dual UnKF estimation results are reliable, with the exception of the ρ parameter (third column, second row), which is very biased with sparse observations. However, all parameters asymptotically converge close to the true values, as the observation density increases.

Similar to the univariate systems, the conditional estimation of the system noise coefficient $\boldsymbol{\Sigma}$, was feasible only with the variational LP approximation algorithm. Because the covariance matrix is assumed diagonal (see Eq.3), we only need to estimate the three diagonal components, which correspond to the noise added in each dimension of the L3D dynamical equations (see Eq.19). Figure 23 suggests that to estimate this very important parameter one has to have dense observations. For the L3D system we observe all three dimensions. Components Σ_x and Σ_z converge close to the true values roughly after 16 observations, per time unit, while the Σ_y parameter converges to a higher value. These results are in agreement with the

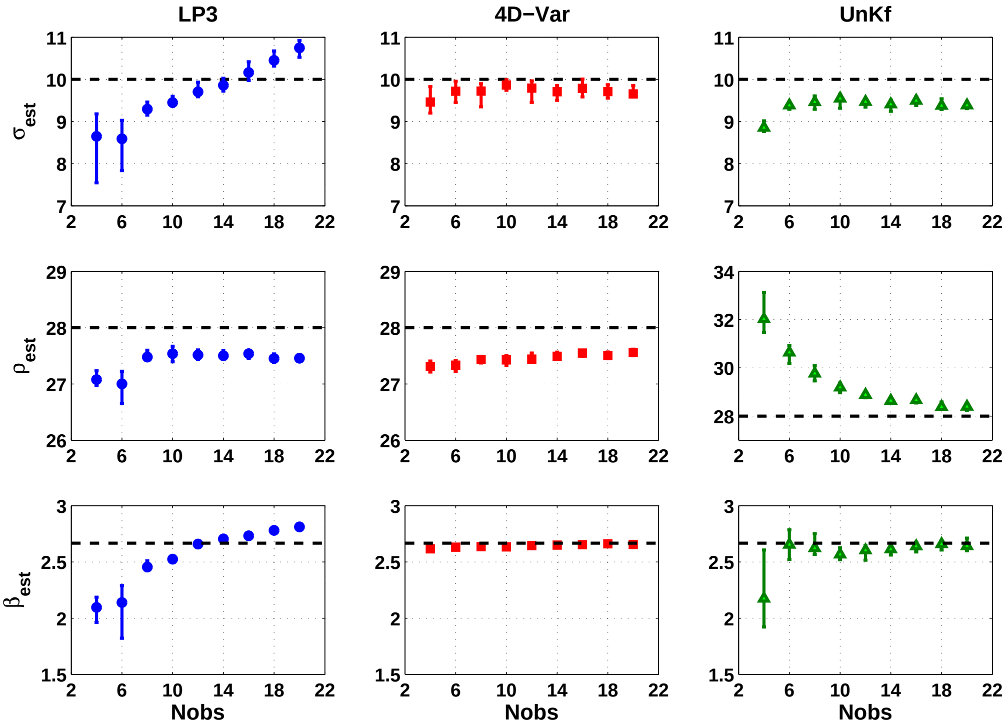


Figure 22: *Drift (conditional) estimation:* The infill asymptotic results for the *L3D* drift parameter vector $\boldsymbol{\theta}$. The summary results when seen horizontally compare the same drift parameter but with different estimation method, while vertically the results are presented for the same estimation method but for all three parameters (σ, ρ and β). The methods tested, from left to right are the *LP* algorithm (3'rd order), the *(full) weak-constraint 4D-Var* and the *dual UnKF* accordingly. In all sub-plots the horizontal dashed lines indicate the true values of the drift parameters that generated the observed trajectories. Where possible the *y-axis* was kept the same for all plots to make comparison easier. All algorithms were tested on the same thirty different realisations of the observation noise.

approximate marginal profiles produced earlier (Fig. 16(b)).

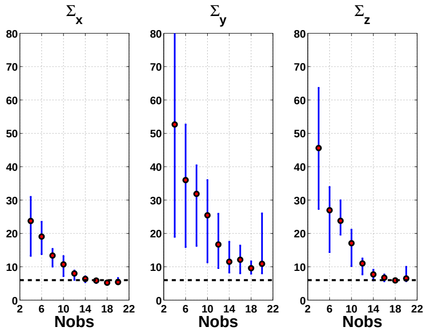


Figure 23: *Noise (conditional) estimation:* Summary results (25'th, 50'th and 75'th percentiles) from thirty different observation realizations, of the *L3D* system, when estimating conditionally the system noise coefficient matrix $\boldsymbol{\Sigma}$. The results were obtained using the *LP* algorithm (3'rd order) and presented as functions of increasing observation density. The estimation of the noise is presented separately in each dimension x, y and z from left panel to right accordingly.

To conclude with the *infill asymptotics* section, we demonstrate the application of the newly proposed LP approximation framework to the joint estimation of the drift and diffusion matrix of the L3D system. In total we estimate six (hyper-) parameters (σ, ρ, β, Σ_x, Σ_y and Σ_z), as shown in Figure 24. The asymptotic behaviour is similar to that observed when estimating the parameters conditionally, which gives us some level of confidence that our algorithm is stable. The general message is that we achieve good estimates when the system is well observed.

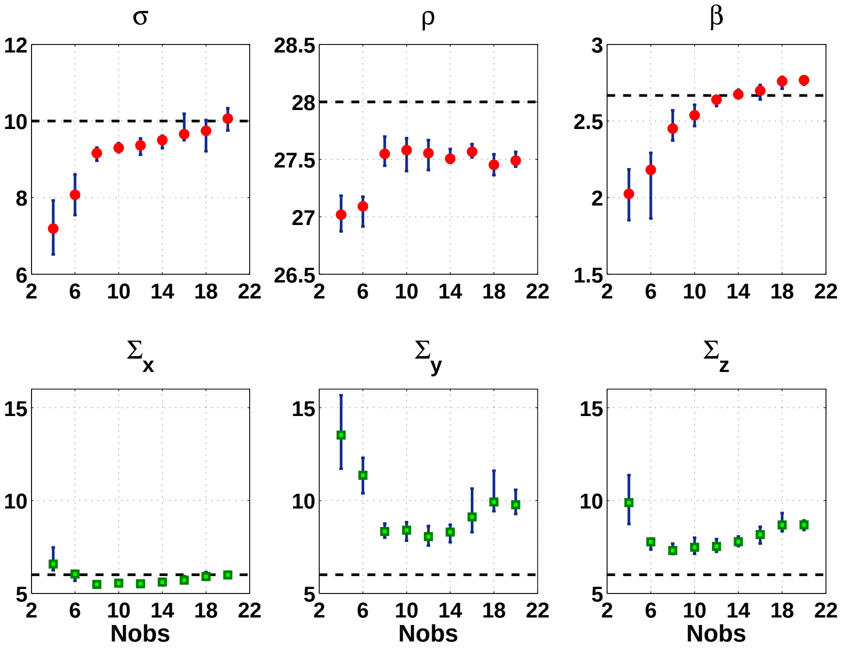


Figure 24: *Joint estimation:* The summary results (25'th, 50'th and 75'th percentiles) when estimating jointly the drift parameters σ, ρ and β (upper three panels), and the system noise coefficients Σ_x, Σ_y and Σ_z (lower three panels), of the $L3D$ system. The same dataset of the thirty different realisations of the observation noise is used, as in the previous experiments.

5.2.2. *Increasing domain asymptotic behaviour, $T \to \infty$*

This section discusses another important asymptotic property; when the observation *density* remains fixed, but the duration that an event (or the random process) is observed, increases to infinity. To explore this behaviour new extended sample paths were created for all the dynamical systems considered in our previous simulations and then the total time-window was split into smaller, but equal, time intervals. TAn example is given on the DW system. As presented in Figure 25, we have a sample path (or history), of the DW system, with time-window $T_{total} = [0, 50]$. The next step consists of measuring the history with fixed observation density (e.g. $N_{obs} = 2$). Then the total time-window is divided in five sub-domains of ten time units to create five time-windows ($T_{10} = [0, 10]$, $T_{20} = [0, 20]$, $\cdots$, $T_{50} = [0, 50]$), including the observations from the previous steps. Finally, the estimation methods are applied on each sub-interval, by introducing the new observations incrementally.

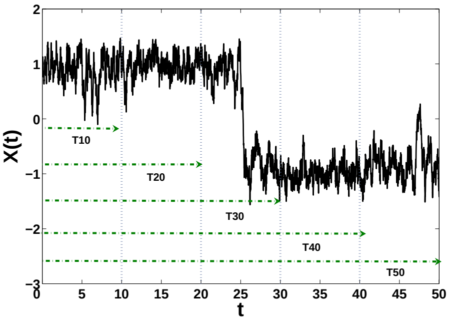
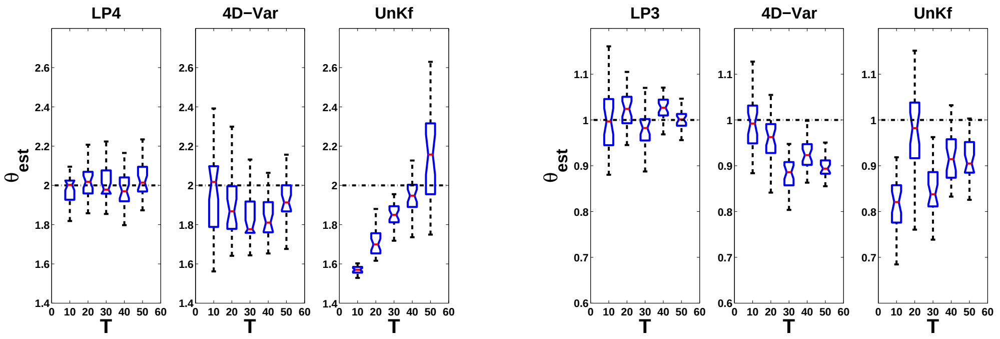


Figure 25: A typical example of a *DW* sample path with an extended time-window that is used for the *increasing domain* asymptotic behaviour of the algorithms. The vertical dotted lines split the total time window in five time domains starting from $T_{10} = [0, 10]$ to $T_{50} = [0, 50]$, which are presented to the estimation methods incrementally.

Figures 26(a) and 26(b), show the results of the conditional drift estimation for the OU and the DW systems respectively, as the time-window of inference increases. As in the *infill asymptotic* simulations, thirty different realizations of the observation noise were generated and the results are presented as summary statistics of the estimation outcomes. Here because the simulations performed were fewer than the previous case all the results are presented with box-plots which provide a richer presentation. It is apparent that in this type of asymptotic convergence, the LP approximation algorithm is remarkably stable with results that are very close to the ones that generated the data. The drop under the true value (as indicated by the horizontal dashed line), in the DW example (Fig. 26(b)), for the third time window (i.e. $T_{30} = [0, 30]$), can be explained by the fact that the transition between the two wells, happens between the 22'nd to 27'th time units, as shown in Figure 25, affecting the estimation. However, when the time-window increases further the algorithm recovers to the initial value. For the same example, the 4D-Var method starts with a higher estimated value but after the transition occurs it settles to a lower value. A similar behaviour can also be observed for the UnKF results, were the method approaches the true value, although it becomes less confident (larger error bars), which was unexpected behaviour.

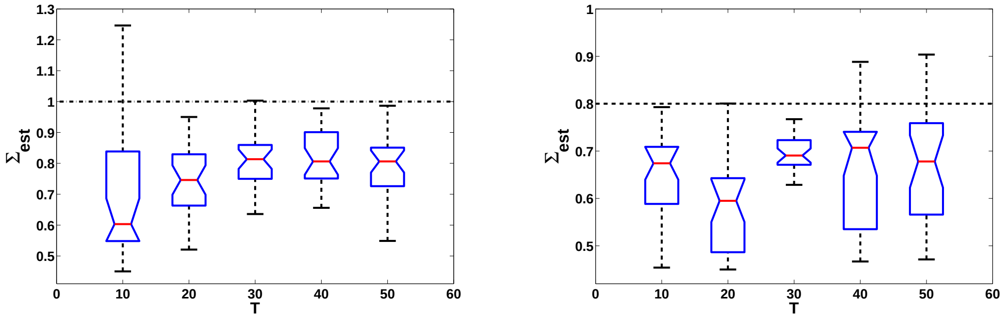
The conditionally estimated diffusion coefficients are presented in Figures 27(a), for the OU and 27(b), for the DW. Here only the LP approximation method was used, as in the previous section. The estimates, for both examples, are stable and improve as the time window increases. Especially for the DW, the results get closer to the true value after the transition has been observed (T_{30}). In a similar way, the results for the joint estimation of the drift θ and diffusion Σ, are consistent and presented in Figures 28(a) and 28(b).



(a) θ_{OU} estimation

(b) θ_{DW} estimation

Figure 26: *Drift (conditional) estimation:* (a) Presents the summary statistics (box-plots) after estimating the drift parameter θ from thirty different realizations, of the observation noise, on the OU system keeping the system noise coefficient Σ fixed to its true value. The left panel presents the results from the *LP* algorithm, while the middle and the right the results from the *(full) weak-constrained 4D-Var* and the *dual UnKF* accordingly. In (b) we repeat the same estimation experiment but for thirty different realizations, of the observation noise, of the DW system. All estimation results are presented as functions of increasing time domain, keeping the observation density fixed.

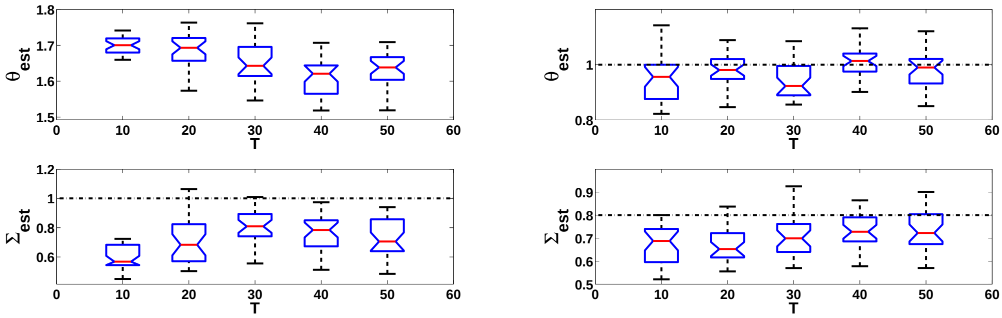


(a) Σ_{OU} estimation

(b) Σ_{DW} estimation

Figure 27: *Noise (conditional) estimation:* (a) shows the conditional estimation of the system noise coefficient Σ, keeping θ to its true value. The plot presents box-plots (5'th, 25'th, 50'th, 75'th and 95'th percentiles), from thirty different realizations, of the observation noise, of the OU system. (b) repeats the same experiment but for the DW system. All results were obtained with the *LP* method (3'rd order) and presented as functions of increasing time domain, keeping the observation density fixed.

This section concludes with the results of the L3D system. Figure 29, presents the summaries of the jointly estimated drift parameter vector $\boldsymbol{\theta} = [\sigma \;\rho\; \beta]^\top$, conditional on the system noise matrix $\boldsymbol{\Sigma}$ set to its true value, from all three estimation methods. All algorithms are stable and produce good results, with 4D-Var having the smallest bias. Once again, the 4D-Var and UnKF methods failed to provide stable results when estimating the system noise coefficients, hence only results from the LP method are shown. The joint estimation of the noise coefficients Σ_x, Σ_y and Σ_z, conditional on the drift vector $\boldsymbol{\theta}$ being fixed to it true value, are illustrated at Figure 30, where it was necessary to observe with quite high density ($N_{obs} = 18$). In addition, the joint estimation of all the (hyper-) parameters, of the L3D system, as the time-window of inference increases, is shown in Figure 31. The results are in accordance with the conditional estimates, although the observation density was set to ten observations, per time unit (i.e. $N_{obs} = 10$).



(a) OU system

(b) DW system

Figure 28: *Joint estimation:* In (a) the drift and diffusion coefficient, of the OU system, are estimated jointly. The left upper panel shows the results for θ, while the left lower panel for Σ. The box-plots present summaries from thirty different observation realizations. (b) shows the same joint estimation but for the DW system. The right upper panel shows the results for θ, while the right lower panel for Σ. All results were obtained with the *LP* method (3'rd order) and fixed observation density to two per time unit ($N_{obs} = 2$).

5.3. Special case: stochastic Lorenz '96 (40D)

In this section the application of the new LP variational approximation framework is illustrated in a forty dimensional system, namely the Lorenz '96 (L40D). An example of this system is given in Figure 32(a), where are shown all forty dimensions for a time period of ten units $T = [0, 10]$. The drift function of the system is given by:

$$\mathbf{f}_{\mathsf{L40D}}(\mathbf{X}_t;\theta) = \begin{pmatrix} (x_t^2 - x_t^{39})x_t^{40} - x_t^1 + \theta \\ (x_t^3 - x_t^{40})x_t^1 - x_t^2 + \theta \\ \vdots \\ (x_t^1 - x_t^{38})x_t^{39} - x_t^{40} + \theta \end{pmatrix}, \quad \theta \in \Re\,. \tag{26}$$

This drift function consists of forty equations:

$$f(x_t^i) = (x_t^{i+1} - x_t^{i-2})x_t^{i-1} - x_t^i + \theta\ ,$$

where $i \in \{1, 2, \ldots, 40\}$, with *cyclic indices* and $\theta \in \Re$ is the forcing (drift) parameter. These equations simulate advection, damping and forcing of some atmospheric variable x^i, therefore it can be seen as a minimalistic weather model (Lorenz and Emanuel [41]).

Figure 32(b), shows the approximate marginal mean $\mathbf{m}_t$ and variance $\mathbf{S}_t$, of three selected dimensions from the $L40D$ system. The mean paths are reasonably smooth and the variances are broad enough to

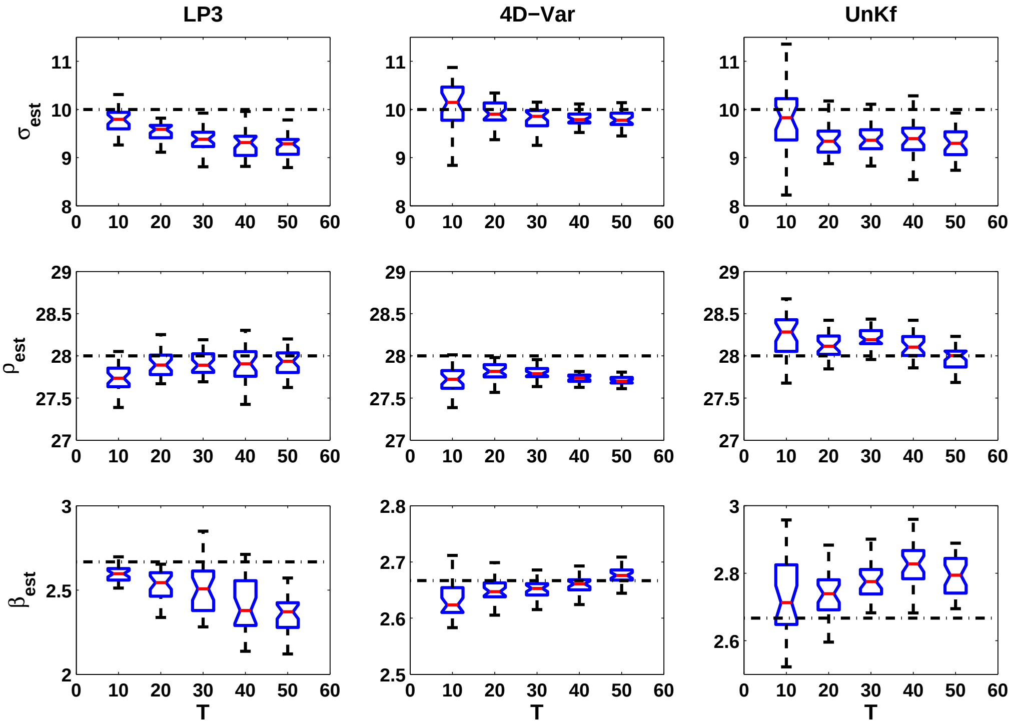


Figure 29: *Drift (conditional) estimation:* This plot compares the increasing domain asymptotic results (fixed observation density), when estimating the *L3D* drift parameter vector $\boldsymbol{\theta}$. The summary results when seen horizontally compare the same drift parameter with different estimation methods, while vertically the results are presented for the same estimation method and all three parameters (σ, ρ and β). The methods tested, from left to right are the *LP* algorithm (3'rd order), the *(full) weak-constrained 4D-Var* and the *dual UnKF* accordingly. In all sub-plots the horizontal dashed lines indicate the true values of the drift parameters that generated the history sample. Where possible the *y-axis* was kept the same for all plots comparing the same parameter to make the comparison easier.

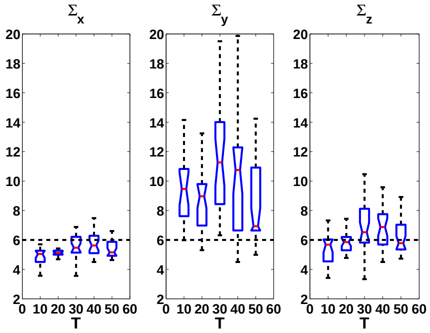


Figure 30: *Noise (conditional) estimation:* Summary results (box-plots) when estimating jointly the noise coefficients Σ_x, Σ_y and Σ_z, of the *L3D* system. The results were obtained with the *LP* method (3'rd order) and presented as functions of increasing time domain, keeping the observation density fixed ($N_{obs} = 18$).

enclose the observations. Similar results were also obtained for the other dimensions of the system.

Finally the new approach was compared against the original VGPA algorithm, in producing conditional

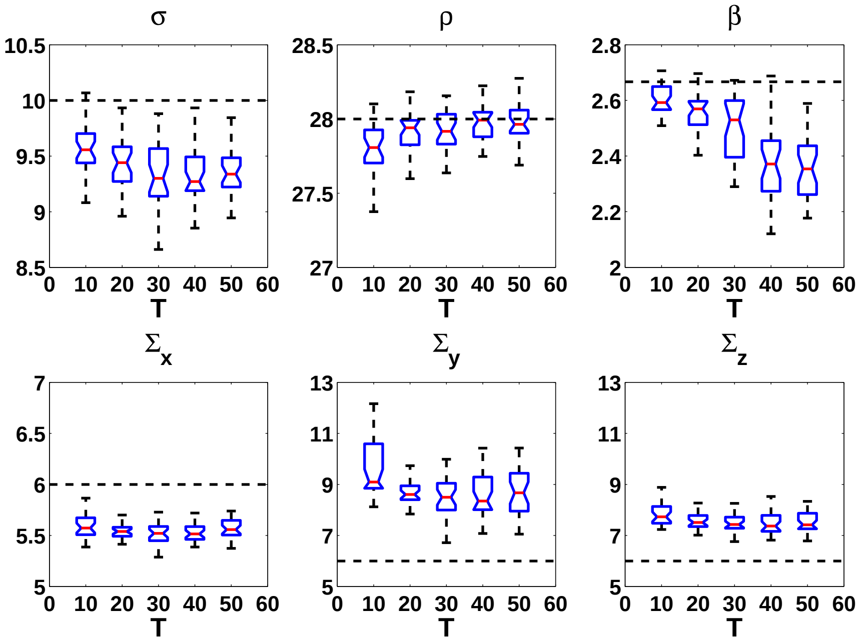
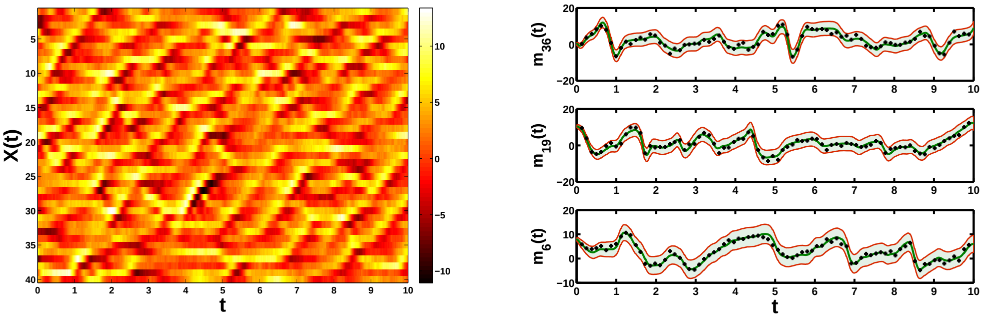


Figure 31: *Joint estimation:* Summary results (box-plots) when estimating jointly the drift parameters σ, ρ and β (upper three panels), and the system noise coefficients Σ_x, Σ_y and Σ_z (lower three panels), of the $L3D$ system. The results were obtained with the LP method (3'rd order) and presented as functions of increasing time domain, keeping the observation density fixed ($N_{obs} = 10$).

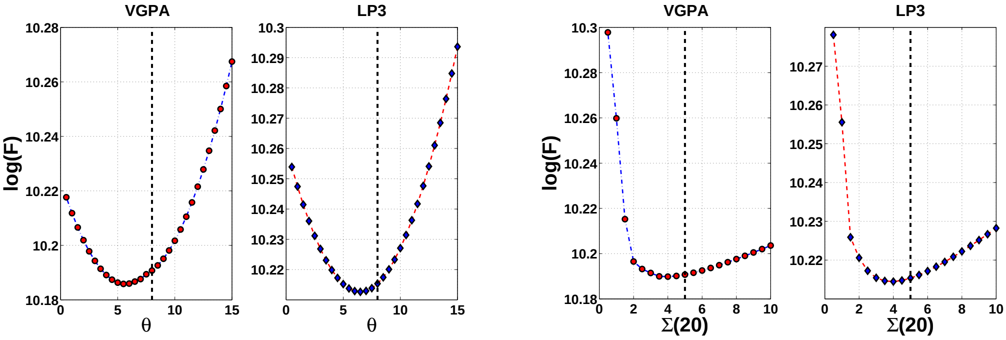


(a) $L40D$ sample path

(b) marginal $\mathbf{m}_t$ and $\mathbf{S}_t$

Figure 32: *Lorenz 40D:* In (a) all forty dimensions (top to bottom) of a ten units time-window ($T = [0, 10]$), of the stochastic Lorenz $40D$ system, used for the experiments. (b) presents three examples (3'rd, 19'th and 36'th dimension) of the marginal means (solid green line) and variances (shaded light green area) obtained with the LP algorithm (3'rd order), at convergence. The crosses indicate the noisy observations. Similar result were also acquired for the remaining dimensions.

profiles for the forcing (drift) parameter $\boldsymbol{\theta}$ (see Figure 33(a)) and system noise coefficients $\boldsymbol{\Sigma}$ (see Figure 33(b), for the system noise in the 20'th dimension). Both algorithms produce smooth profiles, with the new approach identifying the minimum slightly better. However, more important is that these results were obtained by achieving a significant reduction of 67.6% in optimisation space. For this example the observation



(a) $\boldsymbol{\theta}$ profile

(b) $\boldsymbol{\Sigma}$ profile

Figure 33: *Marginal (approximate) profiles:* In (a) the approximate marginal profile log likelihood of the drift parameter θ, obtained with the original $VGPA$ algorithm (left panel, red circles) is compared against the one obtained with the LP algorithm with 3'rd order polynomials (right panel, blue diamonds). In this example the system noise covariance matrix $\boldsymbol{\Sigma}$ is fixed to its true value. (b) presents similar results but for the conditional estimation of the system noise on the 20'th dimension, assuming the drift is known. Similar profiles were also generated for other dimensions. In all sub-plots the vertical dashed lines represent the true values of the parameters that generated the data.

noise variance was set to $\mathbf{R} = 1.0$, with eight observations per time unit (hence $J = 81$) and third order polynomials (hence $M = 3$), we have to infer $L_{total} = 531,360$ variables, comparing to $N_{total} = 1,640,000$, of the original VGPA. Joint estimation of the drift and diffusion coefficients for this system is also possible and produces similar results, albeit at a slightly higher computational cost.

6. Conclusions and discussion

This paper has presented an alternative parametrisation of the VGPA algorithm [4, 5] for Bayesian approximate inference in partially observed diffusions with additive noise. The general case of arbitrary state dependent diffusions (multiplicative noise) is not covered in this work. This is related to limitations that follow the original variational framework proposed in [4]. To be more specific, the VGPA algorithm requires the true and the approximating posterior processes (p_t and q_t respectively) to share the same diffusion coefficient, otherwise the bound on the true negative log marginal likelihood would not be finite. In other words the integral 31, as shown in Appendix A, goes to infinity in the limiting case of $\delta t \rightarrow 0$. However there is a cure to this problem and we are currently working towards a version of the variational algorithm that will overcome this limitation. The main idea is to work entirely in discrete time, therefore instead of computing integrals that go to infinity one will have to work with sums (possibly large) but still bounded to a finite number. This will allow us to relax the constraint of using the same diffusion coefficient for both processes p_t and q_t and will enable the treatment of state dependent diffusions. We also note that for the class of diffusion processes that can be mapped into an additive noise process Kloeden and Platen

[35], the VGPA methods will work effectively. Finally, in some cases it might be possible to capture much of the structure of the model discrepancy / model error in the drift (deterministic) part of the dynamic model, for which our methods have no limitations, leaving the residual discrepancy well approximated by additive noise. This is an area also which should be further explored.

This new approach uses local polynomials to approximate the variational parameters $\mathbf{A}_t$ and $\mathbf{b}_t$ of the linear drift approximation (Eq. 6) to control the complexity of the algorithm and reduce the number of variables need to be optimized. The LP algorithm is validated on a range of different systems to test its convergence behaviour w.r.t. the original VGPA and shows remarkable stability. In most of the examples it requires 3'rd order polynomials to match the original algorithm, although the order is likely to increase as the observations become more sparse (i.e. the time between observations increases).

Despite the notable reduction in optimized variables the LP approach does not produce similar results in computational time. This is mostly because the new gradients of the cost function (Equation 13) w.r.t. the coefficients of the polynomial approximations, have to be computed separately in each sub-interval where each polynomial is defined. In our implementation priority was not given to the computational cost, hence a simple serial approach was chosen. However, a parallel implementation in which the necessary gradients are computed simultaneously is straightforward and could reduce dramatically the execution time, especially for treating long time windows. Another advantage with the LP framework is that different classes of polynomials can be used. In this work we also experiment with different classes of polynomials, mostly orthogonal, such as Chebyshev and Legendre however the results were not significantly different in the systems tested here hence were omitted.

The new LP algorithm can be used to construct, computationally cheap, discrete approximations to the posterior distribution of the (hyper-) parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ (Section 5) matching the results of the HMC sampling rather well, in the examples tested. In exploring the infill and increasing domain behaviour on estimating the parameters of the OU, DW and L3D, all methods show biases and the response was different over the range of the systems. The methods are largely comparable with the UnKF being less stable and slightly more biased. LP and weak constraint 4D-Var are more comparable (since both provide smoothing solutions to the inference problem) but there was no clear preference from one another, except in the case of estimating the system noise parameters $\boldsymbol{\Sigma}$. In this case both 4D-Var and UnKF failed to provide satisfactory results, giving the LP a clear advantage. However, as discussed in DelSole and Yang [8], estimation of drift (deterministic) and diffusion (stochastic) parameters are fundamentally different problems. As shown in Equations (7) and (8), the system noise coefficient $\boldsymbol{\Sigma}$ directly affects the marginal variance. This means that when one conditions the estimation of this parameter only on the, rather smooth, mean path (or the mode in the 4D-Var case) all the information on the roughness of the true trajectory is lost. Therefore, the UnKF and the 4D-Var method were unable to estimate this important parameter accurately. On the contrary, the VGPA approximations base their estimation on a bound to the complete marginal likelihood, as a function of both drift and diffusion parameters, allowing for joint estimation. A particularly difficult case is the noise estimation in the L3D system where the process has to be observed very frequently. We believe that this is related to the chaotic behaviour of the L3D system which makes identification of noise using infrequent observations very challenging.

Comparing the results on the two asymptotic regimes reveals that *increasing domain* is more promising than *infill* and suggests that in order to identify a model parameter, is better to observe an event over a large period of time, rather than observe it more densely in a short period of time. Moreover, another appealing asymptotic behaviour, that is not covered here but is worth exploring, is with the system noise $\mathbf{\Sigma}$ fixed and the observation noise going to zero ($\mathbf{R} \to 0$). An interesting question that is raised is how the parameter estimates are affected if the process is not observed uniformly (at equidistant times), as was the case here, but rather with different densities over different periods of time. An example, on a DW trajectory, would be the estimation of the system noise $\mathbf{\Sigma}$ by having more frequent observations around the transition time than the rest of the sample path.

We believe the range of systems on which these methods have been applied (OU, DW, L3D, L40D) show their generic utility. The systems cover frequently used exemplars in synthetic data assimilation experiments, and include non-linear systems that are often used as surrogates for the sorts of models used in operational weather and climate modelling. The nature of the non-linear interactions in the systems is similar to the interactions seen in more realistic models. The range of observation densities chosen is comparable to those in realistic settings. We note that the length of assimilation window considered in this work is longer than is typical in data assimilation studies, this being related to our aim of learning about model parameters. It seems likely that the results we find in this paper would generalise to more operational settings, although considerable work remains to be done to address the computational cost of the VGPA methods.

Although the application of our variational approach to the forty dimensional Lorenz '96 system (L40D) is very encouraging, there is still an open question on how we can apply this algorithm to very high dimensional models (such as those used for numerical weather prediction). We believe that the LP approximation is a step towards that direction. In most of the examples presented here the computational resources were reduced more than 60% (in terms of optimizing variables) comparing to the original VGPA. By imposing further assumptions on the Gaussian process approximation (e.g. by defining a special class of linear drift functions) it is possible to control the complexity of the posterior variational approximation and reduce the number of variables even further. Finally, a drawback of our algorithm is that it remains quite complex and is our intention to provide more guidance on the usage of the VGPA based algorithms in the future.

Acknowledgements

The authors appreciate the two anonymous reviewers for their valuable comments and suggestions that helped to improve the article. This work has been funded by EPSRC as part of the Variational Inference for Stochastic Dynamic Environmental Models (VISDEM) project (EP/C005848/1).

A. Variational Free Energy

As shown earlier in Section 2.2, the definition of the so called "*variational free energy*", is given by Equation (5). The derivation of the free energy leads to the following expressions:

$$\begin{aligned}
\mathcal{F}(q(\mathbf{X}),\boldsymbol{\theta},\boldsymbol{\Sigma}) &= -\left\langle \ln\frac{p(\mathbf{Y},\mathbf{X}|\boldsymbol{\theta},\boldsymbol{\Sigma})}{q(\mathbf{X}|\boldsymbol{\theta},\boldsymbol{\Sigma})}\right\rangle_{q(\mathbf{X})} &(27)\\
&= -\int_{t0}^{tf} q(\mathbf{X}_t)\ln\frac{p(\mathbf{Y}_{t_k},\mathbf{X}_t)}{q(\mathbf{X}_t)}d\mathbf{X}_t &(28)\\
&= \int_{t0}^{tf} q(\mathbf{X}_t)\ln\frac{q(\mathbf{X}_t)}{p(\mathbf{Y}_{t_k},\mathbf{X}_t)}d\mathbf{X}_t &(29)\\
&= \underbrace{\int_{t0}^{tf} q(\mathbf{X}_t)\ln\frac{q(\mathbf{X}_t)}{p(\mathbf{X}_t)}d\mathbf{X}_t}_{(\mathbf{I1})} - \underbrace{\int_{t0}^{tf} q(\mathbf{X}_t)\ln p(\mathbf{Y}_{t_k}|\mathbf{X}_t)d\mathbf{X}_t}_{(\mathbf{I2})}\ , &(30)
\end{aligned}$$

where $\mathbf{X}=\{\mathbf{X}_t, t_0\leq t\leq t_f\}$ is the diffusion process, $\mathbf{Y}=\{\mathbf{Y}_{t_k}\}_{k=1}^{K}$ the observations and the conditioning on the (hyper) parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ has been omitted for notational simplicity. Solving the integrals $I1$ and $I2$, results in the following expressions:

A.1. Energy term from the SDE.

Using of the fact that both processes p and q are Markovian yields:

$$I1 = \frac{1}{2}\int_{t_0}^{t_f}\left\langle(\mathbf{f}(t,\mathbf{X}_t)-\mathbf{g}_L(t,\mathbf{X}_t))^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{f}(t,\mathbf{X}_t)-\mathbf{g}_L(t,\mathbf{X}_t))\right\rangle_{q(\mathbf{X}_t)}\ dt + \mathrm{KL}\left[q(\mathbf{X}_{t0})\|p(\mathbf{X}_{t0})\right]\ , \qquad (31)$$

where $\mathbf{f}(t, \mathbf{X}_t) \in \Re^D$ is the drift function, $\mathbf{g}_L(t, \mathbf{X}_t) \in \Re^D$ is the linear approximation, $< \cdot >_{q(\mathbf{X}_t)}$ denotes the expectation with respect to measure $q(\mathbf{X}_t)$ and $\mathrm{KL}[q(\mathbf{X}_{t0})\|p(\mathbf{X}_{t0})]$ is the KL divergence at initial time $\mathbf{X}_{t=t0}$.

A.2. Energy term from the observations (likelihood).

Assuming that the measurements are i.i.d. with zero mean and covariance matrix $\boldsymbol{R}$, we have:

$$I2 = -\frac{1}{2}\int_{t_0}^{t_f} \left\langle (\mathbf{Y}_t - h(\mathbf{X}_t))^\top \mathbf{R}^{-1} (\mathbf{Y}_t - h(\mathbf{X}_t)) \right\rangle_{q(\mathbf{X}_t)} \sum_{k=1}^{K} \delta(t - t_k)\ dt + \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{R}|\ , \tag{32}$$

where $|\mathbf{R}|$ is the determinant of matrix $\mathbf{R}$ (observation noise covariance) and $\delta(t)$ is Dirac's delta function, which is added due to the discrete time nature of the actual observations. For a complete derivation of the above equations we refer to [60].

B. Weak constraint 4D-Var cost function

In a Bayesian framework, if one is interested in estimating the system states $\boldsymbol{X}$ as well as the model parameters[6] $\boldsymbol{\Theta}$, then is interested in the joint posterior distribution of the states and the parameters, given the observations (i.e. $p(\boldsymbol{X}, \boldsymbol{\Theta}|\boldsymbol{Y})$). Via Bayes rule this posterior is given by:

$$\begin{aligned} p(\boldsymbol{X}, \boldsymbol{\Theta}|\boldsymbol{Y}) &= \frac{p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta})p(\boldsymbol{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})}{p(\boldsymbol{Y})} \\ &\propto p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta})p(\boldsymbol{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}) \end{aligned} \tag{33}$$

where $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta})$ is the likelihood of the observations given the current state of the system and the (hyper-) parameters, $p(\boldsymbol{X}|\boldsymbol{\Theta})$ is the prior distribution over the system states, $p(\boldsymbol{\Theta})$ is the prior over the (hyper-) parameters and $p(\boldsymbol{Y})$ is the marginal likelihood.

Having discretised the continuous time sample path $\boldsymbol{X} = \{\mathbf{X}_t, t_0 \leq t \leq t_f\}$, using the Euler-Maruyama method (see Section 2), one has to compute the following posterior distribution:

$$p(\boldsymbol{X}_{0:N}, \boldsymbol{\Theta}|\boldsymbol{Y}_{1:K}) \propto \underbrace{p(\boldsymbol{Y}_{1:K}|\boldsymbol{X}_{0:N})}_{B1}\underbrace{p(\boldsymbol{X}_{0:N})}_{B2}\underbrace{p(\boldsymbol{\Theta})}_{B3} \tag{34}$$

[6] Within our framework it includes all the parameters in the drift and the system noise covariance matrix (i.e. $\boldsymbol{\Theta} = \{\boldsymbol{\theta},\boldsymbol{\Sigma}\}$).

where the dependencies on the parameters have been omitted for simplicity.

B.1. Likelihood of the observations

Assuming that the measurements are i.i.d. with zero mean and covariance matrix $\boldsymbol{R}$, we have:

$$\begin{aligned} p(\boldsymbol{Y}_{1:K}|\boldsymbol{X}_{0:N}) &= \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{Y}_k - \boldsymbol{X}_{t_k}|\boldsymbol{R}) \\ &= \prod_{k=1}^{K} (2\pi)^{-D/2}|\boldsymbol{R}|^{-1/2} \exp\{-0.5(\boldsymbol{Y}_k - \boldsymbol{X}_{t_k})^\top \boldsymbol{R}^{-1}(\boldsymbol{Y}_k - \boldsymbol{X}_{t_k})\} \\ &= \left[(2\pi)^{-D/2}|\boldsymbol{R}|^{-1/2}\right]^K \exp\{-0.5\sum_{k=1}^{K}(\boldsymbol{Y}_k - \boldsymbol{X}_{t_k})^\top \boldsymbol{R}^{-1}(\boldsymbol{Y}_k - \boldsymbol{X}_{t_k})\} \qquad (35) \end{aligned}$$

where all the assumptions about the state and observation vector dimensions are the same as introduced in Section 2.

B.2. Prior over the states

Using the assumption that the process is Markov, we have:

$$\begin{aligned} p(\boldsymbol{X}_{0:N}) &= p(\boldsymbol{X}_0) \prod_{k=0}^{N-1} p(\boldsymbol{X}_{k+1}|\boldsymbol{X}_k) \qquad (36) \\ &= p(\boldsymbol{X}_0) \prod_{k=0}^{N-1} \mathcal{N}(\boldsymbol{X}_{k+1}|\boldsymbol{X}_k + \mathbf{f}(\boldsymbol{X}_k)\delta t, \boldsymbol{\Sigma}\delta t) \qquad (37) \\ &= p(\boldsymbol{X}_0) \prod_{k=0}^{N-1} (2\pi)^{-D/2}|\boldsymbol{\Sigma}\delta t|^{-1/2} \exp\{-0.5(\delta\boldsymbol{X}_{k+1} - \mathbf{f}(\boldsymbol{X}_k)\delta t)^\top (\boldsymbol{\Sigma}\delta t)^{-1}(\delta\boldsymbol{X}_{k+1} - \mathbf{f}(\boldsymbol{X}_k)\delta t)\} \qquad (38) \\ &= p(\boldsymbol{X}_0) \left[(2\pi)^{-D/2}|\boldsymbol{\Sigma}\delta t|^{-1/2}\right]^N \exp\{-0.5\delta t \sum_{k=0}^{N-1} (\frac{\delta\boldsymbol{X}_{k+1}}{\delta t} - \mathbf{f}(\boldsymbol{X}_k))^\top \boldsymbol{\Sigma}^{-1}(\frac{\delta\boldsymbol{X}_{k+1}}{\delta t} - \mathbf{f}(\boldsymbol{X}_k))\}, \qquad (39) \end{aligned}$$

where $\delta\boldsymbol{X}_{k+1} = \boldsymbol{X}_{k+1} - \boldsymbol{X}_k$ and $\delta t = t_{k+1} - t_k$. For the initial state $\boldsymbol{X}_0$, we either assume that it is given by fixed values (i.e. $\boldsymbol{X}_0 = x_0$), or that we know its distribution. In this case we chose an initial state that is normally distributed such as $\boldsymbol{X}_0 \sim \mathcal{N}(\boldsymbol{\tau_0}, \boldsymbol{\Lambda_0})$. Notice also the unusual scaling of the system noise coefficient $\boldsymbol{\Sigma}$, with the time increment δt. This comes from the discrete version of the SDE (see Eq.20), where the scaling is necessary to achieve the limit of the diffusion process as $\delta t \to 0$.

B.3. Prior over the parameters

For this prior density we assume that the parameters have no dependencies between them, hence we can write their joint density as the product of their marginal densities:

$$\begin{aligned} p(\boldsymbol{\Theta}) &= p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \\ &= p(\boldsymbol{\theta}) p(\boldsymbol{\Sigma}) \ , \end{aligned} \tag{40}$$

where $p(\boldsymbol{\theta})$ is the prior marginal distribution of the drift parameters and $p(\boldsymbol{\Sigma})$ is the same but for the system noise coefficient. We do not extend any derivation here because these densities can be parametrized with any distribution of choice. In our framework we use the same prior distributions as in the HMC and the variational framework. That is $p(\boldsymbol{\theta}) = \mathcal{G}(\alpha, \beta)$ and $p(\boldsymbol{\Sigma}) = \mathcal{G}^{-1}(a, b)$.

B.4. $\mathcal{J}_{cost}$ - Total cost function

It is common practice in optimisation when one wants to find the minimum (or maximum), of a cost function to look for the minimum (or maximum) of the logarithm of the cost function (due to the monotonicity of the logarithmic function). Hence instead of maximizing the posterior $p(\boldsymbol{X}_{0:N}, \boldsymbol{\Theta} | \boldsymbol{Y}_{1:M})$, we can minimize the negative $\ln p(\boldsymbol{X}_{0:N}, \boldsymbol{\Theta} | \boldsymbol{Y}_{1:M})$, which has some nice characteristics. Therefore, the complete cost function is given by:

$$\begin{aligned} \mathcal{J}_{cost} = \quad & - \quad \underbrace{\ln p(\boldsymbol{X}_0)}_{\mathcal{J}_{X_0}} \ + \ \underbrace{0.5\delta t \sum_{k=0}^{N-1} \left(\frac{\delta \boldsymbol{X}_{k+1}}{\delta t} - \mathbf{f}(\boldsymbol{X}_k) \right)^\top \boldsymbol{\Sigma}^{-1} \left(\frac{\delta \boldsymbol{X}_{k+1}}{\delta t} - \mathbf{f}(\boldsymbol{X}_k) \right)}_{\mathcal{J}_f} \\ & + \quad \underbrace{0.5 \sum_{k=1}^{K} (\boldsymbol{Y}_k - \boldsymbol{X}_{t_k})^\top \boldsymbol{R}^{-1} (\boldsymbol{Y}_k - \boldsymbol{X}_{t_k})}_{\mathcal{J}_{obs}} \underbrace{- \ln p(\boldsymbol{\theta}) - \ln p(\boldsymbol{\Sigma})}_{\mathcal{J}_{hp}} \\ & + \quad \underbrace{0.5 \left[K \ln |\boldsymbol{R}| + N \ln |\boldsymbol{\Sigma} \delta t| + KND \ln(2\pi) \right]}_{\mathcal{C}} , \end{aligned} \tag{41}$$

where $K > 0$ is the total number of observations, $N > 0$ is the number of the discrete time states and $D > 0$ is dimensions of the system states and observations.

References

[1] Alexander, F. J., Eyink, G., Restrepo, J., 2005. Accelerated Monte Carlo for optimal estimation of time-series. Journal of Statistical Physics 119, 1331–1345.

[2] Annan, J. D., Hargreaves, J. C., Edwards, N. R., Marsh, R., 2005. Parameter estimation in an intermediate complexity earth system model using an ensemble Kalman filter. Ocean Modelling 8, 135–154.

[3] Apte, A., Hairer, M., Stuart, A., Voss, J., 2007. Sampling the posterior: An approach to non-Gaussian data assimilation. Physica D 230, 50–64.

[4] Archambeau, C., Cornford, D., Opper, M., Shawe-Taylor, J., 2007. Gaussian Process Approximations of Stochastic Differential Equations. In: Journal of Machine Learning Research, Workshop and Conference Proceedings. Vol. 1. pp. 1–16.

[5] Archambeau, C., Opper, M., Shen, Y., Cornford, D., Shawe-Taylor, J., 2008. Variational Inference for Diffusion Processes. In: Platt, C., Koller, D., Singer, Y., Roweis, S. (Eds.), Annual Conference on Neural Information Processing Systems (NIPS). Vol. 20. The MIT Press, pp. 17–24.

[6] Beskos, A., Papaspiliopoulos, O., Roberts, G., 2006. Retrospective exact simulation of diffusion sample paths with applications. Bernoulli 12 (6), 1077–1098.

[7] Beskos, A., Papaspiliopoulos, O., Roberts, G. O., Fearnhead, P., 2006. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. Journal of Royal Statistical Society 68 (3), 333–382.

[8] DelSole, T., Yang, X., 2010. State and Parameter Estimation in Stochastic Dynamical Models. Physica D 239, 1781–1788.

[9] Dembo, A., Zeitouni, O., 1986. Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm. Stochastic Processes and their Applications 23, 91–113.

[10] Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39, 1–38.

[11] Derber, J. C., 1989. A variational continuous assimilation technique. Monthly Weather Review 117, 2437–2446.

[12] Dimet, F. L., Talagrand, O., 1986. Variational algorithms for analysis and assimilation of meteorological observations: Theorical aspects. Tellus 38(A), 97–110.

[13] Duane, S., Kennedy, A. D., Pendeleton, B. J., Roweth, D., September 1987. Hybrid Monte Carlo. Physics Letters B 195 (2), 216–222.

[14] Durham, G. B., Gallant, A. R., 2002. Numerical techniques for maximum likelihood estimation of continuous time diffusion processes. Journal of Business and Economic Statistics 20, 297–338.

[15] Elerian, O., Chib, S., Shephard, N., 2001. Likelihood inference for discretely observed non-linear diffusions. Econometrica 69, 959–993.

[16] Eraker, B., 2001. MCMC analysis of diffusion models with application to finance. Journal of Business and Economic Statistics 19, 177–191.

[17] Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dynamics 53, 343–367.

[18] Evensen, G., van Leeuwen, P. J., 2000. An Ensemble Kalman Smoother for Non-linear Dynamics. Monthly Weather Review 128, 1852–1867.

[19] Eyink, G., Restrepo, J. M., Alexander, F. J., 2004. A mean field approximation in data assimilation for non-linear dynamics. Physica D 194, 347–368.

[20] Fearnhead, P., Papaspiliopoulos, O., Roberts, G. O., 2008. Particle filters for partially observed diffusions. Journal of the Royal Statistical Society 70 (B), 755–777.

[21] Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs Distributions and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–74.

[22] Golightly, A., Wilkinson, D. J., 2006. Bayesian Sequential Inference for Non-linear Multivariate Diffusions. Statistics and Computing 16, 323–338.

[23] Golub, G. H., van Loan, C. F., 1996. Matrix Computations. The Johns Hopkins University Press.

[24] Gove, J. H., Hollinger, D. Y., 2006. Application of a dual unscented Kalman filter for simultaneous state and parameter estimation in problems of surface-atmosphere exchange. Journal of Geophysical Research 111 (D08S07), DOI:10.1029/2005JD006021.

[25] Hansen, J. N., Penland, C., 2007. On stochastic parameter estimation using data assimilation. Physica D 230, 88–98.

[26] Higham, D. J., 2001. An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations. Society for Industrial and Applied Mathematics 43, 525–546.

[27] Honerkamp, J., 1993. Stochastic Dynamical Systems: Concepts, Numerical Methods, Data Analysis. Wiley - VCH.

[28] Jaakkola, T., 2001. Advanced Mean Field Methods: Theory and Practise. The MIT Press, Ch. Tutorial on Variational Approximation methods.

[29] Julier, S., Uhlmann, J., Durrant-Whyte, H. F., March 2000. A New Method for Non-linear Transformation of Means and Covariances in Filters and Estimators. IEEE Transactions on Automated Control, Technical Notes and Correspondence 45 (3), 477–482, accepted for publication as technical note.

[30] Kalman, R. E., 1960. A new approach to linear filter and prediction problems. Transactions of the ASME - Journal of Basic Engineering 82 (Series D), 35–45.

[31] Kalman, R. E., Bucy, R. S., 1961. New results in linear filtering and prediction theory. Journal of Basic Engineering 83 (Series D), 95–108.

[32] Kalnay, E., 2003. Atmospheric Modelling, Data Assimilation and Predictability. Cambridge University Press.

[33] Kitagawa, G., 1987. Non-Gaussian state space modelling of non-stationary time series. Journal of the American Statistical Association, Theory and Methods 82, 1032–1041.

[34] Kivman, G. A., 2003. Sequential parameter estimation for stochastic systems. Non-linear Processes in Geophysics 10, 253–259.

[35] Kloeden, P. E., Platen, E., 1999. Numerical Solution of Stochastic Differential Equations, 3rd Edition. Springer, Applications of Mathematics.

[36] Kullback, S., Leibler, R. A., 1951. On information and sufficiency. Annal of Mathematical Statistics 22, 79–86.

[37] Kushner, H. J., 1962. On the differential equations satisfied by conditional probability densities of markov processes, with applications. SIAM Control A 2, 106–119.

[38] Kushner, H. J., 1967. Approximation to optimal non-linear filters. IEEE Trans. Auto. Control 12, 546–556.

[39] Kushner, H. J., 1967. Dynamical equations for optimal non-linear filtering. Journal of Differential Equations 3, 179–190.

[40] Lorenz, E. N., 1963. Deterministic non-periodic flow. Journal of Atmospheric Science 20, 130–141.

[41] Lorenz, E. N., Emanuel, K. A., February 1998. Optimal Sites for Supplementary Weather Observations: Simulations with a Small Model. Journal of the Atmospheric Science 55, 399–414.

[42] Maybeck, P. S., 1979. Stochastic models, estimation and control, (Volume 1). Academic Press.

[43] Miller, R. N., 2007. Topics in data assimilation: Stochastic Processes. Physica D 230, 17–26.

[44] Miller, R. N., Ghil, M., Gauthiez, F., April 1994. Advanced data assimilation in strongly non-linear dynamical systems. Journal of the Atmospheric Sciences 51 (8), 1037–1056.

[45] Nabney, I. T., 2002. NETLAB: Algorithms for pattern recognition. Advances in Pattern Recognition. Springer.

[46] Navon, I. M., 1997. Practical and theoretical aspects of adjoint parameter estimation and identifiability in meteorology and oceanography. Dyn. Atmos. Ocean 27, 55–79.

[47] Neal, R. M., September 1993. Probabilistic inference using Markov chain Monte Carlo methods. Tech. Rep. CRG-TR-93-1, Department of Computer Science, University of Toronto.

[48] Papaspiliopoulos, O., Roberts, G., 2008. Retrospective MCMC methods for Dirichlet process hierarchical models. Biometrika 95, 169–186.

[49] Pardoux, E., 1982. Equations du filtrage non lineaire de la prediction et du lissage. Stochastics 6, 193–231.

[50] Pedersen, A. R., 1995. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. Scandinavian Journal of Statistics 22, 55–71.

[51] Rasmussen, C. E., Williams, C. K. I., 2006. Gaussian Processes for Machine Learning. The MIT press, Cambridge.

[52] Roberts, G. O., Stramer, O., 2001. On inference for partially observed non-linear diffusion models using the Metropolis-Hastings algorithm. Biometrika 88 (3), 603–621.

[53] Sasaki, Y., 1970. Some basic formalisms in numerical variational analysis. Monthly Weather Review 98, 875–883.

[54] Sorensen, H., 2004. Parametric inference for diffusion processes observed at discrete points in time: A survey. International Statistics Review 72 (3), 337–354.

[55] Stratonovich, R. L., 1960. Conditional Markov Processes. Theory of Probability and its Application 5, 156–178.

[56] Stuart, A. M., Voss, J., Wiberg, P., 2004. Conditional path sampling of SDEs and the Langevin MCMC method. Communications in Mathematical Science 2, 685–697.

[57] Tremolet, Y., 2006. Accounting for an imperfect model in 4D-Var. Quarterly Journal of the Royal Meteorological Society 132 (621), 2483–2504.

[58] Uhlenbeck, G. E., Ornstein, L. S., 1930. On the theory of Brownian motion. Physical Review 36, 823–841.

[59] Vrettas, M. D., Cornford, D., Opper, M., Shen, Y., 2010. A new variational radial basis function approximation for inference in multivariate diffusions. Neurocomputing 73, 1186–1198.

[60] Vrettas, M. D., Shen, Y., Cornford, D., March 2008. Derivations of Variational Gaussian Process Approximation Framework. Tech. Rep. NCRG/2008/002, Neural Computing Research Group (NCRG), Aston University, Birmingham, B4 7ET, UK.

[61] Wan, E. A., van der Merwe, R., 2000. The unscented Kalman filter for non-linear estimation. In: IEEE Symposium.

[62] Wan, E. A., van der Merwe, R., Nelson, A. T., 2000. Dual estimation and the unscented transformation. In: Neural Information Processing Systems (NIPS).

[63] Zupanski, D., 1996. A general weak constraint applicable to operational 4D-VAR data assimilation systems. Monthly Weather Review 125, 2274–2292.