Available online at www.sciencedirect.com
ScienceDirect
Physica D 230 (2007) 37–49
PHYSICA D
www.elsevier.com/locate/physd

Approximate importance sampling Monte Carlo for data assimilation

L. Mark Berliner[a,*], Christopher K. Wikle[b]

[a] Department of Statistics, Ohio State University, OH, United States
[b] Department of Statistics, University of Missouri, MO, United States

Available online 11 October 2006
Communicated by C.K.R.T. Jones

Abstract

Importance sampling Monte Carlo offers powerful approaches to approximating Bayesian updating in sequential problems. Specific classes of such approaches are known as particle filters. These procedures rely on the simulation of samples or ensembles of the unknown quantities and the calculation of associated weights for the ensemble members. As time evolves and/or when applied in high-dimensional settings, such as those of interest in many data assimilation problems, these weights typically display undesirable features. The key difficulty involves a collapse toward approximate distributions concentrating virtually all of their probability on an implausibly few ensemble members.

After reviewing ensembling, Monte Carlo, importance sampling and particle filters, we present some approximations intended to moderate the problem of collapsing weights. The motivations for these suggestions are combinations of (i) the idea that key dynamical behavior in many systems actually takes place on a low dimensional manifold, and (ii) notions of statistical dimension reduction. We illustrate our suggestions in a problem of inference for ocean surface winds and atmospheric pressure. Real observational data are used.

Keywords: Bayesian analysis; Dimension reduction; Dynamics; Ensemble forecasting; Geostrophy; Hierarchical models; Importance sampling; Particle filter; Sufficient statistic

1. Introduction

Data assimilation (DA) involves the use of a scientifically-based numerical prediction model driven by observations. Consider a state process $\mathbf{X}$ and assume the following discrete-time model

$$\mathbf{X}_t = \mathcal{M}(\mathbf{X}_{t-1}) + \boldsymbol{\epsilon}_t, \tag{1}$$

where $\boldsymbol{\epsilon}_t$ is a vector of model errors. The function $\mathcal{M}$ is typically a numerical representation of a dynamical model developed from physical reasoning, usually represented via a system of partial differential equations.

We begin with a disclaimer. Efficient DA methods involve complicated interactions among the form and type of observations available, the numerical methods used to formulate and integrate the approximation, and the treatment of model error. Nevertheless, we will not delve into these issues. Rather, general propositions of the problem and general, rather than case-by-case, approaches are presented here.

To transition to probabilistic or statistical formulations of DA, consider a time series of state variables $\mathbf{X}_{0:T} = (\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_T)$. Employing a Markov assumption, we have that the joint density function of $\mathbf{X}_{0:T}$ is

$$p(\mathbf{x}_{0:T}) = p(\mathbf{x}_0) \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{2}$$

where p generically denotes a probability density function (PDF) of the indicated argument, while a vertical bar in a term such as $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ indicates a conditional PDF, in this case, for $\mathbf{X}_t$ given $\mathbf{X}_{t-1} = \mathbf{x}_{t-1}$. We follow conventional notation here in that capitalized letters, such as $\mathbf{X}$, refer to random quantities and corresponding lower-case versions refer to their possible values. Note that $p(\mathbf{x}_0)$ is assumed to be based on all data through $t = 0$.

Suppose observations $\mathbf{Y}_{1:T} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_T)$ are available. (Observing at every model time-step is neither standard, nor necessary, but we do not introduce extra notation to treat the

* Corresponding address: Department of Statistics, Ohio State University, 1958 Neil Ave., 43210 Columbus, OH, United States. Tel.: +1 614 2920291; fax: +1 614 2922096.
E-mail address: mb@stat.ohio-state.edu (L.M. Berliner).

0167-2789/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physd.2006.07.031

general case.) Assume that the observations (i) are conditionally independent given $\mathbf{X}_{0:T}$, and (ii) the conditional PDFs of the $\mathbf{Y}_t$ depend on $\mathbf{x}_{0:T}$ only through $\mathbf{x}_t$. These PDFs are denoted by $q(\mathbf{y}_t|\mathbf{x}_t)$.

Bayes' Theorem provides a solution to the DA problem (e.g., [11,14]). Namely, to update the *prior* model for $\mathbf{X}_{0:T}$ reflected in (2) in light of the observations, we compute the *posterior* PDF of $\mathbf{X}_{0:T}$:

$$p(\mathbf{x}_{0:T}|\mathbf{y}_{1:T}) \propto q(\mathbf{y}_{1:T}|\mathbf{x}_{0:T})p(\mathbf{x}_{0:T}) \tag{3}$$

$$\propto p(\mathbf{x}_0)\prod_{t=1}^{T} q(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1}). \tag{4}$$

The first expression is Bayes' Theorem; the second indicates simplifications based on the Markov and condition independence assumptions of the previous paragraph.

Extending the Markov assumption beyond time T, probability theory provides the solution to the prediction or forecasting problem. For example, to forecast $\mathbf{X}_{T+1}$ based on data through time T, we compute

$$p(\mathbf{x}_{T+1}|\mathbf{y}_{1:T}) = \int p(\mathbf{x}_{T+1}|\mathbf{x}_T)p(\mathbf{x}_T|\mathbf{y}_{1:T})\mathrm{d}\mathbf{x}_T, \tag{5}$$

where $p(\mathbf{x}_T|\mathbf{y}_{1:T})$ is obtained from (4). If we assume all inputs on the right-hand-side of (4) are correct and can complete the calculations, the DA problem is solved. Of course, these are two very large "If's".

1.1. Bayesian sequential updating

The term *smoothing* refers to the production of $p(\mathbf{x}_{0:T}|\mathbf{y}_{1:T})$ as given in (4), or at least the collection of marginal PDFs $p(\mathbf{x}_0|\mathbf{y}_{1:T})$, $p(\mathbf{x}_1|\mathbf{y}_{1:T})$, $\ldots$, $p(\mathbf{x}_T|\mathbf{y}_{1:T})$. The results of *filtering* are the PDFs $p(\mathbf{x}_1|\mathbf{y}_1)$, $p(\mathbf{x}_2|\mathbf{y}_{1:2})$, $\ldots$, $p(\mathbf{x}_T|\mathbf{y}_{1:T})$. In either case, as in (5), predictive PDFs for times beyond T can be obtained by integration.

Filtering is by nature sequential; the key is to develop $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ from $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$. Sequential smoothing moves from $p(\mathbf{x}_{0:t-1}|\mathbf{y}_{1:t-1})$ to $p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$.

Derivations of the following facts rely on basic probability theory. First, the joint PDF of two variables $p(x, y)$ can be written as

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x), \tag{6}$$

which yields Bayes' Theorem:

$$p(x|y) = p(y|x)p(x)/p(y). \tag{7}$$

Second, simply put, conditional PDFs are PDFs. For example, Bayes' Theorem holds for conditional PDFs:

$$p(x|y, z) = p(y|x, z)p(x|z)/p(y|z). \tag{8}$$

This is foundational in sequential Bayesian analysis. One can view (8) as a prescription for updating the distribution of x based on y having previously updated based on z.

Sequential filtering. From Bayes' Theorem, we have

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) = p(\mathbf{x}_t|\mathbf{y}_{1:t-1}, \mathbf{y}_t) \propto q(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1}). \tag{9}$$

Applying conditional independence of the observations over time, we have

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto q(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1}). \tag{10}$$

The quantity $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ is a predictive or forecasting PDF obtained via integration:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\mathrm{d}\mathbf{x}_{t-1}. \tag{11}$$

In the parlance of DA, (10) corresponds to the *analysis step* based on Bayes' Theorem, and relying on the *forecast step* (11).

Sequential smoothing. Using the conditional independence of the data and the Markov assumption, Bayes' Theorem yields

$$p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}) \propto q(\mathbf{y}_t|\mathbf{x}_t)q(\mathbf{y}_{1:t-1}|\mathbf{x}_{0:t-1})p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{0:t-1}). \tag{12}$$

Noting that

$$p(\mathbf{x}_{0:t-1}|\mathbf{y}_{1:t-1}) \propto q(\mathbf{y}_{1:t-1}|\mathbf{x}_{0:t-1})p(\mathbf{x}_{0:t-1}), \tag{13}$$

we have

$$p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}) \propto q(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{0:t-1}|\mathbf{y}_{1:t-1}). \tag{14}$$

In some settings, modelers may replace (1) by a deterministic model and act as if $\mathcal{M}$ is known, and hence, the only unknown is $\mathbf{X}_0$. It follows that $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ implies $p(\mathbf{x}_s|\mathbf{y}_{1:t})$ for all $s \geq 0$, though the calculations may be unpleasant in high dimensions. See Berliner [3] for some general discussion of issues for highly nonlinear, or "chaotic", $\mathcal{M}$.

While the theory provides a target for analysis, it is rarely implementable in practice. An exception to this difficulty is the Kalman Filter (KF), which is known to be an exact Bayesian procedure under linear and Gaussian assumptions. A useful approximate procedure in the presence of nonlinearity is the *extended* KF in which one linearizes $\mathcal{M}$. See West and Harrison [16] for a general review. Another interesting direction is the Ensemble KF (e.g., [8,9]).

Our main interest here is classes of procedures known as Particle Filters. The underlying idea behind these procedures is the combination of Monte Carlo sampling and discrete approximation to the Bayes' Theorem. At a given time, Monte Carlo samples or ensembles are generated from an approximation to the current distribution. The probabilities of the elements or particles in the ensemble are updated based on data. That is, the particles are weighted in a fashion that mimics Bayes' Theorem. Unfortunately, in high dimensions or as time evolves, these weights typically collapse to favor just a few or even one of the particles. We amplify on this issue in Section 3.3.

1.2. Outline

Section 2 provides an overview of the fundamental bases of Monte Carlo as an approach to ensemble forecasting. The critical notion of importance sampling is developed. Section 3 begins with a brief review of classes of sequential, importance sampling Monte Carlo approaches generally known as particle filters. That section concludes with remarks concerning the value of such methods in processing complex data and their