Advances in Water Resources 70 (2014) 36–50

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Ensemble-based characterization of uncertain environmental features

Rafał Wójcik [a,b,*], Dennis McLaughlin [a], Seyed Hamed Alemohammad [a], Dara Entekhabi [a]

[a] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[b] AIR Worldwide, Financial and Uncertainty Modeling, 3 Copley Place, Boston, MA 02116, USA

ARTICLE INFO

Article history:
Received 31 October 2013
Received in revised form 27 February 2014
Accepted 1 April 2014
Available online 19 April 2014

Keywords:
Ensemble estimation
Image fusion
Importance sampling
Dimensionality reduction
Data assimilation
Precipitation

ABSTRACT

This paper considers the characterization of uncertain spatial features that cannot be observed directly but must be inferred from noisy measurements. Examples of interest in environmental applications include rainfall patterns, solute plumes, and geological features. We formulate the characterization process as a Bayesian sampling problem and solve it with a non-parametric version of importance sampling. All images are concisely described with a small number of image attributes. These are derived from a multidimensional scaling procedure that maps high dimensional vectors of image pixel values to much lower dimensional vectors of attribute values. The importance sampling procedure is carried out entirely in terms of attribute values. Posterior attribute probabilities are derived from non-parametric estimates of the attribute likelihood and proposal density. The likelihood is inferred from an archive of noisy operational images that are paired with more accurate ground truth images. Proposal samples are generated from a non-stationary multi-point statistical algorithm that uses training images to convey distinctive feature characteristics. To illustrate concepts we carry out a virtual experiment that identifies rainy areas on the Earth's surface from either one or two remote sensing measurements. The two sensor case illustrates the method's ability to merge measurements with different error properties. In both cases, the importance sampling procedure is able to identify the proposals that most closely resemble a specified true image.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

In many fields there is a need to characterize uncertain spatial features that cannot be observed directly. Examples of particular interest in environmental applications include characterization of surface rainfall in ungaged areas [59], tracking of subsurface solute plumes, and identification of geological features such as groundwater aquifers, mineral deposits, and oil reservoirs [6,14,26]. Remote sensing measurements such as passive and active satellite microwave, geodetic, or seismic observations can often provide useful but imperfect information about uncertain features. The statistical attributes of these noisy measurements can be estimated by comparing them to more accurate but less readily available "ground truth" measurements. For example, ground-based weather radar and rain gage data can serve as ground truth for an evaluation of errors in satellite microwave measurements that have greater coverage but may be less accurate. In subsurface flow borehole measurements can serve as local ground truth for seismic or other remotely sensed data.

The diverse measurements collected in a typical environmental assessment can be conveniently combined in a Bayesian framework that considers all available sources of information. In a feature-based application images are characterized by vectors of appropriate variables (e.g. pixel values or feature attributes). The Bayesian approach conditions a prior distribution of the uncertain true image/vector on a set of noisy measured images/vectors, yielding a posterior distribution that characterizes the uncertainty remaining after the measurements are taken into account.

Many researchers have investigated methods for using noisy measurements to characterize complex environmental features. One option is to identify a point estimate (a single image) that optimizes an appropriate deterministic performance objective (e.g. a least-squares measure of misfit between the estimate and a measured image). These methods can be viewed as Bayesian a posteriori estimators (i.e. estimators of the posterior mode) if Gaussian assumptions are adopted. Feature-based optimization methods often use level-set techniques to characterize irregular and/or disconnected feature boundaries [3,32]. Level set optimization methods are flexible and popular in the image processing community but they do not generate posterior distributions of uncertain

[*] Corresponding author at: Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

E-mail addresses: raffywojcik@gmail.com (R. Wójcik), dennism@mit.edu (D. McLaughlin), hamed_al@mit.edu (S.H. Alemohammad), darae@mit.edu (D. Entekhabi).

http://dx.doi.org/10.1016/j.advwatres.2014.04.005
0309-1708/© 2014 Elsevier Ltd. All rights reserved.

feature variables. Posterior distributions are needed for probabilistic applications such as ensemble forecasting, risk assessment, extreme value analysis, and stochastic control.

Ensemble methods such as Markov Chain Monte Carlo (MCMC) and importance sampling provide a probabilistic characterization that yields samples from the Bayesian posterior [1,7,17,44]. However, these methods often rely on parametric prior and measurement error distributions (e.g. the multivariate Gaussian distribution) and are generally not applied to feature characterization problems. This reflects the fact that very large sample sizes are needed to properly characterize the non-parametric multivariate probability densities of high-dimensional image vectors [12,65]. Spatial features such as geological facies, solute plumes, and rain storms are usually difficult to describe with parametric probability models. The multipoint statistical approach used in many geological applications [53,57,58] provides an alternative that generates realistic non-parametric prior or proposal samples from training images. However, multipoint methods are generally unable to sample from the Bayesian posterior distribution.

An ideal approach for characterizing uncertain features would combine the level set method's ability to handle complex geometries, the general Bayesian probabilistic framework provided by MCMC and importance sampling, and the non-parametric samples generated by multi-point statistical techniques. This paper describes an approximate method for integrating these different capabilities in a single characterization procedure. It adopts a non-parametric importance sampling approach with proposal samples generated from training images and measurement error samples generated from archived data. The computational limitations of ensemble sampling are partially circumvented by representing each image with a small vector of feature attributes that provides a more concise and efficient description than a classical pixel-based description. Since it is difficult to specify in advance a set of universal attributes that adequately characterize complex features we use a multidimensional scaling technique to derive the attributes directly from image pixel values.

In the following sections we first formulate an attribute-based approach to the feature characterization problem, using importance sampling to generate approximate probability-weighted samples from the Bayesian posterior. We then consider how multidimensional scaling can identify image attributes from pixel-based image vectors and how the priors and likelihoods used in importance sampling can be generated from archived attribute data. The general concepts are illustrated with a virtual experiment based on real rainfall measurements. The paper concludes with a discussion of conceptual and computational issues that identify directions for future research.

2. Image based importance sampling

Importance sampling is a procedure that generates samples from a posterior probability distribution that is related through Bayes theorem to a prior distribution and a likelihood function. Following the approach outlined above, we formulate the feature-based importance sampling problem in terms of a small number of distinctive image attributes. In particular, we define a random vector $\hat{x}$ composed of true image attributes. The true image is observed by one or more sensors that produce noisy images, which we call current operational measurements to distinguish them from the archived operational measurements discussed below. The current measurement from sensor r is described by an attribute vector $\hat{z}_r$. Attributes for all of the measurements are assembled in the global measurement vector $\hat{z}$. Section 3.4 describes how the image attributes are derived.

The objective of importance sampling is to generate a set of samples $\hat{x}_1, \hat{x}_2, \ldots \hat{x}_{N_x}$ of $\hat{x}$ from the posterior probability density $p_{\hat{X}|\hat{Z}}(\hat{x}|\hat{z})$. This density can be expressed, through Bayes theorem, as:

$$p_{\hat{X}|\hat{Z}}(\hat{x}|\hat{z}) = \hat{c}\, p_{\hat{Z}|\hat{X}}(\hat{z}|\hat{x}) p_{\hat{X}}(\hat{x}) \tag{1}$$

where $p_{\hat{Z}|\hat{X}}(\hat{z}|\hat{x})$ is the likelihood function, $p_{\hat{X}}(\hat{x})$ is a prior probability density that does not depend on the measurements, and $\hat{c}$ is a proportionality constant selected to insure that the posterior density integrates to unity. The prior density quantifies our prior uncertainty about the true image while the likelihood function describes the effects of measurement error. Since it is difficult to sample directly from the posterior density we work instead with a more convenient set of equally likely random samples from a proposal density $q(\hat{x})$. Unlike the prior, the proposal density can depend on the current measurements. We use the multipoint statistical techniques described below to generate proposal samples. An approximate posterior probability distribution can be derived by appropriate weighting of these proposal samples. It is convenient to illustrate the process by considering the following equivalent expressions for the conditional expectation of $\hat{x}$ over $p_{\hat{X}|\hat{Z}}(\hat{x}|\hat{z})$ (for a given $\hat{z}$):

$$\begin{aligned} E_p[\hat{x}|\hat{z}] &= \int \hat{x}\, p_{\hat{X}|\hat{Z}}(\hat{x}|\hat{z}) d\hat{x} = \int q(\hat{x}) \left[\hat{x}\, \frac{p_{\hat{X}|\hat{Z}}(\hat{x}|\hat{z})}{q(\hat{x})}\right] d\hat{x} \\ &= E_q\left[\hat{x}\, \frac{p_{\hat{X}|\hat{Z}}(\hat{x}|\hat{z})}{q(\hat{x})}\right] \end{aligned} \tag{2}$$

This mean can be estimated from the N_x proposal samples:

$$\begin{aligned} E_q\left[\hat{x}\, \frac{p_{\hat{X}|\hat{Z}}(\hat{x}|\hat{z})}{q(\hat{x})}\right] &\approx \frac{\hat{c}}{N_x} \sum_{i=1}^{N_x} \hat{x}_i \frac{p_{\hat{Z}|\hat{X}}(\hat{z}|\hat{x}_i) p_{\hat{X}}(\hat{x}_i)}{q(\hat{x}_i)} \\ &= \int \hat{x} \left[\frac{\hat{c}}{N_x} \sum_{i=1}^{N_x} \frac{p_{\hat{Z}|\hat{X}}(\hat{z}|\hat{x}_i) p_{\hat{X}}(\hat{x}_i)}{q(\hat{x}_i)} \delta(\hat{x} - \hat{x}_i)\right] d\hat{x} \end{aligned} \tag{3}$$

where $\hat{x}_i$ is sampled from $q(\hat{x})$. The final bracketed expression in (3) is a discrete approximation to the posterior density $p_{\hat{X}|\hat{Z}}(\hat{x}|\hat{z})$ that appears in the second term in (2). This expression can be re-written as:

$$p_{\hat{X}|\hat{Z}}(\hat{x}|\hat{z}) \approx \sum_{i=1}^{N_x} w_i\, \delta(\hat{x} - \hat{x}_i) \tag{4}$$

where:

$$w_i = \frac{\hat{c}}{N_x} \frac{p_{\hat{Z}|\hat{X}}(\hat{z}|\hat{x}_i) p_{\hat{X}}(\hat{x}_i)}{q(\hat{x}_i)} \tag{5}$$

and $\hat{c}$ is selected such that $\sum_i w_i = 1$. Eq. (4) tells us that the proposal image attribute vectors $\widehat{x_1}, \widehat{x_2}, \ldots, \widehat{x_{N_x}}$ can be interpreted as samples from the posterior density if they are assigned the weights $\widehat{w_1}, \widehat{w_2}, \ldots, \widehat{w_{N_x}}$ computed in (5) rather than the equal weights that apply when they are treated as samples from the proposal density. So the proposal and posterior samples have the same values but different probabilities. Note that, in the special case where the proposal density is the same as the prior the $p_{\hat{X}}(\hat{x}_i)$and $q(\hat{x}_i)$ terms in (5) cancel and the weights only depend on the likelihood. However, it is usually best to draw proposal samples from a distribution that depends on the current measurement rather than from the prior distribution, which does not.

3. Generating the information needed for importance sampling

The importance sampling approach outlined in Section 2 needs a set of proposals $\widehat{x_1}, \widehat{x_2}, \ldots, \widehat{x_{N_x}}$, the likelihood function $p_{\hat{X}|\hat{Z}}(\hat{x}|\hat{z})$, and the prior $p_{\hat{X}}(\hat{x})$ and proposal $q(\hat{x})$ probability densities. The

Download English Version:

https://daneshyari.com/en/article/4525490

Download Persian Version:

https://daneshyari.com/article/4525490

Daneshyari.com