2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

# Bayesian Prediction and Adaptive Sampling Algorithms for Mobile Sensor Networks

Yunfei Xu, Jongeun Choi, Sarat Dass and Taps Maiti

*Abstract*— **In this paper, we formulate a full Bayesian approach for spatio-temporal Gaussian process regression under practical conditions such as measurement noise and unknown hyperparmeters (particularly, the bandwidths). Thus, multifactorial effects of observations, measurement noise and prior distributions of hyperparameters are all correctly incorporated in the computed predictive distribution. Using discrete prior probabilities and compactly supported kernels, we provide a way to design sequential Bayesian prediction algorithms that can be computed (without using the Gibbs sampler) in constant time as the number of observations increases. Both centralized and distributed sequential Bayesian prediction algorithms have been proposed for mobile sensor networks. An adaptive sampling strategy for mobile sensors, using the maximum a posteriori (MAP) estimation, has been proposed to minimize the prediction error variances. Simulation results illustrate the effectiveness of the proposed algorithms.**

## I. Introduction

Recently, there has been increasing exploitation of the mobile sensor networks in environmental monitoring [1], [2], [3], [4]. Gaussian process regression (or Kriging in geostatistics) has been widely used to draw statistical inferences from geostatistical and environmental data [5], [6], [7]. Gaussian process modeling enables us to predict physical values, such as temperature or harmful algae bloom biomass, at any point and time with a predicted uncertainty level. For example, near-optimal static sensor placements with a mutual information criterion in Gaussian processes were proposed in [8]. A distributed Kriged Kalman filter for spatial estimation based on mobile sensor networks was developed in [4]. Multi-agent systems that are versatile for various tasks by exploiting predictive posterior statistics of Gaussian processes were developed in [9], [10].

The unknown hyperparameters in the covariance function of a Gaussian process can be estimated by the maximum likelihood (ML) estimator and used in the prediction as the true hyperparameters by mobile sensor networks [11]. In a full Bayesian approach, however, the uncertainty in the hyperparameters shall be incorporated in the prediction [12]. In [13], Gaudard et al. presented a Bayesian method that uses importance sampling for analyzing spatial data sampled from a Gaussian random field whose covariance function was not known. However, the complexity and the assumptions made in [13] such as noiseless observations and the time-invariance of the field limit the applicability of the approach on mobile sensors in practice. A distributed adaptive sampling approach was proposed by [14] for sensor networks to find locations that maximize the information contents for prediction purpose. In [14], an iterative prediction algorithm without a Markov chain Monte Carlo (MCMC) method has been developed based on the analytical closed-form solutions from the result in [13] by assuming that the covariance function of the spatio-temporal random field is known up to a constant. Regression analysis for Gaussian processes requires growing computational complexity since the size of the covariance matrix increases as the number of observations increases. This problem in the context of the mobile sensor networks has been tackled in different directions [15], [16]. Computational complexity of a full Bayesian prediction algorithm for spatio-temporal Gaussian processes with unknown covariance functions grows in a prohibitively fast rate as the observation number increases due to the MCMC method. These have been the main hurdles for resource-constrained robots to efficiently use full Bayesian approaches for Gaussian process regression.

The contribution of this paper is as follows. First, we provide a full Bayesian approach for spatio-temporal Gaussian process regression under more practical conditions such as measurement noise and unknown hyperparmeters (particularly, the bandwidths). We also present an approach to compute the predictive distribution using the Gibbs sampler [17] (Section II). Thus, multifactorial effects of observations, measurement noise and prior distributions of hyperparameters are all correctly incorporated in the prediction of the Gaussian process by this full Bayesian approach. Using discrete prior probabilities and compactly supported kernels [18], we then provide a way to design sequential Bayesian prediction algorithms that can be computed in constant time as the number of observations increases. In particular, the sequential Bayesian prediction is developed in the forms of centralized and distributed algorithms (Section IV). An adaptive sampling strategy for mobile sensors, utilizing the maximum a posteriori (MAP) estimation of the hyperparameters, is proposed to minimize the prediction error variances (Section V). Finally, the proposed Bayesian prediction algorithms and the adaptive sampling strategy are tested under spatio-temporal Gaussian processes.

Standard notation is used throughout the paper. Let $\mathbb{R}$, $\mathbb{R}_{\geq 0}$, $\mathbb{R}_{>0}$, $\mathbb{Z}$, $\mathbb{Z}_{\geq 0}$, $\mathbb{Z}_{>0}$ denote, respectively, the sets of real, non-negative real, positive real, integer, non-negative

Yunfei Xu is with the Department of Mechanical Engineering, Michigan State University, East Lansing, MI 48824, USA. E-mail: xuyunfei@egr.msu.edu.

Jongeun Choi is with the Departments of Mechanical Engineering and Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824, USA. E-mail: jchoi@egr.msu.edu.

Sarat Dass and Taps Maiti are with the Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA. Emails: { sdass, maiti }@msu.edu.

978-1-4577-0079-8/11/$26.00 ©2011 AACC

integer, and positive integer numbers. Let E, Var and Corr denote, respectively, the operators of expectation, variance and correlation. Other notation will be explained in due course.

## II. A FULL BAYESIAN APPROACH FOR GAUSSIAN PROCESS REGRESSION

In this section, we present a model of a spatio-temporal Gaussian process and its observations, (which will be used throughout the paper,) by expanding the model used in [13] to include the temporal process and measurement noise. We then show how to perform Bayesian inference using a MCMC technique for our formulation. Let us consider a spatio-temporal Gaussian process $z(\mathbf{x}) \sim \mathcal{GP}\left(\mu(\mathbf{x}), \sigma_f^2 \mathcal{K}(\mathbf{x}, \mathbf{x}')\right)$, where $z \in \mathbb{R}$ and $\mathbf{x} := \begin{bmatrix} \mathbf{s}^T & t \end{bmatrix}^T \in \mathcal{Q} \times \mathbb{R}_{\geq 0}$ contains the sampling location $\mathbf{s} \in \mathcal{Q} \subset \mathbb{R}^D$ and the sampling time $t \in \mathbb{R}_{\geq 0}$. The mean function is assumed to be $\mu(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T \boldsymbol{\beta}$, where $\mathbf{f}(\mathbf{x}) := \begin{bmatrix} f_1(\mathbf{x}) & \cdots & f_p(\mathbf{x}) \end{bmatrix}^T \in \mathbb{R}^p$ is a known regression function, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector of regression coefficients. The covariance between $z(\mathbf{x})$ and $z(\mathbf{x}')$ is assumed to have the form of $\sigma_f^2 \mathcal{K}(\mathbf{x}, \mathbf{x}')$. The correlation function $\mathcal{K}(\mathbf{x}, \mathbf{x}')$ is defined by

$$\mathcal{K}(\mathbf{x}, \mathbf{x}') = \phi_s\left(\frac{\|\mathbf{s} - \mathbf{s}'\|}{\sigma_s}\right) \phi_t\left(\frac{|t - t'|}{\sigma_t}\right), \qquad (1)$$

where we assume that $\phi_s(\cdot)$ and $\phi_t(\cdot)$ are decreasing kernel functions for space and time, respectively. The signal variance $\sigma_f^2$ gives the overall vertical scale relative to the mean of the Gaussian process in the output space. These parameters together with the unknown mean $\boldsymbol{\beta}$ play the role of hyperparameters. We defined the hyperparameter vector as $\boldsymbol{\theta} := \begin{bmatrix} \boldsymbol{\beta}^T & \sigma_f^2 & \sigma_s & \sigma_t \end{bmatrix}^T \in \mathbb{R}^p \times \mathbb{R}_{>0}^3$.

Suppose we have a collection of observations $\mathcal{D} = \left\{(\mathbf{x}^{(i)}, y^{(i)}) | i = 1, \ldots, n\right\}$ where $\mathbf{x}^{(i)}$ denotes an input vector of dimension $D + 1$ and $y^{(i)}$ denotes a scalar value of the noise corrupted measurement, i.e., $y^{(i)} = z(\mathbf{x}^{(i)}) + \epsilon^{(i)}, \quad \epsilon^{(i)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2)$. We assume that the *signal-to-noise ratio* $\gamma = \sigma_f^2 / \sigma_w^2$ is known, which is necessary for identifiability. Let $\mathbf{y} \in \mathbb{R}^n$ denote the collection of the measurements.

The spatio-temporal Gaussian process regression provides the prediction (or predictive distribution) of $z_* := z(\mathbf{x}_*)$ at location $\mathbf{s}_* \in \mathcal{Q}$ and time $t_* \in \mathbb{R}_{\geq 0}$ for given noisy measurements $\mathbf{y}$. For the known hyperparameter vector $\boldsymbol{\theta}$, the prediction of $z_*$ at location $\mathbf{s}_*$ and time $t_*$ can be obtained by $z_* | \mathbf{y} \sim \mathcal{N}\left(\hat{z}_*, \sigma_{\hat{z}_*}^2\right)$, where $\hat{z}_* = \mathrm{E}(z_* | \mathbf{y}) = \mathbf{f}(\mathbf{x}_*)^T \boldsymbol{\beta} + \mathbf{k}^T \mathbf{C}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})$, and $\sigma_{\hat{z}_*}^2 = \mathrm{Var}(z_* | \mathbf{y}) = \sigma_f^2(1 - \mathbf{k}^T \mathbf{C}^{-1} \mathbf{k})$. We have defined

$$\begin{aligned} \mathbf{F} &:= \begin{bmatrix} \mathbf{f}(\mathbf{x}^{(1)}) & \cdots & \mathbf{f}(\mathbf{x}^{(n)}) \end{bmatrix}^T \in \mathbb{R}^{n \times p}, \\ \mathbf{k} &:= \mathrm{Corr}(\mathbf{y}, z_*) = [\mathcal{K}(\mathbf{x}^{(i)}, \mathbf{x}_*)] \in \mathbb{R}^n, \\ \mathbf{C} &:= \mathrm{Corr}(\mathbf{y}, \mathbf{y}) = [\mathcal{K}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) + \frac{1}{\gamma}\delta_{ij}] \in \mathbb{R}^{n \times n}, \end{aligned} \qquad (2)$$

where $\delta_{ij}$ is the Dirac delta function which equals one when $i = j$ and zero otherwise.

TABLE I

THE GIBBS SAMPLER.

```
Initialize $\boldsymbol{\beta}^{(1)}, {\sigma_f^2}^{(1)} \sigma_s^{(1)}, \sigma_t^{(1)}$
**for** $i = 1$ to $m$ **do**
    sample $\boldsymbol{\beta}^{(i+1)}$ from $\pi(\boldsymbol{\beta} | {\sigma_f^2}^{(i)}, \sigma_s^{(i)}, \sigma_t^{(i)}, \mathbf{y})$
    sample ${\sigma_f^2}^{(i+1)}$ from $\pi(\sigma_f^2 | \boldsymbol{\beta}^{(i+1)}, \sigma_s^{(i)}, \sigma_t^{(i)}, \mathbf{y})$
    sample $\sigma_s^{(i+1)}, \sigma_t^{(i+1)}$ from $\pi(\sigma_s, \sigma_t | \boldsymbol{\beta}^{(i+1)}, {\sigma_f^2}^{(i+1)}, \mathbf{y})$
**end for**
```

In Bayesian statistics, the unknown hyperparameter vector $\boldsymbol{\theta}$ is considered to be a random vector and hence its *prior* has to be defined. In this paper, we use the prior distribution of the hyperparameter vector that satisfies

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta}, \sigma_f^2, \sigma_s, \sigma_t) = \pi(\boldsymbol{\beta} | \sigma_f^2) \pi(\sigma_f^2) \pi(\sigma_s) \pi(\sigma_t),$$

along with $\pi(\boldsymbol{\beta} | \sigma_f^2) \propto 1$, and $\pi(\sigma_f^2) = \mathrm{IG}(a_f, b_f)$, where $\mathrm{IG}(a_f, b_f)$ denotes the *inverse gamma distribution* with mean $b_f / (a_f - 1)$. We may choose default priors for $\pi(\sigma_s)$ and $\pi(\sigma_t)$ that ensures posterior property, which would mimic the properties of the ML estimate in absence of meaningful prior information.

The *posterior distribution* of $\boldsymbol{\theta}$ is proportional to the likelihood times the prior, i.e., $\pi(\boldsymbol{\theta} | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta})$. The inference on $\boldsymbol{\theta}$ can be carried out by sampling from the posterior distribution via the Gibbs sampler which is shown in Table I.

To implement the Gibbs sampler in Table I to generate samples from the posterior distribution, we exploit the following proposition.

*Proposition 1:* (i) For given $\sigma_f^2$, $\sigma_s$, $\sigma_t$, and $\mathbf{y}$, we have $\boldsymbol{\beta} | \sigma_f^2, \sigma_s, \sigma_t, \mathbf{y} \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}^2\right)$, where $\hat{\boldsymbol{\beta}} = (\mathbf{F}^T \mathbf{C}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \mathbf{C}^{-1} \mathbf{y}$, and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \sigma_f^2 (\mathbf{F}^T \mathbf{C}^{-1} \mathbf{F})^{-1}$. $\mathbf{F}$ and $\mathbf{C}$ are defined in (2).

(ii) For given $\boldsymbol{\beta}$, $\sigma_s$, $\sigma_t$, and $\mathbf{y}$, we have $\sigma_f^2 | \boldsymbol{\beta}, \sigma_s, \sigma_t, \mathbf{y} \sim \mathrm{IG}\left(\tilde{a}_f, \tilde{b}_f\right)$, where $\tilde{a}_f = a_f + \frac{n}{2}$, and $\tilde{b}_f = b_f + \frac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^T \mathbf{C}^{-1} (\mathbf{y} - \mathbf{F}\boldsymbol{\beta})$.

(iii) For given $\boldsymbol{\beta}$, $\sigma_f^2$, and $\mathbf{y}$, we have $\pi(\sigma_s, \sigma_t | \boldsymbol{\beta}, \sigma_f^2, \mathbf{y}) \propto \frac{1}{|\mathbf{C}|^{1/2}} \exp\left(-\frac{(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^T \mathbf{C}^{-1} (\mathbf{y} - \mathbf{F}\boldsymbol{\beta})}{2\sigma_f^2}\right) \pi(\sigma_s) \pi(\sigma_t)$.

*Proof:* The proof is omitted due to the page limit. ■

*Remark 2:* The hyperparameters $\boldsymbol{\beta}$ and $\sigma_f^2$ in $p(\mathbf{y} | \boldsymbol{\theta})$ can be marginalized [13] such that $\pi(\sigma_s, \sigma_t | \mathbf{y}) \propto p(\mathbf{y} | \sigma_s, \sigma_t) \pi(\sigma_s, \sigma_t)$. Notice that $p(\mathbf{y} | \sigma_s, \sigma_t)$ has an analytical form which is no longer Gaussian. By this way, the dimensionality of the Gibbs sampler can be reduced to 2.

The *predictive distribution* of $z_*$ at location $\mathbf{s}_*$ and time $t_*$ can be obtained by

$$p(z_* | \mathbf{y}) = \int p(z_* | \mathbf{y}, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} | \mathbf{y}) \mathrm{d}\boldsymbol{\theta}. \qquad (3)$$

If we draw $m$ samples $\left\{\boldsymbol{\theta}^{(i)}\right\}_{i=1}^m$ according to the posterior distribution $\pi(\boldsymbol{\theta} | \mathbf{y})$ using the Gibbs sampler, the predictive distribution in (3) can then be approximated by

$$p(z_* | \mathbf{y}) \approx \frac{1}{m} \sum_{i=1}^m p(z_* | \mathbf{y}, \boldsymbol{\theta}^{(i)}). \qquad (4)$$

The predictive mean and variance can be now obtained by

$$\begin{aligned}\hat{z}_* &= \mathrm{E}(z_*|\mathbf{y}) \approx \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}(z_*|\mathbf{y},\boldsymbol{\theta}^{(i)}),\\ \sigma_{\hat{z}_*}^2 &= \mathrm{Var}(z_*|\mathbf{y}) \approx \frac{1}{m}\sum_{i=1}^{m}\mathrm{Var}(z_*|\mathbf{y},\boldsymbol{\theta}^{(i)})\\ &\quad + \frac{1}{m}\sum_{i=1}^{m}\left[\mathrm{E}(z_*|\mathbf{y},\boldsymbol{\theta}^{(i)}) - \mathrm{E}(z_*|\mathbf{y})\right]^2.\end{aligned} \tag{5}$$

In the next section, we will formulate a problem that seeks a sequential Bayesian prediction algorithm to deal with the complexity issue in the Gibbs sampler.

## III. Problem Formulation

From here on, we focus on the spatio-temporal Gaussian process which has zero mean and known signal variance $\sigma_f^2$, i.e., the unknown hyperparameter vector is now $\boldsymbol{\theta} = [\ \sigma_s \quad \sigma_t\ ]^T$. As discussed in Remark 2, the unknown hyperparameters $\boldsymbol{\beta}$ and $\sigma_f^2$ can be treated efficiently exploiting the analytical closed-form solutions [13].

Consider the problem of Bayesian prediction with sampling by a mobile sensor network, which consists of $N$ mobile agents distributed over the surveillance region $\mathcal{Q} \subset \mathbb{R}^D$. The identity of each agent is indexed by $\mathcal{J} := \{1, 2, \cdots, N\}$. Let $\mathbf{q}_i(t) \in \mathcal{Q}$ be the position of agent $i$ at time $t \in \mathbb{R}_{\geq 0}$. At time $t$, agent $i$ makes a noise corrupted observation $y(\mathbf{q}_i(t), t) := z(\mathbf{q}_i(t), t) + w(\mathbf{q}_i(t), t)$, where the sensor noise $w(\mathbf{q}_i(t), t)$ is Gaussian, i.e., $w(\mathbf{q}_i(t), t) \sim \mathcal{N}(0, \sigma_w^2)$.

Suppose agents start making observations every $t_s$ from time $t_1 = 0$. Let $\mathbf{y}_k$ denote the collection of observations by all agents at time $t_k$, i.e., $\mathbf{y}_k := [\ y(\mathbf{q}_1(t_k), t_k) \quad \cdots \quad y(\mathbf{q}_N(t_k), t_k)\ ]^T$. Suppose at time $t_k$, we have a set of observations $\{\boldsymbol{\xi}_j\}_{j=1}^{c_k}$, where $M$ is a constant, $c_k := c(k)$ is a non-negative integer and $c(\cdot)$ is a non-decreasing function. For notational simplicity, we define $\boldsymbol{\xi}_{1:c_k} := [\ \boldsymbol{\xi}_1^T \quad \cdots \quad \boldsymbol{\xi}_{c_k}^T\ ]^T$, where $\boldsymbol{\xi}_{1:0} := \varnothing$ is an empty array. Given extra observations $\boldsymbol{\psi}_k$, our objective is to make prediction of $z_* := z(\mathbf{s}_*, t_*)$ at location $\mathbf{s}_*$ and time $t_* = t_k$. The predictive posterior distribution can be obtained by

$$p(z_*|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi}_k) = \int p(z_*|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi}_k, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi}_k)\mathrm{d}\boldsymbol{\theta}. \tag{6}$$

In (6), the posterior probability distribution of the hyperparameters $\boldsymbol{\theta}$ can be obtained by using the prior probability distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_k})$ as follows.

$$\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi}_k) \propto p(\boldsymbol{\psi}_k|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_k}). \tag{7}$$

To utilize all observations up to $t_k$, the standard choices are $c(k) = k - 1$, $\boldsymbol{\xi}_j = \mathbf{y}_j$, and $\boldsymbol{\psi}_k = \mathbf{y}_k$. However, as pointed out in Remark **??**, when $k$ is large, finding the predictive distribution in (6) is computationally prohibited in this Bayesian framework. Therefore, our objective is to design sequential Bayesian prediction algorithms which can be implemented in constant time (i.e., does not grow with the time index $k$) by choosing an appropriate set of $c_k$, $\boldsymbol{\xi}_{1:c_k}$, and $\boldsymbol{\psi}_k$, without compromising our Bayesian framework. This problem is formally stated as follows.

*Problem 3:* Design sequential Bayesian prediction algorithms which can be computed in constant time as $k$ increases, i.e., find $c_k$, $\{\boldsymbol{\xi}_j\}_{j=1}^{c_k}$, and $\boldsymbol{\psi}_k$ such that the predictive distribution $p(z_*|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi}_k)$ in (6) can be computed in constant time as $k$ increases.

## IV. Sequential Bayesian prediction algorithms

The following proposition provides a way to tackle Problem 3 by carefully choosing a set of sequential observations.

*Proposition 4:* For a given prior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_k})$ in (7), if the following conditions are satisfied

*C1:* $\boldsymbol{\psi}_k$ and $\boldsymbol{\xi}_{1:c_k}$ are uncorrelated, and
*C2:* $z_*$ and $\boldsymbol{\xi}_{1:c_k}$ are uncorrelated,

then $p(z_*|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi}_k, \boldsymbol{\theta})$ in (6) and $p(\boldsymbol{\psi}_k|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\theta})$ in (7) used for Problem 3 can be computed in constant time, i.e., the computational power required will not grow with $k$.

*Proof:* The proof is straightforward and is omitted. ■

In what follows, using the result from Proposition 4, we present sequential Bayesian prediction for Problem 3 in the forms of centralized and distributed algorithms.

### A. A centralized algorithm

Consider a case in which all agents transmit their observations to a central station, which has high computation and memory power.

Assume that we know the range of hyperparameters, i.e., $\sigma_s \in [\ \underline{\sigma}_s \quad \overline{\sigma}_s\ ]$ and $\sigma_t \in [\ \underline{\sigma}_t \quad \overline{\sigma}_t\ ]$, where $\underline{a}$ and $\overline{a}$ denote the known lower-bound and upper-bound of the random variable $a$, respectively. To avoid the computationally demanding MCMC as outlined, we assign discrete uniform probability distributions to $\sigma_s$ and $\sigma_t$ as priors instead of continuous probability distributions. Hence, $\pi(\boldsymbol{\theta})$ is now a probability. By this way, the possible choices of $\boldsymbol{\theta}$ are constrained on a finite set of grid points denoted by $\Theta$.

To satisfy conditions $C$1-2 in Proposition 4, we consider a class of spatio-temporal Gaussian processes generated by a compactly supported kernel function for time ($\phi_t(h)$ in (1)) such that the correlation vanishes when the time difference between two inputs is larger than $\sigma_t$, i.e., $\phi_t(h) = 0, \forall h > 1$.

The following theorem shows how to select $\boldsymbol{\xi}_j$, $c_k$ and $\boldsymbol{\psi}_k$ to satisfy conditions $C$1-2 in Proposition 4.

*Theorem 5:* Consider the aforementioned prior probability $\pi(\boldsymbol{\theta})$ and the compactly supported kernel. If we choose $\eta \in \mathbb{Z}_{>0}$ such that $t_s \geq \overline{\sigma}_t/\eta$, and

$$\begin{aligned} c_k &:= \max\left(\lfloor (k/\eta - 1)/2 \rfloor, 0\right),\\ \boldsymbol{\xi}_j &:= \mathbf{y}_{2(j-1)\eta+1:(2j-1)\eta},\\ \boldsymbol{\psi}_k &:= \mathbf{y}_{k-\eta+1:k},\end{aligned}$$

where $\lfloor\cdot\rfloor$ is the floor function defined by $\lfloor x \rfloor := \max\{n \in \mathbb{Z} | n \leq x\}$, then the predictive distribution in Problem 3 can be computed in constant time, i.e., the computational power required does not grow with the time index $k$.

*Proof:* By construction, conditions $C$1-2 in Proposition 4 are satisfied. Hence, by the result of Proposition 4,

TABLE II

THE CENTRALIZED BAYESIAN PREDICTION ALGORITHM.

| | |
|---|---|
| **Input:** | The number of agents $N$, initial positions of agents $\{\mathbf{q}_i(t_1)\}_{i=1}^N$, the discrete prior distributions $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$, the sampling period $\eta$, the sampling rate $t_s$, an empty array $\boldsymbol{\psi} = \varnothing$, $c_k$ = 0 |
| **Output:** | The prediction at location $\mathbf{s}_*$ and time $t_* = t_k$ |

```
At time $t_k$, agent $i$ does:
  make an observation $y(\mathbf{q}_i(t_k), t_k)$
  send the observation to the central station
At time $t_k$, the central station does:
  collect observations from all agents, i.e., $\mathbf{y}_k$
  set $\boldsymbol{\psi} = [\ \boldsymbol{\psi}^T \quad \mathbf{y}_k^T\ ]^T$
  **if** mod$(k, 2\eta) = \eta$ and $k > \eta$ **then**
      set $c_k = c_k + 1$
  **end if**
  **for** each $\boldsymbol{\theta} \in \Theta$ **do**
      compute $p(\boldsymbol{\psi}|\boldsymbol{\theta})$
      compute $p(z_*|\boldsymbol{\psi}, \boldsymbol{\theta})$
  **end for**
  compute the $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi})$
  **if** mod$(k, 2\eta) = \eta$ **then**
      store $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi})$ as $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_k+1})$
  **end if**
  compute $p(z_*|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi})$ at location $\mathbf{s}_*$ and time $t_* = t_k$
  **if** $k \geq \eta$ **then**
      discard the first $N$ elements in $\boldsymbol{\psi}$, i.e., $\mathbf{y}_{k-\eta+1}$
  **end if**
  compute the next sampling positions for agents $\{\mathbf{q}_i(t_{k+1})\}_{i=1}^N$
    (e.g., using the adaptive sampling strategy proposed in Section V)
  send position commands to agents
At time $t_k$, agent $i$ does:
  receive the position command from the central station
  move to the new positions $\mathbf{q}_i(t_{k+1})$
```

the predictive distribution can be computed in constant time as the time index $k$ increases. ■

By choosing the discrete prior probability distribution on $\boldsymbol{\theta}$, the integration in the predictive distribution in (6) can be reduced to the following summation

$$p(z_*|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi}_k) = \sum p(z_*|\boldsymbol{\psi}_k, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi}_k),$$

where $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi}_k) \propto p(\boldsymbol{\psi}_k|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_k})$.

Using Theorem 5, we provide a centralized sequential Bayesian prediction algorithm as summarized in Table II.

### B. A distributed algorithm

Consider a case in which spatially distributed $M$ groups of agents sample a spatio-temporal Gaussian process over a large region $\mathcal{Q}$. Each group is in charge of its sub-region of $\mathcal{Q}$. The identify of each group is indexed by $\mathcal{I} := \{1, \cdots, M\}$. Each agent in group $i$ is indexed by $\mathcal{J}^{[i]} := \{1, \cdots, N\}$. The leader of group $i$ is referred to as leader $i$, which implements the centralized scheme to make prediction on its sub-region using local observations and the globally updated posterior distribution of $\boldsymbol{\theta}$. Therefore, in this sensor network structure, the posterior distribution of $\boldsymbol{\theta}$ shall be updated correctly using all observations from all groups (or agents) in a distributed fashion.

Let $G(t) := (\mathcal{I}, \mathcal{E}(t))$ be an undirected communication graph such that an edge $(i, j) \in \mathcal{E}(t)$ if and only if leader $i$ can communicate with leader $j$ at time $t$. We define the neighborhood of leader $i$ at time $t$ by $N_i(t) := \{j \in \mathcal{I}|(i, j) \in \mathcal{E}(t), j \neq i\}$. Let $\mathbf{a}^{[i]}$ denote the quantity as $\mathbf{a}$ in the centralized scheme for group $i$. To develop a distributed scheme for data fusion in Bayesian statistics, we exploit the compactly supported kernel for space. Let $\phi_s(h)$ also be a compactly supported kernel function as $\phi_t(h)$. We then have the following Theorem.

*Theorem 6:* Assume that $\boldsymbol{\psi}_k^{[i]}$ and $\boldsymbol{\xi}_{c_k}^{[i]}$ for leader $i$ are selected accordingly to Theorem 5 in time-wise. Let $\boldsymbol{\psi}_k$ be defined by $\boldsymbol{\psi}_k := [(\boldsymbol{\psi}_k^{[1]})^T, \cdots, (\boldsymbol{\psi}_k^{[M]})^T]^T$. If the following condition is satisfied

*C3:* $\|\mathbf{q}_\ell^{[i]}(t) - \mathbf{q}_\nu^{[j]}(t')\| \geq \overline{\sigma}_s, \forall i \neq j, \forall \ell \in \mathcal{J}^{[i]}, \forall \nu \in \mathcal{J}^{[j]}$,

in space-wise, then the global posterior probability distribution of the hyperparameter vector $\boldsymbol{\theta}$, based on all observations from all agents, can be obtained via

$$\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi}_k) \propto \pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_k}) \prod_{i=1}^{M} p(\boldsymbol{\psi}_k^{[i]}|\boldsymbol{\theta}). \tag{8}$$

*Proof:* If the condition $C3$ is satisfied, then we have $\text{Corr}(\boldsymbol{\psi}_k^{[i]}, \boldsymbol{\psi}_k^{[j]}) = 0, \forall i \neq j$, for all possible $\boldsymbol{\theta}$. We then have $p(\boldsymbol{\psi}_k|\boldsymbol{\theta}) = \prod_{i=1}^{M} p(\boldsymbol{\psi}_k^{[i]}|\boldsymbol{\theta})$, which concludes the proof. ■

Suppose that the communication graph $G(t)$ is connected for all time $t$. Then $\prod_{i=1}^{M} p(\boldsymbol{\psi}_k^{[i]}|\boldsymbol{\theta})$ in (8) can be achieved asymptotically via *belief consensus algorithm* [19].

*Theorem 7 ([19]):* Consider a connected (undirected) network of leaders that exchange the likelihood $p(\boldsymbol{\psi}_k^{[i]}|\boldsymbol{\theta})$, then a group product-consensus value $\left(\prod_{i=1}^{M} p(\boldsymbol{\psi}_k^{[i]}|\boldsymbol{\theta})\right)^{1/M}$ can be achieved asymptotically via the updating rule

$$p(\boldsymbol{\psi}_k^{[i]}|\boldsymbol{\theta}) \leftarrow p(\boldsymbol{\psi}_k^{[i]}|\boldsymbol{\theta})^{\beta_i} \prod_{j \in N_i} p(\boldsymbol{\psi}_k^{[i]}|\boldsymbol{\theta})^{\gamma},$$

where $\beta_i = 1 - \gamma d_i > 0$ for all $i$ and $0 < \gamma < 1/\Delta$ ($d_i$ is the degree of node $i$ and $\Delta = \max_i d_i$).

## V. ADAPTIVE SAMPLING

In the previous section, we designed sequential Bayesian prediction algorithms for mobile sensor networks. In this section, we consider how to adaptively deploy mobile sensors at each time step such that the quality of the prediction is improved. At time $t_k$, the goal of the navigation of agents is to improve the quality of prediction of the field $\mathcal{Q}$ at the next sampling time $t_{k+1}$. Therefore, mobile agents should move to the most informative sampling locations $\{\mathbf{q}_1(t_{k+1}), \cdots, \mathbf{q}_N(t_{k+1})\}$ at time $t_{k+1}$ in order to reduce the prediction error [8].

Suppose at time $t_{k+1}$, agents move to a new set of positions $\{\tilde{\mathbf{q}}_1, \cdots, \tilde{\mathbf{q}}_N\}$. The mean squared prediction error is defined as

$$J(\{\tilde{\mathbf{q}}\}_{i=1}^N) = \int_{\mathbf{s} \in \mathcal{Q}} \mathrm{E}\left[(z(\mathbf{s}, t_{k+1}) - \hat{z}(\mathbf{s}, t_{k+1}))^2\right] \mathrm{d}\mathbf{s}, \tag{9}$$

where $\hat{z}(\mathbf{s}, t_{k+1})$ is obtained as in (5). Due to the fact that $\boldsymbol{\theta}$ has a distribution, the evaluation of (9) becomes computationally prohibited. To simplify the optimization, we propose to utilize a *maximum a posteriori* (MAP) estimate of $\boldsymbol{\theta}$ at time $k$, denoted by $\hat{\boldsymbol{\theta}}_{\text{MAP}}(k)$ (i.e., $\boldsymbol{\theta}$ in $\Theta$ which has the highest posterior probabilty at time $t_k$). Hence, (9) can be simplified as

$$J(\{\tilde{\mathbf{q}}\}_{i=1}^N) = \int_{\mathbf{s}\in\mathcal{Q}} \text{Var}(z(\mathbf{s}, t_{k+1})|\boldsymbol{\xi}_{1:c_k}, \boldsymbol{\psi}_k, \hat{\boldsymbol{\theta}}_{\text{MAP}}(k))\text{d}\mathbf{s}.$$

Therefore, the next sampling positions can be obtained by solving the following optimization problem

$$\{\mathbf{q}_i(t+1)\}_{i=1}^N = \arg \min_{\{\tilde{\mathbf{q}}_i\}_{i=1}^N \subset \mathcal{Q}} J(\{\tilde{\mathbf{q}}_i\}_{i=1}^N). \tag{10}$$

This problem can be solved using standard constrained nonlinear optimization techniques (e.g., the conjugate gradient algorithm).

*Remark 8:* The proposed control algorithm in (10) is truly *adaptive* in the sense that the new sampling positions in (10) are functions of all collected observations. On the other hand, if all hyperparameters are known, the optimization in (10) can be performed offline without taking any measurements.

## VI. Simulation results

In this section, we apply our approach to a spatio-temporal Gaussian process with a correlation function in (1). The Gaussian process was numerically generated through circulant embedding of the covariance matrix for the simulation. Assume we know $\boldsymbol{\beta} = \mathbf{0}$ and $\sigma_f^2 = 1$. The signal to noise ratio $\gamma$ is chosen to be 26dB which corresponds to $\sigma_w = 0.05$.

### *A. 1-D scenario using the centralized scheme*

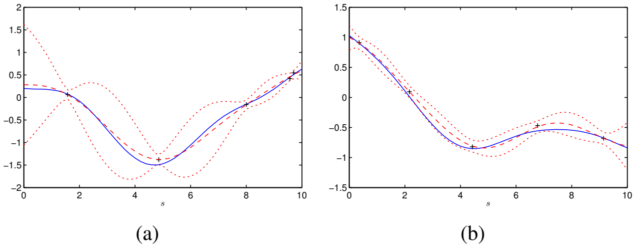
We consider a scenario in which 5 agents sample the spatio-temporal Gaussian process in 1-D space and the central station performs Bayesian prediction. The surveillance region $\mathcal{Q}$ is given by $\mathcal{Q} = [\ 0 \quad 10\ ]$. The hyperparameters used in the simulation was chosen to be $\boldsymbol{\theta} = [\ \sigma_s \quad \sigma_t\ ]^T = [\ 2 \quad 8\ ]^T$.

Here, we can afford a general case where the correlation function for space (i.e., $\phi_s(\cdot)$) is not compactly supported. In particular, we choose the *squared exponential* function $\phi_s(h) = -\frac{1}{2}h^2$. However, the correlation function for time (i.e., $\phi_t(\cdot)$) has to be compactly supported to satisfy the condition in Theorem 5. In particular, we choose [18]

$$\phi_t(h) = \begin{cases} \frac{(1-h)\sin(2\pi h)}{2\pi h} + \frac{1-\cos(2\pi h)}{\pi \times 2\pi h}, & 0 \le h \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Assume we know the bounds of $\boldsymbol{\theta}$, viz. $\sigma_s \in [\ 1.6 \quad 2.4\ ]$ and $\sigma_t \in [\ 4 \quad 12\ ]$. Agents make observations at a fixed sampling rate $t_s = 1$. $\eta$ is chosen to be 12 such that $t_s \geq \overline{\sigma}_t/\eta$. We choose a discrete uniform probability distribution for $\pi(\boldsymbol{\theta})$ as shown in Fig. 2-(a). The prediction is evaluated at each time step for 51 uniform grid points within $\mathcal{Q}$. The prediction result at time $t_1$ is shown in Fig. 1-(a). The predictive variances are large due to the uniform prior distribution (Fig. 2-(a)) for $\boldsymbol{\theta}$ and the small number of observations.

At time $t_{100}$, the prior distribution was updated in a recursive manner based on the observations $\boldsymbol{\xi}_{1:c_{100}}$ and it is shown in Fig. 2-(b). With more number of observations, the support for the posterior distribution of $\boldsymbol{\theta}$ becomes smaller and the peak gets closer to the true value. As shown in Fig 1-(b), the quality of the prediction at time $t_{100}$ is significantly improved by a combination of our Bayesian prediction algorithm and the adaptive sampling strategy. At time $t_{300}$, the prior distribution was further updated which is shown in Fig. 2-(c). At this time, $\boldsymbol{\theta} = [\ 2 \quad 8\ ]^T$, which is the true value, has the highest possibility. This demonstrates the correctness of our algorithm. The running time at each time step is fixed, which is around 12s using Matlab, R2008a (MathWorks) in a PC (2.4GHz Dual-Core Processor).



(a)

(b)

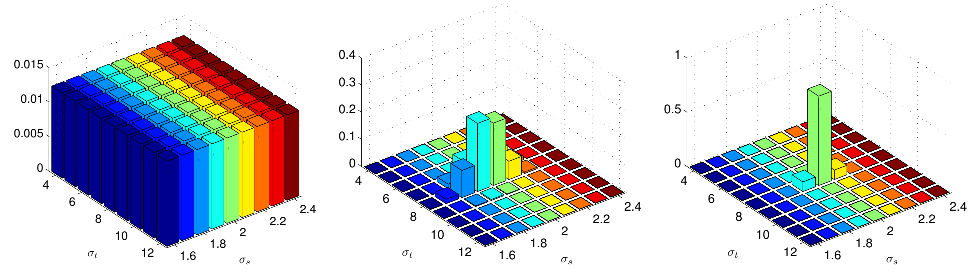
Fig. 1. The prediction at time (a) $t_1$, and (b) $t_{100}$. In each subfigure, the true field is plotted in a blue solid line; the predicted field is plotted in a red dashed line; the area between two red dotted lines indicates the 95% confidence interval.

### *B. 2-D scenario using the distributed scheme*

We consider a scenario in which there are 4 groups, each of which contain 10 agents sampling the spatio-temporal Gaussian process in 2-D space. The surveillance region $\mathcal{Q}$ is given by $\mathcal{Q} = [\ 0 \quad 10\ ] \times [\ 0 \quad 10\ ]$. The hyperparameters used in the simulation were chosen to be $\boldsymbol{\theta} = [\ \sigma_s \quad \sigma_t\ ]^T = [\ 2 \quad 8\ ]^T$. To use the distributed scheme, we have to choose compactly supported kernel functions for both space and time. In particular, we choose $\phi_s(h) = \phi_t(h)$ as in (11).

Assume we know that $\sigma_s \in [\ 1.6 \quad 2.4\ ]$ and $\sigma_t \in [\ 4 \quad 12\ ]$. Agents make observations at a fixed sampling rate $t_s = 1$. $\eta$ is chosen to be 12 such that $t_s \geq \overline{\sigma}_t/\eta$. The region $\mathcal{Q}$ is divided into 4 square sub-regions with equal size as shown in Fig. 4-(a). Distance between any two sub-regions is enforced to be greater than 2.4, which enables the distributed Bayesian prediction. The same uniform prior distribution for $\boldsymbol{\theta}$ as in the centralized version (see Fig. 2-(a)) is chosen.

The globally updated prior distribution of $\boldsymbol{\theta}$ at time $t_{100}$ based on observations $\{\boldsymbol{\xi}_{1:c_{100}}^{[i]}\}_{i=1}^4$ is shown in Fig. 3. It has a peak near the true $\boldsymbol{\theta}$ which show the correctness of the distributed algorithm. The predicted field compared with the true field at time $t_{100}$ is shown in Fig.4. Due to the construction of sub-regions, the interface areas between any of two sub-regions are not predicted. Notice that the prediction is not as good as in the 1-D scenario due to the effect of curse of dimensionality when we move from 1-D to 2-D spaces. The running time of the distributed algorithm in this scenario is about several minutes due to the complexity of the 2-D problem under the same computational environment as the one used for the 1-D scenario. Thanks to



(a) (b) (c)

Fig. 2. (a) The updated prior distribution of $\boldsymbol{\theta}$ at different time using the centralized algorithm. (a) $\pi(\boldsymbol{\theta})$, (b) $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_{100}})$, and (c) $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_{300}})$.

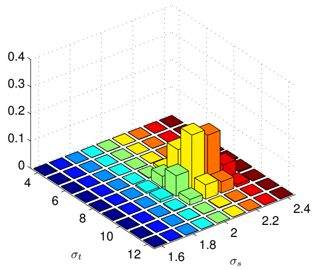
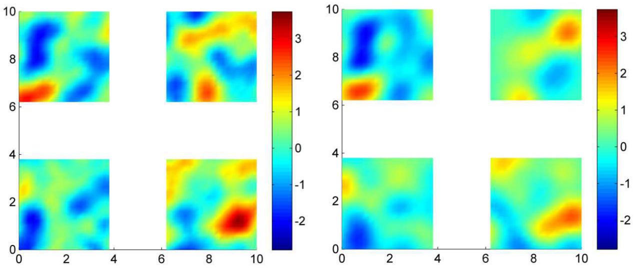


Fig. 3. The updated prior distribution of $\boldsymbol{\theta}$ at time $t_{100}$, i.e., $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}_{1:c_{100}})$, using the distributed algorithm.



(a) (b)

Fig. 4. The comparison of (a) the true field at $t_{100}$ and (b) the predicted field at $t_{100}$ using the distributed algorithm.

our proposed schemes, the running time does not grow with the time increases.

## VII. Conclusion

In this paper, we formulated a full Bayesian approach for spatio-temporal Gaussian process regression under practical conditions such as measurement noise and unknown hyperparmeters. We designed sequential Bayesian prediction algorithms for spatio-temporal Gaussian processes that can be implemented in constant time as the number of observations increases. An adaptive sampling strategy was also provided in order to improve the quality of prediction. Simulation results showed the effectiveness of the proposed algorithms in the context of environmental monitoring by mobile sensor networks.

## Acknowledgment

This work has been supported by the National Science Foundation through CAREER Award CMMI-0846547. This support is gratefully acknowledged.

## References

[1] K. M. Lynch, I. B. Schwartz, P. Yang, and R. A. Freeman, "Decentralized environmental modeling by mobile sensor networks," *IEEE Transactions on Robotics*, vol. 24, no. 3, pp. 710–724, June 2008.

[2] N. E. Leonard, D. A. Paley, F. Lekien, R. Sepulchre, D. M. Fratantoni, and R. Davis, "Collective motion, sensor networks, and ocean sampling," *Proceedings of the IEEE*, vol. 95, no. 1, January 2007.

[3] J. Choi, S. Oh, and R. Horowitz, "Distributed learning and cooperative control for multi-agent systems," *Automatica*, vol. 45, pp. 2802–2814, 2009.

[4] J. Cortés, "Distributed Kriged Kalman filter for spatial estimation," *IEEE Transactions on Automatic Control*, vol. 54, no. 12, pp. 2816–2827, 2010.

[5] N. Cressie, "Kriging nonstationary data," *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 625–634, 1986.

[6] D. J. C. MacKay, "Introduction to Gaussian processes," *NATO ASI Series F Computer and Systems Sciences*, vol. 168, pp. 133–165, 1998.

[7] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. The MIT Press, Cambridge, Massachusetts, London, England, 2006.

[8] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in Gaussian processes: theory, efficient algorithms and empirical studies," *The Journal of Machine Learning Research*, vol. 9, pp. 235–284, 2008.

[9] J. Choi, J. Lee, and S. Oh, "Swarm intelligence for achieving the global maximum using spatio-temporal Gaussian processes," in *Proceedings of the 27th American Control Conference (ACC)*, 2008.

[10] ——, "Biologically-inspired navigation strategies for swarm intelligence using spatial Gaussian processes," in *Proceedings of the 17th International Federation of Automatic Control (IFAC) World Congress*, 2008.

[11] Y. Xu and J. Choi, "Mobile sensor networks for learning anisotropic Gaussian processes," in *Proceedings of the 2009 American Control Conference (ACC)*, 2009, pp. 5049–5054.

[12] C. M. Bishop, *Pattern recognition and machine learning*. Springer, New York, 2006.

[13] M. Gaudard, M. Karson, E. Linder, and D. Sinha, "Bayesian spatial prediction," *Environmental and Ecological Statistics*, vol. 6, no. 2, pp. 147–171, 1999.

[14] R. Graham and J. Cortés, "Cooperative adaptive sampling of random fields with partially known covariance," *International Journal of Robust and Nonlinear Control*, vol. 1, pp. 1–2, 2009.

[15] S. Oh, Y. Xu, and J. Choi, "Explorative navigation of mobile sensor networks using sparse Gaussian processes," in *Proceedings of the 49th IEEE Conference on Decision and Control (CDC)*, 2010.

[16] Y. Xu, J. Choi, and S. Oh, "Mobile sensor setwork coordination using Gaussian processes with truncated observations," *IEEE Transactions on on Robotics*, 2010, conditionally accepted as a Regular Paper.

[17] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, no. 1, pp. 5–43, 2003.

[18] T. Gneiting, "Compactly supported correlation functions," *Journal of Multivariate Analysis*, vol. 83, no. 2, pp. 493–508, 2002.

[19] R. Olfati-Saber, R. Franco, E. Frazzoli, and J. S. Shamma, "Belief consensus and distributed hypothesis testing in sensor networks," *Networked Embedded Sensing and Control*, pp. 169–182, 2006.