Sequential 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 fully Bayesian approach for spatio-temporal Gaussian process regression such that multifactorial effects of observations, measurement noise and prior distributions are all correctly incorporated in the predictive distribution. Using discrete prior probabilities and compactly supported kernels, we provide a way to design sequential Bayesian prediction algorithms in which exact predictive distributions can be computed in constant time as the number of observations increases. For a special case, a distributed implementation of sequential Bayesian prediction algorithms has 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 practical usefulness of the proposed theoretically-correct algorithms.

I. Introduction

Recently, there has been an increasing exploitation of mobile sensor networks in environmental monitoring [1]–[4]. Gaussian process regression (or kriging in geostatistics) has been widely used to draw statistical inference from geostatistical and environmental data [5], [6]. For example, near-optimal static sensor placements with a mutual information criterion in Gaussian processes were proposed in [7]. 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 [8], [9].

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 & Probability, Michigan State University, East Lansing, MI 48824, USA. E-mails: {sdass, maiti}@msu.edu.

The significant computational complexity in Gaussian process regression due to the growing number of observations (and hence the size of covariance matrix) has been tackled in different ways. In [10], the authors analyzed the conditions under which near-optimal prediction can be achieved using only truncated observations. This motivates the usage of sparse Gaussian process proposed in [11]. However, they both assumed the covariance function is known *a priori*, which is unrealistic in practice. On the other hand, unknown parameters in the covariance function can be estimated by the maximum likelihood (ML) estimator. Such ML estimates may be regarded as the true parameters and then used in the prediction [12]. However, the point estimate itself needs to be identified using sufficient amount of measurements. Instead, a maximum a posterior (MAP) estimate can use the prior to provide the point estimate with a small number of measurements. However, it fails to incorporate the uncertainty in the estimate into the prediction.

The advantage of a fully Bayesian approach, which will be adopted in this work, is that the uncertainty in the model parameters are incorporated in the prediction [13]. In [14], 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 unknown. However, the assumptions made in [14], such as noiseless observations and time-invariance of the field, limit the applicability of the approach on mobile sensors in practice. The computational complexity of a fully Bayesian prediction algorithm has been the main hurdle for applications in resource-constrained robots. In [15], an iterative prediction algorithm without resorting to Markov Chain Monte Carlo (MCMC) methods has been developed based on analytical closed-form solutions from results in [14], by assuming that the covariance function of the spatio-temporal Gaussian random field is known up to a constant. Our work builds on such Bayesian approaches used in [14], [15] and explores new ways to synthesize practical algorithms for mobile sensor networks under more relaxed conditions.

The contributions of this paper are as follows. First, we provide a fully Bayesian approach for spatio-temporal Gaussian process regression under more practical conditions such as measurement noise and the unknown covariance function (Section III). In this way, multifactorial effects of observations, measurement noise, the noninformative prior on regression coefficients, and prior distributions of parameters are all correctly incorporated in the prediction. Using discrete prior probabilities and compactly supported kernels [16], we provide a way to design sequential Bayesian prediction algorithms in which the exact predictive distributions can be computed in

constant time as the number of observations increases. In particular, a centralized sequential Bayesian prediction algorithm is developed (Section IV-A) and its distributed implementation among sensor groups is provided for a special case (Section IV-B). To the best of our knowledge, no such exact sequential Bayesian prediction algorithms under our practical and relaxed conditions have been found to date. An adaptive sampling strategy for mobile sensors, utilizing the maximum a posteriori (MAP) estimation of the parameters, is proposed to minimize the prediction error variances (Section IV-C). Finally, the proposed sequential Bayesian prediction algorithms and the adaptive sampling strategy are tested under practical conditions for spatio-temporal Gaussian processes (Section V).

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 integer, and positive integer numbers. Let E, Var and Corr denote, respectively, the operators of expectation, variance and correlation. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\mathbf{I}$ and $\mathbf{0}$ denote, respectively, the identity and zero matrices with appropriate dimensions. Other notation will be explained in due course.

II. PRELIMINARIES

Let $z(\mathbf{s}, \tau)$ be the spatio-temporal field of interest (e.g., water temperature of a lake) at location $\mathbf{s} \in \mathcal{Q} \subset \mathbb{R}^D$ and time $\tau \in \mathbb{R}_{\geq 0}$ modeled by a spatio-temporal Gaussian process denoted by

$$z(\mathbf{x}) \sim \mathcal{GP}\left(\mu(\mathbf{x}), \sigma_f^2 \mathcal{K}(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})\right), \tag{1}$$

where we have defined $\mathbf{x} := (\mathbf{s}^T, \tau)^T$ for notational simplicity. The mean function is assumed to be $\mu(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T \boldsymbol{\beta}$, where $\mathbf{f}(\mathbf{x}) := (f_1(\mathbf{x}), \cdots, f_p(\mathbf{x}))^T \in \mathbb{R}^p$ is a known (multivariate) regression function of $\mathbf{x}$, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector of regression coefficients. The term σ_f^2 denotes the signal variance which gives the overall vertical scale relative to the mean of the Gaussian process in the output space [6]. The correlation between $z(\mathbf{x})$ and $z(\mathbf{x}')$ is given by $\mathcal{K}(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})$ in (1). In this paper, we model the correlation function $\mathcal{K}(\cdot, \cdot)$ as

$$\mathcal{K}(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) = \phi_s\left(\frac{\|\mathbf{s} - \mathbf{s}'\|}{\sigma_s}\right) \phi_t\left(\frac{|\tau - \tau'|}{\sigma_t}\right), \tag{2}$$

where $\boldsymbol{\theta} := (\sigma_s, \sigma_t)^T \in \mathbb{R}^2$, governed by the product of spatial and temporal distance functions $\phi_s(\cdot)$ and $\phi_t(\cdot)$ which are decreasing kernel functions over space and time, respectively. The rate

of decrease depends on the spatial and time bandwidth σ_s, and σ_t, respectively. The tensor product form used in (2) is suggested in [6] and the references therein. In this paper, we consider the class of spatio-temporal Gaussian processes generated by $\phi_t(\cdot)$ that is compactly supported; hence, temporal correlations vanish when the time difference $|\tau - \tau'|$ is larger than σ_t (i.e., $\phi_t(h) = 0, \forall h > 1$). Subsequently we show that compactly supported $\phi_t(\cdot)$ is crucial for developing exact sequential Bayesian approaches justifying its choice here.

Let N mobile sensing agents be distributed over the surveillance region $\mathcal{Q}$ with labels in the set $\mathcal{I} := \{1, 2, \cdots, N\}$. Assume that mobile sensing agents are equipped with identical sensors. At time $t \in \mathbb{Z}_{>0}$, agent i makes a point observation of the spatio-temporal field of interest $z(\mathbf{q}_i(t), t)$ at it's position $\mathbf{q}_i(t) \in \mathcal{Q}$. The noise corrupted observation of $z(\mathbf{q}_i(t), t)$ is

$$y_i(t) := z(\mathbf{q}_i(t), t) + \epsilon_i, \tag{3}$$

where ϵ_i is the random sensor noise considered to be independent and identically distributed according to $\mathcal{N}(0, \sigma_w^2)$ with unknown variance $\sigma_w^2 > 0$. We assume that the *signal-to-noise ratio* $\gamma = \sigma_f^2/\sigma_w^2$ is known and fixed, which is necessary for identifiability of the model that gives rise to the noise corrupted observations in (3).

III. A Fully Bayesian Prediction Approach

Given the collection of noise corrupted observations from mobile sensing agents up to time t, we want to predict $z(\mathbf{s}_*, t_*)$ at a prespecified location $\mathbf{s}_* \in \mathcal{S} \subset \mathcal{Q}$ and current (or future) time t_*. To do this, suppose we have a collection of n observations $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)}) \,|\, i = 1, \ldots, n\}$ from N mobile sensing agents up to time t. Here $\mathbf{x}^{(i)}$ denotes the i-th input vector of dimension $D + 1$ (i.e., the sampling position and time of the i-th observation) and $y^{(i)}$ denotes the i-th noise corrupted measurement. If all observations are considered, we have $n = tN$. Notice that the number of observations n grows with the time t. For notational simplicity, let $\mathbf{y} := (y^{(1)}, \cdots, y^{(n)})^T \in \mathbb{R}^n$ denote the collection of noise corrupted observations. Based on the spatio-temporal Gaussian process, the distribution of the observations given the parameters $\boldsymbol{\beta}$, σ_f^2, and $\boldsymbol{\theta}$ is Gaussian, i.e., $\mathbf{y}|\boldsymbol{\beta}, \sigma_f^2, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{F}\boldsymbol{\beta}, \sigma_f^2\mathbf{C})$ with $\mathbf{F}$ and $\mathbf{C}$ defined as

$$\mathbf{F} := \begin{bmatrix} \mathbf{f}(\mathbf{x}^{(1)}) & \cdots & \mathbf{f}(\mathbf{x}^{(n)}) \end{bmatrix}^T \in \mathbb{R}^{n \times p}, \quad \mathbf{C} := \mathrm{Corr}(\mathbf{y}, \mathbf{y}|\boldsymbol{\theta}) = \left[\mathcal{K}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}; \boldsymbol{\theta}) + \frac{1}{\gamma}\delta_{ij} \right] \in \mathbb{R}^{n \times n}, \tag{4}$$

where δ_{ij} is the Kronecker delta which equals to one when $i = j$, and zero, otherwise.

A. Prior selection

To infer the unknown parameters $\boldsymbol{\beta}$, σ_f^2, and $\boldsymbol{\theta}$ in a Bayesian framework, the collection of them is considered to be a random vector with a *prior* distribution reflecting the *a priori* belief of uncertainty for them. In this paper, we use the prior distribution given by

$$\pi(\boldsymbol{\beta}, \sigma_f^2, \boldsymbol{\theta}) = \pi(\boldsymbol{\beta}|\sigma_f^2)\pi(\sigma_f^2)\pi(\boldsymbol{\theta}), \tag{5}$$

where $\boldsymbol{\beta}|\sigma_f^2 \sim \mathcal{N}(\boldsymbol{\beta}_\mathbf{0}, \sigma_f^2\mathbf{T})$. The prior for $\pi(\sigma_f^2)$ is taken to be the *inverse gamma distribution*, chosen to guarantee positiveness of σ_f^2 and a closed-form expression for the posterior distribution of σ_f^2 for computational ease of the proposed algorithms. To cope with the case where no prior knowledge on $\boldsymbol{\beta}$ is available, which is often the case in practice, we propose to use a noninformative prior. In particular, we take $\boldsymbol{\beta}_0 = \mathbf{0}$, $\mathbf{T} = \alpha\mathbf{I}$, and subsequently, let $\alpha \to \infty$. Any proper prior $\pi(\boldsymbol{\theta})$ that correctly reflects the prior knowledge of $\boldsymbol{\theta}$ can be used.

B. Posterior predictive distribution

The *posterior predictive distribution* of $z_* := z(\mathbf{s}_*, t_*)$ can be written as

$$p(z_*|\mathbf{y}) = \int p(z_*|\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \tag{6}$$

where $\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$, is the posterior distribution of $\boldsymbol{\theta}$, by integrating out analytically the parameters $\boldsymbol{\beta}$ and σ_f^2. We have the following proposition.

Proposition 3.1: For a prior distribution given in (5) with the noninformative prior on $\boldsymbol{\beta}$, we have

i) $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto w(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta})$ with

$$\log w(\boldsymbol{\theta}|\mathbf{y}) = -\frac{1}{2}\log\det(\mathbf{C}) - \frac{1}{2}\log\det(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F}) - \tilde{a}\log\tilde{b}, \tag{7}$$

where $\tilde{a} = a + \frac{n}{2}$, and $\tilde{b} = b + \frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y} - \frac{1}{2}(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{y})^T(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F})^{-1}(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{y})$.

ii) $p(z_*|\mathbf{y}, \boldsymbol{\theta})$ is a shifted student's t-distribution with location parameter μ, scale parameter λ, and ν degrees of freedom, i.e.,

$$p(z_*|\mathbf{y}, \boldsymbol{\theta}) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}}\left(1 + \frac{\lambda(z_*-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{8}$$

where $\nu = 2\tilde{a}$, and

$$\mu = \mathbf{k}^T\mathbf{C}^{-1}\mathbf{y} + (\mathbf{f}(\mathbf{x}_*) - \mathbf{F}^T\mathbf{C}^{-1}\mathbf{k})^T(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F})^{-1}(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{y}),$$

$$\lambda = \frac{\tilde{b}}{\tilde{a}}\big((1 - \mathbf{k}^T\mathbf{C}^{-1}\mathbf{k}) + (\mathbf{f}(\mathbf{x}_*) - \mathbf{F}^T\mathbf{C}^{-1}\mathbf{k})^T(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F})^{-1}(\mathbf{f}(\mathbf{x}_*) - \mathbf{F}^T\mathbf{C}^{-1}\mathbf{k})\big).$$

Proof: See Appendix A. ■

The results in Proposition 3.1 are different from those obtained in [14] by using a noninformative prior on $\boldsymbol{\beta}$. For a special case where $\boldsymbol{\beta}$ and σ_f^2 are known *a priori*, we have the following corollary which will be exploited to derive a distributed implementation among sensor groups in Section IV-B.

Corollary 3.2: In the case where $\boldsymbol{\beta}$ and σ_f^2 are known a priori, (7) and (8) can be simplified as

$$\log w(\boldsymbol{\theta}|\mathbf{y}) = -\frac{1}{2}\log\det(\mathbf{C}) - \frac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^T\mathbf{C}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}),$$

$$z_*|\mathbf{y}, \boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{f}(\mathbf{x}_*)^T\boldsymbol{\beta} + \mathbf{k}^T\mathbf{C}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}), \sigma_f^2(1 - \mathbf{k}^T\mathbf{C}^{-1}\mathbf{k})\right).$$

If we draw m samples $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^m$ from the prior distribution $\pi(\boldsymbol{\theta})$, the posterior predictive distribution in (6) can then be approximated by

$$p(z_*|\mathbf{y}) \approx \frac{\sum w(\boldsymbol{\theta}^{(i)}|\mathbf{y})p(z_*|\mathbf{y}, \boldsymbol{\theta}^{(i)})}{\sum w(\boldsymbol{\theta}^{(i)}|\mathbf{y})}.$$

It follows that the predictive mean and variance can be obtained by

$$\mathrm{E}(z_*|\mathbf{y}) \approx \frac{\sum w(\boldsymbol{\theta}^{(i)}|\mathbf{y})\mathrm{E}(z_*|\mathbf{y}, \boldsymbol{\theta}^{(i)})}{\sum w(\boldsymbol{\theta}^{(i)}|\mathbf{y})},$$

$$\mathrm{Var}(z_*|\mathbf{y}) \approx \frac{\sum w(\boldsymbol{\theta}^{(i)}|\mathbf{y})\mathrm{Var}(z_*|\mathbf{y}, \boldsymbol{\theta}^{(i)})}{\sum w(\boldsymbol{\theta}^{(i)}|\mathbf{y})} + \frac{\sum w(\boldsymbol{\theta}^{(i)}|\mathbf{y})\left(\mathrm{E}(z_*|\mathbf{y}, \boldsymbol{\theta}^{(i)}) - \mathrm{E}(z_*|\mathbf{y})\right)^2}{\sum w(\boldsymbol{\theta}^{(i)}|\mathbf{y})},$$

where the mean and variance of the student's t-distribution $p(z_*|\mathbf{y}, \boldsymbol{\theta})$ are given by $\mathrm{E}(z_*|\mathbf{y}, \boldsymbol{\theta}) = \mu$, and $\mathrm{Var}(z_*|\mathbf{y}, \boldsymbol{\theta}) = \frac{\tilde{a}}{\tilde{a}-1}\lambda$, respectively.

C. Further simplification

To further reduce the computational demands from the Monte Carlo approach, we assign discrete uniform probability distributions to σ_s and σ_t as priors instead of continuous probability distributions. Assume that we know the range of parameters in $\boldsymbol{\theta}$, i.e., $\sigma_s \in [\underline{\sigma}_s, \overline{\sigma}_s]$ and $\sigma_t \in [\underline{\sigma}_t, \overline{\sigma}_t]$, where $\underline{\sigma}$ and $\overline{\sigma}$ denote the known lower-bound and upper-bound of the random variable σ, respectively. We constrain the possible choices of $\boldsymbol{\theta}$ on a finite set of grid points denoted

by Θ. Hence, $\pi(\boldsymbol{\theta})$ is now a probability mass function (i.e., $\sum_{\boldsymbol{\theta}\in\Theta}\pi(\boldsymbol{\theta})=1$) as opposed to a probability density. The integration in (6) is reduced to the following summation

$$p(z_*|\mathbf{y})=\sum_{\boldsymbol{\theta}\in\Theta}p(z_*|\mathbf{y},\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}), \tag{9}$$

where the posterior distribution of $\boldsymbol{\theta}$ is evaluated on the grid points in Θ by

$$\pi(\boldsymbol{\theta}|\mathbf{y})=\frac{w(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}\in\Theta}w(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta})}. \tag{10}$$

In order to obtain the posterior predictive distribution in (9), the computation of $p(z_*|\mathbf{y},\boldsymbol{\theta})$ and $w(\boldsymbol{\theta}|\mathbf{y})$ for all $\boldsymbol{\theta}\in\Theta$ using the results from Proposition 3.1 (or Corollary 3.2 for a special case) are necessary. Note that these quantities are available in closed-form which reduces the computational burden significantly.

IV. Sequential Bayesian prediction algorithms for mobile sensor networks

Although the aforementioned efforts in Sections III-B and III-C reduce the computational cost significantly, the number of observations (that mobile sensing agents collect) n increases with the time t. For each $\boldsymbol{\theta}\in\Theta$, an $n\times n$ positive definite matrix $\mathbf{C}$ needs to be inverted which requires time $O(n^3)$ using standard methods. This motivates us to design scalable sequential Bayesian prediction algorithms by using subsets of observations.

A. A scalable Bayesian prediction algorithm

Let $\mathbf{y}_t\in\mathbb{R}^N$ be the collection of noise corrupted observations by all agents at time t, i.e., $\mathbf{y}_t:=(y_1(t),\cdots,y_N(t))^T$, and let $\mathbf{y}_{1:t}\in\mathbb{R}^{tN}$ be the cumulative observations, i.e., $\mathbf{y}_{1:t}:=(\mathbf{y}_1^T,\cdots,\mathbf{y}_t^T)^T$. The computation of $p(z_*|\mathbf{y}_{1:t})$ soon becomes infeasible as t increases. To overcome this drawback while maintaining the Bayesian framework, we propose to use subsets of all observations $\mathbf{y}_{1:t}$. However, instead of using truncated local observations only as in [10], Bayesian inference will be drawn based on two sets of observations: First, a set of local observations near target points $\tilde{\mathbf{y}}$ which will improve the quality of the prediction, and a second cumulative set of observations $\bar{\mathbf{y}}$ which will minimize the uncertainty in the estimated parameters. Taken together, they improve the quality of prediction as the number of observations increases. We formulate this idea in detail in the following paragraph. For notational simplicity, we define $\mathbf{y}$ as a subset of all observations $\mathbf{y}_{1:t}$ which will be used for Bayesian prediction. We partition $\mathbf{y}$ into two

subsets, namely $\bar{\mathbf{y}}$ and $\tilde{\mathbf{y}}$. Let $\bar{\mathbf{F}}$ and $\tilde{\mathbf{F}}$ be the counterparts of $\mathbf{F}$ defined in (4) for $\bar{\mathbf{y}}$ and $\tilde{\mathbf{y}}$, respectively. The following lemma provides the conditions under which any required function of $\mathbf{y}$ in Proposition 3.1 can be decoupled.

Lemma 4.1: For a given $\boldsymbol{\theta} \in \Theta$, let $\mathbf{C} = \text{Corr}(\mathbf{y}, \mathbf{y}|\boldsymbol{\theta})$, $\bar{\mathbf{C}} = \text{Corr}(\bar{\mathbf{y}}, \bar{\mathbf{y}}|\boldsymbol{\theta})$, $\tilde{\mathbf{C}} = \text{Corr}(\tilde{\mathbf{y}}, \tilde{\mathbf{y}}|\boldsymbol{\theta})$, $\mathbf{k} = \text{Corr}(\mathbf{y}, z_*|\boldsymbol{\theta})$, $\bar{\mathbf{k}} = \text{Corr}(\bar{\mathbf{y}}, z_*|\boldsymbol{\theta})$, and $\tilde{\mathbf{k}} = \text{Corr}(\tilde{\mathbf{y}}, z_*|\boldsymbol{\theta})$. If the following conditions are satisfied

$C1$: $\text{Corr}(\tilde{\mathbf{y}}, \bar{\mathbf{y}}|\boldsymbol{\theta}) = \mathbf{0}$, i.e., $\tilde{\mathbf{y}}$ and $\bar{\mathbf{y}}$ are uncorrelated, and

$C2$: $\text{Corr}(\bar{\mathbf{y}}, z_*|\boldsymbol{\theta}) = \mathbf{0}$, i.e., $\bar{\mathbf{y}}$ and z_* are uncorrelated,

then we have the following results:

$$\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F} = \bar{\mathbf{F}}^T\bar{\mathbf{C}}^{-1}\bar{\mathbf{F}} + \tilde{\mathbf{F}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{F}} \in \mathbb{R}^{p\times p}, \quad \mathbf{F}^T\mathbf{C}^{-1}\mathbf{y} = \bar{\mathbf{F}}^T\bar{\mathbf{C}}^{-1}\bar{\mathbf{y}} + \tilde{\mathbf{F}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{y}} \in \mathbb{R}^p,$$

$$\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y} = \bar{\mathbf{y}}^T\bar{\mathbf{C}}^{-1}\bar{\mathbf{y}} + \tilde{\mathbf{y}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{y}} \in \mathbb{R}, \quad \log\det\mathbf{C} = \log\det\bar{\mathbf{C}} + \log\det\tilde{\mathbf{C}} \in \mathbb{R},$$

$$\mathbf{F}^T\mathbf{C}^{-1}\mathbf{k} = \tilde{\mathbf{F}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{k}} \in \mathbb{R}^p, \quad \mathbf{k}^T\mathbf{C}^{-1}\mathbf{k} = \tilde{\mathbf{k}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{k}} \in \mathbb{R}.$$

Proof: The results follow by noting the correlation matrix $\mathbf{C}$ can be decoupled such that $\mathbf{C} = \text{diag}(\bar{\mathbf{C}}, \tilde{\mathbf{C}})$ and $\bar{\mathbf{k}} = \mathbf{0}$. ■

Remark 4.2: In order to compute the posterior predictive distribution $p(z_*|\mathbf{y})$ (or the predictive mean and variance) in (9), $p(z_*|\mathbf{y}, \boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\mathbf{y})$ for all $\boldsymbol{\theta} \in \Theta$ need to be calculated. Notice that the posterior distribution of $\boldsymbol{\theta}$ can be obtained by computing $w(\boldsymbol{\theta}|\mathbf{y})$ in (7). Suppose $\bar{\mathbf{F}}^T\bar{\mathbf{C}}^{-1}\bar{\mathbf{F}} \in \mathbb{R}^{p\times p}$, $\bar{\mathbf{F}}^T\bar{\mathbf{C}}^{-1}\bar{\mathbf{y}} \in \mathbb{R}^p$, $\bar{\mathbf{y}}^T\bar{\mathbf{C}}^{-1}\bar{\mathbf{y}} \in \mathbb{R}$, and $\log\det\bar{\mathbf{C}} \in \mathbb{R}$ are known for all $\boldsymbol{\theta} \in \Theta$. If $\tilde{\mathbf{F}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{F}} \in \mathbb{R}^{p\times p}$, $\tilde{\mathbf{F}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{y}} \in \mathbb{R}^p$, $\tilde{\mathbf{y}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{y}} \in \mathbb{R}$, and $\log\det\tilde{\mathbf{C}} \in \mathbb{R}$ for all $\boldsymbol{\theta} \in \Theta$ have fixed computation times, then (7) and (8) can be computed in constant time due to decoupling results of Lemma 4.1.

The following theorem provides a way to design scalable sequential Bayesian prediction algorithms.

Theorem 4.3: Consider the discrete prior probability $\pi(\boldsymbol{\theta})$ and the compactly supported kernel function $\phi_t(\cdot)$. If we select $\eta \geq \lfloor \overline{\sigma}_t/t_s \rfloor \in \mathbb{Z}_{>0}$, $\Delta \in \mathbb{Z}_{>0}$ and define

$$\begin{aligned} c_t &:= \max\left(\left\lfloor \frac{t-\Delta}{\Delta+\eta} \right\rfloor, 0\right) \in \mathbb{R}, \quad \boldsymbol{\xi}_j := \mathbf{y}_{(j-1)(\Delta+\eta)+1:(j-1)(\Delta+\eta)+\Delta} \in \mathbb{R}^{\Delta N}, \\ \bar{\mathbf{y}} &:= (\boldsymbol{\xi}_1^T, \cdots, \boldsymbol{\xi}_{c_t}^T)^T \in \mathbb{R}^{\Delta N c_t}, \quad \tilde{\mathbf{y}} := \mathbf{y}_{t-\Delta+1:t} \in \mathbb{R}^{\Delta N}, \end{aligned} \tag{11}$$

where $\lfloor\cdot\rfloor$ is the floor function defined by $\lfloor x \rfloor := \max\{m \in \mathbb{Z} \,|\, m \leq x\}$, then the posterior predictive distribution in (9) can be computed in constant time (i.e., does not grow with the time t).

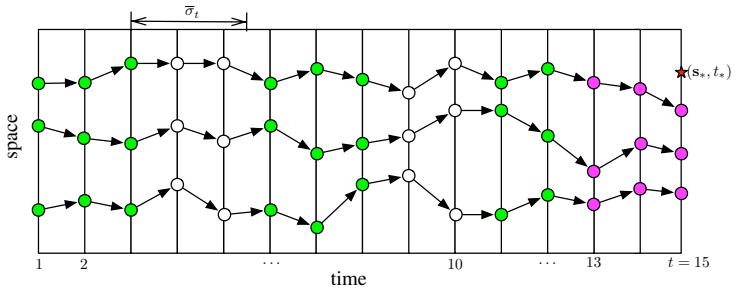


Fig. 1. An example with three agents sampling the spatio-temporal Gaussian process in 1-D space and performing Bayesian inference. In this example, $\overline{\sigma}_t = 2.5$, $\eta = 2$, $\Delta = 3$, $t = 15$, $c_t = 2$, $\bar{\mathbf{y}} = (\mathbf{y}_{1:3}^T, \mathbf{y}_{6:8}^T)^T$ and $\tilde{\mathbf{y}} = \mathbf{y}_{13:15}$.

Proof: By construction, conditions $C1$-2 in Lemma 4.1 are satisfied. Hence, it follows from Remark 4.2 that the posterior predictive distribution can be computed in constant time. ■

Remark 4.4: In Theorem 4.3, $\eta \geq \lfloor \overline{\sigma}_t / t_s \rfloor$ guarantees the time distance between $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_{i+1}$ is large enough such that the conditions in Lemma 4.1 are satisfied. Notice that Δ is a tuning parameter for users to control the trade-off between the prediction quality and the computation efficiency. A large value for Δ yields a small predictive variance but long computation time, and vice versa. An illustrative example with three agents sampling the spatio-temporal Gaussian process in 1-D space is shown in Fig. 1.

Based on Theorem 4.3, we provide the centralized sequential Bayesian prediction algorithm as shown in Table I.

B. A distributed implementation for a special case

In this subsection, we will show a distributed way (among agent groups) to implement the proposed algorithm for a special case in which $\boldsymbol{\beta}$ and σ_f^2 are assumed to be known *a priori*. The assumption for this special case is the exact opposite of the one made in [15] where $\boldsymbol{\beta}$ and σ_f^2 are unknown and $\boldsymbol{\theta}$ is known *a priori*.

To develop a distributed scheme among agent groups for data fusion in Bayesian statistics, we exploit the compactly supported kernel for space. Let $\phi_s(h)$ in (2) also be a compactly supported kernel function as $\phi_t(h)$ so that the correlation vanishes when the spatial distance between two inputs is larger than σ_s, i.e., $\phi_s(h) = 0, \forall h > 1$.

Consider a case in which M groups of spatially distributed agents sample a spatio-temporal

TABLE I

A CENTRALIZED BAYESIAN PREDICTION ALGORITHM.

Input: (1) prior distribution on σ_f^2, i.e., $\pi(\sigma_f^2) = \text{IG}(a, b)$; (2) prior distribution on $\boldsymbol{\theta} \in \Theta$, i.e., $\pi(\boldsymbol{\theta})$; (3) tuning variables Δ and η; (4) number of agents N; (5) $\mathcal{M}(\boldsymbol{\theta}).A = 0 \in \mathbb{R}^{p\times p}$, $\mathcal{M}(\boldsymbol{\theta}).B = 0 \in \mathbb{R}$, $\mathcal{M}(\boldsymbol{\theta}).C = 0 \in \mathbb{R}^p$, $\mathcal{M}(\boldsymbol{\theta}).D = 0 \in \mathbb{R}$, $\mathcal{M}_0(\boldsymbol{\theta}) = \mathcal{M}(\boldsymbol{\theta})$, $\forall \boldsymbol{\theta} \in \Theta$

Output: (1) The predictive mean at location $\mathbf{s}_* \in \mathcal{S}$ and time $t_* = t$, i.e., $\text{E}(z_*|\mathbf{y})$; (2) The predictive variance at location $\mathbf{s}_* \in \mathcal{S}$ and time $t_* = t$, i.e., $\text{Var}(z_*|\mathbf{y})$

At time t, the central station does:

1: receive observations $\mathbf{y}_t$ from agents, set $\tilde{\mathbf{y}} = \mathbf{y}_{t-\Delta+1:t}$ and $n = N\Delta$

2: compute $\tilde{\mathbf{F}} = (\mathbf{f}(\tilde{\mathbf{x}}^{(1)}), \cdots, \mathbf{f}(\tilde{\mathbf{x}}^{(n)}))^T \in \mathbb{R}^{n\times p}$ where $\tilde{\mathbf{x}}^{(i)}$ is the input of the i-th element in $\tilde{\mathbf{y}}$

3: **for** each $\boldsymbol{\theta} \in \Theta$ **do**

4: compute $\tilde{\mathbf{C}} = \text{Corr}(\tilde{\mathbf{y}}, \tilde{\mathbf{y}}) \in \mathbb{R}^{n\times n}$

5: compute the key values

$$\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F} = \mathcal{M}(\boldsymbol{\theta}).A + \tilde{\mathbf{F}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{F}} \in \mathbb{R}^{p\times p}, \quad \mathbf{y}^T\mathbf{C}^{-1}\mathbf{y} = \mathcal{M}(\boldsymbol{\theta}).B + \tilde{\mathbf{y}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{y}} \in \mathbb{R}$$

$$\mathbf{F}^T\mathbf{C}^{-1}\mathbf{y} = \mathcal{M}(\boldsymbol{\theta}).C + \tilde{\mathbf{F}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{y}} \in \mathbb{R}^p, \quad \log\det\mathbf{C} = \mathcal{M}(\boldsymbol{\theta}).D + \log\det\tilde{\mathbf{C}} \in \mathbb{R}$$

6: compute $\tilde{a} = a + \frac{n}{2}$ and $\tilde{b} = b + \frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y} - \frac{1}{2}(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{y})^T(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F})^{-1}(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{y})$

7: update weights via

$$\log w(\boldsymbol{\theta}|\mathbf{y}) = -\frac{1}{2}\log\det\mathbf{C} - \frac{1}{2}\log\det(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F}) - \tilde{a}\log\tilde{b}$$

8: **for** each $\mathbf{s}_* \in \mathcal{S}$ **do**

9: compute $\mathbf{f}(\mathbf{x}_*) \in \mathbb{R}^p$, $\tilde{\mathbf{k}} = \text{Corr}(\tilde{\mathbf{y}}, z_*) \in \mathbb{R}^n$

10: compute predictive mean and variance for given $\boldsymbol{\theta}$

$$\text{E}(z_*|\mathbf{y}, \boldsymbol{\theta}) = \tilde{\mathbf{k}}\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{y}} + (\mathbf{f}(\mathbf{x}_*) - \tilde{\mathbf{F}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{k}})^T(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F})^{-1}(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{y}),$$

$$\text{Var}(z_*|\mathbf{y}, \boldsymbol{\theta}) = \frac{\tilde{b}}{\tilde{a}-1}\left((1 - \tilde{\mathbf{k}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{k}}) + (\mathbf{f}(\mathbf{x}_*) - \tilde{\mathbf{F}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{k}})^T(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F})^{-1}(\mathbf{f}(\mathbf{x}_*) - \tilde{\mathbf{F}}^T\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{k}})\right)$$

11: **end for**

12: **if** $\text{mod}(t, \Delta + \eta) = \Delta$ **then**

13: set $\mathcal{M}(\boldsymbol{\theta}) = \mathcal{M}_0(\boldsymbol{\theta})$, then $\mathcal{M}_0(\boldsymbol{\theta}).A = \mathbf{F}^T\mathbf{C}^{-1}\mathbf{F}$, $\mathcal{M}_0(\boldsymbol{\theta}).B = \mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}$, $\mathcal{M}_0(\boldsymbol{\theta}).C = \mathbf{F}^T\mathbf{C}^{-1}\mathbf{y}$, and $\mathcal{M}_0(\boldsymbol{\theta}).D = \log\det\mathbf{C}$

14: **end if**

15: **end for**

16: compute the posterior distribution

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{w(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}} w(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta})}$$

17: compute the predictive mean and variance

$$\text{E}(z_*|\mathbf{y}) = \textstyle\sum_{\boldsymbol{\theta}} \text{E}(z_*|\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}),$$

$$\text{Var}(z_*|\mathbf{y}) = \textstyle\sum_{\boldsymbol{\theta}} \text{Var}(z_*|\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}) + \sum_{\boldsymbol{\theta}} \left(\text{E}(z_*|\mathbf{y}, \boldsymbol{\theta}) - \text{E}(z_*|\mathbf{y})\right)^2 \pi(\boldsymbol{\theta}|\mathbf{y}).$$

Gaussian process over a large region $\mathcal{Q}$. Each group is in charge of its sub-region of $\mathcal{Q}$. The identity of each group is indexed by $\mathcal{V} := \{1, \cdots, M\}$. Each agent in group i is indexed

by $\mathcal{I}^{[i]} := \{1, \cdots, N\}$. 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, the posterior distribution of $\boldsymbol{\theta}$ shall be updated correctly using all observations from all groups (or agents) in a distributed fashion.

Let $G(t) := (\mathcal{V}, \mathcal{E}(t))$ 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 $N_i(t) := \{j \in \mathcal{V} \,|\, (i, j) \in \mathcal{E}(t), j \neq i\}$. Let $\mathbf{a}^{[i]}$ denote the quantity as $\mathbf{a}$ in the centralized scheme for group i. We then have the following theorem.

Theorem 4.5: Assume that $\bar{\mathbf{y}}^{[i]}$ and $\tilde{\mathbf{y}}^{[i]}$ for leader i are selected accordingly to Theorem 4.3 in time-wise. Let $\tilde{\mathbf{y}}$ defined by $\tilde{\mathbf{y}} := ((\tilde{\mathbf{y}}^{[1]})^T, \cdots, (\tilde{\mathbf{y}}^{[M]})^T)^T$. If the following condition is satisfied

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

in the spatial domain, then the weights $w(\boldsymbol{\theta}|\mathbf{y})$, based on all observations from all agents, can be obtained from

$$\log w(\boldsymbol{\theta}|\mathbf{y}) = \log w(\boldsymbol{\theta}|\bar{\mathbf{y}}) + \sum_{i=1}^{M} \log w(\boldsymbol{\theta}|\tilde{\mathbf{y}}^{[i]}). \tag{12}$$

Proof: The result follows by noting $\text{Corr}(\tilde{\mathbf{y}}^{[i]}, \tilde{\mathbf{y}}^{[j]}|\boldsymbol{\theta}) = 0, \forall i \neq j$, when the condition $C3$ is satisfied. ■

Suppose that the communication graph $G(t)$ is connected for all time t. Then $\frac{1}{M}\sum_{i=1}^{M} \log w(\boldsymbol{\theta}|\tilde{\mathbf{y}}^{[i]})$ can be achieved asymptotically via discrete-time *average-consensus algorithm* [17]:

$$\log w(\boldsymbol{\theta}|\tilde{\mathbf{y}}^{[i]}) \leftarrow \log w(\boldsymbol{\theta}|\tilde{\mathbf{y}}^{[i]}) + \epsilon \sum_{j \in N_i} \left(\log w(\boldsymbol{\theta}|\tilde{\mathbf{y}}^{[j]}) - \log w(\boldsymbol{\theta}|\tilde{\mathbf{y}}^{[i]})\right),$$

with $0 < \epsilon < 1/\Delta(G)$ that depends on the maximum node degree of the network $\Delta(G) = \max_i |N_i|$.

C. Adaptive sampling strategies

At time t, the goal of the navigation of agents is to improve the quality of prediction of the field $\mathcal{Q}$ at the next sampling time $t + 1$. Therefore, mobile agents should move to the most informative sampling locations $\{\mathbf{q}_1(t + 1), \cdots, \mathbf{q}_N(t + 1)\}$ at time $t + 1$ in order to reduce the prediction error [7].

Suppose at time $t+1$, 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

$$J(\{\tilde{\mathbf{q}}\}_{i=1}^N) = \int_{\mathbf{s}\in\mathcal{S}} \mathrm{E}\left[(z(\mathbf{s}, t+1) - \hat{z}(\mathbf{s}, t+1))^2\right] d\mathbf{s}, \tag{13}$$

where $\hat{z}(\mathbf{s}, t+1)$ is obtained as in (9). Due to the fact that $\boldsymbol{\theta}$ has a distribution, the evaluation of (13) becomes computationally prohibitive. To simplify the optimization, we propose to utilize a *maximum a posteriori* (MAP) estimate of $\boldsymbol{\theta}$ at time t, denoted by $\hat{\boldsymbol{\theta}}_{\text{MAP}}(t)$, i.e., $\hat{\boldsymbol{\theta}}_{\text{MAP}}(t) = \arg\max_{\boldsymbol{\theta}\in\Theta} \pi(\boldsymbol{\theta}|\mathbf{y})$, where $\mathbf{y}$ is the subset of all observations used up to time t. The next sampling positions can be obtained by solving the following optimization problem

$$\{\mathbf{q}_i(t+1)\}_{i=1}^N = \arg \min_{\{\tilde{\mathbf{q}}_i\}_{i=1}^N \subset \mathcal{Q}} \int_{\mathbf{s}\in\mathcal{S}} \text{Var}(z(\mathbf{s}, t+1)|\mathbf{y}, \hat{\boldsymbol{\theta}}_{\text{MAP}}(t)) d\mathbf{s}. \tag{14}$$

This problem can be solved using standard constrained nonlinear optimization techniques (e.g., the conjugate gradient algorithm), possibly taking into account mobility constraints of mobile sensors.

Remark 4.6: The proposed control algorithm in (14) is truly *adaptive* in the sense that the new sampling positions are functions of all collected observations. On the other hand, if all parameters are known, the optimization in (14) can be performed offline without taking any measurements.

V. Simulation results

In this section, we apply the proposed sequential Bayesian prediction algorithms to spatio-temporal Gaussian processes with a correlation function in (2). The Gaussian process was numerically generated through circulant embedding of the covariance matrix for the simulation study [18].

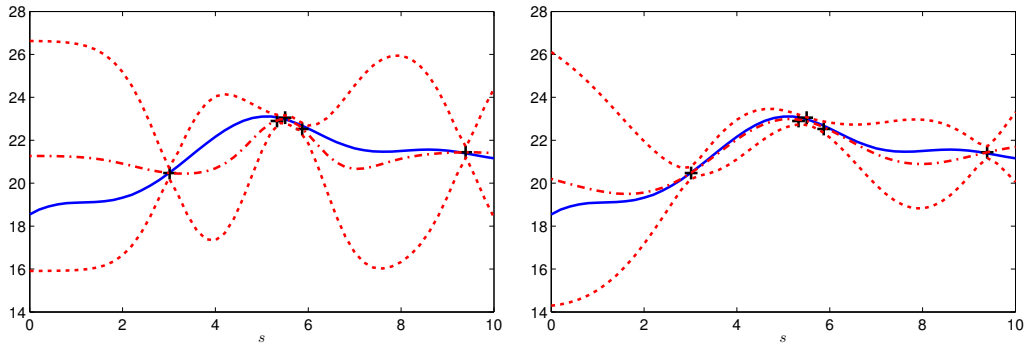
We consider a scenario in which $N = 5$ agents sample the spatio-temporal Gaussian process in 1-D space and the central station performs Bayesian prediction. The surveillance region $\mathcal{Q}$ is given by $\mathcal{Q} = [0, 10]$. We consider the *squared exponential* function $\phi_s(h) = \exp(-\frac{1}{2}h^2)$ for space correlation and a compactly supported correlation function [16] for time as

$$\phi_t(h) = \begin{cases} \frac{(1-h)\sin(2\pi h)}{2\pi h} + \frac{1-\cos(2\pi h)}{\pi \times 2\pi h}, & 0 \le h \le 1, \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

The signal-to-noise ratio γ is set to be 26dB which corresponds to $\sigma_w = 0.158$. The true values for the parameters used in simulating the Gaussian process are given by $(\beta, \sigma_f^2, \sigma_s, \sigma_t) = (20, 10, 2, 8)$. Notice that the mean function is assumed to be an unknown random variable, i.e., the dimension of the regression coefficient $\boldsymbol{\beta}$ is 1. We assume that $\boldsymbol{\beta}|\sigma_f^2$ has the noninformative prior and $\sigma_f^2 \sim \text{IG}(3, 20)$. We also assume the bounds of $\boldsymbol{\theta}$, viz. $\sigma_s \in [1.6, 2.4]$ and $\sigma_t \in [4, 12]$ are known. $\Delta = 12$ is used and $\eta = 11$ is selected satisfying the condition in Theorem 4.3. We use a discrete uniform probability distribution for $\pi(\boldsymbol{\theta})$ as shown in Fig. 3-(a). The adaptive sampling strategy was used in which agents make observations at each time $t \in \mathbb{Z}_{>0}$. The prediction was evaluated at each time step for 51 uniform grid points within $\mathcal{Q}$.

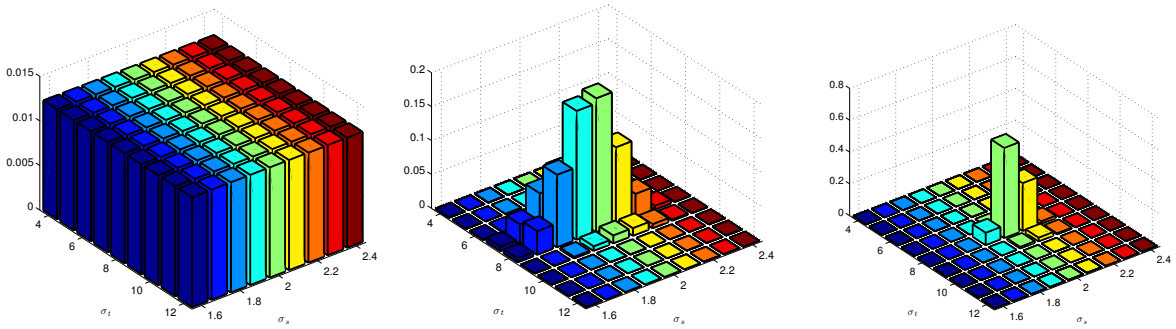
Fig. 2 shows the comparison between predictions at time $t = 1$ using (a) the maximum likelihood (ML) based approach, and (b) the proposed fully Bayesian approach. The ML based approach first generates a point estimate of the hyperparameters and then uses them as true ones for computing the prediction and the prediction error variance. In this simulation, a poor point estimate on $\boldsymbol{\theta}$ was achieved by maximizing the likelihood function. As a result, the prediction and the associated prediction error variance are incorrect and are far from being accurate for a small number of observations. On the other hand, the fully Bayesian approach which incorporates the prior knowledge of $\boldsymbol{\theta}$ and uncertainties in $\boldsymbol{\theta}$ provides a more accurate prediction and an exact confidence interval.

Using the proposed sequential Bayesian prediction algorithm along with the adaptive sampling strategy, the prior distribution was updated in a sequential manner. At time $t = 100$, the posterior distribution of $\boldsymbol{\theta}$ is shown in Fig. 3-(b). With a larger 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. 4-(a), the quality of the prediction at time $t = 100$ is significantly improved. At time $t = 300$, the prior distribution was further updated which is shown in Fig. 3-(c). At this time, $\boldsymbol{\theta} = (2, 8)^T$, which is the true value, has the highest probability. The prediction is also shown in Fig. 4-(b). This demonstrates the usefulness and 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). The distributed algorithm was implemented under a compactly supported correlation function for space. These promising 2-D simulation results can be found in the preliminary version of this paper [19].



(a) (b)

Fig. 2. The prediction at $t = 1$ using (a) the maximum likelihood based approach, and (b) the proposed fully Bayesian approach. The true fields are plotted in blue solid lines. The predicted fields are plotted in red dash-dotted lines. The area between red dotted lines indicates the 95% confidence interval.

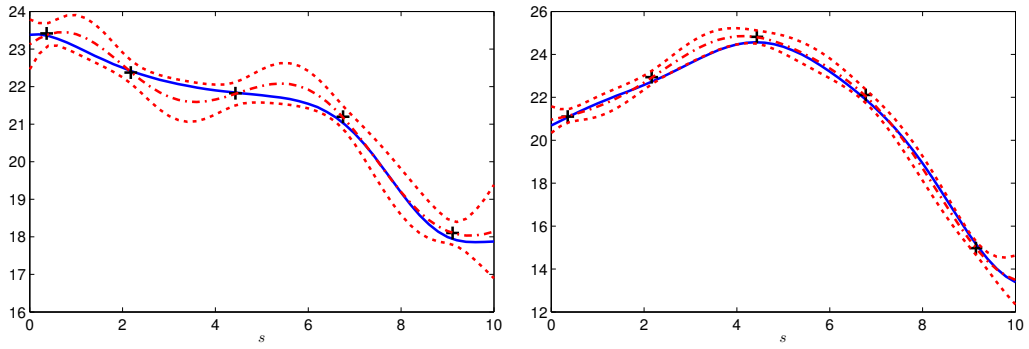


(a) (b) (c)

Fig. 3. (a) The prior distribution $\boldsymbol{\theta}$, (b) the posterior distribution of $\boldsymbol{\theta}$ at time $t = 100$, (c) the posterior distribution of $\boldsymbol{\theta}$ at time $t = 300$.

VI. Conclusion

In this paper, we formulated a fully Bayesian approach for spatio-temporal Gaussian process regression under practical conditions. We designed sequential Bayesian prediction algorithms to compute exact predictive distributions in constant time as the number of observations increases. An adaptive sampling strategy was also provided to improve the quality of prediction. Simulation results showed the practical usefulness of the proposed theoretically-correct algorithms in the context of environmental monitoring by mobile sensor networks.



(a) (b)

Fig. 4. The prediction at (a) $t = 100$, and (b) $t = 300$ using the centralized sequential Bayesian approach. The true fields are plotted in blue solid lines. The predicted fields are plotted in red dash-dotted lines. The area between red dotted lines indicates the 95% confidence interval.

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

[7] 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.

[8] 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.

[9] ——, "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.

[10] Y. Xu, J. Choi, and S. Oh, "Mobile sensor setwork coordination using Gaussian processes with truncated observations," *IEEE Transactions on Robotics*, 2011, accepted as a Regular Paper.

[11] 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.

[12] Y. Xu and J. Choi, "Adaptive sampling for learning Gaussian processes using mobile sensor networks," *Sensors*, vol. 11, no. 3, pp. 3051–3066, 2011.

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

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

[15] R. Graham and J. Cortés, "Cooperative adaptive sampling of random fields with partially known covariance," *International Journal of Robust and Nonlinear Control*, 2009.

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

[17] 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.

[18] C. R. Dietrich and G. N. Newsam, "Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix," *SIAM Journal on Scientific Computing*, vol. 18, no. 4, pp. 1088–1107, 1997.

[19] Y. Xu, J. Choi, S. Dass, and T. Maiti, "Bayesian Prediction and Adaptive Sampling Algorithms for Mobile Sensor Networks," in *Proceedings of American Control Conference (ACC), San Francisco, California*, 2011.

APPENDIX

A. Proof of Proposition 3.1

Proof: i) For given $\boldsymbol{\theta}$, we have

$$\begin{aligned}
p(\mathbf{y}|\boldsymbol{\theta}) &= \iint p(\mathbf{y}|\boldsymbol{\beta},\sigma_f^2,\boldsymbol{\theta})\pi(\boldsymbol{\beta},\sigma_f^2)d\boldsymbol{\beta}d\sigma_f^2 \\
&= \iint p(\mathbf{y}|\boldsymbol{\beta},\sigma_f^2,\boldsymbol{\theta})\pi(\boldsymbol{\beta}|\sigma_f^2)\pi(\sigma_f^2)d\boldsymbol{\beta}d\sigma_f^2 \\
&= \frac{b^a}{\Gamma(a)(2\pi)^{n/2}\det(\mathbf{C})^{1/2}\det(\mathbf{T})^{1/2}\det(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F}+\mathbf{T}^{-1})^{1/2}}\int \frac{\exp\left\{-\frac{b+\frac{RSS}{2}}{\sigma_f^2}\right\}}{(\sigma_f^2)^{n/2+a+1}}d\sigma_f^2 \\
&= \frac{\Gamma(\frac{n+2a}{2})b^a}{\Gamma(a)(2\pi)^{n/2}\det(\mathbf{C})^{1/2}\det(\mathbf{T})^{1/2}\det(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F}+\mathbf{T}^{-1})^{1/2}}\left(b+\frac{RSS}{2}\right)^{-\frac{n+2a}{2}}
\end{aligned}$$

where

$$RSS = \mathbf{y}^T\left(\mathbf{C}^{-1}-\mathbf{C}^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F}+\mathbf{T}^{-1})^{-1}\mathbf{F}^T\mathbf{C}^{-1}\right)\mathbf{y}.$$

As $\alpha\to\infty$, we have

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\mathbf{y}) &= \lim_{\alpha\to\infty}\frac{p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \\
&\propto \det(\mathbf{C})^{-1/2}\det(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F})^{-1/2}\left(b+\frac{1}{2}\mathbf{y}^T\boldsymbol{\Sigma}\mathbf{y}\right)^{-\frac{n+2a}{2}},
\end{aligned}$$

where $\boldsymbol{\Sigma} = \mathbf{C}^{-1}-\mathbf{C}^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{C}^{-1}$.

ii) For given $\boldsymbol{\theta}$ and $\mathbf{y}$, we have

$$\begin{aligned}
p(z_*|\mathbf{y},\boldsymbol{\theta}) &= \iint p(z_*|\mathbf{y},\boldsymbol{\beta},\sigma_f^2,\boldsymbol{\theta})\pi(\boldsymbol{\beta},\sigma_f^2|\boldsymbol{\theta},\mathbf{y})d\boldsymbol{\beta}d\sigma_f^2 \\
&= \iint p(z_*|\mathbf{y},\boldsymbol{\beta},\sigma_f^2,\boldsymbol{\theta})\pi(\boldsymbol{\beta}|\sigma_f^2,\boldsymbol{\theta},\mathbf{y})\pi(\sigma_f^2|\boldsymbol{\theta},\mathbf{y})d\boldsymbol{\beta}d\sigma_f^2,
\end{aligned}$$

where

$$\begin{aligned}
z_*|\mathbf{y},\boldsymbol{\beta},\sigma_f^2,\boldsymbol{\theta} &\sim \mathcal{N}\left(\mathbf{f}(\mathbf{x}_*)^T\boldsymbol{\beta}+\mathbf{k}^T\mathbf{C}^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}),\sigma_f^2(1-\mathbf{k}^T\mathbf{C}^{-1}\mathbf{k})\right), \\
\boldsymbol{\beta}|\sigma_f^2,\boldsymbol{\theta},\mathbf{y} &\sim \mathcal{N}(\hat{\boldsymbol{\beta}},\sigma_f^2\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}), \\
\sigma_f^2|\boldsymbol{\theta},\mathbf{y} &\sim \text{IG}\left(a+\frac{n}{2},b+\frac{RSS}{2}\right).
\end{aligned}$$

Then, it can be shown that

$$p(z_*|\mathbf{y},\boldsymbol{\theta}) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}}\left(1+\frac{\lambda(z_*-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

when $\alpha\to\infty$. ■