BIOCHEMICAL TRANSPORT MODELING, ESTIMATION AND DETECTION IN REALISTIC ENVIRONMENTS

Mathias Ortner and Arye Nehorai

Department of Electrical and Systems Engineering,
Washington University
St. Louis, MO 63130, USA

ABSTRACT

Simulation, detection and estimation of the spread of a biochemical substance are key elements in environmental monitoring. Solving these problems are important for efficient decontamination purposes and prediction of the cloud evolution.We present a set of tools describing the measurements of an array of biochemical sensors through a physical dispersion model, which is amenable to statistical analysis. We first approximate the dispersion model of a contaminant in a realistic environment (for instance urban) through numerical simulations of reflected stochastic diffusions describing the microscopic transport phenomena due to wind and chemical diffusion using the Feynmann-Kac formula. Second, we propose a Bayesian framework based on a random field for localizing multiple dispersive sources with small amounts of measurements. Third, we present a sequential detector allowing on-line analysis and detecting whether a change has occurred, based on realistic numerical simulation. Numerical examples illustrate our results for a dispersion among buildings.

Index Terms— Array signal processing, biochemical diffusion, Feynmann-Kac formula, Bayesian estimation, Sequential detection.

1. INTRODUCTION

Simulation, detection and estimation of the spread of a biochemical substance are key elements in environmental monitoring. In order to exploit a physical knowledge on the biochemical dispersion phenomenon, we employ a forward physical dispersion model relating the source to the measurements given by an array of biochemical sensors in realistic complex environments such as urban or indoor scenarios. In our previous work, we presented detection and estimation techniques for simple scenarios where analytical solutions to the transport equations are available [1], or employed numerical solutions given by finite elements methods [2].

We overview our results [3, 4, 5] that uses Monte-Carlo simulations for computing the transport model in realistic scenario. Our framework is amenable to the inclusion of complex geometries as well as ad hoc stochastic models for wind turbulence, Moreover it is efficient for large setup since its computational load increases linearly with the number of sensors and the time of diffusion. First, we present how to solve the inverse problem. Localizing the origin of the spread of the contaminant is indeed a major issue for environemental monitoring. We consider cases involving multiple sources and propose to use a generic Bayesian approach based on a Random field. Second we show how to detect a biochemical release as early as possible using a sequential detector. We propose using a sequential generalized likelihood ratio test (GLRT) as advocated by Lai [6] since some parameters of the diffusion may be unkown.

This paper is organized as follows: in Section 2 we review briefly the physical dispersion model as well as the measurement model and present the setup for the numerical approximations we proposed in our earlier work [3]. In Section 3 we present the Bayesian approach we adopted for solving the inverse problem and in Section 4 we present the sequential detector we developed in [4].

2. NUMERICAL APPROXIMATION FOR COMPUTING PHYSICAL AND MEASUREMENT MODELS

We assume that both the geometry and the average wind distribution are known. We assume that the wind has a known main direction and that we have a software capable of computing the wind distribution over the area. We also assume that we know the diffusion properties of the contaminant (diffusion coefficient) and that the sensors have been calibrated, resulting in a known noise variance.

2.1. Physical and measurement model

We consider a bounded open domain $D \subset R^3$. Let $\boldsymbol{r} = (x, y, z)$ be a point in D. Denote by $c(\boldsymbol{r}, t)$ the dispersive substance concentration at a point $\boldsymbol{r}$ and time t. The transport equation in the presence of a wind field $\boldsymbol{v}(\boldsymbol{r}, t) \in R^3$ is given by the equation $\frac{\partial c}{\partial t} = div(\boldsymbol{\mathcal{K}} \nabla c) - \nabla c \cdot \boldsymbol{v}$ when

This work was supported by the National Science Foundation Grants CCR-0330342 and CCF-0630734

the medium is assumed to be incompressible [1] where $\mathcal{K}$ is a matrix of conduction (or diffusivity). We suppose that $\mathcal{K}$ is function of the space variables. Let ∂D be the boundaries of the domain D. We assume[3] two kinds of domain boundary conditions and divide ∂D into two disjoint subsets denoted by $\partial D_{\mathcal{N}}$ and $\partial D_{\mathcal{D}}$ corresponding to Neumann conditions ($\nabla c(\boldsymbol{r}) \cdot \boldsymbol{n}(\boldsymbol{r}) = 0$, for all $\boldsymbol{r} \in \partial D_{\mathcal{N}} 0$ and Dirichlet's ($c(\boldsymbol{r}) = 0$, for all $\boldsymbol{r} \in \partial D_{\mathcal{D}}$) We also assume that the sources have released a certain amount of a substance into the environment when the diffusion begins (instantaneous sources) and denote by $c_0(\boldsymbol{r})$ the substance concentration at time $t = 0$ (release time).

To model the measurements, we suppose a spatially distributed array of m biochemical sensors located at known positions $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m$. We assume that each sensor takes measurements at times $t_0, \ldots, t_n$. Referring to our earlier work[7], we adopt the measurement model $y(\boldsymbol{r}_i, t_j) = c(\boldsymbol{r}_i, t_j) + \epsilon$ for $i = 1, \ldots, m$ and $j = 0, \ldots, n$ with $\epsilon \sim \mathcal{N}(0, \sigma_e^2)$ representing measurement noise and modeling errors.

2.2. Transport modeling using a Monte Carlo approximation

We present the Monte Carlo approach we adopted[3] to numerically solve the transport equation in the presence of turbulence. We use a discrete version of the domain D. Denote by Λ a set of sites in D. We partition D into small elements $\Delta D(s)$, i.e. $D = \bigcup_{s \in \Lambda} \Delta D(s)$ with $\Delta D(s) \cap \Delta D(s') = \emptyset$ for all sites $s \neq s'$ For a given sensor, located in $\boldsymbol{r}_i$, consider the following random walk :

$$X_0^i = r_i, \qquad dX_t^i = -\boldsymbol{v}(\boldsymbol{X^i}_t)dt + \sqrt{2\mathcal{K}}(\boldsymbol{X^i}_t)dW_t \quad (1)$$

We consider m stochastic processes started in the sensor locations. According to the Feynman-Kac formula (see our paper[3] and references therein), the result of the diffusion equation at a given time t_j and location r_i is $c(\boldsymbol{r}_i, t_j) = \mathbf{E}\left[c_0(X_{t_j}^i)\right]$. We use suitable Monte Carlo simulations of the processes X_t^i to obtain, for each sensor, N final points denoted by $X_{\boldsymbol{r}_i, t_j}^1, \ldots, X_{r_i, t_j}^N$. Let $p_{i,j,s} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}(X_{\boldsymbol{r}_i, t_j}^k \in \Delta D(s))$ be the average number of such points falling in the element $\Delta D(s)$ For a given initial value function $\boldsymbol{r} \to c_0(\boldsymbol{r})$, the Feynman-Kac formula yields

$$c_{i,j} \approx \sum_{s \in \Lambda} p_{i,j,s} c_0(s) \quad (2)$$

where $c_{i,j}$ is the calculated estimate of the concentration at location $\boldsymbol{r_i}$ and time t_j.

We denote by $\boldsymbol{y}$ all the measurements $y_{i,t}$ lumped into a single $m(n+1)$ dimensioned vector. Denote by $\boldsymbol{c}$ the vector of the initial concentrations $c_0(s)$ for all point $s \in \Lambda$. By assuming independent measurements and Gaussian noise we obtain the following likelihood from equation (2)

$$f(\boldsymbol{y}/\sigma_e, \boldsymbol{c}) = \frac{1}{(\sqrt{2\pi}\sigma_e)^{m(n+1)}} \cdots$$
$$\cdots \exp -\frac{1}{2\sigma_e^2} \sum_{i=1}^{m} \sum_{t=0}^{n} \left(y_{i,t} - \sum_{s \in \Lambda} p_{i,t,s} c_0(s) \right)^2 .$$

We developed[3] an ad-hoc procedure to account for wind turbulence based on the wind direction assumption and a program [8] dedicated to the Navier Stokes Equations. Our approach is enabled by the Monte-Carlo approach we employ .

3. LOCALIZING THE SOURCES

We describe briefly in this section the Bayesian approach we employed[3] for inferring the source location from the measurements. This task is useful for predicting the cloud evolution in space and time dispersion by applying the transport model to the estimated source(s) location(s).

3.1. Bayesian model

Prior model: We use the following mixture as a prior model for the (μ_z). We state that μ_z should be equal to 0 with a probability $1 - \rho$ and uniformly distributed in $[c_{\min}, c_{\max}]$ with a probability ρ. The prior term can be written as $f_{\text{prior}}((\mu_z)_{z \in \Lambda}/\rho) = \sum_{z \in \Lambda} (1-\rho)\mathbf{1}[\mu_z = 0] + \rho \mathbf{1}[\mu_z \in c_{\min}, c_{\max}]]$ The mixing parameter ρ should be chosen according to the size of the domain D and the number of sites $|\Lambda|$. A way to choose ρ is to make a prior decision about the average surface of the release. In practice we took $\rho = 0.01$, meaning that we state that the source surface is expected to be 1% of the overall domain area.

Posterior density: The likelihood of the measurements, the prior model, and Bayes formula result in the following *a posteriori* distribution: $f_{\text{post}}\left((\mu)_{z \in \Lambda}/\boldsymbol{y}_{t_0}, \ldots, \boldsymbol{y}_n, \sigma_e, \rho, t_0\right) = C\left(2\pi\sigma_2^2\right)^{-\frac{q}{2}} \left((1-\rho)\Upsilon + \rho\Psi\right) f(\boldsymbol{y}/\sigma_e, \boldsymbol{\mu})$ where $\Upsilon = \text{card}\{z \in \Lambda : \mu_z = 0\}$ and $\Psi = \text{card}\{s \in \Lambda : \mu_z > 0\}$, and C is the normalizing constant.

Estimator: For each site we consider the posterior probability of having a source at the location z, $\mathbf{P}_{\text{post}}(\mu_z > 0)$ as well as the posterior conditional expectation of the source concentration $\mathbf{E}_{\text{post}}(\mu_z | \mu_z > 0)$.

Algorithm: We employ [3] a Monte Carlo Markov chain method to sample the posterior distribution, and more precisely a Metropolis Hastings approach. This kind of sampler is especially suitable for our case, since we do not know the normalizing constant C. The estimated normalizing constant value is used for determining the diffusion starting time t_0 (see [3]).

3.2. Results

We present in Figure 3.2 the result of the sampler. On the left, we show the posterior probability of having a source in each considered location (note that the gray-scale is logarithmic). The true locations of the sources were correctly found. On the right we show the *a posteriori* expectation of the initial intensity in each location conditioned by the event that there is a release. Note that in the locations were the probability of having a source is high, the estimated intensity is close to the real value (we recall that we used an initial intensity $c_0 = 5$). For that particular example, we assumed the initial time to be known ($t = t_0$). In the following section, we provide a result using the Bayesian evidence, for selecting a relevant initial time hypothesis. These results show that the random field approach is powerful for finding several sources.

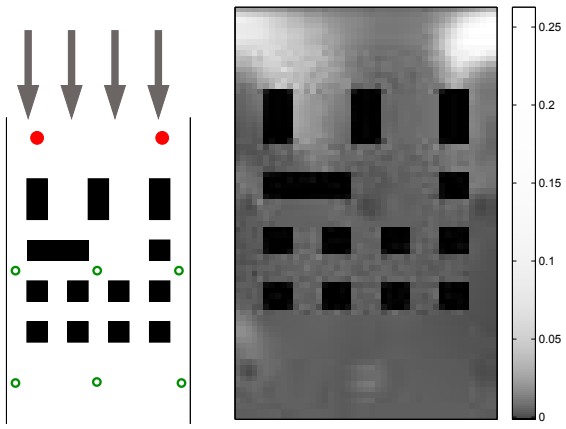

Fig. 1. Results of the source localization estimation given by the Bayesian approach. Left: simulated setup, right $\mathbf{P}_{\text{post}}(\mu_z > 0)$, the probability of having a source in each location (logarithmic scale).

4. SEQUENTIAL DETECTION

In this section we describe a framework [4] for detecting a biochemical release using the incoming measurements. In sequential analysis, the measurements are considered as an incoming flow, and the goal is to select the hypothesis of interest as soon as possible. For a detailed review of sequential detection, see Lai[9]. The two major competitive procedures mostly used today are the Shiryaev-Roberts-Girshik-Rubin [10] algorithm and Page's cumulative sum (CUSUM) algorithm [11].

4.1. A sequential detector

In detection theory, a natural idea when dealing with unknown parameters is to use the likelihood of the best hypothesis under each assumption [12] resulting in GLRT.

We consider three unknown parameters: the initial time δ, the location $s \in \Lambda$, and the intensity μ of an impulse substance source. We assume that the variance σ_e of the noise is known through a calibration step. Let $\boldsymbol{y_t} = (y_{1,t}, \ldots, y_{m,t})^T$ be the vector of m measurements given by the m sensors at time t. We then obtain the following sequential generalized likelihood ratio

$$\tilde{L}_n(\boldsymbol{y_0}, \ldots, \boldsymbol{y_n}) = \max_{0 \le \delta \le n} \max_{s \in \Lambda} \sup_{\mu \ge 0} \frac{f^1_{s,\mu}(\boldsymbol{y_\delta}, \ldots, \boldsymbol{y_n})}{f^0(\boldsymbol{y_\delta}, \ldots, \boldsymbol{y_n})} \quad (3)$$

Denoting by $\gamma = n - \delta + 1$ the number of measurements available at time n under the hypothesis that the release occurred at time δ and obtain the following expression and incorporating the maximum likelihood estimator in the detector expression, we obtain[4] the following ratio value $l_n^{\delta,s}(\boldsymbol{y_\delta}, \ldots, \boldsymbol{y_n}) = \left(\max\left\{0, T_\gamma^{\delta,s}(\boldsymbol{y_\delta}, \ldots, \boldsymbol{y_n})\right\}\right)^2$ where $T_n^{\delta,s}(\boldsymbol{y_\delta}, \ldots, \boldsymbol{y_n}) = \frac{\sum_{i=1}^m \sum_{t=0}^\gamma p_{i,t,s} y_{i,t+\delta}}{\sqrt{\sum_{i=1}^m \sum_{t=0}^\gamma p_{i,t,s}^2}}$. We derived a recursive formulation [4] of the test that is useful in practice. The resulting test can be described as follows: we consider the stopping time $\tau = \inf\{n \ge 0 \quad s.t. \quad L_n \ge \eta\}$ with $L_n = \max_{n-\gamma_{\max}+1 \le \delta \le n} \max_{s \in \Lambda} l_n^{s,\delta}$ where η is the test threshold.

4.2. Threshold, false alarm rate and performance

In [4], we focused on how to select a threshold. In a change-point detection framework the goal is to keep on testing while new measurements are arriving. Instead of fixing a false alarm probability, the usual approach is to decide the average run length (ARL) before a false alarm denoted $\tau_0 = \mathbb{E}_{\mathcal{H}_0}[\tau]$ which is the expected duration before a false alarm. We provide in [4] an analytical result in terms on a bound on the average run length before false alarm for fixing η .

We provide in [4] three performance measures : we examine the probability of detection and show how to compute the minimum signal intensity level that achieves a desired performance as a function of the release location. We also consider the average delay before detection.

4.3. Simulation

We present a result corresponding to the outdoor setup described in Figure 1 with six sensors and two initial release locations. We take the noise as $\sigma_e = 0.3$. In Figure 2 we present an example of a detection scenario. The first six rows correspond to measurements by the six sensors. Until time $t = 90$ the measurements are given by the null hypothesis ($\sigma_e = 0.3$). After time $t = 90$, we use the measurements predicted by the model. The last row shows the test statistic T and the threshold computed to achieve a false-alarm rate of $\alpha = 10^{-5}$. Note that this simulation included two sources, whereas the detector has been designed under a single-source hypothesis.

5. CONCLUSION

We have presented a new way to compute chemical transport equations in realistic environments and proposed a Bayesian framework to solve the inverse problem. The results are potentially useful for array optimal design. The proposed method allows the inclusion of a realistic stochastic wind distribution accounting for turbulence that proved to be powerful in practice. Our results are particularly useful for complex environments such as urban. In a future work we plan to work on optimal design for configuring the sensor array and obtain optimal monitoring.

6. REFERENCES

[1] A. Nehorai, B. Porat, and E. Paldi, "Detection and localization of vapor-emitting sources," *IEEE Trans. on Signal Processing*, vol. SP-43, pp. 243–253, Jan. 1995.

[2] A. Jeremić and A. Nehorai, "Detection and estimation of biochemical sources in arbitrary 2D environments," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Philadelphia, PA, March 2005.

[3] M. Ortner, A. Nehorai, and A. Jeremic, "Biochemical transport modeling and bayesian source estimation in realistic environments.," *IEEE Transactions on Signal Processing*, 2006, Accepted for publication.

[4] M. Ortner and A. Nehorai, "A sequential detector for biochemical release in realistic environments.," *IEEE Transactions on Signal Processing*, 2007, accepted for publication.

[5] M. Ortner and A. Nehorai, "Biochemical transport modeling and estimation in realistic environments," in *Proc. SPIE Vol. 6201, Sensors, Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense V*, Orlando, FL, April 2006, vol. 62010S, 12 pages.

[6] H.P. Chang and T.L. Lai, "Importance Sampling for Generalized Likelihood Ratio Procedures in Sequential Analysis," *Sequential Analysis*, To appear.

[7] B. Porat and A. Nehorai, "Localizing vapor-emitting sources by moving sensors," *IEEE Trans. on Signal Processing*, vol. SP-44, pp. 1018–1021, Apr. 1996.

[8] "Gerris Flow Solver," `http://gfs.sourceforge.net`.

[9] T.L. Lai, "Sequential analysis: some classical problems and new challenges," *Statistica Sinica*, vol. 11, pp. 303–408, 2001.

[10] M. Pollack, "Optimal Detection of a change in distribution," *Annal. of Statistics*, vol. 13, pp. 206–227, 1986.

[11] G. Lorden, "Procedures for reacting to a change in distribution," *Annal of Mathematical Statistics*, vol. 42, pp. 1897–1908, 1971.

[12] S. M. Kay, *Fundamentals of statistical signal processing: Volume II, detection theory*, Prentice Hall PTR, New Jersey, 1998.

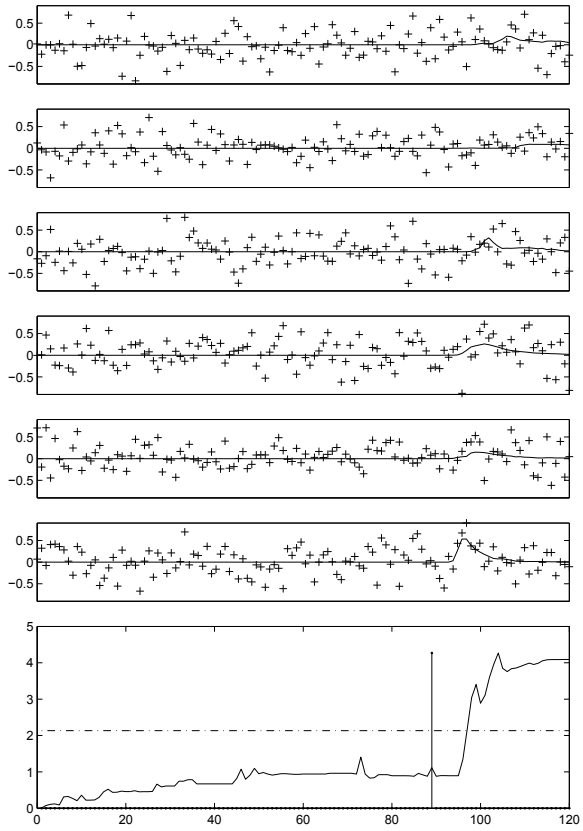

Fig. 2. On-line detection by an array of sensors illustrated by a simulated release on the framework of Figure 1 at time $t = 90$. The horizontal axis corresponds to time. The top six figures show the simulated measurements for each of the six sensors (see Figure 1). The noiseless measurements (solid lines) are given by the null hypothesis until $t = 90$. At time $t = 90$, a chemical diffusion has occurred and we use the measurements given by the diffusion simulation augmented with white noise ($\sigma_e = 0.3$). The bottom figure shows the test value (solid line) and threshold (dashed line), computed to achieve a false-alarm rate $\alpha = 10^{-5}$. The vertical line of the last row ($t = 90$) corresponds to the chemical release instant. The release is detected when the test value is above the threshold ($t = 98$).