51st IEEE Conference on Decision and Control
December 10-13, 2012. Maui, Hawaii, USA

Clustering Large Networks of Parametric Dynamic Generative Models

Yunwen Xu, Sanggyun Kim, Srinivasa M. Salapaka, Carolyn L. Beck, Todd P. Coleman

Abstract— **Analysis, prediction and control of parametric generative models for stochastic processes arise in numerous applications, such as in biology, telecommunications, geography, seismology and finance. In many of these applications, it is desirable to obtain an aggregated behavior from an underlying network of stochastic interactions. This paper focuses on the simplification of parametric models describing multiple stochastic processes, by aggregating the processes that have similar input-output behaviors in an ensemble. We propose a clustering-based method, which is general in the sense that the similarity metric upon which the aggregation relies can accommodate processes characterized by a variety of generative models. To illustrate our aggregation framework, we investigate an example system comprised of a set of point process models for earthquakes. Simulations are presented.**

I. Introduction

Research in many disciplines that include biology, economics, social sciences, computer science, and seismology, requires studying stochastic, dynamic networks of interacting processes. parametric generative models based on historical observations are typically used to model natural or engineered processes to make policies or operational decisions (e.g., in networked queueing systems and high-frequency finance [1], [2]), to avoid hazards (e.g., in earthquake prediction and wildfire control [3], [4]) and to gain a stronger understanding of underlying physical mechanisms (e.g., in neuroscience studies [5]). These models, in addition to the inherent locational information of these processes, characterize the interconnections within the systems, and hence naturally induce a graph representation. For instance, each stochastic process is modeled as a node, and the mutual interactions (inhibition or excitation) are modeled as edges between a pair of nodes.

In many cases, the system contains numerous interacting processes, and directional relations are necessary to describe the system behavior; these usually result in a large directed graph model. Therefore, it is useful to have an aggregated representation which captures the dominant interrelations in the system. Clustering is an effective method for partitioning a large number of elements by aggregating *similar* elements, and provides a methodology to succinctly elucidate the intrinsic interactions. In particular, when the networked processes are represented by graphs of parametric models, a natural approach would be to cluster these *models* based on the estimated model parameters.

This work was partially supported by the NSF grant CMMI - 1100257.

Yunwen Xu, S. Salapaka and C. L. Beck are from University of Illinois at Urbana-Champaign, Urbana, Illinois. Sanggyun Kim and Todd P. Coleman are from UCSD, San Diego, CA. Email: [xu27, salapaka, beck3]@illinois.edu,[tpcoleman, s2kim]@ucsd.edu

This paper aims to address the development of methods for the aggregation (via clustering) of a large network of parametric generative models. The main approach used in these methods is a form of clustering analysis that classifies individual objects into several subclasses, according to prespecified similarity metrics. This type of technique has been adapted previously for numerous clustering problems in image processing [6], statistical learning [7], and multi-agent systems [8]. Recently, clustering approaches have been generalized to reduce large connected graphs and Markov chain models [9], in order to facilitate the analysis of the underlying dominant dynamics of the system.

The main focus of the work presented herein is to investigate the functional aggregation of a class of parametrized generative models. More specifically, we aggregate similar *random processes* and obtain a set of simplified *representative processes* in an ensemble, where the precise notion of similarity will be defined. To implement this aggregation process, we extend the maximum entropy principle (MEP) based methods developed for graph models to discover the underlying structure of functional units in networked random processes. As described in [9], the functional aggregation objective can be formulated as a combinatorial optimization problem which aims to minimize the dissimilarity between the original and the aggregated networked processes. The application of the MEP allows us to successfully avoid local optima, overcome dependence on the initialization, and provide multi-scale models for the original system. After aggregation, processes with similar functions are identified as belonging to a single functional unit. Thus the goal of studying functional interactions among Q random processes is recast as a study of the relations of K functional units, where $K < Q$, therefore providing a visualization of a coarser representation of the ensemble.

We employ techniques from optimization theory, point process theory, clustering analysis, and statistics towards providing a robust, statistically sound, and scalable methodology that can be used to better understand the complexities of interactions amongst numerous random process statistical models.

II. Preliminaries

Throughout this paper, we consider Q random processes that have been recorded simultaneously. Let the ith random process be $\mathbf{X}_i = (X_i(1), \ldots, X_i(M), \ldots)$ and $x_i(\tau)$ be the realization of X_i at the τth time unit, $1 \le i \le Q, \tau \ge 1$. We denote the collection of Q random processes as $\underline{\mathbf{X}} = \{\mathbf{X}_1, \ldots, \mathbf{X}_Q\}$,

978-1-4673-2066-5/12/$31.00 ©2012 IEEE S.

with realization $\underline{x}$. The joint distribution of $\underline{\mathbf{X}}$ is given by

$$\mathbb{P}_{\underline{\mathbf{X}}}(\underline{x}) = \prod_{i=1}^{Q} \prod_{\tau \geq 1} \mathbb{P}_{X_i(\tau)|\underline{\mathbf{X}}(1:\tau-1)}(x_i(j)|\underline{x}(1:\tau-1)), \qquad (1)$$

where $\underline{\mathbf{X}}(1:M) \triangleq \{X_{i',\tau'} : 1 \leq i' \leq Q,\ 1 \leq \tau' \leq M\}$ denotes all the processes up to the Mth time unit. By defining the history $\mathscr{H}_M \triangleq \underline{\mathbf{X}}(1:M-1)$, the joint distribution (1) becomes

$$\mathbb{P}_{\underline{\mathbf{X}}}(\underline{x}) = \prod_{i=1}^{Q} \prod_{\tau \geq 1} \mathbb{P}_{X_i(\tau)|\mathscr{H}_\tau}(x_i(\tau)|\mathscr{H}_\tau). \qquad (2)$$

In this paper, we are interested in parametric statistical models for random processes, whose representations at time t depend solely on a *finite* history $\mathscr{H}_\tau$ $(\tau < t)$, as shown in the following two examples.

Example 1 (Gauss-Markov processes): A network of Gauss-Markov processes with Q individual processes and M-step history dependence are modeled as

$$X_i(t) = \gamma_{i,0,0} + \sum_{j=1}^{Q} \sum_{\tau=1}^{M} \gamma_{i,j,\tau} X_j(t-\tau) + W_i(t), \quad 1 \leq i \leq Q$$

where $W_i(t)$ are i.i.d. and Gaussian $(\mathscr{N}(0,\varepsilon_i))$. For each process $\mathbf{X}_i$, the model parameters $\{\gamma_{i,\cdot,\cdot}\} \in \mathbb{R}^{Q \times M}$ indicate how the future of $\mathbf{X}_i$ depends on the past of Q processes through M previous time units.

Example 2 (Renewal process models for earthquake): Let us first recall the definition of a Poisson process:

A *Poisson process* $N(t)$ of rate $\lambda \geq 0$ is a *counting process* satisfying the following conditions:

1) For any interval $(t, t+\Delta t](t > 0), \Delta N_{(t,t+\Delta t]} = N(t+\Delta t) - N(t)$ has a Poisson distribution $Poi(\lambda \Delta t)$;
2) (Independence) For any non-overlapping intervals $(t, t+\Delta t]$ and $(s, s+\Delta s]$, the counts $\Delta N_{(t,t+\Delta t]}$ and $\Delta N_{(s,s+\Delta s]}$ are independent.

A general renewal process extends the Poisson process by allowing the rate to depend on time and history. Suppose there are Q renewal processes in an ensemble and $\mathscr{H}(t)$ contains the information of all processes up to time $t-1$, the rate function of a renewal process is $\lambda(t|\mathscr{H}(t))$.

It is well known that a renewal process is completely characterized by $\lambda(t|\mathscr{H}(t))$. Though there are some non-parametric methods to estimate these rate functions [10], parametric models are more popular and can provide tools to analyze the behaviors and interactions of the individual processes.

Earthquakes over time and space provide an important example that can be modeled as a renewal process [11], [12], in which the aftershocks of a major earthquake exceeding certain magnitudes are modeled as events. The conditional rate of this process at the ith location, $\lambda_i(t|\mathscr{H}_t)$, depends on the background seismicity rate at that location μ_i, and the aftershock counts in other locations $\Delta N_j, j = 1, 2, \cdots, Q$ up to time $t-1$. One of the most common and basic parametrizations is called the epidemic-type model first introduced in [11]:

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{Q} \int_0^t g_{ij}(t-\tau) dN_j(\tau), \quad 1 \leq i \leq Q \qquad (3)$$

in which $g_{ij}(t)$ is a function of the aftershock magnitude observed at location j and time t. These models show the dependence between different processes, and enable comparison between processes in terms of their responses under the same inputs. That is, suppose that aftershock rates in two locations are modeled by (3), the prediction results would be indistinguishable if the two models have the same parameters, regardless of their geological locations. Therefore, we can classify the processes by aggregating them according to their responses under the same inputs, providing models of dominant earthquake propagation trends.

III. Aggregation on Parameterized Models

As is introduced in Section II, our goal of identifying the functional interactions among processes can be interpreted from a clustering perspective: partition all processes of interest into (non-overlapping) cells by aggregating *similar* processes, then determine the interactions among these cells.

We first set up a mapping from a networked generative model to a weighted directed graph, and formulate the problem as clustering the nodes in the graph. Then we define a function that evaluates the "dissimilarity" of the input-output behaviors of two generative models in terms of their parameters. Finally, we adapt the graph reduction method proposed in [9] and hence obtain an aggregation of the corresponding parametric random processes.

A. Graph representation

Generative parametric models are fully determined by fitting observation data. For example, in earthquake studies wheres the conditional rate functions $\lambda_i(t|\mathscr{H}(t))$ are parametrized by (3). The random processes at multiple locations can easily be cast in a graph structure, whose ith node represents the random process in the ith location, and the edge weight on directed edge (i,j) reflects the influence of the process from the jth location to the i th location. Note that the edge weights may vary with time. Therefore, aggregating similar random processes can be converted to aggregating similar nodes in the graph, where the aggregation varies with time.

For example, suppose there are Q conditional rate models (3) for earthquakes in Q different locations. Let $\boldsymbol{\gamma}_i(t) \triangleq [\mu_i, g_{i1}(t), \cdots, g_{iQ}(t)]$ be the model parameters for the ith process, in which $g_{im}(t) = \int_0^t g_{im}(t-\tau) dN_m(\tau)$ is the instantaneous influence from the aftershock process in the mth location. Then in the directed graph describing the relations of these Q processes, $g_{im}(t)$ represents the (dynamic) weight on the edge (m,i) at time t. Therefore, the parametric vector $\boldsymbol{\gamma}_i$ completely characterizes the statistical model of ith process.

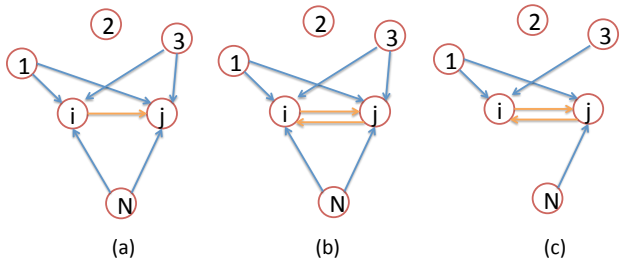

Fig. 1. Illustration of the selection of distance function. Assume all edge weights are the same. The ith and jth nodes in case (a) and case (b) have the same outside influences, but $d_{ij}(b) < d_{ij}(a)$ because $d_{ij}(pair)(a) > d_{ij}(pair)(b)$. In cases (b) and (c), i and j have the same pairwise influence between each other, but the outside influences are different, so $d_{ij}(b) < d_{ij}(c)$.

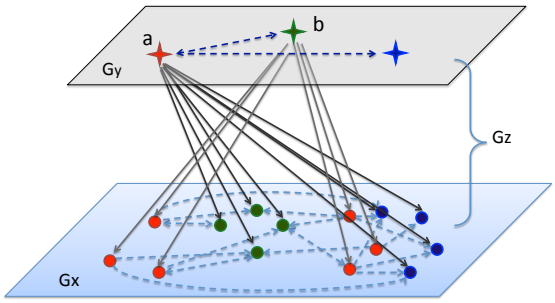

Fig. 2. $\mathscr{G}_x$ on the lower level represents the original graph $\mathscr{G}$ with $Q = 14$ nodes, $\mathscr{G}_y$ on the upper level represents the reduced graph $\hat{\mathscr{G}}$ with $K = 3$ supernodes. $\mathscr{G}_z$ with all nodes in both layers, and edges only from the upper layer to the lower layer, represents the *intermediate graph* $\tilde{\mathscr{G}}$.

B. Similarity function

In order to use the graph aggregation framework proposed in [9], we need to define the the notion of similarity, or equivalently of *dissimilarity* between two nodes. The setting considered in [9] refers to static graphs, with constant edge weights. Here, in general, the edge weights can be in vector form (e.g., when the history information is discretized) and varying with time.

We say that two processes are *similar* if given the same inputs, they produce similar outputs. We choose the dissimilarity function d_{ij} between two processes such that the following conditions are satisfied:

1) d_{ij} is a function of $\boldsymbol{\gamma_i}$ and $\boldsymbol{\gamma_j}$, such that d_{ij} is increasing as $\|\boldsymbol{\gamma_i} - \boldsymbol{\gamma_j}\|$ increases, and $d_{ij} = 0$ only if $\boldsymbol{\gamma_i} = \boldsymbol{\gamma_j}$, here $\|\cdot\|$ stands for some norm compatible with the model parameters.
2) When considering the ith and jth processes, influences between these two processes are treated separately from influences from an outside process k $(k \neq i, j)$. We denote these two types of dissimilarities by $d_{ij}(pair)$ and $d_{ij}(out)$. For example, in Figure 1 (without loss of generality, assume that all edges have the same weight), the influences between node i and node j are the same (symmetric) in cases (b) and (c), but are asymmetric in case (a), that is $d_{ij}(pair)(b) = d_{ij}(pair)(c) = 0 < d_{ij}(pair)(a)$. On the other hand, nodes i and node j in cases (a) and (b) are equally influenced by outside nodes, that is $d_{ij}(out)(a) = d_{ij}(out)(b) = 0$. So only in case (b), we have $d_{ij}(out)(b) + d_{ij}(pair)(b) = 0$.
3) Geometric influence. Let $d_{ij}(loc)$ be the locational distance between the ith and jth processes. d_{ij} can depend on $d_{ij}(loc)$ for geographic-dependent models such as earthquakes.

Therefore we can choose a dissimilarity function with the following structure

$$d_{ij} = \alpha_1 d_{ij}(out) + \alpha_2 d_{ij}(pair) + \alpha_3 d_{ij}(loc), \tag{4}$$

in which the vector $\alpha = [\alpha_1\ \alpha_2\ \alpha_3] \geq \mathbf{0}$ weights the three terms, and the specific forms of $d_{ij}(out)$ and $d_{ij}(pair)$ depend on the structure of edge weights.

C. Aggregating the parametrized models

In the previous section, we have modeled a set of parametrized stochastic processes as a graph. After defining a dissimilarity function between individual processes as in (4), the aggregation of similar parameterized models reduces to the aggregation of the corresponding nodes in a graph. To proceed, we will use the maximum entropy principle (MEP) based aggregation framework we proposed for reducing general weighted directed graphs in [9].

The objective of this clustering problem is as follows: Given Q parametric random process models $\lambda_1, \lambda_2, \cdots, \lambda_Q$, with model parameters $\boldsymbol{\gamma_1}, \cdots, \boldsymbol{\gamma_Q}$, find K *cluster representatives* $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_K$ with model parameters $\hat{\boldsymbol{\gamma}}_1, \cdots, \hat{\boldsymbol{\gamma}}_K$, such that the $\hat{\lambda}_i$ with $\hat{\boldsymbol{\gamma}}_i$ represents the parametrized model for the *process corresponding to the ith cluster*. In terms of model reduction for graphs, this can be stated as: Given a graph $\mathscr{G}(V, E, W)$ with Q nodes (V), edges E and edge weights W being specified by $W_{ij} = \boldsymbol{\gamma_{i,j}} = \{\gamma_{i,j,t} | t > 0\}$ for continuous models, (or $W_{ij} = \boldsymbol{\gamma_{i,j}} = [\gamma_{i,j,1}, \cdots, \gamma_{i,j,\tau}, \cdots]$, for discretized models). Find a reduced graph $\hat{\mathscr{G}}(\hat{V}, \hat{E}, \hat{W})$ with K *supernodes*, such that the dissimilarity between the two graphs is minimized. (See Figure 2 for illustration.)

We define a set of *intermediate graphs* $\tilde{\mathbf{G}}$ as the set of graphs $\tilde{\mathscr{G}}$, with each element $\tilde{\mathscr{G}}$ consisting of all nodes in the original graph $\mathscr{G}$ and the reduced graph $\hat{\mathscr{G}}$ (that is, $\tilde{\mathscr{G}}$ contains $Q + K$ nodes), and with edges only initiating from a supernode in $\hat{\mathscr{G}}$ and ending at a regular node in $\mathscr{G}$. We extend the two step clustering approach for graphs with constant scalar edge weights proposed in [9] to accommodate a general class of edge weights:

1) (Node aggregation) For a reduced graph with a desired size, say K nodes, search over the set of intermediate graphs $\tilde{\mathscr{G}}$ and find a $\tilde{\mathscr{G}}^\star$ that achieves the minimal dissimilarity between $\mathscr{G}$ and the set $\tilde{\mathbf{G}}$. In other words, this step calculates an intermediate parametric model for each aggregated process, as a function of the original processes.
   For example, in the earthquake example (3), node aggregation results in the conditional rate functions,

given by

$$\tilde{\lambda}_i(t) = \tilde{\mu}_i + \sum_{j=1}^{Q} \int_0^t \tilde{g}_{ij}(t-\tau) dN_j(\tau),\ 1 \le i \le K.$$

2) (Edge aggregation) Aggregate the edges of $\tilde{\mathscr{G}}^\star$ to obtain a reduced graph $\hat{\mathscr{G}}^\star$. This reduced graph corresponds to the parametrized models of representative processes in terms of the history of the supernodes activities.

In the earthquake example, these are the conditional rate models among different cluster centers:

$$\hat{\lambda}_i(t) = \hat{\mu}_i + \sum_{j=1}^{K} \int_0^t \hat{g}_{ij}(t-\tau) d\hat{N}_j(\tau),\ 1 \le i \le K,$$

where $d\hat{N}_j(\tau) \stackrel{\triangle}{=} \sum_{j'} d\hat{N}_{j'}(\tau)$, for all processes j' associated with cluster j. (See [9] for details.)

The optimization problem associated with the first step is given by minimizing the following *distortion* function:

$$\min_{\tilde{\mathscr{G}} \in \tilde{\mathbf{G}}} \rho(\mathscr{G}, \tilde{\mathscr{G}}) \stackrel{\triangle}{=} \sum_{i=1}^{Q} p_i \min_{1 \le j \le K} d_{ij}, \tag{5}$$

in which d_{ij} is defined in (4), and p_i is simply $\frac{1}{Q}$ since we treat the models equally. In general, p_i can be any (non-negative) normalized weighting parameters for each node. For example, when the model parameters are estimated with different errors, then p_i can be chosen as $\exp(\sigma_i)/[\sum_{k=1}^{Q} \exp \sigma_k]$, where σ_i denotes the error standard deviation in estimating λ_i.

The main contribution of the MEP is to allow soft clustering [13], which means allowing a node v_i to be associated with all supernodes $\hat{v}_j$ with some nonnegative *association weights*, defined as $p(\hat{v}_j|v_i)$ for $i = 1, \cdots, Q$ and $j = 1, \cdots, K$. By introducing these soft associations, the distortion ρ in (5) becomes

$$\rho'(\mathscr{G}, \tilde{\mathscr{G}}) \stackrel{\triangle}{=} \sum_{i=1}^{Q} p_i \sum_{j=1}^{K} p(\hat{v}_j|v_i) d_{ij}, \tag{6}$$

with an entropy term evaluating the non-uniqueness (uncertainty) of the association, given by

$$H(\tilde{\mathscr{G}}|\mathscr{G}) = -\sum_{i=1}^{Q} p_i \sum_{j=1}^{K} p(\hat{v}_j|v_i) \log p(\hat{v}_j|v_i). \tag{7}$$

Therefore we minimize (6) under constraint $H = H_k$, for a series of decreasing entropy values, which is equivalent to minimizing the Lagrangian:

$$L(\tilde{\mathscr{G}}, \mathscr{G}) = \rho'(\mathscr{G}, \tilde{\mathscr{G}}) - TH(\tilde{\mathscr{G}}|\mathscr{G}), \tag{8}$$

with T being the Lagrange multiplier. As was derived in [9], [13] for the scalar case, solving the optimization problem (8) at a fixed T value results in the following optimal association weights and edge weights:

$$p(\hat{v}_j|v_i) = \frac{\exp\left\{-\frac{1}{T} d_{ij}\right\}}{\sum_{h=1}^{K} \exp\left\{-\frac{1}{T} d_{ih}\right\}} \tag{9}$$

$$\tilde{\gamma}_j = \sum_{i=1}^{Q} p(v_i|\hat{v}_j) \gamma_i, \tag{10}$$

in which $p(v_i|\hat{v}_j)$ is the posterior distribution of (9). Using these results, the second edge aggregation step is computed as

$$\hat{\gamma}_{il} = \sum_{i=1}^{Q} p(\hat{v}_l|v_j) \tilde{\gamma}_{ij}. \tag{11}$$

In the iterative computational algorithm, we drive the Lagrange multipleier T from a very high value to a very low value. Note that, from (9), the association weights are more uniform at high T and become increasingly deterministic as T decreases. We essentially begin the process with a convex Lagrangian (8) (which has a unique global optimizer) far from the original objective (4), and gradually move to a less convex, but closer approximation of (4). As noted in [13], [14], in this algorithm, the number of supernodes automatically increases from one to many, to track the local minima of (4).

IV. Example and Simulation

To illustrate our aggregation framework, we consider epidemic-type earthquake models (3), for the mutually excitatory effects of aftershock activities in different locations. The model depends on the history of the aftershocks, represented by a counting process for $0 < \tau \le t$. The goal of our study is to obtain coarser models that describe the propagation of aftershocks across spatial locations and over time using a clustering approach.

To avoid extensive (unnecessary) computational expense, we first discretize the time horizon into intervals of length Δt, such that the probability of having more than one aftershock within a time interval is $o(\Delta t)$. Moreover, we assume the parametrized model has *finite* history dependence. Therefore, the epidemic models in (3) for Q interactive processes are approximated by summation over finite past information, that is,

$$\lambda_i(N\Delta t) \approx \mu_{i,0} + \sum_{j=1}^{Q} \sum_{s=1}^{M} \mu_{i,j,s} dN_j((N-s)\Delta t), \quad 1 \le i \le Q, \tag{12}$$

where $dN_j(s) = N_j((s+1)\Delta t) - N_j(s\Delta t) \in \{0,1\}$, and the history dependence is assumed to be M time steps. For each process, $1 \le i \le Q$, the aftershock rate is fully captured by coefficients $\boldsymbol{\gamma_i} = (\mu_{\mathbf{i},\mathbf{0}}, \cdots, \mu_{\mathbf{i},\mathbf{j},\mathbf{s}}, \cdots)$ for $1 \le j \le Q, 1 \le s \le M$, in which $\mu_{i,0}$ represents the base activity level at location i, and $\mu_{i,j,s}$ indicates the mutually exciting level.

In the experiment, we first select 25 geographical locations in a domain (see the red circles in Figure 3) and construct epidemic models for each location: we assume a maximum history dependence of 8 steps, randomly select the model coefficients $\mu_{\mathbf{i}}$ for $1 \le i \le 25$. We then begin with an arbitrarily chosen initial condition $\underline{N}(0) \in \{0,1\}^{25}$, use the prespecified models to compute the aftershock rates λ_i, and the counts $\Delta N_i(k\Delta t)$ for all i and $k > 0$. Repeating this process, we obtain a realization of counting processes (aftershock counts) for each location. These processes are run for 2000 time steps, and the last 500 steps are used for parametric model fitting. Models in the form of (12) with $M = 5$ are selected using the Akaike information criterion

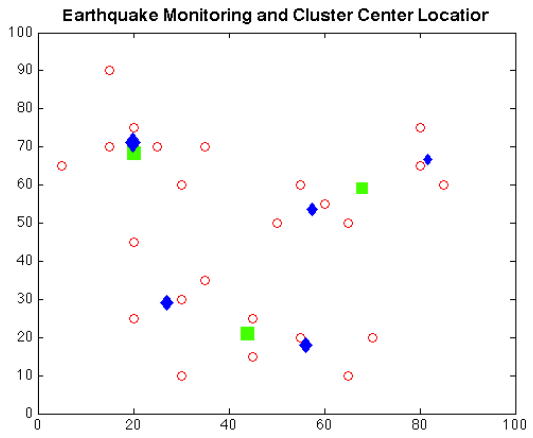

Fig. 3. The 25 locations where the parametric models are built in a domain (red circles); the equivalent representative locations of the 3-cluster reduction (green squares) and 5-cluster reduction (blue diamonds). In latter two cases, the area of each symbol is proportional to the number of processes represented by that cluster.

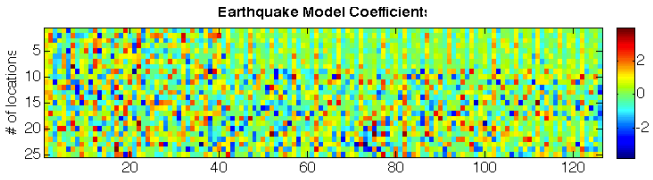

Fig. 4. The model parameters of the earthquake model for the aftershock activities at the 25 locations. A warmer color indicates a high value and a colder color indicates a low value.

(AIC), and the coefficients are shown in Figure 4. Note that μ_i for each process contains $1+5\times 25 = 126$ parameters.

The clustering algorithm is applied to the resulting generative models, or simply the model coefficients γ. We define the dissimilarity function between ith and jth processes as (4), with

$$d_{ij}(out) \stackrel{\triangle}{=} \|\gamma_i^{[i,j]} - \gamma_j^{[i,j]}\|_2^2 \text{ and } d_{ij}(pair) \stackrel{\triangle}{=} \|\gamma_i^{[j]} - \gamma_j^{[i]}\|_2^2, \quad (13)$$

in which $\gamma_i^{[i,j]} \in \mathbb{R}^{1+M(Q-2)}$ is obtained from vector γ_i by removing $\gamma_{i,j,s}, s = 1,\cdots,M$, and $\gamma_i^{[j]} = \{\gamma_{i,j,s}\}(s = 1,\cdots,M) \in \mathbb{R}^M$ for $Q = 25$ and $M = 5$. Thus, $d_{ij}(out)$ accounts for all *outside* model coefficients in process i and process j and $d_{ij}(pair)$ penalizes the asymmetry in the influences between process i and process j. This dissimilarity function is used as the basis of the cluster analysis.

We define a graph $\mathcal{G}(V,E,W)$ for these models of aftershock processes where the node $v_i \in V$ represents the process at location i, the edge weight $w_{ij} \stackrel{\triangle}{=} \{\gamma_{i,j,s}\}(s = 1,\cdots,M) \in \mathbb{R}^M$. As discussed in the general framework in Section III, we first aggregate the models by the dissimilarity function defined in (4) and (13). After choosing $\alpha = [1,1,0.5]$ and implementing the node aggregation step, we obtain a set of intermediate epidemic models for cluster representatives (in terms of the original process models), given by

$$\tilde{\lambda}_i(N\Delta t) = \tilde{\mu}_{i,0} + \sum_{j=1}^{Q}\sum_{s=1}^{M} \tilde{\gamma}_{i,j,s} dN_i((N-s)\Delta t)), \quad (14)$$
$$i = 1,\cdots,K..$$

The second edge aggregation step yields the parametric models that characterize the interactions among different cluster representatives, that is

$$\hat{\lambda}_i(N\Delta t) = \hat{\mu}_{i,0} + \sum_{j=1}^{K}\sum_{s=1}^{M} \hat{\gamma}_{i,j,s} d\hat{N}_i(s\Delta t)), \; j = 1,\cdots,K.$$

where $\hat{N}_i(s\Delta t) = \sum_{h\in C_i} R_h(s\Delta t)$ is the total number aftershocks in the locations associated with the ith cluster.

The proper number of supernodes is unknown at the outset and determined by the nature of the problem. In this simulation, we let the desired number of supernodes increase from 1 to 10 and plot the curve of distortion (Figure 5). The distortion improved rapidly when the number of desired clusters surpasses 3 and 5. Figure 6 shows the partition results of the 25 processes into 3 and 5 clusters. For example, when we separate the processes into 3 cells of similar processes, each cell contains 8, 10 and 7 locations respectively, which is the correct partition with our design. Figure 7 and 8 show the coefficients of the resulting epidemic models with 3 and 5 clusters.

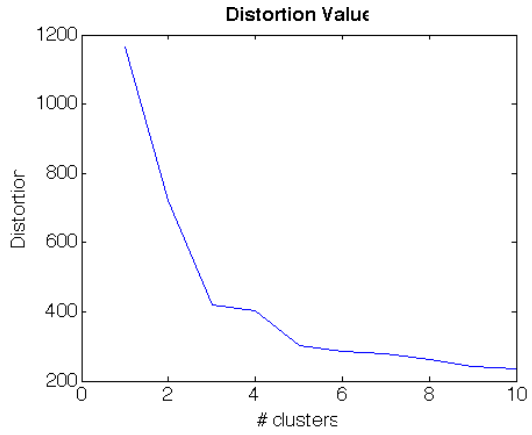

Fig. 5. The distortion curve vs. the number of clusters.

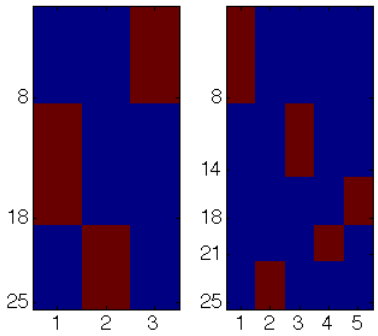

Fig. 6. The partition results of the 25 process when the objective model sizes are 3 and 5. The vertical axis marked with the cut-off number of processes in each cluster.

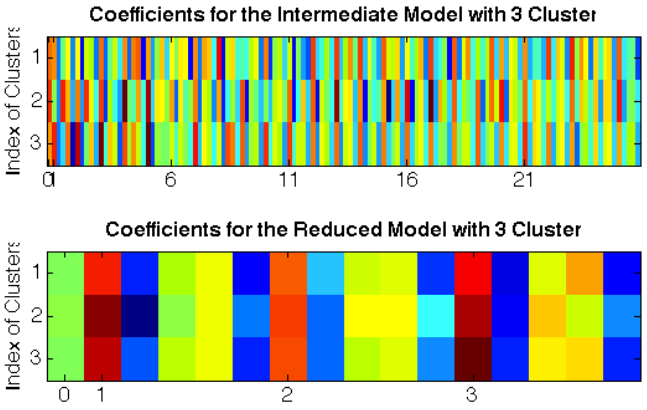

Fig. 7. The 3-cluster reduction results: the model coefficients of the intermediate models (upper) and reduced models (bottom).

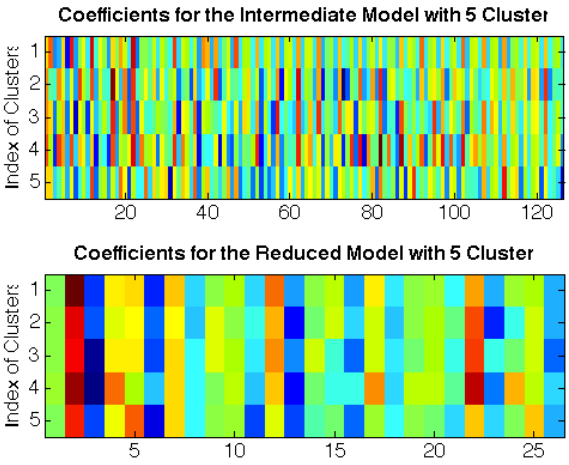

Fig. 8. The 5-cluster reduction results: the model coefficients of the intermediate models (upper) and reduced models (bottom).

V. Discussion

One of the most important contributions of clustering analysis on parametrized stochastic process models is to improve scalability in studying the functional connections among different processes, such as the causality analysis. For example in neuroscience research, the neuron spiking activities are typically modeled as one-dimensional point processes with generalized linear models [15]. In order to investigate the information propagation among neurons in the brain, various type of causality tests are developed for each pair of neurons. In particular, when checking the Granger causality (one of the most widely studied definitions of causality) [16] from the jth spiking process to the ith spiking process, generalized linear models of process i are estimated without and with the information of process j. The extent of causal relationship is determining from the improvement of the model prediction capacity achieved by including the information of process j [17]–[19]. This is a computationally expensive step, and the resulting causal relations are sometimes more detailed than needed. Suppose, for example, there are Q point processes under consideration, then the number of parameters we need to estimate is proportional to Q^2 and the number of causality tests required is Q^2. If after aggregation, the resulting network consists of $K(<Q)$ processes of cluster representatives, then implementing K^2 causality analyses on a network of representatives, and estimating the model parameters will provide a significant improvement in computational efficiency.

References

[1] S. Asmussen, *Applied Probability and Queues*, ser. Applications of mathematics. Springer, 2003.

[2] R. F. Engle and A. Lunde, "Trades and quotes: A bivariate point process," *Journal of Financial Econometrics*, vol. 1, no. 2, pp. 159–188, 2003.

[3] Y. Ogata, "Space-time point-process models for earthquake occurrences," *Annals of the Institute of Statistical Mathematics*, vol. 50, pp. 379–402, 1998, 10.1023/A:1003403601725. [Online]. Available: http://dx.doi.org/10.1023/A:1003403601725

[4] H. Xu and F. P. Schoenberg, "Point process modeling of wildfire hazard in los angeles county, california," *Annals of Applied Statistics*, vol. 5, no. 2A, pp. 684–704, 2011.

[5] C. Quinn, N. Kiyavash, and T. Coleman, "Equivalence between minimal generative model graphs and directed information graphs," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*,, 2011.

[6] L. Mitiche, A. B. Adamou-Mitiche, and D. Berkani, "Low-order model for speech signals," *Signal Processing*, vol. 84, no. 10, pp. 1805 – 1811, 2004.

[7] D. Parikh and T. Chen, "Hierarchical semantics of objects (hsos)," *IEEE International Conference on Computer Vision*, vol. 0, pp. 1–8, 2007.

[8] W. Ren, R. Beard, and E. Atkins, "Information consensus in multivehicle cooperative control," *IEEE Control Systems*, vol. 27, no. 2, pp. 71 –82, april 2007.

[9] Y. Xu, S. Salapaka, and C. L. Beck, "On reduction of graphs and markov chain models," in *the 50th IEEE Conference on Decision and Control and European Control Conference, 2011 (CDC/ECC'11)*, Dec. 2011, pp. 2317–2322.

[10] P. Guttorp and M. L. Thompson, "Nonparametric estimation of intensities for sampled counting processes," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 52, no. 1, pp. h157–173, 1990.

[11] A. G. Hawkes and L. Adamopoulos, "Cluster models for earthquakes - regional comparisons." *Bull. Int. Statist. Inst.*, vol. 45, pp. 454 – 461, 1973.

[12] Y. Ogata, "Statistical models for earthquake occurrences and residual analysis for point processes," *Journal of the American Statistical Association*, vol. 83, no. 401, pp. pp. 9–27, 1988.

[13] K. Rose, "Deterministic annealing for clustering, compression, classification, regression and related optimization problems," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2210–39, Nov. 1998.

[14] P. Sharma, S. Salapaka, and C. Beck, "A scalable approach to combinatorial library design for drug discovery," *Journal of Chemical Information and Modeling*, vol. 48, no. 1, pp. 27–41, 2008.

[15] A. J. Dobson, *An Introduction to Generalized Linear Models*. London: CRC Press, 2002.

[16] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[17] E. Brown, R. Kass, and P. Mitra, "Multiple neural spike train data analysis: state-of-the-art and future challenges," *Nature Neuroscience*, vol. 7, no. 5, pp. 456–461, 2004.

[18] M. Okatan, M. A. Wilson, and E. N. Brown, "Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity," *Neural Comput.*, vol. 17, pp. 1927–1961, Sep. 2005.

[19] S. Kim, D. Putrino, S. Ghosh, and E. N. Brown, "A granger causality measure for point process models of ensemble neural spiking activity," *PLoS Comput Biol*, vol. 7, no. 3, p. e1001110, 03 2011.