A Fast and Memory-Efficient Algorithm for Learning and Retrieval of Phenotypic Dynamics in Multivariate Cohort Time Series

Shamim Nemati[1] and Mohammad M. Ghassemi[2]

Abstract— **Robust navigation and mining of physiologic time series databases often requires finding similar temporal patterns of physiological responses. Detection of these complex physiological patterns not only enables demarcation of important clinical events but can also elucidate hidden dynamical structures that may be suggestive of disease processes. Some specific examples where this physiological signal search may be useful include real-time detection of cardiac arrhythmias, sleep staging or detection of seizure onset. In all these cases, being able to identify a cohort of patients who exhibit similar physiological dynamics could be useful in prognosis and informing treatment strategies. However, pattern recognition for physiological time series is complicated by changes between operating regimes and measurement artifacts. Here we briefly describe an approach we have developed for distributed identification of dynamical patterns in physiological time series using a switching linear dynamical system (SLDS). We present a fast and memory-efficient algorithm for learning and retrieval of phenotypic dynamics in large clinical time series databases. Through simulation we show that the proposed algorithm is at least an order of magnitude faster that the state of the art, and provide encouraging preliminary results based on real recordings of vital sign time series from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) database.**

I. INTRODUCTION

Pattern recognition in time series data has a broad range of applications from finance to medical informatics, however, robust algorithms for finding predictive patterns in long sequences of nonstationary multivariate time series are sparse [1]. We recently developed a machine learning algorithm for identification of dynamical patterns in multivariate cohort time series of physiological systems [2]. A central premise of our approach was that even within heterogeneous patient cohorts there are common phenotypic dynamics that a patient's vital signs may exhibit, reflecting underlying pathologies or temporary physiological state changes (e.g., postural changes or sleep/wake related changes in physiology), and used a switching linear dynamical system (SLDS) to model these dynamics. While previous works on the application of SLDS to physiological data [3] primarily relied on hand-annotated training data and expert knowledge for parameter estimation in a small cohort (tens of patients), we utilized a fully automated approach based on the switching Kalman filter (SKF) algorithm [4] on a large cohort involving hundred of patients [5]. This approach allowed for automatic learning of a collection of time series dynamics within a patient cohort and simultaneous segmentation of each time series in terms of those dynamics, using an iterative procedure known as the expectation maximization (EM) algorithm [4].

However, EM does not scale well to long sequences and large time series cohorts. In this work, we propose a singular value decomposition (SVD)-based algorithm for learning and inference in the SLDS models, which is at least an order of magnitude faster than the EM algorithm. Moreover, given that the SLDS framework allows for defining a notion of "similarity" among multivariate physiological time series based on their underlying shared dynamics [2], we propose a search engine for multivariate physiological time series, which allows for fast indexing and retrieval of dynamical patterns in large time series cohorts.

*Manuscript received September 1, 2014. This work was supported in part by the James S. McDonnell Foundation Postdoctoral grant and the Salerno Foundation. The content of this article is solely the responsibility of the authors.

[1] S. Nemati is with the Harvard School of Engineering and Applied Sciences, 33 Oxford Street, Cambridge, MA 02138, USA. Correspondence email: shamim@seas.harvard.edu

[2] M. Ghassemi is with the Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

II. METHODS

Assume we are given a collection of N nonstationary multivariate time series $\{\mathbf{y}^{(1)},\mathbf{y}^{(2)},\cdots,\mathbf{y}^{(N)}\}$ and the associated outcomes $\{O^{(1)},O^{(2)},\cdots,O^{(N)}\}$, where the n-th time series $\mathbf{y}^{(n)}$ is of length $T^{(n)}$, and may include M channels. Here we consider the case where the corresponding label $O^{(n)}$ is a binary or multinomial outcome variable.Our objective is to learn shared features across the cohort for time series classification.

A. Learning Switching Dynamics in Cohort time series

Switching Linear Dynamical Systems: The switching linear dynamical system (SLDS) [4] models time series using two layers of hidden state evolution. The generative model is as follows: a discrete latent process for each time series $s_t^{(n)} \in \{1,\cdots,J\}$ evolves according to a Markovian dynamic with initial distribution $\pi^{(n)}$ and $J \times J$ transition matrix Z. Each of the n series has an unobserved continuous state variable $\mathbf{x}_t^{(n)} \in \mathbb{R}^D$ that evolves according to linear dynamics which are determined by the current latent state $s_t^{(n)}$, and produces observations $\mathbf{y}_t^{(n)}$. The jth linear system has state dynamics $A^{(j)}$, observation matrix $C^{(j)}$, state noise covariance $Q^{(j)}$, and observation noise covariance $R^{(j)}$:

$$\mathbf{x}_t^{(n)} = A^{(s_t^{(n)})}\mathbf{x}_{t-1}^{(n)} + \mathbf{v}_t \qquad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, Q^{(s_t^{(n)})}) \tag{1}$$

$$\mathbf{y}_t^{(n)} = C^{(s_t^{(n)})}\mathbf{x}_t^{(n)} + \mathbf{w}_t \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, R^{(s_t^{(n)})}). \tag{2}$$

We refer to these state-specific dynamics together as $\Delta^{(j)} = \{A^{(j)},Q^{(j)},C^{(j)},R^{(j)}\}$ (also known as a *mode*), and we use $\Theta = \{\{\Delta^{(j)}\}_{j=1}^{J}, Z, \pi^{(n)}\}$ to denote the set of all model parameters defining the SLDS.

978-1-4799-5666-1/14/$31.00 ©2014 IEEE

EM for Parameter Learning in SLDS models: A comprehensive treatment of the EM algorithm for switching Kalman filters (SKF) is presented in Murphy (1998) [4]. In practice we neither know the set of switching variables nor the parameters that define the modes. EM is a two-pass iterative algorithm: (1) in the expectation (E) step we obtain the expected values of the latent variables $\{\{\mathbf{x}_t^{(n)}, S_t^{(n)}\}_{t=1}^{T_n}\}_{n=1}^{N}$ using a modified Kalman smoother [4], and (2) in the maximization (M) step we find the model parameters $\{\Theta^{(j)}\}_{j=1}^{J}$, Markov dynamics Z and initial conditions $\pi^{(n)}$ that maximize the expected complete data log likelihood. In our implementation of the EM algorithm, we achieve *shared* dynamics by pooling together all subjects' inferred variables in the M step. Iteration through several steps of the EM algorithm results in learning a set of J shared modes and a global transition matrix Z for all the patients.

Practical inference in SKF is approximate, since the exact inference requires tracking an exponentially increasing number of states. Since starting from time $t = 1$, each of the J states has to be propagated forward using J possible modes, resulting in J^2 states at $t = 2$, and so on. In practice, the J^2 modes are often collapsed down to J states using moment matching [4]. Therefore, each inference step of the EM algorithm requires $T \times J^2$ evaluations of the Kalman filter per time series, which has computational and memory complexity of $\mathcal{O}(D^3)$ and $\mathcal{O}(D^2)$ per time-step, respectively. To obtain smoothed estimates a backward pass through the data using the Rauch-Tung-Striebel (RTS) smoother is required, with similar computational complexity. Moreover, cost of the forward-backward algorithm for inference in the discrete hidden Markov model (HMM) is $\mathcal{O}(J^2 \times T)$. Finally, the HMM layer requires $T \times J^2$ calculations of likelihood of the observations, with each likelihood evaluation requiring calculations of the inverse and determinant of the measurement covariance matrices, which is $\mathcal{O}(M^3)$ in the dimension of observations. When the observation dimension is very large (e.g., modeling pixel values video sequences, firing rates of hundreds to thousands of simultaneously recorded neurons, or observations of several vital signs across multiple timescales), the cost of running the SKF becomes prohibitively large. Here we propose an algorithm that avoids running a Kalman filter altogether, and significantly reduces the overall computational complexity of inference and learning in the SLDS.

B. Singular Value-based System Identification

It is well-known that the choice of the matrices in a given mode $\Delta^{(j)}$ is not unique, in the sense that any rotation of the state vector $\mathbf{x}_t$ results in a different set of matrices, but leaves the input-output relationship unchanged. Our work is motivated by a specific parametrization of a state-space model under the assumption that $M >> D$, and $rank(C) = D$, and choosing a canonical model that makes the columns of the C orthogonal: $C^T C = I_D$, where I_D is the identity matrix of the size $D \times D$ [6]. This parametrization is particularly useful when dealing with high-dimensional observations. For completeness, we first summarize the algorithm presented by Soatto et al. [6] for learning such state-space models, and then provide an extension to the case of the switching linear dynamical system. Let $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_T]$ with $T > M$, $X = [\mathbf{x}_1, \cdots, \mathbf{x}_T]$, and $W = [\mathbf{w}_1, \cdots, \mathbf{w}_T]$. The algorithm of Soatto et al. finds the best C and X minimizers of the Frobenius norm of W such that

$$Y = CX + W;\ C \in R^{M \times D};\ C^T C = I_D. \tag{3}$$

This is done by calculating the singular value decomposition (SVD) of $Y = U\Sigma V^T$ and setting $C = U$ and $X = \Sigma V^T$. Moreover, it can be shown that $V_t = E(\mathbf{x}_t \mathbf{x}_t^T)$ asymptotically approaches Σ^2. The major cost of the algorithm presented in the previous section is running the SVD on a $M \times T$ matrix. This is particularly costly when dealing with a cohort of N (> 1000) time series, each consisting of several thousands of samples.

C. Extension to the Switching Linear Dynamical System

We next extend the algorithm of Soatto et al. to the problem of inference and learning in SLDS models. We propose a iterative procedure that alternates between inferring hidden layer variables and updating the model parameters (henceforth, we refer to this algorithm as SVD-SLDS).

1) Updating Model Parameters: Assuming an approximate segmentation of all the time series is known, i.e., $\{s_t^{(n)}, t = 1, \cdots, T^{(n)}\}_{n=1}^{N}$, we make use of the fact that if $U\Sigma V^T$ is the SVD of Y then columns of U are the eigenvectors of YY^T, and the non-zero singular values of Y (diagonal elements of Σ) are the square roots of the non-zero eigenvalues of YY^T. This allows us to do eigenvalue decomposition on $M \times M$ matrices rather than singular value decompositions on $M \times T^{(n)}$ matrices.

Let $Y_i^{(n)} = Y^{(n)} diag(P_\Theta(s_1^{(n)} = i), \cdots, P_\Theta(s_T^{(n)} = i))$, where $Y_i^{(n)}$ is a weighted observation matrix for the i-th mode. The average covariance matrix for the i-th mode is given by:

$$P^{(i)} = 1/N \sum_{n=1}^{N} Y_i^{(n)} {Y_i^{(n)}}^T / T_i^{(n)} \ ,$$

and $T_i^{(n)} = \sum_{t=1}^{T^{(n)}} P_\Theta(s_t^{(n)} = i)$ is the effective size of the n-th weighted time series for the i-th mode. Note that $P^{(i)}$ is $M \times M$. Let the eigenvalue decomposition of of $P^{(i)} = U^{(i)} \Lambda^{(i)} {U^{(i)}}^T$, where $\Lambda = diag(\lambda_1, \cdots, \lambda_M)$ are eigenvalues in descending order. We can solve for parameters of the i-th mode:

$$C^{(i)} = [u_1^{(i)}, \cdots, u_D^{(i)}] \ , \tag{4}$$

$$R^{(i)} = \sum_{m=D+1}^{M} \lambda_m^{(i)} u_m^{(i)} {u_m^{(i)}}^T \ , \tag{5}$$

$$X_i^{(n)} = {C^{(i)}}^T Y^{(n)} \ , \tag{6}$$

$$V_i^{(n)} = diag(\lambda_1, \cdots, \lambda_D) \ , \tag{7}$$

$$A^{(i)} = \frac{1}{N} \sum_{n=1}^{N} (X_{i,\tau+1}^{(n)} {X_{i,\tau}^{(n)}}^T)(X_{i,\tau}^{(n)} {X_{i,\tau}^{(n)}}^T)^{-1} \ , \tag{8}$$

$$Q^{(i)} = \frac{1}{N} \sum_{n=1}^{N} \varepsilon_\tau \varepsilon_\tau^T / T^{(n)} \ , \tag{9}$$

where $X_{i,\tau}^{(n)} = [\mathbf{x}_1, \cdots, \mathbf{x}_{T^{(n)}-1}]$, $X_{i,\tau}^{(n)} = [\mathbf{x}_2, \cdots, \mathbf{x}_{T^{(n)}}]$, and $\varepsilon_\tau = X_{i,\tau+1}^{(n)} - A^{(i)} X_{i,\tau}^{(n)}$. Note that, if D is not known *a priori*, an empirical determination of the state dimension can be made by choosing D as the cutoff where the eigenvalues λ_m drop below a threshold.

2) Inference Step: Since we estimated the continuous latent variables in the previous step, we only need to run the forward-backward algorithm [4] to estimate $P_\Theta(s_t^{(n)})$ for all time series (for clarity we drop the index n):

$$\begin{aligned} L_t(j) &= \mathit{Likelihood}(\mathbf{y}_t,\ \mathbf{x}_t,\ V_t,\ C^{(j)},\ R^{(j)}) \quad (10) \\ \mathbf{e}_t &= \mathbf{y}_t - C^{(j)}\mathbf{x}_t\ , \\ \mathscr{S}_t &= C^{(j)} V_t {C^{(j)}}^T + R^{(j)}\ , \quad (11) \\ L_t(j) &= \mathscr{N}(\mathbf{e}_t;\ \mathbf{0},\ \mathscr{S}_t)\ . \end{aligned}$$

$$\begin{aligned} M_t^f(i) &= \mathit{Forward}(M_{t-1}^f,\ M_{t+1}^s,\ Z) \quad (12) \\ a_t^f(j) &= L_t(j) M_{t-1}^f(j) Z(i,j)\ , \\ M_t^f(j) &= a_t^f(j) / \sum_{j'} a_t^f(j')\ , \end{aligned}$$

for $t = 1, \cdots, T$. Note, $M_t^f(i) = \mathit{Prob}(s_t = i | y_{1:t})$, with the initial condition $M_0^f = \pi$.

$$\begin{aligned} M_t^s(i) &= \mathit{Backward}(M_t^f,\ M_{t+1}^f,\ M_{t+1}^s,\ Z) \quad (13) \\ a_t^s(i,j) &= M_t^f(i) Z(i,j) M_{t+1}^s(j) / M_{t+1}^f(j)\ , \\ M_t^s(i) &= \sum_j a_t^s(i,j)\ , \end{aligned}$$

for $t = T-1, \cdots, 1$. Note, $M_t^s(i) = \mathit{Prob}(s_t = i | y_{1:T})$, with the initial condition $M_T^s = M_T^f$. This has the computational complexity of $\mathscr{O}(NTJ^2)$, however, since the inference is performed independently we can parallelize this step on N cores, with a cost of $\mathscr{O}(TJ^2)$ per core. Moreover, substituting Eq. (7) into Eq. (11) we see that the innovations covariance matrix $\mathscr{S}_t$ is independent of time, and therefore both the inverse and determinants of the covariance matrices across all modes can be calculated before entering the loop; this has the computational complexity of $\mathscr{O}(JM^3)$ as apposed to $\mathscr{O}(TJM^3)$. This reduction in complexity is a direct consequence of choosing an orthogonal basis for the state-space model (see Eq. (3)), which results in asymptotically diagonal and time-invariant state covariance matrices. However, it is still possible to obtain a $\mathscr{O}(TJD^3)$ complexity in the general case where the state covariance matrices (and thus the innovation covariance matrices) are allowed to change over time (this is particularly useful when $D << M$). This can be achieved by applying the matrix inversion and the determinant inversion lemmas:

$$\begin{aligned} \mathscr{S}_t^{-1} &= {R^{(j)}}^{-1} - {R^{(j)}}^{-1} C^{(j)} {\gamma_t^{(j)}}^{-1} {R^{(j)}}^T {R^{(j)}}^{-1} \quad (14) \\ \left|\mathscr{S}_t\right| &= \left|\gamma_t^{(j)}\right|\ \left|V_t\right|\ \left|R^{(j)}\right|\ , \quad (15) \end{aligned}$$

where $\gamma_t^{(j)} = V_t^{-1} + {C^{(j)}}^T {R^{(j)}}^{-1} C^{(j)}$.

3) Initialization: We used a clustering approach to initialize all J modes. This was done by selecting short random segments from each time series, and fitting a state-space model – as described previously – to each segment. Next, using an appropriate similarity kernel on the space of linear dynamical systems [7] we constructed a similarity matrix among the fitted dynamical systems, and performed spectral clustering with the number of clusters set to J [8]. Next, all the time series belonging to the same cluster were pooled together and were fit to a single linear dynamic system. This step was repeated for all J clusters.

III. Experiments and Results

4) Simulated Time Series with Switching Dynamics: We simulated 100 bivariate time series with a duration of 300 samples and dynamic switching among four modes ($J = 4$). All four dynamical modes were stable bivariate ($M = 2$) autoregression (AR) models of order two. We used two different Markov transition matrices (Z_1 and Z_2) with different stationary distributions to create a balanced binary classification problem.

We assumed the state dimension and the number of modes are known *a priori* and compared the performance of the SKF and SVD-SLDS algorithms for classifying the time series as belonging to one of two classes. After 5 iterations of EM, we used the average time spent within each mode (i.e., time average of $s_t^{(n)}$) as the feature vector to represent each time series, and used a logistic regression classifier (with elastic net regularization) to perform classification. All reported results are 10-fold cross-validated (random draws; 70% training and 30% testing), and the performances are based on the held-out testing sets. We report time complexity of each algorithm for both learning and inference in units of seconds per time series (so, the actual time of learning is 70 and 30 times the numbers reported in the table below for learning and inference, respectively).

Performance of the SKF-based inference and learning versus the proposed SVD-SLDS technique is summarized in Table I. Both algorithms perform equally well (AUC of 0.91 for SKF versus 0.90 for SVD-SLDS), however SVD-SLDS is roughly an order of magnitude faster than the SKF algorithm. Fig. 1 provides a qualitative comparison of the marginal probabilities of each mode using SKF (panel B) versus SVD-SLDS (panel C).

A. MIMIC Dataset

We used a subset of the blood pressure (BP) time series from the MIMIC II database [9], as described in Lehman et al. [5]. Briefly, the cohort included ICU patients with at least 8 hours of continuous minute-by-minute invasive BP trends during the first 24 hours of their ICU stays. Patients with more than 15% missing or invalid (i.e., outside physiologically plausible bounds of 20 to 200 mmHg for mean pressures) BP samples were excluded, resulting in 453 patients (with 16% hospital mortality). The data set contained approximately 9,700 hours of minute-by-minute

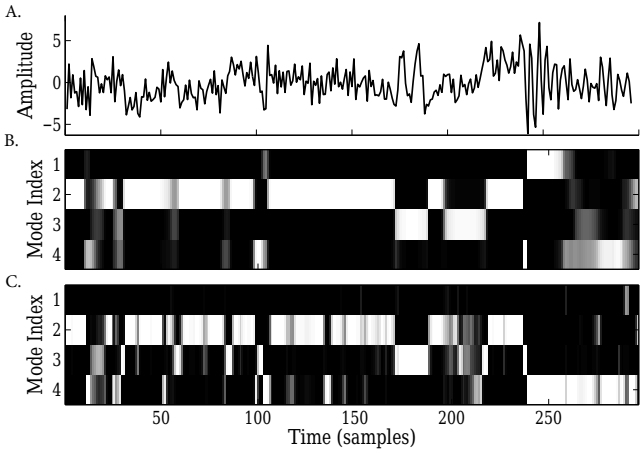

Fig. 1. An example of the simulated time series (only one channel is shown) is shown in panel A. The inferred marginal probabilities of each of the four modes using the SKF and the proposed technique after five iterations of EM are shown in panels B and C, respectively. For a given mode, lighter colors in the grey-scale indicate higher probabilities.

systolic BP measurements (20.2 hours per patient on average, or 1212 samples).

The results presented in Table I are based on using a state dimension $D = 3$ and $J = 10$ modes, with 5 iterations of EM. Here we do not report the performance of the SKF algorithm since the related experiments did not complete after 8 hours of runtime. However, we compared the performance of the SVD-SLDS against the switching vector autoregressive (SVAR) technique of Lehman et al. [5] on the same dataset, which shows negligible difference in performance (AUC of 0.69 versus 0.70). Moreover, the average computational time for leaning on the training dataset (316 time series) and inference on the testing set (137 time series) were 4.8 minutes and 0.3 minutes, respectively.

IV. DISCUSSION AND FUTURE DIRECTION

We presented a novel technique for learning and inference in an SLDS model for cohort time series. The algorithm allows for fast and efficient discovery of shared multivariate dynamical patterns within a large time series cohort. Our simulation studies and an experiment based on real data indicates that the proposed algorithm fares well against the alternative EM-based technique, with at least an order of magnitude improvement in run-time. The SLDS framework is particularly advantageous over other methods (such as the SVAR [5]) when the observation dimension is much larger than the state dimension. Such high-dimensional time series arise in many neuro/physiological recordings involving multi-sensor measurements of a sparse set of underlying sources (e.g., dense electroencephalogram recordings). The inferred continuous latent variables can be interpreted as a low-dimensional representation of the original time series.

Given the sheer volume of multivariate time-series recorded in modern clinical databases, inference over sophisticated models and extraction of multivariate dynamic features are often computationally intensive. The algorithm proposed here can be employed as a fast and memory-efficient technique for dynamics-based time series search,

TABLE I

COMPARISON OF SKF AND SVD-SLDS

Method	performance (AUC)	Learning (sec)	Inference (sec)
	Simulations (N=100)		
SKF	0.91 ± 0.07	1.462 ± 0.084	0.226 ± 0.006
SVD-SLDS	0.90 ± 0.07	0.151 ± 0.016	0.018 ± 0.001
	MIMIC (N=453)		
Li-wei et al. [5]	0.70 ± 0.20	–	–
SVD-SLDS	0.69 ± 0.11	0.926 ± 0.051	0.135 ± 0.024

Performances are based on 10-fold cross-validated area under the curve (AUC) performance on the testing folds. Time complexity for learning and inference are presented in units of seconds per time-series.

with two main components. First, learning of multivariate dynamics and construction of a library of phenotypic dynamical behaviors - this step also involves segmentation and indexing of time-series within the database. Second, given the learned library, the inference step will involve assignment of newly presented patient time series to the most likely dynamic cluster, with the aim of event classification and predictive monitoring. This would allow a clinician to compare incoming patients to those which exhibited similar dynamical activity in the past. We are currently implementing such a search engine using a Hadoop MapReduce framework. Under this framework, one can run the inference for each time-series independently on a separate mapper to calculate the partial latent variable posteriors. Similarly, the maximization-step yields itself to an efficient parallelization over the number of dynamical modes that constitute our library of possible dynamical behaviors (using M reducers running in parallel). Our ultimate aim in showcasing such tool is to facilitate and encourage utilization of high-resolution time series features in clinical studies by the the greater research community.

REFERENCES

[1] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," *ACM SIGKDD*, vol. 12, no. 1, pp. 40–48, 2010.

[2] S. Nemati, L.-w. H. Lehman, R. P. Adams, and A. Malhotra, "Discovering shared cardiovascular dynamics within a patient cohort," in *Engineering in Medicine and Biology Society*, 2012, pp. 6526–6529.

[3] C. Williams, J. Quinn, and N. Mcintosh, "Factorial switching kalman filters for condition monitoring in neonatal intensive care," in *Advances in Neural Information Processing Systems 18*, 2006, pp. 1513–1520.

[4] K. P. Murphy, "Switching kalman filter," *Compaq Cambridge Research Laboratory, Tech. Rep. 98-10.*, 1998, cambridge, MA.

[5] L. Lehman, R. Adams, L. Mayaud, G. Moody, A. Malhotra, R. Mark, and S. Nemati, "A physiological time series dynamics-based approach topatient monitoring and outcome prediction," *Journal of Biomedical and Health Informatics*, 2014.

[6] S. Soatto, G. Doretto, and Y. N. Wu, "Dynamic textures," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 439–446.

[7] S. Vishwanathan, A. J. Smola, and R. Vidal, "Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes," *International Journal of Computer Vision*, vol. 73, no. 1, pp. 95–119, 2007.

[8] A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[9] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter intelligent monitoring in intensive care ii: a public-access intensive care unit database." *Crit Care Med*, vol. 39, no. 5, pp. 952–960, May 2011.