

OPTIMAL TRANSPORT BASED CHANGE POINT DETECTION AND TIME SERIES SEGMENT CLUSTERING

Kevin C. Cheng^{*}, Shuchin Aeron^{*}, Michael C. Hughes^{*}, Erika Hussey[†], Eric L. Miller^{*}

^{*}Tufts University, [†]CCDC-Soldier Center
^{*}[†]{*first name*}.{*last name*} @ tufts.edu

ABSTRACT

Two common problems in time series analysis are the decomposition of the data stream into disjoint segments that are each in some sense “homogeneous” - a problem known as Change Point Detection (CPD) - and the grouping of similar nonadjacent segments, a problem that we call Time Series Segment Clustering (TSSC). Building upon recent theoretical advances characterizing the limiting *distribution-free* behavior of the Wasserstein two-sample test (Ramdas et al. 2015), we propose a novel algorithm for unsupervised, distribution-free CPD which is amenable to both offline and online settings. We also introduce a method to mitigate false positives in CPD and address TSSC by using the Wasserstein distance between the detected segments to build an affinity matrix to which we apply spectral clustering. Results on both synthetic and real data sets show the benefits of the approach.

Index Terms— change point detection, time series segment clustering, Wasserstein two-sample, optimal transport.

1. INTRODUCTION

Change point detection (CPD) is a fundamental problem in data analysis with implications in many real world applications including the analysis of financial [1], electrocardiogram (ECG) [2], and human activity data [3]. Given a collection of change points, time series segment clustering (TSSC) seeks to group nonadjacent periods of activity which are, in some sense, “similar,” in an unsupervised manner. Applications here overlap with those of CPD [4].

In this paper, we focus on the use of statistical methods for CPD which are broadly classified as either parametric (model-based) or non-parametric [5,6]. The problem formulation employed by the majority of methods takes the observations as a sequence of random variables whose distribution changes abruptly at unknown points in time. The processing goal for CPD is to determine when the switches occur and, in those

instances where TSSC is required, use a similarity measure to cluster like segments.

Parametric methods employ a specific model for the dynamics of the time series (either assumed [7] or learned from data [8]) and then make use of decision theory to identify change points. Classically, ARMA-type models and their state-space generalizations were the basis for parametric efforts starting in [9] with recent work focusing on hierarchical models such as switching linear-dynamical systems (SLDS) [10]. Generally, parametric methods are effective when the modelling assumptions hold. For example, SLDS assumes geometric state duration distributions and Gaussian observation models. When these assumptions are not applicable, performance will likely suffer.

When the dynamics or observations cannot be easily modeled, we can consider distribution-free methods that do not assume any particular parametric family of distributions. Change points are then estimated from sample distributions using density-ratio estimates [11, 12] or through two-sample tests like maximum mean discrepancy (MMD) [13], which was recently used for non-parametric CPD [3].

Similar to CPD, parametric TSSC methods have been explored using ARMA based models [14] or HMMs [15]. Non-parametric TSSC methods generally use alternate representations of time series such as frequency-based wavelet decompositions [16] or distribution-based methods [17].

In this work, we contribute a new set of non-parametric CPD and TSSC methods based on recent statistical results in the theory of Optimal Transport (OT). Assuming independent and identically distributed (IID) data, Ramdas et al. [18] provides a theoretical analysis of the asymptotic distribution of an OT-based two-sample test under the null hypothesis for deciding whether two empirical probability density functions are from the same distribution. We use this result as the basis for a sliding window test for identifying change points in a scalar time series. Another novel aspect of our method is the development of a statistically-derived “matched filter” for post-processing our OT statistic to reduce false positives. Given the identified change points, we develop an OT-based spectral clustering scheme for TSSC.

To organize this paper, we start with an overview of optimal transport concepts followed by problem formulation for

This research was supported by funding from Army Research Center Natick via Tufts Center for Applied Brain and Cognitive Sciences (CABCS) under ARM994. Shuchin Aeron was supported in part by NSF CAREER award.

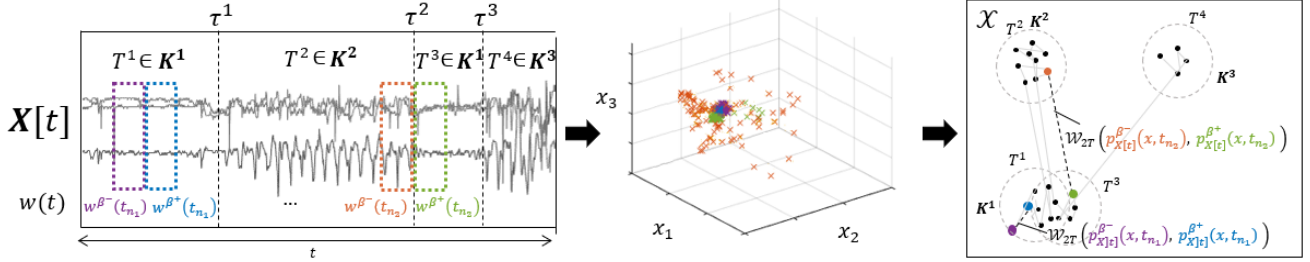


Fig. 1. The time series $X[t] \in \mathbb{R}^3$ can be decomposed at change points τ^i into 4 segments T^i represented by 3 actions. Sample windows of time series data (left) are represented as point cloud in \mathbb{R}^3 (center) which in turn corresponds to a single point in the space of all probability measures \mathcal{X} (right), which is an estimate of the process distribution. The proposed method uses the Wasserstein two-sample test between adjacent windows on each dimension independently as the change point test statistic. At t_{n_1} the two windows belong to similar distributions and thus no change is detected. However at t_{n_2} spanning the change point τ^2 places the distributions of adjacent windows in different clusters K^2, K^1 thus resulting in a high CPD statistic.

CPD and TSSC. We then detail our proposed method and evaluate our techniques on toy and real-world data sets. We show improved precision and recall for CPD (as summarized in F1 scores) compared to state of the art. We also show improved label accuracy in TSSC for human activity data.

2. OPTIMAL TRANSPORT BACKGROUND

Given two probability distributions $p(x), q(y)$, where $x, y \in \mathbb{R}^d$, the 2-Wasserstein distance, or earth mover's distance, $\mathcal{W}_2(p(x), q(y))$ is defined as the minimum expected squared Euclidean cost required to transport $p(x)$ to $q(y)$. Formally,

$$\mathcal{W}_2(p(x), q(y)) = \min_{\pi \in \Pi} \int_x \int_y \|x - y\|_2^2 \pi(x, y) dx dy, \quad (1)$$

$$\int_y \pi(x, y) = p(x), \quad \int_x \pi(x, y) = q(y)$$

where Π denotes the set of all joint distributions. It is well-known that (1) is a linear program. Further, $\mathcal{W}_2(\cdot, \cdot)$ is a metric on the set of probability distributions [19] and metrizes weak convergence of probability measures.

We employ a distribution-free, non-parametric Wasserstein two-sample test (W2T) as a discrepancy measure between two sets of points. To this end, we note the following:

Theorem 2.1 (From [18]) *Under the null hypothesis $H_0 : P = Q$, given empirical CDF's P_m, Q_n consisting of m, n IID samples from scalar distributions P, Q , $\mathcal{W}_{2T}(P_m, Q_n) = \frac{mn}{m+n} \int_0^1 (P_m(Q_n^{-1}(x)) - x)^2 dx \xrightarrow{d} \int_0^1 \mathcal{B}^2(x) dx = \mathbb{B}_2$*

where $\mathcal{B}(x)$ denotes the standard Brownian motion. From [20], \mathbb{B}_2 has mean $\mu_{\mathbb{B}_2} = 0.166$ and that we reject the null with confidence $\alpha = 0.05$ using a threshold of $\lambda = 0.462$.

3. PROBLEM FORMULATION

As detailed in Fig. 1 and throughout, we consider a time series $X[t] \in \mathbb{R}^d$, $t = 1, 2, \dots$, where the data consists of dis-

tinct time segments $[0, \tau^1], [\tau^1 + 1, \tau^2], \dots, [\tau^{S-1} + 1, \tau^S]$, with $\tau^1 < \tau^2 < \dots$ such that within each time segment, $X[t], t \in [\tau^{i-1} + 1, \tau^i], i = 1, 2, \dots, S$, are IID samples from one of $k = 1, 2, \dots, K$ unknown distributions, where we assume here that K is known *a priori*. The problem of change point detection (CPD) is to estimate τ^i , and the problem of time series segment clustering (TSSC) is to cluster the S segments into K classes.

4. PROPOSED METHOD

4.1. Change Point Detection

Given time-series $X[t]$, we define two empirical probability density functions (PDFs) at each time t generated from the sum of dirac-delta functions supported on a window of β samples collected before and after t yielding $p_{X[t]}^{\beta\pm}(x) = \frac{1}{\beta} \sum_{\tau=1}^{\beta} \delta(x - X[t \pm \tau])$. After transforming each PDF p into a cumulative distribution function (CDF) P , we can compute a change point statistic from the Wasserstein two-sample test (W2T) between the CDFs of the two windows:

$$\sigma[t] = \mathcal{W}_{2T} \left(P_{X[t]}^{\beta-}(x), P_{X[t]}^{\beta+}(x) \right). \quad (2)$$

The nominal, offline approach to CPD is to label local maxima of $\sigma[t]$ that exceed some threshold parameter as change points [3]. Shown through empirical analysis on both simulated and real data, we find this is problematic. Fig. 2 indicates the presence of spurious local maxima leading to a large number of false alarms and ambiguity in the change point locations. Moreover, the sliding-window nature of the processing causes a change point at time t to create an extended signature in $\sigma[t]$ over the interval $[t - \beta, t + \beta]$.

These observations suggest the benefit of a matched filtering approach to reduce the spurious maxima and better localize true changes. In Fig. 3 we estimate the shape of this signature empirically by averaging over ensembles of simulated IID data with a known change points separating sam-

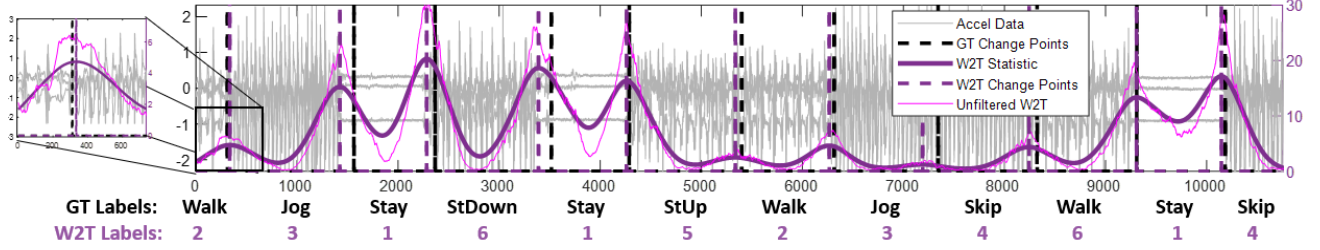


Fig. 2. CPD and TSSC results from HASC2016-PAC data (black, left axis). For CPD, we plot both unfiltered (thin purple) and match filtered (thick purple) change point statistics (right axis). The left subplot shows how the matched filter removes false positives and improves localization of the change point. For TSSC, the bottom row shows our method’s assigned cluster labels, which make only one mistake relative to ground truth (GT) by grouping the stair-down segment with the last walk segment.

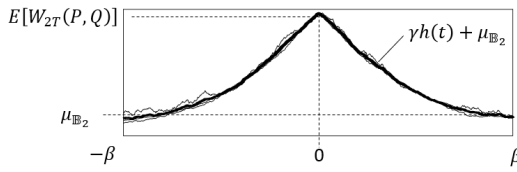


Fig. 3. Un-normalized empirically estimated matched filter. Given a change point at $t = 0$ and window size β , the effects of the change point are reflected in the W2T statistic on the interval $[-\beta, \beta]$. Thin traces represent the ensemble average of 200 IID sequences with different simulated change points ($N(0, 1) \rightarrow N(0.2, 1)$, $N(0, 1) \rightarrow N(0, 1.2)$ and $N(0, 1) \rightarrow L(0, \frac{1}{\sqrt{2}})$). The matched filter $h[t]$ is normalized by removing the bias $\mu_{\mathbb{B}_2}$ and scaling by γ to have unit area.

ples from different pairs of distribution. From this plot, we observe that the structure of this function across a number of distributional changes is remarkably consistent. The theoretical analysis and discussion of this filter is left to future work. We derive the filter $h[t]$ by removing the bias and normalizing by a constant such that the peaks of $\sigma[t]$ signal are preserved. Change points are the set of local maxima where $\sigma[t] \otimes h[t]^1$ exceeds a threshold: $\{\tau\} = \{t \mid \text{peaks}(\sigma[t] \otimes h[t]) > \lambda\}^2$.

4.2. Time Series Segment Clustering

Given change points $\{\tau^i\}$ and time segments $T^i = \{X[t] \mid \tau^i < t \leq \tau^{i+1}\}$, the process distribution in this time segment is estimated by $p^{T^i} = (\sum_j w_j)^{-1} \sum_{j=\tau^i}^{\tau^{i+1}} w_j \delta(x - X[j])$. This represents a weighted point cloud generated from the data points over the time interval. Samples are weighted by a windowing function that down-weights samples around transition boundaries, mitigating the effects of segmentation errors and non-instantaneous transitions. To this effect, we use a half Hamming window of length 2β for samples within β of either boundary. Samples outside this range have weights $w_j = 1$.

The similarity matrix between time segments $A[i, j] =$

¹Where \otimes denotes the convolution operation

² $t \in \{\text{peaks}(f[t])\} \iff f[t] > f[t-1] \ \& \ f[t] > f[t+1]$

Algorithm 1 Wasserstein Change Point Detection and Time Series Segment Clustering

Input: $X[t], \beta, K, h[t], \lambda, \{w_j\}_{j=1}^\beta$

Output: $\{\tau\}, c$

for all t **do** # CPD

$$p_{X[t]}^{\beta^-}(x) = \frac{1}{\beta} \sum_{i=1}^{\beta} \delta(x - X[t - i])$$

$$p_{X[t]}^{\beta^+}(x) = \frac{1}{\beta} \sum_{i=1}^{\beta} \delta(x - X[t + i])$$

$$\sigma[t] = \mathcal{W}_{2T}(P_{X[t]}^{\beta^-}, P_{X[t]}^{\beta^+})$$

end for

$\{\tau\} = \{t \mid \text{peaks}(\sigma[t] \otimes h[t]) > \lambda\}$

$$p^{T^i} = \frac{1}{\sum_j w_j} \sum_{j=\tau^i}^{\tau^{i+1}} w_j \delta(x - X[j]) \quad \# \text{TSSC}$$

for all $0 \leq i, j < |\tau|$ **do**

$$A(i, j) = \exp(-\mathcal{W}_2(p^{T^i}, p^{T^j}))$$

end for

$c = \text{SpectralClustering}(A, K)$

$\exp(-\mathcal{W}_2(p^{T^i}, p^{T^j}))$, uses the 2-Wasserstein distance between their respective empirical distributions as the distance measure. Given the number of action clusters K , we utilize the similarity graph structure under the Wasserstein metric by clustering time segments via spectral clustering [21] into the optimal action clusters.

5. EVALUATION

5.1. Evaluation Criteria

We use the area under the ROC curve (CP-AUC) to evaluate change point performance, following previous work [3, 11, 22]. We also report the F1 score (CP-F1) for offline multiple CPD [23] using a margin of error δ for the acceptable offset to the true label.

For TSSC, cluster labels are mapped onto the ground truth labels using the standard Munkres algorithm and evaluated using the Hamming distance. Performance is reported in Tab. 1 separately using ground truth change points (GT) and learned change points (W2T or MStat).

Data	K	β	δ	CP-AUC		CP-F1		Label Acc		
				W2T	MStat	W2T	MStat	GT	W2T	MStat
Beedance	3	14	14	0.527	0.549	0.647	0.625	0.705	0.651	0.646
HASC-PAC2016	6	500	250	0.689	0.658	0.748	0.713	0.789	0.658	0.675
HASC2011	6	500	250	0.576	0.585	0.824	0.770	0.565	0.498	0.382
ECG200	2	100	50	0.585	0.584	0.637	0.582	0.864	0.708	0.716

Table 1. CPD evaluation using AUC and F1 for proposed W2T method and MStat for given number of labels K , window size β , and detection delay δ . TSSC is evaluated with label Hamming accuracy using ground truth, W2T, and MStat change points.

5.2. Experimental Setup

We compare the performance of our algorithm to the M-Statistic (MStat) [3], setting parameters $N = 1$, $M = \beta$. For fair comparison, we employ a MStat matched filter $h_M[t]$ using a method analogous to that outlined in Sec. 4.1. The only hyperparameters to the CPD model are the window size β and detection threshold λ . Since the window size controls the width of the matched filter, we utilize domain knowledge to set β based on the expected frequency of changes. We also set the threshold parameter $\lambda = 0$ as the distribution of the MStat under the null is not known. The hyperparameters, along with the true positive detection window δ used for F1 can be found in Tab. 1. For vectored time series we computed the W2T over each dimension and averaged the result. We evaluate on the following datasets:

HASC-PAC2016: [24] consists of over 700 three-axis accelerometer sequences of subjects performing six actions: 'stay', 'walk', 'jog', 'skip', 'stairs up', and 'stairs down'. We evaluate on the 92 longest sequences.

HASC-2011: three-axis accelerometer data from 6 actions: 'stay', 'walk', 'escalator up', 'elevator up', 'stairs up', and 'stairs down'.

Beedance: [25] movements of dancing honeybees who communicate through three actions: "turn left", "turn right" and "waggle". We use the gradient of the data as our input.

ECG200: [2] detection of abnormal heartbeats in ECG.

5.3. Results

The proposed algorithm demonstrates robust results for CPD and TSSC. Fig. 2 shows clear detection of change points on HASC-PAC2016, strong efficacy of the matched filter in reducing false positives, and a single label mis-classification.

The CPD performance for the W2T and MStat are comparable under the AUC metric, however, under the F1 metric, W2T consistently performs better. We note that the computation complexity of the W2T ($O(\beta \log(\beta))$) is an improvement compared to that of the MStat ($O(\beta^2)$) and that the OT measures show tighter clustering in the low-dimensional embedding of various simulated measures (Fig. 4).

Comparing to results reported in [22], our unsupervised method shows competitive results with an AUC of 0.527 on Beedance compared to supervised parametric models such as

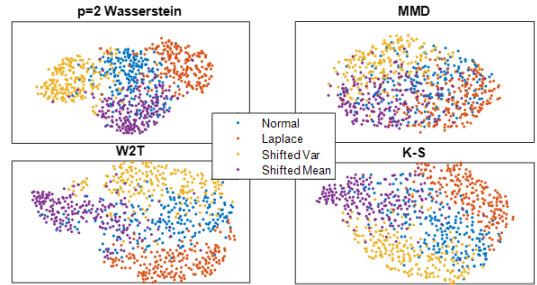


Fig. 4. t-SNE embedding of simulated data of 200 windows of IID 100 samples from 4 simulated distributions: $\mathcal{N}(0, 1)$ (blue), $\mathcal{L}(0, \frac{1}{\sqrt{2}})$ (orange), $\mathcal{N}(0.2, 1)$ (yellow), $\mathcal{N}(0, 1.2)$ (purple) using Wasserstein metric, and two-sample tests: MMD, W2T, and Kolmogorov-Smirnov

ARMA (0.537) and ARGP (0.583). We observe that since $h[t]$ smooths the test statistic, its inclusion decreases AUC for a better F1 score, which we see as a positive tradeoff. For example, when including $h[t]$ for HASC2011, our AUC drops from 0.630 to 0.576 while F1 improves from 0.720 to 0.824.

In terms of TSSC, using our unsupervised, distribution-free approach, we are able to achieve a 65% label accuracy on the Beedance data. For comparison, a state of the art supervised parametric model [25] achieves an 87.7% label accuracy, and a parametric unsupervised model using switching vector autoregressive HMMs [26] achieves a label accuracy of 66.8%. HASC also shows strong performance given that a total of six possible assignments were available.

6. DISCUSSION

We propose a distribution-free, unsupervised approach to CPD and TSSC for time-series data. In our experiments, we run the CPD in an offline manner. Applied in an online setting, the minimum detection delay would be 2β .

We approach CPD and TSSC with a weak set of assumptions: that change points occur when the process distribution changes, and actions can be clustered based on their respective empirical distributions. However, clearly time series data is rarely IID. In future work, we will expand these methods for CPD and TSSC beyond IID assumptions.

7. REFERENCES

- [1] Carlos M. Carvalho and Hedibert F. Lopes, "Simulation-based sequential analysis of Markov switching stochastic volatility models," *Computational Statistics & Data Analysis*, vol. 51, no. 9, May 2007.
- [2] Hoang Anh Dau et al., "The ucr time series classification archive," October 2018, https://www.cs.ucr.edu/~eamonn/time_series_data_2018.
- [3] Shuang Li, Yao Xie, Hanjun Dai, and Le Song, "M-Statistic for Kernel Change-Point Detection," in *Advances in Neural Information Processing Systems* 28. 2015.
- [4] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah, "Time-series clustering A decade review," *Information Systems*, vol. 53, Oct. 2015.
- [5] Abraham Wald, *Sequential analysis*, J. Wiley & Sons ; Chapman & Hall, New York; London, 1947.
- [6] Michèle Basseville and Igor V. Nikiforov, *Detection of Abrupt Changes - Theory and Application*, Prentice Hall, Inc., 1993.
- [7] F. Chamroukhi, S. Mohammed, D. Trabelsi, L. Oukhelou, and Y. Amirat, "Joint segmentation of multivariate time series with hidden process regression for human activity recognition," *Neurocomputing*, vol. 120, Nov. 2013.
- [8] Wei-Han Lee, Jorge Ortiz, Bongjun Ko, and Ruby Lee, "Time Series Segmentation through Automatic Feature Learning," *arXiv:1801.05394 [cs, stat]*, Jan. 2018, arXiv: 1801.05394.
- [9] A. Willsky and H. Jones, "A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems," *IEEE Transactions on Automatic Control*, vol. 21, no. 1, Feb. 1976.
- [10] Kevin P. Murphy, "Switching Kalman Filters," Tech. Rep., 1998.
- [11] Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama, "Change-point detection in time-series data by relative density-ratio estimation," *Neural Networks*, vol. 43, July 2013.
- [12] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama, "A least-squares approach to direct importance estimation," *J. Mach. Learn. Res.*, vol. 10, pp. 13911445, Dec. 2009.
- [13] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, Mar. 2012.
- [14] Marcella Corduas and Domenico Piccolo, "Time series clustering and classification by the autoregressive metric," *Computational Statistics & Data Analysis*, vol. 52, no. 4, Jan. 2008.
- [15] Tim Oates, Laura Firoiu, and Paul R. Cohen, "Clustering Time Series with Hidden Markov Models and Dynamic Time Warping," 1999.
- [16] Yk Huhtala, Juha Krkkinen, and Hannu Toivonen, *Mining for Similarities in Aligned Time Series Using Wavelets*, 1999.
- [17] R. Dahlhaus, "On the Kullback-Leibler information divergence of locally stationary processes," *Stochastic Processes and their Applications*, vol. 62, no. 1, Mar. 1996.
- [18] Aaditya Ramdas, Nicols Garca Trillos, and Marco Cuturi, "On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests," *Entropy*, vol. 19, 2015.
- [19] Gabriel Peyr and Marco Cuturi, "Computational Optimal Transport," *arXiv:1803.00567 [stat]*, Mar. 2018, arXiv: 1803.00567.
- [20] Leonid Tolmatz, "On the Distribution of the Square Integral of the Brownian Bridge," *The Annals of Probability*, vol. 30, no. 1, Jan. 2002.
- [21] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, 2007.
- [22] Wei-Cheng Chang, Chun-Liang Li, Yiming Yang, and Barnabs Pczos, "Kernel Change-point Detection with Auxiliary Deep Generative Models," *arXiv:1901.06077 [cs, stat]*, Jan. 2019, arXiv: 1901.06077.
- [23] Charles Truong, Laurent Oudre, and Nicolas Vayatis, "Selective review of offline change point detection methods," *arXiv:1801.00718 [cs, stat]*, Jan. 2018, arXiv: 1801.00718.
- [24] Haruyuki Ichino et al., "HASC-PAC2016: large scale human pedestrian activity corpus and its baseline recognition," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Heidelberg, Germany, 2016.
- [25] Sang Min Oh, James M. Rehg, Tucker Balch, and Frank Dellaert, "Learning and Inferring Motion Patterns using Parametric Segmental Switching Linear Dynamic Systems," *International Journal of Computer Vision*, vol. 77, no. 1-3, May 2008.
- [26] Emily Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky, "Nonparametric Bayesian Learning of Switching Linear Dynamical Systems," in *Advances in Neural Information Processing Systems* 21. 2009.