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
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, \( W_2(p(x), q(y)) \) is defined as the minimum expected squared Euclidean cost required to transport \( p(x) \) to \( q(y) \). Formally,

\[
W_2(p(x), q(y)) = \min_{\pi \in \Pi} \int \int \| x - y \|^2 \pi(x, y) dx dy,
\]

where \( \Pi \) denotes the set of all joint distributions. It is well-known that (1) is a linear program. Further, \( W_2(\cdot, \cdot) \) 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 \( P, Q \), \( W_{2T}(P_m, Q_n) = \min_{\beta \in \mathbb{R}} \int_0^{1/2} \left( \int_0^1 P_{X[t]}(x) - Q_{X[t]}(x) dx \right)^2 dx = \mathbb{B}_2 \)

where \( B(x) \) 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 \( X[t] \in \mathbb{R}^d \), \( t = 1, 2, \ldots \), where the data consists of distinct time segments \([0, \tau_1], [\tau_1 + 1, \tau_2], \ldots, [\tau^{S-1} + 1, \tau^S]\), with \( \tau_1 < \tau_2 < \cdots \) such that within each time segment, \( X[t], t \in [\tau^{i-1} + 1, \tau^i], i = 1, 2, \ldots, S \), are IID samples from one of \( k = 1, 2, \ldots, K \) unknown distributions, where we assume here that \( K \) is known a priori. The problem of change point detection (CPD) is to estimate \( \tau^i \), 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 \( \beta \) samples collected before and after \( t \) yielding \( p_{X[t]}(x) = \frac{1}{\beta} \sum_{\tau = -1}^{\tau = 1} \delta(x - X[t \pm \tau]) \). After transforming each PDF 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:

\[
\sigma[t] = W_{2T} \left( P_{X[t]}^\beta(x), P_{X[t]}^{\beta+}(x) \right).
\]

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-
Given change points exceed a threshold: 

\[
\{ \{ \tau \} \}
\]

of 200 IID sequences with different simulated change points \(\tau\). 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. 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.

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

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

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

\[
\sigma[t] = W_2(P^\beta_X[t], P^\beta_X[t])
\]

end for

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

\[
p^{T^*} = \sum_{i=1}^{\lambda} \sum_{j=i+1}^{\beta} w_j \delta(x - X[j])
\]

\[
\text{TSSC}
\]

for all \(0 \leq i, j < |\tau|\) do

\[
A(i, j) = exp(-W_2(p^{T^*}, p^{T^*}))
\]

end for

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

exp \((-W_2(p^{T^*}, p^{T^*}))\), 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 \(\delta\) 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).
### Table 1. CPD evaluation using AUC and F1 for proposed W2T method and MStat for given number of labels $K$, window size $\beta$, and detection delay $\delta$. TSSC is evaluated with label Hamming accuracy using ground truth, W2T, and MStat change points.

<table>
<thead>
<tr>
<th>Data</th>
<th>$K$</th>
<th>$\beta$</th>
<th>$\delta$</th>
<th>CP-AUC W2T</th>
<th>MStat</th>
<th>CP-F1 W2T</th>
<th>MStat</th>
<th>Label Acc GT</th>
<th>W2T</th>
<th>MStat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beedance</td>
<td>14</td>
<td>14</td>
<td></td>
<td>0.527</td>
<td>0.549</td>
<td>0.647</td>
<td>0.625</td>
<td>0.705</td>
<td>0.651</td>
<td>0.646</td>
</tr>
<tr>
<td>HASC-PAC2016</td>
<td>6</td>
<td>500</td>
<td>250</td>
<td>0.689</td>
<td>0.658</td>
<td>0.748</td>
<td>0.713</td>
<td>0.789</td>
<td>0.658</td>
<td>0.675</td>
</tr>
<tr>
<td>HASC2011</td>
<td>6</td>
<td>500</td>
<td>250</td>
<td>0.576</td>
<td>0.585</td>
<td>0.824</td>
<td>0.770</td>
<td>0.565</td>
<td>0.498</td>
<td>0.382</td>
</tr>
<tr>
<td>ECG200</td>
<td>2</td>
<td>100</td>
<td>50</td>
<td>0.585</td>
<td>0.584</td>
<td>0.637</td>
<td>0.582</td>
<td>0.864</td>
<td>0.708</td>
<td>0.716</td>
</tr>
</tbody>
</table>

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 $\beta$ and detection threshold $\lambda$. Since the window size controls the width of the matched filter, we utilize domain knowledge to set $\beta$ 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 $\delta$ 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:


**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 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.

### 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.

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 $2/\beta$.

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


