Outlier-Robust Convex Segmentation

Itamar Katz, Koby Crammer
Department of Electrical Engineering
The Technion – Israel Institute of Technology
Haifa, 32000 Israel
(itamark@tx,koby@ee).technion.ac.il

Abstract

We address the task of segmenting sequential data using convex optimization problem, which is specifically designed to work in the context of outliers in the data. We propose two algorithms for solving this problem, one exact and one a top-down hierarchical approach. Robustness to outliers is evaluated on a real-world task related to speech segmentation. Our algorithms outperform baseline segmentation algorithms.

1 Introduction

Segmentation of sequential data, also known as change-point detection, is a fundamental problem in the field of unsupervised learning, and has applications in diverse fields such as speech processing [13] and bioinformatics [12], to name just a few. We are interested in formulating segmentation as a convex optimization problem that avoids issues such as local-minima or sensitivity to initialization. In addition, we want to explicitly incorporate robustness to outliers. Our starting point is a convex objective that minimizes the sum of squared distances of samples $\mathbf{x}_i$ from each sample's associated 'centroid', $\boldsymbol{\mu}_i$. Identical adjacent $\boldsymbol{\mu}_i$s identify their corresponding samples as belonging to the same segment. In addition, some of the samples are identified as outliers, allowing reduced loss on these samples. Two regularization terms are added to the objective, in order to constrain the number of detected segments and outliers, respectively.

We propose two algorithms based on this formulation. The first algorithm, denoted by Outlier-Robust Convex Sequential (ORCS) segmentation, solves the optimization problem exactly, while the second is a top-down hierarchical version of the algorithm, called TD-ORCS.

We evaluate the performance of the proposed algorithms on a speech segmentation task, for both clean source and source contaminated with added non-stationary noise. Our algorithms outperform other algorithms in both the clean and outlier-contaminated setting.

Notation The samples to be segmented are denoted by $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$, and their associated quantities are $\boldsymbol{\mu}_i, \mathbf{z}_i \in \mathbb{R}^d$. The same notation with no subscript, $\boldsymbol{\mu}$, denotes the collection of all $\boldsymbol{\mu}_i$s. The same holds for $\mathbf{x}, \mathbf{z}$.

2 Outlier-Robust Convex Segmentation

Segmentation is the task of dividing a sequence of n data samples $\{\mathbf{x}_i\}_{i=1}^n$ into K groups of consecutive samples, or segments, such that each group is homogeneous with respect to some criterion. A common choice of such a criterion often involves minimizing the squared Euclidean distance of a sample to some representative sample $\boldsymbol{\mu}_i$. Since this criterion is highly sensitive to outliers, it is desirable to incorporate robustness to outliers into the model. We achieve this by allowing some of the input samples $\mathbf{x}_i$ to be identified as outliers, for which we do not require $\boldsymbol{\mu}_i$ to be close to these samples. To this end we propose to

minimize:

$$\min_{\boldsymbol{\mu},\boldsymbol{z}} \left\{ \frac{1}{2}\sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{z}_i - \boldsymbol{\mu}_i\|^2 + \lambda \sum_{i=1}^{n-1} w_i \left\|\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}_i\right\|_p \right. \quad (1)$$
$$\left. + \gamma \sum_{i=1}^{n} \|\boldsymbol{z}_i\|_q \right\},$$

where $p, q \geq 1$ and $\lambda, \gamma > 0$ are parameters, and w_i are weights to be set. Considering samples $\mathbf{x}_i$ for which $\boldsymbol{z}_i = 0$, the first term measures the loss of replacing a point $\mathbf{x}_i$ with some shared point $\boldsymbol{\mu}_i$, and can be thought of as minus the log-likelihood under Gaussian noise. Samples i with $\boldsymbol{z}_i \neq 0$ are intuitively identified as outliers and contribute a reduced loss to the objective. The second and third terms are ℓ_1 regularizations, which are a convex relaxation of counting the number of segments and outliers, respectively. Since ℓ_1 norm induces sparsity, for some samples $\{i, i+1\}$ it will hold that $\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}_i = 0$ exactly, identifying such samples as belonging to the same segment. Similarly for the $\boldsymbol{z}_i$s, some will satisfy $\boldsymbol{z}_i = 0$, identifying them as non-outliers. We note that a similar approach to robustness was employed to robust clustering [7] and robust PCA [11]. The parameter λ indirectly controls the number of detected segments, via the tradeoff between the first and second terms. Similarly, γ controls the amount of outliers. In what follows we set $q = p = 2$. Note that $q = 1$ encourages sparsity of coordinates of $\boldsymbol{z}_i$, and not of the vector as a whole.

2.1 Algorithms

Eq. (1) can be optimized in an alternating manner, and we call this algorithm Outlier-Robust Convex Sequential (ORCS) segmentation. Holding $\boldsymbol{\mu}$ constant, optimizing over $\boldsymbol{z}$ is done analytically. Holding $\boldsymbol{z}$ constant, optimizing over $\boldsymbol{\mu}$ is done by defining $\hat{\mathbf{x}}_i \triangleq \mathbf{x}_i - \boldsymbol{z}_i$, which results in the following optimization:

$$\min_{\boldsymbol{\mu}} \left\{ \sum_{i=1}^{n} \|\hat{\mathbf{x}}_i - \boldsymbol{\mu}_i\|^2 + \lambda \sum_{i=1}^{n-1} w_i \left\|\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}_i\right\|_2 \right\}. \quad (2)$$

Eq. (2) can be solved exactly in a few manners. The proximal-gradient approach [3] for solving non-smooth convex problems suffers in this case from step size which decreases with the number of samples n. This issue can be avoided using the dual formulation of Eq. (2) [2]. Yet another approach was proposed by Bleakley and Vert [4] for the task of change-point detection, who showed that Eq. (2) can be formulated as group-LASSO regression [15].

In addition to the exact solution, we now derive an alternative top-down, greedy algorithm for solving Eq. (2), which enables direct user-control of the resulting number of segments K and can be implemented in $\mathcal{O}(nK)$. We found empirically that this algorithm performs better in some situations. The algorithm works in rounds. On each round it goes over all segments of a current segmentation, and finds the optimal segmentation of each segment into two subsequences. The algorithm then splits the segment which results in a maximal decrease of the sum-of-squared-errors criterion. The optimal split into two segments is found analytically, by analyzing the transition of the solution to Eq. (2) from $K = 1$ to $K = 2$ segments. For the unweighted case, that is $w_i = 1$ for all i, the analytical solution is biased towards segments of approximately the same length. This bias is somewhat alleviated by setting $w_i = \sqrt{i(n-i)}$. We note that the same choice for w_i was derived by Bleakley and Vert [4] from different considerations based on a specific noise model.

Robust top-down algorithm Based on the algorithms described above, we propose a robust top-down algorithm for approximately optimizing Eq. (1), where the number of segments K and outliers M are user-controlled parameters. The algorithm alternates between splitting a segment of current segmentation into two subsequences, as described above, and detecting outliers which is done analytically. In each iteration the algorithm chooses the segment-split which results in the maximal decrease in the squared loss, and whenever a segment is split, the number of outliers belonging to each sub-segment is kept and used in the next iteration, so the number of outliers equals M at all iterations. The algorithms stops when K segments are detected.

3 Empirical Study

We used a 35 minutes, hand-annotated audio recording of a radio talk show, composed of different sections such

as opening title, monologues, dialogs, and songs. A detected segment boundary is considered a true positive if it falls within a tolerance window of two frames around a ground-truth boundary. Segmentation quality is commonly measured using the F measure, which is defined as $2pr/(p+r)$, where p is the precision and r is the recall. However, we used the R measure introduced by Räsänen et al. [14], which is more robust to over-segmentation. The R measure satisfies $R \leq 1$, and $R = 1$ only if $p = r = 1$.

Signal representation The common MFCC representation for speech analysis is typically computed over time windows of tens of milliseconds, and therefore it is not designed to capture phenomena in the order of seconds or minutes. We therefore used the following representation. First, the raw audio is divided into N non-overlapping, 5 seconds blocks, and the MFCC coefficients are computed for all blocks $\{S_j\}_{j=1}^{N}$. Then a Gaussian Mixture Model (GMM) T_j with 10 components and a diagonal covariance matrix is fitted to the jth block S_j. We then define the matrix $A_{ij} = \log \mathbb{P}(S_j|T_i)$. Which is shown in Fig. 1(a). Since using the columns of A as features yields a dimension growing with N, we randomly choose a subset of $d = 100$ rows of A, and the columns of the resulting matrix $X \in \mathbb{R}^{d \times N}$ are the input to the segmentation algorithm. We repeat the experiment for outlier percentage ranging between 0% and 16% with intervals of 2%. A given percentage of outliers refers to the relative number of blocks randomly selected as outliers, to which we add a 5 seconds recording of repeated hammer strokes, normalized to a level of 0dB SNR.

Algorithms We consider the Outlier-Robust Convex Sequential (ORCS) segmentation, and its top-down versions (weighted and unweighted) which we denote by WTD-ORCS and TD-ORCS, respectively. We compare the performance to three other algorithms. The first is a greedy bottom-up (BU) segmentation algorithm, which minimizes the sum of squared errors on each iteration, and which was successfully used in tasks of speech segmentation [8]. The second algorithm is the W-LARS algorithm of Bleakley and Vert [4] [1]. The third algorithm is a Bayesian change-point detection algorithm (BCP), as formulated by Erdman and Emerson [6] [2]. For the ORCS algorithm, the solution over a grid of λ, γ values was calculated. For the TD-ORCS, W-LARS, and BU algorithms, $K = 2, \ldots, 150$ number of segments were used as an input to the algorithms. For the TD-ORCS algorithm, where the number of required outliers is an additional input parameter, the correct number of outliers was used. For the BCP algorithm, a range of thresholds on the posterior probability of change-points was used to detect a range of number of segments. For each algorithm, the maximal R measure over all parameters range was used to compare all algorithms.

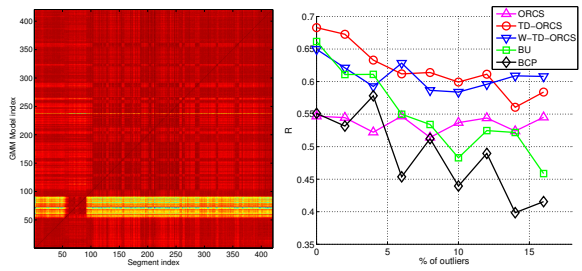
Results Results are shown in Fig. 1(b) as the maximal R measure achieved versus the percentage of outliers, for each of the algorithms considered. It is evident that the performance of the BU and BCP algorithms decreases significantly as more outliers are added, while the outlier-robust ORCS algorithm keeps an approximately steady performance. Our unweighted and weighted TD-ORCS algorithms achieve the best performance for all levels of outliers. Results for LARS algorithm are omitted as it did not perform better then other algorithms. We verified the ability of our algorithms to correctly detect outliers by calculating the R measure of the outliers detection of the ORCS algorithm, with zero length tolerance window, i.e a detection is considered a true-positive only if it exactly pinpoints an outlier. The R measure was evaluated on the γ, λ parameter grid, as well as the corresponding numbers of detected outliers. We found that a high R measure (> 0.9) is attained on a range of parameters that yield around the true number of outliers. We conclude that one does not need to know the exact number of outliers in order to use the ORCS algorithm, and a rough estimate is enough. Some preliminary results suggest that such an estimate can be approximated from the histogram of the number of detected outliers.

4 Related Work and Conclusion

There is a large amount of literature on change-point detection [1, 5]. Optimal segmentation can be found using

[1] `http://cbio.ensmp.fr/~jvert/svn/GFLseg/html/`

[2] `http://cran.r-project.org/web/packages/bcp/index.html`



(a) Likelihood matrix (b) Maximal R measure vs. % of outliers

Figure 1: (a) A_{ij} is the log-probability of segment j given the GMM fitted to segment i. (b) Maximal R measure vs. percentage of outliers. See text for algorithms details.

dynamic programming [9]; however, the complexity of this approach is quadratic in the number of samples n, and therefore might be infeasible for large data sets. Some approaches which achieve complexity linear in n [10] treat only one dimensional data. Some related work is concerned with the objective Eq. (2) we presented in Sec. 2.1. In Levy-leduc et al. [10] it was suggested to reformulate Eq. (2) for the one dimensional case as a LASSO regression problem [15], while Bleakley and Vert [4] extended this approach to multidimensional data, although not treating outliers directly. Another common approach is deriving an objective from a maximum likelihood criterion of a generative model, and then either optimize the objective or use it as a criterion for a top-down or a bottom-up approach [13, 8, 12]. Finally, we note that all these approaches do not directly incorporate outliers into the model.

We formulated the task of segmenting sequential data and detecting outliers using convex optimization, which can be solved in an alternating manner. We showed that a specific choice of weighting can empirically enhance the performance. We also derived a top-down, outlier-robust hierarchical segmentation algorithm which minimizes the objective in a greedy manner. This algorithm allows for directly controlling both the number of desired segments K and number of outliers M. Experiments with real-world audio data with outliers added manually to the raw audio demonstrated the superiority of our algorithms. We consider a few possible extensions to the current work. One is deriving algorithms that will work on-the-fly. Another direction is to investigate more involved noise models, such as noise which corrupts a single feature along all samples, or noise which corrupts a consecutive set of samples. Yet another interesting question is how to identify that different segments come from the same source, e.g. that the same speaker is present at different locations in a recording. We plan to investigate these directions in future work.

References

[1] Michéle Basseville and Igor V Nikiforov. Detection of abrupt changes: theory and application. 1993.

[2] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *Image Processing, IEEE Transactions on*, 18(11):2419 –2434, nov. 2009.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2(1), 2009.

[4] Kevin Bleakley and Jean-Philippe Vert. The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199*, 2011.

[5] Boris E Brodsky and Boris S Darkhovsky. *Nonparametric methods in change point problems*, volume 243. Springer, 1993.

[6] Chandra Erdman and John W Emerson. A fast bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*, 24(19):2143–2148, 2008.

[7] Pedro A Forero, Vassilis Kekatos, and Georgios B Giannakis. Outlier-aware robust clustering. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 2244–2247. IEEE, 2011.

[8] C. Gracia and X. Binefa. On hierarchical clustering for speech phonetic segmentation. In *Eusipco*, 2011.

[9] M. Lavielle and G. Teyssière. Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287–306, 2006.

[10] Céline Levy-leduc et al. Catching change-points with lasso. In *Advances in Neural Information Processing Systems*, pages 617–624, 2007.

[11] Gonzalo Mateos and Georgios B Giannakis. Robust PCA as bilinear decomposition with outlier-sparsity regularization. *Signal Processing, IEEE Transactions on*, 60(10):5176–5190, 2012.

[12] Adam B Olshen, ES Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, 2004.

[13] Yu Qiao, N. Shimomura, and N. Minematsu. Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons. In *ICASSP*, 2008.

[14] Okko J. Räsänen, Unto K. Laine, and Toomas Altosaar. An improved speech segmentation quality measure: the r-value. In *INTERSPEECH*, 2009.

[15] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.