Supplementary Material for Temporally Coherent CRP: A Bayesian Non-Parametric Approach for Clustering Tracklets with applications to Person Discovery in Videos

Adway Mitra* Soma Biswas† Chiranjib Bhattacharyya‡

1 General Problem Definition

In the paper, we addressed the problem of entity discovery in videos through tracklet clustering. For this purpose we proposed a Bayesian nonparametric approach, with a model (TC-CRP). However, our approach is not tied to tracklet clustering or person discovery, and is a solution to a more general in Data Mining, as follows:

We consider sequential data, where each datapoint has a predecessor and a successor datapoint. There are an unknown number of entities, and each datapoint is associated with one of these entities. We use the term *coherent* to indicate the property that each datapoint is likely to be associated with the same entity as its predecessor or successor. Additionally each datapoint has a set of conflicting datapoints with which it cannot share the associated entity. Both datapoints and entities can be represented as vectors (or matrices), and datapoint vectors may have missing values. The task is to discover the *repeated entities*, which are associated with many datapoints, and simultaneously reject *outlier entities*, which are significantly different from the rest. It is of interest to solve the problem online (i.e. with a single pass over the data).

2 Experiments

2.1 Comparison with HMRF : Regarding baselines we compared against an alternative Bayesian Nonparametric Model (sHDP-HMM), a recent face clustering approach based on subspace clustering (WB-SLRR) and a constrained clustering approach. The state-of-the-art for face clustering and tracklet linking is HMRF [11]. Unfortunately, we find that this method [11] (authors' implementation) runs into severe numerical issues on these large videos, due to the matrix computations involved. In fact it fails to yield valid results when the number of clusters to be formed is more than 10. However for the sake of completeness we provide a comparson of TC-CRP and HMRF on the two videos (Frontal and BBTs1e1) on which tracklet clustering results have been reported in [11].

In [11] the clustering is evaluated by *clustering accuracy*. This requires a ground-truth clustering, against which the inferred clustering needs to be compared. Most of the videos on which we evaluated our method are large, and frame-level or tracklet-level ground truth labels are difficult to annotate. So in the main paper we judged the clusters themselves by means of purity, entity coverage and tracklet coverage. For the task of comparison with HMRF, we annotated the ground truth manually on both the videos. The comparisons are given in Table 1. It can be seen that HMRF performs slightly better on the shorter Frontal video, while TC-CRP is slightly better on the longer BBTs1e1 video.

2.2 Experiments on Videos with Missing Pixels User-generated videos are often noisy and grainy, as they are often shot directly from the television. The quality of the camera can also be an issue. Such videos may have random pixels grossly corrupted, i.e. effectively missing. We find that if more than 20% of the pixels are missing at random, the face detector itself often fails, and hence the person and tracklet discovery will not work. So we test the performance of our method with 20% pixels missing at random. As already discussed in Section 3.5, TC-CRP can recover missing entries in the tracklet vectors. As benchmark, we consider the tracklet matrix (formed by juxtaposing the tracklet vectors), and note that because of similarity of tracklets belonging to the same entity, the tracklet matrix must be *approximately low-rank*. In presence of the missing pixels, estimating this low-rank matrix is the well-know problem of *low-rank matrix completion*, for which we consider existing methods like SBMR [4] and OPTSPACE [1]. However, on our long videos SBMR is found to run out of memory, and OPTSPACE produces matrices with very low rank (5 or 6), which is clearly unrealistic as the number of entities are much larger. But TC-CRP's performance remains similar to those already reported in Tables 2,3,4.

*CSA Department, IISc
†EE Department, IISc
‡CSA Department, IISc

Dataset	Face-level		Track-level	
	HMRF	TCCRP	HMRF	TCCRP
Frontal	**0.907**	0.843	**0.905**	0.859
BBTs1e1	0.665	**0.693**	0.668	**0.698**

Table 1: Clustering Accuracy at Face-level and Tracklet-level

3 Connection to Low-rank Matrix Completion

In this section, we explore the scope of the proposed model beyond tracklet association.

3.1 Low-rank Matrices with sets of identical Columns Low-rank matrices are quite commonly used in computer vision. They have been used for both still images [12] and for videos [2]. In case of still images, each column generally corresponds to the face of a person. In case of videos, it corresponds to a frame (for background subtraction), a subwindow in a frame (for denoising) or detector outputs from a frame (in this work). In all cases, the low-rank matrix approximation is considered because several columns of the data matrix are near-identical. For example, in case of background subtraction all columns of the low-rank matrix should be identical, as they correspond to the static background. In TV series or movie videos, due to the property of Temporal Coherence, successive frames generally contain the same entities (for example, persons), and changes occur only at shot boundaries. Hence, if we represent a video with a matrix where each column corresponds to an entity detection (arranged in order of track/frame indices), we can expect to find a *low-rank approximation where succcessive columns to be identical except at the shot/track boundaries*. We investigate if the low-rank matrices recovered by various methods for matrix recovery (completion and extraction) actually do have successive columns identical. The existing methods mostly proceed by regularizing the nuclear norm, i.e. shrinking smaller singular values to 0. This reduces the rank and entry-wise error, but does not necessarily capture the structural property on the columns.

Synthetic Matrices We generate 50 basis vectors $\{\phi_k\}_{k=1}^{50}$ by sampling from the standard multivariate spherical Gaussian. Next, each column is generated by drawing from a basis vector from a multinomial distribution. In one version, all columns are drawn IID from this distribution **(no temporal coherence)**. In another version each column is drawn from a multinomial that emphasizes on the previous draw. In particular, if the column X_i corresponds to basis vector ϕ_k, then for column X_{i+1} we sample ϕ_k with probability 0.9, and any of the basis vectors uniformly with probability 0.002, and thus **temporal coherence exists**. These columns constitute the original matrix $X_{original}$. We study matrices of dimensions (200×1000), as in most applications the number of datapoints is far larger than the dimension. We study the sensitivity of the methods to the fraction of missing values. We try various levels of incompleteness, and vary the fractions of missing entries from 0.1 to 0.7. The matrices are corrupted by additive zero-mean noise with variance 0.1 independently on the observed entries.

Video Face Matrix Next, we consider a small matrix Y of face detections (reshaped to 900-dimensional vectors), taken from a user-uploaded Youtube video. Y has 1000 columns. Due to temporal coherence of videos, successive frames contain the same character, except at the shot change points. However, between the change points the face vectors are near-identical. A set of detections from this video are shown in Figure 1. The matrix Y has rank 900, because of small movements and variations in noise levels across the frames. However, noting that there are only 3 characters and 12 change-points, between which the vectors are almost identical, it is expected that a low-rank approximation X of Y should clearly have rank at most 12. Also, between these change-points, the columns of X should be *identical*.

Rank-column Plot: We consider the quantity $\tilde{X}_i = rank(X_{1:i})$- the rank of the submatrix formed by the first i columns of X. If X has identical columns between the change-points, this quantity should remain fixed between these changepoints, and may increase by 1 only at the changepoints. Hence the plot of $\tilde{X}_i$ versus i should be a *step-function*, as shown in Figure 1. We call this plot as the *Rank-column plot* of X. We study the *rank-column plot*(Figure 2,3) of the estimated low-rank matrix X returned by three recent methods for low-rank matrix approximation: Robust PCA [2], Bayesian Robust PCA [3] and Sparse Bayesian Robust PCA [4]. Surprisingly for all three methods, we observe: **1)** The rank-column plot for none of the methods comes close to the expected step function. All three show similar plots: the rank rises monotonically and then flattens out. **2)** For all three methods, the estimated "low-rank" matrix has rank much higher than the number of characters, and even the number of shot-changepoints. Moreover, *if the estimated matrix had rank r, then the submatrix formed by any set of m columns had rank equal to $min(r, m)$*. Such behavior of the rank-column plot clearly shows that the existing low-rank matrix recovery methods are completely incapable of capturing the temporal coherence of videos.

TC-CRP for Matrix Recovery A better idea is to use a *discrete distribution* on the columns, where each column vector is chosen from a set of vectors. This is very similar to the TC-CRP model proposed in this paper. It models temporal coherence through the change variable that ensures successive columns to

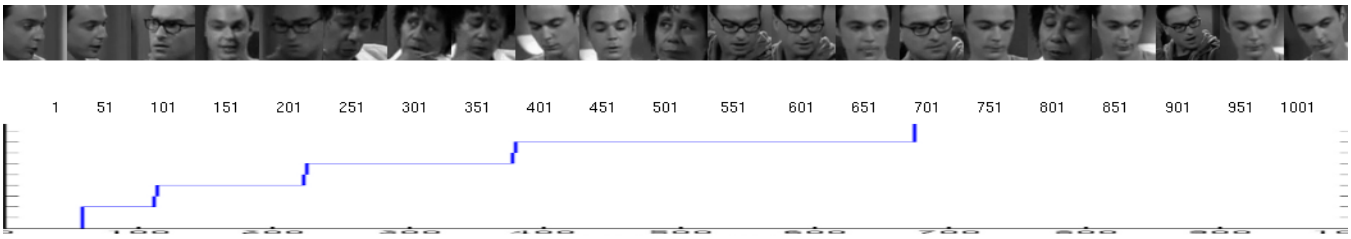

Figure 1: Face detections from the test video and the expected rank-column plot of its low-rank matrix representation: a step function which may increase at shot change-points. In case of the low-rank matrices learnt by RPCA, BRPCA and SBMR, this behavior is not observed.

be identical, but if not desired, this property can be abolished by setting the Bernoulli parameter κ to 1 (i.e. $C_i = 1 \forall i$). Note that at any column i the rank $\tilde{X}_i$ increases from $\tilde{X}_{i-1}$ if a new vector (different from $X_1, \ldots, X_{i-1}$) is sampled. The value α in the PPF (Equation 3.6) regulates the probability of sampling of a new vector from the base distribution, so a *smaller value of α ensures a lower rank*.

Evaluation: We evaluate TC-CRP's performance against the existing methods, for both the synthetic matrices (with and without TC) and the video face matrix. We measure the *Frobenius norm error*(FE) $\frac{||X_{recovered} - X_{original}||_F}{||X_{original}||_F}$, the *rank error*(RE) $\frac{|rank(X_{recovered} - X_{original})|}{|rank(X_{original})|}$. Also, as the original matrices have sets of identical columns, and the ones recovered by TC-CRP also have the same property, we compute the RAND index to evaluate the matching. As none of the existing low-rank recovery methods provide a matrix with identical columns, we compare TC-CRP's clustering against Spectral Clustering [7], which requires a *similarity matrix* between pairs of datapoints. We define pairwise similarity $S(i,j) = exp(-||X_i - X_j||_{\Omega_i \cap \Omega_j})$ (Ω_i: set of observed entries of X_i), and try out different values of K. The results for synthetic data are shown Tables VII and VIII. The rank-column plots are provided in Figure 2 for synthetic data and Figure 3 for faces. We see that on the synthetic data, not only does TC-CRP provide the perfect rank-column plots (which coincide with the true plots), but even in terms of Frobenius norm error, Rank error and RAND index, its performance is way ahead of the existing methods. For the face data also, its rank-column plot is roughly accurate, and increments around the shot changepoints.

3.2 Subspace Clustering

A problem related to low-rank matrix completion is Subspace Clustering, where each column vector of the data matrix Y is considered to lie in an *Union of Subspaces*. The data matrix is expressed as $Y = YC + B$, where C is the coefficient matrix where the column-vector C_i indicates the subspace membership of column Y_i (the i-th datapoint). This representation has also been used in Computer Vision, most

missing	TCCRP		SVT		OPTSPACE		SBMR	
fraction	FE	RE	FE	RE	FE	RE	FE	RE
0.1	**0.002**	**0**	**0.031**	0.055	0.138	0.98	0.038	0.24
0.3	**0.008**	**0**	0.040	0.03	0.138	0.98	0.049	0.66
0.5	**0.040**	**0.02**	0.048	0.09	0.137	0.98	0.068	0.71
0.7	**0.059**	**0.02**	0.116	0.40	0.136	0.98	0.103	0.70

Table 2: Comparison of Low-rank Matrix Completion techniques with varying fractions of missing entries, in absence of TC. FE: Frobenius Norm Error, RE: Rank Error, RAND: Rand index for clustering

missing	TCCRP		SVT		OPTSPACE		SBMR	
fraction	FE	RE	FE	RE	FE	RE	FE	RE
0.1	0.008	**0**	0.03	0.14	0.169	0.97	**0.007**	0.15
0.3	**0.013**	**0.01**	0.03	0.14	0.169	0.98	0.037	0.60
0.5	**0.035**	**0.04**	0.038	0.09	0.178	0.98	0.056	0.71
0.7	**0.048**	**0.05**	0.105	0.41	0.178	0.98	0.095	0.83

Table 3: Comparison of Low-rank Matrix Completion techniques with varying fractions of missing entries, in presence of TC. FE: Frobenius Norm Error, RE: Rank Error, RAND: Rand index for clustering

missing	without TC		with TC	
fraction	TCCRP	NCUT	TCCRP	NCUT
0.1	**1.0000**	0.998	**1.0000**	0.998
0.3	**1.0000**	0.988	**1.0000**	0.994
0.5	**0.9994**	0.985	**0.9999**	0.987
0.7	**0.9990**	0.980	**0.9996**	0.976

Table 4: RAND index for clustering columns for varying fractions of missing entries, in presence and absence of TC

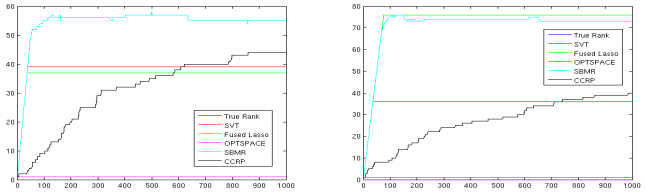

Figure 2: Rank-column plots for various methods. Left figure is for a matrix with 10% missing entries, and right figure for 50% missing entries. The Blue Line (True Plot) and the Black Line (proposed method) coincide

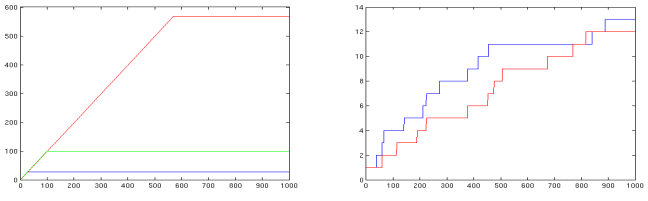

Figure 3: Left: Rank-column plots for SBMR(blue), RPCA(red) and BRPCA(green) for the test video. The estimated matrices all have rank much more than the number of shot segments (12), and do not exhibit the expected step function-like behavior. Right: Rank-column plot for TC-CRP(blue), and the shot number(red) which increments at the shot changepoints, and the matrix rank is 13

notably for motion segmentation. Once again, as several datapoints are almost identical, their corresponding coefficient vectors are also expected to be similar, and hence the coefficient matrix C should again have sets of identical columns. No wonder, C has been modelled as low-rank in the LRR formulation [6]. Also, in case of sequential data like videos, the successive datapoints are very similar and likely to have same subspace coefficients, which is handled in [8] by an additional penalty term. However, these methods also model the rank of C using the nuclear norm, and as we have seen already, this cannot recover a coefficient matrix with sets of identical columns, as needed for clustering. Hence, they perform the clustering as a separate step, using *spectral clustering with an affinity matrix constructed from* C. However, spectral clustering is very slow.

A better option can be to use TC-CRP on the recovered C to obtain its approximation having sets of identical columns. Sequential data can easily be taken care of, because TC-CRP models temporal coherence. In case of non-sequential data also, TC-CRP (with $\kappa = 1$) can produce sets of identical columns because it generates the columns from a discrete distribution. Note that this does not obviate the need to estimate the coefficient matrix C. This is because, TC-CRP in its current form models each datapoint Z_i as a draw from a single Gaussian, and not as the linear combination of several such draws, which is required for this purpose.

References

[1] Keshavan,R., Montanari,A. & Oh,S. (2010) *Matrix Completion from a few entries*, IEEE Transactions on Information Theory

[2] Wright,J., Ssatry,S., Peng,Y., Ganesh,A. & Ma, Y. (2009) *Robust Principal Component Analysis*, Advances in Neural Information Processing Systems (NIPS)

[3] Ding, X., He, L. & Carin, L. *Bayesian Robust Principal Component Analysis*, IEEE Transactions on Image Processing

[4] Babacan,S.D., Luessi,M., Molina,R. & Katsaggelos,A.K. (2012) *Sparse Bayesian Methods for Low-Rank Matrix Estimation*, IEEE Transactions on Signal Processing

[5] Elhamifar,E. & Vidal,R. (2009) *Sparse Subspace Clustering*, IEEE Intl. Conf on Computer Vision and Pattern Recognition (CVPR)

[6] Liu,G., Lin,Z., Yan,S., Sun,J, Yu,Y. & Ma,Y. (2013) *Robust recovery of subspace structures by low-rank representation*, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol 35(1)

[7] Shi,J. & Malik,J. (2000) *Normalized cuts and image segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol 22(8)

[8] Tierney,S., Gao,J. & Yi Guo,Y. (2014) *Subspace Clustering for Sequential Data*, IEEE Intl. Conf on Computer Vision and Pattern Recognition (CVPR)

[9] Kawale,J. & Boley,D. (2013) *Constrained Spectral Clustering using L1 Regularization*, SIAM Conference on Data Mining (SDM)

[10] Xiao1,S., Tan,M., & Xu,D. (2014) *Weighted Block-Sparse Low Rank Representation for Face Clustering in Videos*, European Conference on Computer Vision (ECCV)

[11] Wu,B., Lyu,S., Hu,B-G & Ji,Q. (2013) *Simultaneous Clustering and Tracklet Linking for Multi-Face Tracking in Videos*, IEEE Intl. Conf on Computer Vision (ICCV)

[12] Peng,Y., Ganesh,A., Wright,J., Xu,W. & Ma,Y. (2010) *RASL: Robust Batch Alignment of Images by Sparse and Low-Rank Decomposition*, IEEE Intl Conf Computer Vision and Pattern Recognition (CVPR)