Subset Selection and Summarization in Sequential Data

Ehsan Elhamifar
Computer and Information Science College
Northeastern University
Boston, MA 02115
eelhami@ccs.neu.edu

M. Clara De Paolis Kaluza
Computer and Information Science College
Northeastern University
Boston, MA 02115
clara@ccs.neu.edu

Abstract

Subset selection, which is the task of finding a small subset of representative items from a large ground set, finds numerous applications in different areas. Sequential data, including time-series and ordered data, contain important structural relationships among items, imposed by underlying dynamic models of data, that should play a vital role in the selection of representatives. However, nearly all existing subset selection techniques ignore underlying dynamics of data and treat items independently, leading to incompatible sets of representatives. In this paper, we develop a new framework for sequential subset selection that finds a set of representatives compatible with the dynamic models of data. To do so, we equip items with transition dynamic models and pose the problem as an integer binary optimization over assignments of sequential items to representatives, that leads to high encoding, diversity and transition potentials. Our formulation generalizes the well-known facility location objective to deal with sequential data, incorporating transition dynamics among facilities. As the proposed formulation is non-convex, we derive a max-sum message passing algorithm to solve the problem efficiently. Experiments on synthetic and real data, including instructional video summarization, show that our sequential subset selection framework not only achieves better encoding and diversity than the state of the art, but also successfully incorporates dynamics of data, leading to compatible representatives.

1 Introduction

Subset selection is the task of finding a small subset of most informative items from a ground set. Besides helping to reduce the computational time and memory of algorithms, due to working on a much smaller representative set [1], it has found numerous applications, including, image and video summarization [2, 3, 4], speech and document summarization [5, 6, 7], clustering [8, 9, 10, 11, 12], feature and model selection [13, 14, 15, 16], sensor placement [17, 18], social network marketing [19] and product recommendation [20]. Compared to dictionary learning methods such as Kmeans [21], KSVD [22] and HMMs [23], that *learn* centers/atoms in the input-space, subset selection methods *choose* centers/atoms from the given set of items.

Sequential data, including time-series such as video, speech, audio and sensor measurements as well as ordered data such as text, form an important large part of modern datasets, requiring effective subset selection techniques. Such datasets contain important structural relationships among items, often imposed by underlying dynamic models, that should play a vital role in the selection of representatives. For example, there exists a logical way in which segments of a video or sentences of a document are connected together and treating segments/sentences as a bag of randomly permutable items results in losing the semantic content of the video/document. However, existing subset selection methods
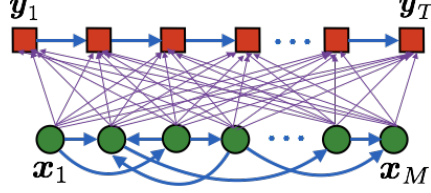
Figure 1: We propose a framework for the summarization of sequential data. Given a source set of items $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$, with a dynamic transition model, and a target set of sequential items $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$, we find a sequence of representatives from the source set with high transition probability that well encode the target set.

ignore these relationships and treat items independent from each other. Thus, there is a need for sequential subset selection methods that, instead of treating items independently, use the underlying dynamic models of data to select high-quality, diverse and compatible representatives.

Prior Work: A subset selection framework consists of three main components: i) the inputs to the algorithm; ii) the objective function to optimize, characterizing the informativeness and diversity of selected items; iii) the algorithm to optimize the objective function. The inputs to subset selection algorithms are in the form of either feature vector representations or pairwise similarities between items. Several subset selection criteria have been studied in the literature, including maximum cut objective [24, 25], maximum marginal relevance [26], capacitated and uncapacitated facility location objectives [27, 28], multi-linear coding [29, 30] and maximum volume subset [6, 31], which all try to characterize the informativeness/value of a subset of items in terms of ability to represent the entire distribution and/or having minimum information overlap among selected items. On the other hand, optimizing almost all subset selection criteria is, in general, NP-hard and non-convex [25, 32, 33, 34], which has motivated the development and study of approximate methods for optimizing these criteria. This includes greedy approximate algorithms [28] for maximizing submodular functions, such as graph-cuts and facility location, which have worst-case approximation guarantees, as well as sampling methods from Determinantal Point Process (DPP) [6, 31], a probability measure on the set of all subsets of a ground set, for approximately finding the maximum volume subset. Motivated by the maturity of convex optimization and advances in sparse and low-rank recovery, recent methods have focused on convex relaxation-based methods for subset selection [8, 9, 2, 35, 36].

When it comes to sequential data, however, the majority of subset selection methods ignore the underlying dynamics and relationships among items and treat items independent from each other. Recent results in [37, 3] have developed interesting extensions to DPP-based subset selection, by capturing representatives in a sequential order such that newly selected representatives are diverse with respect to the previously selected ones. However, sequential diversity by itself is generally insufficient, especially, when the sequence of diverse selected items are unlikely to follow each other according to underlying dynamic models. For example, in a video/document on a specific topic with intermediate irrelevant scenes/sentences to the topic, promoting sequential diversity results in selecting irrelevant scenes/sentences. [38] extends submodular functions to capture ordered preferences among items, where ordered preferences are represented by a directed acyclic graph over items, and presents a greedy algorithm to pick edges instead of items. The method, however, cannot deal with arbitrary graphs, such as Markov chains with cycles. On the other hand, while Hidden Markov Models (HMMs) [23, 39] and dynamical systems [40, 41] have been extensively studied for modeling sequential data, they have not been properly exploited in the context of subset selection.

Paper Contributions: In this paper we develop a new framework for sequential subset selection that incorporates the dynamic model of sequential data into subset selection. We develop a new class of objective functions that promotes to select not only high-quality and diverse items, but also a sequence of representatives that are compatible with the dynamic model of data. To do so, we propose a dynamic subset selection framework, where we equip items with transition probabilities and design objective functions to select representatives that well capture the data distribution with a high overall transition probability in the sequence of representatives, see Figure 1. Our formulation generalizes the facility location objective [27, 28] to sequential data, by incorporating transition dynamics among facilities. Since our proposed integer binary optimization is non-convex, we develop a max-sum message passing framework to solve the problem efficiently. By experiments on synthetic and real data, including instructional video summarization, we show that our method outperforms the state of the art in terms of selecting representatives with better encoding, diversity and dynamic compatibility.

2 Subset Selection for Sequential Data

Sequential data, including time-series and ordered data contain important structural relationships among items, often imposed by underlying dynamic models of data, that should play a vital role in the selection of representatives. In this section, we develop a new framework for sequential subset selection that incorporates underlying dynamic models and relationships among items into subset selection. More specifically, we propose a dynamic subset selection framework, where we equip items with transition probabilities and design objectives to select representatives that capture the data distribution with a high transition probability in the sequence of representatives. In the next section, we develop an efficient algorithm to solve the proposed optimization problem.

2.1 Sequential Subset Selection Formulation

Assume we have a source set of items $\mathbb{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$, equipped with a transition model, $p(\boldsymbol{x}_{i'}|\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_n})$, between items, and a target set of sequential items $\mathbb{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$. Our goal is to find a small representative subset of $\mathbb{X}$ that well encode $\mathbb{Y}$, while the set of representatives are compatible according to the dynamic model of $\mathbb{X}$. Let $\boldsymbol{x}_{r_t}$ be the representative of $\boldsymbol{y}_t$ for $t \in \{1, \ldots, T\}$. We propose a potential function $\Psi(r_1, \ldots, r_T)$ whose maximization over all possible assignments $(r_1, \ldots, r_T) \subseteq \{1, \ldots, M\}^T$, i.e.,

$$\max_{(r_1, \ldots, r_T) \subseteq \{1, \ldots, M\}^T} \Psi(r_1, \ldots, r_T) \tag{1}$$

achieves the three goals of i) minimizing the encoding cost of $\mathbb{Y}$ via the representative set; ii) selecting a small set of representatives from $\mathbb{X}$; iii) selecting an ordered set of representatives $(\boldsymbol{x}_{r_1}, \ldots, \boldsymbol{x}_{r_T})$ that are compatible with the dynamics on $\mathbb{X}$.

To tackle the problem, we consider a decomposition of the potential function Ψ into the product of three potentials, corresponding to the three aforementioned objectives, as

$$\Psi(r_1, \ldots, r_T) \triangleq \Phi_{\mathrm{enc}}(r_1, \ldots, r_T) \times \Phi_{\mathrm{card}}(r_1, \ldots, r_T) \times \Phi_{\mathrm{dyn}}(r_1, \ldots, r_T), \tag{2}$$

where $\Phi_{\mathrm{enc}}(r_1, \ldots, r_T)$ denotes the encoding potential that favors selecting a representative set from $\mathbb{X}$ that well encodes $\mathbb{Y}$, $\Phi_{\mathrm{card}}(r_1, \ldots, r_T)$ denotes the cardinality potential that favors selecting a small number of distinct representatives. Finally, $\Phi_{\mathrm{dyn}}(r_1, \ldots, r_T)$ denotes the dynamic potential that favors selecting an ordered set of representatives that are likely to be generated by the underlying dynamic model on $\mathbb{X}$. Next, we study each of the three potentials.

Encoding Potential: Since the encoding of each item of $\mathbb{Y}$ depends on its own representative, we assume that the encoding potential function factorizes as

$$\Phi_{\mathrm{enc}}(r_1, \ldots, r_T) = \prod_{t=1}^{T} \phi_{\mathrm{enc},t}(r_t), \tag{3}$$

where $\phi_{\mathrm{enc},t}(i)$ characterizes how well $\boldsymbol{x}_i$ encodes $\boldsymbol{y}_t$ and becomes larger when $\boldsymbol{x}_i$ better represents $\boldsymbol{y}_t$. In this paper, we assume that $\phi_{\mathrm{enc},t}(i) = \exp(-d_{i,t})$, where $d_{i,t}$ indicates the dissimilarity of $\boldsymbol{x}_i$ to $\boldsymbol{y}_t$. A lower dissimilarity $d_{i,t}$ means that $\boldsymbol{x}_i$ better encodes/represents $\boldsymbol{y}_t$.

Cardinality Potential: Notice that maximizing the encoding potential alone results in selecting many representatives. Hence, we consider a cardinality potential to restrict the total number of representatives. Denoting the number of representatives by $|\{r_1, \ldots, r_T\}|$, we consider

$$\Phi_{\mathrm{card}}(r_1, \ldots, r_T) = \exp(-\lambda|\{r_1, \ldots, r_T\}|), \tag{4}$$

which promotes to select a small number of representatives. The parameter $\lambda > 0$ controls the effect of the cardinality on the global potential Ψ, where a close to zero λ ignores the effect of cardinality potential, resulting in many representatives, and a larger λ results in a smaller representative set.

Dynamic Potential: While encoding and cardinality potentials together promote selecting a few representatives from $\mathbb{X}$ that well encode $\mathbb{Y}$, there is no guarantee that the sequence of representatives $(\boldsymbol{x}_{r_1}, \ldots, \boldsymbol{x}_{r_T})$ is compatible with the underlying dynamic of $\mathbb{X}$. Thus, we introduce a dynamic potential that measures the compatibility of the sequence of representatives. To do so, we consider

3

an n-th order Markov Model to represent the dynamic relationships among the items in $\mathbb{X}$, where the selection of the representative $\boldsymbol{x}_{r_t}$ depends on the m previously selected representatives, i.e., $\boldsymbol{x}_{r_{t-1}}, \ldots, \boldsymbol{x}_{r_{t-n}}$. More precisely, we consider

$$\Phi_{\text{dyn}}(r_1, \ldots, r_T) = \left(\prod_{t=1}^{n} p_t(\boldsymbol{x}_{r_t}) \times \prod_{t=n+1}^{T} p_t(\boldsymbol{x}_{r_t}|\boldsymbol{x}_{r_{t-1}}, \ldots, \boldsymbol{x}_{r_{t-n}}) \right)^{\beta}, \quad (5)$$

where $p_t(\boldsymbol{x}_i)$ indicates the probability of selecting $\boldsymbol{x}_i$ as the representative of $\boldsymbol{y}_t$ and $p_t(\boldsymbol{x}_{i'}|\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_n})$ denotes the probability of selecting $\boldsymbol{x}_{i'}$ as the representative of $\boldsymbol{y}_t$ given that $\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_n}$ has been selected as the representative of $\boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-n}$, respectively. The regularization parameter $\beta > 0$ determines the effect of the dynamic potential on the overall potential Ψ, where a close to zero β results in discounting the effect of the dynamic of $\mathbb{X}$. As a result, maximizing the dynamic potential promotes to select a sequence of representatives that are highly likely to follow the dynamic model on the source set.

2.2 Optimization Framework for Sequential Subset Selection

In the rest of the paper, we consider a first order Markov model, which performs well in the application studied in the paper (our proposed optimization can be generalized to n-th order Markov models as well). Putting all three potentials together, we consider maximization of the global potential function

$$\Psi = \prod_{t=1}^{T} \phi_{\text{enc},t}(r_t) \times \Phi_{\text{card}}(r_1, \ldots, r_T) \times \left(p_1(\boldsymbol{x}_{r_1}) \times \prod_{t=2}^{T} p_t(\boldsymbol{x}_{r_t}|\boldsymbol{x}_{r_{t-1}}) \right)^{\beta}. \quad (6)$$

over all possible assignments $(r_1, \ldots, r_T) \subseteq \{1, \ldots, M\}^T$. To do so, we cast the problem as an integer binary optimization. We define binary assignment variables $\{z_{i,t}\}_{i=1,\ldots,M}^{t=1,\ldots,T}$, where $z_{i,t} \in \{0, 1\}$ indicates if $\boldsymbol{x}_i$ is a representative of $\boldsymbol{y}_t$. Since each item $\boldsymbol{y}_t$ is associated with only a single representative, we have $\sum_{i=1}^{M} z_{i,t} = 1$. Also, we define variables $\{\delta_i\}_{i=1,\ldots,M}$ and $\{u_{i',i}^t\}_{i,i'=1,\ldots,M}^{t=1,\ldots,T}$, where $\delta_i \in \{0, 1\}$ indicates if $\boldsymbol{x}_i$ is a representative of $\boldsymbol{y}_1$ and $u_{i',i}^t \in \{0, 1\}$ indicates if $\boldsymbol{x}_{i'}$ is a representative of $\boldsymbol{y}_t$ given that $\boldsymbol{x}_i$ is a representative of $\boldsymbol{y}_{t-1}$. As we will show, $\{\delta_i\}$ and $\{u_{i',i}^t\}$ are related to $\{z_{i,t}\}$, hence, the final optimization only depends on $\{z_{i,t}\}$.

Using the variables defined above, we can rewrite the global potential function in (6) as

$$\Psi = \prod_{t=1}^{T} \prod_{i=1}^{M} \phi_{\text{enc},t}(i)^{z_{i,t}} \times \Phi_{\text{card}}(r_1, \ldots, r_T) \times \prod_{i=1}^{M} p_1(\boldsymbol{x}_i)^{\beta \delta_i} \times \prod_{t=2}^{T} \prod_{i'=1}^{M} \prod_{i=1}^{M} p_t(\boldsymbol{x}_{i'}|\boldsymbol{x}_i)^{\beta u_{i',i}^t}. \quad (7)$$

We can equivalently maximize the logarithm of Ψ, which is to maximize

$$\sum_{t=1}^{T} \sum_{i=1}^{M} -z_{i,t} d_{i,t} + \log \Phi_{\text{card}}(r_1, \ldots, r_T) + \sum_{i=1}^{M} \delta_i \log p_1(\boldsymbol{x}_i) + \sum_{t=2}^{T} \sum_{i,i'=1}^{M} u_{i',i} \log p_t(\boldsymbol{x}_{i'}|\boldsymbol{x}_i), \quad (8)$$

where we used $\log \phi_{\text{enc},t}(i) = -d_{i,t}$. Notice that $\{\delta_i\}$ and $\{u_{i',i}^t\}$ can be written as functions of the assignment variables $\{z_{i,t}\}$. Denoting the indicator function by $1_{(\cdot)}$, which is one when its argument is true and is zero otherwise, we can write $\delta_i = 1_{(r_1 = i)}$ and $u_{i',i}^t = 1_{(r_t = i', r_{t-1} = i)}$. Hence, we have

$$\delta_i = z_{i,1}, \quad u_{i',i}^t = z_{i,t-1} z_{i',t}. \quad (9)$$

As a result, we can rewrite the maximization in (8) as the equivalent optimization

$$\begin{aligned}
\max_{\{z_{i,t}\}} \quad &\sum_{t=1}^{T} \sum_{i=1}^{M} -z_{i,t} d_{i,t} + \log \Phi_{\text{card}}(r_1, \ldots, r_T) + \beta \Big(\sum_{i=1}^{M} z_{i,1} \log p_1(\boldsymbol{x}_i) \\
&+ \sum_{t=2}^{T} \sum_{i,i'=1}^{M} z_{i,t-1} z_{i',t} \log p_t(\boldsymbol{x}_{i'}|\boldsymbol{x}_i) \Big) \quad \text{s.t.} \quad z_{i,t} \in \{0,1\}, \quad \sum_{i=1}^{M} z_{i,t} = 1, \, \forall \, i, t.
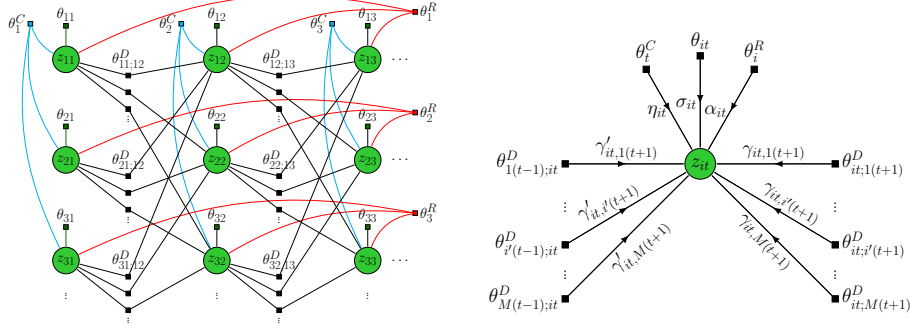\end{aligned} \quad (10)$$

Figure 2: Left: Factor graph representing (12). Right: Messages from each factor to a variable node $z_{i,t}$.

It is important to note that if $\boldsymbol{x}_i$ becomes a representative of some items in $\mathbb{Y}$, then $\| [z_{i,1} \; \cdots \; z_{i,T}] \|_\infty$ would be 1. Hence, the number of representatives is given by $\sum_{i=1}^M \| [z_{i,1} \; \cdots \; z_{i,T}] \|_\infty$. As a result, we can rewrite the cardinality potential in (4) as

$$\Phi_{\text{card}}(r_1, \ldots, r_T) = \exp(-\lambda \sum_{i=1}^M \| [z_{i,1} \; \cdots \; z_{i,T}] \|_\infty). \tag{11}$$

Finally, considering a homogeneous Markov Model on the dynamics of the source set, where $p_t(\cdot|\cdot) = p(\cdot|\cdot)$, i.e., transitioning from $\boldsymbol{x}_i$ as the representative of $\boldsymbol{y}_{t-1}$ to $\boldsymbol{x}_{i'}$ as the representative of $\boldsymbol{y}_t$ does not depend on t, we propose to solve the optimization

$$\max_{\{z_{i,t}\}} \; \sum_{t=1}^T \sum_{i=1}^M -z_{i,t} d_{i,t} - \lambda \sum_{i=1}^M \| [z_{i,1} \; \cdots \; z_{i,T}] \|_\infty + \beta (\sum_{i=1}^M z_{i,1} \log p_1(\boldsymbol{x}_i)$$
$$+ \sum_{t=2}^T \sum_{i,i'=1}^M z_{i,t-1} z_{i',t} \log p(\boldsymbol{x}_{i'}|\boldsymbol{x}_i)) \quad \text{s. t.} \quad z_{i,t} \in \{0,1\}, \;\; \sum_{i=1}^M z_{i,t} = 1, \; \forall \, i, t. \tag{12}$$

In our proposed formulation above, we assume that the dissimilarities $\{d_{i,t}\}$ and the dynamic models, i.e., the probabilities $p_1(\cdot)$ and $p(\cdot|\cdot)$, are known. These models can be given by prior knowledge or by learning from training data, as we show in the experiments. It is important to notice that the optimization in (12) is non-convex, due to binary optimization variables and quadratic terms in the objective function, which is not necessarily positive semi-definite (this can be easily seen when $p(x_{i'}|x_i) \neq p(x_i|x_{i'})$ for some i, i'). In the next section, we treat (12) as a MAP inference on binary random variables and develop a message passing algorithm to find the hidden values $\{z_{i,t}\}$.

Once we solve the optimization in (12), we can obtain the representatives as the items of $\mathbb{X}$ for which $z_{i,t}$ is non-zero for some t. Moreover, we can obtain the segmentation of the sequential items in $\mathbb{Y}$ according to their assignments to the representatives. In fact, the sequence of representatives obtained by our proposed optimization in (12) not only corresponds to diverse items that well encode the sequential target data, but also is compatible with the underlying dynamic of the source data.

Remark 1 *Without the dynamic potential, ie, with $\beta = 0$, our proposed optimization in* (12) *reduces to the uncapacitated facility location objective. Hence, our framework generalizes the facility location to sequential data by considering transition dynamics among facilities (source set items).*

3 Message Passing for Sequential Subset Selection

In this section, we develop an efficient message passing algorithm to solve the proposed optimization in (12). To do so, we treat the sequential subset selection as a MAP inference, where $\{z_{i,t}\}$ correspond to binary random variables whose joint log-likelihood is given by the objective function in (12). We represent the log-likelihood, i.e., the objective function in (12), with a factor graph [42], which is shown in Figure 2. Recall that a factor graph is a bipartite graph that consists of variable nodes and factor nodes, where every factor evaluates a potential function over variables it is connected to. The log-likelihood is then proportional to the sum of all factor potentials.

To form the factors corresponding to the objective function in (12), we define $m_{i,i'} \triangleq \log p(\boldsymbol{x}_{i'}|\boldsymbol{x}_i)$ and $\bar{d}_{i,t} \triangleq d_{i,t} - \log p_1(\boldsymbol{x}_i)$ if $t = 1$ and $\bar{d}_{i,t} \triangleq d_{i,t}$ for all other values of t. Denoting $\boldsymbol{z}_{i,:} \triangleq [z_{i,1} \cdots z_{i,T}]^\top$ and $\boldsymbol{z}_{:,t} \triangleq [z_{1,t} \cdots z_{M,t}]^\top$, we define factor potentials corresponding to our framework, shown in Figure 2. More specifically, we define the encoding and dynamic potentials, respectively, as $\theta_{i,t}(z_{i,t}) \triangleq -\bar{d}_{i,t} z_{i,t}$ and $\theta^D_{i,t-1;i',t}(z_{i,t-1}, z_{i',t}) \triangleq m_{i,i'} z_{i,t-1} z_{i',t}$. Moreover we define the cardinality and constraint potentials, respectively, as

$$\theta^R_i(\boldsymbol{z}_{i,:}) \triangleq \begin{cases} -\lambda, & \|\boldsymbol{z}_{i,:}\|_\infty > 0 \\ 0, & \text{otherwise} \end{cases}, \quad \theta^C_t(\boldsymbol{z}_{:,t}) \triangleq \begin{cases} 0, & \sum_{i=1}^M z_{i,t} = 1 \\ -\infty, & \text{otherwise} \end{cases}.$$

The MAP formulation of our sequential subset selection is then given by

$$\max_{\{z_{i,t}\}} \sum_{t=1}^T \sum_{i=1}^M \theta_{i,t}(z_{i,t}) + \sum_{i=1}^M \theta^R_i(\boldsymbol{z}_{i,:}) + \sum_{t=1}^T \theta^C_l(\boldsymbol{z}_{:,t}) + \beta \sum_{t=1}^{T-1} \sum_{i'=1}^M \sum_{i=1}^M \theta^D_{i,t-1;i',t}(z_{i,t-1}, z_{i',t}). \quad (13)$$

To perform MAP inference, we use the max-sum message passing algorithm, which iteratively updates messages between variable and factor nodes in the graph. In our framework, the incoming messages to each variable node $z_{i,t}$ are illustrated in Figure 2. Messages are computed as follows (please see the supplementary materials for the derivations).

$$\sigma_{i,t} \leftarrow -\bar{d}_{i,t} \tag{14}$$

$$\gamma_{i,t;j,t+1} \leftarrow \max\{0, m_{i,j} + \rho\} - \max\{0, \rho\} \tag{15}$$

$$\gamma'_{i,t-1;j,t} \leftarrow \max\{0, m_{i,j} + \rho'\} - \max\{0, \rho'\} \tag{16}$$

$$\eta_{i,t} \leftarrow -\max_{k \neq i', i} \{\alpha_{i',1} - \bar{d}_{i',1} + \sum_{j=1}^M \gamma_{i',t;j,t+1} + \sum_{j=1}^M \gamma'_{j,t-1;i',t}\} \tag{17}$$

$$\alpha_{i,t} \leftarrow \min\{0, -\lambda + \sum_{k \neq t} \max\{0, -\bar{d}_{i,k} + \eta_{i,k} + \sum_{j=1}^M (\gamma_{i,k;j,k+1} + \gamma'_{j,k-1;i,k})\}\} \tag{18}$$

where, for brevity of notation, we have defined ρ and ρ' as

$$\rho \overset{\triangle}{=} -\bar{d}_{j,t+1} + \alpha_{j,t+1} + \eta_{j,t+1} + \sum_{k=1}^M \gamma_{j,t+1;k,t+2} + \sum_{k \neq i} \gamma'_{k,t;j,t+1}, \tag{19}$$

$$\rho' \overset{\triangle}{=} -\bar{d}_{i,t-1} + \alpha_{i,t-1} + \eta_{i,t-1} + \sum_{k \neq i} \gamma_{i,t-1;k,t} + \sum_{k=1}^M \gamma'_{k,t-2;i,t-1}. \tag{20}$$

The update of messages continues until convergence, when each variable $z_{i,t}$ is assigned to the value that maximizes the sum of its incoming messages. It is important to note that the max-sum algorithm always converges to the optimal MAP assignment on trees, and has shown good performance on graphs with cycles in many applications, including our work. We also use a dampening factor $\lambda \in [0, 1)$ on message updates as so that a message μ is computed as $\mu^{(\text{new})} \leftarrow \lambda \mu^{(\text{old})} + (1 - \lambda) \mu^{(\text{update})}$.

4   Experiments

In this section, we evaluate the performance of our proposed method as well as the state of the art for subset selection on synthetic and real sequential data. For real applications, we consider the task of summarizing instructional videos to learn the key steps of the task described in the videos. In addition to our proposed message passing (MP) algorithm, we have implemented the optimization in (12) using an ADMM framework [43], where we have relaxed the integer binary constraints to $z_{i,t} \in [0, 1]$. In practice both MP and ADMM algorithms achieve similar results.

We compare our proposed method, Sequential Facility Location (SeqFL), with several subset selection algorithms. Since we study the performance of methods as a function of the size of the representative set, we use the fixed-size variant of DPP, called kDPP [44]. In addition to kDPP, we evaluate the performance of Markov kDPP (M-kDPP) [37], in which successive representatives are diverse among themselves and with respect to the previously selected representatives, as well as Sequential kDPP
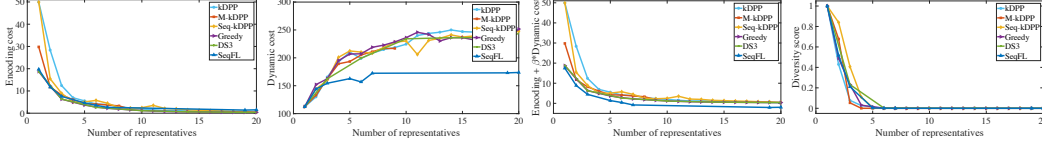
Figure 3: Encoding cost, dynamic cost, total cost and diversity score of different algorithms as a function of the number of selected representatives. The size of the source set is $M = 50$.
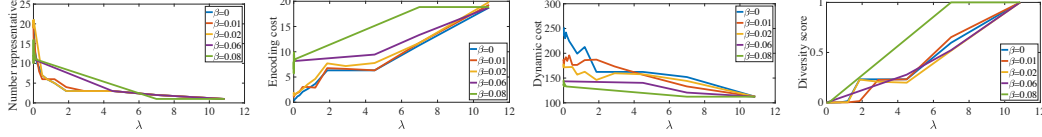

Figure 4: Number of representatives, encoding cost, dynamic cost and diversity score of our proposed method (SeqFL) as a function of the parameters (β, λ).

(Seq-kDPP) [3], which divides a time-series into multiple windows and successively selects diverse samples from each window conditioned on the previous window.[1] We also compare our method against DS3 [8] and the standard greedy method [28], which optimize the conventional facility location, which has no dynamic cost, via convex relaxation and greedy selection, respectively.

To compare the performance of different methods, we evaluate several costs and scores that demonstrate the effectiveness of each method in terms of encoding, diversity and dynamic compatibility of the set of selected representatives. More specifically, given dissimilarities $\{d_{i,t}\}$, the dynamic model $p_1(\cdot)$ and $p(\cdot|\cdot)$, representative set Λ, and the assignment of points to representatives $\{z^*_{i,t}\}$, we compute the encoding cost as $\sum_{t=1}^{T} \sum_{i=1}^{M} d_{i,t} z^*_{i,t}$, the dynamic cost as $-\sum_{i=1}^{M} \log p_1(\boldsymbol{x}_i) z^*_{i,1} - \sum_{t=2}^{T} \sum_{i,i'=1}^{M} \log p(\boldsymbol{x}_{i'}|\boldsymbol{x}_i) z^*_{i,t-1} z^*_{i',t}$ and the diversity score as $\det(\boldsymbol{K}_\Lambda)$. Here, $\boldsymbol{K}$ corresponds to the kernel matrix, used in DPP and its variants, and $\boldsymbol{K}_\Lambda$ denotes the submatrix of $\boldsymbol{K}$ indexed by Λ. In this paper, we use Euclidean distances as dissimilarities and compute the corresponding inner-product kernel to run DPPs. Notice that the diversity score, which is the volume of the parallelotope spanned by the representatives, is what DPP methods aim to (approximately) maximize. As DPP methods only find representatives and not assignment of points, we compute $z^*_{i,t}$'s by assigning each point to the closest representative in Λ, according to the kernel.

4.1 Synthetic Data

To demonstrate the effectiveness of our proposed method for sequential subset selection, we generate synthetic data where, for a source set $\mathbb{X}$ with M items corresponding to means of M Gaussians, we generate a transition probability matrix among items and an initial probability vector. We draw a sequence of length T from the corresponding Markov model to form the target set $\mathbb{Y}$ and run different algorithms to generate k representatives. We then compute the average encoding and transition costs as well as the diversity scores for sequences drawn from the Markov model, as a function of $k \in \{1, 2, \ldots, M\}$. In the experiments we set $M = 50$, $T = 100$. For a fixed β, we run SeqFL for different values of λ to select different number of representatives.

Figure 3 illustrates encoding and transition costs and the diversity scores of different methods, where for SeqFL we have set $\beta = 0.02$. Notice that our proposed method consistently obtains lower encoding, dynamic and total costs for all numbers of representatives, demonstrating its effectiveness for obtaining a sequence of high-quality and compatible representatives according to the dynamics. It is important to notice that although our method does not maximize the diversity score, used in kDPP and its variants, it achieves slightly better diversity scores (higher is better) than kDPP and M-kDPP. Figure 4 demonstrates the effect of the parameters (β, λ) on the solution of our proposed method. Notice that for a fixed β, as λ increases, we select a smaller number of representatives, hence, the encoding cost increases. Also, for a fixed λ, as β increases, we put more more emphasis on dynamic compatibility of representatives, hence, the dynamic cost decreases. On the other hand, the diversity score decreases for smaller λ, as we select more representatives which become more redundant. The results in Figure 4 also demonstrate the robustness of our method to the change of parameters.

---

[1]To have a fair comparison and to select a fixed number of representatives, we modify the SeqDPP method [3] and implement Seq-kDPP where k representatives are chosen in each window.

Task		kDPP	M-kDPP	Seq-kDPP	DS3	SeqFL
Change tire	(P, R)	(0.56, 0.50)	(0.55, 0.60)	(0.44, 0.40)	(0.56, 0.50)	(0.60, 0.60)
	F-score	0.53	0.57	0.42	0.53	**0.60**
Make coffee	(P, R)	(0.38, 0.33)	(0.50, 0.44)	(0.63, 0.56)	(0.50, 0.56)	(0.50, 0.56)
	F-score	0.35	0.47	**0.59**	0.53	0.53
CPR	(P, R)	(0.71, 0.71)	(0.71, 0.71)	(0.71, 0.71)	(0.71, 0.71)	(0.83, 0.71)
	F-score	0.71	0.71	0.71	0.71	**0.77**
Jump car	(P, R)	(0.50, 0.50)	(0.56, 0.50)	(0.56, 0.50)	(0.50, 0.50)	(0.60, 0.60)
	F-score	0.50	0.53	0.53	0.50	**0.60**
Repot plant	(P, R)	(0.57, 0.67)	(0.60, 0.50)	(0.57, 0.67)	(0.57, 0.67)	(0.80, 0.67)
	F-score	0.62	0.55	0.62	0.62	**0.73**
All tasks	(P, R)	(0.54, 0.54)	(0.58, 0.55)	(0.58, 0.57)	(0.57, 0.59)	(0.67, 0.63)
	F-score	0.54	0.57	0.57	0.58	**0.65**

Table 1: Precision (P), Recall (R) and F-score for summarization of instructional videos for five tasks.
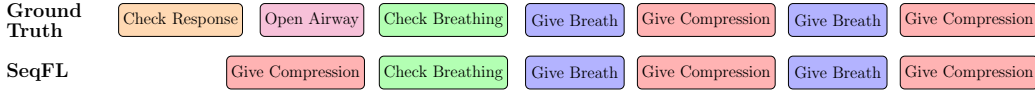

Figure 5: Ground-truth and automatic summarization results of our method for the CPR task.

4.2 Instructional Video Summarization

We apply SeqFL to the task of summarization of intructional videos to automatically learn the sequence of key actions to perform a task. We use videos from the instructional video dataset [45], which consists of 30 instructional videos for each of five activities. The dataset also provides labels for frames which contain the main steps required to perform that task. We preprocess the videos by segmenting each video into superframes [46] and obtain features using a deep neural network that we have constructed for feature extraction for summarization tasks. We use 60% of the videos from each task as the training set to build an HMM model whose states form the source set, $\mathbb{X}$. Using the learned dynamic model, we apply our method to summarize each of the remaining videos. The summaries for each video are sets of elements of $\mathbb{X}$, states in the HMM model. For evaluation, we map the representative states into actions in the ground truth by using the labels of the five nearest neighbors in the training set to each state. The summary for each video is an assignment of each superframe in the video to one of the action labels in the training set. Since each video may have shown each action performed for a different length of time, we remove consecutive repeated labels to form a list of actions performed, removing the length of time each action was performed. To construct the final summary for a task, we align the lists of actions obtained by each method for all the test videos, following the alignment method of [45] for several number of slots. For each method, we choose the number of HMM states and the alignment of videos that give the best performance (see the supplementary materials for more details). Table 1 shows the precision, recall and the F-score of various methods. Notice that existing methods perform similar to each other for most tasks, suggesting that sequential diversity promoted by Seq-kDPP and M-kDPP is not sufficient for capturing the important steps of tasks. On the other hand, for most tasks and over the entire dataset, our method (SeqFL) significantly outperforms other algorithms, better producing the sequence of important steps to perform a task, thanks to the ability of our framework to incorporate the underlying dynamics of the data. Figure 5 shows the ground-truth and the automatic summary produced by our method for the CPR task, demonstrating that we can sufficiently well capture the main steps and the sequence of steps to perform the task.

5 Conclusions

We developed a new framework for sequential subset selection that takes advantage of the underlying dynamic models of data, promoting to select a set of representatives that are compatible according to the dynamic models of data. By experiments on synthetic and real data, we showed the effectiveness of our method for summarization of sequential data. Our ongoing research include development of fast greedy algorithms for our sequential subset selection objective that can also deal with n-th order dynamic models as well as investigation of the theoretical guarantees of our proposed formulations.

References

[1] S. Garcia, J. Derrac, J. R. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 417–435, 2012.

[2] E. Elhamifar and M. C. D. P. Kaluza, "Online summarization via submodular and convex optimization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[3] B. Gong, W. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in *Neural Information Processing Systems*, 2014.

[4] I. Simon, N. Snavely, and S. M. Seitz, "Scene summarization for online image collections," in *IEEE International Conference on Computer Vision*, 2007.

[5] H. Lin and J. Bilmes, "Learning mixtures of submodular shells with application to document summarization," in *Conference on Uncertainty in Artificial Intelligence*, 2012.

[6] A. Kulesza and B. Taskar, "Determinantal point processes for machine learning," *Foundations and Trends in Machine Learning*, vol. 5, 2012.

[7] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, 2007.

[8] E. Elhamifar, G. Sapiro, and S. S. Sastry, "Dissimilarity-based sparse subset selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[9] E. Elhamifar, G. Sapiro, and R. Vidal, "Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery," *Neural Information Processing Systems*, 2012.

[10] G. Kim, E. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *International Conference on Computer Vision*, 2011.

[11] A. Shah and Z. Ghahramani, "Determinantal clustering process – a nonparametric bayesian approach to kernel based semi-supervised clustering," in *Conference on Uncertainty in Artificial Intelligence*, 2013.

[12] R. Reichart and A. Korhonen, "Improved lexical acquisition through dpp-based verb clustering," in *Conference of the Association for Computational Linguistics*, 2013.

[13] E. Elhamifar, S. Burden, and S. S. Sastry, "Adaptive piecewise-affine inverse modeling of hybrid dynamical systems," in *World Congress of the International Federation of Automatic Control (IFAC)*, 2014.

[14] E. Elhamifar and S. S. Sastry, "Energy disaggregation via learning 'powerlets' and sparse coding," in *AAAI Conference on Artificial Intelligence*, 2015.

[15] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, 2003.

[16] I. Misra, A. Shrivastava, and M. Hebert, "Data-driven exemplar model selection," in *Winter Conference on Applications of Computer Vision*, 2014.

[17] A. Krause, H. B. McMahan, C. Guestrin, and A. Gupta, "Robust submodular observation selection," *Journal of Machine Learning Research*, vol. 9, 2008.

[18] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Transactions on Signal Processing*, vol. 57, 2009.

[19] J. Hartline, V. S. Mirrokni, and M. Sundararajan, "Optimal marketing strategies over social networks," in *World Wide Web Conference*, 2008.

[20] D. McSherry, "Diversity-conscious retrieval," in *Advances in Case-Based Reasoning*, 2002.

[21] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley-Interscience, October 2004.

[22] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[23] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, 1989.

[24] F. Hadlock, "Finding a maximum cut of a planar graph in polynomial time," *SIAM Journal on Computing*, vol. 4, 1975.

[25] R. Motwani and P. Raghavan, "Randomized algorithms," *Cambridge University Press, New York*, 1995.

[26] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *SIGIR*, 1998.

[27] P. B. Mirchandani and R. L. Francis, *Discrete Location Theory*. Wiley, 1990.

[28] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions," *Mathematical Programming*, vol. 14, 1978.

[29] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[30] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin, "A convex model for non-negative matrix factorization and dimensionality reduction on physical space," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3239–3252, 2012.

[31] A. Borodin and G. Olshanski, "Distributions on partitions, point processes, and the hypergeometric kernel," *Communications in Mathematical Physics*, vol. 211, 2000.

[32] U. Feige, "A threshold of ln n for approximating set cover," *Journal of the ACM*, 1998.

[33] T. Gonzalez, "Clustering to minimize the maximum intercluster distance," *Theoretical Computer Science*, vol. 38, 1985.

[34] A. Civril and M. Magdon-Ismail, "On selecting a maximum volume sub-matrix of a matrix and related problems," *Theoretical Computer Science*, vol. 410, 2009.

[35] P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, and R. Ward, "Relax, no need to round: Integrality of clustering formulations," in *Conference on Innovations in Theoretical Computer Science (ITCS)*, 2015.

[36] A. Nellore and R. Ward, "Recovery guarantees for exemplar-based clustering," in *Information and Computation*, 2015.

[37] R. H. Affandi, A. Kulesza, and E. B. Fox, "Markov determinantal point processes," in *Conference on Uncertainty in Artificial Intelligence*, 2012.

[38] S. Tschiatschek, A. Singla, and A. Krause, "Selecting sequences of items via submodular maximization," *AAAI*, 2017.

[39] Z. Ghahramani and M. I. Jordan, "Factorial hidden markov models," *Machine Learning*, vol. 29, no. 2-3, 1997.

[40] Z. Ghahramani and S. Roweis, "Learning nonlinear dynamical systems using an em algorithm," *NIPS*, 2008.

[41] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2007.

[42] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[43] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[44] A. Kulesza and B. Taskar, "k-dpps: Fixed-size determinantal point processes," in *International Conference on Machine Learning*, 2011.

[45] J.-B. Alayrac, P. Bojanowski, N. Agrawal, I. Laptev, J. Sivic, and S. Lacoste-Julien, "Unsupervised learning from narrated instruction videos," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[46] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool, "Creating summaries from user videos," in *European Conference on Computer Vision*, 2014.