# Learning Robust Representations for Data Analytics

**Sheng Li**
Advisor: Yun Fu
Northeastern University, Boston, MA, USA
shengli@ece.neu.edu

## Abstract

Learning compact representations from high-dimensional and large-scale data plays an essential role in many real-world applications. However, many existing methods show limited performance when data are contaminated with severe noise. To address this challenge, we have proposed several effective methods to extract robust data representations, such as balanced graphs, discriminative subspaces, and robust dictionaries. In addition, several topics are provided as future work.

## 1 Introduction

Extracting knowledge from data is a critical task in many intelligent systems, which mitigates the gap between low-level observed data and high-level semantic information. Traditional machine learning methods usually pose strong assumptions on the distribution of data. For instance, linear discriminant analysis (LDA) assumes that the samples follow Gaussian distribution [Belhumeur *et al.*, 1997]. However, in practice, the data might be corrupted or contaminated with severe noise, which violates the assumptions on data distribution. As a result, traditional methods may have limited performance.

Recent advances on low-rank and sparse modeling have shown promising performance on recovering clean data from noisy observations [Candès *et al.*, 2011], which motivate us to develop new models to learn robust representations for various data analytic tasks. In particular, we have proposed methodologies on robust graph construction, robust subspace learning, robust dictionary learning, and robust multi-view learning. These methods have obtained remarkable improvements on many real-world applications, such as image classification [Li and Fu, 2015a], human motion segmentation [Li *et al.*, 2015a], person re-identification [Li *et al.*, 2015b], etc.. We will introduce the key ideas of these methods in the following sections, and also discuss several research topics as future work.

## 2 Completed Work

### 2.1 Robust Graph Construction

Graph based machine learning has shown promising performance in many tasks, such as classification, clustering, and semi-supervised learning. Graph provides a very effective way of representing underlying relationships in data. However, how to accurately measure these relationships during graph construction is always a challenging problem. In addition, sparsity in graphs is also preferred since sparse graphs have much less misconnections among dissimilar data points. We focus on addressing these two fundamental problems in graph construction, which are *similarity metric* and *graph sparsification*.

We propose practical graph construction algorithms in [Li and Fu, 2015a]. First, we represent the high-dimensional data in a low-rank coding space, $X = XZ + E$, in order to model and remove the noisy information, where $X$ is the data matrix, $Z$ is the low-rank representation coefficient matrix, and $E$ is the noise matrix. We then obtain the compact representations of data. Second, the fully connected graph is built using the compact representations $Z$. Third, the graph is sparsifid using two constraints, including the $k$-nearest neighbor ($k$-NN) constraint and the $b$-matching constraint. The former one leads to an unbalanced graph, while the latter one results in a balanced graph. Non-convex optimization algorithms are designed to solve the models efficiently. Experiments on image datasets show promising results on clustering and semi-supervised classification [Li and Fu, 2015a].

### 2.2 Robust Subspace Learning

Subspace learning is widely used in extracting discriminative features for classification. However, conventional subspace learning methods usually have strong assumptions on the data distribution, and therefore they are sensitive to the noisy data. The learned subspace has limited discriminability. To address this problem, we propose to exploit a discriminative and robust subspace, which is insensitive to noise or pose/illumination variations, for dimensionality reduction and classification. In particular, we propose a novel linear subspace approach named Supervised Regularization based Robust Subspace (SRRS) for pattern classification [Li and Fu, 2014; 2015b].

SRRS seeks low-rank representations from the noisy data, and learns a discriminative subspace from the recovered clean data jointly. A supervised regularization function based on Fisher criterion is designed to make use of the class label information and therefore to enhance the discriminability of subspace.

### 2.3 Robust Dictionary Learning

In the above methods, we assume that the data matrix $X$ is self-expressive, and therefore $X$ is simply used as the dictionary. A more flexible way is to learn an adaptive dictionary $D$, and the data reconstruction can be rewritten as $X = DZ + E$.

By taking advantages of the robust dictionary learning, we propose a temporal subspace clustering (TSC) method for human motion segmentation [Li *et al.*, 2015a]. We adopt the least-square regression based formulation to learn compact codings for each data point. To obtain more expressive codings, we learn a non-negative dictionary from data, instead of using the data self-expressive model. In addition, a temporal Laplacian regularization function is used to encode the sequential relationships in time series data. The objective of TSC is:

$$\begin{array}{ll} \min\limits_{Z,D} & \|X - DZ\|_{\mathrm{F}}^2 + \lambda_1 \|Z\|_{\mathrm{F}}^2 + \lambda_2 \mathrm{tr}(ZL_{\mathrm{T}}Z^\top), \\ \text{s.t.} & Z \geq 0,\ D \geq 0,\ \|d_i\|_2^2 \leq 1,\ i = 1, \cdots, r. \end{array} \quad (1)$$

where $L_{\mathrm{T}}$ is a temporal Laplacian matrix [Li *et al.*, 2015a].

After constructing an affinity graph using the codings, multiple temporal segments can be automatically grouped via spectral clustering. Experimental results on three action and gesture datasets show that TSC outperforms the related methods, which validates the effectiveness of robust dictionary learning [Li *et al.*, 2015a].

### 2.4 Robust Multi-View Learning

Nowadays information about objects can be collected from multiple views, due to the increasingly large amount of various sensors. The collected multi-view data could lead to significant improvement of machine learning tasks like classification. We study two real-world problems in multi-view settings, including *person re-identification* and *outlier detection*.

Person re-identification is the problem of matching pedestrian images observed from multiple non-overlapping cameras. It saves a lot of human efforts in many safety-critical applications such as video surveillance. We propose a cross-view projective dictionary learning (CPDL) approach for person re-identification [Li *et al.*, 2015b]. Two objectives are designed based on the CPDL framework, which extract compact representations for each pedestrian in the patch-level and the image-level, respectively. The proposed objectives can capture the intrinsic relationships of different representation coefficients in various settings. CPDL adopts the projective dictionary learning strategy, which is more efficient than the traditional $l_1$ optimization problem. We also design a strategy to fuse the similarity scores estimated in two levels. The effectiveness of CPDL has been validated in [Li *et al.*, 2015b].

Outlier detection, or anomaly detection, is a fundamental problem in data analytics. Conventional outlier detection algorithms are mainly designed for single-view data. Detecting outliers from multi-view data is still a very challenging problem, as the multi-view data usually have more complicated distributions and exhibit inconsistent behaviors in different views. To address this problem, we propose a multi-view low-rank analysis (MLRA) framework for outlier detection [Li *et al.*, 2015c]. MLRA pursuits outliers from the perspective of robust data representation. The cross-view low-rank coding is performed to reveal the intrinsic structures of data. Different from the existing multi-view outlier detection methods, MLRA is able to detect two different types of outliers from multiple views simultaneously. To this end, we design a criterion to estimate the outlier scores by analyzing the obtained representation coefficients.

## 3 Conclusions and Future Work

We have been exploring the usefulness of robust data representations for various data analytic tasks, including graph construction, subspace learning, dictionary learning, and multi-view learning, and have obtained quite promising results in several real-world applications.

In our future work, we will design robust data representation models for more extensive scenarios, such as multi-view streaming data analysis and transfer learning. In addition, we would like to provide more rigorous theoretical analysis to justify the effectiveness of the proposed methods.

## References


[Belhumeur *et al.*, 1997] Peter N Belhumeur, João P Hespanha, and David J Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[Candès *et al.*, 2011] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[Li and Fu, 2014] Sheng Li and Yun Fu. Robust subspace discovery through supervised low-rank constraints. In *SIAM International Conference on Data Mining (SDM)*, pages 163–171. SIAM, 2014.

[Li and Fu, 2015a] Sheng Li and Yun Fu. Learning balanced and unbalanced graphs via low-rank coding. *IEEE Trans. Knowledge and Data Engineering*, 27(5):1274–1287, 2015.

[Li and Fu, 2015b] Sheng Li and Yun Fu. Learning robust and discriminative subspace with low-rank constraints. *IEEE Trans. Neural Networks and Learning Systems*, 2015.

[Li *et al.*, 2015a] Sheng Li, Kang Li, and Yun Fu. Temporal subspace clustering for human motion segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4453–4461, 2015.

[Li *et al.*, 2015b] Sheng Li, Ming Shao, and Yun Fu. Cross-view projective dictionary learning for person re-identification. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2155–2161, 2015.

[Li *et al.*, 2015c] Sheng Li, Ming Shao, and Yun Fu. Multi-view low-rank analysis for outlier detection. In *SIAM International Conference on Data Mining (SDM)*. SIAM, 2015.