Unsupervised Feature Selection using a Blue Noise Graph Spectrum

Jayaraman J. Thiagarajan[†], Rushil Anirudh[†], Rahul Sridhar[‡] and Peer-Timo Bremer[†]

[†]Lawrence Livermore National Laboratory, [‡]Walmart Labs

Abstract

Dimension selection aims to solve the combinatorial problem of identifying the top-k dimensions for effective experiment design, reducing data while keeping it interpretable, and designing better sensing mechanisms. We develop a novel approach based on analysis of synthetic graph signals with blue noise spectra to select the most relevant dimensions. Using experiments in supervised learning and image masking, we demonstrate the superiority our approach in capturing crucial characteristics of high-dimensional spaces, using only a small subset of the original features.

Introduction

Dimension Selection: Select the most relevant dimensions from high-dimensional (HD) data, such that both complexity and robustness of downstream analysis is improved.

Applications: (i) Improve experiment design in ML pipelines, (ii) Adhere to communication, computation, and storage constraints in sensing systems.

Feature selection in unsupervised learning is challenging, yet a crucial problem in machine learning:

- Need to perform reduction in the input domain directly – different from dimension reduction.
- No access to target signals – cannot use heuristics based on *predictability* or *uncertainty*.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DEAC52- 07NA27344

Spectral Analysis of Signals Defined on Undirected Graphs

For an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ with node list $\mathcal{V}$, edge list $\mathcal{E}$ and adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, we construct the graph-shift operator as follows:

$$\tilde{s}_i = \sum_{j=1}^{N} \mathbf{L}_{i,j} s_j \implies \tilde{\mathbf{s}} = \mathbf{W}\mathbf{s}, \quad \text{where } s_i \text{ is the signal at node } v_i, \text{ and } \mathbf{L} = \mathbf{D}^{-1}\mathbf{A} \text{ is the normalized Laplacian matrix.}$$

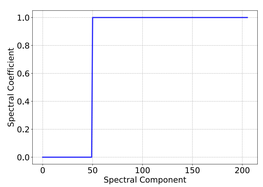
Spectral decomposition of a signal space corresponds to identifying subspaces that are invariant to the choice of filtering. The set of generalized eigenvectors of the graph Laplacian, $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{U} \in \mathbb{R}^{N \times N}$, is referred as the graph Fourier basis.

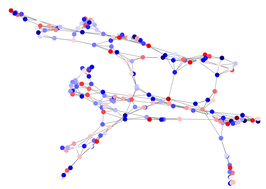
Graph Fourier Transform: $\mathbf{s} = \mathbf{U}\hat{\mathbf{s}}$, with the expansion coefficients $\hat{\mathbf{s}} = \mathbf{U}^{-1}\mathbf{s}$.

Blue Noise Spectrum

Blue Noise: (a) spectrum should be close to zero in the low-frequency region (no aliasing); (b) spectrum should be a constant in mid and high-frequency regions (reduce aliasing).

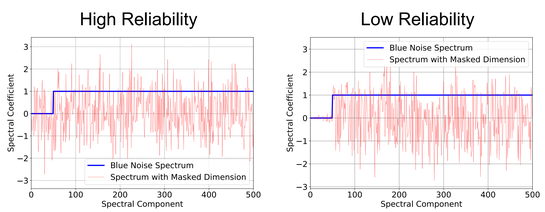
$$\hat{s}_k^b = \begin{cases} 0 \text{ if } k \leq k_0, \\ 1 \text{ if } k > k_0. \end{cases}$$





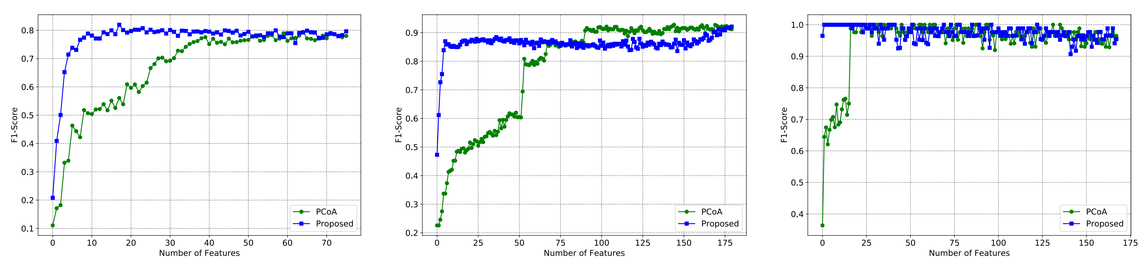
Core Idea

A predictable signal is characterized by *smoothness* with respect to the neighborhoods.



Impact of Feature Selection on Classifier Design

Unsupervised feature selection as a pre-processing step for classifier design in *small data* scenarios.



Designing Masking Patterns for Image Sensors

Acquire only a subset of pixel locations for a given distribution of images, such that it provides enough information to recover the local topology.