Canonical Correlation Analysis in Tensor Representation

Abstract

Canonical Correlation Analysis (CCA) is a multivariate statistical method used to identify relationships between two sets of variables. While traditionally applied to matrix data, modern applications often involve high-dimensional structured data, which can be more naturally represented as tensors. This paper presents a mathematical formulation of CCA in the context of tensor representation, offering a framework for exploring correlations between datasets structured as tensors.

1 Introduction

CCA was first introduced by Hotelling in 1936 and has since been widely used in various fields such as signal processing, neuroscience, and machine learning. The goal of CCA is to find linear combinations of variables from two datasets, $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathbb{R}^{n \times q}$, such that the correlation between the resulting projections is maximized. Given the growing use of tensor data (multi-dimensional arrays), CCA has been extended to handle tensors in order to exploit the structural information embedded in these multi-way data.

2 Tensor Representation of Data

Tensors are generalizations of matrices to higher dimensions. A matrix is a second-order tensor, whereas tensors of order 3 or more are termed as higher-order tensors. Let $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ and $\mathcal{Y} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_d}$ represent two d-way tensors, which generalize the traditional CCA where $\mathbf{X}$ and $\mathbf{Y}$ are matrices.

The extension of CCA to tensors involves finding projection matrices $\mathbf{A} \in \mathbb{R}^{n_1 \times k}$ and $\mathbf{B} \in \mathbb{R}^{m_1 \times l}$ that maximize the correlation between the projected tensor modes. The objective of Tensor CCA can be formulated as:

$$\max_{\mathbf{A},\mathbf{B}} \operatorname{corr}\left(\mathbf{A}^\top \mathcal{X}(1), \mathbf{B}^\top \mathcal{Y}(1)\right) \tag{1}$$

where $\mathcal{X}(1)$ and $\mathcal{Y}(1)$ are the mode-1 unfoldings of tensors $\mathcal{X}$ and $\mathcal{Y}$, respectively.

3 Mathematical Formulation

Given tensors $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ and $\mathcal{Y} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_d}$, the goal is to find linear transformations that maximize the correlation between corresponding modes of the tensors. The generalized objective is:

$$\max_{\mathbf{A},\mathbf{B}} \frac{\mathrm{tr}\left(\mathbf{A}^\top \mathcal{X}(1) \mathcal{Y}(1)^\top \mathbf{B}\right)}{\left(|\mathbf{A}^\top \mathcal{X}_{(1)}|F|\mathbf{B}^\top \mathcal{Y}(1)|_F\right)} \tag{2}$$

where $\mathrm{tr}(\cdot)$ represents the trace of a matrix and $|\cdot|_F$ denotes the Frobenius norm. This maximization seeks the best projection matrices $\mathbf{A}$ and $\mathbf{B}$ that align the two tensors in terms of their shared latent structure.

4 Discussion

CCA in tensor form allows us to generalize the correlations to multi-dimensional data, retaining the intrinsic relationships across multiple modes (dimensions). This makes Tensor CCA particularly powerful for analyzing complex data, such as those encountered in neuroimaging, genomics, and social network analysis. By projecting tensors instead of matrices, Tensor CCA can capture richer interactions that are not available in traditional CCA models.

5 Conclusion

Canonical Correlation Analysis has been extended to tensor data to handle multi-way structures inherent in modern datasets. Tensor CCA offers a more flexible and powerful approach for exploring correlations between higher-order data representations, and it has wide applications in areas involving complex, structured datasets. Further research could focus on computational efficiency and application to specific domains.