Scalable and Flexible Multiview MAX-VAR Canonical Correlation Analysis

Abstract:

Generalized canonical correlation analysis (GCCA) aims at finding latent low-dimensional common structure from multiple views (feature vectors in different domains) of the same entities. Unlike principal component analysis that handles a single view, (G)CCA is able to integrate information from different feature spaces. Here we focus on MAX-VAR GCCA, a popular formulation that has recently gained renewed interest in multilingual processing and speech modeling. The classic MAX-VAR GCCA problem can be solved optimally via eigen-decomposition of a matrix that compounds the (whitened) correlation matrices of the views; but this solution has serious scalability issues, and is not directly amenable to incorporating pertinent structural constraints such as nonnegativity and sparsity on the canonical components. We posit regularized MAX-VAR GCCA as a nonconvex optimization problem and propose an alternating optimizationbased algorithm to handle it. Our algorithm alternates between inexact solutions of a regularized least squares subproblem and a manifold-constrained nonconvex subproblem, thereby achieving substantialmemory and computational savings.An important benefit of our design is that it can easily handle structure-promoting regularization. We show that the algorithm globally converges to a critical point at a sublinear rate, and approaches a global optimal solution at a linear rate when no regularization is considered. Judiciously designed simulations and large-scale word embedding tasks are employed to showcase the effectiveness of the proposed algorithm.