Probabilistic PCA Progress Report

Michael Montemurri, Ahmed Mhedhbi
December 2, 2024

Summary of Relevant Papers and Connections

Our project examines three foundational papers that extend Principal Component Analysis (PCA) into probabilistic frameworks. The first, *Probabilistic Principal Component Analysis* by Tipping and Bishop, frames PCA as a maximum likelihood problem using a Gaussian latent variable model, introducing the Expectation-Maximization (EM) algorithm for parameter estimation. Building on this, their second paper, *Mixtures of Probabilistic Principal Component Analyzers*, generalizes PCA to datasets with complex, multimodal structures by introducing a mixture model approach. Finally, *A Generalization of Principal Components Analysis to the Exponential Family* by Collins, Dasgupta, and Schapire extends PCA to handle data distributions from the exponential family (e.g., Bernoulli, Poisson) rather than assuming Gaussian noise. Together, these papers form a cohesive narrative: PPCA provides a probabilistic foundation for PCA, mixture models expand its applicability to multimodal data, and exponential family PCA broadens the scope to diverse data types.

Current Progress and Implementation Plan

While the majority of the algorithms outlined in these papers have already been implemented in Scikit-learn, we plan to implement the algorithms from scratch so that we can adapt them to our proposed extensions. We will begin with the EM algorithm for PPCA and its mixture extension. Results will be compared to traditional eigen-decomposition methods, which, while exact, are computationally expensive for high dimensional data. For exponential family PCA, we will adapt optimization techniques based on Bregman distances as introduced in the paper to align with the distributional assumptions of the exponential family. We will apply these implemented methods to various datasets in order to highlight and compare their relative strengths and weaknesses.

Extensions and Next Steps

In addition to these implementations, we will explore the following original extensions. We will explore the possibility of combining mixture models with exponential family PCA to handle multimodal and non-Gaussian data simultaneously. This will involve modifying the likelihood functions in the mixture models algorithm to accommodate exponential family distributions (e.g., Poisson, Bernoulli) and adapting the clustering and subspace learning accordingly. In addition to this, we will explore the idea of extending the probabilistic framework to the *Kernel* PCA and multidimensional scaling through Bayesian optimization techniques discussed in class.