Dimensionality Reduction using Probabilistic Principal Component Analysis

Auden Cote-L'Heureux, Quentin Moliterno, Tom Barber

Introduction

Probabilistic principal component analysis (PPCA) is an extension of the traditional principal component analysis (PCA) technique of dimension reduction that incorporates probabilistic methods. PPCA is used to reduce high-dimensional data into a lower dimensional `latent' space that captures a maximal amount of variance. This latent space can then be used to generate data that approximates the original experimental data, or can be used itself for visual representation of the data (e.g. in 2 or 3 dimensions) or for data storage and compression.

Motivation

Probabilistic PCA was introduced in 1999, building off of earlier work that laid the foundation for PPCA and was soon expanded by introducing Bayesian inference methods. This work demonstrated the potential of the probabilistic approach and discussed its application to modelling complex (non-linear) relationships. Diverse dimension-reduction models have since been applied to a wide variety of problems, from effective data visualization tools to machine learning.

Theory

PPCA is applied to an observed data set *X* of dimensionality *D x N* with the goal of reducing data to a latent space of dimension *M* using a linear transformation. The following relationship is key:

$$\boldsymbol{X = WZ + \mu + \varepsilon}$$

Where *W* is a *D x M* matrix of principal axes, *Z* is the latent vector space, *μ* is the mean vector, and *ε* is the noise matrix.

Probabilistic Assumptions:

- *Z* follows a multivariate Gaussian distribution with zero mean and diagonal covariance matrix *I*
- *ε* is Gaussian noise with zero mean and diagonal covariance matrix $\sigma^2 I$
- These key assumptions enable the derivation of the log-likelihood function

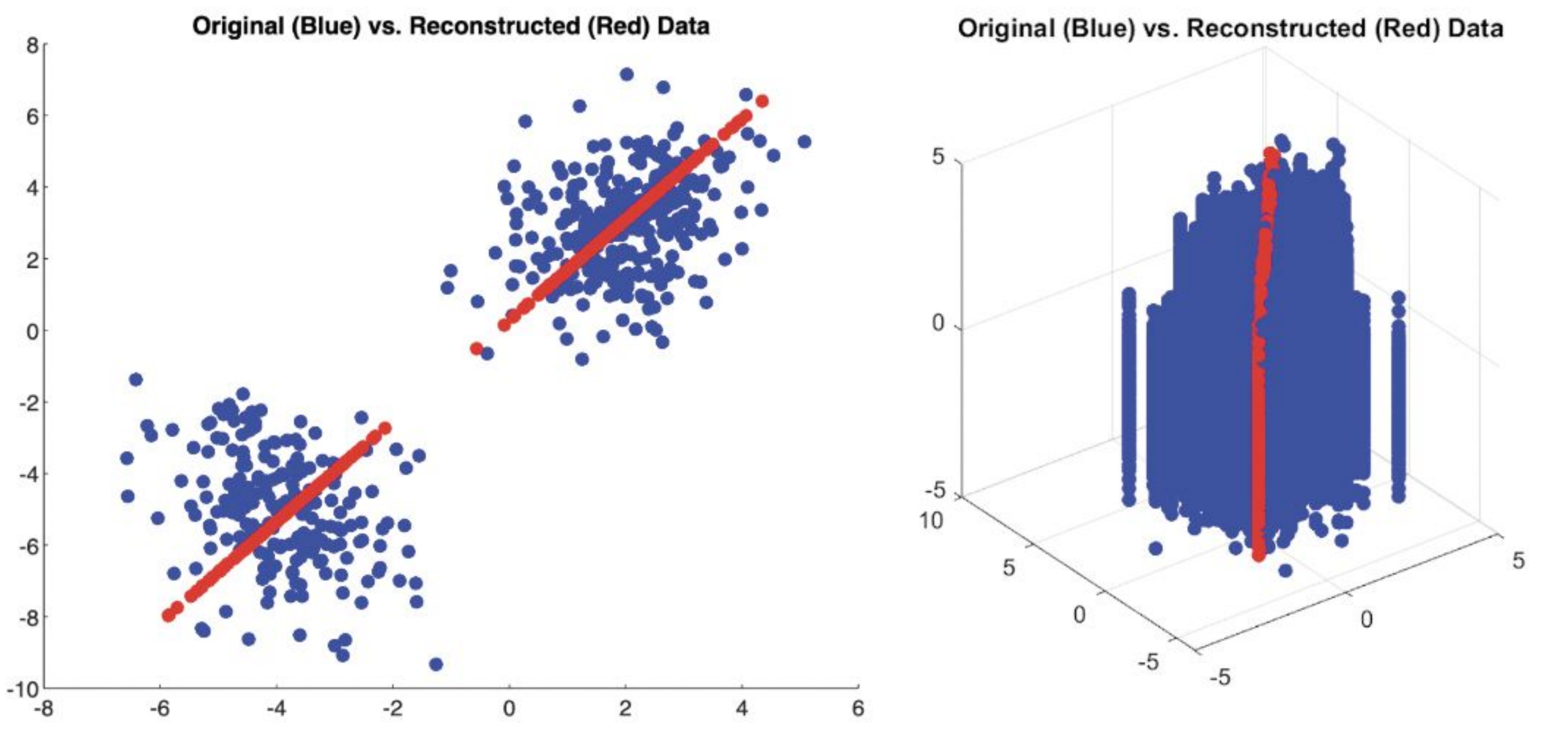
Optimization

- Parameters (W, Z, *ε*) estimated through expectation maximization (EM) or maximum likelihood estimation (MLE).
- The log-likelihood function that is iteratively optimized is shown below

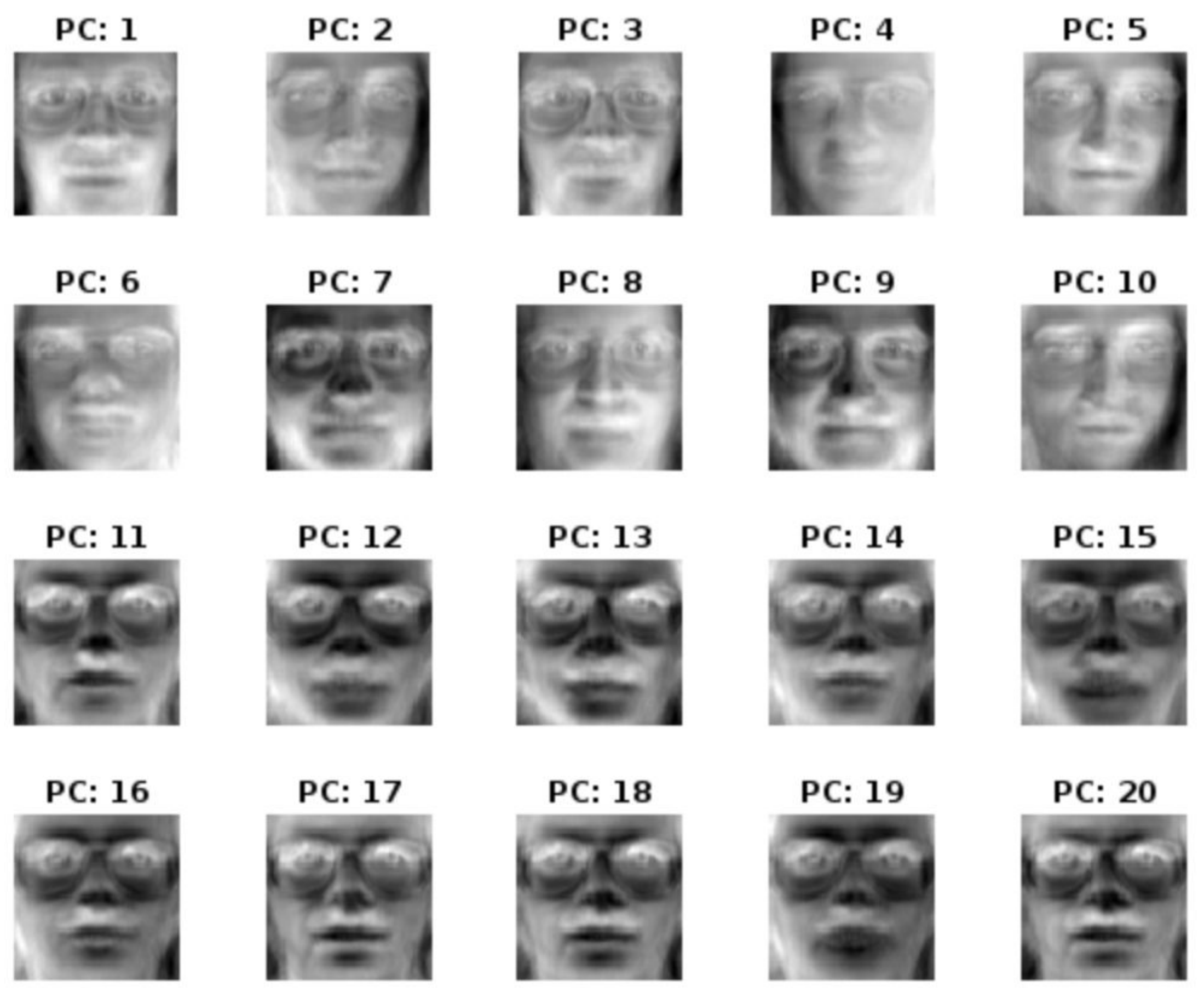
$$\mathcal{L} = -\frac{N}{2}\left[D * log(2\pi) + log(Det(C)) + tr(C^{-1}S)\right]$$

$$C = WW^T + \sigma^2 I \qquad S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T$$

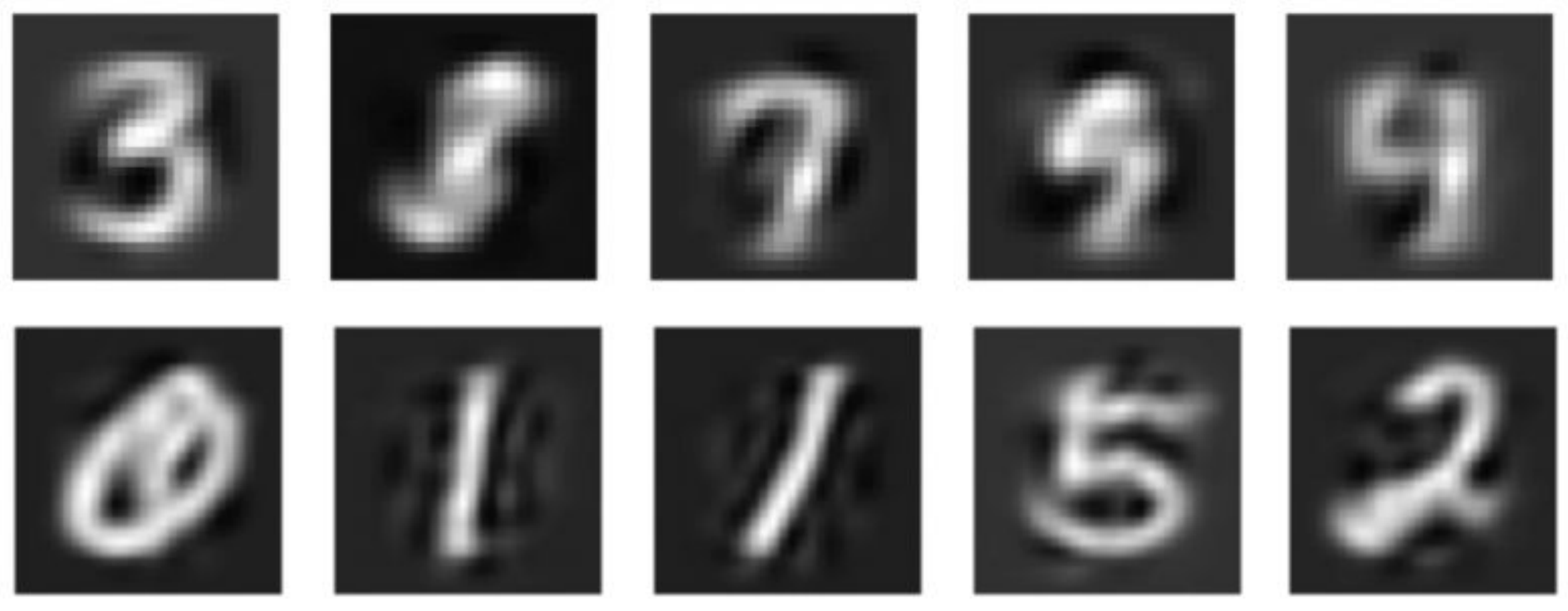
Results



Dimension reduction: We first ran a PPCA example in MATLAB on a simple mixture model of two 2-dimensional Gaussian distributions (left). The original dimensionality of 2 (blue points) is reduced to a latent space that is 1-dimensional (red points). Similarly, a 3-dimensional set of points (right, blue points) can be reduced to a plane (red points).



Generative potential: We applied PPCA to two datasets of images: the Olivetti Faces dataset of 400 grayscale 64 × 64–pixel images of faces, and the MNIST dataset of 70,000 28 x 28 images of handwritten digits from 0 to 9. For the Olivetti Faces, we used PPCA as implemented in MATLAB to reduce the experimental data (photographs) down to 20 latent dimensions (shown cumulatively above). We also generated MNIST images from 10 latent dimensions (top row below) and 50 latent dimensions (bottom row) using a Python script:



Acknowledgements

Tipping, M.E. and Bishop, C.M. (1999), Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61: 611-622. https://doi.org/10.1111/1467-9868.00196

Sam Roweis. 1998. EM algorithms for PCA and SPCA. In Proceedings of the 1997 conference on Advances in neural information processing systems 10 (NIPS '97). MIT Press, Cambridge, MA, USA, 626–632.

A. T. Cemgil, B. Kappen and D. Barber, "Generative model based polyphonic music transcription," 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684), New Paltz, NY, USA, 2003, pp. 181-184, doi: 10.1109/ASPAA.2003.1285861.

S. Velliangiri, S. Alagumuthukrishnan, S Iwin Thankumar joseph,
A Review of Dimensionality Reduction Techniques for Efficient Computation, Procedia Computer Science, Volume 165, 2019, Pages 104-111, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.01.079.

Olivetti Faces as downloaded from Sam Roweis (https://cs.nyu.edu/roweis/), originally supplied by AT\&T Laboratories Cambridge.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6), 141–142.

K. P. Murphy, Probabilistic Machine Learning: Advanced Topics, The MIT Press, 2023

ChatGPT 3.5 and following prompts: "Can you write a Matlab code to perform probabilistic principle component analysis on a gaussian mixture model?", "write Matlab code to run probabilistic PCA on a GMM of 3-dimensional Gaussians and visualize the principal components and the amount of variance explained by each component.", "rewrite [github] code as is but use MLE instead of EM to find parameters.", "summarize our Theory into bullet points