VS265 Problem 16: PCA, ICA and Sparse-Coding

Archit Gupta

1 Problem description

Over the years, scientists have debated on processing strategies employed by the brain to process the vast stream of sensory information in the natural world. Exploiting the underlying structure in the sensory information arising from natural scenarios is a recurring theme that has gained wide consensus. Principal Component Analysis (PCA), Independent Component Analysis (ICA) and sparse-coding are some of the methods that allow us to find underlying or latent representations in data. Furthermore, under appropriate circumstances, all three methods can be applied for reducing the data dimensionality.

In this challenge problem, we compare and contrast the three methods. First, in Section 2 we look at the mathematical definitions of these modeling paradigms and comment on their similarities and difference. In Section 3, we look at simplified datasets that illustrate the similarities of the three methods. Section 4 illustrates how these methods differ, again on simplified datasets. Finally, we look at some real-world applications of these methods in Section **??** and summarize our observations.

2 Definitions

The models under discussion here assume that the underlying generative model is linear. In particular, the data, given by $\mathbf{x} \in \mathbb{R}^n$ arises from the model[1] given in Equation 1.

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n} \tag{1}$$

The matrix $\mathbf{A}$ has n rows and m columns with each column representing a basis vector. $\mathbf{x}$ can now be represented as a weighted sum of the basis vectors in $\mathbf{A}$. If we call the m basis vectors $\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_m}$, where $\mathbf{A_k} \in \mathbb{R}^n$, we are using a weighted combination $s_1\mathbf{A_1} + s_2\mathbf{A_2} + \ldots, s_m\mathbf{A_m}$ to represent $\mathbf{x}$. The basis coefficients $s_1, s_2, \ldots, s_m$ constitute the coefficient vector $\mathbf{s} \in \mathbb{R}^m$. Additionally, we incorporate $\mathbf{n}$ to model noise in measurement or modeling. Typically, $\mathbf{n}$ is assumed to be comprised of Independent and Identically Distributed (IID) Gaussian random variables.

Next we will look at how the three approaches PCA, ICA and sparse-coding find the model parameters $\mathbf{A}$ and $\mathbf{s}$.

2.1 PCA

PCA gives us a complete basis of n basis vectors, called the principal components. Because of the strict constraints on the PCA solution, it is mathematically tractable and offers a lot of insight into various properties which can then be generalized to the other algorithms. In PCA, we constrain the basis vectors $\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_n}$ to be orthonormal, *i.e.*,

$$\mathbf{A_i} \cdot \mathbf{A_j} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \tag{2}$$

The basis vectors (which are now called principal components) being orthogonal results in the matrix $\mathbf{A}_{n \times n}$ being a unitary matrix.

$$\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^* = \mathbf{I}, \tag{3}$$

where $\mathbf{I}$ is the identity matrix of size n. Assuming that the data $\mathbf{x}$ is centered, the resultant matrix $\mathbf{A}$ can be found by diagonalizing the covariance matrix $\mathbf{X}\mathbf{X}^T$, where $\mathbf{X}$ is an $n \times m$ matrix comprised of the data entries $\mathbf{x}$ stacked along its columns.

[1] We assume that the data lies in an n-dimensional vector space

Since the covariance matrix is square and symmetric, it is always diagonalizable. Diagonalizing the covariance matrix gives us

$$\mathbf{X}\mathbf{X}^T = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^* \tag{4}$$

The matrix $\mathbf{\Lambda}$ is vital for an understanding of PCA as the diagonal entries represent the total amount of variance explained by the principal component.

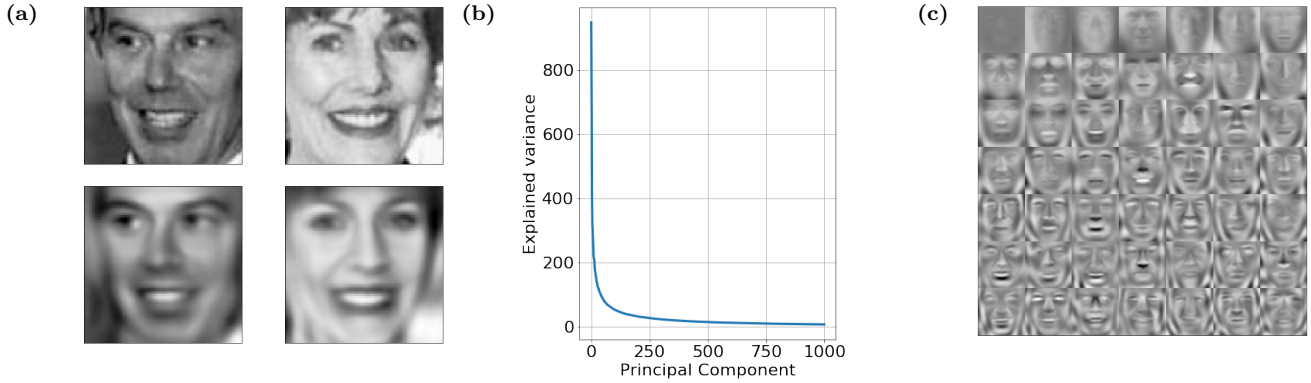

Figure 1: Illustration of image compression with PCA. (a) Example (Top) images of faces and their reconstruction (Bottom) using the first 100 of 13233 principal components. Variance explained by each of the first 1000 principal components. Top 50 eigenfaces (principal components).

With this, we can list some of the important properties of PCA that we will compare and contrast against the other methods.

1. The basis vectors and obtained from PCA are complete and orthonormal.
2. Different basis vectors explain different amounts of variance in the data - In some sense, they are not equally important.
3. The weights (or variance explained) for different basis vectors drops rapidly in most natural datasets.
4. Basis vectors are unique up to permutation and sign - This follows naturally from diagonalization not being unique. In practice, principal components are often arranged in a decreasing order of explained variance.

2.2 ICA

Similar to PCA, ICA produces a complete basis of n basis vectors, or independent components. While in PCA, the basis vectors $\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_m}$ are constrained to be orthonormal, in ICA, this constraint is relaxed to $\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_m}$ being linearly independent. Mathematically

$$\mathbf{As} \neq \mathbf{0}, \text{ for } \mathbf{s} \in \mathbb{R}^n - \mathbf{0} \tag{5}$$

Such a relaxation is more useful in practice, especially when we have prior information that $\mathbf{x}$ arises from a mixture of linearly-independent sources. Solving for the best solutions $\mathbf{A}$ and $\mathbf{s}$ is computationally expensive and several numerical solutions exist for finding local minima that minimize the error [1, 2]. More importantly, the PCA solution is a valid ICAsolution as well.

2.3 Sparse-Coding

In sparse-coding, we usually find an over-complete basis or dictionary of vectors, $m > n$. The constraint changes to a minimization of sparsity of the coefficient vector $\mathbf{s}$.

$$\min_{\mathbf{A},\mathbf{s}} \ \|\mathbf{s}\|_0, \text{ s.t. } \mathbf{x} = \mathbf{As} + \mathbf{n} \tag{6}$$

For a fixed number of basis vectors (m), this often results in a trade-off between

1. Accurately modeling the data, i.e. reducing the error $\|\mathbf{e}_2\|^2 = \|\mathbf{x} - \mathbf{As}\|_2^2$, and
2. The desired sparsity or $\|\mathbf{s}\|_0$, which is the typical number of non-zero coefficients required to express a data point $\mathbf{x}$.

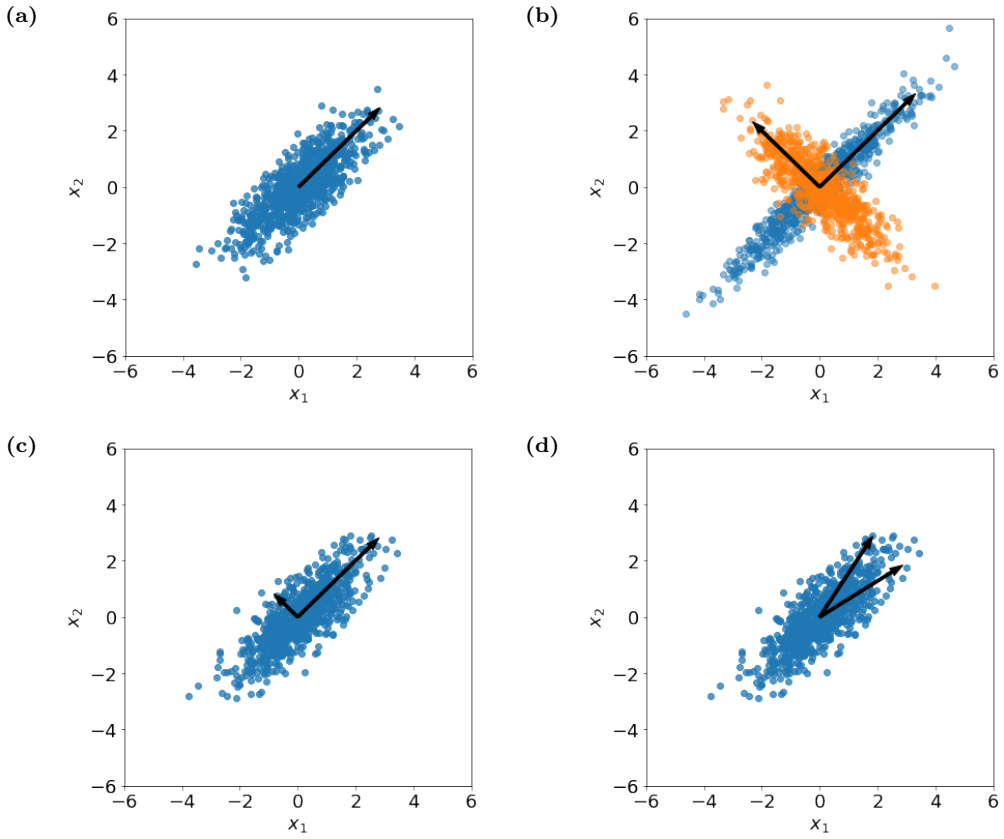

Figure 2: PCA vs. ICA *vs.* Sparse-Coding on data generated from orthogonal sources.

3 Similar properties

4 Differences

5 Summary

References

[1] Erkki Oja and Zhijian Yuan. The fastica algorithm revisited: Convergence analysis. *IEEE Transactions on Neural Networks*, 17(6):1370–1381, 2006.

[2] Aapo Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

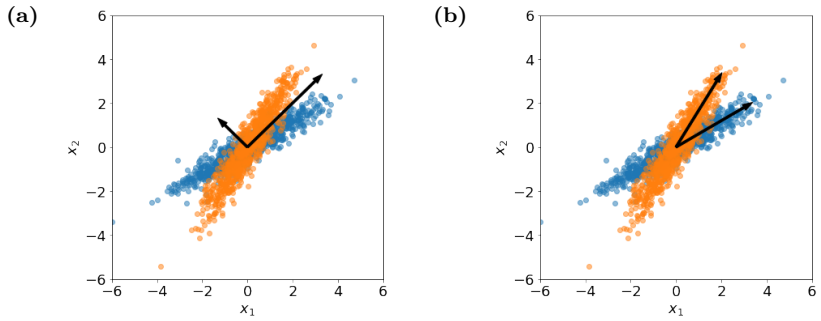

Figure 3: PCA vs. ICA *vs.* Sparse-Coding on data generated from non-orthogonal sources.

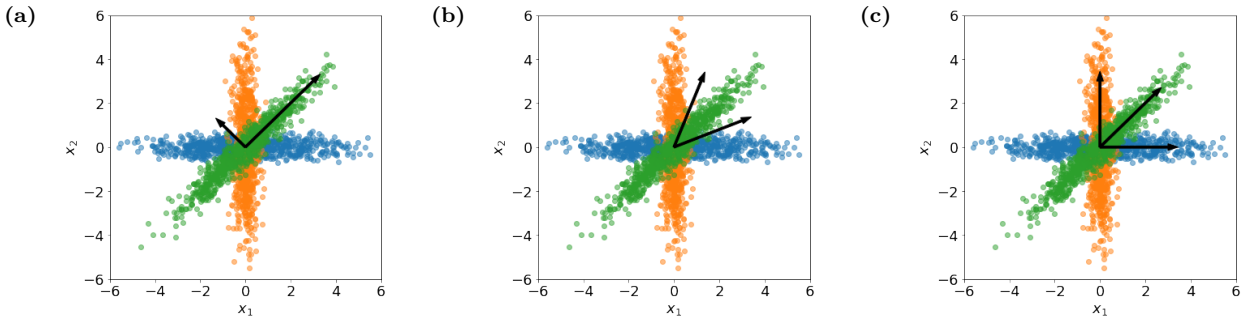

Figure 4: PCA vs. ICA *vs.* Sparse-Coding on data generated from redundant sources.