PCA, t-SNE, and Orthogonal Procrustes Problem

anton.selitskii

January 2025

1 PCA

1.1 Problem Formulation

In this notes, we derive PCA from the dimensionality reduction perspective.

We assume that we have n data points with p features:

$$x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_p^{(i)} \end{bmatrix}, \qquad i = 1, \dots, n. \tag{1}$$

We want to find a low-dimensional representation of the data

$$z^{(i)} = \begin{bmatrix} z_1^{(i)} \\ \vdots \\ z_\ell^{(i)} \end{bmatrix}, \qquad i = 1, \dots, n, \tag{2}$$

with a linear reconstruction

$$\hat{x}^{(i)} = W z^{(i)}, \qquad W \in M(p \times \ell). \tag{3}$$

We denote by $M(n_1 \times n_2)$ the space of real-valued matrices with n_1 rows and n_2 columns.

We want to minimize the reconstruction loss

$$L(W, z^{(1)}, \dots, z^{(n)}) = \sum_{i=1}^{n} \|x^{(i)} - \hat{x}^{(i)}\|^2 = \sum_{i=1}^{n} \|x^{(i)} - W z^{(i)}\|^2 \tag{4}$$

under the additional assumption that the columns W_k of the matrix W are orthonormal:

$$\|W_i\|^2 = 1, \qquad (W_i, W_j) = 0, \qquad i, j = 1, \dots, \ell, \quad i \neq j. \tag{5}$$

Denote by X a matrix with vectors $x^{(i)}$ in rows:

$$X = \begin{bmatrix} -x^{(1)T}- \\ -x^{(2)T}- \\ \dots \\ -x^{(n)T}- \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_p^{(2)} \\ \dots & \dots & \dots & \dots \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_p^{(n)} \end{bmatrix}. \tag{6}$$

Similarly, denote by Z a matrix with vectors $z^{(i)}$ in its rows. Then the loss function (4) can be written as follows[1]:

$$L(W, Z) = \|X^T - WZ^T\|_F^2. \tag{7}$$

Condition (5) is equivalent to the following equations:

$$W_i^T W_j = \delta_{ij}, \quad i, j = 1, \dots, \ell, \tag{8}$$

or in matrix notation

$$W^T W = I. \tag{9}$$

We arrive to the minimization problem:

$$\begin{cases} L(W, Z) = \|X^T - WZ^T\|_F^2 \to \min\limits_{W,Z}, \\ W^T W = I. \end{cases} \tag{10}$$

Because it contains only equality constrains, we can use the Lagrange multipliers method. Compose the Lagrange function:

$$\mathcal{L}(W, Z, \Lambda) = \|X^T - WZ^T\|_F^2 - \sum_{i=1}^{\ell} \lambda_{ii}(W_i^T W_i - 1) - \sum_{\substack{i,j=1 \\ i<j}}^{\ell} \lambda_{ij} W_i^T W_j. \tag{11}$$

Fist, we take the derivatives[2] with respect to $z^{(i)}$:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial z^{(i)}} &= \frac{\partial}{\partial z^{(i)}} \sum_{k=1}^{\ell} \|x^{(k)} - Wz^{(k)}\|^2 \\ &= \frac{\partial}{\partial z^{(i)}} \sum_{k=1}^{\ell} (x^{(k)} - Wz^{(k)})^T (x^{(k)} - Wz^{(k)}) \\ &= \frac{\partial}{\partial z^{(i)}} \sum_{k=1}^{\ell} (x^{(k)T} - z^{(k)T} W^T)(x^{(k)} - Wz^{(k)}) \\ &= \frac{\partial}{\partial z^{(i)}} \sum_{k=1}^{\ell} (x^{(k)T} x^{(k)} - 2x^{(k)T} W z^{(k)} + z^{(k)T} W^T W z^{(k)}) \\ &= -2W^T x^{(i)} + 2z^{(i)} = 0. \end{aligned} \tag{12}$$

---

[1] Recall that the Frobenius norm of a matrix A is the square root from the sum of squared elements of this matrix: $\|A\|_F = \sqrt{\sum\limits_{i,j} a_{ij}^2}$

[2] We use the following equalities $\frac{\partial a^T x}{\partial x} = a$ and $\frac{\partial x^T A x}{\partial x} = 2Ax$. See, e.g., `https://en.wikipedia.org/wiki/Matrix_calculus`

We find that

$$z^{(i)} = W^T x^{(i)}. \tag{13}$$

Substituting this representation in loss function (7), we obtain a new optimization problem:

$$\begin{cases} L(W) = L(W, Z(W)) = \|X^T - WW^TX^T\|_F^2 \to \min\limits_W, \\ W^TW = I. \end{cases} \tag{14}$$

Rewrite the target function:[3]

$$\begin{aligned} \|X^T - WW^TX^T\|_F^2 &= Tr\left((X^T - WW^TX^T)^T(X^T - WW^TX^T)\right) \\ &= Tr\left((X - XWW^T)(X^T - WW^TX^T)\right) \\ &= Tr\left(XX^T - 2XWW^TX^T + XWW^TWW^TX^T\right) \\ &= Tr\left(XX^T - 2XWW^TX^T + XWW^TX^T\right) \\ &= Tr\left(XX^T - XWW^TX^T\right). \end{aligned} \tag{15}$$

Because the first term does not depend on W, problem (14) can be reformulated as follows:

$$\begin{cases} L(W) = Tr\left(-XWW^TX^T\right) \to \min\limits_W, \\ W^TW = I. \end{cases} \tag{16}$$

Or equivalently[4] as

$$\begin{cases} L(W) = Tr\left(\frac{1}{n}X^TXWW^T\right) \to \max\limits_W, \\ W^TW = I. \end{cases} \tag{17}$$

The matrix $\frac{1}{n}X^TX$ is known as a covariance matrix and will be denoted by

$$\Sigma = \frac{1}{n}X^TX. \tag{18}$$

Compose the Lagrange function for this problem:

$$\mathcal{L}(W, \underset{\ell\times\ell}{\Lambda}) = Tr\left(\Sigma WW^T\right) - \sum_{i=1}^{\ell} \lambda_{ii}(W_i^TW_i - 1) - \sum_{\substack{i,j=1 \\ i<j}}^{\ell} \lambda_{ij}W_i^TW_j. \tag{19}$$

Here, we used a lower-triangular matrix $\underset{\ell\times\ell}{\Lambda} \in M(\ell\times\ell)$ to store the Laplace multipliers:

$$\underset{\ell\times\ell}{\Lambda} = \begin{bmatrix} \lambda_{11} & 0 & \cdots & 0 \\ \lambda_{21} & \lambda_{22} & \cdots & 0 \\ \dots & \dots & \dots & \dots \\ \lambda_{\ell 1} & \lambda_{\ell 2} & \cdots & \lambda_{\ell\ell} \end{bmatrix}. \tag{20}$$

[3]We use the following identity $\|A\|_F^2 = Tr(A^TA)$. The trace of a matrix is defined as the sum of its diagonal elements.

[4]We use the following property of the trace: $Tr(AB) = Tr(BA)$.

Taking the derivatives with respect to columns W_k and λ_{ij}, we get the following system of equations:[5]

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial W_k} = 2\Sigma W_k - 2\lambda_{kk} W_k - \sum\limits_{i=1}^{k-1} \lambda_{ik} W_i - \sum\limits_{j=k+1}^{\ell} \lambda_{kj} W_j = 0, \quad k = 1, \ldots, \ell, \\ \dfrac{\partial \mathcal{L}}{\partial \lambda_{kk}} = W_k^T W_k - 1 = 0, \quad k = 1, \ldots, \ell, \\ \dfrac{\partial \mathcal{L}}{\partial \lambda_{ij}} = W_i^T W_j = 0, \quad i, j = 1, \ldots, \ell, \ i < j. \end{cases} \tag{21}$$

If we multiply the first equation by W_k^T from the left, we will get

$$2W_k^T \Sigma W_k - 2\lambda_{kk} W_k^T W_k - \sum_{i=1}^{k-1} \lambda_{ik} W_k^T W_i - \sum_{j=k+1}^{\ell} \lambda_{kj} W_k^T W_j = 0, \quad k = 1, \ldots, \ell. \tag{22}$$

Using the orthogonality conditions, we get

$$W_k^T \Sigma W_k - \lambda_{kk} = 0, \quad k = 1, \ldots, \ell. \tag{23}$$

And multiplying the last equation by W_k from the left, it takes form

$$\Sigma W_k = \lambda_{kk} W_k, \quad k = 1, \ldots, \ell, \tag{24}$$

which is the eigenvalue problem for the matrix Σ.

Because the matrix $\Sigma \in M(p \times p)$ is symmetric,[6] it has p orthogonal eigenvectors[7] and corresponding eigenvalues are real and nonnegative.[8] We can choose $\ell \leq p$ columns of the matrix W as normalized eigenvectors of the matrix Σ and λ_{kk} as corresponding eigenvalues. They satisfy equation (24) and conditions (9).

Due to the linear independence of the eigenvectors W_k from the first equation of (21) it follows that off-diagonal elements $\lambda_{ij} = 0$ with $i < j$. Therefore, matrix Λ has form

$$\underset{\ell \times \ell}{\Lambda} = \begin{bmatrix} \lambda_{11} & 0 & \cdots & 0 \\ 0 & \lambda_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_{\ell\ell} \end{bmatrix}. \tag{25}$$

If we substitute this solution to the target function, we will get

$$L(W) = Tr\left(\Sigma W W^T\right) = Tr\left(W^T \Sigma W\right) = Tr(\underset{\ell \times \ell}{\Lambda}) = \lambda_{11} + \ldots + \lambda_{\ell\ell}. \tag{26}$$

To get the maximal value, we should use the first ℓ maximal eigenvalues of the matrix Σ.

---

[5] To calculate the derivative $\dfrac{\partial \mathcal{L}}{\partial W_k}$ we use formula $\dfrac{\partial Tr(\Sigma W W^T)}{\partial W} = 2\Sigma W$ and then take kth column.

[6] For $A = X^T X$ $A^T = (X^T X)^T = X^T (X^T)^T = X^T X = A$.

[7] Theorem

[8] If h is an eigenvector of the matrix $X^T X$ with the eigenvalue λ, $X^T X h = \lambda h$. Therefore, $\lambda h^T h = h^T X^T X h = (Xh)^T (Xh) = \|Xh\|^2 \geq 0$. By definition of the eigenvector, $h \neq 0$, that means that $h^T h = \|h\|^2 > 0$; this leads to $\lambda \geq 0$.

1.2 Statistical Interpretation

Before applying PCA algorithm, i.e., calculating ℓ eigenvectors of the matrix Σ corresponding to the ℓ greatest eigenvalues, we it is recommended to subtract the mean:

$$x \to x - \bar{x}, \tag{27}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_1^{(i)} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_p^{(i)} \end{bmatrix}. \tag{28}$$

If we assume the zero mean of the points $x^{(i)}$, then the low-dimensional points

$$z^{(i)} = W^T x^{(i)} = \begin{bmatrix} -W_1^T- \\ \cdots \\ -W_\ell^T- \end{bmatrix} \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_p^{(i)} \end{bmatrix} = \begin{bmatrix} W_1^T x^{(i)} \\ \cdots \\ W_\ell^T x^{(i)} \end{bmatrix} \tag{29}$$

will have zero mean as well. Then their variance can be calculated as follows:

$$\begin{aligned} S_{z_k}^2 &= \frac{1}{n}\sum_{i=1}^{n}\big(z_k^{(i)}\big)^2 = \frac{1}{n}\sum_{i=1}^{n}\big(W_k^T x^{(i)}\big)^2 = \frac{1}{n}\sum_{i=1}^{n} W_k^T x^{(i)} x^{(i)T} W_k \\ &= W_k^T\left(\frac{1}{n}\sum_{i=1}^{n} x^{(i)} x^{(i)T}\right) W_k = W_k^T \Sigma W_k = W_k^T \lambda_{kk} W_k = \lambda_{kk}. \end{aligned} \tag{30}$$

The total variance of the data points $z^{(i)}$ is equal to

$$\sum_{k=1}^{n} S_{z_k}^2 = \lambda_{11} + \ldots + \lambda_{\ell\ell}. \tag{31}$$

The total variance of the data points $x^{(i)}$ is equal to[9]

$$\sum_{k=1}^{n} S_{x_k}^2 = Tr(\Sigma) = Tr(W \underset{p\times p}{\Lambda} W^T) = Tr(\underset{p\times p}{\Lambda}) = \lambda_{11} + \ldots + \lambda_{pp}. \tag{32}$$

This means that if we take $\ell = p$ components, then we reconstruct the data without any losses. It will correspond to the rotation of the original data. The fraction

$$\frac{\lambda_{11} + \ldots + \lambda_{\ell\ell}}{\lambda_{11} + \ldots + \lambda_{pp}} \tag{33}$$

represents the fraction of the explained variance.

[9] We assume that the eigenvalues are sorted in descending order.

1.3 Geometric Interpretation

We will show that reconstruction of the projection $z^{(i)}$ is the orthogonal projection of the point $x^{(i)}$ on the subspace spanned by vectors $W_1, \dots, W_\ell$.

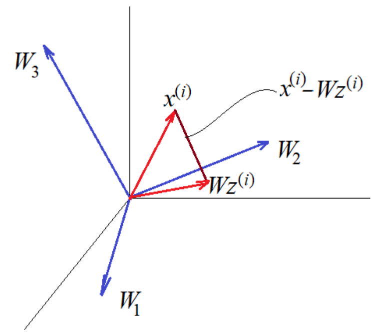


Figure 1: Orthogonal projection of a 3D vector $x^{(i)}$ on 2D plane of two principal components W_1 and W_2

It is sufficient to show that the residuals are orthogonal for each component:

$$x^{(i)} - Wz^{(i)} \perp W_k, \qquad k = 1, \dots, \ell. \tag{34}$$

Equivalently, their dot product should be zero:

$$\begin{aligned}
W_k^T(x^{(i)} - Wz^{(i)}) &= W_k^T x^{(i)} - W_k^T \begin{bmatrix} | & & | \\ W_1 & \cdots & W_\ell \\ | & & | \end{bmatrix} z^{(i)} \\
&= W_k^T x^{(i)} - [\underset{1}{0}, \dots, \underset{k}{1}, \dots \underset{\ell}{0}] z^{(i)} \\
&\overset{(29)}{=} W_k^T x^{(i)} - [\underset{1}{0}, \dots, \underset{k}{1}, \dots \underset{\ell}{0}] \begin{bmatrix} W_1^T x^{(i)} \\ \cdots \\ W_\ell^T x^{(i)} \end{bmatrix} \\
&= W_k^T x^{(i)} - W_k^T x^{(i)} = 0.
\end{aligned} \tag{35}$$

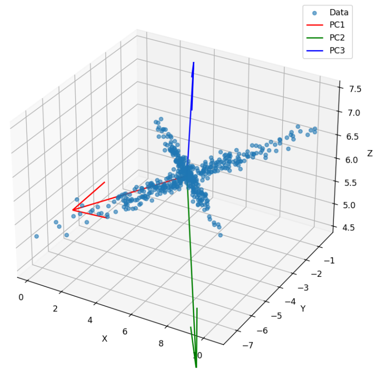
The explained variance ratio (33) provides another algorithm of PCA. Find the first component as a vector, such that projection of the data on it has the highest variance (see Fig. 2b):

$$\lambda_{11} = \max_{\|W_1\|=1} W_1^T \Sigma W_1. \tag{36}$$

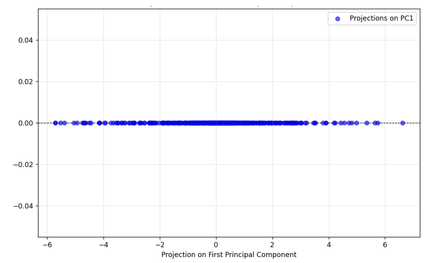
After obtaining the first component W_1, we should find the second component orthogonal to W_1, such that it captures the highest variance of the residuals (see Fig. 2c) $\tilde{x}^{(i)} = x^{(i)} - W_1 z^{(i)}$. This can be formulated as the following maximization problem:

$$\lambda_{22} = \max_{\|W_2\|=1} W_2^T \tilde{\Sigma} W_2, \tag{37}$$

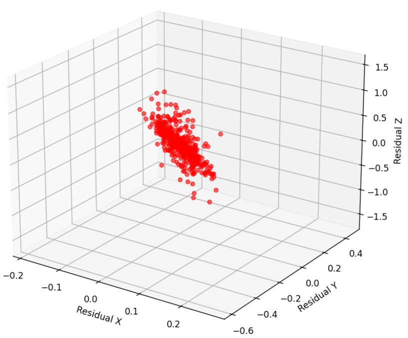
where $\tilde{\Sigma} = \frac{1}{n}\tilde{X}^T\tilde{X}$. And so on.



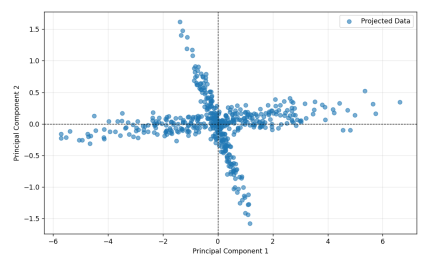
(a) Data with principal components.



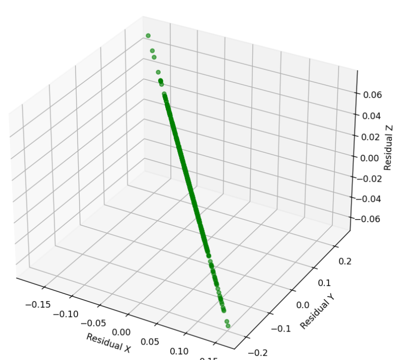
(b) Projection on W_1.



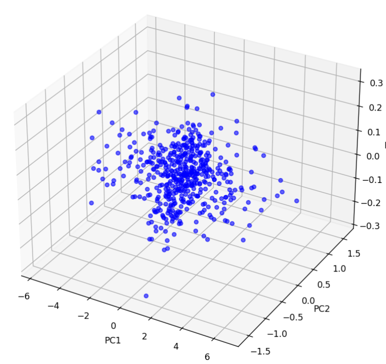
(c) Residuals lie on a plane.



(d) Projection on W_1 and W_2.



(e) Residuals lie on a line.



(f) Projection on W_1, W_2, and W_3.

Figure 2: Illustration to sequential building of PCA.

1.4 Connection with Singular Value Decomposition (SVD)

Each matrix $A \in M(n_1 \times n_2)$ allows the singular value decomposition:

$$\underset{n_1\times n_2}{A} = \underset{n_1\times n_1}{U}\,\underset{n_1\times n_2}{\Sigma}\,\underset{n_2\times n_2}{V^T}, \tag{38}$$

with depending on $n_1 \geq n_2$ or $n_1 \leq n_2$

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_{n_2} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \quad \text{or} \quad \Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_{n_1} & 0 & \cdots & 0 \end{bmatrix}. \tag{39}$$

Where σ_i are eigenvalues of AA^T or A^TA. Columns of U are normalized eigenvectors of AA^T and olumns of V are normalized eigenvectors of A^TA.

Consider the singular value decomposition of the matrix X :

$$X = USV^T. \tag{40}$$

Then

$$\frac{1}{n}X^TX = \frac{1}{n}VSU^TUSV^T = V\left(\frac{1}{\sqrt{n}}S\right)^2 V^T. \tag{41}$$

Comparison with

$$\frac{1}{n}X^TX = W\underset{\ell\times\ell}{\Lambda}W^T \tag{42}$$

demonstrates that variances can be calculated via squared singular values of the matrix X

$$\lambda_{ii} = \frac{\sigma_{ii}^2}{n}, \quad i = 1, \ldots, \ell, \tag{43}$$

and the principle components W are its right singular vectors V.

2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

In some cases PCA doesn't reflect the data structure. For example, in Fig. 3, the projections of the data on the first component overlap. To address this problem, we can use linear discriminant analysis (LDA), but for data visualization the t-distributed stochastic neighbor embedding algorithm (t-SNE) is typically used.

`https://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf`

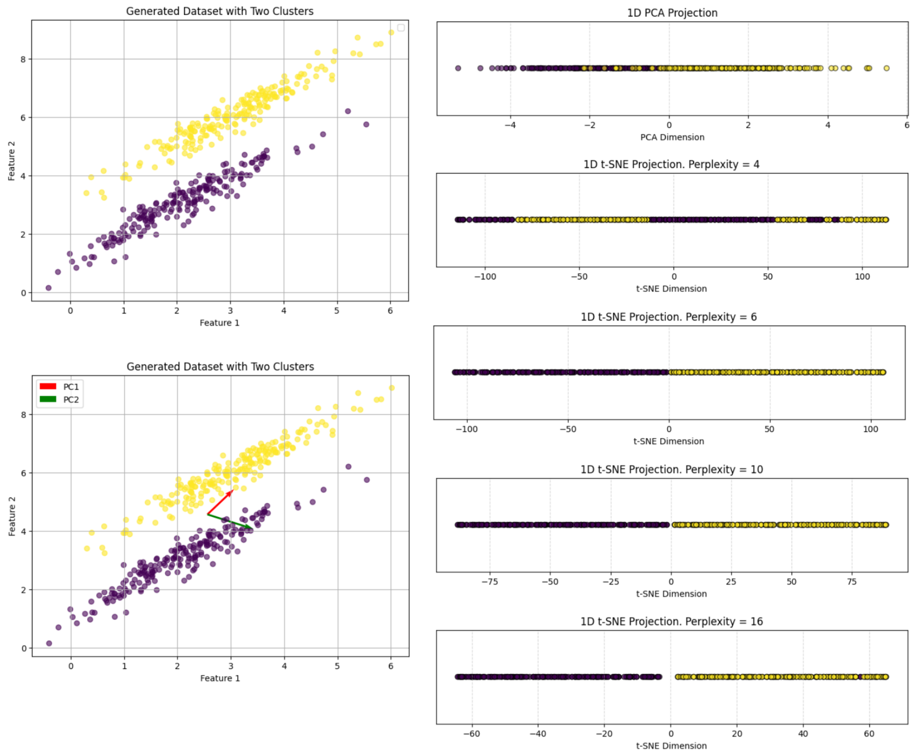


Figure 3: Comparison of PCA and t-SNE

2.1 One Degree of Freedom in t-Distribution

Introduce the distribution on the discrete space of pairwise distances. The easiest way to do it is for the point $x^{(j)}$ to find k nearest neighbors (excluding itself)

$$x^{(j)}, x^{(j_1)}, \ldots, x^{(j_k)}, \quad j_0 = j,$$

and define numbers

$$p_{j|j} = 0, \quad p_{j_1|j} = \ldots = p_{j_k|j} = \frac{1}{k}, \quad p_{j_{k+1}|j} = \ldots = p_{j_{n-1}|j} = 0. \tag{44}$$

Then for each pair $x^{(i)}$ and $x^{(j)}$ define probabilities

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}. \tag{45}$$

It is easy to check that

$$\sum_{i,j=1}^{n} p_{ij} = \frac{1}{2n} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} p_{j|i} + \sum_{j=1}^{n} \sum_{i=1}^{n} p_{i|j} \right) = 1. \tag{46}$$

By *perplexity* we mean

$$Perplexity = 2^H, \tag{47}$$

where H is the Shannon's entropy.

For one point $x^{(j)}$ with probabilities () the perplexity is equal to the number of the neighbors:

$$Perplexity = 2^{-\sum\limits_{i=1}^{n} p_{i|j} \log_2 p_{i|j}} = k. \tag{48}$$

The common choice for the numbers $p_{i|j}$ is

$$p_{i|j} = \frac{e^{-\frac{\left|x^{(i)}-x^{(j)}\right|^2}{2\sigma_j^2}}}{\sum\limits_{k=1}^{n} e^{-\frac{\left|x^{(k)}-x^{(j)}\right|^2}{2\sigma_k^2}}}, \tag{49}$$

where σ_j is chosen such that distribution $p_{i|j}$ has the desired perplexity.[10] The standard choice of perplexity is $5 - 50$.

The corresponding points $z^{(i)}$ in the lower-dimensional space are initialized randomly or using PCA.

The pairwise density is defined as follows:

$$q_{ij} = \frac{1}{Z} \frac{1}{1 + \left|z^{(i)} - z^{(j)}\right|^2}, \tag{50}$$

where Z is the normalizing constant:

$$Z = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \left(1 + \left|z^{(i)} - z^{(j)}\right|^2\right)^{-1}. \tag{51}$$

The positions of the points $z^{(i)}$ are defined by minimization of the Kullback — Leibler divergence

$$KL(P||Q) = \sum_{r,s=1}^{n} p_{rs} \log \frac{p_{rs}}{q_{rs}}, \tag{52}$$

using gradient[11] descent:

$$z^{(i)} = z^{(i)} - \eta \nabla_{z^{(i)}} KL(P||Q), \quad i = 1, \ldots, n. \tag{53}$$

[10] In practice, perplexity is chosen in the range 5—50.

[11] By definition, the gradient is a vector of partial derivatives: $\nabla_{z^{(i)}} F(z^{(i)}) = \begin{bmatrix} \frac{\partial F}{\partial z_1^{(i)}} \\ \vdots \\ \frac{\partial F}{\partial z_\ell^{(i)}} \end{bmatrix}$. We use for it also notation of a matrix derivative $\frac{\partial F}{\partial z^{(i)}}$.

To calculate the gradient with respect to $z^{(i)}$ denote by

$$d_{ij} = \|z^{(i)} - z^{(j)}\|^2 = z^{(i)T} z^{(i)} - 2z^{(i)T} z^{(j)} + z^{(j)T} z^{(j)} \tag{54}$$

and rewrite the minimizing function:

$$\begin{aligned} KL(P||Q) &= \sum_{r,s=1}^{n} p_{rs} \log p_{rs} - \sum_{r,s=1}^{n} p_{rs} \log q_{rs} \\ &= \sum_{r,s=1}^{n} p_{rs} \log p_{rs} - \sum_{r,s=1}^{n} p_{rs} \log(1+d_{rs})^{-1} + \sum_{r,s=1}^{n} p_{rs} \log Z \\ &= \sum_{r,s=1}^{n} p_{rs} \log p_{rs} - \sum_{r,s=1}^{n} p_{rs} \log(1+d_{rs})^{-1} + \log Z. \end{aligned} \tag{55}$$

Then

$$\frac{\partial KL(P||Q)}{\partial z^{(k)}} = \sum_{i,j=1}^{n} \frac{\partial KL(P||Q)}{\partial d_{ij}} \frac{\partial d_{ij}}{\partial z^{(k)}}. \tag{56}$$

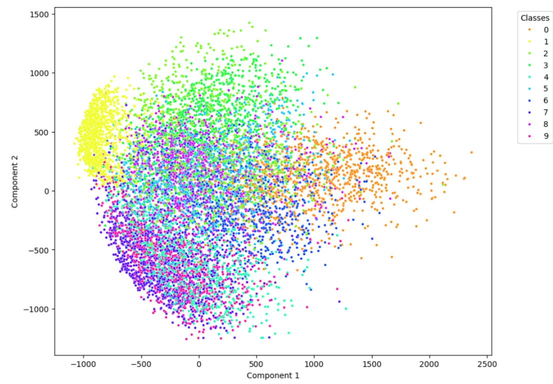
From (54) it follows that

$$\frac{\partial d_{ij}}{\partial z^{(k)}} = \begin{cases} 0, \quad i = j \text{ or } i \neq k, j \neq k, \\ 2z^{(k)} - 2z^{(j)}, \quad i = k, j \neq k, \\ -2z^{(i)} + 2z^{(k)}, \quad i \neq k, j = k. \end{cases} \tag{57}$$

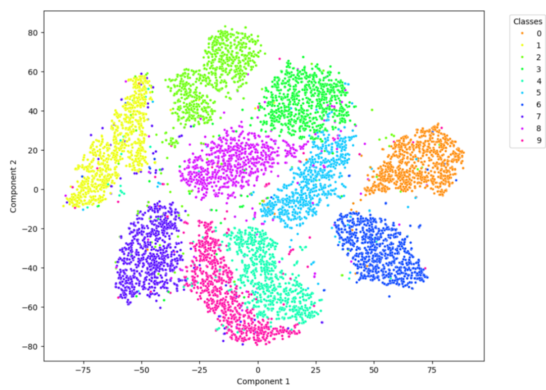
From (56) it follows that

$$\begin{aligned} \frac{\partial KL(P||Q)}{\partial d_{ij}} &= \frac{p_{ij}}{1+d_{ij}} + \frac{1}{Z}\frac{\partial(1+d_{ij})^{-1}}{\partial d_{ij}} = p_{ij}(1+d_{ij})^{-1} - \frac{1}{Z}(1+d_{ij})^{-2} \\ &= \frac{p_{ij}(1+d_{ij})^{-1}}{Z} Z - q_{ij}(1+d_{ij})^{-1} = p_{ij} q_{ij} Z - q_{ij}^2 Z \\ &= (p_{ij} - q_{ij}) q_{ij} Z. \end{aligned} \tag{58}$$

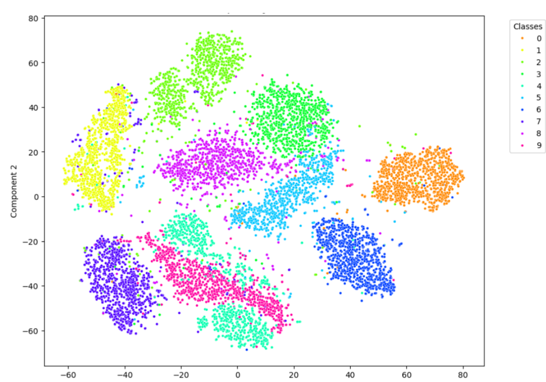
$$\begin{aligned} \frac{\partial KL(P||Q)}{\partial z^{(k)}} &= \sum_{i \neq k} \frac{\partial KL(P||Q)}{\partial d_{ik}} \left(-2z^{(i)} + 2z^{(k)}\right) + \sum_{j \neq k} \frac{\partial KL(P||Q)}{\partial d_{kj}} \left(2z^{(k)} - 2z^{(j)}\right) \\ &= \sum_{i \neq k} (p_{ik} - q_{ik}) q_{ik} Z \left(-2z^{(i)} + 2z^{(k)}\right) + \sum_{j \neq k} (p_{kj} - q_{kj}) q_{kj} Z \left(2z^{(k)} - 2z^{(j)}\right) \\ &= 4 \sum_{i \neq k} (p_{ik} - q_{ik}) q_{ik} Z \left(z^{(k)} - z^{(i)}\right). \end{aligned} \tag{59}$$



(a) PCA.



(b) t-SNE with perplexity 30.



(c) t-SNE with perplexity 50.

Figure 4: Comparison of different dimensionality reduction algorithms on MNIST.

3 Orthogonal Procrustes Problem

This problem arises when we need to align two sets of points using only rotations and reflections. As in PCA we assume the data is centered.

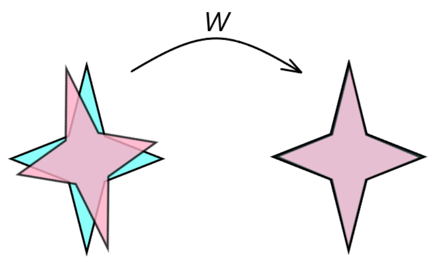


Figure 5: Illustration of the orthogonal alignment

Mathematically we want to find an orthogonal matrix W that maps a set of points X onto the points Y:

$$\begin{cases} L(W) = \sum_{i=1}^{n} \|Wx^{(i)} - y^{(i)}\|^2 = \|WX^T - Y^T\|_F^2 \to \min_W, \\ W^T W = I. \end{cases} \tag{60}$$

Rewrite the target function using trace:

$$\begin{aligned} L(W) &= Tr\left((WX^T - Y^T)^T (WX^T - Y^T)\right) \\ &= Tr\left((XW^T - Y)(WX^T - Y^T)\right) \\ &= Tr\left(XW^T WX^T - XW^T Y^T - YWX^T + YY^T\right) \\ &= Tr(XX^T) - Tr(XW^T Y^T) - Tr(YWX^T) + Tr(YY^T). \end{aligned} \tag{61}$$

Thus, the minimization problem (60) can be reformulated as a maximization problem:[12]

$$\begin{cases} L(W) = Tr(WX^T Y) \to \max_W, \\ W^T W = I. \end{cases} \tag{62}$$

If we consider SVD decomposition of the matrix $X^T Y$

$$X^T Y = U\Sigma V^T, \tag{63}$$

then[13]

$$Tr(WX^T Y) = Tr(WU\Sigma V^T) \leq Tr(V\Sigma V^T) = Tr(\Sigma). \tag{64}$$

Therefore, the matrix W^* delivering the maximum satisfies equation

$$W^* U = V \tag{65}$$

[12]We use the fact $Tr(A^T) = \sum_{i=1}^{m} a_{ii} = Tr(A)$ for a matrix $A \in M(m \times m)$.

[13]Theorem

or

$$W^* = VU^T. \tag{66}$$