Factor and Component Models

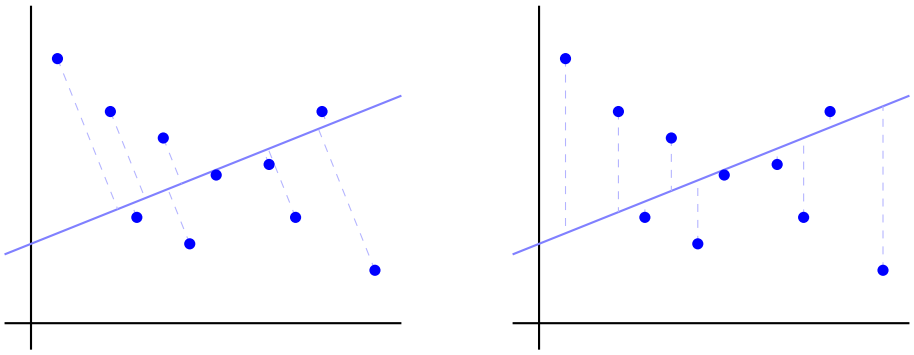
Muchang Bahng

Spring 2025

Contents

Principal component analysis (PCA) and factor analysis (FA) originated independently by Pearson in 1901 and Spearman in 1904 [Pea01, Spe04]. Pearson gave the first formal treatment of it not to compute principal components, but to give a new measure of what a "best fit" line means. On the other hand, Spearman—frustrated by the lack of rigorous analyses on nontrivial in psychology—attempted to model the correlation between mental aptitude and sensory tasks. Though their discoveries were independent, the similarity of their models had inevitably caused their developments to coincide.

Note that PCA is similar to linear regression in that it fits some line (or hyperplane) of best fit to some data. However, linear regression—as a model that tries to use the covariates x to predict the response y—attempts to minimize the *residual* $(y - \hat{y})^2$. If we were to flip the model and try to predict x with y, then the best fit line would not be the same. As Pearson puts it, *the most probable stature of a man with a given length of leg l being s, the most probable length of a leg for a man of stature s will not be l* [Pea01]. This is further motivated by the fact that in many data collecting procedures, you do not collect a perfect measurement of x first and then a noisy measurement of y. Rather, you are usually collecting both x and y together at the same time, in which they may both be perceptive to error.



(a) PCA minimizes the orthogonal distance to the subspace.

(b) Linear regression minimizes the residual distance to the subspace.

Figure 1: Note that this is in fact different from linear regression as it minimizes the expected *orthogonal distance* to the subspace, rather than the residual distance to the subspace as in linear regression.

1 Principal Component Analysis

Say that we have a random vector $x = (x_1, \dots, x_d)$. These d covariates will naturally be correlated, and we want to ask whether some more fundamental set of independent variables exist [Hot33] such that we can express

$$x = f(v_1, \dots, v_k) \tag{1}$$

Naturally, we think of f as a linear function.

We can think of PCA doing two things. First, it is a dimensionality-reduction algorithm where it takes samples $x \in \mathbb{R}^d$ and projects them into some smaller subspace $\mathcal{L}$ of dimension k. Second, it identifies an orthonormal basis of $\mathcal{L}$ that act as uncorrelated low-dimensional features. Because the projection map is linear and we are working in a lower-dimensional subspace, these new basis vectors are linear combinations of the original basis, which may reduce redundancy. Furthermore, by approximately modeling the original x as a linear combination of these features, we are able to get a more parsimonious representation.

In PCA literature, it is more common to work with row vectors $x \in \mathbb{R}^{1\times d}$, so linear mappings are realized through right matrix multiplication xA. Furthermore, we will assume that the data are 0-mean.

1.1 L2 Residual Minimization Approach

To give some motivation, we try to find a best fit line in $\mathbb{R}^d$. A line ℓ can be parameterized by a unit vector u, and so given some sample x, its projection onto ℓ is $\text{proj}_\ell(x) = \langle x, u\rangle u$. Therefore, the residual is

$$\begin{aligned} \|x - \langle x, u\rangle u\|^2 &= \|x\|^2 - 2\langle x, \langle x, u\rangle u\rangle + \|\langle x, u\rangle u\|^2 && (2)\\ &= \|x\|^2 - 2\langle x, u\rangle^2 + \langle x, u\rangle^2\|u\|^2 && (3)\\ &= \|x\|^2 - \langle x, u\rangle^2 && (4)\end{aligned}$$

since $\|u\|^2 = 1$.

Now given a random variable x, our risk is

$$R(u) = \mathbb{E}_x\big[\|x - (x\cdot u)u\|\big] = \mathbb{E}_x\big[\|x\|^2\big] - \mathbb{E}_x\big[\langle x, u\rangle^2\big] \tag{5}$$

In practice, we want to minimize our empirical risk. Assume that we have sampled data $x^{(1)}, \dots, x^{(n)} \sim x$. Then,

$$\begin{aligned} \underset{u\in\mathbb{R}^d, \|u\|=1}{\text{argmin}}\ \hat{R}(u) &= \underset{u\in\mathbb{R}^d, \|u\|=1}{\text{argmin}}\ \frac{1}{n}\left(\sum_{i=1}^n \|x^{(i)}\|^2 - \sum_{i=1}^n \langle x^{(i)}, u\rangle^2\right) && (6)\\ &= \underset{u\in\mathbb{R}^d, \|u\|=1}{\text{argmax}}\ \frac{1}{n}\sum_{i=1}^n \langle x^{(i)}, u\rangle^2 && (7)\end{aligned}$$

We have our loss function! Now what if we wanted to look for best fitting subspaces in general? Let's first rigorously define such a space.

Definition 1.1 (Principal Subspace)

Let x be a 0-mean random variable in $\mathbb{R}^d$ and let $\mathcal{L}^k$ denote all k-dimensional linear subspaces of $\mathbb{R}^d$. The k**th principal subspace** is defined as

$$\ell_k = \underset{\ell\in\mathcal{L}_k}{\text{argmin}}\, \mathbb{E}_{\tilde{x}}\big[\|x - \text{proj}_\ell\, x\|_2\big] \tag{8}$$

This isn't a big step from what we had before. We just want to construct a subspace ℓ that minimizes the expected L^2 distance between x and ℓ. Now how do we do such a thing? The most natural extension would

be to identify an orthonormal basis $u_1, \dots, u_k$, and since

$$\text{proj}_\ell \, x = \sum_{i=1}^{k} \text{proj}_{u_i} \, x \tag{9}$$

our loss can be simplified to

$$\begin{aligned} R(\ell) = R(u_1, \dots, u_k) &= \mathbb{E}\left[\|x - \text{proj}_\ell \, x\|^2 \right] && (10) \\ &= \mathbb{E}\left[\|x\|^2 - 2\langle x, \sum_{i=1}^{k} \text{proj}_{u_i} \, x \rangle + \| \sum_{i=1}^{k} \text{proj}_{u_i} \, x\|^2 \right] && (11) \\ &= \mathbb{E}\left[\|x\|^2 - 2\sum_{i=1}^{k} \langle x, \text{proj}_{u_i} \, x \rangle + \sum_{i=1}^{k} \| \text{proj}_{u_i} \, x\|^2 \right] && (12) \\ &= \mathbb{E}\left[\|x\|^2 - 2\sum_{i=1}^{k} \langle x, u_i \rangle^2 + \sum_{i=1}^{k} \langle x, u_i \rangle^2 \|u_i\|^2 \right] && (13) \\ &= \mathbb{E}\left[\|x\|^2 - \sum_{i=1}^{k} \langle x, u_i \rangle^2 \right] && (14) \end{aligned}$$

and in the empirical case, we can get rid of the fixed x and find

$$\underset{u_i \in \mathbb{R}^d}{\text{argmax}} \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{k} \langle x^{(j)}, u_i \rangle, \text{ subject to } \|u_i\|^2 = 1, \langle u_i, u_j \rangle = 0 \text{ for } i \neq j \tag{15}$$

By stacking the u_i's left-to-right in matrix $U \in \mathbb{R}^{d \times k}$, we can get a cleaner form of the loss function.

Theorem 1.1 (Constrained Empirical Risk of PCA)

The empirical risk, or loss function, of PCA is

$$\underset{U \in \mathbb{R}^{d \times k}, U^T U = I_k}{\text{argmax}} \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{k} \langle x^{(j)}, u_i \rangle \tag{16}$$

or equivalently,

$$\underset{U \in \mathbb{R}^{d \times k}, U^T U = I_k}{\text{argmax}} \frac{1}{n} \|X - XUU^T\|^2 \tag{17}$$

1.2 Variance Maximization Approach

But we can turn this into a variance maximization problem. Note that $\text{Var}_x[\langle x, u \rangle] = \mathbb{E}[\langle x, u \rangle^2] - \mathbb{E}[\langle x, u \rangle]^2$, and so we can rewrite our true risk as

$$\underset{u \in \mathbb{R}^d, \|u\|=1}{\text{argmin}} \; R(u) = \underset{u \in \mathbb{R}^d, \|u\|=1}{\text{argmin}} \; \mathbb{E}_x\left[\|x\|^2\right] - \text{Var}_x[\langle x, u \rangle] - \mathbb{E}[\langle x, u \rangle]^2 \tag{18}$$

where the last term vanishes since x is 0-mean, and hence by linearity of expectation $\mathbb{E}_x[\langle x, u \rangle] = \langle \mathbb{E}[x], u \rangle = \langle 0, u \rangle = 0$. In parallel the empirical risk reduces to simply the sample variance.

$$\underset{u \in \mathbb{R}^d, \|u\|=1}{\text{argmax}} \; \hat{\text{Var}}[\langle x, u \rangle] = \underset{u \in \mathbb{R}^d, \|u\|=1}{\text{argmax}} \; \frac{1}{n} \left(\sum_{i=1}^{n} \langle x^{(i)}, u \rangle^2 \right) \tag{19}$$

Therefore, we can think of the L^2 minimization problem as equivalent to a variance maximization approach.

Lemma 1.1 (Variance Maximization Approach)

Minimizing the L^2 distance of a random variable x to a line ℓ in $\mathbb{R}^d$ is equivalent to maximizing the scalar variance in the projected space.

$$\underset{u\in\mathbb{R}^d,\|u\|=1}{\operatorname{argmin}} \mathbb{E}[\|x - \operatorname{proj}_u(x)\|_2] = \underset{u\in\mathbb{R}^d,\|u\|=1}{\operatorname{argmax}} \operatorname{Var}_x[\langle x, u\rangle] \tag{20}$$

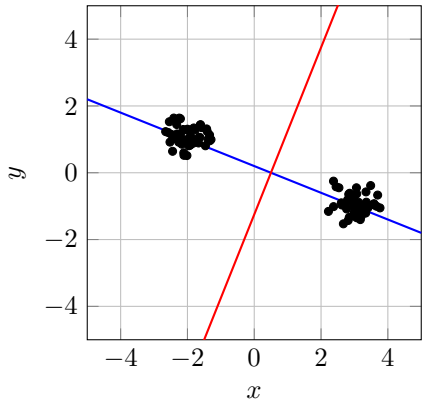


Figure 2: Projecting the dataset onto the blue line seems to retain more variance than projecting onto the red line.

Let's in fact try to directly maximize the variance. If we vertically stack our n data points into a matrix $X \in \mathbb{R}^{n\times d}$, then the projections are simply Xu. Again, since this is 0-mean, the variance is

$$\hat{\operatorname{Var}}(Xu) = \frac{1}{n}(Xu)^T(Xu) \tag{21}$$
$$= \frac{1}{n}u^T X^T X u \tag{22}$$
$$= u^T \frac{X^T X}{n} u \tag{23}$$
$$= u^T \hat{\Sigma} u \tag{24}$$

where $\hat{\Sigma}$ is the empirical covariance matrix of X. We want to find

$$\max_u u^T \hat{\Sigma} u \text{ subject to } \|u\|^2 = 1 \tag{25}$$

This is a classic Lagrange multiplier problem. We construct the Lagrangian and compute its partial derivatives to set equal to 0.

$$\mathcal{L}(u, \lambda) = u^T \hat{\Sigma} u - \lambda(\|u\|^2 - 1) \tag{26}$$
$$\frac{\partial \mathcal{L}}{\partial u} = 2\hat{\Sigma}u - 2\lambda u = 0 \tag{27}$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = u^T u - 1 = 0 \tag{28}$$

which gives us

$$\hat{\Sigma}u = \lambda u, \qquad u^T u = 1 \tag{29}$$

This tells us that u is a unit eigenvector, and the maximizing vector will be the one corresponding to the largest eigenvalue. Essentially, we have reduced this to an eigenvalue problem.

Theorem 1.2 (Principal Component as Eigenvector)

The first principal subspace of data matrix $X \in \mathbb{R}^{n \times d}$ is spanned by the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix $\hat{\Sigma} = \frac{1}{n} X^T X$.

Now for higher dimensional subspaces, we take the same approach. Going through the same derivation gives the expected risk in terms of the variance

$$R(u_1, \dots, u_k) = \mathbb{E}[\|x\|^2] - \sum_{i=1}^{k} \mathbb{E}[\langle x, u_i \rangle^2] \tag{30}$$

$$= \mathbb{E}[\|x\|^2] - \sum_{i=1}^{k} \text{Var}[\langle x, u_i \rangle] - \mathbb{E}[\langle x, u_i \rangle]^2 \tag{31}$$

and by fixing the x's, we get our equivalent empirical risk.

Theorem 1.3 (Empirical Risk of PCA as Variance-Maximizer)

The empirical risk tells us to find an orthonormal basis that maximizes the sum of the variance of projections.

$$\underset{u_i \in \mathbb{R}^d}{\text{argmax}} \sum_{i=1}^{k} \hat{\text{Var}}[\langle x, u_i \rangle] \text{ subject to } \|u_i\|^2 = 1, \langle u_i, u_j \rangle = 1 \text{ for } i \neq j \tag{32}$$

1.3 Solving PCs with Singular Value Decomposition

The variance-maximization loss is very insightful, and we may naively think of just taking the unit eigenvectors corresponding to the top k largest eigenvalues. Surprisingly, this greedy approach turns out to be correct.

Theorem 1.4 (2nd Principal Component is 1st Principal Component of Residuals)

Theorem 1.5 (Construction of the kth Principle Subspace)

Let $X \in \mathbb{R}^{n \times d}$ be a 0-mean data matrix. Given the SVD with the singular values listed in decreasing order[a]

$$X = U \Sigma V^T, \qquad U \in \mathbb{R}^{n \times n}, \Sigma \in \mathbb{R}^{n \times d}, V \in \mathbb{R}^{d \times d} \tag{33}$$

[a] We can make it decreasing by permuting the rows/columns of the unitary matrices U, V.

Definition 1.2 (Terminology)

The terminology

1. The columns $v_1, \dots, v_d$ of V are called the **principal component axes**. $\ell_k = \text{span}\{v_1, \dots, v_k\} \subset \mathbb{R}^d$, i.e. the subspace spanned by the columns of V. They are the right singular vectors or the eigenvectors of the covariance matrix living in $\mathbb{R}^d$.
2. The columns of $U\Sigma$ are called the **principal component scores**.

Now let's take the first k principal component axes $v_1, \dots, v_k$ and denote the truncated matrix as V_k. Then,

1. $XV_k \in \mathbb{R}^{n \times k}$ is the projection of the dataset into $\mathbb{R}^k$ spanned by the principal components.

1. The projection P is defined

$$\hat{x} = P_k(x) = \mu + \sum_{j=1}^{k} \langle x - \mu, v_j \rangle \, v_j = \sum_{j=1}^{k} \text{proj}_{v_j}(x - \mu) = \mu + \text{proj}_{\ell_k}(x - \mu) \tag{34}$$

   where we can rewrite it as the projection operator since the v_j's are orthonormal.
2. The change of basis T is defined with the mapping $\hat{x} \in \mathcal{L}_k \mapsto \sigma_j v_j \in \mathcal{L}_k$. Note that the v_j's form an orthogonal basis of $\mathcal{L}_k$.

Now let $V_k \in \mathbb{R}^{d \times k}$ represent the first k columns of V (aka first k principal axes), $U_k \in \mathbb{R}^{n \times k}$ represent the first k columns of U, and $\Sigma_k \in \mathbb{R}^{k \times k}$ represent the upper-left $k \times k$ matrix of Σ.[a] The product $U_k \Sigma_k$ represents the matrix containing the first k principal components. The matrix $\tilde{X}_k = U_k \Sigma_k V^T$, which is the low-rank approximation of $\tilde{X}$, is called the **denoised matrix** of $\tilde{X}$.

[a] Note that V^T, which was originally surjective, is now just injective.

Proof.

For notational convenience let $X = \tilde{X}$. We see that

$$X^T X = V \Sigma^T \Sigma V^T \tag{35}$$

Note that $X \neq V \Sigma^T$ in general. Now let $v_1, \dots, v_d$ be the columns of V. Then

$$X^T X [v_1, \dots, v_d] = X^T X V = V \Sigma^T \Sigma = [\sigma_1^2 v_1, \dots, \sigma_d^2 v_d] \tag{36}$$

Therefore, we can see that the way $X^T X$ acts on V That the v_i's are the eigenvectors of $X^T X$, with σ_i^2 the associated eigenvalues.

Let's take a few moments to appreciate what U and V really represent. In some sense, $U_k \in \mathbb{R}^{n \times k}$ can be considered the dimension-reduced form of $\tilde{X} \in \mathbb{R}^{n \times d}$. To see why consider the following. Let's label the *rows* of U_k as $u^{(1)}, \dots, u^{(n)} \in \mathbb{R}^k$. By transposing the equation of the denoised matrix, we get $\tilde{X}_k^T = V_k \Sigma_k U_k^T$, and so

$$x^{(i)} - \mu = V_k \Sigma_k u^{(i)} \tag{37}$$

for $i = 1, \dots, n$. As an immediate consequence, since T^{-1} maps e_j to $\sigma_j v_j$, we can interpret $U_k \Sigma_k V_k^T$ with the decomposition

$$\overbrace{u^{(i)} \in \mathbb{R}^k \xrightarrow{\Sigma_k} \Sigma_k u^{(i)} \in \mathbb{R}^k \xrightarrow{V_k} \overbrace{\hat{x} \in \mathcal{L}_k \xleftarrow{P} x \in \mathbb{R}^d}^{P}}^{T^{-1}}$$

This is very revealing. To embed the low-rank $u^{(i)}$ representation, it must go through some scaling Σ_k followed by the injective map V_k. Now let's interpret V_k and consider its *columns*, labeled $v_1, \dots, v_k \in \mathbb{R}^d$. These represent the basis vectors that span the subspace $\mathcal{L}_k$, i.e. the upscaled features in the higher-dimensional space. Therefore, V_k represents the injection $e_i \in \mathbb{R}^k \mapsto v_i \in \mathcal{L}_k \subset \mathbb{R}^d$. This means that if we would like to pick a point with some combination of these features, we are really picking a point

$$z = \sum_i z_i v_i \in \mathcal{L}_k \tag{38}$$

Algorithm 1.1 (Fitting)

Given a dataset $X \in \mathbb{R}^{n \times d}$, let us denote the rows as x_i, and say that we are looking for a subspace of dimension k.

1. Compute the mean
$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \in \mathbb{R}^d \tag{39}$$
2. Standardize the data $\tilde{X} = X - \mu$, i.e. $\tilde{x}_i = x_i - \mu$.
3. Compute the SVD $\tilde{X} = U \Sigma V^T$.
4. Compute the submatrices $V_k \in \mathbb{R}^{k \times k}$ and $\Sigma_k \in \mathbb{R}^{D \times k}$.
5. Define the projection operator $P_k(x) = \mu + \sum_{j=1}^{k} \langle x - \mu, v_j \rangle \, v_j$, the change of basis operator T, and the embedding operator $T^{-1}(z) = \mu + V_k \Sigma_k z$.

A demonstration is done here.

Example 1.1 (Eigenfaces)

In 1991, Turk and Pentland presented an eigenface method of face recognition by taking the low-rank approximation of a dataset of face images.

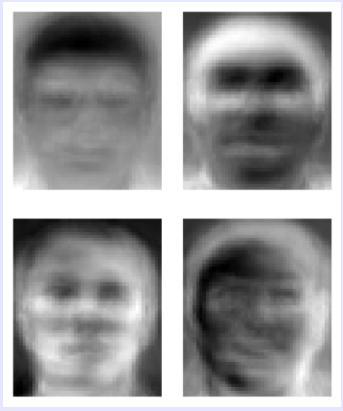


Figure 3: Some eigenfaces from AT&T Labs.

Now a question arises: how do we know that this sample decomposition is a good approximation to the true decomposition? It comes from the fact that the sample covariance $\hat{\Sigma}$ is a good approximation of the true covariance Σ, which we will later prove using concentration of measure.

Theorem 1.6 (Risk)

The risk satisfies
$$R(k) = \mathbb{E}[||x - P_k(x)||^2] = \sum_{j=k+1}^{D} \lambda_j \tag{40}$$

It is essential that you plot the spectrum in decreasing order. This allows you to analyze how well PCA is working. People often use the "elbow" technique to determine where to choose K, and we value
$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{d} \lambda_j} \tag{41}$$
accounts for the **variance explained**, which should be high with K low. If you have to go out to dimension $K = 50$ to explain 90% of the variance, then PCA is not working. It may not work because of many reasons,

such as there being nonlinear structure within the data.

It turns out that the elements of $\hat{\Sigma}$ are close entry-wise to those of Σ. But if this is true, then does it mean that the eigenvalues of the sample covariance matrix are close to the true eigenvalues of the covariance matrix? It turns out that the answer is no, and we need a proper metric to satisfy this assumption. The metric, as we can guess from linear algebra, is the operator norm, and we will show some results from matrix perturbation theory.

Lemma 1.2 ()

It turns out that

$$||\hat{\Sigma} - \Sigma|| = O_p\left(\frac{1}{\sqrt{n}}\right) \tag{42}$$

where $||\cdot||$ is the operator norm.

Theorem 1.7 (Weyl's Theorem)

If $\hat{\Sigma}$ and Σ are close in the operator norm, then their eigenvalues are close.

$$||\hat{\Sigma} - \Sigma|| = O_p\left(\frac{1}{\sqrt{n}}\right) \implies |\hat{\lambda}_j - \lambda_j| = O_p\left(\frac{1}{\sqrt{n}}\right) \tag{43}$$

This only talks about their eigenvalues, but this does not necessarily imply that the eigenvalues are close. We need an extra condition.

Theorem 1.8 (David-Kahan Theorem)

If $\hat{\Sigma}$ and Σ are close in the operator norm, and if the eigenvectors of Σ are well-conditioned, then the eigenvectors of $\hat{\Sigma}$ are close to the eigenvectors of Σ. More specifically,

$$||\hat{v}_j - v_j|| \leq \frac{2^{3/2}||\hat{\Sigma} - \Sigma||}{\lambda_j - \lambda_{j+1}} \tag{44}$$

1.4 Iterative Methods

1.5 Old

To begin with some motivation, let a linear map $A : \mathbb{R}^D \to \mathbb{R}^D$ be full rank, which maps some set of n data points to the space of features. Then it is injective, and therefore for all data $x \in \mathbb{R}^D$ there exists a feature vector $z \in \mathbb{R}^D$ such that $z = Ax$. Generally, real-world data does not span the full space of D dimensions.[1] In fact, if we further assume that the data lies in a linear subspace, we want to compress it into a lower-dimensional vector such that the covariates in this lower dimensional space are also orthogonal, i.e. uncorrelated. We tackle both problems in 2 steps.

1. To compress this representation, we can take a data point $x \in \mathbb{R}^D$ and *approximate* it as a point $\hat{x} \in L_k$ for some k-dimensional subspace $L_k \subset \mathbb{R}^d$ (say that this is done with some function $P : \mathbb{R}^D \to L_k \subset \mathbb{R}^D$).
2. After this projection, we then want to extract the k features such that they are *orthogonal* (i.e. no correlation). This is done with a simple change of basis, which we denote $T : L_k \to \mathbb{R}^k$, giving us $\hat{z} = T\hat{x} = T(P(x))$. We can invert this map $T^{-1} : \mathbb{R}^k \to L_k$ to go from the orthogonalized compressed version $\hat{z}$ to the approximate full version $\hat{x}$.

[1] The *manifold hypothesis* that real-world data in high-dimensions actually lies on a lower-dimensional manifold.

We can see that by definition the properties of the principal subspace allows us to construct the best approximation of the points in a lower-dimensional subspace. This seems like a hard optimization problem, but it turns out that the theorem gives a simple solution. Note that we need to do 3 things:

1. Find such a subspace $\mathcal{L}_k \subset \mathbb{R}^D$.
2. Find the projection $P_k : \mathbb{R}^D \to \mathcal{L}_k \subset \mathbb{R}^D$. Note that by definition of the principal subspace P_k should be an *orthogonal* projection.
3. Find the bijection $T_k : \mathcal{L}_k \to \mathbb{R}^k$.

2 Factor Analysis

Note that in PCA, we have taken some data x in high-dimension D and reduced it to a lower-dimensional orthogonal representation in $\mathbb{R}^k$. In other words, the corresponding z represents x in another space, which we call a **latent space**. A model that represents data from the original space $\mathcal{X}$ to a latent space $\mathcal{Z}$ is called a *latent variable model*. We will extend on this.

Say that we have some covariates $x^{(i)} \sim X$ and we want to find its true distribution p^*. In density estimation so far, what we have done is define a family of distributions $\{p_\theta\}$ and optimize the loss by maximizing the MLE or something else.

$$\min_\theta L(p_\theta, p^*) = \max_\theta \prod_i p_\theta(x^{(i)}) \tag{45}$$

In order to do this we work with explicitly parameterized distribution families (e.g. Gaussian, Gamma, multinomial, etc.), but this is too simple to model complex things in real like (e.g. the distribution of faces). Therefore, we consider *implicitly parameterized* probability distributions by "adding" a latent distribution Z, creating the joint distribution (X, Z). This may look more complicated, but it captures a much richer family of distributions.

Definition 2.1 (Generative Latent Variable Model)

A **latent variable model** is a model of a distribution $p^*(x)$ over a space $\mathcal{X}$ using implicitly parameterized probability distributions p_θ constructed as such:

1. We define a simple random variable Z over $\mathcal{Z}$ with its distribution $p(z)$, called the **prior**.[a]
2. We define a family of functions $\{f_\theta\}$ defined over z and parameterized by θ.
3. We define a way to convert any $f_\theta(z)$ into a distribution $p(x \mid z)$, called the **likelihood** or **generative component**. There are generally two ways to do this:
   (a) Let the random variable $X \mid Z = z$ be an explicitly parameterized distribution, and have $f_\theta(z)$ be the parameters of $X \mid Z = z$. Therefore, we take the output of $f_\theta(z)$ and plug in these values as the parameters of $X \mid Z = z$.[b]
   (b) Have f_θ be a transformation of random variables, i.e. $X = f(Z)$. This may result in a conditional pdf that is not explicitly parameterizable.

This defines the family of joint distributions p_θ over $(\mathcal{Z}, \mathcal{X})$. It is easy to sample $(x, z) \sim p_\theta$: sample $z \sim p$, then compute $f_\theta(z)$, use this to define $p_\theta(x \mid z)$, and finally sample from the likelihood. Therefore, the joint is also of a simple nature.

While we assume simple, explicitly parameterized forms for the prior and the likelihood, we do not assume anything about

1. the **marginal** $p_\theta(x)$. Usually this is an extremely complicated distribution, which is equivalent to
$$p_X(x) = \int_{z \in \mathbb{R}^k} p(x \mid z)\, p_Z(z)\, dz = \mathbb{E}_Z[p(X \mid Z)] \tag{46}$$
from marginalizing but is computationally impossible to integrate.
2. the **posterior** $p_\theta(z \mid x)$ that describes the hidden features given some data point. This is also known as the **inference component**. By Bayes rule, we have
$$p_\theta(z \mid x) = \frac{p_\theta(x \mid z)\, p(z)}{p_\theta(z \mid x)} \iff p_\theta(z \mid x) \propto p_\theta(x \mid z)\, p(z) \tag{47}$$
which we might be able to sample from using MCMC.

[a] Almost always a uniform or normal distribution suffices. If not, we can constrain it to be factorable (i.e. is the product of its marginal distributions: $p(z) = \prod_i p(z_i)$) so that it is easy to sample from. Occasionally, the stronger assumption of the z_i's being iid is made.

[b] For example, let $f_\theta(z) = (f_1(z), f_2(z))$. Then we define the corresponding distribution $X \mid Z = z \sim \mathcal{N}(f_1(z), e^{f_2(z)})$.

Like we do with everything else in math, we take a look at the simplest example: when the class $\{f_\theta\}$ are

linear functions that represent *transformations*[2] $X = f(Z)$ of the random variable Z. This is known as **linear latent variable modeling**.

$$X = \mu + WZ + \epsilon \tag{48}$$

where the noise ϵ is typically Gaussian and diagonal (but not necessarily the same component-wise variances). Finally, we can use techniques like MLE to estimate W, μ, and the parameters of ϵ. The entire reason we want to do this is that we are hoping that we can construct a complex distribution X from a simple distribution Z with $d >> k$, connected by some well-studied function $X = f(Z)$. In the linear case, $W \in \mathbb{R}^{d \times k}$, and the latent variables z give a more compact, parsimonious explanation of dependencies between the components of the observations x.

Definition 2.2 (Factor Analysis)

Factor analysis is a specific case of a linear latent variable model where

$$X = \mu + WZ + \epsilon, \text{ where } z \in \mathcal{N}(0, I), \epsilon \sim \mathcal{N}\big(0, \text{diag}(\sigma_1^2, \dots, \sigma_k^2)\big) \tag{49}$$

It should be clear to us that X should be Gaussian[a] and that $\mathbb{E}[X] = \mu$, with

$$\begin{aligned}
\text{Var}[X] &= \mathbb{E}[(X-\mu)(X-\mu)^T] && (50)\\
&= \mathbb{E}[(WZ+\epsilon)(Z^TW^T+\epsilon^T)] && (51)\\
&= \mathbb{E}[WzZ^TW^T] + \mathbb{E}[\epsilon\epsilon^T] && (52)\\
&= W\mathbb{E}[zz^T]W^T + \mathbb{E}[\epsilon\epsilon^T] && (53)\\
&= WW^T + \text{diag}(\sigma_1^2, \dots, \sigma_d^2) && (54)
\end{aligned}$$

The W, μ, and σ_i's can be estimated using MLE methods.

[a] Since linear transformations of Gaussians are Gaussian

[2] Not have its output parameterize $X \mid Z = z$.

3 Robust PCA

4 Sparse PCA

5 Dynamic PCA

6 Functional PCA

7 Kernel PCA

Definition 7.1 (Kernel PCA)

Let N_i be the neighborhood around X_i. Then, we want to find a mapping $W : \mathbb{R}^n \to \mathbb{R}^k$ that minimizes

$$\min_W \sum_{i=1}^n \left|\left| X_i - \sum_{j \in N_i} W_{ij} X_j \right|\right|^2 \text{ where } \sum_j W_{ij} = 1 \tag{55}$$

We can constrain the weights in W so that anything that is not in the neighborhoods are 0.

8 Group PCA

9 Probabilistic PCA

We want to take PCA and extend it to be a *generative model*, which allows you to sample data. In regular PCA, we saw that for some $z \in \mathbb{R}^k$ in the latent space, $\hat{x} = \mu + V_k \Sigma_k z$. Therefore, if we just change z from a point to a probability distribution (e.g. Gaussian), we can take a random variable $z \sim \mathcal{N}(0, I)$ from $\mathbb{R}^k$, and then transform it to get a random variable $x = \mu + U_k \Sigma_k z$, which will give a density.

$$x \sim \mathcal{N}\big(\mu, (V_k\Sigma_k)(V_k\Sigma_k)^T\big) = \mathcal{N}\big(\mu, V_k\Sigma_k U_k^T U_k \Sigma_k V_k^T\big) = \mathcal{N}\big(\mu, X_k^T X_k\big) \tag{56}$$

Note that in here, x is a random variable that we are trying to fit to the data X_k. However, $X_k \in \mathbb{R}^{n \times d}$ with $d << n$, and so $X_k^T X_k \in \mathbb{R}^{d \times d}$ is not full rank, and so the distribution is restricted to strictly the k-dimensional subspace $L_k \subset \mathbb{R}^D$. We want to add a bit of noise beyond the subspace, so we add an extra small Gaussian ϵ around it. In general factor analysis above, we set ϵ to have an arbitrary diagonal Gaussian, but for PPCA we just use an isotropic one $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, giving us

$$x = \mu + U_k \Sigma_k z + \epsilon \implies X \sim \mathcal{N}(\mu, X_k^T X_k + \sigma^2 I) \tag{57}$$

Now rather than treating X_k^T as a data matrix that we use to calculate the principal subspace, we treat it as a parameter matrix $W \in \mathbb{R}^{d \times n}$ that we want to fit [TB99]. Note that PPCA is really a specific instance of factor analysis, and we assume that the latent variable z follows a standard Gaussian $\mathcal{N}(0, 1)$.

Definition 9.1 (Probabilistic PCA)

The **probabilistic PCA** model is a latent factor model with $Z \sim \mathcal{N}(0, I)$ and

$$X = f_\theta(Z) = \mu + (WW^T + \sigma^2 I)^{1/2} Z \tag{58}$$

and $\theta = \{\mu, W, \sigma\}$, which gives

$$X \sim \mathcal{N}(\mu, WW^T + \sigma^2 I) \tag{59}$$

Optimizing this model is actually quite easy.

Theorem 9.1 (MLE of PPCA Model)

Given $x^{(i)} \sim X$ iid, the MLEs for W, μ, σ are

$$\mu_{MLE} = \frac{1}{N} \sum_{i=1}^N x^{(i)} \implies \hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^N x^{(i)} \tag{60}$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{d-k} \sum_{j=k+1}^d \lambda_j \tag{61}$$

$$W_{MLE} = U_q (\Lambda_d - \hat{\sigma}^2_{MLE} I_d)^{1/2} R \tag{62}$$

Proof.

Given $x^{(i)} \sim X$ iid, the MLEs for W, μ, σ have a closed form, and model parameter estimation can be performed iteratively and efficiently. We have

$$\mu_{MLE} = \frac{1}{N} \sum_{i=1}^N x^{(i)} \implies \hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^N x^{(i)} \tag{63}$$

and setting the biased MLE estimator of the variance,

$$\widehat{\mathrm{Var}}_{MLE}(\mu_{MLE}) = S = \frac{1}{N} \sum_{i=1}^N (x^{(i)} - \mu_{MLE})(x^{(i)} - \mu_{MLE})^T \tag{64}$$

we can derive the MLE of W.[a] We can find the MLE estimate of σ first by taking a look at $C = \text{Var}[X] = WW^T + \sigma^2 I$. It is the sum of positive semidefinite patrices that are also symmetric, so by the spectral theorem it is diagonalizable and has full rank d. But WW^T is rank k, so $d-k$ of the eigenvalues of WW^T is 0, indicating that the same $d-k$ smallest eigenvalues of C is σ^2. Therefore, we can take the smallest $d-k$ eigenvalues of our MLE estimator of C, which is S, and average them to get our MLE for σ.

$$\hat{\sigma}^2_{MLE} = \frac{1}{d-k} \sum_{j=k+1}^{d} \lambda_j \tag{65}$$

We can approximate $WW^T = C - \sigma^2 I \approx S - \hat{\sigma}^2_{MLE} I$, and by further taking the eigendecomposition $C = U\Sigma U^T \implies WW^T = U(\Sigma - \sigma^2 I)U^T$ and cutting off the last $d-k$ smallest eigenvalues and their corresponding eigenvectors, we can get

$$W_{MLE} = U_q(\Lambda_d - \hat{\sigma}^2_{MLE} I_d)^{1/2} R \tag{66}$$

where the R just accounts for any unitary matrix.

[a] Note that W_{MLE} is not unique. Say that W^* is an MLE, then, for any unitary $U \in \mathbb{R}^{k \times k}$, we have $W^* W^{*T} = (W^* U)(W^* U)^T$.

Now as $\sigma \to 0$, the density model defined by PPCA becomes very sharp around these d dimensions spanned by the columns of W. At 0, our MLE of W is simplified and we have

$$X = W_{MLE} z + \mu_{MLE} + \epsilon = U_q \Lambda_q^{1/2} z + \mu_{MLE} \tag{67}$$

which essentially reduces to regular PCA. That is, the conditional expected value of z given X becomes an orthogonal projection of $X - \mu$ onto the subspace spanned by the columns of W. Intuitively, we can see that we are estimating the Gaussian, which corresponds to the mean squared distance from each $x^{(i)}$ to ℓ_k.

10 Linear Independent Component Analysis

ICA is a method to separate a multivariate signal into additive, statistically independent components. It does come with a lot of assumptions, and is a specific instance of a linear factor model where $\mu = 0$ and $\epsilon = 0$.

Definition 10.1 (Linear ICA)

In **linear ICA**, we have the simple model.

$$x = Wz \tag{68}$$

In here, $X \in \mathbb{R}^d$ is a mixture vector and $W \in \mathbb{R}^{d \times k}$ is a **mixing matrix**. Both W and z are unknown, and we need to recover them given x. We have 2 strong assumptions.

1. Each component of z is independent (not just uncorrelated). This is an easy enough assumption to intuit.
2. Independent components of z must *not* be Gaussian.[a]

[a] This is needed for us to be able to "unmix" the signals. To see why, just suppose z was Gaussian, and so the vector Rz is also Gaussian for any invertible R. Therefore, we could find an infinite number of solutions of form $x = WR^{-1}Rz$ and have no way to separate them.

Algorithm 10.1 (Fitting)

Now let's see how linear ICA actually estimates W and z. Once W is estimated, the latent components of a given test mixture vector, x^* is computed by $z^* = W^{-1}x^*$. So now all there's left to do is to estimate W, which we want to estimate so that $W^{-1}x$ is far from Gaussian. The reason for this is that given a bunch of independent non-Gaussian h_i's, if we mix them with a matrix that is not $\pm I$, then by CLT, a linear combination of random variables will tend to be Gaussian, and so for an arbitrary W we would expect x to be Gaussian. Therefore, what we want to do is guess some matrix A, and compute

$$Ax = AWh \tag{69}$$

and if we get things right, $A \approx W^{-1}$, and the result of Ax would look pretty non-Gaussian. If it it not the case, then AW will still be some mixing matrix, and so Ax would look Gaussian. So now the question reduces to how do we choose this A? There are multiple ways to measure non-Gaussianity:

1. The absolute or squared kurtosis, which is 0 for Gaussians. This is a differentiable function w.r.t. W, so we can try maximizing it. This is done for the sample kurtosis, of course.
2. Another measure is by maximizing the neg-entropy.

There are further ambiguities with ICA regarding uniqueness of a best representation. For one, we can only estimate the latent components up to a scaling factor since we will still get

$$x = (\alpha W)(\frac{1}{\alpha}z) \text{ for some } \alpha > 0 \tag{70}$$

We can fix this by forcing $\mathbb{E}[z_i^2] = 1$. However, there is still an ambiguity for the sign of hidden components, but this is insignificant in most applications. Second, we can estimating the components up to permutation. We have

$$x = WP^{-1}Pz \tag{71}$$

for some permutation matrix P.

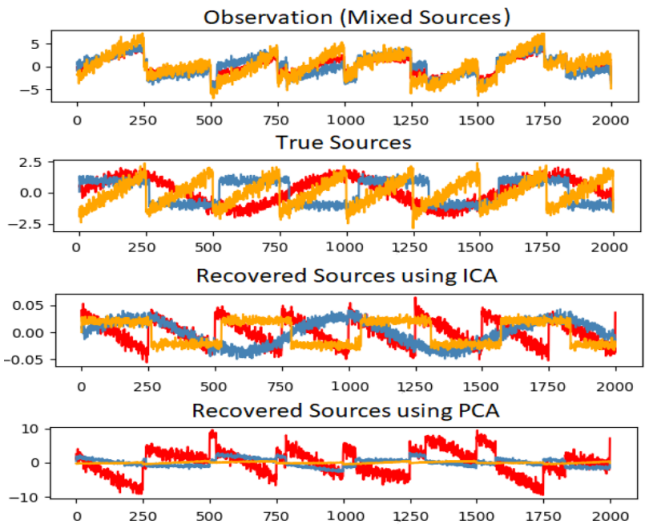


Figure 4: We can perform this on three mixed signals with additive noise, and ICA does very well, though again some recovered signals are scaled or permuted weirdly.

11 Slow Feature Analysis

Slow feature analysis also another special case of a linear factor model that uses information from time signals to learn invariant features. It is motivated by a general principle called the **slowness principle**. The idea is that the important characteristics of scenes change very slowly compared to the individual measurements that make up a description of a scene. For example, in computer vision, individual pixels can change very rapidly. If a zebra moves from left to right across the image, an individual pixel wil rapidly change from black to white. By comparison, the feature indicating whether a zebra is in the image will not change at all, and the feature describing the zebra's position will change slowly. Therefore, we want to regularize our model to learn features that change slowly over time.

We can apply the slowness principle to any differentiable model trained with gradient descent. That is, we can add the following term to the loss function:

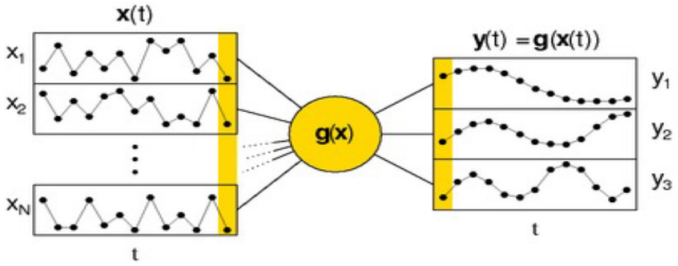
$$\lambda \sum_i d\big(f(x^{(t+1)}), f(x^{(t)})\big) \tag{72}$$

where λ is a hyperparameter determining the strength of the slowness regularization term, t is the time index, f is the feature extractor to be regularized, and d is the distance between $f(x^{(t)})$ and $f(x^{(t+1)})$. A common choice for d is the mean squared difference.

Essentially, given a set of time-varying input signals $x^{(t)}$, SFA learns a nonlinear function f that transforms x into slowly-varying output signals y. Obviously, we can't just take some trivial function like $f = 0$, so we have the following constraints

$$\mathbb{E}_t[f(x^{(t)})_i] = 0 \tag{73}$$

$$\mathbb{E}_t[f(x^{(t)})_i^2] = 1 \tag{74}$$



We can restrict the nonlinear f to some subspace of functions, and this becomes a standard optimization problem where we solve

$$\min_\theta \mathbb{E}_t\left[\big(f(x^{(t+1)})_i - f(x^{(t)})_i\big)^2\right] \tag{75}$$

12 Sparse Dictionary Learning

Latent variables can help us represent data in lower dimensions, but another advantage is that we can get *sparse* representations as well. What we want to do in sparse coding is that for each input $x^{(i)}$, we want to find a latent representation $z^{(i)}$ such that it is sparse (i.e. has many 0s) and also we can reconstruct the original input $x^{(i)}$ well. We have basically two things to optimize: the latent representations z and the decoding mechanism, which we can do with a *dictionary matrix* D. Note that we are optimizing for *both* the latent encodings and the decoding mechanism, and so this isn't a generative model.

Definition 12.1 (Sparse Dictionary Encoding Model)

The **sparse dictionary encoding model** is a representation model defined

$$X = g_D(Z) = DZ \tag{76}$$

where $D \in \mathbb{R}^{d \times k}$ is a **dictionary matrix** that decodes the latent $Z \in \mathbb{R}^k$ to $X \in \mathbb{R}^d$. Note that both the $z^{(i)}$'s and D are optimized, so we want to perform the *joint* optimization[a]

$$\min_D \frac{1}{N} \sum_{i=1}^{N} \min_{z^{(i)}} \underbrace{\frac{1}{2}||x^{(i)} - Dz^{(i)}||_2^2}_{\text{reconstruction error}} + \underbrace{\lambda ||z^{(i)}||_1}_{\text{sparsity penalty}} \tag{77}$$

[a] To break this term down, let's just assume that we have a fixed dictionary D. Then, we just need to minimize with respect to each $h^{(t)}$. Now we can add the dictionary parameter back again.

Note that the reconstruction, or decoding, of $x = Dz$ is linear and explicit, but if we want to encode $x \mapsto z$, we need to substitute the x into the term above and minimize it w.r.t. D and z to solve it. Therefore, this encoder is an implicit and *nonlinear* function of x.

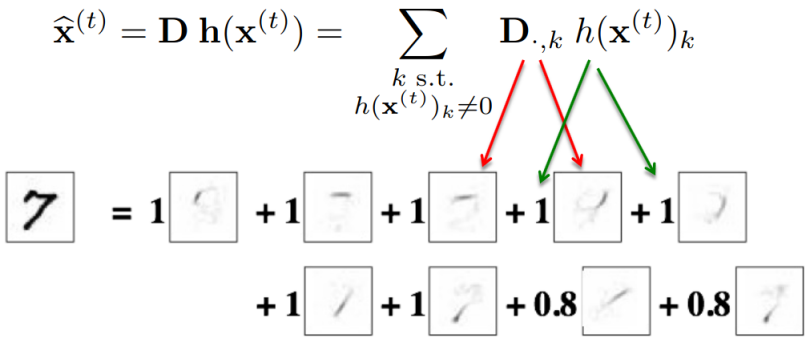


Figure 5: We can reconstruct an image of a seven as a linear combination of a set of images. Note that each of the images of strokes are columns of W and the coefficients make up the sparse vector h.

Let's think about how we can optimize the objective function w.r.t. h, keeping D constant. We can do stochastic gradient descent, which gives us the steps

$$\nabla_{h^{(t)}} \mathcal{L}(x^{(t)}) = D^T(Dh^{(t)} - x^{(t)}) + \lambda \operatorname{sign}(h^{(t)}) \tag{78}$$

but this wouldn't achieve sparsity since it overshoots the 0 all the time. Therefore, we can clip it, or we can use proximal gradient descent/ISTA to take a step, and shrink the parameters according to the L1 norm.

$$h^{(t)} = h^{(t)} - \alpha D^T(Dh^{(t)} - x^{(t)}) \tag{79}$$

$$h^{(t)} = \operatorname{shrink}(h^{(t)}, \alpha\lambda) \tag{80}$$

where $\text{shrink}(a, b) = [\ldots, \text{sign}(a_i)\ \max(|a_i| - b_i, 0), \ldots]$. This is guaranteed to converge if $1/\alpha$ is bigger than the largest eigenvalue of $D^T D$.

Bibliography

[Hot33] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:498–520, 1933.

[Pea01] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[Spe04] C. Spearman. "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

[TB99] Michael E. Tipping and Christopher M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B*, 61(3):611–622, 1999.