Principal Component Analysis

Jong-Han Kim

EE787 Fundamentals of machine learning
Kyung Hee University

Subspace data model

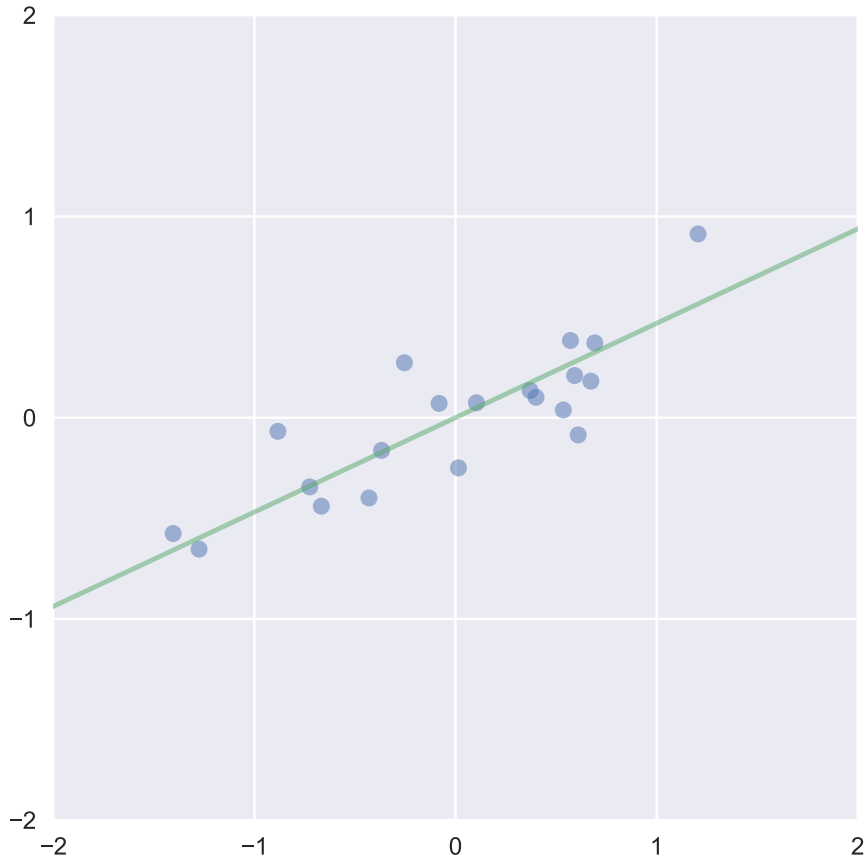
- data model: *x is near to a linear combination of the vectors $\theta_1, \ldots, \theta_r \in \boldsymbol{R}^d$*
- $d \times r$ matrix parameter $\theta = [\theta_1 \cdots \theta_r]$ parametrizes the model
- $r < d$ is called the *rank* of the model
- $\theta_1, \ldots, \theta_r$ are called the *principal components* or *archetypes*
- called *principal component analysis (PCA) model* or *low rank model*
- the implausibility or loss function is

$$\ell_\theta(x) = \min_a \|x - \theta a\|^2$$

  i.e., the minimum distance squared to a linear combination of $\theta_1, \ldots, \theta_r$

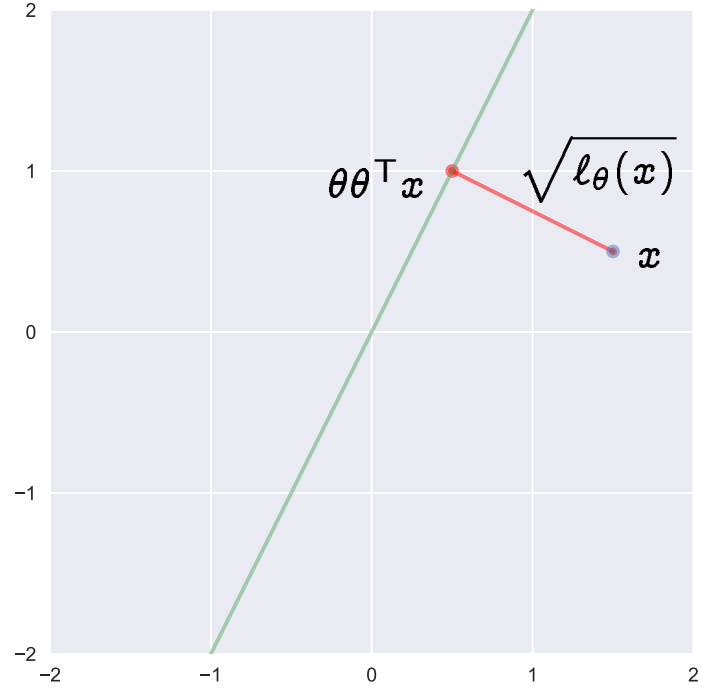
- we can assume that θ has orthonormal columns, *i.e.*, $\theta^T\theta = I$

Example



- plot shows case when $r = 1$
- line is $\{a\theta \mid a \in \mathbf{R}\}$

PCA loss function



- set of all linear combinations of $\theta_1, \ldots, \theta_r$ is called a *subspace* $\mathcal{S}$ of $\mathbf{R}^d$
- PCA loss function $\min_a ||x - \theta a||^2$ is distance squared to the subspace $\mathcal{S}$
- since we assume $\theta^{\mathsf{T}}\theta = I$, optimal a is $a = \theta^{\dagger}x = \theta^{\mathsf{T}}x$, so

$$\ell_\theta(x) = ||(I - \theta\theta^T)x||^2 = ||x||^2 - ||\theta\theta^{\mathsf{T}}x||^2 = ||x||^2 - ||\theta^{\mathsf{T}}x||^2$$

- if x is a linear combination of $\theta_1, \ldots, \theta_r$, $\ell_\theta(x) = 0$

PCA empirical loss

- given data set $x^1, \ldots, x^n$, form $n \times d$ data matrix

$$X = \begin{bmatrix} (x^1)^\mathsf{T} \\ \vdots \\ (x^n)^\mathsf{T} \end{bmatrix}$$

- empirical PCA loss is

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (||x^i||^2 - ||\theta^\mathsf{T} x^i||^2) = ||X||_F^2 - ||X\theta||_F^2$$

where $||B||_F^2 = \sum_{i,j} B_{ij}^2$ is the Frobenius norm squared of a matrix B

Fitting a PCA model

- we choose θ to minimize $\mathcal{L}(\theta)$ subject to $\theta^{\mathsf{T}}\theta = I$
- same as maximizing $||X\theta||_F^2$ subject to $\theta^{\mathsf{T}}\theta = I$

- this can be done *exactly* (non-heuristically) by several algorithms (singular value decomposition, eigenvalue decomposition)
- `theta=pca_fit(X,r)`
- complexity of simple methods is order nd^2 flops
- other methods are more efficient when $r \ll d$

Imputing with subspace data model

- find coefficients a to minimize $\sum_{i \in \mathcal{K}} (x_i - (\theta a)_i)^2$
- roughly speaking, find the closest linear combination of $\theta_1, \ldots, \theta_r$ to x, considering only the known entries
- guess $\hat{x}_i = (\theta a)_i$ for $i \notin \mathcal{K}$
- *i.e.*, use the same linear combination of $\theta_1, \ldots, \theta_r$ to guess the unknown entries

Approximate matrix factorization interpretation

- $a^i = \theta^\mathsf{T} x^i$ minimizes $||x^i - \theta a||^2$
- write as $A = X\theta$, where A has rows $a_1^\mathsf{T}, \ldots, a_n^\mathsf{T}$ (using $\theta^\dagger = \theta^\mathsf{T}$ since $\theta^\mathsf{T}\theta = I$)
- A is an $n \times r$ matrix
- $\tilde{x}^i = \theta a^i = \theta\theta^\mathsf{T} x^i$ is closest point to x^i in subspace
- write as $\tilde{X} = A\theta^\mathsf{T}$, where $\tilde{X}$ has rows $\tilde{x}_1^\mathsf{T}, \ldots, \tilde{x}_n^\mathsf{T}$
- $\tilde{X}$ is an $n \times d$ matrix; it is *tall-wide product*
- empirical loss is

$$\mathcal{L}(\theta) = ||X - \tilde{X}||_F^2 = ||X - X\theta\theta^\mathsf{T}||_F^2 = ||X - A\theta^\mathsf{T}||_F^2$$

- so PCA finds the closest matrix to X that is a product of an $n \times r$ and an $r \times n$ matrix

PCA for embedding and dimension reduction

- the mapping $a = \theta^T x$ gives *compressed features*
  - $x \in \mathbf{R}^d$ is the *original feature vector*
  - $a \in \mathbf{R}^r$ is the associated *compressed feature vector*
  - since (usually) $r \ll d$, this is *dimension reduction*

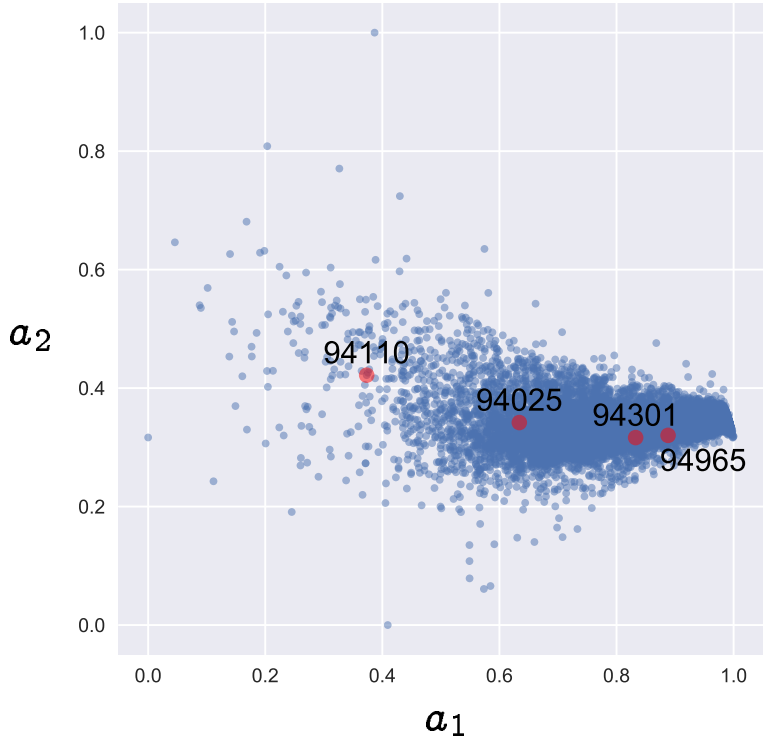
- the mapping $a = \theta^T x$ is a (linear) *embedding* from $\mathbf{R}^d$ into $\mathbf{R}^r$
  - the embedding is based on the data set
  - roughly speaking, it preserves the distances between the original feature vectors, to the extent possible, *i.e.*, we have $||a - \tilde{a}|| \approx ||x - \tilde{x}||$ for typical data

Approximate isometry property

- a mapping $F : \mathbf{R}^p \to \mathbf{R}^q$ is called an *isometry* if it preserves distances, *i.e.*, $\|F(x) - F(\tilde{x})\| = \|x - \tilde{x}\|$ for all x, $\tilde{x}$
- classic example: $F(x) = Qx$, where $Q^T Q = I$ (so Q is square or tall)

- recall that $\ell_\theta(x) = \|x\|^2 - \|\theta^{\mathsf{T}} x\|^2$ is the distance squared to the subspace $\mathcal{S}$
- so if this is small, *i.e.*, the data model is good, we have $\|x\| \approx \|a\|$
- in other words, the embedding $x \mapsto a = \theta^{\mathsf{T}} x$ is an *approximate isometry*
- useful for plotting or visualization with $r = 2$ or 3

Example: Census data



- U.S. census data 2010
- x has dimension $d = 186$
- $n = 33120$ rows, one per zipcode
- use PCA with $r = 2$; plot shows reduced features

Example: Census data

- plot shows entries of θ_1 and θ_2
- three large positive entries of θ_1 are median ages of total pop., male pop., and female pop.
- largest entries of θ_2 are racial population counts

Latent semantic indexing

Features from text

- each record u^i is a document
- d unique words in corpus of all documents
- embedding maps documents to d-vectors
- embed so that $\phi(u^i)_j > 0$ if word j is in document i

Embedding

- for a document u, *term frequency* of word j is

$$f_{\text{term}}(u, j) = \frac{\text{number of occurrences of word } j \text{ in } u}{\text{the number of words in } u}$$

- for a set of documents, the *document frequency* of word j is

$$f_{\text{doc}}(j) = \frac{\text{the number of documents in which the word occurs}}{n}$$

- TFIDF embedding

$$\phi(u)_j = f_{\text{term}}(u, j) \log(1/f_{\text{doc}}(j))$$

Example: Distinguishing texts

- *The Critique of Pure Reason* by Immanuel Kant and *The Problems of Philosophy* by Bertrand Russell
- 50 excerpts from each book
- each excerpt is approximately 3000 characters
- split into words, remove punctuation, capitalization
- $d = 3566$ unique words
- TFIDF embedding, standardize, PCA

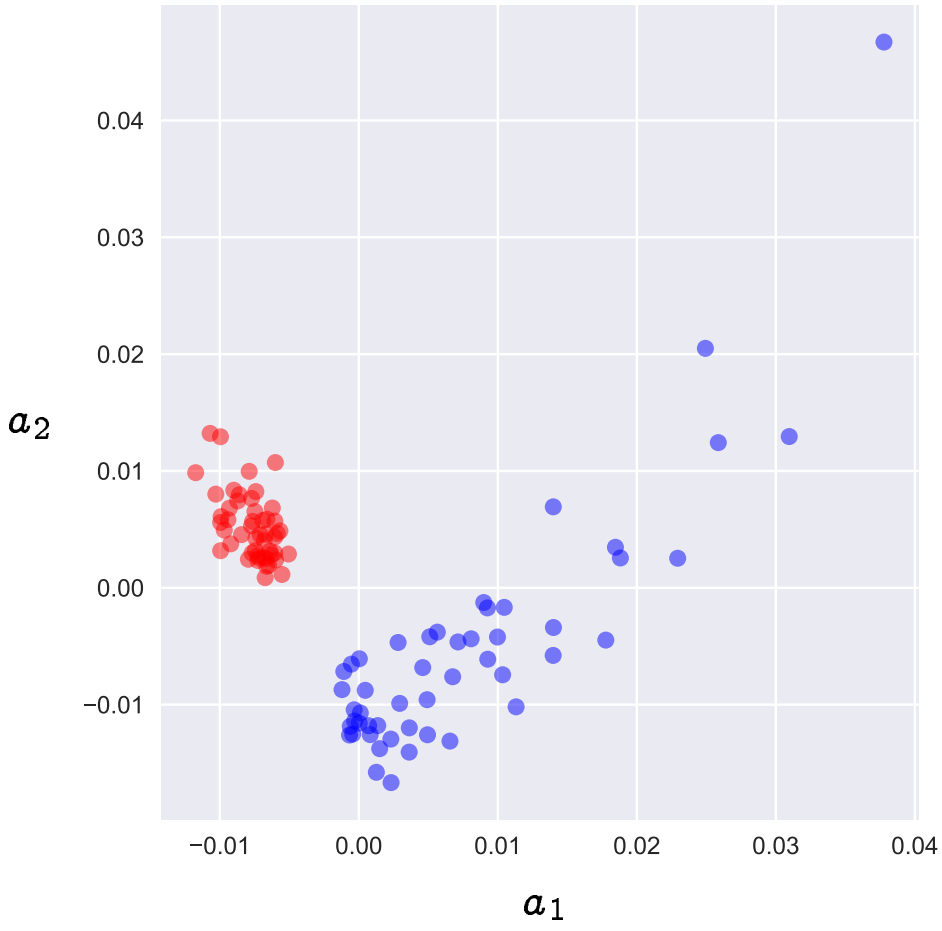
Example: 1000 characters of Kant

for these must be contemplated not as properties of things, but only as changes in the subject, changes which may be different in different men. For, in such a case, that which is originally a mere phenomenon, a rose, for example, is taken by the empirical understanding for a thing in itself, though to every different eye, in respect of its colour, it may appear different. On the contrary, the transcendental conception of phenomena in space is a critical admonition, that, in general, nothing which is intuited in space is a thing in itself, and that space is not a form which belongs as a property to things; but that objects are quite unknown to us in themselves, and what we call outward objects, are nothing else but mere representations of our sensibility, whose form is space, but whose real correlate, the thing in itself, is not known by means of these representations, nor ever can be, but respecting which, in experience, no inquiry is ever made.

Example: 1000 characters of Russell

intrinsic nature, and continues to exist when I am not looking, or is the table merely a product of my imagination, a dream-table in a very prolonged dream? This question is of the greatest importance. For if we cannot be sure of the independent existence of objects, we cannot be sure of the independent existence of other people's bodies, and therefore still less of other people's minds, since we have no grounds for believing in their minds except such as are derived from observing their bodies. Thus if we cannot be sure of the independent existence of objects, we shall be left alone in a desert–it may be that the whole outer world is nothing but a dream, and that we alone exist. This is an uncomfortable possibility; but although it cannot be strictly proved to be false, there is not the slightest reason to suppose that it is true. In this chapter we have to see why this is the case. Before we embark upon doubtful matters, let us try to find some more or less fixed point from which

Example: Distinguishing texts



- X is 100×2262
- russell in red, kant in blue

Example: Distinguishing texts