probabilistic principal component analysis (probabilistic PCA)

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

Probabilistic principal component analysis (probabilistic PCA) is a probabilistic model for the feature vectors of a dataset: a low-dimensional Gaussian random vector, the code, is mapped into the feature space by a matrix, and isotropic Gaussian noise is added. The probability distribution of the feature vector is then a multivariate normal distribution whose covariance matrix is a matrix of rank d' plus a multiple of the identity matrix, so the model gives a probability density function (pdf) value to every point of the feature space. That value is what principal component analysis (PCA) does not deliver: a data point outside the training set can be ranked by how probable it is, and features missing from a feature vector are filled in by an expectation. The maximum likelihood estimate of the model parameters has a closed form in the eigenvalues and eigenvectors of the sample covariance matrix: the columns of the matrix span the subspace that PCA keeps, and the noise variance is the average of the discarded eigenvalues. The expectation–maximization (EM) algorithm computes the same estimate as a fixed-point iteration, without an eigenvalue decomposition (EVD). The posterior mean of the code pulls the PCA coordinates toward the origin, and becomes the orthogonal projection on the subspace as the noise variance tends to zero, which recovers PCA. The isotropic noise is the model's main restriction: it makes the learned subspace depend on the units of the features, and factor analysis gives each measurement its own noise variance instead.

Definition

A weather station at Krems an der Donau (Austria) records eight numbers for each day of 2024: the lowest, highest, and mean air temperature, precipitation, sunshine duration, humidity, air pressure, and wind speed. Principal component analysis (PCA) replaces the eight by two, few enough for a scatterplot, and says nothing about how probable a day is: 14 September 2024, which brought 81.7 mm of rain, is placed in the plane like any other day. Probabilistic PCA keeps the subspace that PCA delivers and adds a probability distribution over the eight numbers, given by a probability density function (pdf) on the feature space. The rainy day then carries a density value far below that of every other day of the year.

Probabilistic PCA is a latent variable probabilistic model. Each data point of a dataset carries a feature vector $\mathbf{x}^{(r)} \in \mathbb{R}^d$, for $r = 1, \ldots, m$, with $d = 8$ and $m = 366$ for the days of 2024. The model builds it from a low-dimensional vector $\mathbf{z}^{(r)} \in \mathbb{R}^{d'}$, the code, whose length $d' < d$ is a hyperparameter:

$$\mathbf{x}^{(r)} = \mathbf{W}\mathbf{z}^{(r)} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}^{(r)}, \text{ for } r = 1, \ldots, m. \tag{1}$$

Here, $\mathbf{W} \in \mathbb{R}^{d \times d'}$ is a matrix and $\boldsymbol{\mu} \in \mathbb{R}^d$ a mean vector. The codes and the noise vectors of the model (1) are realizations of independent random vectors, $\mathbf{z}^{(r)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\varepsilon}^{(r)} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$, with the identity matrix $\mathbf{I}$ and a noise variance $\sigma^2 > 0$. The noise is isotropic: it has the same variance in every direction and no covariance between features. The matrix $\mathbf{W}$ maps a code to a feature vector, the role the decoder plays in PCA, and the model parameters are $\mathbf{W}$, $\boldsymbol{\mu}$, and σ^2.

Integrating the code out of the model (1) leaves a multivariate normal distribution for the feature vector itself,

$$\mathbf{x}^{(r)} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}), \text{ where } \mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}. \tag{2}$$

The covariance matrix $\mathbf{C}$ of the multivariate normal distribution (2) is a matrix of rank at most d' plus a multiple of the identity matrix. That is the one restriction separating probabilistic PCA from an unconstrained multivariate normal distribution: the variance along each of the d' retained directions is modeled on its own, while every direction orthogonal to all of them is given the same variance σ^2. Because the model (1) states how a data point comes about, it is a generative model: drawing a code and adding noise produces a new day. Fig. 1 draws the probability distribution (2) learned from the days at Krems taken with two features, the lowest and the highest temperature of the day, as a level curve one standard deviation out from the mean vector. For a code of one number, the subspace that PCA keeps is a straight line, and the probability distribution fills out an ellipse around it.

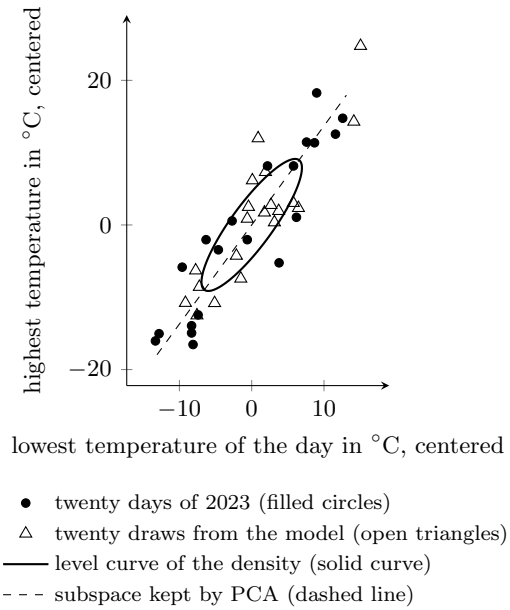


Figure 1: Twenty of the 365 days of 2023 at Krems (every 19th day), each a data point with the lowest and the highest temperature of the day as its two features, centered (filled circles). The solid curve is the level curve one standard deviation out of the pdf (2) learned from all 365 days with $d' = 1$, and the dashed line marks the subspace that PCA keeps. The open triangles are twenty feature vectors drawn from the learned model. Data generated by `pythondemos/ppca.py`

What the probability distribution (2) delivers and PCA does not is a number for every point of the feature space. A day that did not belong to the training set receives a density value, so days can be ranked by how probable the model finds them. With the model parameters learned from the 365 days of 2023 and $d' = 3$, the least probable day of 2024 at Krems is 14 September, where the logarithm of the pdf (2) is -440.8, against -8.5 for the median day and -105.6 for the next lowest (Fig. 2). Its 81.7 mm of rain are 20.4 sample standard deviations above the 2023 average of 1.5 mm. Ranking data points by their density value is one route to anomaly detection, and it is closed to PCA, which returns coordinates and no probability.

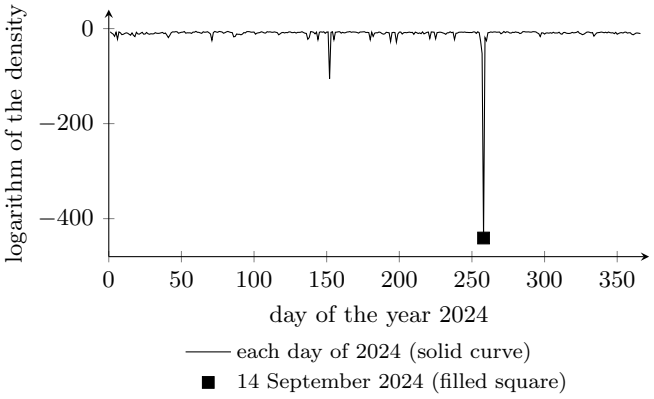


Figure 2: The logarithm of the pdf (2) at each of the 366 days of 2024 at Krems, under the model parameters learned from the 365 days of 2023 with $d' = 3$ and all eight measurements (solid curve). The filled square marks 14 September 2024, the day of 81.7 mm of rain, at -440.8, while the median day of the year is at -8.5, where the solid curve runs for most of the year. Data generated by `pythondemos/ppca.py`

The model parameters are learned by maximum likelihood, and the estimate has a closed form. The eigenvalues of the sample covariance matrix $\widehat{\mathbf{C}}$ of the training set, ordered as $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$, come with orthonormal eigenvectors $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(d)}$. The estimate of the mean vector is the sample mean $\widehat{\boldsymbol{\mu}}$, the estimate of the noise variance is

$$\widehat{\sigma}^2 = \frac{1}{d - d'} \sum_{j=d'+1}^{d} \lambda_j, \tag{3}$$

and the estimate $\widehat{\mathbf{W}}$ has the columns $\sqrt{\lambda_j - \widehat{\sigma}^2}\, \mathbf{u}^{(j)}$, for $j = 1, \ldots, d'$ (Bishop, 2006, Sect. 12.2.1, Eqs. (12.45) and (12.46); Tipping and Bishop, 1999, Sect. 3.2, Eqs. (7) and (8)). The average (3) of the discarded eigenvalues is the variance the retained directions leave unexplained, spread evenly over the $d - d'$ directions that carry it. The columns of $\widehat{\mathbf{W}}$ span the subspace that PCA keeps, the one spanned by the first d' eigenvectors of $\widehat{\mathbf{C}}$. Rotating those columns among themselves leaves the covariance matrix $\mathbf{C}$ unchanged, so maximum likelihood determines the subspace and not the individual columns.

The expectation–maximization (EM) algorithm reaches the same estimate without an eigenvalue decomposition (EVD) of $\widehat{\mathbf{C}}$, which is the cheaper route when d is large and few directions are wanted (Bishop, 2006, Sect. 12.2.2). Collecting the model parameters $\mathbf{W}$ and σ^2 in $\mathbf{w}$, one iteration is a fixed-point iteration

$$\mathbf{w}^{(t+1)} = T\big(\mathbf{w}^{(t)}\big),$$

whose operator T re-estimates $\mathbf{W}$ from $\widehat{\mathbf{C}}$ and the posterior distribution of the codes, then sets σ^2 to the variance the new $\mathbf{W}$ leaves unexplained (Tipping

and Bishop, 1999, Appendix B, Eqs. (29) and (30)). No iteration decreases the likelihood function, and the fixed points of T are the stationary points of that likelihood function. The closed-form estimate is the one among them at which the likelihood function attains its maximum; the stationary points built from eigenvectors other than the leading d' are saddle points (Tipping and Bishop, 1999, Appendix A.3).

In the vocabulary of empirical risk minimization (ERM), probabilistic PCA changes two of the three parts of PCA and keeps the third. The hypothesis space gains one number. PCA searches over the products of an encoder and a decoder, which are the matrices of rank at most d'. Probabilistic PCA searches over the multivariate normal distributions (2), whose covariance matrices are such a matrix plus $\sigma^2\mathbf{I}$. The columns of $\mathbf{W}$ need not be orthonormal: those of $\widehat{\mathbf{W}}$ have squared lengths $\lambda_j - \widehat{\sigma}^2$, for $j = 1, \ldots, d'$, so they fix the variance along each retained direction and not only the subspace. The loss function is the negative logarithm of the pdf of the probability distribution $\mathbb{P}$ (2), in place of the squared error loss of PCA,

$$L\left(\mathbf{x}, \mathbb{P}\right) = -\ln \mathcal{N}\left(\mathbf{x}\,;\, \boldsymbol{\mu}, \mathbf{C}\right),$$

so minimizing the empirical risk is maximum likelihood. The third part is unchanged: the EVD of $\widehat{\mathbf{C}}$ that solves PCA delivers this minimum in closed form as well, and the EM algorithm is the route that avoids it. The two loss functions meet as the noise variance tends to zero: $2\sigma^2$ times the empirical risk tends to the minimum reconstruction error $\sum_{j=d'+1}^{d} \lambda_j$ of PCA (Tipping and Bishop, 1999, Appendix A.2, Eq. (17)). That sum is 1.636 for the eight standardized measurements of the 365 days of 2023 with $d' = 3$. Multiplied by $2\sigma^2$, the empirical risk reaches 0.979 of it at $\sigma^2 = 10^{-2}$, 0.998 at 10^{-4}, and 1.000 at 10^{-6}.

A data point is placed in the subspace through the posterior distribution of its code. The posterior mean of the code is

$$\mathbb{E}\big\{\mathbf{z} \mid \mathbf{x}\big\} = \big(\widehat{\mathbf{W}}^\top \widehat{\mathbf{W}} + \widehat{\sigma}^2 \mathbf{I}\big)^{-1} \widehat{\mathbf{W}}^\top \big(\mathbf{x} - \widehat{\boldsymbol{\mu}}\big), \tag{4}$$

whose j-th entry, for the closed-form estimate, is the j-th PCA coordinate $\langle \mathbf{u}^{(j)}, \mathbf{x} - \widehat{\boldsymbol{\mu}} \rangle$ multiplied by $\sqrt{\lambda_j - \widehat{\sigma}^2}/\lambda_j$, a factor below one, for $j = 1, \ldots, d'$. The code is therefore pulled toward the origin, by the factors 0.50, 0.71, and 0.78 for the three retained directions at Krems. As $\widehat{\sigma}^2$ tends to zero each factor tends to one, and the posterior mean (4) becomes the orthogonal projection of the feature vector on the subspace spanned by the first d' eigenvectors, which is what PCA computes (Bishop, 2006, Sect. 12.2.1, Eq. (12.50); Tipping and Bishop, 1999, Sect. 3.4, Eq. (9)). At that limit the covariance matrix $\mathbf{C}$ is singular and the pdf ceases to exist, so the limit that recovers PCA is the one that gives up the probability distribution.

A probability distribution over the feature space also handles missing data. Conditioned on the entries of a feature vector that were recorded, the probability distribution (2) is again Gaussian in the entries that were not, and its expectation is the filled-in value. Deleting the highest temperature of each of

the 366 days of 2024 and filling it in this way, from the model parameters learned on 2023, leaves a mean squared error (MSE) whose square root is 2.28 degrees, against 9.21 degrees for filling in the 2023 average instead. The projection that PCA applies needs every entry of the feature vector and offers no such rule, while the same conditional reading lets the model parameters be learned from feature vectors with entries missing (Tipping and Bishop, 1999, Sect. 4.1).

The model makes two assumptions, and the days at Krems show what each one costs. The noise is isotropic, so every direction orthogonal to the retained subspace is given the same variance $\widehat{\sigma}^2 = 0.33$ for the eight standardized measurements with $d' = 3$. The mean air temperature of a day is the average of its lowest and its highest to within 0.05 degrees on every day of 2023. The sample standard deviation of the days along the last eigenvector of $\widehat{\mathbf{C}}$ is therefore 0.0027, where the model expects $\widehat{\sigma} = 0.57$. The isotropic noise also makes the learned subspace depend on the units of the measurements: multiplying the precipitation reading by ten turns it by 51 degrees, the largest angle between the subspace before and after the rescaling. That is why the eight measurements are standardized before the model parameters are learned. Factor analysis differs from probabilistic PCA only in giving the noise a covariance matrix that is a diagonal matrix, one variance per measurement, and a rescaling of one measurement leaves it unchanged (Bishop, 2006, Sect. 12.2.4; Tipping and Bishop, 1999, Sects. 2.1 and 3.3). With its model parameters learned from the same 365 days of 2023, it raises the median logarithm of the pdf over the 366 days of 2024 from -8.5 to -3.2, and what it gives up is the closed form (Tipping and Bishop, 1999, Sect. 2.1).

The second assumption is the Gaussian shape of the pdf (2), on which the ranking of days by density value depends. If the days of 2024 were realizations of i.i.d. random vectors with the learned multivariate normal distribution, the quadratic form in the exponent of that pdf would be a realization of a chi-squared random variable (RV) with eight degrees of freedom. Its 99th percentile is 20.1, which 3.7 of the 366 days would exceed on average. Twenty-six do, as do 16 of the 365 days of 2023 the model parameters were learned from. For 14 September 2024, the quadratic form is 870.6, which the same chi-squared RV exceeds with probability $1.2 \cdot 10^{-182}$, and that day occurred. The density value therefore ranks days, and it does not say how often a day of that kind occurs. Where one multivariate normal distribution is too simple, the standard extension is a mixture of probabilistic PCA models, a Gaussian mixture model (GMM) whose component covariance matrices are each of the form $\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$ (Tipping and Bishop, 1999, Sect. 4.2). What the two assumptions buy in return is a probability distribution over the whole feature space in closed form: it gives a density value to a day outside the training set and a filled-in value to a feature that was not recorded.

Cross References

PCA, robust PCA, sparse PCA, dimensionality reduction, ERM, probabilistic model, multivariate normal distribution, maximum likelihood, EM, latent

variable, generative model, sample covariance matrix, missing data, anomaly detection, GMM

References

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

2. Tipping and Bishop (1999). Probabilistic Principal Component Analysis. J. Roy. Statist. Soc.: Ser. B (Statist. Methodology), 61(3), pp. 611–622.