Principal Component Analysis

Kaiwen Zhou

Contents

Readings

In addition to the lecture notes, the following are required readings:

- Chapter 2.3.1, 2.3.2, 2.3.5, 2.4.1-2.4.4, 2.5.1, 2.5.2, 5.5.1-5.5.2 in in Golub and Van Loan (2013). Each one of these sections is short and quick to read.
- Chapter 14.1, 14.5.1, 14.7.2 in Friedman, Hastie, and Tibshirani (2017)
- Chapter 12.2 in Bishop (2006)

The Singular Value Decomposition

The Singular Value Decomposition I

Proposition (The Singular Value Decomposition)

For any matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, there exist orthogonal matrices

$$\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_m] \in \mathbb{R}^{m \times m}$$
$$\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n] \in \mathbb{R}^{n \times n}$$

and a $m \times n$ diagonal matrix

$$\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m \times n}, p := \min\{m, n\}$$

with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$, such that

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$$

The Singular Value Decomposition II

The decomposition

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$$

is called the singular value decomposition (SVD) of $\mathbf{X}$. The following naming convention is common

$$\begin{aligned} \sigma_i &: \text{ singular value }, \\ \mathbf{u}_i &: \text{ left singular vector }, \\ \mathbf{v}_i &: \text{ right singular vector }. \end{aligned}$$

The Singular Value Decomposition III

Remark

Note that we can also think of the SVD as decomposing a rank-r as a sum of rank-1 matrices:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i'$$

If $r = \min(m, n)$ then (9) corresponds to the full-rank SVD. Otherwise it would correspond to (9) a truncated SVD decomposition.

Properties of the SVD I

We denote by $\|\cdot\|_2$ and $\|\cdot\|_F$ the I^2 - and Frobenius norms. The following proposition shows that the l^2 - and Frobenius norms of a matrix can be easily determined from its SVD.

Proposition

If $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $p := \min\{m, n\}$, then the following identities hold

$$\|\mathbf{X}\|_F := \sqrt{\sum_{i=1,j=1}^{m,n} |a_{ij}|^2} = \sqrt{\sigma_1^2 + \ldots + \sigma_p^2},$$
$$\|\mathbf{X}\|_2 := \max \frac{\|\mathbf{X}\mathbf{v}\|_2}{\|\mathbf{v}\|_2} = \sigma_1,$$
$$\min \frac{\|\mathbf{X}\mathbf{v}\|_2}{\|\mathbf{v}\|_2} = \sigma_n (m \geq n).$$

Properties of the SVD II

Proposition (Eckart-Young Theorem)

Suppose $\mathbf{X} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i', k < r = \operatorname{rank}(\mathbf{X})$ and denote by $\mathbf{X}_k := \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i'$ the k-truncated SVD of $\mathbf{X}$. Then $\mathbf{X}_k$ is the best rank- k approximation of $\mathbf{X}$ both in the I^2-norm and the Frobenius norm, that is

$$\min_{\mathbf{Y} \text{ s.t. } \operatorname{rank}(\mathbf{Y})=k} \|\mathbf{X} - \mathbf{Y}\|_2 = \|\mathbf{X} - \mathbf{X}_k\|_2 = \sigma_{k+1}$$

and

$$\min_{\mathbf{Y} \text{ s.t. } \operatorname{rank}(\mathbf{Y})=k} \|\mathbf{X} - \mathbf{Y}\|_F^2 = \|\mathbf{X} - \mathbf{X}_k\|_F^2 = \sum_{k+1}^{N} \sigma_i^2$$

Properties of the SVD III

Suppose we now interpret $\mathbf{X}$ as a linear map $\mathbf{X} : \mathbb{R}^m \to \mathbb{R}^n$ which takes a vector $\mathbf{y} \in R^m$ and maps it to $\mathbf{z} = \mathbf{X}'y \in \mathbb{R}^n$.

In this case it turns out the SVD of $\mathbf{X}$ can be used in order to compute the orthogonal projection of the linear map onto the range and null spaces of $\mathbf{X}$ as follows

Properties of the SVD IV

Proposition

Suppose $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$ with $r = \operatorname{rank}(\mathbf{X})$. Writing $\mathbf{U}, \mathbf{V}$ in the form

$$\mathbf{U} = \left[\mathbf{U}_r \widetilde{\mathbf{U}}_r\right] \text{ where } \mathbf{U}_r \in \mathbb{R}^{m \times r} \text{ and } \widetilde{\mathbf{U}}_r \in \mathbb{R}^{m \times (m-r)},$$
$$\mathbf{V} = \left[\mathbf{V}_r \widetilde{\mathbf{V}}_r\right] \text{ where } \mathbf{V}_r \in \mathbb{R}^{n \times r} \text{ and } \widetilde{\mathbf{V}}_r \in \mathbb{R}^{n \times (n-r)},$$

then [2]

$\mathbf{U}_r \mathbf{U}_r'$: orth proj onto range $(\mathbf{X})$,

$\widetilde{\mathbf{V}}_r \widetilde{\mathbf{V}}_r'$: orth proj onto $(\mathbf{X})$,

$\mathbf{V}_r \mathbf{V}_r'$: orth proj onto null $(\mathbf{X})^{\perp}$ = range $(\mathbf{X}')$,

$\widetilde{\mathbf{U}}_r \widetilde{\mathbf{U}}_r'$: orth proj onto range $(\mathbf{X})^{\perp} = \operatorname{null}(\mathbf{X}')$.

Properties of the SVD V

In other words, we have the following picture.

$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}'$

Properties of the SVD VI

Proposition (SVD and Linear Regression)

Suppose $\mathbf{b} \in \mathbb{R}^m$ and $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' \in \mathbb{R}^{m \times n}$ with $r = \operatorname{rank}(\mathbf{A})$ and

$$\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$$
$$\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$$

Then:

1. The solution to

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$$

is given by

$$\mathbf{x}^* = \sum_{i=1}^{r} \frac{\mathbf{u}_i'\mathbf{b}}{\sigma_i}\mathbf{v}_i$$

and

$$\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2^2 = \sum_{i=r+1}^{m} \left(\mathbf{u}_i'\mathbf{b}\right)^2$$

Properties of the SVD VII

2. If $m \geq n$ and $\text{rank}(\mathbf{X}) = n$, we have that

$$\mathbf{x}^* = \left(\mathbf{A}'\mathbf{A}\right)^{-1}\mathbf{A}'\mathbf{b}$$

Properties of the SVD VIII

Remark

Note that #2 is the "standard situation" we encountered in previous lectures where $\mathbf{A}$ has full column rank, so that $\mathbf{A}'\mathbf{A}$ is invertible and the solution can be computed from the normal equations.

[2] Note that null $(\mathbf{X})^{\perp} = \text{range}\,(\mathbf{X}')$ and $\text{range}(\mathbf{X})^{\perp} = \text{null}\,(\mathbf{X}')$.

The Moore-Penrose Generalized Inverse I

The previous proposition motivates the following definition.

Suppose $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i$ with $r = \text{rank}(\mathbf{A})$. Then we define the Moore-Penrose inverse of $\mathbf{A}$ by

$$\mathbf{A}^+ := \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}'$$

where

$$\mathbf{\Sigma}^+ := \text{diag}\left(\sigma_1^{-1}, \ldots, \sigma_r^{-1}, 0, \ldots, 0\right) \in \mathbb{R}^{n \times m}$$

It is easy to verify that $\mathbf{A}^+$satisfies the four Moore-Penrose conditions

$$\begin{aligned}
\mathbf{A}\mathbf{A}^+\mathbf{A} &= \mathbf{A} \\
\mathbf{A}^+\mathbf{A}\mathbf{A}^+ &= \mathbf{A}^+ \\
\left(\mathbf{A}\mathbf{A}^+\right)' &= \mathbf{A}\mathbf{A}^+ \\
\left(\mathbf{A}^+\mathbf{A}\right)' &= \mathbf{A}^+\mathbf{A}.
\end{aligned}$$

The Moore-Penrose Generalized Inverse II

The Moore-Penrose inverse [3] is an example of a pseudo inverse or generalized inverse of a matrix. There is an infinite number of ways in which one can define a pseudo inverse. What makes the Moore-Penrose inverse useful is that it is unique. In other words, there is no other inverse that satisfies the Moore-Penrose conditions above.

Note that from the Moore-Penrose conditions it follows that $\mathbf{A}\mathbf{A}^+$ and $\mathbf{A}^+\mathbf{A}$ are orthogonal projections onto range $(\mathbf{A})$ and range $(\mathbf{A}')$, respectively.

[3] The interested reader might want to consult Moore (1920) and Penrose (1955), the original papers on this topic.

Principal Component Analysis

PCA - Setup and Motivation I

The goal of Principal Component Analysis (PCA) of a set of n i.i.d. p-dimensional observations $\{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^{p \times 1}$ is to find the best rank- k linear approximation to this data:

$$\mathbf{x}_i = \boldsymbol{\mu} + \mathbf{W}_k\mathbf{z}_i + \varepsilon_i$$

where $\boldsymbol{\mu} := n^{-1}\sum_{i=1}^{n}\mathbf{x}_i, \mathbf{W}_k \in \mathbb{R}^{p\times k}$ with $\mathbf{W}_k'\mathbf{W}_k = \mathbf{I}_k \in \mathbb{R}^{k\times k}$, $\{\varepsilon\}_{i=1}^{n} \in \mathbb{R}^{p\times 1}$ and $\{\mathbf{z}_i\}_{i=1}^{n} \in \mathbb{R}^{k\times 1}$.

Note that because we only measure the n-observations of the r.v. $\mathbf{x}$, the k-components of the r.v. $\mathbf{z}$, are typically called latent or hidden factors.

PCA - Setup and Motivation II

Because the observations are i.i.d., it does not matter in which order they arrive. However in time-series applications the index i is time t.

One common application of (33) is in risk modeling where $\mathbf{x}_t = \mathbf{r}_t \in \mathbb{R}^p$ is the returns of a set of p assets at time t and one seeks to identify the top- k systemic linear factors that correlate with these returns.

As standard portfolio theory tells us that only idiosyncratic exposure diversifies, by varying k in the latent factor model (33) one then hopes to find the maximal number of systemic factors in the asset universe in order to control the systemic exposure and understand all non-diversifiable risks.

PCA - Setup and Motivation III

Therefore, PCA in risk modeling attempts to address several goals:

1. Reconstruction and prediction

1.1 For a fixed k find the optimal factors and their loading that capture most of the reconstruction error in-sample

1.2 Predict likely factor realizations and reconstruction error from new data out-of-sample.

2. Probabilistic inference and generative modeling

2.1 Identify confidences for both error reconstruction and prediction

2.2 Generate future scenarios as a function of systemic risks

3. Missing Data inference

PCA - Reconstruction I

Note that in (33) we can always assume that $\boldsymbol{\mu} \equiv 0$. If that is not the case we demean our data by letting $\mathbf{x}_i \leftarrow \mathbf{x}_i - \boldsymbol{\mu}$.

For simplicity, let us assume our data is demeaned, and we are seeking a linear model such that

$$\mathbf{x}_i = \mathbf{W}_k\mathbf{z}_i + \varepsilon_i$$

PCA - Reconstruction II

We can think about fitting such a model by minimizing the reconstruction error

$$\mathbf{z}_i^*, \mathbf{W}_k^* = \min_{\{\mathbf{z}_i\},\mathbf{W}_k} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{W}_{\mathbf{z}_i}\|_2^2$$

Or in matrix notation:

$$\mathbf{Z}^*, \mathbf{W}_k^* = \min_{\mathbf{Z},\mathbf{W}_k} \left\|\mathbf{X} - \mathbf{Z}\mathbf{W}_k'\right\|_F^2.$$

PCA - Reconstruction III

By the Ekhart-Young theorem we know that $\mathbf{X}_k = \mathbf{U}_k\boldsymbol{\Sigma}_k\mathbf{V}_k'$ optimizes the Frobenius norm so one possible solution is:

$$\mathbf{Z}^* = \mathbf{U}_k\boldsymbol{\Sigma}_k, \quad \mathbf{W}_k^* = \mathbf{V}_k$$

Note, that the solution is not unique (Question: what is the full set of solutions?).

With this choice of solution, the $k-$ unit vectors $\mathbf{W}_k^*$ are called the top- k principal components of the data $\mathbf{X}$. We will drop the subscript k for simplicity.

PCA - Prediction

For fixed $\mathbf{W}$, the optimal latent factor realizations are:

$$\mathbf{Z}_{\mathbf{W}}^* = \mathbf{X}\mathbf{W}\left(\mathbf{W}'\mathbf{W}\right)^{-1} = \mathbf{X}\mathbf{W}$$

or in vector notation, with non-zero μ :

$$\mathbf{z}^*_{\mathbf{W}} = \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'(\mathbf{x}-\boldsymbol{\mu}) = \mathbf{W}'(\mathbf{x}-\boldsymbol{\mu})$$

If $\mathbf{W}$ happens to be the optimal loading $\mathbf{W} = \mathbf{W}^*$ and $\mathbf{x}$ is a newly observed out-of-sample point, then the above two formulas can also be used for prediction. So how do we find $\mathbf{W}^*$?

Computing the PCA I

If we substitute $\mathbf{Z}^*_{\mathbf{W}}$ back into the loss function, the resulting loss only depends on $\mathbf{W}$ and $\mathbf{X}$ and the optimization reduces to:

$$\begin{aligned}\mathbf{W}^* &= \min_{\mathbf{W},\mathbf{W}'\mathbf{W}=\mathbf{I}_k} \|\mathbf{X}\mathbf{P}_{\mathbf{W}}\|_F^2, \quad \mathbf{P}_{\mathbf{W}} := \left(\mathbf{I}-\mathbf{W}\mathbf{W}'\right)\\ &:= \min_{\mathbf{W},\mathbf{W}'\mathbf{W}=\mathbf{I}_k} J^{PCA}_{MSE}(\mathbf{X},\mathbf{W})\end{aligned}$$

or in vector notation:

$$\{\mathbf{w}^*_i\} = \min_{\{\mathbf{w}_i\},\mathbf{w}_i'\mathbf{w}_j=\delta_{ij}} \mathbb{E}\left\|\mathbf{x}-\sum_{\ell=1}^{k}\mathbf{w}_\ell'\mathbf{x}\mathbf{w}_\ell\right\|^2$$

Computing the PCA II

Note that because $\mathbf{P}_w$ is a projection operator onto the space spanned by $\{\mathbf{w}_i\}$ then the principal components $\{\mathbf{w}_i\}_{i=1}^k$ optimize the mean square error compression of the original p - dimensional signal x onto a k-dimensional subspace. In other words $\{\mathbf{w}_i\}_{i=1}^k$ span the k - dimensional subspace with the optimal mean-square reconstruction error J^{PCA}_{MSE} of the original dataset.

How do we find the top k - principal components?

Computing the PCA III

One can show that the reconstruction error can be rewritten as

$$\begin{aligned}J^{PCA}_{MSE} &= \mathrm{Tr}\left[\mathbf{C}_x\right] - \mathrm{Tr}\left[\mathbf{C}_x\mathbf{W}\mathbf{W}'\right]\\ &= \mathrm{Tr}\left[\mathbf{C}_x\right] - \mathrm{Tr}\left[\mathbf{W}'\mathbf{C}_x\mathbf{W}\right]\end{aligned}$$

where $\mathbf{C}_x := \mathbf{X}'\mathbf{X} = \mathbb{E}\left[\mathbf{x}'\mathbf{x}\right]$ is the covariance matrix of the data.

Computing the PCA IV

As the first term in the rhs of (44) does not depend on $\mathbf{W}$ then the PCA are the set of k - unit vectors which optimize the second term. In vector notation this amounts to the following quadratic optimization:

$$\{\mathbf{w}^*_i\}_{i=1}^k = \max_{\{\mathbf{w}_i\},\mathbf{w}_i'\mathbf{w}_j=\delta_{ij}} \sum_{\ell=1}^{k}\mathbf{w}_\ell'\mathbf{C}_x\mathbf{w}_\ell$$

One can show (how?) that the necessary conditions for optimiality in (45) amount to the problem of finding the k-eigenvectors with the top eigenvalues of $\mathbf{C}_x$.

Computing the PCA $\vee$

Another way to see this is to note that since $\mathbf{C}_x$ is a symetric positive semidefinite matrix, then it can be decomposed in terms of its eigenvectors and eigenvalues as $\mathbf{C}_x = \sum_{\ell=1}^p \lambda_\ell \mathbf{e}_\ell\mathbf{e}_\ell'$ where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$ and $\{\mathbf{e}_i\}_{i=1}^p$ form an orthonormal basis of $\mathbb{R}^p$.

As a result, any $\mathbf{w} \in \mathbb{R}^p$ can be decomposed into the eigenvector basis as $\mathbf{w} = \sum_{i=1}^p b_p\mathbf{e}_p$. By substituting into (45) one can show that $\mathbf{w}_\ell = \mathbf{e}_\ell, \ell = 1,\ldots,k$ is a solution.

Computing the PCA VI

The top- k PCA can be identified with the unit eigenvectors corresponding to the top- k eigenvalues of the covariance of the data, and the optimal MSE is therefore:

$$J^{PCA,*}_{MSE} = \sum_{\ell=1}^{p}\lambda_\ell - \sum_{\ell=1}^{k}\lambda_\ell = \sum_{\ell=k+1}^{p}\lambda_\ell$$

Question: Are the PCA components unique? How about the optimal reconstruction error above? More on this later.

Computing the SVD

By now it should be clear that computing the rank- r truncated SVD approximation of a matrix in the Eckart-Young Theorem (14)

$$\begin{aligned}\mathbf{X} &= \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i' + \mathbf{E}_r \\ &= \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r' + \mathbf{E}_r\end{aligned}$$

can be reduced to a PCA computation. We leave the details as an exercise.

Probabilistic PCA - Motivation

So far we have no notion of confidence for our estimate of the PCA components, latent factor realizations, or our predictions. For that, we need a probabilistic approach.

The confidences of our in-sample estimates are what one usually associates with probabilistic statistical inference. The probabilistic theory of predictions is what is typically associated with generative modeling.

We will next develop the probabilistic PCA framework while exposing its salient features. However, we will postpone discussing the Expectation-Maximization Algorithm used in missing data probabilistic inference, for a later lecture.

Probabilistic PCA - Formalism I

A probabilistic formulation of the linear model (33) reproduced below (we drop the subscript of $\mathbf{W}_k$ for simplicity):

$$\mathbf{x}_i = \boldsymbol{\mu} + \mathbf{W}\mathbf{z}_i + \varepsilon_i$$

can easily be achieved by assigning distributional priors to the k-dimensional latent variable $\mathbf{z} \in \mathbb{R}^k$ and the noise $\varepsilon \in \mathbb{R}^p$. However, the linear model considered so far corresponds to a correlated factor model due to the fact that the factor loadings are constrained to be unitary $\mathbf{W}'\mathbf{W} = \mathbf{I}_k$.

Probabilistic PCA - Formalism II

An equivalent formulation analyzed by Tipping & Bishop (see Bishop (2006) 12.2) is to relax the unitary constraint of the loadings, and instead consider an arbitrary $\mathbf{W} \in \mathbb{R}^{p \times k}$ while constraining the factor realizations $\{\mathbf{z}_i\}$ correspond to an uncorrelated random variable $\mathbf{z} \in \mathbb{R}^k$.

As a result, the unconditional distributions of the latent variables and the noise are:

$$\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_k) = p(\mathbf{z}), \quad \varepsilon \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}_p\right) = p(\varepsilon)$$

Probabilistic PCA - Formalism III

Note: If in the matrix formulation $\mathbf{X} = \mathbf{Z}\mathbf{W}' + \mathbf{E} = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k' + \mathbf{E}$ one were to optimize for $\mathbf{W}$ and $\mathbf{Z}$ by minimizing the reconstruction error, then the above choice of constraints would correspond to:

$$\mathbf{Z}^* = \mathbf{U}_k, \quad \mathbf{W}^* = \mathbf{V}_k \mathbf{\Sigma}_k$$

Probabilistic PCA - Formalism IV

From (50) and (49) we can read off the distribution of $\mathbf{x}$ conditioned on the latent variable $\mathbf{z}$:

$$\mathbf{x} \mid \mathbf{z} \sim \mathcal{N}\left(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2 \mathbf{I}\right) = p\left(\mathbf{x} \mid \mathbf{z}; \mathbf{W}, \sigma^2, \boldsymbol{\mu}\right)$$

as well as the unconditional distribution of x :

$$\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{C} := \mathbf{W}\mathbf{W}' + \sigma^2 \mathbf{I}\right) = p\left(\mathbf{x}; \mathbf{W}, \sigma^2, \boldsymbol{\mu}\right)$$

One can verify that $\int p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{z} = p(\mathbf{x})$.

Probabilistic PCA - Formalism V

Because we have a probabilistic formulation, we can now find the optimal $\mathbf{W}$ and σ^2 which maximize the log-likelihood of the data:

$$\begin{aligned}\ln p\left(\mathbf{X} \mid \mu, \mathbf{W}, \sigma^2\right) &= \sum_{i=1}^{n} \ln p\left(\mathbf{x}_i \mid \mathbf{W}, \boldsymbol{\mu}, \sigma^2\right) \\ &= -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{x}_i - \boldsymbol{\mu}\right)' \mathbf{C}^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}\right) \\ &= -\frac{n}{2}\left[p\ln(2\pi) + \ln|\mathbf{C}| + \operatorname{Tr}\left[\mathbf{C}^{-1}\mathbf{C}_{\mathbf{x}}\right]\right]\end{aligned}$$

where $\mathbf{C}_x = \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})'(\mathbf{x} - \boldsymbol{\mu})\right]$ is the data covariance matrix which we encountered earlier in the reconstruction error formulation.

Probabilistic PCA - Formalism VI

Minimizing (56) is not straightforward due to the non-linear nature of the log-likelihood. However, by using of the Woodbury matrix identity one can show

$$\mathbf{C}^{-1} = \sigma^{-1} - \sigma^{-2}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}'$$

where $\mathbf{M} := \mathbf{W}'\mathbf{W} + \sigma^2\mathbf{I}$ is a $k \times k$ matrix with typically $k \ll \min(n, p)$

It turns out, that despite the non-linearities of the log-likelihood, an exact solution exists (see Bishop (2006) eq (12.45)) which reduces to (51) up to a k-dimensional rotation (why?) in the $\sigma^2 \to 0$ limit.

Probabilistic PCA - Formalism VII

Finally, once the factor loadings $\mathbf{W}$, noise σ^2 and expectation μ are learned from history, one often wants to know what a given new set of data $\mathbf{x}$ implies about the distribution of the latent variables $\mathbf{z}$. This information is captured by the posterior distribution of $\mathbf{z}$ given x and the model parameters, which one can show (how?) is given by:

$$\mathbf{z} \mid \mathbf{x} \sim \mathcal{N}\left(\mathbf{M}^{-1}\mathbf{W}'(\mathbf{x} - \boldsymbol{\mu}), \sigma^{-2}\mathbf{M}\right) = p\left(\mathbf{z} \mid \mathbf{x}; \mathbf{W}, \sigma^2, \boldsymbol{\mu}\right)$$

Note that in the limit of $\sigma^2 \to 0$,

$$\mathbb{E}[\mathbf{z} \mid \mathbf{x}] \to \left(\mathbf{W}'_{ML}\mathbf{W}_{ML}\right)^{-1}\mathbf{W}'_{ML}(\mathbf{x} - \mu)$$

which is the projection to the principal components (38) which we had derived in the matrix notation.

References

[1] Rasmussen, et al. (2006). Gaussian Processes In Machine Learning.

[2] Sheffield, S. (2007). "Gaussian Free Fields for Mathematicians". In: 139 (3-4), pp. 521-541.