Dimensionality reduction-principal component analysis[1]

Ningxin Yang

March 5, 2024

This document describes a specific type of dimensionality reduction technique-principal component analysis

Principal component analysis (PCA) is a widely covered machine learning method on the web. And while there are some great articles about it, many go into too much detail. Below we cover how principal component analysis works in a simple step-by-step way.

Dimensionality reduction (*DR*)

Let us consider a computational model $\mathcal{M}$ with a specific realization of input sets $\boldsymbol{x} = (x_1, \cdots, x_M)$. This model enables the analyst to predict certain *QoI*, represented in a vector $\boldsymbol{y} \in \mathbb{R}^N$ as a function of input parameters $\boldsymbol{x}$:

$$\mathcal{M} : \boldsymbol{x} \in D_{\boldsymbol{X}} \subset \mathbb{R}^M \rightarrow \boldsymbol{y} = \mathcal{M}(\boldsymbol{x}) \in \mathbb{R}^N \tag{1}$$

One possible set of high-dimensional outputs can be represented as $\boldsymbol{y} = (y_1, \cdots, y_N)^\intercal$. Considering different realizations in the experimental design $\mathcal{X} = \{\boldsymbol{x}^1, \cdots, \boldsymbol{x}^K\}$, where $\boldsymbol{x}^k \in \mathbb{R}^M$ for $k = 1, \cdots, K$, the computed responses can be collected into a data matrix as:

$$\boldsymbol{y} = \begin{bmatrix} (\boldsymbol{y}^1)^\intercal \\ (\boldsymbol{y}^2)^\intercal \\ \cdots \\ (\boldsymbol{y}^K)^\intercal \end{bmatrix} = \begin{bmatrix} y_1^1 & y_2^1 & \cdots & y_N^1 \\ y_1^2 & y_2^2 & \cdots & y_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ y_1^K & y_2^K & \cdots & y_N^K \end{bmatrix} \tag{2}$$

The superscript denotes the realizations of *experimental design* (or *observed samples*) and the subscript denotes the number of outputs. In an abstract form, the transformation from the *original space* $\mathcal{D}_{\boldsymbol{Y}} \subseteq \mathbb{R}^N$ to a *reduced space* $\mathcal{D}_{\boldsymbol{Z}} \subseteq \mathbb{R}^{N'}$ ($N' \ll N$) is the general form of a DR mapping:

$$\mathcal{T}_{DR} : \mathcal{D}_{\boldsymbol{Y}} \rightarrow \mathcal{D}_{\boldsymbol{Z}} \tag{3}$$

where the underlying assumption is that $\mathcal{D}_{\boldsymbol{Z}}$ is embedded inside $\mathcal{D}_{\boldsymbol{Y}}$. And the nature and number of the parameters N' is depending on the specific DR technique chosen. So far, there exists a large set of DR techniques ranging from linear to nonlinear approaches. To clarify the principles of DR, this study will adopts a simple but effective linear technique called *princepal component analysis* (PCA) [2]. PCA is selected

[1] Motivated by Dr Truong Le to write this document

[2] P-R Wagner, Reto Fahrni, Michael Klippel, Andrea Frangi, and Bruno Sudret. Bayesian calibration and sensitivity analysis of heat transfer models for fire insulation panels. *Engineering structures*, 205:110063, 2020; Paul-Remo Wagner. *Stochastic Spectral Embedding in Forward and Inverse Uncertainty Quantification*. PhD thesis, ETH Zurich, 2021; Joseph B Nagel, Jörg Rieckermann, and Bruno Sudret. Principal component analysis and sparse polynomial chaos expansions for global sensitivity analysis and model calibration: Application to urban drainage simulation. *Reliability Engineering & System Safety*, 195:106737, 2020; and Manav Vohra, Paromita Nath, Sankaran Mahadevan, and Yung-Tsun Tina Lee. Fast surrogate modeling using dimensionality reduction in model inputs and field output: Application to additive manufacturing. *Reliability engineering & system safety*, 201:106986, 2020

not only for its simplicity but also for its enduring favor across diverse disciplines, underscoring its continued widespread use. It is closely related to the *Karhunen-Loève expansion*, *Hotelling transform*, and *proper orthogonal decomposition*.

In practical applications, a PCA process is mainly calculating the estimation of the expectation $\boldsymbol{\mu_Y}$ and the covariance matrix $\boldsymbol{\Sigma_Y}$:

$$\boldsymbol{\mu_Y} = \mathbb{E}[\boldsymbol{Y}] = \begin{bmatrix} \mu_{y_1} \\ \mu_{y_2} \\ \cdots \\ \mu_{y_N} \end{bmatrix}^{\mathsf{T}} \quad \text{with } \mu_{y_i} = \frac{1}{K}\sum_{k=1}^{K} y_i^k,\ i = 1, \cdots, N \tag{4}$$

$$\boldsymbol{\Sigma_Y} \approx Cov[\boldsymbol{Y}] = \mathbb{E}[(\boldsymbol{Y} - \boldsymbol{\mu_Y})(\boldsymbol{Y} - \boldsymbol{\mu_Y})^{\mathsf{T}}] \tag{5}$$

Since $\boldsymbol{\Sigma_Y}$ is symmetric and positive definite, one can find linearly independent eigenvectors $\boldsymbol{\phi}_i$ with positive eigenvalues λ_i for $i = 1, \cdots, N$. The characteristic vectors and values should satisfy:

$$\Sigma_{\boldsymbol{Y}}\boldsymbol{\phi}_i = \lambda_i \boldsymbol{\phi}_i \tag{6}$$

The N eigenvectors of this covariance matrix (new coordinations to be projected on) are collected into a matrix $\boldsymbol{\Phi}_N = \{\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_N\}$ for $i = 1, \cdots, N$. The corresponding eigenvalue λ_i signifies the variance of $\boldsymbol{Y}$ in direction of the i-th principle component showing the descending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$. Since original data $\boldsymbol{Y}$ now has been centered and decorrelated in Equations (4) to (6), e.g., the linearly transformed random vector has mean zero $\mathbb{E}[\boldsymbol{Z}] = 0$ and the diagonal covariance matrix $Cov[\boldsymbol{Z}] = \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}] = \boldsymbol{\Lambda}$. Orthogonal components can be then represented as:

$$\boldsymbol{Z}_{\text{full}} = \boldsymbol{\Phi}_N^{\mathsf{T}}(\boldsymbol{Y} - \boldsymbol{\mu_Y}) \tag{7}$$

Figure 1 shows an example of PCA process centering and decorrelating on a two-dimensional data. In which, linear non-correlated coordinations (e.g., PC1 and PC2) can be found based on the descending orders of total variance.

By retaining only those N' principal components with the highest variance, the model output $\boldsymbol{Y}$ can then be compressed to a lower dimensional subspace $\boldsymbol{Z}$.

$$\boldsymbol{Z} = \boldsymbol{\Phi}_{N'}^{\mathsf{T}}(\boldsymbol{Y} - \boldsymbol{\mu_Y}) \approx \boldsymbol{Z}_{\text{full}} \tag{8}$$

Each realization (*observation sample*) can reduce the number of outputs though Equation (8). A specific realization $\boldsymbol{x}^k$ can be visualized in Figure 2 based on the selected eigenvectors $\boldsymbol{\Phi}_{N'}^{\mathsf{T}}$.

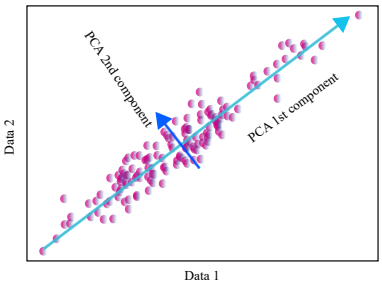

Figure 1: *An example of PCA process on a two-dimensional data*

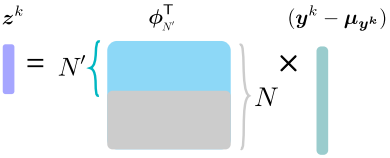

Figure 2: *A visualized realization of data compressing with selected eigenvectors*

All dataset can be then compressed while retaining most of the total variation by:

$$\boldsymbol{z} = \begin{bmatrix} (\boldsymbol{z}^1)^\mathsf{T} \\ (\boldsymbol{z}^2)^\mathsf{T} \\ \cdots \\ (\boldsymbol{z}^K)^\mathsf{T} \end{bmatrix} = \begin{bmatrix} z_1^1 & z_2^1 & \cdots & z_{N'}^1 \\ z_1^2 & z_2^2 & \cdots & z_{N'}^2 \\ \vdots & \vdots & \ddots & \vdots \\ z_1^K & z_2^K & \cdots & z_{N'}^K \end{bmatrix} \tag{9}$$

The selection of the number N' is determined such that $\sum_{i=1}^{N'} \lambda_i = (1 - \varepsilon_{\text{DR}}^{threshold}) \sum_{i=1}^{N} \lambda_i$, where $\varepsilon_{\text{DR}}^{threshold}$ is a prespecified threshold. An adequate number of principal components is to represent the system in an optimal way. If few PCs are selected than required, a poor model will be obtained. On the contrary, if more PCs than necessary are selected, the model will still face the pressure from high dimensionality. In real high-dimensional data outputs, it may also encounter the problems of over-parameterized and will include noise. To avoid these problems, several criteria for selecting the optimum number of PCs were proposed such as scree plot, explain-variance, permutation test, cross-validation and variance of reconstruction error [3]. With a suitable threshold and choosing criteria, the original space can be reconstructed as $\boldsymbol{Y}^{Re}$ through its optimum N' principal components using Equation (10):

$$\boldsymbol{Y} = \boldsymbol{\mu}_{\boldsymbol{Y}} + \sum_{i=1}^{N} \boldsymbol{z}_i \boldsymbol{\phi}_i \approx \boldsymbol{Y}^{Re} = \boldsymbol{\mu}_{\boldsymbol{Y}} + \sum_{i=1}^{N'} \boldsymbol{z}_i \boldsymbol{\phi}_i \tag{10}$$

Similar to PCA, most DR techniques share the same ingredients as illustrated in Figure 3: (1) calculate the reduced latent spaces $\boldsymbol{Z}$ (e.g., principal components in PCA) and (2) based on a predefined threshold $\varepsilon_{\text{DR}}^{threshold}$ and transform process $\mathcal{T}_{DR}^{-1}$, the outputs $\boldsymbol{Y}^{Re}$ can be reconstructed. A suitable number of latent spaces N' can be retained. While PCA is a highly effective and powerful compression tool, the main idea of PCA is still based on linear decomposition. As data complexity increases, the limitations of PCA become apparent. In such cases, more advanced DR techniques, like *MDS*, *kPCA*, and *autoencoder*, come into play.

[3] Sergio Valle, Weihua Li, and S Joe Qin. Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Industrial & Engineering Chemistry Research*, 38(11): 4389–4401, 1999; Edoardo Saccenti and José Camacho. Determining the number of components in principal components analysis: A comparison of statistical, crossvalidation and approximated methods. *Chemometrics and Intelligent Laboratory Systems*, 149:99–116, 2015; S Joe Qin and Ricardo Dunia. Determining the number of principal components for best reconstruction. *Journal of process control*, 10(2-3):245–250, 2000; and Baligh Mnassri, Bouchra Ananou, Mustapha Ouladsine, et al. A generalized variance of reconstruction error criterion for determining the optimum number of principal components. In *18th Mediterranean Conference on Control and Automation, MED'10*, pages 868–873. IEEE, 2010

References

Baligh Mnassri, Bouchra Ananou, Mustapha Ouladsine, et al. A generalized variance of reconstruction error criterion for determining the optimum number of principal components. In *18th Mediterranean Conference on Control and Automation, MED'10*, pages 868–873. IEEE, 2010.

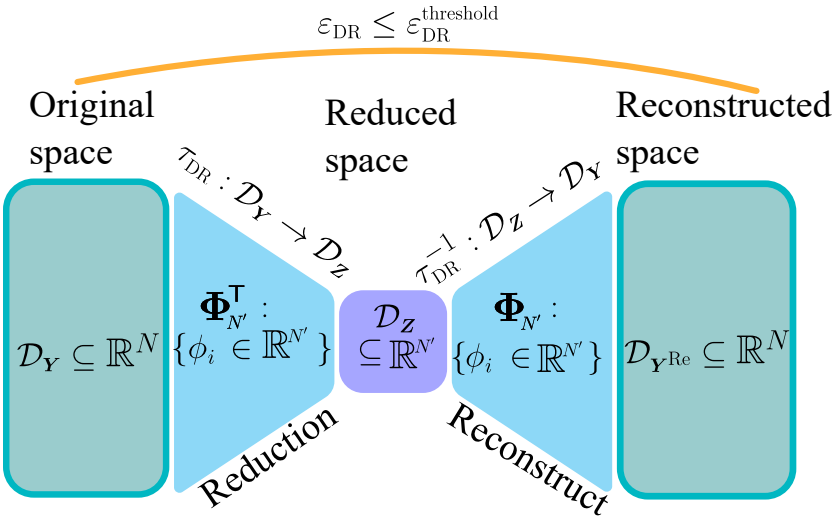

Figure 3: *DR-flowchart*

Joseph B Nagel, Jörg Rieckermann, and Bruno Sudret. Principal component analysis and sparse polynomial chaos expansions for global sensitivity analysis and model calibration: Application to urban drainage simulation. *Reliability Engineering & System Safety*, 195: 106737, 2020.

S Joe Qin and Ricardo Dunia. Determining the number of principal components for best reconstruction. *Journal of process control*, 10 (2-3):245–250, 2000.

Edoardo Saccenti and José Camacho. Determining the number of components in principal components analysis: A comparison of statistical, crossvalidation and approximated methods. *Chemometrics and Intelligent Laboratory Systems*, 149:99–116, 2015.

Sergio Valle, Weihua Li, and S Joe Qin. Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Industrial & Engineering Chemistry Research*, 38(11):4389–4401, 1999.

Manav Vohra, Paromita Nath, Sankaran Mahadevan, and Yung-Tsun Tina Lee. Fast surrogate modeling using dimensionality reduction in model inputs and field output: Application to additive manufacturing. *Reliability engineering & system safety*, 201:106986, 2020.

P-R Wagner, Reto Fahrni, Michael Klippel, Andrea Frangi, and Bruno Sudret. Bayesian calibration and sensitivity analysis of heat transfer models for fire insulation panels. *Engineering structures*, 205:110063, 2020.

Paul-Remo Wagner. *Stochastic Spectral Embedding in Forward and Inverse Uncertainty Quantification*. PhD thesis, ETH Zurich, 2021.