Principle Component Analysis

Songyan Hou

July 6, 2025

Abstract

This is an overview of principle component analysis in different aspects.

1 Principle Component Analysis

Let $\boldsymbol{X}$ random variable on $\mathbb{R}^d$, we would like to reduce the dimension of data while keeping information as much as possible.

1.1 Minimize residual

Denote the k dimension subspace $\mathcal{U} = \operatorname{span}\{u_1, \dots, u_k\}$ and the projection $\operatorname{proj}_{\mathcal{U}} \colon \mathbb{R}^d \to \mathbb{R}^k$. We minimize

$$\mathbb{E}\big[\|\boldsymbol{X} - \operatorname{proj}_{\mathcal{U}} \boldsymbol{X}\|^2\big]$$

By definition of projection, $\|\operatorname{proj}_{\mathcal{U}} \boldsymbol{X}\|^2 = \langle \boldsymbol{X}, \operatorname{proj}_{\mathcal{U}} \boldsymbol{X}\rangle$. Thus we get

$$\min_{\mathcal{U}} \mathbb{E}\big[\|\boldsymbol{X} - \operatorname{proj}_{\mathcal{U}} \boldsymbol{X}\|^2\big] = \min_{\mathcal{U}} \mathbb{E}\big[\|\boldsymbol{X}\|^2 + \|\operatorname{proj}_{\mathcal{U}} \boldsymbol{X}\|^2 - 2\langle \boldsymbol{X}, \operatorname{proj}_{\mathcal{U}} \boldsymbol{X}\rangle\big] = \min_{\mathcal{U}} \mathbb{E}\big[\|\boldsymbol{X}\|^2 - \|\operatorname{proj}_{\mathcal{U}} \boldsymbol{X}\|^2\big].$$

Equivalently, we only need to solve

$$\max_{\mathcal{U}} \mathbb{E}\big[\|\operatorname{proj}_{\mathcal{U}} \boldsymbol{X}\|^2\big].$$

1.2 Maximize variance

This turns minimizing residual into maximizing variance of the compressed data. Let $\operatorname{proj}_{\mathcal{U}} = x \mapsto A_{\mathcal{U}}x$. We have

$$\max_{\mathcal{U}} \mathbb{E}\big[\|A_{\mathcal{U}}\boldsymbol{X}\|^2\big] = \max_{\mathcal{U}} \mathbb{E}\big[\operatorname{tr}(A_{\mathcal{U}}\boldsymbol{X}\boldsymbol{X}^\top A_{\mathcal{U}}^\top)\big] = \max_{\mathcal{U}} \operatorname{tr}(A_{\mathcal{U}}\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]A_{\mathcal{U}}^\top) = \max_{\mathcal{U}} \operatorname{tr}(A_{\mathcal{U}}\Sigma A_{\mathcal{U}}^\top),$$

where $\Sigma := \operatorname{Cov}(\boldsymbol{X}) = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$ and the second equality is by linearity of expectation and trace operator. Finally, by von-Neumann inequality we get

$$\max_{\mathcal{U}} \operatorname{tr}(A_{\mathcal{U}}\Sigma A_{\mathcal{U}}^\top) = \max_{\mathcal{U}} \operatorname{tr}(A_{\mathcal{U}}^\top A_{\mathcal{U}}\Sigma) = \max_{\mathcal{U}} \operatorname{tr}((A_{\mathcal{U}}U)^\top (A_{\mathcal{U}}U)\Lambda) \le \sum_{i=1}^d \rho_i\sigma_i = \sum_{i=1}^k \sigma_i,$$

where the maximum is obtained by $A_{\mathcal{U}} = U_{1:k}U_{1:k}^\top$ where $U\Lambda U^\top$ is the symmetric decomposition of Σ.

Remark 1.1. PCA neither depends on Gaussian assumption, nor samples representation. PCA only depends on the covariance structure of data.

1.3 Samples and singular value decomposition

Now we consider $X = [x_1, \dots, x_n] \in \mathbb{R}^{d\times n}$ samples. The un-biased covariance estimator of X is given by

$$\Sigma = \frac{1}{n-1} XX^\top.$$

Thus plugging this into our discussion above would yield the solution. From another perspective, we can consider the singular value decomposition of

$$X = USV^\top, \quad U = [u_1, \dots, u_d], \quad V = [v_1, \dots, v_n]$$

Therefore, all x_i are represented by $u_1, \dots, u_d$:

$$[x_1, \dots, x_n] = [u_1, \dots, u_d]SV^\top.$$

Therefore, by picking the first k largest ones, we get the principle directions. The reconstructed $\hat{x}_i$ are

$$[\hat{x}_1, \dots, \hat{x}_n] = [u_1, \dots, u_k]I_{k,d}SV^\top = U_{1:k}U_{1:k}^\top USV^\top = U_{1:k}U_{1:k}^\top[x_1, \dots, x_n]$$

1.4 Probabilistic PCA

Let $\boldsymbol{Z} \sim \mathcal{N}(I_k)$, $\epsilon \sim \mathcal{N}(I_d)$, then $\boldsymbol{Y} = W\boldsymbol{Z} + \sigma\epsilon \sim \mathcal{N}(WW^\top + \sigma^2 I_d)$. Minimizing DL-divergence

$$\min D_{\mathrm{KL}}(\mu_{\boldsymbol{X}} | \mu_{\boldsymbol{Y}}) = \min \mathbb{E}[\log p_{\boldsymbol{X}}(\boldsymbol{X}) - \log p_{\boldsymbol{Y}}(\boldsymbol{X})]$$

is equivalent to maximizing log likelihood

$$\max \mathbb{E}[\log p_{\boldsymbol{Y}}(\boldsymbol{X})] \approx \max \log(|WW^\top + \sigma^2 I_d|) + \mathrm{tr}\left((WW^\top + \sigma^2 I_d)^{-1}\Sigma\right),$$

which has a closed-form solution $W = U_{1:k}(S_{1:k} - \sigma^2 I_k)^{\frac{1}{2}}\mathcal{O}$, where $USV^\top = X$ and $O \in \mathbb{R}^{k\times k}$ orthogonal; see [Bis06]. Similarly, you could replace the D_{KL} by entropic Wasserstein distance and get something similar; see [Col+23].

EM algorithm

We can view $\boldsymbol{Z}$ as missing data and solve it by EM: (E step) estimate $\mathbb{E}[\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top]$ with fixed W; (M step): find the optimal W given $\mathbb{E}[\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top]$; see [Row97].

Factor model

By assuming $\epsilon \sim \mathcal{N}(\Psi)$ where $\Psi \in \mathbb{R}^{d\times d}$ is diagonal, we have the factor model; see [Gho+21].

1.5 High-dimension regime

Computationally, we can use this trick: $XX^\top v = \lambda v \Rightarrow X^\top X(X^\top v) = \lambda(X^\top v)$ to reduce the computational dimension to $\min(d, n)$. Statistically, interesting things happens when $d = n$. In this case, law of large number fails and Σ is not a good enought estimate of $\mathbb{E}[\boldsymbol{X}]$. Given $\boldsymbol{X} \sim \mathcal{N}(0, I_n)$, the distribution of eigenvalues of Σ tends to Marchenko-Pastur distribution as $n \to \infty$.

References

[Bis06] Christopher M Bishop. "Pattern recognition and machine learning". In: *Springer google schola* 2 (2006), pp. 1122–1128.

[Col+23] Antoine Collas, Titouan Vayer, Rémi Flamary, and Arnaud Breloy. "Entropic Wasserstein component analysis". In: *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2023, pp. 1–6.

[Gho+21] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. "Factor analysis, probabilistic principal component analysis, variational inference, and variational autoencoder: Tutorial and survey". In: *arXiv preprint arXiv:2101.00734* (2021).

[Row97] Sam Roweis. "EM algorithms for PCA and SPCA". In: *Advances in neural information processing systems* 10 (1997).