Dimensionality Optimization and Spectral Thresholding in High-Dimensional Datasets

Mika Han*and Bona Han

The University of Hong Kong

Abstract

High-dimensional datasets suffer from sparsity, high computational cost and the curse of dimensionality. This paper presents a self-contained mathematical treatment of dimensionality reduction by principal component analysis (PCA). We first show, using Lagrange multipliers, that maximizing the variance of a unit-length projection reduces to the eigenvalue problem for the covariance matrix, so that the principal directions are its eigenvectors and the variance captured along each direction is the corresponding eigenvalue. We then derive the spectral decomposition of the sample covariance matrix and turn to the question of how many components to retain. Drawing on random matrix theory, we review the Wigner semicircle law and the Marchenko–Pastur law and use the Baik–Ben Arous–Péché phase transition to obtain a threshold that separates eigenvalues carrying signal from those attributable to noise.

1 Introduction

With the continued development of science and technology, the need for techniques to analyze high-dimensional data is growing in many fields, including computational science and engineering, machine learning, artificial intelligence, finance and economics, and management science. In this context, the *dimensions* of a dataset are simply the features, or variables, that describe each data point; for example, if the data concern people, possible dimensions include age, height, weight and income. Data are high-dimensional when the number of features is large relative to the number of observations. Such data are sparse, costly to process, and subject to the curse of dimensionality. Low-dimensional data, by contrast, have few features relative to the number of observations, so that patterns are clearer and algorithms run more efficiently.

In this paper, we study mathematically how the dimensionality of a high-dimensional dataset can be reduced in an optimal way. Section 2 derives the first principal component as the solution of a constrained variance-maximization problem. Section 3 obtains the remaining principal components from the spectral decomposition of the covariance matrix. Section 4 uses results from random matrix theory to determine how many components are statistically significant, and Section 5 concludes.

2 Variance Maximization and the First Principal Component

To reduce the number of dimensions without discarding important information, we seek directions along which the projected data have high variance. A projection onto a direction $\mathbf{w} \in \mathbb{R}^d$ with high variance retains more information, because the data points are more spread out along that axis.

*Corresponding author.

We show below that the direction of maximum variance is the eigenvector $\mathbf{w}_1$ of the covariance matrix associated with its largest eigenvalue λ_1.

2.1 Variance of a projection

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be the data points, where d is the number of dimensions, and regard them as realizations of a random vector $\mathbf{x}$ with mean $\boldsymbol{\mu}$. Our aim is to find the direction $\mathbf{w} \in \mathbb{R}^d$ along which the projected data have maximum variance, without the variance being inflated simply by increasing the length of $\mathbf{w}$. The projection of a data point $\mathbf{x}_i$ onto $\mathbf{w}$ is

$$\mathbf{w}^\top \mathbf{x}_i. \tag{1}$$

For a fixed, non-zero vector $\mathbf{w}$, the variance of the projection is

$$\operatorname{Var}\left(\mathbf{w}^\top \mathbf{x}\right) = \mathbb{E}\left[\left(\mathbf{w}^\top (\mathbf{x} - \boldsymbol{\mu})\right)^2\right]. \tag{2}$$

Since $\mathbf{a}^\top \mathbf{b}$ is a scalar for any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, we have the identity

$$\left(\mathbf{a}^\top \mathbf{b}\right)^2 = \left(\mathbf{b}^\top \mathbf{a}\right)\left(\mathbf{a}^\top \mathbf{b}\right) = \mathbf{b}^\top \mathbf{a}\mathbf{a}^\top \mathbf{b}. \tag{3}$$

Taking $\mathbf{a} = \mathbf{w}$ and $\mathbf{b} = \mathbf{x} - \boldsymbol{\mu}$, the variance can be rewritten as

$$\operatorname{Var}\left(\mathbf{w}^\top \mathbf{x}\right) = \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{w}\mathbf{w}^\top (\mathbf{x} - \boldsymbol{\mu})\right]. \tag{4}$$

The quantity inside the expectation is a scalar and is therefore equal to its own trace. By the cyclic property of the trace,

$$(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{w}\mathbf{w}^\top (\mathbf{x} - \boldsymbol{\mu}) = \operatorname{Tr}\left[\mathbf{w}\mathbf{w}^\top (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\right]. \tag{5}$$

Because $\mathbf{w}$ is deterministic and both the trace and the expectation are linear, $\mathbf{w}\mathbf{w}^\top$ can be taken outside the expectation:

$$\operatorname{Var}\left(\mathbf{w}^\top \mathbf{x}\right) = \operatorname{Tr}\left(\mathbf{w}\mathbf{w}^\top \, \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\right]\right). \tag{6}$$

The covariance matrix Σ is defined as

$$\Sigma = \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\right], \tag{7}$$

which is symmetric ($\Sigma = \Sigma^\top$) and positive semi-definite. The variance therefore simplifies to

$$\operatorname{Var}\left(\mathbf{w}^\top \mathbf{x}\right) = \operatorname{Tr}\left(\mathbf{w}\mathbf{w}^\top \Sigma\right). \tag{8}$$

Applying the cyclic property of the trace once more,

$$\operatorname{Tr}\left(\mathbf{w}\mathbf{w}^\top \Sigma\right) = \operatorname{Tr}\left(\mathbf{w}^\top \Sigma \mathbf{w}\right) = \mathbf{w}^\top \Sigma \mathbf{w}, \tag{9}$$

since $\mathbf{w}^\top \Sigma \mathbf{w}$ is a scalar. Hence

$$\operatorname{Var}\left(\mathbf{w}^\top \mathbf{x}\right) = \mathbf{w}^\top \Sigma \mathbf{w}, \tag{10}$$

which is a quadratic form in $\mathbf{w}$. Consequently, the variance depends quadratically on the magnitude of $\mathbf{w}$: replacing $\mathbf{w}$ by $\alpha\mathbf{w}$ for a scalar α gives

$$\operatorname{Var}\left((\alpha\mathbf{w})^\top \mathbf{x}\right) = \alpha^2 \operatorname{Var}\left(\mathbf{w}^\top \mathbf{x}\right) = \alpha^2 \, \mathbf{w}^\top \Sigma \mathbf{w}. \tag{11}$$

2.2 The unit-length constraint

The Euclidean norm of $\mathbf{w}$ is

$$\|\mathbf{w}\| = \sqrt{w_1^2 + w_2^2 + \cdots + w_d^2}, \tag{12}$$

which implies

$$\|\mathbf{w}\|^2 = \mathbf{w}^\top \mathbf{w} = w_1^2 + w_2^2 + \cdots + w_d^2. \tag{13}$$

In principal component analysis (PCA), the variance of the projection is maximized subject to the constraint

$$\mathbf{w}^\top \mathbf{w} = 1, \tag{14}$$

that is, $\mathbf{w}$ is required to have unit length. This constraint is necessary because the objective of PCA is to find the *direction* of greatest variance, not a vector of any particular length. By (11), the variance grows quadratically as $\mathbf{w}$ is scaled, so without a constraint, letting $\|\mathbf{w}\| \to \infty$ would send the variance to infinity and the optimization problem would have no solution. Since rescaling $\mathbf{w}$ does not change the direction it represents, we may normalize it to unit length without loss of generality, so that the optimization concerns only the direction that maximizes variance.

2.3 Constrained optimization

Let $f(\mathbf{w})$ denote the variance function,

$$f(\mathbf{w}) = \mathrm{Var}\left(\mathbf{w}^\top \mathbf{x}\right) = \mathbf{w}^\top \Sigma \mathbf{w}, \tag{15}$$

where $\Sigma = (a_{ij})$ is the covariance matrix with entries a_{ij}. The vector $\Sigma\mathbf{w}$ has components

$$(\Sigma\mathbf{w})_i = \sum_{j=1}^{d} a_{ij} w_j, \qquad i = 1, \ldots, d, \tag{16}$$

and taking the inner product with $\mathbf{w}$ yields the scalar

$$f(\mathbf{w}) = \mathbf{w}^\top \Sigma \mathbf{w} = \sum_{i=1}^{d} w_i \left(\sum_{j=1}^{d} a_{ij} w_j \right) = \sum_{i=1}^{d} \sum_{j=1}^{d} w_i a_{ij} w_j. \tag{17}$$

We now differentiate with respect to a component w_k. A term $w_i a_{ij} w_j$ depends on w_k only if $i = k$ or $j = k$. In the first case,

$$\frac{\partial}{\partial w_k} \left(w_k a_{kj} w_j \right) = a_{kj} w_j, \tag{18}$$

and in the second,

$$\frac{\partial}{\partial w_k} \left(w_i a_{ik} w_k \right) = w_i a_{ik}. \tag{19}$$

(The diagonal term $i = j = k$, namely $a_{kk} w_k^2$, has derivative $2a_{kk} w_k$, which is accounted for by one contribution from each case.) Summing over all i and j gives

$$\frac{\partial f}{\partial w_k} = \sum_{j=1}^{d} a_{kj} w_j + \sum_{i=1}^{d} w_i a_{ik}. \tag{20}$$

Collecting the partial derivatives for $k = 1, \ldots, d$, the gradient can be written in matrix notation as

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \Sigma\mathbf{w} + \Sigma^\top \mathbf{w}. \tag{21}$$

Since the covariance matrix is symmetric, this simplifies to

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = 2\Sigma\mathbf{w}. \tag{22}$$

To maximize $f(\mathbf{w})$ subject to (14), we introduce a Lagrange multiplier λ and form the Lagrangian

$$\mathcal{L}(\mathbf{w}, \lambda) = f(\mathbf{w}) - \lambda \left(\mathbf{w}^\top \mathbf{w} - 1 \right). \tag{23}$$

Differentiating $\mathcal{L}$ with respect to $\mathbf{w}$ gives

$$\nabla_{\mathbf{w}} \mathcal{L} = 2\Sigma\mathbf{w} - 2\lambda\mathbf{w}, \tag{24}$$

and setting the gradient equal to zero to find the critical points yields

$$2\Sigma\mathbf{w} - 2\lambda\mathbf{w} = \mathbf{0}, \tag{25}$$

which simplifies to

$$\Sigma\mathbf{w} = \lambda\mathbf{w}. \tag{26}$$

Thus every critical point $\mathbf{w}$ is an eigenvector of Σ, and the Lagrange multiplier λ is the corresponding eigenvalue. Substituting (26) into the variance function gives

$$\mathbf{w}^\top \Sigma\mathbf{w} = \mathbf{w}^\top (\lambda\mathbf{w}) = \lambda\, \mathbf{w}^\top \mathbf{w}, \tag{27}$$

and since $\mathbf{w}^\top \mathbf{w} = 1$,

$$\mathbf{w}^\top \Sigma\mathbf{w} = \lambda. \tag{28}$$

This derivation shows that maximizing the variance under the unit-length constraint reduces to finding an eigenvector of the covariance matrix. By (28), the variance of the projection onto an eigenvector equals its eigenvalue, so the maximum is attained at the eigenvector $\mathbf{w}_1$ associated with the largest eigenvalue λ_1. This eigenvector is the *first principal component*, and the eigenvalue λ_1 is the variance explained in that direction. More generally, each eigenvalue measures the variance captured along its eigenvector, so the largest eigenvalues identify the most significant sources of variation in the data.

3 Subsequent Principal Components

Consider a dataset of n observations on d features, arranged as the $n \times d$ data matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}, \tag{29}$$

in which each row is an observation and each column is a feature. Let $\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$ denote the mean of the jth feature across all observations, and define the $n \times d$ matrix of means

$$\bar{X} = \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_d \\ \mu_1 & \mu_2 & \cdots & \mu_d \\ \vdots & \vdots & \ddots & \vdots \\ \mu_1 & \mu_2 & \cdots & \mu_d \end{pmatrix}. \tag{30}$$

To remove the effect of the means on the result of PCA, we subtract $\bar{X}$ from X, which produces the centered data matrix

$$Y = X - \bar{X} = \begin{pmatrix} x_{11} - \mu_1 & x_{12} - \mu_2 & \cdots & x_{1d} - \mu_d \\ x_{21} - \mu_1 & x_{22} - \mu_2 & \cdots & x_{2d} - \mu_d \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \mu_1 & x_{n2} - \mu_2 & \cdots & x_{nd} - \mu_d \end{pmatrix}. \tag{31}$$

Denote the columns of Y (the centered features) by $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d \in \mathbb{R}^n$. The sample covariance matrix, which estimates Σ, is

$$\hat{\Sigma} = \mathrm{Cov}(Y) = \frac{1}{n-1} Y^\top Y, \tag{32}$$

or, entry by entry,

$$\hat{\Sigma} = \begin{pmatrix} \frac{1}{n-1}\mathbf{v}_1^\top \mathbf{v}_1 & \frac{1}{n-1}\mathbf{v}_1^\top \mathbf{v}_2 & \cdots & \frac{1}{n-1}\mathbf{v}_1^\top \mathbf{v}_d \\ \frac{1}{n-1}\mathbf{v}_2^\top \mathbf{v}_1 & \frac{1}{n-1}\mathbf{v}_2^\top \mathbf{v}_2 & \cdots & \frac{1}{n-1}\mathbf{v}_2^\top \mathbf{v}_d \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n-1}\mathbf{v}_d^\top \mathbf{v}_1 & \frac{1}{n-1}\mathbf{v}_d^\top \mathbf{v}_2 & \cdots & \frac{1}{n-1}\mathbf{v}_d^\top \mathbf{v}_d \end{pmatrix}. \tag{33}$$

The diagonal entries are the sample variances of the individual features and the off-diagonal entries are the sample covariances between pairs of features. The matrix $\hat{\Sigma}$ is square ($d \times d$), symmetric and positive semi-definite. By the spectral theorem for real symmetric matrices, all of its eigenvalues are real, its eigenvectors can be chosen to be orthonormal, and it can be factorized as

$$\hat{\Sigma} = PDP^\top, \tag{34}$$

where P is an orthogonal matrix whose columns are eigenvectors and D is a diagonal matrix of eigenvalues. Positive semi-definiteness further ensures that every eigenvalue is non-negative.

To obtain D, we find the eigenvalues of $\hat{\Sigma}$ by solving the eigenvalue problem (26) derived above,

$$\hat{\Sigma}\mathbf{w} = \lambda\mathbf{w} \quad \Longleftrightarrow \quad \hat{\Sigma}\mathbf{w} - \lambda\mathbf{w} = \mathbf{0}. \tag{35}$$

Since $\lambda\mathbf{w} = (\lambda I)\mathbf{w}$, where I is the $d \times d$ identity matrix, this can be rearranged as

$$(\hat{\Sigma} - \lambda I)\,\mathbf{w} = \mathbf{0}. \tag{36}$$

Hence $\mathbf{w}$ lies in the null space of $\hat{\Sigma} - \lambda I$. For a non-zero $\mathbf{w}$ to exist, this matrix must be singular, so its determinant must vanish. This gives the characteristic equation

$$\det(\hat{\Sigma} - \lambda I) = 0, \tag{37}$$

a polynomial equation of degree d in λ. Its roots are the eigenvalues, which we order as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. (For large d, the eigenvalues are computed in practice by numerical methods such as the QR algorithm rather than by finding polynomial roots.) We then form

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}. \tag{38}$$

Given the eigenvalues, the eigenvectors are found from (36). This homogeneous linear system can be solved by Gaussian elimination: reducing $\hat{\Sigma} - \lambda_j I$ to reduced row echelon form identifies

the free variables, in terms of which the eigenvectors can be expressed. Each eigenvector is then normalized to unit length, $\mathbf{w}_j^\top \mathbf{w}_j = 1$; where an eigenvalue is repeated, an orthonormal basis of its eigenspace is chosen (for instance, by the Gram–Schmidt process). The eigenvectors are collected as the columns of

$$P = \begin{pmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_d \end{pmatrix}, \tag{39}$$

where $\mathbf{w}_j$ corresponds to λ_j.

These eigenvectors are the subsequent principal components. Applying the argument of Section 2 with the additional constraints $\mathbf{w} \perp \mathbf{w}_1, \ldots, \mathbf{w}_{k-1}$ shows that the unit vector of maximum variance orthogonal to the first $k-1$ principal components is $\mathbf{w}_k$, and that the variance it captures is λ_k. Substituting P and D into (34) gives

$$\hat{\Sigma} = \begin{pmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_d \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix} \begin{pmatrix} \mathbf{w}_1^\top \\ \mathbf{w}_2^\top \\ \vdots \\ \mathbf{w}_d^\top \end{pmatrix}, \tag{40}$$

which can be rewritten as the sum of rank-one matrices

$$\hat{\Sigma} = \lambda_1 \mathbf{w}_1 \mathbf{w}_1^\top + \lambda_2 \mathbf{w}_2 \mathbf{w}_2^\top + \cdots + \lambda_d \mathbf{w}_d \mathbf{w}_d^\top. \tag{41}$$

For dimensionality reduction, PCA retains the terms with the largest eigenvalues and truncates the rest. This requires a threshold that distinguishes statistically significant eigenvalues from the remainder, which we determine in the next section.

4 Determining the Number of Significant Components

Random matrix theory describes how the eigenvalues of a matrix behave when its entries are pure noise. Eigenvalues that fall within the range predicted for noise can be discarded, while those lying clearly outside it indicate genuine structure in the data. We first review the classical Wigner semicircle law, which illustrates this idea, and then state the corresponding result for sample covariance matrices, which is the setting relevant to PCA.

4.1 Wigner matrices and the semicircle law

A (real) Wigner matrix is a large, symmetric $n \times n$ random matrix H whose entries on and above the diagonal are independent random variables with mean zero. When these entries are Gaussian, H belongs to the Gaussian Orthogonal Ensemble (GOE). If the off-diagonal entries have variance $1/n$, the empirical distribution of the eigenvalues of H converges, as $n \to \infty$, to the semicircle density of radius $R = 2$:

$$\rho_{\mathrm{sc}}(\lambda) = \frac{1}{2\pi}\sqrt{4 - \lambda^2}, \qquad \lambda \in [-2, 2]. \tag{42}$$

This result, due to Wigner [6], is a universality theorem: the limit does not depend on the particular distribution of the entries. In this respect it plays a role for random matrices similar to that of the central limit theorem for sums of random variables, although it is proved by different methods (for example, the method of moments).

For the unscaled GOE, in which the diagonal entries are $N(0,1)$ and the off-diagonal entries are $N(0, \frac{1}{2})$, the joint probability density of the eigenvalues is

$$p(\lambda_1, \ldots, \lambda_n) \propto \prod_{i<j} |\lambda_i - \lambda_j| \cdot \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \lambda_i^2 \right). \tag{43}$$

The product term expresses the mutual repulsion of the eigenvalues, which makes coincident eigenvalues unlikely, while the exponential term confines them near zero. The balance between these two effects spreads the eigenvalues over the interval $[-2\sqrt{n},\, 2\sqrt{n}]$; after rescaling by $1/\sqrt{n}$, their density converges to the semicircle on $[-2, 2]$ [3].

The semicircle law also explains how a signal can be detected. If a rank-one signal $\theta \mathbf{u}\mathbf{u}^\top$ with $\|\mathbf{u}\| = 1$ and $\theta > 0$ is added to a scaled Wigner matrix, then for $\theta \le 1$ the largest eigenvalue remains at the edge of the semicircle, at 2, and the signal cannot be distinguished from noise. For $\theta > 1$, however, the largest eigenvalue detaches from the bulk and converges to $\theta + 1/\theta > 2$ [4]. An eigenvalue lying outside the support of the semicircle therefore indicates a meaningful signal.

4.2 Sample covariance matrices and the Marchenko–Pastur law

PCA operates on the sample covariance matrix $\hat{\Sigma}$, not on a Wigner matrix. Since $\hat{\Sigma}$ is positive semi-definite, its eigenvalues are non-negative, and their distribution under noise is not the semicircle but the Marchenko–Pastur law [5]. Suppose the $n \times d$ data consist purely of independent noise with mean zero and variance σ^2, and let $n, d \to \infty$ with $d/n \to \gamma \in (0, 1]$. Then the empirical distribution of the eigenvalues of $\hat{\Sigma}$ converges to the density

$$\rho_{\mathrm{MP}}(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\sigma^2\gamma\lambda}, \qquad \lambda \in [\lambda_-, \lambda_+], \tag{44}$$

supported between the edges

$$\lambda_\pm = \sigma^2 \left(1 \pm \sqrt{\gamma}\right)^2 . \tag{45}$$

The analogue of the detection result above for covariance matrices is the phase transition of Baik, Ben Arous and Péché [1], extended to real-valued data by Baik and Silverstein [2]. Suppose the population covariance matrix has a small number of "spiked" eigenvalues $\ell_1 \ge \cdots \ge \ell_r > \sigma^2$, with all remaining eigenvalues equal to σ^2. If $\ell_k \le \sigma^2(1+\sqrt{\gamma})$, the corresponding sample eigenvalue converges to the upper edge λ_+ and cannot be distinguished from noise. If $\ell_k > \sigma^2(1 + \sqrt{\gamma})$, it separates from the bulk and converges to

$$\ell_k \left(1 + \frac{\gamma\sigma^2}{\ell_k - \sigma^2}\right) > \lambda_+ . \tag{46}$$

4.3 Truncation rule

These results yield a practical threshold. Sample eigenvalues no larger than $\lambda_+ = \sigma^2(1+\sqrt{d/n})^2$ are consistent with pure noise and can be truncated, whereas eigenvalues exceeding λ_+ are considered significant, and their eigenvectors are retained as principal directions. When the noise variance σ^2 is unknown, it must be estimated, for example from the bulk of the eigenvalue spectrum.

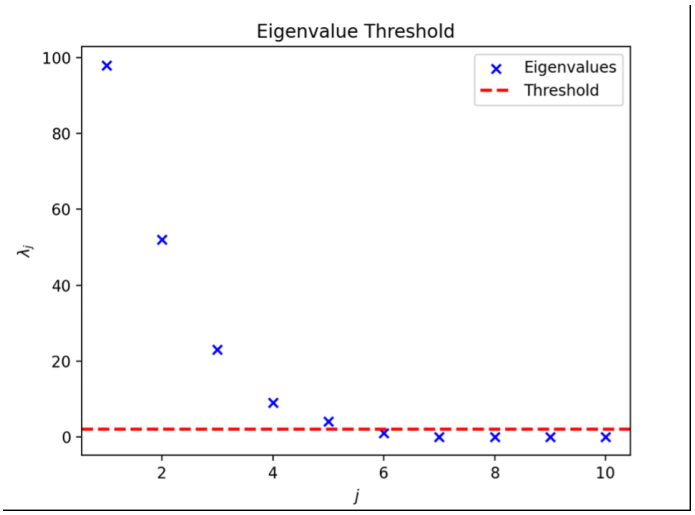


Figure 1: Illustrative scree plot of sample eigenvalues λ_j against their index j. The dashed line marks the noise threshold; the first five eigenvalues exceed it and are retained.

Figure 1 shows a case in which the first five eigenvalues exceed the threshold. The covariance matrix is then approximated by the first five terms of (41):

$$\hat{\Sigma} \approx \lambda_1 \mathbf{w}_1 \mathbf{w}_1^\top + \lambda_2 \mathbf{w}_2 \mathbf{w}_2^\top + \cdots + \lambda_5 \mathbf{w}_5 \mathbf{w}_5^\top. \tag{47}$$

In general, if r eigenvalues exceed the threshold, the retained eigenvectors $W_r = \begin{pmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_r \end{pmatrix}$ form a new basis onto which the centered data are projected:

$$Z = YW_r \in \mathbb{R}^{n \times r}. \tag{48}$$

This final step of PCA identifies the dimensions of greatest significance and discards those attributable to noise. As a result, the data are represented in a lower-dimensional space that retains most of their variance, as illustrated in Figure 2.

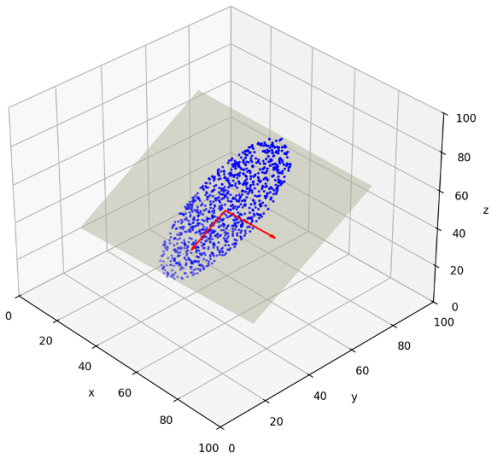


Figure 2: Projection of three-dimensional data onto the two-dimensional subspace (plane) spanned by the two leading eigenvectors (arrows).

5 Conclusion

In this paper, we showed that maximizing the variance of a unit-length projection reduces to finding an eigenvector of the covariance matrix, and that the maximum is attained at the eigenvector associated with the largest eigenvalue. This result provides the theoretical justification for PCA: the principal directions are the eigenvectors of the covariance matrix, and the variance captured along each direction is the corresponding eigenvalue.

We then examined the spectral behavior of large random matrices to decide how many components to retain. The eigenvalues of a Wigner matrix follow the semicircle law on $[-2, 2]$ as $n \to \infty$, and the eigenvalues of a pure-noise sample covariance matrix follow the Marchenko–Pastur law with upper edge $\sigma^2(1 + \sqrt{d/n})^2$. By the Baik–Ben Arous–Péché phase transition, eigenvalues above this edge indicate signal, while those within the bulk are attributable to noise. Discarding the latter and projecting the data onto the eigenvectors of the former reduces the number of dimensions of the original dataset while preserving most of its variance.

References

[1] Baik, J., Ben Arous, G., & Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, *33*(5), 1643–1697. https://doi.org/10.48550/arXiv.math/0403022

[2] Baik, J., & Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, *97*(6), 1382–1408.

[3] Benaych-Georges, F., & Knowles, A. (2016). *Lectures on the local semicircle law for Wigner matrices* (arXiv:1601.04055). https://doi.org/10.48550/arXiv.1601.04055

[4] Féral, D., & Péché, S. (2007). The largest eigenvalue of rank one deformation of large Wigner matrices. *Communications in Mathematical Physics*, *272*(1), 185–228.

[5] Marčenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, *1*(4), 457–483.

[6] Wigner, E. P. (1958). On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, *67*(2), 325–327.

[7] Yu, P. (n.d.). *Lecture 05: Principal components analysis (Sections 12.2–3 & Section 6.3)* [Lecture slides]. HKU Business School, The University of Hong Kong. Retrieved December 19, 2025, from https://pweb.fbe.hku.hk/~pingyu/6086/Lecture5_Principal%20Components%20Analysis_ver2.pdf