Глава 7 Снижение размерности

В главе 6 мы рассмотрели некоторые характеристики многомерных данных, которые подчас противоречат здравому смыслу. Например, в пространствах высокой размерности точки чаще рассредоточены по поверхности пространства или его углам, а центр практически пуст. Можно наблюдать явное увеличение ортогональных осей. Как следствие, данные высокой размерности создают определённые проблемы для анализа, хотя в некоторых применениях (например, в задачах классификации) высокие размерности могут быть полезны. Тем не менее, важно проверить, возможно ли понижение размерности без потери основных свойств исходной матрицы. Это может помочь как в интеллектуальном анализе данных, так и при их визуализации. В этой главе мы изучим методы, которые позволяют получить оптимальные проекции данных на меньшие размерности.

7.1 Предпосылки

Пусть данные представляют из себя матрицу $\mathbf{D}$ размера $n \times d$ (n точек с d признаками):

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right).$$

Каждая точка $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$ — это вектор в d-мерном пространстве с базисом $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d$, в котором компонент $\mathbf{e}_i$ соотносится с i-тым признаком X_i. Напомним, что стандартным базисом называется ортонормированный базис для пространства данных, то есть, базисные векторы попарно ортогональны: $\mathbf{e}_i^T \mathbf{e}_j = 0$, и имеют единичную длину: $\|\mathbf{e}_i\| = 1$.

Таким образом, для любого набора d ортонормированных векторов $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d$ такого, что $\mathbf{u}_i^T \mathbf{u}_j = 0$ и $\|\mathbf{u}_i\| = 1$ (или $\mathbf{u}_i^T \mathbf{u}_i = 1$), мы можем представить каждую точку $\mathbf{x}$ как линейную комбинацию:

$$\mathbf{x} = a_1 \mathbf{u}_1 + a_2 \mathbf{u}_2 + \cdots + a_d \mathbf{u}_d, \tag{7.1}$$

где вектор $\mathbf{a} = (a_1, a_2, \ldots, a_d)^T$ — это координаты $\mathbf{x}$ в новом базисе. Линейную комбинацию выше можно записать в виде умножения матриц:

$$\mathbf{x} = \mathbf{U}\mathbf{a}, \tag{7.2}$$

где $\mathbf{U}$ — матрица размера $d \times d$, i-ый столбец которой содержит i-тый базисный вектор:

$$\mathbf{U} = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ | & | & & | \end{pmatrix}.$$

Матрица $\mathbf{U}$ является *ортогональной*, все её столбцы, будучи базисными векторами, *ортогональны*, то есть, они попарно ортогональны и имеют единичную длину:

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, \text{если } i = j \\ 0, \text{если } i \neq j \end{cases}.$$

Поскольку $\mathbf{U}$ ортогональна, её обратная матрица равна её транспонированной:

$$\mathbf{U}^{-1} = \mathbf{U}^T,$$

следовательно: $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, где $\mathbf{I}$ — единичная матрица размера $d \times d$.

Умножив ур. (7.2) слева на $\mathbf{U}^T$, получим выражение для вычисления координат $\mathbf{x}$ в новом базисе:

$$\begin{aligned} \mathbf{U}^T\mathbf{x} &= \mathbf{U}^T\mathbf{U}\mathbf{a} \\ \mathbf{a} &= \mathbf{U}^T\mathbf{x}. \end{aligned} \tag{7.3}$$

Поскольку теоретически существует бесконечное число наборов ортонормальных базисных векторов, возникает вполне естественный вопрос: существует ли оптимальный базис для конкретного толкования оптимальности? Кроме того, часто входная размерность d достаточно велика, что может вызвать различные проблемы из-за проклятия размерности (см. главу 6). Вполне естественно поставить вопрос о поиске пространства меньшей размерности, которое сохраняло бы основные характеристики данных. То есть, нам интересно найти r-мерное представление $\mathbf{D}$ такое, чтобы $r \ll d$, или, иными словами, рассматривая произвольную точку $\mathbf{x}$ и полагая, что базисные векторы отсортированы в порядке уменьшения значимости, мы могли бы усечь выражение 7.1 до r членов и получить следующее:

$$\mathbf{x}' = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_r\mathbf{u}_r = \sum_{i=1}^{r} a_i\mathbf{u}_i\,. \tag{7.4}$$

Здесь $\mathbf{x}'$ — это проекция $\mathbf{x}$ на первые r базисных векторов, что в терминах матричных операций можно записать следующим образом:

$$\mathbf{x}' = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r \\ | & | & & | \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_r \end{pmatrix} = \mathbf{U}_r\mathbf{a}_r, \tag{7.5}$$

где $\mathbf{U}_r$ — матрица, состоящая из r первых базисных векторов, а $\mathbf{a}_r$ — вектор, содержащий первые r координат. Далее, ограничив ур-е 7.3 первыми r членами, получим:

$$\mathbf{a}_\mathrm{r} = \mathbf{U}_r^T\mathbf{x}. \tag{7.6}$$

Подставив 7.6 в 7.5, можно получить окончательное выражение для проекции $\mathbf{x}$ на первые r базисных векторов:

$$\mathbf{x}' = \mathbf{U}_r\mathbf{U}_r^T\mathbf{x} = \mathbf{P}_r\mathbf{x}, \tag{7.7}$$

где $\mathbf{P}_r = \mathbf{U}_r\mathbf{U}_r^T$ — *ортогональная проекционная матрица* для подпространства над первыми r базисными векторами. То есть, $\mathbf{P}_r$ симметрична, и $\mathbf{P}_r^2 = \mathbf{P}_r$. Это легко

проверить, поскольку: $\mathbf{P}_r^T = (\mathbf{U}_r\mathbf{U}_r^T)^T = \mathbf{U}_r\mathbf{U}_r^T = \mathbf{P}_r$ и $\mathbf{P}_r^2 = (\mathbf{U}_r\mathbf{U}_r^T)(\mathbf{U}_r\mathbf{U}_r^T) = \mathbf{U}_r\mathbf{U}_r^T = \mathbf{P}_r$. Здесь мы использовали тот факт, что $\mathbf{U}_r^T\mathbf{U}_r = \mathbf{I}_{r\times r}$, единичная матрица размера $r \times r$. Проекционную матрицу $\mathbf{P}_r$ можно легко записать как декомпозицию:

$$\mathbf{P}_r = \mathbf{U}_r\mathbf{U}_r^T = \sum_{i=1}^{r} \mathbf{u}_i\mathbf{u}_i^T . \tag{7.8}$$

Из выражений 7.1 и 7.4 понятно, что проекция $\mathbf{x}$ на оставшиеся размерности составляет *вектор ошибки*:

$$\epsilon = \sum_{i=r+1}^{d} a_i\mathbf{u}_i = \mathbf{x} - \mathbf{x}'.$$

Следует заметить, что $\mathbf{x}'$ и ϵ — ортогональные векторы:

$$\mathbf{x}'^T\epsilon = \sum_{i=1}^{r}\sum_{j=r+1}^{d} a_i a_j\, \mathbf{u}_i^T\mathbf{u}_j = 0,$$

что следует из ортонормированности базиса. Более того, мы можем сделать ещё более сильное утверждение: подпространство над первыми r базисными векторами:

$$S_r = span(\mathbf{u}_1, \dots, \mathbf{u}_r)$$

и подпространство над оставшимися базисными векторами:

$$S_{d-r} = span(\mathbf{u}_{r+1}, \dots, \mathbf{u}_d)$$

являются *ортогональными подпространствами*, то есть, любая пара векторов $\mathbf{x} \in S_r$ и $\mathbf{y} \in S_{d-r}$ должна быть ортогональной. Подпространство S_{d-r} называют также *ортогональным дополнением* S_r.

Цель понижения размерности — найти такой r-мерный базис, который даст наилучшее приближение $\mathbf{x}_i'$ для всех точек $\mathbf{x}_i \in \mathbf{D}$, либо уменьшить ошибку $\epsilon = \mathbf{x} - \mathbf{x}'$ для всех точек.

7.2 Метод главных компонент

Метод главных компонент (PCA) – это метод, который ищет r-мерный базис, наилучшим образом описывающий дисперсию данных. Направление с наибольшей проецируемой дисперсией называется первым главным компонентом. Ортогональное направление, которое захватывает вторую по величине проецируемую дисперсию, называется вторым главным компонентом и т.д. Как мы увидим, направление, которое максимизирует дисперсию, также минимизирует среднеквадратичную ошибку.

7.2.1 Наилучшая линейная аппроксимация

Мы начнем с $r = 1$, т.е. с одномерного подпространства или прямой $\boldsymbol{u}$, которая наилучшим образом аппроксимирует $\boldsymbol{D}$, с точки зрения дисперсии проецируемых точек. Это приведет к общей методике PCA для наилучшего $1 \leq r \leq d$ размерного базиса для $\boldsymbol{D}$.

Без потери общности, мы предполагаем, что $\boldsymbol{u}$ имеет величину $\|\boldsymbol{u}\|^2 = \boldsymbol{u}^T\boldsymbol{u} = 1$; в противном случае можно продолжать увеличивать проецируемую дисперсию, просто увеличивая величину $\boldsymbol{u}$. Мы также предполагаем, что данные центрированы так, что среднее значение $\boldsymbol{\mu} = \boldsymbol{0}$.

Проекция $\boldsymbol{x_i}$ на вектор $\boldsymbol{u}$ берется как

$$\boldsymbol{x_i'} = \left(\frac{\boldsymbol{u}^T\boldsymbol{x}_i}{\boldsymbol{u}^T\boldsymbol{u}}\right)\boldsymbol{u} = (\boldsymbol{u}^T\boldsymbol{x}_i)\boldsymbol{u} = a_i\boldsymbol{u}$$

где скаляр

$$a_i = \boldsymbol{u}^T\boldsymbol{x}_i$$

дает координату $\boldsymbol{x_i'}$ вдоль прямой $\boldsymbol{u}$. Заметим, что поскольку средняя точка $\boldsymbol{\mu} = \boldsymbol{0}$, ее координата вдоль $\boldsymbol{u}$ равна $\mu_{\boldsymbol{u}} = 0$.

Необходимо выбрать направление $\boldsymbol{u}$ так, чтобы дисперсия проецируемых точек была максимальной. Проецируемая дисперсия вдоль $\boldsymbol{u}$ берется как:

$$\begin{aligned}
\sigma_{\boldsymbol{u}}^2 &= \frac{1}{n}\sum_{i=1}^{n}(a_i - \mu_{\boldsymbol{u}})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{u}^T\boldsymbol{x}_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}^T(\boldsymbol{x}_i\boldsymbol{x}_i{}^T)\boldsymbol{u} \\
&= \boldsymbol{u}^T\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i{}^T\right)\boldsymbol{u} \\
&= \boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}
\end{aligned} \tag{7.9}$$

где $\boldsymbol{\Sigma}$ – это ковариационная матрица для центрированных данных $\boldsymbol{D}$.

Чтобы максимизировать проецируемую дисперсию, мы должны решить задачу ограниченной оптимизации, а именно, чтобы максимизировать $\sigma_{\boldsymbol{u}}^2$ с учетом ограничения $\boldsymbol{u}^T\boldsymbol{u} = 1$. Это может быть решено путем введения множителя Лагранжа α для ограничения, чтобы получить задачу безусловной максимизации.

$$\max_{\boldsymbol{u}} J(\boldsymbol{u}) = \boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u} - \alpha(\boldsymbol{u}^T\boldsymbol{u} - 1) \tag{7.10}$$

Приравнивая производную $J(\boldsymbol{u})$ по $\boldsymbol{u}$ к нулевому вектору, получаем

$$\frac{\partial}{\partial \boldsymbol{u}} J(\boldsymbol{u}) = \mathbf{0}$$

$$\frac{\partial}{\partial \boldsymbol{u}} \left(\boldsymbol{u}^T \boldsymbol{\Sigma} \boldsymbol{u} - \alpha (\boldsymbol{u}^T \boldsymbol{u} - 1) \right) = \mathbf{0}$$

$$2\boldsymbol{\Sigma} \boldsymbol{u} - \mathbf{2}\alpha \boldsymbol{u} = \mathbf{0}$$

$$\boldsymbol{\Sigma} \boldsymbol{u} = \alpha \boldsymbol{u} \tag{7.11}$$

Это означает, что α является собственным значением ковариационной матрицы $\boldsymbol{\Sigma}$, с соответствующим собственным вектором $\boldsymbol{u}$. Далее, взяв скалярное произведение с $\boldsymbol{u}$ по обе стороны от уравнения (7.11) получаем

$$\boldsymbol{u}^T \boldsymbol{\Sigma} \boldsymbol{u} = \boldsymbol{u}^T \alpha \boldsymbol{u}$$

Из уравнения (7.9) имеем

$$\sigma_{\boldsymbol{u}}^2 = \alpha \boldsymbol{u}^T \boldsymbol{u}$$

$$\text{или } \sigma_{\boldsymbol{u}}^2 = \alpha \tag{7.12}$$

Таким образом, чтобы максимизировать проецируемую дисперсию $\sigma_{\boldsymbol{u}}^2$, необходимо выбрать наибольшее собственное значение матрицы $\boldsymbol{\Sigma}$. Другими словами, доминирующий собственный вектор $\boldsymbol{u}_1$ задает направление наибольшей дисперсии, также называемое первым главным компонентом, т.е. $\boldsymbol{u} = \boldsymbol{u}_1$. Кроме того, наибольшее собственное значение λ_1 определяет проецируемую дисперсию, т.е. $\sigma_{\boldsymbol{u}}^2 = \alpha = \lambda_1$.

Минимальная квадратичная ошибка

Теперь покажем, что направление, которое максимизирует проецируемую дисперсию, также минимизирует среднеквадратичную ошибку. Как и раньше, предположим, что набор данных $\boldsymbol{D}$, был центрирован путем вычитания среднего значения из каждой точки. Для точки $\boldsymbol{x}_i \in \boldsymbol{D}$ пусть $\boldsymbol{x}_i'$ обозначает ее проекцию вдоль направления $\boldsymbol{u}$, и пусть $\boldsymbol{\epsilon}_i = \boldsymbol{x}_i - \boldsymbol{x}_i'$ обозначает вектор ошибки. Среднеквадратичная ошибка (MSE) условия оптимизации определяется как

$$MSE(\boldsymbol{u}) = \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{\epsilon}_i\|^2 \tag{7.13}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x}_i'\|^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{x}_i')^T (\boldsymbol{x}_i - \boldsymbol{x}_i')$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(\|\boldsymbol{x}_i\|^2 - 2\boldsymbol{x}_i^T \boldsymbol{x}_i' + (\boldsymbol{x}_i')^T \boldsymbol{x}_i' \right) \tag{7.14}$$

Заметив, что $\boldsymbol{x}_i' = (\boldsymbol{u}^T \boldsymbol{x}_i)\boldsymbol{u}$, имеем

$$= \frac{1}{n} \sum_{i=1}^{n} \left(\|\boldsymbol{x}_i\|^2 - 2\boldsymbol{x}_i^T (\boldsymbol{u}^T \boldsymbol{x}_i)\boldsymbol{u} + ((\boldsymbol{u}^T \boldsymbol{x}_i)\boldsymbol{u})^T (\boldsymbol{u}^T \boldsymbol{x}_i)\boldsymbol{u} \right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\|\boldsymbol{x}_i\|^2 - 2(\boldsymbol{u}^T\boldsymbol{x}_i)(\boldsymbol{x}_i^T\boldsymbol{u}) + (\boldsymbol{u}^T\boldsymbol{x}_i)(\boldsymbol{x}_i^T\boldsymbol{u})\boldsymbol{u}^T\boldsymbol{u}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\|\boldsymbol{x}_i\|^2 - (\boldsymbol{u}^T\boldsymbol{x}_i)(\boldsymbol{x}_i^T\boldsymbol{u})\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^2 - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}^T(\boldsymbol{x}_i\boldsymbol{x}_i^T)\boldsymbol{u}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^2 - \boldsymbol{u}^T\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{u}$$

$$= \sum_{i=1}^{n}\frac{\|\boldsymbol{x}_i\|^2}{\boldsymbol{n}} - \boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u} \tag{7.15}$$

Заметим, что из уравнения (1.4) общая дисперсия центрированных данных (т.е. $\boldsymbol{\mu} = \mathbf{0}$) получается как

$$\mathrm{var}(\boldsymbol{D}) = \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i - \mathbf{0}\|^2 = \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^2$$

Далее, из уравнения (2.28) имеем

$$\mathrm{var}(\boldsymbol{D}) = tr(\boldsymbol{\Sigma}) = \sum_{i=1}^{d}\sigma_i^2$$

Таким образом, можно переписать уравнение (7.15) как

$$MSE(\boldsymbol{u}) = \mathrm{var}(\boldsymbol{D}) - \boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u} = \sum_{i=1}^{d}\sigma_i^2 - \boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}$$

Поскольку первый член $\mathrm{var}(\boldsymbol{D})$ является константой для данного набора данных $\boldsymbol{D}$, вектор $\boldsymbol{u}$, который минимизирует $MSE(\boldsymbol{u})$, максимизирует второй член, проецируемую дисперсию $\boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}$. Поскольку мы знаем, что $\boldsymbol{u}_1$, доминирующий собственный вектор матрицы $\boldsymbol{\Sigma}$, максимизирует проецируемую дисперсию, имеем

$$MSE(\boldsymbol{u}_1) = \mathrm{var}(\boldsymbol{D}) - \boldsymbol{u}_1{}^T\boldsymbol{\Sigma}\boldsymbol{u}_1 = \mathrm{var}(\boldsymbol{D}) - \boldsymbol{u}_1{}^T\lambda_1\boldsymbol{u}_1 = \mathrm{var}(\boldsymbol{D}) - \lambda_1 \tag{7.16}$$

Таким образом, главный компонент $\boldsymbol{u}_1$, являющийся направлением, которое максимизирует проецируемую дисперсию, также является направлением, которое минимизирует среднеквадратичную ошибку MSE.

7.2.2 Наилучшая двумерная аппроксимация

Теперь нас интересует наилучшее двумерное приближение к $\boldsymbol{D}$. Как и раньше, предположим, что $\boldsymbol{D}$ уже центрирован, т.ч. $\boldsymbol{\mu} = \mathbf{0}$. Мы уже вычислили направление с наибольшей дисперсией, а именно $\boldsymbol{u}_1$, которое является собственным вектором с соответствующим наибольшим собственным значением λ_1 матрицы $\boldsymbol{\Sigma}$. Теперь мы хотим найти другое направление $\mathbf{v}$, которое также максимизирует проецируемую дисперсию, но ортогонально $\boldsymbol{u}_1$. Согласно формуле (7.9) проецируемая на $\mathbf{v}$ дисперсия определяется как

$$\sigma_{\mathbf{v}}^2 = \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}$$

Далее мы потребуем, чтобы $\mathbf{v}$ был единичным вектором, ортогональным $\boldsymbol{u}_1$, т.е.

$$\mathbf{v}^T\boldsymbol{u}_1 = 0$$

$$\mathbf{v}^T\mathbf{v} = 1$$

Тогда условие оптимизации определяется как

$$\max_{\mathbf{v}} J(\mathbf{v}) = \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} - \alpha(\mathbf{v}^T\mathbf{v} - 1) - \beta(\mathbf{v}^T\boldsymbol{u}_1 - 0) \tag{7.17}$$

Взяв производную $J(\mathbf{v})$ по $\mathbf{v}$ и прировняв ее к нулевому вектору, получаем

$$2\boldsymbol{\Sigma}\mathbf{v} - 2\alpha\mathbf{v} - \beta\boldsymbol{u}_1 = \mathbf{0} \tag{7.18}$$

Если умножить левую часть на $\boldsymbol{u}_1^T$, то получим

$$2\boldsymbol{u}_1^T\boldsymbol{\Sigma}\mathbf{v} - 2\alpha\boldsymbol{u}_1^T\mathbf{v} - \beta\boldsymbol{u}_1^T\boldsymbol{u}_1 = 0$$

$2\mathbf{v}^T\boldsymbol{\Sigma}\boldsymbol{u}_1 - \beta = 0$, откуда следует, что

$$\beta = 2\mathbf{v}^T\lambda_1\boldsymbol{u}_1 = 2\lambda_1\mathbf{v}^T\boldsymbol{u}_1 = 0$$

При выводе выше мы использовали тот факт, что $\boldsymbol{u}_1^T\boldsymbol{\Sigma}\mathbf{v} = \mathbf{v}^T\boldsymbol{\Sigma}\boldsymbol{u}_1$, и что $\mathbf{v}$ ортогонален $\boldsymbol{u}_1$. Подставляя $\beta = 0$ в уравнение (7.18), получаем

$$2\boldsymbol{\Sigma}\mathbf{v} - 2\alpha\mathbf{v} = \mathbf{0}$$

$$\boldsymbol{\Sigma}\mathbf{v} = \alpha\mathbf{v}$$

Это означает, что $\mathbf{v}$ – это еще один собственный вектор $\boldsymbol{\Sigma}$. Так же, как в уравнении (7.12), имеем $\sigma_{\mathbf{v}}^2 = \alpha$. Чтобы максимизировать дисперсию по $\mathbf{v}$, мы должны выбрать $\alpha = \lambda_2$, второе по величине собственное значение $\boldsymbol{\Sigma}$, причем второй главный компонент задается соответствующим собственным вектором, т.е. $\mathbf{v} = \boldsymbol{u}_2$.

Общая проецируемая дисперсия

Пусть $\boldsymbol{U}_2$ – матрица, столбцы которой соответствуют двум главным компонентам

$$\boldsymbol{U}_2 = \begin{pmatrix} | & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 \\ | & | \end{pmatrix}$$

Для точки $\boldsymbol{x}_i \in \boldsymbol{D}$ координаты в двумерном подпространстве, натянутом на $\boldsymbol{u}_1$ и $\boldsymbol{u}_2$, могут быть вычислены по формуле (7.6) следующим образом

$$\boldsymbol{a}_i = \boldsymbol{U}_2^T\boldsymbol{x}_i$$

Предположим, что каждая точка $\boldsymbol{x}_i \in \mathbb{R}^d$ в $\boldsymbol{D}$ была спроецирована для получения ее координат $\boldsymbol{a}_i \in \mathbb{R}^d$, что дало новый набор данных $\boldsymbol{A}$. Кроме того, поскольку предполагается, что $\boldsymbol{D}$ центрирован ($\boldsymbol{\mu} = \mathbf{0}$), координаты спроецированного среднего также равны нулю, т.к. $\boldsymbol{U}_2\boldsymbol{\mu} = \boldsymbol{U}_2^T\mathbf{0} = \mathbf{0}$.

Общая дисперсия для $\boldsymbol{A}$ определяется как

$$var(\boldsymbol{A}) = \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{a}_i - \mathbf{0}\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{U}_2^T\boldsymbol{x}_i)^T\,(\boldsymbol{U}_2^T\boldsymbol{x}_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^T (\boldsymbol{U}_2 \boldsymbol{U}_2^T)\boldsymbol{x}_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{P}_2 \boldsymbol{x}_i \tag{7.19}$$

Где $\boldsymbol{P}_2$ – матрица ортогональной проекции [уравнение (7.8)], которая определяется как

$$\boldsymbol{P}_2 = \boldsymbol{U}_2 \boldsymbol{U}_2^T = \boldsymbol{u}_1 \boldsymbol{u}_1^T + \boldsymbol{u}_2 \boldsymbol{u}_2^T$$

Подставляя это в (7.19), проецируемая общая дисперсия определяется как

$$var(\boldsymbol{A}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{P}_2 \boldsymbol{x}_i \tag{7.20}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^T (\boldsymbol{u}_1 \boldsymbol{u}_1^T + \boldsymbol{u}_2 \boldsymbol{u}_2^T)\boldsymbol{x}_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{u}_1^T \boldsymbol{x}_i)(\boldsymbol{x}_i^T \boldsymbol{u}_1) + \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{u}_2^T \boldsymbol{x}_i)(\boldsymbol{x}_i^T \boldsymbol{u}_2)$$

$$= \boldsymbol{u}_1^T \boldsymbol{\Sigma} \boldsymbol{u}_1 + \boldsymbol{u}_2^T \boldsymbol{\Sigma} \boldsymbol{u}_2 \tag{7.21}$$

Поскольку $\boldsymbol{u}_1$ и $\boldsymbol{u}_2$ – собственные вектора $\boldsymbol{\Sigma}$, имеем $\boldsymbol{\Sigma}\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1$ и $\boldsymbol{\Sigma}\boldsymbol{u}_2 = \lambda_2 \boldsymbol{u}_2$, т.е.

$$var(\boldsymbol{A}) = \boldsymbol{u}_1^T \boldsymbol{\Sigma} \boldsymbol{u}_1 + \boldsymbol{u}_2^T \boldsymbol{\Sigma} \boldsymbol{u}_2 = \boldsymbol{u}_1^T \lambda_1 \boldsymbol{u}_1 + \boldsymbol{u}_2^T \lambda_2 \boldsymbol{u}_2 = \lambda_1 + \lambda_2 \tag{7.22}$$

Таким образом, сумма собственных значений представляет собой общую дисперсию проецируемых точек, а первые два главных компонента максимизируют эту дисперсию.

Среднеквадратичная ошибка

Теперь мы продемонстрируем, что первые два главных компонента также минимизируют среднеквадратичную ошибку. Среднеквадратичная ошибка задается как

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x}_i'\|^2$$

$$= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\big(\|\boldsymbol{x}_i\|^2 - 2\boldsymbol{x}_i^T \boldsymbol{x}_i' + (\boldsymbol{x}_i')^T \boldsymbol{x}_i'\big)$$, используя (7.14)

$$= var(\boldsymbol{D}) + \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(-2\boldsymbol{x}_i^T \boldsymbol{P}_2 \boldsymbol{x}_i + (\boldsymbol{P}_2 \boldsymbol{x}_i)^T \boldsymbol{P}_2 \boldsymbol{x}_i)$$, используя (7.7), что $\boldsymbol{x}_i' = \boldsymbol{P}_2 \boldsymbol{x}_i$

$$= var(\boldsymbol{D}) - \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i^T \boldsymbol{P}_2 \boldsymbol{x}_i)$$

$$= var(\boldsymbol{D}) - var(\boldsymbol{A}), \text{используя } (7.20) \tag{7.23}$$

Таким образом, MSE минимизируется именно тогда, когда общая проецируемая дисперсия $var(\boldsymbol{A})$ максимальна. Из (7.22) имеем

$$MSE = var(\boldsymbol{D}) - \lambda_1 - \lambda_2$$

7.2.3 Наилучшая r-мерная аппроксимация

Теперь нас интересует наилучшая r-мерная аппроксимация к $\boldsymbol{D}$, где $2 < r \leq d$. Предположим, что мы уже вычислили первые $j-1$ главных компонент или собственных векторов, $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{j-1}$,

соответствующих $j-1$ наибольшим собственным значениям матрицы $\mathbf{\Sigma}$, для $1 \leq j \leq r$. Чтобы вычислить j-й новый базисный вектор $\mathbf{v}$, мы должны убедиться, что он нормализован до единичной длины, т.е. $\mathbf{v}^T\mathbf{v} = 1$, и ортогонален всем предыдущим компонентам $\boldsymbol{u}_i$, т.е. $\boldsymbol{u}_i^T\mathbf{v} = 0$, для $1 \leq i < j$. Как и до этого, проецируемая дисперсия по $\mathbf{v}$ задается как

$$\sigma_{\mathbf{v}}^2 = \mathbf{v}^T\mathbf{\Sigma}\mathbf{v}$$

В сочетании с ограничениями на $\mathbf{v}$ это приводит к следующей задаче максимизации с множителями Лагранжа:

$$\max_{\mathbf{v}} J(\mathbf{v}) = \mathbf{v}^T\mathbf{\Sigma}\mathbf{v} - \alpha(\mathbf{v}^T\mathbf{v} - 1) - \sum_{i=1}^{j-1}\beta_i(\boldsymbol{u}_i{}^T\mathbf{v} - 0)$$

Взяв производную $J(\mathbf{v})$ по $\mathbf{v}$ и приравняв ее к нулевому вектору, получаем

$$2\mathbf{\Sigma}\mathbf{v} - 2\alpha\mathbf{v} - \sum_{i=1}^{j-1}\beta_i\boldsymbol{u}_i = \mathbf{0} \tag{7.24}$$

Если умножить левую часть на $\boldsymbol{u}_k^T$ для $1 \leq k < j$, то получим

$$2\boldsymbol{u}_k^T\mathbf{\Sigma}\mathbf{v} - 2\alpha\boldsymbol{u}_k^T\mathbf{v} - \beta_k\boldsymbol{u}_k^T\boldsymbol{u}_k - \sum_{\substack{i=1 \\ i\neq k}}^{j-1}\beta_i\boldsymbol{u}_k^T\boldsymbol{u}_i = 0$$

$$2\mathbf{v}^T\mathbf{\Sigma}\boldsymbol{u}_k - \beta_k = 0$$

$$\beta_k = 2\mathbf{v}^T\lambda_k\boldsymbol{u}_k = 2\lambda_k\mathbf{v}^T\boldsymbol{u}_k = 0$$

где мы использовали тот факт, что $\mathbf{\Sigma}\boldsymbol{u}_k = \lambda_k\boldsymbol{u}_k$, т.к. $\boldsymbol{u}_k$ является собственным вектором, которому соответствует k-е наибольшее собственное значение λ_k матрицы $\mathbf{\Sigma}$. Таким образом, мы находим, что $\beta_i = 0$ для всех $i < j$ в уравнении (7.24), откуда следует

$$\mathbf{\Sigma}\mathbf{v} = \alpha\mathbf{v}$$

Чтобы максимизировать дисперсию по $\mathbf{v}$, мы устанавливаем $\alpha = \lambda_j$, j-е наибольшее собственной значение матрицы $\mathbf{\Sigma}$, где $\mathbf{v} = \boldsymbol{u}_j$ задает j-ю главную компоненту.

Таким образом, чтобы найти наилучше r-мерное приближение к $\boldsymbol{D}$, мы вычисляем собственные значения матрицы $\mathbf{\Sigma}$. Поскольку $\mathbf{\Sigma}$ является положительно полуопределенной, все ее собственные числа должны быть неотрицательными, и мы можем отсортировать их в порядке убывания следующим образом:

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_r \geq \lambda_{r+1} \ldots \geq \lambda_d \geq 0$$

Затем мы выбираем r наибольших собственных значений и соответствующие им собственные векторы, чтобы сформировать лучшее r-мерное приближение.

Общая проецируемая дисперсия

Пусть $\boldsymbol{U}_r$ – r-мерная матрица базисных векторов

$$\boldsymbol{U}_r = \begin{pmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_r \\ | & | & & | \end{pmatrix}$$

с матрицей проекции, заданной как

$$\boldsymbol{P}_r = \boldsymbol{U}_r \boldsymbol{U}_r^T = \sum_{i=1}^{r} \boldsymbol{u}_i \boldsymbol{u}_i^T$$

Пусть $\boldsymbol{A}$ обозначает набор данных, сформированный координатами проецируемых точек в r-мерном подпространстве, т.е. $\boldsymbol{a}_i = \boldsymbol{U}_r^T \boldsymbol{x}_i$, и пусть $\boldsymbol{x}_i' = \boldsymbol{P}_r \boldsymbol{x}_i$ обозначает спроецированную точку в исходном d-мерном пространстве. Из (7.19), (7.21) и (7.22) проецируемая дисперсия определяется как

$$var(\boldsymbol{A}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{P}_r \boldsymbol{x}_i = \sum_{i=1}^{r} \boldsymbol{u}_i^T \boldsymbol{\Sigma} \boldsymbol{u}_i = \sum_{i=1}^{r} \lambda_i$$

Таким образом, общая проецируемая дисперсия – это просто сумма r наибольших собственных значений матрицы $\boldsymbol{\Sigma}$.

Среднеквадратичная ошибка

На основании вывода для уравнения (7.23), среднеквадратичная ошибка в r-измерениях может быть записана как

$$\begin{aligned} \text{MSE} &= \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x}_i'\|^2 \\ &= var(\boldsymbol{D}) - var(\boldsymbol{A}) \\ &= var(\boldsymbol{D}) - \sum_{i=1}^{r} \boldsymbol{u}_i^T \boldsymbol{\Sigma} \boldsymbol{u}_i \\ &= var(\boldsymbol{D}) - \sum_{i=1}^{r} \lambda_i \end{aligned}$$

Первые r главных компонент максимизируют проецируемую дисперсию $var(\boldsymbol{A})$ и, таким образом, они также минимизируют MSE.

Общая дисперсия

Обратите внимание, что общая дисперсия $\boldsymbol{D}$ инвариантна к изменению базисных векторов. Следовательно, имеем следующее тождество:

$$var(\boldsymbol{D}) = \sum_{i=1}^{d} \sigma_i^2 = \sum_{i=1}^{d} \lambda_i$$

Выбор размерности

Часто мы можем не знать, сколько измерений r использовать для хорошей аппроксимации. Одним из критериев выбора r является вычисление доли общей дисперсии, охваченной первыми r главными компонентами

$$f(r) = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_r}{\lambda_1 + \lambda_2 + \cdots + \lambda_d} = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{d} \lambda_i} = \frac{\sum_{i=1}^{r} \lambda_i}{var(\boldsymbol{D})} \tag{7.25}$$

Учитывая определенный желаемый порог дисперсии, скажем α, начиная с первого главного компонента, мы продолжаем добавлять дополнительные компоненты и останавливаемся на наименьшем значении r, для которого $f(r) \geq \alpha$. Другими словами, мы выбираем наименьшее количество измерений так, что подпространство, охватываемое этими r измерениями, захватывает

по крайней мере долю α общей дисперсии. На практике α обычно устанавливается равным 0.9 или выше, т.ч. уменьшенный набор данных фиксирует не менее 90% общей дисперсии.

Алгоритм 7.1 демонстрирует псевдокод для алгоритма метода главных компонент PCA. Центрируются входные данные $\boldsymbol{D} \in \mathbb{R}^{n \times d}$ вычитанием среднего значение из каждой точки. Затем вычисляются собственные векторы и собственные значения ковариационной матрицы $\boldsymbol{\Sigma}$. Учитывая желаемый порог дисперсии α, выбирается наименьший набор измерений r, который охватывает по меньшей мере долю α общей дисперсии. В заключении, вычисляются координаты каждой точки в новом r-мерном подпространстве главных компонентов, чтобы получить новую матрицу данных $\boldsymbol{A} \in \mathbb{R}^{n \times r}$.

АЛГОРИТМ 7.1. Метод главных компонент PCA

$\boldsymbol{PCA}(\boldsymbol{D}, \alpha)$:

1. $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ // вычисление среднего
2. $\boldsymbol{Z} = \boldsymbol{D} - \boldsymbol{1} \cdot \boldsymbol{\mu}^T$ // центрирование данных
3. $\boldsymbol{\Sigma} = \frac{1}{n}(\boldsymbol{Z}^T\boldsymbol{Z})$ // вычисление ковариационной матрицы
4. $(\lambda_1, \lambda_2, \ldots, \lambda_d) = eigenvalues(\boldsymbol{\Sigma})$ // вычисление собственных значений
5. $\boldsymbol{U} = (\boldsymbol{u}_1 \quad \boldsymbol{u}_2 \quad \ldots \quad \boldsymbol{u}_d) = eigenvectors(\boldsymbol{\Sigma})$ // собственных векторов
6. $f(r) = \frac{\sum_{i=1}^{r}\lambda_i}{\sum_{i=1}^{d}\lambda_i}$, for all r = 1, 2, …, d // доля от общей дисперсии
7. $Choose\ smallest\ r\ so\ that\ f(r) \geq \alpha$ // выбор размерности
8. $\boldsymbol{U}_r = (\boldsymbol{u}_1 \quad \boldsymbol{u}_2 \quad \ldots \quad \boldsymbol{u}_r)$ // уменьшенный базис
9. $\boldsymbol{A} = \{\boldsymbol{a}_i | \boldsymbol{a}_i = \boldsymbol{U}_r^T\boldsymbol{x}_i, for\ i = 1, \ldots, n\}$ // данные уменьшенной размерности

7.2.4 Геометрия метода главных компонент

Геометрически, когда $r = d$, PCA соответствует ортогональному изменению базиса, так что общая дисперсия фиксируется суммой дисперсий по каждому из главных направлений $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d$, и далее, все ковариации равны нулю. Это можно увидеть, посмотрев на коллективное действие полного набора главных компонентов, которые могут быть организованы в ортогональную матрицу $d \times d$.

$$\boldsymbol{U} = \begin{pmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_d \\ | & | & & | \end{pmatrix}$$

$$\boldsymbol{U}^{-1} = \boldsymbol{U}^T$$

Каждый главный компонент $\boldsymbol{u}_i$ соответствует собственному вектору ковариационной матрицы $\boldsymbol{\Sigma}$, т.е.

$$\boldsymbol{\Sigma}\boldsymbol{u}_i = \lambda_i\boldsymbol{u}_i \text{ for all } 1 \leq i \leq d$$

Который можно компактно записать в матричных обозначениях следующим образом:

$$\boldsymbol{\Sigma}\begin{pmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_d \\ | & | & & | \end{pmatrix} = \begin{pmatrix} | & | & & | \\ \lambda_1\boldsymbol{u}_1 & \lambda_2\boldsymbol{u}_2 & \ldots & \lambda_d\boldsymbol{u}_d \\ | & | & & | \end{pmatrix}$$

$$\boldsymbol{\Sigma}\mathbf{U} = \mathbf{U}\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}$$

$$\boldsymbol{\Sigma}\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda} \tag{7.26}$$

Если мы умножим уравнение (7.26) слева на $\boldsymbol{U}^{-1} = \boldsymbol{U}^T$, то получим

$$\boldsymbol{U}^T\boldsymbol{\Sigma}\mathbf{U} = \boldsymbol{U}^T\mathbf{U}\boldsymbol{\Lambda} = \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}$$

Это означает, что если мы изменим базис на $\boldsymbol{U}$, то ковариационная матрица $\boldsymbol{\Sigma}$ изменится на аналогичную матрицу $\boldsymbol{\Lambda}$, которая фактически является ковариационной матрицей в новом базисе. Тот факт, что $\boldsymbol{\Lambda}$ является диагональной, подтверждает, что после изменения базиса все ковариации исчезают, и остается только дисперсия по каждой из главных компонент, причем дисперсия по каждому новому направлению $\boldsymbol{u}_i$ задается соответствующим собственным значением λ_i.

Стоит отметить, что в новом базисе уравнение

$$\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = 1 \tag{7.27}$$

Определяет d-мерный эллипсоид (или гиперэллипс). Собственные векторы $\boldsymbol{u}_i$ матрицы $\boldsymbol{\Sigma}$, т.е. главные компоненты, являются направлениями главных осей эллипсоида. Квадратные корни собственных значений, т.е. $\sqrt{\lambda_i}$, являются длинами полуосей эллипса.

Умножая (7.26) справа на $\boldsymbol{U}^{-1} = \boldsymbol{U}^T$, имеем

$$\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T \tag{7.28}$$

Предполагая, что $\boldsymbol{\Sigma}$ обратимая или невырожденная, имеем

$$\boldsymbol{\Sigma}^{-1} = (\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T)^{-1} = (\boldsymbol{U}^{-1})^T\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T$$

Где

$$\boldsymbol{\Lambda}^{-1} = \begin{pmatrix} \frac{1}{\lambda_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\lambda_d} \end{pmatrix}$$

Подставив $\boldsymbol{\Sigma}^{-1}$ в уравнение (7.27) и используя тот факт, что $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{a}$ из (7.2), где $\boldsymbol{a} = (a_1, a_2, \ldots, a_d)^T$ представляет координаты $\boldsymbol{x}$ в новом базисе, получаем

$$\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = 1$$

$$(\boldsymbol{a}^T\boldsymbol{U}^T)\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T(\boldsymbol{U}\boldsymbol{a}) = 1$$

$$\boldsymbol{a}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{a} = 1$$

$$\sum_{i=1}^{d}\frac{a_i^2}{\lambda_i} = 1$$

что и есть уравнение эллипса с центром в 0 и длинной полуосей $\sqrt{\lambda_i}$. Таким образом $\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = 1$ или эквивалентно $\boldsymbol{a}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{a} = 1$ в новом базисе главных компонентов определяет d-мерный эллипсоид, где длины полуосей равны стандартным отклонениям (квадратный корень из дисперсии $\sqrt{\lambda_i}$) вдоль каждой оси. Точно так же уравнение $\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = s$ или эквивалентное $\boldsymbol{a}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{a} = s$, для различных значений скалярной величины s, представляет концентрические эллипсоиды.

7.3 Ядерный анализ главных компонент

Анализ главных компонент может быть расширен, чтобы искать нелинейные “направления” в данных с использованием ядерных методов. Ядерный PCA ищет направления наибольшей дисперсии в пространстве признаков вместо входного пространства. То есть, вместо поиска линейных комбинаций во входном пространстве, алгоритм ищет линейные комбинации в пространстве признаков, полученном из нелинейного преобразования входного пространства. Таким образом, линейные главные компоненты пространства признаков соответствуют нелинейным направлениям во входном пространстве. Как мы увидим, с использованием *ядерного трюка* все операции могут быть выполнены в терминах входного пространства, без необходимости преобразования данных в пространстве признаков.

Пусть $\boldsymbol{\phi}$ – отображение из входного пространства в пространство признаков. Каждой точке в пространстве признаков соответствует образ $\boldsymbol{\phi}(\boldsymbol{x}_i)$ точки $\boldsymbol{x}_i$ во входном пространстве. Во входном пространстве первая главная компонента показывает направление наибольшей дисперсии; её собственный вектор совпадает с наибольшим собственным значением в матрице ковариантности. Таким же образом, через поиск собственного вектора, соответствующий наибольшему собственному значению в матрице ковариантности в пространстве признаков, можно найти первую главную компоненту в пространстве признаков $\boldsymbol{u}_1$ (с $\boldsymbol{u}_1^T\boldsymbol{u}_1 = 1$):

$$\boldsymbol{\Sigma}_{\boldsymbol{\phi}} = \lambda_1 \boldsymbol{u}_1 \tag{7.29}$$

где $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ – матрица ковариантности в пространстве признаков, определенная как

$$\boldsymbol{\Sigma}_{\boldsymbol{\phi}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_i)^T \tag{7.30}$$

Здесь предполагается, что точки центрированы, т.е. $\boldsymbol{\phi}(\boldsymbol{x}_i) = \boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\mu}_{\boldsymbol{\phi}}$, где $\boldsymbol{\mu}_{\boldsymbol{\phi}}$ – среднее в пространстве признаков.

Подставляя (7.30) в (7.29) получаем:

$$\begin{aligned}
\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_i)^T\right)\boldsymbol{u}_1 &= \lambda_1 \boldsymbol{u}_1 \\
\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i)(\boldsymbol{\phi}(\boldsymbol{x}_i)^t\boldsymbol{u}_1) &= \lambda_1 \boldsymbol{u}_1 \qquad (7.31) \\
\sum_{i=1}^{n}\left(\frac{\boldsymbol{\phi}(\boldsymbol{x}_i)^T\boldsymbol{u}_1}{n\lambda_1}\right)\boldsymbol{\phi}(\boldsymbol{x}_i) &= \boldsymbol{u}_1 \\
\sum_{i=1}^{n} c_i \boldsymbol{\phi}(\boldsymbol{x}_i) &= \boldsymbol{u}_1 \qquad (7.32)
\end{aligned}$$

где $c_i = \frac{\boldsymbol{\phi}(\boldsymbol{x}_i)^T\boldsymbol{u}_1}{n\lambda_1}$ – скалярная величина. Из (7.32) видно, что лучшее направление в пространстве признаков, $\boldsymbol{u_1}$ есть линейная комбинация преобразованных точек, где скаляры c_i показывают важность каждой точки в направлении наибольшей дисперсии.

Теперь можно подставить (7.32) назад в (7.31) и получить

$$\left(\frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^T\right)\left(\sum_{i=1}^{n}c_j\phi(\boldsymbol{x}_i)\right)=\lambda_1\sum_{i=1}^{n}c_i\phi(\boldsymbol{x}_i)$$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}c_j\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j)=\lambda_1\sum_{i=1}^{n}c_i\phi(\boldsymbol{x}_i)$$

$$\sum_{i=1}^{n}\left(\phi(\boldsymbol{x}_i)\sum_{j=1}^{n}c_j\phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j)\right)=n\lambda_1\sum_{i=1}^{n}c_i\phi(\boldsymbol{x}_i)$$

В предыдущем уравнении можно заменить скалярное произведение в пространстве признаков, $(\boldsymbol{x}_i)^T\boldsymbol{\phi}(\boldsymbol{x}_j)$, соответствующей ядерной функцией во входном пространстве, а именно $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$

$$\sum_{i=1}^{n}\left(\phi(\boldsymbol{x}_i)\sum_{j=1}^{n}c_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)=n\lambda_1\sum_{i=1}^{n}c_i\phi(\boldsymbol{x}_i) \tag{7.33}$$

Здесь предполагается, что точки в пространстве признаков уже центрированы, т.е. ядерная матрица была центрирована через уравнение (5.14):

$$\boldsymbol{K}=\left(\boldsymbol{I}-\frac{1}{n}\mathbf{1}_{n\times n}\right)\boldsymbol{K}\left(\boldsymbol{I}-\frac{1}{n}\mathbf{1}_{n\times n}\right)$$

где $\boldsymbol{I}-$ единичная матрица, у которой элементы главной диагонали 1, а остальные 0, и $\mathbf{1}_{n\times n}$ – матрица $n\times n$, у которой все элементы – 1.

Мы заменили одно скалярное произведение ядерной функцией. Чтобы убедится, что все вычисления в пространстве признаков выполняются только в терминах скалярного произведение, возьмем любую точку $\boldsymbol{\phi}(\boldsymbol{x}_k)$ и умножим (7.33) на неё с обеих сторон чтобы получить

$$\begin{aligned}\sum_{i=1}^{n}\left(\phi(\boldsymbol{x}_k)\phi(\boldsymbol{x}_i)\sum_{j=1}^{n}c_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)&=n\lambda_1\sum_{i=1}^{n}c_i\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_k)\\ \sum_{i=1}^{n}\left(K(\boldsymbol{x}_k,\boldsymbol{x}_i)\sum_{j=1}^{n}c_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)&=n\lambda_1\sum_{i=1}^{n}c_iK(\boldsymbol{x}_k,\boldsymbol{x}_i)\end{aligned} \tag{7.34}$$

Затем, обозначим $\boldsymbol{K}_i$ строку i центрированной ядерной матрицы, записанную как вектор-столбец:

$$\boldsymbol{K}_i=(K(\boldsymbol{x}_i,\boldsymbol{x}_1)\quad K(\boldsymbol{x}_i,\boldsymbol{x}_2)\quad\cdots\quad K(\boldsymbol{x}_i,\boldsymbol{x}_n))^T$$

Пусть $\boldsymbol{c}$ обозначает вектор-столбец весов

$$\boldsymbol{c}=(c_1\quad c_2\quad\cdots\quad c_n)^T$$

Теперь можно записать (7.34) как:

$$\sum_{i=1}^{n}K(\boldsymbol{x}_k,\boldsymbol{x}_i)\boldsymbol{K}_i^T\boldsymbol{c}=n\lambda_1\boldsymbol{K}_k^T\boldsymbol{c}$$

На самом деле, поскольку можно выбрать любые из n точек $\boldsymbol{\phi}(\boldsymbol{x}_k)$ в пространстве признаков, чтобы получить (7.34), имеется набор уравнений:

$$\begin{aligned}
\sum_{i=1}^{n} K(\boldsymbol{x}_1, \boldsymbol{x}_i)\boldsymbol{K}_i^T\boldsymbol{c} &= n\lambda_1\boldsymbol{K}_1^T\boldsymbol{c} \\
\sum_{i=1}^{n} K(\boldsymbol{x}_2, \boldsymbol{x}_i)\boldsymbol{K}_i^T\boldsymbol{c} &= n\lambda_1\boldsymbol{K}_2^T\boldsymbol{c} \\
\vdots &= \vdots \\
\sum_{i=1}^{n} K(\boldsymbol{x}_2, \boldsymbol{x}_i)\boldsymbol{K}_i^T\boldsymbol{c} &= \sum_{i=1}^{n} K(\boldsymbol{x}_k, \boldsymbol{x}_i)\boldsymbol{K}_n^T\boldsymbol{c}
\end{aligned}$$

Можно представить эти $\boldsymbol{n}$ уравнений как:

$$\boldsymbol{K}^2\boldsymbol{c} = n\lambda_1\boldsymbol{K}\boldsymbol{c}$$

где $\boldsymbol{K}$ – центрированная ядерная матрица. Умножением на $\boldsymbol{K}^{-1}$ получаем:

$$\begin{aligned}
\boldsymbol{K}^{-1}\boldsymbol{K}^2\boldsymbol{c} &= n\lambda_1\boldsymbol{K}^{-1}\boldsymbol{K}\boldsymbol{c} \\
\boldsymbol{K}\boldsymbol{c} &= n\lambda_1\boldsymbol{c} \\
\boldsymbol{K}\boldsymbol{c} &= \eta_1\boldsymbol{c}
\end{aligned} \tag{7.35}$$

где $\eta_1 = n\lambda_1$. Таким образом, вектор весов $\boldsymbol{c}$ есть собственный вектор, соответствующий наибольшему собственному значению η_1 ядерной матрицы $\boldsymbol{K}$.

Как только найдено $\boldsymbol{c}$, её можно подставить обратно в (7.32), чтобы получить первую ядерную основную компоненту $\boldsymbol{u}_1$. Единственное ограничение – надо нормализовать $\boldsymbol{u}_1$ до единичного вектора следующим образом:

$$\begin{aligned}
\boldsymbol{u}_1^T\boldsymbol{u}_1 &= 1 \\
\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j) &= 1 \\
\boldsymbol{c}^T\boldsymbol{K}\boldsymbol{c} &= 1
\end{aligned}$$

Из (7.35) имеем:

$$\begin{aligned}
\boldsymbol{c}^T(\eta_1\boldsymbol{c}) &= 1 \\
\eta_1\boldsymbol{c}^T\boldsymbol{c} &= 1 \\
\|\boldsymbol{c}\|^2 &= \frac{1}{\eta_1}
\end{aligned}$$

Однако, поскольку $\boldsymbol{c}$ – собственный вектор $\mathbf{K}$, он будет иметь единичную норму. То есть, чтобы удостоверится, что $\boldsymbol{u}_1$ – единичный вектор, надо масштабировать вектор весов $\boldsymbol{c}$ так, что его норма $\|c\| = \sqrt{\frac{1}{\eta_1}}$, что можно получить, домножив $\boldsymbol{c}$ на $\sqrt{\frac{1}{\eta_1}}$.

В общем случае, поскольку мы не отображаем входные точки в пространство признаков через ϕ, невозможно непосредственно вычислить главное направление в терминах $\phi(\boldsymbol{x}_i)$, как показано в (7.32). Однако, здесь важно, что любую точку $\phi(\boldsymbol{x})$ можно отобразить на главное направление $\boldsymbol{u}_1$ следующим образом:

$$\boldsymbol{u}_1^T\phi(\boldsymbol{x}) = \sum_{i=1}^{n} c_i\phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}) = \sum_{i=1}^{n} c_i K(\boldsymbol{x}_i, \boldsymbol{x})$$

что требует только ядерные операции. Когда $\boldsymbol{x} = \boldsymbol{x}_i$ – одна из входных точек, проекция $\phi(\boldsymbol{x}_i)$ на главную компоненту $\boldsymbol{u}_1$ может быть записана как скалярное произведение

$$\boldsymbol{a}_i = \boldsymbol{u}_1^T\phi(\boldsymbol{x}_i) = \boldsymbol{K}_i^T\boldsymbol{c} \tag{7.36}$$

где $\boldsymbol{K}_i$ – вектор-столбец, соответствующий i-й строке в ядерной матрице. Таким образом, мы показали, что все вычисления, или для решения главной компоненты, или для проекции точек, могут быть выполнены с использованием одной только ядерной функции. Затем, мы можем получить дополнительные главные компоненты путём решения 7.35 для других собственных значений и собственных векторов. Другими словами, отсортировав собственные значения $\boldsymbol{K}$ в порядке убывания $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_n \geq 0$, можно получить j-ю главную компоненту как соответствующий собственный вектор $\boldsymbol{c}_j$, который должен быть нормализован по норме $\|\boldsymbol{c}_j\| = \sqrt{\frac{1}{n}}$, при условии $\eta_j > 0$. И поскольку $\eta_j = n\lambda_j$, дисперсия по j-й главной компоненте определяется как $\lambda_j = \frac{\eta_j}{n}$. Алгоритм 7.2 приводит псевдокод для метода ядерного PCA.

Алгоритм 7.2. Ядерный анализ главных компонент

KernelPCA$(\boldsymbol{D}, K, \alpha)$:

1. $\boldsymbol{K} = \{K(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{\{i,j=1\ldots n\}}$ // вычислить ядерную матрицу $n \times n$
2. $\boldsymbol{K} = \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}_{n\times n}\right)\boldsymbol{K}\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}_{n\times n}\right)$ // центрировать ядерную матрицу
3. $(\eta_1, \eta_2, \ldots, \eta_d) = \text{eigenvalues}(\boldsymbol{K})$ // собственные значения
4. $(\boldsymbol{c}_1 \quad \boldsymbol{c}_2 \quad \cdots \quad \boldsymbol{c}_n) = \text{eigenvectors}(\boldsymbol{K})$ // собственные вектора
5. $\lambda_i = \frac{\eta_i}{n}, \forall i = 1, \ldots, n$ // вычислить дисперсию каждой компоненты
6. $\boldsymbol{c}_i = \sqrt{\frac{1}{\eta_i}} \cdot \boldsymbol{c}_i, \forall i = 1, \ldots, n$ // убедится, что $\boldsymbol{u}_i^T \boldsymbol{u}_i = 1$
7. $f(r) = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{d} \lambda_i}, \forall r = 1,2, \ldots, d$ // часть общей дисперсии
8. Выбрать наименьшую r такую, что $f(r) \geq \alpha$ // выбор размерности
9. $\boldsymbol{C}_r = (\boldsymbol{c}_1 \quad \boldsymbol{c}_2 \quad \cdots \quad \boldsymbol{c}_r)$ // Уменьшенный базис
10. $\boldsymbol{A} = \{\boldsymbol{a_i} | \boldsymbol{a_i} = \boldsymbol{C}_r^T \boldsymbol{K}_i, \forall i = 1, \ldots, n\}$ // Данные меньшей размерности

7.4. Сингулярное разложение

Анализ главных компонентов является частным случаем более общего метода разложения матриц, называемого сингулярным разложением (Singular Value Decomposition - SVD). Мы знаем, что по ур-ю 7.28 PCA дает следующее разложение ковариационной матрицы:

$$\Sigma = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T \textbf{ 7.37}$$

где ковариационная матрица была представлена через ортогональную матрицу $\boldsymbol{U}$, содержащую ее собственные вектор и диагональную матрицу $\boldsymbol{\Lambda}$, содержащую ее собственные числа (отсортированные в порядке убывания). SVD обобщает приведенную выше факторизацию для любой матрицы. В частности, для матрицы данных $\boldsymbol{D}$ размера $n \times d$ с n точками и d столбцами SVD факторизует $\boldsymbol{D}$ следующим образом:

$$\boldsymbol{D} = \boldsymbol{L}\Delta\boldsymbol{R}^{\boldsymbol{T}} \textbf{ 7.38}$$

где $\boldsymbol{L}$ - ортогональная матрица размера $n \times n$, $\boldsymbol{R}$ - ортогональная матрица размера $d \times d$, а $\boldsymbol{\Delta}$ - «диагональная» матрица $n \times d$. Столбцы $\boldsymbol{L}$ называются левыми сингулярными векторами, а столбцы $\boldsymbol{R}$ (или строки $\boldsymbol{R}^{\boldsymbol{T}}$) называются правыми сингулярными векторами. Матрица $\boldsymbol{\Delta}$ определяется как

$$\boldsymbol{\Delta}(\mathrm{i},\mathrm{j}) = \begin{cases} \delta_i, & \text{если } i = j \\ 0, & \text{если } i \neq j \end{cases}$$

где $i = 1, \dots, n$ и $j = 1, \dots, d$. Элементы $\boldsymbol{\Delta}(\mathrm{i},\mathrm{i}) = \delta_i$ вдоль главной диагонали $\boldsymbol{\Delta}$ называются сингулярными числами $\boldsymbol{D}$, и все они неотрицательны. Если ранг $\boldsymbol{D}$ равен $r \leq \min(n, d)$, то существует только r ненулевых особых значений, которые, как мы предполагаем, упорядочены следующим образом:

$$\delta_1 \geq \delta_2 \geq \cdots \geq \delta_r \geq 0$$

Можно отбросить эти левые и правые сингулярные вектора, которые соответствуют нулевым сингулярным значениям, чтобы получить сокращенный SVD в следующем виде:

$$\boldsymbol{D} = \boldsymbol{L_r}\boldsymbol{\Delta_r}\boldsymbol{R_r^T} \textbf{ 7.39}$$

где $\boldsymbol{L_r}$ - матрица левых сингулярных векторов размера $n \times r$, $\boldsymbol{R_r}$ - матрица правых сингулярных векторов размера $d \times r$, а $\boldsymbol{\Delta_r}$ - диагональная матрица размера $r \times r$, содержащая положительные сингулярные вектора. Уменьшенный SVD приводит непосредственно к спектральному разложению $\boldsymbol{D}$, заданному как

$$\boldsymbol{D} = \boldsymbol{L_r}\boldsymbol{\Delta_r}\boldsymbol{R_r^T} = \begin{pmatrix} | & | & & | \\ l_1 & l_2 & \dots & l_r \\ | & | & & | \end{pmatrix} \begin{pmatrix} \delta_1 & 0 & \cdots & 0 \\ 0 & \delta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_r \end{pmatrix} \begin{pmatrix} - & r_1^T & - \\ - & r_2^T & - \\ - & \vdots & - \\ - & r_r^T & - \end{pmatrix}$$

$$= \delta_1 l_1 r_1^T + \delta_2 l_2 r_2^T + \cdots + \delta_r l_r r_r^T = \sum_{i=1}^{r} \delta_i l_i r_i^T$$

Спектральное разложение $\boldsymbol{D}$ представляет собой сумму матриц первого ранга вида $\delta_i l_i r_i^T$. Выбрав q наибольших сингулярных значений $\delta_1, \delta_2, \dots, \delta_q$ и соответствующие левый и правый сингулярные вектора, мы получаем наилучшее приближение ранга q к исходной матрице $\boldsymbol{D}$. То есть, если $\boldsymbol{D}_q$ - матрица, определенная как

$$\boldsymbol{D}_q = \sum_{i=1}^{r} \delta_i l_i r_i^T$$

то можно показать, что $\boldsymbol{D}_q$ - это матрица ранга q, которая минимизирует следующее выражение

$$\left\| \boldsymbol{D} - \boldsymbol{D}_q \right\|_F$$

где $\|\boldsymbol{A}\|_F$ - норма Фробениуса матрицы $\boldsymbol{A}$ размера $n \times d$, определяемая как

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{d} \boldsymbol{A}(i,j)^2}$$

7.4.1. Геометрия сингулярного разложения

В общем, любая матрица $\boldsymbol{D}$ размера $n \times d$ представляет собой линейное преобразование $\boldsymbol{D}$: $\mathbb{R}^d \rightarrow \mathbb{R}^n$ из пространства d-мерных векторов в пространство n-мерных векторов, поскольку для любого $x \in \mathbb{R}^d$ существует $y \in \mathbb{R}^n$ такое, что

$$\boldsymbol{Dx} = \boldsymbol{y}$$

Множество всех векторов $y \in \mathbb{R}^n$ таких, что $\boldsymbol{Dx} = \boldsymbol{y}$ по всем возможными $x \in \mathbb{R}^d$, называется пространством столбцов $\boldsymbol{D}$, а множество всех векторов $x \in \mathbb{R}^d$, таких что $\boldsymbol{D^T y} = \boldsymbol{x}$ по всем $y \in \mathbb{R}^n$, называется пространство строк $\boldsymbol{D}$, что эквивалентно пространству столбцов $\boldsymbol{D^T}$. Другими словами, пространство столбцов $\boldsymbol{D}$ — это набор всех векторов, которые могут быть получены как линейные комбинации столбцов $\boldsymbol{D}$, а пространство строк $\boldsymbol{D}$ - это набор всех векторов, которые могут быть получены как линейные комбинации строк $\boldsymbol{D}$ (или столбцов $\boldsymbol{D^T}$). Также обратите внимание, что множество всех векторов $x \in \mathbb{R}^d$, таких что $\boldsymbol{Dx} = \boldsymbol{0}$, называется нулевым пространством $\boldsymbol{D}$, и, наконец, множество всех векторов $y \in \mathbb{R}^n$, таких что $\boldsymbol{D^T y} = \boldsymbol{0}$, называется левым нулевым пространством $\boldsymbol{D}$.

Одним из основных свойств сингулярного разложения является то, что оно дает основу для каждого из четырех фундаментальных пространств, связанных с матрицей $\boldsymbol{D}$. Если $\boldsymbol{D}$ имеет ранг r, это означает, что у него есть только r независимых столбцов и r независимых строк. Таким образом, r левых сингулярных векторов $l_1, l_2, \ldots, l_r$, соответствующие r ненулевым сингулярным значениям $\boldsymbol{D}$ в ур-и (7.38), представляют собой основу для пространства столбцов $\boldsymbol{D}$. Остальные $n - r$ левых сингулярных векторов $l_{r+1}, l_{r+2}, \ldots, l_n$, представляют собой основу для левого нулевого пространства $\boldsymbol{D}$. Для пространства строк r правых сингулярных векторов $r_1, r_2, \ldots, r_r$, соответствующие r ненулевым сингулярным значениям, представляют собой базис для пространства строк $\boldsymbol{D}$, а оставшиеся $d - r$ правых сингулярных векторов r_j $(j = r+1, \ldots, d)$ представляют собой базис для нулевое пространство $\boldsymbol{D}$.

Рассмотрим сокращенное представление сингулярного разложения из ур-я (7.39). Умножив обе части уравнения справа на $\boldsymbol{R_r}$ и зная, что $\boldsymbol{R_r^T R_r} = \boldsymbol{I_r}$, где $\boldsymbol{I_r}$ - единичная матрица размера $r \times r$, имеем

$$\boldsymbol{DR_r} = \boldsymbol{L_r \Delta_r R_r^T R_r}$$

$$\boldsymbol{DR_r} = \boldsymbol{L_r \Delta_r}$$

$$\boldsymbol{DR}_r = \boldsymbol{L_r}\begin{pmatrix} \delta_1 & 0 & \cdots & 0 \\ 0 & \delta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_r \end{pmatrix}$$

$$\boldsymbol{D}\begin{pmatrix} | & | & & | \\ r_1 & r_2 & \ldots & r_r \\ | & | & & | \end{pmatrix} = \begin{pmatrix} | & | & & | \\ \delta_1 l_1 & \delta_2 l_2 & \ldots & \delta_r l_r \\ | & | & & | \end{pmatrix}$$

Из вышесказанного получаем, что

$$\boldsymbol{DR}_i = \delta_i l_i \text{ для всех } i = 1, \ldots, r$$

Другими словами, SVD — это специальная факторизация матрицы $\boldsymbol{D}$, так что любой базисный вектор r_i для пространства строк отображается в соответствующий базисный вектор l_i в пространстве столбцов, масштабируемый сингулярным значением δ_i. Таким образом, мы можем рассматривать SVD как отображение ортонормированного базиса $(r_1, r_2, \ldots, r_r)$ $\mathbb{R}^d$ (пространство строк) в ортонормированный базис $(l_1, l_2, \ldots, l_r)$ $\mathbb{R}^n$ (пространство столбцов) с соответствующими осями, масштабируемыми согласно сингулярным значениям $(\delta_1, \delta_2, \ldots, \delta_r)$.

7.4.2. Связь между SVD и PCA

Предположим, что матрица $\boldsymbol{D}$ центрирована, и, что она разложена с помощью SVD [ур-е. (7.38)] как $\boldsymbol{D} = \boldsymbol{L\Delta R^T}$. Рассмотрим матрицу рассеяния для $\boldsymbol{D}$, заданную как $\boldsymbol{D^T D}$. Имеем

$$\boldsymbol{D^T D} = \left(\boldsymbol{L\Delta R^T}\right)^{\boldsymbol{T}}\left(\boldsymbol{L\Delta R^T}\right) = \boldsymbol{R}\Delta^T\boldsymbol{L^T L\Delta R^T} = \boldsymbol{R}(\Delta^T\Delta)\boldsymbol{R^T} = \boldsymbol{R}\left(\Delta_d^2\right)\boldsymbol{R^T} \textbf{ 7.40}$$

где Δ_d^2 - диагональная матрица $d \times d$, определенная как $\Delta_d^2(i,i) = \delta_i^2$ для $i = 1, \ldots, d$. Причем только $r \leq \min(d, n)$ из этих собственных чисел положительны, а остальные - нули.

Поскольку ковариационная матрица центрированного $\boldsymbol{D}$ задается как $\Sigma = \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{D}^T\boldsymbol{D}$ и поскольку ее можно разложить как $\Sigma = \boldsymbol{U\Lambda U}^T$ через PCA [ур-е. (7.37)], имеем

$$\boldsymbol{D^T D} = n\Sigma = n\boldsymbol{U\Lambda U}^T = \boldsymbol{U}(n\boldsymbol{\Lambda})\boldsymbol{U}^T \textbf{ 7.41}$$

Приравнивая уравнение (7.40) и уравнение (7.41), мы получаем, что правые сингулярные вектора $\boldsymbol{R}$ совпадают с собственными векторами Σ. Кроме того, соответствующие сингулярные значения $\boldsymbol{D}$ связаны с собственными числами Σ через выражение

$$n\lambda_i = \delta_i^2 \text{ или } \lambda_i = \frac{\delta_i^2}{n} \text{ для } i = 1, \ldots, d \textbf{ 7.42}$$

Рассмотрим теперь матрицу $\boldsymbol{DD^T}$. Имеем

$$\boldsymbol{DD^T} = \left(\boldsymbol{L\Delta R^T}\right)\left(\boldsymbol{L\Delta R^T}\right)^{\boldsymbol{T}} = \boldsymbol{L\Delta R^T R}\Delta^T\boldsymbol{L^T} = \boldsymbol{L}(\Delta\Delta^T)\boldsymbol{L^T} = \boldsymbol{L}\Delta_n^2\boldsymbol{L^T}$$

где Δ_n^2 - диагональная матрица $d \times d$, определенная как $\Delta_n^2(i,i) = \delta_i^2$ для $i = 1, \ldots, d$. Только r из этих сингулярных значений положительны, а все остальные - нули. Таким образом, левые сингулярные вектора $\boldsymbol{L}$ являются собственными векторами матрицы $n \times n$ $\boldsymbol{DD^T}$, а соответствующие собственные числа задаются как δ_i^2.