Глава 4

Анализ главных компонент

В главе 3 рассматривался метод наименьших квадратов, как способ оценки неизвестных параметров линейной модели. В разделе 3.4 использовалось сингулярное разложение матрицы. С сингулярным разложением матрицы ассоциируется *анализ (метод) главных компонент*[1] — одновременно мощный и изящный подход к анализу измерений на основе линейных моделей. В отличии от главы 3, преимущественно речь идёт об эмпирических моделях.

Итак, пусть $\mathbf{x} \in \mathbf{R}^n$ — случайный вектор с нулевым средним, т.е. $\mathrm{E}[\mathbf{x}] = 0$. Ковариационную матрицу вектора $\mathbf{x}$ будем обозначать Σ. Про матрицу Σ сразу известно, что она симметричная и положительно определённая, как и любая ковариационная матрица. Ковариационная матрица полагается неизвестной, и при необходимости оценивается на практике с использованием доступной выборки реализаций случайного вектора $\mathbf{x}$.

Требование о равенстве нулю среднего значения вектора не сильно ограничит наши рассуждения, обычно считается, что на практике, если известна выборка реализаций вектора $\mathbf{x}$ достаточного размера, то можно оценить среднее значение с удовлетворительной точностью, и вычесть оценку из каждой реализации.

4.1 Проекция с максимальной дисперсией

Зададимся вопросом, существует ли такой детерминированный вектор $\mathbf{q} \in \mathbf{R}^n$, чтобы проекция случайного вектора $\mathbf{x}$ на вектор $\mathbf{q}$, равная $(\mathbf{x}, \mathbf{q})$, имела бы максимально возможную дисперсию? Понятно, что эта проекция, т.е. скалярное произведение с участием случайного вектора — случайное число, которое имеет своё распределение, среднее и дисперсию. Не трудно заметить, что дисперсия может быть сделана сколь угодно большой за счёт умножения вектора $\mathbf{q}$ на произвольную константу, поэтому, чтобы вопрос о максимуме дисперсии имел смысл, необходимо ограничить норму вектора: $\|\mathbf{q}\|^2 = 1$.

[1] *principal component analysis (PCA)*

Отметим, что максимальная дисперсия интересует нас из соображений так называемого информационного содержания. Например, дифференциальная энтропия нормального распределения:

$$H = -\int p(x)\ln p(x)dx = \frac{1}{2}\sigma^2 \ln 2\pi e, \tag{4.1}$$

т.е. чем больше дисперсия, тем больше энтропия, и тем больше информации несёт реализация случайной величины.

Средняя проекция очевидно равна нулю:

$$\mathrm{E}\left[(\mathbf{x},\mathbf{q})\right] = (\mathrm{E}\left[\mathbf{x}\right],\mathbf{q}) = 0, \tag{4.2}$$

т.е. выражение для дисперсии проекции можно записать в следующем виде:

$$\mathrm{E}\left[(\mathbf{x},\mathbf{q})^2\right] = (\mathbf{q},\Sigma\mathbf{q}). \tag{4.3}$$

Воспользуемся подходом множителей Лагранжа для отыскания условного максимума этой функции относительно вектора $\mathbf{q}$. Нужно максимизировать следующую функцию

$$\Phi(\mathbf{q}) = (\mathbf{q},\Sigma\mathbf{q}) + \lambda(1 - \|\mathbf{q}\|^2), \tag{4.4}$$

где λ неизвестный множитель Лагранжа. Приравняв градиент целевой функции к нулю получим следующее уравнение:

$$\Sigma\mathbf{q} = \lambda\mathbf{q}. \tag{4.5}$$

Это задача на собственные значения и собственные вектора матрицы Σ. Известно, что её решением являются n собственных значений λ_k и соответствующие им собственные вектора $\mathbf{q}_k$. Заметим, что Σ положительно определённая матрица, следовательно все собственные значения $\lambda_k > 0$. Для удобства будем считать, что собственные значения λ_k пронумерованы в порядке убывания значения: λ_1 — максимальное собственное значение, а λ_n — минимальное.

Пусть $\mathbf{q}_k$ — собственный вектор матрицы Σ, отвечающий собственному значению λ_k, тогда дисперсия проекции на вектор $\mathbf{q_k}$ записывается следующим образом:

$$\mathrm{E}\left[(\mathbf{x},\mathbf{q}_k)^2\right] = (\mathbf{q}_k,\Sigma\mathbf{q}_k) = \lambda_k(\mathbf{q}_k,\mathbf{q}_k) = \lambda_k. \tag{4.6}$$

Откуда становится очевидным, что решением интересующей нас задачи является собственный вектор матрицы Σ, отвечающий максимальному собственному значению этой матрицы. Кроме того, собственное значение λ_k имеет смысл дисперсии проекции случайного вектора $\mathbf{x}$ на собственный вектор $\mathbf{q}_k$. Итак, собственный вектор $\mathbf{q}_1$ отвечающий максимальному собственному значению матрицы Σ называют *первым главным компонентом* вектора $\mathbf{x}$.

Попробуем обобщить задачу и построить такой ортогональный базис некоторой фиксированной размерности (т.е. задать линейное подпространство), чтобы проекция случайного вектора $\mathbf{x}$ на это линейное подпространство имело бы максимально возможную дисперсию в смысле суммы дисперсий проекций на отдельные орты. Решим задачу по индукции. Пусть Q — матрица, столбцы которой сформированы из ортонормированного базиса некоторого линейного подпространства. Выше мы доказали, что в случае одномерного подпространства единственный столбец этой матрицы — собственный вектор матрицы Σ, отвечающий максимальному собственному значению матрицы. Найдём максимум дисперсии проекции $(\mathbf{x}, \mathbf{q})$ среди всех векторов $\mathbf{q}$ с единичной нормой и ортогональных линейному подпространству, т.е. $Q^T\mathbf{q} = 0$.

Используя подход множителей Лагранжа, запишем целевую функцию:

$$\Phi(\mathbf{q}) = (\mathbf{q}, \Sigma\mathbf{q}) + \lambda(1 - \|\mathbf{q}\|^2) + (\mu, Q^T\mathbf{q}), \tag{4.7}$$

где λ — неизвестный множитель Лагранжа, а μ — целый вектор множителей Лагранжа, количество компонент которого равно размерности линейного подпространства. Приравняв градиент целевой функции к нулю получим следующее уравнение:

$$\Sigma\mathbf{q} - \lambda\mathbf{q} + \frac{1}{2}Q\mu = 0. \tag{4.8}$$

Покажем, что третье слагаемое на самом деле всегда равно нулю. Для этого домножим (4.8) слева на Q^T:

$$Q^T\Sigma\mathbf{q} - \lambda Q^T\mathbf{q} + \frac{1}{2}Q^TQ\mu = Q^T\Sigma\mathbf{q} + \frac{1}{2}\mu = 0. \tag{4.9}$$

Откуда видно, что вектор μ всегда равен нулю:

$$\mu = -2Q^T\Sigma\mathbf{q} = -2(\Sigma Q)^T\mathbf{q} = -2\Lambda Q^T\mathbf{q} = 0, \tag{4.10}$$

здесь Λ — диагональная матрица, составленная из собственных значений матрицы Σ, отвечающих соответствующим колонкам Q.

Итак, задача сводится к предыдущей:

$$\Sigma\mathbf{q} = \lambda\mathbf{q}, \tag{4.11}$$

однако на этот раз в качестве ответа нужно выбрать собственное значение λ максимальное среди всех не использованных собственных значений λ_k, т.е. таких, чьи собственные вектора q_k не содержатся в Q. Поэтому, собственный вектор $\mathbf{q}_2$ отвечающий второму по величине собственному значению матрицы Σ называют *вторым главным компонентом*, по аналогии все собственные вектора матрицы Σ называют главными компонентами. Аналогично, собственный вектор $\mathbf{q}_n$ отвечающий минимальному собственному значению матрицы Σ иногда называют *первым побочным компонентом*.

Во-первых, нетрудно видеть, что координаты проекции вектора $\mathbf{x}$ на базис Q некореллированны, это следствие ортогональности базиса Q:

$$\mathrm{E}\left[(\mathbf{x}, \mathbf{q}_k)(\mathbf{x}, \mathbf{q}_l)\right] = (\mathbf{q}_k, \Sigma\mathbf{q}_l) = \lambda_l(\mathbf{q}_k, \mathbf{q}_l). \tag{4.12}$$

Это свойство зачастую используется в рамках подходов машинного обучения, когда метод анализа главных компонент используется в качестве начальных шагов при подготовке данных.

Во-вторых, анализ главных компонент часто используется в качестве метода *понижения размерности* данных, так как для любой заданной размерности подпространства позволяет найти базис, сохраняющий максимум информационного содержания случайного вектора $\mathbf{x}$. Понижение размерности требуется, например, для визуализации данных; либо для последующего применения к данным методов, эффективность которых существенно зависит от размерности входных данных; либо в случаях, когда размерность входных данных сильно больше размера выборки, т.е. числа измерений.

В-третьих, метод можно рассматривать как способ фильтрации шума в данных. Задав размерность подпространства, мы считаем, что главные компоненты отвечают за информационное содержание данных, т.е. задают возможное разнообразие измерений, а побочные компоненты отвечают за шум, т.е. задают флуктуации измерений. Таким образом, спроецировав вектор $\mathbf{x}$ в линейное подпространство меньшей размерности, можно отфильтровать случайный шум, сохранив важную информацию. Заметим, что, аналогично методу регуляризации на основе усечённого сингулярного разложения из раздела 3.4, выбирая размерность подпространства мы балансируем между сохранением информации, содержащейся в данных, и уровнем шума.

4.2 Проекция с минимальной ошибкой

Пусть $\mathbf{x} \in \mathbf{R}^n$ всё тот же случайный вектор, на этот раз мы хотим найти такой вектор $\mathbf{q}$, проекция на который минимизирует ошибку проекции в смысле квадрата нормы разности:

$$R(\mathbf{x}) = \|(\mathbf{x}, \mathbf{q})\mathbf{q} - \mathbf{x}\|^2. \tag{4.13}$$

Иными словами, нам хотелось бы заменить случайный вектор на одно случайное число, по возможности потеряв при этом минимум информации, однако теперь мы формулируем эту задачу по другому.

Легко заметить, что средняя ошибка выражается как:

$$\mathrm{E}\left[\|(\mathbf{x}, \mathbf{q})\mathbf{q} - \mathbf{x}\|^2\right] = \operatorname{tr}\Sigma - (\mathbf{q}, \Sigma\mathbf{q}). \tag{4.14}$$

Используя подход множителей Лагранжа для того, чтобы ограничить норму $\mathbf{q}$, запишем целевую функцию в следующем виде:

$$\Phi(\mathbf{q}) = \operatorname{tr}\Sigma - (\mathbf{q}, \Sigma\mathbf{q}) + \lambda(\|\mathbf{q}\|^2 - 1), \tag{4.15}$$

откуда после взятия градиента вновь получается задача на собственные значения матрицы Σ:

$$\Sigma\mathbf{q} = \lambda\mathbf{q}. \tag{4.16}$$

Средняя ошибка проекции выражается через собственное значение матрицы λ_k и собственный вектор $\mathbf{q}_k$ следующим образом:

$$\mathrm{E}\left[\|(\mathbf{x},\mathbf{q}_k)\mathbf{q}_k-\mathbf{x}\|^2\right]=\operatorname{tr}\Sigma-(\mathbf{q}_k,\Sigma\mathbf{q}_k)=\operatorname{tr}\Sigma-\lambda_k=\sum_{j=1}^{n}\lambda_j-\lambda_k. \quad (4.17)$$

Т.е. из суммы положительных собственных значений предлагается вычесть одно слагаемое так, чтобы результат был минимально возможным. Очевидно, что в качестве решения нужно брать максимальное собственное значение λ_1 и соответствующий ему собственный вектор $\mathbf{q}_1$, т.е. первый главный компонент является решением задачи.

Аналогичным образом задача решается для проекции на многомерное подпространство. Пусть Q — матрица, m колонок которой составлены из векторов ортонормированного базиса некоторого линейного подпространства, нас интересует такой базис, чтобы средняя ошибка проекции вектора $\mathbf{x}$ в соответствующее линейное подпространство, вычисленная следующем образом

$$\mathrm{E}\left[\|QQ^T\mathbf{x}-\mathbf{x}\|^2\right]=\operatorname{tr}\Sigma-\operatorname{tr}Q^T\Sigma Q, \quad (4.18)$$

была минимально возможной.

На этот раз придётся ввести целую матрицу множителей Лагранжа $H=H^T$ и записать целевую функцию в следующем виде:

$$\Phi(Q)=\operatorname{tr}\Sigma-\operatorname{tr}Q^T\Sigma Q+\operatorname{tr}H\left(Q^TQ-I\right). \quad (4.19)$$

Нетрудно взять производную от этого выражения по каждому из компонент искомой матрицы Q и получить следующее уравнение для условного экстремума:

$$\Sigma Q=QH, \quad (4.20)$$

используя симметрию H и разложение $H=V\Lambda V^T$, где V - ортогональная матрица, а Λ - диагональная, перепишем уравнение в следующем виде:

$$\Sigma(QV)=(QV)\Lambda, \quad (4.21)$$

произведение QV говорит нам о том, что искомый базис можно поворачивать и отражать как угодно, и на результат задачи (т.е. минимальность средней ошибки) такая операция не повлияет. Это достаточно очевидное и ожидаемое свойство, поэтому положим $V=I$:

$$\Sigma Q=Q\Lambda, \quad (4.22)$$

Так как матрица Λ диагональная, то для каждого столбца $\mathbf{q}$ матрицы Q можно записать отдельно знакомое нам уравнение:

$$\Sigma\mathbf{q}=\lambda\mathbf{q}. \quad (4.23)$$

Значение средней ошибки проекции равно:

$$\mathrm{E}\left[\|QQ^T\mathbf{x}-\mathbf{x}\|^2\right]=\operatorname{tr}\Sigma-\operatorname{tr}Q^T\Sigma Q=\sum_{j=1}^{n}\lambda_j-\sum_{k:\ \mathbf{q}_k\in Q}\lambda_k. \quad (4.24)$$

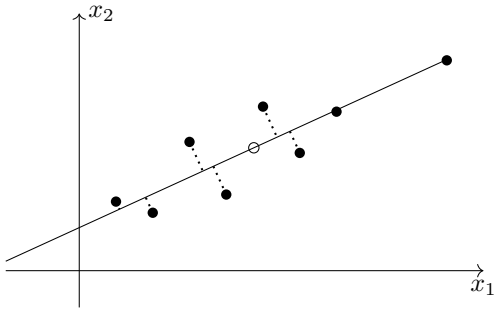

Рис. 4.1: Применение анализа главных компонент для проекции выборки двумерных векторов $\mathbf{x}_i$ на одномерное линейное подпространство. Точками показаны положения выборки $\mathbf{x}_i$; кружком показано положение выборочного среднего; сплошной линией, проходящей через положение выборочного среднего, показано направление первого главного компонента $\mathbf{q}_1$; пунктирные линии показывают минимизируемое расстояние между измерениями и проекцией на первый главный компонент. По сравнению с методом наименьших квадратов на Рис. 3.1, минимизируемые расстояния измеряются не параллельно вертикальной оси, а перпендикулярно к модельной прямой.

Понятно, что решение задачи состоит в том, чтобы взять m максимальных по значению собственных значений, а соответствующие им собственные вектора (т.е. главные компоненты) использовать в качестве базиса искомого линейного подпространства. На Рис. 4.1 приведён графический пример применения анализа главных компонент к выборке точек на плоскости.

Мы ещё раз подтвердили свойство главных компонент: сохранение информации при понижении размерности. Понятно, что вектор после проекции на линейное подпространство меньшей размерности остаётся похож на самого себя с наименьшей ошибкой в смысле квадрата нормы невязки.

4.3 Сингулярное разложение и анализ главных компонент

На практике обычно известна некоторая конечная выборка из N реализаций случайного вектора $\mathbf{x}$, поэтому нам интересно, как вычислять главные компоненты, и проекции реализаций случайного вектора на линейное подпространство меньшей размерности.

Наивный способ мог вы состоять в том, чтобы сначала найти выборочную оценку матрицу ковариации этого вектора:

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \otimes \mathbf{x}_i, \tag{4.25}$$

где $\otimes$ обозначает внешнее векторное произведение, а затем вычислить собственные значения и собственные вектора этой матрицы тем или иным алгоритмом.

Однако, на практике предпочитают другой способ, основанный на вычислении сингулярного разложения. Назовём *матрицей данных* или *матрицей измерений* матрицу X, строки которой состоят из доступных реализацией случайного вектора $\mathbf{x}$. Таким образом X имеет N строк и n столбцов. Тогда оценку ковариации (4.25) можно записать в матричном виде:

$$\hat{\Sigma} = \frac{1}{N} X^T X \tag{4.26}$$

Запишем сингулярное разложение матрицы X:

$$X = USV^T, \tag{4.27}$$

где U — ортогональная матрица размера N, V — ортогональная матрица размера n, S — прямоугольная матрица, все элементы которой равны нулю, кроме элементов, стоящих на её диагонали. Заметим, что

$$\hat{\Sigma} = \frac{1}{N} X^T X = \frac{1}{N} V S^2 V^T, \tag{4.28}$$

откуда видно, что матрица V состоит из колонок, соответствующих главным компонентам, а диагональная матрица S^2 состоит из соответствующих им собственных значений. Значит матрица данных, записанная в базисе главных компонент, выражается следующим образом:

$$XV = USV^T V = US. \tag{4.29}$$

Напомним, что все элементы матрицы S ниже строки n равны нулю, значит для вычисления произведения US достаточно вычислить и хранить в памяти только первые n колонок матрицы U. Алгоритмы вычисления сингулярного разложения, реализованные в популярных библиотеках, как правило позволяют использовать эту оптимизацию.