**Использование метода главных компонент для анализа надежности цепей поставок**

**Кузнецов В.О.**

Аспирант

Департамент логистики и управления цепями поставок

Национальный Исследовательский Университет «Высшая Школа экономики» (Санкт-Петербург, Россия)

**Kuznetsov V.O.**

Postgraduate student

Department of Logistics and Supply Chain Management

National Research University Higher School of Economics (St.-Petersburg, Russia)

**vokuznetsov@hse.ru**

**Ключевые слова:** обучение без учителя, машинное обучение, метод главных компонент, надежность цепей поставок.

**Keywords**: unsupervised learning, machine learning, principal component analysis, supply chain reliability.

**АННОТАЦИЯ**

Одним из вариантов более гибкого подхода к анализу надежности цепей поставок нам представляется метод главных компонент (PCA). Учитывая большое количество переменных, описывающих цепь поставок, является сложной задачей - проанализировать в двумерном пространстве структуру переменных. Метод PCA позволяет перейти, в рамках анализа зависимостей переменных, от многомерного пространства к маломерному, оставляя для анализа саму полезную информацию, находящуюся в массиве данных. На основе сгенерированного набора данных, в данной работе демонстрируется возможность применения PCA относительно анализа надежности цепей поставок. Сгенерированный набор данных включает в себя наблюдения по 50-ти цепям поставок, описанный пятью переменными. На основе массива данных, максимизировав линейную комбинацию параметров по каждому наблюдению, мы определили коэффициенты нагрузки и оценки каждой из главных компонент. Расчет этих коэффициентов позволил перейти от многомерного пространства к двумерному. Двумерное отображение всех данных, осями которого являются первые две главные компоненты, объясняя 84% дисперсии, позволило увидеть структуру всех цепей поставок, выделить аутсайдеров и лидеров в данном наборе.

**ABSTRACT**

 One of the options for a more flexible approach to analyzing the reliability of supply chains is the principal component analysis (PCA). Given the large number of variables describing the supply chain, it is a difficult task to analyze the structure of variables in a two-dimensional space. Within the analysis of the dependencies of variables, PCA allows you to go from a multidimensional space to a low-dimensional space, leaving for analysis the most informative data that is in the array. Based on the generated data set, this paper demonstrates the possibility of applying PCA to the supply chain reliability analysis. The generated data set includes observations on 50 supply chains, described by five variables. Based on the array, maximizing the linear combination of parameters for each observation, we determined the load coefficients and the estimates of each of the main components. The calculation of these coefficients made possible moving from a multidimensional space to a two-dimensional one. The two-dimensional representation of data (explaining 84% of the variance), whose axes are the first two main components, allows to see the structure of all supply chains in the order to identify outsiders and leaders in this set.

**Введение**

На данный момент нельзя переоценить важность различных метрик для оценки надежности цепей поставок. Функционирование последних становится все более сложным механизмом с набором различных статических и динамических параметров, а также колоссальным количеством взаимосвязей между элементами всей цепи. На таком уровне сложности, учитывая динамический характер цепи поставок, является довольно трудоемким, по нашему мнению, приведение к расчету одного конкретного показателя надежности. В данной работе рассмотрено применение метода главных компонент применительно к анализу надежности цепей поставок.

Данный метод был изобретен К. Пирсоном в 1901 году, но более широкое применение получил в связи с бурным развитием машинного обучения в последние 20 лет. В связи с ростом вычислительных мощностей, и параллельно – объемов данных, накапливаемых у бизнеса, интеллектуальный анализ данных приобрел высокую потребность.

Применение методов машинного обучения применительно к цепям поставок, а в частности, к надежности цепей поставок, является, по нашему мнению, потенциально мощным инструментом, позволяющим избегать грядущие сбои в цепях поставок, понимать, в какой ситуации находится конкретная цепь поставок, и, в соответствии с этим, принимать необходимые про-активные действия. Также, в рамках детального анализа одной цепи поставок, применяя, например, в рамках дескриптивного анализа данных метод PCA, мы можем отслеживать динамику развития одной цепи поставок и видеть варианты нашего развития, какие были совершены ошибки, что позволит успешно провести, при необходимости, реинжиниринг цепи поставок.

**Методология**

В соответствии с [Samuel, 1959], машинное обучение – это класс методов искусственного интеллекта в области компьютерных наук, основанных на статистических методах, которые дают возможность моделям «обучаться» на данных, не будучи явно запрограммированными. Другими словами, область машинного обучения исследует конструирование алгоритмов, которые могут обучаться на наборе данных, а затем делать прогнозы на основе обученной модели.

В соответствии с [James, 2013], задачи машинного обучения делятся на две категории, в зависимости от того, доступен ли обучающий «сигнал» или ответ ($Y$) (также называемый «учителем») в конкретном наборе данных ($X$):

* Обучение с учителем (Supervised Learning, SL)

В контексте данного подхода, нам доступен набор $(p)$ переменных (или предикторов) $X\_{1},X\_{2},…,X\_{p}$, измеренный по $n$ наблюдений. Также, мы имеем вектор зависимой переменной $(Y)$ (ответ или «сигнал» на соответствующее наблюдение), измеренный по $n$ наблюдений. Задача заключается в получении прогноза $Y$, на основе $X\_{1},X\_{2},…,X\_{p}$

* Обучение без учителя (Unsupervised Learning, UL)

Использовать данную группу методов мы можем тогда, когда нам доступен только набор $(p)$ переменных $X\_{1},X\_{2},…,X\_{p}$, измеренный по $n$ наблюдений. Мы не заинтересованы в получении прогноза, т.к. не имеем в данных соответствующего «ответа» на $X\_{1},X\_{2},…,X\_{p}$, по которому будет строиться обучающий алгоритм. Основная цель таких методов – получение скрытых паттернов в данных, т.е. нахождение закономерностей в самих входных данных.

В контексте оценки надежности цепей поставок представляется релевантным использование UL методов, т.к. рассматривая конкретную цепь поставок мы имеем колоссальный объем данных, описывающих саму цепь поставок, но не имеем выражения, что при конкретном наборе входных данных $X\_{1},X\_{2},…,X\_{p}$, надежность будет принимать какое-то значение.

Очевидно, что по сравнению с SL, задачи UL имеют ряд трудностей, связанных с проверкой полученного результата анализа, т.к. при SL мы можем проверить наш обученный алгоритм и качество прогноза на тестовой выборке, которая не входила в первоначальный набор данных. При использовании методов UL мы не сможем оценить правильность нашего алгоритма, т.к. мы не знаем истинного ответа по каждому наблюдению: в задаче нет учителя. Поэтому, задача сводится к определению некой структуры или подгрупп (кластеров) среди переменных или наблюдений.

В применении к концепции надежности цепей поставок может быть применим один из методов UL – анализ главных компонент (Principal Component Analysis, PCA). Учитывая большое количество переменных-предикторов (количество может доходить до нескольких сотен), которые описывают цепь поставок (такими переменными могут быть: количество уровней, количество узлов/объектов, тип транспортировки, использование собственного/стороннего транспорта, доля затрат на логистику от общих затрат компании и др.), мы обязательно столкнемся с ситуацией, когда переменные будут коррелировать между собой. Главные компоненты (Principal Components, PC) позволяют уменьшить количество репрезентативных переменных, которые объясняют основную вариабельность в данных [Jolliffe, 2002]. Анализ главных компонент заключается в расчете главных компонент и последующем их использовании в понимании данных и извлечении информации из данных.

Предположим, что мы хотим визуализировать $n$ наблюдений, измеренных по $p$ переменным, $X\_{1},X\_{2},…,X\_{p}$, в контексте первичного исследования данных. Мы можем сделать это путем анализа двумерных точечных графиков, каждый из которых будет содержать $n$ наблюдений по двум переменным. Количество диаграмм будет описываться следующей формулой:

$\left(\begin{matrix}p\\2\end{matrix}\right)=\frac{p\left(p-1\right)}{2}$ (1)

, где p – количество независимых переменных.

Однако, если в нашем обучающем наборе данных, например, 20 переменных, количество диаграмм будет равно 190, и все, очевидно, не поддаются корректному анализу. Более того, большинство из этих диаграмм будут неинформативны, т.к. каждый из них содержит лишь малую часть информации, представленной в данных. Другими словами, нужно найти малоразмерное пространство представления данных, которое бы отражало максимум полезной информации. PCA позволяет найти такое пространство. Математически, PCA является ортогональной линейной трансформацией, которая, фактически, преобразует данные в новую систему координат таким образом, что первая координатная ось отражает переменную с максимальной дисперсией (первая главная компонента), вторая ось отражает вторую переменную с максимальной дисперсией (вторая главная компонента) и т.д.

Каждое из $n$ наблюдений находится $p$-мерном пространстве, но не все измерения будут информативными. PCA ищет некоторое количество измерений, которые являются наиболее информативными, где информативность, в свою очередь, определяется через то, насколько наблюдения варьируются в контексте каждого измерения [James, 2013]. Каждое из измерений, найденное методом PCA, является линейной комбинацией $p$ переменных.

Первая главная компонента набора $X\_{1},X\_{2},…,X\_{p}$ является нормализованной комбинацией переменных

$M\_{1}=φ\_{11}X\_{1}+φ\_{21}X\_{2}+…+φ\_{p1}X\_{p}$ *,* (2)

которые имеют максимальную дисперсию. Под нормализованностью подразумевается, что $\sum\_{j=1}^{p}φ\_{j1}^{2}=1$.

$φ\_{11},…,φ\_{p1} $являются коэффициентами нагрузки первой главной компоненты, которую образуют вектор главной компоненты, $φ\_{1}=\left(\begin{matrix}φ\_{11}&φ\_{21}&…&φ\_{p1}\end{matrix}\right)^{T}$. Описанное выше ограничение на сумму квадратов обосновывается тем, что большие значения данных коэффициентов будет отражаться в больших значениях дисперсии.

Поскольку мы заинтересованы в дисперсии переменных, мы должны стандартизировать значения переменных по наблюдениям:

$x\_{ij}=\frac{x\_{ij}-μ}{σ}$,

где $x\_{ij}$ – значение по i-наблюдению j-переменной;

 $μ$ – среднее значение j-переменной среди n наблюдений;

$σ$ – среднеквадратическое отклонение j-переменной.

Учитывая, что линейная комбинация по каждому наблюдению $m\_{i1}=φ\_{11}x\_{i1}+φ\_{21}x\_{i2}+…+φ\_{p1}x\_{ip}$ должна иметь максимальную дисперсию, с ограничением $\sum\_{j=1}^{p}φ\_{j1}^{2}=1$, мы получаем целевую функцию:

$\frac{1}{n}\sum\_{i=1}^{n}\left(\sum\_{j=1}^{p}φ\_{j1}x\_{ij}\right)^{2}=\frac{1}{n}\sum\_{i=1}^{n}m\_{i1}^{2} \rightarrow \begin{matrix}max\\φ\_{11,…,}φ\_{p1}\end{matrix}$, где $\sum\_{j=1}^{p}φ\_{j1}^{2}=1$ (3)

Так как, $\frac{1}{n}\sum\_{i=1}^{n}x\_{ij}=0$, среднее $m\_{11},…,m\_{n1}$ будет также равно нулю. Следовательно, целевая функция является дисперсией $n$ значений $m\_{n1}$. Значения $m\_{11},…,m\_{n1}$ являются оценками первой главной компоненты.

После того, как была определена первая главная компонента $M\_{1}$, представляется возможным найти вторую главную компоненту $M\_{2}$. Вторая главная компонента является линейной комбинацией $X\_{1},X\_{2},…,X\_{p}$, которая имеет максимальную дисперсию из всех возможных линейных комбинаций, которые не коррелируют с $M\_{1}$. Оценки $m\_{12},…,m\_{n2}$ второй главной компоненты принимают форму

$$m\_{i2}=φ\_{12}x\_{i1}+φ\_{22}x\_{i2}+…+φ\_{p2}x\_{ip}$$

где $φ\_{2}$ – вектор коэффициентов нагрузки второй главной компоненты.



Рисунок 1 – Отображение двух первых главных компонент для двух переменных (зеленая прямая – первая главная компонента, пурпурная – вторая главная компонента)

Ограничение на отсутствие корреляции $M\_{2}$ с $M\_{1}$ равносильно тому, что направление $φ\_{2}$ перпендикулярно направлению $φ\_{1}$ (Рисунок 1). Расчет $φ\_{2}$ аналогичен формуле 3 с добавлением дополнительного ограничения: $φ\_{2}$ ортогонален $φ\_{1}$.

После того, как были рассчитаны главные компоненты, появляется возможность визуализации данных в маломерном пространстве. Например, мы можем построить график в двумерном пространстве, где по оси X будет проложена первая главная компонента, а по оси Y – вторая.

**Применение PCA для анализа надежности цепей поставок**

В целях применения метода PCA для анализа надежности цепей поставок был сгенерирован набор данных по 50 цепям поставок. Массив описывают 5 переменных:

- *LevelsQuantity* (количество уровней цепи. Например, производство может быть один из уровней, если у компании есть собственное производства, или, также, система распределительных центров тоже будет одним из уровней. Последним уровнем, как правило является уровень, где клиент получает товар)

- *NodesQuantity* (количество узлов в цепи. Данный показатель общее количество конкретных объектов компании. Например, отдельный распределительный центр, магазин или производство будет конкретным узлом)

- *NodeDistance* (среднее расстояние между узлами цепи)

- *ShareOwnTransport* (доля собственного транспорта в парке)

- *LogisticsCost* (доля логистических затрат от общих затрат компании)

В целях анализа зависимостей между переменными мы можем построить попарную матрицу диаграмм рассеяния (Рисунок 2), которая может помочь изучить поведение конкретных переменных при изменении других. Как уже было сказано, данная диаграмма не дает общего представления о структуре объекта. Также, построение 5-мерного пространства не представляется возможным. На рисунке 2, к примеру, в первой строке, можно увидеть зависимость параметра LevelsQuantity с параметрами NodesQuantity, NodeDistance, ShareOwnTransport, LogisticsCost, соответственно.



Рисунок 2 – Попарные точечные диаграммы зависимостей по каждой переменной

В целях получения визуального представления маломерного пространства были рассчитаны необходимые показатели для определения главных компонент: коэффициенты нагрузки (Таблица 1) и оценки главных компонент (Рисунок 4). Строки в таблице 1 отражают название переменной, а столбцы значение коэффициентов нагрузки по каждой главной компоненте (Principal Component, PC). Значения были рассчитаны в среде для статистических вычислений R путем максимизации функции из формулы 3.

Таблица 1 – Коэффициенты нагрузки

|  |
| --- |
|  PC1 PC2 PC3 PC4 PC5 |
| LevelsQuantity 0.2484345 -0.6853636 0.3197587 -0.05646428 -0.60259707 |
| NodesQuantity 0.4340322 -0.5333056 -0.1598087 -0.05670258 0.70600791 |
| nodeDistance -0.5074199 -0.1333310 0.7219576 -0.28219018 0.35198554 |
| shareOwnTransport -0.5052185 -0.3689527 -0.1535928 0.76264634 0.05837759 |
| logisticsCost 0.4870069 0.3032467 0.5721906 0.57648579 0.10548857 |

Дальнейшим шагом является определение весов каждой компоненты, которые рассчитываются исходя из того, какую долю дисперсии каждая из главных компонент объясняет. Видно (Рисунок 3), что первых двух главных компонент может быть достаточно для анализа структуры объектов и поиска новой информации на основе исходного массива данных, т.к. накопленная доля объясняемой дисперсии по первым двум компонентам равна около 0,84.



Рисунок 3 – Кумулятивная дол объясненной дисперсии начиная с первой главной компоненты

Таким образом, мы имеем вектора оценок каждой главной компоненты длиной $n=50$ и вектора коэффициентов нагрузки длиной $p=5$. До расчета данных векторов все переменные были стандартизированы в целях получения среднего, равного нулю, и среднеквадратического отклонения, равного одному. На рисунке отображены две главные компоненты на основе рассчитанных данных. По осям снизу и слева расположены оценки главных компонент, а по осям сверху и справа, соответственно, коэффициенты нагрузки. Данный график является искомым маломерным представлением первоначальных многомерных данных. Анализ структуры цепи поставок и ее положения среди аналогичных объектов теперь становится возможным, т.к. мы имеем двумерное представление данных, которое объясняет 84% всей дисперсии.

Для примера, проанализируем 3 различные цепи поставок.

- Если мы посмотрим на 21-ю цепочку, то увидим, что цепочка поставок в целом функционирует достаточно хорошо. При всей разветвленности сети (большое количество уровней и узлов), цепочка имеет затраты на логистику ниже среднего, при довольно большом уровне собственного парка.

- Анализируя 44-ю цепочку, можно заключить, что это один из лидеров в представленной выборке, т.к. имея очень разветвленную сеть узлов и уровней, данная компания имеет минимальные затраты на логистику. Кроме того, узлы, в среднем, очень удалены друг от друга, что, во-первых, позволяет говорить о глобальном масштабе функционирования компании, во-вторых, такая отдаленность узлов говорит о низкой плотности цепи, что снижает фактор риска сбоя, например из-за природных катаклизмов.

- Безусловным аутсайдером является 35-я цепочка, т.к. имеет один из максимальных уровней затрат на логистику, имея низкую разветвленность цепи.

В общем, по эффективности функционирования визуально можно выделить, например, следующие группы цепочек: лидеры (1, 3, 5, 6, 44), аутсайдеры (4, 35, 38) и «корпорации» (20, 25, 26, 31, 33).



Рисунок 4 – Визуальное представление оценок и коэффициентов нагрузок первых двух главных компонент

**Заключение**

Таким образом, метод главных компонент, по нашему мнению, является одним из потенциально важных шагов в рамках исследования и анализа данных для оценки надежности цепи поставок. Также один из вариантов использования метода PCA – исследование одной конкретной цепи поставок, где в качестве переменных, например, будут те же переменные, которые мы описали а выше, а в роли наблюдений – значение параметра на конкретный момент времени. Это позволит понять, как изменяется совокупная структура показателей цепи поставок в динамике, и, в общем, предпринять необходимые шаги в рамках реинжиниринга цепей поставок.

**Список литературы**

1. Hunter, J. D., Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, Volume: 9, [Issue: 3](https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=4160244), May-June 2007
2. James, G. et al., Introduction to statistical learning, Springer Science and Business
Media, New York, 2013
3. Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28 illus.
4. Samuel, A. L, Some Studies in Machine Learning Using the Game of Checkers, 1959
5. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.