Fachhochschule Südwestfalen
University of Applied Sciences

South Westphalia University

Department of Engineering and Economics

Statistics (Prof. Dr. Buchwitz)

Supervisor: Prof. Dr. Buchwitz

Math stuff for pyspark

Patrick Adrian Ulbrich

Abstract
A brief summary of our ideas.

Keywords: Statistics, Regression, Forecasting

Meschede

9th September 2023

Wir geben Impulse

Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. This paper was not previously presented to another examination board and has not been published.

I am aware that the thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the thesis as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future theses submitted.

Meschede, 9th September 2023.

Patrick Adrian Ulbrich
MatNr: 123454678
Email: curie.marie@fh-swf.de
Corresponding Author

This Declaration must be signed by all listed Authors to be valid.

Checklist

I declare that in this work. . .

. . . the following criteria concerning the research question(s) are fulfilled:

- ☐ This work contains (at least) one clearly defined research question.
- ☐ All research questions will be extensively answered in the end of this work.
- ☐ (Analytical) methods which have been used are essential to answer the research question. A connection between the (analytical) methods and the research question(s) is obvious.
- ☐ All used variables are described. No variables will be analyzed which have not been described. All described variables are important to comprehend the argumentation.

. . . the following criteria concerning the representation of results are fulfilled:

- ☐ Equations of teaching materials (slides, scripts, transcript (dt. *Studienbuch*) will not be repeated. Own calculations (e.g. to generate new variables) will be presented by own equations.
- ☐ This work does not contain any "recipes of calculations", i.e. each calculation is presented in a comprehensible manner.
- ☐ Tables and figures have a caption. Each caption ends with a point.
- ☐ Each figure and table of this work is described, explained and interpreted in form of a text. Always cross-reference to figures and tables. Figures and tables complement text but they do not substitute text.
- ☐ No figure is a pie chart.
- ☐ Listed numbers are based on the notation used in R, i.e. they use a point as decimal point (All numbers are consistently formatted).
- ☐ Presented results consist of four decimal places.
- ☐ To each mentioned number belongs a unit.
- ☐ Additional notes, calculations, sketches and drawings which are written by hand will not be graded. In order to show calculations use the LaTeX notation in RMarkdown. Substitute handwritten sketches and drawings by figures which have been generated for example in R.

. . . the language is used as follows:

- ☐ This work is free of swelling and embellishing words and phrases. This means the focus of the text lies on the object of investigation and not on the *linguistic design* (Academic Rigor).
- ☐ This work does not consist of phrases which are imprecise since you could not decide which word to take (Often slashes "/" are used).
- ☐ This work is free of words that pretend to be precise but which are not precise (e.g. different, some, certain time points).
- ☐ Avoid using modal verbs (especially shall and should; Instead of "In this work the following questions should be examined." –> "In this work the following questions will be examined.")
- ☐ The statements in the text do not contain any or just a few superlatives. Superlatives (= assertions) have to be proven by scientific (!) sources.

. . . and the following is fulfilled:

- ☐ Single sentences do not form paragraphs. Each paragraph consists of several sentences. Paragraphs are forming units in terms of content.
- ☐ Each division of text is subdivided into at least two parts (e.g. chapter 3.1 will only exist if chapter 3.2 exists).
- ☐ All sources of the bibliography are used in this work. All sources which have been cited in this work are listed in the bibliography.
- ☐ All authors have personally signed the declaration of authorship.

All list elements must be checked.

Contents

1 Singular Value Decomposition (SVD)

In the following two chapters the singular value decomposition (SVD) will be briefly explained. In the first subchapter the mathematical background will be layed out. In the second subchapter short references to the implementation of SVD in PySpark will be made. The focus is set on the main things that are important for understanding the general concept of SVD and the implementation in PySpark. References to additional mathematical proofs are made.

1.1 Mathematical Background

A singular value decomposition (SVD) is mainly used to determine the pseudo-inverse of a matrix to solve the linear system of equations that is represented by the matrix. A pseudo-inverse is a generalized inverse matrix. According to Burg et al. (2012, p. 354, definition 3.37), a matrix G must satisfy the following conditions (1) and (2) to be referred to as a pseudo-inverse:

$$AGA = A \tag{1}$$

$$GAG = G \tag{2}$$

To be called a *Moore-Penrose-Inverse* the following condition (3) also has to be met.

$$AG \text{ und } GA \text{ are symmetrical} \tag{3}$$

The Moore-Penrose inverse is denoted by $A^{\dagger}$.

Furthermore, the general form of a SVD can be written as shown in equation (4) (Burg et al. 2012, p. 354, equation 3.425).

$$A = U \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} V^T \tag{4}$$

An alternative way of writing this equation is shown in equation (5) (Duvvuri & Singhal 2016, p. 251 - 252).

$$A = U\Sigma V^T \tag{5}$$

The matrices have the following properties:

- A is the original matrix with *m-rows* and *n-columns*
- U is a column-orthonormal matrix with *m-rows* and *r columns*
- V^T is the transpose of a column-orthonormal matrix with *n-rows* and *r columns*

- Σ is an $r \times r$ diagonal matrix containing non-negative real numbers

The vectors in U are also called the left-singular vectors of A. Respectively, the vectors in V are called the right-singular vectors of A (Apache Spark 2017). The elements of $\Sigma \in \text{Mat}(r; R)$ are non-negative and arranged in descending order. These diagonal values are called the singular values of Matrix A, which is why the equation (4) is called the singular value decomposition of A.

It is further assumed that for each matrix $A \in \text{Mat}(m, n; R)$ there is exactly one Moore-Penrose inverse. The following equation (6) is from Burg et al. (2012, p.355 equation 3.427). The complete mathematical proof of this assumption is not part of this study and can be found in Burg et al. (2012, p. 355 - p. 357).

$$A^{\dagger} = V \begin{bmatrix} S^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T \in \text{Mat}(n, m; \mathbb{R}) \tag{6}$$

The final step in solving the system of linear equations is to find the optimal solution by utilizing the Moore-Penrose-Inverse. According to Burg et al. (2012, p. 357 Satz 3.86), with the Moore-Penrose-Inverse $A^{\dagger} \in \text{Mat}(n, m; \mathbb{R})$, an original matrix $A \in \text{Mat}(n, m; \mathbb{R})$ and a given $b \in \mathbb{R}^m$, the following equation (7) is the solution set of the linear optimization problem.

$$x = A^{\dagger}b + y - A^{\dagger}Ay \quad \text{mit} \quad y \in \mathbb{R}^n \tag{7}$$

Derived from that the optimal solution is shown in equation (8).

$$\hat{x} = A^{\dagger}b \tag{8}$$

As shown, the SVD is mainly a way to calculate the Moore-Penrose-Inverse, which then is used to find the optimal solution for the given matrix. There are multiple methods to calculate the SVD to determine the corresponding matrices shown in equation (4). Typical methods are:

1. Jacobi Method
2. Golub-Kahan-Reinsch algorithm
3. Divide-and-Conquer method

One way of thinking about the singular value decomposition is that the matrix Σ in equation (5) contains the strength of the corresponding components in the two other matrices (Duvvuri & Singhal 2016, p. 252). So one additional way of approximately solving numerical problems (or doing lossy image or data compression in general) is to set the values in the matrix Σ of lower magnitude to zero to reduce the number of relevant rows in the remaining two matrices.

1.2 implementation in PySpark

Apache Spark uses two ways to perform the SVD, depending on the absolute size of the number of rows n or the size of n compared to the number of columns k (Apache Spark 2023c). In the case that n is small (n < 100) or n is small compared to k (n/2 < k) "the Gramian matrix (is computed) first and then the top eigenvalues and eigenvectors are locally computed on the driver" (Apache Spark 2023c). In all other cases $(A^T A)v$ is calculated "in a distributive way and send (...) to ARPACK to compute (ATA)'s top eigenvalues and eigenvectors on the driver node" (Apache Spark 2023c).

It is possible to use an additional optimization step to decrease the calculation time by only taking the top k singular values into consideration as described in Duvvuri & Singhal (2016, p. 252) by setting the parameter k to a specific value (Apache Spark 2017). In our implementation we chose to not use this optimization to arrive at the most accurate solution (k=k, referred to as just k in the first parameter in our function call).

2 QR Dekomposition:

This section describes the theoretical basics of QR decomposition. Following on from that, the second part deals with the mathematical basics. In the third chapter, instructions for the implementation of QR decomposition in PySpark are given. The focus here is on the central aspects that are important for a basic understanding of the QR concept as well as the implementation in PySpark.

2.1 Theoretical basics

Note that the Gram-Schmidt method is used to transform a linearly independent set of vectors into an orthonormal vectorset. In other words, a vector set that has the standard of unity and is orthogonal to each other.

Given a **K x L** matrix **A**, its columns are labeled

$$A_1, ..., A_L.$$

When these columns are linearly independent, they can be transformed into a set of orthonormal column vectors

$$Q_1, ..., Q_L$$

using the Gram-Schmidt method, in which normalization and projection steps alternate. These steps will be presented in the next chapter about mathematical basics (Taboga n.d.).

2.2 Mathematical basics

As already mentioned in the theoretical basics, the QR decomposition is used to describe a matrix with linear independent columns as a product of a matrix Q with orthonormal columns and an upper triangular matrix. According to Burg et al. (2012, p. 310, definition 3.69), a QR decomposition can be performed under the following conditions:

Any regular matrix A can be decomposed into a product **A = QR**, where **Q** is an orthogonal matrix and **R** is a regular triangular matrix. Q is a product of at most **(n - 1)** reflections.

Proof:

Assume that (9) and (10). If (11), then set **S(1) := E**. If (12), then we form with (13) the reflection (14).

$$A = [a_1, ..., a_n] \tag{9}$$

$$E = [e_1, ..., e_n] \tag{10}$$

$$a_1 = |a_1|e_1 \tag{11}$$

$$a_1 = |a_1|e_1 \tag{12}$$

$$u = \frac{a_1 - |a_1|e_1}{|a_1 - |a_1|e_1|} \tag{13}$$

$$S^{(1)} := S_u \tag{14}$$

For this we calculate (15) and thereof (16) with (17).

$$S^{(1)}a_1 = |a_1|e_1 \tag{15}$$

$$A^{(2)} := S^{(1)}A = \begin{bmatrix} r_{11} & * \\ 0 & A \end{bmatrix} \tag{16}$$

$$r_{11} = |a_1| \tag{17}$$

The same step is now performed for **A2**, which means that a mirror **S2** is formed in (18) (or **S2 = unit matrix**), so that in **S2 A2** the first column is filled only with an **r22 > 0**. All the other elements of this column are zero. With (19) follows (20).

Proceeding in this way, in the end we obtain (21), where R is a right triangular matrix. It is regular because the left side is regular. With (22) follows **A = QR** and therefore the proof of the theorem.

$$\mathbb{R}^{n-1} \tag{18}$$

$$S^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & S_2 \end{bmatrix} \tag{19}$$

$$A^{(3)} := S^{(2)}A^{(2)} = \begin{bmatrix} r_{11} & * & ...* \\ 0 & r_{22} & *...* \\ & & A_3 \end{bmatrix} \tag{20}$$

$$S^{(n-1)}S^{(n-2)}...S^{(2)}S^{(1)}A = R \tag{21}$$

$$Q = S^{(1)}S^{(2)}...S^{(n-1)} \tag{22}$$

2.3 Implementation in PySpark

In the implementation in PySpark, according to the information from (Apache Spark 2023), a RowMatrix is created from a vector instance. With this RowMatrix, it is possible to perform various statistical summaries of the columns as well as decompositions. An important decomposition in this scope is the QR decomposition, which takes the form A = QR. Here Q stands for an orthogonal matrix and R for an upper triangular matrix. This type of decomposition enables efficient calculations and analysis of large datasets in Spark environments.

In Spark, there are several functions for calculating QR decomposition, depending on the property of the matrix at hand. In the present use case of a RowMatrix, the tallSkinnyQR() function is best suited because it is optimized specifically for RawMatrices. The computeQR() method is suitable as a generalist for any matrix, but is not optimized for any particular shape and is therefore misfit. The tallSkinnyQR() function has the boolean parameter computeQ as input parameter. With computeQ = True, both R-matrix and Q-matrix are computed. With computeQ = False only the R-matrix is calculated. For the calculation of the betas both matrices are needed, therefore the tallSkinnyQR method is passed the boolean TRUE as input parameter. The Apache Spark documentation was used as a literature source, in particular the page on the RowMatrix class (Apache Spark 2023c). Since the dataMatrix has dimensions n x k, the QR.Q matrix has dimensions n x n and the upper triangular matrix has dimensions n x k.

In the next step the inverse of the R matrix is calculated. Since PySpark is specialized for the calculation of large data sets, there is no direct method to calculate the inverse. For the local calculation of the inverse the method np.linalg.inv() of the numpy library is suitable, because numpy is optimized for the numerical calculations of matrices, vectors and arrays. To use the np.linalg.inv() method correctly, the R matrix is converted to a numpy matrix using the np.asmatrix() function. The np.asmatrix() again expects a numpy array as input parameter to perform the conversion. For this reason the result is passed to QR.R.toArray() which converts the R matrix into a numpy array. The dimension for the inverse of the R matrix is k x k. The inverse is a smaller dimension compared to “n”.

In the next step, the transpose of the Q matrix is formed. Since the RowMatrix in PySpark does not have a transpose method, other distributed approaches are needed. For this, a CoordinateMatrix is created using so-called “MatrixEntry” objects. With QR.Q.rows.zipWithIndex() an index is passed to each vector in the RDD. This is necessary to correctly assign the rows and columns of the transposed matrix later. With flatMap() a function is applied to all elements of the RDD. Since as described “MatrixEntry” objects are necessary for the creation of the CoordinateMatrix, the transformation of the elements in the RDD into a list of “MatrixEntry” objects is done with the help of flatMap() (Apache Spark 2023b). This approach allows efficient computation of the

transposed Q-matrix in a distributed Spark environment, especially to ensure scalability and performance. For the transposed Q-matrix the dimension k x n follows.

Subsequently, the values of the dependent variable “y” are represented as a single-column matrix. In order to multiply “y” with a RawMatrix, a compatible data structure is required. PySpark offers the DenseMatrix as a suitable multiplicant. To create the DenseMatrix accordingly, the data array is required in addition to the input parameters numRows and numCols (Apache Spark 2023a). To achieve this, dataDF.select(“y”).toPandas().to_numpy().ravel() converts the column “y” from the DataFrame “dataDF” into a pandas dataframe and finally into a Numpy array. With ravel() an exclusively one-dimensional vector is stored.

Finally, the matrix multiplications are performed. For this, in the first step “Q_T” and “y” are multiplied with multiply, a function from the PySpark framework. With rows.collect() the calculations of the Spark driver nodes are returned to a local data structure, in this case a Python list is suitable. With np.matmul(), the matrices are multiplied together locally after appropriate conversion. Thus the matrix multiplication k x k * k x 1 is available. In the end, the results of the OLS estimation, the true betas and the total execution time are output. The implementation enables efficient and distributed processing of matrix operations by using Apache Spark, and local matrix multiplication by applying the Numpy library.

3 LU Decomposition

The first paragraph explains the mathematical approach, with particular emphasis on its use in linear systems. The second paragraph explains the divide and conquer approach to LU decomposition of large matrices and how the PySpark and Scipy libraries are used.

3.1 Mathematical Background

In LU decomposition, a matrix A is transformed into the product of matrices L and U. The mathematical formula is:

$$A = LU$$

If problems arise during the application of the transformations, such as a division by 0, a permutation matrix can be used. This permutation matrix also increases the robustness with limited accuracy as well as the numerical stability (Lu 2022, p. 23). The corresponding form is:

$$A = PLU$$

The matrices A, P, L and U are defined as follows:

- A is the origin matrix
- L is a lower triangular matrix with 1 at the diagonal, and 0 above the diagonal
- U is an upper triangular matrix
- P is a permutation matrix

The LU decomposition is often used to calculate the inverse of nonsingular matrices or to calculate the determinant of a matrix. It is also used for solving linear systems (Lu 2022, p. 31-33).

To solve a linear system like $Ax = b$ using LU decomposition, the following steps must be performed as in (Furlan 1997, p. 4):

- 1. Calculate the LU decomposition of $A : A = PLU$
- 2. Solve $P\vec{z} = \vec{b}$ with $\vec{z} = \mathbf{P}^\top \vec{b}$
- 3. Solve $L\vec{y} = \vec{z}$ recursive, start with y_1
- 4. Solve $U\vec{x} = \vec{y}$ recursive, start with x_n

Then the coefficients can be taken from the solution vector.

3.2 Implementation in PySpark

Since there is no direct function for LU decomposition like for QR or SVD included in PySpark, a data parallelism or divide and conquer approach is taken for this. The dataset is divided into equal parts and then the LU decomposit is performed separately for each of these parts. As the variables have a moderately pronounced covariance structure, a weighted average of all parts can be generated.

In the program, the first step is to create the function which calculates the coefficients using the LU decomposition. As an input the function gets a pandas dataframe with the matrix A (features) and the corresponding values b (y). To use the *lu_factor* function (SciPy 2023a) and the *lu_solve* function (SciPy 2023b) from the Scipy library, first a Numpy array is created from A. Following the principle from (3.1), the function *lu_factor* first calculates the LU decomposition from A and stores the LU matrix and the P matrix. Then the *lu_solve* function performs steps 2,3 and 4 from (3.1). The calculated coefficient values x (betas) are returned with the number of rows of the partial data set A (sampleCounts) as Pandas DataFrame.

To split the dataset, the function *.groupBy(spark_partition_id)* is used. This function splits the dataset into n equal sized partitions. n corresponds to the number of different partitions of the RDD. For using the function *.applyInPandas()* each of these data partitions is passed as a Pandas data frame to the function described above. The computation now takes place paralelly in the individual Spark instances.

The return of the function *.applyinpands()* is a DataFrame which contains the coefficients (betas) for each of the parts of the data set, calculated by *lu_solve*, and the number of records (sampleCounts) of the data part. The number of records is required to calculate a weighted average of the result coefficients of the return DataFrame. This weighted average is calculated at the end of the program and contains the result vector of the linear equation system.

4 Citation

References can be cited in three different ways.

Fahrmeir et al. (2016)

Fahrmeir et al. (2016, p. 1058)

(Fahrmeir et al. 2016, p. 1058)

Technical Appendix

```
Sys.time()
```

```
## [1] "2023-09-09 15:03:17 CEST"
```

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22621)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=German_Germany.utf8  LC_CTYPE=German_Germany.utf8
## [3] LC_MONETARY=German_Germany.utf8 LC_NUMERIC=C
## [5] LC_TIME=German_Germany.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] fhswf_0.0.3
##
## loaded via a namespace (and not attached):
##  [1] bookdown_0.29   digest_0.6.29   magrittr_2.0.3  evaluate_0.17
##  [5] rlang_1.0.6     stringi_1.7.8   cli_3.4.1       rstudioapi_0.14
##  [9] rmarkdown_2.17  tools_4.2.2     stringr_1.4.1   xfun_0.33
## [13] yaml_2.3.5      fastmap_1.1.0   compiler_4.2.2  htmltools_0.5.3
## [17] knitr_1.40
```

References

Apache Spark (2017). Apache Spark pyspark.mllib.linalg.distributed.IndexedRowMatrix documentation. https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.mllib.linalg.distributed.RowMatrix.html#pyspark.mllib.linalg.distributed.RowMatrix.computeSVD.

Apache Spark (2023a). Apache Spark pyspark.ml.linalg.DenseMatrix. https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.linalg.DenseMatrix.html.

Apache Spark (2023b). Apache Spark pyspark.mllib.linalg.distributed.CoordinateMatrix. https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.mllib.linalg.distributed.CoordinateMatrix.html.

Apache Spark (2023c). Dimensionality Reduction - RDD-based API. https://spark.apache.org/docs/latest/mllib-dimensionality-reduction.html#performance.

Burg, K, H Haf, F Wille & A Meister (2012). *Höhere Mathematik für Ingenieure Band II - Lineare Algebra*. Berlin Heidelberg New York: Springer-Verlag.

Duvvuri, S & B Singhal (2016). *Spark for Data Science* -. Birmingham: Packt Publishing Ltd.

Fahrmeir, L, C Heumann, R Künstler, I Pigeot & G Tutz (2016). *Statistik. Der Weg zur Datenanalyse*. 8th ed. Springer-Lehrbuch. Berlin: Springer. https://doi.org/10.1007/978-3-662-50372-0.

Furlan, P (1997). Zusätze zum gelben Rechnenbuch LU Zerlegung. *Verlag Martina Furlan Dortmund*.

Lu, J (2022). *Matrix Decomposition and Applications*.

SciPy (2023a). *Scipy API referenc lu factor*. Version 1.11.2. https://docs.scipy.org/doc/scipy/reference/generated/scipy.linalg.lu_factor.html.

SciPy (2023b). *Scipy API referenc lu solve*. Version 1.11.2. https://docs.scipy.org/doc/scipy/reference/generated/scipy.linalg.lu_solve.html#scipy.linalg.lu_solve.

Taboga, M (n.d.). QR decomposition (). https://www.statlect.com/matrix-algebra/QR-decomposition.