Analyse dünnbesetzter Hauptachsen für Frequenzdaten

Tobias Bork

Geboren am 21. November 1997 in Reutlingen

21. Oktober 2019

Bachelorarbeit Mathematik

Betreuer: Prof. Dr. Jochen Garcke

Zweitgutachter: Prof. Dr. X Y

Mathematisches Institut für Numerische Simulation

Mathematisch-Naturwissenschaftliche Fakultät der

Rheinischen Friedrich-Wilhelms-Universität Bonn

Danksagung

The acknowledgments and the people to thank go here, don't forget to include your project advisor…

Inhaltsverzeichnis

Kapitel 1

Einführung

[1] [6] [10] [2] [5] [3] [12] [4] [7] [9] [13] [14] [15] [8]

1.1 Motivation

1.2 Dimensionsreduktionsverfahren

High dimensionality means that the dataset has a large number of features. The primary problem associated with high-dimensionality in the machine learning field is model overfitting, which reduces the ability to generalize beyond the examples in the training set. Richard Bellman described this phenomenon in 1961 as the Curse of Dimensionality where “Many algorithms that work fine in low dimensions become intractable when the input is high-dimensional. “

1.3 Sparse Approximations / Representations

1.4 Interpretierbarkeit

1.5 Compressed Sensing Beispiel

Kapitel 2

Mathematische Grundlagen

2.1 Normen und deren Effekte

2.1.1 l0-Norm

2.1.2 l1-Norm

2.1.3 l2-Norm

2.2 Regression

Lineare Regression (Least Squares)

2.2.1 LASSO

2.2.2 Ridge Regression

2.3 Orthogonalprojektion

Definition 2.1. Zwei Vektoren $\vec{a}$ und $\vec{b}$ sind genau dann orthogonal, wenn ihr Skalarprodukt null ist, also

$$\vec{a} \perp \vec{b} \iff \vec{a} \cdot \vec{b} = 0.$$

Was sind orthogonale, orthonormale Matrizen, orthogonale, orthonormale Basis? Skalarprodukt? Von einem Skalarprodukt induzierte Norm? Projektionsmatrizen?

Allgemeine orthogonale Projektionsmatrix falls keine ONB gegeben ist.

$$\mathbf{P_A} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}$$

Von Wikipedia:

Definition 2.2. Eine Orthogonalprojektion auf einen Untervektorraum U eines Vektorraums V ist eine lineare Abbildung $P_U \colon V \to V$, die für alle Vektoren $v \in V$ die beiden Eigenschaften

- $P_U(v) \in U$ (Projektion)
- $\langle P_U(v) - v, u \rangle = 0$ für alle $u \in U$ (Orthogonalität)

erfüllt.

Allgemeine orthogonale Projektion auf einen affinen linearen Unterraum.

$$P_{U_0}(v) = r_0 + \sum_{i=1}^{k} \frac{\langle v - r_0, w_i \rangle}{\langle w_i, w_i \rangle} w_i$$

WÖRTLICH VON WIKIPEDIA: Der orthogonal projizierte Vektor minimiert den Abstand zwischen dem Ausgangsvektor und allen Vektoren des Untervektorraums bezüglich der von dem Skalarprodukt abgeleiteten Norm $\|\cdot\|$, denn es gilt mit dem Satz des Pythagoras für Skalarprodukträume

$$\|u - v\|^2 = \|u - P_U(v)\|^2 + \|P_U(v) - v\|^2 \geq \|P_U(v) - v\|^2$$

2.4 Matrixzerlegungen

Diagonalisierbarkeit?

2.4.1 Eigenwertzerlegung

Eigenwerte, Eigenvektoren

2.4.2 Singulärwertzerlegung

Singulärwerte

2.5 Signaltheorie

2.5.1 Fouriertransformation

2.5.2 Nyquist-Shannon Abtasttheorem

2.6 Statistik

Varianz, Erwartungswert

2.6.1 Empirische Kovarianzmatrix

2.7 Mannigfaltigkeit

2.8 Dictionary Learning

Kapitel 3

Hauptkomponentenanalyse

3.1 Motivation

Die Hauptkomponentenanalyse ist ein weitverbreitetes multivariates statistisches Verfahren zur Dimensionsreduktion. Multivariate Verfahren zielen darauf ab, die in einem Datensatz enthaltene Zahl der Variablen zu verringern, ohne die darin enthaltene Information wesentlich zu reduzieren. Dadurch können umfangreiche Datensätze strukturiert, veranschaulicht und vereinfacht werden. Somit ist das Verfahren Teil der explorativen Statistik, welche Datensätze hinsichtlich ihrer Zusammenhänge analysiert, da nur ein geringes Wissen darüber vorliegt.

Oftmals können sich Gruppierungen / Cluster bilden.

In vielen Bereichen hat die Hauptkomponentenanalyse erfolgreich Anwendung gefunden. Darunter fallen zum Beispiel die Erkennung handgeschriebener Zahlen, welche zum Beispiel zur automatischen Sortierung von Briefen nach Postleitzahl genutzt wird [10], Gesichtserkennung CITE oder in der Genexpressionsanalyse.

Das dahinterstehende mathematische Problem kann auf verschiedene Weisen beschrieben werden. Zunächst wollen wir die Hauptkomponentenanalyse so konstruieren, dass die Idee des minimalen Informationsverlust im Vordergrund steht. Anschließend werden wir das Problem auf eine Singulärwertzerlegung zurückführen, die auch zur effizienten Implementierung genutzt wird. Des Weiteren werden wir die Hauptkomponentenanalyse als Regressionsproblem betrachten und die geometrische Interpretation weiter verdeutlichen. Zu Schluss werden wir einige theoretische Aussagen zeigen.

3.2 Konstruktion

Gegeben sei ein Datensatz mit n samples und p Variablen. Die zentrale Idee der Hauptkomponentenanalyse besteht darin, die p bestehenden Variablen in r neue, unkorrelierte Variablen zu überführen. Um eine Reduktion der Dimension, also $r < p$ zu erreichen, müssen die bestehenden Variablen ßusammengefasst"werden. Idealerweise sollte bei diesem Prozess möglichst wenig Information verloren gehen. Als Maß für den Informationsgehalt der Daten wird hierbei die Varianz verwendet. Das heißt, je größer die Varianz einer Variable, desto mehr Information birgt sie und desto "wichtigerïst sie. Konkret suchen wir also sukzessive nach einer Linearkombination der bestehenden Variablen. Die entstehenden Vektor zeigen also in die Richtung größter Varianz in unserem Datensatz. Wir nennen sie die Hauptachsen bzw. Hauptrichtungen.

Abbildung Höhe Gewicht

Um dieses Prinzip zu veranschaulichen, wenden wir uns nun einem simplem Beispiel zu. Gegeben seien die Größe [cm] und das Gewicht [kg] zu 1000 Personen

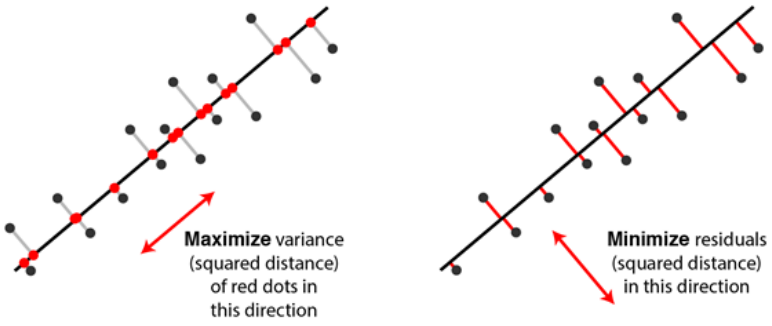

ABBILDUNG 3.1: Die obenstehende Abbildung zeigt die Äquivalenz von Varianzmaximierung und Projektionsdistanzminimierung

(Daten sind simuliert, keine real-world-data) (siehe dazu Abbildung). In diesem Fall ist also $n = 1000$ und $p = 2$. Bei Betrachtung der Abbildung fällt schnell auf, dass die beiden Variablen positiv korreliert sind, d.h. prinzipiell gibt es die folgende Tendenz: Je größer eine Person, desto schwerer ist sie. Wir können

Konkret konstruieren wir Variablen, die sich aus Linerakombinationen der Alten zusammensetzen. Dabei sollen die neuen Variablen der Wichtigkeit nach sortiert sein. In anderen Worten enthält die erste Variable die meiste Information bzw. die größte Varianz, dann die zweite, usw. Die eigentliche Dimensionsreduktion findet dann durch Selektierung statt. Je nach Komplexität des Modells und Informationsverlust können so mehr oder weniger ausgewählt werden. Somit haben wir eine kleine, neue Menge an Variablen, die aber trotzdem den Großteil an Informationen / Varianz beinhaltet.

Bevor wir die Hauptkomponentenanalyse auf den Datensatz anwenden können gibt es noch einen wichtigen Bearbeitungsschritt zu beachten. Wenn eine Variable weniger variiert als eine Andere aufgrund der verwendeten Einheit oder Skala (meter oder kilo) kann dies zu ungewollten Ergebnissen führen. Ohne eine Vorbehandlung der Daten hat so zum Beispiel eine Änderung von 1m die gleiche Bedeutung wie eine Änderung von 1kg. Daher werden die Daten oft einem sog. preprocessing unterzogen. Ein oft verwendetes Verfahren ist die Standardisierung (auch z-Transformation genannt). In diesem Schritt werden die Variablen so transfomiert, dass sie "vergleichbarer"werden. Seien dazu X_i die Zufallsvariablen mit Erwartungswert $E[X_i] = \mu$ und Varianz $Var[X_i] = \sigma^2$. So erhält man die zugehörigen standardisierten Zufallsvariablen Z_i durch Zentrierung und anschließender Division durch die Standardabweichung $Z = \frac{X-\mu}{\sigma}$. Somit gilt:

- $E[Z_i] = 0$ für alle $1 \leq i \leq p$
- $Var[Z_i] = 1$ für alle $1 \leq i \leq p$

Diese Dinge müssen noch irgendwo untergebracht werden im Text:

- Mathematisch wendet man das Verfahren also nicht auf die Kovarianzmatrix, sondern auf die Korrelationsmatrix an.
- Anordnung nach absteigender Varianz bzw. Information
- Hauptkomponente vs Hauptachse

- Wir gehen von einem vollständigem Datensatz aus bei dem keine Einträge fehlen bzw. korrupt sind. Entsprechende Erweiterungen existieren in cite und cite.
- Annahmen: Keine Außreiser, Lineare Verteilung
- No, PCA is not selecting some characteristics and discarding the others. Instead, it constructs some new characteristics that turn out to summarize our list of wines well. Of course these new characteristics are constructed using the old ones
- Indeed, imagine that you come up with a property that is the same for most of the wines. This would not be very useful, wouldn't it? Wines are very different, but your new property makes them all look the same! This would certainly be a bad summary. Instead, PCA looks for properties that show as much variation across wines as possible.
- ortogonale Projectionen

3.2.1 Problemformulierung als Varianzmaximierung

Wir wollen nun die Intuition des minimalen Informationsverlust mathematisch formulieren. Gegeben sei dazu eine Matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, wobei n die Anzahl der Samples bzw. Beobachtungen und p die Anzahl der Variablen ist. Ohne Beschränkung der Allgemeinheit nehmen wir im Folgenden an, dass die Daten zuvor standardisiert wurden. Aufgabe der Hauptkomponentenanalyse ist es nun sukzessive die Richtungen größter Varianz zu finden und anschließend die Daten auf die neu konstruierten Variablen zu projizieren. Mathematisch können erhalten wir die erste Hauptachse dann durch:

$$v_1 = \underset{\|v\|_2=1}{\arg\max} \sum_{i=1}^{n} (X_i v)^2 = \underset{\|v\|_2=1}{\arg\max} \|Xv\|_2^2 = \underset{\|v\|_2=1}{\arg\max} v^T \mathbf{K}_{xx} v$$

wobei $\mathbf{K_{xx}}$ die Kovarianzmatrix ist. Die erste Hauptkomponente, d.h. die Projektion der Daten auf die erste Hauptachse erhält man durch $Z_1 = \sum_{j=1}^{p} v_{1j} X_j$, wobei $v_1 = (v_{11}, \ldots, v_{1p})^T$. Die restlichen Hauptachsen und Komponenten können sukzessive definiert werden.

$$v_{k+1} = \underset{\|v\|=1}{\arg\max} v^T \mathbf{K}_{xx} v$$

unter der Nebenbedingung, dass $v_{k+1}^T v_l = 0, \forall 1 \leq l \leq k$. Man sucht also unter den Richtungen, die orthogonal zu allen bisherigen Hauptachsen sind, diejenige, die die Varianz maximiert. [15] CITE JOLLIFE

Wie wir bereits in THEOREM gesehen haben, entsprechen die Eigenvektoren der Kovarianzmatrix genau den Richtungen maximaler Varianz wie oben definiert. Daher können wir anstatt sukzessiver Berechnung einzelner Hauptachsen die Kovarianzmatrix $\mathbf{K}_{xx}$ diagonalisieren. Da $\mathbf{K}_{xx}$ symmetrisch ist

$$\mathbf{K}_{xx} = \mathbf{V}\mathbf{L}\mathbf{V}^T$$

wobei $\mathbf{V}$ die Matrix der Eigenvektoren ist (Jede Spalte ist ein Eigenvektor) und $\mathbf{L}$ ist eine Diagonalmatrix mit Eigenwerten λ_i in absteigender Reihenfolge. Die Eigenvektoren entsprechen den Hauptachsen und die Projektion der Daten auf die

Hauptachsen wird erreicht durch Multiplikation des Datensatz mit den Eigenvektoren $\mathbf{Z} = \mathbf{XV}$. Die Spalten in $\mathbf{Z}$ sind also die Hauptkomponenten. Die i-te Beobachtung bezüglich der neuen Variablen sind die Zeilen von $\mathbf{XV}$.

Wie wir bereits in CITE gesehen haben, entsprechend diese Richtungen genau den Eigenvektoren der Stichprobenkovarianzmatrix. Die Stichprobenkovarianzmatrix ist gegeben durch $\mathbf{K}_{xx} = \frac{(\mathbf{X}^T\mathbf{X})}{n}$.

If we now perform singular value decomposition of X, we obtain a decomposition X=USV, where U is a unitary matrix and S is the diagonal matrix of singular values si. From here one can easily see that C=VSUUSV/(n1)=VS2n1V, meaning that right singular vectors V are principal directions and that singular values are related to the eigenvalues of covariance matrix via i=s2i/(n1). Principal components are given by XV=USVV=US

.

To summarize:

If X=USV

, then columns of V are principal directions/axes. Columns of US are principal components (ßcores"). Singular values are related to the eigenvalues of covariance matrix via i=s2i/(n1) . Eigenvalues i show variances of the respective PCs. Standardized scores are given by columns of n1U and loadings are given by columns of VS/n1 . See e.g. here and here for why "loadingsßhould not be confused with principal directions. The above is correct only if X is centered. Only then is covariance matrix equal to XX/(n1) . The above is correct only for X having samples in rows and variables in columns. If variables are in rows and samples in columns, then U and V exchange interpretations. If one wants to perform PCA on a correlation matrix (instead of a covariance matrix), then columns of X should not only be centered, but standardized as well, i.e. divided by their standard deviations. To reduce the dimensionality of the data from p to k<p, select k first columns of U, and kk upper-left part of S. Their product UkSk is the required nk matrix containing first k PCs. Further multiplying the first k PCs by the corresponding principal axes Vk yields Xk=UkSkVk matrix that has the original np size but is of lower rank (of rank k). This matrix Xk provides a reconstruction of the original data from the first k PCs. It has the lowest possible reconstruction error, see my answer here. Strictly speaking, U is of nn size and V is of pp size. However, if n>p then the last np columns of U are arbitrary (and corresponding rows of S are constant zero); one should therefore use an economy size (or thin) SVD that returns U of np size, dropping the useless columns. For large np the matrix U would otherwise be unnecessarily huge. The same applies for an opposite situation of np.

3.2.2 Formulierung als Singulärwertzerlegung

$$\mathbf{X} = \mathbf{UDV}^T$$

wobei $\mathbf{D}$ eine Diagonalmatrix mit Elementen $d_1, \ldots, d_p$ in absteigender Reihenfolge, $\mathbf{U}$ eine $n \times p$ und $\mathbf{V}$ eine $p \times p$ orthogonale Matrix. $\mathbf{UV}$ sind die Hauptkomponenten und die Spalten von $\mathbf{V}$ sind die Eigenvektoren von $\mathbf{X}$.

3.2.3 Formulierung als Regressionsproblem

$$\hat{\mathbf{V}}_k = \arg\min_{\mathbf{V}_k} \sum_{i=1}^{n} \left\| x_i - \mathbf{V}_k \mathbf{V}_k^T x_i \right\|^2 + \lambda \sum_{j=1}^{k} \|\beta_j\|^2$$

subject to $\mathbf{V}_k^T \mathbf{V}_k = I_{k \times k}$
[14]

Man projiziert die Daten auf einen k-dimensionalen linearen Unterraum. Man kann zeigen, dass die Lösung dieses Problem genau die ersten k Hauptachsen sind.

Ausgehend von dieser Formulierung als Regressionsproblem werden wir im nächsten Kapitel die Variante der dünnbesetzten Hauptkomponentenanalyse beschreiben.

3.3 Theoretische Aussagen

Theorem 3.1. *PCA always gives unique solution.*

Theorem 3.2 ([11])**.** *Sei* $\mathbf{X} \in \mathbb{R}^n$ *und* $\mathbf{A}_{p,k} = [\alpha_1, \ldots \alpha_k]$

Theorem 3.3. *PCA inconsistent for* $p \gg n$.

3.4 Limitations

PCA Limitations

Model performance: PCA can lead to a reduction in model performance on datasets with no or low feature correlation or does not meet the assumptions of linearity.

Classification accuracy: Variance based PCA framework does not consider the differentiating characteristics of the classes. Also, the information that distinguishes one class from another might be in the low variance components and may be discarded.

Outliers: PCA is also affected by outliers, and normalization of the data needs to be an essential component of any workflow.

Interpretability: Each principal component is a combination of original features and does not allow for the individual feature importance to be recognized.

Kapitel 4

Dünnbesetzte Hauptkomponentenanalyse

Ein Nachteil der Hauptkomponentenanalyse ist, dass sich die neuen Variablen meist aus einer Linearkombination aller bestehenden Variablen zusammensetzt. Dies macht es besonders für hochdimensionale Daten schwierig die Hauptachsen zu interpretieren. Oft können somit nicht die relevanten features/Variablen herausgelesen werden. Es kann durchaus passieren, dass nicht alle Variablen relevant zur Strukturerkennung sind.

4.1 Motivation

4.2 Problemformulierung

NP-schwere Formulierung

4.3 Relaxation / Approximation Ideen

4.4 Konstruktion

Sparse PCA Kriterium.

$$(\hat{\mathbf{A}}\hat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} \left\| x_i - \mathbf{A}\mathbf{B}^T x_i \right\|^2 + \lambda \sum_{j=1}^{k} \left\| \beta_j \right\|^2 + \sum_{j=1}^{k} \lambda_{1,j} \left\| \beta_j \right\|_1$$

subject to $\mathbf{A}^T\mathbf{A} = I_{k\times k}$

4.5 Theoretische Aussagen Sparse PCA

z.B. wie werden neue Varianzen berechnet

Kapitel 5

Implementierung

5.1 Implementierung nach original paper

Eigene Implementierung

5.2 Implementierung in scikit-learn in python

5.3 Laufzeitvergleich

Kapitel 6

Anwendung

6.1 Anwendung auf Simulationsdaten

Vergleich Tabelle PCA / Sparse PCA (Loadings)

6.2 Der Datensatz

6.3 Anwedung auf Frequenzdaten

6.4 Auswertung der Ergebnisse

6.5 Vergleich mit PCA Resultaten

6.6 Hyperparameter

Veränderung des Hyperparameters und dessen Effekte

6.6.1 Zeit

6.6.2 Effekt auf Resultate

Kapitel 7

Ausblick / Zusammenfassung

7.1 Einsetzbarkeit

Wann ist die Methode sinnvoll einzusetzen?

7.2 Übertragbarkeit

Übertragbarkeit auf andere Datensätze

7.3 Ongoing Research / Weitere Techniken

Literatur

[1] Michael Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. English. Bd. 1. Springer Science+Business Media, 2010, S. 376.

[2] Rémi Gribonval, Rodolphe Jenatton und Francis R. Bach. „Sparse and spurious: dictionary learning with noise and outliers“. In: *CoRR* abs/1407.5155 (2014). arXiv: 1407.5155. URL: http://arxiv.org/abs/1407.5155.

[3] Iain M. Johnstone und Arthur Yu Lu. „On Consistency and Sparsity for Principal Components Analysis in High Dimensions“. In: *Journal of the American Statistical Association* 104.486 (2009). PMID: 20617121, S. 682–693. DOI: 10.1198/jasa.2009.0121. eprint: https://doi.org/10.1198/jasa.2009.0121. URL: https://doi.org/10.1198/jasa.2009.0121.

[4] Jean Ponce Guillermo Sapiro Julien Mairal Francis Bach. „Online Dictionary Learning for Sparse Coding“. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. ACM, 2009, S. 689–696. DOI: 10.1145/1553374.1553463. URL: http://doi.acm.org/10.1145/1553374.1553463.

[5] Francis R. Bach Rodolphe Jenatton Guillaume Obozinski. „Structured Sparse Principal Component Analysis“. In: *Artificial Intelligence and Statistics (AISTATS)* 9 (2010). URL: https://arxiv.org/abs/0909.1440.

[6] Holger Rauhut Simon Foucart. *A Mathematical Introduction to Compressive Sensing*. English. Bd. 1. Birkhäuser Basel, 2013, S. 625.

[7] Robert Tibshirani. „Regression Shrinkage and Selection via the Lasso“. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), S. 267–288. URL: http://www.jstor.org/stable/2346178.

[8] Robert Tibshirani u. a. „Least angle regression“. In: *The Annals of Statistics* 32.2 (2004), 407–499. DOI: 10.1214/009053604000000067. URL: http://dx.doi.org/10.1214/009053604000000067.

[9] Ryan J. Tibshirani. „The Lasso Problem and Uniqueness“. In: (2012). URL: https://arxiv.org/abs/1206.0313.

[10] Jerome Friedman Trevor Hastie Robert Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. English. Bd. 2. Springer-Verlag New York, 2009, S. 745.

[11] R. Vidal, Y. Ma und S. Sastry. *Generalized Principal Component Analysis*. Bd. 1. Interdisciplinary Applied Mathematics. Springer New York, 2016. ISBN: 9780387878119. DOI: 10.1007/978-0-387-87811-9. URL: https://books.google.de/books?id=I9H7CwAAQBAJ.

[12] Kazuyoshi Yata und Makoto Aoshima. „Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations“. In: *Journal of Multivariate Analysis* 105.1 (2012), S. 193–215. URL: https://doi.org/10.1016/j.jmva.2011.09.002.

[13] Hui Zou und Trevor Hastie. „Regularization and Variable Selection via the Elastic Net“. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67.2 (2005), S. 301–320. URL: http://www.jstor.org/stable/3647580.

[14] Hui Zou, Trevor Hastie und Robert Tibshirani. „Sparse Principal Component Analysis“. In: *Journal of Computational and Graphical Statistics* 15.2 (2006), S. 265–286. DOI: 10.1198/106186006X113430. URL: https://doi.org/10.1198/106186006X113430.

[15] Hui Zou und Lingzhou Xue. „A Selective Overview of Sparse Principal Component Analysis“. In: *Proceedings of the IEEE* 106.8 (2018), S. 1311–1320. DOI: 10.1109/JPROC.2018.2846588. URL: https://ieeexplore.ieee.org/document/8412518.