Analyse dünnbesetzter Hauptachsen für Frequenzdaten

Tobias Bork

Geboren am 21. November 1997 in Reutlingen

18. 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

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 weitverbreites 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. Als Teil der explorativen Statistik ...

Somit findet die Hauptkomponentenanalyse in vielen Bereichen Anwendung. Ein paar Beispiele (hand written zip code classification or human face recognition).

Das dahinterstehende mathematische Problem kann auf mehrere 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 ein 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 die Äquivalenz dieser Formulierungen und einige theoretische Aussagen zeigen.

3.2 Einführung

Was heißt überhaupt Hauptkomponente und was ist eine Hauptachse vielleicht an dieser Stelle.

Die zentrale Idee der Hauptkomponentenanalyse besteht darin, die bestehenden Variablen in neue, unkorrelierte Variablen zu überführen, ohne dabei Information zu verlieren. Als Maß für den Informationsgehalt der Daten wird hierbei die Varianz verwendet. 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.

Wie müssen die Daten aufbereitet sein? Zentriert und skaliert? Erklären verschiedener Methoden und deren Auswirkungen für die Daten. Korrelationsmatrix oder Kovarianzmatrix?

3.3 Konstruktion

Gegeben sei eine Matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, wobei n die Anzahl der Samples bzw. Beobachtungen und p die Anzahl der Variablen. Ohne Beschränkung der Allgemeinheit nehmen wir im Folgenden an, dass die Spaltendurchschnitte der Matrix 0 sind, d.h. jede Variable einzeln zentriert ist. Falls dies nicht der Fall ist können wir die Variablen einfach zentrieren. (Eventuell erwähnen, was passiert wenn man die Variablen nicht zentriert) Aufgabe der Hauptkomponentenanalyse ist es nun

3.3.1 Problemformulierung als Varianzmaximierung

Die erste Hauptachse wird definiert als

$$v_1 = \max_{\|v\|_2=1} v^T \sum v$$

. Die Hauptkomponente wird dann definiert durch $Z_1 = \sum_{j=1}^{p} \alpha_{1j} X_j$, wobei $\alpha_1 = (\alpha_{11}, \ldots, \alpha_{1p})^T$ wobei $\sum = \frac{(\mathbf{X}^T\mathbf{X})}{n}$ die Kovarianzmatrix ist. Anschließend werden die restlichen Hauptachsen sequentiell definiert.

$$\alpha_{k+1} = \arg\max_{\|\alpha\|=1} \alpha^T \sum \alpha$$

unter der Bedingung, dass $\alpha^T \alpha_l = 0, \forall 1 \leq l \leq k$.

[15]

Fasst man diese Schritte zusammen kann man eine Eigenwertzerlegung vornehmen. Theorem, dass die Eigenwerte von $X^T X$ die Varianz maximieren.

3.3.2 Formulierung als Singulärwertzerlegung

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

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 $\mathbf{X}$.

3.3.3 Formulierung als Regressionsproblem

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

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

[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.

3.4 Theoretische Aussagen

Theorem 3.1. *PCA always gives unique solution.*

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

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

Kapitel 4

Dünnbesetzte Hauptkomponentenanalyse

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

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.