Ludwig–Maximilians–Universität

Institut für Statistik

Theorie und Anwendung der partiellen kleinste Quadrate Regression (PLS)

Masterarbeit

Autor: Uwe Pipiorke

Betreuer: Prof. Dr. Christian Heumann

München, 5.9.2018

Inhaltsverzeichnis

1 Einleitung

Die partielle kleinste Quadrate Regression (Partial Least Squares Regression – kurz: PLS, bzw. PLSR) schätzt lineare Regressionsmodelle, mithilfe der Projektion auf die latente (unsichtbare) Struktur. Diese latente Struktur von Zielgrößen – oft nur eine Variable – zu den p Einflußgrößen, also die Kovarianz, bildet die Ausgangssituation für die Schätzung. Hier unterscheidet sich die PLS von der Faktorenanalyse, welche Varianz basiert arbeitet, d.h. ausschließlich in den Einflußgrößen die Kovarianz auswertet.

Mit einer PLS ergeben sich zudem erweiterte Möglichkeiten bei der Schätzung von Modellen. So ist es möglich das Problem $n < p$, welches ansonsten bei Regressionstechniken ein ernsthaftes Problem darstellen kann, auf angenehme Weise anzugehen. Der Preis für diesen Vorzug ist bereits ein erheblich erschwerter Zugang zum zweiten Moment der Schätzung. Das mag ein Grund dafür sein, daß die Methode in der Statistiker-Welt eher ein Nischen-Dasein führt.
Falls $n \ll p$ vorliegt, stellt Multikollinearität in den Einflußgrößen ein massives Problem dar. Eine Möglichkeit für die Lösung des Problems kann in funktionalen Datenanalysen zu suchen sein, deren Zweck im Verarbeiten sehr vieler Daten-Spalten liegt. Die PLS, als Alternative hierzu, hat sich offenbar bewährt. Beispielsweise stellt sie in der Chemometrie, zu der auch die Nahinfrarotspektroskopie zählt, eine etablierte Methode dar.

Insbesondere wird die PLS für Vorhersagen verwendet. Dafür ist das erste Moment der Schätzung, also der Erwartungswert, i.d.R. bereits ausreichend.

Die Gesellschaft Deutscher Chemiker[1] beklagt den „black-box“ Charakter, und damit den Verlust an Kontrolle, von Analyse-Software. Der Vorwurf ist sicher nicht ungerechtfertigt und nachvollziehbar. Beispielsweise ist Literatur, welche die PLS konsequent bis auf die numerische Ebene herunter durchdringt, nach dem heutigen Stand dennoch als rar einzustufen.

Der Zweck dieser Arbeit besteht darin, ein tieferes Verständnis in die klassischen PLS-Techniken zu erhalten und ihre Anwendung zu zeigen.

Die verwendeten Rohdaten stammen vom Institut für Pflanzenernährung der TUM aus Weihenstephan und wurden von Herrn Prof. Urs Schmidhalter, dem Lehrstuhlinhaber, genehmigt.

—

Zu einigen Theorie-Kapiteln existieren flankierend Minimal-Beispiele als Matlab- bzw. R-Skripte, die zu einem guten Verständnis beitragen sollen. In Kapitel 2 werden die verwendeten Daten beschrieben und auf einige ihrer Besonderheiten eingegangen.

Das Kapitel 3 versteht sich im Wesentlichen als Grundlage bzw. Vorbereitung für die Regressionstechnik in der PLS. Am Anfang wird das klassische lineare Modell erläutert. Der Ridge-Schätzer ist als Erweiterung des klassischen Modells ein erster vager Hinweis

[1] https://www.gdch.de/ (30.11.2017)

in Richtung PLS und mit der Hauptkomponenten-Regression als dimensionsreduzierendes Verfahren wird eine Methoden-Kombination zum Beseitigen von Multikollinearität motiviert. Abgeschlossen wird das Kapitel mit einer Glättungstechnik.

In Kapitel 4 sollen grundlegende Techniken der Matrixzerlegung verstanden werden. Es dient als Vorbereitung. Schrittweise wird auf die Faktorenanalyse und ihre Eigenheiten hingeleitet, die sich bei der Hauptfaktorenanalyse offenbaren. Die Vorstellung der Begriffe *latente Struktur* und *Faktor* sind über die Faktorenanalyse offenbar intuitiver erfaßbar. So hat die Hauptkomponentenmethode bzgl. der Daten-Projektion einige Ähnlichkeiten zur PLS, vor allem aber bzgl. der Notation. Die Diskussion anhand der Hauptfaktorenanalyse zeigt auch in Richtung Reliabiltät von Faktoren.
Das Kapitel wird mit der Moore-Penrose-Inverse abgerundet, die als Verbindung auf dem Weg von der Ridge-Regression zur PLS verstanden werden kann.

Die PLS vereinigt die Faktorenanalyse und die lineare Regression gleichsam in sich. Kapitel 5 erklärt die Theorie der uni-/ und multivariaten PLS. Herleitungen sollen das Konzept beleuchten und zeigen, wie man zu einer Lösung kommen kann. Von den Begrifflichkeiten ist im vorherigen Kapitel das Grundlegende beschrieben. Trotzdem die Gleichungen im Vergleich zum Kapitel 4 bisweilen identisch erscheinen, der Matrixzerlegungsprozeß geschieht auf eine teilweise vollkommen abweichende Art und Weise.
Genau wie die Hauptkomponentenmethode offenbart sich die PLS als sehr robustes Verfahren. Erwähnenswert sind die automatisch enthaltenen Berechnungen, die zur Erhaltung der Faktoren-Reliabilität beitragen. Die faktorenanalytische Sichtweise soll auch zum Verstehen einiger Formeln beitragen.

Das Kapitel 6 beschreibt den praktischen Teil. Anfangs wird auf Korrekturmaßnahmen – wie z.B. Glättung – eingegangen, die auf dem Weg zu einem geeigneten Modell hilfreich sein werden. Dabei wird viel Aufmerksamkeit der Beurteilung von Ausreißern beigemessen. Da bivariate Zielgrößen zur Verfügung stehen wird neben der univariaten Technik (PLS1) auch die Erweiterung im Multivariaten (PLS2) gezeigt.

Mitunter gibt es zu den Abbildungen ergänzende bzw. alternative Darstellungen, die in den R-Skripten verfügbar gehalten sind. Man kann es als weitgehend kommentarlosen „elektronischen Anhang“ verstehen, dessen Abbildung vielmehr Redundanz als weiteren Aufschluß erbringen würde.

—

> Zu der deutschen Selbstabneigung gehört leider auch, dass Deutsche vielfach ihre eigene Sprache nicht mehr sprechen wollen, dass sie sich Gott weiß wie weltläufig vorkommen, wenn sie Englisch sprechen. Das ist zumal in der Wissenschaft verhängnisvoll, wo das Deutsche immer mehr verschwindet. Für mich ist mein »geliebtes Deutsch«, um mit Goethes *Faust* zu reden, von dem unablösbar, was ich denke und schreibe.[2]

[2] Dieter Borchmeyer, Theater- und Literaturwissenschaftler. (2018, 7, 27). Süddeutsche Zeitung Magazin. Nr. 30, S. 21.

2 Datenbeschreibung

Die vorliegenden elektronischen Daten sind quasi reellwertig. Bei gebrochenen, insbesondere bei den reellen Zahlen werden auf dem Computer Fehler in ihrer Darstellung auftreten, da sie unvermeidbar in ihrer Stellenzahl abgebrochen (gerundet) werden. Auch durch mathematische Operationen, z.B. $\sqrt{2} \approx 1,414213\ldots$, können reelle Zahlen entstehen, die einen bereits vorhandenen Fehler im Mittel weiter vergrößern.

2.1 Zahlen in Gleitkommadarstellung

Die Zahl 123,45 kann in Exponentialdarstellung umgeschrieben werden: $1,2345 * 10^2$, wobei die Ziffernfolge 12345 vor der 10er-Potenz als Mantisse bezeichnet wird. Das Komma wurde in dieser Darstellung um zwei Stellen nach links verschoben, welches mit dem Exponenten 2 ausgeglichen wird. Abgekürzt kann man 1.2345e2 bzw. 1.2345e+2 schreiben, welches ebenso 123,45 entspricht. Diese e-Notation nennt man auch Fließkommadarstellung.
Reelle Zahlen haben in ihrem gebrochenen Anteil keine Periode. Deshalb ist es unmöglich, sie in endlicher Länge fehlerfrei darzustellen. Spätestens, seit Computer mit 32 Bit Architektur (theoretisch maximal adressierbarer Hauptspeicher: 2^{32} Byte = 4 GB) existieren, ist das *long real* Zahlenformat etabliert, welches eine 64-Bit Darstellung von reellen Zahlen gestattet. In der Software R ist das Format mit *double* umschrieben.
Diese 64 Bit sind aufgeteilt in ein Vorzeichen-Bit, 11 Bit für den Exponenten und den verbleibenden 52 Bit für die Mantisse. Daraus ergeben sich folgende Konsequenzen:

Bei einer Breite von 11 Bit ergeben sich für den Exponenten $2^{11} - 1 = 2047$ Darstellungen. Für kleine Zahlendarstellungen nahe der Null muß er auch negative Werte annehmen können. Um das Exponenten-Vorzeichen einzusparen, wird im Prozessor der Exponent mit dem offset $2^{10} - 1$ subtrahiert, welches einer Verschiebung seines Wertebereichs entspricht. Unter Beachtung der eingeschlossenen Null ist der Exponent auf dem Intervall $[2-2^{10}\,,\, 2^{10}]$ definiert, womit ein Wertebereich von -1022 bis 1024 realisiert wird.
D.h., das halboffene Intervall der reellwertigen Maschinenzahlen beträgt $[2^{-1022}\,,\, 2^{1024})$ bzw. umformuliert in das dezimale Zahlensystem für die untere Intervallgrenze:
$10^{-1022*\log_{10}(2)} = 10^{-307,6526556} = 2.22507386e-308$ und für die obere Grenze:
$10^{1024*\log_{10}(2)} = 10^{308,2547156} = 1.79769313e+308$,
welches in R mit c(.Machine$double.xmin , .Machine$double.xmax) abrufbar ist.
Die Mantissenbreite von 52 Bit realisiert eine Genauigkeit der Gleitkommazahl von $\log_{10}(2^{52}) = 52 * \log_{10}(2) = 15,65$ effektiven Dezimalstellen. Im Vergleich zur 32-Bit Darstellung (vgl. (Faires & Burden, 1994, Seite 12)), bei welcher eine 24 Bit Mantisse vorliegt, d.h. 7,22 effektive Dezimalstellen resultieren, entspricht es einer doppelten Genauigkeit bzgl. der Dezimalstellenzahl.

Als mittlere Grenze der Zahlenauflösung kann die Maschinengenauigkeit[3] ϵ_m herangezogen werden. ϵ_m ist abschätzbar mit dem kleinsten darstellbaren Wert größer Null:
$\epsilon_m = 2^{-52} = 10^{-52*\log_{10}(2)} = 10^{-15,65\ldots} = 10^{0,3464402255} * 10^{-16} = 2.22044605e-16.$

[3] https://de.wikipedia.org/wiki/Maschinengenauigkeit (15.5.2018)

In R ist sie über .Machine$double.eps abrufbar.
Dann gilt: $1 + \epsilon_m = 1$. Die Abweichung ϵ_m geht sicher im Rundungsfehler unter. Das muß bedeuten, daß reelle Zahlen in ihrer Gleitkommadarstellung nur innerhalb eines gröberen Intervalls beschreibbar sind. Sie auf 16 Dezimalstellen genau zu beschreiben, entspricht einem „best case“ Szenario, welches i.A. bei der Darstellung fast sicher nicht erreicht wird.

Beim Auffinden von numerischen – nicht algebraisch erhältlichen – Lösungen x über Iteration kann ein Abbruchkriterium mithilfe des relativen Näherungsfehler ε definiert werden:

$$\varepsilon = \left| \frac{x_i - x_{i+1}}{x_i} \right| \ , \ x_i \neq 0. \tag{1}$$

Aufgrund der begrenzten Zahlenauflösung kann der relative Fehler nicht beliebig klein gerechnet werden. Als Toleranz eine maximale Genauigkeit von 15 Dezimalstellen vorzugeben, liegt – bei sonst keinen weiteren Fehler verursachenden Operationen – gerade noch im Rahmen des Möglichen. Eine Toleranz von 10^{-8} anzustreben kann mitunter schon kritisch sein, weil mit kleiner werdendem Fehler auch der der Grenznutzen einer Approximation abnimmt. D.h., wird die Toleranz zu klein gewählt, kann eine Näherung viel Rechenzeit beanspruchen bzw. im ungünstigsten Fall nicht zur Konvergenz führen.

2.2 Inhaltliche Beschreibung

Die Zielgrößen Y können als bivariate Ausprägungen ($q = 2$) betrachtet werden. Hier existieren Bodenprobenkonzentrationen, im Umfang von 180, für Kohlenstoff (C) und Stickstoff (N). Bei den Einflußgrößen X handelt es sich um Nahinfrarotspektroskopie-Daten[4] (NIRS). Mit *Nah* liegt die Betonung auf dem kurzwelligen Infrarotlichtbereich[5], welcher unmittelbar oberhalb des sichtbaren Rot-Spektrums beginnt und eine Spannweite von etwa 760 nm (Nanometer) bis 3 μm (Mikrometer) aufweist. Dieser Wellenlängenbereich ist für das menschliche Auge nicht mehr sichtbar.
Es sind in den Einflußgrößen (Spektren genannt) *funktionale Daten*[6] aufgezeichnet, die auf den Wellenzahlen von 12000 bis 3699 basieren. Diese liegen auf einem regulären Gitter der Breite vier, welches jeweils auf Vielfachen von sieben folgenden Knoten eine Breite von drei aufweist. Das ergibt mit der Abschätzung $p = 1 + 7 * \frac{12000-3699}{6*4+3}$ ca. 2153 geschätzte Knoten, die für diesen Datensatz als Meßpunkte bzw. Kovariablen für die aufgezeichneten Wellenlängen tatsächlich vorliegen. Die Wellenzahlen in X sind in Wellenlängen konvertierbar:

$$\text{Wellenlänge } [nm] = 10^7/\text{Wellenzahl}.$$

Nach der Konvertierung ergibt sich eine Spannweite von ca. 833 bis 2703 nm.
Beim Bestrahlen einer Stoff-Probe mit einer Wellenlänge (idealtypisch monochromatisches Licht[7]) interessiert der Grad der Absorption, also die Eindringtiefe des Lichts. Der Grad der Absorption wird für jede einzelne Wellenlänge als Koeffizient protokolliert. Daraus resultiert eine Zeile in X, die Absorptions-Kennlinie.

[4] http://www.analytik.de/content/view/4329/851/ (4.6.2017)
[5] https://de.wikipedia.org/wiki/Infrarotstrahlung (4.6.2017)
[6] üblicherweise als Funktion auf die Zeit gemeint; aber auch für andere Skalen, wie hier z.B. Spektren möglich
[7] http://www.chemie.de/lexikon/Monochromatisches_Licht.html (5.6.2017)

Ein Problem in der Spektroskopie ist, daß selbst für ein und dieselbe Probe – allein nur durch das Herausnehmen und Wiedereinsetzen der Probe – andere Kennlinien entstehen. In (Tillmann, 1996) auf Seite 13 wird das als Parallelverschiebungen der Spektren aufgrund von Streulichteffekten, d.h. entlang der Absorptions-Skala, beschrieben. Die Verschiebung läßt sich als systematischer Fehler auffassen. Auch Messungen bei unterschiedlichen Umgebungstemperaturen tragen zur Entstehung des systematischen Fehlers bei. Er läßt sich verifizieren mit einer Serie wiederholter Messungen.
Ein zweites Problem besteht darin, daß jedes Meßgerät eine andere Charakteristik besitzt. Selbst baugleiche Meßgeräte vom selben Hersteller produzieren auf dieselbe Probe leicht unterschiedliche Absorptions-Kennlinien, die nicht einfach durch Parallelverschiebung auszugleichen sind. Es ist deshalb wichtig, zu wissen, auf welchem Gerät eine Messung durchgeführt wurde. Jedes Gerät besitzt seine spezifische Kennlinie. Diese ist sogar noch einmal abhängig von der verbauten Strahlungsquelle. Wird eine Lampe ausgetauscht, entsteht eine dauerhaft neue Kennlinie, die auf die Messung moduliert. Dieses Problem stellt sich bei den vorliegenden Daten nicht. Alle Messungen wurden mit demselben Spektrographen *Bruker Vector 22/N* erhoben.
Ein drittes Problem bei der Messung ist im zufälligen Meßfehler zu suchen, welcher sich als Rauschen offenbart. Er ist unsystematisch und als unbekannt anzusehen. Folglich sind die Einflußgrößen stochastisch. Für die in Kapitel 3 beschriebenen Regressionstechniken bedeutet es eine unbekannte Verzerrung in den Regressionskoeffizienten inkludiert zu haben. Meßwiederholungen der Einflußgrößen können die Verzerrung verkleinern.

Der Datensatz enthält 1347 Einträge. Sie stammen von vier Versuchsplänen aus Weihenstephan, die teils zu verschiedenen Zeiten und in verschiedenen Bodentiefen erhoben wurden. Die Abkürzungen OB und UB entsprechen Oberboden bzw. Unterboden.

Etikett	*Jahr*	*Versuchs− plan*	*Bodentiefe [cm]*	*Anzahl Proben*	*Messwie− derholung*	*# Ziel− grösse Y*
04-021	2004	021	OB 0-20	96	3*	96
14-021	2014	021	OB 0-25	96	3	21
14-022	2014	022	OB 0-25	96	3	23
14-020	2014	020	OB 0-20	64	3	22
15-016	2015	016	OB 0-25	36	4	10
15-016u	2015	016	UB 25-50	36	4	8
			Summen:	424		180

(*) Messung Nr. 191 liegt in sechsfacher Wiederholung vor.

Tab. 1: Struktur der Daten

Es liegen klassierte Daten vor, die in sechs Kategorien unterteilt sind. Addiert man die Anzahl der Proben, erhält man $n = 424$ gruppierte Mittelwerte. Das ist die relevante Zeilenzahl an unabhängigen Einträgen. D.h., der Datensatz hat ca. fünfmal so viele Spalten wie Zeilen: $n \ll p$. Die Summe aus Probenanzahl∗Meßwiederholung ergibt dann die Zahl der physischen Einträge von 1347. In den Zielgrößen Y finden sich insgesamt aber nur 180 Einträge. Beim Verbinden der Zielgrößen mit den Einflußgrößen gehen von den 424 gruppierten Mittelwerten in X noch einmal 244 Einträge verloren. Auf Seite 35 schreibt (Tillmann, 1996): „In der Praxis haben sich 100 bis 200 Proben als ausreichend erwiesen.“.

Im folgenden Plot sind alle 180 Zielgrößeneinträge für Kohlenstoff (C) und Stickstoff (N) klassenweise entsprechend Tab. 1 inkl. einer Regressionsgeraden /-gleichung abgebildet:

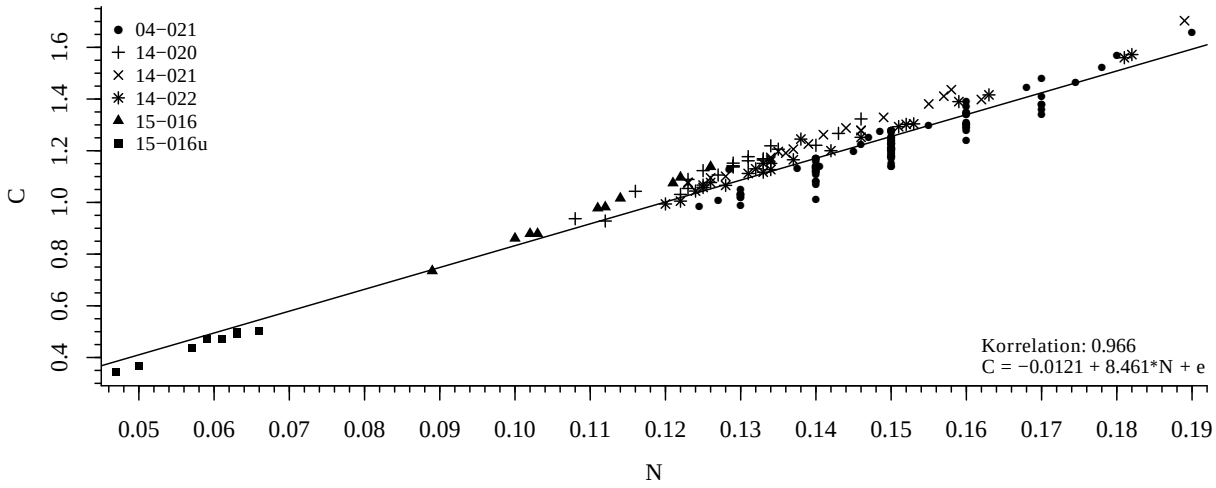


Abb. 1: Zusammenhang zwischen Kohlenstoff- und Stickstoff-Konzentrationen.

Der Anstieg der Gerade ist signifikant. Alle acht Einträge der Unterbodenklasse sind von den restlichen Daten vollständig separiert, verhalten sich aber konform zur angepaßten Funktion. In dieser tieferen Bodenregion sind die Vorkommen für C und N geringer.

Aus dem Versuchsplan 016 sei eine Messung beispielhaft abgebildet:

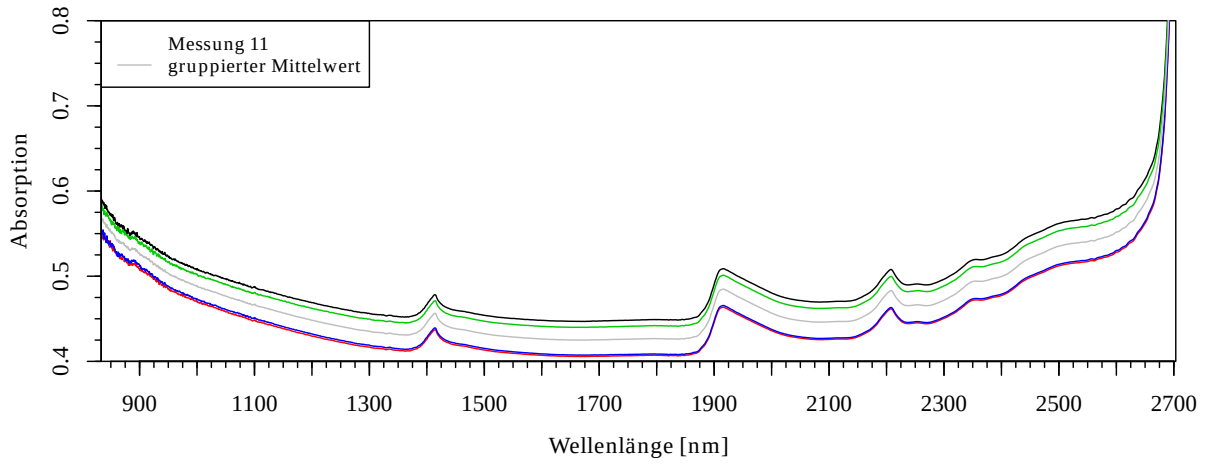


Abb. 2: Spektren in Meßwiederholung.

Bei den Maxima im Spektrum handelt es sich um Resonanzspitzen, die (nach Augenmaß) Hinweise auf die gemessenen chemischen Verbindungen geben können. Im kurzwelligen Nahinfrarot, also bei hohen Frequenzen, erscheinen die Absorptionen zunehmend rauh. Dem Anschein nach sieht es wie ein Rauschen aus. Die Aufnahme einer Probe benötigt eine gewisse Zeitdauer. Je feiner die Auflösung vorgegeben ist, umso mehr Daten müssen bei der Aufnahme abgespeichert werden. Das verlängert das Zeitfenster für eine Aufnahme. Während die Zeit verstreicht dreht sich der Behälter mit der Probe aber um einen von Null verschiedenen Winkel weiter. Gelingt die Aufnahme also nicht schnell genug, macht sich dieses „Verwackeln“ als Rauschen bei den höchsten Frequenzen, also den kleinsten

Wellenlängen, zuerst bemerkbar: Diese modellieren die unerwünschte Bewegung als Unschärfe mit, während bei niedrigeren Frequenzen größere Drehwinkel erforderlich wären, um ein Rauschen herbeizuführen.
Auf dem gruppierten Mittelwert (graue Kurve in Abb. 2) wird die Rauheit geglättet. Das ist ein Hinweis für einen zumindest gering ausgeprägten stochastischen Störprozeß. Meßwiederholungen können das Rauschen vermindern. Das Ausmitteln kann als Registrierung – vgl. (Schmid, 2016a) – aufgefaßt werden, welches die Auflösung verbessert.

Folgende Übersicht[8] beschreibt die Wellenlängen-Bereiche der Resonanzen – Absorptionsbanden genannt – von organischen Verbindungen. R ist eine Abkürzung für das Restmolekül, welches für die Molekül-Schwingung unbedeutend ist:

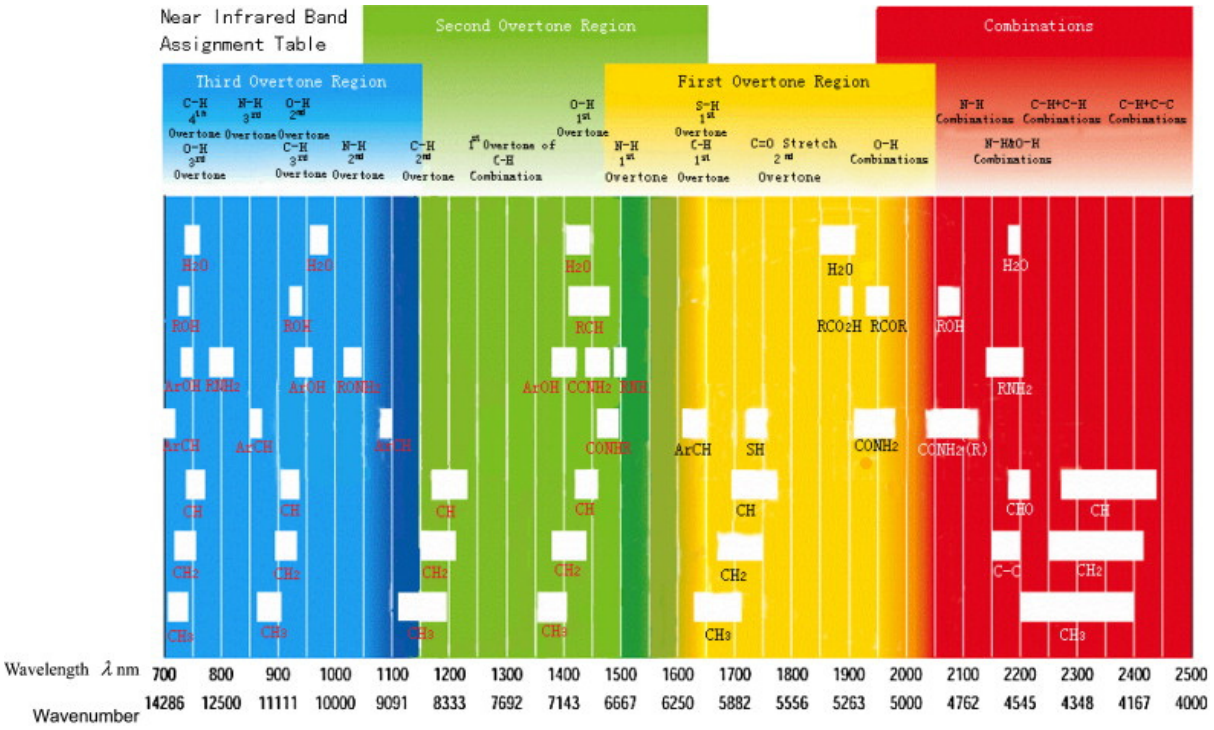


Abb. 3: Absorptionsbanden von organischen Verbindungen im nahen Infrarot.

Mithilfe der Karte läßt sich bzgl. Abb. 2 grob abschätzen, daß bei Wellenlängen von knapp über 1400 nm Kohlenwasserstoffe mit dem Vorsatz CH_3 bzw. CH_2, bei 1920 nm $CONH_2$[9] – also eine Variante mit simultan C und N – und bei ca. 2200 nm Kohlenstoffverbindungen, aber auch Verbindungen mit NH_2-Anteil angeregt werden. Die Grundschwingungen der Kohlenwasserstoffvorsätze CH_i erstrecken sich breitbandig von 2200 – 2440 nm. Verbindungen mit Kohlenstoff überdecken demnach einen Bereich von 2200 bis ca. 2500 nm.
Im langwelligen Bereich oberhalb von 2000 nm liegen die Grundschwingungen der Moleküle. Darunter, d.h. bei höheren Frequenzen, entstehen Oberwellen der Grundschwingungen, die aufgrund ihrer kleineren Amplitude weniger hervorscheinen. In Kapitel 6 wird auf die spezifischen Anforderungen der Daten beim Modellieren ausführlich eingegangen.

[8] durch Claudia Buchhart vom Institut für Pflanzenernährung zur Verfügung gestellt
[9] https://de.wikipedia.org/wiki/Amide (25.3.2018)

In folgendem Korrelogramm sind die Bodenproben, d.h. die Zielgrößen, mit dem Spektrum – den Einflußgrößen – korreliert:

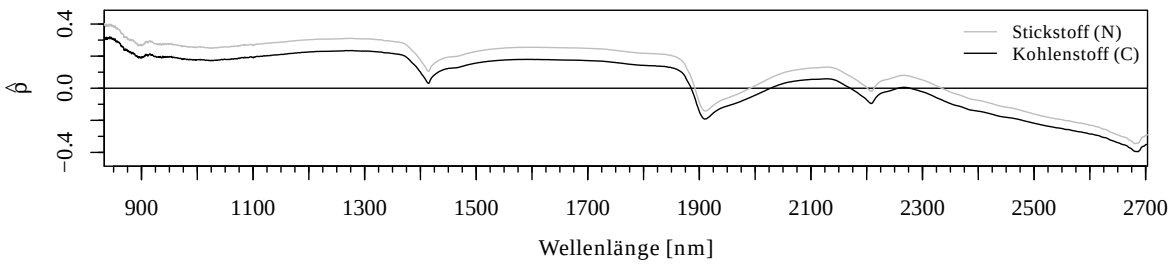


Abb. 4: Korrelationsfunktionen der Bodenproben {N,C} mit dem Spektrum.

Aus der Sicht *funktionaler Daten* sind die Zielgrößen skalar, aber die Spektren funktional. Funktionale Daten sind durch besondere Eigenschaften gekennzeichnet. Die Zahl der Datenspalten in den Spektren ist lediglich der Endlichkeit der Auflösung der Meßvorrichtung geschuldet. Ein Eintrag entspricht einer Funktion, die auf den Spalten definiert wird. Rein gedanklich befindet man sich in einem stetigen Raum, der in der Realität nur durch ein möglichst feinmaschiges Gitter, also diskretisiert, abgebildet werden kann. Die Spalten sind deswegen miteinander korreliert. Die Korrelation ρ hat über das Gitter hinweg eine gewisse Reichweite: Räumliche Effekte wirken ein. D.h., die Korrelation ist abhängig von der Distanz der Spalten zueinander. (Schmid, 2017) beschreibt drei Klassen von Korrelationsfunktionen, in denen die Distanz ausgewertet wird. Unmittelbar benachbarte Spalten sollten maximale Korrelationen aufweisen. Das erinnert an Zeitreihen und damit wird deutlich, daß die Spalten zueinander in einer Ordnung stehen und nicht vertauscht werden dürfen. Zwischen benachbarten Spalten beträgt die Korrelation oft mindestens 0,995.

Die nächste Grafik soll die Eigenschaft illustrieren. Abgebildet ist ein regelmäßiges Gitter (65×65) der $\{X, X\}$-Grundfläche, basierend auf der Korrelationsmatrix $(p \times p)$, in welche $p = 2153$ Einflußgrößen einfließen. Auf den Stützpunkten des regelmäßigen Gitters liegt die Oberfläche („Response“) der symmetrischen Korrelationsfunktion. Da fast überall die Korrelationen nur eine Nuance unterhalb Eins liegen, würde bei linearer Abbildung der Höhenwerte eine nahezu homogen schwarze Fläche resultieren.
Die mit 100 multiplizierten Höhenwerte werden zur Basis 100 logarithmiert, womit eine Spreizung des Gradienten, d.h. der Grau-Verteilung hin zu weiß erreicht wird:

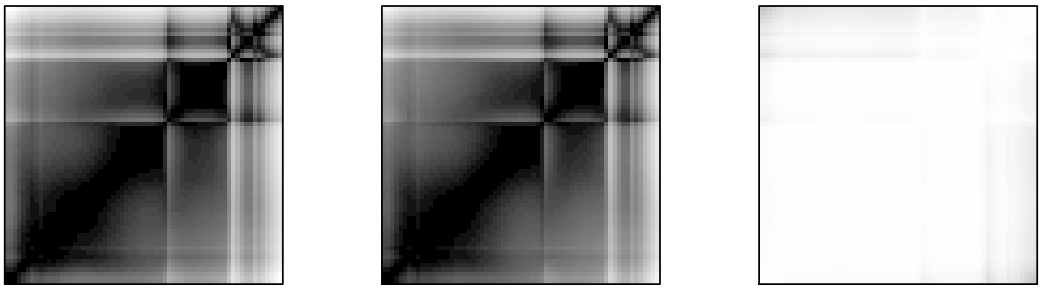


Abb. 5: Korrelation über das Spektrum – links: $n = 424$, Mitte: $n = 180$ Einträge; rechts: Betrag der Differenz beider Oberflächen im Kontrast 5× verstärkt.

Die Grauabstufung deckt von weiß bis schwarz eine Spannweite der Korrelation $\hat{\rho}\ \epsilon\ [0,57; 1]$ ab. Bei stationären Zeitreihen klingt die Korrelation mit steigendem Versatz (lag) exponentiell ab. Allerdings ist hier ein persistenter Verlauf erkennbar, welcher sich in einem zähen Abfall offenbart, vor allem für $n = 180$. Der Rand x könnte evtl. hinreichend mit der Matérn-Korrelationsfunktion (vgl. (Schmid, 2016a), Seite 31-32) modelliert werden. Man kann es aber auch so deuten, daß das Spektrum aufgrund der hohen Korrelationen wenig informativ ist. Das steht auch im Einklang zur geringen Fallzahl. Bei Durchführung einer Spektralzerlegung (Kapitel 4.2) auf die Korrelationsmatrix beschreibt der größte Eigenwert λ für $n = 424$ bereits 95,2% und für $n = 180$ sogar 96,3% der gesamten Variation. Bemerkenswert ist der Anstieg der Korrelation bei Fallzahlverringerung. Um das zu untersuchen, kann zwischen der Korrelationsmatrix, basierend auf den 180 mit derjenigen von 424 Einträgen die Differenz gebildet werden. Die um die Hauptdiagonale bereinigte Dreiecksmatrix der Differenzen ist für deren Interpretation geeignet. Quantile der Differenzen zeigt der folgende Boxplot. Er beschreibt eine sogenannte linkssteile Verteilung, bei welcher der Median mit 0,00255 leicht positiv ausfällt. Er ist nicht Null.

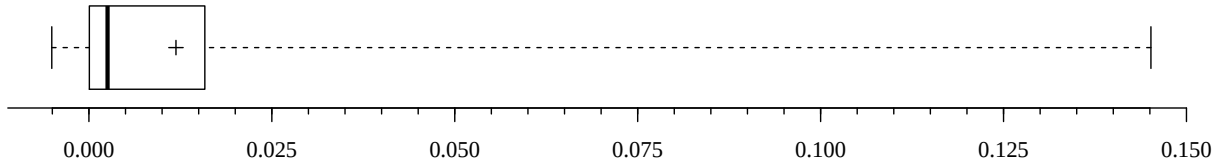


Abb. 6: Boxplot der Differenz der Korrelationen inkl. arithmetischem Mittel (+)

Das arithmetische Mittel beträgt ca. 0,012. Das entspricht in etwa dem Anteil, um welchen der größte Eigenwert bei 180 gegenüber 424 Einträgen mehr an Korrelation erklärt. Das Ausdünnen der Fallzahl geschieht über das Bilden einer Untermenge, welches einer deterministischen Auswahl entspricht. Unter der Annahme, man hätte keine Kenntnis über den datengenerierenden Prozeß, kann man stochastisch über eine Stichprobe argumentieren. Besitzt jedes Objekt die gleiche Wahrscheinlichkeit in eine Stichprobe gezogen zu werden und geschieht die Ziehung ohne Zurücklegen, so handelt es sich nach (Kauermann & Küchenhoff, 2010) um eine einfache Zufallsstichprobe.
Bei einer gemeinsamen Normalverteilung (MVN) ist die bedingte Verteilung[10] ebenfalls normal. Von der bedingten Normalverteilung ist mithilfe der Randverteilungen der Rückschluß auf die gemeinsame Verteilung möglich. D.h., beliebige Schnitte durch die gemeinsame Verteilung bewirken stets Normalverteilungen. Der Rückschluß stellt ein Alleinstellungsmerkmal der Normalverteilung dar. Bei einer Stichprobenziehung sollte sich deswegen die Korrelation im Mittel nicht ändern, falls den Daten eine gemeinsame Normalverteilung zugrunde liegt, d.h., die Information wird durch die Auswahl nicht abgeschwächt. In einer Stichprobe befinden sich weniger Objekte, woraus ein Stichprobenfehler resultiert, der sich umgekehrt proportional zur Größe der Stichprobe verhält. Falls die gemeinsame Verteilung keiner MVN entspricht, sollte die Korrelation der Stichprobe, aufgrund des Informationsverlusts im Mittel ansteigen.

[10] http://www.wiwi.uni-muenster.de/ (11.10.2017)

Mithilfe einer Simulation, bei welcher als Einstellgrößen der Median der Korrelation bei Ausgangsdatenlage $\rho = 0,98$ und die Fallzahlen der vollen bzw. reduzierten Daten eingehen, wird mit einer einfachen Zufallsstichprobe nach jedem Durchlauf die sich einstellende Korrelation notiert. Danach wird eine neue Verteilung über einen Zufallsgenerator aufgebaut. Am Schluß wird über alle Durchläufe gemittelt. Dabei interessiert, in welche Richtung der simulierte Wert von der Vorgabe driftet. Entsteht eine Verzerrung (Bias)? Dies wurde einerseits für eine bivariate Normalverteilung und andererseits für eine Poissonverteilung ($\lambda = 5$) gegen eine Exponentialverteilung ($\lambda = 2$) versucht.
Am Streuplot ist das Resultat jeder Ziehung für die Normalverteilung eingetragen, inkl. des Durchschnittswertes (waagerechte Linie) der Korrelationen:

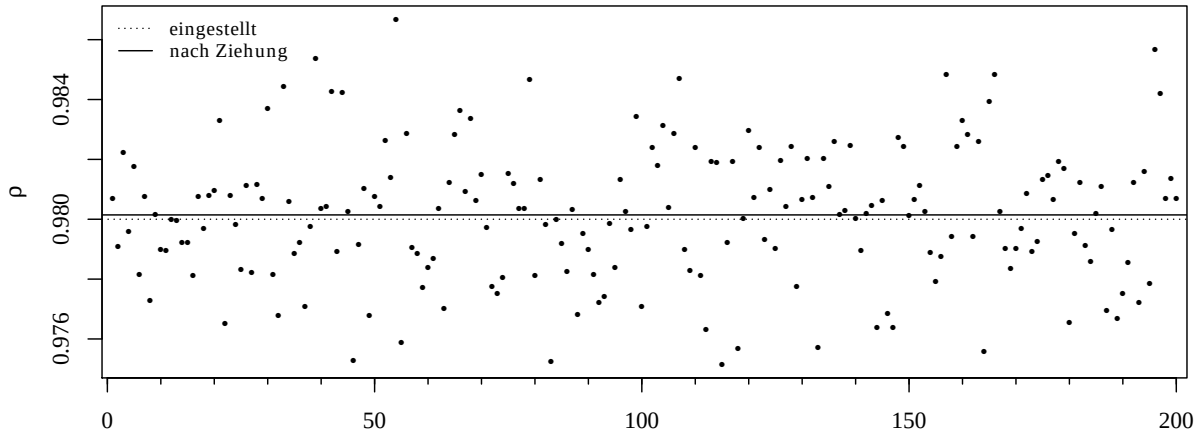


Abb. 7: 200 simulierte Korrelationen für $\rho = 0,98$.

Die Grafik zeigt lediglich das Bild einer möglichen Ausprägung bei 200 Ziehungen. Mit jedem Neustart einer Simulation sind andere Verläufe für ρ vorstellbar. Hier im Bild – des ersten von 15 Neustarts – übersteigt die mittlere Korrelation der reduzierten Daten die der vollen Daten. Eine Unterschreitung von ρ ist auch möglich. Werden die mittleren Korrelationen gemeinsam ausgemittelt, sollte sich jenes globale Mittel dann möglichst nahe an ρ befinden. Für $\rho = 0,98$ beträgt die relative Abweichung ca. $2 * 10^{-5}$.
Bei Simulationen unter Abwesenheit der Normalverteilung an den Rändern kann die eingestellte Korrelation ρ nicht mehr punktgenau vorgehalten werden. Um gleiche Startbedingungen zu erhalten, wird die Korrelation der normalverteilten Daten auf das resultierende ρ abgeglichen. Immerhin kann damit noch die Differenz der Korrelation reduzierte vs. volle Daten ausgewertet werden. Im besten Fall beträgt sie dann Null. Die gemittelte Differenz, also der Bias, fällt bei nicht normalverteilten Daten fast immer positiv und vor allem größer als bei der Normalverteilung aus. Hingegen ist der MSE der Differenzen bei der bivariaten Normalverteilung defacto größer. Für aussagekräftigere Beurteilungen bzgl. des MSE müßte mit verschiedensten Nichtnormal-Verteilungen experimentiert werden.

Schlußfolgerung:
Der Anstieg der Korrelation bei Fallzahlausdünnung ist offenbar nicht ungewöhnlich. Wird die Normalverteilung durch Daten ohnehin nicht perfekt angenähert, ist ein Anstieg der Korrelation durchaus vorstellbar.

3 Regression

Der Begriff kann als Synonym für ein inverses Problem aufgefaßt werden. I.d.R. soll ein überbestimmtes Gleichungssystem[11] gelöst werden. Es werden nun ausführlich zwei Regressionsmethoden motiviert, Kleinste Quadrate und Ridge-Regression. Für das Beschreiben der Ridge-Regression ist das Verstehen der ersten Methode hilfreich.
Die kurze Erwähnung eines alternativen Ansatzes im dritten Unterkapitel dient der Überleitung zur Faktorenanalyse. Im vierten Unterkapitel wird eine Glättungstechnik erklärt, deren Grundlage das erste Unterkapitel ist.
Wenn in der Matrix X auf p Spalten argumentiert wird, müßte wegen der zusätzlich vorhandenen Scheinvariable genauer auf $p+1$ Spalten eingegangen werden. Dies ist in den Darlegungen insofern berücksichtigt, daß p in diesem Kapitel für $p+1$ steht. Damit sind die Dimensionen etwas übersichtlicher schreibbar.

3.1 Kleinste-Quadrate Schätzung (KQ)

Das Modell

$$\vec{y} = X\vec{\beta} + \vec{\epsilon} \tag{2}$$

soll den Zusammenhang zwischen der Zielgröße $\vec{y}$ und den festen Einflußgrößen X $(n \times p)$ erklären. In der ersten Spalte der Designmatrix X ist ein Vektor mit Einsen enthalten, welcher als Scheinvariable für die Berechnung des Absolutgliedes notwendig ist. Vektor $\vec{y}$ wird als eine unkorrelierte Zufallsgröße aufgefaßt und beinhaltet demnach keine Meßwiederholungen[12] für eine Beobachtung. Die Zielgröße $\vec{y}$ kann als Linearkombination von $\vec{\beta}$ $(p \times 1)$, in Abhängigkeit von X, plus Fehler aufgefaßt werden.

Modellannahmen:

- Der Erwartungswert der Fehler ist Null: $\mathbb{E}(\vec{\epsilon}) = \vec{0}_n$.
  Das ist eine Mindestanforderung an das Modell.
- Die Fehler sind voneinander unabhängig, d.h., sie sind unkorreliert:
  $\operatorname{cov}(\epsilon_i, \epsilon_j) = \mathbb{E}(\epsilon_i, \epsilon_j) = 0 \ , \forall i \neq j$.
- Die Fehler sind identisch, d.h. die Varianz der Fehler ist konstant und damit skalar als σ^2 separierbar:
  $$\operatorname{cov}(\vec{\epsilon}) = \mathbb{E}\left(\vec{\epsilon}\vec{\epsilon}^{\,T}\right) = \sigma^2 I_n. \tag{3}$$

Das sind die Annahmen des klassischen Modells, vgl. (Mittnik, 2015, Kapitel 1).
Wird $\vec{\beta}$ mit dem Maximum-Likelihood Ansatz (ML) berechnet kommt zwingend eine Verteilungsannahme der Residuen hinzu. Sind die Residuen normalverteilt: $\vec{\epsilon} \sim \mathcal{N}\left(\vec{0}_n, \sigma^2 I_n\right)$, dann entspricht die ML-Lösung einer KQ-Lösung.
Mit dem Instrumentarium der Regression wird $\vec{y}$ in einen deterministischen und zufälligen Anteil zerlegt. Dies geschieht über ein Varianzkriterium, den Quadratsummen, auch

[11] http://www.tm-mathe.de/Themen/html/uberbestimmte_gls_theorie.html (1.8.2017)

[12] „Wiederholte“ Messungen z.B. bei redundanten Zeilen in der Designmatrix, indem die Zielgrößeneinträge vervielfacht werden. Eine Regression darauf produziert bzgl. der Schätzer zu optimistische Standardfehler (beachte Skript Regression.r).

als Totalen bezeichnet. Das Kriterium ist als Gütemaß geeignet und wird beispielsweise in (Fahrmeir, Heumann, Künstler, Pigeot & Tutz, 2016) im Kapitel 3.6.3 über die Streuungszerlegung beschrieben.

Die Zerlegung wird definiert als Gesamtstreuung = Modellstreuung + Fehlerstreuung:

$$(n-1)\,\mathbb{V}(\vec{y}) = (\vec{y}-\bar{y})^T(\vec{y}-\bar{y}) = \left(\bar{y}-X\vec{\beta}\right)^T\left(\bar{y}-X\vec{\beta}\right) + \left(\vec{y}-X\vec{\beta}\right)^T\left(\vec{y}-X\vec{\beta}\right). \quad (4)$$

Wobei die Fehlerstreuung minimal sein soll, im Sinne eines Verlustes[13], welcher nicht durch das Modell erklärt wird. Gemäß Formel (2) ist $\vec{\epsilon} = \vec{y} - X\vec{\beta}$. Die Varianz $\mathbb{V}(\vec{y})$ impliziert ein quadratisches Verlustkriterium. Hiermit läßt sich eine Verlustfunktion definieren:

$$f\left(\vec{\beta}\right) = \vec{\epsilon}^{\,T}\vec{\epsilon} = \left(\vec{y}-X\vec{\beta}\right)^T\left(\vec{y}-X\vec{\beta}\right) \to \min_{\vec{\beta}}.$$

Für das Auffinden eines Minimums wird die Ableitung Null gesetzt:

$$\frac{\partial f}{\partial \vec{\beta}} = -2X^T\left(\vec{y}-X\vec{\beta}\right) \stackrel{!}{=} \vec{0}_p \quad \Leftrightarrow \quad -X^T\vec{y} + X^TX\vec{\beta} = \vec{0}_p.$$

Die Lösung für $\vec{\beta}$ entspricht einem Kleinste-Quadrate-Schätzer:

$$\hat{\vec{\beta}}_{KQ} = \left(X^TX\right)^{-1}X^T\vec{y}\,,\ (p\times 1). \quad (5)$$

Als Lösung der zweiten Ableitung resultiert eine quadratische Form: $2X^TX$. Negative Eigenwerte können aufgrund der strikt positiven Hauptdiagonale nicht auftreten – das sichert das Minimum der Optimierung ab. Die Schätzung für $\vec{\beta}$ erfordert zwingend ein invertierbares Kreuzprodukt X^TX. Damit es regulär wird muß voller Spaltenrang für X gelten, d.h. $n \geq p$ ist eine notwendige Bedingung. Das aus (2) nach $\vec{\epsilon}$ umgestellte Modell liefert bei Einsetzen von (5) für die geschätzten Residuen:

$$\hat{\vec{\epsilon}} = \vec{y} - X\hat{\vec{\beta}} \stackrel{(5)}{=} \vec{y} - X\left(X^TX\right)^{-1}X^T\vec{y} = \vec{y} - H\vec{y} = (I_n - H)\vec{y} = Q\vec{y}. \quad (6)$$

Die Matrizen H, Q entsprechen idempotenten $n \times n$ Projektionsmatrizen. Mithilfe der Spur (Summe der Diagonalelemente) kann zügig deren Rang ermittelt werden. Q mit dem Rang $n-p$ ist orthogonal zu H, welche vom Rang p ist: $QH = 0$ $(n \times n)$. Aufgrund des Verlustes in den Rängen ergeben sich Konsequenzen für die Schätzung:

$$\begin{aligned} \operatorname{cov}\left(\hat{\vec{\epsilon}}\right) &\stackrel{(6)}{=} \operatorname{cov}(Q\vec{y}) \stackrel{(2)}{=} \operatorname{cov}\left(Q\left(X\vec{\beta}+\vec{\epsilon}\right)\right) \stackrel{QX=0}{=} \operatorname{cov}(Q\vec{\epsilon}) \\ &= Q\,\mathbb{E}\left(\vec{\epsilon}\vec{\epsilon}^{\,T}\right)Q^T = Q\sigma^2 I_n Q^T = \sigma^2 QQ^T = \sigma^2 Q. \end{aligned} \quad (7)$$

Die Folgen sind beachtlich: In den geschätzten Residuen $\hat{\vec{\epsilon}}$ spiegeln sich heterogene Varianzen und auch Korrelationen wider. Von den obigen Modellannahmen verbleibt bei einer Schätzung nur noch die Mindestanforderung. Die Varianz der Beobachtungen auf die geschätzte Regressionsfunktion ist schätzbar:

$$\hat{\sigma}^2 = \frac{\hat{\vec{\epsilon}}^{\,T}\hat{\vec{\epsilon}}}{n-p} \stackrel{(6)}{=} \frac{(Q\vec{y})^TQ\vec{y}}{n-p} = \frac{\vec{y}^{\,T}Q^TQ\vec{y}}{n-p} = \frac{\vec{y}^{\,T}Q\vec{y}}{n-p} = \frac{\vec{y}^{\,T}\left(I_n - X\left(X^TX\right)^{-1}X^T\right)\vec{y}}{n-p}. \quad (8)$$

[13] Der Begriff *Verlust* impliziert eine entscheidungstheoretische Sichtweise.

Diese Notation findet sich u.a auch in (Hartung & Elpelt, 1999, Kapitel II) auf Seite 84.

Geometrische Deutung:
Die Zielgröße $\vec{y}$ wird in den p-dimensionalen Designraum projiziert. Dann resultiert für die Projektion: $\hat{\vec{y}} = H\vec{y}$. Beim univariaten Modell kann man sich die Projektion anhand Abb. 11 von Seite 23 vorstellen: $\vec{b}$ alias $\vec{y}$ wird orthogonal auf $\vec{a}$ alias X projiziert. $\vec{c}$, die Projektion, entspricht der Schätzung $\hat{\vec{y}}$. Dann entspricht der Anstieg in $\vec{d}$ der Differenz der Ortsvektoren $\left(\vec{b}, \vec{c}\right)$, den geschätzten Residuen $\hat{\vec{\epsilon}}$. D.h., diese Residuen stehen orthogonal auf dem Design: $X \perp \hat{\vec{\epsilon}}$. Allerdings wird $\hat{\vec{\epsilon}} = Q\vec{\epsilon}$ von n in $n - p$ Dimensionen projiziert. $\hat{\vec{\epsilon}}$ kann nicht mehr unkorreliert sein. Mit zunehmenden Umfang an Einträgen ($n \gg p$) wird die Problematik der Korrelationen aber gemildert.

Wird in (4) mit dem Term der linken Seite dividiert verbleibt links Eins. Anschließend wird der Term, welcher die Fehlerstreuung beinhaltet, subtrahiert. Man erhält die prozentual erklärte Varianz, auch Bestimmtheitsmaß R^2 genannt. Für die wahren Werte müssen Schätzungen eingesetzt werden. Bei Substitution der Fehler $\vec{\epsilon}$ mithilfe von (6) folgt schließlich für den Anteil der erklärten Varianz: $\mathrm{R}^2 = 1 - \vec{y}^{\,T} Q \vec{y}$.

Die Verzerrung von $\hat{\vec{\beta}}$:

$$\mathrm{Bias}\left(\hat{\vec{\beta}}, \vec{\beta}\right) = \mathbb{E}\left(\hat{\vec{\beta}}\right) - \vec{\beta}. \tag{9}$$

$\hat{\vec{\beta}}$ ist ein unverzerrter Schätzer, denn:

$$\mathbb{E}\left(\hat{\vec{\beta}}\right) \overset{(5)}{=} \mathbb{E}\left(\left(X^T X\right)^{-1} X^T \vec{y}\right) \overset{(2)}{=} \mathbb{E}\left(\left(X^T X\right)^{-1} X^T \left(X\vec{\beta} + \vec{\epsilon}\right)\right) \overset{X^T\vec{\epsilon} = \vec{0}}{=} \mathbb{E}\left(\vec{\beta}\right) = \vec{\beta}.$$

Dann ist die Verzerrung $\mathbb{E}\left(\hat{\vec{\beta}}\right) - \vec{\beta} = \vec{\beta} - \vec{\beta} = \vec{0}_p$ nicht existent.

Die Kovarianzmatrix von $\hat{\vec{\beta}}$ beträgt:

$$\begin{aligned}
\mathrm{cov}\left(\hat{\vec{\beta}}\right) &\overset{(5)}{=} \mathrm{cov}\left(\left(X^T X\right)^{-1} X^T \vec{y}\right) \overset{(2)}{=} \mathrm{cov}\left(\left(X^T X\right)^{-1} X^T \left(X\vec{\beta} + \vec{\epsilon}\right)\right) \\
&= \mathrm{cov}\left(\vec{\beta} + \left(X^T X\right)^{-1} X^T \vec{\epsilon}\right) \overset{\vec{\beta} \perp X^T\vec{\epsilon}}{=} \mathrm{cov}\left(\vec{\beta}\right) + \mathrm{cov}\left(\left(X^T X\right)^{-1} X^T \vec{\epsilon}\right) \\
&\overset{\mathbb{V}(\vec{\beta}) = \vec{0}}{=} \left(X^T X\right)^{-1} X^T \mathrm{cov}\left(\vec{\epsilon}\right) X \left(X^T X\right)^{-1} \overset{(3)}{=} \sigma^2 \left(X^T X\right)^{-1}, \ (p \times p).
\end{aligned}$$

Approximativ, also mit wachsendem n, konvergiert der Schätzer $\hat{\vec{\beta}}$ gegen eine Normalverteilung:

$$\hat{\vec{\beta}} \overset{a}{\sim} \mathcal{N}\left(\vec{\beta}, \sigma^2 \left(X^T X\right)^{-1}\right). \tag{10}$$

Definition des mittleren quadratischen Fehlers (MSE)

Die Kovarianzmatrix in (10) läßt sich bei Anwenden des Varianz-Verschiebungssatzes schreiben als: $\mathbb{V}\left(\vec{\hat{\beta}}\right) = \mathbb{E}\left(\left(\vec{\hat{\beta}} - \mathbb{E}\left(\vec{\hat{\beta}}\right)\right)\left(\vec{\hat{\beta}} - \mathbb{E}\left(\vec{\hat{\beta}}\right)\right)^T\right)$.

Bei Unverzerrtheit resultiert in (9) ein Nullvektor, d.h. $\mathbb{E}\left(\vec{\hat{\beta}}\right) = \vec{\beta}$.

Für diesen erwartungstreuen Fall beträgt der mittlere quadratische Fehler:

$$\mathrm{MSE}\left(\vec{\hat{\beta}}, \vec{\beta}\right) = \mathbb{E}\left(\left(\vec{\hat{\beta}} - \vec{\beta}\right)\left(\vec{\hat{\beta}} - \vec{\beta}\right)^T\right). \tag{11}$$

Varianz und MSE stimmen also überein. Diese Formel ist verträglich zu (Rao, Toutenburg, Shalabh & Heumann, 2007, Formel (3.41)). Die MSE-Matrix in (11) entspricht wegen der Unverzerrtheit von $\vec{\hat{\beta}}$ gleichzeitig der Kovarianzmatrix in (10).

Falls $\mathbb{E}\left(\vec{\hat{\beta}}\right) \neq \vec{\beta}$, liegt Inkonsistenz vor und der MSE in (11) kann mit

$0 = -\mathbb{E}\left(\vec{\hat{\beta}}\right) + \mathbb{E}\left(\vec{\hat{\beta}}\right)$ erweitert werden ('Null-Trick'):

$$\begin{aligned}\mathrm{MSE}\left(\vec{\hat{\beta}}, \vec{\beta}\right) &= \mathbb{E}\left(\left[\left(\vec{\hat{\beta}} - \mathbb{E}\left(\vec{\hat{\beta}}\right)\right) + \left(\mathbb{E}\left(\vec{\hat{\beta}}\right) - \vec{\beta}\right)\right]\left[\left(\vec{\hat{\beta}} - \mathbb{E}\left(\vec{\hat{\beta}}\right)\right) + \left(\mathbb{E}\left(\vec{\hat{\beta}}\right) - \vec{\beta}\right)\right]^T\right) \\ &= \mathbb{E}\left(\left(\vec{\hat{\beta}} - \mathbb{E}\left(\vec{\hat{\beta}}\right)\right)\left(\vec{\hat{\beta}} - \mathbb{E}\left(\vec{\hat{\beta}}\right)\right)^T + 2\left(\vec{\hat{\beta}} - \mathbb{E}\left(\vec{\hat{\beta}}\right)\right)\left(\mathbb{E}\left(\vec{\hat{\beta}}\right) - \vec{\beta}\right)^T + \left(\mathbb{E}\left(\vec{\hat{\beta}}\right) - \vec{\beta}\right)\left(\mathbb{E}\left(\vec{\hat{\beta}}\right) - \vec{\beta}\right)^T\right).\end{aligned}$$

Beim Realisieren der Erwartungswerte für den ausmultiplizierten mittleren Term entsteht: $2\left(\vec{\hat{\beta}}\vec{\hat{\beta}}^T - \vec{\hat{\beta}}\vec{\beta}^T - \vec{\hat{\beta}}\vec{\hat{\beta}}^T + \vec{\hat{\beta}}\vec{\beta}^T\right) = 0_{p\times p}$. Im MSE verbleibt für den linken Term die Varianz und für den rechten Term der quadrierte Bias, so daß

$$\mathrm{MSE}\left(\vec{\hat{\beta}}, \vec{\beta}\right) = \mathbb{V}\left(\vec{\hat{\beta}}\right) + \mathrm{Bias}\left(\vec{\hat{\beta}}, \vec{\beta}\right)\mathrm{Bias}\left(\vec{\hat{\beta}}, \vec{\beta}\right)^T, \tag{12}$$

welches in Übereinstimmung zu (Rao et al., 2007, Formel (3.44)) ist.

Für die Interpretierbarkeit des MSE ist es vorteilhaft, eine skalare Größe zu beurteilen. Mit einer Matrixnorm, z.B. der Frobenius-Norm, wird dies derart bewerkstelligt, daß zumindest alle Elemente der MSE-Matrix in die Norm eingehen: Vom Kreuzprodukt der MSE-Matrix mit sich selbst wird die Wurzel der Spur berechnet. Anmerkung: die Hauptdiagonale ist beim Kreuzprodukt einer reellen Matrix stets positiv besetzt.

Die MSE-Matrix beinhaltet bereits die quadratische Form: Einen eleganteren Ansatz beschreiben (Rao et al., 2007) über ein Spurkriterium – *MDE I criterion* in Formel (3.46) – auf die Differenz. Direkt auf (11) angewandt, würde sich die Verzerrung ebenfalls in einer Null übersteigenden Spur offenbaren.

Sobald der Fall $n < p$ eintritt kann das Regressionsproblem mithilfe von (5) nicht mehr gelöst werden. Eine Singularität von X^TX tritt wegen ungünstiger Rechteckform von X ein. Bei funktionalen Daten ist dieses Problem defacto immer vorhanden.

Für eine Lösung genügt dann irgendein p-dimensionaler Koeffizientenvektor, der alle n Bedingungen erfüllt: Beliebig viele Möglichkeiten sind vorstellbar.

3.2 Verzerrte Schätzung mit dem Ridge-Schätzer

Immerhin könnte man sich bei Unterlaufen des Gauß-Markov Theorems (d.h. Verlust einer erwartungstreuen Schätzung) über Regularisierung behelfen. Dazu wird dem linearen Minimierungsproblem ein λ-Gewichtsterm als „Penalty“ beigfügt:

$$f\left(\vec{\beta}_R\right) \;=\; \vec{\epsilon}^{\,T}\vec{\epsilon} \;=\; \left(\vec{y}-X\vec{\beta}_R\right)^T\left(\vec{y}-X\vec{\beta}_R\right)+\lambda\vec{\beta}_R^{\,T}D\,\vec{\beta}_R \to \min_{\vec{\beta}_R}.$$

Für den Erhalt eines Extremwertes (ein Minimum) wird die Ableitung Null gesetzt:

$$\frac{\partial f}{\partial\vec{\beta}_R} = -2X^T(\vec{y}-X\vec{\beta}_R)+2\lambda D\beta_R \overset{!}{=} \vec{0}_p \;\;\Leftrightarrow\;\; -X^T\vec{y}+X^TX\vec{\beta}_R+\lambda D\vec{\beta}_R \;=\; \vec{0}_p.$$

Die Lösung für $\vec{\beta}_R$ entspricht einem Ridge-Schätzer:

$$\vec{\hat{\beta}}_R = \left(X^TX+\lambda D\right)^{-1}X^T\vec{y}\;,\;\lambda\geq 0\;,\;(p\times 1). \tag{13}$$

Üblicherweise setzt man für die Penalty-Matrix D die Einheitsmatrix ein. Setzt man $\lambda = 0$, so degeneriert $\vec{\hat{\beta}}_R$ zu $\vec{\hat{\beta}}_{KQ}$. Bei (Mittnik, 2015, Formel (2.2.3)) ist im Ridge-Schätzer $D = \operatorname{diag}(X^TX)$ definiert. Damit gewichtet er zusätzlich die Varianz der $\hat{\beta}_i$.

Die geschätzten Residuen:

$$\vec{\hat{\epsilon}} \;=\; \vec{y}-X\vec{\hat{\beta}}_R \;\overset{(13)}{=}\; \vec{y}-X\left(X^TX+\lambda D\right)^{-1}X^T\vec{y} \;=\; \left(I_n - X\left(X^TX+\lambda D\right)^{-1}X^T\right)\vec{y} \tag{14}$$

und die Varianz der Beobachtungen auf die geschätzte Regressionsfunktion:

$$\hat{\sigma}^2 = \frac{\vec{\hat{\epsilon}}^{\,T}\vec{\hat{\epsilon}}}{n-p} \overset{(14)}{=} \frac{\vec{y}^{\,T}\left(I_n - X\left(X^TX+\lambda D\right)^{-1}X^T\right)^2\vec{y}}{n-p}. \tag{15}$$

Mit λ verfügt man über einen Tuning-Parameter. Ist $\lambda > 0$, wird das Multikollinearitätsproblem in Matrix X^TX um den Preis der Erwartungstreue ausgehebelt. Daraus resultiert eine verzerrte Schätzung, bei gleichzeitigem Absinken der Varianz. Aber immerhin ist wenigstens eine Schätzung denkbar. Je positiver λ eingestellt wird, umso verzerrter erfolgt die Schätzung. Die Verzerrung ist verifizierbar. Sei nun

$$Z = \left(X^TX+\lambda D\right)^{-1}. \tag{16}$$

Der Erwartungswert beträgt: $\mathbb{E}\left(\vec{\hat{\beta}}_R\right) \overset{(13)}{=} \mathbb{E}\left(\left(X^TX+\lambda D\right)^{-1}X^T\vec{y}\right) \overset{(16)}{=} \mathbb{E}\left(ZX^T\vec{y}\right)$.

Mit der Substitution $\vec{y} = X\vec{\beta}+\vec{\epsilon}$ wird $\vec{y}$ zerlegt und es folgt:

$$\mathbb{E}\left(\vec{\hat{\beta}}_R\right) \overset{(2)}{=} \mathbb{E}\left(ZX^T\left(X\vec{\beta}+\vec{\epsilon}\right)\right).$$

Da $X\perp\vec{\epsilon}$, d.h. $X^T\vec{\epsilon}=\vec{0}_p$, vereinfacht sich der Erwartungswert: $\mathbb{E}\left(\vec{\hat{\beta}}_R\right) = ZX^TX\vec{\beta}$.

Die Verzerrung beträgt:

$$\operatorname{Bias}\left(\vec{\hat{\beta}}_R\right) \;=\; \mathbb{E}\left(\vec{\hat{\beta}}_R\right)-\vec{\beta} \;=\; \left(ZX^TX-I_p\right)\vec{\beta} \;\overset{(16)}{=}\; \left(\left(X^TX+\lambda D\right)^{-1}X^TX-I_p\right)\vec{\beta}. \tag{17}$$

Die Kovarianzmatrix von $\vec{\hat{\beta}}_R$ beträgt:

$$\operatorname{cov}\left(\vec{\hat{\beta}}_R\right) \overset{(13)}{=} \operatorname{cov}\left(\left(X^TX+\lambda D\right)^{-1}X^T\vec{y}\right) \overset{(16)}{=} \operatorname{cov}\left(ZX^T\vec{y}\right) \overset{(2)}{=} \operatorname{cov}\left(ZX^T\left(X\vec{\beta}+\vec{\epsilon}\right)\right)$$

$$\overset{X\vec{\beta}\perp\vec{\epsilon}}{=} \operatorname{cov}\left(ZX^TX\vec{\beta}\right)+\operatorname{cov}\left(ZX^T\vec{\epsilon}\right) \overset{\mathbb{V}(\vec{\beta})=\vec{0}}{=} ZX^T\operatorname{cov}\left(\vec{\epsilon}\right)XZ^T,$$

$$\operatorname{cov}\left(\vec{\hat{\beta}}_R\right) \overset{(3)}{=} \sigma^2ZX^TXZ^T \overset{(16)}{=} \sigma^2\left(X^TX+\lambda D\right)^{-1}X^TX\left(X^TX+\lambda D\right)^{-1}, \ (p\times p). \tag{18}$$

Im Fall von $\lambda = 0$ verschwindet die Verzerrung und Varianzformel (18) reduziert sich zu Formel (10) des KQ-Schätzers. Beliebiges Justieren von λ ist nicht empfehlenswert, denn der Ridge-Schätzer schrumpft mit zunehmendem λ die Koeffizienten Richtung Null; vgl. (Brockhaus, 2015, Seite 5). Er kann dennoch ein effizienter Schätzer sein: Bei guter Wahl von λ kann er den MSE des KQ-Schätzers unterschreiten. Die Auswirkungen sind anhand der Grafik in (Rao et al., 2007, Seite 80, Abb. 3.6) ablesbar.

$$\operatorname{MSE}\left(\vec{\hat{\beta}}_R\right) = \mathbb{V}\left(\vec{\hat{\beta}}_R\right) + \operatorname{Bias}\left(\vec{\hat{\beta}}_R\right)\operatorname{Bias}\left(\vec{\hat{\beta}}_R\right)^T \tag{19}$$

Dabei entsteht ein Optimierungsproblem bzgl. λ, welches den MSE minimieren soll. Der Bias in (17) beinhaltet das wahre unbekannte $\vec{\beta}$. Die erwartungstreue Schätzung für $\vec{\beta}$ ist (5) entnehmbar. Das bedeutet für den Bias, daß er als Faktor die inverse Matrix $\left(X^TX\right)^{-1}$ enthält. Diese Matrix läßt sich nicht herauskürzen und damit ist der Bias im Fall $n < p$ wegen deren Singularität nicht verfügbar. Somit ist die Lösung bzgl. eines optimalen λ nicht auffindbar.

Bzw., λ müßte gedanklich lediglich eine Nuance (kleine positive Gleitkommazahl $> \epsilon_m$, vgl. Kapitel 2.1) eingestellt werden, um eine reguläre Inverse zu erhalten. Damit wird de-facto eine Rekonstruktion von $\vec{y}$ herbeigeführt. Die geschätzte Varianz $\hat{\sigma}^2$ der Residuen ist für so einen speziellen „hauchdünnen“ Fall gerade noch herleitbar: Das Maschinenepsilon degeneriert $X\left(X^TX+\lambda D\right)^{-1}X^T$ in (15) nahezu zu einer n-dimensionalen Einheitsmatrix, so daß im Zähler die quadrierte Differenz von zwei Einheitsmatrizen auftritt. Die Division mit einem negativen Nenner $(n-p)$ ändert am Ergebnis nichts mehr, d.h. die Residuenvarianz $\hat{\sigma}^2$ beträgt Null.

Für $\lambda \gg \epsilon_m$ kann im Falle $n < p$ die Varianz nicht mehr identifiziert werden, d.h. Verlust des zweiten Momentes. Man verfügt dann über eine perfekt angepaßte und gleichzeitig nicht sinnvoll interpretierbare Lösung. Diese Lösung versteht sich als eine aus dem Kontinuum von beliebig vielen Lösungen.

Bei einem numerisch so hochproblematischen Fall kann eine andere Lösungsstrategie erfolgversprechend sein, bei der sogar der Tuningparameter λ in (13) obsolet wird: X^TX über das Konzept der verallgemeinerten Inverse (vgl. Kapitel 4.5) zu invertieren.

3.3 Regression auf Hauptkomponenten

Es existiert eine Alternative in einem strikt zweistufigen Ansatz. Im ersten Schritt wird ein faktorenanalytisches Verfahren auf die Kovariablen in X ausgeführt. Dabei wird die Zerlegung gemäß (31) auf Seite 29 durchgeführt: $X = \vec{1}\,\vec{\bar{x}}^T + FA^T + E$.

Bei der Hauptkomponentenmethode (PCA, vgl. Kapitel 4.4.1) werden die Faktoren F auch Hauptkomponenten genannt. Die Zerlegung von X entschärft mithilfe der Faktoren F $(n \times g)$ das Rangproblem bzgl. X. Für die Anzahl g der Faktoren F ist möglichst $g \ll p$ anzustreben.

Im zweiten Schritt wird eine Regression auf die g extrahierten Faktoren durchgeführt. Hierbei regressiert man Linearkombinationen. Ein weiterer Vorteil dieser Methode liegt im orthogonalen Designraum, den diese Faktoren aufspannen. Die Multikollinearität, welche in starker Ausprägung (Verwenden der rohen Einflußgrößen) die Schätzung besonders verzerren kann, bisweilen fragwürdige Koeffizienten produziert, ist vollständig ausgeschaltet. Im zweiten Schritt wird die Zielgröße $\vec{y}$ zerlegt: $\vec{y} = F\vec{\beta} + \vec{\epsilon}$.

Das Rangproblem wird zwar entschärft aber die Interpretierbarkeit insgesamt etwas erschwert. Vor allem der Rückweg von den Faktoren zu den Originalwerten X dürfte im Regressionskontext gedanklich kaum gelingen. Um die systematische Verzerrung der Regression klein zu halten, darf die Faktorenzahl g nicht beliebig klein gewählt werden.
Wünschenswert wäre ein Ansatz, der „in Anlehnung“ zur Regression – möglichst unverzerrt – agiert, und vor allem die Ergebnisse quasi in einem Schritt herbeiführt. Das soll mit einer partiellen kleinste Quadrate Regression (PLS) später in Kapitel 5 diskutiert werden.

3.4 Glättung

Funktionale Daten sind, hier in ihrer Realisation, ebenfalls mit stochastischen Meßfehlern behaftet: $f\left(\vec{t}\right) = x\left(\vec{t}\right) + \epsilon\left(\vec{t}\right)$. Wie bereits weiter oben erwähnt, stehen die Realisationen f_i der Spalten t_i (entsprechen den Knoten) in einem Zusammenhang, auch über h nächste Nachbarn/Knoten hinweg. Es ist mit den Methoden der Regression möglich, den Meßfehler zu begradigen, d.h. zu glätten, indem die Datenmatrix ($n < p$) transponiert und darauf ein multivariates Regressionsproblem auf n Zielgrößen gelöst wird.
Datentreue und Glattheit stehen im Widerstreit, ein Zielkonflikt besteht.

Als etabliertes Verfahren sei die B-Spline Modellierung genannt, bei der die Wahl des Spline-Polynomgrades, auch die Zahl der einfließenden Knoten bestimmt. B-Splines umschließen stets einen Flächeninhalt von Eins. Üblich ist ein kubischer Spline, d.h. vom Polynomgrad 3. Dieser Spline-Typ besteht aus vier zusammengesetzten kubischen Parabelstücken und deckt damit fünf Knoten ab, an denen er zweimal stetig differenzierbar ist: Die Stoßenden der Kurvenstücke sind *glatt* miteinander verbunden. Wegen der Glattheitseigenschaft – bei gleichzeitig geringer Komplexität – kommt dem kubischen Spline eine besondere Bedeutung zu.
Für den Aufbau einer Designmatrix Z gleitet das Polynom über alle Knoten hinweg. An der Spaltensumme von Z läßt sich der Flächeninhalt der Splines ablesen. Für jeden Eintrag beträgt er gleich Eins. Die Ausrichtung der Spline-Funktionen auf die Zielgröße (d.h. Anpassen des Flächeninhaltes) wird über die zugehörigen Koeffizienten γ_i vorgenommen: $\vec{y} = Z\vec{\gamma} + \vec{\epsilon}$. Bzgl. $\vec{\epsilon}$ gelten die Annahmen wie in (3) formuliert.
Eine Glättung kann mit penalisierter Regression vollzogen werden, bei der die γ-Koeffizienten über einen Penalisierungsparameter λ gewichtet werden – vgl. (Heumann & Schmid, 2016, Kap. 7.4.2, Seite 99). Der Schätzer bzgl. $\vec{\gamma}$ weist hierbei diesselbe Struktur wie der Ridge-Schätzer (13) auf.

In der analytischen Chemie ist eine Glättungs-Methode von (Savitzky & Golay, 1964) populär. Offenbar handelt es sich um einen Standard-Ansatz, denn die Autoren werden auch heuzutage oft erwähnt. Diese Glättung läßt sich nicht so komfortabel wie Splines, d.h. nicht rein vektoriell implementieren, aber im Unterschied zur Spline-Modellierung kann ein Polynom beliebig viele Knoten abdecken. Mit dem Paramter h wird die Zahl der Knoten ausgewählt, welches mit einer Penalisierung assoziierbar ist. Zu den Randknoten hin muß unweigerlich der Grad des Polynoms vermindert werden, um die Identifizierbarkeit abzusichern. Wie bei der Glättung i.A. üblich, d.h. auch bei B-Spline-Glättung, werden die Ränder dann mit einem linearen Term abgebildet.

Schematische Beschreibung der Glättung[14]:
– transponiere Zeilenvektor $\vec{y}$ zu Spaltenvektor
– wähle Polynomgrad Pm und Anzahl h der Nachbarknoten
– berechne Zeilenzahl $p = 2h + 1$ *%p beschreibt im Eigentlichen Spalten*
– $X_g = \vec{1}$ *%Startwert Designmatrix, g bedeutet global*
– iteriere von $k = 1 \ldots Pm$ und erzeuge $X_g \stackrel{!}{=} \left(X_g\,,\left((1:p)^T\right)^k\right)$ *%Designmatrix-Vorlage*
– erzeuge leeres Objekt y_{glatt} *%daraus entsteht durch Iteration $\vec{y}_{glatt}$*
– iteriere von $k = 1 \ldots p$
- $a = (k + h + 1 - |h + 1 - k|)\,/2 - 1$ *%a schrittweise von 0 auf h erhöhen*
- $b = (p - h - k - |p - h - k|)\,/2 + h$ *%b von h auf 0 mindern, so daß für $k \stackrel{!}{=} p$: b=0*
- $q = a + b + 1$ *%q: Zeilenzahl von Designmatrix X*
- $X = X_g\,[1:q\,,1:min(q-1\,,Pm+1)]$ *%X: adaptive Designmatrix X*
- $\vec{\hat{\beta}} = \left(X^T X\right)^{-1} X^T \vec{y}\,[k-a:k+b]$ *%[k − a : k + b]: gleitendes Fenster durch $\vec{y}$*
- $\mathrm{tmp} = X\vec{\hat{\beta}}$ *%temporäre Prognose*
- wenn $b - a \leq 0\,,\ \mathrm{tmp} \stackrel{!}{=} \mathrm{tmp}[a+1]\,,$ sonst $\mathrm{tmp} \stackrel{!}{=} \mathrm{tmp}[\mathrm{end} - b]$
  %selektiert Punktschätzer für k-ten Knoten; end ≙ letzten Eintrag im Vektor $X\vec{\hat{\beta}}$
- $\vec{y}_{glatt} \stackrel{!}{=} (\vec{y}_{glatt}\,;\mathrm{tmp})$ *%füge neue Prognose als nächsten Eintrag an*

Am Ende der Prozedur stimmt die Länge von $\vec{y}$ und seiner Glättung $\vec{y}_{glatt}$ überein.
Die hier vorgeschlagene Rechenvorschrift ist abweichend von der Originalschrift und kommt deshalb ohne voreinzustellende Koeffizienten aus. Das macht sie bequemer in der Handhabung. Der ursprüngliche Algorithmus beherrscht allerdings auch das (nicht stetige) Ableiten der Reihe.
In der folgenden Abbildung soll die Wirkungsweise der Glättung skizziert werden.

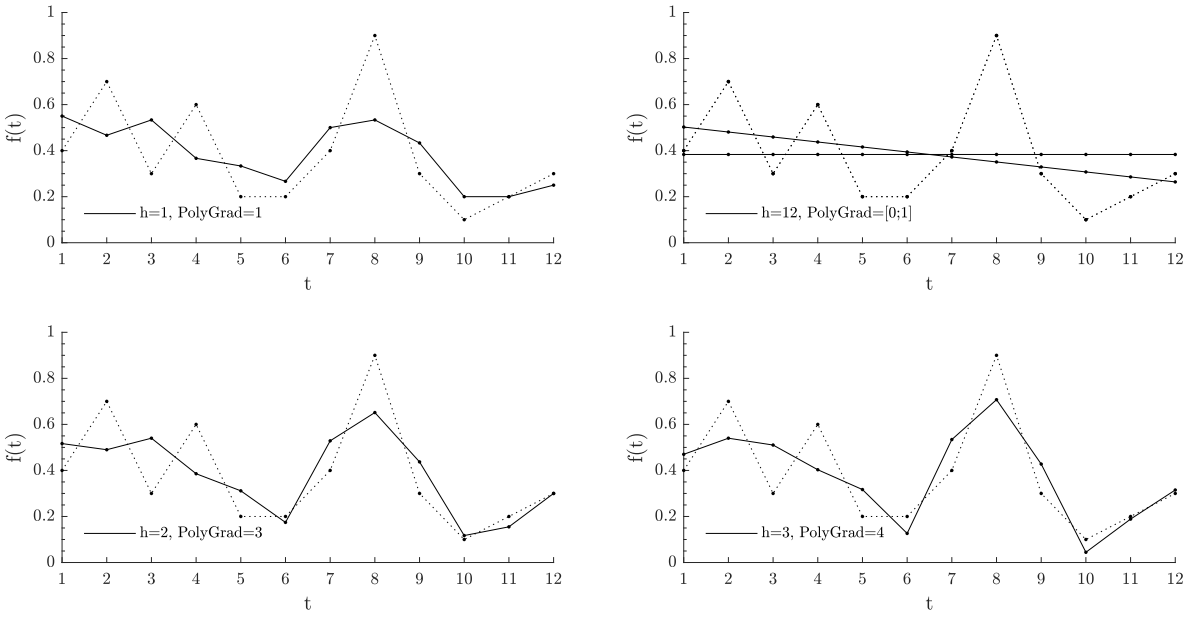


Abb. 8: Auswirkungen der Parameter-Variationen auf dieselbe Datenreihe (……)

[14]Die Details für einen numerischen Lösungsansatz lassen sich im Skript glaetter.m nachvollziehen.

Die linke obere Grafik beschreibt in ihrer Parametrisierung einen gleitenden Mittelwert[15]. Hingegen zeigt die rechte obere Grafik bei Hinzunahme sämtlicher Knoten das Verhalten eines linearen Modells. Das Intercept-Modell wurde ebenfalls mit einbeschrieben. Die beiden unteren Grafiken zeigen einen typischen Fall: Glätten von schnellen Wechseln im Anstieg, aber bewahren von Mustern. Mit dem Polynomgrad 4 beginnt ein leichtes Überschwingen (beachte Knoten 6 und 10), welches mit einem breiteren Fenster gemildert werden soll. Hier im Beispiel gelingt der Kompromiß zufriedenstellend. Die Variationen sind mit Bedacht anzuwenden, um Daten nicht zu stark zu verändern und möglicherweise unbrauchbar zu machen. Denn diese Sichtweise ließe sich auch einnehmen: Wiederherstellung des Originals aufgrund der Glättung der verrauschten Kurve.

Evtl. ungewohnt erscheint für eine Glättung der Begriff *Filter*[16]. Dabei handelt es sich um einen Ausdruck aus der Signaltheorie. In Anlehnung soll in der nächsten Abbildung ein Problem überzeichnet skizziert werden: Auf einen reinen Sinuston wird ein amplitudenstarker höherfrequenter Sinuston aufmoduliert. Beispielsweise bei einem Hörtest würde sich diese Tonfolge als unbrauchbar erweisen, sie wäre auch konträr im Sinneseindruck.

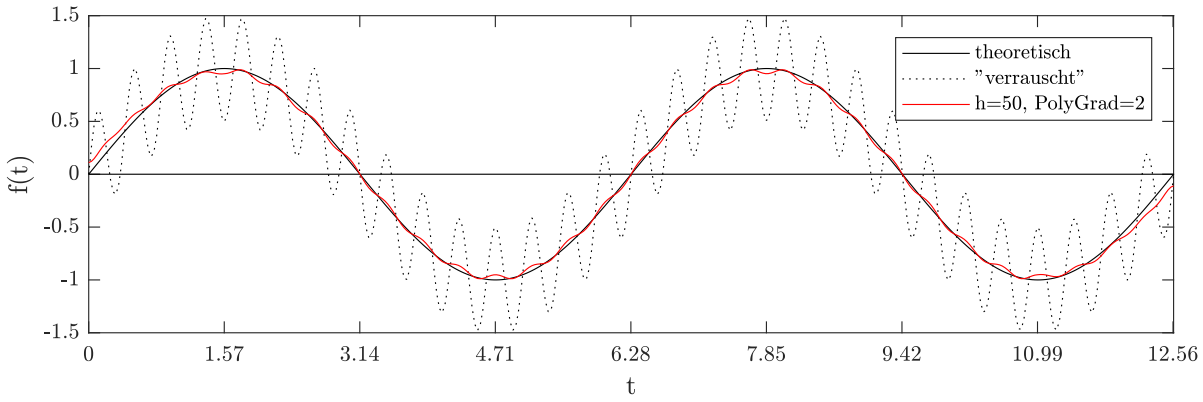


Abb. 9: Tiefpaß gefiltertes Signal auf 400 Knotenpunkten

Das Originalsignal kann mit der Glättung von (Savitzky & Golay, 1964) weitgehend zurück gewonnen werden. Die verbliebene Verunreinigung (Artefakt) sollte die Grundwelle kaum hörbar beeinträchtigen. Tiefpaßfilterung bedeutet, daß ab einer Grenzfrequenz höhere Frequenzanteile zunehmend stark gedämpft werden. Bei günstiger Wahl der Zahl h an Nachbarknoten und des Polynomgrades lassen sich Rauheiten (entsprechen höheren Frequenzen) regelrecht „wegglätten“. Filtern ist also ein Synonym für extremes Glätten. Dafür existieren spezielle Methoden, wie z.B. die Fouriertransformation (Schmid, 2016a), welche das Original-Signal noch besser restaurieren kann.

[15] https://de.wikipedia.org/wiki/Gleitender_Mittelwert (10.3.2018)
[16] http://www.statistics4u.info/fundstat_germ/cc_filter_savgolay.html (10.3.2018)

4 Matrix – Zerlegungstechniken I

Die Beschreibungen beziehen sich stets auf reellwertige Matrizen.
Die Faszination an nicht verlustbehafteten Zerlegungen besteht weiterhin an einer möglicherweise spürbaren Datenreduktion. Daten lassen sich damit sparsamer vorhalten. Dabei wird die Dimension reduziert, bei gleichzeitiger Erhaltung der Information in den Matrizen. Um den Preis von CPU-Zeit Verbrauch müssen allerdings bei einem Zugriff die ursprünglichen Daten durch Komposition erst wieder rekonstruiert werden. Diese Aspekte sollen nicht weiter vertieft werden. Die in dieser Arbeit besprochenen Datensätze sind für solche Fragestellungen aus heutiger Sicht (big data Sichtweise) bereits als sehr klein anzusehen.
Für die Zerlegungen erweist sich eine Zentrierungsmatrix als sehr hilfreich:

$$K = I_n - \frac{1}{n}\vec{1}\,\vec{1}^T,\ (n \times n). \tag{20}$$

K ist vollbesetzt aber idempotent und hat deshalb keinen Vollrang.
Mithilfe der Spur von K ist der Rang bestimmbar. Der Rangabfall von K beträgt Eins.

4.1 Projektion

Eine Projektion meint hier das Abbilden eines räumlichen Körpers. Die durch das „Hinwerfen“ entstehende Figur, welche man als Schattenbild interpretieren kann, entspricht der Projektion. Das Ziel besteht i.d.R. vorrangig darin, mit dem Schattenbild einen Unterraum geringerer Dimension aufzuspannen. Das sei an folgendem Foto gezeigt:

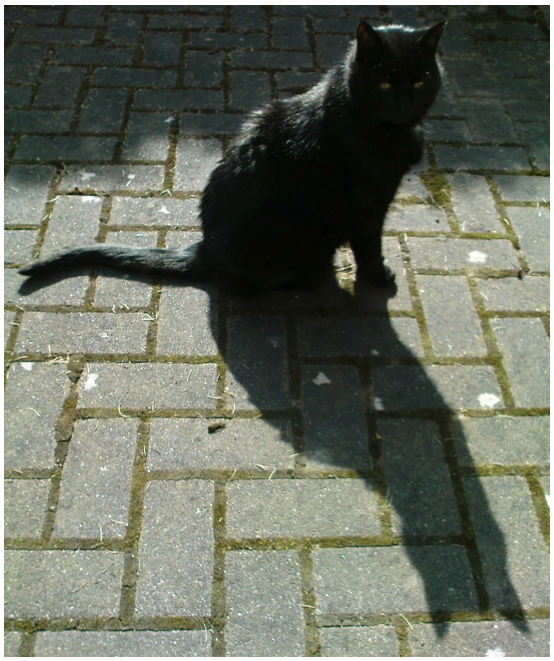


Abb. 10: linear verzerrte Projektion vom $\mathbb{R}^3$ in den $\mathbb{R}^2$

Der Körper der Katze, welcher die sichtbaren Sonnenstrahlen absorbiert und von ihnen gewärmt wird, erscheint auf der Ebene als Umriß / Schatten.

Bei dieser Projektion verschwindet die Rauminformation der Tiefe. Da die Ebene, auf welcher der Schatten entsteht, nicht orthogonal zu den Sonnenstrahlen ausgerichtet ist, entsteht bei dieser Projektion ein verzerrter (gestreckter) Schatten. Wäre der Winkel zur Sonne bekannt, könnte die Projektion nachträglich entzerrt werden.

Das Konzept der Projektion kann mittels Ortsvektoren beschrieben werden:
Vektor $\vec{b}$ soll auf Vektor $\vec{a}$ projiziert werden. Der entstehende Vektor $\vec{c}$ stimmt dann bzgl. Lage und Richtung mit $\vec{a}$ überein. Die resultierende Länge von $\vec{c}$ kann man sich über eine Normale, welche orthogonal durch $\vec{a}$ hin zur Vektorspitze $\vec{b}$ verläuft, vorstellen.
Die Vektoren-Projektion als Formel (Merziger & Wirth, 2006, Seite 131):

$$\vec{c} = \frac{\vec{a}^T \vec{b}}{\vec{a}^T \vec{a}} \vec{a}. \tag{21}$$

In der Formel ist der Quotient der Skalarprodukte ein Linearfaktor: $\vec{c}$ ist eine Linearkombination von $\vec{a}$.
Die Anwendung von Formel (21) wird mithilfe der nächsten Abbildung illustriert:

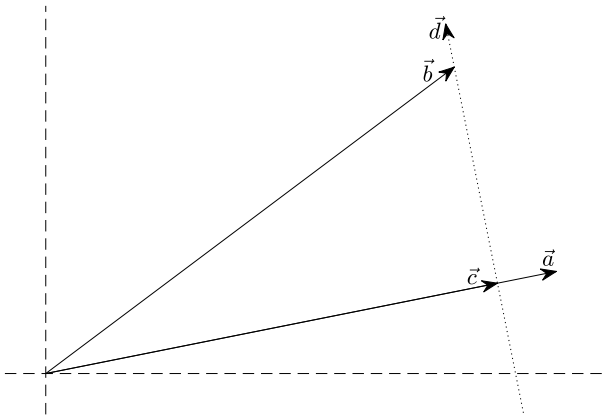


Abb. 11: Projektion von $\vec{b}$ auf $\vec{a}$ in der Ebene: $\vec{c}$ entsteht.

Das ebene Koordinatensystem ist gestrichelt angedeutet. Der punktiert gezeichnete Vektor $\vec{d}$, welcher die Vektorspitze $\vec{b}$ berührt, entspricht einer Normale auf $\vec{a}$.
Am Schnittpunkt $\left(\vec{a}, \vec{d}\right)$ befindet sich die Spitze von $\vec{c}$. Demzufolge entstand $\vec{c}$ aufgrund einer orthogonalen Projektion. Diese Art der Projektion wird in den Kapiteln zur Matrizen-Zerlegung die Grundlage darstellen.
Der Ortsvektor, welcher aus dem Schnittpunkt von Normale $\vec{d}$ mit der Abszisse resultieren würde, wäre vergleichbar mit der Projektion, wie in Abb. 10 gezeigt.

4.2 Spektralzerlegung

Der Begriff könnte aus der Optik[17] stammen: Dort meint eine Spektralzerlegung die Zerlegung des Lichts mithilfe eines Prismas in seine Bestandteile, die Spektralfarben.

Vorausgesetzt wird zwingend eine quadratische Matrix. Dafür ist eine Kovarianzmatrix geeignet, welche stets symmetrisch aufgebaut ist. Sie soll in ihre Bestandteile, die Eigenwerte und eine orthogonale Matrix der Eigenvektoren, zerlegt werden. Wobei die Eigenwerte das Spektrum beschreiben. Eigenvektoren spannen in ihrer Eigenschaft als Basisvektoren einen kartesischen Raum auf. Manchmal wird die Zerlegung auch *Eigenwertzerlegung* genannt. Die gefundenen Eigenwerte werden zu einer Diagonalmatrix zusammengestellt.
Die Symmetrie der Kovarianzmatrix garantiert strikt reellwertige Eigenwerte und Eigenvektoren. Komplexwertige Lösungen kommen hier also nicht vor. Desweiteren ist aufgrund der Dominanz der Varianz gegenüber der Kovarianz die Hauptdiagonale der Kovarianzmatrix mindestens so stark wie die betragsmäßigen Nebendiagonalen besetzt. D.h., die Kovarianzmatrix ist positiv. Präziser: eine Kovarianzmatrix ist mindestens positiv semidefinit. Dies ist an der normierten Kovarianzmatrix, der Korrelationsmatrix R ersichtlich:

$$R \;=\; \operatorname{cor}(X) \;=\; \operatorname{cov}\left(X\mathbb{V}(X)^{-1/2}\right) \;=\; \mathbb{V}(X)^{-1/2}\operatorname{cov}(X)\,\mathbb{V}(X)^{-1/2}. \tag{22}$$

Die Hauptdiagonale von R ist nur mit Einsen besetzt. Aufgrund der Definitheit der Kovarianzmatrix C folgen keine negativen Lösungen der Eigenwerte.

$$C = \widehat{\operatorname{cov}}(X) \;=\; \frac{(KX)^T KX}{n-1} \;=\; \frac{X^T KX}{n-1} \;=\; P\Lambda P^T,\ (p\times p) \tag{23}$$

Schließlich ist $P\Lambda P^T$ die Spektralzerlegung der Kovarianz. Wenn m den Rang von C beschreibt, dann gilt: $m \leq p$. P ist dementsprechend eine $p\times m$ Matrix der m Eigenvektoren und Λ damit die $m \times m$ Diagonalmatrix der Eigenwerte. Λ läßt sich aufspalten:

$$\Lambda \;=\; \Lambda^{1/2} I_m \Lambda^{1/2} \;=\; \frac{\Lambda^{1/2} F^T F \Lambda^{1/2}}{n-1} \;=\; \frac{T^T T}{n-1} \;=\; \widehat{\operatorname{cov}}(T). \tag{24}$$

Auf die orthonormale Matrix F wird in Kapitel 4.4 noch näher eingegangen. T beschreibt die Matrix der Hauptachsen. Die Aufgabe besteht nun darin, die Lösungen für P und Λ zu erhalten.
Die p Spalten der zentrierten Matrix KX fließen ein in die m-fache Maximierung der Varianz des Ausgangsraumes. Der unbekannte Basisvektor $\vec{p}_i$ $(p\times 1)$ enthält die Gewichte der Spalten von X. Als Basisvektor erfüllt er die Eigenschaft $\vec{p}_i^{\,T}\vec{p}_i = 1$ (Normiertheit).
Es resultiert als Optimierungsproblem ein Maximierungsproblem.

Die erste Richtung entlang der maximalen Varianz $\vec{t}_1 = KX\vec{p}_1$ $(n\times 1)$ ergibt

$$\mathbb{V}\left(\vec{t}_1\right) = \frac{(KX\vec{p}_1)^T KX\vec{p}_1}{n-1} = \frac{\vec{p}_1^{\,T}X^T KX\vec{p}_1}{n-1} \overset{(23)}{=} \vec{p}_1^{\,T}\operatorname{cov}(X)\,\vec{p}_1 \;=\; \vec{p}_1^{\,T}C\vec{p}_1 \to \max_{\vec{p}_1}. \tag{25}$$

C versteht sich als feste Matrix, denn Matrix X sei deterministisch angenommen.
Zum Lösen von (25) ist eine Nebenbedingung erforderlich, die Normierung von $\vec{p}_1$.

[17] http://mathematikalpha.de/regenbogen (10.6.2017)

Ost, einer der Autoren aus (Fahrmeir & Hamerle, 1984), beschreibt auf die Hauptachsentransformation eine Methodik zum Ausrichten des neuen Koordinatensystems mithilfe von Langrange-Multiplikatoren. Diese Methode wird hier leicht modifiziert beschrieben.

Gemäß (23) ist festgelegt: $C = P\Lambda P^T$. Wenn P die Matrix der Eigenvektoren sein soll, dann folgt, daß in P die Spalten als Basisvektoren definiert sind:
P ist eine orthogonale Matrix. Demzufolge entsteht aus dem Kreuzprodukt P^TP eine $m \times m$ Einheitsmatrix, und damit gilt: $P^TCP = \Lambda$ $(m \times m)$.
Als äquivalent versteht sich deshalb auch diese Darstellung: $P^TCP = \Lambda P^TP$.
D.h., über die Differenz $P^TCP - \Lambda P^TP$ soll in Abhängigkeit von P und Λ eine Nullmatrix resultieren. Schreibt man das Problem eindimensional, so erhält man gestaffelte Lagrange-Funktionen, die man Lagrange-Multiplikatorenansatz nennt, für die erste Dimension:

$$\varphi\left(\vec{p}_1; \lambda\right) = \vec{p}_1^{\,T} C \vec{p}_1 - \lambda \vec{p}_1^{\,T} \vec{p}_1. \tag{26}$$

Hierin ist λ der Lagrange-Multiplikator. Viele Lösungen für $\vec{p}_1$ und λ würden die Gleichung simultan erfüllen. Es soll aber die Kombination für beide Unbekannte gefunden werden, die λ maximiert. Das führt auf ein Extremwertproblem. Dazu wird (26) partiell nach dem Basisvektor $\vec{p}_1$ und der Nebenbedingung λ differenziert:

$$\frac{\partial \varphi}{\partial \vec{p}_1} = 2C\vec{p}_1 - 2\lambda\vec{p}_1 \stackrel{!}{=} \vec{0} \quad \Leftrightarrow \quad \left(C - \lambda I_p\right)\vec{p}_1 = \vec{0} \tag{26.1}$$

$$\frac{\partial \varphi}{\partial \lambda} = -\vec{p}_1^{\,T}\vec{p}_1 = -1 \quad \Leftrightarrow \quad \vec{p}_1^{\,T}\vec{p}_1 \stackrel{!}{=} 1. \tag{26.2}$$

Der erste Eigenwert λ ist die Lösung des Maximierungsproblems. In (26.1) ist die Lösung faktorisiert dargestellt. Eigenwert und Eigenvektor können dadurch unabhängig voneinander berechnet werden, weil die Determinante $|(C - \lambda I_p)|$ für einen eingesetzten Eigenwert λ Null beträgt. Mit der in (26.1) eingesetzten Lösung für λ resultiert ein homogenes Gleichungssystem. Dessen Lösung führt zum zugehörigen Eigenvektor $\vec{p}_1$, welcher in Richtung der maximalen Varianz ausgerichtet ist.
Addiert man nun, bevor in (26.1) faktorisiert wird, den Term $\lambda\vec{p}_1$ auf die rechte Seite, so erhält man $C\vec{p}_1 = \lambda\vec{p}_1$. Mit dem linksseitigen Multiplizieren von $\vec{p}_1^{\,T}$ wird eine quadratische Form erreicht. Damit entsteht unter Beachtung von (25):

$$\mathbb{V}\left(\vec{t}_1\right) \;=\; \vec{p}_1^{\,T} C \vec{p}_1 \;=\; \vec{p}_1^{\,T}\lambda\vec{p}_1 \;=\; \lambda\vec{p}_1^{\,T}\vec{p}_1 \stackrel{(26.2)}{=} \lambda. \tag{26.3}$$

D.h., der erste Eigenwert λ entspricht der Varianz bzgl. der ersten Hauptachse $\vec{t}_1$.

Für die zweite Hauptachse entlang der verbleibenden Varianz: $\vec{t}_2 = KX\vec{p}_2$ $(n \times 1)$ wird der Index in (25) um Eins erhöht: $\mathbb{V}\left(\vec{t}_2\right) = \vec{p}_2^{\,T} C \vec{p}_2 \rightarrow \max_{\vec{p}_2}$.

Gleichzeitig entsteht eine weitere Nebenbedingung: $\vec{p}_1 \perp \vec{p}_2 \;\Leftrightarrow\; \vec{p}_1^{\,T}\vec{p}_2 = 0$.
Ein zweiter Lagrange-Multiplikator γ muß eingesetzt werden:

$$\varphi\left(\vec{p}_2; \lambda, \gamma\right) \;=\; \vec{p}_2^{\,T} C \vec{p}_2 - \lambda\vec{p}_2^{\,T}\vec{p}_2 - \gamma\vec{p}_1^{\,T}\vec{p}_2 \tag{27}$$

$$\frac{\partial \varphi}{\partial \vec{p}_1} = -\gamma \vec{p}_2 \stackrel{!}{=} \vec{0} \quad \Leftrightarrow \quad \gamma = 0 \tag{27.1}$$

$$\frac{\partial \varphi}{\partial \vec{p}_2} = 2C\vec{p}_2 - 2\lambda \vec{p}_2 - \gamma \vec{p}_1 \stackrel{!}{=} \vec{0} \quad \Leftrightarrow \quad (C - \lambda I_p)\, \vec{p}_2 = \vec{0} \tag{27.2}$$

$$\frac{\partial \varphi}{\partial \lambda} = -\vec{p}_2^{\,T} \vec{p}_2 = -1 \quad \Leftrightarrow \quad \vec{p}_2^{\,T} \vec{p}_2 \stackrel{!}{=} 1. \tag{27.3}$$

Die Lösung für den zweiten Eigenwert λ wird äquivalent zur Methodik des ersten Eigenwertes erhalten. Dann zeigt der Eigenvektor $\vec{p}_2$ in Richtung maximaler Varianz, mit der Restriktion orthogonal zu $\vec{p}_1$ angeordnet zu sein. Wird in (26.3) der Index um Eins erhöht, entspricht der zweite Eigenwert λ der Varianz für Hauptachse $\vec{t}_2$.

Mit jeder weiteren Lösung i steigt die Anzahl der Ortogonalitäts-Restriktionen in der Nebenbedingung auf $\binom{i}{2}$ Kombinationen an. Insgesamt entsteht ein orthogonaler Raum, bei dem laut Voraussetzung ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$) die erste Dimension den stärksten Informationsgehalt beinhaltet.

Aus den Überlegungen lassen sich Matrizen formulieren: $P = (\vec{p}_1, \ldots, \vec{p}_m)$ bzw.
$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$. Beide Matrizen finden sich in (23) wieder. P, die Matrix der Eigenvektoren, spannt einen kartesischen m-dimensionalen Raum auf, in welchem der p-dimensionale Ursprungsraum orthogonal hineinprojiziert wird. Mithilfe der Matrix P verfügt man dann über eine Gewichtsmatrix der ursprünglichen Beobachtungen.
Die Hauptachsen-Matrix $T = \left(\vec{t}_1, \ldots, \vec{t}_m\right)$ $(n \times m)$ enthält die projizierten Beobachtungen. Demzufolge handelt es sich bei T um Linearkombinationen der zentrierten Matrix X.

Der Lagrange-Ansatz zeigt auf eine algebraische Lösung hin, das charakteristische Polynom beim Berechnen von λ, welches für kleine Dimensionen handhabbar ist. Es fehlt noch die Idee, wie man eine Lösung für ein hochdimensionales Problem erhält.
(Collins, 2010) beschreibt schematisiert auf Seite 12 einen iterativen Weg zum Finden der Lösung. Der Algorithmus[18] wurde für diese Arbeit übernommen und erweitert. Es handelt sich dabei um eine modifizierte Variante des NIPALS-Algorithmus (Nonlinear Iterative Partial Least Squares[19]).
Ein entscheidender Nachteil bei der Kovarianz basierten Herangehensweise liegt, insbesondere bei funktionalen Daten mit ihren Hunderten bzw. Tausenden von Merkmalen, in einer sehr großen Kovarianzmatrix begründet: Bei einer 2000 Spalten umfassenden Ausgangsmatrix erhält man eine Kovarianzmatrix mit vier Millionen Einträgen. Der Aufbau einer derartigen Kreuzproduktmatrix beansprucht auch heutzutage noch fühlbare CPU-Zeit.
Besonders bei einer funktionalen Datenstruktur, bei der i.d.R. $n \ll p$ vorliegt, folgt für die Kovarianzmatrix ein Rangverlust von mindestens $p - n$. Die Spektralzerlegung ist dann bzgl. der Verarbeitungsgeschwindigkeit nicht die beste Wahl: Zum einen wird viel Zeit für den Aufbau der Kreuzproduktmatrix benötigt, andererseits ist diese Matrix dann improper, also wenig informativ. Das Umgehen des Problems soll im nächsten Kapitel beschrieben werden.

[18] Die Details für einen numerischen Lösungsansatz lassen sich im Skript plp.m nachvollziehen.
[19] http://www.statistics4u.info/fundstat_germ/dd_nipals_algo.html (10.5.2017)

4.3 Singulärwertzerlegung (SVD)

Bzgl. der Gestalt einer zu zerlegenden Matrix gibt es keine Einschränkungen. Damit kann der Umweg der Daten über (23) vermieden werden.

$$KX = UDV^T,\ (n \times p) \tag{28}$$

Dabei ist die Dekomposition UDV^T die Singulärwertzerlegung der zentrierten Matrix X. Ost (Fahrmeir & Hamerle, 1984) bezeichnet die dort mit (3.7) markierte äquivalente Zerlegung auch als *Grundstruktur*. Wenn m den Rang vom Matrizenprodukt KX beschreibt, folgt daraus: $m \leq \min\{n, p\}$. Falls $n < p$ vorliegt, gewinnt die SVD gegenüber der Spektralzerlegung zunehmend durch weitere Einsparung ($m \leq n - 1$) an Rechenzeit.
U ist die orthogonale $n \times m$ Matrix der Links-Singulärvektoren, D die $m \times m$ Diagonalmatrix der Singulärwerte und V die orthogonale $p \times m$ Matrix der Rechts-Singulärvektoren. Die projizierten Beobachtungen:

$$T = UD\ ,\ (n \times m). \tag{29}$$

Schematische Beschreibung der Dekomposition[20] mithilfe des NIPALS-Algorithmus:
– erzeuge temporäre zentrierte Matrix $E = KX$; erzeuge leere Matrizen U, D, V
– berechne $m = \text{Rang}\,(E)$
– iteriere m fach

- untersuche spaltenweise Matrix E mithilfe der l_∞-Vektornorm und markiere die Spalte mit dem betragsmäßig maximalen Wert
- normiere die markierte Spalte und bezeichne sie mit $\vec{u}$
- wiederhole folgende Berechnungen bis zur Konvergenz (d.h. Abweichungen in $\vec{u}$ und in $\vec{v}$ sinken unterhalb einer Fehlerschranke)
  – $\vec{v} = E^T\vec{u}$ *%Rechts-Singulärvektor*
  – normiere $\vec{v}$ *%Einheitsvektor*
  – $\vec{u} = E\vec{v}$ *%Links-Singulärvektor*
  – normiere $\vec{u}$
- $\lambda = \vec{u}^T E\vec{v}$ *%Singulärwert d*
- reihe $\vec{u}$ als neue Spalte in Matrix U und $\vec{v}$ ebenso in Matrix V ein
- reihe λ in Matrix D als neues erweitertes Diagonalelement ein
- $E \stackrel{!}{=} E - \vec{u}\lambda\vec{v}^T$ *%reduzierte Matrix: Rest*

Wird die Dekomposition nicht vorher abgebrochen, entspricht die Fehlermatrix E am Prozedurende bis auf numerische Unzulänglichkeiten einer Nullmatrix.

Gemeinsamkeiten zur Spektralzerlegung:
Die SVD kann als Kern der Spektralzerlegung betrachtet werden, wenn es gelingt ihre Lösungen in die einer Spektralzerlegung zu überführen.

$$C = \widehat{\text{cov}}\,(X) = \frac{X^TKX}{n-1} \stackrel{(28)}{=} \frac{VDU^TUDV^T}{n-1} = \frac{VD^2V^T}{n-1} = V\Lambda V^T = P\Lambda P^T \tag{30}$$

[20] Die Details für einen numerischen Lösungsansatz lassen sich im Skript udv.m nachvollziehen.

Wegen der Orthogonalität von U entsteht aus U^TU eine Einheitsmatrix. In der Tat ist Matrix V, bis auf evtl. Vorzeichenumpolungen einzelner Spalten, äquivalent zu Matrix P. Das resultiert aus der Nichteindeutigkeit eines Eigenvektors bzgl. seines Gegenvektors und ist üblicherweise unkritisch. Die Singulärwerte aus Matrix D lassen sich durch Quadrierung und Division mit $(n-1)$ direkt in die Eigenwerte von Λ umrechnen.

4.4 Faktorenanalyse (FA)

Historie, Entwicklung

Die Faktorenanalyse ist bereits seit 114 Jahren eingeführt. Damals ging Spearman von einem Intelligenz-Faktor, dem Generalfaktor[21] aus. Seitdem wurde die Methodik ständig weiterentwickelt und verfeinert.
In (Fahrmeir & Hamerle, 1984, Seite 575) schreibt Ost im zweiten Absatz

> Unter Faktorenanalyse versteht man nicht ein bestimmtes statistisches Verfahren, sondern es handelt sich bei dieser Bezeichnung um einen Sammelbegriff für viele, zum Teil sehr unterschiedliche Techniken. Das Ziel einer Faktorenanalyse ist stets die Zurückführung einer größeren Menge beobachtbarer Variablen auf möglichst wenige hypothetische Variablen, die *Faktoren*.

.
Am Anfang des 20. Jahrhunderts war die Dimension der Daten mit dem heutigen Umfang nicht im Ansatz vergleichbar. Mit höherem Umfang wurden die Berechnungen aufwendiger. Im Wesentlichen hat man es bei der FA mit Problemen der linearen Algebra zu tun. Über numerische Methoden konnten die Berechnungen automatisiert und katalogisiert werden. Mit steigender Rechnerleistung, die ab ca. den 60er Jahren in zunehmend kürzeren Zeitintervallen verfügbar gemacht wurde, konnten schrittweise immer höherdimensionale Probleme bewältigt werden.

Vor etwa 20 Jahren erlebte die Informatik-Welt einen Boom an heuristischen Verfahren, welche für damalige Verhältnisse einen hohen Bedarf an Arbeitsspeicher (RAM) und hohe Ansprüche an die Prozessoren (CPU) stellten, die man der Neuroinformatik zuordnet – auch *machine learning* bzw. *Neuronale Netze* genannt. Die Taktratensteigerung des RAM und besonders der CPU wirkten sich unmittelbar auf die Verkürzung der Rechenzeit aus, so daß die Antwortzeiten einer Berechnung auf ein erträgliches Maß zurückgingen. Mit diesen – besonders für die Prognose geeigneten – Methoden, denen großteils ein black-box Charakter innewohnt, die auf Daten trainiert werden müssen und die durch die damals rasant anwachsenden Prozessortakt-Raten im Personal-Computer Bereich auch als Software-Lösungen verfügbar gemacht wurden, konnte man endlich von speziell verschalteter teurer Hardware abgehen. Damals überschätzte man die Möglichkeiten der neuartigen Verfahren, bis hin zu Deutungen, eine Überlegenheit gegenüber älteren statistischen Methoden sei automatisch inkludiert. Das hat sich keinesfalls generell bewahrheitet. Bei einer FA wird manchmal das Verstehen der zugrunde liegenden Mathematik eher als lästig empfunden.

[21] http://deacademic.com/dic.nsf/dewiki/427447 (11.12.2017)

Im ersten Jahrzehnt dieses Jahrhunderts flachte der Boom dann ab. Die Taktraten-Steigerung war in eine Sättigung gekommen, aber die Datenmengen wuchsen weiter. Damit wurden die Methoden, in Relation zu den Daten betrachtet, langsamer.
Man besann sich zurück, auch auf die FA als schlanke Alternative. In Zukunft könnten *machine learner* für normale und kleinere Problemstellungen unattraktiver werden, wenn man den Kosten einer Berechnung, neben der Zeit, auch die Kosten für die verbrauchte Energie hinzurechnet.
Attraktiv ist die FA z.B. in der Psychologie. Dort steht man vor dem Problem nicht direkt meßbare Eigenschaften über manifeste Stellvertreter-Merkmale zu ergründen. Mithilfe von Faktoren versucht man die latenten Eigenschaften zu beschreiben.

Eine FA arbeitet auf die Kovarianz- bzw. Korrelationsmatrix. Wenn im Folgenden auf die Kovarianz argumentiert wird, so gilt es auch für die Korrelation.
Mit $\vec{1}\,\vec{\bar{x}}^T$ $(n \times p)$ erhält man diejenige Matrix, die beim Zentrieren von X herausgerechnet wird: $KX = X - \vec{1}\,\vec{\bar{x}}^T$. Sei F $(n \times g)$ die Matrix der ersten g Faktoren, A $(p \times g)$ die Matrix der zu F gehörenden Faktorladungen (Ladungsmatrix) und E $(n \times p)$ die Matrix der unbeobachteten spezifischen Faktoren, die als Fehler mit eingehen. Dann kann Matrix X $(n \times p)$ zerlegt werden:

$$X = \vec{1}\,\vec{\bar{x}}^T + FA^T + E, \tag{31}$$

mit folgenden Modellannahmen:

- Der Erwartungswert der gemeinsamen Faktoren ist Null: $\mathbb{E}(F) = 0_{n\times g}$.
- Der Erwartungswert der Fehler ist Null: $\mathbb{E}(E) = 0_{n\times p}$.
- Die Kovarianz von F entspricht der Einheitsmatrix: $\operatorname{cov}(F) = \mathbb{E}\left(F^TF\right) = I_g$.
- Die Kovarianz der Fehler: $\operatorname{cov}(E) = \mathbb{E}\left(E^TE\right) = V_p = \operatorname{diag}\left(v_1^2\ ,\ \ldots\ ,\ v_p^2\right)$.
- Die Kovarianz der Faktoren mit den Fehlern: $\operatorname{cov}(F,E) = \mathbb{E}\left(E^TF\right) = 0_{p\times g}$.

Das Modell (31) hat dasselbe Aussehen wie ein multivariates Regressionsmodell, bei dem A^T mit den Regressionskoeffizienten B assoziierbar ist. Allerdings ist hier die Matrix der orthonormalen „Einflußgrößen“ F im Vorfeld zusätzlich unbekannt.

Die Kovarianzmatrix von X:

$$\begin{aligned}\operatorname{cov}\left(X - \vec{1}\,\vec{\bar{x}}^T\right) &\overset{(31)}{=} \operatorname{cov}\left(FA^T + E\right) = \mathbb{E}\left[\left(FA^T + E\right)^T\left(FA^T + E\right)\right] \\ &= \mathbb{E}\left(AF^TFA^T\right) + \mathbb{E}\left(AF^TE\right) + \mathbb{E}\left(E^TFA\right) + \mathbb{E}\left(E^TE\right).\end{aligned}$$

Wobei die Matrix der Mittelwerte $\vec{1}\,\vec{\bar{x}}^T$ fest ist und deswegen in der Kovarianz gestrichen werden kann. In der Gleichung reduziert sich $\mathbb{E}\left(F^TF\right)$ zu I_g ; $V = \mathbb{E}\left(E^TE\right)$. Außerdem verschwinden beide mittleren Erwartungswert-Terme, da die Kovarianz von F mit E nicht existiert. Übrig bleibt das Fundamentaltheorem der Faktorenanalyse:

$$\operatorname{cov}(X) = \Sigma = AA^T + V\ ,\ (p \times p). \tag{32}$$

Mithilfe von V, der Einzelrestvarianzmatrix + Meßfehler, werden die Varianzen spezifischer Faktoren beschrieben. Gleichzeitig reduziert V den Rang von Σ, welcher sich auf AA^T auswirkt.

Die Ladungsmatrix A enthält Kovarianzeinträge (Ladungen genannt), über welche p Einflußgrößen mit g Faktoren verbunden sind. Hierbei beschreibt AA^T die reproduzierte (reduzierte) Kovarianz von X. Auf ihrer Hauptdiagonale sind die quadrierten und summierten Kommunalitäten h_i^2 abzulesen, der Anteil der gemeinsamen Faktoren an der Varianz:

$$\operatorname{diag}(\Sigma - V) = \operatorname{diag}\left(AA^T\right). \tag{33}$$

I.d.R. wird V mithilfe einer Schätzung ermittelt. Das Aussehen der Diagonalmatrix V entscheidet über Anzahl und Art der Faktoren. $\Sigma - V$ ist positiv semidefinit. Es gibt verschiedene FA-Verfahren, eine geeignete V-Matrix zu erhalten/schätzen.

Es kann vorkommen, daß eine Kovarianzmatrix Σ keinen Vollrang p, sondern nur über einen Rang $m < p$ verfügt. Die Maximum-Likelihood-FA ist für dieses Problem nicht konditioniert, da der Zugriff auf die Inverse Σ^{-1} erforderlich ist.

Die in den Kapiteln 4.2 bzw. 4.3 beschriebenen Zerlegungen sind u.a. für zwei Verfahren der FA geeignet, die Hauptkomponentenmethode (Principal component analysis) und die Hauptfaktorenanalyse (Principal factor analysis, auch Principal axis factoring genannt). Von der Methodik sind beide Verfahren gleich. Der Unterschied liegt im Faktorenmodell. Darauf wird zum Schluß näher eingegangen.

4.4.1 Hauptkomponentenmethode (PCA)

Möglichst wenige Faktoren ($g \ll p$) sollen die Varianz in den Daten erklären, bei Vorgabe eines maximal zu akzeptierenden Verlustes an Varianz. Eine anschaulich bildbezogene Einführung zur PCA für einen groben Überblick findet man bei ChemgaPedia[22] auf 10 Seiten. (Chen, 2003) etikettiert die Lösung der PCA u.a. mit „Karhunen“ bzw. „Loève“. Bisweilen findet man dafür den Begriff *Karhunen–Loève-Transformation*[23]. Dies steht als Synonym für eine SVD.

Für Matrix V in (32) wird eine Nullmatrix angesetzt. D.h., V wird nicht gemeinsam mit A geschätzt. Die spezifischen Faktoren werden nicht herausgerechnet. Bzgl. der Fundamentalgleichung (32) resultiert auf Grundlage von (23) als Spektralzerlegung:

$$\Sigma = AA^T = P\Lambda^{1/2}\Lambda^{1/2}P^T = P\Lambda P^T.$$

Mit einer einmaligen vollständigen Zerlegung können die Auswirkungen beliebiger Faktorenzahlen $g \leq m$ untersucht werden. Die Art der Zerlegung ist konform zu (Eckey & Rengers, 2002).
Weil V nicht existiert, ist die geschätzte Kovarianz $\widehat{\Sigma} = C \overset{(30)}{=} X^TKX/(n-1) \overset{(32)}{=} AA^T$, d.h. lediglich die Komposition AA^T steht auf der rechten Seite. Deshalb darf die Zerlegung mit der SVD direkt auf KX (28) vorgenommen werden:

$$X - \vec{1}\vec{\bar{x}}^T = KX = TP^T + E, \tag{34}$$

[22] http://www.chemgapedia.de/ (30.11.2017)
[23] https://de.wikipedia.org/wiki/Singulärwertzerlegung (15.11.2017)

wobei in (28) der Ausdruck UD mit T (29) substituiert wurde.
E degeneriert zur Nullmatrix, falls bei der Zerlegung der Rang m vollständig ausgeschöpft wird. T entspricht den Hauptachsen, der Matrix der projizierten Objekte aus KX. Allerdings entspricht P der Matrix der Eigenvektoren, welche nicht die Ladungen darstellen. Mittels einer nach außen hin neutralen Matrizenoperation, unter Bezug auf (24), kann (34) durch Hauptachsennormierung in (31) überführt werden, nämlich:

$$TP^T = \left(T\Lambda^{-1/2}\right)\left(\Lambda^{1/2}P^T\right) = FA^T, \tag{35}$$

in die orthonormalen Hauptkomponenten/Faktoren F und die Ladungen A. F entspricht somit einer orthogonalen Matrix. Die Matrix $KX = F\Lambda^{1/2}P^T$ entspricht dann der SVD. Diese Darstellung ist äquivalent zu (Fahrmeir & Hamerle, 1984, (3.7), Seite 598).

Bei unvollständiger Zerlegung kommt es zur Rangaufspaltung:

$$m = \text{Rang}\left(AA^T\right) + \text{Rang}\left(E^TE\right),$$

wobei der $\text{Rang}\left(AA^T\right) = g$ beträgt.

In Anlehnung an das Theorem (32) kann formuliert werden: $\text{cov}\,(X) = \Sigma = AA^T + W$. W als Komplement ist nach der Faktoren-Extraktion bekannt. Als Schätzung für W resultiert $\widehat{W} = E^TE/(n-1)$, die restliche empirische Kovarianz. $\widehat{W}$ ist nicht diagonal, denn das würde den Vollrang m implizieren.

Üblicherweise arbeitet man bei einer PCA ohne Verteilungsannahme. Bei einem Kovarianz basierten Verfahren ist eine multivariate Normalverteilung in den Daten der Methode zuträglich und wird wohlwollend akzeptiert: Dann entspricht die gemeinsame Dichte einer Hyper-Ellipse. Im Raum der Hauptkomponenten existiert rein rechnerisch keine Kovarianz mehr. Bei Abwesenheit der multivariaten Normalverteilung ist aber damit noch nicht sichergestellt, tatsächlich über stochastisch unabhängige Hauptkomponenten zu verfügen. (Schmid, 2016b) führt in Satz 10.8. sinngemäß aus, daß bei unabhängigen Zufallsvariablen die gemeinsame Dichte aus dem Produkt der Einzeldichten resultiert. Die Umkehrung gilt i.A. nicht, mit Ausnahme der multivariaten Normalverteilung.

Wiederum beschreibt (Collins, 2010) einen iterativen Weg – auf Seite 13.
Die numerische Zerlegung[24] (Alternative: NIPALS[25], Seite 6) ist hier geringfügig vereinfacht im Vergleich zum Kapitel 4.3 ausgeführt. Denn das Hauptinteresse liegt nun lediglich an der Matrix der Rechts-Singulärvektoren, also V alias P und an der Matrix T.
Die Eigenwerte $\Lambda = T^TT/\,(n-1) = A^TA$ beschreiben die Varianz $\mathbb{V}\,(T)$. Mitunter besteht ein Interesse am Anteil der aufgeklärten Varianz durch die Faktoren:

$$\text{R}^2 \;=\; \text{diag}\left(T^TT\right) / \,\text{Spur}\left(X^TKX\right) \;=\; \text{diag}\left(A^TA\right) / \,\text{Spur}\left(\widehat{\mathbb{V}}(X)\right). \tag{36}$$

Anmerkung:
Bei einer Hauptkomponentenanalyse sind die Eigenwerte skalenabhängig. Man sollte deswegen möglichst auf standardisierte Daten arbeiten. Durch Standardisierung wird aus der Kovarianzmatrix eine Korrelationsmatrix, wie in (22) auf Seite 24 beschrieben.

[24] Die Details für einen numerischen Lösungsansatz lassen sich im Skript hka.m nachvollziehen.
[25] http://www.camo.com/TheUnscrambler/Appendices/ (5.4.2018)

4.4.2 Hauptfaktorenanalyse (HFA)

Angenommen wird, daß zu den Daten p (unbeobachtete) spezifische Faktoren in E gehören, deren Varianzen über Matrix V (32) beschrieben werden sollen. V kommt nun eine zentrale Rolle zu: Sie ist diagonal und beschreibt die Einzelrestvarianzen v_i^2.
In (32) wird über das Setzen der Einzelrestvarianzen ($v_i^2 > 0$) eine Rangreduktion der Kovarianz Σ herbeigeführt, welche mit AA^T als reduzierte Kovarianz $\Sigma - V$ ausgedrückt wird. Mithilfe der Rangreduktion wird die Zahl der gemeinsamen Faktoren eingestellt: $g = \text{Rang}\,(\Sigma - V)$. Negative Einträge in V sind unzulässig (Ultra-Heywood Fall). Bei Nulleinträgen in V (Heywood Fall) kann ggfs. korrigiert werden, indem z.B. das betreffende v_i^2 mit einem $\epsilon > 0$ fixiert wird. Das wird unter anderem in (Fahrmeir & Hamerle, 1984, Seite 591) vorgeschlagen.

Eine HFA ist als zweistufiger Vorgang vorstellbar: Im ersten Schritt wird festgelegt, wieviel Faktoren ins Modell eingehen sollen. V ist adäquat zu wählen, um den Rang von AA^T auf $g < m$ zu reduzieren. Im zweiten Schritt wird auf die reduzierte Kovarianz eine Spektralzerlegung durchgeführt, bei der g Hauptkomponenten, die *Hauptfaktoren*, extrahiert werden:

$$\Sigma - V = P\Lambda^{1/2}\Lambda^{1/2}P^T = AA^T. \tag{37}$$

Das Problem wird i.d.R. iterativ gelöst, da meistens ein geeignetes V zu gegebenem g unbekannt sein wird. Sobald die Konvergenz erreicht wird, ist eine V-Matrix verfügbar und die Zerlegung kann final durchgeführt werden.

Schematischer Ablauf der Faktorenextraktion[26]:

- schätze unter Zuhilfenahme der Zentrierungsmatrix K die Kovarianz in den Daten $C = X^TKX/(n-1)$; wähle Faktorenzahl g; verwende für V z.B. die Nullmatrix
- zerlege symmetrische Matrix $C - V$ in Matrix P der Eigenvektoren und Diagonalmatrix L der Eigenwerte, so daß $C - V = PLP^T$
- behalte in L die g größten Eigenwerte und die in P zugehörigen Eigenvektoren, so daß L $(g \times g)$ und P $(p \times g)$
- wiederhole folgende Berechnungen bis zur Konvergenz (d.h. der Fehler sinkt unterhalb einer Fehlerschranke ε)
  - $L_{\text{alt}} = L$
  - Rest $= C - PLP^T$
  - $V = \text{diag}\,(\text{Rest})$
  - berechne Eigenvektoren P und Eigenwerte L von $(C - V)$
  - Fehler $= \sqrt{\text{diag}(L_{\text{alt}} - L)^T \text{diag}(L_{\text{alt}} - L)}$ $\qquad$ *%l^2 – Vektornorm*
- $A = PL^{1/2}$; $\quad \hat{F} = KXA\left(A^TA\right)^{-1} = KXPL^{-1/2}$ $\qquad$ *%$KX = FA^T + E$*
  %$A\left(A^TA\right)^{-1} = PL^{1/2}\left(L^{1/2}P^TPL^{1/2}\right)^{-1} = PL^{1/2}\left(L^{1/2}I_gL^{1/2}\right)^{-1} = PL^{1/2}L^{-1} = PL^{-1/2}$

[26] Die Details für einen numerischen Lösungsansatz lassen sich im Skript hfa.m nachvollziehen.

Die Faktorenmatrix $\hat{F}$ ist nicht orthogonal, d.h. $\widehat{\text{cov}}\left(\hat{F}\right) = \hat{F}^T\hat{F}/(n-1) \neq I_g$.
Moderate Abweichungen von der Einheitsmatrix I_g treten auf, d.h. die resultierende Matrix weicht von deren Hauptdiagonale ab und ist i.d.R. vollbesetzt. Der Grund ist in V zu suchen: Die Einzelrestvarianzen werden aus C herausgerechnet, aber in den Daten X verbleibt die Information.

Probleme:

- Aufgrund nur in endlicher Mantisse darstellbaren Fließkommazahlen (siehe Kapitel 2.1), d.h. unvermeidbar numerischer Ungenauigkeit, ist die final angezeigte Rangreduktion bzgl. $(C - V)$ evtl. geringer als über g voreingestellt.
- Abweichende Startwerte für V können u.U. eine andere finale V-Matrix hervorbringen, die letztendlich zu anderen Kommunalitäten führt.
  Wünschenswert wäre aber über diejenige V-Matrix zu verfügen, die in der Summe maximale Kommunalitäten realisiert.
- Abgesehen von dem Fall der Rangreduktion um Eins, auf eine propere Kovarianzmatrix C, kann für beliebige Rangreduktionen ansonsten nicht garantiert werden, auf der Diagonale von V über ausschließlich positive Einträge zu verfügen.
- Die Faktorenzahl g sollte im Vorfeld gut begründet/nachvollziehbar sein.
  Das Ändern von g führt zu Änderungen in den Kommunalitäten, dahingehend, daß mit steigendem g im Mittel auch eine Informationserhöhung in den Faktoren einhergeht, da mit steigendem g die Hauptdiagonale weniger stark reduziert wird. AA^T wird dadurch positiver. Das Interpretieren von Faktoren würde bei nicht festgelegtem g zusätzlich erschwert.
- Falls $n < p$ vorliegt, resultiert eine impropere Kovarianzmatrix. Deren Rangdefizit beeinträchtigt die Konvergenz, so daß eine widersprüchliche Lösung auftreten kann.

4.4.3 Zusammenhang zwischen PCA und HFA

Während die PCA die Dimensionen des abbildenden Raumes Varianz optimal ausrichtet, versucht die HFA latente Merkmale/Eigenschaften zu ergründen.

Es sei $g = \text{Rang}\left(AA^T\right) < m$. Dennoch gelingt bei der HFA die Abkürzung über die SVD nicht: Zwar existiert V, aber die zugrunde liegende Matrix, welche die reduzierte Kovarianz $(C - V)$ erzeugt, ist unbekannt. Das Problem stellt sich auch bei der Multidimensionalen Skalierung[27] (MDS), bei der auf Grundlage einer Skalarproduktmatrix $(n \times n)$, welche Eigenschaften von n Koordinaten beschreibt, eine Karte bzgl. der n Einträge im g-dimensionalen Unterraum geplottet wird.

Wählt man $g \stackrel{!}{=} m = \text{Rang}\,(C)$, tritt der Grenzfall $V = 0_p$ ein – die HFA degeneriert zur PCA und es gilt folgender Zusammenhang:
Bei der klassischen MDS von Torgerson, der Haupt-Koordinaten-Methode auf Euklidische Distanzen, kann der Skalarproduktmatrix – ohne Kenntnis von X – dieselbe Rang-Information wie der Kovarianz C entnommen werden. Die zentrierten Koordinaten, welche

[27] http://www.faes.de/ (5.1.2018)

den Hauptachsen T in (34) adäquat sind, werden bei der MDS mit der Spektralzerlegung auf die Skalarproduktmatrix erhalten. Die PCA kann somit als ein Spezialfall der MDS betrachtet werden.

Angenommen, die Lösung einer HFA mit g extrahierten Faktoren F inkl. der Ladungen A liegt vor: FA^T, aber die Ur-Daten X seien nicht zugänglich. Nachträglich wird nun eine PCA-Lösung benötigt. Mit der Cholesky-Zerlegung[28] kann eine symmetrische positiv definite Matrix, die Kovarianz der Faktoren, zerlegt werden: $F^TF/(n-1) = JJ^T$, bei der J eine untere Dreiecksmatrix darstellt. Die orthogonalisierten und normierten PCA-Faktoren sind über die Transformation $G = F\left(J^T\right)^{-1}$ erhältlich.
Wurde die HFA auf die Korrelation, d.h. die normierte Kovarianz, von X berechnet, so gelingt über AA^T die Rekonstruktion der normierten Kovarianz C, indem die Hauptdiagonale zu 1 ergänzt wird. Darauf eine Spektralzerlegung (Eigenvektoren P_2, Eigenwerte L_2) ausgeführt, ergibt für die Ladungsmatrix: $B = P_2L_2^{1/2}$. Nur die Einträge, die zu den g größten Eigenwerten gehören sind dabei interessant. Die erklärte Varianz B^TB ist der PCA-Lösung äquivalent und es tritt eine Erhöhung der Kommunalitäten $\text{diag}\left(BB^T\right)$ im Vergleich zu $\text{diag}\left(AA^T\right)$ auf Grundlage von (33) ein.
In guter Näherung beschreibt GB^T jetzt die PCA-Lösung für g extrahierte Faktoren.

Generell ist die Transformation von F zu orthogonalem G möglich. Mit Bezug auf G kann die reduzierte Kovarianz erweitert werden:

$$C - V \;=\; AA^T \;=\; AG^TGA^T/\,(n-1) \;=\; \widehat{\text{cov}}\left(GA^T\right).$$

D.h., GA^T entspricht der Datengrundlage, die $C-V$ äquivalent ist. Die Darstellung GA^T wäre eine Möglichkeit die Auswirkungen des Entfernens der spezifischen Faktoren E auch auf der Datenseite zu erfahren.

4.5 Moore-Penrose-Inverse (MPI)

Sei Matrix A quadratisch und von vollem Rang, dann ist A invertierbar und es folgt $AA^{-1} = A^{-1}A = I$, die Einheitsmatrix.
Daraus folgt weiterhin $AA^{-1}A = A$ und $A^{-1}AA^{-1} = A^{-1}$.
Die Inverse für eine beliebige reelle Matrix beschreiben (Schmidt & Trenkler, 2015, Kapitel 5) als die verallgemeinerte Inverse A^- (g-Inverse) von A, wenn gilt: $AA^-A = A$. Diese Inverse muß nicht eindeutig sein, d.h. verschiedene Lösungen können existieren.
Es sei festgestellt, daß das Problem der Invertierung von rechteckigen Matrizen nicht vorliegt. Damit genügt es, den Spezialfall von verallgemeinerten Inversen, die Moore-Penrose-Inverse (MPI), näher zu betrachten.
(Schmidt & Trenkler, 2015) schreiben in Kapitel 6:

> Fordert man darüber hinaus, dass $A^-AA^- = A^-$ ist, und dass sowohl A^-A als auch AA^- symmetrisch sind, so wird dies nur von einer einzigen g-Inversen, nämlich der Moore-Penrose-Inversen A^+, erfüllt.

[28] kompakte allgemeine Einführung (Stoer, 1979, Seite 146 – 149) zu entnehmen

In dem Zitat sind drei Bedingungen erwähnt. Gemeinsam mit der eingangs erwähnten Bedingung für die g-Inverse entspricht es vier Bedingungen für die Definition einer MPI. Die obigen Autoren schlagen für die Berechnung der MPI einen iterativen Ansatz, den Greville-Algorithmus vor. Im Artikel von (Saraev, 2013) wird eine Modifikation jenes Algorithmus anhand einer Intervallhalbierung mithilfe einer Intervallmatrix T illustriert. Hinweis: An dieser Stelle wird im Rahmen des Kapitel 4 eine andere Lösung[29] favorisiert, eine Ranguntersuchung mithilfe der Spektralzerlegung von Matrix A.

Schematische Beschreibung der Invertierung[30]:

- zerlege symmetrische Matrix A in Matrix P der Eigenvektoren und Diagonalmatrix L der Eigenwerte, so daß $A = PLP^T$; $\quad L^+ = L$
- identifiziere in L^+ jene Eigenwerte, die betragsmäßig unter einer Null-Toleranzschwelle liegen und setze diese Eigenwerte auf ∞
- $L^+ = (L^+)^{-1}$ $\quad$ *%Matrix-Inverse,* $^1/_\infty = 0$
  %LL^+ entspricht einer „Einheitsmatrix“, auf deren Hauptdiagonale die Zahl der %Nulleinträge dem Rangverlust von A entsprechen $\rightarrow$ *Rang* $(A) =$ *Spur* (LL^+)
- berechne die Moore-Penrose-Inverse $A^+ = PL^+P^T$

Der Rang von A spiegelt sich getreu im Rang von A^+ wider. Verfügt A über vollen Rang, so ist $A^+ = A^{-1}$. Falls $A = X^TX$ zu berechnen ist, kann alternativ eine SVD direkt auf X berechnet werden und unter Zuhilfenahme der Singulärvektor-Matrizen die MPI dann aufgebaut werden. Mithin spart man hiermit bei großen und gleichzeitig stark Rang reduzierten A-Matrizen Rechenzeit ein.

Regression bei singulärem X^TX
Insofern beim Ridge-Schätzer (13) die Penalty-Matrix D einer Einheitsmatrix entspricht und der Tuningparameter λ kleinstmöglich – aber größer Null – gewählt wird, führt das zur Äquivalenz: $\left(X^TX + \lambda D\right)^{-1} = \left(X^TX\right)^+$. Eine kleinste-Quadrate-Lösung mittels

$$\vec{\hat{\beta}}_+ = \left(X^TX\right)^+ X^T\vec{y}\,, \ (p \times 1) \tag{38}$$

zu erhalten, beinhaltet im regulären Fall den KQ-Schätzer aus (5).

[29] https://de.mathworks.com/help/matlab/ref/pinv.html? (9.11.2017)
[30] Die Details für einen Lösungsansatz lassen sich im Skript mpi.m nachvollziehen.

5 Matrix – Zerlegungstechniken II (PLS)

Mithilfe der Zerlegungen soll verstanden werden, wie eine partielle kleinste Quadrate Regression (PLS) gerechnet werden kann. Auf der PLS liegt der Schwerpunkt dieses Kapitels.

Im Unterschied zur PCA, bei der die Varianz der Datenwolke mittels Linearkombinationen auf X maximiert wird, wird bei einer PLS mit Linearkombinationen die Kovarianz der Zielgröße zu X maximiert. Dies ist erforderlich, um im Anschluß die Art des Zusammenhangs zwischen Zielgröße und Kovariable zu identifizieren – über ein lineares Regressionsproblem. D.h., mit dem PLS-Verfahren sollen simultan zwei Probleme bewältigt werden. Im Unterschied zu einer Regression auf Hauptkomponentenwerte, bei der die Zahl der Faktoren die Zeilenzahl der Regressions-Koeffizienten vorgibt, beträgt bei einer PLS die Zeilenzahl in den Koeffizienten der Anzahl p der Ursprungsvariablen – wie bei einer Regression direkt auf eine Designmatrix.
Wenn aber $n < p$ auftritt, fällt die Vorstellung evtl. besonders schwer, inwieweit ein direkter Zusammenhang, im Sinne eines kleinsten Verlustes, auf p Merkmale herstellbar sein kann: Man steht vor dem Problem eines unterbestimmten Gleichungssystems[31]. Bei einer Regression auf die puren Kovariablen wäre eine eindeutige bzw. bei Regularisierung eine MSE-minimale Lösung bisher unmöglich. Das Problem ist jetzt lösbar.

Für die Berechnungen wird, wie in den vorangegangenen Kapiteln, der NIPALS-Algorithmus (Nonlinear Iterative Partial Least Squares) eingesetzt. Die Gemeinsamkeiten in den Begrifflichkeiten *NIPALS* und *PLS* deuten es schon an: der Algorithmus ist für die PLS-Schätzung das klassische Werkzeug.
Herman Wold darf als Vater von NIPALS gelten und hat beim Design des NIPALS-Algorithmus entscheidende Impulse gesetzt. Seine ersten Veröffentlichungen zu diesem Thema liegen bereits mehr als 50 Jahre zurück. NIPALS wird seitdem ständig weiterentwickelt[32] und in seinem Funktionsumfang verbreitert. Sein Sohn Svante ist ebenso involviert. Offensichtlich wird die PLS im skandinavischen Raum, vor allem in Norwegen, besonders wertgeschätzt.

5.1 PLS auf eine univariate Zielgröße (PLS1)

Beschrieben wird die PLS, welche auf einen Zielgrößen-Vektor optimiert. Aufgrund der vektoriellen Zielgröße $\vec{y}$ nennt man dieses Verfahren abgekürzt PLS1.
Bei einer Spektralzerlegung kann für die Einflußgrößen die Kovarianzmatrix $C = \mathrm{cov}\,(X)$ gemäß (23) geschrieben werden. Daraus resultiert eine quadratische Matrix. Die maximale Varianz kann unter Einbezug des ersten Eigenvektors $\vec{p}_1$ äquivalent zu Formel (25) beschrieben werden:

$$\mathrm{cov}\,(X\vec{p}_1) = \vec{p}_1^{\,T}\mathrm{cov}\,(X)\,\vec{p}_1 \rightarrow \max_{\vec{p}_1}.$$

[31] https://www.matopt.de/grundlagen/loesung-unterbestimmte-gleichungssysteme.html (25.7.2017)
[32] http://sagaofpls.github.io/chapter4.html (11.5.2017)

Kommt die Zielgröße mit hinzu, geht die quadratische Form der Kovarianz verloren:

$$\vec{c} = \widehat{\mathrm{cov}}\,(X, \vec{y}\,) = \frac{1}{n-1} X^T K \vec{y}\;,\ (p \times 1). \tag{39}$$

Bei Hinzunahme des Eigenvektors $\vec{p}_1$, zum Erhalten der maximalen Varianz, entsteht:

$$\mathrm{cov}\,(X\vec{p}_1, \vec{y}\,) = \vec{p}_1^{\,T} \mathrm{cov}\,(X, \vec{y}\,) \to \max_{\vec{p}_1}.$$

Das Kovarianz-Problem kann nun nicht über eine Spektralzerlegung gelöst werden. (Krämer, Boulesteix & Tutz, 2006) schlagen zum Maximieren mit Lagrange-Multiplikatoren die quadrierte Kovarianz vor. Als quadrierte Kovarianz wird dort sinngemäß definiert:

$$(n-1)^2\, \widehat{\mathrm{cov}}^2(X, \vec{y}\,) \overset{(39)}{=} X^T K \vec{y}\, \vec{y}^{\,T} K X\;,\ (p \times p). \tag{40}$$

Auf die resultierende quadratische Matrix ist eine Spektralzerlegung anwendbar. Das täuscht aber nicht darüber hinweg, daß der Rang dieser Matrix lediglich Eins beträgt. Um Faktoren extrahieren zu können, ist offenbar ein abgewandelter Ansatz anzuwenden. Damit kommt man zum eigentlichen Zweck des NIPALS-Algorithmus. D.h., der Ablauf der Varianzmaximierung gestaltet sich im Detail anders als in den Kapiteln 4.2 – 4.4.1 beschrieben. Bei einer „klassischen“ Spektralzerlegung gibt der Rang m der Kovarianzmatrix C die maximale Dimensionszahl des Unterraums vor.
Da der Fall $n < p$ vorliegt, ist es ohnehin empfehlenswert, von der Spektralzerlegung abzusehen. Wie bei einer SVD, im Kapitel 4.3 gezeigt, soll die Zerlegung direkt auf die Daten erfolgen, wobei $g \leq m$ zur Auswahl steht. Im Unterschied zu Kapitel 3.3 werden Einflußgrößen und Zielgröße aber simultan zerlegt:

$$\begin{aligned} X &= \vec{1}\,\vec{\bar{x}}^{\,T} + TP^T + E \\ \vec{y} &= \vec{1}\,\bar{y}\ + T\vec{Q}^T + \vec{f}. \end{aligned} \tag{41}$$

Die Diagonalmatrix Λ der Eigenwerte kann über die Kovarianz auf T gebildet werden:

$$\Lambda \overset{(24)}{=} \widehat{\mathrm{cov}}\,(T) = \frac{1}{n-1} T^T T\;,\ (g \times g). \tag{42}$$

Für das Verständnis der später zu beschreibenden Formeln kann es hilfreich sein, auch die Sicht des faktorenanalytischen Modells einzunehmen und folgende Transformationen auf (41) durchzuführen:

$$\begin{aligned} TP^T &= \left(T\Lambda^{-1/2}\right)\left(\Lambda^{1/2} P^T\right) \overset{(35)}{=} FA^T \\ T\vec{Q}^T &= \left(T\Lambda^{-1/2}\right)\left(\Lambda^{1/2}\vec{Q}^T\right)\ = F\vec{b}^{\,T}. \end{aligned} \tag{43}$$

Im Sinne eines faktorenanalytischen Modells handelt es sich bei T, unter Bezug auf Kapitel 4.4, um die Matrix der Hauptachsen. Ihre Normierung via (43) produziert, im Eigentlichen, die Faktoren F. – Die Matrix T sollte nicht *Faktoren* genannt werden. Bzgl. P und $\vec{Q}$ könnte die Bezeichnung *skalierte Ladungsmatrix /-vektor* präziser (aber umständlicher) sein, da es sich um Größen handelt, welche eine Assoziation zu den Eigenvektoren zulassen.

Beschreibung der Matrizen:

- PLS-Hauptachsenmatrix: T $(n \times g)$, orthogonal
- Ladungsmatrix: P $(p \times g)$, nicht orthogonal
  P^TP nicht diagonal sondern eine symmetrische Band-Matrix
- Ladungsvektor: $\vec{Q}$ $(1 \times g)$
- Fehlermatrix: E $(n \times p)$ und Fehlervektor: $\vec{f}$ $(n \times 1)$
- PLS-Faktorenmatrix: F $(n \times g)$, orthogonal
- Ladungsmatrix: A $(p \times g)$, nicht orthogonal
- Ladungsvektor: $\vec{b}$ $(1 \times g)$.

Bei vollständiger Faktoren-Extraktion gelten folgende Zusammenhänge:

$$\widehat{\mathrm{cov}}\,(X) \overset{(23)}{=} \tfrac{1}{n-1} X^T K X \overset{(41)}{=} \mathrm{cov}\left(TP^T + E\right) \overset{E=0}{=} \mathrm{cov}\left(TP^T\right) \overset{(42)}{=} P\Lambda P^T,$$

$$\widehat{\mathbb{V}}\,(\vec{y}) = \tfrac{1}{n-1} \vec{y}^{\,T} K \vec{y} \overset{(41)}{=} \mathbb{V}\left(T\vec{Q}^T + \vec{f}\right) \overset{\vec{f}=\vec{0}}{=} \mathbb{V}\left(T\vec{Q}^{\,T}\right) \overset{(42)}{=} \vec{Q}\Lambda\vec{Q}^T,$$

$$\widehat{\mathrm{cov}}\,(X, \vec{y}) \overset{(39)}{=} \tfrac{1}{n-1} X^T K \vec{y} \overset{(41)}{=} \mathbb{E}\left(\left(TP^T + E\right)^T \left(T\vec{Q}^T + \vec{f}\right)\right) \overset{\vec{f}=\vec{0}}{=} \tfrac{1}{n-1} PT^T T\vec{Q}^T \overset{(42)}{=} P\Lambda\vec{Q}^T.$$

Vertritt man dabei die Sichtweise eines Kovarianz-Strukturmodells, kann in Anlehnung an (Long, 1983, Seite 24) aus den Einzelgleichungen eine Kovarianz-Gleichung Σ notiert werden:

$$\Sigma = \mathbb{E}\left[\begin{pmatrix}\vec{y}^{\,T}\\ X^T\end{pmatrix} K\,(\vec{y}\ ,\ X)\right] = \begin{bmatrix}\mathbb{E}\left(\vec{y}^{\,T}K\vec{y}\right) & \mathbb{E}\left(\vec{y}^{\,T}KX\right)\\ \mathbb{E}\left(X^TK\vec{y}\right) & \mathbb{E}\left(X^TKX\right)\end{bmatrix}$$

$$= \begin{bmatrix}\mathbb{E}\left(\left(T\vec{Q}^{\,T} + \vec{f}\right)^T\left(T\vec{Q}^{\,T} + \vec{f}\right)\right) & \mathbb{E}\left(\left(T\vec{Q}^{\,T} + \vec{f}\right)^T\left(TP^T + E\right)\right)\\ \mathbb{E}\left(\left(TP^T + E\right)^T\left(T\vec{Q}^{\,T} + \vec{f}\right)\right) & \mathbb{E}\left(\left(TP^T + E\right)^T\left(TP^T + E\right)\right)\end{bmatrix} = \begin{bmatrix}\vec{Q}\Lambda\vec{Q}^T & \vec{Q}\Lambda P^T\\ P\Lambda\vec{Q}^T & P\Lambda P^T\end{bmatrix}, \begin{pmatrix}1\times 1\ ,\ 1\times p\\ p\times 1\ ,\ p\times p\end{pmatrix}.$$

Bei einer PLS ist die Kovarianz, also die Nebendiagonale von Σ interessant.
Vorgetragen wurde die Zerlegung $\mathrm{cov}\,(X) = P\Lambda P^T$ in (23) auf Seite 24. Wenn auch die Zerlegungen bzgl. der Gleichung und Symbolik übereinstimmen, ist doch der Unterschied fundamental: In (Burns & Ciurczak, 2008) schreiben die Autoren Bjørsvik und Martens auf Seite 195, daß die Ladungsmatrix P nicht orthogonal ist. Dementsprechend sind auch die Λ-Matrizen der PCA und PLS nicht direkt miteinander vergleichbar.
Die Anteilswerte der durch die Faktoren aufgeklärten Varianz für $(X, \vec{y})$ betragen:

$$\begin{aligned}
\mathrm{R}^2_X &= \frac{\mathrm{diag}\left(A^TA\right)}{\mathrm{Spur}\left(\widehat{\mathbb{V}}(X)\right)} \overset{(43)}{=} \frac{\mathrm{diag}\left(\Lambda^{1/2}P^TP\Lambda^{1/2}\right)}{\mathrm{Spur}\left(\widehat{\mathbb{V}}(X)\right)} = \frac{\mathrm{diag}\left(\Lambda P^TP\right)}{\mathrm{Spur}\left(\widehat{\mathbb{V}}(X)\right)} \overset{(42)}{=} \frac{\mathrm{diag}\left(T^TTP^TP\right)}{\mathrm{Spur}\left(X^TKX\right)}\\
\mathrm{R}^2_{\vec{y}} &= \frac{\mathrm{diag}\left(\vec{b}^T\vec{b}\right)}{\widehat{\mathbb{V}}\,(\vec{y})} \overset{(43)}{=} \frac{\mathrm{diag}\left(\Lambda^{1/2}\vec{Q}^T\vec{Q}\Lambda^{1/2}\right)}{\widehat{\mathbb{V}}\,(\vec{y})} = \frac{\mathrm{diag}\left(\Lambda\vec{Q}^T\vec{Q}\right)}{\widehat{\mathbb{V}}\,(\vec{y})} \overset{(42)}{=} \frac{\mathrm{diag}\left(T^TT\vec{Q}^T\vec{Q}\right)}{\vec{y}^{\,T}K\vec{y}}.
\end{aligned} \tag{44}$$

(Collins, 2010) verweist bei der Angabe seiner Quellen u.a. auf (Jørgensen & Goegebeur, 2007a). Darin wird der PLS1-Algorithmus beschrieben.

Schematische Beschreibung der Dekomposition[33] mithilfe des NIPALS-Algorithmus:
– erzeuge temporäre zentrierte Matrix $E = KX$ und zentrierten Vektor $\vec{f} = K\vec{y}$
– berechne $m = \text{Rang}\,(E)$
– wähle Zahl g zu extrahierender Faktoren, unter der Nebenbedingung $g \leq m$
– erzeuge leere Matrizen W, T, P und leeren Vektor $\vec{Q}$
– initialisiere Laufvariable $k = 0$
– wiederhole folgende Operationen solange Iterationszahl $k < g$

- $k \stackrel{!}{=} k + 1$ *%Nr. des Faktors*
- $\vec{w} = E^T\vec{f}\,; \;\; \vec{w} = \frac{\vec{w}}{\sqrt{\vec{w}^T\vec{w}}}$ *%Ladungsgewichte/Eigenvektoren* $(p \times 1)$ *für Faktor k*
- $\vec{t} = E\vec{w}$ *%k. Hauptachse* $(n \times 1)$
- $\vec{p} = E^T\vec{t}/\left(\vec{t}^{\,T}\vec{t}\right)$ *%Ladung* $(p \times 1)$ *für Faktor k*
- $q = \vec{f}^{\,T}\vec{t}/\left(\vec{t}^{\,T}\vec{t}\right)$ *%Gewicht* (1×1)
- reihe Vektor $\vec{w}$ als neue Spalte in Matrix W, $\vec{t}$ in T, $\vec{p}$ in P und Skalar q in $\vec{Q}$ ein
- $E \stackrel{!}{=} E - \vec{t}\vec{p}^{\,T}$ *%Residuen* $(n \times p)$ *für KX bei k Faktoren*
- $\vec{f} \stackrel{!}{=} \vec{f} - \vec{t}\,q$ *%Residuen* $(n \times 1)$ *für* $K\vec{y}$ *bei k Faktoren*

Wird die Faktoren-Extraktion nicht vorher abgebrochen ($g \stackrel{!}{=} m$), entspricht die Fehlermatrix E am Prozedurende bis auf numerische Unzulänglichkeiten einer Nullmatrix und $\vec{f}$ einem Nullvektor.
Unverkennbar ist mit dem Schritt $\vec{w} = E^T\vec{f}$ $(p \times 1)$ der Bezug zur Kovarianz (39). D.h., im Algorithmus taucht eine bisher nicht erwähnte orthonormale Ladungsgewichts-Matrix W $(p \times g)$ auf: $W^TW = I_g$. Sie beeinflußt durch ihre dominante Position im Algorithmus alle anderen Matrizen bzw. Vektoren und wird außerdem für die Regression benötigt.

5.1.1 Regression – Herleitung des KQ-Schätzer $\hat{\vec{\beta}}_{PLS}$

In den Kapiteln 3.1 und 3.2 wurde die Unmöglichkeit einer optimalen Schätzung gezeigt und diskutiert, falls das Problem $n < p$ auftritt. Dieses Problem soll nun gelöst werden. Unter Bezug auf (41) kann man die Daten mithilfe von (20) auch einzentriert darstellen: $KX = TP^T + E$ und $K\vec{y} = T\vec{Q}^T + \vec{f}$. Die Fehler E und $\vec{f}$ sollen ab jetzt ignoriert werden. Ihre Erwartungswerte betragen Null. Es ist klar, daß bei nicht vollständiger Faktoren-Extraktion das Ignorieren von E und $\vec{f}$ Auswirkungen auf die Regressions-Schätzung haben wird. Nun verbleibt:

$$\begin{aligned} KX &= TP^T \\ K\vec{y} &= T\vec{Q}^T. \end{aligned} \tag{45}$$

Die erste Gleichung in (45) wird nach den Hauptachsen aufgelöst und es resultiert:

$$T = KXP\left(P^TP\right)^{-1}.$$

[33] Die Details für einen numerischen Lösungsansatz lassen sich im Skript pls1.m nachvollziehen.

Solange die Matrix X vollständig zerlegt wird, ist gegen diesen Ansatz soweit nichts einzuwenden. Generell ist bei Faktoren-Extraktionen ein striktes Hierarchie-Prinzip einzuhalten: Faktoren, welche bereits extrahiert wurden, sind unabhängig bzgl. weiterer Extraktionen. Für die Darlegungen in Kapitel 4.4.1 gilt dies ohnehin, da in (34) Matrix P in ihrer Eigenschaft orthonormal ist. Dort kollabiert P^TP zur Einheitsmatrix und für die Hauptachsen verbleibt: $T = KXP$, aus denen gemäß (43) die Faktoren geformt werden können: $F = T\Lambda^{-1/2}$.
Bei einer PLS offenbart sich im Kreuzprodukt P^TP eine Tridiagonalstruktur (Bandmatrix mit Bandweite 3 und Vollrang), ein Hinweis auf Autokorrelation erster Ordnung (AR(1)[34] genannt). Das deutet auf einen autoregressiven Stör-Prozeß, als zugrunde liegenden Prozeß, hin.
P^TP kann als Präzisionsmatrix gedeutet werden. Darin lassen sich Nachbarschaftsbeziehungen ablesen. Ihre Invertierung führt zur Autokorrelationsmatrix. Mit der Autokorrelation wird die Markov-Eigenschaft abgebildet, welche – aufgrund der Beschaffenheit der vorliegenden Präzisionsmatrix – einen Horizont von einer Periode aufweist.
Im Zeitreihenkontext argumentiert: Solange die Zukunft – also ein neu zu extrahierender Faktor – von der Vergangenheit abhängt, ist das unproblematisch. Mit dem Problem wird man auch in Kapitel 4.4.2 konfrontiert. D.h., falls hinzukommende Faktoren die Vergangenheit – also bereits extrahierte Faktoren – ändern, dann ist das System nicht kausal. Die Extraktionsmethode wäre unreliabel[35]. Bjørsvik und Martens führen sinngemäß auf Seite 195 weiterhin aus, daß gewöhnlich W und P sich sehr ähnlich sind.
Das Autokorrelationsproblem wird bei der PLS mit der orthonormalen Ladungsmatrix W begradigt:

$$KX = TP^T \quad \Leftrightarrow \quad KXW = TP^TW \quad \Leftrightarrow \quad T = KXW\left(P^TW\right)^{-1}. \tag{46}$$

P^TW entspricht einer oberen Dreiecksmatrix mit einer durch Einsen besetzten Hauptdiagonale und ihrer ersten besetzten Nebendiagonale. Obwohl bei vollständig extrahierten Faktoren gilt: $P^TWW^TP = P^TP$, entspricht P^TW nicht der Cholesky-Wurzel von P^TP, sondern $\left(P^TW\right)^{-1}$ – die vollbesetzte obere Dreiecksmatrix – entspricht der Cholesky-Zerlegung von $\left(P^TP\right)^{-1}$.
Bereits extrahierte Faktoren werden durch hinzukommende Faktoren nicht mehr geändert. Das wird über die Dreiecksform von P^TW abgesichert. Demzufolge wird die schädliche Wirkung der Autokorrelation vollständig ausgeschaltet.
Setzt man die Hauptachsenmatrix T aus (46) zurück in die Gleichungen von (45) ein, folgt daraus:

$$\begin{aligned} KX &= KXW\left(P^TW\right)^{-1}P^T \\ K\vec{y} &= KXW\left(P^TW\right)^{-1}\vec{Q}^T. \end{aligned} \tag{46.1}$$

Außerdem gilt:

$$\left(P^TW\right)^{-1}P^TW = I_k \quad \text{und} \quad W^TW = I_k \quad \Leftrightarrow \quad W^T = \left(P^TW\right)^{-1}P^T, \tag{46.2}$$

[34] im univariaten Modell: $y_t = \varphi y_{t-1} + \epsilon_t$ mit $|\varphi| < 1$
[35] https://www.psychomeda.de/lexikon/reliabilitaet.html (21.5.2017)

welches in die erste Gleichung (46.1) substituiert wird: $KX \overset{(46.2)}{=} KXWW^T \overset{(45)}{=} TP^T$. Damit wäre für die erste Gleichung in (45) eine wahre Aussage gezeigt.
Der Hinweis auf (45) ist gerechtfertigt, denn die Erweiterung der ersten Gleichung mit W in (46) bleibt wahr: $KXW = TP^TW$.

Der KQ-Schätzer für die zentrierten Größen basiert auf der Struktur von Formel (5):

$$\vec{\hat{\beta}}_{PLS} \overset{(45)}{=} \Big(\big(TP^T\big)^T TP^T\Big)^{-1}\big(TP^T\big)^T\vec{y} \overset{(46.1)}{=} \big(PT^TTP^T\big)^{-1}PT^TTP^TW\big(P^TW\big)^{-1}\vec{Q}^T.$$

Beim letzten Substitutionsschritt wurde automatisch unter Bezug auf (45) gleichzeitig die Ersetzung $KX = TP^T$ vorgenommen. Im KQ-Schätzer offenbart sich aber in der Inverse ein Problem, dahingehend, daß T^TT eine Matrix der Dimension $g \times g$, Matrix PT^TTP^T die Dimension $p \times p$ besitzt, welches einem Rangverlust von $p-g$ entspricht. Die Singularität kann durch einfaches Herauskürzen der Inverse mit ihrem nicht invertierten Pedant umgangen werden. Zurück bleibt ein identifizierbarer Regressions-Schätzer:

$$\vec{\hat{\beta}}_{PLS} = W\big(P^TW\big)^{-1}\vec{Q}^T,\ (p \times 1). \tag{46.3}$$

Bei einer PLS regressiert man nicht direkt auf die Designmatrix und verfügt deshalb über keine Scheinvariable. Der Achsenabschnitt (Kalibration genannt) wird separat berechnet:

$$\hat{\beta}_0 = \frac{1}{n}\vec{1}^T\Big(\vec{y} - X\vec{\hat{\beta}}_{PLS}\Big)\ ,\ (1 \times 1) \tag{46.4}$$

und man erhält somit die Prognosefunktion:

$$\vec{\hat{y}} = \hat{\beta}_0 + X\vec{\hat{\beta}}_{PLS}\ ,\ (n \times 1). \tag{46.5}$$

Es ist möglich ein Abbruchkriterium zu definieren, beispielsweise über die Norm des mittleren quadratischen Vorhersagefehlers (RMSEP):

$$\text{RMSEP} = \sqrt{\frac{1}{n}\left(\vec{y} - \vec{\hat{y}}\right)^T\left(\vec{y} - \vec{\hat{y}}\right)}. \tag{47}$$

Da der Fehler der Zerlegung mit jeder weiteren Iteration monoton sinkt, kann man beim bloßen Anwenden eines solchen Kriteriums die benötigte Faktorenzahl nur grob schätzen. Sinnvoller wäre sein Einsatz im Rahmen einer Kreuzvalidierung.
Kreuzvalidierungen können aber rechenzeitintensiv sein.

PLS-Regression als Verallgemeinerung der multiplen linearen Regression
Verfügt man mit X^TX über eine invertierbare Matrix ($n \geq p$) und schöpft den Rang bei der PLS-Regression voll aus ($g \overset{!}{=} p$), so verstehen sich die Lösungen für $\vec{\beta}$ bzgl. der Regression nach Kapitel 3.1 als äquivalent. Lediglich der Achsenabschnitt ist bei der PLS-Regression vom β-Vektor separiert. Das läßt sich nachträglich durch Zusammenfügen zu einem Vektor vereinen, um auf die klassische Designmatrix operieren zu können.

Numerisch stabiler ist die PLS-Regression. Starke Abhängigkeiten in den Kovariablen sind ja gerade das Hauptanliegen der PLS. Desweiteren kann bei Anwenden des NIPALS-

Algorithmus auf das Kreuzprodukt der Einflußgrößen verzichtet werden, welches die Effizienz des Algorithmus gegenüber dem einfachen KQ-Schätzer aufbessert. Sicher ist das Konstrukt bei der PLS insgesamt rechenintensiver.

Wenn man lediglich an der Prognose interessiert ist, sollte die PLS- generell der KQ-Schätzung vorgezogen werden. Denn bei regulärem X^TX und vollständiger Faktoren-Extraktion ist für den PLS-Schätzer die Kovarianz genau wie beim KQ-Schätzer mit der Matrix $\hat{\sigma}^2\left(X^TX\right)^{-1}$ schätzbar. Das bedeutet auch, daß Formel (10) ihre Gültigkeit behält. Numerisch effizienter ist die Umsetzung mithilfe von (41) zu bewerkstelligen. Bei vollständiger Extraktion entspricht Fehlermatrix E einer Nullmatrix. Benötigt wird der p-dimensionale Vektor der Mittelwerte von X.

$$X^TX \;=\; \left(\vec{1}\,\vec{\bar{x}}^{\,T} + TP^T\right)^T\left(\vec{1}\,\vec{\bar{x}}^{\,T} + TP^T\right) \;=\; \vec{\bar{x}}\,\vec{1}^{\,T}\vec{1}\,\vec{\bar{x}}^{\,T} + PT^TTP^T \tag{48}$$

Die beiden nicht abgebildeten gemischten Produkt-Terme verschwinden, weil das Produkt der Mittelwerte mit T einer gewichteten Summe entspricht, wobei die Zentrierung von T deren Verschiebung zu Null bewirkt.
Für singuläres X^TX ist keine Kovarianz-Matrix (benötigt zweites Moment) berechenbar, während der Erwartungswert (erstes Moment) weiterhin existiert. Eine vollständig extrahierte PLS bewahrt demnach die Eigenschaften einer multiplen KQ-Schätzung und stellt im singulären Fall immerhin eine Alternative zur Verfügung.

PLS-Regression vs. dem auf der MPI basierenden KQ-Schätzer im Fall $n < p$
In Kapitel 4.5 ist in (38) der KQ-Schätzer notiert, welcher unter Anwendung der Moore-Penrose-Inverse (MPI) erhalten werden kann. Bei der PLS wird auf die Hauptachsenmatrix T gearbeitet. Diese Hauptachsen sind zentriert. Werden die, von einer evtl. Scheinvariable bereinigten, Einflußgrößen X zentriert und darauf Formel (38) angewendet, so erhält man $\vec{\hat{\beta}}_+ = \left(X^TKX\right)^+X^TK\vec{y}$, worauf automatisch die Zielgröße $\vec{y}$ einzentriert wird. K entspricht der Zentrierungsmatrix (20) von Seite 22. Die Lösungen dieser Koeffizienten sind identisch zu (46.3) und der Achsenabschnitt ist äquivalent zu (46.4) zu erhalten. Diese Erkenntnis ist allerdings interessant, denn sie erlaubt eine alternative, u.U. intuitive Darstellung der vollständig extrahierten PLSR, falls an den Hauptachsen T ansonsten kein weiteres Interesse besteht.
Unter gewissen Voraussetzungen ($\lambda \to 0\,, D = I_p$), welche abschließend in Kapitel 4.5 erwähnt sind, läßt sich die Brücke von der Ridge-Regression hin zur PLSR schlagen.

Anmerkung:
Mit einem MPI basierten KQ-Schätzer wird die Komplexität, d.h. die Monitor-Funktion der PLSR, bereits für den univariaten Fall nicht erreicht. Der Aufwand der Zerlegung der PLS wird in Kapitel 5.2 nochmals erhöht.

5.1.2 Krylov-Sequenz

Alternativ zum NIPALS-Algorithmus kann die PLS über eine Krylov-Sequenz[36] ausgeführt werden. Diese Methode weist, im Vergleich zu NIPALS, Vor- und Nachteile auf. Im weiteren Text wird darauf kurz eingegangen.

Die Krylov-Sequenz stellt einen analytischen Zugang[37] zum Erhalten der Regressionskoeffizienten dar. Allerdings stellt sie keine PLS-Faktoren zur Verfügung.

Es gilt $m = \text{Rang}\left(X^TKX\right)$. Die Krylov-Sequenz K_g $(p \times g)$, mit der Breite $g \leq m$, ist definiert:

$$K_g = \left(X^TK\vec{y}\ ,\ X^TKXX^TK\vec{y}\ ,\ \left(X^TKX\right)^2X^TK\vec{y}\ ,\ \ldots\ ,\ \left(X^TKX\right)^{g-1}X^TK\vec{y}\right), \tag{49}$$

wobei g festzulegen ist. K entspricht der Zentrierungsmatrix (20) und g beschreibt die Zahl extrahierter Faktoren.

Das Potenzieren des Kreuzproduktes ist aus computationaler Sicht mit Risiken verbunden: Die Potenzen können bei umfangreicher Datenbreite p für hinreichend großes g extreme Wertebereiche erreichen, wodurch die numerische Genauigkeit eingeschränkt sein kann; besonders problematisch bei Invertierungen.

(Phatak, Reilly & Penlidis, 2002) definieren die Hat-Matrix, welche über einen Rang von g verfügt:

$$H_g = K_g\left(K_g^TX^TKXK_g\right)^{-1}K_g^T\ ,\ (p \times p) \tag{49.1}$$

und es resultiert als Schätzer der Regressionskoeffizienten

$$\vec{\hat{\beta}}_g = H_gX^TK\vec{y}\ ,\ (p \times 1). \tag{49.2}$$

Für den Achsenabschnitt β_0 erhält man genau wie in (46.4):

$$\hat{\beta}_0 = \frac{1}{n}\vec{1}^T\left(\vec{y} - X\vec{\hat{\beta}}_g\right)\ ,\ (1 \times 1). \tag{49.3}$$

Die Prognosefunktion ist äquivalent zu (46.5) zu erhalten:

$$\vec{\hat{y}} = \hat{\beta}_0 + X\vec{\hat{\beta}}_g\ ,\ (n \times 1). \tag{49.4}$$

Diskussion

$g = 1: \quad K_1 \overset{(49)}{=} X^TK\vec{y} \quad \Leftrightarrow \quad H_1 \overset{(49.1)}{=} X^TK\vec{y}\left(\vec{y}^TKXX^TKXX^TK\vec{y}\right)^{-1}\vec{y}^TKX$

Der Ausdruck der Inversen entspricht einem Skalar und deshalb ist $H_1 \propto X^TK\vec{y}\,\vec{y}^TKX$. Diese Form der Hat-Matrix ist, bis auf einen abweichenden Faktor, proportional zu (40) von Seite 37, der quadrierten Kovarianz von X mit $\vec{y}$. Bzgl. H_1 verfügt man über diejenige Matrix, welche die maximierte Kovarianz auf den ersten PLS-Faktor projiziert.

$g = 2: \quad K_2 = \left(X^TK\vec{y}\ ,\ X^TKXX^TK\vec{y}\right)$

Die Substitutionen, proportional zur Kovarianz (23), $S = X^TKX$ bzw. für die Kovarianz X mit $\vec{y}$, proportional zu (39), mit $\vec{c} = X^TK\vec{y}$ angewandt, ergibt: $K_2 = (\vec{c}\ ,\ S\vec{c}) \quad \Leftrightarrow$

$$H_2 = (\vec{c}\ ,\ S\vec{c})\left[\begin{pmatrix}\vec{c}^T\\ \vec{c}^TS\end{pmatrix}S\,(\vec{c}\ ,\ S\vec{c})\right]^{-1}\begin{pmatrix}\vec{y}^T\\ \vec{c}^TS\end{pmatrix}.$$

[36] Der Begriff ist nicht eindeutig gefaßt. Eine geeignete, kurze Definition findet man in (Stoer & Bulirsch, 1978, Seite 10).

[37] Die Details für einen Lösungsansatz lassen sich im Skript krylov.m nachvollziehen.

Im ersten Schritt die Lösung der Inverse in H_2:

$$\left[\begin{pmatrix}\vec{c}^{\,T}\\ \vec{c}^{\,T}S\end{pmatrix} S\,(\vec{c}\ ,\ S\vec{c})\right]^{-1} = \begin{pmatrix}\vec{c}^{\,T}S\vec{c} & \vec{c}^{\,T}S^2\vec{c}\\ \vec{c}^{\,T}S^2\vec{c} & \vec{c}^{\,T}S^3\vec{c}\end{pmatrix}^{-1} = \left(\vec{c}^{\,T}S\vec{c}\vec{c}^{\,T}S^3\vec{c} - \vec{c}^{\,T}S^2\vec{c}\vec{c}^{\,T}S^2\vec{c}\right)^{-1}\begin{pmatrix}\vec{c}^{\,T}S^3\vec{c} & -\vec{c}^{\,T}S^2\vec{c}\\ -\vec{c}^{\,T}S^2\vec{c} & \vec{c}^{\,T}S\vec{c}\end{pmatrix} =$$

$\left[\vec{c}^{\,T}S\left(\vec{c}\vec{c}^{\,T}S - S\vec{c}\vec{c}^{\,T}\right)S^2\vec{c}\right]^{-1}\begin{pmatrix}\vec{c}^{\,T}S\\ -\vec{c}^{\,T}\end{pmatrix}S(S\vec{c}\ ,\ -\vec{c})$. Der inverse Vorfaktor ist skalar.

Im zweiten Schritt die eigentliche Berechnung:

$$\begin{aligned}H_2 &= \left[\vec{c}^{\,T}S\left(\vec{c}\vec{c}^{\,T}S - S\vec{c}\vec{c}^{\,T}\right)S^2\vec{c}\right]^{-1}(\vec{c}\ ,\ S\vec{c})\begin{pmatrix}\vec{c}^{\,T}S\\ -\vec{c}^{\,T}\end{pmatrix}S(S\vec{c}\ ,\ -\vec{c})\begin{pmatrix}\vec{c}^{\,T}\\ \vec{c}^{\,T}S\end{pmatrix}\\ &= \frac{\left(\vec{c}\vec{c}^{\,T}S - S\vec{c}\vec{c}^{\,T}\right)S\left(S\vec{c}\vec{c}^{\,T} - \vec{c}\vec{c}^{\,T}S\right)}{\vec{c}^{\,T}S(\vec{c}\vec{c}^{\,T}S - S\vec{c}\vec{c}^{\,T})S^2\vec{c}}.\end{aligned}$$

Die Differenzen der Matrizen-Terme $S\vec{c}\vec{c}^{\,T}$, $\vec{c}\vec{c}^{\,T}S$ bewirken lediglich Null-Haupdiagonalen. H_2 projiziert die maximierte Kovarianz auf die ersten beiden Faktoren. Die Information wächst u.U. an. D.h., $H_2 - H_1$ kann nicht negativ definit sein.

$g > 2$: Wenn auch die Komplexität des Ausdrucks mit steigendem g überlinear ansteigt – das Problem bleibt in geschlossener Form, also analytisch beschreibbar.

Anmerkungen:
Der Begriff *Hat-Matrix* für H_g in (49.1) ist ungewohnt. Dabei handelt es sich um eine Projektionsmatrix, basierend auf der Krylov-Sequenz. Die Inverse in (49.1) existiert stets, denn deren Dimension beträgt $g \times g$. Es gilt weiterhin: Rang $\left(X^TKX\right) = m \geq g$. Aufgrund der Substitution $\vec{c} = X^TK\vec{y}$ kann die rechte Seite in (49.2) zu $H_g\,\vec{c}$ vereinfacht werden.
H_g kann in eine andere Projektionsmatrix, in Anlehnung zu H in (6) auf Seite 13, transformiert werden. Die Hat-Matrix des klassischen Modells lautet $H = X\left(X^TX\right)^{-1}X^T$.
Assoziiert man X mit KXK_g, folgt für die Gestalt dieser Hat-Matrix:

$$H_{kry} = KXK_g\left(K_g^TX^TKXK_g\right)^{-1}K_g^TX^TK \overset{(49.1)}{=} KXH_gX^TK\ ,\ (n \times n). \tag{50}$$

H_{kry} ist allerdings nicht direkt mit der Projektionsmatrix H aus Kapitel 3.1 vergleichbar, da im klassischen Modell ausschließlich die Designmatrix X – mit Scheinvariable – ausgewertet wird. In H_{kry} geht $\vec{y}$ mit ein. Für den Grenzfall $g \overset{!}{=} m$ gilt für die Projektionsmatrix: $H_{kry} = K$, die idempotente Zentrierungsmatrix resultiert.
Desweiteren wird mit $H\vec{y}$ die Prognose $\hat{\vec{y}}$ berechnet. Wegen des fehlenden Absolutgliedes in X beim Aufbau von K_g bekommt man mit $H_{kry}\vec{y}$ die zentrierte Prognose $K\hat{\vec{y}}$ zurück, bzw. $\hat{\vec{y}} = H_{kry}\vec{y} + \bar{y}$.

Extraktion der benötigten Faktorenzahl mithilfe des GCV-Kriteriums
Mit H_{kry} eröffnet sich eine zusätzliche Option: Die generalisierte Kreuzvalidierung (GCV) stellt eine Approximation der einfachen Kreuzvalidierung dar. Statt in n Schritten die Kreuzvalidierung für eine Ausprägung k durchzuführen, kann über eine Rechenvorschrift quasi in einem Schritt die Berechnung pro k erfolgen. Bei (Heumann & Schmid, 2016, Kap. 7.5) wird das GCV-Kriterium auf Seite 114 definiert. Angepaßt auf die hiesige Fragestellung:

$$\text{GCV} = \frac{1}{n}\left(\vec{y} - \hat{\vec{y}}\right)^T\left(\vec{y} - \hat{\vec{y}}\right) / \left(1 - \text{Spur}\,(H_{kry})/n\right)^2 \rightarrow \min_{k=1,\,\ldots\,,g}. \tag{51}$$

Die Formel ist auch für den Fall $n < p$ anwendbar, da die Inverse in (50) stets existiert, welches weiter oben bereits für (49.1) angemerkt wurde. Zu beachten ist, daß bei $g \stackrel{!}{=} m$ extrahierten Faktoren das GCV für $k = g$ auf Null absinken würde. Dieser Grenzfall ist i.d.R. uninteressant und auszuschließen, bei der Suche nach dem globalen Minimum.

5.1.3 Kovarianzmatrix der Parameter bei nicht vollständiger Extraktion

I.d.R. wird das Augenmerk auf einer möglichst sparsamen Faktoren-Extraktion, also einer gewollt Verlust behafteten Schätzung, liegen. Dann wäre die Schätzung der Kovarianzmatrix mittels $\hat{\sigma}^2\left(X^TX\right)^{-1}$ zu optimistisch, da dies einer vollständigen Faktoren-Lösung äquivalent ist. Für eine unvollständige Faktoren-Extraktion beschreiben (Phatak et al., 2002), bei Vorliegen der Voraussetzungen Spalten-Vollrang bzgl. KX – d.h. $n > p$, eine Methodik zum Erhalten einer adäquaten Kovarianzmatrix. Ihr Ansatz beruht auf der Definition der Krylov-Sequenz, deren Breite zu Beginn voreingestellt wird. Auf Seite 248 definieren die Autoren eine Ableitungsmatrix, die Jacobi-Matrix J der Dimension $(p \times n)$:

$$J \equiv \frac{\partial \vec{\hat{\beta}}_g(\vec{y})}{\partial \vec{y}^T} = \left(\vec{c}^T \otimes I_p\right)(I_{p^2} + C)\left[K_g\left(K_g^T S K_g\right)^{-1} \otimes (I_p - H_g S)\right]U_g^T + H_g X^T K. \quad (52)$$

Das Symbol $\otimes$ beschreibt das Kronecker-Produkt[38]. Die Matrix U_g wird ähnlich wie K_g – rekursiv – aufgebaut, indem $U_g = (KX\ ,\ KXS\ ,\ KXS^2\ ,\ \ldots\ ,\ KXS^{g-1})\ ,\ (n \times pg)$.
Die quadratische Matrix C $(p^2 \times p^2)$ besteht aus Nullen und enthält in allen Zeilen/Spalten jeweils eine Eins. Sie entspricht einer speziellen Vertauschungsmatrix (commutation matrix).
Beispiel: Eine quadratische Matrix, der Dimension p, wird vektorisiert: $\vec{a} = \mathrm{vec}\,(A)$. Nun stehen in $\vec{a}$ $(p^2 \times 1)$ die aufeinander abwärts gestapelten Spalten von A. Angenommen, es ist die Vektorisierung der transponierten Matrix A^T gewünscht – die Operation ist direkt auf den Vektor $\vec{a}$ anwendbar: $\vec{b} = C\vec{a}$ $(p^2 \times 1)$.

Der Sinn hinter der Umschichtung eines Vektors, statt die Vektorisierung auf die Transponierte auszuführen, kann auch darin begründet sein, daß Transpositionen und Vektorisierungen sehr großer Matrizen viel Rechenzeit beanspruchen. Hingegen ist der Aufbau einer Vertauschungsmatrix C vergleichsweise schnell erledigt: In der anfänglichen Nullmatrix müssen lediglich, nach einem genau festgelegtem Muster (i.d.R. rekursiv), p^2 Einsen einbeschrieben werden. Einen umarrangierten neuen Vektor zu erzeugen, der aus dem ursprünglichen Vektor hervorgeht, benötigt weniger Zeit. Eine sehr ausführliche Einführung und Beschreibung zur *commutation matrix* geben (Magnus & Neudecker, 1979).

Die Krylov-Sequenz K_g enthält die Zielgröße $\vec{y}$. Über den Ausdruck $\left(K_g^T S K_g\right)^{-1}$ geht $\vec{y}$ nichtlinear in die Hat-Matrix H_g (49.1) und gemeinsam mit H_g in die Jacobi-Matrix J (52) ein. Für (Phatak et al., 2002) ist die Nichtlinearität von $\vec{y}$ ein Grund, die Zahl der Freiheitsgrade, statt $d = n - g$, abweichend zu definieren:

$$d = \mathrm{Spur}\left(\left(I_n - J^T X^T K\right)(I_n - KXJ)\right),\ d \in \mathbb{R}^+. \quad (52.1)$$

Bei unvollständiger Zerlegung wird die Spur die Differenz oft übersteigen. Das wird gern

[38] http://www.spektrum.de/lexikon/physik/kronecker-produkt/8573 (3.12.2017)

gesehen: Eine größere Zahl an Freiheitsgraden schmälert die Residuenvarianz. Bei vollständiger Faktoren-Extraktion, d.h. bei $g \stackrel{!}{=} m$, entspricht die Definition für d mittels der Spur gleich der Differenz. Im Übrigen gilt dann $H_g = JJ^T$.
Mit den Residuen $\vec{\hat{\epsilon}} = \vec{y} - \vec{\hat{y}}$ kann – in Anlehnung zu (8) auf Seite 13 – die Residuenvarianz des Modells geschätzt werden:

$$\hat{\sigma}^2 = \frac{\vec{\hat{\epsilon}}^{\,T}\vec{\hat{\epsilon}}}{d-1} \tag{52.2}$$

und daraus die Kovarianz-Matrix der PLS-Regressionskoeffizienten ermittelt werden:

$$\widehat{\mathrm{cov}}\left(\vec{\hat{\beta}}_g\right) = \hat{\sigma}^2 JJ^T,\ (p \times p). \tag{52.3}$$

Jetzt ist das zweite Moment, und damit die Varianz, verfügbar. In dieser angenehmen Situation ist der Weg für die analytische Inferenz (Inferenz $\widehat{=}$ Schlußfolgerung) offen, allerdings mit den Restriktionen $n > p$ und für den Rang $m = p$.

Falls zusätzlich eine vollständige Faktoren-Extraktion ($g \stackrel{!}{=} m$) und als Designmatrix $Z = \left(\vec{1}\,, X\right)$ vorliegt, gilt der Zusammenhang $\left(Z^TZ\right)^{-1}[2\!:\!(p+1)\ ,\ 2\!:\!(p+1)] = JJ^T$.
Wegen des fehlenden Absolutgliedes $\hat{\beta}_0$ kann JJ^T nur eine Teilmatrix von $\left(Z^TZ\right)^{-1}$ sein.

Für $n \leq p$ verfügt JJ^T $(p \times p)$ über keinen Vollrang mehr, welches eine impropere Kovarianzmatrix zur Folge hat. Während beim verzerrten linearen Modell in (15) auf Seite 16 der Nenner ohnehin Null bzw. negativ sein würde, ist das hier anders. Der Nenner $d-1$ in (52.2) kann weiterhin positiv ausfallen. Dann wäre in (52) nur JJ^T problematisch. Auch wenn eine singuläre Matrix JJ^T vorliegt, eine Invertierung ist jedoch nicht erforderlich. Überlegenswert wäre dann, inwieweit die Varianzschätzung noch als zuverlässig gelten kann. Vermutlich wird mit zunehmend überwiegendem p, besonders für $n \ll p$, die Verzerrung stark zunehmen.

5.2 PLS auf eine multivariate Zielgröße (PLS2)

Dieses Kapitel versteht sich als Ergänzung zur PLS1. Augenmerk gilt hier den Änderungen und Erweiterungen.

Festlegungen:
Mit der Matrix C ist in der PLS2 nicht die Kovarianzmatrix (23) von Seite 24 gemeint. In (53) beschreibt $\widetilde{F}$ die Fehlermatrix der Zielgrößen-Zerlegung und ist nicht mit den Faktoren F zu assoziieren. Außerdem beschreibt $\underline{B}$ die Zielgrößen-Ladungsmatrix und hat keinen Bezug zur Koeffizientenmatrix B der PLS-Regression.

Verfügt man über ein ganzes Bündel an Zielgrößen, mindestens jedoch über zwei Stück, wird die Extraktionsmethode, die eine Zielgrößenmatrix Y zerlegt, PLS2 genannt. Man unterscheidet nach der äußeren und inneren Beziehung. Die äußere Beziehung ist in der PLS1 durch (41) beschrieben. Sie ist in ihrer Endform; unter Beachtung von (42), (43):

$$\begin{aligned} X &= \vec{1}\vec{\bar{x}}^T + TP^T + E = \vec{1}\vec{\bar{x}}^T + \left(T\Lambda^{-1/2}\right)\left(\Lambda^{1/2}P^T\right) + E = \vec{1}\vec{\bar{x}}^T + FA^T + E \\ Y &= \vec{1}\vec{\bar{y}}^T + TCQ^T + \widetilde{F} = \vec{1}\vec{\bar{y}}^T + \left(T\Lambda^{-1/2}\right)\left(\Lambda^{1/2}CQ^T\right) + \widetilde{F} = \vec{1}\vec{\bar{y}}^T + F\underline{B}^T + \widetilde{F}. \end{aligned} \tag{53}$$

Eine Diagonalmatrix C[39], welche aus der inneren Beziehung hervorgeht, ist hier separat ausgewiesen. Matrix Y besitzt die Dimension $n \times q$. Q entspricht einer $q \times g$ Matrix, deren g Spalten Einheitsvektoren entsprechen. Mit dem Matrixaufruf $QC\Lambda CQ^T$ bzw. $\underline{B}\underline{B}^T$ läßt sich die geschätzte Kovarianzmatrix von Y beschreiben.
Die innere Beziehung verbindet die latenten Datenräume T, U miteinander:

$$\begin{aligned} KX &= TP^T + E \\ KY &= UQ^T + Y_{err}, \end{aligned} \tag{53.1}$$

indem

$$U = TC + U_{err} \tag{53.2}$$

als ein multivariates Regressionsproblem verstanden wird, bei dem C zu schätzen ist. Den Regressionsfehler fängt Matrix U_{err} auf. T und U haben dieselbe Dimension, wobei U nicht orthogonal ist.
Die Erwartungswerte der Fehlermatrizen E, Y_{err} und U_{err} entsprechen Nullmatrizen, bei vollständiger Zerlegung gilt es für Matrix E auch empirisch. Das Kreuzprodukt $Y_{err}^T E$ soll approximativ einer Nullmatrix entsprechen, d.h. als Annahme steht Unkorreliertheit in den Fehlern.
Setzt man Gleichung (53.2) in (53.1) ein, resultiert: $KY = \left(TC^T + U_{err}\right) Q^T + Y_{err}$.
Bei einer vollständigen Zerlegung gilt $U_{err}Q^T + Y_{err} = 0$ und daraus folgt: $KY = TCQ^T$. Das ist konform zu (53).

Nur bei vollständiger Zerlegung könnte durch Anwenden von folgendem KQ-Schätzansatz Matrix C und darauf aufbauend U_{err} sinnvoll berechnet werden:

$$\hat{C} = \left(T^T T\right)^{-1} T^T U.$$

[39] Im R PLS-Paket v2.5 von Mevik & Wehrens wird die Separation unterdrückt: $Q \stackrel{!}{=} QC$.

Die Hauptachsenmatrix T spannt einen orthogonalen Raum auf und deshalb ist die Inverse besonders einfach mit den Kehrwerten der Diagonale von T^TT abzubilden. $\hat{C}$ entspricht zunächst einer unteren Dreiecksmatrix. Allerdings wäre die KQ-Lösung damit so vollumfänglich, daß U_{err} in (53.2) verschwindet. Zerlegt man im Anschluß $\hat{C}$ in eine Diagonalmatrix C plus eine untere Dreiecksmatrix C_u, gelingt die Berechnung. D.h.,

$$\hat{C} \stackrel{!}{=} \text{diag}\left(\hat{C}\right) + C_u = C + C_u.$$

Somit kann der Regressionsfehler $U_{err} = TC_u$ ermittelt werden und aufgrund der Dreieckstruktur von C_u wird nachvollziehbar, daß in der Matrix U_{err} die letzte Spalte bei vollständiger Zerlegung stets Null beträgt.

Für Λ gilt derselbe Zusammenhang, wie bereits in (42) vorgetragen. Die Anteilswerte der durch die Faktoren aufgeklärten Varianz für (X, Y) betragen:

$$\begin{aligned} \text{R}^2_X &= \frac{\text{diag}\left(A^TA\right)}{\text{Spur}\left(\widehat{\mathbb{V}}(X)\right)} \stackrel{(53)}{=} \frac{\text{diag}\left(\Lambda^{1/2}P^TP\Lambda^{1/2}\right)}{\text{Spur}\left(\widehat{\mathbb{V}}(X)\right)} = \frac{\text{diag}\left(\Lambda P^TP\right)}{\text{Spur}\left(\widehat{\mathbb{V}}(X)\right)} \stackrel{(42)}{=} \frac{\text{diag}\left(T^TTP^TP\right)}{\text{Spur}\left(X^TKX\right)} \\ \text{R}^2_Y &= \frac{\text{diag}\left(\underline{B}^T\underline{B}\right)}{\text{Spur}\left(\widehat{\mathbb{V}}(Y)\right)} \stackrel{(53)}{=} \frac{\text{diag}\left(\Lambda^{1/2}CQ^TQC\Lambda^{1/2}\right)}{\text{Spur}\left(\widehat{\mathbb{V}}(Y)\right)} \stackrel{(42)}{=} \frac{\text{diag}\left(T^TTCQ^TQC\right)}{\text{Spur}\left(Y^TKY\right)}. \end{aligned} \tag{54}$$

In (Jørgensen & Goegebeur, 2007b) wird der PLS2-Algorithmus beschrieben. Im Vergleich zur PLS1 steigt seine Komplexität weiter an.

Schematische Beschreibung der Dekomposition[40] mithilfe des NIPALS-Algorithmus:
– erzeuge temporäre zentrierte Matrizen $E = KX;\ \widetilde{F} = KY$
– berechne $m = \text{Rang}\,(E)$
– wähle Zahl g zu extrahierender Faktoren, unter der Nebenbedingung $g \leq m$
– erzeuge leere Matrizen W, T, P, Q, U und leeren Vektor $\vec{C}$
– initialisiere Vektor $\vec{u}$ mit der ersten Spalte von $\widetilde{F}$ *%erste Spalte von $\widetilde{F}$ immer verfügbar*
– initialisiere Laufvariable $k = 0$
– wiederhole folgende Operationen solange Iterationszahl $k < g$

- $k \stackrel{!}{=} k + 1$ *%Nr. des Faktors*
- iteriere solange bis Veränderung in $\vec{u}$ kleiner einem vorgegebenen Fehler
  - $\vec{w} = E^T\vec{u}\,;\ \ \vec{w} = \dfrac{\vec{w}}{\sqrt{\vec{w}^T\vec{w}}}$ *%Ladungsgewichte/Eigenvektoren $(p \times 1)$ für Faktor k*
  - $\vec{t} = E\vec{w}$ *%k. Hauptachse $(n \times 1)$*
  - $\vec{q} = \widetilde{F}^T\vec{t}\,;\ \ \vec{q} = \dfrac{\vec{q}}{\sqrt{\vec{q}^T\vec{q}}}$ *%Gewichte $(q \times 1)$ der äußeren Beziehung*
  - $\vec{u} = \widetilde{F}\vec{q}$ *%k. Zielgröße $(n \times 1)$ der inneren Beziehung*
- $c = \vec{t}^{\,T}\vec{u}\,/\left(\vec{t}^{\,T}\vec{t}\right)$ *%k. Regressionskoeffizient (1×1) für innere Beziehung*
- $\vec{p} = E^T\vec{t}\,/\left(\vec{t}^{\,T}\vec{t}\right)$ *%Ladung $(p \times 1)$*
- reihe $\vec{w}$ als neue Spalte in W, $\vec{t}$ in T, $\vec{q}$ in Q, $\vec{u}$ in U, c in $\vec{C}$ und $\vec{p}$ in P ein
- $E \stackrel{!}{=} E - \vec{t}\,\vec{p}^{\,T}$ *%Residuen $(n \times p)$ für KX bei k Faktoren*

[40] Die Details für einen numerischen Lösungsansatz lassen sich im Skript pls2.m nachvollziehen.

- $\widetilde{F} \stackrel{!}{=} \widetilde{F} - \vec{t}\, c\, \vec{q}^{\,T}$ *%Residuen* $(n \times q)$ *für KY bei k Faktoren*

– $C = \operatorname{diag}\left(\vec{C}\right)$ *%Regressionsmatrix* $(g \times g)$

Bei der Initialisierung von $\vec{u}$ kann an sich eine beliebige Spalte von Y ausgewählt werden. Denn mithilfe der inneren Schleife im Algorithmus wird die Verzerrung von $\vec{u}$ herausiteriert. Solange man die Faktoren-Extraktion nicht vorzeitig abbricht, d.h. $g \stackrel{!}{=} m$, entsprechen die Residualmatrizen E und $\widetilde{F}$ am Prozedurende – bis auf numerische Unzulänglichkeiten – Nullmatrizen.
Der Regressions-Schätzer ist auf gleichem Wege wie bei der PLS1 (46.3) auf Seite 41 zu erhalten. Die Endform ohne Herleitung:

$$\hat{B}_{PLS} = W\left(P^T W\right)^{-1} C Q^T,\ (p \times q). \tag{55}$$

Der Achsenabschnitt (Kalibration) wird zu einem Vektor:

$$\vec{\hat{\beta}}_0 = \frac{1}{n}\vec{1}^{\,T}\left(Y - X\hat{B}_{PLS}\right)\ ,\ (1 \times q) \tag{55.1}$$

und für die Prognosefunktion resultiert:

$$\hat{Y} = \vec{1}^{\,T}\vec{\hat{\beta}}_0 + X\hat{B}_{PLS}\ ,\ (n \times q). \tag{55.2}$$

PLS-Regression als Verallgemeinerung der multivariaten linearen Regression
Die Aussagen sind äquivalent zu denen, die bei der PLS1 getroffen wurden. Der Unterschied besteht nur in der Zielgröße, die nun in einer Matrix Y multivariat vorliegt. In der Koeffizientenmatrix $\hat{B}$ befinden sich q Koeffizientenvektoren.
Die Kovarianzen für den PLS-Schätzer ergeben sich für die q Komponenten in einem Schritt: $\left(X^T X\right)^{-1}\left(\hat{\sigma}_1^2 I_p\ ,\ \ldots\ ,\ \hat{\sigma}_q^2 I_p\right)$. Für $X^T X$ gelten die gleichen Aussagen wie in (48). Solange das Kreuzprodukt regulär bleibt, bewahrt eine vollständig extrahierte PLS die Eigenschaften einer multivariaten KQ-Schätzung. Im singulären Fall steht immerhin der Punktschätzer $\hat{B}$ weiterhin zur Verfügung.

6 Anwendung

In Kapitel 2.2 wurde bereits kurz auf die Ausgangsdatenlage eingegangen. Erwähnt sei, daß Korrelationen zwischen Zielgrößen und Einflußgrößen im Mittel wesentlich geringer ausgeprägt sind, als in den Einflußgrößen selbst (vgl. Abb. 4 & 5, Seite 9 ff.), welche annähernd multikollinear vorliegen.
Für ein gutes Modell ist eine derartige Konstellation ungünstig. Einerseits soll nicht das gesamte Spektrum als Einflußgröße für lediglich zwei ausgewählte chemische Elemente erforderlich sein, andererseits sollen die Zielgrößen aber möglichst gut erklärt werden. Angenehmer wäre die Situation, wenn die Einflußgrößen untereinander so wenig wie nötig korreliert sind, hingegen die Zielgrößen zu den relevanten Spalten des Spektrums hohe Korrelationen aufweisen. Solche bedeutsamen Spalten der Einflußgrößen werden in der analytischen Chemie *Banden* genannt.
In Anlehnung an (Drechsler, 2018) wird die Priorität für das Zustandekommen eines Modells zuerst unter Ignorieren der Zielgröße gesetzt. Ansonsten könnte man der Vesuchung erliegen, auf die Zielgröße „hin zu optimieren“. Beispielsweise kann ein derart überangepaßtes Modell die vorliegenden Daten mit sehr kleinem Fehler beschreiben, aber für die Prognose völlig ungeeignet sein. Das andere Extrem wäre dann der umgekehrte Sachverhalt.
Dem Problem der Nichtidentifizierbarkeit, in den Einflußgrößen X (180×2153), wird formal mit der PLS begegnet; welche durch Multikollinearität nicht beeinträchtigt wird.

6.1 Datenvorbehandlung

Spektraldaten sollten wegen ihrer Eigenheiten nicht roh in die Auswertung eingehen. Für ein aussagekräftiges Modell ist es wichtig, die relevanten Bereiche des Spektrums zu erfahren. Entscheidende Hinweise in den Spektren geben die sichtbar ausgeprägten Resonanzspitzen in den Kurven zurück. Überlagerungen sind nach Augenmaß schlimmstenfalls überhaupt nicht erkennbar. (Tillmann, 1996) geht auf Seite 16/17 kurz auf die *Multiplicative scatter correction* (MSC) ein. Es handelt sich dabei um eine Mittelwertbereinigung in den Einflußgrößen der Form

$$X_{MSC} = \vec{1}^T X \vec{1} \,/\, (np) + XK, \tag{56}$$

wobei in $\vec{1}^T X \vec{1}$ der linke Vektor aus $n = 180$ Einsen besteht und der rechte Vektor über $p = 2153$ Einträge verfügt. Aufgrund des Produktes im Divisor werden beide Vektoren zu Einheitsvektoren normiert, mit denen der globale Mittelwert von X berechnet wird, welcher nun als Intercept der Kurven fungieren kann. Denn das Zentrieren der Einflußgrößen mit einer Zentrierungsmatrix K (2153×2153) von rechts bewirkt deren Verschiebung zu Null. Mithilfe des Intercepts werden sie in einer definierten mittleren Position ausgerichtet. Es entspricht insgesamt einer Streulichtkorrektur. So werden durch Formel (56) alle Kurven bestmöglich ineinander verschoben, welches ihre Varianz zueinander reduziert. Kurven, die aufgrund einer weit entfernten Lage unter Ausreißerverdacht stehen, sind hiermit gleichfalls begradigt. In der nächsten Abbildung befinden sich die vier Kurven,

wie in Abb. 2 gezeigt, unverändert bzgl. ihrer Form:

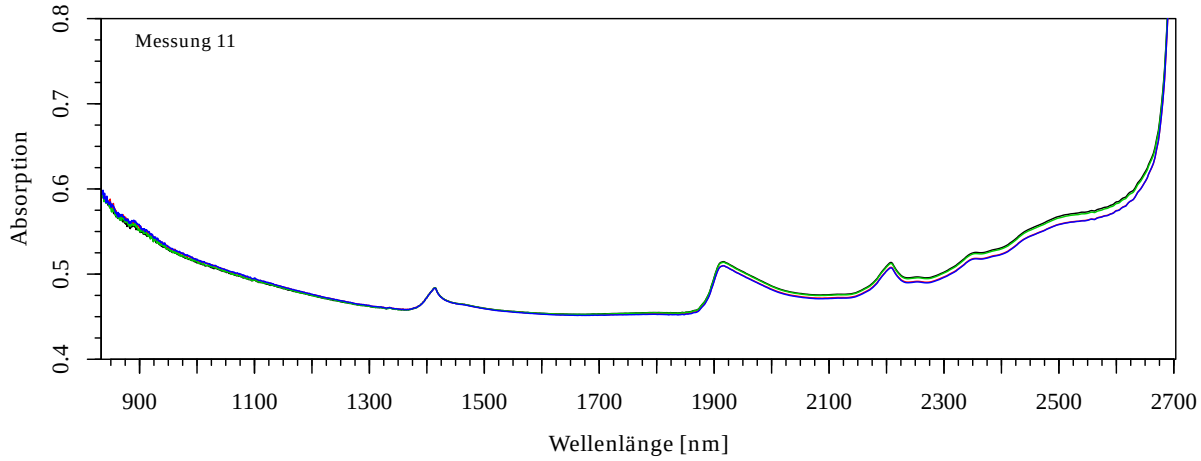


Abb. 12: MSC auf Spektren in Meßwiederholung.

Offenbar liegt den Einflußgrößen ein überlinearer Zusammenhang zugrunde, dem Anschein nach mit quadratischem Anteil, wenn von den Wellenlängen oberhalb 2650 nm abgesehen wird. Ein quadratisches Ausgleichspolynom $\widehat{f(\vec{x})} = \hat{\beta}_0 + \hat{\beta}_1\vec{x} + \hat{\beta}_2\vec{x}^2$, auf das globale Mittel $\vec{x}$ des Spektrums X angewandt, führt zu signifikanten Koeffizienten. In schematischer Abbildung:

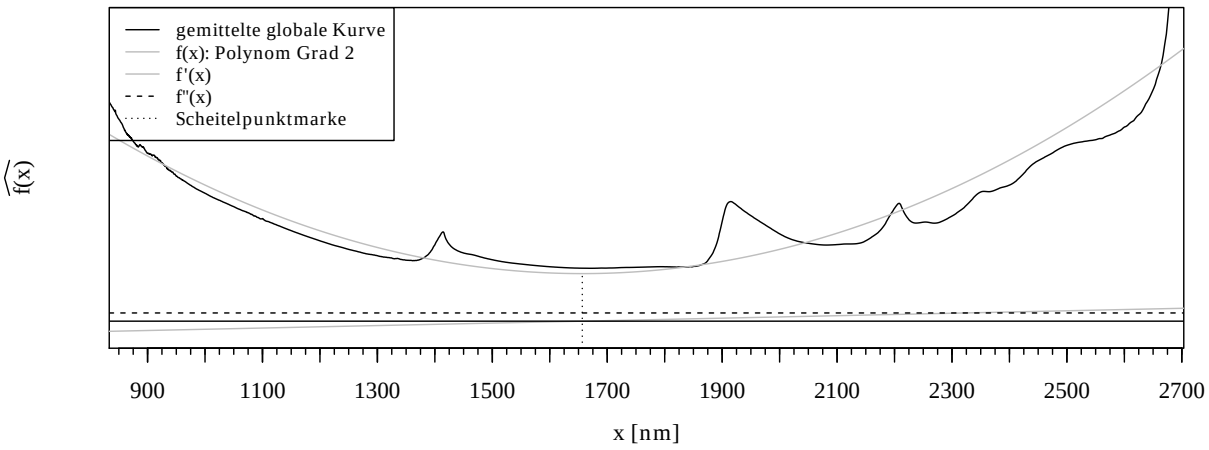


Abb. 13: Trendanalyse auf die Kurve des ausgemittelten Spektrums.

Der steile Anstieg der empirischen Kurve in den obersten Wellenlängen hat eine Hebelwirkung auf die Parabel und „zieht“ schließlich den rechten Parabelzug weitab vom Scheitel aus den Daten etwas heraus. Durch zweimaliges Differenzieren sollte es möglich sein, den allgemeinen Trend im Spektrum weitgehend zu entfernen, um die Resonanzpunkte hervorzuheben.
(Tillmann, 1996, Kapitel 1.2) beschreibt die *Derivativspektroskopie* zum Beheben von Streulichteffekten. Auf Seite 14 schreibt er: „Durch die Bildung von Ableitungen können auch feine Absorptionsbanden in dem Spektrum betont werden.“.
Die Ableitungen auf empirische Kurven sind rein numerisch und können mithilfe von Differenzenmatrizen – vgl. (Heumann & Schmid, 2016, Kap. 7.4.1, Seite 91) – vorgenommen werden. Die Differenzenmatrix erster Ordnung:

$$D_1 = \begin{pmatrix} -1 & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{pmatrix}, ((p-1) \times p).$$

Das numerische Differenzieren von X geschieht über den Aufruf $X D_1^T$, $(n \times (p-1))$. Dabei verliert man eine Spalte aus X.
(Rinnan, Van Der Berg & Engelsen, 2009, Kapitel 4, Seite 1213) geben bei Verwenden dieser Art der Ableitung einen Hinweis auf *Rauschen-Inflation*.

Über die erste Ableitung wird der Achsenabschnitt einer Kurve entfernt, welches ebenso einer Streulichtkorrektur – ähnlich der MSC – entspricht. Extrema werden zu Nullstellen, wie auch in Abb. 13 angedeutet.
Anhand der folgenden Abbildung wird ersichtlich, weshalb eine Glättung (hier: Savitzky & Golay) der Reihe empfehlenswert ist:

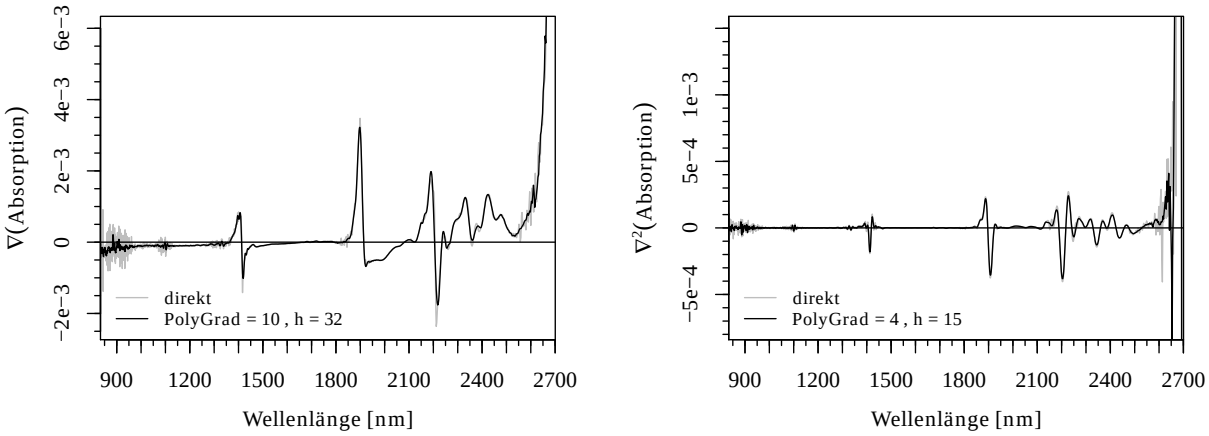


Abb. 14: zweimal differenzierte Kurve des ausgemittelten Spektrums.

Ungeglättet differenziert tritt das Rauschen, welches in der ursprünglichen Reihe nur wenig auffällt, stärker hervor. Die Glättung wiederum verursacht selbst Fehler, die das Signal an den informativen Punkten – im günstigsten Fall – aber nicht beeinträchtigen soll. Auf Seite 273 schreiben (Biener, Steinkämper, Masuch, Wolf & Hofmann, 2017), daß z.B. Acetat im Nahinfrarot eine zu geringe analytische Bandbreite im Spektralbereich aufweist.

Im Kapitel 3.4 in Abb. 8 auf Seite 20 wurde u.a. das Überschwingverhalten eines höheren Glättungs-Polynomgrades angedeutet. Diese an sich unerwünschte Eigenschaft wird für die empirische Reihe mit ihren scharfkantigen Resonanzspitzen vorteilhaft ausgenutzt: Das Rauschen maximal dämpfen, aber die Form der Peaks möglichst unverfälscht bewahren.

Bei Verwenden des Polynomgrades 10 und der Knotenzahl 32 ist die Kurve des gemittelten Spektrums (Abb. 13) im Vergleich zu ihrer Glättung nach Augenmaß defacto nicht mehr unterscheidbar. Anhand der linken Grafik in Abb. 14 werden die Auswirkungen des Glättens beim Differenzieren sichtbar. Für die darin rechts dargestellte zweite Ableitung wird die Reihe dann noch einmal nachgeglättet, allein um das Überschwingen im Rauschen zu begrenzen. Der obige Hinweis auf Rauschen-Inflation ist berechtigt.

Mithilfe der zweiten Ableitung wird die leicht steigende Tendenz (vgl. Abb. 13) der ersten Ableitung beseitigt und die Resonanzspitzen werden mit umgekehrtem Vorzeichen zurückgewonnen, welche auch teilweise in den Ausgangsdaten sichtbar waren:

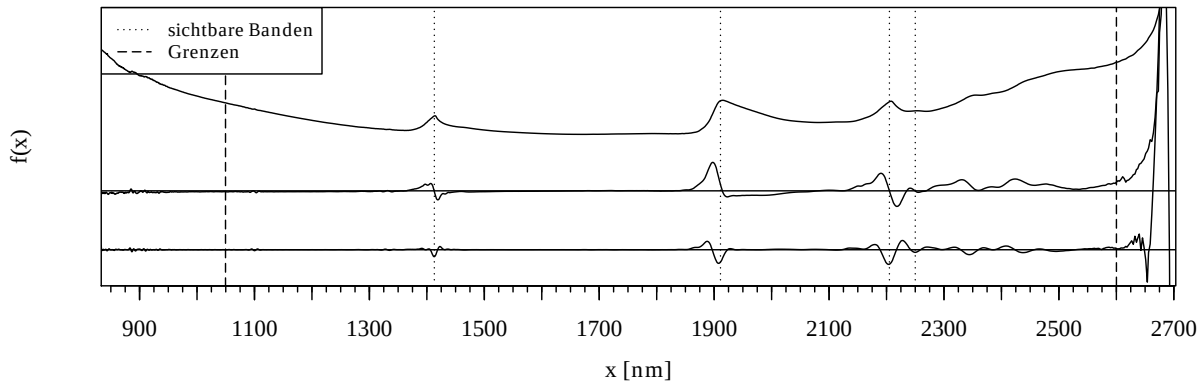


Abb. 15: ausgemitteltes Spektrum inkl. zweimalige Differentiation.

Markante Banden sind markiert, um die Wirkungsweise der Ableitungen zu illustrieren (siehe auch (Rinnan et al., 2009, Seite 1213, Figure 14)). Die eingezeichneten Grenzen stehen für eine Bandpaß-Charakteristik. Denn beim Auswerten der Standardabweichung auf alle 180 Spektren kommt es jenseits dieser Grenzen zu einem zunehmend steilen Anstieg der Streuung:

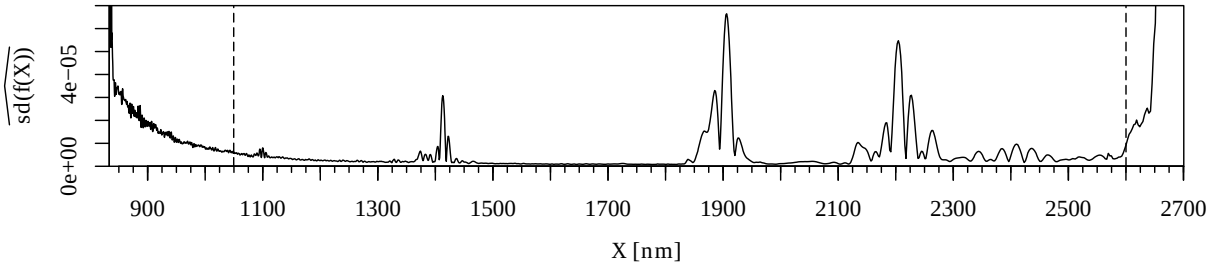


Abb. 16: geschätzte Standardabweichung der zweimal differenzierten, geglätteten Spektren.

Insbesondere jenseits der linken Grenze trägt das Spektrum keine nennenswerte Information mehr. Der Anstieg der Streuung ist vor allem dem Rauschen geschuldet.

Die Auswirkungen der Transformation des Spektrums auf die Korrelationen zur Zielgröße:

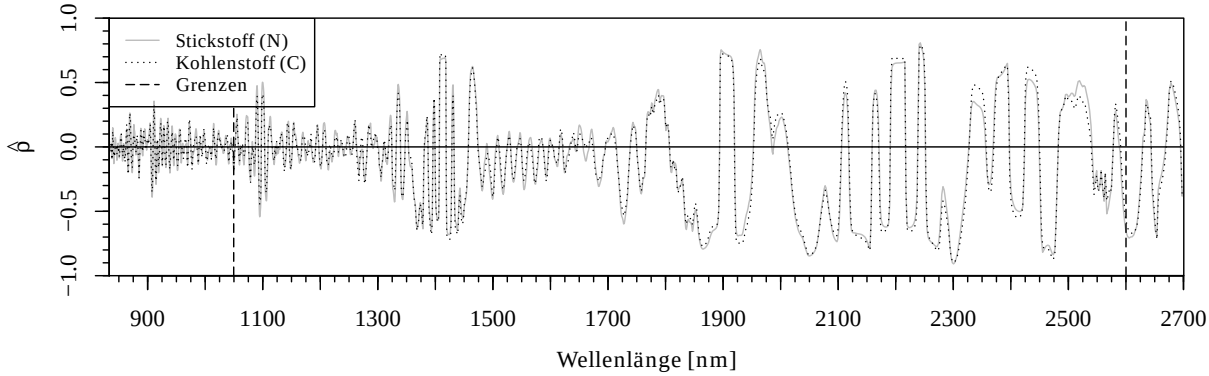


Abb. 17: Korrelationsfunktionen der Bodenproben {N,C} mit transformiertem Spektrum.

In schneller Abfolge findet ein Vorzeichenwechsel statt. Im Vergleich zu Abb. 5 auf Seite 9 haben sich die Korrelationen zum Rauschen (kurzwelliger Bereich jenseits der Grenze) verringert und in markant sichtbaren Banden (1413 1911 2205 2250) nm erhöht. Die Wirkung der Transformationen offenbart sich zumindest visuell als Zugewinn an Information.

Da Spektren auf den positiven reellen Zahlenbereich begrenzt sind, können sie optional auch mithilfe der logistischen Funktion[41] transformiert werden, die den Definitionsbereich beibehält.

$$\widetilde{X} \stackrel{!}{=} \frac{1}{(1+10^{\,aX})^2}\ ,\ a>0 \tag{57}$$

Die Idee bzw. Auswirkungen dieser nichtlinear bijektiven Transformation auf die Einflußgrößen, werden in den Kapiteln 6.3 – 6.5 diskutiert.

6.2 Dichte-Transformation

In Vorbereitung zu Kapitel 6.4 wird eine bijektive Transformation der Exponentialverteilung $F_\lambda(\vec{x}\,) = 1 - e^{-\lambda\vec{x}}$ vorgestellt. Der Träger der Dichte $f_\lambda(\vec{x}\,) = \lambda e^{-\lambda\vec{x}}$ soll hierzu quadriert werden. Dabei wird die neu entstehende Funktion $\tilde{f}_\lambda(\vec{x}\,)$ ebenfalls die Eigenschaften einer Dichte aufweisen.
Die Umkehrfunktion zu $\vec{y} = f_\lambda(\vec{x}\,)$, welche Abszisse mit Ordinate vertauscht, lautet $\vec{x} = f_\lambda^{-1}(\vec{x}\,) = -1/\lambda \ln\!\left(\frac{\vec{y}}{\lambda}\right)$. Durch elementweises Quadrieren wird die Transformation des Trägers vorbereitet und es resultiert $\vec{x} \stackrel{!}{=} \left(-1/\lambda \ln\!\left(\frac{\vec{y}}{\lambda}\right)\right)^2$.
Erneutes Invertieren führt auf eine vorläufige Form $f_\lambda^v(\vec{x}\,) = \lambda e^{-\lambda\sqrt{\vec{x}}}$, mit v als deren Synonym. Mithilfe der vorläufigen Form verfügt man über eine positive Funktion, bei welcher der Flächeninhalt aber noch zu Eins normiert werden muß.

Das uneigentliche Integral

$$a\int_0^\infty f_\lambda^v(\vec{x})d\vec{x} \;=\; a\int_0^\infty \lambda e^{-\lambda\sqrt{\vec{x}}}d\vec{x} \;=\; 1 \quad \text{mit}\ \ \vec{z}=\sqrt{\vec{x}}\ \ \Rightarrow\ \ \frac{d\vec{z}}{d\vec{x}} = \frac{1}{2\sqrt{\vec{x}}}\ \ \Leftrightarrow\ \ d\vec{x} = 2\vec{z}\,d\vec{z}$$

ist bzgl. a aufzulösen. D.h., das Einsetzen der Substitution $\vec{z} = \sqrt{\vec{x}}$ redefiniert das Integral

$$a\int_0^\infty \lambda e^{-\lambda\vec{z}}\,2\vec{z}\,d\vec{z} = 1,$$

welches sich in einem Produkt mit linearer Integrationsvariable bzgl. des Integranden offenbart, also durch partielle Integration analytisch gelöst werden kann:

$$a\int_0^\infty 2\vec{z}\,\lambda e^{-\lambda\vec{z}}\,d\vec{z} = a\left(-2\,\vec{z}\,e^{-\lambda\vec{z}} + \int_0^\infty 2e^{-\lambda\vec{z}}\,d\vec{z}\right) = -2a\left(\vec{z}\,e^{-\lambda\vec{z}} + 1/\lambda e^{-\lambda\vec{z}}\right) = -2a(\vec{z}+1/\lambda)\,e^{-\lambda\vec{z}}$$

und mittels Resubstitution zu $F_\lambda^v(\vec{x}\,) = c - 2a\left(\sqrt{\vec{x}}+1/\lambda\right)e^{-\lambda\sqrt{\vec{x}}}$ geschrieben werden kann. Das Einsetzen der Grenzen $\{0;\infty\}$ für $\vec{x}$ in $F_\lambda^v(\vec{x}\,)$ – beachten, daß die Exponentialfunktion schneller schwindet als die Wurzelfunktion wächst – beim Berechnen des zu Eins gleichgesetzten Flächeninhaltes führt zur Lösung $a = \lambda/2$.
Da die Verteilungsfunktion im Ursprung beginnt, kann somit auf die Anfangsbedingung c rückgeschlossen werden: $c = 1$. Als Verteilungsfunktion[42] wird final realisiert:

[41] Wenn im Folgenden auf $a = 0$ argumentiert wird, so meint es das Umgehen der Transformation.
[42] Diese Verteilung wird im Folgenden $\mathrm{Exp}^{0.5}$-Verteilung genannt.

$$\tilde{F}_\lambda(\vec{x}\,) \;=\; 1 - \lambda\left(\sqrt{\vec{x}} + 1/\lambda\right) e^{-\lambda\sqrt{\vec{x}}} \;=\; 1 - \left(\lambda\sqrt{\vec{x}} + 1\right) e^{-\lambda\sqrt{\vec{x}}}. \tag{58}$$

Dann folgt für die Dichte $\frac{\partial \tilde{F}_\lambda(\vec{x}\,)}{\partial \vec{x}} \;=\; \tilde{f}_\lambda(\vec{x}\,) \;=\; \frac{\lambda^2}{2} e^{-\lambda\sqrt{\vec{x}}}$.

Erwartungswert und Varianz lassen sich über Integrale (ohne Herleitung) notieren:

$$\mathbb{E}(\vec{x}\,) = \int_0^\infty \vec{x}\, \tilde{f}_\lambda(\vec{x}\,)\, d\vec{x} \;=\; \frac{\lambda^2}{2} \int_0^\infty \vec{x}\, e^{-\lambda\sqrt{\vec{x}}}\, d\vec{x} \;=\; \frac{6}{\lambda^2},$$

$$\mathbb{V}(\vec{x}\,) = \int_0^\infty \vec{x}^{\,2}\, \tilde{f}_\lambda(\vec{x}\,)\, d\vec{x} - \left(\mathbb{E}(\vec{x}\,)\right)^2 \;=\; \frac{120}{\lambda^4} - \left(\frac{6}{\lambda^2}\right) \;=\; \frac{84}{\lambda^4}.$$

Wird eine normierte Verteilung benötigt, beträgt $\mathbb{V}(\vec{x}\,) = 1$, welches mit $\lambda = \sqrt[4]{84} \approx 3$ möglich ist.

Während für die Exponentialverteilung die Quantile durch bloßes Umstellen der Verteilungsfunktion erreichbar sind: $\vec{x}_q \;=\; -\ln\left(1 - F_\lambda(\vec{x}\,)\right)/\lambda$, ist es für ihre transformierte Darstellung nicht analytisch möglich. Allerdings lassen sich mit Newtonscher Näherung die Lösungen schnell auffinden: Die Lösung des aktuellen Quantils ist gleichzeitig der Startwert für das nachfolgende Quantil.

6.3 Hauptkomponentenmethode auf das Spektrum

Für die weitere Analyse wird der Bandpaß beschnittene Datensatz in zweiter geglätteter Ableitung verwendet, wie in Kapitel 6.1 beschrieben und hergeleitet. Die Spaltenzahl p des Spektrums lautet nicht mehr 2153, sondern fortan 1473.

(Tillmann, 1996, Kapitel 4) empfiehlt eine Hauptkomponentenzerlegung des Spektrums zum Auffinden von Fehlern. Auf Seite 27 zeigt er einen beispielhaften Streuplot des ersten mit dem zweiten Faktor, in welchem die Objekte zufällig und symmetrisch angeordnet sind. Dies entspricht dem Idealtyp. Abweichungen von diesem Muster gelten als Fehler.

In den folgenden drei überlagerten Streuplots beschreiben graue Symbole die Lösungen der mit (57) logistisch transformierten Spektren. Schwarze Symbole entsprechen der untransformierten Voreinstellung $a = 0$. Die beiden prozentualen Angaben zu den Faktorenachsen beschreiben die erklärte Varianz (schwarz, grau):

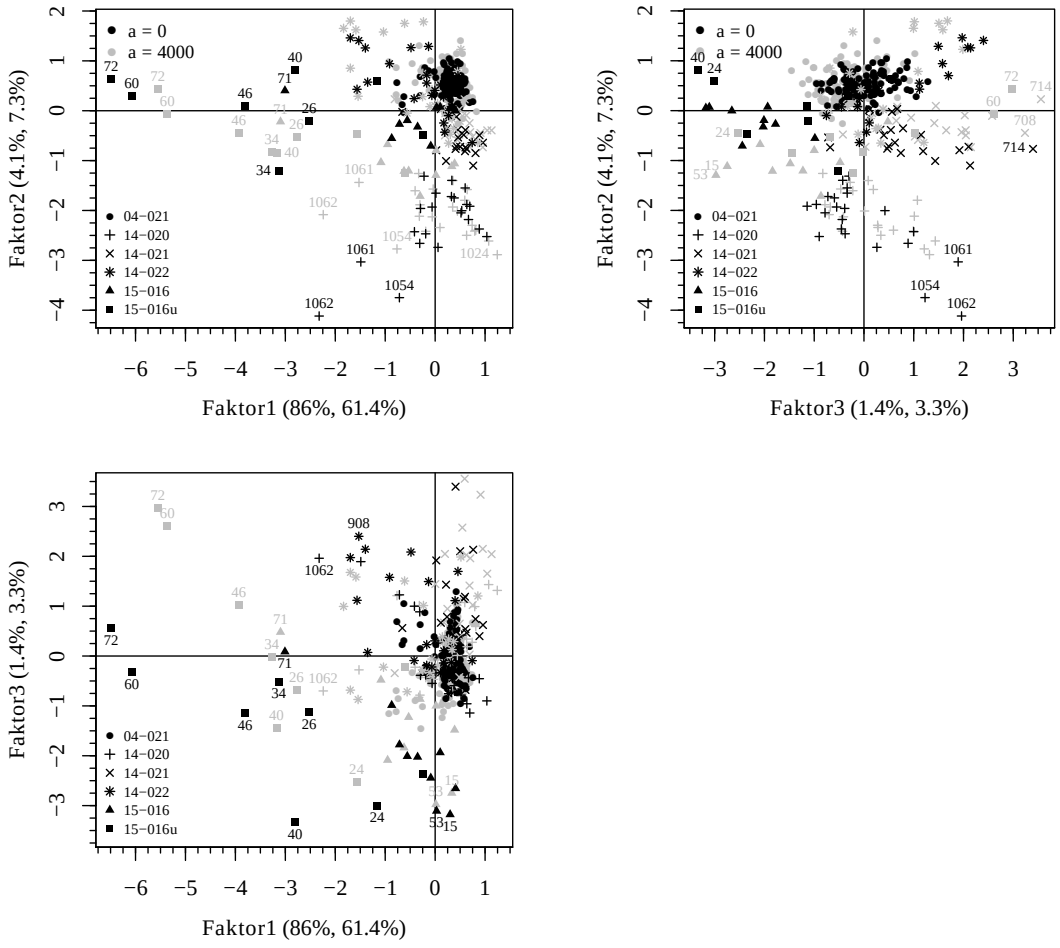


Abb. 18: Projektionen der ersten drei Faktoren, $a \in \{0\,;4000\}$.

Die hier vorliegenden zweimal differenzierten Spektren verletzen die Symmetrie. Einerseits tritt die Unterbodenklasse hervor, andererseits ist die Verteilung in der Projektion des ersten mit dem zweiten Faktor extrem schief.

Interpretation für $a = 0$ (schwarz):
Die Bodenklasse 04–021 wird in jedem Plot besonders kompakt abgebildet. Insgesamt

wirkt die Population heterogen, da die Klassen in den höheren Faktoren teilweise auch separiert erscheinen. Eine getrennte Auswertung der Bodenklassen ist aber nicht empfehlenswert. Schließlich ist die Fallzahl für die Population mit $n = 180$ eher niedrig.
Eine Alternative kann das Entfernen von weit entlegenen Objekten sein, da sie die Symmetrie stören. Dieses Vorgehen würde beinahe zum Totalverlust der Unterbodenklasse führen, aber das Schiefe-Problem nicht einmal im Ansatz bewältigen: Auch beim zusätzlichen Entfernen weniger weit entfernter Objekte bleibt die Schiefe hartnäckig bestehen. D.h., werden aufgrund der Projektion von Faktor1 mit Faktor2 alle zehn numerierten Einträge entfernt und auf die Restdaten eine Zerlegung angewandt, offenbaren sich wiederum weitere Objekte als untypisch.
Darauf könnte als Kritik eingewendet werden, es gibt Ausreißertests. Das weitere Vorgehen zu Ausreißern wird in Kapitel 6.4 beschrieben.

Interpretation für $a = 4000$ (grau):
Wenn räumlich eng eingegrenzte Regionen mit vielen benachbarten Objekten existieren, wie es die Projektionen offenbaren, dann verursachen einzelne weit entfernte Objekte eine betragsmäßig hohe Kovarianz in den Gesamtdaten, eine unangenehme Situation.
Das simultane Lösen des Problems der Ausreißer und der Schiefe kann mit einer nichtlinear bijektiven Datentransformation versucht werden. Erschwerend kommt aber hinzu, daß die Transformation nicht auf die puren Faktoren angwendet werden sollte. Denn ab $g > 2$ Faktoren kommt es zu Interaktionen zwischen den Faktoren, die unvorhersehbare Wirkungen haben. Ohnehin würde von vornherein die Orthogonalitätseigenschaft der Faktoren verloren gehen. Günstiger ist es die Transformation bereits auf die Einflußgrößen durchzuführen.
Welche nichtlineare Transformationsfunktion kann die gewünschten Eigenschaften herbeiführen? Mit der Exponentialfunktion durchgeführte Transformationen haben kaum Auswirkungen auf die Lageänderungen der Objekte im Faktorenraum. Hingegen ist der Logarithmus nicht verfügbar, da sich die Ableitung des Spektrums auch in den negativen reellen Zahlenbereich erstreckt – vgl. Abb. 14.
Die Transformation mit (57) auf das Spektrum vorzunehmen, ändert die Faktor-Lösungen. Insgesamt wirkt die Verteilung weniger schief. Augenscheinlich verringern sich die Abstände der Daten zu den beiden entferntest gelegenen Objekten {60 72}, welches beabsichtigt wird. Und das hat Auswirkungen auf die Varianz, wie man sofort am ersten Faktor erkennt: Nach der bijektiven Transformation erklärt er ca. 61% statt 86% an gesamter Variation. Die Differenz nehmen die höheren Faktoren auf.

Bei einer Extraktion von $g = 11$, der maximal $n = 180 - 1$ möglichen Faktoren, kommt es zu einem gewollten Informationsverlust. Mit den ersten 11 Faktoren wird auf die untransformierten Daten 95,5% und den logistisch transformierten Daten 84,7% der Varianz erklärt. Eine verlustbehaftete Darstellung ist mit (31) von Seite 29 möglich, indem der Rest der Fehlermatrix zugewiesen wird.
Die informativste Eigenfunktion aus der ersten Dimension (in Realisation der erste Eigenvektor) als Verbindung zwischen dem Spektren- und Hauptachsenraum wird durch die Transformation wenig verformt. Auf den informativen Wellenlängen wird ihre Amplitude

teilweise abgeschwächt aber auch verstärkt:

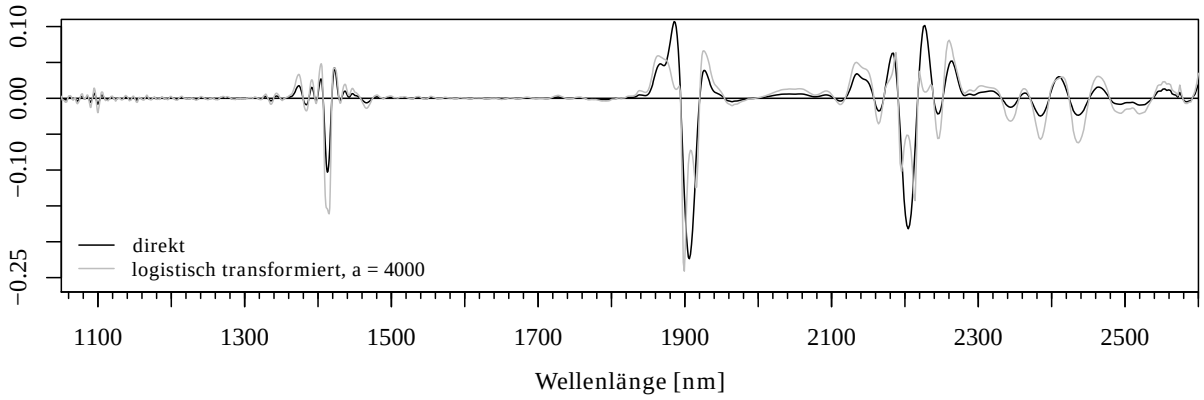


Abb. 19: Eigenfunktion des Spektrums für die erste Hauptachse.

Wäre die Eigenfunktion der transformierten Spektren deformiert, müßte der Nutzen der Transformation kritisch hinterfragt werden. Symmetrische Projektionen lassen sich z.B. mit $a = 4{*}10^5$ erzeugen. Diese Datenstruktur hätte mit den Ausgangsdaten keine Gemeinsamkeiten mehr. Die in (57) beschriebene Transformation sollte vorsichtig angewendet werden, um die Kovarianz-Struktur weitgehend zu erhalten.

Für eine PLS mit dem Design $n < p$ ist der erklärte Anteil an Einflußgrößen nicht unbedingt Modell entscheidend. Das liegt in der Struktur von (46.3) begründet, mit welcher ein unterbestimmtes Gleichungssystem gelöst wird. Interessanter wird der erklärte Anteil der Zielgrößen sein. Letztendlich soll ein hinreichend angepaßtes Modell aufgestellt werden. Als Optimalitäts-Kriterium kann der MSE in (47) auf die Zielgröße zu ihrer Prognose verwendet werden.

6.4 Erkennen von Ausreißern

Die Hebelwirkung der beiden Objekte {60 72} ist nach der vorgestellten Transformation geringer (vgl. Projektion Faktor1 mit Faktor2 der Abb. 18). Sind es Ausreißer bzw. existieren noch mehr Ausreißer?

Hierzu benötigt man i.A. eine Verteilungsannahme, z.B. die Normalverteilung: Über die t-Verteilung werden dann im univariaten Fall verdächtige Objekte identifiziert, bei denen ihr empirischer Quantilswert betragsmäßig ein festgelegtes theoretisches (tabelliertes) Quantil $1 - \alpha/2$ übersteigt – für ein multivariates Problem äquivalent über die Hotellings T^2-Verteilung[43] zu bewerkstelligen. Auch mit dem multivariaten Test können untypische Objekte bereits in den Ausgangsdaten entdeckt werden, falls die Inverse der Kovarianzmatrix von X exisitiert. Diese Voraussetzung ist wegen $n < p$ nicht erfüllt, ließe sich aber im g-dimensionalen Faktorenraum realisieren. Zu beachten ist, daß die Hotellings T^2-Verteilung eine quadratische Form beinhaltet, wie man sie im Kern der multivariaten Normalverteilung wiederfindet. Damit wird der euklidische Raum über die Kovarianz gewichtet – quadrierte Mahanalobis-Distanz genannt.

[43] https://de.wikipedia.org/wiki/Hotellings_T-Quadrat-Verteilung (25.4.2018)

Das nächste Problem: Als Voraussetzung für den Test steht die p-dimensionale Normalverteilungsannahme. Anhand der Projektionen in Abb. 18 ist ersichtlich, sie ist verletzt.

Wie aber können dann multivariate Ausreißer identifiziert werden? Angenehmer wäre es über ein univariates Problem zu gestalten. In den Projektionen zeigen sich beide Objekte {60 72} bzgl. der Projektionen Faktor1 – Faktor2 und Faktor1 – Faktor3 extrem vom Ursprung liegend. Bei der abschließenden Ausreißer-Entscheidung muß man bedenken, der endgültige Faktorraum wird höherdimensional sein. Die richtige Faktorenzahl wird für die PLS relevant sein. Unter Vorwegnahme der Kenntnis des Kapitels 6.5 ist ein 11-Faktorenmodell für die PLS naheliegend.
Bei der Ausreißerbewertung ist es notwendig, nicht auf die Faktoren F, welche normiert sind, zu arbeiten; sondern auf die Hauptachsen T, die exakte Abstände reproduzieren. Da Hauptachsen – wie auch Faktoren – orthogonal zueinander stehen, sind sie unkorreliert, womit eine Gewichtung entfallen kann. Die Distanzberechnung aller Objekte zum Ursprung wird im euklidischen Hauptachsenraum vollzogen:

$$\text{Distanz} = \sqrt{\text{diag}\,(TT^T)}. \tag{59}$$

Darin sind alle 11 extrahierten Hauptachsen berücksichtigt. Die Prozedur muß ein zweites Mal für die logistisch transformierten Daten ausgeführt werden. Damit man die Auswirkungen direkt miteinander vergleichen kann, werden die Distanzen normiert.

Die folgende Abbildung zeigt im oberen Boxplot die Auswirkungen bei Vorlage der Ausgangsdaten und im unteren Boxplot bei den logistisch transformierten Daten. Mit dem Symbol + wird der Schwerpunkt einbeschrieben:

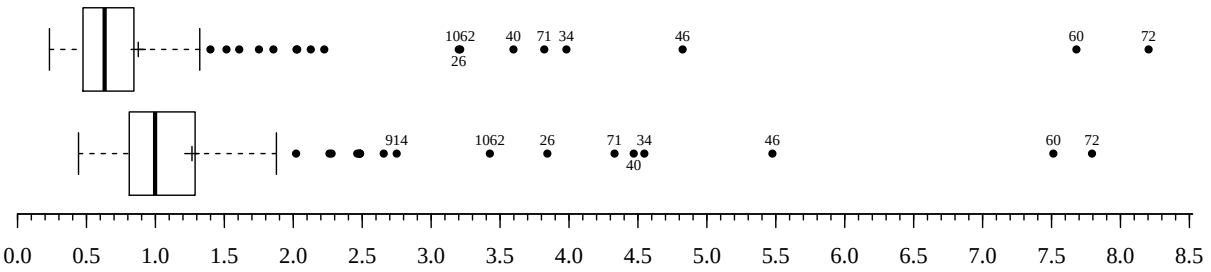


Abb. 20: Boxplot der Distanzen im 11-dim. Hauptachsenraum, oben: $a = 0$, unten: $a = 4000$.

Der obere Boxplot hat engere Quartilsabstände und die Objekte {60 72} liegen extremer als im darunterliegenden Boxplot, welcher auch weniger Schiefe andeutet. Die Tatsachen waren schon in den Projektionen ablesbar. Markierte Objekte sind bei beiden Darstellungen in etwa in derselben Reihenfolge gelistet. Insgesamt wirken beide Boxplots recht ähnlich. Nur allein über das Beurteilen der Projektionen ist es mühevoll, die Distanzverringerung zu beiden Objekten sicher zu erkennen. Mithilfe der eindimensionalen Boxplots läßt es sich intuitiver erfassen.

Ausreißer zu identifizieren ist oftmals mit viel Unsicherheit unterlegt. Verteilungen sind theoretisch mit den ersten beiden Momenten hinreichend beschreibbar. Kennt man nur beide Momente, aber nicht die Verteilung, wie es bei empirischen Daten die Regel darstellt, ist ein Rückschluß auf die Verteilung mit beiden Momenten sehr kritisch. Eigentlich

benötigt man dafür alle Momente. Diese Forderung erscheint wiederum unrealistisch. Weit gelegene Objekte automatisch als Ausreißer zu deklarieren mag bequem sein, aber nicht immer nachvollziehbar.
Mithilfe von (59) wird ein Distanz-Vektor berechnet. Unter der Wurzel befindet sich ein quadratischer Ausdruck. Unterstellt man normalverteilte Hauptachsen, wäre mit quadrierten Distanzen ein Hinweis auf die $\chi^2(1)$-Verteilung gegeben, welche eine quadrierte Standardnormalverteilung surjektiv abbildet. Allerdings wird durch die Wurzel-Operation keine Normalverteilung zurückgewonnen. Reellwertige Distanzen sind außerdem nichtnegativ.
Verschiedene Verteilungs-Szenarien:

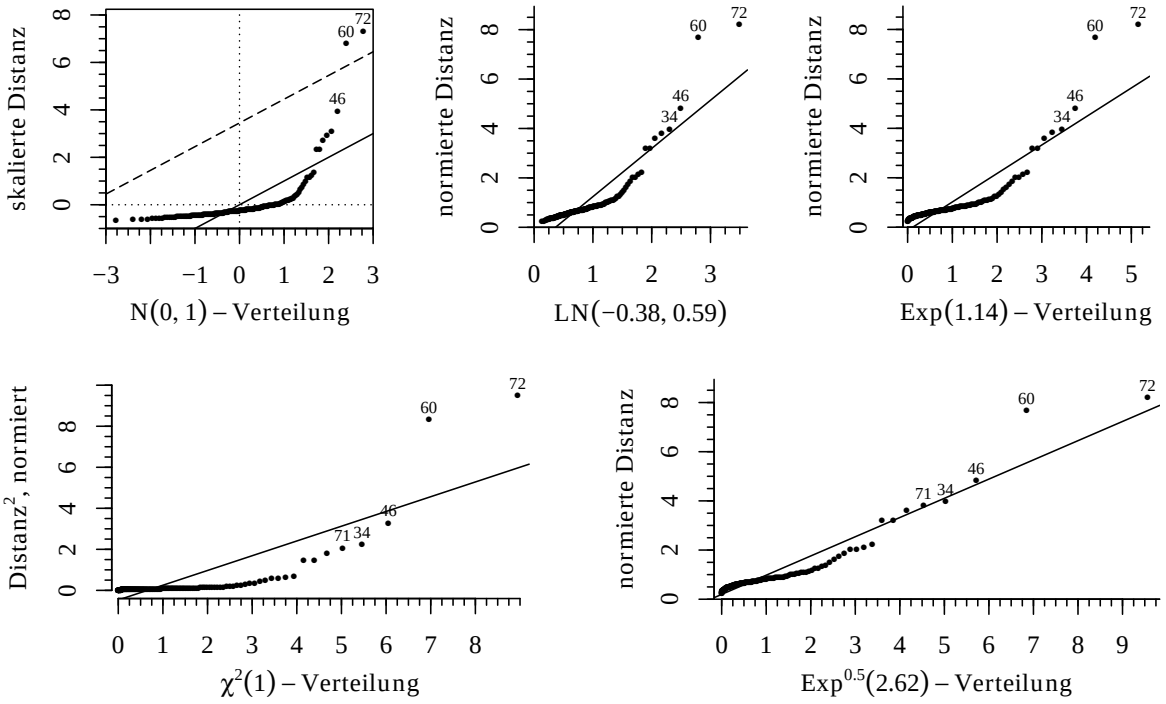


Abb. 21: QQ-Plots – Quantile der Distanzen gegen fünf theoretische Quantile, $n = 180$.

Falls sich die Objekte entlang der einbeschriebenen Idealgeraden aufreihen, entspräche ihre Verteilung perfekt der theoretischen Verteilung. Abweichungen von der Geraden stellen somit Verletzungen der Verteilungsannahme dar. Am Beispiel der Normalverteilung wäre zu beachten, daß sie gemäß zentralem Grenzwertsatz[44] nur für $n \to \infty$ perfekt realisiert wird. Abweichungen von der Idealgeraden sind bei endlicher Population zu erwarten. D.h., die Ausreißergrenze sollte großzügiger ausgelegt werden. Für die anderen hier gezeigten Verteilungen darf die Richtlinie ebenfalls unterstellt werden.
Mit einer Normalverteilungsannahme kommen auch kleine Distanzen unter Ausreißerverdacht, eine widersprüchliche Aussage. Für die Berechnung der nach oben einseitigen Grenze wird der invertierten Verteilungsfunktion als Argument $1-\alpha/(n-2)$ mit $\alpha = 0,05$ und $n = 180$ übergeben. Damit wird eine Spannweite von knapp $3^1/_2$ Standardabweichungen von der Idealgeraden erreicht: Zwei Objekte mit den größten Distanzen liegen oberhalb

[44] https://de.wikipedia.org/wiki/Zentraler_Grenzwertsatz (29.4.2018)

der Grenze und werden als Ausreißer deklariert. Im QQ-Plot entspricht der stark gekrümmte Verlauf nicht annähernd einer Normalverteilung. Er gibt einen Hinweis auf eine rechtsschiefe Verteilung.
Über eine logarithmische Normalverteilung werden die Distanzen stärker linear gereiht. Das ist ein eindeutiges Indiz für eine schiefe Verteilung. Eine logarithmische Normalverteilung gewinnt man via Dichtetransformation aus einer Normalverteilung, indem der Träger exponiert wird. Diese Abbildung ist bijektiv.
In der Tat ergibt sich mit einer Exponentialverteilung ebenso ein Bild, welches den Eigenheiten der Daten besser entspricht: kleine Distanzen bleiben unkritisch und zwei bzw. acht Objekte wären kritisch bzgl. ihrer Werte.
Die χ^2-Verteilung könnte, wegen der Spezifika bei Distanzen, zur Darstellung geeignet sein und gibt einen Hinweis auf zwei bzw. acht Ausreißer. Wären die Hauptachsen annähernd multivariat normalverteilt, sollte eine χ^2-Verteilung die erste Wahl darstellen.
Beim Plot mit der $\text{Exp}^{0.5}$-Verteilung ergibt sich insgesamt ein grob geradliniger Verlauf. Daran sei noch einmal gezeigt, wie wichtig eine Verteilungsannahme für das Interpretieren von Ausreißern ist. Lediglich Nr. 60 weicht stärker von der Ideallinie ab – Verdacht auf einen Ausreißer. Wird allerdings Nr. 60 entfernt, rückt Nr. 72 von der neuen Idealgerade weiter ab.
In allen Plots zeigen sich zwei Brüche. Die größten acht Distanzen sind sichtbar von den Restdaten separiert und innerhalb der acht Werte ragen die beiden größten Werte nocheinmal nach oben heraus. Es könnte eine Mischverteilung vorliegen.

Wie soll entschieden werden, was passiert beim Entfernen von den acht extremsten Objekten?

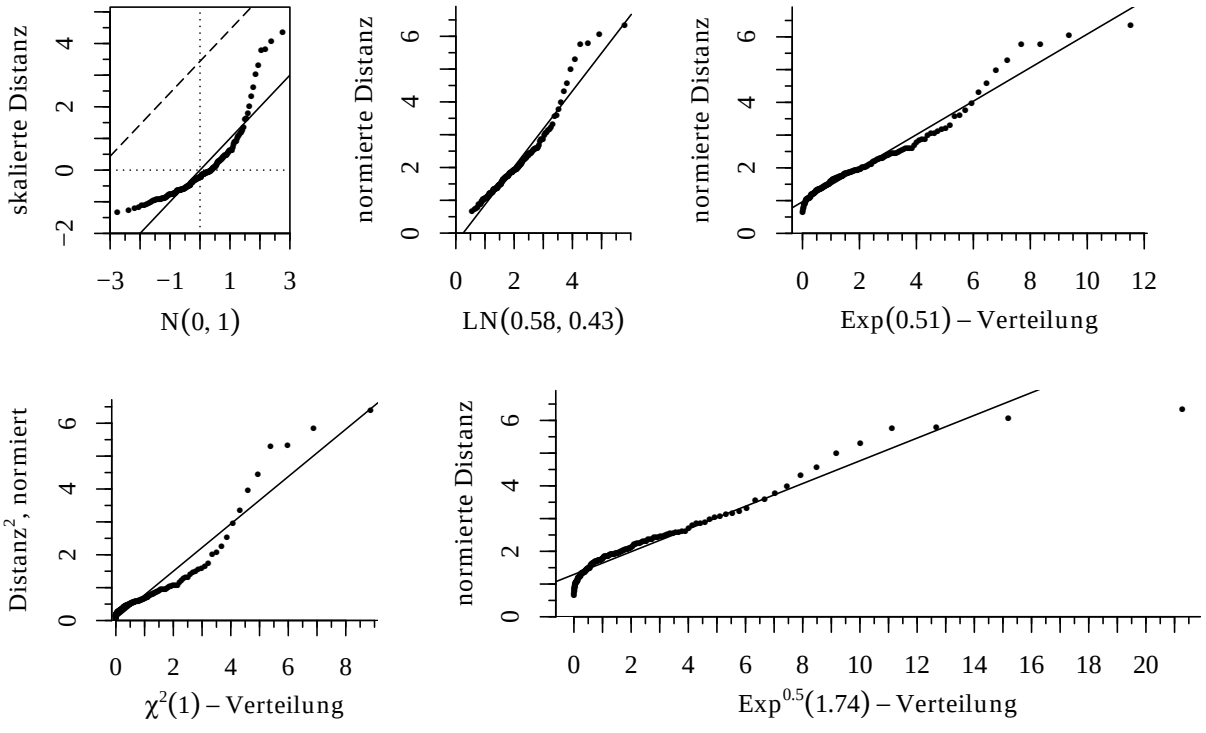


Abb. 22: QQ-Plots – Quantile der Distanzen gegen fünf theoretische Quantile, $n = 172$.

Abgesehen von der $\text{Exp}^{0.5}$-Verteilung, bei der mit den acht offenbar zuviele Objekte entfernt und die Verteilung bereits gestuzt wurde, werden die Darstellungen der anderen vier QQ-Plots weitgehend linearisiert, indem die Masse der Objekte an die Idealgerade heranrückt. Gute Näherungen werden mit der logarithmischen Normalverteilung, der Exponentialverteilung und der χ^2-Verteilung realisiert.
U.U. kann ein Ausreißerverdacht vorliegen. Für Distanzen nahe der Null existieren bei der Exponentialverteilung im Vergleich zu Abb. 21 weiterhin Abweichungen, lediglich mit umgekehrtem Vorzeichen. Bei der $\text{Exp}^{0.5}$-Verteilung tritt der Effekt noch drastischer hervor. Hier kann die χ^2-Verteilung auf Basis quadrierter Distanzen besser zufriedenstellen.

In QQ-Plots mit umfangreich vorliegenden Objekten ist die Auflösung der Punkte allerdings eingeschränkt. Die Vorstellung über die Häufung der Daten bleibt abstrakt. Zum Verständnis kann ein zusätzlicher Plot der Dichten beitragen:

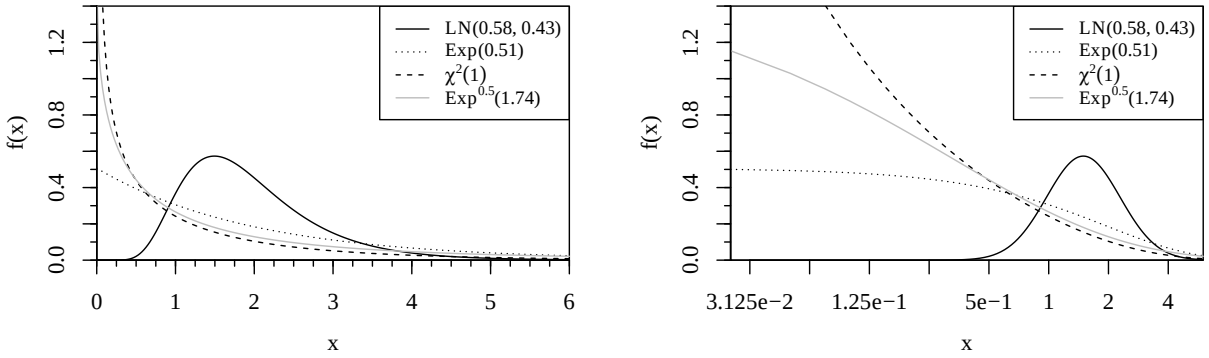


Abb. 23: Dichten von theoretischen Verteilungen, Abszisse: links linear, rechts $\log_2$-Skala.

Mithilfe einer log-Skala können besonders kleine Skalenwerte in erhöhter Auflösung, zu Lasten der großen Werte, dargestellt werden.
Die $\text{Exp}^{0.5}$-Verteilung konzentriert, stärker als die Exponentialverteilung, mehr Wahrscheinlichkeitsmasse in der Umgebung von Null. Offensichtlich liegen jedoch eingipflig verteilte Distanzen vor, welche in guter Näherung mit einer logarithmischen Normalverteilung beschreibbar sind, sobald die acht größten Objekte entfernt werden. Unter Berücksichtigung von Abb. 22 kann der Modalwert der einfachen Distanzen oberhalb Null gefolgert werden.
Unter Ausreißerverdacht stehen die Objekte {26 34 40 46 60 71 72 1062}. Mit Ausnahme der Nummern {71 1062} gehören die übrigen sechs verdächtigen Objekte der Unterbodenklasse an, welche selbst insgesamt acht Objekte beinhaltet.

6.5 PLS1 auf die Zielgrößen Stickstoff und Kohlenstoff

Die für eine weitere Analyse relevante Spaltenzahl des abgeleiteten Spektrums wurde zu Anfang in Kapitel 6.3 auf 1473 eingegrenzt. Einen ersten Eindruck der PLS für die ersten 50 Faktoren, getrennt nach beiden Zielgrößen, geben die Güte-Grafiken:

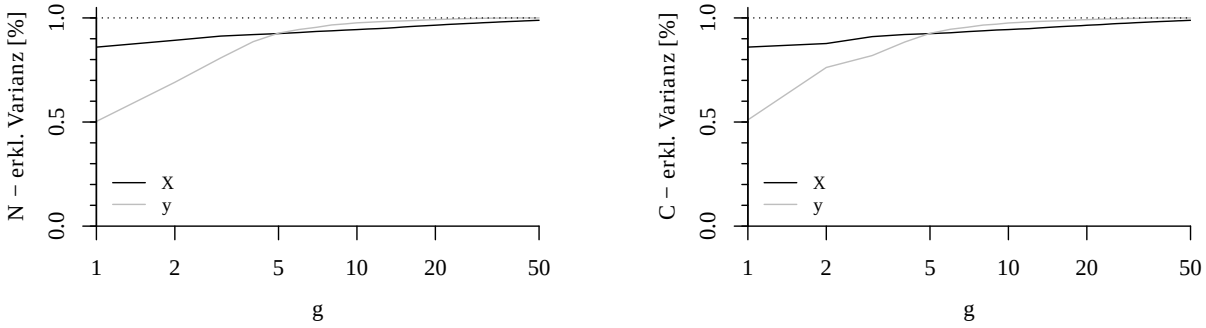


Abb. 24: Zunahme an erklärter Varianz für jeden weiteren Faktor – Abszisse: $\log_2$-Skala.

Die Kurven in beiden Plots geben einen Überblick über die Zunahme der erklärten Varianz bei steigender Faktorenzahl, für Einfluß- und Zielgrößen.

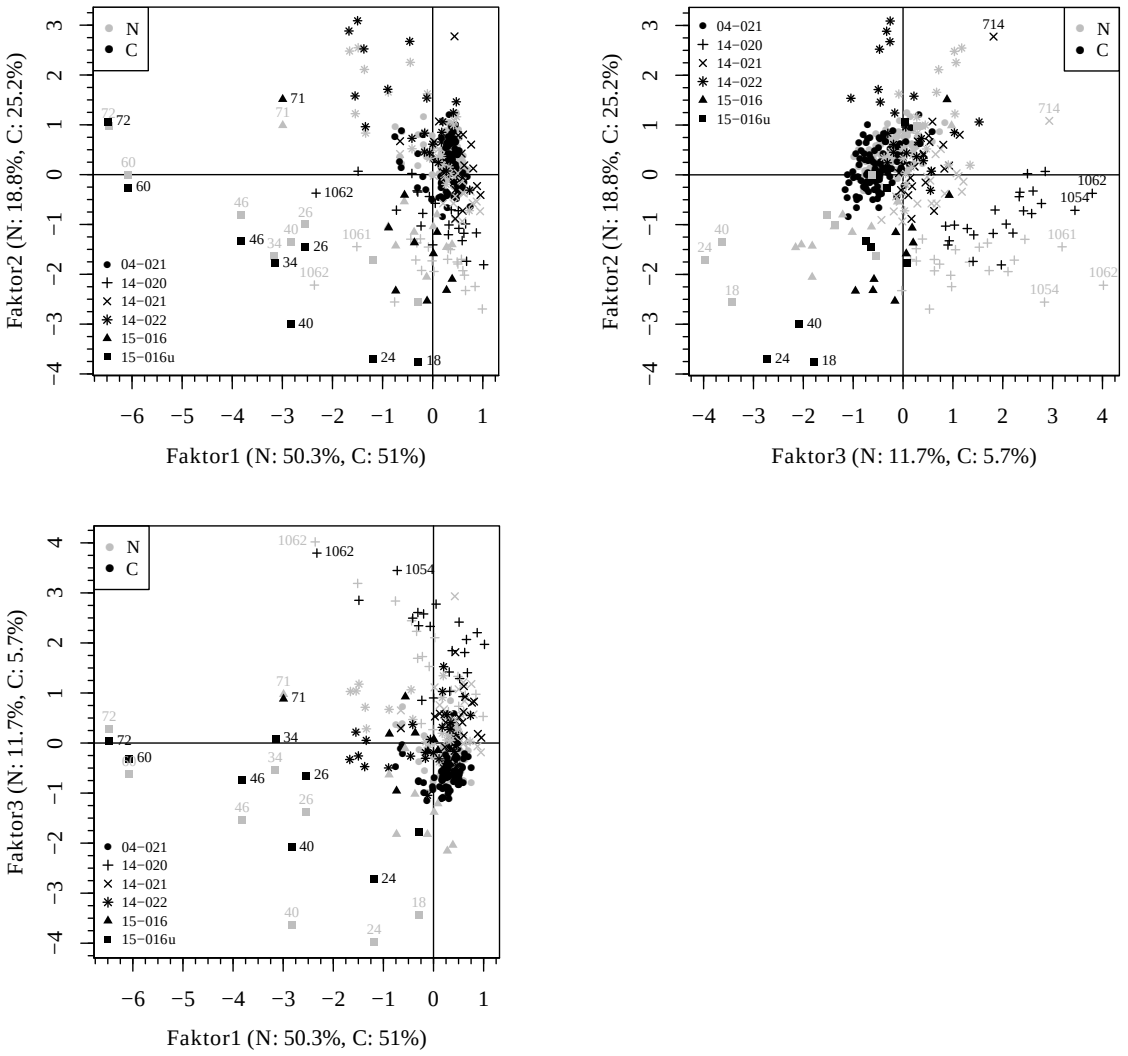


Abb. 25: N – C überlagerte Projektionen der ersten drei PLS-Faktoren

An den Faktorenachsen in Abb. 25 sind die aufgeklärten Varianzanteile der Zielgrößen abgetragen. Grobe Ähnlichkeiten in der Darstellung im Vergleich zu Abb. 18 lassen sich nicht verneinen. Die Gemeinsamkeiten von Faktorenanalyse und PLS beschränken sich nicht nur auf die rein formale Darstellung des Zerlegungsproblem.
Als neue Erkenntnis kommt hinzu, für Stickstoff (N) läßt sich die Unterbodengruppe in der Projektion Faktor1 – Faktor2 vollständig linear separieren, für Kohlenstoff (C) in der Projektion Faktor1 – Faktor3.
In den Projektionen Faktor1 – Faktor2 bzw. Faktor1 – Faktor3 liegen im Mittel keine außergewöhnlichen Unterschiede bei der simultanen Darstellung der numerierten Paare von Stickstoff und Kohlenstoff vor. Aber im überlagerten Plot Faktor2 – Faktor3 zeigen sich zwischen beiden Schichten alle numerierten Objekte systematisch in eine Richtung zueinander versetzt positioniert. Das ist unkritisch, denn durch Rotation einer Schicht ließen sich die Paare in bessere Übereinstimmung bringen. Orthogonales Rotieren von Faktoren ändert in der Summe nicht die Kommunalitäten – beachte (32) auf Seite 29. Diese Art einer optimalen Verdrehung gegeneinander, zum Minimieren der paarweisen Abstände, kann mit Procrustes[45]-Techniken vorgenommen werden, welche an dieser Stelle nicht weiter vertieft werden.

Nur so wenig Faktoren wie nötig sollen in das finale Modell einfließen. Unter Berücksichtigung der Abb. 20 bis 22 wurden bis zu acht Objekte unter Ausreißer-Verdacht gesetzt. Die Zielgrößen in Abb. 1 auf Seite 7 zeigen besonders für die Unterbodenklasse kleine Werte an. Sind Einfluß- und Zielgrößen miteinander mittel bis hoch korreliert (vgl. Abb. 17, Seite 53), wird das Entfernen von Ausreißern kaum eine Wirkung entfalten, da die verdächtigen Punkte vermutlich im Einklang zum Trend sind. Durch Verringern der Stichprobe verliert man in den Daten an Information, ohne Hinzugewinn von Modellgüte – im ungünstigsten Fall.

Unsymmetrien in den Projektionen der Spektren wurden mit der Hauptkomponentenmethode entdeckt und versucht abzumildern. Wenn die Spektren mit (57) logistisch transformiert werden, wäre der Nutzen der Transformation in einem besser erklärenden Modell zu erkennen, bei dem der Einfluß von Ausreißern abgeschwächt wäre. Dabei steht als Problem, die beste Lösung für a zu errechnen. Das ist über den MSE (mittlerer quadratischer Fehler) möglich:

$$f(a) = RMSEP_a = \sqrt{\frac{1}{n}\left(\vec{y} - \vec{\hat{y}}(a)\right)^T\left(\vec{y} - \vec{\hat{y}}(a)\right)} \ , \tag{60}$$

bei dem die Wurzel-Operation aus dem *MSE* den *RMSEP* (Wurzel des mittleren quadratischen Vorhersagefehlers) generiert. Zur Vereinfachung soll im Folgenden (60) als MSE bezeichnet werden.
Beim Berechnen des MSE der Regression muß zumindest irgendein $a > 0$ vorausgesetzt werden, um ihn mit dem MSE des Nullmodells zu vergleichen. Das führt auf ein Minimierungsproblem, bei dem die bzgl. a zu minimierende MSE-Funktion a priori unbekannt ist. Allerdings muß zu jedem interessierenden Wert von a ein PLS-Schritt, d.h. die PLS-

[45] https://de.mathworks.com/help/stats/procrustes.html (5.8.2018)

Prozedur einmal pro Zielgröße durchgeführt werden.
Wenn das Funktional der MSE-Funktion $f(a)$ unbekannt ist, so auch ihre Ableitung $f'(a)$. Bei einer Intervallschachtelung wird üblicherweise ein Nullstellenproblem gelöst, welches mit der Ableitung auch vorläge. Zum Definieren eines Intervalls kann das Kontinuum der positiven reellen Zahlen für a grob abgetastet werden. Mit einem modifizierten Ansatz der Intervallhalbierung, bei dem quasi simultan auf Intervallviertel/ -dreiviertel gerechnet wird, kann trotz fehlender Nullstelle wenigstens in jeder Iteration das Intervall halbiert werden, in welchem sich argmin $(f(a))$ aufhält.
Es muß bedacht werden, mit a wird ein weiterer Parameter eingeführt, der evtl. keinen Nutzen hat. Er wirkt dennoch nicht schädlich auf die Güte des PLS-Modells, denn es existiert vorab seine Alternative mit $a = 0$. Wird dieser Parameter nicht kontrolliert (i.A. regularisiert), unterliegt man u.U. der Versuchung sich das Modell „zurecht zu biegen“ bzw. realisiert einen Wert für a, der nicht optimal erklärt.
Unkontrolliert optimiert liegt bei 11 Faktoren das realisierte a in einen Bereich um 10^9 bis 10^{10}, in welchem das MSE-Minimum vermutet wird und das dann etwa vier bis fünfmal kleiner ausfällt, als der MSE für $a = 0$. In der Tat retuschiert dieser extreme Wertbereich alle verdächtigen Objekte und erzeugt eine homogenisierte Faktorenwolke der 180 Spektreneinträge, welche sich perfekt darbietet. Die Präzision von min(MSE) ist gering, da auch noch in größerer Umgebung von a der Anstieg der MSE-Funktion betragsmäßig innerhalb von 10^{-16} bis 10^{-15} liegt. In diesem Modell haben weder die Eigenfunktionen bzw. die Position der Objekte in den Faktoren-Projektionen noch irgendeinen Bezug zur Realität, aber die Prognose wird bestmöglich.
Für die Größenordnung von a ist es nicht mehr von Belang, ob die verdächtigen Objekte tatsächlich entfernt wurden. Damit kommt die Frage auf, wie kritisch die Ausreißer überhaupt zu bewerten sind. Ebenso ist nicht sichergestellt, daß der Minimalpunkt lediglich aufgrund von Rundungsfehlern (vgl. Kapitel 2) in den Fließkommazahlen resultiert. Mit einer Sensitivitätsanalyse kann es überprüft werden.

Mithilfe der Kreuzvalidierung[46] (CV – *cross validation*) ist es möglich, in die Nähe eines Optimums zu gelangen. Kreuzvalidierung bezeichnet eine Technik, die man dem *Maschinellen Lernen*[47] zuordnen kann. Vorab steht die Entscheidung, welche Art der CV durchgeführt werden soll. Populär sind drei Varianten: 5-fache, 10-fache, n-fache[48] CV, wobei die n-fach Variante bei Existenz einer Hat-Matrix verallgemeinerbar wäre – generalisiertes Kreuzvalidierungskriterium (gcv) genannt (vgl. (51) auf Seite 44).
Eine Aufteilung der Einträge in Trainings- und Testdaten muß vorgenommen werden. Dabei wird ein Datensatz in möglichst gleichgroße Schichten gesplittet. Der Vorfaktor der Kreuzvalidierungsart beschreibt die Zahl der Überkreuzungen/Wiederholungen an Berechnungen und seine Inverse den relativen Umfang eines Testdatensatzes. D.h., bei einer 5-fachen CV existieren vier Trainingsdatensätze und es sind fünf Durchläufe erforderlich, um jedes Objekt mit Sicherheit auch einmal im Testdatensatz vorliegen zu haben. Der Umfang der Testdaten beträgt hier $1/5 = 20\%$. Läßt sich die n-fache Kreuzvalidierung

[46] https://de.mathworks.com/discovery/kreuzvalidierung.html (27.5.2018)
[47] https://www.kaspersky.de/blog/machine-learning-explained/9245/ (27.5.2018)
[48] https://en.wikipedia.org/wiki/Jackknife_resampling (27.5.2018)

nicht generalisieren, ist sie mit Abstand die Rechenzeit teuerste Variante, da n Durchläufe zu absolvieren sind.

Auf die Trainingsdaten wird das Modell angepaßt. Angewandt wird es auf die Testdaten, indem die geschätzten Modellparameter der Trainingsdaten eingesetzt werden und Schätzungen der Testdaten berechnet werden. Das entspricht einer Vorhersage der Testdaten, welche mit den wahren Testdaten über (60) validiert wird. Bei fünf Durchläufen werden ebensoviele MSE-Werte erzeugt, die zu einem MSE gemittelt werden.
Um die Unsicherheit des MSE zu verkleinern, kann es sinnvoll sein, Kreuzvalidierungen, die einen von n abweichenden Vorfaktor haben, mehrmals zu permutieren, indem die Schichten variiert werden. Beispielsweise über zufällige einfache Stichprobenziehungen können mehrere Sets an Trainings- und Testdaten nacheinander generiert werden.

Überlegenswert ist eine dauerhafte Entzerrung der Spektren mit a einzustellen. Dazu muß zum Auffinden einer geeignet eingeschränkten Faktorenzahl das zur Faktorenzahl optimale a ermittelt werden. D.h., für jede Faktorenzahl ist eine andere Realisation von a zu erwarten, die den MSE minimiert. Diese beiden Probleme sollen mit der Kreuzvalidierung simultan bewältigt werden. Es gewinnt jene Faktorenzahl g, die mit ihrem optimalen Wert bzgl. a den kleinsten MSE produziert. Für eine Sensitivitätsanalyse steht zum Vergleich das Nullmodell mit fixiertem $a = 0$ für jede Faktorenzahl als Konkurrenzmodell gegenüber.

Die folgenden Grafiken sind zusammengesetzte Plots, bei denen die linke Ordinate die MSE-Skala der 5-fachen Kreuzvalidierung und die rechte Ordinate die Skala bzgl. a beschreibt – jeweils getrennt für Stickstoff (N) und Kohlenstoff (C):

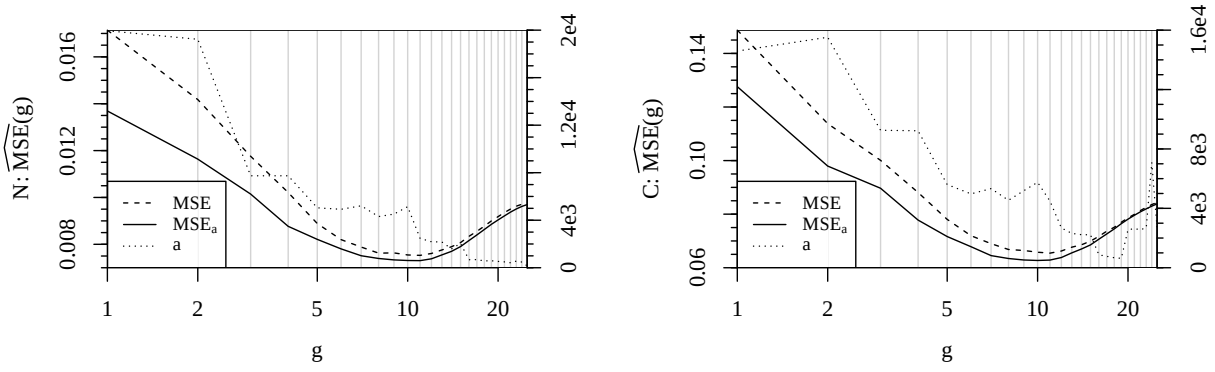


Abb. 26: arithmetischer Mittelwert der Kreuzvalidierungen für {N,C}: $20\times$ 5-fache CV (-----) und $50\times$ 5-fache CV simultan auf a; $n = 180$ – Abszisse: $\log_2$-Skala.

Abgesehen von der n-fachen Kreuzvalidierung kann das Minimum um den wahren Faktor stärker streuen, so daß mit einem einzigen Durchlauf die richtige Faktorenzahl nicht in jedem Fall getroffen werden muß. Die 5-fache CV wurde deshalb 20 mal permutiert. Das ist mithin nicht viel, wenn man beachtet, daß rechnerisch $\binom{180}{5} \approx 1,49 * 10^9$ Permutationen möglich wären. Je mehr Auswahlsätze – d.h. Permutationen – ausgeführt werden, umso präziser wird der Verlauf der mittleren Kurve erwartet. Mit 20 absolvierten Durchläufen ist die resultierende mittlere MSE-Funktion dennoch stabil.

Bei einer Kreuzvalidierung mit simultaner Optimierung von a wird bei 50 Durchläufen dann ebenfalls eine stabile Lösung erreicht.

Die MSE-Kurven für die – um acht Ausreißer – bereinigten Daten:

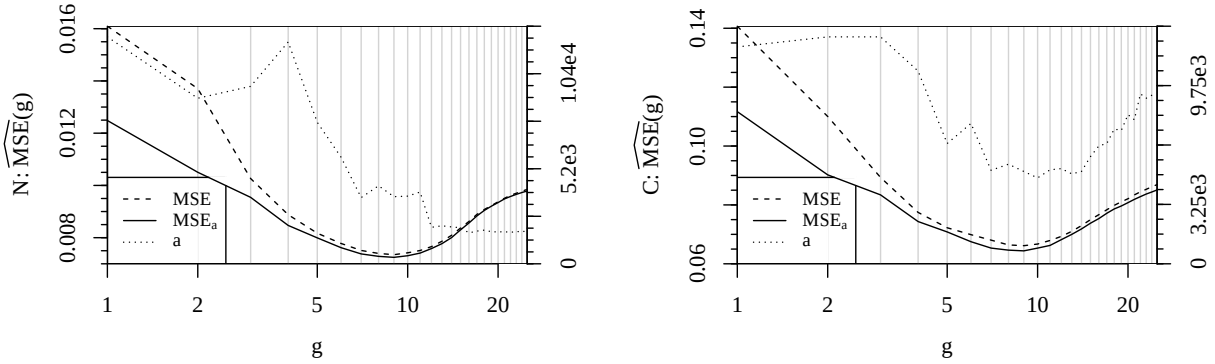


Abb. 27: arithmetischer Mittelwert der Kreuzvalidierungen für {N,C}: $20\times$ 5-fache CV (-----) und $50\times$ 5-fache CV simultan auf a; $n = 172$ – Abszisse: $\log_2$-Skala.

Tabellarische Zusammenfassung der Ergebnisse aus den Abb. 26 & 27:

		N		C	
	n	g	a	g	a
$\star$	180	11	0	11	0
	180	11	2470	10	5754

		N		C	
	n	g	a	g	a
$\star$	172	9	0	9	0
	172	9	3672	9	5051

Tab. 2: Ergebnisse der Kreuzvalidierungen; $\star$ als Synonym für fixiertes a

In den Plots sind lediglich Funktionen von Mittelwerten abgebildet. Weiteren Aufschluß gibt die Streuung der a-simultanen mittleren MSE-Funktion:

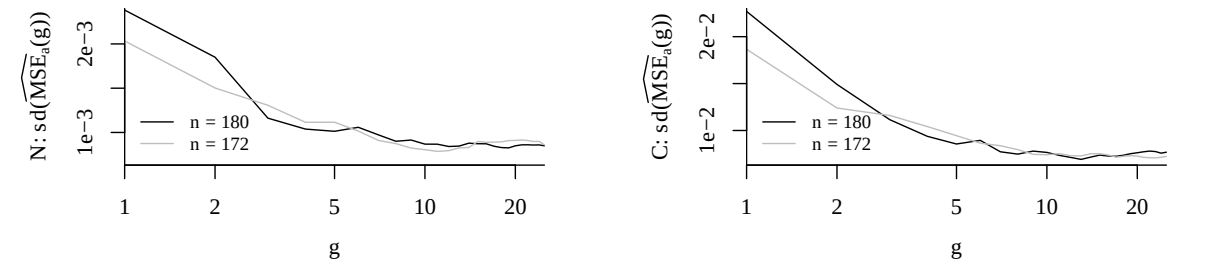


Abb. 28: Standardabweichung der Kreuzvalidierungen für {N,C}: $50\times$ 5-fache CV simultan auf a – Abszisse: $\log_2$-Skala.

Beim Auswerten der Streuung der MSE-Kurven auf pur optimierte Faktoren (-----) klingen die Resultate nicht – wie hier – mit zunehmendem g gegen eine untere Schranke ab, sondern es stellen sich chaotische Verläufe der Streuungen ein; vgl. Abb. 33.

Interpretation:
Die Krümmung – d.h. die Präzision – der MSE-Funktionen in der Umgebung des Minimums ist mäßig, so daß die Entscheidung für eine bestimmte Faktorenzahl g mit erhöhter Unsicherheit einkalkuliert werden muß.

Bei abnehmender Faktorenzahl, besonders unterhalb des MSE-Minimums, wird die Abweichung zwischen beiden MSE-Funktionen größer: Eine logistische Transformation senkt den MSE deutlich ab. Offenbar lassen sich mit einer symmetrischeren Verteilung ($a \gg 0$) der Einflußgrößen die Zielgrößen besonders für wenige Faktoren etwas besser prognostizieren.
Ohne Ausreißer-Bereinigung ($n = 180$) verbleibt in der Umgebung des Minimums immerhin ein geringer Nutzen der Transformation. Beim bereinigten Datensatz ($n = 172$) schwindet der Nutzen der Transformation ab vier und mehr extrahierten Faktoren – noch vor Erreichen des Minimums. Für $n \epsilon \{172\,;180\}$ gilt gleichermaßen: Mit zunehmender Faktorenzahl nähern sich die Ergebnisse beider MSE-Varianten immer weiter an.

Tab. 2 offenbart, nicht die Transformation des Spektrums, sondern das Entfernen von untypisch erscheinenden Einträgen verringert wesentlich die erforderliche Zahl g an Faktoren, nämlich von 11 auf 9.
Obwohl beide Zielgrößen miteinander hoch korreliert sind (vgl. Abb. 1, Seite 7), unterscheiden sich die Funktionen a der logistischen Entzerrung nach der Bereinigung beträchtlich hinsichtlich ihres Verlaufs. Während in Abb. 27 bei höherer Faktorenzahl für N der Wert in a gegen eine untere Schranke > 0 strebt, wächst a bei C nach Erreichen eines Plateus in der Umgebung des MSE-Minimums schließlich mit der Faktorenzahl an.
In Abb. 28 erreichen die Funktionen der Standardabweichung ab etwa 8 Faktoren eine untere Schranke. Bezüglich des Erwartungswertes – in Realisation: arithmetisches Mittel – ist der unsicherste Bereich ab einer Faktorenzahl $g > 8$ überwunden. In ihrem Verlauf sind alle vier Streuungskurven ähnlich geformt.
Wird auch berücksichtigt, daß die Kreuzvalidierung auf einer nichtstetigen Skala der Faktorenzahl zur Anwendung kommt und generell zur Unterschätzung neigt, sollte eine 9-Faktorenlösung als Minimallösung kritisch betrachtet und nicht unterschritten werden.

Ergebnisse der Regression
Die Funktion der Regressionskoeffizienten zeigt in den rechten Bildhälften, für größere Wellenlängen bei beiden Zielgrößen, höhere Auslenkungen:

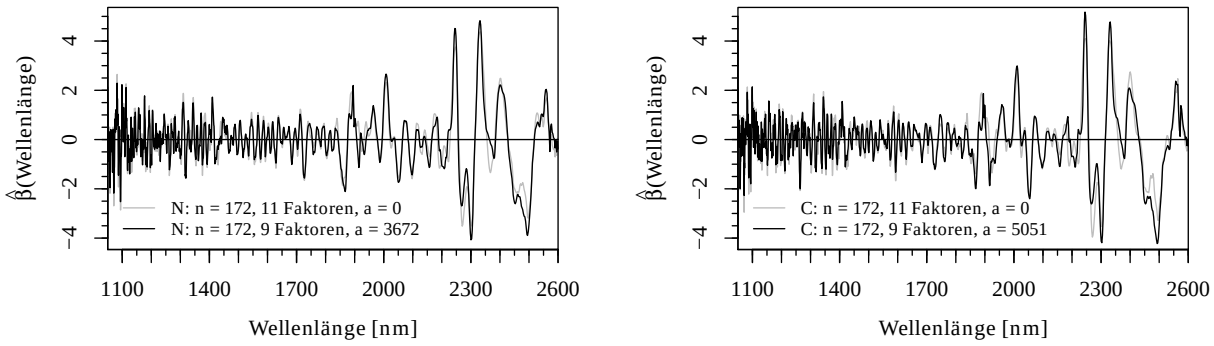


Abb. 29: Regressionskoeffizienten standardisiert – für bessere Vergleichbarkeit.

Auf Seite 8 sind in Abb. 3 Nahinfrarotbanden organischer Verbindungen auf ihren spezifischen Wellenlängen illustriert. Hohen Erklärungsgehalt haben die Koeffizienten in Abb. 29 demnach im Grundschwingungs- und dem anschließenden halben ersten Obertonbereich. Darüberhinaus – in den kleinen Wellenlängen – wirkt das Muster regelmäßiger,

gleichförmiger und gedämpfter.
Beim Vergleich der Koeffizientenfunktionen zeigen sich zwischen verschiedenen Datenumfängen, bei unterschiedlicher Faktorenzahl, aber auch über den Vergleich beider Zielgrößen, hohe Ähnlichkeiten im Muster.

Was kann das bedeuten?
Auch wenn das Koeffizientenmuster auf Wellenlängen bis ca. 1800 nm dem Anschein nach wie ein weißes Rauschen ausschaut, muß es das nicht sein. D.h., die Koeffizienten tragen auch für diese Wellenlängen noch Informationen. Hinweis: Vor der PLS-Berechnung wurden bereits wenig informative Wellenlängen abgeschnitten.
Wenn eine Datentransformation und/oder die Entfernung von Einträgen die Koeffizientenfunktion kaum ändert, wäre ein Hinweis auf Nichtwirksamkeit einer derartigen Vorbehandlung angezeigt.
Da sich die Muster in Abb. 29 über die Zielgrößen hinweg ähneln, kann es überlegenswert sein, für beide Zielgrößen auf dem gleichen Design zu arbeiten. Dann wäre die multivariate Erweiterung (PLS2) des Problems denkbar.

Zuletzt die Auswirkungen der Faktorenreduktion auf die Zielgrößen:

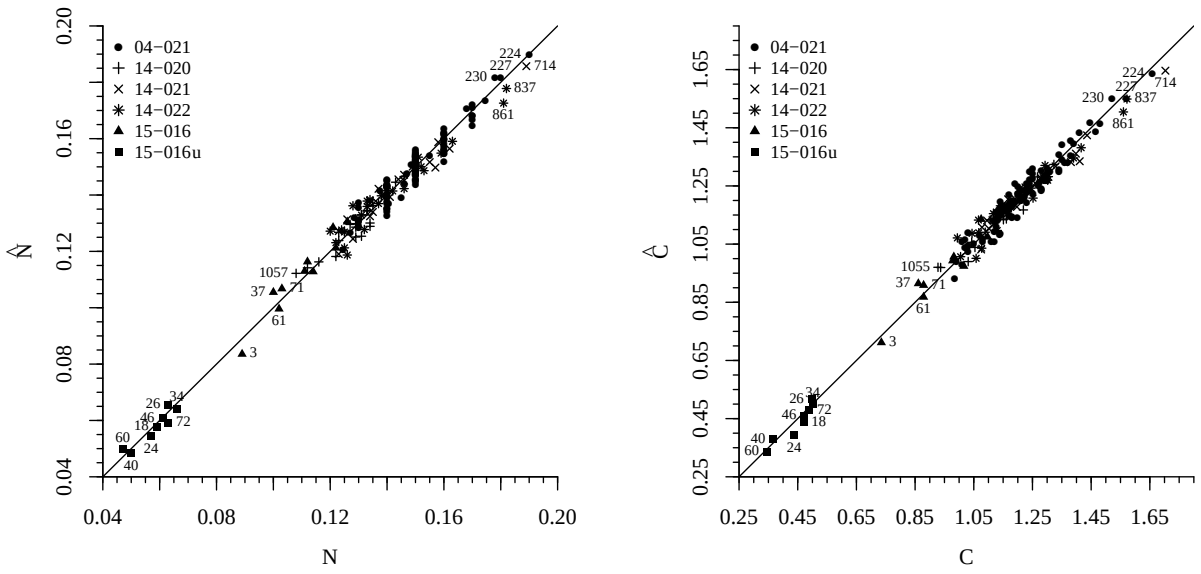


Abb. 30: Prognose gegen Zielgröße – $g = 11$, $a = 0$, inkl. Winkelhalbierende.

Bereinigt um die acht Objekte, entsprechend den Voreinstellungen gemäß Tab. 2, würde sich defacto die gleiche Situation (ohne Abbildung) noch einmal zeigen: fehlende Koordinaten der verdächtigen Objekte – vor allem am unteren linken Bildrand, aber an der Gesamtsituation ändert sich im Wesentlichen nichts. Das korrespondiert zu den Aussagen bzgl. Abb. 29.
Mit 11 Faktoren resultiert eine Genauigkeit in der Schätzung der Zielgrößen, so daß die Prognosen zu den Ausgangs-Zielgrößen bereits hohe Ähnlichkeiten aufweisen.

6.6 PLS2 auf standardisierte Zielgrößen

Auf Transformationen wird aufgrund der Erfahrungen aus der PLS1 verzichtet. D.h., die Zahl der Einträge beträgt nun konstant $n = 180$ und der Parameter der logistischen Entzerrung verbleibt in seiner Voreinstellung $a = 0$, womit (57) nicht weiter zur Anwendung kommt.
Für Kovarianz basierte Methoden ist im Multivariaten die Skala von Variablen in aller Regel kritisch. Bei den Spektren entstand bisher kein Problem, da alle Spalten semantisch miteinander im Einklang sind – das Beschreiben von Wellenlängen in der Größenordnung Nanometer. Anders sieht es bei den Zielgrößen aus. In Abb. 30 zeigen beide Plots voneinander verschiedene Skalen. Das ist für univariate Analysen vollkommen unproblematisch.

Wenn alle Zielgrößen gleichzeitig analysiert werden, wird diejenige mit der größeren Varianz, also oftmals auch die größere Skala, die Lösungen dominieren. In den Zielgrößen beträgt das Varianzverhältnis $\mathbb{V}(\mathrm{N}) : \mathbb{V}(\mathrm{C}) \approx 1 : 77$. Unbeachtet dieser Varianzschieflage verliert Stickstoff fast seinen gesamten Einfluß auf das Modell und bei der Auswertung einer PLS2 würde nahezu eine PLS1 auf Kohlenstoff beurteilt werden, während der Prognosefehler für Stickstoff überproportional ansteigt.
Das Standardisieren der Zielgrößen beseitigt die Ungleichheit in den Varianzen und somit gehen beide Zielgrößen gleichberechtigt in die Analyse ein. Auch für eine gute Vergleichbarkeit zu den Lösungen der PLS1 werden die auf standardisierte Zielgrößen geschätzten Prognosen mit den ursprünglichen Standardabweichungen multipliziert, womit die Skala der Ausgangsdatenlage zurückgewonnen wird.

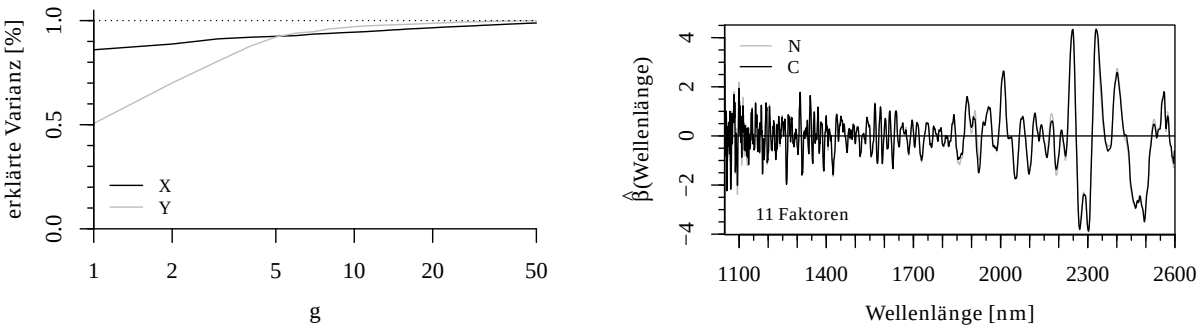


Abb. 31: links: Zunahme an erklärter Varianz für jeden weiteren Faktor – Abszisse: $\log_2$-Skala
rechts: Regressionskoeffizienten standardisiert – für bessere Vergleichbarkeit.

Im Vergleich zur Abb. 24 fallen keine besonderen Veränderungen auf. Die Grafiken sind sich sogar so ähnlich, daß sie zum eindeutigen Herausfinden von Unterschieden überlagert sein sollten. In der Umgebung um 2250 nm zeigt sich das Muster der Koeffizienten der PLS2 dennoch anders als bei beiden separaten PLS1-Lösungen in Abb. 29 – nur die schwarzen Kurven ($a = 0$) beachten.

Defacto verhalten sich die Faktoren wie bei der PLS1 in Abb. 25:

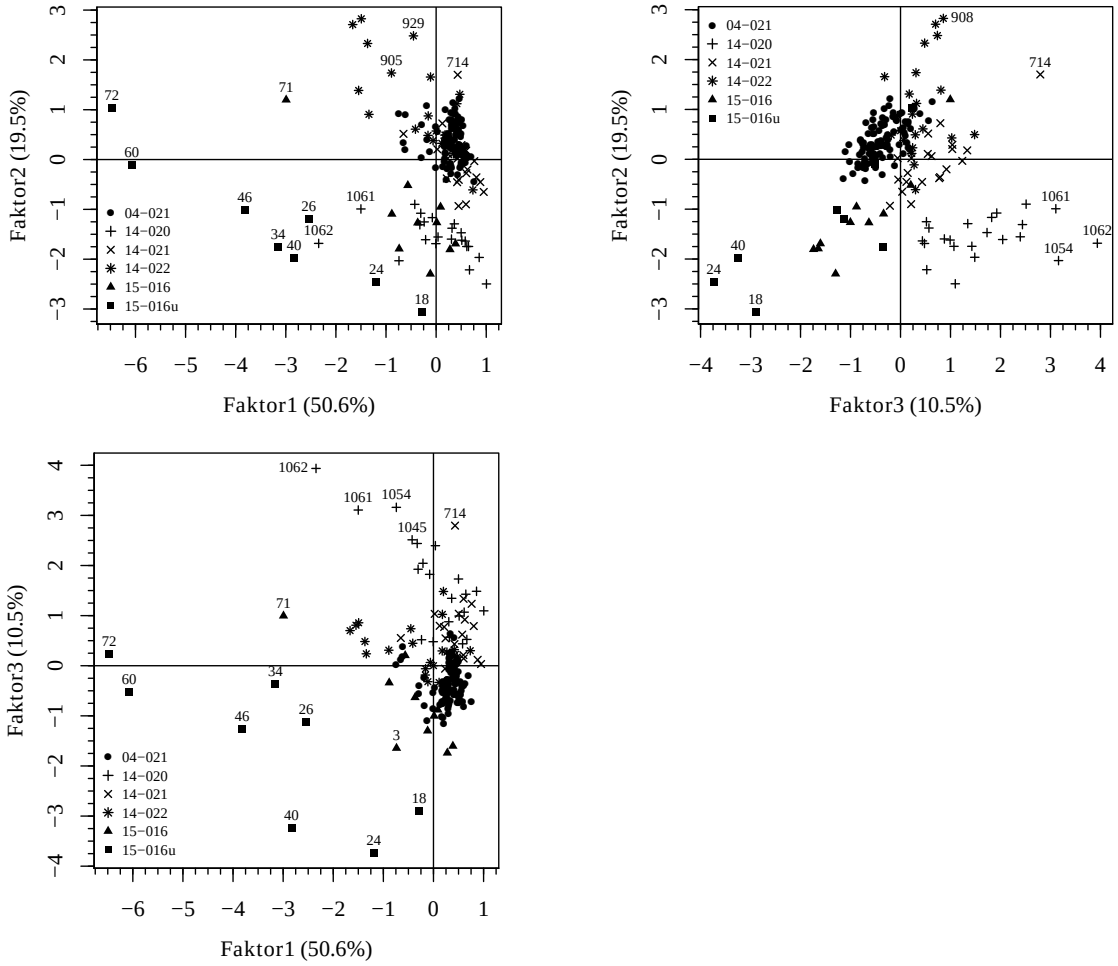


Abb. 32: Projektionen der ersten drei PLS-Faktoren.

In der Projektion Faktor1 – Faktor3 erkennt man nochmals die vollständige lineare Separierung der Unterbodengruppe.

Bzgl. der optimalen Faktorenzahlen liegen die Minima der Kreuzvalidierungen jeweils bei 11 Faktoren. Die rechte Ordinate beschreibt die Streuung des MSE:

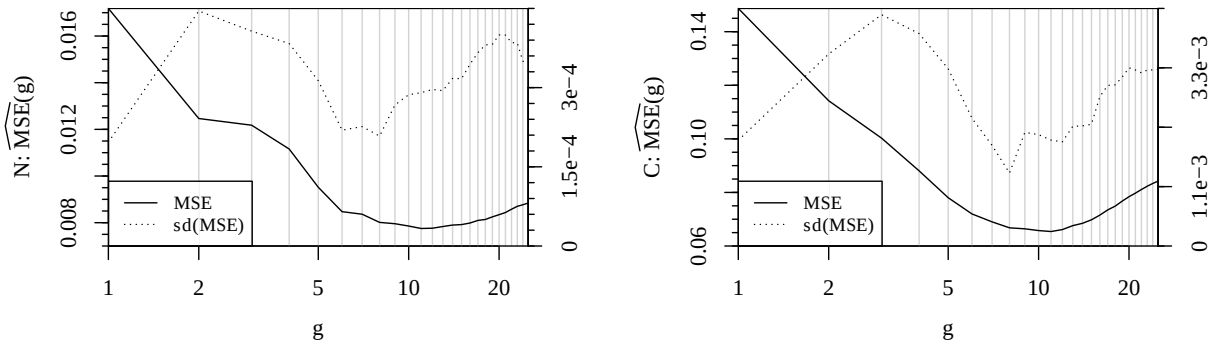


Abb. 33: arithmetischer Mittelwert der Kreuzvalidierungen für {N,C}: 20 × 5-fache CV inkl. der Standardabweichung – Abszisse: $\log_2$-Skala.

Wird statt 20 mit 50 Wiederholungen gerechnet, bleibt der Verlauf der Streuung im Wesentlichen derselbe.

Auf Seite 47 wird die innere Beziehung dargestellt, die den Zusammenhang der latenten Datenräume von den inneren Zielgrößen (U) auf die Einflußgrößen (T), den Hauptachsen, beschreibt. U^TU ist vollbesetzt, als Folge des Fehlers beim Regressieren. Üblicherweise beschränkt man sich beim Interpretieren der PLS nur auf die transformierten Hauptachsen, die Faktoren; weniger auf den inneren Zusammenhang.
Die diagonale Koeffizientenmatrix C übernimmt multivariat in (53.2) die Verbindung beider Datenräume: $U = TC + U_{Err}$, wobei in C auf der Diagonale die Anstiege notiert sind. Aufgrund zentrierter Größen ist das Absolutglied stets Null.
Bei diesem Regressionsproblem ist man in der angenehmen Situation $n > g$ – entspricht der Notation $n > p$ in Kapitel 3, d.h. aufgelöst im Kontext der klassischen Regression können äquivalent mit (2) in einem Schritt g Einfachregressionen berechnet werden.
Damit kommen Hoffnungen bzgl. des zweiten Moments für die Inferenz auf. Immerhin kann um den Erwartungswert (Punktprognose) der Regression ein symmetrisches Prognoseintervall berechnet werden, welches den wahren Wert mit einer Wahrscheinlichkeit von $1 - \alpha$ überdeckt.

Die grafische Darstellung der Regressionen für die ersten vier Dimensionen:

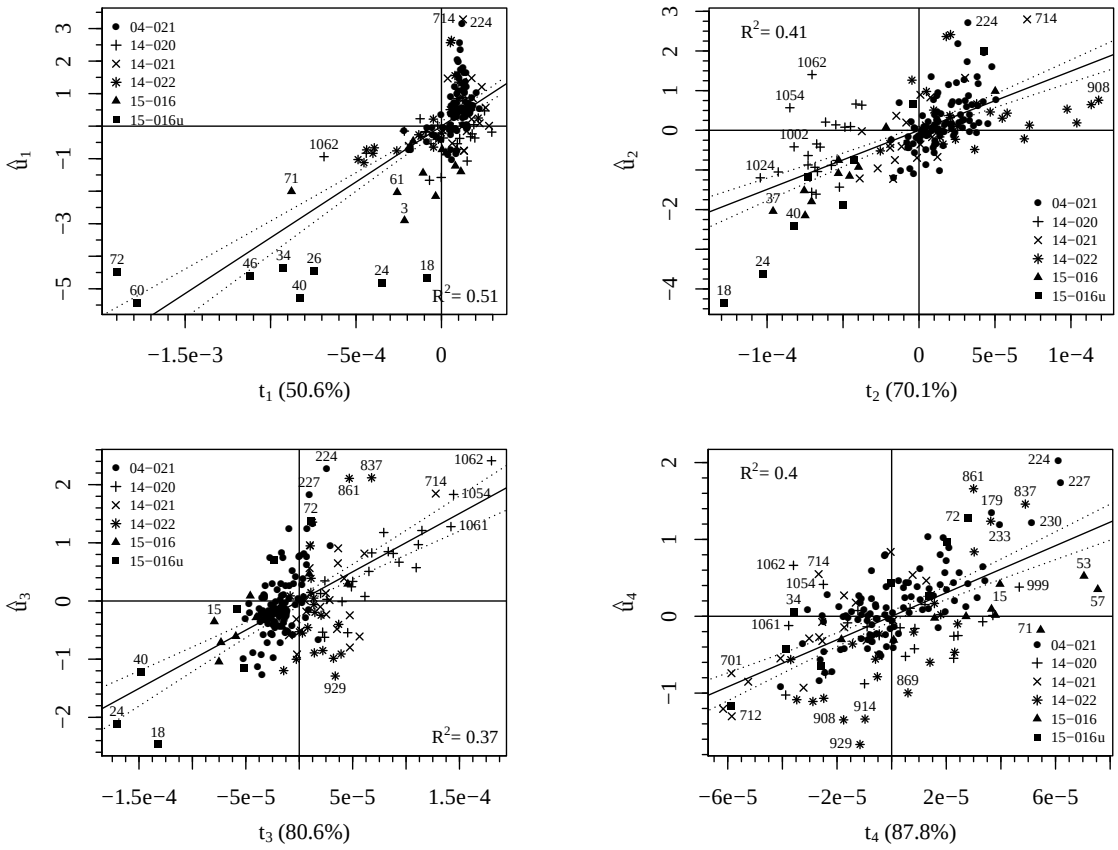


Abb. 34: Regression: Zielgrößen U der inneren Beziehung auf Hauptachsen T, inkl. kumulierter erklärter Varianz R^2 und 95%-Prognoseintervall (……)

Besonders auf die erste Dimension zeigt sich der Zusammenhang hochgradig nichtlinear. Ca. 88% der Zielgrößen-Varianz erklären die ersten vier Faktoren. Mit steigender Dimensionszahl gruppieren sich die Objekte um die jeweilige Regressionsgerade zunehmend linear, um spätestens in der letzten Dimension $g = n - 1$ fehlerfrei entlang der Gerade zu liegen.

Wollte man durch Entfernen der kompletten Unterbodengruppe den Zusammenhang für die erste Dimension linearisieren, so müßte das Anliegen enttäuschen. Es würde sich nichts ändern: Die am Rand verbleibenden Objekte würden die Funktion der entfernten Objekte übernehmen. Das Problem wurde auf Seite 57 schon einmal aus Sicht der Hauptkomponentenmethode diskutiert.
Bei der Beurteilung der Signifikanz von Regressionen der inneren Beziehung stellen sich sämtliche Regressionen bereits von Anfang an als (hoch) signifikant heraus. Über diesen Ansatz erscheint es nicht plausibel, eine möglichst adaptive Faktorenzahl-Einschränkung zu erhalten.
Dennoch, aufschlußreich ist die Darstellung der inneren Beziehung. Eine Regression auf die g-te Hauptachse – inkl. Streuplot – gibt einen Hinweis auf die insgesamt erreichte Linearität des Problems.
Ab vier extrahierten Faktoren wirken die Objekte in der Umgebung der Regressionsgeraden linear angeordnet und weitgehend Varianz homogen. Ebenso wie die Linearität wird die Varianzhomogenität für steigende Faktorenzahlen immer idealtypischer.

Zum Schluß die Auswirkungen der Faktorenreduktion auf die Zielgrößen:

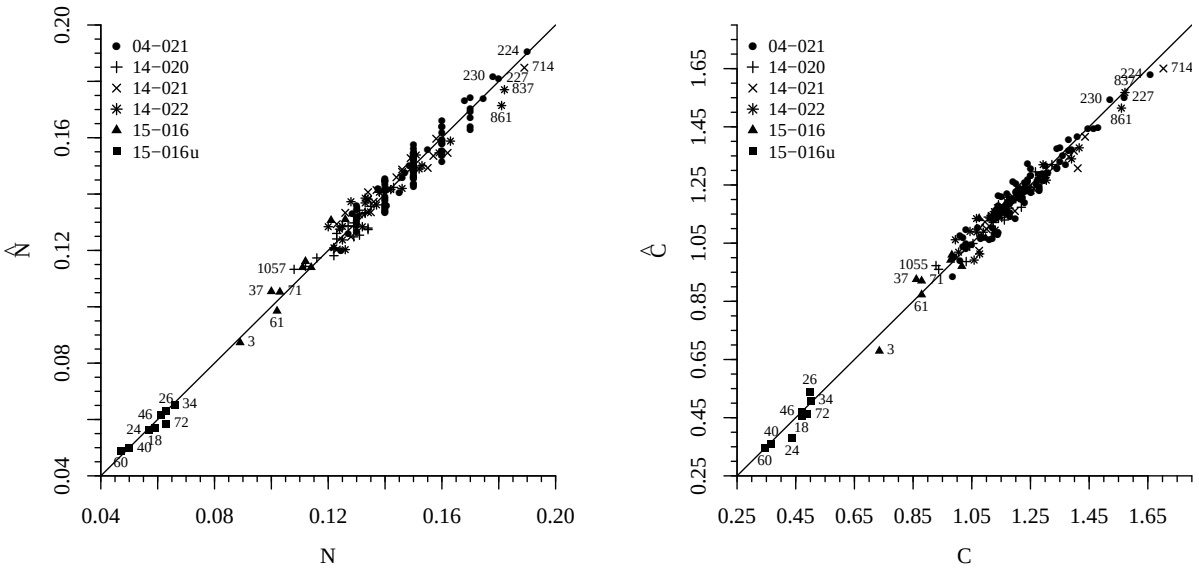


Abb. 35: Prognose gegen Zielgröße – $g = 11$, inkl. Winkelhalbierende; erklärte Varianz: 97,4%.

Die grafische Lösung unterscheidet sich kaum von Abb. 30. Bzgl. der Zielgrößen-Prognose wird also mit beiden Methoden das gleiche Ergebnis erreicht.

In den Zielgrößen-Streuplots ist die Separierung der Unterbodengruppe sehr anschaulich. Hingegen bei den Streuplots der inneren Beziehung, die als Verbindung der multivariaten Einfluß- zu den multivariaten Zielgrößen aufgefaßt werden kann, wird die Separierung mit steigender Faktorenzahl zunehmend beseitigt. Es spricht gegen die Annahme von Ausreißern, bei genügend extrahierten Faktoren.

7 Fazit & Ausblick

Bei den zur Verfügung gestellten Datensätzen sind nicht nur die Zielgrößen, sondern die Spektren – also die Einflußgrößen – ebenfalls mit einem zufälligen Fehler behaftet. Nicht deterministische Einflußgrößen[49] sind bei Regressionstechniken i.A. unerwünscht, denn sie verzerren die festen Koeffizienten in eine unbekannte Richtung. Deshalb war zuerst dafür zu sorgen, den Fehler aus den Einflußgrößen mit einer Glättungstechnik herauszurechnen, zumindest zu verkleinern. Dabei passiert schematisiert folgendes:

$$\mathbb{E}\left(\vec{y}\right) = \mathbb{E}\left(X\vec{\beta} + \vec{\epsilon}\right) = \mathbb{E}\left((Z+G)\vec{\beta} + \vec{\epsilon}\right) = \mathbb{E}\left(Z\vec{\beta}\right) + \mathbb{E}\left(G\vec{\beta}\right) + \mathbb{E}\left(\vec{\epsilon}\right) = Z\vec{\beta}.$$

D.h., in Matrix G sind die Fehler vorkommend, die aufgrund der Glättung separiert werden. Der Erwartungswert von Fehlermatrix G entspricht einer Nullmatrix und deswegen folgt aus dem Ausdruck $\mathbb{E}\left(G\vec{\beta}\right)$ ein Nullvektor. Als Resultat offenbart sich eine Transformation der Einflußgrößen von X nach Z.

Da die Einflußgrößen mit $n < p$ eine ungünstige Rechteckstruktur aufweisen, ist die Methodenwahl stark eingeschränkt. Die funktionale Regression ist vom Prinzip her eine Alternative zur PLS. Bei ausschließlich linearer Modellierung ist sie mit der hier vorgestellten Wirkweise der PLS vergleichbar. Dennoch wäre ein Glätten des Spektrums ebenso erforderlich. Attraktiv bei der PLS hingegen ist die gleichzeitig Faktoren-bezogene Sichtweise und damit einhergehend die Faktorenreduktion.

Durch die Glättung der Spektren entstand ein Zielkonflikt, welcher sich einerseits im Dämpfen von Rauschen auftat; andererseits sollte die Signalqualität aufrechterhalten werden. Besonders die Resonanzstellen mußten möglichst unbeeinträchtigt bleiben, aber das Rauschen sollte später die Regression nur minimal beeinträchtigen.
Zum Hervorheben von Resonanzstellen wurde die Derivativspektroskopie angewandt, indem das Spektrum zweimal differenziert wurde und mit der zweiten Ableitung final weiter gearbeitet wurde. Als günstig erwies sich zunächst die Glättung mit einem Polynom 10. Grades, bei einer Knotenbreite von 32, auf das vorliegende Spektrum vorzunehmen; dann die Ableitungen zu bilden, um danach nochmals mit einem Polynom 4. Grades, auf einer Breite von 15 Knoten, nachzuglätten. Der verwendete Algorithmus von Savitzky & Golay wurde angepaßt, um bei der Wahl des Polynomgrades – befreit von formellen Restriktionen – entscheiden zu können. Die Glättungen verringerten das Rauschen, konnten aber auch scharfkantige Spitzen modellieren.
Zum Verkleinern der Varianz wurde abschließend eine Mittelwertbereinigung (Multiplicative scatter correction) der Ableitungen durchgeführt: Jede der 180 Reihen ist genau um Null zentriert. Damit wurde sichergestellt, daß Faktoren keine Scheinvarianz erklären.

Mithilfe der Hauptkomponentenmethode wurden zweidimensionale Faktorenprojektionen nach ungewöhnlichen Mustern betrachtet. Rein visuell zeigte sich die Unterbodenklasse in den ersten Faktoren a-typisch weitgehend von den anderen Objekten separiert. Das kann trotzdem nur als vager Ausreißer-Hinweis aufgefaßt werden, da diese Analyse unter

[49] http://www.empiwifo.uni-freiburg.de/ (21.8.2018)

Ausschluß der Zielgrößen durchgeführt wird. Um die Wirkung von Ausreißern zu mindern, wurde eine logistische Transformation auf die Spektren versucht, bei der ein zusätzlicher Paramter zum Einstellen der Verzerrung erforderlich ist. Die Faktoren, der auf diese Art vorbehandelten Spektren, zeigten sich in der Projektion symmetrischer.
Auch wurde darüber nachgedacht, welche parametrische Verteilung den Ausreißern zugrunde liegen könnte. Mittels einer Dichte-Transformation wurde dann eine für die Ausreißer passendere Verteilung berechnet.

Ausreißer zu entfernen sollte die Prognose der Zielgrößen verbessern. Ausreißer, die sich allerdings konform – ohne Hebelwirkung – verhalten, sind für den Erwartungswert des Modells unschädlich. Lediglich „leere“ Bereiche entstehen bei der Abbildung der Prognosen. Eine PLS auf dem Design $n \leq p$ kann mit der Variation der Faktorenzahl die Zielgrößen beliebig genau – bei einer Faktorenzahl, die mindestens dem Rang der Designmatrix entspricht, immer perfekt – prognostizieren. Wurden verdächtige Objekte entfernt, reichten demzufolge weniger extrahierte Faktoren aus. In den Prognosen änderte sich faktisch nichts. Da der Umfang an Einträgen ($n = 180$) nur gering ist, wurden alle Einträge als kostbar angesehen und bewahrt.
Ausreißer nur separat in den Spektren bzw. Zielgrößen zu suchen, würde der vorliegenden komplexen Datensituation überhaupt nicht gerecht werden.

Für das Erkennen der optimalen Faktorenzahl wurde die 5-fache Kreuzvalidierung angewandt. Da dieses Verfahren heuristisch optimiert, war es geboten, mehrere unabhängige Wiederholungen mit verschiedenen Durchmischungen durchzuführen. Bei 20 Wiederholungen stabilisierte sich der Mittelwert der Lösungen.
Die Kreuzvalidierung mit simultaner Optimierung des Verzerrungs-Parameters der logistischen Transformation ist extrem zeitaufwendig, im Vergleich zur Kreuzvalidierung allein auf die Faktoren. Zudem mußte die Zahl der Wiederholungen auf 50 erhöht werden. Der sich einstellende Nutzen ist im Vergleich zum Zeitverbrauch zu klein und rechtfertigt nicht so einen hohen Aufwand, jedenfalls nach heutigem Stand der Computertechnik. Deswegen wurde später von der logistischen Transformation als Vorbehandlung abgesehen.

Durch die Vielzahl an Spalten in den Einflußgrößen, ist es nach einer PLS-Analyse schwer möglich, den Zusammenhang zwischen Einfluß- und Zielgrößen grafisch wiederzugeben, bzw. es kann nur ungenügend gelingen. Bei wenigen Zielgrößen können diese mit ihren Prognosen in Bezug gesetzt werden, wie es mit beiden Zielgrößen gezeigt wurde. Die Ergebnisse von PLS1 und PLS2 sind sich in hohem Maße ähnlich; doch muß bei einer PLS2 unbedingt beachtet werden, mit standardisierten Zielgrößen zu rechnen.
Liegen ohnehin viele Zielgrößen vor, ist eine PLS2 zu bevorzugen, da u.a. die Darstellung eines Zusammenhangs über die innere Beziehung mit einem bivariatem Streuplot, entsprechend der extrahierten Faktorenzahl, möglich ist. Man könnte dann auf die Abbildung der puren Zielgrößen ebenso verzichten. Ein weiterer Vorteil läge in nur einer PLS2-Berechnung, statt vieler einzelner PLS1-Analysen.

Literatur

Biener, Steinkämper, Masuch, Wolf & Hofmann. (2017, 5). Acetatmessung mit MIR-Transmissionsspektroskopie bei der E. coli-Kultur. In *BIO spektrum* (Bd. 3, S. 273–275). Springer.

Brockhaus. (2015, 11). Kurze Einführung in Regularisierung und penalisierte Schätzansätze in der Statistik. In *Schätzen und Testen I.* Institut für Statistik, Ludwig–Maximilians–Universität München.

Burns & Ciurczak. (2008). Handbook of Near-Infrared Analysis. In (Bd. 3, Kap. 9). CRC Press.

Chen. (2003). *Dreidimensionale Vermessung kreisförmiger Objekte mittels Luminanz und Tiefendaten* (Dissertation, Fakultät für Elektrotechnik, Informatik und Mathematik der Universität Paderborn). (Kap. 2.3). (Online erhältlich unter http://d-nb.info/969398433/34; 22.5.2017.)

Collins. (2010). Partial Least Squares Regression. In (S. 9 – 19, 36, 37). (Online erhältlich unter http://vision.cse.psu.edu/seminars/talks/PLSpresentation.pdf; 23.4.2017.)

Drechsler. (2018). Grundlegende Methoden der Sozialstatistik A. In *Sommersemester 2018. Part II, Causal Inference for Observational Studies* (S. 86).

Eckey & Rengers. (2002). Multivariate Statistik. In (1. Aufl., Kap. 2). Gabler.

Fahrmeir & Hamerle. (1984). Multivariate statistische Verfahren. In (1. Aufl., Kap. 11). de Gruyter.

Fahrmeir, Heumann, Künstler, Pigeot & Tutz. (2016). Statistik – Der Weg zur Datenanalyse. In (8. Aufl., Kap. 3.6). Springer.

Faires & Burden. (1994). Numerische Methoden – Näherungsverfahren und ihre praktische Anwendung. In (Kap. 1, 2). Spektrum.

Hartung & Elpelt. (1999). Multivariate Statistik. In (6. Aufl., Kap. II , VIII). Oldenbourg.

Heumann & Schmid. (2016). Schätzen und Testen II. In *Sommersemester 2016.* Institut für Statistik, Ludwig–Maximilians–Universität München.

Jørgensen & Goegebeur. (2007a). Module 7: Partial least squares regression I. In *ST02: Multivariate Data Analysis and Chemometrics.* Department of Statistics, Syddansk Universitet. (Online erhältlich unter http://statmaster.sdu.dk/courses/ST02/module07/index.html; 23.4.2017.)

Jørgensen & Goegebeur. (2007b). Module 8: Partial least squares regression II. In *ST02: Multivariate Data Analysis and Chemometrics.* Department of Statistics, Syddansk Universitet. (Online erhältlich unter http://statmaster.sdu.dk/courses/ST02/module08/index.html; 23.4.2017.)

Kauermann & Küchenhoff. (2010). Stichproben. In (Kap. 2.5). Springer.

Krämer, Boulesteix & Tutz. (2006). Penalized Partial Least Squares Based on B-Splines Transformations. In (Kap. 2). Institut für Statistik, Ludwig–Maximilians–Universität München, Sonderforschungsbereich 386. (Online erhältlich unter https://epub.ub.uni-muenchen.de/1853/1/paper_485.pdf; 14.5.2017.)

Long. (1983). *Covariance Structure Models.* Sage.

Magnus & Neudecker. (1979, 3). The Commutation Matrix: Some Properties and

Applications. *The Annals of Statistics*, *7* (2), 381–394. (Online erhältlich unter http://www.jstor.org/stable/2958818; 18.11.2017.)

Marinell. (1990). Multivariate Verfahren. In (3. Aufl., Kap. C). Oldenbourg.

Merziger & Wirth. (2006). *Repititorium der höheren Mathematik* (5. Aufl.). Binomi.

Mittnik. (2015). Ökonometrie – Grundlagen. In *SS 2015.* Seminar für Finanzökonometrie, Institut für Statistik, Universität München.

Phatak, Reilly & Penlidis. (2002). The asymptotic variance of the univariate PLS estimator. *Linear Algebra and its Applications*, *354*, 245–253. (Online erhältlich unter https://doi.org/10.1016/S0024-3795(01)00357-3; 21.11.2017.)

Rao, Toutenburg, Shalabh & Heumann. (2007). Linear Models and Generalizations. In (Bd. 3, Kap. 3). Springer.

Rinnan, Van Der Berg & Engelsen. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry*, *28* (10), 1201–1222.

Saraev. (2013). Interval Pseudo-Inverse Matrices and Interval Greville Algorithm. In (Bd. 18, S. 147–156). Lipetsk State Technical University, Lipetsk, Russia. (Online erhältlich unter https://pdfs.semanticscholar.org/b130/67ab2bb8c411170ded10099a7ade27d37732.pdf; 3.12.2017.)

Savitzky & Golay. (1964, 7). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. In *Analytical Chemistry* (Bd. 36, S. 1627–1639). (Online erhältlich unter https://wenku.baidu.com/view/7738e76aa26925c52cc5bf9b.html; 8.3.2018.)

Schmid. (2016a). Bioimaging. In *Medizinische und Biologische Bildgebungsverfahren, Sommersemester 2016.* Institut für Statistik, Ludwig–Maximilians–Universität München. (Online erhältlich unter https://moodle.lmu.de/course/view.php?id=960; 6.6.2017.)

Schmid. (2016b). Wahrscheinlichkeitstheorie und Inferenz I. In *Wintersemester 2016/17* (Kap. 10). Institut für Statistik, Ludwig–Maximilians–Universität München. (Online erhältlich unter https://moodle.lmu.de/course/view.php?id=1126; 17.6.2017.)

Schmid. (2017). Räumliche Statistik. In *Sommersemester 2017* (Kap. 3.3). Institut für Statistik, Ludwig–Maximilians–Universität München. (Online erhältlich unter https://moodle.lmu.de/course/view.php?id=1374; 26.6.2017.)

Schmidt & Trenkler. (2015). Einführung in die Moderne Matrix-Algebra. In (3. Aufl., Kap. 5,6). Springer Gabler.

Stoer. (1979). Einführung in die Numerische Mathematik I. In (3. Aufl., Kap. 4.3). Springer.

Stoer & Bulirsch. (1978). Einführung in die Numerische Mathematik II. In (2. Aufl., Kap. 6.3). Springer.

Tillmann. (1996). Kalibrationsentwicklung für NIRS-Geräte. Cuvillier.

Danksagungen

Herrn Prof. Dr. Christian Heumann möchte ich für den enormen Handlungs- und Gestaltungspielraum danken, den er mir beim Erstellen der Masterarbeit gewährte und welchen ich auch sehr gern in Anspruch nahm. Durch das Nichtspüren eines „Korsetts“, konnte ich mich ausprobieren und entwickeln, welches für mich eine bereichernde Erfahrung war. Die Zusammenarbeit empfand ich als sehr angenehm, geprägt von Fachlichkeit, Fairneß und Kollegialität.

Herrn Prof. Dr. Volker Schmid möchte ich danken, daß er nach der Vorlesung stets ein offenes Ohr für Fragen hatte und mir manchen Tip mit auf den Weg gab.

Ab und an habe ich mir bei Herrn Dr. Fabian Scheipl zu weniger themenrelevanten Fragestellungen eine Meinung eingeholt. Seine Sichtweisen halfen mir beim tieferen Verstehen der Datensituation. Für seine Hilfsbereitschaft möchte ich ihm danken.

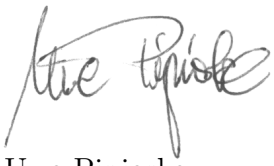
Frau Claudia Buchhart möchte ich dafür danken, daß sie mir beim Übergang von der Theorie zur Praxis zu den Daten und zur organischen Chemie einige Hinweise gab, die mir beim Verstehen halfen.

Zuletzt möchte ich einer guten Freundin, Frau Uta Hoffmann danken, welche während meiner Abwesenheit vom Heimatort unsere gemeinsamen Katzen Molly (vgl. Abb. 10) und Flitzer liebevoll versorgt hat.
Außerdem hat sie Ausdruck und Grammatik dieser Arbeit hinterfragt und manche verschlungene Satz-Konstellation entschärft.

Deklaration

Hiermit erkläre ich, daß ich die vorliegende Arbeit ohne fremde Hilfe verfaßt und nur die im Literaturverzeichnis aufgeführten Quellen verwendet habe.

f.d.R.



Uwe Pipiorke